University of Regensburg

Methods of Econometrics
Handout

Prof. Dr. Rolf Tschernig

Winter semester 2023/2024

Version of January 15, 2024

Note: This handout has emerged from an extensive set of slides. A substantial part of the slides was produced in cooperation with Harry Haupt, University of Passau. I would like to thank Kathrin Kagerer and Stefan Rameseder for their excellent contributions, important corrections and substantial suggestions for improvement. I also thank Christoph Rust for his important comments and Florian Greindl for corrections. I would also like to thank Andreas Kelbel very much for transferring the handout to the beamer package. I would also like to thank Dominik Ammon for further corrections. I ask that any errors be sent to rolf.tschernig@ur.de.

Please cite as: Rolf Tschernig, Methods of Econometrics - Handout, University of Regensburg, October 2023. Downloaded on [day month year].

Organisation

Note

Apart from corrections of organisational notes, links and typing errors and a few improvements to the content, this handout corresponds to the handout of the winter semester 2018/2019, version of 29 November 2018, which in turn corresponds to the handout of the winter semester 2015/2016, version of 20 October 2015, apart from corrections and improvements to the content, in particular in sections 5.2 and 5.3.

Contact

Prof. Dr. Rolf Tschernig

Building RW(L), 5th floor, room 514
Universitätsstr. 31, 93040 Regensburg
Phone (+49) 941/943 2737, Fax (+49) 941/943 4917
Email: rolf.tschernig@ur.de

Dominik Ammon, M.Sc. in Economics

Building RW(L), 5th floor, room 516
Universitätsstr. 31, 93040 Regensburg
Phone (+49) 941/943 2738, Fax (+49) 941/943 4917
Email: dominik.ammon@ur.de

https://www.uni-regensburg.de/business-economics-and-management-information-systems/economics-tschernig/homepage/index.html

Schedule and Scope

Schedule for the current winter term

- Week 1 and 2: **Math Camp Part 3**
  (Part of the mandatory module **Methods of Econometrics**)
- Week 3 to 15: Mandatory module **Methods of Econometrics**
- for semester-accompanying performances see course homepage https://www.uni-regensburg.de/business-economics-and-management-information-systems/economics-tschernig/teaching/master/methods-of-econometrics/index.html

Format and ECTS

- **4 h lecture** and 2 h tutorial
- **10 ECTS**: corresponds to approx. 250 h to 300 h expenditure of time for the entire module

Organisation

Contents, Dates and Rooms, Downloads, News

https://www.uni-regensburg.de/business-economics-and-management-information-systems/economics-tschernig/teaching/master/methods-of-econometrics/index.html

Prerequisite for course participation

- Knowledge of the contents of the Math Camp Part 1 and 2.
- Helpful, but not required: knowledge of an introductory econometrics course, e. g. of the bachelor course *Introduction to Econometrics*.

Aims of this course

(Basic) knowledge to answer the following questions

- **How** do I do a careful empirical/econometric analysis?
- What are the econometric methods?
- How can I assess the quality of an empirical analysis?
- **Why** and under which assumptions does an econometric method work?

- How can I conduct empirical analyses with the free software `R`?

Benefit

During studying

- Basis for master programme, especially for the specialisation *Data Science and Econometrics.*
- Basis for understanding advanced econometric textbooks.
- Understanding empirical analyses in other courses.
- Be able to conduct empirical analyses in the master's thesis or a seminar paper (Cassar, Engl, Gürtzgen, Jerger, Kindermann, Knoppik, Lee, Roider, Tschernig, Weber).

At work

- Data analyses increasingly important (**Big Data**, **Open Data**)!
- Programming skills are helpful in many professional activities.

Grade Composition and Exam

Grade composition

- Study-accompanying performances (25%):
  Type of performances see course homepage
- Final exam (75%)

Final exam

- Date: Exam period
- Duration: 90 minutes
- Contains tasks on part 3 of the mathematical pre-course

Note — relevant if master studies started in Oct 2021 or later and Prüfungsordnung of 2021 applies)

To pass the module, the overall grade in the module must be 4.0 or better.

Note — relevant if older Prüfungsordnung of 2015 applies)

To pass the module, a grade of 4.0 in the final exam is **not** sufficient if you have an overall grade worse than 4.0 in the study-accompanying performances.

Software

In the course: Use of the R software.

- Advantages of R:
  - very flexible mathematical-statistical programming language.
  - free software: `http://www.r-project.org/`.
  - is used in science and business.
  - fast growing library of `R` packages for various tasks.
  - anyone can program packages oneselve and make them available to the general public.
  - wide distribution according to TIOBE Programming Index: `http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html`.
- Intensive use of R in the module. All R programs in Appendix A of the handout.
- Use of R:
  - in the master courses **Advanced Econometrics**, **Applied Financial Econometrics**, **Quantitative Economics II**, etc.
  - in master's theses at the chair of econometrics.
- R course:
  - Programming with R (offered by the chair of econometrics, runs parallel to this course during the winter term. Old versions available as screencast on GRIPS)

Alternative software for econometric analyses — Overview

Graphical user interface

- **EViews** (EViews courses (Christoph Knoppik), programmable, available in the CIP pool, also used in master course: Quantitative Economic Research II)
- **Gretl** (programmable, free software: http://gretl.sourceforge.net/)
- **Stata** (Stata course (see SPUR), available in the CIP pool)
- **JMulTi** (free software: http://www.jmulti.de/, master course: Quantitative Economic Research II)

(Statistical) programming languages with ready-to-use program modules

- **R**, see above.
- **Gauss** (commercial)
- **Ox** (commercial)
- **Matlab** (commercial)
- **Python** (Data Science and Machine Learning)
- **Fortran, C, C++** (General programming languages with extensive numeric libraries)

Computer Algebra Languages

- **Maple** (UR licence)
- **Maxima** (free software)
- **Mathematica**

Mandatory literature

Davidson, R. & MacKinnon, J.G. (2004). *Econometric Theory and Methods*, Oxford University Press (http://econ.queensu.ca/ETM/)

Literature for math camp for linear algebra

- Schmidt, K. & Trenkler, G. (2015). *Einführung in die Moderne Matrix-Algebra. Mit Anwendungen in der Statistik*, 3. Auflage, Springer. Compact, easy-to-read German textbook with many examples of arithmetic with matrices. (from the network of the university full text available)
- Gentle, J.E. (2007) *Matrix Algebra Theory, Computations, and Applications in Statistics*, Springer. Chapter 2 interesting for econometricians: detailed introduction to vector spaces (from the network of the university full text available)
- Fischer, G. (2014) *Lineare Algebra*, 18. Auflage, Vieweg & Teubner. Section 1.4 basic introduction for mathematicians, physicists, engineers, etc.(from the network of the university full text available)
- Lütkepohl, H. (1996) *Handbook of Matrices*, John Wiley & Sons, Chichester. Excellent reference book on linear algebra and its various matrices and associated calculation rules and transformation possibilities.

Literature for math camp on probability theory

- Casella, G. & Berger, R.L. (2002). *Statistical Inference*, Duxbury - Thomson. Very detailed, formal introduction to probability theory.
- Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I.& Tutz, G. (2016). *Statistik. Der Weg zur Datenanalyse*, 8. Auflage, Springer. Simple introduction to statistics (from the network of the university full text available)
- Steland, A. (2016). *Basiswissen Statistik: Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 4. Auflage, Spinger. Well-written, concise, technically precise introduction to statistics (from the network of the university full text available)

Literature for review and supplementary literature

- Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R Springer*, Springer. Very good introduction to `R` (from the network of the university full text available)
- Steland, A. (2013). *Basiswissen Statistik: Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 3. Auflage, Springer. (from the network of the university here)
- Stock, J.H. & Watson, M.W. (2012). *Introduction to Econometrics*, 3. ed., Person, Addison-Wesley. `https://scholar.harvard.edu/stock/pages/introduction-econometrics`
- Wooldridge, J.M. (2013). *Introductory Econometrics. A Modern Approach*, 5. Ed., Thomson South-Western.

Literature for further reading (in alphabetical order)

- Angrist, J. & Pischke, J. (2009). *Mostly Harmless Econometrics. An Empiricist's Companion*, Princeton University Press.
  (Well-readable introduction to the empirical evaluation literature)
  http://press.princeton.edu/titles/8769.html
- Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics*, Cambridge University Press.
  (Methodology for microeconometric problems)
  http://cameron.econ.ucdavis.edu/mmabook/mma.html
- Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics.* Oxford University Press.
  Many details on the methodology for non-linear regression models,
  http://qed.econ.queensu.ca/dm-book/
- Greene, W. (2012). *Econometric Analysis.* 7e, Prentice Hall.
  Comprehensive reference book with moderate methodological depth,
  http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm
- Hayashi, F. (2000). *Econometrics*, Princeton University Press.
  Formally very clearly structured.
  http://fhayashi.fc2web.com/hayashi_econometrics.htm
- Hansen, B. (2015). *Econometrics* http://www.ssc.wisc.edu/~bhansen/econometrics/
- Peracchi, F. (2001). *Econometrics*, John Wiley & Sons.
  The statistical approach to regression with methodological depth,
  http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471987646,descCd-tableOfContents.html
- Ruud, P.A. (2000). *An Introduction to Classical Econometric Theory.* Oxford University Press.
  The geometric approach with methodological depth
- Verbeek, M. (2012). *A Guide to Modern Econometrics*, 4th. ed., Wiley.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd. ed., MIT Press.
  A lot of intuition and methodical depth
  http://mitpress.mit.edu/books/econometric-analysis-cross-section-and-panel-data

More organisational issues

- **PC introduction to the computer language R** either as a screencast or as part of the first lecture after the pre-course in mathematics, see course homepage under "Current issues”

- Information events on **studies abroad** always take place at the beginning of the winter semester. Exact dates on the Homepage of the International Offices
- Nobel Lecture, see Homepage of the Department of Economics and Econometrics

Contents

Part I.

Math Camp

1. Linear Algebra

Why do you need linear algebra?

- For the analysis of the properties of the solutions of linear systems of equations
- Representation of multivariate data
- Representation of linear relationships
- To solve systems of linear equations
  Example: Normal equations to derive the least squares estimator
- Part of non-linear optimisation algorithms
- Representation of geometric relations by systems of numbers (analytic geometry)
  Matrices as geometric transformations
- All together: in econometrics and beyond as in machine learning

Overview

- Vectors
- Vector spaces
- Euclidean space and subspaces
- Matrices, calculation rules, special matrices and measures
- Matrices and linear mappings
- (Semi-)definite matrices
- Rules for the derivative of vector-valued functions
- Partitioned matrices

Literature references

- Schmidt & Trenkler (2015) Compact, easy-to-read German textbook with many examples for calculating with matrices.
- Gentle (2007, Chapter 2) (Full text access in the UR area): detailed introduction to vector spaces
- Fischer (2014, Section 1.4) Basic introduction for mathematicians, physicists, engineers, etc.
- Lütkepohl (1996) Excellent reference work on linear algebra and its various matrices and associated calculation rules and transformation possibilities. Often helpful when reading technical articles.

1.1. Vectors

Overview

- Space
- Euclidean space
- Vectors
- Dimension, length of a vector

Euclidean Space and Vectors

Space

In mathematics, a space is a **set** of mathematical objects **with** an additional **structure**. I.e. operations are possible with respect to the elements of the set. (http://de.wikipedia.org/wiki/Raum_(Mathematik))

Examples:

- Vector space
- Euclidean space (= vector space with scalar product)
- Probability space (="set with set system and probability mapping")

Euclidean Space, n-dimensional space

- The set underlying a Euclidean space is the set of ordered **n-tuples $\mathbf{x}$** of real numbers:

$$\mathbb{R}^n = \{\mathbf{x} = (x_1, \ldots, x_n) : x_1, \ldots, x_n \in \mathbb{R}\}.$$

  The ordered n-tuples $\mathbf{x} = (x_1, \ldots, x_n)$ are also called **n-vectors** or **vectors** for short. Each ordered n-vector represents a point in the n-dimensional Euclidean space $\mathbb{R}^n$, in short: $\mathbf{x} \in \mathbb{R}^n$.

- The associated structure includes
  - addition,
  - scalar multiplication and

- the scalar product.

as operations between the elements.

Examples:

- $n = 1$: $x \in \mathbb{R}^1$ Set corresponds to number line, elements are scalars
- $n = 2$: $\mathbf{x} \in \mathbb{R}^2$ Set corresponds to the plane, elements are two-dimensional vectors.
- $n = 3$: $\mathbf{x} \in \mathbb{R}^3$ Set corresponds to space with length, width and height.

Further terms

- n: **Dimension** of $\mathbf{x}$. n is also called **length** in linear algebra (in R too!).

  Attention: The length of a vector often also denotes the Euclidean norm of a vector. See Section 1.2.

- x_i: **Element** or **Component** of $\mathbf{x}$.

1.2. Vector spaces

Overview

- Vector space
- Addition and subtraction
- Zero vector
- Inverse vector
- Linear combination
- Line
- Scalar product or dot product (or inner product)

Definition

A set $\mathcal{V}$ with the operations

- **Addition** $\mathcal{V} \times \mathcal{V} \to \mathcal{V}$

- **Multiplication by a scalar (Scalar Multiplication)** $\mathbb{R} \times \mathcal{V} \to \mathcal{V}$

is called a **linear vector space**, if, in addition

1. a **zero vector** and an **inverse element** exist for the addition and associativity and commutativity apply and
2. **distributivity** and **associativity** apply to multiplication by a number, and **multiplication by one** results in the same element again, i. e. multiplication and addition are compatible.

(adapted from Fischer (2010, S. 76))

Remarks

- **Linear Combination**: Combination of the operations of addition and multiplication with scalars:

$$\alpha, \beta \text{ scalar}, \mathbf{x}, \mathbf{y} \in \mathcal{V}: \quad \alpha\mathbf{x} + \beta\mathbf{y} \in \mathcal{V}.$$

Every linear combination of the vectors is contained in $\mathcal{V}$.

Therefore, a vector space is a **linear space**.

Real-valued vectors and vector space

The set of real-valued n-vectors $\mathbf{x} \in \mathbb{R}^n$ forms a linear vector space.

Verification – Operations

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

- **Addition** of vectors of length n

$$\begin{gathered} \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n : \\ (x_1, \dots, x_n) + (y_1, \dots, y_n) := (x_1 + y_1, \dots, x_n + y_n) = \mathbf{z} \\ \mathbf{x} + \mathbf{y} = \mathbf{z} \end{gathered}$$

- **Multiplication by a number** $\lambda \in \mathbb{R}$

$$\mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n : \lambda \cdot \mathbf{x} = \lambda \cdot (x_1, \dots, x_n) := (\lambda \cdot x_1, \dots, \lambda \cdot x_n)$$

Verification – Conditions for addition

- **Zero vector**: There exists a zero vector $\mathbf{0} := (0, \dots, 0)$ such that it holds:

$$\mathbf{0} + \mathbf{x} = \mathbf{x}, \qquad 0 \cdot \mathbf{x} = \mathbf{0}$$

- **Inverse vector**: For every vector $\mathbf{x} \in \mathbb{R}^n$ there exists an inverse element $\mathbf{z} \in \mathbb{R}^n$ that maps it to the zero vector with the operation. The inverse vector is the **negative vector** $-\mathbf{x} = -(x_1, \dots, x_n) := (-x_1, \dots, -x_n)$. Check!

Verification – Conditions for scalar multiplication

- **Distributivity for scalar multiplication**: For $\alpha, \beta \in \mathbb{R}$, it holds that:

$$\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}, \quad (\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$$

- **Associativity of scalar multiplication**

$$(\alpha\beta) \cdot \mathbf{x} = \alpha \cdot (\beta \cdot \mathbf{x})$$

Other properties

- **Associativity** of addition:

$$(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$$

- **Subtraction** results from addition and multiplication by a number:

$$\mathbf{z} - \mathbf{y} = \mathbf{z} + (-\mathbf{y}) = \mathbf{x}$$

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ of length n are equal if and only if $x_1 = y_1, \dots, x_n = y_n$ holds.

Note: The **set of real numbers** $\mathbb{R}$ with the mentioned operations also forms a vector space (= special case for $n = 1$).

Straight line in $\mathbb{R}^n$

Definition

- Two distinct points $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^n$ determine a **straight line**
- In $\mathbb{R}^n$: Let $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^n$ be fixed. All points on the straight line defined by $\mathbf{v}$ and $\mathbf{v}'$ are given by

$$L = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{v} + \lambda\mathbf{w}, \lambda \in \mathbb{R}\}.$$

where $\mathbf{w} = \mathbf{v}' - \mathbf{v}$. The set L is the image of the mapping $\Phi : \mathbb{R} \to L \subset \mathbb{R}^n : \lambda \to \mathbf{v} + \lambda\mathbf{w}$ and is called **parameterisation of the straight line**.

Straight line in $\mathbb{R}^2$ (in the plane)

- Special case for $n = 2$.
- All points $\mathbf{x}$ of a straight line in $\mathbb{R}^2$ can be represented as an equation with two unknowns $\mathbf{x} = (x_1, x_2)$ and three fixed coefficients a_1, a_2, b

  $$a_1 x_1 + a_2 x_2 = b.$$

  The three coefficients a_1, a_2, b determine the position of the straight line and can be determined from two given points of the straight line $\mathbf{v}$, $\mathbf{v}'$ and vice versa.
- Determination of two points on the straight line for given coefficients: For given x_1, x_2 can be determined uniquely and vice versa, provided $a_1 \neq 0, a_2 \neq 0$. Example:

  $$\begin{aligned} x_1 = 0 :& x_2 = \frac{b}{a_2} \quad \text{a point of the straight line} \\ x_2 = 0 :& x_1 = \frac{b}{a_1} \quad \text{second point of the straight line} \end{aligned}$$
- Two straight lines intersect at exactly one point, unless they are equal or parallel $\Leftrightarrow$ a two-dimensional linear system of equations has one, infinitely many or no solution.
- **Intersection of two straight lines**: Solution $\mathbf{x}$ of the linear system of equations

  $$\begin{aligned} a_1 x_1 + a_2 x_2 &= b \\ c_1 x_1 + c_2 x_2 &= d \end{aligned}$$

  Solve for x_1, x_2 by substitution or use of **matrix algebra**. See section 1.4.

Vector space: Scalar product

Scalar product or dot product (or inner product)

The mapping $\mathcal{V} \times \mathcal{V} \to \mathbb{R}$ gives a scalar as a result.

The additional existence of the scalar product for a vector space enables

1. a unique characterization of the relationship between the elements,
2. the characterization of the individual elements by determining their length.

Note: The scalar product is a special type of an inner product. Inner products can also be defined for functions, for example. In general, an inner product $< \cdot, \cdot >$ always yields a real or complex quantity as a result (Gentle 2007, Sections 2.1.4, 3.2.6).

In general, a

- unique characterization of the relationship between the elements of a vector space is given if a **metric** exists for the vector space,
- unique characterization of the relationship of individual elements of a vector space is given if a **norm** exists for the vector space.

1.3. Euclidean space

Overview

- Vector space in $\mathbb{R}^n$
- Euclidean space
- Norm
- Normed vector space
- Euclidean norm
- Metric
- Metric space
- Orthogonal vectors
- Linear independence

Scalar product in the vector space $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} : \quad < \mathbf{x}, \mathbf{y} >:= \sum_{i=1}^{n} x_i y_i \tag{1.1}$$

Definition Euclidean space

The vector space of all real-valued n-vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, in which additionally the scalar product $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} :< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i$ is defined, is called Euclidean space.

The existence of the scalar product provides a descriptive geometric characterisation of the Euclidean space.

Norm and normed vector space

A **norm** allows, in general terms, the quantitative evaluation of individual elements of a set and, as can be shown, their relations to each other.

Norm for a vector space

The mapping $||\cdot|| : \mathcal{V} \to [0, \infty)$: assigns to each element $\mathbf{x}$ of the vector space a non-negative real number $||\mathbf{x}||$ and satisfies the following properties:

1. If $\mathbf{x} \neq 0$, then $||\mathbf{x}|| > 0$ and if $||\mathbf{x}|| = 0 \Leftrightarrow \mathbf{x} = 0$.
2. $||\alpha \mathbf{x}|| = |\alpha| \, ||\mathbf{x}||$.
3. $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (Triangle inequality).

(Vgl. Gentle 2007, Section 2.1.5)

Normed vector space

a vector space whose elements can be evaluated/measured with a norm.

Various vector norms

- L_2**-norm** or **Euclidean norm**:

$$||\mathbf{x}||_2 := \sqrt{\sum_{t=1}^{n} x_t^2}$$

  The Euclidean norm **measures** the **length of a** n**-dimensional vector**:

$$||\mathbf{x}||_2 := \left(\sum_{t=1}^{n} x_t^2\right)^{1/2}$$

The **modulus** of a real number $|x|, x \in \mathbb{R}$ is the Euclidean norm in $\mathbb{R}$.

- $\sharp$ L_∞**-norm** or **Chebyshev-norm**: $||\mathbf{x}||_\infty := \max_{t \in n} |x_t|$. E. g. relevant when loading vehicles, when no edge of an object to be transported may exceed a maximum length.
- $\sharp$ L_p**-norm**:

$$||\mathbf{x}||_P := \left(\sum_{t=1}^{n} |x_t|^p\right)^{1/p},$$

  contains both cases already mentioned as special cases.

Metric and metric space

Metric

A metric is a distance function $d : \mathcal{V} \times \mathcal{V} \to [0, \infty)$ satisfying the following conditions, where $\mathcal{V}$ denotes a vector space. For two objects x and y in $\mathcal{V}$ holds:

1. $d(y, x) > 0$, if $x \neq y$ and $d(y, x) = 0$, if $x = y$,
2. $d(x, y) = d(y, x)$,
3. $d(x, z) \leq d(x, y) + d(y, z)$.

(Gentle 2007, Section 2.1.7)

Metric space

A normed vector space is automatically a **metric space**, since the induced metric $d(x, y) := ||x - y||$ satisfies all requirements.

Scalar product, norm, metric

In general (not only for Euclidean space)

$$\begin{array}{ccccc} \text{Scalar product} & \Longrightarrow & \text{Norm} & \Longrightarrow & \text{Metric} \\ < \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i & \Longrightarrow & < \mathbf{x}, \mathbf{x} >= \sum_{i=1}^{n} x_i^2 = ||\mathbf{x}||_2^2 & \Longrightarrow & d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_2 \end{array}$$

Geometry of vectors in two-dimensional Euclidean vector space

- **Notation**: In the following we write: $||\mathbf{x}|| = ||\mathbf{x}||_2$.
- Geometry of **addition** of vectors: result is diagonal in parallelogram.
- Geometry of **multiplication with a scalar** α: $\alpha\mathbf{x}$ is vector parallel to $\mathbf{x}$ with different length and possibly with opposite direction.
- Geometry of the **scalar product** or **inner product** of two vectors:
  Product of the lengths of the two vectors and the cosine of the angle θ between them (without proof)

$$< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i = ||\mathbf{x}|| \; ||\mathbf{y}|| \cos\theta. \tag{1.2}$$

for validity: Given two special vectors in E^2:

$$\mathbf{w} = \begin{pmatrix} 1 & 0 \end{pmatrix},$$
$$\mathbf{z} = \begin{pmatrix} \cos\theta & \sin\theta \end{pmatrix}.$$

Scalar multiplication gives two more vectors:

$$\begin{aligned} \mathbf{x} &= \alpha\mathbf{w}, \quad \alpha > 0, \\ \mathbf{y} &= \gamma\mathbf{z}, \quad \gamma > 0. \end{aligned}$$

Then the inner products or scalar products are obtained

$$\begin{aligned} ||\mathbf{w}|| &= 1, \\ ||\mathbf{z}|| &= \left(\cos^2\theta + \sin^2\theta\right)^{1/2} = 1, \\ < \mathbf{w}, \mathbf{z} > &= w_1 z_1 + w_2 z_2 = \cos\theta \end{aligned}$$

and

$$\begin{aligned} ||\mathbf{x}|| &= |\alpha|||\mathbf{w}|| = \alpha, \\ ||\mathbf{y}|| &= |\gamma|||\mathbf{z}|| = \gamma, \\ < \mathbf{x}, \mathbf{y} > &=< \alpha\mathbf{w}, \gamma\mathbf{z} >= \alpha w_1 \gamma z_1 + \alpha w_2 \gamma z_2 = \alpha\gamma < \mathbf{w}, \mathbf{z} > \\ &= \alpha\gamma\cos\theta \\ &= ||\mathbf{x}||\ ||\mathbf{y}||\cos\theta. \end{aligned}$$

Orthogonal vectors

- The inner product of two vectors is zero if and only if the two vectors are **orthogonal to each other** (perpendicular to each other), since $\cos 90^o = 0$. I. e.:
  $< \mathbf{x}, \mathbf{y} >= 0 \quad \Longleftrightarrow$ the vectors $\mathbf{x}$ und $\mathbf{y}$ are orthogonal to each other.
- **Cauchy-Schwarz Inequality**

  $$| < \mathbf{x}, \mathbf{y} > | \leq ||\mathbf{x}||\ ||\mathbf{y}|| \quad \text{or} \quad < \mathbf{x}, \mathbf{y} >^2 \leq < \mathbf{x}, \mathbf{x} > < \mathbf{y}, \mathbf{y} > .$$

  This follows from (1.2) and $-1 \leq \cos\theta \leq 1$.

Linear independence

- **Linear independence**: k vectors $\mathbf{x}_i$, $i = 1, \ldots, k$, (with positive length) are linearly independent, if there are *no* $k-1$ scalars c_i such that

  $$\mathbf{x}_j = \sum_{\substack{i=1 \\ i \neq j}}^{k} c_i \mathbf{x}_i, \quad 1 \leq j \leq k$$

  holds.

  Example: Let the columns of the $n \times k$ matrix $\mathbf{X}$ be linearly independent. Then, there exists only a zero vector $\boldsymbol{\gamma}$, i.e. no $\boldsymbol{\gamma}$ with positive length, such that

  $$\sum_{i=1}^{k} \mathbf{x}_{ji}\boldsymbol{\gamma}_i = 0, \quad j = 1, \ldots, n.$$

1.4. Matrices

Overview

- Definition
- Addition of matrices
- Zero matrix
- Scalar multiplication
- Subtraction of matrices

Matrices

Definition

- A **matrix A** is a rectangular scheme of nm numbers, $n, m \in \mathbb{N}$,

$$\mathbf{A} := \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = (a_{ij})_{i=1,\dots,m,j=1,\dots,n} = (a_{ij})$$

- **Dimension** of a matrix: Number of rows m and number of columns n.
  Short notation: $(m \times n)$-matrix or $m \times n$-matrix.
- The entries a_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n$ are called **elements** or **coefficients** of a matrix.

Remarks

- **Note**: Often, in addition to the dimension n, a vector is defined as a column or row vector.

$$\mathbf{x} = \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}_{\text{column vector}} \quad \text{or} \quad \mathbf{x} = \underbrace{\begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}}_{\text{row vector}}$$

  However, in `R`, no column or row property is assigned to the class of vectors. This only happens with the class of matrices. Remember this!

- In `R` the class `matrix` exists. It is needed to define column or row vectors.
- A $(m \times n)$ matrix consists of n column vectors of length m, or m row vectors of length n. In the case of real numbers as elements one writes

  $$\mathbf{A} \in \mathbb{R}^{m \times n}.$$

  Since n vectors of dimension m are present.

Remarks – continued

- Two matrices of the same dimension are identical if all coefficients are equal.
- The elements can come from different sets: e. g. $\mathbb{N}$, $\mathbb{R}$, the complex numbers $\mathbb{C}$ or polynomials.
- Each table corresponds to a matrix.
- A column vector of length m corresponds to a $(m \times 1)$ matrix. A row vector of length n corresponds to a $(1 \times n)$ matrix.

Basic operations with matrices

- The basic operations **addition** and **multiplication by a number** from section 1.1 can also be applied to matrices.
- All further properties concerning these operations from section 1.1 apply accordingly to matrices as can be seen in the following.

1.4.1. Addition of matrices

The addition of two matrices $\mathbf{A}$, $\mathbf{B}$ with the same dimension m and n again gives a $(m \times n)$ matrix.

The (i, j)-th element is just the sum of the (i, j)-th elements of the matrices to be added.

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & b_{ij} & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & a_{ij}+b_{ij} & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{pmatrix}$$

Example:

$$\begin{pmatrix} 3 & 4 & 1 \\ 6 & 7 & 0 \\ -1 & 3 & 8 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 7 \\ 6 & 5 & 1 \\ -1 & 7 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 8 \\ 12 & 12 & 1 \\ -2 & 10 & 8 \end{pmatrix}$$

Example: Be careful:

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & -2 \end{pmatrix} + \begin{pmatrix} 3 & 6 \\ 1 & 4 \end{pmatrix} \text{ is not defined!}$$

Zero matrix

Definition

A $(m \times n)$-matrix $\mathbf{0}$ is called **zero matrix** if all entries are 0.

It follows:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} 0_{11} & \cdots & 0_{1n} \\ \vdots & 0_{ij} & \vdots \\ 0_{m1} & \cdots & 0_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

bzw.

$$\mathbf{A} + \mathbf{0} = \mathbf{A} \qquad \forall \mathbf{A}$$

1.4.2. Scalar multiplication

Multiplying a matrix $\mathbf{A}$ by a number λ again gives a matrix. Here, each element a_{ij} is multiplied by λ.

$$\lambda \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & \lambda a_{ij} & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{pmatrix}.$$

1.4.3. Subtraction of matrices

Definition of a negative matrix

The matrix $-\mathbf{A}$ results from multiplying a matrix by a scalar or from the matrix that must be added to $\mathbf{A}$ to get the zero matrix.

$$-\mathbf{A} = -\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} -a_{11} & \cdots & -a_{1n} \\ \vdots & -a_{ij} & \vdots \\ -a_{m1} & \cdots & -a_{mn} \end{pmatrix}.$$

Subtraction

$\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$.

It follows:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} - \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & b_{ij} & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} - b_{11} & \cdots & a_{1n} - b_{1n} \\ \vdots & a_{ij} - b_{ij} & \vdots \\ a_{m1} - b_{m1} & \cdots & a_{mn} - b_{mn} \end{pmatrix}$$

1.5. Other operations with matrices

Overview

- Matrix multiplication
- Element-wise multiplication or Hadamard product
- Transpose of a matrix
- Calculation rules
- Multiplication of vectors: Inner and outer product

1.5.1. Matrix multiplication

- **Requirement for matrix multiplication AB** of two matrices: a $(k \times r)$ matrix **A** and a $(m \times n)$ matrix **B**:
  - **The number of columns r of A is equal to the number of rows m of B**, i.e. **A** must have dimension $(k \times m)$ and **B** dimension $(m \times n)$.
  - This condition is necessary and sufficient.
- The order of multiplication:
  - cannot be interchanged if $k \neq n$,
  - can be interchanged if $k = n$, but generally with different results.
- Matrix multiplication is based on the scalar product
- **Notation**: $(\mathbf{AB})_{ij}$ denotes the (i, j)-th element of the matrix **AB**.

Calculation of the matrix product

The (i,j)-th entry of the matrix product $\mathbf{AB}$ is defined as the scalar product of the i-th row of $\mathbf{A}$ (a row vector) with the j-th column of $\mathbf{B}$ (a column vector):

$$\begin{aligned}(\mathbf{AB})_{ij} &= \begin{pmatrix} a_{i1} & a_{i2} & \cdots & a_{im}\end{pmatrix} \cdot \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{mj}\end{pmatrix} \\ &= a_{i1}b_{1j} + a_{i2}b_{2j} \cdots a_{im}b_{mj} \\ &= \sum_{h=1}^{m} a_{ih}b_{hj}\end{aligned}$$

Example:

$$\begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \\ eA + fC & eB + fD \end{pmatrix}.$$

Note: The product in reverse order is not defined!

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}$$

Dimension of a matrix product

The matrix product $\mathbf{AB}$ inherits the number of rows r from $\mathbf{A}$ and the number of columns n from $\mathbf{B}$:

$$\begin{array}{ccccc} \mathbf{A} & \cdot & \mathbf{B} & = & \mathbf{C} \\ (k \times m) & \cdot & (m \times n) & = & (k \times n). \end{array}$$

- **Good practice**: Check the dimensions of the matrices before each matrix multiplication! Especially when programming!
- **Note in `R`**: The matrix product is given as `A %*% B`. In other languages, however, often with `A*B`.

No commutative property for matrix multiplication

- Given a $(n \times m)$ matrix $\mathbf{A}$ and a $(m \times n)$ matrix $\mathbf{B}$

$$\begin{array}{ccccc} \mathbf{A} & \cdot & \mathbf{B} & = & \mathbf{C} \\ (n \times m) & \cdot & (m \times n) & = & (n \times n). \\ \mathbf{B} & \cdot & \mathbf{A} & = & \mathbf{D} \\ (m \times n) & \cdot & (n \times m) & = & (m \times m). \end{array}$$

- Even if **A** and **B** are quadratic, i.e. number of rows and columns are equal, $m = n$, $\mathbf{AB} \neq \mathbf{BA}$ can occur.

No commutative property for matrix multiplication

Example:

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix},$$

while

$$\begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 2 \end{pmatrix}.$$

Identity matrix

The $(n \times n)$ matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

where $a_{ii} = 1$, $\forall i$ and $a_{ij} = 0$, $\forall i \neq j$, is called the identity matrix.

Properties of the identity matrix **I**:

- Multiplicative identity for matrix multiplication:
  - for each $(m \times n)$ matrix **A**, it holds that
  $$\mathbf{AI} = \mathbf{A}, \tag{1.3}$$
  - for each $(n \times m)$ matrix **B**, it holds that
  $$\mathbf{IB} = \mathbf{B}. \tag{1.4}$$
- **I** corresponds to 1 in the real numbers.

Element-wise multiplication (Hadamard product)

For two $(m \times n)$ matrices $\mathbf{A}$ and $\mathbf{B}$, element-wise multiplication for the (i, j)-th entry of the Hadamard product $\mathbf{A} \odot \mathbf{B}$ yields

$$(\mathbf{A} \odot \mathbf{B})_{ij} = a_{ij} b_{ij}$$

The resulting matrix again has dimension $(m \times n)$ like $\mathbf{A}$ and $\mathbf{B}$.

Example:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \odot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{pmatrix}$$

Remarks:

- In `R`, `A*B` is used for the element-wise product! In other languages this notes the matrix product!

1.5.2. Calculation rules for matrices

- **Associativity for addition and matrix multiplication:**

$$\begin{aligned} (\mathbf{A} + \mathbf{B}) + \mathbf{C} &= \mathbf{A} + (\mathbf{B} + \mathbf{C}), \\ (\mathbf{A}\mathbf{B})\mathbf{C} &= \mathbf{A}(\mathbf{B}\mathbf{C}) \end{aligned}$$

- **Commutative property for addition**

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

- **Distributivity for matrix multiplication**

$$\begin{aligned} \mathbf{A}(\mathbf{B} + \mathbf{C}) &= \mathbf{A}\mathbf{B} + \mathbf{A}\mathbf{C}, \\ (\mathbf{A} + \mathbf{B})\mathbf{C} &= \mathbf{A}\mathbf{C} + \mathbf{B}\mathbf{C} \end{aligned}$$

Reminder: In general: $\mathbf{AB} \neq \mathbf{BA}$: Matrix multiplication not commutative!

Transpose of a matrix

Definition: Transpose of a matrix A

Notation: $\mathbf{A}^T$ or $\mathbf{A}'$.
The transpose of a $(k \times n)$-matrix $\mathbf{A}$ is a $(n \times k)$-matrix obtained by transposing each row so that (i, j)-th element of $\mathbf{A}$ becomes the (j, i)-th element of $\mathbf{A}^T$.

The calculation can also be done by interchanging rows and columns:

First column of $\mathbf{A}$ becomes first row of $\mathbf{A}^T$,

Second column of $\mathbf{A}$ becomes second row of $\mathbf{A}^T$,

Example:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix},$$

$$\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} \end{pmatrix}.$$

Calculation rules with transposed matrices

Addition and multiplication by a number

Given two $(m \times n)$ matrices $\mathbf{A}$ and $\mathbf{B}$, and a scalar α

$$\begin{aligned} (\mathbf{A}+\mathbf{B})^T &= \mathbf{A}^T + \mathbf{B}^T \\ (\mathbf{A}-\mathbf{B})^T &= \mathbf{A}^T - \mathbf{B}^T \\ (\mathbf{A}^T)^T &= \mathbf{A} \\ (\alpha\mathbf{A})^T &= \alpha\mathbf{A}^T \end{aligned}$$

It is a good exercise to prove these rules!

Matrix multiplication

Given a $(k \times m)$ matrix $\mathbf{A}$ and a $(m \times n)$ matrix $\mathbf{B}$. It holds that

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{1.5}$$

- *Note the swapping of the order!!*
- **Notation**: $\left((\mathbf{AB})^T\right)_{ij}$ denotes the (i,j)-th element of $(\mathbf{AB})^T$. Analogous to previous notation.
- R-command: `t(A)`.

Proof of (1.5):

$$\begin{aligned} \left((\mathbf{AB})^T\right)_{ij} &= (\mathbf{AB})_{ji} && \text{(Definition of the transpose)} \\ &= \textstyle\sum_h \mathbf{A}_{jh} \cdot \mathbf{B}_{hi} && \text{(Definition of matrix multiplication)} \\ &= \textstyle\sum_h (\mathbf{A}^T)_{hj} \cdot (\mathbf{B}^T)_{ih} && \text{(Definition of the transpose, twice)} \\ &= \textstyle\sum_h (\mathbf{B}^T)_{ih} \cdot (\mathbf{A}^T)_{hj} && (a \cdot b = b \cdot a) \text{ for scalars} \\ &= (\mathbf{B}^T\mathbf{A}^T)_{ij} && \text{(Definition of matrix multiplication.)} \end{aligned}$$

Therefore, it holds that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. **QED**

Multiplication of vectors

Special case of the transposition of a matrix

- If a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as a column vector, the transposition of $\mathbf{x}$ results in a row vector with the same n-tuple

$$\mathbf{x}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}^T := \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$$

- The notation $\mathbf{x}'$ is also frequently used instead of $\mathbf{x}^T$.

Inner product

The scalar product $<\mathbf{x}, \mathbf{y}>= \sum_{i=1}^n x_i y_i$ implies the following vector multiplications

$$<\mathbf{x}, \mathbf{y}>= \sum_{i=1}^n x_i y_i$$

and if $\mathbf{x}, \mathbf{y}$ column vectors

$$<\mathbf{x}, \mathbf{y}>= \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$$

- Another option is the outer product.

Outer product

For two column vectors $\mathbf{x}, \mathbf{y}$ of length n one obtains for $\mathbf{x}\mathbf{y}^T$ a $(n \times n)$ matrix (see section 1.5.1)

$$\mathbf{x}\mathbf{y}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix} = \begin{pmatrix} x_1y_1 & \cdots & x_1y_n \\ \vdots & \ddots & \vdots \\ x_ny_1 & \cdots & x_ny_n \end{pmatrix}.$$

Be careful:

- In both cases, **only vectors with the same length** can be **multiplied** with each other.
- Vectors of different lengths cannot be multiplied with each other.

1.6. Important special matrices

Overview

- Square matrix
- Diagonal matrix
- Symmetric Matrix
- Upper/lower triangular matrix
- Idempotent matrix

Square matrix

A $(n \times n)$ matrix is a square matrix. The number of columns and rows is equal.

Diagonal matrix

A square $(n \times n)$ matrix $\mathbf{A}$ is a diagonal matrix, if all nondiagonal elements a_{ij}, $i \neq j$, $i, j = 1, \dots, n$ are zero.

Symmetric matrix

A square $(n \times n)$ matrix $\mathbf{A}$ is symmetric, if for all $i, j = 1, \dots, n$ $a_{ij} = a_{ji}$, or

$$\mathbf{A} = \mathbf{A}^T$$

holds.

Upper triangular matrix

A square matrix $\mathbf{A}$ is an upper triangular matrix if for all $i > j$, $i, j = 1, \dots, n$, it holds that: $a_{ij} = 0$.

Lower triangular matrix

A square matrix $\mathbf{A}$ is a lower triangular matrix if for all $i < j$, $i, j = 1, \dots, n$, it holds that: $a_{ij} = 0$.

Idempotent matrix

A square matrix $\mathbf{A}$ is called **idempotent**, if

$$\mathbf{A}\mathbf{A} = \mathbf{A}.$$

holds.

Examples:

Diagonal matrices

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Symmetric matrices

$$\begin{pmatrix} a & b \\ b & d \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}$$

Upper triangular matrix

$$\begin{pmatrix} a & c \\ 0 & b \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 2 & 9 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Lower triangular matrix

$$\begin{pmatrix} a & 0 \\ c & d \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & 0 \\ 2 & 4 & 0 \\ 3 & 5 & 6 \end{pmatrix}$$

1.7. Measures of matrices

Overview

- Trace
- Rank
- Determinant

1.7.1. Trace of a matrix

Definition

The **trace** of a square matrix $\mathbf{A}$ is the sum of the elements a_{ii} on the diagonal

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

The trace is a mapping from $\mathbb{R}^{n \times n}$ to $\mathbb{R}$.

Example:

$$\operatorname{tr}(\mathbf{I}) = n, \quad \operatorname{tr}\begin{pmatrix} 1 & 3 \\ a & b \end{pmatrix} = 1 + b$$

Calculation rules

Given $(n \times n)$ matrices $\mathbf{A}$, $\mathbf{B}$ and a scalar $\alpha \in \mathbb{R}$:

- $\operatorname{tr}(\mathbf{A}) = \operatorname{tr}\left(\mathbf{A}^T\right)$
- $\operatorname{tr}(\alpha \mathbf{A}) = \alpha \operatorname{tr}(\mathbf{A})$
- $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$
- further rules in Schmidt & Trenkler (2006, Section 3.1).

Check!

1.7.2. Rank of a matrix

Definition

The **rank** $\text{rk}(\mathbf{A})$ is a mapping from $\mathbb{R}^{m\times n} \to \mathbb{N}$ that assigns to a $(m \times n)$ matrix $\mathbf{A}$ the maximum number of vectors (either row or column vectors) that are linearly independent.

- A $(m \times n)$ matrix $\mathbf{A}$ has **full rank**, if the rank of the matrix is equal to the smaller dimension, i.e.

$$\text{rk}(\mathbf{A}) = \begin{cases} m, & \text{if } m \leq n \text{ and all } m \text{ rows are linearly independent,} \\ n, & \text{if } m \geq n \text{ and all } n \text{ columns are linearly independent.} \end{cases}$$

- A $(m \times n)$ matrix $\mathbf{A}$ has **full column rank** if $\text{rk}(\mathbf{A}) = n$.
- A $(m \times n)$ matrix $\mathbf{A}$ has **full row rank** if $\text{rk}(\mathbf{A}) = m$.

Remarks

- A matrix that does not have full rank has a **rank deficiency**.
- The rank is smaller than the column number k of $\mathbf{X}$ if columns of $\mathbf{X}$ are linearly dependent. Then
  - a matrix $\mathbf{X}'$ can be formed consisting of k' linearly independent columns of $\mathbf{X}$ such that $\text{rk}(\mathbf{X}) = k' < k$ and
  - for the subspaces, see definition in section 1.9, it holds that $\delta(\mathbf{X}) = \delta(\mathbf{X}')$,
  - also $\mathbf{X}^T\mathbf{X}$ has a rank deficiency, since $\text{rk}(\mathbf{X}) = \text{rk}(\mathbf{X}^T\mathbf{X}) = k'$, and is singular. (Cf. MLR.3 in Introduction to Econometrics).
- **R code** : `rankMatrix()` in `R` package `Matrix`

Calculation rules

Given $(m \times n)$ matrices $\mathbf{A}$, $\mathbf{B}$:

- $0 \leq \text{rk}(\mathbf{A}) \leq \min(m, n)$
- $\text{rk}(\mathbf{A}) = \text{rk}(\mathbf{A}^T) = \text{rk}(\mathbf{A}^T\mathbf{A}) = \text{rk}(\mathbf{A}\mathbf{A}^T)$
- $\text{rk}(\mathbf{A} + \mathbf{B}) \leq \text{rk}(\mathbf{A}) + \text{rk}(\mathbf{B})$
- $\text{rk}(\mathbf{AC}) \leq \min(\text{rk}(\mathbf{A}), \text{rk}(\mathbf{C}))$
- further rules in Schmidt & Trenkler (2006, Section 3.2).

1.7.3. Determinants

Determinant

- A **determinant** is a mapping $\mathbb{R}^{n\times n} \to \mathbb{R}$, which assigns a real number to a square matrix $\mathbf{A}$.
- The determinant has an important role in determining the solutions of linear systems of equations but also in geometry. Fischer (2010, Section 3.1.1)
- The determinant is noted as $|\mathbf{A}|$ or as $\det(\mathbf{A})$.
- The calculation of a determinant can be done recursively. Gentle (2007, Section 3.1.5) or Schmidt & Trenkler (2006, Section 3.3).
- For $n \leq 3$ there are simple calculation formulas.

$\sharp$ **Geometric Interpretation**: The $(n \times 1)$ vector defines in the n-dimensional Euclidean space E^n an n-dimensional parallelepiped (= parallelogram for $n = 2$) for which a volume (for $n = 2$ an area) can be calculated.

If a $(n \times 1)$ vector $\mathbf{x}$ is multiplied from the left by the matrix $\mathbf{A}$, this corresponds to a mapping of

$$E^n \longrightarrow E^n : \mathbf{x} \longrightarrow \mathbf{z} = \mathbf{A}\mathbf{x}.$$

The determinant $|\mathbf{A}|$ indicates by how much the volumes determined by $\mathbf{x}$ and $\mathbf{z}$ respectively differ (An example for $n = 2$ can be found in Davidson & MacKinnon 2004, Section 12.2, pp. 511-512).

Calculation of the determinant for $n = 2, 3$

- (2×2) matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \det \mathbf{A} = |\mathbf{A}| = ad - bc.$$

- (3×3) matrix (Sarrus' Rule)

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \\ \det \mathbf{A} &= |\mathbf{A}| = aei + bfg + cdh - gec - hfa - idb \\ & \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{matrix} a & b \\ d & e \\ g & h \end{matrix} \end{aligned}$$

1.8. Matrix inversion

Overview

- Definition of an inverse matrix
- Calculation for (2 × 2) matrix
- Existence
- Calculation rules

The inverse of a matrix

- is defined only for square matrices.
- results from the solution of a linear system of equations.
- plays a central role in matrix algebra.

Inverse of a matrix

A square matrix $\mathbf{A}$ is called **invertible** if there exists a square matrix $\mathbf{B}$ such that it holds:

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}.$$

The matrix $\mathbf{B}$ is called the **inverse** $\mathbf{A}^{-1}$.

- The inverse is a mapping $\mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$.
- A non-invertible matrix $\mathbf{A}$ is called **singular**.
- An invertible matrix $\mathbf{A}$ is called **regular** or **nonsingular**.

Calculation of the inverse for $n = 2, 3$

- (2 × 2) matrix

$$\mathbf{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

- (3 × 3) matrix

$$\mathbf{A}^{-1} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} ei - fh & ch - bi & bf - ce \\ fg - di & ai - cg & cd - af \\ dh - eg & bg - ah & ae - bd \end{pmatrix}$$

- For $n > 3$, complicated procedures are necessary, which are best left to the computer.

Existence of the inverse

- **Existence of the inverse**: The inverse $\mathbf{A}^{-1}$ exists if and only if the determinant of $\mathbf{A}$ is non-zero, $|\mathbf{A}| \neq 0$. This holds for all n!
- **Important**: If the determinant is close to zero in calculations, large numerical inaccuracies can occur. Therefore, when programming, the use of the inverse is avoided if possible.
- If **the inverse exists**, a **linear system of equations**

  $$\mathbf{Ax} = \mathbf{b}$$

  is **uniquely solvable**:

  $$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

- For non-square and non-invertible matrices there are **generalised inverses**.

Calculation rules for inverses

Let $\mathbf{A}$ be regular.

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $\left(\mathbf{A}^T\right)^{-1} = (\mathbf{A}^{-1})^T$
- Let $\mathbf{B}$ be regular. Then $\mathbf{AB}$ is regular and $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- If $\mathbf{A}$ is a diagonal matrix, then $\mathbf{A}^{-1} = (1/a_{ii})$.

Calculation rules for determinants

Given $(n \times n)$ matrices $\mathbf{A}$, $\mathbf{B}$ and a scalar $\lambda \in \mathbb{R}$:

- $|\mathbf{A}| = 0 \quad \Longleftrightarrow \quad \text{rk}(\mathbf{A}) < n \quad \Longleftrightarrow \mathbf{A}$ is singular
- $|\mathbf{A}| \neq 0 \quad \Longleftrightarrow \quad \text{rk}(\mathbf{A}) = n \quad \Longleftrightarrow \mathbf{A}$ is regular
- $|\lambda\mathbf{A}| = \lambda^n|\mathbf{A}|$
- $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- $|\mathbf{A}^T| = |\mathbf{A}|$

- $|\mathbf{A}| = \prod_{i=1}^{n} a_{ii}$, if $\mathbf{A}$ is a diagonal or a triangular matrix.

- further rules e. g. in Schmidt & Trenkler (2006, Section 3.3).

1.9. Euclidean subspaces

Overview

- Basis vectors in E^n
- Euclidean subspaces
- Column space of a matrix
- Orthogonal complement

Basis vectors in E^n

Definition

n different $(n \times 1)$ vectors are **basis vectors** if no basis vector can be represented as a linear combination of the other $(n-1)$ basis vectors. I. e., the basis vectors are linearly independent.

Remarks

- Each element in Euclidean space E^n can be represented as a **linear combination** of n **basis vectors**.

- One then says: **The** n **basis vectors span** E^n, i. e. form an Euclidean space E^n. If one denotes the n basis vectors by $\mathbf{x}_i$, then the set of all vectors in E^n is given by

$$\left\{ \mathbf{z} \in E^n \,\middle|\, \mathbf{z} = \sum_{i=1}^{n} b_i \mathbf{x}_i, \; b_i \in \mathbb{R}, i = 1, \dots, n \right\}.$$

Euclidean subspaces

Definition

If one reduces the number of basis vectors to $k < n$, only a subset of the vectors can be represented in E^n. Such a subset forms an **Euclidean subspaces**.

Notation and ways of speaking

- We denote the subspace spanned by k basis vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ by $\delta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$, or $\delta(\mathbf{X})$, if all basis vectors are combined in the matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$.

Column space of a matrix

- The set of vectors $\mathbf{z}$ contained in the subspace, i. e. all linear combinations of the columns of the $(n \times k)$ matrix $\mathbf{X}$, can be described as

$$\delta(\mathbf{X}) = \delta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k) := \left\{ \mathbf{z} \in E^n \,\middle|\, \mathbf{z} = \sum_{i=1}^{k} b_i \mathbf{x}_i,\ b_i \in \mathbb{R} \right\}. \tag{1.6}$$

- One says that the subspace $\delta(\mathbf{X})$ corresponds to the **column space of the matrix X**.

Orthogonal complement

- The **orthogonal complement** to the subspace $\delta(\mathbf{X})$ is another subspace in E^n, for which it holds that

$$\begin{aligned} \delta^{\perp}(\mathbf{X}) &= \delta^{\perp}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k) \\ &:= \left\{ \mathbf{w} \in E^n \,\middle|\, < \mathbf{w}, \mathbf{z} >= \mathbf{w}^T \mathbf{z} = 0 \text{ for all } \mathbf{z} \in \delta(\mathbf{X}) \right\}. \end{aligned} \tag{1.7}$$

Question: Let $\dim \delta(\mathbf{X}) = k$ be the dimension of $\delta(\mathbf{X})$. Then what is $\dim \delta^{\perp}(\mathbf{X})$?

1.10. Matrices and linear mappings

Overview

- Mapping between two vector spaces
- Linear mapping between two vector spaces
- Kernel and image of a linear mapping

Given two Euclidean vector spaces that can have different dimensions n and m. Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. The **mapping**

$$\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m, \qquad \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

for short

$$\mathbf{F}(\mathbf{x}) = \mathbf{y} = \mathbf{A}\mathbf{x}$$

assigns to each point $\mathbf{x}$ in the n-dimensional Euclidean space $\mathbb{R}^n$ a point $\mathbf{y}$ in the m-dimensional Euclidean space $\mathbb{R}^m$.

Linear mapping

A mapping

$$\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m, \quad \mathbf{F}(\mathbf{x}) = \mathbf{y} = \mathbf{A}\mathbf{x}$$

is called **linear**, if the following properties hold:

1. $\mathbf{F}(\mathbf{x} + \mathbf{z}) = \mathbf{F}(\mathbf{x}) + \mathbf{F}(\mathbf{z})$
2. $\mathbf{F}(\lambda\mathbf{x}) = \lambda\mathbf{F}(\mathbf{x})$

for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$.

Kernel and image of a linear mapping

Let $\mathcal{V} \in \mathbb{R}^n$ and $\mathcal{W} \in \mathbb{R}^m$. For the mapping $\mathbf{F} : \mathcal{V} \to \mathcal{W}$

- $\operatorname{Im} \mathbf{F} := \mathbf{F}(\mathcal{V})$ denotes the **image** of this mapping,

- $\operatorname{Ker}\mathbf{F} := \mathbf{F}^{-1}(\mathbf{0})$ denotes the **kernel** of this mapping.

- The kernel can be determined by $\mathbf{F}^{-1}(\mathbf{y}) = \mathbf{A}^{-1}\mathbf{y}$ with $\mathbf{y} = \mathbf{0}$ if the inverse exists.
- The kernel determines the set of all $\mathbf{x} \in \mathcal{V}$, whose image is just the origin in $\mathcal{W}$.

1.11. Matrix representation of linear systems of equations

Overview

- Definition of suitable matrices
- System of equations in matrix form
- Unique solution

Consider a typical system of linear equations:

$$\begin{array}{cccc} a_{11}x_1+ & \cdots & +a_{1n}x_n & = & b_1 \\ a_{21}x_1+ & \cdots & +a_{2n}x_n & = & b_2 \\ \vdots & \vdots & & & \vdots \\ a_{k1}x_1+ & \cdots & +a_{kn}x_n & = & b_k. \end{array}$$

The linear system of equations can be represented more compactly with matrices.

Definition of suitable matrices: $(k \times n)$-**coefficient matrix** $\mathbf{A}$, $(n \times 1)$-**variable vector** $\mathbf{x}$ and $(k \times 1)$-**parameter vector** $\mathbf{b}$

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{k1} & \cdots & a_{kn} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{and} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

System of equations in matrix form

The system of equations is then:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{k1} & \cdots & a_{kn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

In compact form

$$\mathbf{Ax} = \mathbf{b}.$$

The matrix product $\mathbf{Ax}$ yields a $(k \times 1)$ vector equal to the $(k \times 1)$ parameter vector $\mathbf{b}$ when $\mathbf{x}$ is a solution of the system of equations.

Unique solution

If $\mathbf{A}$ is regular, i. e. invertible, then there exists a unique solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

1.12. (Semi-)definite matrices

Overview

- Quadratic form
- Positive definite and positive semi-definite matrices
- Negative definite and negative semi-definite matrices
- Indefinite matrices

Quadratic form

$\mathbf{x}^T\mathbf{Ax} = \sum_{i=1}^{k}\sum_{j=1}^{k} x_i x_j A_{ij}$ is a quadratic form. The result is a scalar.

Positive definite und semidefinite matrices

- A $(k \times k)$ matrix $\mathbf{A}$ is called **positive definite**, if for any $(k \times 1)$ vector $\mathbf{x}$ with positive norm

  $$\mathbf{x}^T\mathbf{Ax} > 0$$

  holds.
- A $(k \times k)$ matrix $\mathbf{A}$ is called **positive semidefinite**, if for any $(k \times 1)$ vector $\mathbf{x}$ with

positive norm it holds that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0.$$

Negative definite and semidefinite matrices

- A $(k \times k)$ matrix $\mathbf{A}$ is called **negative definite**, if for any $(k \times 1)$ vector $\mathbf{x}$ with positive norm it holds that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} < 0.$$

- A $(k \times k)$ matrix $\mathbf{A}$ is called **negative semidefinite**, if for any $(k \times 1)$ vector $\mathbf{x}$ with positive norm it holds that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \leq 0.$$

Indefinite matrices

Matrices that are neither positive nor negative (semi-)definite are called **indefinite**.

- If $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, then $\mathbf{A}$ is always positive semidefinite, since

$$\mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} = (\mathbf{B}\mathbf{x})^T(\mathbf{B}\mathbf{x}) = ||\mathbf{B}\mathbf{x}||^2 \geq 0. \tag{1.8}$$

  If $\mathbf{B}$ has full rank, $\mathbf{A}$ is positive definite. Why?

- The diagonal elements of a positive definite matrix are positive. Moreover, for every positive definite matrix $\mathbf{A}$ there exists a matrix $\mathbf{B}$ such that $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ holds. Here $\mathbf{B}$ is not unique.

- A matrix $\mathbf{A}$ is called **negative (semi-)definite**, if $-\mathbf{A}$ is positive (semi-)definite.

Example:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

is positive definite. This is because for every vector $\mathbf{z} = \begin{pmatrix} z_0 \\ z_1 \end{pmatrix}$ with $||\mathbf{z}|| > 0$, it holds that

$$\begin{pmatrix} z_0 & z_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \end{pmatrix} = z_0^2 + z_1^2 > 0.$$

Example: The matrix $\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is indefinite, since it is neither positive nor negative semidefinite. This is because for $\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ one obtains

$$\begin{pmatrix} z_1 & z_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} z_2 & z_1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 2z_1z_2.$$

Depending on the choice of z_1 and z_2, the result is positive, zero or negative.

1.13. Calculation rules for the derivative of vector-valued functions

Overview

- First partial derivatives of scalar products
- First partial derivatives of linear combinations
- First partial derivatives for quadratic forms
- ♯ Jacobian matrix

First partial derivatives of scalar products

Given are the $(n \times 1)$ column vectors $\mathbf{v}$ and $\mathbf{w}$. For the first partial derivative of the scalar product $z =< \mathbf{v}, \mathbf{w} >= \mathbf{v}^T\mathbf{w} = \mathbf{w}^T\mathbf{v} = \sum_{i=1}^n v_i w_i$ with respect to w_i, it holds that $\partial z/\partial w_i = v_i$. Collecting all first partial derivatives with respect to $\mathbf{w}$ in a column vector

$$\frac{\partial z}{\partial \mathbf{w}} = \begin{pmatrix} \frac{\partial z}{\partial w_1} \\ \frac{\partial z}{\partial w_2} \\ \vdots \\ \frac{\partial z}{\partial w_n} \end{pmatrix},$$

results in

$$\frac{\partial z}{\partial \mathbf{w}} = \mathbf{v}.$$

First partial derivatives of linear combinations

For $\mathbf{z} = \mathbf{A}\mathbf{w}$ with

$$\mathbf{z} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}$$

one obtains the first partial derivatives

$$\frac{\partial \mathbf{z}}{\partial \mathbf{w}^T} = \mathbf{A}$$

First partial derivatives for quadratic forms

For the quadratic form $v = \mathbf{w}^T \mathbf{A} \mathbf{w}$

$$v = \begin{pmatrix} w_1 & w_2 & \cdots & w_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}$$

one obtains the first partial derivatives

$$\frac{\partial v}{\partial \mathbf{w}} = \left(\mathbf{A} + \mathbf{A}^T \right) \mathbf{w}.$$

♯ Jacobian matrix

Let a vector-valued function be given for $\mathbf{x} \in \mathbb{R}^n$

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^m : \mathbf{x} \longrightarrow \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} g_1(\mathbf{x}) \\ \dots \\ g_m(\mathbf{x}) \end{pmatrix}.$$

The $(m \times n)$ matrix

$$\mathbf{J}(\mathbf{x}) \equiv \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}^T} \equiv \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m(\mathbf{x})}{\partial x_1} & \frac{\partial g_m(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_m(\mathbf{x})}{\partial x_n} \end{pmatrix} \tag{1.9}$$

of the first order partial derivatives is called the **Jacobian matrix**. If the Jacobi matrix is square, the determinant of the Jacobian matrix exists (often called the **Jacobian determinant**):

$$|\mathbf{J}(\mathbf{x})| = \left| \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}^T} \right|. \tag{1.10}$$

1.14. Partitioned matrices

Overview

- Addition, subtraction and matrix multiplication
- Inversion of a partitioned matrix

Partitioned matrices

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where the submatrices $\mathbf{A}_{ij}$ have dimension $(m_i \times n_j)$ and $m_1 + m_2 = m$, $n_1 + n_2 = n$ holds.

Calculation rules

Pay attention to the correct dimensions of the matrices and submatrices!

- $\mathbf{A}^T = \begin{pmatrix} \mathbf{A}_{11}^T & \mathbf{A}_{21}^T \\ \mathbf{A}_{12}^T & \mathbf{A}_{22}^T \end{pmatrix}$
- Addition: replace in standard addition elements with submatrices.
- Matrix multiplication: replace elements by corresponding submatrices

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix},$$

Inversion of a partitioned matrix

The inverse of a partitioned matrix can be calculated as follows

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BWCA}^{-1} & -\mathbf{A}^{-1}\mathbf{BW} \\ -\mathbf{WCA}^{-1} & \mathbf{W} \end{pmatrix}$$

with $\mathbf{W} = (\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}$.

2. Fundamentals of Probability Theory

Overview

- Important basic concepts
- Why do we need probability theory?
- Random variables
- Distribution and density functions (univariate and multivariate)
- Conditional probabilities
- Expected values and moments
- Conditional expected values and moments
- Important probability distributions

Literature references

- Davidson & MacKinnon (2004, Section 1.2): concise overview of the basics of probability theory.
- Casella & Berger (2002): very detailed, formal introduction to probability theory.
- Fahrmeir et al. (2016): simple introduction to statistics.
- Steland (2013): well-written, concise, technically precise introduction to statistics. (Available from the university network as full text here.)

2.1. Important basic concepts

Overview

- Population
- Sample

Important basic concepts

Definition: Population

"set of all statistical units about which one aims to gain information".

The population (Fahrmeier et al. 2004, Abschnitt 1.3.1, S. 14)

- depends on the question of interest,
- can be finite (proportions of a production with quality defects), infinite (set of all possible rail delays) or hypothetical (set of all potential buyers).

The population

- can in principle be observable (all students of the UR, amount of organically produced grain in a region within one year) or
- be unobservable (e. g. effect of a measure for a *single* individual)

Definition: Sample

A sample is typically a subset of the population that can be or has been observed and can be used to analyse the population.

Examples:

- Participant of a lecture
- 1 kg of grain per 100 randomly selected fields within a region
- Participant in the socio-economic panel

2.2. Why do we need probability theory?

Overview

- Illustrative task
- Inductive statistics, descriptive statistics, exploratory data analysis
- Inductive statistics and probability theory

Why do we need probability theory?

Task on gender distribution

What is the gender distribution of students at the beginning of the master's programme in Economics at the University of Regensburg?

- **Population**: All students starting a master's programme in Economics this semester.
- **Sample**: All students starting a master's programme in Economics this semester and sitting in this lecture room.

Inductive statistics versus descriptive statistics versus exploratory data analysis

Statements about

- sample/data:
  - Description of key data indicators: **descriptive statistics**
  - Looking for what else the data might reveal about formal models or hypothesis testing: **exploratory data analysis**
- population: **inductive statistics**

What statements can be made about the population?

To what extent can statements be made about the gender proportions in the population based on the information in this sample?

Possible answers without probability theory

On the basis of the sample, interval statements about the proportion of female students are possible. However, the larger the population is compared to the sample, the less precise these statements become.

Possible answers with probability theory

- Point predictions
- Interval predictions with shorter intervals and coverage probabilities
- always require additional assumptions. **More detailed statements** than statements about the possible range of the gender proportion in the population **require additional assumptions**!

  Examples:

  – The gender ratio in the population corresponds to that in the sample.

  – A random sample is present.

Continuation of the task on gender distribution

Responses without probability theory

The following table allows interval statements without probability theory after the actual sample data have been completed.

Subpopulation	Total number	Number female	Proportion female		
			possible range	point prediction based on sample	correct
Sample: present students of Methods of Econometrics who start a master's programme in Economics this semester					
present + 1 missing students of Methods of Econometrics who start a master's programme in Economics this semester					
all students who start a master's programme in Economics this semester at the UR					

Answers with probability theory

still require some patience and the study of probability theory!

2.3. Probability space

Overview

- Sample space
- Event
- Elementary event
- Sigma-algebra
- Probability function
- Probability space
- Calculation rule for probabilities

Sample space (outcome space)

Definition Sample space

The sample space Ω is the set of all possible outcomes of a random experiment.

The set can contain **countably** many or **uncountably** many outcomes.

Examples:

- Gender of a student: $\Omega = \{\text{female},\ \text{male}\}$
- Urn with 4 balls of different colours: $\Omega = \{\text{yellow}, \text{red}, \text{blue}, \text{green}\}$
- future monthly income of a household: $\Omega = [0, \infty)$

Remarks

- If the outcomes are finitely many, then the individual outcomes are often denoted by ω_i. For S outcomes, Ω is then
$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_S\}.$$
- If there are infinitely (more precisely: uncountably) many outcomes, then a single one of them is often denoted by ω.

Events

Definitions

- When a specific outcome occurs, it is referred to as an **event**.
- If the event contains exactly one element of the sample space, it is called an **elementary event**.
- An event is a **subset of the sample space** Ω, i. e. any set of possible elementary events = any subset of the set Ω including Ω itself.
- The sample space Ω is a **certain event**.
- The **complementary event** A^c to the event A contains all events that are in the sample space Ω but not in A.

Examples:

- Urn: Possible events are e. g. $\{\text{yellow}, \text{red}\}$ or $\{\text{red}, \text{blue}, \text{green}\}$. Complementary event to the event $A = \{\text{yellow}, \text{red}\}$ is $A^c = \{\text{blue}, \text{green}\}$.
- Household income: Possible events are all possible subintervals and combinations thereof, e. g. $(0, 5000]$, $[1000, 1001)$, $(400, \infty)$, 4000, etc.

Remarks

If one uses the general notation with the ω's, then we get

- in the case of S elementary events: $\{\omega_1, \omega_2\}$, $\{\omega_S\}$, $\{\omega_3, \ldots, \omega_S\}$, etc.
- in the case of infinitely (more precisely: uncountably) many elementary events within an interval $\Omega = (-\infty, \infty)$: $(a_1, b_1]$, $[a_2, b_2)$, $(0, \infty)$, etc., where the lower limit is always less than or equal to the upper limit, i. e. $(a_i \leq b_i)$.

Sigma-algebra

Preliminary remarks: Let's consider our example with the 4 balls in different colours. To make the example even more general, we denote $\omega_1 = \text{yellow}$, $\omega_2 = \text{red}$, $\omega_3 = \text{blue}$, $\omega_4 = \text{green}$: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Let us now suppose that we are interested in particular in whether the following events occur in *one* draw:

$$\mathcal{C} = \{\{\omega_1\}, \{\omega_1, \omega_3, \omega_4\}\},$$

which are combined in the set of subsets $\mathcal{C}$. If we now take a closer look at this collection of subsets $\mathcal{C}$, we notice that the elementary event $\{\omega_1\}$ can occur, but what do we do if it *does not* occur. Then inevitably the event $\{\omega_2, \omega_3, \omega_4\}$ must occur, but it is *not* included in the collection. This means that we cannot assign a probability to this event. Since this makes no sense, we have to extend the set $\mathcal{C}$ at least by the event $\{\omega_2, \omega_3, \omega_4\}$. It follows that a collection of subsets, for each of which we want to define probabilities, must have certain properties. For example, at least the complement of an event must always be contained in the collection of subsets. We can also consider that any union of subsets must also be included in the collection. If a collection of subsets fulfils these requirements, then it is called a sigma-algebra.

Note: A σ-algebra is a set of events (subsets) that allows the assignment of probabilities with respect to all contained events. For those interested, the definition:

♯ Definition of a sigma algebra

A set of subsets of Ω is called **sigma algebra** or σ-algebra (σ-field), if the following properties hold for this set of subsets. A σ-algebra is often denoted by $\mathcal{F}$:

1. $\emptyset \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Remark: In the case of finitely many elementary events, the σ-algebra is identical to the power set. In the case of infinitely many elementary events, for example in the case of the possibe intervals of real numbers, the σ-algebra is smaller than the power set. This concept was developed precisely for this case, as the power set would be "too big".

Sigma algebra and probability function

Probability function

Let there be given a set Ω and a σ-algebra $\mathcal{F}$. Then a probability function P is a function with domain of definition $\mathcal{F}$ that satisfies the following conditions:

1. $P(A) \geq 0$ for all $A \in \mathcal{F}$
2. $P(\Omega) = 1$, $P(\emptyset) = 0$.
3. If $A_1, A_2, \ldots$ are pairwise disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

The probability function assigns a probability to each possible event in the σ-algebra.

For more on the σ-algebra, see e. g. Steland (2010, Section 2.1.3) or A somewhat condensed introduction to probability theory.

It can be seen that the definition of a probability function is only possible with respect to a sample space Ω and a suitable σ-algebra. Strictly speaking, one would always have to say to a probability function P to which Ω and $\mathcal{F}$ it belongs.Then one obtains a

Probability space

The triple $(\Omega, \mathcal{F}, \mathcal{P})$ is called **probability space**.

If no ambiguities arise, the specification of the probability space is often omitted. That is what we do here as well.

Calculation rule for probabilities

Let $A, B \in \mathcal{F}$. Then, it holds that

$$P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B) \tag{2.1}$$

2.4. Random variables

Overview

- Definition and examples of a random variable
- Realisation of a random variable
- Notations
- Discrete and continuous random variables
- Probability space of random variables

Random variables

Definition

A **real random variable** X is a **function** from a sample space Ω to $\mathbb{R}$ that assigns a number $X(\omega)$ to each elementary event $\omega \in \Omega$. For $X(\omega) \in \mathbb{R}$

$$X : \Omega \mapsto \mathbb{R} : \omega \mapsto X(\omega).$$

Each event $A \in \mathcal{F}$ can be mapped to a set $\{X(\omega) \in \mathbb{R} | \omega \in A \in \mathcal{F}\}$.

Examples:

- Students: $X(\omega = \text{female}) = 0$, $X(\omega = \text{male}) = 1$.
- Urn example: $X(\omega_1) = 0$, $X(\omega_2) = 3$, $X(\omega_3) = 17$, $X(\omega_4) = 20$.
- Household income: $X(\cdot) \geq 0$

Realisation of a random variable

Specification x of a random variable $X(\omega)$ observed in a sample such that $x = X(\omega)$.

Important: A random variable as such cannot be observed because it is a function of all possible outcomes.

Notations of random variables

- In this section we will write X instead of $X(\omega)$. Realisations or possible specifications are denoted by x.
- In the econometric literature, due to a lack of sufficient symbols, a distinction is generally not made between a random variable X and a possible realisation x, but both are denoted with the same symbol (examples: dependent variable y_t, error term u_t in the linear regression model).

Types of random variables

- **Discrete random variables**: They can take on finitely many (e. g. binary random variables) or infinitety, but countably many values (e. g. count data $\Omega = \mathbb{N}$)

$$\sum_{i=1}^{\infty} P(X(\omega) = x_i) = \sum_{i=1}^{\infty} P(X = x_i) = 1$$

- **Continuous random variables**:
  - Examples: $X \in \mathbb{R}$, $X \in [0, \infty)$.
  - Note: $P(X = x) = 0$. Why?
  - Instead, one considers probabilities for intervals, e. g. $P(X \leq x)$, $P(a < X \leq b)$, $P(0 < X) \Rightarrow$ cumulative probability distribution.

Probability space of random variables

A probability function for the random variable $X(\omega)$ on Ω can only be determined, if

1. there is a new set of elementary events Ω' corresponding to the image set of the random variable for the elementary events and
2. a new σ-algebra $\mathcal{F}'$ that can be obtained from $\mathcal{F}$.

♯ Details

The **probability function for the random variable** has as argument $A \in \mathcal{F}'$ in the case of

- discrete random variables numbers,

$$P(X = x) = P(X(\omega) = x) = P(\{\omega \in \Omega | X(\omega) = x\}).$$

- continuous random variables intervals of (real) numbers with

$$P(X \in A) = P(X(\omega) \in A) = P(\{\omega \in \Omega | X(\omega) \in A\})$$

Urn example:

- $\Omega' = \{X(\omega_1), X(\omega_2), X(\omega_3), X(\omega_4)\} = \{0, 3, 17, 20\}$
- Possible σ-algebra: $\mathcal{F}' = \{\emptyset, \{0,3\}, \{17,20\}, \Omega'\}$
- $P(X \in \{0,3\}) = 3/8$, $P(X \in \{17,20\}) = 5/8$.

Then a **new probability space** results. For $X \in \mathbb{R}$ one writes $(\mathbb{R}, \mathcal{B}, \mathcal{P}_X)$. The σ-algebra $\mathcal{B}$ is an appropriate set of all real intervals, called the Borel algebra.

However, to simplify the notation, we still often write $\{\Omega, \mathcal{F}, \mathcal{P}\}$.

2.5. Distribution and density functions

Overview

- Univariate probability distribution (cumulative distribution function (CDF))
- Multivariate distribution and density functions

2.5.1. Univariate distribution and density functions

Overview

- Univariate probability distribution (cumulative distribution function (CDF))
- Properties of distribution functions
- Probability density functions
- Interpretation of probability density function
- Standard normal distribution and normal distribution
- CDF of a binary random variable
- Support
- Quantiles and quantile functions

Univariate probability distribution

probability distribution (cumulative distribution function (CDF))

A probability function for a scalar random variable X is defined by

$$F : \mathbb{R} \mapsto [0,1] : \quad \begin{aligned} F(x) \equiv & P(X \leq x) \\ = & P(X(\omega) \in (-\infty, x]). \end{aligned} \tag{2.2}$$

Properties of distribution functions

- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$

- $F(x)$ is monotonically nondecreasing
- $P(a < X \leq b) = F(b) - F(a)$
- $F(x) = P(X \leq x) = P(X < x)$, if X is continuous.

Probability density functions

Motivation of probability densities

For a continuous random variable Y, the probability 'Y takes the value y' is just zero, i.e. $P(Y = y) = 0$. Intuition: Area under an integral at a point is zero.

Instead, one must consider an interval for Y, e.g. $[a, b]$ or frequently $(-\infty, y]$. For the latter, one obtains the probability distribution

$$F(y) = P(Y \leq y) \overset{Y \text{ continuous}}{=} P(Y < y),$$

which increases monotonically in y. Thus, one can also consider the change in probability when the interval length increases by a marginal amount $\delta > 0$. This gives the absolute change in probability

$$P(Y \leq y + \delta) - P(Y \leq y)$$

and the relative change

$$\frac{P(Y \leq y + \delta) - P(Y \leq y)}{\delta}.$$

Now, by letting the marginal change δ of the interval length go towards 0, we obtain the **probability density function**

$$f(y) = \lim_{\delta \to 0} \frac{P(Y \leq y + \delta) - P(Y \leq y)}{\delta},$$

which must be positive at some y, because otherwise there would be no change in the probability when the interval length is changed. The probability density thus indicates the **rate** at which the probability changes when the interval is marginally changed.

Since

$$P(y < Y \leq y + \delta) = P(Y \leq y + \delta) - P(Y \leq y),$$

one obtains, to put it crudely,

$$P(y < Y \leq y + \delta) \approx f(y)\delta.$$

One can therefore approximate the probability that a realisation of Y is observed in a certain interval $(y, y + \delta]$ with the product of the density and the interval length. This approximation is better the smaller δ is. **The density is approximately proportional to the probability that Y is observed in a very small interval around y.**

Probability density function (PDF)

For a continuous random variable with a differentiable probability distribution $F(x)$, the first order derivative is called **probability density function**.

$$f(x) \equiv \frac{dF(x)}{dx}, \tag{2.3}$$

$$\int_{-\infty}^{x} f(z)dz = F(x). \tag{2.4}$$

Important probability distributions

Standard normal distribution

$x \sim N(0,1)$ for $x \in \mathbb{R}$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \quad \Phi(x) = \int_{-\infty}^{x} \phi(z)dz. \tag{2.5}$$

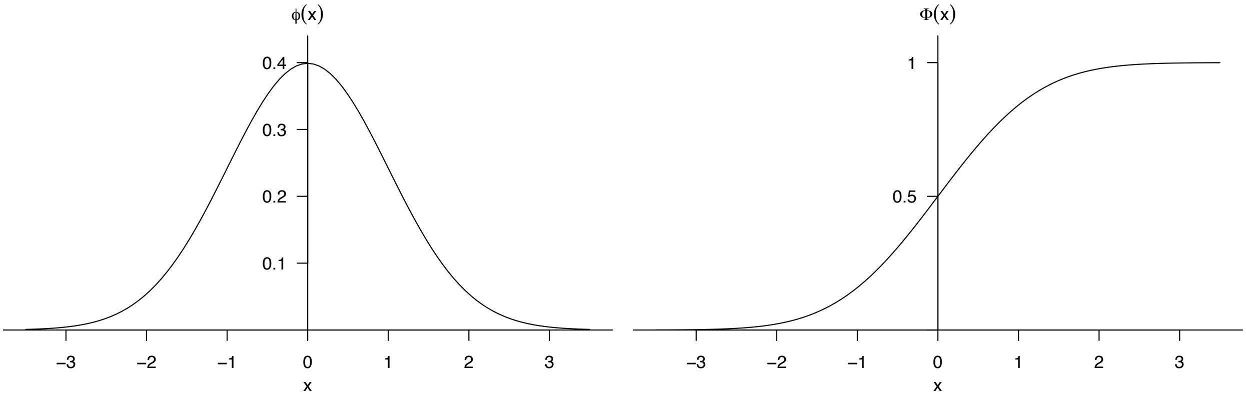


Figure 2.1.: PDF and CDF of the standard normal distribution (R program see section A.1, page 324)

Normal distribution

$x \sim N(\mu, \sigma^2)$ with density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right). \tag{2.6}$$

Note: (2.6) can be derived using change of variables (2.39).

CDF of a binary random variable

$$F(x) = \begin{cases} 0 & \text{für } x < 0 \\ p & \text{for } 0 \le x < 1 \\ 1 & \text{for } x \ge 1. \end{cases} \tag{2.7}$$

Further remarks

- CDFs can have jump discontinuities, CDFs can also be defined for random variables that are partly continuous and party discrete (e. g. in the case of censored variables).
- **Support**: Let a random variable X be given. The domain on which a density function $f_X(x)$ is positive is called **support** $\mathcal{X} \subset \mathbb{R}$ of a density function:
$$\mathcal{X} = \{x : f_X(x) > 0\}.$$
- See section 2.9 for details of important probability distributions.

  A tabular overview of many probability distributions can be found on the course homepage.

Quantiles

Quantile

The α-**quantile** q_α of a distribution for a random variable X is defined by

$$F(q_\alpha) = P(X \le q_\alpha) = \alpha. \tag{2.8}$$

Die **quantile function** is:

$$q_\alpha = F^{-1}(\alpha). \tag{2.9}$$

`R` commands

Calculating a quantile of the standard normal distribution: with `qnorm()`.

Example: The $P(X \le q_{0.85}) = 0.85$ quantile of the standard normal distribution is $q_{0.85} = 1.036433$. It is obtained with the Rcommand `qnorm(0.85)= 1.036433`. It is plotted vertically in the graphs and in red. The blue shaded area under the density is just $\alpha = 0.85$.

Important quantiles

- **Median**: $q_{0.5}$
- **Quartiles**: q_α with $\alpha = 0.25, 0.5, 0.75$
- **Quintiles**: q_α with $\alpha = 0.2, 0.4, 0.6, 0.8$

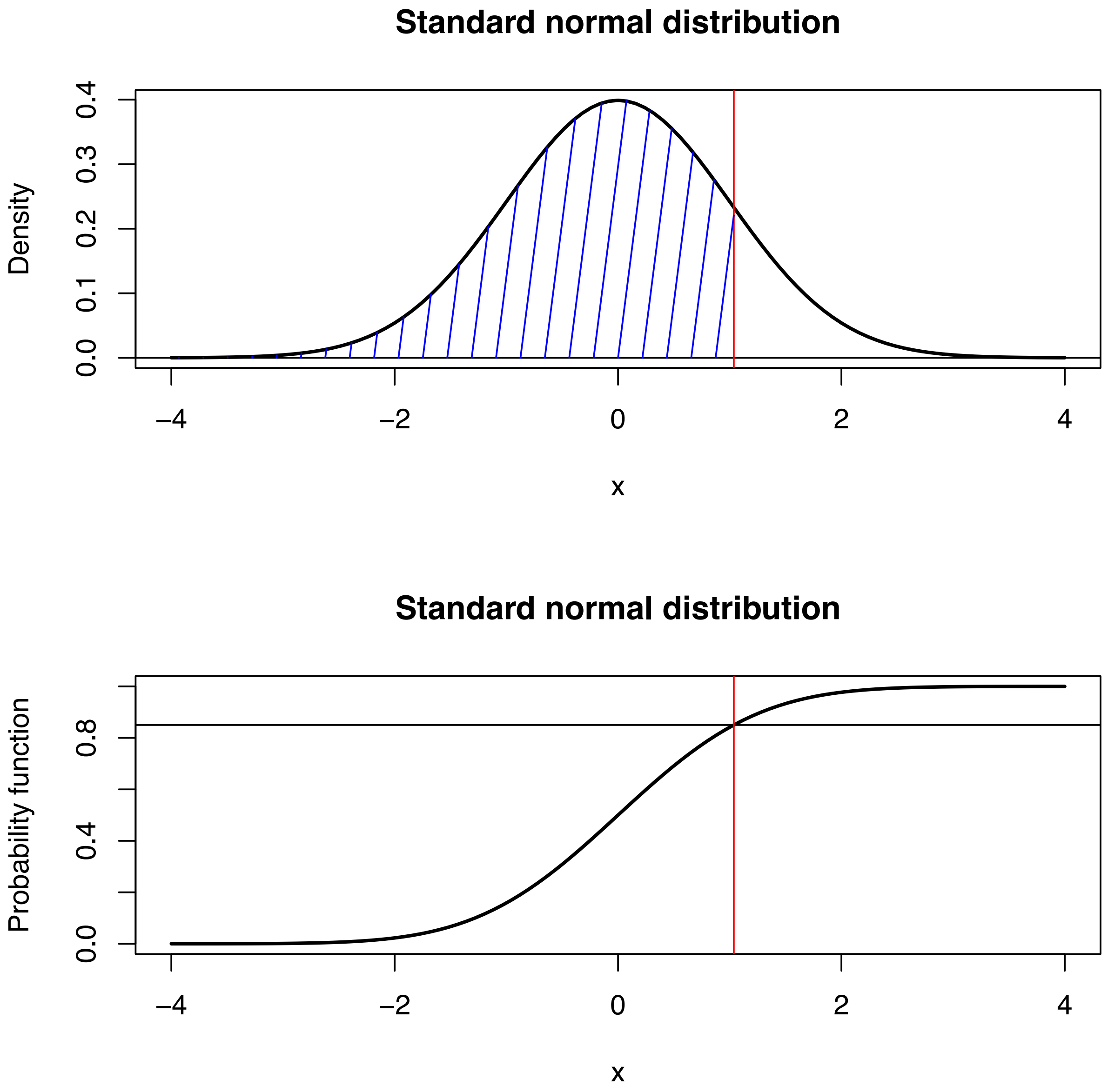


Figure 2.2.: 0.85 quantile of the standard normal distribution (R program see section A.1, page 326)

- **Deciles**: q_α with $\alpha = 0.1, 0.2, \dots, 0.8, 0.9$
- **Percentiles**: q_α with $\alpha = 0.01, 0.02, \dots, 0.98, 0.99$

2.5.2. Multivariate distribution and density functions

Overview

- Multivariate distribution and density functions
- Joint probability distribution
- Marginal probability distribution
- Joint density function

Joint probability distribution function

for two or more random variables $X_1, \dots, X_m$

$$\begin{aligned} F_{X_1,X_2,\dots,X_m}(x_1, x_2, \dots, x_m) &\equiv P\left((X_1 \leq x_1) \cap \dots \cap (X_m \leq x_m)\right) \\ &= P(X_1 \leq x_1, \dots, X_m \leq x_m). \end{aligned} \tag{2.10}$$

Marginal probability distribution

$$F_{X_i}(x_i) \equiv P(X_i \leq x_i). \tag{2.11}$$

Marginal probability density function for a continuous random variable X_i

$$f_{X_i}(x_i) \equiv \frac{dF_{X_i}(x_i)}{dx_i}. \tag{2.12}$$

Joint probability density function

for two or more continuous random variables $X_1, \dots, X_m \in \mathbb{R}$ with partially differentiable CDF:

$$f_{X_1,X_2,\dots,X_m}(x_1, x_2, \dots, x_m) \equiv \frac{\partial^m F_{X_1,X_2,\dots,X_m}(x_1, x_2, \dots, x_m)}{\partial x_1 \partial x_2 \cdots \partial x_m}, \tag{2.13}$$

$$\begin{gathered} F_{X_1,X_2,\dots,X_m}(x_1, \dots, x_m) \\ = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_m} f_{X_1,X_2,\dots,X_m}(z_1, z_2, \dots, z_m)\, dz_1 dz_2 \cdots dz_m, \\ F_{X_1}(x_1) = F_{X_1,X_2,\dots,X_m}(x_1, \infty, \dots, \infty). \end{gathered}$$

Relationship between marginal and joint densities

The following applies, e.g. in the case of three random variables

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1, z_2, z_3)\, dz_2 dz_3. \tag{2.14}$$

Notation: Davidson & MacKinnon (2004) omit the indexing of F and f. With the exception of this section, this is also done in these documents to simplify the notation if the indexing can be easily inferred from the context.

Bivariate normal distribution

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 \right.\right. \\ \left.\left. -2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\} \tag{2.15}$$

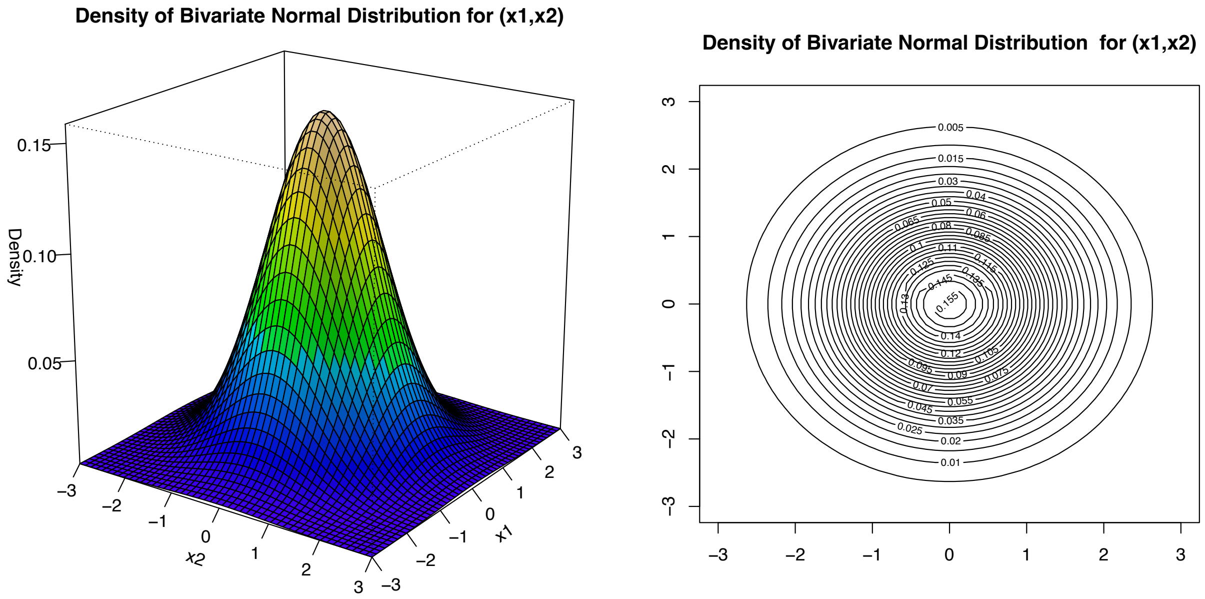


Figure 2.3.: PDF of bivariate normal distribution (R program see section A.1, page 327)

Multivariate normal distribution: see (2.31) in section 2.9.1.

2.6. Conditional probabilities

Overview

- Motivation
- Relationship with joint probability
- Basic rule for conditional probabilities
- Relationship with unconditional probabilities
- Conditioning on random variables
- Conditional probability density
- Conditional normal distribution
- Relationship between marginal and conditional density
- Stochastic independence and conditional density / distribution

- **Example of motivation**: Let the random variable $X \in [0, \infty)$ denote the payout amount in a lottery. The probability function or distribution function $P(X \leq x) = F_X(x)$ gives the probability for a maximum winning amount of x. It is further known that 2 machines are available to determine the payout amount, machine A and machine B.

  Question: What is the probability of winning the maximum amount of x *if* machine A is used?

  In other words, what is the probability we are looking for if the condition "Machine A in use" holds? Therefore, the probability we are looking for is also called **conditional probability** and we write

  $$P(X \leq x|A).$$

  Analogously, if the condition "Machine B in use" holds, one writes $P(X \leq x|B)$.

- **Relationship** with **joint probability** Let E be the event $(X \leq x)$. If someone is only happy if getting a payoff of at most x from machine B, then this person wants to determine the probability $P(E \cap B)$. This probability is given by the

Multiplication rule

For two events E, B from the collection of all possible events $\mathcal{F}$ the following applies:

$$P(E \cap B) = P(B)P(E|B), \quad P(B) > 0.$$

Knowledge about the realisation of B can help to make more precise statements about the possible realisation of E.

The multiplication rule results from the

Definition of a conditional probability

For two events E, B from the collection of all possible events $\mathcal{F}$, it holds that:

$$P(E|B) = \frac{P(E \cap B)}{P(B)}, \quad P(B) > 0.$$

Examples:

- $B \in E$: $P(E|B) = 1$; E. g. machine B always pays a minimum amount greater than zero, but a maximum less than x.
- E and B are disjoint: $P(E|B) = 0$.

Bayes' theorem

For two events E, B from the collection of all possible events $\mathcal{F}$, it holds that:

$$P(E|B) = \frac{P(B|E)P(E)}{P(B)}, \quad P(B), P(E) > 0.$$

- Relationship between the **unconditional probability** $P(X \leq x)$ and the two **conditional probabilities** $P(X \leq x|A)$ and $P(X \leq x|B)$?

  To answer this, we need to know the probability of machine A or machine B being used. If we denote these probabilities by $P(A)$ and $P(B)$, then we can answer the above question:

  Relationship between unconditional and conditional probabilities

$$\begin{aligned} P(E) &= P(E \cap A) + P(E \cap B) \\ P(X \leq x) &= P(X \leq x|A)P(A) + P(X \leq x|B)P(B) \\ F_X(x) &= F_{X|machine}(x|A)P(A) + F_{X|machine}(x|B)P(B) \end{aligned}$$

(The sample space with the elementary events for the machine choice is $\Omega = \{A, B\}$.)

Replacing the event B by the event A^c which is complementary to A, one obtains the generally applicable

Law of total probability

$$
\begin{aligned}
P(E) &= P(E \cap A) + P(E \cap A^c) \\
P(E) &= P(E|A)P(A) + P(E|A^c)P(A^c)
\end{aligned}
$$

- **Condition on random variables**: So far, we have defined the condition in the form of events rather than in the form of random variables. An example of the latter would be if only one machine is available to determine the payout amount, but its functioning depends on the previous payout amount Z. Then the conditional distribution function is $F_{X|Z}(x|Z = z)$, where $Z = z$ means that the condition is that the random variable Z takes exactly the realisation z. If Z is continuous and $Z \in [0, \infty)$, we have to replace the sum by an integral and the probability of the condition by the corresponding density function in order to get the relationship between the unconditional and the conditional probabilities, since Z can take on infinitely many values. For our example, this then results in:

$$F_X(x) = \int_0^\infty F_{X|Z}(x|Z = z) f_Z(z) dz = \int_0^\infty F_{X|Z}(x|z) f_Z(z) dz$$

or in general:

Relationship between unconditional and conditional distribution functions

$$F_X(x) = \int F_{X|Z}(x|Z = z) f_Z(z) dz = \int F_{X|Z}(x|z) f_Z(z) dz \tag{2.16}$$

Conditional probability distribution function

for random variable X_1 given one random variable X_2 or several random variables $X_2, \dots, X_m$:

$$
\begin{aligned}
f_{X_1|X_2}(x_1|x_2) &\equiv \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}, && (2.17)\\
&\text{provided that } f_{X_2}(x_2) > 0, \\
f_{X_1|X_2,\dots,X_m}(x_1|x_2, \dots, x_m) &\equiv \frac{f_{X_1,\dots,X_m}(x_1, x_2, \dots, x_m)}{f_{X_2,\dots,X_m}(x_2, \dots, x_m)}, && (2.18)\\
&\text{provided that} \\
&f_{X_2,\dots,X_m}(x_2, \dots, x_m) > 0.
\end{aligned}
$$

Conditional normal distribution:

Let $\mu(X) = E[Y|X]$ and $\sigma^2(X) = Var(Y|X)$. Then the following notations are equivalent:

$$
\begin{aligned}
Y|X &\sim N(\mu(x), \sigma^2(x)) \\
f_{Y|X}(y|x) &= \frac{1}{\sqrt{2\pi\sigma^2(x)}} \exp\left(-\frac{1}{2}\frac{(y - \mu(x))^2}{\sigma^2(x)}\right) && (2.19)
\end{aligned}
$$

Important properties:

Calculating the marginal density from the conditional density

$$f_X(x) = \int f_{X|Z}(x|Z = z) f_Z(z) dz = \int f_{X|Z}(x|z) f_Z(z) dz. \tag{2.20}$$

Stochastic independence

If

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1,X_2}(x_1, \infty) F_{X_1,X_2}(\infty, x_2) = P(X_1 \leq x_1)\, P(X_2 \leq x_2) \tag{2.21}$$

holds, the random variables X_1 and X_2 are called **stochastically independent** or **independent** and it holds that

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2). \tag{2.22}$$

Corresponding factorisations hold for more than two random variables. If the random numbers X_1 and X_2 are stochastically independent, then it holds that

$$F_{X_1|X_2}(x_1|x_2) = F_{X_1}(x_1), \tag{2.23a}$$
$$f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1). \tag{2.23b}$$

2.7. Expected values and moments

Overview

- Definitions and rules
- Inequalities for expected values
- Second order moments: variance, covariance, correlation
- Rules
- Higher moments: uncentred and centred moments
- Skewness, kurtosis

Expected values, or first moments

- **Expected value of a discrete random variable X with finitely** many possible realisations x_i, $m < \infty$,
$$E[X] = \sum_{i=1}^{m} x_i P(X = x_i)$$

- **Expected value of a discrete random variable X with infinitely** many realisations x_i
$$E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i)$$
Note: This expected value only exists if
$$\sum_{i=1}^{\infty} |x_i| P(X = x_i) < \infty.$$

- **Expected value of a continuous random variable $X \in \mathbb{R}$**
$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$
Note: This expected value only exists if
$$\int_{-\infty}^{\infty} |x| \, f(x) dx < \infty.$$

- **Expected value of a continuous random variable X** on support $\mathcal{X} = (a, b) \subset \mathbb{R}$
$$E[X] = \int_{a}^{b} x f(x) dx$$
This expected value always exists provided $f(x) < \infty$ for $x \in \mathcal{X}$.

Rules for the expected value

z. B. Wooldridge (2009, Appendix B)

1. For any constant c it holds that
$$E[c] = c.$$

2. For all constants a and b and random variables X and Y
$$E[aX + bY] = aE[X] + bE[Y]$$
holds.

3. - If the random variables X and Y are stochastically independent, it holds that
$$E[XY] = E[X]E[Y]$$

- or more generally: If the random variables X and Y are stochastically independent and it holds for all functions $f(x)$ and $g(y)$ that $E\left[|f(X)|\right] < \infty$ and $E\left[|g(Y)|\right] < \infty$, then it holds that
$$E[f(X)g(Y)] = E[f(X)]E[g(Y)].$$

Inequalities for expected values

1. $E\left[|X+Y|\right] \leq E\left[|X|\right] + E\left[|Y|\right]$
2. **Jensen inequality**: If $g(x)$ is convex, then $E[g(X)] \geq g\left(E[X]\right)$ holds. The inequality sign holds strictly if $g(x)$ is strictly convex. If $g(x)$ is concave, the inequality sign reverses.

Second order moments

Variance, covariance, correlation

$$\begin{aligned} Var(X) &= E\left[(X-E[X])^2\right] = \int_{-\infty}^{\infty} (x-E[X])^2 f(x)dx, \\ Cov(X,Y) &= E\left[(X-E[X])(Y-E[Y])\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-E[X])(y-E[Y]) f_{X,Y}(x,y)dxdy, \\ Corr(X,Y) &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \end{aligned} \tag{2.24}$$

Rules

- $Var(X) = E[(X-E[X])^2] = E[X^2] - E[X]^2$ ("Verschiebungssatz"),
- $Var(a+bX) = Var(bX) = b^2 Var(X)$,
- $Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E[X\,Y] - E[X]\,E[Y]$
- $Cov(aX,bY) = ab\; Cov(X,Y)$,

Higher moments

- **second (uncentred) moment**: $m_2(X) = \int_{-\infty}^{\infty} x^2 f(x)dx$
- Let $\mu = E[X] = m_1(X)$ and $\sigma = \sqrt{Var(X)} = \sqrt{\tilde{m}_2(X)}$.
- **k-th (uncentred) moment**:
$$m_k(X) = E\left[X^k\right] = \int_{-\infty}^{\infty} x^k f(x)dx$$

- ***k*-th centred moment**:

$$\tilde{m}_k(X) = E\left[(X - E(X))^k\right] = \int_{-\infty}^{\infty} (x - m_1(X))^k f(x)dx$$

- **Skewness** (third centred moment)

$$\frac{E\left[(X - E[X])^3\right]}{\sigma^3} = \frac{\int_{-\infty}^{\infty} (x-\mu)^3 f(x)dx}{\sigma^3}.$$

- **Kurtosis**

$$\frac{E\left[(X - E[X])^4\right]}{\sigma^4} = \frac{\int_{-\infty}^{\infty} (x-\mu)^4 f(x)dx}{\sigma^4}.$$

  Examples:

  - The skewness of symmetrical densities is 0.
  - The kurtosis of a standard normally distributed random variable is 3.

2.8. Conditional expected values and moments

Overview

- Definitions and rules
- Law of iterated expectations
- Rules for conditional expectations
- Rules
- Rules for conditional variances and covariances

Conditional expected value

- **Definition**: So far, we have not paid attention to which machine is used in the payout determination. However, if we are interested in the expected payout when machine A is in use, then we need to calculate the **conditional expected value**

$$E[X|A] = \int_0^{\infty} xf(x|A)dx.$$

This is done simply by replacing the unconditional density $f(x)$ by the conditional density $f(x|A)$ and specifying the condition in the notation of the expected value. Accordingly the expected payout for machine B can be calculated as

$$E[X|B] = \int_0^\infty xf(x|B)dx.$$

If it is not yet "realised" which $M = A, B$ is in use, the **conditional expected value**

$$E[X|M] = \int_0^\infty xf(x|M)dx = g(M)$$

is a **function** with argument M. Thus the conditional expected value is a **random variable**. This holds in general.

Depending on whether the condition or X are continuous or discrete, the **calculation of the conditional expected value** differs slightly

X	$=$	continuous	discrete	Condition
$E[X\|A]$	$=$	$\int xf(x\|A)dx$	$\sum x_i P(X = x_i\|A)$	discrete
$E[X\|Z = z]$	$=$	$\int xf(x\|Z = z)dx$	$\sum x_i P(X = x_i\|z)$	continuous

Note: Often the short forms are used, as in Wooldridge (2009), e. g.

$$E[X|z] = \int xf(x|z)dx.$$

- **Law of iterated expectations (LIE)**: Corresponding to the relationship between unconditional and conditional probabilities, a similar relationship also exists between the unconditional and the conditional expected values. It is

  $$E[X] = E\left[E(X|Z)\right] = E\left[g(Z)\right], \quad g(Z) = E(X|Z)$$

  and is called **law of iterated expectations**.

  Proof sketch:

  $$\begin{aligned} E[X] =& \int xf(x)dx \\ =& \int x \left[\int f(x|z)f(z)dz \right] dx \quad \text{(Substituting (2.20))} \\ =& \int \int xf(x|z)f(z)dzdx \\ =& \int \underbrace{\int xf(x|z)dx}_{E[X|z]} f(z)dz \quad \text{(Interchanging } dx \text{ and } dz\text{)} \\ =& \int E[X|z]f(z)dz \\ =& E\left[E(X|Z)\right]. \end{aligned}$$

In our **example** with the 2 machines the law of iterated expectations yields

$$\begin{aligned} E[X] &= E[X|A]P(A) + E[X|B]P(B), \\ E[X] &= g(A)P(A) + g(B)P(B). \end{aligned} \tag{2.25}$$

This example makes it clear once again that the conditional expected values $E[X|A]$ and $E[X|B]$ are random numbers that, weighted with their probabilities of occurrence $P(A)$ and $P(B)$, yield the expected value $E[X]$. Imagine that before the game starts you only know the two conditional expected values, but not which machine will be used. Then the expected payout is just $E[X]$ and we have to consider the two conditional expected values as random variables. Once we know which machine has been used, the corresponding conditional expected value is the realisation of the random variable.

Rules for conditional expectations

(e. g. Wooldridge (2009, Appendix B))

1. For each function $c(\cdot)$ it holds that

$$E[c(X)|X] = c(X).$$

2. For all functions $a(\cdot)$ and $b(\cdot)$ it holds that

$$E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X).$$

3. If the random variables X and Y are independent, it holds that

$$E[Y|X] = E[Y].$$

4. **Law of iterated expectations** (**LIE**)

$$\begin{aligned} E[Y] &= E[E(Y|X)] \\ E[Y|X] &= E[E(Y|X,Z)|X] \end{aligned}$$

5. If $E[Y^2] < \infty$, $E[g(X)^2] < \infty$ for an arbitrary function $g(\cdot)$, then:

$$\begin{aligned} E\left\{[Y - E(Y|X)]^2 |X\right\} &\leq E\left\{[Y - g(X)]^2 |X\right\} \\ E\left\{[Y - E(Y|X)]^2\right\} &\leq E\left\{[Y - g(X)]^2\right\}. \end{aligned}$$

Rules for conditional variances and covariances

"Verschiebungssatz" for (co-)variances, etc.

$$\begin{aligned}
Var(Y|X) &= E[(Y - E[Y|X])^2|X] = E[Y^2|X] - E[Y|X]^2, \qquad (2.26)\\
Cov(Y, X|Z) &= E[(Y - E[Y|Z])(X - E[X|Z])|Z]\\
&= E[Y\,X|Z] - E[Y|Z]\,E[X|Z], \qquad (2.27)\\
Var(Y) &= E\left[Var(Y|X)\right] + Var\left(E[Y|X]\right). \qquad (2.28)
\end{aligned}$$

A proof for (2.28) can be found for the multivariate variant (9.6).

Relationships between conditional expectations and covariances

It holds for two random variables Y and X:

$$\begin{aligned}
E[Y|X] = E[Y] \quad &\Longrightarrow \quad Cov(Y, X) = 0, && (2.29a)\\
E[Y|X] = 0 \quad &\Longrightarrow \quad E[Y] = 0 \text{ and } Cov(Y, X) = 0, && (2.29b)\\
Cov(Y, X) \neq 0 \quad &\Longrightarrow \quad E[Y|X] \neq 0, && (2.29c)\\
Cov(Y, X) = 0 \;\&\; E[Y] = 0 \quad &\Longrightarrow \quad E[Y X] = E[XE(Y|X)] = 0, && (2.29d)\\
E[Y] = 0 \quad &\not\Longrightarrow \quad E[Y|X] = 0, && (2.29e)\\
Cov(Y, X) = 0 \quad &\not\Longrightarrow \quad E[Y|X] = 0. && (2.29f)
\end{aligned}$$

Example: For $Y = X^2$ and $E(X) = E(X^3) = 0$ it holds that $Cov(Y, X) = 0$, since $Cov(X^2, X) = E[X^3] - E[X^2]E[X] = 0$, but $E[Y|X] = X^2 \neq 0$.

Evidence via

$$\begin{aligned}
Cov(Y, X) &= E[Y X] - E[Y]E[X] = E\left[E[Y X|X]\right] - E\left[E[Y|X]\right]E[X]\\
&= E\left[XE[Y|X]\right] - E\left[E[Y|X]\right]E[X]
\end{aligned}$$

- (2.29b): If $E[Y|X] = 0$, it must follow that $Cov(Y, X) = 0$.
- (2.29c): If this statement were false and $E[Y|X] = 0$ followed, $Cov(Y, X) = 0$ follows from (2.29b), which leads to a contradiction.
- (2.29f): From $Cov(Y, X) = 0$ only $E\left[wE[Y|X]\right] = E\left[E[Y|X]\right]E[X]$ follows, but not $E[Y|X] = 0$.

2.9. Important probability distributions

Overview

- Normal distribution
- χ^2, t-, F-distribution
- $\sharp$ Change of variables

2.9.1. Normal distribution

Overview

- Standard normal distribution
- Normal distribution
- Multivariate standard normal distribution
- Multivariate normal distribution
- Bivariate normal distribution
- Linear combinations of multivariate normally distributed random vectors
- IID and NID

- **Standard normal distribution**: $x \sim N(0,1)$ with density function (2.5)

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right). \tag{2.5}$$

- **Normal distribution**: $x \sim N(\mu, \sigma^2)$ with density

$$f(x) = \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right). \tag{2.6}$$

  Note: (2.6) can be derived using change of variables in the one-dimensional case (2.39).

- **Multivariate standard normal distribution**: $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_n)$ with density

$$\phi(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right). \tag{2.30}$$

  Note that this representation is equivalent to (cf. (2.22))

$$\phi(\mathbf{z}) = \phi(z_1)\phi(z_2)\cdots\phi(z_n).$$

  A multivariate standard normally distributed random vector $\mathbf{z}$ is thus composed of independently and identically distributed (more precisely standard normally distributed) random variables $z_1, \dots, z_n$. Conversely: n i.i.d. standard normally distributed random numbers can be written as a multivariate standard normally distributed random vector. Note: This does not work without the i.i.d. requirement!

- **Multivariate normal distribution**:

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \quad \text{where } \boldsymbol{\Omega} = \mathbf{A}\mathbf{A}^T \tag{2.31}$$

and for the $(r \times n)$ matrix $\mathbf{A}$, $r \leq n$, $rk(\mathbf{A}) = r$ holds. Density function:

$$f(x_1, x_2, \dots, x_r) = f(\mathbf{x}) = \frac{1}{(2\pi)^{r/2}} \left(\det(\boldsymbol{\Omega})\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^T \boldsymbol{\Omega}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right). \tag{2.32}$$

- **Bivariate normal distribution** (2.15). See section 2.5 for plot.

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

- **Linear combinations of multivariate normally distributed random vectors**

$$\begin{aligned} \text{For} \qquad & \mathbf{w} = \mathbf{b} + \mathbf{B}\mathbf{x} \quad \text{with } \mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ it holds that :} \\ & \mathbf{w} \sim N\left(\mathbf{b} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T\right). \end{aligned} \tag{2.33}$$

- **Notation**

  - The random variables v_t, $t = 1, \dots, n$ are **independently and identically distributed (IID)**:

$$v_t \sim IID(E[v_t], Var(v_t)).$$

  - The random variables v_t, $t = 1, \dots, n$ are **independently and identically normally distributed (NID)**:

$$v_t \sim NID(E(v_t), Var(v_t)).$$

In matrix notation and using $\mu_v = E[v_t]$, $\sigma_v^2 = Var(v_t)$, this corresponds to

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_v \\ \mu_v \\ \vdots \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & 0 & \cdots & 0 \\ 0 & \sigma_v^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_v^2 \end{pmatrix}\right),$$

$$\mathbf{v} \sim N(\mu_v \boldsymbol{\iota}, \sigma_v^2 \mathbf{I}).$$

Cf. for the definition of $\boldsymbol{\iota}$ (7.8).

2.9.2. χ^2-, t-, F-distribution

Overview

- χ^2-distribution
- Student's t-distribution
- F-distribution

χ^2-distribution

- If $z_1, \dots, z_m$ are i.i.d. standard normally distributed, $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_m)$, then the sum of the squared random variables
$$y = \sum_{i=1}^{m} z_i^2 = \mathbf{z}^T\mathbf{z} = ||\mathbf{z}||^2$$
is χ^2**-distributed with** m **degrees of freedom**. In short form:
$$y \sim \chi^2(m).$$

- **Expected value**: $E(y) = m$,
$$\text{since} \quad E\left(\sum_{i=1}^{m} z_i^2\right) = \sum_{i=1}^{m} E(z_i^2) = m.$$

- **Variance**: $Var(y) = 2m$, since
$$\begin{aligned} E\left[(y-m)^2\right] &\overset{\text{Independence}}{=} mVar(z_i^2) \\ &= mE\left[\left(z_i^2 - 1\right)^2\right] \\ &= m\left(E[z_i^4] - 2 + 1\right) \\ &= 2m. \end{aligned}$$

- If $y_1 = \sum_{i=1}^{m_1} z_i^2 \sim \chi^2(m_1)$ and $y_2 = \sum_{i=m_1+1}^{m} z_i^2 \sim \chi^2(m_2)$, $m = m_1 + m_2$, are independent, it holds that
$$y = y_1 + y_2 \sim \chi^2(m).$$

- If $\mathbf{x}$ is a multivariate normal distributed $(m \times 1)$ vector with nonsingular covariance matrix $\mathbf{\Omega}$, $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Omega})$, then
$$y = \mathbf{x}^T\mathbf{\Omega}^{-1}\mathbf{x} \sim \chi^2(m). \tag{2.34}$$

Proof: Since $\mathbf{\Omega}$ is regular, there exists a decomposition $\mathbf{\Omega} = \mathbf{A}\mathbf{A}^T$ such that $\mathbf{z} = \mathbf{A}^{-1}\mathbf{x}$ has covariance matrix $\mathbf{I}$. Then $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ and

$$E\left[\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^T\left(\mathbf{A}^{-1}\right)^T\right] = \mathbf{A}^{-1}\mathbf{\Omega}\left(\mathbf{A}^{-1}\right)^T = \mathbf{A}^{-1}\mathbf{A}\mathbf{A}^T\left(\mathbf{A}^T\right)^{-1} = \mathbf{I}.$$

- If $\mathbf{P}$ is a projection matrix with $\operatorname{rk}\mathbf{P} = r < m$ and $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, it holds that

$$\mathbf{z}^T\mathbf{P}\mathbf{z} \sim \chi^2(r). \tag{2.35}$$

Proof: Assume $\mathbf{P}$ projects onto the r linearly independent columns of the $(m \times r)$ matrix $\mathbf{Z}$. Then $\mathbf{P} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ and one obtains

$$\mathbf{z}^T\mathbf{P}\mathbf{z} = \underbrace{\mathbf{z}^T\mathbf{Z}}_{\mathbf{w}^T} \underbrace{\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}}_{\text{inverse covariance matrix with rank } r} \underbrace{\mathbf{Z}^T\mathbf{z}}_{\mathbf{w}}.$$

Since for the $(r \times 1)$ vector $\mathbf{w} \sim N\left(\mathbf{0}, \mathbf{Z}^T\mathbf{Z}\right)$ holds, it holds that

$$\mathbf{w}^T\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{w} \sim \chi^2(r)$$

because of (2.34).

- For $m \to \infty$ it holds that a $\chi^2(m)$-distributed random variable converges in distribution to a normally distributed random variable $N(m, 2m)$.

Student's t-distribution

- Given a standard normally distributed random variable $z \sim N(0,1)$ and a χ^2-distributed random variable $y \sim \chi^2(m)$ with m degrees of freedom, where z and y are stochastically independent. Then the random variable

$$t = \frac{z}{(y/m)^{1/2}} \sim t(m) \tag{2.36}$$

is ***t*-distributed with m degrees of freedom**.

- The density of the t-distribution is symmetrical and bell-shaped.

- All moments of the t-distribution exist up to the $m-1$ moment. The t-distribution with $m = 1$ is also called **Cauchy distribution**. Note that neither expected value nor variance exist because too much mass of the distribution concentrates in the tails.

- **Expected value**: For $m > 1$: $E(t) = 0$, **Variance**: For $m > 2$: $Var(t) = m/(m-2)$.

- The t-distribution approaches the standard normal distribution with increasing number of degrees of freedom. One can argue asymptotically here: With $m \to \infty$ it holds that $\operatorname{plim}_{m\to\infty} y/m = 1$, since y is a sum of m squared independent standard normal distributed random variables. Thus, using Slutsky's theorem, $\operatorname{plim}_{m\to\infty}(y/m)^{1/2} = 1$ also holds and thus

$$\operatorname*{plim}_{m\to\infty} \frac{z}{(y/m)^{1/2}} = z \sim N(0,1).$$

F-distribution

- Given two stochastically independent χ^2-distributed random variables $y_1 \sim \chi^2(m_1)$ and $y_2 \sim \chi^2(m_2)$. Then the random variable
$$F = \frac{y_1/m_1}{y_2/m_2} \sim F(m_1, m_2) \tag{2.37}$$
follows a **F-distribution with m_1 and m_2 degrees of freedom**.
- For $m_2 \to \infty$, the random variable $m_1 F$ approaches a $\chi^2(m_1)$-distribution, since $\operatorname{plim}_{m_2\to\infty} y_2/m_2 = 1$. If $t \sim t(m_2)$, it holds that $t^2 \sim F(1, m_2)$.

2.9.3. Supplement: Change of variables

Change of variables

- $\sharp$ **Change of variables in the one-dimensional case**: Given is a continuous random variable $X \in \mathbb{R}$ with density function $f_X(x) > 0$.

  Let there also be given a random variable $Y = g(X)$, where the function $g(\cdot)$ is continuous and invertible, such that
$$x = g^{-1}(y). \tag{2.38}$$
  Furthermore, let $g(\cdot)$ and $g^{-1}(\cdot)$ be differentiable once.

  Then for the random variable Y, the density function $f_Y(y)$ can be calculated through
$$f_Y(y) = \left|\frac{d}{dy} g^{-1}(y)\right| f_X\left(g^{-1}(y)\right) \tag{2.39}$$
  (Casella & Berger 2002, Theorem 2.1.5).
- $\sharp$ **Change of variables in the multi-dimensional case**: Given a continouos $(m \times 1)$ random vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^m$ with density function $f_{\mathbf{x}}(\mathbf{x}) > 0$. Further, let an $(m \times 1)$ random vector
$$\mathbf{y} = \mathbf{g}(\mathbf{x}) = \mathbf{a} + \mathbf{A}\mathbf{x} \tag{2.40}$$
  be given.

  If $\mathbf{A}$ is invertible (see Casella & Berger (2002, Section 4.6, p. 185) for conditions for the case that $\mathbf{g}(\mathbf{x})$ in (2.40) is nonlinear), it holds that
$$\mathbf{x} = \mathbf{h}(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{a})$$
  and (see section 6.2.2)
$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}^T} = \frac{\partial \mathbf{h}(\mathbf{y})}{\partial \mathbf{y}^T} = \mathbf{A}^{-1}.$$

Then, for the random vector $\mathbf{y}$, the density function $f_{\mathbf{y}}(\mathbf{y})$ can be calculated through

$$f_{\mathbf{y}}(\mathbf{y}) = \left|\frac{\partial \mathbf{h}(\mathbf{y})}{\partial \mathbf{y}^T}\right| f_{\mathbf{x}}\left(\mathbf{h}(\mathbf{y})\right) = \left|\mathbf{A}^{-1}\right| f_{\mathbf{x}}\left(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{a})\right), \tag{2.41}$$

where $\left|\frac{\partial \mathbf{h}(\mathbf{y})}{\partial \mathbf{y}^T}\right|$ denotes the determinant of the Jacobian matrix $\frac{\partial \mathbf{h}(\mathbf{y})}{\partial \mathbf{y}^T}$, see (1.10) for more details. (Davidson 2000, Theorem B.9.2)

3. Convergence and limits

Convergences occur in many areas of mathematics without it being clear in applied use that the constructs are limit processes. Consider the function $f(x) = x^2$. The derivative function is $f'(x) = 2x$, simply "‘derived’" by algebraic formulae. But the actual process of taking the derivative would look like this:

Example: Derivative of a simple function $f(x) = x^2$:

$$\begin{aligned} f'(x) :&= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h\to 0} \frac{(x+h)^2 - x^2}{h} \\ &= \lim_{h\to 0} \frac{2xh + h^2}{h} = \lim_{h\to 0} \frac{2xh}{h} + \lim_{h\to 0} \frac{h^2}{h} = 2x + 0 = 2x \end{aligned}$$

No matter whether derivative, integral, continuity or sequences of functions, limit values occur in very different forms. In order to understand the "‘new"’ forms of convergence that are important for econometrics, we will briefly review the standard cases.

In mathematics and probability theory, there are several types of convergence, of which we will discuss the following:

Overview: types of convergence

1. Convergence of sequences of numbers - basic framework of all theories of convergence.
2. Convergence of sequences of functions
3. Convergence of sequences of random variables
   - Almost sure convergence
   - Convergence in probability
   - Convergence in distribution

For econometrics, 3) with its forms is particularly relevant. To understand them, one must understand 1) and 2)

3.1. Convergence of sequences

Let (a_n) be a sequence of real numbers, i. e., a mapping $f : \mathbb{N} \to \mathbb{R}$ with $n \mapsto a_n \in \mathbb{R}$ (imagine $a_n := \frac{1}{n}$). Instead of $f(n) = a_n$ one also often writes $(a_n) = (a_n)_{n\in\mathbb{N}} = \{a_1, a_2, a_3, ...\}$ for the set of members of the sequence.

Examples:

$f(n) = (a_n)_{n\in\mathbb{N}} = (\frac{1}{n})_{n\in\mathbb{N}} = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$.

$f(n) = (a_n)_{n\in\mathbb{N}} = (n^2)_{n\in\mathbb{N}} = \{1, 4, 9, 16, \ldots\}$.

$f(n) = a_n = \frac{f(x+1/n)-f(x)}{1/n} \overset{h:=1/n}{=} \frac{f(x+h)-f(x)}{h}$ für fixed x.

Convergent sequence

A sequence (a_n) is called convergent in $\mathbb{R}$, if there exists a number $a \in \mathbb{R}$, with the following property

- For all $\epsilon > 0$, there is a member of the sequence indexed by $N \in \mathbb{N}$ (more precisely $N(\epsilon)$, since it depends on ϵ), such that $\mid a_n - a \mid< \epsilon$ holds for all subsequent members of the sequence $n > N$.

The number a is called **limit** of the sequence and one writes

$$\lim_{n\to\infty} a_n = a \text{ or } a_n \to a \text{ for } n \to \infty$$

Note: In the case of convergence, the limit of the sequence is uniquely determined!

Example: $f(n) := (a_n) := \frac{50}{n} \cdot \mid \sin(0.1 \cdot n) \mid$. The members of the sequence are plotted as circles in Figure 3.1. One now thinks that the limit is $a = a_\infty = 0$.
To show this, one must prove for all $\epsilon > 0$ that there is a member of the sequence $N(\epsilon)$ from which all further members of the sequence are within the ϵ-distance from a.
If we now imagine a $\epsilon_1 = 1$, we see that the members of the sequence 26 to 43 already have a distance of less than 1 from a_∞, but the members of the sequence 44 to 50 are outside this range. One now chooses a $N(\epsilon_1) = 51$ (or a higher member of the sequence) and hopes that all subsequent ones lie within this distance of the limit and could continue this with further, smaller ϵ_i. However, this does not show the convergence of the sequence! The definition clearly says "‘**for all**"’ $\epsilon > 0$!
A proof would look as follows:
Since one has to show that it holds for all ϵ, one gives oneself a small $\epsilon > 0$, which can become arbitrarily small, but is fixed for a moment. Now one has to show

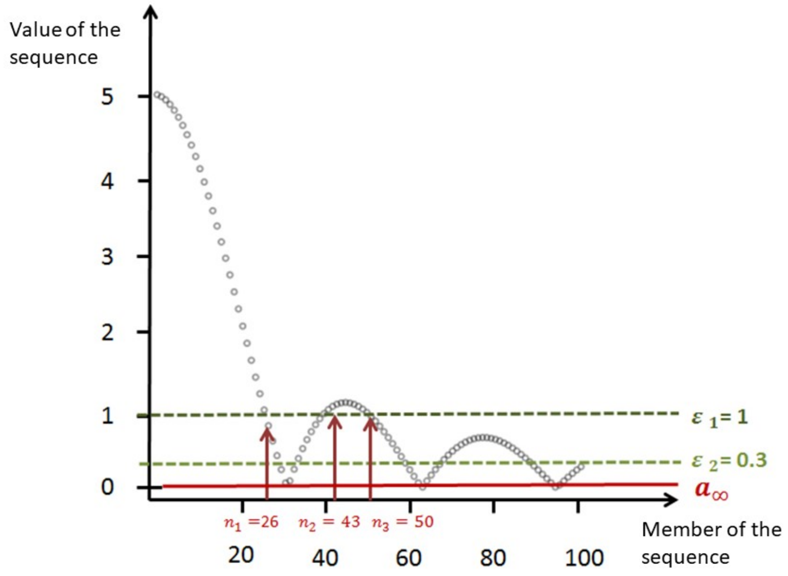


Figure 3.1.: The first 100 members of the sequence of $\frac{50}{n} \cdot \mid \sin(0.1 \cdot n) \mid$

that there is a member of the sequence $N(\epsilon)$ depending on the ϵ just given, from which all further members of the sequence a_n with $n > N(\epsilon)$ lie in the ϵ-distance of $a_\infty = 0$. So the task is to find this $N(\epsilon)$ such that it holds that:

$$\mid a_{N(\epsilon)} - 0 \mid < \epsilon$$

Rearranging the condition gives the desired result:

$$\mid a_N \mid < \epsilon \Longleftrightarrow \frac{50}{N} \cdot \underbrace{\mid \sin(0.1 \cdot N) \mid}_{\leq 1} < \epsilon \Longleftrightarrow \frac{50}{N} < \epsilon \Longleftrightarrow N > \frac{50}{\epsilon}$$

Now, given a ϵ, we know that all members of the sequence with index greater than $\frac{50}{\epsilon}$ lie in the ϵ region around the limit.
A proof would now look like this:
Let $\epsilon > 0$ be arbitrarily small but fixed.
Then choose the index of sequence $N(\epsilon)$ such that on the one hand $N(\epsilon) > \frac{50}{\epsilon}$ and $N(\epsilon) \in \mathbb{N}$.
Thus, for the following members of the sequence with index $n \geq N(\epsilon)$ it holds that:

$$\begin{aligned} \mid a_n - a \mid &\overset{a=0}{=} \mid a_n - 0 \mid \overset{\text{Def.}}{=} \mid \frac{50}{n} \cdot \sin(0.1 \cdot n) \mid = \mid \frac{50}{n} \mid \cdot \mid \sin(0.1 \cdot n) \mid \\ &\overset{\mid \sin(0.1 \cdot n) \mid \leq 1}{\leq} \mid \frac{50}{n} \mid \overset{n \geq N(\epsilon)}{\leq} \mid \frac{50}{N(\epsilon)} \mid \overset{N(\epsilon) > \frac{50}{\epsilon}}{<} \mid \frac{50}{\frac{50}{\epsilon}} \mid = \mid \epsilon \mid \overset{\epsilon \geq 0}{=} \epsilon \end{aligned}$$

Altogether one has thus shown:
For arbitrary $\epsilon > 0$ there is an $N(\epsilon) \in \mathbb{N}$, such that for all subsequent members of

the sequence with index $n \geq N(\epsilon)$ it holds that:

$$\mid a_n - a \mid < \epsilon$$

Calculation rules of convergent sequences of numbers x_n, y_n:

- If $\lim_{n\to\infty}\ x_n = x$ and $\lim_{n\to\infty}\ y_n = y$, then it also follows: $\lim_{n\to\infty}\ x_n + \lim_{n\to\infty}\ y_n = \lim_{n\to\infty}\ (x_n + y_n)$.
- If $\lim_{n\to\infty}\ x_n = x$ and $\lim_{n\to\infty}\ y_n = y$, then it also follows: $\lim_{n\to\infty}\ x_n \cdot \lim_{n\to\infty}\ y_n = \lim_{n\to\infty}\ (x_n \cdot y_n)$.
- If $\lim_{n\to\infty}\ x_n = x$ and $\lim_{n\to\infty}\ y_n = y$ and $y \neq 0$, then: $\lim_{n\to\infty}\ x_n / \lim_{n\to\infty}\ y_n = \lim_{n\to\infty}\ (x_n/y_n)$.

Note: The inversions do not generally apply : Since $0 = (-1)^n + (-1)^{n+1}$, it does not follow from $\lim_{n\to\infty}\ 0 = 0$ that the limits $\lim_{n\to\infty}\ (-1)^n$ and $\lim_{n\to\infty}\ (-1)^{n+1}$ exist, which is clearly not the case.

3.2. Convergence of functions

While previously defining a scalar sequence, one now goes a step further and considers sequences or families of functions $f_n(x)$ (think of $f_n(x) = x + \frac{1}{n}$).

For every fixed $n \in \mathbb{N}$, the expression $f_n(x)$ is a function $f_n(x) : \mathbb{D} \to \mathbb{R}$, where $\mathbb{D}$ is the domain of definition of the function, and for every fixed $x_0 \in \mathbb{D}$, the expression $f_n(x_0)$ is a sequence $\mathbb{N} \to \mathbb{R}$. One now defines the (**pointwise** - as opposed to the **uniform** -) convergence of a sequence of functions as the limit of the sequence $f_n(x_0)$ for fixed x_0:

Pointwise convergence

A sequence of functions $(f_n(x))$ is called pointwise convergent if for every fixed point $x_0 \in \mathbb{D}$ the sequence of numbers $(f_n(x_0))_{n\in\mathbb{N}}$ converges.

By

$$f(x) := \lim_{n\to\infty}\ f_n(x)$$

then a function $f : \mathbb{D} \to \mathbb{R}$ is defined. Properties of f_n such as continuity, integrability, differentiability, i.e. properties that also require limit values, do not transfer **in general!** (hence the term uniform convergence).

Example: $f_n(x) := x^n$ with domain of definition $\mathbb{D} = [0,1]$. Some of the elements of the family of sequences can be seen in figure 3.2. Since for all $n \in \mathbb{N}$ and a number $x \in (0,1)$ just $x^n < x$ holds, but $1^n = 1$, it is clear that the limit of the sequence of functions $f_n(x)$ drawn in red looks like this

$$f(x) = \begin{cases} 0 & \text{falls } x < 1 \\ 1 & \text{falls } x = 1 \end{cases}$$

This is proved again with the $\epsilon - N$-definition for sequences of numbers from above, doing this for "‘each"’ $x \in \mathbb{D} = [0,1]$ individually. But here it reduces to a case differentiation of $x \in [0,1)$ and $x = 1$. What is striking here is that each member of the sequence $f_n(x)$ represents a continuous function, while the limit has a point of discontinuity at point 1.

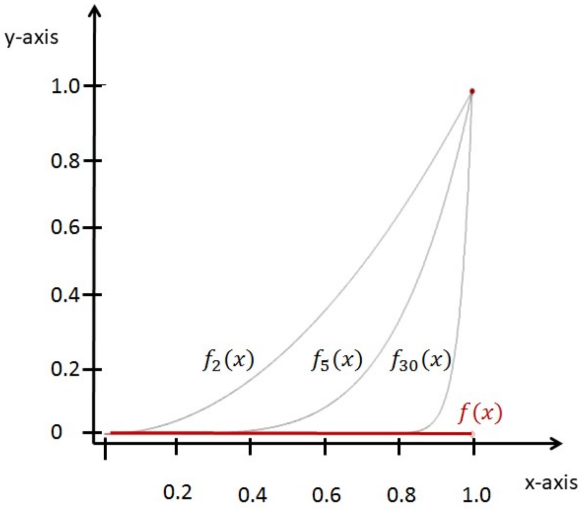


Figure 3.2.: Three members of the sequence of the family of functions $f_n(x) = x^n$

3.3. Almost sure convergence

If one considers random variables as functions to $\mathbb{R}$ for which there is a probability function for evaluation, one can form a special pointwise convergence for random variables, the almost sure convergence. Let $X_n(\omega)$ be a sequence of random variables. $X_n(\omega)$ converges almost surely to $X(w)$, in symbols $X_n \xrightarrow{a.s.} X$ (almost surely), if the probability of the set of outcomes ω at which $\lim_{n\to\infty} X_n$ and X differ is 0. In formal notation:

$$
\begin{aligned}
X_n \xrightarrow{a.s.} X \quad &\Longleftrightarrow \quad P\left(\lim_{n\to\infty} X_n = X\right) = P\left(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1 \\
&\Longleftrightarrow \quad P\left(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}\right) = 0
\end{aligned}
$$

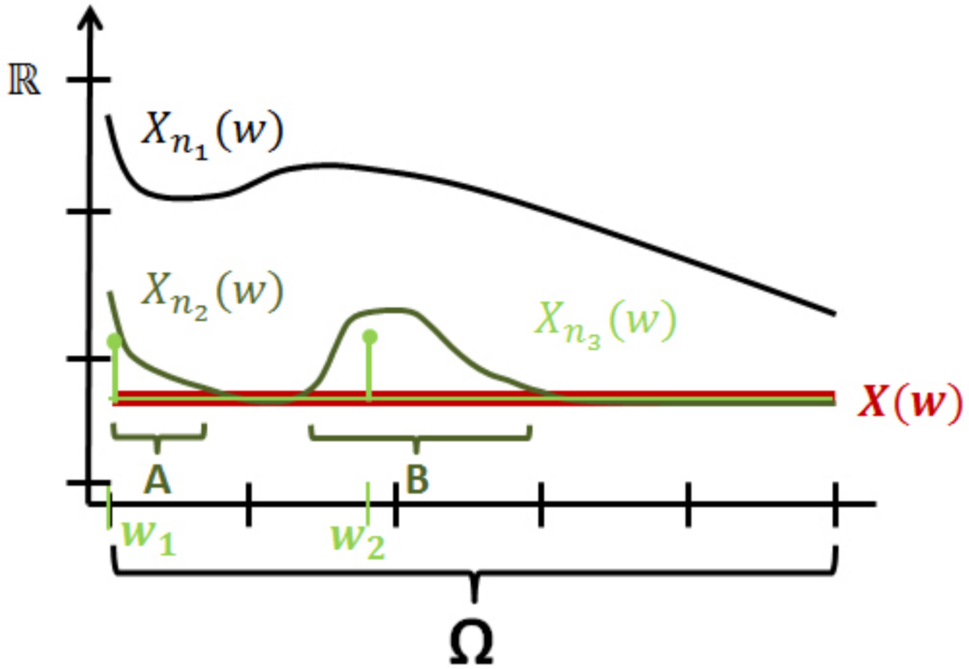


Figure 3.3.: Three members of the sequence of random variables with $n_1 < n_2 < n_3$

While pointwise convergence required that for every $x \in \mathbb{R}$ $\lim_{n\to\infty} f_n(x) = f(x)$ must hold, probability gives us the possibility to require that the limit of the sequence and the limit differ "‘by probability 0"’. Thus, one first lets the sequence tend to infinity and then considers the probability at which the two differ. If we look at the member of the sequence $X_{n_1}(\omega)$ in the picture on the right, it differs from $X(\omega)$ in every ω, and since $P(\Omega) = 1$ this is not yet an approximation. If we take a subsequent member of the sequence $X_{n_2}(\omega)$, $X_{n_2}(\omega)$ and $X(\omega)$ differ only in $A \cup B$ with probability $P(A \cup B) = p < 1$ and $X_{n_3}(\omega)$ and $X(\omega)$ differ only in

$\{\omega_1, \omega_2\}$.
Supposing that X is a continuous random variable, thus Ω has uncountably many elements, then the probability is $P(\{\omega_1, \omega_2\}) = P(\{\omega_1\}) + P(\{\omega_2\}) = 0$ and (the limit) $X_{n_3}(\omega)$ and $X(\omega)$ differ in a set $\{\omega \in \Omega \mid X_{n_3}(\omega) \neq X(\omega)\}$ with probability 0. We say that they are almost everywhere the same or that the so sketched sequence $X_n(\omega)$ almost surely converges to $X(\omega)$.

Calculation rules of almost surely convergent sequences of random variables X_n, Y_n:

- If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then it also follows that: $X_n + Y_n \xrightarrow{a.s.} X + Y$.
- If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then it also follows that: $X_n \cdot Y_n \xrightarrow{a.s.} X \cdot Y$.

Remark: Addition and multiplication preserve almost sure convergence.

3.4. Convergence in probability

2nd idea: instead of considering the likelihood of the limit, consider the limit of the likelihood of the differences.

Convergence in probability

Let $X_n(\omega)$ be a sequence of random variables. X_n converges in probability to X, in symbols $X_n \xrightarrow{P} X$ (in probability), if for all $\epsilon > 0$ it holds that

$$\begin{aligned} &\lim_{n\to\infty} p_n := \lim_{n\to\infty} P\left(|X_n - X|_{\mathbb{R}} > \epsilon\right) = 0 \\ \iff &\lim_{n\to\infty} P\left(\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)|_{\mathbb{R}} > \epsilon\right) = 0 \end{aligned}$$

Whereas previously one first formed the limit of the sequence of random variables and then evaluated the differences to the limit with the probability, one now proceeds differently. For each member of the sequence $X_n(\omega)$ and for each ϵ one can calculate the probability $p_n := P\left(\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| > \epsilon\}\right)$ and thus get a sequence of numbers p_n on $[0, 1]$. If this sequence of numbers goes towards 0, X_n lies with probability 1 in a ϵ neighbourhood around X and X_n is said to converge in probability towards X, in symbols $X_n \xrightarrow{P} X$ or $\operatorname{plim}_{n\to\infty} X_n = X$. Expressed formally:
A sequence of random variables X_n converges in probability to X if for all $\epsilon > 0$ it holds that

$$\lim_{n\to\infty} p_n := \lim_{n\to\infty} P\left(|X_n - X|_{\mathbb{R}} > \epsilon\right) = \lim_{n\to\infty} P\left(\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)|_{\mathbb{R}} > \epsilon\right) = 0$$

If we consider figure 3.4 with two members of the sequence and two different ϵ, this no longer makes a statement about the identity of two random variables. It only makes a statement

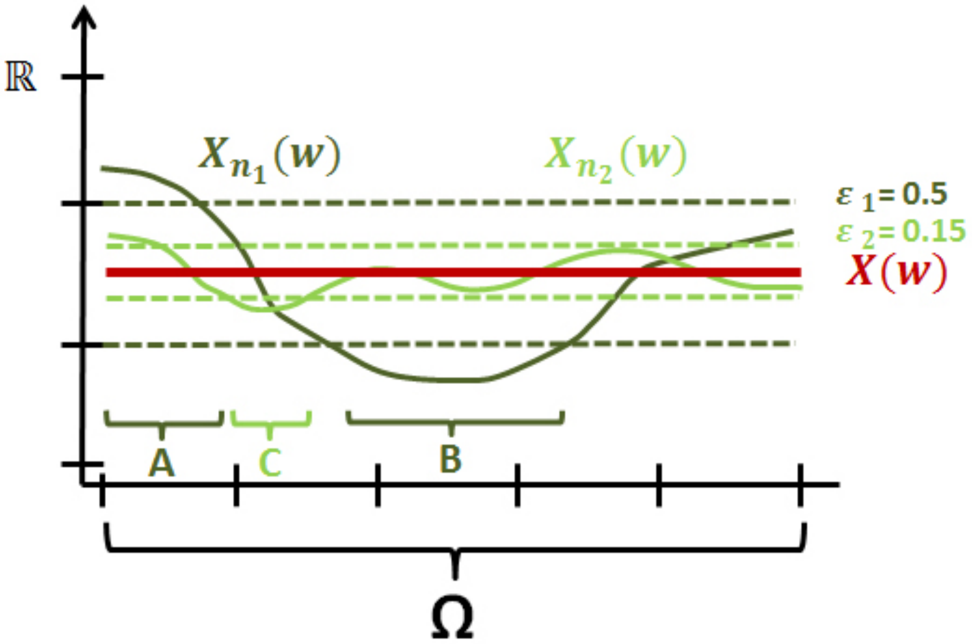


Figure 3.4.: Two members of the sequence of random variables with $n_1 < n_2$

that if you want to know it more precisely (make ϵ smaller), you will find a member of the sequence such that the probability of outcomes ω where the distance is greater than ϵ is 0.
In the example $X_{n_1}(\omega)$ and ϵ_1 it is just the set $A \cup B$ in which $X_{n_1}(\omega)$ differed from $X(\omega)$ by more than ϵ_1. One could now use $X_{n_2}(\omega)$ with the same ϵ_1; since $X_{n_2}(\omega)$ is in the ϵ_1 neighbourhood, the probability of outcomes outside is 0.
If one now narrows the neighbourhood to ϵ_2, $X_{n_2}(\omega)$ and $X(\omega)$ differ only in C. If the set C has probability 0, one would not have to find a new index, if the probability is not 0, one would again find a subsequent member of the sequence, from which the set of "‘outliers"’ has probability 0.

Calculation rules of random variable sequences X_n, Y_n converging in probability:

- If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then it also follows that: $X_n + Y_n \xrightarrow{P} X + Y$.
- If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then it also follows that: $X_n \cdot Y_n \xrightarrow{P} X \cdot Y$.

Remark: Addition and multiplication preserve convergence in probability.

Convergence in probability for random vectors

Let $\mathbf{y}^n$ denote a $(n \times 1)$ random vector whose dimension varies with n.

A vector function $\boldsymbol{a}_n : \mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{a}_n := \boldsymbol{a}(\mathbf{y}^n)$ converges in probability to $\boldsymbol{a}_0 \in \mathbb{R}^m$ if

$$\lim_{n\to\infty} P\left(||\boldsymbol{a}(\mathbf{y}^n) - \boldsymbol{a}_0||_{\mathbb{R}^m} < \epsilon\right) = 1.$$

Remark: $||\cdot||_{\mathbb{R}^m} : \mathbb{R}^m \to \mathbb{R}$ (cf. definition from above).

Example: $\hat{\mu} : \mathbb{R}^n \to \mathbb{R} : \hat{\mu}(\mathbf{y}^n) = \frac{1}{n}\sum_{t=1}^n y_t$. So here $m = 1$. For $\hat{\boldsymbol{\beta}}$, the OLS estimator with k regressors, $m = k$.

Rules of calculation for random vectors converging in probability

Let $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ be sequences of random vectors. If $\operatorname{plim}_{n\to\infty} \mathbf{x}_n$, and $\operatorname{plim}_{n\to\infty} \mathbf{y}_n$, then it holds that:

$$\operatorname*{plim}_{n\to\infty} (\mathbf{x}_n \pm \mathbf{y}_n) = \operatorname*{plim}_{n\to\infty} \mathbf{x}_n \pm \operatorname*{plim}_{n\to\infty} \mathbf{y}_n, \tag{3.1a}$$

$$\operatorname*{plim}_{n\to\infty} (\mathbf{x}_n^T \mathbf{y}_n) = (\operatorname*{plim}_{n\to\infty} \mathbf{x}_n)^T (\operatorname*{plim}_{n\to\infty} \mathbf{y}_n), \tag{3.1b}$$

These rules also apply if the random vectors are replaced by matrices with random variables with corresponding properties.

3.5. Convergence in distribution

3rd idea: Instead of considering the random variables as functions, consider only the distribution functions of the random variables.

Convergence in distribution

Let $X_n(\omega)$ be a sequence of random variables. X_n converges in distribution to X, in symbols $X_n \xrightarrow{d} X$ (in distribution), if for the sequence of functions of the distributions F_n of X_n and the distribution F of X it holds that

$$\lim_{n\to\infty} F_n = F \text{ pointwise}$$

Example: Let $\{X_n\}$ be the sequence of random variables defined above. Recall that: $X_n \xrightarrow{P} X$, where $X \sim N(\mu, \sigma^2)$. Let Z be a normally distributed random variable with expected value μ and variance σ^2. Then it holds that $X_n \xrightarrow{d} Z$. Thus X and Z have the same distribution, but are different random variables!

Relation between the concepts of convergence

$$X_n \xrightarrow{ptw.} X \Longrightarrow X_n \xrightarrow{a.s} X \Longrightarrow X_n \xrightarrow{P} X \Longrightarrow X_n \xrightarrow{d} X. \tag{3.2}$$

(For an example of why the inversion of the third sequence arrow does not hold, see the

BA Time Series Econometrics module, section 5.1.4., for an example of the inversion of the second sequence arrow, see the sliding hills example.)

Continuous Mapping Theorem (CMT)

- Let $h(\cdot)$ be a continuous function.

$$\text{If } X_n \xrightarrow{a.s.} X, \text{ then } h(X_n) \xrightarrow{a.s.} h(X) \text{ holds.}$$
$$\text{If } X_n \xrightarrow{p} X, \text{ then } h(X_n) \xrightarrow{p} h(X) \text{ holds.}$$
$$\text{If } X_n \xrightarrow{d} X, \text{ then } h(X_n) \xrightarrow{d} h(X) \text{ holds.}$$

  Remark: Continuous transformations preserve the convergence concepts.

- Accordingly, for sequences of $(k \times 1)$ random vectors $\mathbf{x}_n$ it holds that: Given a continuous vector-valued function $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^m$, then for $\star \in \{a.s., p, d\}$ it holds that:

$$\text{If } \mathbf{x}_n \xrightarrow{\star} \mathbf{x}, \text{ then } \mathbf{h}(\mathbf{x}_n) \xrightarrow{\star} \mathbf{h}(\mathbf{x}) \text{ holds.} \tag{3.3}$$

  (Cf. e. g. Vaart (1998, Theorem 3.2).)

For clarification, consider the function

$$f(x) = \begin{cases} 0 & x < 2 \\ 1 & x \geq 2 \end{cases}$$

and the sequence of numbers $a_n = 2 - \frac{1}{n}$. Then obviously $\lim_{n\to\infty} a_n = 2$ is monotonic from the left. We now quickly see that the limit of the function values is not equal to the function value of the limit:

$$\lim_{n\to\infty} f(a_n) \overset{a_n \leq 2 \forall n}{=} \lim_{n\to\infty} 0 = 0 \neq 1 = f(2) = f(\lim_{n\to\infty} a_n)$$

So far it has not been investigated in which way convergence in distribution is preserved, very famous is the following theorem:

Slutzky's theorem

Let $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{c}$ with $\mathbf{c} \in \mathbb{R}^p$ constant. Then it holds that

$$\mathbf{y}_n^T \mathbf{x}_n \xrightarrow{d} \mathbf{c}^T \mathbf{x}. \tag{3.4}$$

Vgl. Vaart (1998, Theorem 3.6).

The theorem also holds if $\mathbf{y}$ and $\mathbf{c}$ are replaced by appropriately dimensioned matrices.

Note: The term Slutzky's theorem is not used consistently in the statistical and econometric literature. In his theorem 3.1.3, Davidson (2000) denotes the statement (3.3) for scalars and convergence in probability as Slutzky's theorem.

Conclusion:

- Convergence terms that measure different "‘distances"' properties of (random) variables (and their interrelationships within and among them).
- For econometric statements, generally only the two "‘last"' ones, convergence in probability and in distribution, are interesting
- For a better understanding of econometric theory, one also needs knowledge of the other convergence terms.

3.6. Example of sliding hills: convergence in probability $\not\Rightarrow$ Almost sure convergence

Considering initially for simplicity the doubly indexed random variable sequence X_{nk} on $\Omega = [0,1]$, where $n, k \in N$, $k \leq n$ and $I_{[a,b]}(\omega) := \begin{cases} 1 & \text{if } \omega \in [a,b] \\ 0 & \text{if } \omega \notin [a,b] \end{cases}$,

$$\begin{array}{lll} X_{11} = I_{[0,1]} & & \\ X_{21} = I_{[0,1/2]} & X_{22} = I_{[1/2,1]} & \\ X_{31} = I_{[0,1/3]} & X_{32} = I_{[1/3,2/3]} & X_{33} = I_{[2/3,1]} \\ \dots & & \end{array}$$

one quickly recognises their structure. The graphs of these random variables are "‘sliding hills’" which become narrower and narrower as n increases (n is the number of intervals in which one subdivides the unit interval, k is the interval in which one is located). According to the lexicographic order, one can generate a random variable Y_n from X_{nk}:

$$Y_1 = X_{11}, Y_2 = X_{21}, Y_3 = X_{22}, Y_4 = X_{31}, Y_5 = X_{32}, \dots$$

Thus one recognises:

- Y_n converges in probability to 0,

  because for an arbitrary $\epsilon \in (0,1)$ it holds that $P\left(|X_{nk} - 0| > \epsilon\right) = P\left(X_{nk} > \epsilon\right) \overbrace{=}^{n \text{ equal intervals}} \frac{1}{n} \xrightarrow{n\to\infty} 0$

  (in words: the probability of being on a hill goes towards 0)

- Y_n does not converge almost surely to 0, because the random variable sequence Y_n does not converge pointwise at any point (if it does not converge pointwise anywhere, then trivially it does not converge almost surely either, since in this case the set of all outliers have probability 0):

To be shown: for all $\omega \in \Omega$ there exists $\epsilon > 0$ such that for all $N \in \mathbb{N}$ there exists an $n > N$ with $|Y_n - 0| = Y_n > \epsilon$.
Let $\omega \in \Omega$ be fixed but arbitrary, $\epsilon = 0.5$, $N \in \mathbb{N}$ be fixed but arbitrary and without restriction $Y_N(\omega) = 0$. Let $Y_N = X_{n'k'}$ and without restriction $\omega = \frac{k^\star}{n^\star} \notin [k'/n', (k'+1)/n']$ (if irrational, use a rational approximation; if $k^\star = 1$ or $k^\star = n^\star$, trivial). Then ω lies in any interval $[\omega - 1/r, \omega + 1/r] = [\frac{rk^\star - n^\star}{rn^\star}, \frac{rk^\star + n^\star}{rn^\star}]$ for arbitrary $r \in \mathbb{N}$. Now choose r by Archimedean property such that on the one hand $rn^\star > n'$ and on the other hand $[\omega - 1/r, \omega + 1/r] \subset [0, 1]$. Define $\tilde{k} = rk^\star - n^\star \in \mathbb{N}$ (because of "‘on the other hand"’), then $Y_{\hat{n}}(\omega) := X_{rn^\star, \tilde{k}}(\omega) = 1 > \epsilon = 0.5$ for $\hat{n} > N$.

See, for example, Casella & Berger (2002, Example 5.5.8, p. 234-5).

Part II.

Econometric Methods

4. Introduction

Aim of this course:

To learn the basic econometric theory and practice relevant to the careful empirical analysis of economic problems.

Examples of economic questions requiring empirical analysis:

- Does a reduction in class size (school, university) lead to better learning outcomes? (Cf. Stock & Watson 2007, Section 1.1.)
- Effect of training measures by the Federal Employment Office: Does this increase the remaining lifetime income? Does it reduce the duration of unemployment?
- Will there be higher inflation in the coming years?
- What factors influence country-specific imports into Germany?
- Understanding dynamic processes
- What factors influence the length and intensity of business cycles?
- What factors influence economic growth, income and wealth distribution?

In the examples given, it is often a matter of determining causal variables.

4.1. Statements about causal relationships

The knowledge of causal relationships is a **prerequisite for the evaluation** of planned or implemented (economic policy, operational, etc.) measures.

Causality

- Common understanding: "causality means that a specific action leads to a specific, measurable consequence (Stock & Watson 2007, p. 8)
- Considered precisely: The effect of an action is **always** unknown in the individual case,

because

- if individual/unit/variable i is affected by an action/measure, one can only observe the outcome with the action for i, but not the outcome if the action had not been carried out.
- Alternatively, if individual/unit/variable i is not affected by this action, one only knows the result without action, but not the result for this i if it had been affected by the action.

The case that did not occur in each case is the **counterfactual state** and would provide the answer to a **“what if¿‘ question**.

In the language of econometrics: **the individual success** of a measure is **always unobservable** because it always contains a counterfactual.

- Under certain conditions, however, the **average effect of an action** on a group of individuals can be measured.
- **Definition of causality**: In the following, we refer to **an action or measure as causal** if a **causal effect** of an action can be measured.

  A “**causal effect** is defined to be an effect on an outcome of a given action or treatment, as measured in an ideal randomized controlled experiment (Stock & Watson 2007, p. 9)”.
- **Ceteris paribus**: If all other causal variables except the variable of interest are held constant and only the action of interest is performed, one considers the outcome of the action **ceteris paribus (c. p.)**.
- Note: For the existence of a causal effect, it is a prerequisite that an action also has an effect on single individuals.

Measurability of causal relationships

A quantification of the average effect of an action, i.e. the quantification of a **causal relationship** on the basis of an econometric model is only possible if

1. an **ideal controlled random experiment** can be performed or if
2. a sample from a **quasi-experiment** is available and specific **identification assumptions** are made or if
3. the econometric model was derived from a causally interpretable **(economic) model** that is a useful approximation of reality for the research question.

In this course we only consider case 3

Cases 1. and 2. are described in more detail in sections 2.2 and 2.3 in **Advanced Issues in**

Econometrics and are dealt with in detail in the master's course **Impact Evaluation Methods**. They play a major role in **evaluation research**.

Controlled random experiment

A total group of individuals is divided into a **treatment group** and a **control group**. The latter contains all individuals who do not participate in a measure. The central feature of a controlled random experiment is that the participants of the **treatment group** are selected **randomly**.

Example: class size At the beginning of a school year, students (and teachers, etc.) of a school are randomly divided into small and large classes. This avoids that students with certain characteristics are predominantly found in one class size.

Quasi-Experiment / Natural Experiment

Often it is not possible to conduct a controlled random experiment for legal, ethical, economic or other reasons.

Under certain **additional** assumptions on the population, sample observations obtained from them can be treated as if a controlled random experiment were present.

Note: For many economic questions, e. g. macroeconomic questions, neither controlled random experiments are feasible nor can natural experiments be observed. Then the causal interpretation of an econometric model can only be done on the basis of an economic model underlying the econometric model.

Simultaneity and Causality

If one observes two variables y and x in the same time period, e. g. for the year 2014, then it is possible that

1. both variables show a **simultaneous relationship**, i. e. influence each other **simultaneously**,
$$x \longleftrightarrow y$$
or
2. one variable is **causal** for the other, e. g. x for y,
$$x \longrightarrow y$$
or

3. both variables are influenced by a third variable but do not influence each other

$$z \longrightarrow \begin{cases} x \\ y \end{cases} , x \not\longrightarrow y, y \not\longrightarrow x$$

   or

4. no influence exists

$$x \not\longleftrightarrow y.$$

Note:

- In principle, in an empirical analysis with several potential variables in one time period, one must assume that simultaneity, i. e. case 1, is present. Only a theoretically or statistically justified exclusion of a direction of an effect makes it possible to exclude simultaneity and to obtain case 2, precisely a causal relationship. Or to obtain case 4, that there is no causal relationship.
- Whether case 2 exists can be tested statistically under certain conditions (see course **Impact Evaluation Methods**).
- A statistical check of case 4 can be done with methods of model selection, see chapter 10.

Example: factors influencing imports

Objective/scientific question:

Identify the factors that influence imports to Germany and quantify their influence.

First considerations: Which variables of a time period could be relevant and which directions of effect could exist between them?

The $(m \times 1)$ vector $\mathbf{s}_t \in \mathbb{R}^m$ contains all variables that could be relevant for the analysis. The following are examples and incomplete (!):

- Human capital of the exporting country (s_1)
- Colonial past of both countries (s_2)
- Gross domestic product of the importing country (s_3)
- GDP of the exporting country (s_4)
- Distance to exporting country (s_5)
- Area of exporting country (s_6)
- Openness in country (s_7)
- Imports (s_8)
- unspecified (s_9)

Figure 4.1 shows these variables and initially assumes **simultaneity** for all pairs of variables.

To reduce the number of simultaneous relations, the following assumption seems clearly justified:

Assumption: Area and distance are not influenced by other variables.

Figure 4.2 is obtained.

A further reduction of simultaneous relationships is best done by an economic model, which however requires further assumptions. Section 6.3 presents an economic model that allows all remaining simultaneous relationships between imports (s_8) and the other variables to be either eliminated entirely or transformed into a causal direction of effect. Figure 4.3 is obtained in which, compared to figure 4.2, the number and type of relationships between imports and the other variables has changed, but the number and type of relationships between the potential impact variables has not.

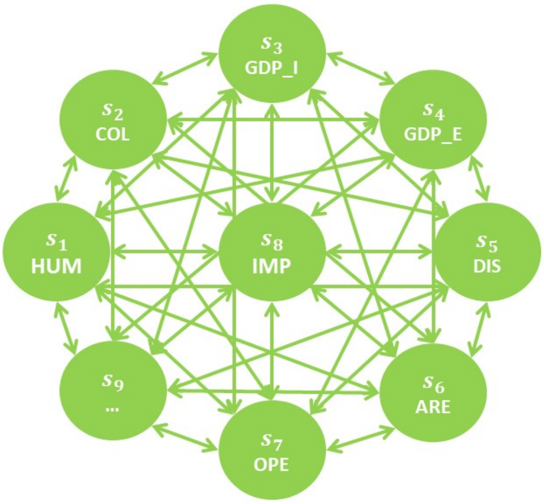


Figure 4.1.: Factors influencing trade flows: possible simultaneous relationships

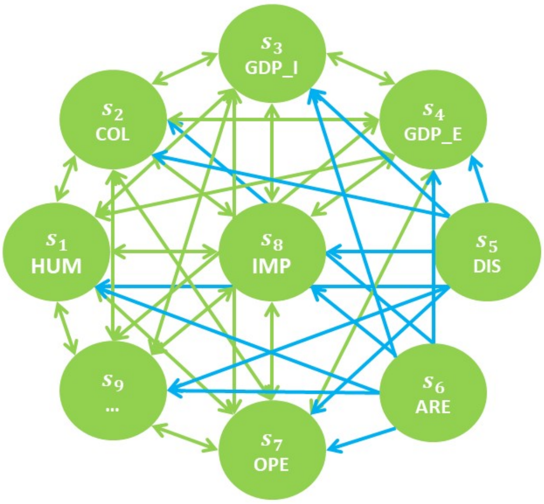


Figure 4.2.: Factors influencing trade flows: first reduction in simultaneous relationships: blue arrows indicate causal relationships

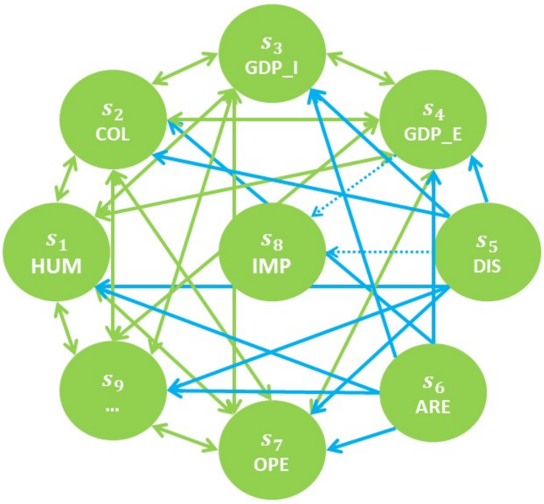


Figure 4.3.: Factors influencing trade flows: causal and simultaneous relationships based on an economic model plus other relevant influencing factors; dashed arrows represent the relevant causal relationships (later model 2).

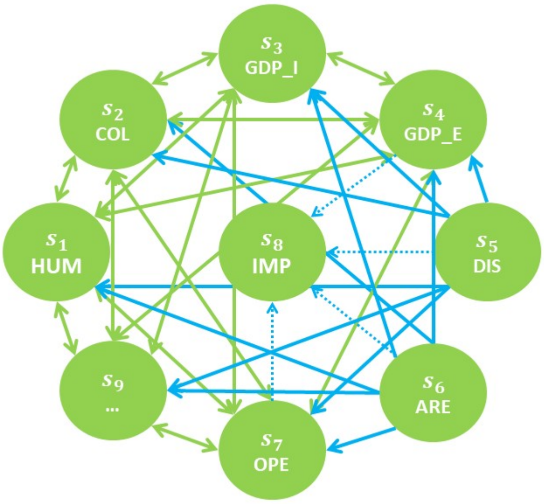


Figure 4.4.: Factors influencing trade flows: causal and simultaneous relationships based on an economic model plus other relevant influencing factors; dashed arrows represent the relevant causal relationships (later model 4).

To answer the initial question, the remaining causal relationships in figure 4.3 must be quantified. In the simplest case, this is done with a multiple linear regression model.

In reality, the assumed economic model probably represents an oversimplification, so that further influencing factors must be appropriately taken into account. An example is shown in figure 4.4. How this is done efficiently is also subject of this course.

Multiple linear regression model

- In order to quantify the (average) effect of an action, in the above example a change in economic power in the exporting country on imports, it is generally not sufficient to consider only these two variables in a model. Instead, all relevant causal variables must be considered in the model.

- If one has well justified that the variables $z_1, \ldots, z_{k-1}$ causally influence the variable y and one is interested in quantifying the causal effect of the variable z_1 on the variable y and assumes a linear relationship (in the parameters $\beta_1, \ldots, \beta_k$), then an example of a **multiple linear regression model** results:

$$y = \beta_1 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u. \tag{4.1}$$

  - The variable u is called **error term**, which contains all non-modellable / non-modelled influences. A possibility to interpret u more precisely is given by (5.29).
  - The variable y is called **endogenous variable** because it is / should be explained by the model.
  - The variables z_j postulated as causal for y are called **exogenous variables**.

  In the following chapters, the properties and assumptions concerning these variables will be specified in more detail.

- A quantification of the causal effect of z_1 on y is done by determining β_2. This requires

  1. a sample with suitable data and
  2. a suitable econometric estimation procedure.

  Example: Factors influencing imports From figure 4.3 the model with $y = s_8, z_1 = s_4, z_2 = s_5$ follows.

$$Imports = \beta_1 + \beta_2 GDP + \beta_3 Distance + u \tag{4.2}$$

- In chapters 12 and 13, dynamic models are also considered. Here it is always assumed that lagged endogenous variables, i.e. variables that lie in prior periods of the period under consideration, are causal. To enable a general notation, the variables on the right-hand side are typically denoted by x_j.

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{4.3}$$

The **multiple linear regression model** (4.3) is a central tool for all the aims of empirical analyses mentioned below and is therefore the focus of this course.

- The variables y and x have different names in the literature.

Names for variables in regression models

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Controll variable
Outcome variable	Predictor variable
Regressand	Regressor
	Covariate

Simultaneous equation models

- In some analyses, there is a simultaneous relationship between the y of interest and other variables s_i. Then there are several endogenous variables that must be modelled explicitly or indirectly together.
- A simple example of a simultaneous equation model can be found in section 5.2, equations (5.8) and (5.9).
- The estimation of simultaneous equilibrium models is covered in the master module **Advanced Econometrics**.

Aims of empirical analyses

1. Identification and quantification of causal relationships
2. Falsification of postulated economic relationships
3. Point and interval forecasting
4. Analysis and evaluation of implemented/planned measures (economical, operational, etc.)
5. Assessment of uncertainty and risk

4.2. What is Econometrics?

Econometrics

- offers solutions to deal with unobserved factors in economic models,
- offers "both a numerical answer to the question and a measure how precise the answer is (Stock & Watson 2007, p. 7)",
- offers tools for disproving economic hypotheses by confronting theories with empirically collected data using statistical methods, *and* offers tools for quantifying the probabilities that such decisions are wrong, (see among others chapter 11)
- allows the **quantification** of the **risks** of predictions, decisions and even their own analysis, (see among others section 9.3 and following)
- allows the **quantification** of **causal relationships** arising from an economic model.

In general:

- Quantitative answers always involve **uncertainty**. Uncertainty exists regarding:
  - the "true" (data generating) mechanism,
  - the choice of variables in the empirical analysis,
  - the measurement of the variables,
  - the choice of the econometric model,
  - the statistical quality of the estimation or forecasting procedure.
- For the quantification of **uncertainty** the toolbox of **probability theory** is very useful, but not only for this ....

4.3. Components of an empirical analysis

An empirical analysis should follow a structured procedure, which will be justified throughout the course. It is structured as follows:

I. **Economic analysis part**

  1. **Scientific issue**
     - Careful formulation of the question of interest or problem.
  2. **Economic model**

- Specification of an economic model.
- Identify causal and simultaneous relationships.
- Obtaining hypotheses which are to be tested empirically.
- Interpretation of model parameters.

3. **Data availability**
   - Which data are required with regard to the economic model and are already available or can be acquired?

II. **Econometric modelling process**

1. **Selecting a class of econometric models**
   - Consider variables from the economic model and their availability.
   - Consider the functional relationships from the economic model or their approximation.
   - Consider whether data generating mechanism (DGP) could be included in model class.
   - If applicable, formulate statistically testable hypotheses regarding the DGP.
   - Choose estimation methods with favourable estimation properties: Which estimation method is suitable and as efficient as possible, i.e. makes the best possible use of the sample information? What are the properties of the chosen estimation method?
2. **Procuring data: Collecting a sample**
   - Characterisation of the sample survey.
3. **Specify, estimate and select one or more econometric models**:
   - Use appropriate estimation methods.
   - Use appropriate model selection methods.
4. **Reviewing the selected models**
   - Is the selected model correctly specified? If yes, no relevant explanatory variables are missing, the functional form is correctly chosen and the assumptions regarding the errors are fulfilled.
   - Are the assumptions for the chosen estimation procedure fulfilled, so that the statistical properties of the estimation procedure apply and the inference is valid?
   - If assumptions are violated, specify and estimate alternative models with different variables if necessary and/or choose alternative estimation procedures $\longrightarrow$ Go back

to step 1 or step 3.

5. **Using the tested models**:

   - Testing the statistical hypotheses: Are the postulated (economic) hypotheses statistically refuted by the data?
   - Predictions
   - Interpretation of parameters of interest

The econometric procedures relevant for the individual steps are discussed in the following chapters.

5. Fundamentals of Estimation and Test Theory

5.1. Samples and data-generating processes

Let $\mathbf{s}_t$ denote a $m \times 1$ vector of random variables.

Samples

- A **sample** is a subset of the population that can be surveyed (=random vector) or has already been surveyed (=realisation of a random vector). A sample of **sample size** n is given by
$$\{\mathbf{s}_t, t = 1, \ldots, n\}.$$
- The stochastic properties of a sample are fully described by the **joint density of all** n **sample observations**:
$$f_{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n}(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n).$$
With respect to this joint density, a sample is a **possible future** or **already occurred realisation** with n observations $\mathbf{s}_t$.
- **Types of samples**:
    - **Random sample**: The n sample observations $\mathbf{s}_t$, $t = 1, \ldots, n$ are drawn **randomly**, i. e. they are **independently and identically distributed (IID)**, i. e. in addition to (2.22) the marginal densities are identical,
$$f_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n){=}f_{\mathbf{s}}(\mathbf{s}_1)f_{\mathbf{s}}(\mathbf{s}_2)\cdots f_{\mathbf{s}}(\mathbf{s}_n) = \prod_{t=1}^{n} f_{\mathbf{s}}(\mathbf{s}_t). \tag{5.1}$$
The great advantage of random sampling is that only the joint / marginal density $f_{\mathbf{s}}(\mathbf{s}_t)$ has to be determined and not the joint density of all sample observations. All sample observations are draws from the same density.
    - There are **stochastic dependencies** between the individual sample observations $\mathbf{s}_t$, i. e. the decomposition (5.1) does not hold. Then it holds that
$$\begin{aligned} f_{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n}(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) &= f_{\mathbf{s}_n|\mathbf{s}_{n-1},\ldots,\mathbf{s}_1}(\mathbf{s}_n|\mathbf{s}_{n-1}, \ldots, \mathbf{s}_1) \\ &\quad f_{\mathbf{s}_{n-1}|\mathbf{s}_{n-2},\ldots,\mathbf{s}_1}(\mathbf{s}_{n-1}|\mathbf{s}_{n-2}, \ldots, \mathbf{s}_1) \\ &\quad \cdots f_{\mathbf{s}_2|\mathbf{s}_1}(\mathbf{s}_2|\mathbf{s}_1)f_{\mathbf{s}_1}(_1) \qquad (5.2) \\ &= \prod_{t=1}^{n} f_{\mathbf{s}_t|\mathbf{s}_{t-1},\ldots,\mathbf{s}_1}(\mathbf{s}_t|\mathbf{s}_{t-1}, \ldots, \mathbf{s}_1) \end{aligned}$$

by applying Bayes' theorem several times. The joint density can be expressed as a product of conditional densities in the case of dependent observations.

- If the index t notes the time, the observations are uniquely sorted. Then the time-ordered collection of random variables $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ is called **stochastic process** and an observed sample is called **time series**. To model them, **time series models** are used. If $m = 1$ and s_t is a scalar, it is a **univariate time series**. If $m > 1$ and $\mathbf{s}_t$ is a vector, one examines a **multivariate time series**. An introduction to univariate time series models can be found in section 12.3.1.

  It is possible that the conditional densities $f_{\mathbf{s}_t|\mathbf{s}_{t-1},\dots,\mathbf{s}_1}(\mathbf{s}_t|\mathbf{s}_{t-1},\dots,\mathbf{s}_1)$ depend on time t. For example, they may depend on seasonal components or a time trend. This is indicated either by suitable indices at the conditional densities or corresponding variables in the condition of the densities. More on this in chapter 13.

- If time series data are available for all units in the cross-section, one speaks of **panel data**, see master course **Applied Microeconometrics**.

Data generating mechanism, data generating process (DGP):

- In econometrics/statistics, the concept **data generating mechanism** or **data generating process (DGP)** is often used instead of **population**. This refers to the **stochastic mechanism** that may have generated the observed sample data $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$ in the real world (Davidson & MacKinnon 2004, Sections 1.5, 3.1).

- The DGP underlying a **random sample** of n observations $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$ is fully determined by the joint / marginal ($m = 1$) density $f_{\mathbf{S}}(\mathbf{s})$.

- In the case of **dependent sample observations**, typically in the case of **time series** generated by a **stochastic process**, due to (5.2) the conditional densities $f_{\mathbf{s}_{t-1}|\mathbf{s}_{t-2},\dots,\mathbf{s}_1}(\mathbf{s}_{t-1}|\mathbf{s}_{t-2},\dots,\mathbf{s}_1)$ have to be considered.

  Example: DGP for daily DAX returns:

  Assumption regarding DGP: The daily DAX returns are independent and identically normally distributed

  $$y_t \sim NID(\mu_0, \sigma_0^2). \tag{5.3}$$

  The expected value μ_0 and variance σ_0^2 are fixed but unknown. Alternative notation, cf. (2.6):

  $$f(y_t; \mu_0, \sigma_0^2) = \frac{1}{\sigma_0^2}\phi\left(\frac{y_t - \mu_0}{\sigma_0}\right) \tag{5.4}$$

- As in the previous DAX example, in this text, parameters of a DGP are always noted with index 0.

If one is only interested in the part of the DGP for the endogenous variables given the causal variables, one decomposes the density $f(\mathbf{s}_t)$ suitably into conditional densities.

It denotes

$$\mathbf{s}_t = \begin{pmatrix} \mathbf{w}_t \\ \mathbf{y}_t \\ \mathbf{z}_t \end{pmatrix} = \begin{cases} \text{variables without direct impact on } \mathbf{y}_t \\ \text{explained/endogenous variables} \\ \text{explanatory/exogenous variables} \end{cases} \tag{5.5}$$

Factors influencing imports $\mathbf{w}_t = \begin{pmatrix} s_1 & s_2 & s_3 & s_9 \end{pmatrix}^T$, $y_t = s_8$, $\mathbf{z}_t = \begin{pmatrix} s_4 & s_5 & s_6 & s_6 \end{pmatrix}$, s_3 is irrelevant if only one importing country and one period is considered.

Then (in general) the following factorisation

$$f_{\mathbf{S}}(\mathbf{s}_t) = f_{\mathbf{W}|\mathbf{Y},\mathbf{Z}}(\mathbf{w}_t|\mathbf{y}_t, \mathbf{z}_t)\, f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}_t|\mathbf{z}_t)\, f_{\mathbf{Z}}(\mathbf{z}_t) \tag{5.6}$$

makes sense.

For the explanation of $\mathbf{y}_t$ **given the explanatory variables** $\mathbf{z}_t$, only the conditional density

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}_t|\mathbf{z}_t)$$

is necessary. Neither the conditional density $f_{\mathbf{W}|\mathbf{Y},\mathbf{Z}}(\mathbf{w}|\mathbf{y}, \mathbf{z})$, nor the joint density $f_{\mathbf{Z}}(\mathbf{z}_t)$ need to be considered, which simplifies the modelling process substantially!

5.2. Econometric Models

To simplify the notation, in this section we only consider models for random samples.

Models for samples with stochastically dependent observations are covered in section 13.2 and are an extension of the models from this section.

- An **econometric model** $\mathbb{M}$ is a family of functions $M(\cdot)$ depending on the data and a $p^* \times 1$ parameter vector $\boldsymbol{\psi}$. The functions can describe (economic) relationships and implicitly or explicitly contain a full or partial description of the DGP or at least an approximation of the DGP (Davidson 2000, Section 4.1.1). The set of possible and allowed parameters is called **parameter space** $\boldsymbol{\Psi}$,

$$\mathbb{M} \equiv \{M(\mathbf{s}_t; \boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}, \quad \boldsymbol{\Psi} \subseteq \mathbb{R}^{p^*}. \tag{5.7}$$

- **Structural form** of a model: Essential parameters of the model can be interpreted (economically).

  Example: supply and demand functions The variable q_A denotes the quantity supplied for a good given its price p, the variable q_N denotes the quantity demanded for the same good given its price p, the variables z_1 and z_2 are exogenous. The error terms are denoted by $\tilde{u}_1$ and $\tilde{u}_2$. A simple example of a supply function and a demand function is then

$$\text{Supply equation:} \quad q_A = \alpha_1 p + \beta_1 z_1 + \tilde{u}_1, \tag{5.8a}$$

$$\text{Demand equation:} \quad q_N = \alpha_2 p + \beta_2 z_2 + \tilde{u}_2, \tag{5.8b}$$

$$\text{Equilibrium condition:} \quad q_A = q_N = q. \tag{5.8c}$$

If market clearing is present, q_N and q_A are in principle unobservable, since only the market clearing quantity q is observable. The parameters α_1 and α_2 can be interpreted economically as the price elasticity of supply and the price elasticity of demand, respectively.

Example: simultaneous equation model

The system of equations (5.8) can be written as a simultaneous equation model by setting $y_1 = q_A$ und $y_2 = p$ and dividing the demand equation by $-\alpha_2$ and rearranging. One then obtains a simultaneous equation model with two endogenous variables

$$y_1 = \alpha_{12}y_2 + \beta_{11} + \beta_{12}z_1 + u_1, \tag{5.9a}$$
$$y_2 = \alpha_{21}y_1 + \beta_{21} + \beta_{23}z_2 + u_2. \tag{5.9b}$$

In (5.9) (and in (5.8)) there are not only causal variables on the right-hand side, as in (4.1), but also endogenous variables. The parameter α_{12} measures the causal effect $y_2 \longrightarrow y_1$ and the parameter α_{21} measures the causal effect $y_1 \longrightarrow y_2$. If one additionally assumes

$$E[u_1|z_1, z_2] = 0, \quad E[u_1|z_1, z_2] = 0, \tag{5.9c}$$
$$\sigma_1^2 = Var(u_1|z_1, z_2), \quad \sigma_1^2 = Var(u_2|z_1, z_2), \tag{5.9d}$$

the simultaneous equation model (5.9) with $\mathbf{s}_t = \begin{pmatrix} y_{1t} & y_{2t} & z_{1t} & z_{2t} \end{pmatrix}^T$ yields the parameter vector

$$\boldsymbol{\psi} = \begin{pmatrix} \alpha_{12} & \beta_{11} & \beta_{12} & \sigma_1^2 & \alpha_{21} & \beta_{21} & \beta_{23} & \sigma_2^2 \end{pmatrix}. \tag{5.10}$$

- It is possible, as in the case of a simultaneous equation model (5.9), that the elements of structural models do not contain a set of conditional densities or parts thereof (such as conditional expectation values). In such cases, the structural model has to be transformed so that the elements of the derived model are conditional densities or parts thereof.

  Notation: This transformation also produces a new parameter vector $\boldsymbol{\theta}$ of length p, which results from the parameter vector $\boldsymbol{\psi}$ of the structural form and typically has fewer parameters, $p \leq p^*$. We then write

  $$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\psi}) \in \boldsymbol{\Theta} \quad \text{for all } \boldsymbol{\psi} \in \boldsymbol{\Phi}. \tag{5.11}$$

  So the corresponding parameter space is $\boldsymbol{\Theta}$.

- If joint densities are considered for $\mathbf{s}_t$ that depend on a parameter vector $\boldsymbol{\psi}$, one writes

  $$f_{\mathbf{S}}(\mathbf{s}_t; \boldsymbol{\theta}) = f_{\mathbf{S}}(\mathbf{s}_t; \boldsymbol{\theta}(\boldsymbol{\psi})).$$

  Thus, the set of all densities implied by a structural model $\mathbb{M}$ can be written as

  $$\mathbb{M}_D \equiv \{f_{\mathbf{S}}(\mathbf{s}_t; \boldsymbol{\theta}(\boldsymbol{\psi})), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}. \tag{5.12}$$

  In many standard cases, the structural model already corresponds to the model definition (5.12). Therefore, Davidson & MacKinnon (2004, Section 3.1) define an econometric model as a set $\mathbb{M}_D$ of possible DGPs. However, the definition (5.7) used here is more general.

- A parameter vector $\boldsymbol{\psi}$ of a model for which the density $f_{\mathbf{S}}(\mathbf{s}_t;\boldsymbol{\theta})$ implied by the model is equal to the density of the DGP $f_{\mathbf{S}}(\mathbf{s}_t)$, i. e.

$$f_{\mathbf{S}}(\mathbf{s}_t;\boldsymbol{\theta}(\boldsymbol{\psi}_0)) = \underbrace{f_{\mathbf{S}}(\mathbf{s}_t)}_{\text{DGP}}, \tag{5.13}$$

  is denoted by $\boldsymbol{\psi}_0$ and is often referred to as **true parameter vector** or **correct parameter vector**. The same applies to $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\psi}_0)$.

- Models $\mathbb{M}_D$ are also called models in **reduced form**. The parameters of interest $\boldsymbol{\theta}$ of a reduced form model are interpretable only if the structural form and reduced form of a model are identical. More on this in section 13.3.

Example: DAX returns – continued: The model with $s_t = y_t$ includes all possible DGPs of the type

$$y_t \sim NID(\mu,\sigma^2), \quad \boldsymbol{\theta} = \begin{pmatrix}\mu\\ \sigma\end{pmatrix} \in \boldsymbol{\Theta} \in \left(\mathbb{R}\times\mathbb{R}^+\right). \tag{5.14}$$

Or.:

$$\mathbb{M} = \mathbb{M}_D = \left\{ f(y_t;\mu,\sigma^2) := \frac{1}{\sigma}\phi\left(\frac{y_t-\mu}{\sigma}\right), \mu\in\mathbb{R}, \sigma\in\mathbb{R}^+ \right\}.$$

The structural model is a set of densities $\mathbb{M}_D$.

Example: simultaneous equation model — reduced form To simplify the following illustrations, the simultaneous equation system (5.9) is written in matrix notation

$$\underbrace{\begin{pmatrix}1 & -\alpha_{12}\\ -\alpha_{21} & 1\end{pmatrix}}_{\mathbf{A}}\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \underbrace{\begin{pmatrix}\beta_{11}\\ \beta_{21}\end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix}\beta_{12} & 0\\ 0 & \beta_{23}\end{pmatrix}}_{\mathbf{B}}\begin{pmatrix}z_1\\ z_2\end{pmatrix} + \begin{pmatrix}u_1\\ u_2\end{pmatrix},$$

so that the **structural form** in matrix notation is

$$\mathbf{A}\mathbf{y} = \boldsymbol{\beta} + \mathbf{B}\mathbf{z} + \mathbf{u}. \tag{5.15}$$

If $\mathbf{A}$ is regular, the **reduced form** results from inversion

$$\begin{aligned}\mathbf{y} &= \underbrace{\mathbf{A}^{-1}\boldsymbol{\beta}}_{\mathbf{b}} + \underbrace{\mathbf{A}^{-1}\mathbf{B}}_{\mathbf{D}}\mathbf{z} + \underbrace{\mathbf{A}^{-1}\mathbf{u}}_{\boldsymbol{\varepsilon}},\\ \mathbf{y} &= \mathbf{b} + \mathbf{D}\mathbf{z} + \boldsymbol{\varepsilon}.\end{aligned}$$

In order to obtain a model of the type $\mathbb{M}_D$ for (5.9), one additionally has to use

– an assumption concerning the joint distribution of the errors u_{1t} and u_{2t} and

– an assumption regarding the joint distribution of the exogenous variables z_{1t} and z_{2t} and their stochastic relation to the errors, e. g.

$$\begin{pmatrix}u_1\\ u_2\end{pmatrix}|z_1,z_2 \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \boldsymbol{\Sigma}_u\right).$$

Then

$$\boldsymbol{\varepsilon}|\mathbf{z} \sim N(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \mathbf{A}^{-1}\boldsymbol{\Sigma}_u(\mathbf{A}^{-1})^T, \tag{5.16}$$

$$\mathbf{y}|\mathbf{z} \sim N(\mathbf{b} + \mathbf{D}\mathbf{z}, \boldsymbol{\Sigma}) \tag{5.17}$$

respectively

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}; \mathbf{b}, \mathbf{D}, \boldsymbol{\Sigma}) = \frac{1}{2\pi}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{b} - \mathbf{D}\mathbf{z})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{b} - \mathbf{D}\mathbf{z})\right)$$

$$\mathbb{M}_D = \left\{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}; \mathbf{b}, \mathbf{D}, \boldsymbol{\Sigma}), \mathbf{b}, \mathbf{D}, \boldsymbol{\Sigma} \text{ elements of the parameter space } \boldsymbol{\Theta}\right\}.$$

This is presented in even more detail in section 13.2 in the more general framework of multivariate time series models..

- **Correctly specified econometric model**: A model is
  - **correctly specified**, if DGP $\in \mathbb{M}_D$,
  - **misspecified**, if DGP $\notin \mathbb{M}_D$

  holds.

  Example: DAX returns – continued: If $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0)^T \in \boldsymbol{\Theta}$, then the model (5.14) also contains the actual DGP (5.3) with μ_0 and σ_0^2 and the model is correctly specified.

  However, if the DGP of the DAX returns is given by a t-distribution

  $$y_t/\sigma_0 \; IID\, t(m_0), \quad m_0 = 5,$$

  the model (5.14) is misspecified.

  Example: DAX returns – continued: The model (5.14) is fully specified. A model $y_t \sim IID(\mu, \sigma^2)$ is not fully specified because a distributional assumption is missing.

- Davidson & MacKinnon (2004, Section 1.3) call a parametric model **fully specified** if it is possible to generate realisations of the dependent variable $\mathbf{y}_t$ after assigning numerical values to all parameters present in the model. Otherwise it is **partially specified**.
- If a model in the reduced form $\mathbb{M}_D$ can be derived from the model $\mathbb{M}$ in structural form, we say that the model $\mathbb{M}$ is **fully specified**.
- If a structural **model** $\mathbb{M}$ **is fully and moreover correctly** specified, there exists a parameter vector $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\psi}_0)$ for which the density in $\mathbb{M}_D$ corresponds to the DGP:

$$\mathbb{M}_D \supseteq M_D(\mathbf{s}_t; \boldsymbol{\theta}_0) \equiv f(\mathbf{s}_t; \boldsymbol{\theta}_0) = \underbrace{f_{\mathbf{S}}(\mathbf{s}_t)}_{\text{DGP}}. \tag{5.18}$$

Example: Simultaneous equation model

If for the simultaneous equation model (5.9) with parameter vector (5.10) no joint distribution of the causal variables z_1 and z_2 is given, but only a joint distribution (5.16) of the errors, then "only" a model for conditional densities can be derived.

- **Model classes**:
  - **Univariate models**: $s_t = y_t$, is a scalar, $m = 1$.
  - **Multivariate models**: $\mathbf{s}_t$ is a vector, $m > 1$.
- Econometric models in which the implied DGPs are distinguished by functions depending on the possible variables and a parameter vector $\boldsymbol{\psi}$ of fixed length $p^* < \infty$ are called **parametric econometric models**.
- However, **semiparametric models** and **nonparametric models** also play a role in econometric theory and practice. A brief introduction is provided by Davidson & MacKinnon (2004, Section 15.5). A detailed presentation is given in the monograph by Li & Racine (2007).

Conditional models

- If one is only interested in explaining the endogenous variables $\mathbf{y}$ given the causal variables $\mathbf{z}$, it is sufficient to consider conditional models. Based on the factorisation of the DGP in (5.6) we obtain a **conditional econometric model** (for conditional densities)

$$\mathbb{M}_D \equiv \left\{ f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}. \tag{5.19}$$

  The variables $\mathbf{z}$ are determined outside the model. The distinction into endogenous and exogenous variables was already made in the first presentation of the multiple linear regression model (4.1) in section 4.1:
  - **Endogenous variable(s)**: variable(s) is/are generated by the mechanism described in the model.
  - **Exogenous variables**: variables that can be determined outside the model (as they have no simultaneous relationship with the endogenous variables).
- **Important**: It is often not clear whether a variable s_j has a simultaneous relationship with y or is causal for y. Then this must be determined in the modelling process (which, however, is only possible under certain conditions). This requires that s_j is first allowed to be simultaneous in the model and that at the same time parameter values exist in the model for which s_j becomes causal. An example of this can be found in 13.2 and 13.3.
- **In the following up to chapter 13 we assume that only *one* endogenous variable is to be explained and all explanatory variables are causal.** Under this assumption, the structural form and the reduced form are identical and $\boldsymbol{\theta} = \boldsymbol{\psi}$ holds.

Example: Factors influencing imports

In order to obtain a density conditional on GDP and $distance$ for $imports$ based on the regression equation (4.2), an assumption must still be made for the distribution of the errors and their relationship to the conditioning random variables GDP and $distance$. Often a conditional normal distribution

$$u|GDP, distance \sim NID(0, \sigma^2)$$

is assumed. Under this assumption, it follows immediately that $imports$ are conditionally normally distributed:

$$Imports|GDP, distance \sim NID(\beta_1 + \beta_2\, GDP + \beta_3\, distance, \sigma^2). \qquad (5.20)$$

Or with $\boldsymbol{\theta} = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \sigma^2 \end{pmatrix}^T$

$$\begin{aligned} \mathbb{M} = \mathbb{M}_D = \{ & f(Imports|GDP, distance; \boldsymbol{\psi}) \\ & \equiv \frac{1}{\sigma^2}\phi\left(\frac{Imports - \beta_1 - \beta_2\, GDP - \beta_3\, distance}{\sigma}\right), \\ & \boldsymbol{\theta} \in \boldsymbol{\Theta} \in \left(\mathbb{R}^3 \times \mathbb{R}^+\right)\right\}. \end{aligned}$$

- For an empirical analysis, often different models $\mathbb{M}_i$, $i = 1, 2, \ldots, I$ are considered. With the help of **model selection methods**, an attempt is then made to select a correct model. More on this later.
- In practice, econometric models are (almost) always misspecified. For conditional fully specified models this means

  $$f(\mathbf{y}|\mathbf{z}) \not\subseteq \mathbb{M}_{D,i} \quad , i = 1, 2, \ldots, I.$$

  One then tries to choose from the different models, $i = 1, 2, \ldots, I$, a model that provides the best possible approximation to the DGP for the purpose of the investigation. However, we ignore the resulting consequences in this course.

Information sets

- The set of all potential variables that could be considered as **causal** variables for a given question and a model to be used for it to explain the endogenous variables $\mathbf{y}_t$ is often referred to as **information set** and abbreviated with Ω_t. The information set typically depends on time t for time series, hence the index t, see section 13.1.
- The set of all variables used as **causal** variables in a given model to explain the endogenous variables $\mathbf{y}_t$ is also an **information set** and will be abbreviated as $\mathcal{I}_t \subset \Omega_t$ in the following.

5.3. Regression models

For many (economic) questions it is not necessary to model the DGP or the conditional density completely.

Notation: All explanatory variables and a constant, if necessary, are combined in the $(1 \times k)$ row vector

$$\mathbf{X}_t \equiv \begin{pmatrix} X_{t1} & \cdots & X_{tk} \end{pmatrix}.$$

If a constant is present, the following applies,

$$\mathbf{X}_t = \begin{pmatrix} 1 & z_{1t} & \cdots & z_{k-1,t} \end{pmatrix}.$$

Very often it is sufficient to model individual characteristics of the (conditional) densities, in particular

- the conditional expected value $E\left[y_t|\mathbf{X}_t\right]$ and/or
- the conditional variance $Var\left(y_t|\mathbf{X}_t\right)$ or also
- conditional quantiles.

Regression models:

- A conditional model for modelling the conditional expected value $E\left[y_t|\mathbf{X}_t\right]$ is called a **regression model**.
- The identity
$$y_t = \underbrace{E\left[y_t|\mathbf{X}_t\right]}_{\text{systematic part}} + \underbrace{y_t - E\left[y_t|\mathbf{X}_t\right]}_{\text{unsystematic part}}$$
becomes a regression model by specifying the conditional expected value function $E\left[y_t|\mathbf{X}_t\right]$.

  The conditional expected value is called the systematic part and is written in the following as $m(\mathbf{X}_t) = E[y_t|\mathbf{X}_t]$, where the conditional expected value is calculated with respect to the density of the DGP. The unsystematic part is called the **error term** or the **disturbance term**. In the context of the correct function $m(\mathbf{X}_t)$, the error term is denoted by ε_t,
$$y_t = m(\mathbf{X}_t) + \varepsilon_t. \tag{5.21}$$

- The function of the conditional expected value $m(\mathbf{X}_t)$ is generally not known. If we assume that the function $m(\mathbf{X}_t)$ is linear in the parameters $\beta_1, \dots, \beta_k$, we obtain
$$m(\mathbf{X}_t, \boldsymbol{\beta}) = x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tk}\beta_k = \mathbf{X}_t\boldsymbol{\beta}, \quad \boldsymbol{\beta} := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{5.22}$$
If this assumption is correct and there exists a $\boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \boldsymbol{\Psi}$ such that
$$m(\mathbf{X}_t, \boldsymbol{\beta}_0) = m(\mathbf{X}_t) = E[y_t|\mathbf{X}_t], \tag{5.23}$$
then the **function of the conditional expected value is correctly specified** and $\boldsymbol{\beta}_0$ is called the **correct parameter vector**.

- For all parameter vectors in the parameter space, $\boldsymbol{\beta} \in \boldsymbol{\Psi}$, one obtains the **multiple linear regression model**

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t \tag{5.24}$$
$$y_t = \mathbf{X}_t \boldsymbol{\beta} + u_t. \tag{5.25}$$

- **Notation**: For a given sample $\{(y_t, \mathbf{X}_t), t = 1, \dots, n\}$, strictly speaking, one would have to write $u_t(\boldsymbol{\beta})$ instead of u_t, i. e.

$$y_t = \mathbf{X}_t \boldsymbol{\beta} + u_t(\boldsymbol{\beta}), \tag{5.26}$$

  since the error term must change if $\boldsymbol{\beta}$ changes, as long as all y_t and $\mathbf{X}_t$ remain the same. We will use this notation later when interpreting the OLS estimator. Otherwise, we simply use u_t, as is common in the literature.

- **Notation**: Unlike Wooldridge (2009), Davidson & MacKinnon (2004) start the index of parameters at 1 and count to k. The course generally follows Davidson & MacKinnon (2004), also in other notational matters.

- **If the conditional expected value function is correctly specified**, then (5.22) holds for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Substituting $\boldsymbol{\beta}_0$ into the multiple linear regression model (5.25) and determining the conditional expected value shows that then

$$E\left[u_t | \mathbf{X}_t\right] = 0 \tag{5.27}$$

  holds. Therefore, it makes sense to use the requirement (5.27) to estimate $\boldsymbol{\beta}_0$. This leads directly to the least squares estimator (6.5), which is discussed in the following chapters.

- **Possible interpretation of the error term** The condition (5.27) can be interpreted as follows. There are other causal factors $\mathbf{v}_t$ for y_t that are stochastically independent of the explicitly considered causal factors $z_{1t}, \dots, z_{k-1,t}$ and also not considered in the vector $\mathbf{s}_t$ in (5.5). If these influence y_t as a linear combination, the equation contains all causal variables

$$y_t = \beta_{20} z_{1t} + \beta_{30} z_{2t} + \cdots + \beta_{k0} z_{k-1,t} + \mathbf{v}_t^T \boldsymbol{\gamma}_0. \tag{5.28}$$

  If it were possible to observe $\mathbf{v}_t$ in addition to $z_{1t}, \dots, z_{k-1,t}$, then y_t could be predicted exactly given that the true parameters are known, since $E[y_t | z_{1t}, \dots, z_{k-1,t}, \mathbf{v}_t]$ just corresponds to the right-hand side of (5.28) and hence y_t.

  If the $\mathbf{v}_t$ are not known, only the conditional expectation value

$$\begin{aligned} E[y_t | z_{1t}, \dots, z_{k-1,t}] &= \beta_{20} z_{1t} + \beta_{30} z_{2t} + \cdots + \beta_{k0} z_{k-1,t} + E[\mathbf{v}_t | z_{1t}, \dots, z_{k-1,t}]^T \boldsymbol{\gamma}_0 \\ &= \beta_{20} z_{1t} + \beta_{30} z_{2t} + \cdots + \beta_{k0} z_{k-1,t} + \underbrace{E[\mathbf{v}_t]^T \boldsymbol{\gamma}_0}_{\equiv \beta_1} \end{aligned}$$

  can be determined. The second equal sign follows because of the stochastic independence of $\mathbf{v}_t$ and the $z_{jt}, j = 1, \dots, k-1$. Thus, the constant β_1 just corresponds to the linear combination of the unconditional expected values of all unconsidered factors $\mathbf{v}_t$ and the parameter vector $\boldsymbol{\gamma}_0$.

Alternatively, it is possible that one is not interested in the individual influences $\mathbf{v}_t$, so that it is sufficient to consider the conditional expected value.

The error term u_t then results from

$$u_t = \mathbf{v}_t^T \boldsymbol{\gamma}_0 - E[\mathbf{v}_t]^T \boldsymbol{\gamma}_0 \tag{5.29}$$

from the individual deviations from this mean. Furthermore, (5.27) holds.

Therefore, this part of the model is called the unsystematic part.

Reminder: In empirical analysis, causality can only be determined in the average effect of an action.

Important: If one were to include some elements of $\mathbf{v}_t$ in the regression, this would lead to a reduction in the variance of the error term. If $\mathbf{v}_t$ could be included completely, the variance disappears, since y_t can be predicted exactly.

- Regression models belong to **conditional models** because the regressors are not explained in the model.
- **Simple linear regression model**:

$$y_t = \beta_1 + \beta_2 x_t + u_t. \tag{5.30}$$

- Regression models are either
  - **correctly specified** (DGP included in the model) or
  - **misspecified** (DGP not included in the model).

Example of misspecified model: DGP

$$y_t = \beta_{10} + \beta_{20} x_t + \beta_{30} x_t^2 + v_t, \quad E[v_t|x_t] = 0, \quad \beta_{30} \neq 0 \tag{5.31}$$

Model: the simple linear regression model (5.30).

The conditional expected value given the DGP is:

$$\begin{aligned} y_t &= \beta_{10} + \beta_{20} x_t + \underbrace{\beta_{30} x_t^2 + v_t}_{u_t} \\ E\left[y_t|x_t\right] &= \beta_{10} + \beta_{20} x_t + \underbrace{E\left[u_t|x_t\right]}_{=\beta_{30} x_t^2 \neq 0}, \end{aligned}$$

such that condition $E[u_t|x_t] = 0$ in (5.30) is violated and the DGP is not included in the model (5.30).

Note: For the analysis of a specific question, it is possible to use misspecified models under certain conditions. This includes the selection of an adequate estimation procedure, e. g. the use of the **instrument variable estimator** (**IV estimator**) or the **GMM estimator** in each case with appropriate instruments, see master course **Advanced Econometrics** or Davidson & MacKinnon (2004, Chapter 8 and 9).

- Regression models are **fully specified** if all conditional density parameters are included in the modelling. I. e. in particular that the distribution of the disturbance term is modelled.

  Examples on fully and partially specified regression models:

  - The regression model
$$\begin{aligned} \ln(Importe_t) &= \beta_1 + \beta_2 \ln(BIP_t) + u_t, \\ u_t|\ln(BIP_t) &\sim NID(0, \sigma^2) \end{aligned} \tag{5.32}$$
    is fully specified.

  - In contrast, if only $u_t|\ln(BIP_t) \sim IID(0, \sigma^2)$ is specified in (5.32) in the model, the distribution of the disturbance term remains open and the model is partially specified. If the DGP is included in the model, the model is partially but correctly specified.

- **Property** of the **true parameter vector** $\boldsymbol{\beta}_0$ in the multiple linear regression model of the **population**:

  - In the correctly specified model, $E[y_t|\mathbf{X}_t] = \mathbf{X}_t\boldsymbol{\beta}_0$ and thus (5.27) $E[u_t|\mathbf{X}_t] = 0$ holds, where according to (5.26) $u_t = u_t(\boldsymbol{\beta}_0)$. This results in
$$E[u_t x_{tj}] = 0 \quad j = 1, \dots, k, \quad E[u_t\mathbf{X}_t] = \mathbf{0}. \tag{5.33}$$
    After multiplying (5.25) with $\mathbf{X}_t^T$, we get
$$\begin{aligned} \mathbf{X}_t^T y_t &= \mathbf{X}_t^T\mathbf{X}_t\boldsymbol{\beta}_0 + \mathbf{X}_t^T u_t \\ E\left[\mathbf{X}_t^T y_t\right] &= E\left[\mathbf{X}_t^T\mathbf{X}_t\right]\boldsymbol{\beta}_0 + \underbrace{E\left[\mathbf{X}_t^T u_t\right]}_{=\mathbf{0}} \\ \boldsymbol{\beta}_0 &= E\left[\mathbf{X}_t^T\mathbf{X}_t\right]^{-1} E\left[\mathbf{X}_t^T y_t\right], \end{aligned} \tag{5.34}$$
    provided $E\left[\mathbf{X}_t^T\mathbf{X}_t\right]$ is invertible. Since (5.33) is only valid for the true parameter vector if specified correctly, this condition can be used to derive an estimator. Since the conditions (5.33) contains second moments of the DGP, they are called **moment conditions**. If the moments change, the parameter vector also changes.

  - Using the **moment conditions**, estimators can be derived in many cases. In section 6.2.1 it is shown that the moment conditions (5.33) imply the OLS estimator.

- Now the question is what happens when the **moment condition is applied to a misspecified model**. We consider the case where (5.21) describes the DGP, where the regression function $m(\cdot)$ can also be non-linear in the parameters, so that

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \underbrace{m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta} + \varepsilon_t}_{u_t(\boldsymbol{\beta})}. \tag{5.35}$$

In the following, the expectation value concerning the densities of the unknown DGP is calculated. The conditional expectation value of the error term $u_t(\boldsymbol{\beta})$ is

$$\begin{aligned} E[u_t(\boldsymbol{\beta})|\mathbf{X}_t] &= E\left[m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta} + \varepsilon_t|\mathbf{X}_t\right] \\ &= m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta} + \underbrace{E[\varepsilon_t|\mathbf{X}_t]}_{=0} \\ &= m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta} = \begin{cases} = 0 & \text{if correctly specified and } \boldsymbol{\beta} = \boldsymbol{\beta}_0, \\ \neq 0 & \text{possible if model misspecified.} \end{cases} \end{aligned} \tag{5.36}$$

In the second case, an **approximation error** may occur for individual observations.

Analogous to the procedure for correct specification, one obtains

$$\mathbf{X}_t^T y_t = \mathbf{X}_t^T\mathbf{X}_t\boldsymbol{\beta} + \mathbf{X}_t^T u_t(\boldsymbol{\beta}), \tag{5.37}$$

$$E[\mathbf{X}_t^T y_t] = E[\mathbf{X}_t^T\mathbf{X}_t]\boldsymbol{\beta} + E[\mathbf{X}_t^T u_t(\boldsymbol{\beta})]. \tag{5.38}$$

Applying the moment condition (5.33) accordingly to $E[\mathbf{X}_t^T u_t(\boldsymbol{\beta})]$ and there exists a $\boldsymbol{\beta}_{00}$, so that

$$E[\mathbf{X}_t^T u_t(\boldsymbol{\beta}_{00})] = 0 \tag{5.39}$$

holds, we get

$$\boldsymbol{\beta}_{00} = E[\mathbf{X}_t^T\mathbf{X}_t]^{-1}E[\mathbf{X}_t^T y_t]. \tag{5.40}$$

The parameter vector $\boldsymbol{\beta}_{00}$ is often referred to as **pseudo-true parameter vector**.

Interpretation of the pseudo-true parameter vector:

1. **If there is a constant in the model**, i.e. $x_{t1} = 1$, then it follows from the moment condition (5.39) that the unconditional expected value of the errors is zero, since $E[x_{t1}u_t(\boldsymbol{\beta}_{00})] = E[u_t(\boldsymbol{\beta}_{00})] = 0$ holds.

   Then we get

$$\begin{aligned} Cov(\mathbf{X}_t, u_t(\boldsymbol{\beta}_{00})) &= E[\mathbf{X}_t^T u_t(\boldsymbol{\beta}_{00})] - E[\mathbf{X}_t^T]\, E[u_t(\boldsymbol{\beta}_{00})] \\ &= E[\mathbf{X}_t^T u_t(\boldsymbol{\beta}_{00})] = E[\mathbf{X}_t^T(m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta}_{00})] + E[\mathbf{X}_t^T\varepsilon_t] \\ &= E[\mathbf{X}_t^T u_t(\boldsymbol{\beta}_{00})] = E[\mathbf{X}_t^T(m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta}_{00})]. \end{aligned} \tag{5.41}$$

The moment condition (5.39) therefore guarantees (given a constant in the model) that for the **pseudo-true parameter vector $\boldsymbol{\beta}_{00}$ the covariance between the regressors $\mathbf{X}_t$ and the approximation errors $m(\mathbf{X}_t) - \mathbf{X}_t\boldsymbol{\beta}_{00}$ is zero**. In other words, the approximation errors and the regressors $\mathbf{X}_t$ are uncorrelated. In this case, (5.40) can also be written as

$$\boldsymbol{\beta}_{00} = Var(\mathbf{X}_t)^{-1}\, Cov(\mathbf{X}_t^T, y_t).$$

$\sharp$ Proof: $Cov(\mathbf{X}_t^T, y_t) = E[\mathbf{X}_t^T y_t] - E[\mathbf{X}_t^T]E[y_t] = E[\mathbf{X}_t^T y_t] - E[\mathbf{X}_t^T]E[\mathbf{X}_t]\boldsymbol{\beta}_{00} - E[\mathbf{X}_t^T \varepsilon_t] = E[\mathbf{X}_t^T y_t] + \big(Var(\mathbf{X}_t) - E[\mathbf{X}_t^T\mathbf{X}_t]\big)\,\boldsymbol{\beta}_{00}$, so that

$$Cov(\mathbf{X}_t^T, y_t) - Var(\mathbf{X}_t)\boldsymbol{\beta}_{00} = E[\mathbf{X}_t^T y_t] - E[\mathbf{X}_t^T\mathbf{X}_t]\boldsymbol{\beta}_{00}.$$

2. A second interpretation of the pseudo-true parameter vector follows from the following reasoning.

   Consider the expected value of the squared deviations of the endogenous variable y_t from its **linear predictions $\mathbf{X}_t\boldsymbol{\beta}_{00}$** (linear in the parameters $\boldsymbol{\beta}$). This expected value will be referred to as the mean squared error (MSE) in the following section.

   First, one considers for a *given DGP* the MSE for a multiple linear regression model for any parameter vector $\boldsymbol{\beta}$

$$\begin{aligned} MSE(y_t, \mathbf{X}_t; \boldsymbol{\beta}) \equiv E[(y_t - \mathbf{X}_t\boldsymbol{\beta})^2] &= E[y_t^2 - 2y_t\mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}_t^T\mathbf{X}_t\boldsymbol{\beta}] \\ &= E[y_t^2] - 2E[y_t\mathbf{X}_t]\boldsymbol{\beta} + \boldsymbol{\beta}^T E[\mathbf{X}_t^T\mathbf{X}_t]\boldsymbol{\beta} \end{aligned} \tag{5.42}$$

   and then searches for the parameter vector that minimises this mean squared error. This is done by deriving the $MSE(y_t, \mathbf{X}_t; \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and setting to zero

$$\frac{\partial MSE(y_t, \mathbf{X}_t; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2E[\mathbf{X}_t^T y_t] + 2E[\mathbf{X}_t^T\mathbf{X}_t]\boldsymbol{\beta} \overset{!}{=} \mathbf{0}. \tag{5.43}$$

   From this follows again

$$\boldsymbol{\beta}_{00} = E\left[\mathbf{X}_t^T\mathbf{X}_t\right]^{-1} E\left[\mathbf{X}_t^T y_t\right]. \tag{5.40}$$

   Thus, the **pseudo-true parameter vector** $\boldsymbol{\beta}_{00}$ also yields the **best linear prediction** of y_t in terms of a minimum mean squared error (MSE) of the regression model.

   Outlook: The equation (5.43) further motivates a possible derivation of the least squares estimator in section 6.2.2.

Obviously, $\boldsymbol{\beta}_{00} = \boldsymbol{\beta}_0$ holds if the regression model is correctly specified.

5.4. Relevant properties of estimators

Notation for expected values of matrices:

$$E[\mathbf{X}] = \begin{pmatrix} E[x_{11}] & E[x_{12}] & \cdots & E[x_{1k}] \\ E[x_{21}] & E[x_{22}] & \cdots & E[x_{2k}] \\ \vdots & \vdots & \ddots & \vdots \\ E[x_{n1}] & E[x_{n2}] & \cdots & E[x_{nk}] \end{pmatrix} \tag{5.44}$$

Estimator and estimate

- The model contains p parameters that are summarised in the $(p \times 1)$ parameter vector $\boldsymbol{\theta}$. An **estimator** $\tilde{\boldsymbol{\theta}}(y_1, \ldots, y_n)$ for the parameter vector $\boldsymbol{\theta}$ is a **(vector-valued) function** that contains as argument only sample observations $(y_1, \ldots, y_n)$ and is used to determine estimates of $\boldsymbol{\theta}$ that are as close as possible to $\boldsymbol{\theta}$ in a sense to be further specified. An estimator $\tilde{\boldsymbol{\theta}}(y_1, \ldots, y_n)$ is a **function of random variables** since the sample observations are random variables before they are collected.

- If an estimator is calculated on the basis of a collected sample, an **estimate** of $\boldsymbol{\theta}_0$ is obtained.

- In general, the estimate deviates from the parameter values $\boldsymbol{\theta}_0$ of the actual DGP available. These deviations are called **estimation errors** $\tilde{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}_0$. The parameter vector of the actual DGP is often referred to as **true parameter vector**, as explained above.

Selection criteria for estimators

- The choice of estimation method depends on the chosen assessment of the estimation errors, which in turn depends on the research question. An assessment of the estimation error for a parameter i is possible using the **loss function** $L\left(\tilde{\theta}_i(\mathbf{y}), \theta_i\right)$. Typical loss functions for scalar parameters θ are:
  - **Quadratic loss function**:

    $$L_{sq}\left(\tilde{\theta}(\mathbf{y}), \theta_0\right) := \left(\tilde{\theta}(\mathbf{y}) - \theta_0\right)^2 \tag{5.45}$$

    The quadratic loss function measures the square of the Euclidean distance (length) between the estimated $\tilde{\theta}$ and the true parameter θ_0.

  - **Absolute value of the estimation error**:

    $$L_{abs}(\left(\tilde{\theta}(\mathbf{y}), \theta_0\right) := \left|\tilde{\theta}(\mathbf{y}) - \theta_0\right|. \tag{5.46}$$

  - **Asymmetric loss function**: Example:

    $$\begin{aligned} L_{abs}(\left(\tilde{\theta}(\mathbf{y}), \theta_0\right) := {} & a\left|\tilde{\theta}(\mathbf{y}) - \theta_0\right| 1\left(\tilde{\theta}(\mathbf{y}) - \theta_0 > 0\right) \\ & + b\left|\tilde{\theta}(\mathbf{y}) - \theta_0\right| 1\left(\tilde{\theta}(\mathbf{y}) - \theta_0 < 0\right), \quad a, b > 0, \end{aligned} \tag{5.47}$$

    where $1(\cdot)$ denotes the indicator function.

- The value of the loss function depends on the sample. To get a sample-independent value, one considers the expected value of the loss function
$$E\left[L\left(\tilde{\theta}(\mathbf{y}),\theta_0\right)\right], \tag{5.48}$$
where the expected value is determined with respect to the sample observations $\mathbf{y}$ that can be generated by the DGP. This expected value measures the expected loss of an estimator and is referred to in statistics as the **risk** of an estimator for parameter θ.

  Interpretation: If the loss function is calculated for a large number of different samples from the same DGP, the average is close to the risk.

- The risk regarding the squared loss function for a scalar parameter is also called **mean squared error** (**MSE**),
$$MSE\left(\tilde{\theta}(\mathbf{y})\right) := E\left[\left(\tilde{\theta}(\mathbf{y})-\theta_0\right)^2\right]. \tag{5.49}$$

  If all p parameters are considered together, the **matrix of mean squared errors** is obtained:
$$MSE\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})\right) := E\left[\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}_0\right)\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}_0\right)^T\right]. \tag{5.50}$$
The MSE matrix can be decomposed into two important components, which are defined below.

- **Expected value minimises MSE**

- **Bias** of an estimator $\tilde{\boldsymbol{\theta}}(\boldsymbol{y})$:
$$B\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})\right) := E\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right]-\boldsymbol{\theta}_0 \tag{5.51}$$

- **Variance-covariance matrix** / **Covariance matrix** / **Variance matrix** of an estimator $\tilde{\boldsymbol{\theta}}(\mathbf{y})$:
$$Var\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})\right) := E\left[\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})-E\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right]\right)\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})-E\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right]\right)^T\right] \tag{5.52}$$
The variance-covariance matrix is as follows in detail (for better readability, the dependence on the sample is not indicated, as is generally the case).
$$\begin{aligned} Var\left(\tilde{\boldsymbol{\theta}}\right) &:= E\left[\left(\tilde{\boldsymbol{\theta}}-E\left[\tilde{\boldsymbol{\theta}}\right]\right)\left(\tilde{\boldsymbol{\theta}}-E\left[\tilde{\boldsymbol{\theta}}\right]\right)^T\right] \\ &= \begin{pmatrix} Var(\tilde{\theta}_1) & Cov(\tilde{\theta}_1,\tilde{\theta}_2) & \cdots & Cov(\tilde{\theta}_1,\tilde{\theta}_p) \\ Cov(\tilde{\theta}_2,\tilde{\theta}_1) & Var(\tilde{\theta}_2) & \cdots & Cov(\tilde{\theta}_2,\tilde{\theta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\tilde{\theta}_p,\tilde{\theta}_1) & Cov(\tilde{\theta}_p,\tilde{\theta}_2) & \cdots & Var(\tilde{\theta}_p) \end{pmatrix}. \end{aligned} \tag{5.53}$$

- **Decomposition of the MSE matrix**: In general, the MSE matrix can be decomposed into the variance-covariance matrix of the estimator and the outer product of the biases:
$$MSE(\tilde{\boldsymbol{\theta}}(\mathbf{y})) = Var\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})\right)+B\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})\right)B\left(\tilde{\boldsymbol{\theta}}(\mathbf{y})\right)^T, \tag{5.54}$$

- If there is no bias, the estimator is called an **unbiased estimator**. Then, the expected value of the estimator regarding all possible samples corresponds to the parameter vector of the actual DGP.

  $$E\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right] = \boldsymbol{\theta}_0. \tag{5.55}$$

  Interpretation: Unbiasedness implies that for a large number of samples the average value of all estimates is very close to the true value.

- If an estimator is unbiased, i.e. $\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right] = \boldsymbol{\theta}_0$, the MSE is equal to the variance of the estimator.

- **Properties of variance-covariance matrices**

  – Variance-covariance matrices are **symmetric** and **always positive semidefinite**, but mostly positive definite, since due to their definition (1.8) holds.

  – The inverse of a variance-covariance matrix

  $$Var\left(\tilde{\boldsymbol{\theta}}\right)^{-1}$$

  is called the **precision matrix**.

  – **Comparison of variance-covariance matrices of two estimators $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$**

  If two estimators are unbiased and the MSE is used as the selection criterion, the estimator with the smaller variance of the two is chosen.

  In the scalar case ($p = 1$) this is easy, as both variances can be easily compared. If $p > 1$, one has to compare two variance-covariance matrices. The estimator $\hat{\boldsymbol{\theta}}$ has a “smaller” variance-covariance matrix than the estimator $\tilde{\boldsymbol{\theta}}$ if the following difference of the precision matrices

  $$Var(\hat{\boldsymbol{\theta}})^{-1} - Var(\tilde{\boldsymbol{\theta}})^{-1}$$

  is positive semidefinite and not zero. If both variance-covariance matrices are positive definite, it holds equivalently that the difference

  $$Var(\tilde{\boldsymbol{\theta}}) - Var(\hat{\boldsymbol{\theta}})$$

  is positive semidefinite and not zero (Davidson & MacKinnon 2004, Section 3.5, page 105 and Exercise 3.8).

  Interpretation: The property of a **positive semidefinite** difference of variance-covariance matrices means that any linear combination of the difference is non-negative. In particular

  $$Var(\tilde{\theta}_j) \geq Var(\hat{\theta}_j), \quad j = 1, \ldots, p. \tag{5.56}$$

- **Correlation matrix** of an estimator $\tilde{\boldsymbol{\theta}}$:

Cf. for the definition of a correlation (2.24)

$$Corr\left(\tilde{\boldsymbol{\theta}}\right) := \left(\frac{Cov(\tilde{\theta}_i, \tilde{\theta}_j)}{\left(Var(\tilde{\theta}_i)Var(\tilde{\theta}_j)\right)^{1/2}}\right)_{i=1,\dots,p,j=1,\dots,p} \tag{5.57}$$

The correlation matrix can also be represented in matrix notation as

$$Corr\left(\tilde{\boldsymbol{\theta}}\right) = \left(diag(Var(\tilde{\boldsymbol{\theta}}))\right)^{-1/2} Var(\tilde{\boldsymbol{\theta}}) \left(diag(Var(\tilde{\boldsymbol{\theta}}))\right)^{-1/2}, \tag{5.58}$$

where $diag(\boldsymbol{A})$ denotes a diagonal matrix that contains the diagonal elements of the matrix $\boldsymbol{A}$ on the diagonal.

Essential to the correlation matrix is that all elements on the diagonal are 1 and all non-diagonal elements are in the interval $[-1, 1]$.

R-commands

Calculating the correlation matrix from a covariance matrix with `cov2cor()`.

- **Desirable requirements** for an estimator:

  1. minimum risk or

  2. minimum risk with unbiasedness, i.e. minimum variance.

- **Efficiency** of an estimator: If the MSE is chosen as the selection criterion for the risk and if one considers estimators from a class that contains exclusively unbiased estimators, an estimator of the considered class is called efficient if it has the smallest possible variance in the sense determined above.

  Specifically: An estimator $\hat{\boldsymbol{\beta}}$ is the **efficient estimator in a class of unbiased estimators** $\tilde{\boldsymbol{\beta}}$, if it holds that the matrix of the difference of variance-covariance matrices $Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}})$ is **positive semidefinite**.

- **Knowledge** of the probability distribution of the estimator for each sample size n. This distribution is called the **exact probability distribution** of an estimator.

Important properties of an estimator for finite samples

- Unbiasedness
- Variance-covariance matrix and correlation matrix
- Efficiency or more generally risk
- Exact probability distribution

Example: the estimator of the expected value μ:

- A possible estimator of the expected value is given by the **arithmetic mean** of all sample observations
$$\hat{\mu}(\mathbf{y}) := \frac{1}{n}\sum_{t=1}^{n} y_t. \tag{5.59}$$
$\hat{\mu}(\mathbf{y})$ is a special case of the least squares estimator (6.5).
  - Calculating the **bias**:
  $$\begin{aligned} E\left[\hat{\mu}(\mathbf{y})\right] - \mu_0 &= E\left[\frac{1}{n}\sum_{t=1}^{n} y_t\right] - \mu_0 = \frac{1}{n}\sum_{t=1}^{n} E\left[y_t\right] - \mu_0 \\ &\overset{IID}{=} \frac{1}{n}\sum_{t=1}^{n}\mu_0 - \mu_0 = \mu_0 - \mu_0 = 0. \end{aligned}$$

    The estimator of the expected value is unbiased in the case of a random sample.
  - Calculating the **variance** of the estimator:
  $$Var\left(\hat{\mu}(\mathbf{y})\right) = Var\left(\frac{1}{n}\sum_{t=1}^{n} y_t\right) \overset{\text{IID}}{=} \frac{1}{n^2}\sum_{t=1}^{n} Var(y_t) = \frac{\sigma_0^2}{n} \tag{5.60}$$
  - **MSE** of the estimator: corresponds to the variance, since the estimator is unbiased. The MSE here also corresponds to the risk with regard to the quadratic loss function.

    Note: The risk of the expected value estimator decreases with increasing sample size n at the rate n.
  - **Distribution** of the estimator: Due to the model assumption (5.14), the estimator $\hat{\mu}(\mathbf{y}) = \frac{1}{n}\sum_{t=1}^{n} y_t$ is a linear combination of independently and identically normally distributed y_t. Therefore
  $$\mathbf{y} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \quad \text{mit } \boldsymbol{\mu} = \mu\boldsymbol{\iota}, \boldsymbol{\Sigma} = \sigma^2\mathbf{I},$$
    where $\boldsymbol{\iota}$ is a $(n \times 1)$ vector with ones. The sum $\boldsymbol{\iota}^T\mathbf{y} = \sum_{t=1}^{n} y_t$ is also normally distributed because of (2.33) with
  $$\boldsymbol{\iota}^T\mathbf{y} \sim N\left(\boldsymbol{\iota}^T\boldsymbol{\mu}, \boldsymbol{\iota}^T\boldsymbol{\Sigma}\boldsymbol{\iota}\right).$$
    Because of $\boldsymbol{\iota}^T\mu\boldsymbol{\iota} = n\mu$ and $\boldsymbol{\iota}^T\boldsymbol{\Sigma}\boldsymbol{\iota} = n\sigma^2$ we get
  $$\begin{aligned} \boldsymbol{\iota}^T\mathbf{y} &\sim N(n\mu, n\sigma^2) \quad \text{und} \\ \hat{\mu}(\mathbf{y}) &\sim N\left(\mu, \frac{\sigma^2}{n}\right). \end{aligned} \tag{5.61}$$
    So the estimator of the expected value $\hat{\mu}(\mathbf{y})$ is also normally distributed.

- Another possible estimator is

  $$\tilde{\mu}(\mathbf{y}) = \frac{1}{2}(y_1 + y_n). \tag{5.62}$$

  Again, determine all the properties and compare them. Show that in the comparison of the arithmetic mean (5.59) and (5.62), the former is efficient.

Asymptotic properties

In principle, the indicators considered so far - bias, variance, risk, MSE and distribution - depend on the sample size and the DGP. The dependence on parameters of the DGP can be very inconvienent, since these are unknown and thus the selection of a suitable estimator is not well possible. For this reason, indicators are also considered which, in such cases, are independent of the DGP in a suitable sense and at least guarantee that the properties of an estimator under consideration approach "‘desirable’" properties with increasing sample size, e.g. unbiasedness. One then "‘operates’" **asymptotics** or **asymptotic theory**: one indexes the estimator function with sample size n and investigates the properties of $\tilde{\boldsymbol{\theta}}_n$ for $n \to \infty$. One thus examines the convergence properties of a sequence of functions, see mathematical pre-course chapter 3.

Important asymptotic properties of an estimator

- Consistency
- Asymptotic variance
- Asymptotic efficiency
- Asymptotic distribution

The properties in detail:

- **Consistency**: if an estimator is biased, one can ask whether the magnitude of the bias decreases as the sample size increases and the estimator converges to the true parameter vector $\boldsymbol{\theta}_0$ when the sample size tends to infinity. “Convergence” here means **convergence of the estimator in probability**

  $$\operatorname*{plim}_{n\to\infty} \tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 \tag{5.63}$$

  or **almost sure convergence**

  $$\tilde{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0. \tag{5.64}$$

  Consistency implies that

  1. the estimator is asymptotically unbiased

     $$\lim_{n\to\infty} E\left[\tilde{\boldsymbol{\theta}}_n\right] = \boldsymbol{\theta}_0.$$

  2. the variance of the estimator for $n \to \infty$ tends to zero.

- If an estimator is not consistent, it is called **inconsistent**.

  Example: The arithmetic mean as an estimator of the expected value: (5.59) is consistent because it is unbiased for any n and the variance approaches zero with $n \to \infty$, see (5.60).

- **Asymptotic variance-covariance matrix**: In general, the variance-covariance matrix depends on the sample size n. If, as necessary for consistency, the variances and covariances of the estimator for $n \to \infty$ tend to zero, the variance-covariance matrices of different estimators cannot be compared for the case $n \to \infty$. Therefore, no non-degenerate probability distribution can exist for the case $n \to \infty$. Both require that the dependence of the variance-covariance matrix $Var(\hat{\boldsymbol{\theta}}_n)$ on the sample size for $n \to \infty$ can be eliminated.

  To prevent this dependence of the variance on $\hat{\boldsymbol{\theta}}$, one must multiply $\hat{\boldsymbol{\theta}}_n$ by a sample size dependent **factor** $r(n)$, **which prevents** $Var(r(n) \cdot \hat{\boldsymbol{\theta}}_n)$ **from converging to zero** or **from diverging to infinity**. It may also be necessary to have a specific factor $r_i(n)$ for each parameter estimator $\hat{\theta}_{in}$. These factors are called **convergence rates**. As a result, one obtains the **asymptotic variance-covariance matrix**, which is often noted as asyVar$(\tilde{\boldsymbol{\theta}}_n)$.

  Example: The arithmetic mean as an estimator of the expected value: The collapse or divergence of the variance of $\hat{\mu}_n - \mu_0$ is prevented by multiplying $\hat{\mu}_n - \mu_0$ by the factor $r(n) = \sqrt{n}$ which depends on the sample size. From $Var(\hat{\mu}_n) = n^{-1}\sigma_0^2$ follows

  $$Var\left(\sqrt{n}\,(\hat{\mu}_n - \mu_0)\right) = nVar\left(\hat{\mu}_n - \mu_0\right) = n\frac{\sigma_0^2}{n} = \sigma_0^2 = \text{asyVar}(\hat{\mu}_n). \tag{5.65}$$

  Thus σ_0^2 is the asymptotic variance of the arithmetic mean and the convergence rate is $\sqrt{n}$.

  Example: The inefficient expected value estimator (5.62)

  Show that for this expected value estimator $r(n) = 1$ holds. Thus, its rate of convergence is smaller than the rate of the arithmetic mean, which is why the latter is preferable.

- **Asymptotic distribution**:

  - The asymptotic distribution is the limit distribution that results for $n \to \infty$. Later this will be defined in more detail.

    Example: Estimator of the expected value: The distribution or density $f(\hat{\mu}; \mu_0, \sigma_0^2/n)$ of the estimator of the expected value $\hat{\mu}$ depends on the sample size, as its variance depends on the sample size, see (5.61).

    The normal distribution becomes independent of the sample size n if the asymptotic variance can be used. This is achieved by considering the sequence of random variables multiplied by the convergence rate:

$$\sqrt{n}\,(\hat{\mu}_n - \mu_0) \sim N(0, \sigma_0^2). \tag{5.66}$$

Since the distribution $N(0, \sigma_0^2)$ is independent of the sample size, it is also valid for $n \to \infty$ and thus the limit distribution

$$\sqrt{n}(\hat{\mu}_n - \mu_0) \xrightarrow{d} N(0, \sigma_0^2). \tag{5.67}$$

– **When is knowledge of the asymptotic distribution useful?**

If the normal distribution assumption cannot be made, the derivation in (5.66) will no longer work. Thus, if only

$$y_t \sim IID(\mu_0, \sigma_0^2), \quad t = 1, 2, \dots, n, \tag{5.68}$$

can be assumed, it is not possible to determine the **exact probability distribution** of the estimator

$$F_n(z) := P(\hat{\mu}_n \leq z).$$

However, if the asymptotic distribution is known in such a case, it can be used approximatively instead of the unknown exact distribution. For the present case, the asymptotic distribution exists, see section 5.5.2.

- **Asymptotic efficiency of an estimator**

  For two asymptotically normally distributed estimators $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$, both with convergence rate $r(n) = \sqrt{(n)}$, $\hat{\boldsymbol{\theta}}_n$ is asymptotically relatively more efficient than $\tilde{\boldsymbol{\theta}}_n$, if the difference of their asymptotic variance-covariance matrices $asyVar(\tilde{\boldsymbol{\theta}}_n) - asyVar(\hat{\boldsymbol{\theta}})_n$ is positive semidefinite (Wooldridge 2010, Definition 3.11). Asymptotic efficiency plays a role in chapter 14.

5.5. Tools for asymptotic analysis

5.5.1. Law of Large Numbers (LLN)

A law of large numbers states conditions under which the arithmetic mean converges in probability or even almost surely to the true mean.

- **Khinchin's Weak Law of Large Numbers (WLLN)** Let z_t, $t = 1, 2, \dots, n$, be an IID sequence of random variables with finite expected value μ. Then, it holds for the arithmetic mean $\hat{\mu}_n = n^{-1} \sum_{t=1}^{n} z_t$ that

$$\hat{\mu} \xrightarrow{P} \mu, \tag{5.69a}$$

$$\text{or} \quad \operatorname{plim}(\hat{\mu}) = \mu. \tag{5.69b}$$

(See e. g. Davidson (1994, Theorem 23.5) — proof too difficult.)

- Two **versions of LLN**
  - Weak LLN (WLLN):
  $$\hat{\mu} \xrightarrow{P} \mu.$$
  - Strong LLN (SLLN):
  $$\hat{\mu} \xrightarrow{a.s.} \mu.$$
- There are also LLN for various non-IID cases, see e. g Davidson (2000, Section 3.2).
- **Note** that z_t can also be a function of another random variable, for example the power of a random variable or the product of two different random variables.

Example: Estimator of the expected value: If the conditions of one of the laws of large numbers are satisfied, the arithmetic mean is an **consistent estimator** of the expected value.

If a random sample is present and the DGP has a finite expected value, then, for example, the weak law of large numbers of Khinchin applies

$$\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} y_t \xrightarrow{P} \mu_0 \quad \text{or} \quad \text{plim}_{n\to\infty}\, \hat{\mu}_n = \mu_0.$$

Example: Comparison of two estimators of the expected value using Monte Carlo simulation

In a Monte Carlo simulation, the estimation properties of the arithmetic mean (5.59) and the inefficient mean estimator (5.62) are compared.

Structure:

- **DGP**

$$y_t = \mu_0 + \sigma_0 u_t, \quad u_t = (\varepsilon_t - m)/\sqrt{2m}, \quad \varepsilon_t \sim i.i.d.\chi^2(m) \tag{5.70}$$
$$\mu_0 = 1, \quad \sigma_0 = 2, \quad m = 1 \tag{5.71}$$

  The density of errors u_t is asymmetric because the ε_t are drawn from a χ^2 distribution with m degrees of freedom and are then standardised.

- $R = 10000$ realisations of samples with sample sizes $n = 10, 50, 100, 500$ respectively. Note: The conditions of the weak law of large numbers (LLN) are fulfilled.
- Calculation of the arithmetic mean and the standard deviation for each sample size and each estimator

R code, see section A.2, page 329.

R output

```
      N mu_hat_mean  mu_hat_sd mu_tilde_mean mu_tilde_sd
[1,]  10   1.0014164 0.31831308     1.0020431   0.7099616
[2,]  50   0.9991498 0.14162425     0.9960143   0.7054847
[3,] 100   0.9990515 0.09997354     0.9900695   0.6896356
[4,] 500   1.0003874 0.04474699     1.0058432   0.7074540
```

It can be seen that both estimators are unbiased. The standard deviation of the alternative estimator `mu_tilde` (5.62) is larger than the standard deviation of the arithmetic mean `mu_hat` for any sample size. Moreover, the standard deviation of the arithmetic mean becomes smaller as the sample size increases. The first result illustrates the efficiency of the arithmetic mean and the second result illustrates that the arithmetic mean is a consistent estimator.

The distributions of the estimators can be seen in the figures 5.1 and 5.2.

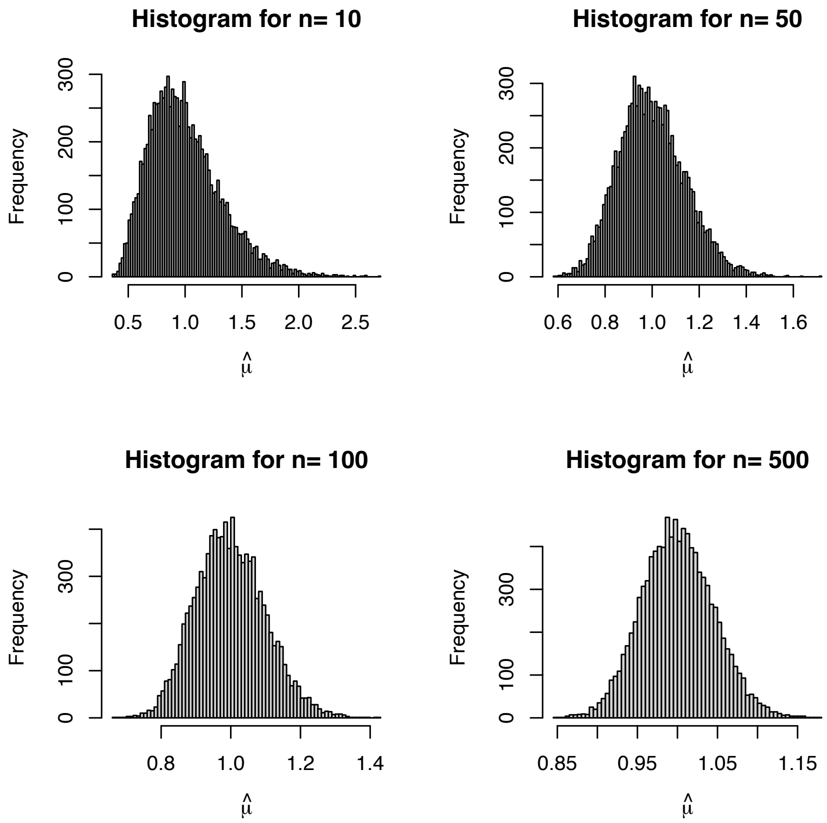


Figure 5.1.: Histograms of the arithmetic mean (R program see section A.2, page 329) DGP see equation (5.70)

It is noticeable that

- the inefficient estimator has a skewed distribution (like the errors) regardless of the sample size, but the density of the arithmetic mean becomes more symmetric as the sample size increases — and, as will be shown in the next

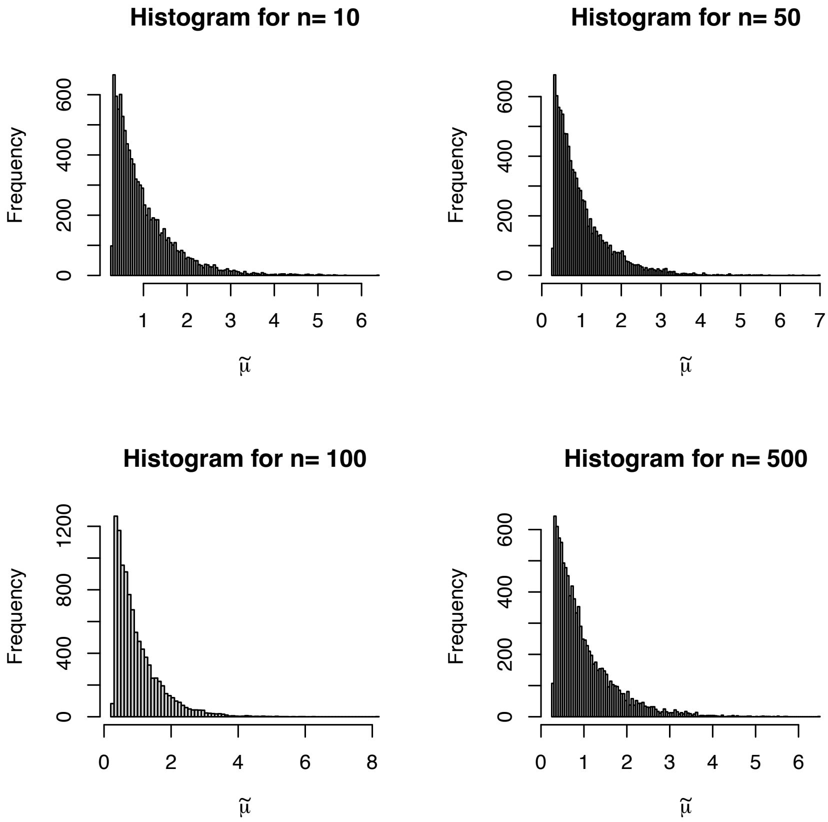


Figure 5.2.: Histograms of the inefficient expected value estimator (5.62) (R program see section A.2, page 329) DGP see equation (5.70)

section, converges to the density of the normal distribution.

5.5.2. Central limit theorems

Preface: In principle, a central limit theorem is of central importance to be able to determine a limit distribution for an estimator in very general cases. There are different versions of central limit theorems, which differ in their assumptions.

Example: Estimator of the expected value:

- If a random sample is available but it is not known which distribution the DGP has, i.e. which distribution e. g. the return of the DAX has, then the derivation of the limit distribution via (5.66) does not work.

- Since the existence of an asymptotic variance requires the rate of convergence $r(n) = \sqrt{(n)}$, we have to ask against which asymptotic distribution the sequence of random variables $\sqrt{n}(\hat{\mu}_n - \mu_0)$ converges if y_t is, for example, IID but not normally distributed?

The answer for this case is provided by the **central limit theorem (CLT)** of Lindeberg and Lévy.

- **Central limit theorem for IID random variables** (Lindeberg-Lévy Theorem) Let $y_t \sim IID(\mu_0, \sigma_0^2)$, $t = 1, 2, \dots$, $|\mu_0| < \infty, 0 < \sigma_0^2 < \infty$. Then, for the estimator of the expected value $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^n y_t$:

$$\sqrt{n}(\hat{\mu}_n - \mu_0) \xrightarrow{d} N(0, \sigma_0^2).$$

Proof: (For a proof idea see, e.g. Hendry (1995, Section A.5)) □

Remarks:

- Alternatively, one can write

$$\sqrt{n}(\hat{\mu}_n - \mu_0) \xrightarrow{d} z, \quad z \sim N(0, \sigma_0^2),$$

but not (as erroneously in Davidson & MacKinnon (2004, Section 4.5, p. 149))

$$\text{plim}_{n\to\infty}\sqrt{n}(\hat{\mu}_n - \mu_0) = z \sim N(0, \sigma_0^2),$$

because this probability limit does not exist; see, for example, Davidson (1994, Section 23.1) for a proof.

- Regardless of the nature of the marginal distribution of y_t, the $\sqrt{n}$ **scaled estimator of the expected value converges in distribution to a normal distribution** as long as y_t has a finite variance. The **estimator of the expected value is said to be asymptotically normally distributed**.

- **Alternative notation of the central limit theorem**: Let denote the random variable $X_n = \hat{\mu}_n$. Then the **exact probability distribution** of the arithmetic mean for the sample size n is given by

$$F_n(z) := P(\hat{\mu}_n \le z).$$

The central limit theorem states that the sequence of distribution functions $F_n(z)$ converges pointwise to the distribution function $F(z) = \Phi(z)$

$$\lim_{n\to\infty} F_n(z) = \Phi(z).$$

- The central limit theorem says nothing about how well the asymptotic distribution approximates the exact distribution $F_n(z)$ for a given sample size n. In order to gain information about this, computer simulations are generally necessary.

Example: Comparison of two estimators of the expected value using Monte Carlo simulations (continued from section 5.5.1) The histograms of the arithmetic mean in figure 5.1 illustrate well the central limit theorem. The histograms for the inefficient estimator in figure 5.2 indicate that no central limit theorem applies. The reason for this is that regardless of the sample size, exactly two observations are always used in the estimation and thus no CLT can apply.

- **Central limit theorem for heterogeneous but stochastically independent random variables** Often the y_t are not IID, but are only independently but not identically distributed, for example, if they have a different variance, $y_t \sim (\mu_0, \sigma_t^2)$, $t = 1, 2 \ldots$. Then the following holds for the variance of $\sqrt{n}\hat{\mu}_n$,

$$Var(\sqrt{n}\hat{\mu}_n) = Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} y_t\right) = \frac{1}{n}\sum_{t=1}^{n} Var(y_t) = \frac{1}{n}\sum_{t=1}^{n}\sigma_t^2.$$

Provided the $Var(y_t)$ satisfy some conditions, e. g $0 < Var(y_t) < c < \infty$, for all $t = 1, 2, \ldots$, a central limit theorem holds

$$\sqrt{n}(\hat{\mu}_n - \mu_0) \xrightarrow{d} N\left(0, \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} Var(y_t)\right). \tag{5.72}$$

Conditions on the sequence of variances are necessary to exclude the following cases:

– For example, if it held for a fixed $0 < a < 1$ that $Var(y_t) = \sigma_0^2 a^t \to 0$ mit $t \to \infty$, then $\sum_{t=1}^{\infty} Var(y_t) = \sigma_0^2 \frac{1}{1-a}$ and thus

$$Var(\sqrt{n}\hat{\mu}_n) = \frac{1}{n}\sigma_0^2\frac{1}{1-a} \to 0 \text{ for } n \to \infty,$$

so the variance of $\sqrt{n}\hat{\mu}_n$ vanishes asymptotically. Thus of course no (meaningful) limit distribution is possible.

– If $Var(y_t) = \sigma_0^2 t \to \infty$ were to apply accordingly, then one obtains

$$Var(\sqrt{n}\hat{\mu}_n) = \frac{1}{n}\sigma_0^2\frac{n(n+1)}{2} \to \infty \text{ with } n \to \infty.$$

Conditions that ensure that a limit distribution exists are often referred to as **regularity conditions**.

> **Example: Estimator of the expected value:** This central limit theorem is useful when the unconditional variance of DAX returns depends on time, for example, the day of the week.

- **Central limit theorems for vectors**

– ♯ **Cramér-Wold Device**: For a sequence of random vectors $\mathbf{x}_n$ it holds that

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x}$$

if and only if for all feasible vectors $\boldsymbol{\lambda}$ holds:

$$\boldsymbol{\lambda}^T\mathbf{x}_n \xrightarrow{d} \boldsymbol{\lambda}^T\mathbf{x}.$$

- **Multivariate limit theorem**: Given the independently distributed $(r \times 1)$ random vectors $\mathbf{v}_t$ with expected value $\boldsymbol{\mu}_0$ and variance-covariance matrix $Var(\mathbf{v}_t)$. Then, under appropriate regularity conditions, it holds for the estimator of the multivariate expected value $\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{t=1}^n \mathbf{v}_t$ that

$$\sqrt{n}\left(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n Var(\mathbf{v}_t)\right). \tag{5.73}$$

5.6. Fundamentals of tests

Statistical tests are used

- to test economic hypotheses,
- in model specification and model validation of econometric models (relevant regressor variables, functional form of the regression function, violation of assumptions,...).

Statistical test:

- Sample-based decision procedure to decide whether a hypothesis must be rejected.
- The hypothesis must relate to properties of probability distributions contained in the model under consideration.
- There are exactly two alternatives for deciding: not to reject the hypothesis or to reject it.

Components of a statistical test

:

1. Pair of hypotheses
2. Test statistic
3. Decision rule
4. Decision and interpretation

Zu 1.: Two **disjoint hypotheses** about one or more elements of the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ denotes the parameters of the probability distributions under consideration.

- **Null hypothesis** $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}$.
- **Alternative hypothesis** $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$.

The union of the two hypotheses $\boldsymbol{\Theta}_{H_0} \cup \boldsymbol{\Theta}_{H_1} = \boldsymbol{\Theta}$ covers the entire parameter space $\boldsymbol{\Theta}$. (Cf. on parameter space section 5.1.)

Example: Test concerning the expected value of the DAX returns:

- Economic question: Is the average daily return of the DAX zero?
- Statistical test is to be carried out within the framework of the model considered so far,
$$y_t \sim NID(\mu, \sigma^2), \quad \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+. \tag{5.14}$$
This assumption defines the set of possible probability distributions: normal distributions with variance $\sigma^2 > 0$ and expected value μ.
- General formulation of the pair of hypotheses concerning the expected value μ:
$$H_0 : \mu = \mu_{H_0} \quad \text{versus} \quad \mathrm{H}_1 : \mu \neq \mu_{H_0}.$$
In the present case, $\mu_{H_0} = 0$ and thus
$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$
- We have not yet said anything about the other model parameter, the variance σ^2. The full formulation of the pair of hypotheses includes the entire parameter vector $\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \in \boldsymbol{\Theta} = (\mathbb{R} \times \mathbb{R}^+)$:
$$H_0 : \boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \in \boldsymbol{\Theta}_{H_0} = \left(\{\mu_{H_0}\} \times \mathbb{R}^+\right) \qquad \text{versus}$$
$$H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1} = \left((\mathbb{R} \backslash \{\mu_{H_0}\}) \times \mathbb{R}^+\right)$$
- If the expected value μ were known, it could be decided immediately whether the null hypothesis is correct.
- In practice, a decision can only be made on the basis of a sample and an *estimate* $\hat{\mu}(y_1, y_2, \ldots, y_n)$ of the expected value. A statistical test provides this decision. It should fulfil certain optimality criteria. More on this later.

Zu 2.: A **test statistic** λ is a function calculated from the sample values $\mathbf{y}$: $\lambda = \lambda(\mathbf{y})$. Note: Before observing a sample, a test statistic is a random variable, after observing a sample it is a realisation of a random variable, i.e. a number.

Zu 3.: A **decision rule** that determines for which values of λ the **null hypothesis** H_0 **is rejected** and for which values the **null hypothesis is not rejected**. More precisely: The range of values of λ is divided into two disjoint subranges:

- **Rejection region, critical region** $\mathcal{C}$ If the test statistic λ is within the critical region, H_0 is rejected:
$$\text{Reject } H_0 \text{ if } \lambda \in \mathcal{C}.$$
- **Non-rejection region** If the test statistic λ is within the non-rejection region, H_0 is *not rejected*:
$$\text{Do not reject } H_0 \text{ if } \lambda \notin \mathcal{C}.$$
- **Critical values**: One or more boundaries c between rejection and non-rejection region.

Note: Instead of the symbol λ, the symbol t is typically used for t-tests or the symbol F is often used for F-tests.

Example: Test concerning the expected value (mean) of DAX returns — continued:

The null hypothesis can be tested with a t-test as follows. The individual elements are then derived and justified:

Zu 2.: Test statistic of the t-test:

$$t(\mathbf{y}) = \frac{\hat{\mu} - \mu_{H_0}}{\hat{\sigma}_{\hat{\mu}}} = \frac{\left(\frac{1}{n}\sum_{t=1}^{n} y_t\right) - \mu_{H_0}}{\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2 \frac{1}{n}}} \tag{5.74}$$

Zu 3.:

- Rejection region: $\mathcal{C} = (-\infty, -1.96) \cup (1.96, \infty)$
- Non-rejection region: $(-1.96, 1.96)$
- Critical values: $c_l = -1.96, c_r = 1.96$.
- Decision rule: Reject H_0 if $t(\mathbf{y}) \in \mathcal{C}$.

Performing the test:

- Sample: Daily returns of the DAX from 25/03/1993 to 30/09/2015, a total of 5652 observations (R program see section A.3, page 330, data `dax19930325_20150930.xlsx`):
- For $H_0 : \mu_{H_0} = 0$ and $\hat{\mu} = 0.00004130056$ and $\hat{\sigma}_{\hat{\mu}} = 0.00002342752$ we get the test statistic

$$t(\mathbf{y}) = \frac{0.00004130056 - 0}{0.00002342752} = 1.762908$$

- $t(\mathbf{y}) \in \mathcal{C} \Rightarrow$ Do not reject H_0.

What is the probability of a wrong decision?

Properties of a test:

- **Type I error**: The type I error of a test indicates the probability with which the null hypothesis H_0 is rejected for a sample that has not yet been collected, although H_0 is correct in the population:

$$\begin{array}{ll} \text{Intuitive (sloppy) notation:} & P\left(\text{Reject } H_0|H_0 \text{ is true}\right) \\ \text{Exact notation:} & \alpha(\boldsymbol{\theta}) = P\left(\lambda \in \mathcal{C}; \boldsymbol{\theta}\right), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}. \end{array} \tag{5.75}$$

  Note: The type I error may depend on $\boldsymbol{\theta}$!

  Example: One-tailed t-test, see later.

- **Type II error or β-error**: The type II error indicates the probability with which H_0 is *not* rejected even though H_0 is false:

$$\begin{array}{ll} \text{Intuitive (...) notation:} & P\left(\text{Do not reject } H_0|H_1 \text{ is true}\right) \\ \text{Exact notation:} & \beta(\boldsymbol{\theta}) = P\left(\lambda \notin \mathcal{C}; \boldsymbol{\theta}\right), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}. \end{array} \tag{5.76}$$

- **Power function** of a test: The power function of a test indicates the rejection probability for a specific parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= P\left(\text{Reject } H_0; \boldsymbol{\theta}\right) \\ &= 1 - P\left(\lambda \notin \mathcal{C}; \boldsymbol{\theta}\right), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}. \end{aligned} \tag{5.77}$$

  Note: The power function is defined for the entire parameter space $\boldsymbol{\Theta}$.

- **Power** π of a test: The power of a test indicates the probability $\pi(\boldsymbol{\theta})$ of rejecting the null hypothesis for a specific $\boldsymbol{\theta}$, **if** $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$:

$$\pi(\boldsymbol{\theta}) = 1 - \beta(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}.$$

- **Size**) of a test: In many cases, the type I error depends on $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}$. The **supremum** of the type I errors for all possible $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}$ is called the **size** of a test:

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}} P\left(\lambda \in \mathcal{C}; \boldsymbol{\theta}\right) \tag{5.78}$$

Determining the critical region $\mathcal{C}$

- **Test distribution**: $P(\lambda \leq x; \boldsymbol{\theta})$ — necessary to determine the power function $\pi(\boldsymbol{\theta})$ (5.77) of a test.
  - under $H_0 : P(\lambda \leq x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}$ — necessary for determining the critical region $\mathcal{C}$.
  - under $H_1 : P(\lambda \leq x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$ — necessary for determining the power of a test.
- **Ideally**, the type I error should be as small as possible and at the same time the power function of a test should be as large as possible. Unfortunately, this is not possible. Therefore, one bounds the type I error and then wants to maximise the power $\pi(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$.

- **Level of significance** (**level**): Therefore, a **level of significance** α is **specified** which bounds the type I error:

$$P(\text{Reject } H_0; \boldsymbol{\theta}) = P(\lambda \in \mathcal{C}; \boldsymbol{\theta}) \leq \alpha \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_0}. \tag{5.79}$$

  From this condition, the rejection region $\mathcal{C} = \mathcal{C}(\alpha)$ can be determined.

- **If there are several tests to choose from** that meet the significance level α, then choose the test that maximises the power function $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$.

- See the following example for the derivation of the t-statistic and the relevant critical region. Tests for testing hypotheses with multiple parameters are derived in 11.3.2.

Example: test concerning the expected value — continued:

Deriving the test statistic for testing a null hypothesis regarding the expected value with a known standard deviation and determining the critical region

1. Under the assumptions made, the estimator of the expected value is normally distributed, see (5.61)

$$\hat{\mu}(\mathbf{y}) \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{5.61}$$

2. However, the distribution depends on unknown parameters. This is avoided by standardizing:

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{5.80}$$

   Case A: $\sigma = \sigma_0$ known: Under $H_0 : \mu = \mu_{H_0}$, $\frac{\hat{\mu}-\mu}{\sigma_0/\sqrt{n}}$ can be calculated and one obtains a standard normally distributed test statistic

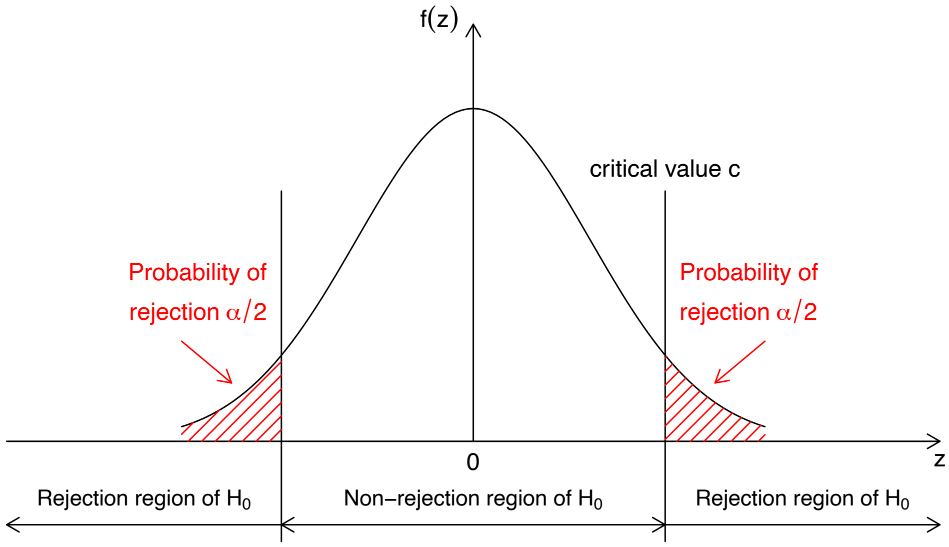
$$z(\mathbf{y}) = \frac{\hat{\mu} - \mu_{H_0}}{\sigma_0/\sqrt{n}} \sim N(0, 1). \tag{5.81}$$

   Case B: σ unknown: see general derivation of (5.74) in the context of the OLS regression model in section 11.3.1.

3. Determining the critical region $\mathcal{C}$ for case A (case B proceeds analogously):

   a) Determining the significance level α.

   b) Figure of the probability density of $z(\mathbf{y})$ under H_0:


f(z)
critical value c
Probability of rejection α/2
Probability of rejection α/2
0
z
Rejection region of H0
Non−rejection region of H0
Rejection region of H0

When should H_0 be rejected?

Intuition: If z is very large (or very small),

i. then the estimated expected value $\hat{\mu}$ is far from μ_{H_0} (under H_0). This could provide evidence for $H_1 : \mu \neq \mu_{H_0}$. One should then reject H_0.

ii. Or the standard deviation $\sigma_{\hat{\mu}} = \sigma_0/\sqrt{n}$ of the estimated deviation is small compared to the difference $\hat{\mu} - \mu_{H_0}$.

That is, one should reject H_0 if z is very large or very small.

The critical region is therefore

$$\mathcal{C} = (-\infty, c_l) \cup (c_r, \infty).$$

Determine the critical values c_l, c_r using the given significance level (5.79). Usually, the given significance level α is divided symmetrically on both sides. The type I error is smaller than or equal to the significance level α if the following applies

$$P(z < c_l; \mu_{H_0}, \sigma_0) \leq \alpha/2 \quad \text{and} \quad P(z > c_r; \mu_{H_0}, \sigma_0) \leq \alpha/2, \tag{5.82}$$

$$F(z; \mu_{H_0}, \sigma_0) \leq \alpha/2 \quad \text{and} \quad 1 - F(z; \mu_{H_0}, \sigma_0) \leq \alpha/2. \tag{5.83}$$

Under H_0, z is standard normally distributed (2.5), such that

$$\Phi(c_l) \leq \alpha/2 \quad \text{and} \quad (1 - \Phi(c_r)) \leq \alpha/2. \tag{5.84}$$

Ideally, the equal sign should apply because then the significance level controls the type I error exactly. The critical value c_l just corresponds to the $\alpha/2$-quantile (2.9) of the standard normal distribution

$$c_l = q_{\alpha/2} = \Phi^{-1}(\alpha/2), \quad c_r = q_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2).$$

Due to the symmetry of the normal distribution density, $c_l = -c_r = -c$ is obtained. Thus, for $\alpha = 0.05$ one obtains the critical values $-c = -1.96$ and $c = 1.96$, respectively. See e. g. Table G.1 in Wooldridge (2009)) or calculate c with the R command `c <- qnorm(p=1-alpha/2)`, where `alpha` indicates the level of significance.

Calculating the power

- General procedure: First determine the power function, i. e. the density function of the test statistic for an arbitrary $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Since the power of the test generally depends on $\boldsymbol{\theta}$, one calculates the power of the test for all or selected $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$. The power is calculated by determining the area under the density function in the critical region.

Example: Test concerning the expected value — continued:

In the following, σ_0 is still assumed to be known:

- Under both H_0 and H_1, given the expected value μ_0 and the standard deviation σ_0 of the DGP, the following applies according to (5.80),

$$\frac{\hat{\mu} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0,1).$$

By expanding one obtains

$$\frac{\hat{\mu} + \mu_{H_0} - \mu_{H_0} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\hat{\mu} - \mu_{H_0}}{\sigma_0/\sqrt{n}} + \frac{\mu_{H_0} - \mu_0}{\sigma_0/\sqrt{n}} = \underbrace{\frac{\hat{\mu} - \mu_{H_0}}{\sigma_0/\sqrt{n}}}_{z(\mathbf{y})} - \underbrace{\frac{\mu_0 - \mu_{H_0}}{\sigma_0/\sqrt{n}}}_{m}$$

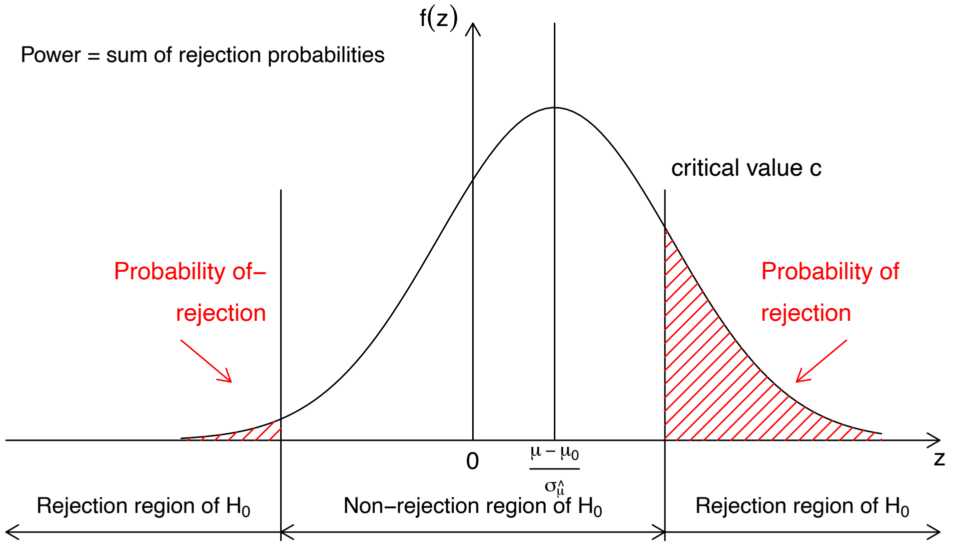
and thus one gets

$$z(\mathbf{y}) = \frac{\hat{\mu} - \mu_{H_0}}{\sigma_0/\sqrt{n}} \sim N\left(\frac{\mu_0 - \mu_{H_0}}{\sigma_0/\sqrt{n}}, 1\right),$$

since $X \sim N(m,1)$ is equivalent to $X - m \sim N(0,1)$.

- **Conclusion**: If H_1 holds, the density as well as the distribution of the test statistic $z(\mathbf{y})$ is shifted by $(\mu_0 - \mu_{H_0})/(\sigma_0/\sqrt{n})$.

- In the figure of the density under H_1 (for a specific $\mu_0 \neq \mu_{H_0}$) the power is obtained from the sum of the two shaded areas: $\pi(\boldsymbol{\theta}) = P(z < -c; \boldsymbol{\theta}) + P(z > c; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1}$.


f(z)
Power = sum of rejection probabilities
critical value c
Probability of–
rejection
Probability of
rejection
0
$\frac{\mu - \mu_0}{\sigma_{\hat{\mu}}}$
z
Rejection region of H_0
Non–rejection region of H_0
Rejection region of H_0

- For a given $\sigma_{\hat{\mu}}$, the power of the test increases with the difference between the value of the null hypothesis μ_{H_0} and the true value μ_0. It is then "easier" to reject a false null hypothesis.
- For given parameter values $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the power function can be calculated and plotted.

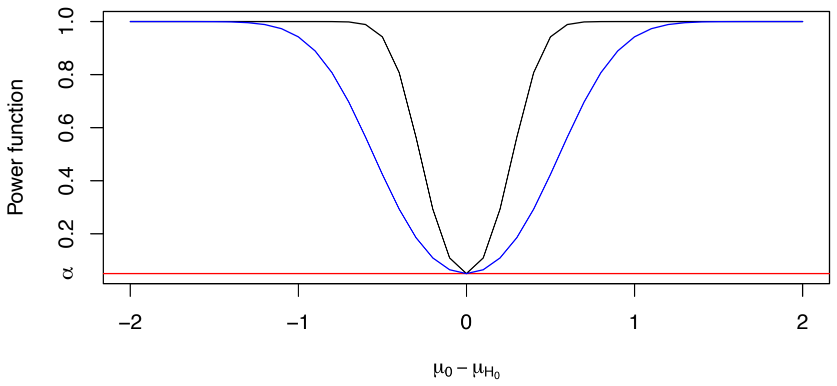


Figure 5.3.: Power function for a test concerning the expected value (R program see section A.3, page 333) Parameters used: level of significance $\alpha = 0.05$ (red horizontal line), $\mu_0 - \mu_{H_0} \in [-2, 2]$, $n = 50$, $\sigma = 1$ (black line), $\sigma = 2$ (blue line).

- **Properties of the power** : The power of a test increases with
  - greater distance between correct value and null hypothesis and/or
  - decreasing standard deviation σ and/or
  - the sample size n.

Conclusion: Statistical tests require at least knowledge of the probability distribution of the

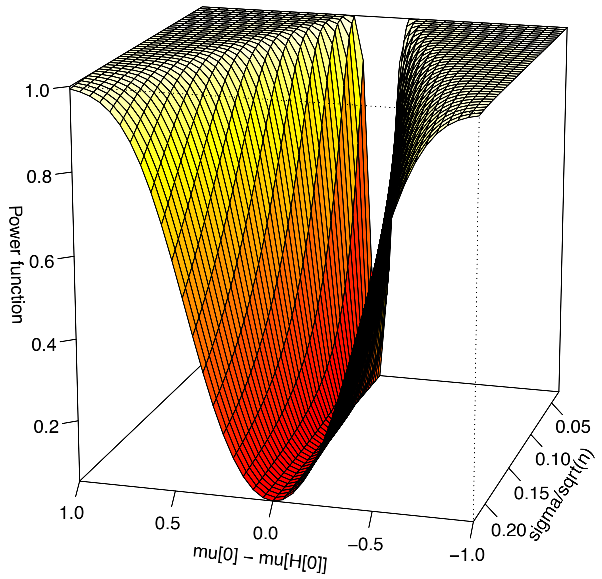


Figure 5.4.: Illustration of the power function for z given $\alpha = 0.05$ on a grid (`R` program see section A.3, page 334) Parameter range $\mu_0 - \mu_{H_0} \in [-1, 1]$, $\sigma_{\hat{\mu}} = \sigma_0/\sqrt{n} \in [1/\sqrt{20}, 1/\sqrt{1000}]$

test statistic under H_0, but to determine the power, knowledge of the probability distribution under H_1 is also required.

Test concerning the expected value with unknown variance: t-statistic (5.74) (case B, page 129)

In practice, the variance σ^2 is unknown. The general procedure is illustrated by the test concerning the expected value.

Example: Test concerning the expected value — continued:

Solution: The variance σ^2 is estimated using

$$s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y})^2$$

(cf. (9.25)). Substituting s into (5.80) results in

$$\hat{\sigma}_{\hat{\mu}} = \frac{s}{\sqrt{n}}$$

and the so-called **t-statistic** (5.74)

$$t(\mathbf{y}) = \frac{\hat{\mu} - \mu_{H_0}}{\hat{\sigma}_{\hat{\mu}}} = \frac{\left(\frac{1}{n}\sum_{t=1}^{n} y_t\right) - \mu_{H_0}}{\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2 \frac{1}{n}}},$$

which has already been used for the test of the expected value. This t-statistic is no longer normally distributed. In section 11.3.1 it is shown that this t-statistic follows a t-distribution with $n-1$ degrees of freedom (t_{n-1} for short).

Thus

$$t(\mathbf{y}) = \frac{\hat{\mu} - \mu_{H_0}}{\hat{\sigma}_{\hat{\mu}}} \sim t_{n-1}.$$

For the properties of the (symmetrical) t-distribution (2.36) see Part I. Mathematical Pre-course. To obtain the critical values

$$P(t < -c|H_0) = \frac{\alpha}{2} \quad \text{and} \quad P(t > c|H_0) = \frac{\alpha}{2},$$

one can look for example in Table G.2 in Wooldridge (2009) or calculate them with the R command `c <- qt(p=1-alpha/2,df=n-k)`, where `alpha` indicates the level of significance and `k` indicates the number of the estimated parameters, here $k = 1$.

One- and two-tailed hypothesis tests with the t-test

The possibility of one-tailed tests exists when **one** element θ_j of the parameter vector $\boldsymbol{\theta}$, such as the expected value μ, is to be tested. For both the one-tailed and the two-tailed test, the t-statistic is generally

$$t(\mathbf{y}) = \frac{\hat{\theta}_j - \theta_{j,H_0}}{\hat{\sigma}_{\hat{\theta}_j}}.$$

- **Two-tailed tests**

$$H_0 : \theta_j = \theta_{j,H_0} \quad \text{versus} \quad H_{H_1} : \theta_j \neq \theta_{j,H_0}.$$

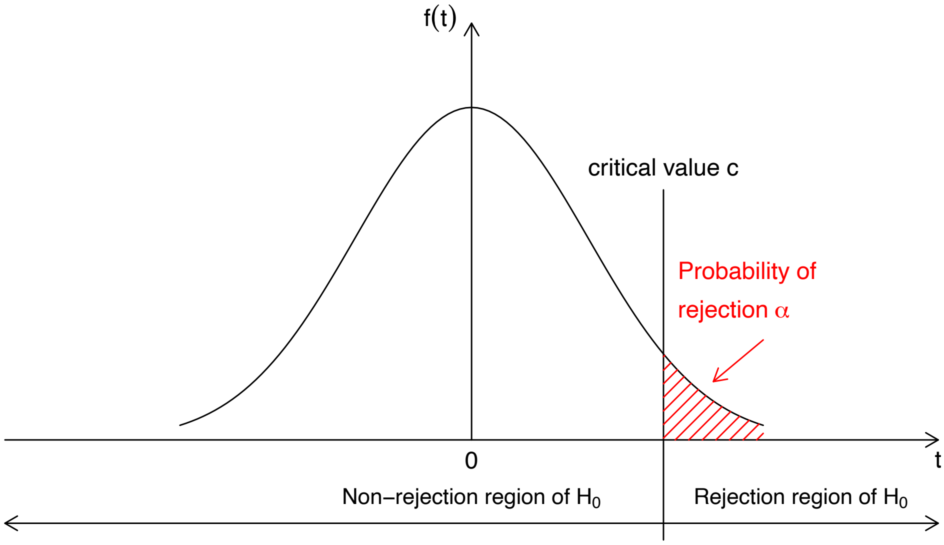
- **One-tailed tests**
  - **Right-tailed test**

$$\mathrm{H}_0 : \theta_j \leq \theta_{j,H_0} \quad \text{versus} \quad \mathrm{H}_1 : \theta_j > \theta_{j,H_0}$$

  Note: Often, as in Wooldridge (2009), one reads $H_0 : \theta_j = \theta_{j,H_0}$ versus $H_1 : \theta_j > \theta_{j,H_0}$. This notation is not precise, since every possible parameter value must belong either to H_0 or to H_1. However, this is not clear from this notation.

  ∗ **Critical value**:

  Density of the t-test statistic for $\theta_{j,0} = \theta_{j,H_0}$, where $\theta_{j,0}$ denotes the parameter value of the DGP:



No rejection region is needed on the left-hand side because all $\theta_j < \theta_{j,H_0}$ are elements of H_0 and thus belong to the non-rejection region.

Type I error and size of a one-tailed test: Assume that $\theta_{j,0} < \theta_{j,H_0}$ holds for $\theta_{j,0}$ of the DGP such that H_0 applies. Since the position of the density of the test statistic $t(\mathbf{y})$ depends on $\frac{\theta_{j,0}-\theta_{j,H_0}}{\sigma_{\hat{\theta}_j}}$ (cf. Figure on page 131 for $\theta_j = \mu$), the density for $\theta_{j,0} - \theta_{j,H_0} < 0$ is to the left of the density for $\theta_j = \theta_{j,H_0}$. Accordingly, the shaded area, i.e. the type I error, is smaller in the first case than in the second case. Thus, the type I error for $\theta_j = \theta_{j,H_0}$ just corresponds to the size (5.78) of the test. Since the chosen level of significance α specifies the size of a test, the critical value for $\theta_j = \theta_{j,H_0}$ is therefore determined.

∗ **Decision rule**:

$$t > c \quad \Rightarrow \text{Reject } H_0.$$

Example: Test concerning the expected value (mean) of the DAX returns — continued: Are the DAX returns positive?

∗ Pair of hypotheses:

$$H_0 : \mu \leq 0 \quad \text{versus} \quad H_1 : \mu > 0$$

∗ Determining the critical value: For $\alpha = 0.05$, the critical value 1.646179 is obtained from the t-distribution with 1151 degrees of freedom `c <- qt(p=0.95,df=1151)`.

∗ Calculation of the test statistic: As in the case of the two-tailed test (5.74):

$$t(\mathbf{y}) = \frac{0.00004130056 - 0}{0.00002342752} = 1.762908$$

∗ **Test decision**: Since

$$t = 1.763 > c = 1.645,$$

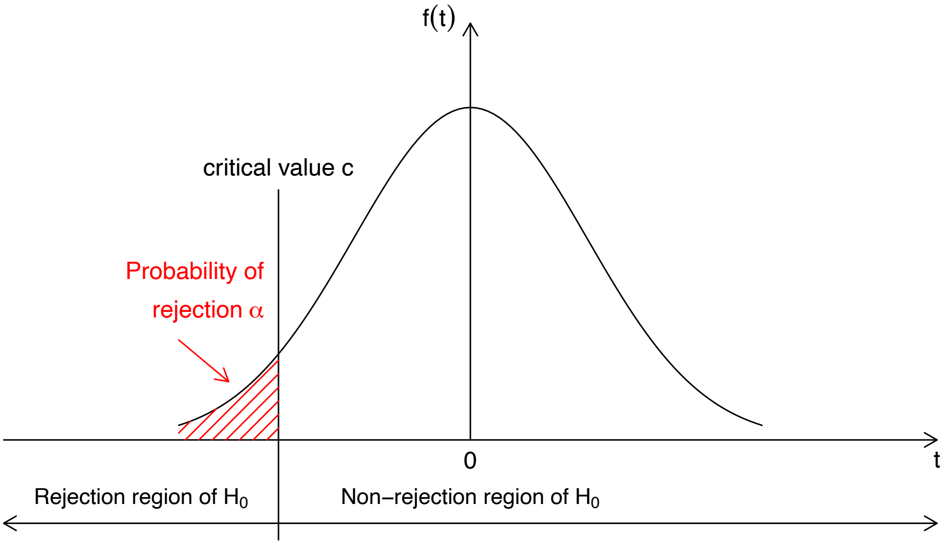
the null hypothesis is rejected. There is statistical evidence for a positive expected value of the daily DAX returns.

∗ What test result do you get for a significance level of 1%?

- **Left-tailed test**

$$H_0 : \theta_j \geq \theta_{j,H_0} \quad \text{versus} \quad H_1 : \theta_j < \theta_{j,H_0}.$$

Density of the t-test statistic for $\theta_{j,0} = \theta_{j,H_0}$:



Proceed as for right-tailed alternative hypothesis, only mirror-inverted.

- **Conclusion**: Difference of one-tailed and two-tailed tests: area of the given significance level is one-tailed concentrated or two-tailed halved.

- **Advantage of one-tailed tests**

  - **Since statistical tests cannot confirm hypotheses, but only reject them**, the alternative hypothesis is usually chosen to reflect the conjecture that is to be statistically "supported".

    Thus, if the conjecture concerns only one side because the other side is not of interest or can be ruled out for economic reasons, a one-tailed test is possible.

  - With the one-tailed test, the given significance level can be concentrated on one side, so that the critical value becomes smaller in absolute value compared to the two-tailed test and a rejection of the null hypothesis becomes more likely and thus the power increases if the null hypothesis is false in the population.

    Example: Test concerning the expected value (mean) of the DAX returns — continued: If one is only interested in whether the DAX returns are positive, a one-tailed test as above is possible. While $H_0 : \mu = 0$ in the two-tailed test given a significance level of 0.05 cannot be rejected, this is possible in the right-tailed test.

  - **Important**: However, a one-tailed test is only justified if the side included in the null hypothesis is not interested or can be excluded for economic reasons.

p-values

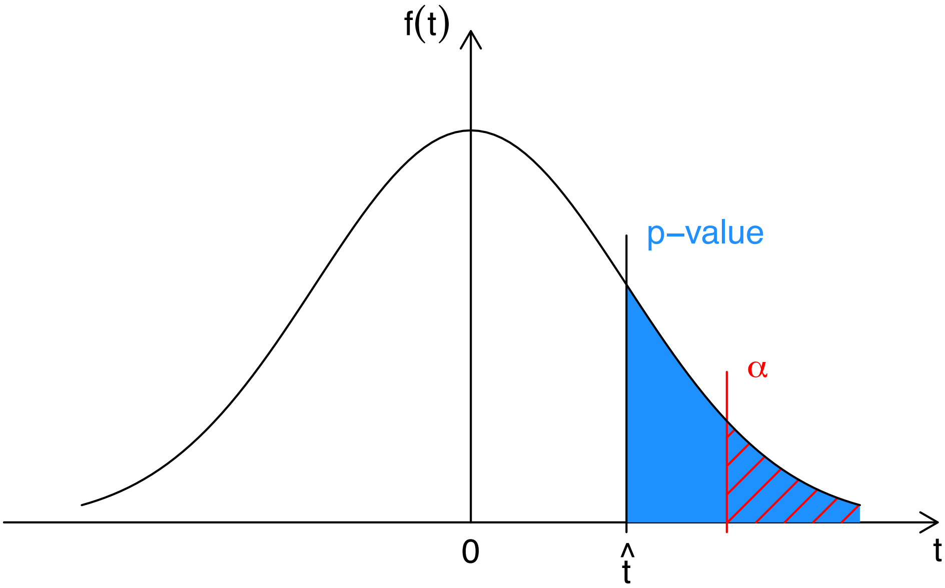
- For a given sample, the *largest* significance level can be calculated for each test statistic at which the calculated test statistic would *just not* have led to a rejection of the null hypothesis. If the significance level were increased further, the null hypothesis would be rejected. This probability is called p**-value** (**probability value**).
- It is also said that the p-value indicates the *smallest* significance level at which the null hypothesis *can just* be rejected. See Davidson & MacKinnon (2004, Section 4.2, pages 126-127) or Wooldridge (2009, Section 4.2, p. 133).
- In the case of a one-tailed t-test with a right-tailed alternative, we get

$$P(X > t(\mathbf{y})|\mathbf{y}, \theta_{j,H_0}) := p, \tag{5.85a}$$

or

$$P(X \leq t(\mathbf{y})|\mathbf{y}, \theta_{j,H_0}) = 1 - p, . \tag{5.85b}$$

since $P(X > t(\mathbf{y})|\mathbf{y}, \theta_{j,H_0}) = 1 - P(X \leq t(\mathbf{y})|\mathbf{y}, \theta_{j,H_0})$.



- **Decision rule with** p**-values**: Instead of checking whether the test statistic is in the critical region, one can compare the p-value with the significance level:
  Reject H_0, if the p**-value is smaller** than the significance level α.

Left-tailed test:	$p = P(t < t(\mathbf{y})|\mathbf{y}, \theta_{j,H_0})$,
Right-tailed test:	$p = P(t > t(\mathbf{y})|\mathbf{y}, \theta_{j,H_0})$,
Two-tailed test:	$p = P(t < -|t(\mathbf{y})||\mathbf{y}, \theta_{j,H_0}) + P(t > |t(\mathbf{y})||\mathbf{y}, \theta_{j,H_0})$

- Many computer programs (such as `R`) routinely provide the p-value for

$$H_0 : \theta_j = 0 \quad \text{versus} \quad H_1 : \theta_j \neq 0.$$

Literature: Davidson & MacKinnon (2004, Section 4.2) or for a start Wooldridge (2009, Appendix C.6).

6. The ordinary least squares estimator: Derivation and an application

Depending on the DGP and the properties of the data of the sample, different estimators are used to estimate the parameters of the **multiple linear regression model**

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t. \qquad (5.24)$$

Review of section 5.4: An **estimator** for the parameter vector $\boldsymbol{\beta}$ is a vector-valued function $\tilde{\boldsymbol{\beta}}(\mathbf{X}_1, y_1, \dots, \mathbf{X}_n, y_n)$ of the sample $\{(\mathbf{X}_t, y_t), t = 1, \dots, n\}$.

Important estimators for the multiple linear regression model

- **Ordinary least squares estimator (OLS estimator)**
  $\Rightarrow$ all chapters up to and including chapter 11
- **Generalized least squares estimator (GLS estimator)**
  $\Longrightarrow$ chapter 14
- **Instrumental variables estimator (IV estimator)**
  $\Longrightarrow$ master course **Advanced Econometrics**
- **Generalized method of moment estimator (GMM estimator)**
  $\Longrightarrow$ master course **Advanced Econometrics**
- **Maximum likelihood estimator (ML estimator)**
  $\Longrightarrow$ master course **Advanced Econometrics**

The estimators differ in their assumptions, properties and possible applications.

6.1. Vector and matrix representation of the multiple linear regression model

- **Notation**:

$$\mathbf{X}_t = \begin{pmatrix} x_{t1} & \cdots & x_{tk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

- **Matrix representation**
  - for one **observation t of the sample**:

$$y_t = \mathbf{X}_t \boldsymbol{\beta} + u_t. \tag{5.25}$$

  - for the **total sample**

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{6.1}$$

- **Vector representation**:

  The regression model for the total sample (6.1) can also be represented as **addition of vectors**:

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \dots + \mathbf{x}_k\beta_k + \mathbf{u}, \tag{6.2}$$

  where the following vectors of variables each consist of all n observations of the sample

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}, \quad i = 1, \dots, k. \tag{6.3}$$

Further matrix notation for later

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}$$

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_k^T \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^T & \mathbf{X}_2^T & \cdots & \mathbf{X}_n^T \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \tag{6.4a}$$

$$= \begin{pmatrix} \sum_{t=1}^n x_{t1}^2 & \sum_{t=1}^n x_{t1}x_{t2} & \cdots & \sum_{t=1}^n x_{t1}x_{tk} \\ \sum_{t=1}^n x_{t2}x_{t1} & \sum_{t=1}^n x_{t2}^2 & \cdots & \sum_{t=1}^n x_{t2}x_{tk} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=1}^n x_{tk}x_{t1} & \sum_{t=1}^n x_{tk}x_{t2} & \cdots & \sum_{t=1}^n x_{tk}^2 \end{pmatrix} \tag{6.4b}$$

$$= \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 & \cdots & \mathbf{x}_1^T\mathbf{x}_k \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 & \cdots & \mathbf{x}_2^T\mathbf{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_k^T\mathbf{x}_1 & \mathbf{x}_k^T\mathbf{x}_2 & \cdots & \mathbf{x}_k^T\mathbf{x}_k \end{pmatrix} \tag{6.4c}$$

$$= \sum_{t=1}^n \mathbf{X}_t^T\mathbf{X}_t \tag{6.4d}$$

6.2. The OLS estimator for multiple linear regression models

- **Ordinary least squares estimator (OLS estimator)** of $\boldsymbol{\beta}$ in the multiple linear regression model (6.1):

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^n \mathbf{X}_t^T\mathbf{X}_t\right)^{-1} \sum_{t=1}^n \mathbf{X}_t^T y_t \tag{6.5a}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{6.5b}$$

  Derivation in matrix notation in section 6.2.2.

- **Regression model of the sample**:
  - **Sample regression function**

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{6.6}$$

  - **Fitted values/OLS estimates/predicted values**: $\hat{\mathbf{y}}$
  - **Residuals**: $\mathbf{u}(\tilde{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$
  - **OLS residuals**: $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$

    In the following, the OLS residuals $\hat{\mathbf{u}}$ are often simply referred to as residuals.

- **Properties of the OLS estimator for the simple multiple regression model**
  - The **statistical** estimation properties depend on the type of data generation, respectively on the properties of the population. They can never be verified because the data

generation is unobservable. Their analysis requires the methods of **probability theory** $\Longrightarrow$ chapter 9 and following.

- The **numerical** properties always apply and are independent of the data generation. They can be investigated by algebraic or geometric methods $\Longrightarrow$ chapter 7.

6.2.1. Derivation of the OLS estimator as a moment estimator

- **Basis of the moment estimator**: Law of large numbers (LLN), cf. section 5.5.1.
- **Simplest case of (5.34)**:

$$y_t = \beta_1 + u_t, \quad E[u_t] = 0, \tag{6.7a}$$

  such that

$$\beta_1 = E[y_t] \tag{6.7b}$$

  just corresponds to the expected value (the first moment) of y_t.

- Under certain conditions (e. g. presence of a random sample, cf. (5.69)), the **law of large numbers** justifies estimating an expected value $E[y_t]$ with the arithmetic mean $\frac{1}{n}\sum_{t=1}^{n} y_t$ of a sample $y_1, \dots, y_n$,

$$\widehat{E}[y_t] = \frac{1}{n}\sum_{t=1}^{n} y_t,$$

  so that the accuracy of the estimator increases with sample size. More on this in section 5.5.1.

- β_1 can thus be estimated by estimating the expected value $E[y_t]$ with the arithmetic mean

$$\hat{\beta}_1 = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

- This principle also works for the OLS estimator of the multiple linear regression model because of (5.34), since the expected values $E\left[\mathbf{X}_t^T\mathbf{X}_t\right]$ and $E\left[\mathbf{X}_t^T y_t\right]$ in (5.34) and (5.40), respectively, can again be estimated by calculating the mean values of the matrices $\mathbf{X}_t^T\mathbf{X}_t$ and the vectors $\mathbf{X}_t^T y_t$, respectively. One obtains:

$$\begin{aligned}
\widehat{E}\left[\mathbf{X}_t^T\mathbf{X}_t\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^T\mathbf{X}_t \\
\widehat{E}\left[\mathbf{X}_t^T y_t\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^T y_t \\
\hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{n}\mathbf{X}_t^T\mathbf{X}_t\right)^{-1}\sum_{t=1}^{n}\mathbf{X}_t^T y_t && (6.5a) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. && (6.5b)
\end{aligned}$$

Here, $1/n$ was cancelled out in the penultimate row. The matrix representation (6.5) of the OLS estimator in the last row follows from the application of the matrix rules.

Assumptions used in the derivation of (6.5):

1. There is a random sample.

2. The matrix $\mathbf{X}^T\mathbf{X}$ is invertible — this requires $rk(\mathbf{X}) = k$.

Thus, β_{00}, as defined in (5.40), is estimated in any case.

If $\hat{\boldsymbol{\beta}}$ is to estimate the correct parameter vector $\boldsymbol{\beta}_0$ of the DGP, (5.34) must hold.

4. For (5.34) to hold, it is also necessary (see section 5.3) that

   a) the multiple linear regression model is correctly specified, i. e. the DGP is included in (**??**), and

   b) the expected value of the errors given the regressors are zero, i. e. $E[u_t|\mathbf{X}_t] = 0$, so that (5.33) holds.

- Literature: Davidson & MacKinnon (Cf. 2004, Section 1.5).

6.2.2. Least squares derivation of the OLS estimator

- Given is the multiple linear regression model (6.1)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

- **Idea** of the ordinary least squares estimator: Minimise the **Sum of Squared Residuals (SSR)**, i. e. the objective function

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n} u_t(\boldsymbol{\beta})^2 = \sum_{t=1}^{n} (y_t - \mathbf{X}_t\boldsymbol{\beta})^2 . \tag{6.8}$$

  This objective function is obtained by estimating the statistical risk (5.48) based on the quadratic loss function (5.46), i.e. (5.42), with the arithmetic mean.

- A possible **alternative** to the OLS objective function (6.8): Minimising the sum of absolute values

$$S_M(\boldsymbol{\beta}) = \sum_{t=1}^{n} |u_t(\boldsymbol{\beta})| = \sum_{t=1}^{n} |y_t - \mathbf{X}_t\boldsymbol{\beta}| \tag{6.9}$$

  provides estimate of the conditional **median**, i. e. the conditional 50% quantile. This objective function is obtained by estimating the statistical risk (5.48) based on the absolute value of the estimation error (5.46) with the arithmetic mean.

- Residual sum of squares in matrix notation:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{t=1}^{n} u_t(\boldsymbol{\beta})^2 \\ &= \mathbf{u}(\boldsymbol{\beta})^T\mathbf{u}(\boldsymbol{\beta}) \\ &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

  Minimise: Derivative with respect to $\boldsymbol{\beta}$, setting equal to zero, ...

- Derivation of the OLS estimator in matrix algebra, see section 1.13 for calculation rules:
  - From the vector of the first-order partial derivatives

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta},$$

    one obtains by setting the equation to zero the **first-order conditions (foc)**

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \tag{6.10}$$

    These are also called the **normal equations**.
  - If $\mathbf{X}^T\mathbf{X}$ is invertible — this requires $rk(\mathbf{X}) = k$ —, the OLS estimator (6.5)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

    is obtained again.
  - $\hat{\boldsymbol{\beta}}$ is an **unique minimum** of the objective function $S(\boldsymbol{\beta})$ if for the rank $\text{rk}(\mathbf{X})$ of the matrix $\mathbf{X}$, it holds that: $\text{rk}(\mathbf{X}) = k$.
- For the interpretation of the OLS estimator $\hat{\boldsymbol{\beta}}$ see the end of section 5.3 and the later chapter 8.

6.3. Empirical analysis of trade flows: part 1 — a brief overview

(A simplified version in some parts can be found in the course material for the bachelor course Introduction to Econometrics in chapter 1 and following.)

The following steps correspond to the section 4.3 **Components of an empirical analysis**

I. Economic analysis part

I.1 Scientific issue:

- Identify the factors that affect imports to Germany and quantify their impact.

- **A first, rough (empirical) attempt**:

  Data: Imports to Germany from 54 countries of origin in 2004 (in current US dollars)

Data description	Unit	Abbreviation	Source
Imports from Germany	current US dollars	`trade_0_d_o`	UN COMTRADE
Country of origin GDP data	current US dollars	`wdi_gdpusdcr_o`	World Bank - World Development Indicators

(See Appendix C for detailed data descriptions.)

R code to generate the scatter plot in Figure 6.1:

The following R code is part of the R program in section A.4, page 333. Remark: The indented commands are only necessary if a PDF file is to be generated.

```
################################################################################
#                   Start main program
################################################################################
save.pdf      <- 1             # 1=create PDFs of graphs, 0=otherwise

# The following libraries are loaded during the process: car,lmtest

# If these are not installed, they will be installed first:
if (!require(car)){
  install.packages("car")
}
if (!require(lmtest)){
  install.packages("lmtest")
}

# Determination of the working directory
# in which the R program and the data are located
WD            <- getwd() # Determine the directory of the R file and
setwd(WD)                # set it as working directory

# Read the data as data frame
daten_all     <-read.table("importe_ger_2004_ebrd.txt", header = TRUE)
# Assign the variable names and
# eliminate the observation export country: GER, import country: GER.
attach(daten_all[-20,])

# To try out, if importe_ger_2004_ebrd.txt has already been read in
stats(trade_0_d_o)

################################################################################
# Section 6.3
################################################################################

############# Scatterplot with (linear) regression line #####################
# I.1 Aim/scientific issue: first empirical attempt

# Define file name for output in PDF format
if (save.pdf)   pdf("plot_wdi_vs_trade.pdf", height=6, width=6)

# OLS estimation of a simple linear regression model, stored in ols
ols           <- lm(trade_0_d_o ~ wdi_gdpusdcr_o)
# Scatterplot of the two variables
```

```
plot(wdi_gdpusdcr_o, trade_0_d_o, col = "blue", pch = 16)
# Plot the linear regression line using abline
abline(ols, col = "red")
# Add a legend
legend("bottomright", "Lineare Regression", col = "red", lty = 1, bty = "n")

# Close device
if (save.pdf) dev.off()
```

Listing 6.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

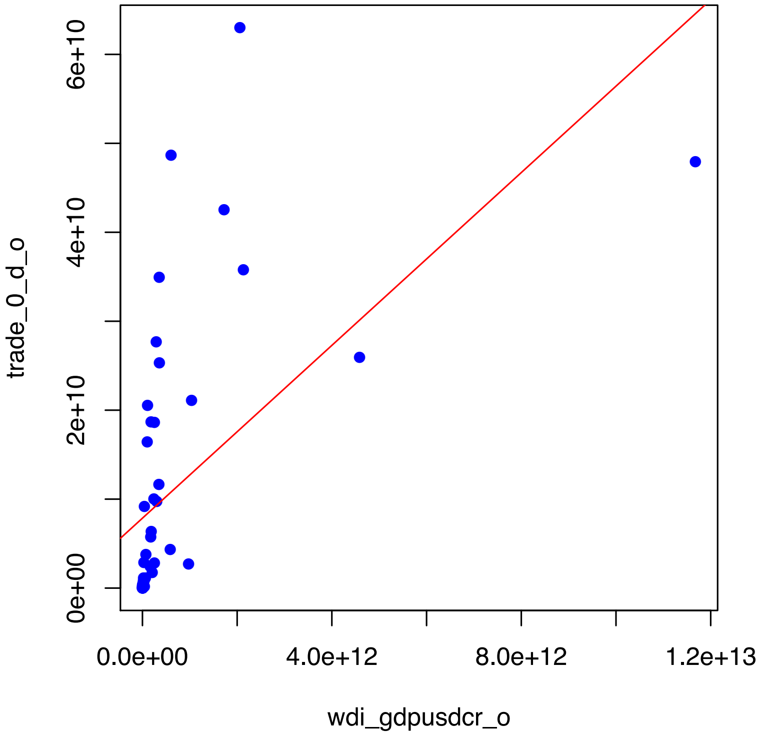


Figure 6.1.: Scatterplot on trade flow data versus GDP

- Some questions:
- What do you see?
- Is there a relationship?
- If so, how can it be quantified?
- Does a causal relationship exist - which variable determines which?
- How do imports from the US change if US GDP changes by 1%?
- Are there other relevant factors that determine imports, e. g. distance?
- Is it possible to forecast future trade flows?
- How do we place the straight line through the point cloud?
- What are the properties of the fitted straight line?
- What do we do with the other relevant factors that were neglected in the current analysis?
- What criteria does one choose to identify a possible relationship?

- Is the possible relationship really linear?
- And: how much may the results differ for another sample, e. g. for 2003?

I.2 Economic Model: Simplest form of a **gravity equation**:

- Short introduction to gravity equations: e. g. in Fratianni (2007). A theoretical foundation of the gravity equation can be found in Anderson & Wincoop (2003).

- Under idealised assumptions such as perfect specialisation of production, identical consumption preferences in the countries, no transport and trade costs, trade flows between pairs of countries are explained as a function of the respective income of the paired countries and their distance from each other:

$$M_{ijt} = A_0 Y_{it}^{\alpha_1} Y_{jt}^{\alpha_2} d_{ij}^{\alpha_3} \tag{6.11}$$

  M_{ijt} :Export from country i to country j in period t

  Y_{it} :Real income in country i in period t

  d_{ij} :Distance between country i and country j (different measures possible)

- From the economic theory of gravity equations, see Fratianni (2007), arise the **hypotheses**

  - that $\alpha_1, \alpha_2 > 0, \alpha_3 < 0$ and,
  - under certain conditions, the **hypothesis that GDP elasticities are equal to 1**

$$\alpha_1 = \alpha_2 = 1.$$

  These **hypotheses** can be **statistically tested** if suitable data are available.

- Double index ij can be converted into *one* index l.

- **Simplification**: **Considering only one time period and one direction**, namely imports of Germany in 2004. A gravity equation simplified in this way reads as follows

$$Imports_i = e^{\beta_1} Y_i^{\beta_2} d_i^{\beta_3}. \tag{6.12}$$

  By logarithmising we obtain

$$\ln(Imports_i) = \beta_1 + \beta_2 \ln(Y_i) + \beta_3 \ln(d_i). \tag{6.13}$$

  Interpretation of the parameters, cf. (8.2):

  - β_2: GDP elasticity of imports.
  - β_3: Distance elasticity of imports

  An **economic hypothesis:**

  The GDP elasticity of imports is 1: $\beta_2 = 1$.

I.3 Data availability

For our example, all available data are listed in Appendix C including detailed data descriptions.

II. Econometric model:

1. **Selection of a class of econometric models**

   - Choice of class **multiple linear regression models**: It is assumed that the logarithmic theoretical model (6.13) after extension by country-specific characteristics and a stochastic error term correctly specifies the systematic part, cf. section 5.3. Together with the unsystematic part (disturbance term) one obtains a **multiple linear regression model**

$$\begin{gathered} \ln(M_{ijt}) = \beta_1 + \beta_2 \ln Y_{it} + \beta_3 \ln Y_{jt} + \beta_4 \ln d_{ij} + \mathbf{F}_{ijt}\boldsymbol{\beta}_5 + u_{ij}, \\ \mathbf{F}_{ij} : \text{specific characteristics for exports from } i \text{ to } j. \end{gathered} \tag{6.14}$$

   - Consideration of different periods requires panel data models, see e. g. Davidson & MacKinnon (2004, Chapter 7.8).

   - The restriction to imports (6.12) to Germany and cross-sectional data results in

$$\ln(Imports_i) = \beta_1 + \beta_2 \ln(GDP_i) + \beta_3 \ln(Distance_i) + \mathbf{F}_i\boldsymbol{\beta}_5 + u_i. \tag{6.15}$$

   - Note: Since the variables in $\mathbf{F}_i$ are not yet chosen, many regression models are conceivable, each with different variables satisfying (6.15). Since at least all models are multiple linear regression models, this is referred to as the choice of a model class.

2. **Procuring data: Collecting a sample**

   - Which goods should be included in imports?

   - How to measure the distance between countries?

   - What variables should be included in $\mathbf{F}_i$? Possible (and available) variables: openness, population, area, colonial past.

   - How to measure openness?, etc.

   A sample with a large number of alternative variables is available for the following estimations. See Appendix C for detailed data descriptions.

   Important: Variable selection and measurement of the variables can substantially influence empirical results.

3. **Specifying, estimating and selecting an econometric model**

   - First, we neglect all variables in $\mathbf{F}_i$ and consider a linear regression model with GDP and distance as explanatory variables: **Model 2**:

$$\ln(Imports_i) = \beta_1 + \beta_2 \ln(GDP_i) + \beta_3 \ln(Distance_i) + u_i. \tag{6.16}$$

(Model 1 only includes GDP as a regressor and is considered in section 10.3.)

- Estimating model 2 with the ordinary least squares estimator (OLS estimator):

  R code (Extract from the R program in section A.4)

```
# The numbering of the regression models is based on
# the models in the script, section 10.3

# Run a linear regression and save the results as an object
mod_2_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))

# Display of the regression results
summary(mod_2_kq)
```

Listing 6.2: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

R output

```
Call:
lm(formula = log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))

Residuals:
Min       1Q   Median       3Q      Max
-1.99289 -0.58886 -0.00336  0.72470  1.61595

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.67611    2.17838   2.147   0.0371 *
log(wdi_gdpusdcr_o)  0.97598    0.06366  15.331  < 2e-16 ***
log(cepii_dist)     -1.07408    0.15691  -6.845 1.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9284 on 46 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.8838,     Adjusted R-squared: 0.8787
F-statistic: 174.9 on 2 and 46 DF,  p-value: < 2.2e-16
```

4. **Validating the estimated model**

   Is the model correctly specified?

   a) Are variables missing?

   b) Is the relationship really linear in the logarithms?

   c) Are the assumptions for using the OLS estimator fulfilled? Is the OLS estimator actually suitable for estimating (6.16)?

   Re a): First check: Do the parameter estimates change when additional variables are included in the regression model, e. g. openness?

Model 3a:

$$\begin{aligned} \ln(Imports) &= \beta_1 + \beta_2 \ln(GDP) + +\beta_3 \ln(Distance) && (6.17) \\ &+ \beta_4\, Openness + \beta_6 \ln Area + u. && (6.18) \end{aligned}$$

R code (Extract from the R program in section A.4)

```
# using the formula command
mod_3a_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                  ebrd_tfes_o
mod_3a_kq      <- lm(mod_3a_formula)
# Display the regression results of the second linear regression model
summary(mod_3a_kq)
```

Listing 6.3: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

R output

```
Call:
lm(formula = log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
ebrd_tfes_o)

Residuals:
Min      1Q  Median      3Q     Max
-2.1999 -0.5587  0.1009  0.5866  1.5220

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.74104    2.17518   1.260   0.2141
log(wdi_gdpusdcr_o)  0.94066    0.06134  15.335  < 2e-16 ***
log(cepii_dist)     -0.97032    0.15268  -6.355 9.26e-08 ***
ebrd_tfes_o          0.50725    0.19161   2.647   0.0111 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8731 on 45 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.8995,     Adjusted R-squared: 0.8928
F-statistic: 134.2 on 3 and 45 DF,  p-value: < 2.2e-16
```

Is the difference in estimates between the two model specifications relevant? A t-test can be used to check, see chapter 11.

Instead of the variable openness or simply as an additional variable, the variable area could be used. The selection of a model can be done with model selection procedures, see section 10.1.

Re b) and c): Test procedures for model diagnostics are discussed in chapter 15.

5. **Using the validated model**

If the model validation reveals no more problems, then we can use the model:

- Interpretation of the parameters of the model. See sections 8.1 and 8.4 for interpretation of parameters in differently specified models.
- Conducting hypothesis tests:

  - Is there a causal relationship between imports and economic output of the exporting country? The condition for this is that $\beta_2 \neq 0$.
  - Testing the hypothesis already formulated: Is the GDP elasticity of imports equal to one?
  - Corresponding tests are carried out in the chapter on asymptotics and testing in part 3 in section 11.7.
- Predictions

Systematic continuation of **Empirical analysis of trade flows: part 1** in the model specification chapter with part 2 in section 10.3.

7. The ordinary least squares estimator and its geometric interpretation

Multiple linear regression model in matrix representation for the total sample

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{6.1}$$

OLS estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{6.5}$$

For a better understanding of the OLS estimator it is **very helpful** to look at the geometry of the OLS estimator. This is done in two steps:

1. Interpretation of the normal equations (6.10) as orthogonality conditions $\Longrightarrow$ section 7.1.1.
2. Interpretation of the so-called **projection matrices $\mathbf{P}_\mathbf{X}$** und **$\mathbf{M}_\mathbf{X}$** $\Longrightarrow$ section 7.1.2

   The **projection matrices $\mathbf{P}_\mathbf{X}$** and **$\mathbf{M}_\mathbf{X}$** occur when predicting the dependent variable y and when calculating the OLS residuals:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}_{\hat{\boldsymbol{\beta}}} := \mathbf{P}_\mathbf{X}\mathbf{y}, \tag{7.1}$$

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} := \mathbf{M}_\mathbf{X}\mathbf{y}.\end{aligned} \tag{7.2}$$

   Definition of the projection matrices:

$$\mathbf{P}_\mathbf{X} := \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T, \tag{7.3}$$

$$\mathbf{M}_\mathbf{X} := \mathbf{I} - \mathbf{P}_\mathbf{X}. \tag{7.4}$$

Application examples of the projection matrices $\mathbf{P}_\mathbf{X}$ and $\mathbf{M}_\mathbf{X}$ in this chapter:

- Decomposition (7.15) of the Total Sum of Squares: $\mathbf{y}^T\mathbf{y} = \hat{\mathbf{y}}^T\hat{\mathbf{y}} + \hat{\mathbf{u}}^T\hat{\mathbf{u}}$
- Scaling of $\mathbf{X}_t$ irrelevant for fitted values.
- Frisch-Waugh-Lovell theorem and partialling-out
- Coefficients of determination

- Analysis of the impact of possible outliers on the OLS estimator (6.5)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Applications of the projection matrices in the following chapters:

- Calculation of the variance of an estimator of a single parameter β_j (9.15) in section 9.3
- χ^2-distribution in section 2.9.2.
- Derivation of the distribution of the t-statistic (11.14) in section 11.3.1
- Derivation of the distribution of the F-statistic (11.28) in section 11.3.2
- Derivation fixed-effects estimator for panel data

7.1. The geometry of the OLS estimator

Reminder: **Vector representation** of the multiple linear regression model:

- The regression model (6.1) corresponds to a **addition of vectors**

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_k\beta_k + \mathbf{u} \tag{6.2}$$

- Accordingly, the following applies to the regression model of the sample

$$\mathbf{y} = \mathbf{x}_1\hat{\beta}_1 + \mathbf{x}_2\hat{\beta}_2 + \cdots + \mathbf{x}_k\hat{\beta}_k + \hat{\mathbf{u}}, \tag{7.5}$$

  where for the OLS residuals $\hat{\mathbf{u}}$ will be shown:

$$\mathbf{x}_i^T\hat{\mathbf{u}} = 0, \quad i = 1, \ldots, k, \tag{7.6b}$$

7.1.1. Orthogonality conditions: Proof, Interpretation

Orthogonality conditions:

$$\mathbf{X}^T\hat{\mathbf{u}} = \mathbf{0}, \tag{7.6a}$$

$$\mathbf{x}_i^T\hat{\mathbf{u}} =< \mathbf{x}_i, \hat{\mathbf{u}} > = 0, \quad i = 1, \ldots, k. \tag{7.6b}$$

Proof: From the normal equations (6.10)

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

follows

$$\mathbf{X}^T\underbrace{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}_{\hat{\mathbf{u}}} = \mathbf{0}$$

and hence (7.6a).

For the i-th row, accordingly

$$\mathbf{x}_i^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = 0, \quad i = 1, \dots, k,$$

applies and thus (7.6b). □

Note: In contrast, the following applies to the disturbance terms u in general, (cf. (1.2))

$$\mathbf{x}_i^T\mathbf{u} = ||\mathbf{x}_i||\ ||\mathbf{u}||\cos(\theta), \quad i = 1, 2, \dots, k,$$

where $||\cdot||$ measures the length (Euclidean norm) of a vector and θ measures the angle between the two vectors $\mathbf{x}_i$ and $\mathbf{u}$. The latter is generally not 90 degrees, so the product is generally not zero.

Which magnitudes are in which spaces?

- Every **linear combination of regressors $\mathbf{Xd}$** with a $(k \times 1)$ vector $\mathbf{d}$ lies in the **subspace of the regressors** $\delta(\mathbf{X})$, thus also the **vector of fitted values $\mathbf{X}\boldsymbol{\beta}$ with known $\boldsymbol{\beta}$** and the **vector of the estimated fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}$**.
  - Reminder:
    Every linear combination of the columns of a matrix $\mathbf{X}$ lies in the subspace $\delta(\mathbf{X})$ spanned by the columns of the matrix $\mathbf{X}$, cf. (1.6).
  - This also applies to the regressor matrix $\mathbf{X}$, so that

    $$\mathbf{X}\boldsymbol{\beta} = \sum_{i=1}^{k}\mathbf{x}_i\beta_i \ \in \ \delta(\mathbf{X}) \subset E^n$$

    for any $\boldsymbol{\beta}$ is contained in the **subspace of regressors** $\delta(\mathbf{X})$.
  - This also applies to the fitted values $\hat{\mathbf{y}}$

    $$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \delta(\mathbf{X}).$$

- Due to (7.6a), the **vector of the OLS residuals $\hat{\mathbf{u}}$** lies in the **orthogonal complement of the subspace of the regressors** $\delta^{\perp}(\mathbf{X})$, cf. (1.7),

  $$\hat{\mathbf{u}} \in \delta^{\perp}(\mathbf{X}) \subset E^n.$$

  The equations (7.6a) and (7.6b), respectively, are therefore called **orthogonality conditions**.
  - The **OLS residual vector $\hat{\mathbf{u}}$ is perpendicular** to the explained/fitted/predicted values $\mathbf{X}\hat{\boldsymbol{\beta}} \in \delta(\mathbf{X})$. ($\mathbf{x}_i^T\hat{\mathbf{u}} = 0$ implies that $\cos(\theta) = 0$ in (1.2).)

- $\hat{\mathbf{u}}$ corresponds to the perpendicular of $\mathbf{y}$ on $\mathbf{X}\boldsymbol{\beta}$, which is given by minimising the Euclidean norm of $\mathbf{u}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$:

  $$\min_{\boldsymbol{\beta}} ||\mathbf{u}(\boldsymbol{\beta})||.$$

  The OLS estimator thus minimises the Euclidean norm of the residual vector!

- **Note**: Minimising a different norm (which would imply a different loss function) would result in a different estimator and the residual vector would no longer be perpendicular to $\mathbf{X}$!

- **Definition**

  - **Unit basis vector**: $\mathbf{e}_t = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots 0 \end{pmatrix}^T$, where there is a 1 in the t-th row. All n unit basis vectors $e_t, t = 1, \ldots, n$ form a **basis** for E^n, where every basis vector has norm $||e_t|| = 1$.

- Additional property if **constant in the model**:
  The regression line, or in the case of $k > 2$ regressors the **regression hyperplane**, **passes** through the **barycentre**, i. e. through $\bar{y}$ and the mean values of the regressors $\bar{x}_i$, $i = 1, \ldots, k$

  $$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_k\bar{x}_k. \tag{7.7}$$

  If the regression contains a constant, $\mathbf{x}_1$ corresponds to a vector $\boldsymbol{\iota}$ with ones

  $$\boldsymbol{\iota} := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{7.8}$$

  Proof: Replacing $\mathbf{x}_1$ in (7.6b) with $\boldsymbol{\iota}$ yields

  $$\boldsymbol{\iota}^T\hat{\mathbf{u}} = 0 \quad \text{bzw.} \quad \boldsymbol{\iota}^T\hat{\mathbf{u}} = \sum_{t=1}^{n} \hat{u}_t = 0, \tag{7.9}$$

  i. e. the deviations of the regression line cancel out on average.

  $$\boldsymbol{\iota}^T\mathbf{y} = \boldsymbol{\iota}^T\mathbf{x}_1\hat{\beta}_1 + \boldsymbol{\iota}^T\mathbf{x}_2\hat{\beta}_2 + \cdots + \boldsymbol{\iota}^T\mathbf{x}_k\hat{\beta}_k + \underbrace{\boldsymbol{\iota}^T\hat{\mathbf{u}}}_{=0 \text{ see above}}$$
  $$n\bar{y} = n\bar{x}_1\beta_1 + n\bar{x}_2\beta_2 + \cdots + n\bar{x}_k\beta_k + 0$$

  yields (7.7) after multiplication with $1/n$. □

7.1.2. Orthogonal projections and their properties

Projection in everyday language:

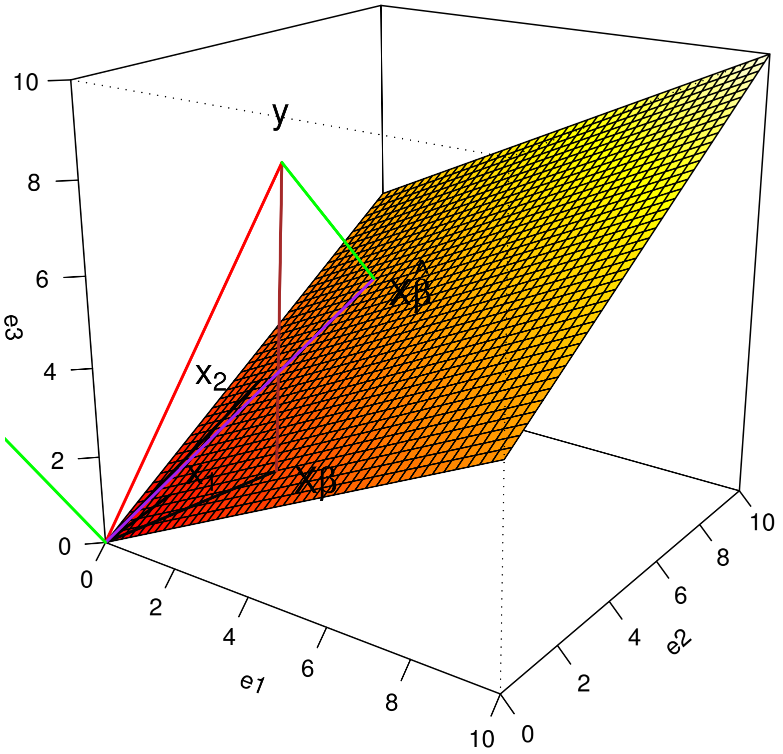


Figure 7.1.: Geometry in E^3 of the OLS estimator, $n = 3$ (R program (allows rotation and tilting of the graph) and calculation notes see section A.5, page 342

- By the action of light, a two-dimensional image of a three-dimensional object is produced on a wall: the three-dimensional object is projected onto a surface, i.e. a two-dimensional object.
- When projecting from three-dimensional space into two-dimensional 'space', information is lost.
- Depending on the position of the light source, the projection on the wall changes.

Definitions

- A **projection** is a mapping from an n-dimensional space into a k-dimensional subspace, $k < n$. Within the subspace, the projection is invariant, since the points do not change through the mapping within the subspace. (Cf. property of idempotence for projection matrices)

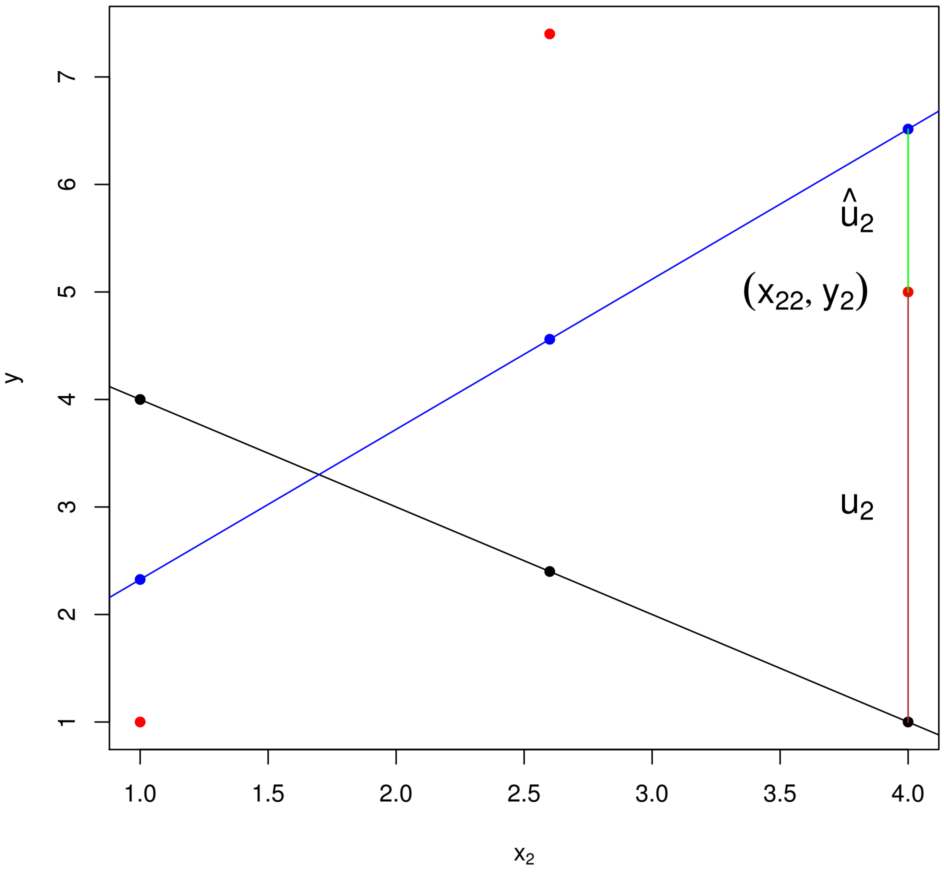


Figure 7.2.: Scatterplot ((x_{t2}, y_t) in red, $(x_{t2}, \mathbf{X}_t\boldsymbol{\beta})$ in black, $(x_{t2}, \hat{y}_t)$ in blue) (R program see section A.5, page 342)

- An **orthogonal projection** is a mapping in which the distances between the points in E^n and the projections in the subspace are minimised. So: the vectors connecting the points in E^n and the orthogonal subspace are perpendicular to the subspace.

Projection in econometrics: the n sample observations $\mathbf{y}$ define a point in a n-dimensional Euclidean space. A Euclidean subspace is defined by the $k \leq n$ regressor variables. The fitted values $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ lie in the said subspace since the OLS estimator is a projection of $\mathbf{y}$ into this subspace, as will be shown below. See section 7.1 for this.

Review: The projection matrices for the OLS estimators are

$$\mathbf{P}_\mathbf{X} := \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T, \tag{7.3}$$

$$\mathbf{M}_\mathbf{X} := \mathbf{I} - \mathbf{P}_\mathbf{X} \tag{7.4}$$

The projections into a k-dimensional subspace require that all regressors are linearly independent, i. e. the dimension of $\delta(\mathbf{X})$ is equal to k. This justifies assumption **(B3)** in section

9.1.1).

OLS projections:
The OLS estimator involves two projections:

- The **OLS estimator of the fitted values $\hat{\mathbf{y}}$**

$$\hat{\mathbf{y}} = \mathbf{P_X y} \tag{7.1}$$

corresponds to a projection from $\mathbf{y} \in E^n$ to $\hat{\mathbf{y}} \in \delta(\mathbf{X})$,

  i. e. from the n-dimensional space into the k-dimensional subspace $\delta(\mathbf{X})$ spanned by the regressors $\mathbf{X}$.

- The **OLS estimator of the residuals $\hat{\mathbf{u}}$**

$$\hat{\mathbf{u}} = \mathbf{M_X y} \tag{7.2}$$

corresponds to a projection from $\mathbf{y} \in E^n$ to $\hat{\mathbf{u}} \in \delta^{\perp}(\mathbf{X})$,

  i. e. from the n-dimensional space into the orthogonal complement of the subspace spanned by the regressors. The dimension of the subspace $\delta^{\perp}(\mathbf{X})$ is equal to $n - k$.

Properties of the OLS projections and the corresponding projection matrices $\mathbf{P_X}$ and $\mathbf{M_X}$, (cf. section 1.6):

- The projection matrices $\mathbf{P_X}$ and $\mathbf{M_X}$ are **idempotent**:

$$\mathbf{P_X P_X} = \mathbf{P_X}, \quad \mathbf{M_X M_X} = \mathbf{M_X}$$

  and thus

$$\mathbf{P_X} \cdot \ldots \cdot \mathbf{P_X} \cdot \mathbf{P_X} = \mathbf{P_X} \quad \text{and} \quad \mathbf{M_X} \cdot \ldots \cdot \mathbf{M_X} \cdot \mathbf{M_X} = \mathbf{M_X}.$$

- The projection matrices $\mathbf{P_X}$ and $\mathbf{M_X}$ are **symmetric**, i. e. $\mathbf{P}_\mathbf{X}^T = \mathbf{P_X}$ and $\mathbf{M}_\mathbf{X}^T = \mathbf{M_X}$.

- $\mathbf{P_X M_X} = \mathbf{0}$.

  Geometric interpretation: the first projection (i. e. single premultiplication with $\mathbf{P_X}$ and $\mathbf{M_X}$, respectively) yields a vector in the invariant subspace which a further projection cannot change.

- $\mathbf{P_X}$ and $\mathbf{M_X}$ imply **complementary projections**.

  This is because, due to $\mathbf{M_X} = \mathbf{I} - \mathbf{P_X}$, their sum equals the output vector:

$$\mathbf{P_X y} + \mathbf{M_X y} = \mathbf{y}. \tag{7.10}$$

- The **OLS method corresponds to orthogonal projections**.

  Proof: For two complementary projections it holds that

$$\mathbf{P_X M_X} = \mathbf{P_X}\left(\mathbf{I} - \mathbf{P_X}\right) = \mathbf{P_X} - \mathbf{P_X} = \mathbf{O}. \tag{7.11}$$

For arbitrary vectors in the two subspaces $\mathbf{z} \in \delta(\mathbf{X})$ and $\mathbf{w} \in \delta^{\perp}(\mathbf{X})$ it holds that $\mathbf{z} = \mathbf{P_X z}$ and $\mathbf{w} = \mathbf{M_X w}$. Since $\mathbf{P_X}$ is symmetric, $\mathbf{z}$ and $\mathbf{w}$ are orthogonal to each other, since

$$\mathbf{z}^T\mathbf{w} = \mathbf{z}^T\mathbf{P}_\mathbf{X}^T\mathbf{M_X w} = 0 \quad \text{bzw.} \quad <\mathbf{z},\mathbf{w}> = <\mathbf{P_X z},\mathbf{M_X w}> = 0.$$

□

In general, if two projections are **complementary** and the corresponding projection matrices are **symmetric**, they define an **orthogonal decomposition**.

Geometric interpretation: $\mathbf{P_X}$ and $\mathbf{M_X}$ define an **orthogonal decomposition** of E^n, so the two vectors $\mathbf{P_X y}$ and $\mathbf{M_X y}$ lie in two orthogonal subspaces.

If one wants to project a vector in $\delta(\mathbf{X})$ onto $\delta^{\perp}(\mathbf{X})$, the perpendicular must be formed into the subspace $\delta^{\perp}(\mathbf{X})$. This leads exactly to the origin. The two projections therefore cancel each other out. $\mathbf{M_X}$ eliminates all vectors in $\delta(\mathbf{X})$ to the origin and correspondingly $\mathbf{P_X}$ eliminates all vectors in $\delta^{\perp}(\mathbf{X})$.

Consequences (of orthogonality) of OLS projections

- **Notation**:

  Total Sum of Squares

$$TSS := ||\mathbf{y}||^2 \neq \sum_{t=1}^{n}(y_t - \bar{y})^2 := SST, \tag{7.12}$$

  Explained Sum of Squares

$$ESS := ||\hat{\mathbf{y}}||^2 = ||\mathbf{P_X y}||^2 \neq \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 := SSE, \tag{7.13}$$

  Sum of Squared Residuals

$$SSR := ||\hat{\mathbf{u}}||^2 = ||\mathbf{M_X y}||^2. \tag{7.14}$$

  SST, SSE were defined in Wooldridge (2009, Section 2.3) or in the course material for the bachelor course Introduction to Econometrics.

  $\mathbf{P_{X,W}}$ projects into the invariant subspace $\delta(\mathbf{X},\mathbf{W})$.

- **Decomposition of the Total Sum of Squares** (TSS)

$$\begin{aligned} ||\mathbf{y}||^2 &= ||\mathbf{X}\hat{\boldsymbol{\beta}}||^2 + ||\hat{\mathbf{u}}||^2 \\ TSS &= ESS + SSR \end{aligned} \tag{7.15}$$

  The decomposition of the TSS (7.15) corresponds to **Pythagoras' theorem**.

Proof:

$$\begin{aligned}||\mathbf{y}||^2 &= ||\mathbf{P_X y} + \mathbf{M_X y}||^2 = < \mathbf{y}, \mathbf{y} > \\ &= < \mathbf{P_X y} + \mathbf{M_X y}, \mathbf{P_X y} + \mathbf{M_X y} > \\ &= \mathbf{y}^T\mathbf{P}_\mathbf{X}^T\mathbf{P_X y} + \mathbf{y}^T\mathbf{P}_\mathbf{X}^T\mathbf{M_X y} + \mathbf{y}^T\mathbf{M}_\mathbf{X}^T\mathbf{P_X y} \\ &+ \mathbf{y}^T\mathbf{M}_\mathbf{X}^T\mathbf{M_X y}.\end{aligned} \tag{7.16}$$

One obtains

$$\begin{aligned}||\mathbf{y}||^2 &= \mathbf{y}^T\mathbf{P_X y} + \mathbf{y}^T\mathbf{M_X y} \\ &= ||\mathbf{P_X y}||^2 + ||\mathbf{M_X y}||^2,\end{aligned}$$

and thus (7.15). □

However:

$$\begin{aligned}||\mathbf{P_X y}||^2 &\leq ||\mathbf{y}||^2 \quad \text{and} \\ ||\mathbf{y}||^2 &\leq ||\mathbf{X}\boldsymbol{\beta}||^2 + ||\mathbf{u}||^2.\end{aligned}$$

- **Fitted values and residuals** are **independent of scaling of the regressors and independent linear combinations** of the regressors with a non-singular $(k \times k)$ matrix $\mathbf{A}$, because $\delta(\mathbf{X}) = \delta(\mathbf{XA})$, since

$$\begin{aligned}\mathbf{P_{XA}} &= \mathbf{XA}\left((\mathbf{XA})^T\mathbf{XA}\right)^{-1}(\mathbf{XA})^T \\ &= \mathbf{XA}\left(\mathbf{A}^T\mathbf{X}^T\mathbf{XA}\right)^{-1}\mathbf{A}^T\mathbf{X}^T \\ &= \mathbf{XAA}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{A}^T)^{-1}\mathbf{A}^T\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{P_X}\end{aligned}$$

  and correspondingly for $\mathbf{M_{XA}}$, i. e.

$$\begin{aligned}\mathbf{y} &= \mathbf{P_X y} + \mathbf{M_X y} \\ \mathbf{y} &= \mathbf{P_{XA} y} + \mathbf{M_{XA} y}.\end{aligned}$$

- **Frisch-Waugh-Lovell theorem**, see next section.

To read: Davidson & MacKinnon (2004, Section 2.3)

7.1.3. Partitioned regression and Frisch-Waugh-Lovell theorem

- The starting point is again the multiple linear regression model (6.1)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

- If one is particularly **interested in** β_k, (6.1) can be written as follows:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_k\beta_k + \mathbf{u} = \begin{pmatrix}\mathbf{X}_1 & \mathbf{x}_k\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \beta_k\end{pmatrix} + \mathbf{u}, \tag{7.17}$$

  where

  - $\mathbf{X}_1$ is a $(n \times (k-1))$ matrix and $\mathbf{x}_k$ is a $(n \times 1)$ vector,
  - $\boldsymbol{\beta}_1$ is a $((k-1)\times 1)$ vector and β_k is a scalar.

  In section 9.6 it is shown that the **OLS estimator of** β_k using

$$\mathbf{y} = \mathbf{x}_k\beta_k + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}$$

  is biased unless the empirical correlation between $\mathbf{x}_k$ and all other regressors $\mathbf{x}_1, \dots, \mathbf{x}_{k-1}$ is zero or $\boldsymbol{\beta} = \mathbf{0}$, cf. (9.32). The empirical correlation is zero if in the regression

$$\mathbf{x}_k = \mathbf{X}_1\boldsymbol{\delta} + \boldsymbol{\eta}$$

  it holds that:

$$\begin{aligned}\hat{\boldsymbol{\delta}} = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{x}_k = \mathbf{0} \quad &\Leftrightarrow \quad \mathbf{X}_1^T\mathbf{x}_k = \mathbf{0}\\ &\Leftrightarrow \quad \mathbf{x}_1^T\mathbf{x}_k = \mathbf{x}_2^T\mathbf{x}_k = \cdots = \mathbf{x}_{k-1}^T\mathbf{x}_k = 0. \end{aligned} \tag{7.18}$$

  Geometric interpretation of (7.18): $\mathbf{x}_k$ is orthogonal to $\mathbf{x}_1, \dots, \mathbf{x}_{k-1}$.

- What to do if (7.18) does not hold? **Orthogonalise**!
  Considering the general case: The regression model is then

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{7.19}$$

  with partitioning of the regressor matrix

$$\mathbf{X} = \begin{pmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{pmatrix}$$

  into the $(n \times k_1)$ matrix $\mathbf{X}_1$ and the $(n \times k_2)$ matrix $\mathbf{X}_2$ $(k = k_1 + k_2)$.

- How to orthogonalise? **Use of orthogonal projections**.
  Orthogonalise by

$$\mathbf{Z} = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2.$$

  Test with $\mathbf{M}_{\mathbf{X}_1} := \mathbf{M}_1$:

$$\mathbf{X}_1^T\mathbf{Z} = \mathbf{X}_1^T(\mathbf{M}_1\mathbf{X}_2) = \mathbf{X}_1^T(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_2 = \mathbf{X}_1^T\mathbf{X}_2 - \mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}.$$

- Thus, to estimate $\boldsymbol{\beta}_2$, one can perform the following regressions:

  - an OLS regression for $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$ or

– an OLS regression for $\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{v}$.

Possible **problem**: The residual vectors are not equal (verify!).
Solution: Multiply all variables by $\mathbf{M}_1$. One obtains

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\mathbf{u}, \tag{7.20}$$
$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{7.21}$$

Frisch-Waugh-Lovell theorem (FWL theorem)

1. The OLS estimators for $\boldsymbol{\beta}_2$ for the regression models

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{7.19}$$

and

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{7.21}$$

are numerically identical.

2. The OLS residuals of the regressions for (7.19) and (7.21) are numerically identical.

Regeln for calculating with projection matrices for partitioned regressions (7.19):

$$\mathbf{P}_X\mathbf{P}_1 = \mathbf{P}_1\mathbf{P}_X = \mathbf{P}_1 \tag{7.22}$$
$$\mathbf{M}_X\mathbf{M}_1 = \mathbf{M}_1\mathbf{M}_X = \mathbf{M}_X \tag{7.23}$$

The multiplication of two different projection matrices, where the subspace of one projection matrix is contained in the subspace of the other projection matrix, corresponds to the projection matrix projecting into the smaller subspace.

Proof of the FWL theorem: Cf. Davidson & MacKinnon (2004, Section 2.5, p. 68-69). Statement 1.: The OLS estimator for (7.21) is

$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2^T\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{M}_1\mathbf{y}. \tag{7.24}$$

Substituting the OLS estimators for the total regression (7.19) yields

$$\mathbf{y} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\mathbf{u}} \tag{7.25}$$

Multiplication from the left by $\mathbf{X}_2^T\mathbf{M}_1$ yields

$$\mathbf{X}_2^T\mathbf{M}_1\mathbf{y} = \mathbf{X}_2^T\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2, \tag{7.26}$$

since $\mathbf{X}_2^T\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ and $\mathbf{X}_2^T\mathbf{M}_1\hat{\mathbf{u}} = \mathbf{X}_2^T\mathbf{M}_1\mathbf{M}_X\mathbf{y} = \underbrace{\mathbf{X}_2^T\mathbf{M}_X}_{=\mathbf{0}}\mathbf{y} = \mathbf{0}$. Solving (7.26) yields (7.24).

Statement 2.: Multiplication of (7.25) by $\mathbf{M}_1$ yields

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \underbrace{\mathbf{M}_1\hat{\mathbf{u}}}_{=\mathbf{M}_1\mathbf{M}_X\mathbf{y}} = \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \underbrace{\mathbf{M}_X\mathbf{y}}_{\hat{\mathbf{u}}} = \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\mathbf{u}}.$$

Thus, the OLS residuals $\hat{\boldsymbol{\varepsilon}}$ for (7.21) just corresponds to the OLS residuals $\hat{\mathbf{u}}$ for the full regression (7.19). □

Interpretation of the Frisch-Waugh-Lovell theorem: The OLS estimator for $\boldsymbol{\beta}_2$ can also be performed sequentially by running different OLS regressions with fewer variables. The regression

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{7.21}$$

corresponds to a regression of residuals on residuals of the following OLS estimations:

- $\mathbf{M}_1\mathbf{y}$ just corresponds to the residuals of a regression of $\mathbf{y}$ on $\mathbf{X}_1$.
- $\mathbf{X}_2$ just contains the regressors $\mathbf{x}_{k_1+1}, \dots, \mathbf{x}_{k_1+k_2}$. Thus, for each $j = k_1+1, \dots, k_1+k_2$ the vector $\mathbf{M}_1\mathbf{x}_j$ just corresponds to the residuals of a regression from $\mathbf{x}_j$ on $\mathbf{X}_1$.

By pre-multiplying $\mathbf{M}_1$ in (7.21), residuals are generated from the respective variable that are perpendicular to the subspace spanned by the regressors in $\mathbf{X}_1$, so that for the OLS estimation of (**??**) the influences of the regressors in $\mathbf{X}_1$ do not matter, since they are each orthogonal to the variables in (7.21).

For reading: Davidson & MacKinnon (2004, Section 2.4)

7.2. Applications of the Frisch-Waugh-Lovell theorem

1. **Adjustment of regressors of no interest**

Examples:

- **Constant**: W.l.o.g, let $\mathbf{x}_1 = \boldsymbol{\iota} = (1, 1, ..., 1)^T$ and thus $\mathbf{M}_\iota := \mathbf{I}_n - \frac{1}{n}\boldsymbol{\iota\iota}^T$. $\mathbf{M}_\iota$ is called the centring matrix, since

$$\mathbf{M}_\iota = \mathbf{I}_n - \frac{1}{n}\boldsymbol{\iota\iota}^T = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & & & 1 \end{pmatrix} = \begin{bmatrix} 1-\frac{1}{n} & & & -\frac{1}{n} \\ & 1-\frac{1}{n} & & \\ & & \ddots & \\ -\frac{1}{n} & & & 1-\frac{1}{n} \end{bmatrix}.$$

Pre-multiplication of a vector with $\mathbf{M}_\iota$ calculates the **deviations from the mean of the vector**. $\mathbf{M}_\iota\mathbf{X}$ yields **centred regressors**. The vector of slope parameters $\boldsymbol{\beta}_2$ can be estimated using the Frisch-Waugh-Lovell theorem:

$$\mathbf{M}_\iota\mathbf{y} = \mathbf{M}_\iota\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_\iota\mathbf{u},$$
$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2^T\mathbf{M}_\iota\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{M}_\iota\mathbf{y}.$$

Interpretation: The point cloud in a scatter plot is shifted by centring $\mathbf{x}$ or $\mathbf{y}$, the slope of the regression line does not change.

- $\sharp$ **Seasonal dummy variables**: In time series, regularly recurring fluctuations can often be modelled by seasonal dummies. If one combines seasonal dummies and constant, if available, in the matrix $\mathbf{S}$ and one is only interested in the parameter vector $\boldsymbol{\beta}$, one can estimate

$$\begin{aligned} \mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{or} \\ \mathbf{M}_\mathbf{S}\mathbf{y} = \mathbf{M}_\mathbf{S}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}_\mathbf{S}\mathbf{u} \end{aligned}$$

where $\mathbf{M}_\mathbf{S} = \mathbf{I} - \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$. For quarterly data starting with the first quarter of a year and ending with the last quarter of a year, $\mathbf{S}$ can be chosen as follows:

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{or} \quad \mathbf{S} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{or} \ \ldots$$

- $\sharp$ **Time trend**

2. $\sharp$ **Representation of the uncentred coefficient of determination**

- Note definitions of SSE, SST, TSS, ESS in (7.12) and (7.13), SSR in (7.14) as well as

$$||\mathbf{y}||^2 = ||\hat{\mathbf{y}}||^2 + ||\hat{\mathbf{u}}||^2 \tag{7.15}$$

- **Uncentred** R^2:

$$R_u^2 := \frac{ESS}{TSS} = \frac{||\hat{\mathbf{y}}||^2}{||\mathbf{y}||^2} = \frac{||\mathbf{P}_\mathbf{X}\mathbf{y}||^2}{||\mathbf{y}||^2} = \cos^2\theta \quad \Rightarrow \quad 0 \le R_u^2 \le 1. \tag{7.27}$$

Proof sketch: The last equal sign in (7.27) follows from the definition of cosine: $\cos\theta = \text{Adjacent}/\text{Hypotenuse} = ||\mathbf{P}_\mathbf{X}\mathbf{y}||/||\mathbf{y}||$. □

From (7.15) it also follows that

$$R_u^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{||\hat{\mathbf{u}}||^2}{||\mathbf{y}||^2} = 1 - \frac{||\mathbf{M}_\mathbf{X}\mathbf{y}||^2}{||\mathbf{y}||^2}. \tag{7.28}$$

Disadvantage of R_u^2: If there is a constant in the regression model, $\mathbf{x}_1 = \boldsymbol{\iota}$, and if the data are not centred, R_u^2 depends on the value of the constant (Davidson & MacKinnon 2004, Section 2.5), since when β_1 is increased, the denominator changes while the numerator remains constant.

3. **Representation of the (centred) coefficient of determination)**

(Centred) coefficient of determination R^2: When we talk about the coefficient of determination, we generally mean the centred R^2. Definitions commonly used in the literature:

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{t=1}^{n} \left(\hat{y}_t - \bar{y}\right)^2}{\sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2}. \tag{7.29}$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{||\hat{\mathbf{u}}||^2}{||\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}||^2} = 1 - \frac{||\mathbf{M}_{\mathbf{X}}\mathbf{y}||^2}{||\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}||^2}. \tag{7.30}$$

$$R^2 = \frac{||\mathbf{M}_{\boldsymbol{\iota}}\hat{\mathbf{y}}||^2}{||\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}||^2} = \frac{||\mathbf{M}_{\boldsymbol{\iota}}\mathbf{P}_{\mathbf{X}}\mathbf{y}||^2}{||\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}||^2}. \tag{7.31}$$

$$R^2 = \frac{||\mathbf{P}_{\mathbf{X}}\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}||^2}{||\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}||^2} \quad \Rightarrow \quad 0 \leq R^2 \leq 1 \qquad \text{(because of (7.10) with } \mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}\text{)}. \tag{7.32}$$

$$\begin{aligned} R^2 = \widehat{Corr}\left(\hat{y}, y\right)^2 &= \frac{\left(\sum_{t=1}^{n} \left(\hat{y}_t - \bar{\hat{y}}\right)\left(y_t - \bar{y}\right)\right)^2}{\left(\sum_{t=1}^{n} \left(\hat{y}_t - \bar{\hat{y}}\right)^2\right)\left(\sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2\right)} \\ &= \frac{\left(\hat{\mathbf{y}}^T\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}\right)^2}{\left(\hat{\mathbf{y}}^T\mathbf{M}_{\boldsymbol{\iota}}\hat{\mathbf{y}}\right)\left(\mathbf{y}^T\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}\right)} \quad \Rightarrow \quad 0 \leq R^2 \leq 1. \end{aligned} \tag{7.33}$$

Notes:

- All definitions are identical if there is a constant in the model.
- **Warning**: If no constant is in the model, not all definitions guarantee that $R^2 \in [0, 1]$, see the following table. Software gives different results depending on the definition used.
- **Properties** of different definitions for OLS:

Definition	used e. g. by	Codomain without constant in $\mathbf{X}$
(7.29)	Wooldridge (2009, Equation (2.38))	≥ 0
(7.30)	Davidson & MacKinnon (2004, Equation (2.55)), Wooldridge (2009, Equation (2.38))	≤ 1
(7.31)	Greene (2008, Equation (3-26))	≥ 0
(7.32)	Davidson & MacKinnon (2004, Equation (2.55))	$[0, 1]$
(7.33)	Greene (2008, Equation (3-27))	$[0, 1]$

- **Valid transformations if X with constant**:

$$\mathbf{P}_{\boldsymbol{\iota}}\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\boldsymbol{\iota}}. \tag{7.34a}$$

$$\mathbf{M}_{\boldsymbol{\iota}}\mathbf{M}_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}. \tag{7.34b}$$

$$\hat{\mathbf{y}}^T\mathbf{M}_{\boldsymbol{\iota}}\hat{\mathbf{y}} = \hat{\mathbf{y}}^T\mathbf{M}_{\boldsymbol{\iota}}\mathbf{y}. \tag{7.34c}$$

$$\boldsymbol{\iota}\,\bar{\hat{y}} = \mathbf{P}_{\boldsymbol{\iota}}\hat{\mathbf{y}} = \mathbf{P}_{\boldsymbol{\iota}}\mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{P}_{\boldsymbol{\iota}}\mathbf{y} = \boldsymbol{\iota}\,\bar{y} \quad \Leftrightarrow \quad \bar{\hat{y}} = \bar{y}. \tag{7.34d}$$

> **Proof sketch:** Apply the rules for calculating with projection matrices (7.22) and (7.23) for $\mathbf{x}_1 = \boldsymbol{\iota}$. (7.34c) holds since $\hat{\mathbf{y}}^T\mathbf{M}_{\boldsymbol{\iota}}\hat{\mathbf{u}} = \hat{\mathbf{y}}^T\mathbf{M}_{\boldsymbol{\iota}}\mathbf{M}_{\mathbf{X}}\hat{\mathbf{u}} = \mathbf{y}^T\mathbf{P}_{\mathbf{X}}^T\mathbf{M}_{\mathbf{X}}\hat{\mathbf{u}} = 0$. □

General remarks

- All definitions of R^2 (all except (7.33)) , which are based on the Pythagorean theorem, are only meaningful when using the OLS estimator. Otherwise, values less than zero or greater than one may occur.

- Since for (7.33) $0 \leq \widehat{Corr}\,(\hat{y}, y)^2 \leq 1$ holds, but the Pythagorean theorem has not been used, the **square of the empirical correlation coefficient** can always be used as a **goodness-of-fit** measure. It is often referred to as **pseudo-R^2**.

For reading: Davidson & MacKinnon (2004, Section 2.5)

4. $\sharp$ **Leverage effect**

- To estimate the effect of a possibly influential sample observation $(y_t, \mathbf{X}_t)$, the OLS estimators for the complete sample are compared with the OLS estimator for the sample without observation t. The latter is obtained by including a suitable dummy variable $\mathbf{e}_t$ in (6.1)
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_t\alpha + \mathbf{u}, \tag{7.35}$$
since $\mathbf{M}_{\mathbf{e}_t}\mathbf{y} = \mathbf{M}_{\mathbf{e}_t}\mathbf{X}\boldsymbol{\beta} + residuals$ (Frisch-Waugh-Lovell theorem) holds and because of $\mathbf{M}_{\mathbf{e}_t} = \mathbf{I} - \mathbf{e}_t\mathbf{e}_t^T$ the t-th observation is dropped.

- $\mathbf{P}_{\mathbf{X}}$ is sometimes called a **hat matrix** and its t-th diagonal element is therefore referred to as h_t. The latter corresponds to
$$0 \leq h_t = \mathbf{e}_t^T\mathbf{P}_{\mathbf{X}}\mathbf{e}_t = ||\mathbf{P}_{\mathbf{X}}\mathbf{e}_t||^2 \leq ||\mathbf{e}_t||^2 = 1. \tag{7.36}$$
It holds that $\sum_{t=1}^n h_t = tr(\mathbf{P}_{\mathbf{X}}) = k$, see tutorial or (Davidson & MacKinnon 2004, Section 2.6), and thus
$$\bar{h} = \frac{k}{n} \tag{7.37}$$
and if $\mathbf{X}$ contains a constant, it holds that
$$h_t \geq 1/n \quad \Leftrightarrow \quad h_t = ||\mathbf{P}_{\mathbf{X}}\mathbf{e}_t||^2 \geq ||\mathbf{P}_{\boldsymbol{\iota}}\mathbf{P}_{\mathbf{X}}\mathbf{e}_t||^2 = ||\mathbf{P}_{\boldsymbol{\iota}}\mathbf{e}_t||^2 = 1/n.$$

- If the OLS estimator for $\boldsymbol{\beta}$ based on (7.35) (without the t-th observation) is denoted with $\hat{\boldsymbol{\beta}}^{(t)}$, the difference of the OLS estimators can be given as
$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(t)} = \hat{\alpha}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{P}_{\mathbf{X}}\mathbf{e}_t = \frac{1}{1-h_t}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_t^T\hat{u}_t. \tag{7.38}$$
The t-th observation is possibly **influential** and thus a **leverage point** if

- h_t is large (close to 1) (refers to x-coordinates),
- $\hat{u}_t$ is large (refers to y-coordinate).

Proof: Verification of (7.38) via multiple applications of the properties of projection matrices etc. (details in Davidson & MacKinnon (2004, Section 2.6)):

$$\begin{aligned}
\mathbf{y} &= \mathbf{P}_{\mathbf{X},\mathbf{e}_t}\mathbf{y} + \mathbf{M}_{\mathbf{X},\mathbf{e}_t}\mathbf{y},\\
\mathbf{y} &= \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\mathbf{e}_t + \mathbf{M}_{\mathbf{X},\mathbf{e}_t}\mathbf{y},\\
\mathbf{P}_{\mathbf{X}}\mathbf{y} &= \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\mathbf{P}_{\mathbf{X}}\mathbf{e}_t + \mathbf{0}\\
\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(t)}\right) &= \hat{\alpha}\mathbf{P}_{\mathbf{X}}\mathbf{e}_t,\\
\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(t)} &= \hat{\alpha}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\underbrace{\mathbf{X}^T\mathbf{P}_{\mathbf{X}}\mathbf{e}_t}_{\mathbf{X}_t^T} = \frac{1}{1-h_t}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_t^T\hat{u}_t,
\end{aligned}$$

where by FWL theorem $\hat{\alpha} = \frac{\mathbf{e}_t^T\mathbf{M}_{\mathbf{X}}\mathbf{y}}{\mathbf{e}_t^T\mathbf{M}_{\mathbf{X}}\mathbf{e}_t} = \frac{\hat{u}_t}{1-h_t}$. □

-

R commands

In R, one obtains the h_t's and $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(t)}$, $t = 1, \ldots, n$ given by (7.36) and (7.38) with `influence(...)`.

For reading: Davidson & MacKinnon (2004, Section 2.6)

More on the geometry of the OLS estimator can be found in Ruud (2000), for example.

8. Multiple Regression: Interpretation

8.1. Parameter interpretation, functional form and data transformation

- The term *linear* in the linear regression model does not mean that there must be a linear relationship between the variables, but that the parameters enter the model linearly.
- Examples of models that are *linear* in the parameters:

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \\ y_t &= \beta_1 + \beta_2 \ln x_{t2} + u_t, \\ \ln y_t &= \beta_1 + \beta_2 \ln x_{t2} + \beta_3 x_{t3}^2 + u_t, \\ \ln y_t &= \beta_1 + \beta_2 x_t + u_t, \\ y_t &= \beta_1 + \beta_2 x_t^2 + u_t. \end{aligned}$$

- Examples of models that are *nonlinear* in the parameters:

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{t2}^{\gamma} + u_t \qquad \text{with parameters } \beta_1, \beta_2, \gamma, \\ y_t^{\gamma} &= \beta_1 + \beta_2 \ln x_{t2} + u_t \qquad \text{with parameters } \gamma, \beta_1, \beta_2, \\ y_t &= \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \frac{1}{1+\exp(\lambda(x_{t2}-\pi))}(\delta_1 + \delta_2 x_{t2} + \delta_3 x_{t3}) + u_t \\ &\qquad \text{with parameters } \beta_1, \beta_2, \beta_3, \lambda, \pi, \delta_1, \delta_2, \delta_3. \end{aligned}$$

- The last example allows smooth switching between two linear systems/regimes. Of course, almost infinitely many arbitrary forms of nonlinearity are conceivable. The estimation requires, for example, the **nonlinear OLS estimator**, which is covered in the master course **Advanced Econometrics**.

  However, linear regression models can approximate nonlinear relationships between dependent and independent variables well if the former provide a good (Taylor)approximation of the nonlinear relationship through variable transformation and/or consideration of terms with powers of higher order.

  Second-order Taylor expansion:

$$\begin{aligned} g(x,z) &= g(x_0,z_0) + g_x(x_0,z_0)(x-x_0) + g_z(x_0,z_0)(z-z_0) \\ &\quad + \frac{1}{2}\left[g_{xx}(x_0,z_0)(x-x_0)^2 + 2g_{xz}(x_0,z_0)(x-x_0)(z-z_0) + g_{zz}(x_0,z_0)(z-z_0)^2\right] \\ &\quad + \mathit{Remainder}(x,z,x_0,z_0), \end{aligned} \tag{8.1}$$

with the following notation of the partial derivatives:

$$
\begin{aligned}
g_x(x_0, z_0) &= \left.\frac{\partial g(x,z)}{\partial x}\right|_{x=x_0, z=z_0}, \\
g_{xz}(x_0, z_0) &= \left.\frac{\partial^2 g(x,z)}{\partial x \partial z}\right|_{x=x_0, z=z_0}.
\end{aligned}
$$

- **The natural logarithm in econometrics**

  Probably the most widely used transformation in empirical economics is the natural logarithm, or ln for short. The interpretation of the slope parameter has to be adjusted accordingly.

  Taylor approximation of the logarithmic function: $\ln(1+z) \approx z$ if z is close to 0.

  From this, an approximation important for the calculation of growth rates or returns can be derived:

$$
\begin{aligned}
(\Delta x_t)/x_{t-1} \qquad &:= (x_t - x_{t-1})/x_{t-1} \\
&\approx \ln\left(1 + (x_t - x_{t-1})/x_{t-1}\right), \\
(\Delta x_t)/x_{t-1} \qquad &\approx \ln(x_t) - \ln(x_{t-1}).
\end{aligned}
$$

  For relative changes $\Delta x_t / x_{t-1}$ close to zero this is a good approximation. Percentage values are obtained by multiplying by 100:

$$
100\Delta \ln(x_t) \approx \%\Delta x_t = 100(x_t - x_{t-1})/x_{t-1}.
$$

  Accordingly, for small $\Delta x_t / x_{t-1}$, percentage changes can be well approximated via $100[\ln(x_t) - \ln(x_{t-1})]$.

Economic interpretation of OLS parameters

- Consider the **ratio of relative changes** of two **non-stochastic** variables y and x

$$
\frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\%\text{-change of } y}{\%\text{-change of } x} = \frac{\%\Delta y}{\%\Delta x}.
$$

  If $\Delta y \to 0$ and $\Delta x \to 0$, then $\frac{\Delta y}{\Delta x} \to \frac{dy}{dx}$.

- Applying this result to the above ratio gives the **elasticity**

$$
\eta(x) = \frac{dy}{dx}\frac{x}{y}. \tag{8.2}
$$

- **Interpretation**: If the relative change of x is 0.01, then the relative change of y is exactly $0.01\eta(x)$.

  Or: If x changes by 1%, then y changes by $\eta(x)$%.

- If y, x are **random variables**, the elasticity is defined based on the conditional expected value of y given x:
$$\eta(x) = \frac{dE[y|x]}{dx}\,\frac{x}{E[y|x]}.$$

  This can be derived by considering
$$\frac{\frac{E[y|x_1=x_0+\Delta x]-E[y|x_0]}{E[y|x_0]}}{\frac{\Delta x}{x_0}} =$$
$$\frac{E[y|x_1 = x_0 + \Delta x] - E[y|x_0]}{\Delta x}\,\frac{x_0}{E[y|x_0]}$$

  and then letting Δx go towards 0.

- **Notation**:
$$\Delta E[y|x_1, \ldots, x_j, \ldots, x_k] := E[y|x_1, \ldots, x_j + \Delta x_j, \ldots, x_k] - E[y|x_1, \ldots, x_j, \ldots, x_k]$$
$$\approx \frac{\partial E[y|x_1, \ldots, x_j, \ldots, x_k]}{\partial x_j}\Delta x_j$$

Different models and interpretations of β_j

For each model, we assume below that it is correctly specified and that the conditional expected value of the errors given the regressors is zero.

- In the following, the index t does not appear because the model of the population is considered.

- **Models that are linear in variables (level-level)**
$$y = \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_k x_k + u.$$

  It is $E[y|x_1, \ldots, x_k] = \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_k x_k$
$$\frac{\partial E[y|x_1, \ldots, x_k]}{\partial x_j} = \beta_j$$

  and thus approximately
$$\Delta E[y|x_1, \ldots, x_k] = \beta_j \Delta x_j.$$

  In words: The slope parameter indicates the *absolute* change in the conditional expected value of the dependent variable y when the independent variable x_j changes by one *unit*, ceteris paribus.

- **Linear-log models (level-log)**

  $$y = \beta_1 \ln x_1 + \ldots + \beta_j \ln x_j + \ldots + \beta_k \ln x_k + u.$$

  It holds that

  $$\frac{\partial E[y|x_1, \ldots, x_k]}{\partial x_j} = \beta_j \frac{1}{x_j}$$

  and thus approximately

  $$\Delta E[y|x_1, \ldots, x_k] \approx \beta_j \Delta \ln x_j = \frac{\beta_j}{100} 100 \Delta \ln x_j \approx \frac{\beta_j}{100} \% \Delta x_j.$$

  In words: The conditional expected value of y changes by $\beta_j/100$ *units*, when x_j changes by 1%.

- **Log-linear models (log-level)**

  $$\ln y = \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_k x_k + u$$

  or

  $$y = e^{\ln y} = e^{\beta_1 x_1 + \ldots + \beta_k x_k + u} = e^{\beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_k x_k} e^u.$$

  Thus

  $$E[y|x_1, \ldots, x_k] = e^{\beta_1 x_1 + \ldots + \beta_k x_k} E[e^u | x_1, \ldots, x_k].$$

  If $E[e^u|x_1, \ldots, x_k]$ is constant, it holds that

  $$\frac{\partial E[y|x_1, \ldots, x_k]}{\partial x_j} = \beta_j \underbrace{e^{\beta_1 x_1 + \ldots + \beta_k x_k} E[e^u|x_1, \ldots, x_k]}_{E[y|x_1, \ldots, x_k]} = \beta_j E[y|x_1, \ldots, x_k].$$

  One obtains approximately

  $$\frac{\Delta E[y|x_1, \ldots, x_k]}{E[y|x_1, \ldots, x_k]} \approx \beta_j \Delta x_j, \quad \text{or} \quad \% \Delta E[y|x_1, \ldots, x_k] \approx 100 \beta_j \Delta x_j$$

  In words: The conditional expected value of y changes by $100\,\beta_j\%$ when x_j changes by one *unit*.

- **Log-log models** are often called **log-linear** or **constant-elasticity** models and are very popular in empirical practice

  $$\ln y = \beta_1 \ln x_1 + \ldots + \beta_2 \ln x_k + u.$$

  Similar to above, it can be shown that the following holds:

  $$\frac{\partial E[y|x_1, \ldots, x_k]}{\partial x_j} = \beta_j \frac{E[y|x_1, \ldots, x_k]}{x_j}, \quad \text{and thus} \quad \beta_j = \eta(x_1, \ldots, x_k),$$

  if $E[e^u|x_1, \ldots, x_k]$ is constant.

  In this model, the slope parameter of the log-log model is just equal to the elasticity for the original level variables $E[y|x_1, \ldots, x_k]$ and x_j. In words: The conditional expected value of y changes by $\beta_j\%$ when x_j changes by 1%.

- The transformations of regressors can be different for different regressors.

Example: $y = \beta_1 + \beta_2 \ln x_{t2} + \beta_3 x_{t3}^2 + u$

Trade flows: (Continuation of the empirical analysis of section 6.3)

R code (Extract from the R program in section A.4)

```
summary(lm(trade_0_d_o ~ wdi_gdpusdcr_o))           #level - level model
summary(lm(trade_0_d_o ~ log(wdi_gdpusdcr_o)))      #level - log model
summary(lm(log(trade_0_d_o) ~ wdi_gdpusdcr_o))      #log - level model
summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o))) #log - log models
summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o)+log(cepii_dist)))
```

Listing 8.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

R output for log-level model

```
Call:
lm(formula = log(trade_0_d_o) ~ wdi_gdpusdcr_o)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6770 -1.4776  0.3231  2.1255  3.4143

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.086e+01  3.776e-01  55.248  < 2e-16 ***
wdi_gdpusdcr_o 5.466e-13  2.010e-13   2.719  0.00915 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.505 on 47 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.1359,     Adjusted R-squared: 0.1175
F-statistic: 7.392 on 1 and 47 DF,  p-value: 0.009148
```

Interpretation: An increase in GDP in the exporting country by \$1 billion $(= 10^9$ US dollar) leads to an average increase in imports of $100\,\hat{\beta}_2\,10^9\% = 5.466 \cdot 10^{-13} \cdot 10^{11}\% = 0.055\%$. Accordingly, an increase of 100 billion, which is roughly equivalent to a 1% increase, results in an average increase of 5.5%.

R output for log-log model

```
Call:
lm(formula = log(trade_0_d_o) ~ log(wdi_gdpusdcr_o))

Residuals:
    Min      1Q  Median      3Q     Max
-2.6729 -1.0199  0.2792  1.0245  2.3754

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -5.77026    2.18493  -2.641   0.0112 *
log(wdi_gdpusdcr_o)  1.07762    0.08701  12.384   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.305 on 47 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.7654,     Adjusted R-squared: 0.7604
F-statistic: 153.4 on 1 and 47 DF,  p-value: < 2.2e-16
```

Interpretation: A 1% increase in the GDP of the exporting country is accompanied by an average increase in imports of 1,077%.

8.2. Data scaling

- **Scaling of the dependent variable**:
  - Initial model:
  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$
  - Variable transformation: $y_i^* = a \cdot y_i$ with scaling factor a. → New transformed regression equation:
  $$\begin{aligned} \underbrace{a\mathbf{y}}_{\mathbf{y}^*} &= \mathbf{X}\underbrace{a\boldsymbol{\beta}}_{\boldsymbol{\beta}^*} + \underbrace{a\mathbf{u}}_{\mathbf{u}^*} \\ \mathbf{y}^* &= \mathbf{X}\boldsymbol{\beta}^* + \mathbf{u}^* \end{aligned} \tag{8.3}$$
  - OLS estimator for $\boldsymbol{\beta}^*$ from (8.3):
  $$\begin{aligned} \hat{\boldsymbol{\beta}}^* &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}^* \\ &= a\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = a\hat{\boldsymbol{\beta}}. \end{aligned}$$
  - Error variance for homoscedastic (see (9.10)) and uncorrelated errors:
  $$Var(\mathbf{u}^*|\mathbf{X}) = Var(a\mathbf{u}|\mathbf{X}) = a^2Var(\mathbf{u}|\mathbf{X}) = a^2\sigma^2\mathbf{I}.$$
  - Variance-covariance matrix:
  $$Var(\hat{\boldsymbol{\beta}}^*|\mathbf{X}) = \sigma^{*2}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = a^2\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = a^2Var(\hat{\boldsymbol{\beta}}|\mathbf{X})$$
  - t-statistic:
  $$t^* = \frac{\hat{\beta}^*{}_j - 0}{\sigma_{\hat{\beta}^*{}_j}} = \frac{a\hat{\beta}_j}{a\sigma_{\hat{\beta}_j}} = t.$$
- **Scaling of explanatory variables**:
  - Variable transformation: $\mathbf{X}^* = \mathbf{X}\mathbf{A}$, where $\mathbf{A}$ is quadratic and in the case of variable scaling diagonal. $\mathbf{A}$ must be invertible, cf. section 7.1.2. New regression equation:
  $$\mathbf{y} = \mathbf{X}\mathbf{A}\mathbf{A}^{-1}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{u}. \tag{8.4}$$

  - OLS estimator for $\boldsymbol{\beta}^*$ from (8.4):

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= \left(\mathbf{X}^{*T}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*T}\mathbf{y} = \left(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{X}^T\mathbf{y}\\ &= \mathbf{A}^{-1}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}.\end{aligned}$$

  - Result: The mere size of β_j does not indicate how relevant the influence of the jth regressor is. One must always take into account the scaling of the variable.

    Example: In (8.1), a simple log-level model was estimated for the impact of GDP on imports. The parameter estimate $\hat{\beta}_{GDP} = 5.466 \cdot 10^{-13}$ seems to be quite small. However, if we take into account that GDP is measured in dollars, this parameter value is not small at all. If we rescale GDP into billions of dollars (using $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 10^{-9}\end{smallmatrix}\right)$), we get $\hat{\beta}^*{}_{GDP} = 5.466 \cdot 10^{-4}$.

- **Scaling of variables in logarithmic form** only changes the constant β_1, since $\ln y^* = \ln ay = \ln a + \ln y$.
- **Standardised coefficients**: see Wooldridge (2009, Section 6.1) or **Introduction to Econometrics**, section 6.2.

8.3. Qualitative data as regressors

8.3.1. Dummy variable or binary variable

A **binary variable** can take exactly **two** different values and allows **two** qualitatively different states to be described.

Examples: female vs. male, employed vs. unemployed, etc.

- Generally, these values are coded as $D = 0$ and $D = 1$. This allows for a very **simple interpretation**:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1}x_{k-1} + \delta D + u,$$

$$\begin{aligned}E[y|x_1, \ldots, x_{k-1}, D = 1] - E[y|x_1, \ldots, x_{k-1}, D = 0] =&\\ \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1}x_{k-1} + \delta&\\ - \left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1}x_{k-1}\right) &= \delta\end{aligned}$$

  The coefficient δ of a dummy variable thus indicates by how much the intercept shifts when $D = 1$ instead of $D = 0$. *All* slope parameters β_i remain *unchanged*, where $i = 1, \ldots, k-1$ (without constant) resp. $i = 2, \ldots, k-1$ (with constant).

  Note: In order to interpret the coefficient of a dummy variable, one must know the reference group. The reference group is the group for which the dummy takes the value zero.

  Example: wage regression:

- Initial question: Is the income of women significantly lower than that of men?
- Data: Sample of $n = 526$ workers in the U.S. from 1976. (Source: Examples 2.4, 7.1 in Wooldridge (2009)).

Data:

- *wage*: Hourly wage in USD,
- *educ*: Duration of education,
- *exper*: Work experience in years,
- *tenure*: Duration of employment with current company,
- *female*: dummy=1 if female, dummy=0 otherwise.

R code (Extract from R program in section A.6)

```
# Specification of the working directory
# in which the R program and the data are located

WD            <- getwd() # set the directory of the R file and
setwd(WD)                # set it as working directory

# Import the data
# The data file "wage1.txt" must be located in the same directory as the
# R file
wage_data     <- read.table("wage1.txt", header = TRUE)
attach(wage_data)

# Wage regression with dummy variable, see section 8.4.1
wage_mod_1_kq <- lm(log(wage) ~ female +
                      educ + exper + I(exper^2) + tenure + I(tenure^2))
summary(wage_mod_1_kq)
```

Listing 8.2: ./R_code/8_4_Interpretationen_Wage_eng.R

R output

```
Call:
lm(formula = log(wage) ~ female + educ + exper + I(exper^2) +
    tenure + I(tenure^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.83160 -0.25658 -0.02126  0.25500  1.13370

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4166910  0.0989279   4.212 2.98e-05 ***
female      -0.2965110  0.0358054  -8.281 1.04e-15 ***
educ         0.0801966  0.0067573  11.868  < 2e-16 ***
exper        0.0294324  0.0049752   5.916 6.00e-09 ***
I(exper^2)  -0.0005827  0.0001073  -5.431 8.65e-08 ***
tenure       0.0317139  0.0068452   4.633 4.56e-06 ***
I(tenure^2) -0.0005852  0.0002347  -2.493    0.013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3998 on 519 degrees of freedom
Multiple R-squared: 0.4408,     Adjusted R-squared: 0.4343
F-statistic: 68.18 on 6 and 519 DF,  p-value: < 2.2e-16
```

The parameter δ corresponds to the difference in logarithmised income between female and male workers, *holding everything else constant* (e. g. duration of education, experience, etc.).

For the interpretation of the parameters of regressors that also occur as quadratic terms in the model, see section 8.4.

- **Approximate partial effect in log-level models** The parameter δ corresponds to an **approximate prediction** of the difference in y when $\ln y$ is modelled and the dummy variable changes ceteris paribus.

  Example: wage regression: The approximate average difference in income between female and male workers is -29.65% in 1976.

  What is the exact average difference in income?

- **Expected value of a log-normally distributed random variable**: If $\ln z \sim N(\mu, \sigma^2)$, then z is log-normally distributed with

$$E[z] = E\left[e^{\ln z}\right] = e^{\mu+\sigma^2/2}. \tag{8.5}$$

  If a **conditionally log-normally distributed random variable**

$$\ln z|x \sim N(g(x), \sigma^2(x))$$

  is given, then

$$E[z|x] = E\left[e^{\ln z}|x\right] = e^{g(x)+\sigma^2(x)/2}. \tag{8.6}$$

- **Exact partial effect in log-level models**

$$\ln y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1}x_{k-1} + \delta D + u,$$

  Assumption for calculation: $u|x_1, \ldots, x_{k-1}, D \sim N(0, \sigma^2)$.

  Then, $E\left[e^u|x_1, \ldots, x_{k-1}, D\right] = e^{\sigma^2/2}$ and

$$\frac{E[y|x_1, \ldots, x_{k-1}, D=1] - E[y|x_1, \ldots, x_{k-1}, D=0]}{E[y|x_1, \ldots, x_{k-1}, D=0]} = \left(e^\delta - 1\right) \tag{8.7}$$

  Proof:

$$\begin{aligned}
&E[y|x_1, \ldots, x_{k-1}, D=1] - E[y|x_1, \ldots, x_{k-1}, D=0] \\
&= e^{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{k-1}x_{k-1}+\delta} E\left[e^u|x_1, \ldots, x_{k-1}, D=1\right] \\
&- e^{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{k-1}x_{k-1}} E\left[e^u|x_1, \ldots, x_{k-1}, D=0\right] \\
&= e^{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{k-1}x_{k-1}} e^\delta e^{\sigma^2/2} \\
&- e^{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{k-1}x_{k-1}} e^{\sigma^2/2} \\
&= E[y|x_1, \ldots, x_{k-1}, D=0]\left(e^\delta - 1\right).
\end{aligned}$$

  Dividing the difference by $E[y|x_1, \ldots, x_{k-1}, D=0]$ yields (8.7). □

Example: wage regression: Question: What is the exact wage difference? Answer: $100(e^{-0.2965} - 1) = -25.66\%$, if one assumes normally distributed errors.

- **Note:** **If Δx_j is not close to zero, the exact partial effect**

$$\frac{E[y|x_1,\dots,x_j+\Delta x_j,\dots,x_k] - E[y|x_1,\dots,x_j,\dots,x_k]}{E[y|x_1,\dots,x_j,\dots,x_k]} = \left(e^{\beta_j \Delta x_j} - 1\right) \qquad (8.8)$$

should always **be calculated**, because then the Taylor approximation does not approximate the logarithm function well and thus the value of the approximate partial effect is not very reliable.

- **Important:** **For comparisons between groups, comparing conditional means is much more meaningful than comparing unconditional means**.

Example: wage regression Comparison of the wages of men and women: Assuming normally distributed errors, the exact partial effect is -25.66%. On average, women earn about 26% less than men after taking into account education, work experience and time in a company.

If, in contrast, one compares the unconditional mean values, e. g. with the

R code (Extract from R program in section A.6)

```
# Relative difference of unconditional mean wages of women and men
(mean(wage[female==1])-mean(wage[female==0]))/mean(wage[female==0])
  # alternative calculation possibility
wage_mean <- lm(wage~0+female+I(1-female))
(wage_mean$coef[1]-wage_mean$coef[2])/wage_mean$coef[2]
```

Listing 8.3: ./R_code/8_4_Interpretationen_Wage_eng.R

then the difference is 35.38%, i. e. it is considerably larger, because men and women obviously also differ in terms of education, work experience and time in a company.

So it is essential to take relevant influencing factors into account!

- **Exact and approximate prediction in log-level model:** Expected value of y given the regressors $x_1, \dots, x_k$ is given by

$$E[y|x_1,\dots,x_k] = e^{\beta_1 x_1 + \dots + \beta_k x_k} \cdot E[e^u|x_1,\dots,x_k].$$

The true value of $E[e^u|x_1,\dots,x_k]$ depends on the probability distribution of u.

If $u|x_1,\dots,x_k \sim N(0,\sigma^2)$, then $E[e^u|x_1,\dots,x_k] = e^{E[u|x_1,\dots,x_k]+\sigma^2/2}$. The **exact prediction** ist thus

$$E[y|x_1,\dots,x_k] = e^{\beta_1 x_1 + \dots + \beta_k x_k + \sigma^2/2}.$$

Example: wage regression — exact prediction: How much does a woman with 12 years of education, 10 years of experience and one year of employment earn? The exact prediction of the hourly wage is

$$\begin{aligned}
&E[wage|female = 1, educ = 12, exper = 10, tenure = 1]\\
&= \exp(0.4167 - 0.2965 \cdot 1 + 0.0802 \cdot 12 + 0.02943 \cdot 10\\
&\quad - 0.0006 \cdot (10^2) + 0.0317 \cdot 1 - 0.0006 \cdot (1^2) + 0.3998^2/2)\\
&= 4.18,
\end{aligned}$$

where σ^2 is estimated by s^2 (9.25). The exact value of the mean hourly wage of the person described is thus about \$4.18.

If one omits the term $e^{s^2/2}$ from the prediction, then one obtains an **approximate prediction**.

Example: wage regression — approximate prediction:

$$\begin{aligned}
&E[\ln(wage)|female = 1, educ = 12, exper = 10, tenure = 1]\\
&= 0.4167 - 0.2965 \cdot 1 + 0.0802 \cdot 12 + 0.0294 \cdot 10\\
&\quad - 0.0006 \cdot (10^2) + 0.0317 \cdot 1 - 0.0006 \cdot (1^2)\\
&= 1.35
\end{aligned}$$

Accordingly, the expected hourly wage is **approximately** $\exp(1.35) = 3.86$ US dollar and thus 30 cents less than the exact value.

Conclusion: for **log-log** and **log-level models**:

- for exact predictions one needs the empirical variance (with normal distribution assumption),
- for approximate predictions it is sufficient to “plug” it into the regression equation.

8.3.2. Multiple subgroups

Illustration with an example:

Example: wage regression: (continued from section 8.3.1)

A worker is female or male and married or unmarried $\Longrightarrow$ 4 subgroups.

1. female and unmarried
2. female and married
3. male and unmarried
4. male and married

Procedure

- We define a group as **reference group**, such as: female and unmarried.
- Define dummy variables for the other subgroups.

R code (Extract from R program in section A.6)

```
femmarr       <- female * married
malesing      <- (1 - female) * (1 - married)
malemarr      <- (1 - female) * married

wage_mod_2_kq <- lm(log(wage) ~ femmarr + malesing + malemarr +
                      educ + exper + I(exper^2) + tenure + I(tenure^2))
summary(wage_mod_2_kq)
```

Listing 8.4: ./R_code/8_4_Interpretationen_Wage_eng.R

R output

```
Call:
lm(formula = log(wage) ~ femmarr + malesing + malemarr + educ +
    exper + I(exper^2) + tenure + I(tenure^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.89697 -0.24060 -0.02689  0.23144  1.09197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2110279  0.0966445   2.184   0.0294 *
femmarr     -0.0879174  0.0523481  -1.679   0.0937 .
malesing     0.1103502  0.0557421   1.980   0.0483 *
malemarr     0.3230259  0.0501145   6.446 2.64e-10 ***
educ         0.0789103  0.0066945  11.787  < 2e-16 ***
exper        0.0268006  0.0052428   5.112 4.50e-07 ***
I(exper^2)  -0.0005352  0.0001104  -4.847 1.66e-06 ***
tenure       0.0290875  0.0067620   4.302 2.03e-05 ***
I(tenure^2) -0.0005331  0.0002312  -2.306   0.0215 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3933 on 517 degrees of freedom
Multiple R-squared: 0.4609,     Adjusted R-squared: 0.4525
F-statistic: 55.25 on 8 and 517 DF,  p-value: < 2.2e-16
```

Examples of interpretation:

- Ceteris paribus, married women earn on average approximately 8.8% less than unmarried women. However, this effect is only significant at the 10% level (with a two-sided test).
- The expected wage difference between married men and women is, ceteris paribus, about $32.3 - (-8.8) = 41.1\%$. No t-statistic can be calculated directly for this, but an F-statistic can. (To be able to perform a t-test, run the estimation again with one of the two subgroups as reference group).

Remarks:

- It is not recommended to create a dummy variable for *all* subgroups because then the differences with respect to the reference group cannot be tested directly.

- If one uses a dummy variable for all subgroups, no constant may be included in the model, otherwise $\mathbf{X}$ has reduced column rank. Why?

Ordinal data in the regression

> **Ranking of universities:**
>
> The differences in quality between ranks 1 and 2, or ranks 11 and 12, can vary enormously. Therefore, rankings are **not** suitable as regressors. Instead, we assign a dummy variable D_j to each university except one (the “reference category”), which means that we have to estimate some new parameters (Therefore, in the foreign trade example, we might need to split the variable *openess* into several dummies...).
>
> Note: The coefficient of a dummy variable D_j now indicates the shift of the intercept between university j and the reference university.
>
> Occasionally the ranking list is too long, so that too many parameters would have to be estimated. It is then usually helpful to combine the data into groups, e. g. ranks 1-10, 11-20, etc..

8.3.3. Interactions and dummy variables

- **Interactions between dummy variables**:
  - e. g. to define subgroups (e. g. married men).
  - Note that a meaningful interpretation and comparison of the influences of the subgroups depends crucially on a correct use of the dummies. For example, we add the dummies *male* and *married* and their interaction to a wage equation
    $$y = \beta_1 + \delta_1 male + \delta_2 married + \delta_3 male \cdot married + \dots.$$
    A comparison between married and unmarried men is thus given by
    $$\begin{aligned}&E[y|male = 1, married = 1] - E[y|male = 1, married = 0]\\&= \beta_1 + \delta_1 + \delta_2 + \delta_3 + \dots - (\beta_1 + \delta_1 + \dots) = \delta_2 + \delta_3.\end{aligned}$$

- **Interactions between dummies and quantitative variables**:
  - This allows for different slope parameters for different groups
    $$y = \beta_1 + \beta_2 D + \beta_3 x + \beta_4 (x \cdot D) + u.$$
    Note: Here, β_2 denotes the differences between the two groups *only* for the case $x = 0$. If $x \neq 0$, this difference is
    $$\begin{aligned}&E[y|D = 1, x] - E[y|D = 0, x]\\&= \beta_1 + \beta_2 \cdot 1 + \beta_3 x + \beta_4 (x \cdot 1) - (\beta_1 + \beta_3 x)\\&= \beta_2 + \beta_4 x.\end{aligned}$$

Even if β_2 is negative, the overall effect may be positive!

Example: wage regression (continued from section 8.3.1)

Interaction of dummy with the regressor `educ`:

R code (Extract from R program in section A.6)

```
wage_mod_3_kq <- lm(log(wage) ~ female +
                      educ + exper + I(exper^2) + tenure + I(tenure^2) +
                      I(female*educ))
summary(wage_mod_3_kq)
```

Listing 8.5: ./R_code/8_4_Interpretationen_Wage_eng.R

R output

```
Call:
lm(formula = log(wage) ~ female + educ + exper + I(exper^2) +
    tenure + I(tenure^2) + I(female * educ))

Residuals:
     Min       1Q   Median       3Q      Max
-1.83264 -0.25261 -0.02374  0.25396  1.13584

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.3888060  0.1186871   3.276  0.00112 **
female           -0.2267886  0.1675394  -1.354  0.17644
educ              0.0823692  0.0084699   9.725  < 2e-16 ***
exper             0.0293366  0.0049842   5.886 7.11e-09 ***
I(exper^2)       -0.0005804  0.0001075  -5.398 1.03e-07 ***
tenure            0.0318967  0.0068640   4.647 4.28e-06 ***
I(tenure^2)      -0.0005900  0.0002352  -2.509  0.01242 *
I(female * educ) -0.0055645  0.0130618  -0.426  0.67028
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4001 on 518 degrees of freedom
Multiple R-squared: 0.441,      Adjusted R-squared: 0.4334
F-statistic: 58.37 on 7 and 518 DF,  p-value: < 2.2e-16
```

The *return to education*, i. e. the average hourly wage difference for an additional year of education, is not gender-specific, as the p-value of the corresponding interaction term is above any common significance level.

- **Conclusion**: If a regression variable occurs in several terms (interactions, quadratic terms) in the model, generally more parameters have to be considered to interpret a relationship.
- **Tests for group differences**
  - are carried out using F-tests.
  - **Chow test**: Allows to test whether group differences exist in terms of group-specific intercepts and/or (at least one) group-specific slope parameter.

    Example:

$$y = \beta_1 + \beta_2 D + \beta_3 x_1 + \beta_4 (x_1 \cdot D) + \beta_5 x_2 + \beta_6 (x_2 \cdot D) + u. \tag{8.9}$$

Pair of hypotheses:

$$\begin{aligned}\mathrm{H}_0 :& \beta_2 = \beta_4 = \beta_6 = 0 \quad \text{vs.}\\ \mathrm{H}_1 :& \beta_2 \neq 0 \text{ and/or } \beta_4 \neq 0 \text{ and/or } \beta_6 \neq 0.\end{aligned}$$

8.4. Models with quadratic regressors

- **Models with quadratic regressors**:
  - As an example, assume the following multiple regression model

    $$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3^2 + u.$$

    The marginal impact of a change in x_3 on the conditional expected value y is

    $$\frac{\partial E[y|x_1, \ldots, x_3]}{\partial x_3} = \beta_3 + 2\beta_4 x_3.$$

    Thus, a change in x_3 by Δx_3 ceteris paribus affects the independent variable y on average by

    $$\Delta E[y|x_1, \ldots, x_3] = (\beta_3 + 2\beta_4 x_3)\Delta x_3.$$

    So the effect obviously depends on the level of x_3 (and thus an interpretation of β_3 alone is not meaningful!).

  - In some empirical applications, one uses quadratic or logarithmic regressors to approximate a **non-linear regression function**.

    Example: Non-constant elasticities can be approximated as follows

    $$\ln y = \beta_1 + \beta_2 x_2 + \beta_3 \ln x_3 + \beta_4 (\ln x_3)^2 + u.$$

    The elasticity of y with respect to x_3 is therefore

    $$\beta_3 + 2\beta_4 \ln x_3$$

    and is constant if and only if $\beta_4 = 0$.

  - **Example: trade flows:** So far we have only considered multiple regression models that were log-log or log-level specified in the initial variables.

    Now we want to consider another specification for modelling imports in which a logarithmised regressor is also present in quadratic form in the equation.

    Model 5: (Models 2 and 3a were estimated in section 6.3. Models 1, 3b and 4 are first introduced in section 10.3.)

    $$\begin{aligned}\ln(Imports) = \beta_1 + \beta_2 \ln(GDP) + \beta_3 \left(\ln(GDP)\right)^2 + \beta_4 \ln(Distance)\\ + \beta_5\, Openness + \beta_6 \ln Area + u.\end{aligned}$$

It was just shown that then for the elasticity of *imports* with respect to GDP one has:

$$\beta_2 + 2\beta_3 \ln(GDP). \tag{8.10}$$

The estimation of model 5 was carried out with the following

R code (Extract from R program in section A.4)

```
mod_5_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) +
  I(log(wdi_gdpusdcr_o)^2) + log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o)

mod_5_kq      <- lm(mod_5_formula)
summary(mod_5_kq)
```

Listing 8.6: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

R output

```
Call:
lm(formula = log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + I(log(wdi_gdpusdcr_o)^2) +
    log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))

Residuals:
    Min      1Q  Median      3Q     Max
-2.0672 -0.5451  0.1153  0.5317  1.3870

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -35.23314   17.44175  -2.020  0.04964 *
log(wdi_gdpusdcr_o)        3.90881    1.32836   2.943  0.00523 **
I(log(wdi_gdpusdcr_o)^2)  -0.05711    0.02627  -2.174  0.03523 *
log(cepii_dist)           -0.74856    0.16317  -4.587 3.86e-05 ***
ebrd_tfes_o                0.41988    0.20056   2.094  0.04223 *
log(cepii_area_o)         -0.13238    0.08228  -1.609  0.11497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8191 on 43 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.9155,     Adjusted R-squared: 0.9056
F-statistic: 93.12 on 5 and 43 DF,  p-value: < 2.2e-16
```

Those already familiar with significance tests will see that the quadratic term is significant at the 5% level. This provides statistical evidence for a non-linear elasticity. Substituting the parameter estimates into (8.10), one gets

$$\eta(GDP) = 3.908811 - 0.057108 \ln(GDP).$$

Figure 8.1 plots the elasticity of $\eta(GDP)$ for each observed value of GDP against GDP (generated with the following R code).

R code (Extract from R program in section A.4)

```
elast_gdp     <- mod_5_kq$coef[2] + 2* mod_5_kq$coef[3]*log(wdi_gdpusdcr_o)
# Create scatterplot
if (save.pdf)  pdf("plot_modell5_elast.pdf.pdf", height=6, width=6)
plot(wdi_gdpusdcr_o, elast_gdp, pch = 16, col = "blue", main = "GDP-Elasticity")
```

Listing 8.7: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

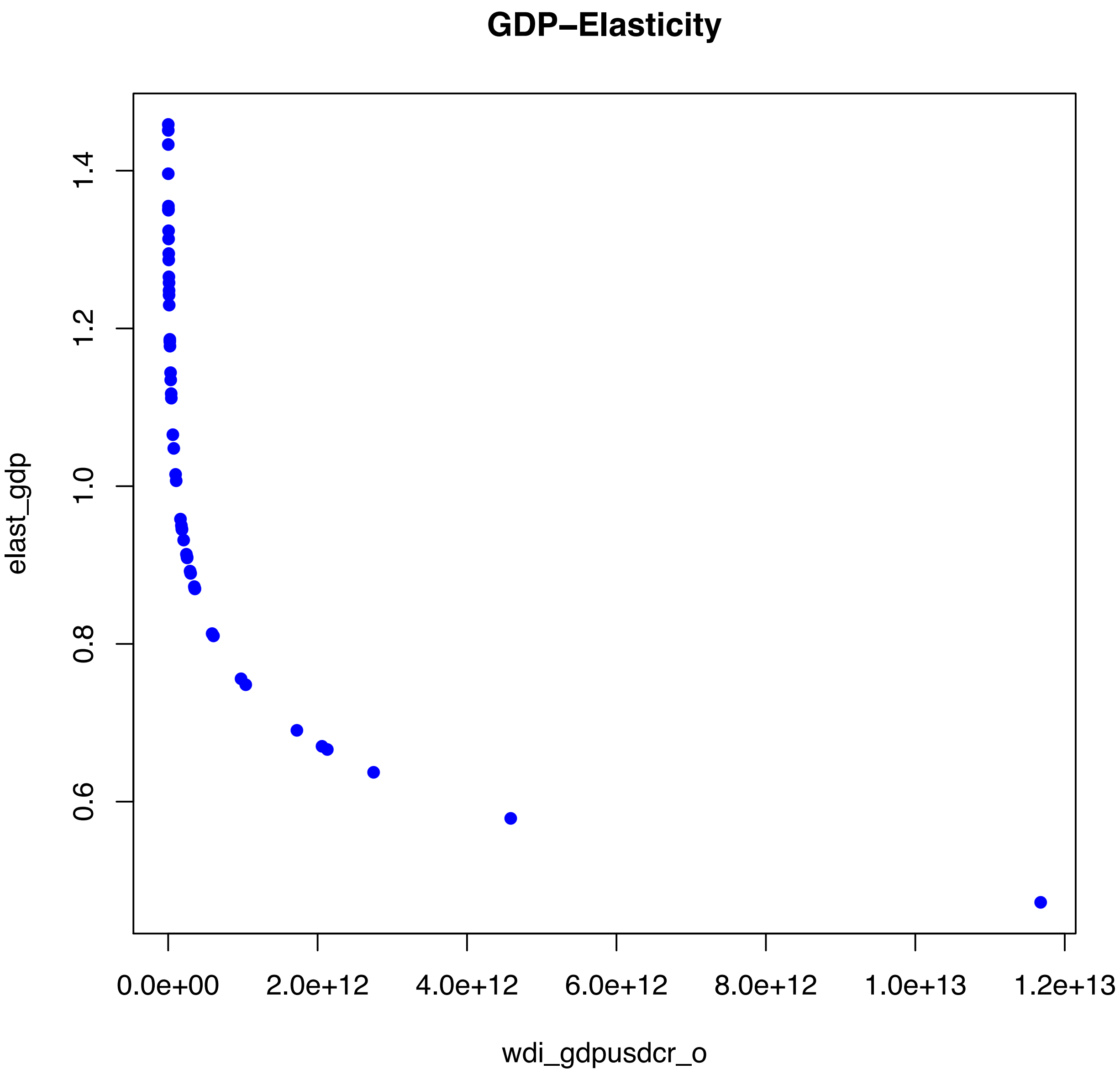


Figure 8.1.: Elasticity of $\eta(GDP)$ (R code see example on trade flows)

The GDP elasticity of imports is much larger for small economies (as measured by GDP) than for large economies. In other words, for small economies, an increase in GDP has a much greater impact on imports than for large economies.

Caution: Non-linearities sometimes result from missing relevant variables. Can you guess which control variable should be added to model 5?

- **Interactions**: Example:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3 x_2 + u.$$

The marginal effect of a change in x_3 is given by

$$\Delta E[y|x_2, x_3] = (\beta_3 + \beta_4 x_2)\Delta x_3.$$

Thus, the marginal effect also depends on the level of x_2!

For reading: Chapter 6 (without section 6.4) and chapter 7 (without sections 7.5 and 7.6) in Wooldridge (2009).

9. Statistical properties of the OLS estimator: expected value and covariance

- The algebraic and geometric properties of the **OLS estimator**

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{6.5}$$

  for the **multiple linear regression model**

  $$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t, \quad t = 1, \ldots, n, \tag{5.24}$$
  $$y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t, \quad t = 1, \ldots, n, \tag{5.25}$$
  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{6.1}$$

  were analysed in chapter 7.

- To answer many questions, knowledge of the algebraic and geometric properties of the OLS estimator is not sufficient, but **knowledge of the statistical properties of the OLS estimator is necessary**.

  Examples:

  - Example based on trade flows
  - Based on the estimated parameter vector $\hat{\boldsymbol{\beta}}$, what can be concluded about the values of the parameter vector $\boldsymbol{\beta}$ for the DGP (for the population)?
  - To what extent can it be verified that the chosen econometric model contains the DGP?
  - Suppose you have another sample with k regressors on the same question.
    - ∗ Why do the two OLS estimates probably differ?
    - ∗ Which of the two OLS estimates do you choose?
    - ∗ Should you merge the OLS results of both samples?

  If statements about the DGP are to be made on the basis of the sample, inductive statistical methods are necessary. In order to be able to say something about the properties of such statements, assumptions about the DGP and the econometric model are necessary.

  Which assumptions lead to which (statistical) properties of the OLS estimator is the subject of this chapter.

If the population corresponded to the sample and we were only interested in key figures such as sample correlation or coefficient of determination, we would already be done.

- **Important properties of an estimator**:
  - **Unbiasedness**
  - **Variance**
  - **Mean squared error (MSE)**
  - **Consistency**
  - **Efficiency**
  - **Exact distribution in finite samples**
  - **Asymptotic distribution**

Analysis of the properties of the OLS estimator Overview of sections

	OLS estimator for		
	$\boldsymbol{\beta}$	σ^2	covariance matrix
Unbiasedness	9.1.1	9.5	9.3
Variance	9.3		
MSE	9.6		
Consistency	9.2		
Efficiency	9.4		
Exact distribution in finite samples	11.1		
Asymptotic distribution	11.2		

9.1. Unbiasedness of the OLS estimator

Review of section 5.4:

- The **bias** of a parameter estimator $\hat{\theta}$ for θ is defined as

$$E[\hat{\theta}] - \theta_0,$$

  where θ_0 is the true parameter value, i. e. the parameter value of the DGP (cf. (5.51)).

- An estimator $\hat{\theta}$ is called **unbiased**, if there is no bias for all feasible values of θ_0.

- **Interpretation**: Unbiasedness implies that for a large number of samples the average value of all estimates is very close to the true value.

- If two estimators are equal in all properties except unbiasedness, the unbiased estimator is preferable. Why?

9.1.1. Conditions for unbiasedness of the OLS estimator

Derivation: It holds, provided $\mathbf{X}$ has full rank and the multiple linear regression model is correctly specified, that

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\end{aligned}$$

and so

$$E[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}_0 = E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\right].$$

Unbiasedness of the OLS estimator holds if at least one of the following assumptions regarding the **regressors** and errors is satisfied:

- all regressors are **non-stochastic** and $E[\mathbf{u}] = \mathbf{0}$:
$$E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{u}] = \mathbf{0}.$$

- Regressors $\mathbf{X}$ are stochastic but **stochastically independent** of the error vector $\mathbf{u}$ with $E[\mathbf{u}] = \mathbf{0}$. Then, it holds that
$$E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\right] = E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right] E[\mathbf{u}] = \mathbf{0}.$$

- A weaker assumption than stochastic independence is
$$E[\mathbf{u}|\mathbf{X}] = \mathbf{0}. \tag{9.1a}$$
Hence
$$E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\middle|\,\mathbf{X}\right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{u}|\,\mathbf{X}] = \mathbf{0}.$$
Alternatively, assumption (9.1a) can be written as
$$E[u_t|\mathbf{X}] = E[u_t|\mathbf{X}_1, \ldots, \mathbf{X}_t, \ldots, \mathbf{X}_n] = 0, \quad t = 1, \ldots, n. \tag{9.1b}$$
Explanatory variables that satisfy (9.1) are called **exogenous**. Very often variables satisfying assumption (9.1) are called **strictly exogenous** (e. g. Wooldridge (2009, Chapter 10)), see also bachelor course Time Series Econometrics, chapter 2.

- Note: From (9.1) follows by applying the iterated expected value that
$$\begin{aligned}&E\left[E[u_t|\mathbf{X}_1, \ldots, \mathbf{X}_t, \ldots, \mathbf{X}_n]|x_{sj}\right] = E[u_t|x_{sj}] = 0 \Longrightarrow Cov(u_t, x_{sj}) = 0 \\ &\text{for all } s = 1, \ldots, n \text{ and all } j = 1, \ldots, k.\end{aligned} \tag{9.2}$$
Strict exogeneity thus implies that the **error** u_t **is uncorrelated with past, present or future regressors**.

- Note: The **assumption (9.1)** is meaningless without specifying a model for the errors $\mathbf{u}$, such as $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, and **gains meaning only by referring to a (parametric) model**. Thus, the condition of (strict) exogeneity implicitly always includes a (parametric) model.

  Example: For the simple (normal) linear regression model resulting from (5.32), (9.1) is satisfied, since for the pair $\beta_1, \beta_2 \in \mathbb{R}$ of the DGP, it holds that:

  $$E[\ln(Imports_t)|GDP_1, GDP_2 \dots, GDP_n] = \beta_1 + \beta_2 GDP_t.$$

- **Summary of assumptions** or **conditions for unbiasedness** of the **OLS estimator** $\hat{\boldsymbol{\beta}}$ for the parameter vector $\boldsymbol{\beta}$:

  - **(B1) Correctly specified model** The DGP is included for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in the multiple linear regression model (6.1)

    $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

    (MLR.1 in Wooldridge (2009)).

  - **(B2a) Exogeneity** resp. **strict exogeneity** (9.1): (follows from MLR.2 and MLR.4 in Wooldridge (2009)).

    $$E[\mathbf{u}|\mathbf{X}] = \mathbf{0}.$$

  - Assumption **(B2b)** will be used and introduced later.

  - **(B3) No perfect collinearity X** (or $\mathbf{X}^T\mathbf{X}$) has full rank (MLR.3 in Wooldridge (2009)).

- Unbiasedness can be "‘checked"’ with Monte Carlo simulation.

  Example: Generate 1000 samples with $n = 50$ and estimate a correctly specified simple linear regression model. The DGP is

  $$y_t = 1 + 0.9x_t + u_t, \quad u_t \sim NID(0, 4), \quad t = 1, 2, \dots, n. \qquad (9.3)$$

  See section 2.9.1 for definition of NID. The x_t are drawn randomly from the set $1, 2, \dots, 20$ with replacement. Using the **R program, see section A.7, page 347** the 1000 replications yield the mean 0.9973185 for β_1 and the mean 0.9004453 for β_2. I. e. the mean values as estimators of the expected value are very close to the true values. The histograms for $\hat{\beta}_1$ and $\hat{\beta}_2$ in figure 9.1 show that the OLS estimates scatter around the true parameters.

9.1.2. Predetermined regressors

- A weaker assumption than strict exogeneity (9.1) is

  $$E[u_t|\mathbf{X}_t] = 0 \quad \text{für } t = 1, \dots, n, \qquad (9.4)$$

  because the error u_t may only *not* depend on the regressors $\mathbf{X}_t$ of the t-th sample observation.

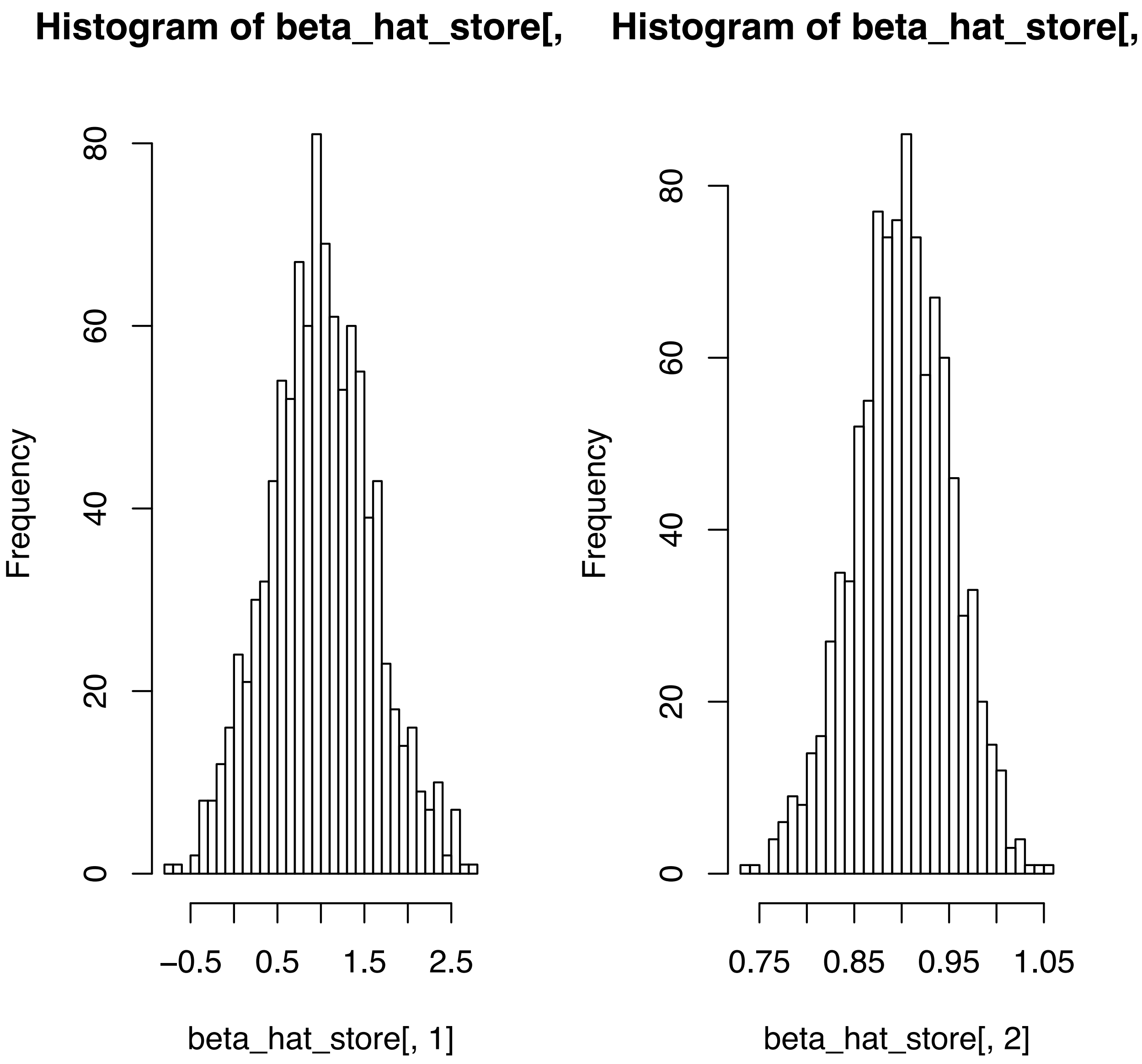


Figure 9.1.: Histograms of OLS estimates for β_1 and β_1 based on 1000 replications (**R program**, see section A.7, page 347)

- Regressors $\mathbf{X}_t$ that satisfy the condition (9.4) are called **predetermined** with respect to the error term u_t.
- In regression models for time series data, the errors u_t are also referred to as **innovations** or **shocks**.
- Wooldridge (2009, Chapter 10) also refers to assumption (9.4) as **contemporaneous exogeneity**.

- Strict exogeneity (9.1) follows from the assumption of predetermined regressors (9.4) (equivalent to Wooldridge 2009, MLR.4) *and* the assumption of a random sample (Wooldridge 2009, MLR.2), because of

$$E[u_t|\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_t, \dots, \mathbf{X}_n] = E[u_t|\mathbf{X}_t].$$

- **Models whose regressors violate the assumption of strict exogeneity but are predetermined with respect to u_t:**
  - **autoregressive models**, see section 12.3.1.
  - **dynamic linear regression models**, see section 13.4.

  Both models include lagged dependent variables as regressors.

- If the assumption of strict exogeneity (9.1) is violoated, the OLS estimator is **biased**. To obtain an unbiased estimator, it is not sufficient for regressors to be predetermined (9.4).

- Revise **relationship between conditional expected value and covariance** (2.29b), (2.29c), (2.29f).

9.2. Consistency of the OLS estimator

- See section 5.4 for the definition and meaning of the consistency of an estimator.

- **Consistency of the OLS estimator**: In addition to **(B1)**, the following **assumptions** apply:
  - **(A1)** There is a LLN for $\mathbf{X}^T\mathbf{X}/n$

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right) &= \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^T\mathbf{X}_t \\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}E\left[\mathbf{X}_t^T\mathbf{X}_t\right] = \mathbf{S}_{\mathbf{X}^T\mathbf{X}}\end{aligned}$$

    and $\mathbf{S}_{\mathbf{X}^T\mathbf{X}}$ has full rank.

    (equivalent to Davidson & MacKinnon 2004, equations (3.17) and (4.49), respectively)

  - **(A2)** There is a LLN for $\mathbf{X}^T\mathbf{u}/n$

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^T u_t = \mathbf{0}.$$

  Then, $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0$ and the OLS estimator is **consistent**.

- **Common procedure for the theoretical derivation of consistency conditions** using the example of the OLS estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ &= \boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\\ &= \boldsymbol{\beta}_0 + \underbrace{\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1}}_{:=\mathbf{A}_n} \underbrace{\frac{\mathbf{X}^T\mathbf{u}}{n}}_{:=\mathbf{a}_n}.\end{aligned}$$

  Applying the calculation rules for plim's (3.1) in section 3.4 yields under the assumption **(B1)** of a correctly specified model

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_n &= \boldsymbol{\beta}_0 + \operatorname*{plim}_{n\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1}\operatorname*{plim}_{n\to\infty}\frac{\mathbf{X}^T\mathbf{u}}{n}\\ &= \boldsymbol{\beta}_0 + \left(\underbrace{\operatorname*{plim}_{n\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)}_{\text{exists and is nonsingular because of } \textbf{(A1)}}\right)^{-1} \underbrace{\operatorname*{plim}_{n\to\infty}\frac{\mathbf{X}^T\mathbf{u}}{n}}_{=\mathbf{0},\text{ since because of } \textbf{(A2)} \text{ a LLN holds}}\\ &= \beta_0\end{aligned}$$

- **Discussion of the assumptions**

  - Simplest case for validity of the assumptions **(A1)** and **(A2)**: $\mathbf{X} = \boldsymbol{\imath}$, a constant is the only regressor, and $u_t \sim IID(0,\sigma^2)$. Then the WLLN of Khinchin (see section 5.5.1) holds, so that **(A2)** holds. **(A1)** can easily be shown.

    Example: arithmetic mean for IID random variables

    DGP: $y_t = \mu_0 + u_t, \quad u_t \sim IID(0,\sigma_0^2)$. Then, $\mathbf{S}_{\mathbf{X}^T\mathbf{X}} = n/n = 1$.

  - If there is a **random sample and (9.4) holds**, then assumption **(B2a)**) holds and **(A1)** and **(A2)** are satisfied.

    Proof: Since the sample elements are independent and identically distributed, $E\left[\mathbf{X}_t^T\mathbf{X}_t\right] = \mathbf{M}$, $t = 1, 2, \dots, n$ holds, so it automatically follows that $\mathbf{M} = \mathbf{S}_{\mathbf{X}^T\mathbf{X}}$ and hence **(A1)**. Moreover, because of (9.4) and the law of iterated expectations, $E\left[\mathbf{X}_t u_t\right] = \mathbf{0}$. Due to random sampling, Khinchin's weak law of large numbers can be applied to each vector element $z_t = \mathbf{X}_{tj}u_t$, from which **(A2)** follows. □

  - Even if there is no random sample, e. g. because there is a sample with time series data, there are assumptions that can be checked more easily than **(A2)**. These can be found in section 13.4.

- There are simple cases for which **(A1)** is violated.

  Example: $x_t = t$.

- Section 13.4 also points out that assumption **(A2)** is weaker than assumption **(B2a)**.

- The assumption **(B3)** is not mentioned because it is allowed that this can be violated for individual realisations of samples. Only in the limit it is required that there is no linear dependence between the regressors, since $\mathbf{S}_{\mathbf{X}^T\mathbf{X}}$ must have full rank in **(A1)**.

Example: Monte Carlo simulation on estimation properties of the OLS estimator with a random sample

- DGP (as in the Monte Carlo simulation in the previous section):

$$y_t = 1 + 0.9x_t + u_t, \quad u_t \sim NID(0,4), \quad t = 1, 2, \ldots, n. \tag{9.3}$$

  See section 2.9.1 for definition of NID. The x_t are drawn randomly from the set $1, 2, \ldots, 20$ with replacement.

- Sample sizes: $n = 50, 100, 500, 1000, 10000, 100000$.

- $R = 10000$ replications.

R code, see section A.8, page 349.

R output

```
         N beta_1_hat_mean beta_1_hat_sd beta_2_hat_mean beta_2_hat_sd
[1,] 5e+01       0.9939493    0.59113380       0.9008026   0.049219333
[2,] 1e+02       0.9979973    0.41867138       0.9005215   0.035010494
[3,] 5e+02       0.9979537    0.18655091       0.9001597   0.015467546
[4,] 1e+03       0.9983807    0.13101124       0.9001677   0.010893364
[5,] 1e+04       0.9996438    0.04134015       0.9000331   0.003431829
[6,] 1e+05       1.0001878    0.01323944       0.8999901   0.001098157
```

One can clearly see the unbiasedness of the OLS estimator and the decrease in the standard deviation of the OLS estimator with increasing sample size. The histograms in figures 9.2 and 9.3 for the parameter estimators and sample sizes $n = 500, 100, 500, 1000$ indicate the validity of the central limit theorem. More on this in section 11.2. Histograms for $n = 10000, 100000$ are generated with the R code but not shown here.

Example: Monte Carlo simulation on estimation properties of the OLS estimator in AR processes In section 13.5, the OLS estimator is used to estimate time series data. In the MC study to determine the bias of the OLS estimator in the AR(1) model, section 12.3.1, increase the sample size N and note your results. Also calculate the variance of the estimates of all replications.

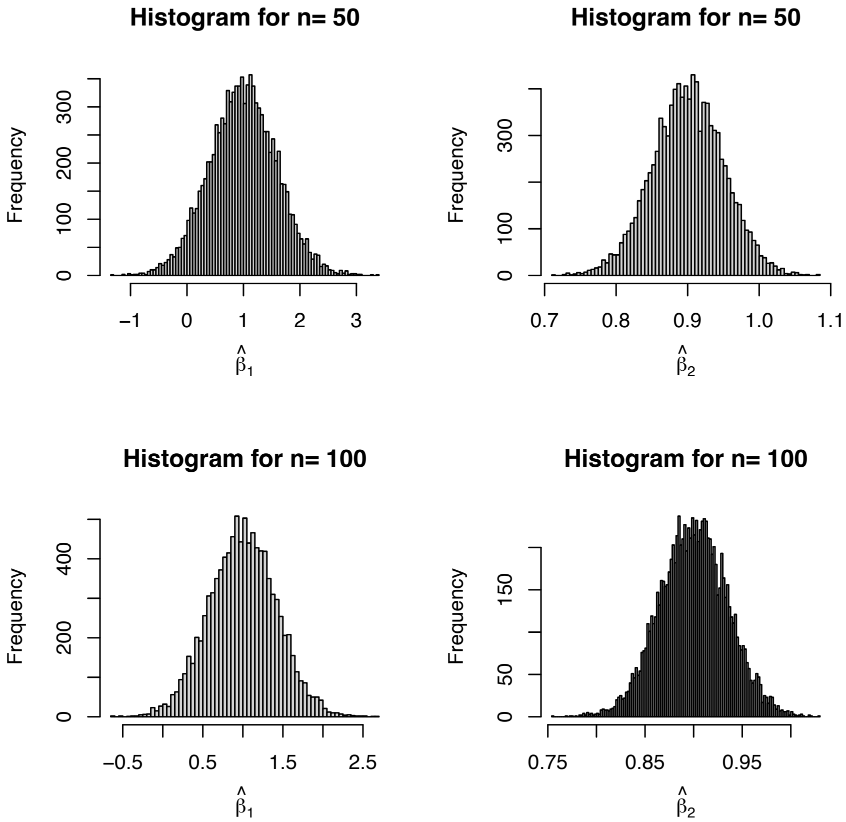


Figure 9.2.: Histograms of the OLS estimator for β for $n = 50, 100$ (R program see section A.8, page 349) DGP see equation (9.3)

9.3. The covariance matrix of the parameter estimators

- **Covariance matrix / Variance covariance matrix / Variance matrix**: see equation (5.52)
- **Conditional variance covariance matrix**: The conditional variance covariance matrix provides the variance of $\hat{\boldsymbol{\theta}}$ associated with the conditional distribution of $\hat{\boldsymbol{\theta}}$ given $\mathbf{X}$:

$$Var(\tilde{\boldsymbol{\theta}}|\mathbf{X}) = E\left[\left(\tilde{\boldsymbol{\theta}} - E\left[\tilde{\boldsymbol{\theta}}|\mathbf{X}\right]\right)\left(\tilde{\boldsymbol{\theta}} - E\left[\tilde{\boldsymbol{\theta}}|\mathbf{X}\right]\right)^T\Big|\mathbf{X}\right] \tag{9.5a}$$

$$= E\left[\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^T|\mathbf{X}\right] - E\left[\tilde{\boldsymbol{\theta}}|\mathbf{X}\right]E\left[\tilde{\boldsymbol{\theta}}|\mathbf{X}\right]^T. \tag{9.5b}$$

- **Relationship between unconditional and conditional variances** (see (2.28) for scalar case)

$$Var(\tilde{\boldsymbol{\theta}}) = E\left[Var(\tilde{\boldsymbol{\theta}}|\mathbf{X})\right] + Var\left(E(\tilde{\boldsymbol{\theta}}|\mathbf{X})\right). \tag{9.6}$$

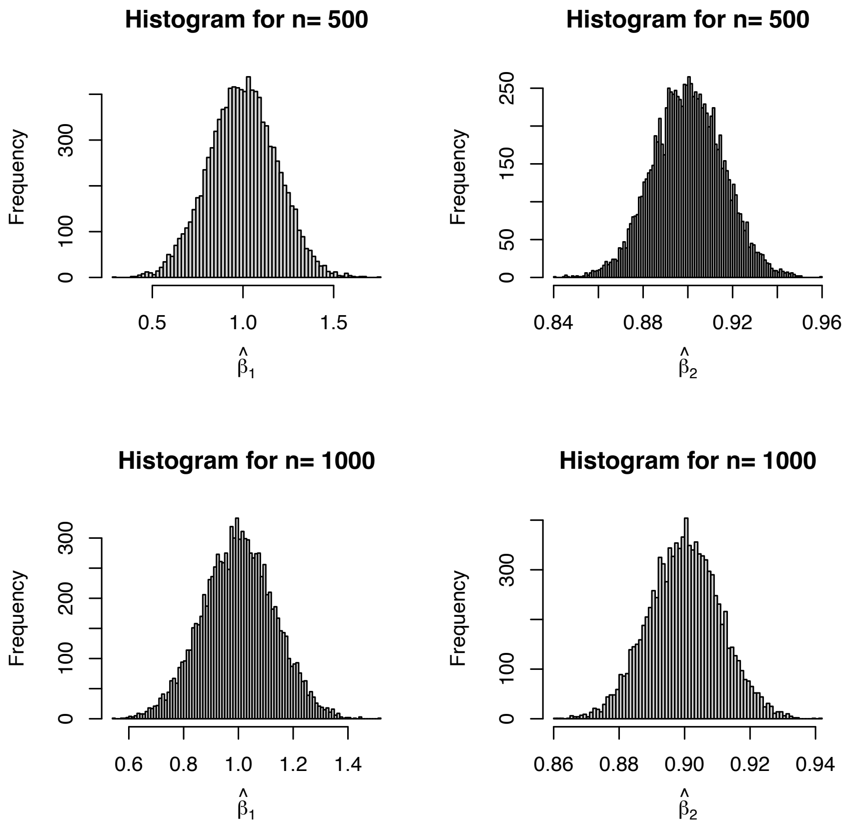


Figure 9.3.: Histograms of the OLS estimator for β for $n = 500, 1000$ (R program see section A.8, page 349) DGP see equation (9.3)

Proof: ♯ **Derivation**:

$$E\left[\left(\tilde{\boldsymbol{\theta}} - E(\tilde{\boldsymbol{\theta}})\right)\left(\tilde{\boldsymbol{\theta}} - E(\tilde{\boldsymbol{\theta}})\right)^T\right]$$

$$= E\left[\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^T\right] - E(\tilde{\boldsymbol{\theta}})E(\tilde{\boldsymbol{\theta}}^T)$$

$$= E\left[E\left(\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^T|\mathbf{X}\right)\right] - E\left[E(\tilde{\boldsymbol{\theta}}|\mathbf{X})\right]E\left[E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X})\right]$$

$$= E\left[E\left(\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^T|\mathbf{X}\right) - E(\tilde{\boldsymbol{\theta}}|\mathbf{X})E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X}) + E(\tilde{\boldsymbol{\theta}}|\mathbf{X})E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X})\right]$$

$$- E\left[E(\tilde{\boldsymbol{\theta}}|\mathbf{X})\right]E\left[E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X})\right]$$

$$= \underbrace{E\left[E\left(\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^T|\mathbf{X}\right) - E(\tilde{\boldsymbol{\theta}}|\mathbf{X})E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X})\right]}_{E\left[Var(\tilde{\boldsymbol{\theta}}|\mathbf{X})\right]} +$$

$$\underbrace{E\left[E(\tilde{\boldsymbol{\theta}}|\mathbf{X})E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X})\right] - E\left[E(\tilde{\boldsymbol{\theta}}|\mathbf{X})\right]E\left[E(\tilde{\boldsymbol{\theta}}^T|\mathbf{X})\right]}_{Var\left(E\left(\tilde{\boldsymbol{\theta}}|\mathbf{X}\right)\right)}$$

□

- **Variance covariance matrix of the unbiased OLS estimator $\hat{\boldsymbol{\beta}}$** (assumptions **(B1)**, **(B2a)**, **(B3)** satisfied):

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T|\mathbf{X}\right] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{u}\mathbf{u}^T|\mathbf{X})\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\mathbf{u}|\mathbf{X})\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned} \tag{9.7}$$

  This is the general variance covariance matrix of the OLS estimator, where heteroscedasticity and correlation between the errors given $\mathbf{X}$ is also allowed, since the conditional variance covariance matrix of the errors $Var(\mathbf{u}|\mathbf{X})$ is not further specified. This general case is discussed in chapter 14.

- **Variance covariance matrix of the OLS estimator with homoscedastic and uncorrelated errors**:

  Additionally, the following **assumption** holds:

  (B2b) Homoscedasticity and uncorrelated errors

$$Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I},$$

  where $\sigma^2 = \sigma_0^2$ holds for the error variance of the DGP.

  - Then the **variance covariance matrix of the OLS estimator** (9.7) simplifies to the well-known form

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{9.8}$$

  - The **unconditional variance covariance matrix** is obtained using (9.6):

$$Var(\hat{\boldsymbol{\beta}}) = \sigma_0^2 E\left[(\mathbf{X}^T\mathbf{X})^{-1}\right], \tag{9.9}$$

    since $Var\left(E[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right) = Var(\boldsymbol{\beta}) = \mathbf{0}$.

    ♯ For the existence of $E\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]$, see the technical supplement at the end of the section 9.4.

  - **Homoscedastic errors**: The errors u_t are called homoscedastic if their variance is constant for all sample observations, i. e. it holds that:

$$Var(u_t|\mathbf{X}_t) = \sigma_t^2 = \sigma^2. \tag{9.10}$$

    A stricter form is $Var(u_t|\mathbf{X}) = \sigma^2$.

  - **Behaviour of the variance covariance matrix for increasing sample size**: An equivalent representation to (9.8) is:

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \left(\frac{1}{n}\sigma_0^2\right)\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{9.11}$$

If condition **(A1)**

$$\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1} \xrightarrow{p} \mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}$$

is also fulfilled, the conditional variances $Var(\hat{\beta}_j|\mathbf{X})$ or covariances $Cov(\hat{\beta}_j, \hat{\beta}_i|\mathbf{X})$ generally decrease if

∗ the sample size n increases,

∗ the error variance σ_0^2 becomes smaller.

– **Asymptotic variance covariance matrix**: If condition **(A1)** holds for (9.11), then one gets

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) \xrightarrow{p} \mathbf{0}, \tag{9.12}$$

since $\sigma_0^2/n \to 0$ for $n \to \infty$. To obtain a variance covariance matrix that is fixed for $n \to \infty$, this very term must converge to a fixed value, which is precisely the case when $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is multiplied by $\sqrt{n}$, since then

$$\operatorname*{plim}_{n\to\infty} Var(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|\mathbf{X}) = \sigma_0^2\, \mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}. \tag{9.13}$$

The expression on the right-hand side is called the **asymptotic variance covariance matrix** of the OLS estimator.

In practice, $\mathbf{S}_{\mathbf{X}^T\mathbf{X}}$ is estimated by $\mathbf{X}^T\mathbf{X}/n$, which after canceling n yields the approximation

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) \approx \sigma_0^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \tag{9.14}$$

where σ_0^2 is again estimated by s^2.

– **Variance of an estimator of a single parameter** β_j: If the regression includes a constant, it holds that

$$Var(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma_0^2}{SST_j(1 - R_j^2)}, \tag{9.15}$$

where R_j^2 denotes the coefficient of determination of a regression of $\mathbf{x}_j$ on all remaining regressors.

Interpretation: The variance of $\hat{\beta}_j$ is larger,

∗ the better $\mathbf{x}_j$ is explained by the remaining regressors in $\mathbf{X}$, i. e. the larger the coefficient of determination of the regression of $\mathbf{x}_j$ on the remaining regressors in $\mathbf{X}$,

∗ the smaller the dispersion of the regressor $\mathbf{x}_j$,

∗ the larger the error variance σ_0^2.

Proof: Derivation of (9.15) (where $j = 1$ is chosen w.l.o.g. for simplicity): Thus the following partitioning

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

is possible and β_1 can be estimated with the Frisch-Waugh-Lovell theorem (cf. section 7.1) on the basis of the regression

$$\mathbf{M}_2\mathbf{y} = \mathbf{M}_2\mathbf{x}_1\beta_1 + residuals,$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T$. One obtains the OLS estimator:

$$\hat{\beta}_1 = \left(\mathbf{x}_1^T\mathbf{M}_2\mathbf{x}_1\right)^{-1}\mathbf{x}_1^T\mathbf{M}_2\mathbf{y}.$$

It can be (easily) shown that

$$Var(\hat{\beta}_1|\mathbf{X}) = \sigma_0^2(\mathbf{x}_1^T\mathbf{M}_2\mathbf{x}_1)^{-1} = \frac{\sigma_0^2}{\mathbf{x}_1^T\mathbf{M}_2\mathbf{x}_1}.$$

Note that the expression $\mathbf{x}_1^T\mathbf{M}_2\mathbf{x}_1 = ||\mathbf{M}_2\mathbf{x}_1||^2 = SSR_1$ (cf. (7.14)) corresponds to the squared length of the residual vector of the regression from $\mathbf{x}_1$ on $\mathbf{X}_2$, or the residual sum of squares of the regression from $\mathbf{x}_1$ on $\mathbf{X}_2$. Since $R_1^2 = SSE_1/SST_1$ and, if $\mathbf{X}_2$ contains a constant, $SST_1 = SSE_1 + SSR_1$, one obtains $SSE_1 = R_1^2\,SST_1$, and consequently, via $SST_1 - R_1^2\,SST_1 = SSR_1$,

$$||\mathbf{M}_2\mathbf{x}_1||^2 = SST_1(1 - R_1^2)$$

as well and thus (9.15) for $j = 1$. □

- **Multicollinearity** or **collinearity** for short:

  As just noticed, the following follows from (9.15): If the vector $\mathbf{x}_j$ is 'almost' linearly dependent on at least one other column in $\mathbf{X}$, the length of the residual vector is short and the **variance for $\hat{\beta}_j$ is large**. In this case, the variable j is said to be **multicollinear** with one or more variables. We then have **multicollinearity** or **collinearity** for short.

  The problem of multicollinearity can only be solved by increasing the sample size n. Omitting the variable j will generally lead to a misspecified model, see section 9.6. However, it is possible to consider the overall effect by looking at the **mean squared error** (5.50) or (9.36).

  In practice it is not necessary to calculate R_j^2 for each variable. Instead, the **correlation matrix** $Corr(\hat{\beta}|\boldsymbol{X})$ is considered. If the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$ is very close to 1 in absolute value, this indicates multicollinearity.

- **Variance of linear functions of parameter estimators**

  If the quantity γ to be estimated is a linear function of the estimated parameters

$$\hat{\gamma} = \mathbf{w}^T\hat{\boldsymbol{\beta}},$$

  where $\mathbf{w}$ is a suitably dimensioned column vector, then the variance of $\hat{\gamma}$ can be determined very simply by

$$\begin{aligned}
Var(\hat{\gamma}|\mathbf{X}) &= Var(\mathbf{w}^T\hat{\boldsymbol{\beta}}|\mathbf{X}) \\
&= E\left[\mathbf{w}^T(\hat{\boldsymbol{\beta}} - E\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right])(\hat{\boldsymbol{\beta}} - E\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right])^T\mathbf{w}|\mathbf{X}\right] \\
&= \mathbf{w}^T E\left[(\hat{\boldsymbol{\beta}} - E\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right])(\hat{\boldsymbol{\beta}} - E\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right])^T|\mathbf{X}\right]\mathbf{w} \\
&= \mathbf{w}^T Var(\hat{\boldsymbol{\beta}}|\mathbf{X})\mathbf{w}. \qquad (9.16)
\end{aligned}$$

And for homoscedastic and uncorrelated errors (assumption **(B2b)**):

$$Var(\hat{\gamma}|\mathbf{X}) = \sigma_0^2\mathbf{w}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}. \qquad (9.17)$$

Example of scale elasticity: $\gamma = \alpha_1 + \alpha_2$ in Cobb-Douglas production function:

$$\begin{aligned}
Y &= AL^{\alpha_1}K^{\alpha_2}e^u \\
\ln Y &= \beta_1 + \alpha_1 \ln L + \alpha_2 \ln K + u \qquad (9.18)
\end{aligned}$$

- **Variance of the prediction error with unbiased prediction** (Application of (9.17))

  If the assumptions **(B1)**, **(B2a)**, **(B3)** are fulfilled and thus the model is correctly specified, the prediction $\hat{y}_s = \mathbf{X}_s\hat{\boldsymbol{\beta}}$ for $(y_s, \mathbf{X}_s)$ from the population is unbiased, since

  $$E[\hat{y}_s|\mathbf{X}, \mathbf{X}_s] = \mathbf{X}_s\boldsymbol{\beta}_0. \qquad (9.19)$$

  This results in the prediction error

  $$y_s - \mathbf{X}_s\hat{\boldsymbol{\beta}} = \mathbf{X}_s\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right) + u_s,$$

  whose expected value is zero. The variance of the prediction error is therefore

  $$\begin{aligned}
  Var(y_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}|\mathbf{X}_s, \mathbf{X}) &= E\left[\left\{\mathbf{X}_s\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right) + u_s\right\}\left\{\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right)^T\mathbf{X}_s^T + u_s\right\}\middle|\mathbf{X}_s, \mathbf{X}\right] \\
  &= \mathbf{X}_s Var(\hat{\boldsymbol{\beta}}|\mathbf{X})\,\mathbf{X}_s^T + E[u_s^2|\mathbf{X}_s] - 2\mathbf{X}_s \quad \underbrace{Cov(\hat{\boldsymbol{\beta}}, u_s|\mathbf{X}_s, \mathbf{X})}_{=0,\text{ with uncorrelatedness}} \\
  &= \sigma_0^2\mathbf{X}_s(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_s^T + \sigma_0^2 \quad \text{(given assumption \textbf{(B2b)})}
  \end{aligned}$$

  ⟶ Variance of prediction error = Variance of the estimator of the dependent variable + Variance of u_s.

- **Summary of the assumptions of the multiple linear regression model with strictly exogenous regressors**

  – **(B1)** Correctly specified model: The DGP is included in the multiple linear regression model for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

- **(B2)**: $\mathbf{u}|\mathbf{X} \sim (\mathbf{0}, \sigma^2\mathbf{I}) \iff \begin{cases} \text{(B2a): } E[\mathbf{u}|\mathbf{X}] = \mathbf{0} \text{ (}\mathbf{X}\text{ is (strictly) exogenous) \&} \\ \text{(B2b): } Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I} \text{ (The errors conditional on }\mathbf{X} \\ \text{are homoscedastic and uncorrelated).} \end{cases}$
- **(B3)** $\mathbf{X}$ has full column rank.

9.4. The efficiency of unbiased OLS estimators

- Cf. for the definition of **efficiency** of an estimator section 5.4 and (5.56). In the following, the class of linear estimators is considered.
- **Linear estimator**: An estimator $\tilde{\boldsymbol{\beta}}$ for the parameter vector $\boldsymbol{\beta}$ in a multiple linear regression model is called **linear** if $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$, where the $(k \times n)$ matrix $\mathbf{A} := \mathbf{A}(\mathbf{X})$ **may only depend on the regressors** $\mathbf{X}$, but not on $\mathbf{y}$, i. e. $E[\mathbf{A}|\mathbf{X}] = \mathbf{A}$ holds.
- The OLS estimator is a linear estimator, since $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
- A linear estimator $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$ is **unbiased** if the assumptions **(B1)**, **(B2a)** apply, and

$$\mathbf{A}\mathbf{X} = \mathbf{I}, \quad \text{since } E[\tilde{\boldsymbol{\beta}}|\mathbf{X}] = \mathbf{A}\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{A}E[\mathbf{u}|\mathbf{X}]. \tag{9.20}$$

- **Comparison of the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ with arbitrary linear and unbiased estimators $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$** with $\mathbf{A}\mathbf{X} = \mathbf{I}$
  - **Gauss-Markov theorem**: Under the assumptions **(B1)**, **(B2)**, **(B3)** the OLS estimator $\hat{\boldsymbol{\beta}}$ is the efficient estimator (**best linear unbiased estimator** (**BLUE**)) among all linear and unbiased estimators $\tilde{\boldsymbol{\beta}}$. This means that the matrix of the difference of the variance-covariance matrices $Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}})$ is **positive semidefinite**.

    Examples of inefficient linear unbiased estimators:

    * Estimator of the expected value (mean) $(y_1 + y_n)/2$.
    * Any OLS estimator applied to a regression model with redundant independent variables, see section 9.6.
    * Instrument variable estimator, see e. g. the bachelor course **Advanced Issues in Econometrics** or the master course **Advanced Econometrics**.

    Proof sketch: Since $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \underbrace{\left(\mathbf{A} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)}_{\mathbf{C}}\mathbf{y} = \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{u} = \mathbf{C}\mathbf{u}$, it holds that

$$Var(\tilde{\boldsymbol{\beta}}) = Var(\hat{\boldsymbol{\beta}} + \mathbf{C}\mathbf{u}) = Var(\hat{\boldsymbol{\beta}}) + Var(\mathbf{C}\mathbf{u}),$$

since $E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\mathbf{Cu})^T\right] = \mathbf{0}$ can be shown when taking (9.20) and **(B2b)** into account. Since every variance-covariance matrix is positive semidefinite, this also applies to $Var(\mathbf{Cu})$. □

- Originally, the Gauss-Markov theorem was proved for non-stochastic regressors $\mathbf{X}$.

- ♯ **Technical supplement**: If $\mathbf{X}$ is stochastic, it is possible in principle that, for example, assumption **(B3)** or (9.20) is violated for a specific realisation of $\mathbf{X}$, i. e. $\mathbf{X}$ does not have full rankand therefore $(\mathbf{X}^T\mathbf{X})$ is not invertible. If the regressors are continuously distributed, then the probability of this is 0.
  - If $P(C) = 1$ applies to an event C, then $P(C^c) = 0$ applies to the complement C^c. It is then said that the event C occurs **almost surely (a.s.)**.
  - Example of an almost sure event: Let $X \in \mathbb{R}$ be a continuous random variable. The event $C = \{X \in (-\infty, a) \cup (a, \infty)\}$ has the complementary event $C^c = \{X = a\}$. Since $P(X = a) = P(C^c) = 0$, it holds for C that $P(C) = 1$.
  - If $\mathbf{X}$ only contains discrete regressors, for example a constant and a dummy variable, then there is a positive probability that a sample will be drawn in which the dummy variable takes the value 1 for all observations and therefore $\mathbf{X}$ has reduced rank and $\mathbf{X}^T\mathbf{X}$ is not invertible. The assumption **(B3)** is therefore not *almost surely* fulfilled for this example. In this case, $E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]$ does not exist either, as there is a positive probability that the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible.
  - The existence of the unconditional expected value and the unconditional variance of the OLS estimator therefore requires that the assumptions **(B1)** to **(B3)** hold almost surely.
  - For **practice** it is generally sufficient to know the distribution properties **given the regressors**. Then you don't need to worry about this problem.
  - However, if you want to perform Monte Carlo simulations in which $\mathbf{X}$ is also redrawn at each realisation, but $\mathbf{X}$ has reduced rank with positive probability, the case of a singular $\mathbf{X}^T\mathbf{X}$ matrix will repeatedly occur and the OLS estimator cannot be calculated.

9.5. Estimating the error variance

- In this section, the assumptions **(B1)** to **(B3)** are assumed.
- In the correctly specified OLS model, the following applies,

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{M_X y}\\ &= \mathbf{M_X X}\boldsymbol{\beta}_0 + \mathbf{M_X u}\\ &= \mathbf{M_X u},\end{aligned} \tag{9.21}$$

  since $\mathbf{M_X X} = \mathbf{0}$.

The residual $\hat{u}_t$ corresponds to a **linear combination** of the error vector $\mathbf{u}$.

- **Variance** of the residual vector:

$$\begin{aligned} Var(\hat{\mathbf{u}}|\mathbf{X}) &= Var(\mathbf{M_X u}|\mathbf{X}) \\ &= E\left[\mathbf{M_X u u}^T\mathbf{M}_\mathbf{X}^T|\mathbf{X}\right] \\ &= \mathbf{M_X}(\sigma_0^2\mathbf{I})\mathbf{M}_\mathbf{X}^T \\ &= \sigma_0^2\mathbf{M_X}. \end{aligned} \tag{9.22}$$

- **Properties of the residuals** $\hat{u}_t$: These result from the variance-covariance matrix of the residuals $Var(\hat{\mathbf{u}}|\mathbf{X})$.

  The residuals are generally

  – correlated and

  – heteroscedastic with $Var(\hat{u}_t|\mathbf{X}) \leq Var(u_t) = \sigma_0^2$.

  Proof: As in section 7.2, $\mathbf{e}_t$ denotes a **unit basis vector**. Then

$$\hat{u}_t = \mathbf{e}_t^T\hat{\mathbf{u}}$$

  and

$$Var(\hat{u}_t|\mathbf{X}) = Var(\mathbf{e}_t^T\hat{\mathbf{u}}|\mathbf{X}) = \mathbf{e}_t^T Var(\hat{\mathbf{u}}|\mathbf{X})\mathbf{e}_t = \sigma_0^2\mathbf{e}_t^T\mathbf{M_X}\mathbf{e}_t = \sigma_0^2||\mathbf{M_X}\mathbf{e}_t||^2.$$

  Due to the orthogonal decomposition

$$||\mathbf{e}_t||^2 = \underbrace{||\mathbf{P_X}\mathbf{e}_t||^2}_{h_t} + \underbrace{||\mathbf{M_X}\mathbf{e}_t||^2}_{1-h_t},$$

  such that $||\mathbf{M_X}\mathbf{e}_t||^2 \leq ||\mathbf{e}_t||^2 = 1$. □

- **Maximum likelihood estimator for the error variance**:

  – The estimator

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{t=1}^n \hat{u}_t^2 \tag{9.23}$$

  is called **maximum likelihood estimator** for the error variance σ^2, as it results from the maximum likelihood approach, see master course **Advanced Econometrics**.

  – **Property**: $\tilde{\sigma}^2$ is **unbiased**.

  Proof:

$$\begin{aligned} E[\tilde{\sigma}^2|\mathbf{X}] &= \frac{1}{n}\sum_{t=1}^n E[\hat{u}_t^2|\mathbf{X}] \\ &= \frac{1}{n}\sum_{t=1}^n Var(\hat{u}_t|\mathbf{X}) \\ &= \sigma_0^2\frac{1}{n}\sum_{t=1}^n ||\mathbf{M_X}\mathbf{e}_t||^2. \end{aligned}$$

From $||\mathbf{P_X e}_t||^2 = h_t$ finally follows

$$E[\hat{\sigma}^2|\mathbf{X}] = \sigma_0^2 \frac{1}{n} \sum_{t=1}^{n} \underbrace{(1 - h_t)}_{\leq 1} \leq \sigma_0^2.$$

With the help of the trace operator it can be shown that

$$\sum_{t=1}^{n} (1 - h_t) = n - k.$$

From this follows

$$E[\hat{\sigma}^2|\mathbf{X}] = \frac{n-k}{n} \sigma_0^2. \tag{9.24}$$

□

- **Unbiased estimator for the error variance**: Considering (9.24) in (9.23) provides the unbiased estimator

  $$s^2 = \frac{1}{n-k} \sum_{t=1}^{n} \hat{u}_t^2. \tag{9.25}$$

  (Note the notation: in many other econometrics books, e. g. Wooldridge (2009), this estimator is denoted by $\hat{\sigma}^2$.)

- The root of s^2 is denoted as **standard error of regression**.

- An **unbiased estimator of the covariance matrix of the OLS estimator** is then

  $$\widehat{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = s^2 (\mathbf{X}^T \mathbf{X})^{-1}. \tag{9.26}$$

Example: trade flows For the OLS estimations of model 3 (6.17), the variance-covariance matrix and correlation matrix of the parameter estimators are given in the following R-output.

R code (Extract from R program in section A.4)

```
summary(mod_3a_kq)$cov

# Estimate the correlation matrix of the OLS estimators for model 3a
cov2cor(summary(mod_3a_kq)$cov)

# Estimate the covariance matrix of sample observations for model 3a
cor(data.frame(log_wdi_gdpusdcr_o = log(wdi_gdpusdcr_o),
               log_cepii_dist=log(cepii_dist),ebrd_tfes_o))
```

Listing 9.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

R output

```
                          (Intercept) log(wdi_gdpusdcr_o) log(cepii_dist)  ebrd_tfes_o
(Intercept)             6.2069332        -0.124749916    -0.315428513 -0.183737444
log(wdi_gdpusdcr_o)   -0.1247499         0.004936052     0.002017428 -0.003353511
log(cepii_dist)       -0.3154285         0.002017428     0.030582699  0.009851900
ebrd_tfes_o           -0.1837374        -0.003353511     0.009851900  0.048163990
```

```
                          (Intercept) log(wdi_gdpusdcr_o) log(cepii_dist) ebrd_tfes_o
(Intercept)                1.0000000          -0.7127084      -0.7239766  -0.3360454
log(wdi_gdpusdcr_o)       -0.7127084           1.0000000       0.1641989  -0.2174947
log(cepii_dist)           -0.7239766           0.1641989       1.0000000   0.2566970
ebrd_tfes_o               -0.3360454          -0.2174947       0.2566970   1.0000000

                  log_wdi_gdpusdcr_o log_cepii_dist ebrd_tfes_o
log_wdi_gdpusdcr_o         1.0000000      -0.233241   0.2723423
log_cepii_dist            -0.2332410       1.000000  -0.3037030
ebrd_tfes_o                0.2723423      -0.303703   1.0000000
```

Note that the correlations between the variables are not greater than 0.26 in absolute value, i.e. relatively low, and there are no signs of multicollinearity.

9.6. Overspecified or misspecified linear regression models

For the definition of the **information set** see section 5.3.

Overspecification

- A model $\mathbb{M}$ is **overspecified** if it contains variables that belong to the information set Ω_t but are not contained in the DGP. (Note: Overspecified models are not misspecified).

  Example: Let the DGP be contained in

  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}, \quad \mathbf{u}|\mathbf{X} \sim (\mathbf{0}, \sigma_0^2\mathbf{I}), \tag{9.27}$$

  (**(B1)**,**(B2)** hold), but

  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}, \quad \mathbf{u}|\mathbf{X}, \mathbf{Z} \sim (\mathbf{0}, \sigma^2\mathbf{I}) \tag{9.28}$$

  is estimated. The ‘unrestricted’ model (9.28) also contains the DGP ($DGP \in \mathbb{M}$), since the parameters $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma} = \mathbf{0}$ and $\sigma^2 = \sigma_0^2$ are possible.

- **Properties of the OLS estimator** $\tilde{\boldsymbol{\beta}}$ of the overspecified model (9.28):

  (i) **unbiased**, since according to the Frisch-Waugh-Lovell theorem, see page 7.1.3, the OLS estimator $\tilde{\boldsymbol{\beta}}$ of the regression

  $$\mathbf{M_Z y} = \mathbf{M_Z X}\boldsymbol{\beta} + residuals$$

  with $\mathbf{M_Z} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ is identical to the OLS estimator for $\boldsymbol{\beta}$ in the overspecified model (9.28). Therefore, the following applies,

  $$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{M_Z X})^{-1}\mathbf{X}^T\mathbf{M_Z u} \quad \Rightarrow \quad E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0.$$

  (ii) in general compared to the OLS estimator $\hat{\boldsymbol{\beta}}$ of the ‘smallest’ correctly specified model (9.27) **not efficient**. This is due to the **Gauss-Markov theorem**, cf. section 9.4. From this follows, among other things, cf. (5.56),

  $$Var(\tilde{\beta}_j|\mathbf{X}, \mathbf{Z}) \geq Var(\hat{\beta}_j|\mathbf{X}), \quad j = 1, \ldots, k.$$

This inequality results, cf. (9.15), also directly from

$$\frac{\sigma_0^2}{SST_j(1-R^2_{j,\mathbf{X},\mathbf{Z}})} \geq \frac{\sigma_0^2}{SST_j(1-R^2_{j,\mathbf{X}})}, \quad j=1,\ldots,k.$$

The probability of **multicollinearity** is also increased by additional, unneeded variables.

- These results apply regardless of the sample size. It can therefore be shown that the estimator of an overspecified model is **asymptotically inefficient**.

Misspecification (also underspecification)

- A model $\mathbb{M}$ is **underspecified** or **misspecified** if the DGP is not included in the model.

  Example: DGP is included in

  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Z}\boldsymbol{\gamma}_0 + \mathbf{u}, \quad \mathbf{u}|\mathbf{X},\mathbf{Z} \sim (\mathbf{0}, \sigma_0^2\mathbf{I}), \quad \boldsymbol{\gamma}_0 \neq \mathbf{0}, \tag{9.29}$$

  with $(n \times k_1)$ regressor matrix $\mathbf{X}$ and $(n \times k_2)$ regressor matrix $\mathbf{Z}$, but the regression model

  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} \tag{9.30}$$

  is estimated. This results in the following for the OLS estimator for (9.30),

  $$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\boldsymbol{\gamma}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}\\ &= \boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\boldsymbol{\gamma}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}.\end{aligned}$$

- **Note**: The first part of the second term on the right-hand side (cf. notation (6.2)) is:

  $$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} = \begin{pmatrix}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}_1 & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}_2 & \cdots & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}_{k_2}\end{pmatrix}.$$

  The l-th column of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$ therefore just contains the OLS estimator $\hat{\boldsymbol{\delta}}_l = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}_l$ of the (auxiliary) regression

  $$\mathbf{z}_l = \mathbf{X}\boldsymbol{\delta}_l + error. \tag{9.31}$$

  The OLS estimator $\hat{\boldsymbol{\beta}}$ can thus be written as

  $$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}\boldsymbol{\gamma}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}. \tag{9.32}$$

  Depending on the choice of condition in the (conditional) expected value, different biases are obtained:

  - Thus, the OLS estimator **for given sample values of all regressors relevant in the DGP** is **biased** if

    $$E[\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Z}] = \boldsymbol{\beta}_0 + \begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}\boldsymbol{\gamma}_0 \neq \boldsymbol{\beta}_0, \tag{9.33}$$

    i. e. the regressors $\mathbf{X}$ and $\mathbf{Z}$ are not orthogonal in a given sample.

- Thus, the OLS estimator **for given sample values of all regressors in X** is **biased** if

$$E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = E\left[E[\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Z}]|\mathbf{X}\right] = \boldsymbol{\beta}_0 + E\left[\begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}|\mathbf{X}\right]\boldsymbol{\gamma}_0 \neq \boldsymbol{\beta}_0, \tag{9.34}$$

  i. e. if $E\left[\begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}|\mathbf{X}\right] \neq 0$.

- Thus, the OLS estimator **is biased** if

$$E[\hat{\boldsymbol{\beta}}] = E\left[E[\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Z}]\right] = \boldsymbol{\beta}_0 + E\left[\begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}\right]\boldsymbol{\gamma}_0 \neq \boldsymbol{\beta}_0, \tag{9.35}$$

  i. e. if the unconditional expected value $E\left[\begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}\right] \neq 0$. In other words, at least one z_{ti} and x_{tj} are **correlated** with each other.

- **Important**: If the expected value $E\left[\begin{pmatrix}\hat{\boldsymbol{\delta}}_1 & \hat{\boldsymbol{\delta}}_2 & \cdots & \hat{\boldsymbol{\delta}}_{k_2}\end{pmatrix}\right] \neq 0$ independent of the sample size n, i. e. also for $n \to \infty$, then the OLS estimator for β_0 is **inconsistent**!

Conclusion:

	Misspecified model	Overspecified model
	OLS estimator is	
finite sample	generally biased	inefficient
asymptotically	generally inconsistent	asymptotically inefficient

Obviously, the choice of a correct but not overspecified model is very important. This is the task of model selection methods, which are described in the next chapter.

Mean squared error:

- The matrix of the **mean squared error** (**MSE**), cf. (5.50), is given all regressors $\mathbf{X}, \mathbf{Z}$:

$$MSE(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Z}) = E\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \middle| \mathbf{X},\mathbf{Z}\right]. \tag{9.36}$$

  As with the bias, a distinction can be made here with regard to the conditions (but this is not done here).

- Note: only for unbiased estimators is the matrix of the mean squared error equal to the variance-covariance matrix.

- It can be shown (possibly as an exercise) that

$$MSE(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Z}) = \underbrace{\sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}}_{\text{variance}} + \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^T\mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}}_{\text{squared bias}}. \tag{9.37}$$

  An unambiguous statement on the comparison of this MSE matrix with that of the unbiased OLS estimator in (9.29), i. e. $MSE(\tilde{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Z}) = \sigma_0^2(\mathbf{X}^T\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1}$, is not possible, but depends on the size of the bias.

- The MSE (9.36) is therefore suitable as a criterion for evaluating different models, since

- in the case of misspecified models, the squared bias is taken into account and
- in the case of overspecified models, the too large OLS estimation variance

is weighed against each other.

- To determine the **accuracy** of the OLS estimator of the misspecified model, it no longer makes sense to use the covariance matrix due to the bias of the estimator.

To read: Davidson & MacKinnon (2004), chapter 3.

10. Model specification

Aims of model selection:

- **Essential aspects of econometric modelling**:

  1. Search for the model that contains the DGP that generated the sample observations.
  2. Avoid too large models.
  3. Search for an efficient estimation procedure.

  In other words, model selection procedures for **model specification** are used to **suitably bound** the **probabilities**

  – of choosing a **misspecified model** and

  – of choosing an **overspecified model**

  or, if possible, to make them asymptotically approach zero.

  The **asymptotic requirements** imply:

  1. **Consistent estimation** of quantities of interest.
  2. **Efficient estimation** of quantities of interest.
  3. Existence of **(asymptotic) test distributions** for performing **hypothesis tests**.

- **In practice** there will rarely be a correctly specified model. Instead, search for the "‘best"’ model for the intended investigation, e. g. to determine the GDP elasticity of exports or $E[y_t|\Omega_t]$. What does "‘best"’ model mean? The quality of the model depends on how an element contained in a model in reduced form can approximate the DGP. For models in structural form, the implied model in reduced form must be considered. See section 5.2 for structural and reduced form. The evaluation of the approximation depends on the loss function, for example (5.45), which was chosen for the problem.

- However, the econometric theory for approximating models is too complicated for this course. Therefore, in **this course** we assume that there is a correctly specified model.

Notes:

The **C** assumptions have not yet been discussed. They include time series data and are discussed in detail in section 13.4. For the sake of completeness, however, they are also mentioned below.

To 1. **Consistency** requires, among other things, that the model is correctly specified, i.e. the

- assumptions **(B1)** and **(B2a)**
- or more generally for time series, the assumptions **(C1)** and **(C2a)**

must be fulfilled. Then the following applies,

$$E[y_t|\Omega_t] = \mathbf{X}_t\boldsymbol{\beta}_0. \tag{10.1}$$

To 2. **Efficiency** requires, among other things, that

- an efficient estimation procedure is selected and
- that superfluous variables are avoided in the model.

For example, the OLS estimator is only

- efficient if, among others, assumption **(B2b)** holds, or
- asymptotically efficient if, among others, the assumption **(C2b)**

holds, i. e. the errors are homoscedastic.

To 3. Deriving **test distributions** requires additional assumptions, e. g. **(B4)** for exact tests or **(C4a)** or **(C4b)** for asymptotic tests. (Cf. chapter 11.)

- The use of **model selection criteria** is intended to ensure that

  a) no superfluous variables are included in the model and thus the efficiency of the estimator is reduced,

  b) all relevant variables are contained in the model, i.e. (10.1) holds and thus a prerequisite for consistency is fulfilled.

  In smaller samples, it may not be possible to include all relevant variables in the model without the estimation variance becoming too large. Model selection criteria allow a "‘trade-off"’ between a) and b).

- **Nested models**: $\mathbb{M}_1$ and $\mathbb{M}_2$ are nested if either $\mathbb{M}_1 \subset \mathbb{M}_2$ or $\mathbb{M}_2 \subset \mathbb{M}_1$ applies.

10.1. Model selection criteria

- **Basic idea of model selection criteria**:

  $$Selectioncriterion = \text{fit measure} + \text{number of parameters} \cdot \text{penalty function(n)} \qquad (10.2)$$

  - **First term: Fit measure**: Measures how well the estimated model fits the data. Fit measures are selected that generally lead to an improvement in fit with an additional parameter, but never to a deterioration in fit. Typically, either the maximum likelihood estimator $\hat{\sigma}^2 = \hat{\mathbf{u}}^T\hat{\mathbf{u}}/n$ of the error variance (9.23) or minus twice the log-likelihood function is selected here, with the latter differing from $\hat{\sigma}^2$ by only a constant for a given sample size, see (10.3).

    It can be shown that the fit of a model in which relevant regressors are missing is asymptotically larger than that of a correctly specified model. This suggests selecting the model with the smallest fit. However, if the model is overspecified and $\hat{\sigma}^2$ is used, the true error variance is typically underestimated. There is therefore a risk of selecting an overspecified model. To reduce the probability of this, a penalty term is used to make it more difficult to add additional irrelevant regressors.

  - **Second term: Penalty term**: Product of the number of estimated parameters k in $\boldsymbol{\beta}$ and penalty function:

    - The **penalty term** penalises the number of parameters in order to avoid including superfluous variables in the model and thus making the estimation procedure inefficient.

    - The penalty term increases with increasing k and the penalty function must be chosen so that it decreases with increasing n. In the latter case, this means that additional parameters in larger samples are penalised relatively less, but this penalty must not approach zero too quickly!

  - This implies a **trade-off**: regressors are included in the model if the penalty is less than the improvement in fit.

    The choice of the penalty function (and thus the criterion) determines how this trade-off is quantified. Three different criteria are commonly used: AIC, HQ and SC/BIC, see below.

  - **Rule**: Among all candidates considered, the specification for which the criterion yields the *smallest* value is selected.

  - It is advisable to check AIC, HQ and SC/BIC. In favourable cases, all criteria provide the same result. Note that SC penalises additional parameters for sample sizes $n > 8$ more than HQ, and HQ again more than AIC.

  - It is possible to use selection criteria to select from non-nested models as long as the dependent variable is identical, see empirical example in section 10.3.

- For information: **Log-likelihood function**, more precisely concentrated log-likelihood function

$$l(\hat{\boldsymbol{\beta}}, \hat{\sigma}|\mathbf{y}, \mathbf{X}) = -\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2 \tag{10.3}$$

cf. for explanation and derivation Davidson & MacKinnon (2004, Equation (10.12)) or course **Advanced Econometrics**, section 5.5

- **Alternative definitions of model selection criteria**:

$$\begin{array}{cccc} \text{Criterion} & \text{Fit measure} & \text{Number of par.} & \text{Penalty function}(n) \end{array}$$

$$AIC = \ln\hat{\sigma}^2 + k \cdot \frac{2}{n}, \tag{10.4}$$

$$HQ = \ln\hat{\sigma}^2 + k \cdot \frac{2\ln(\ln(n))}{n}, \tag{10.5}$$

$$SC = \ln\hat{\sigma}^2 + k \cdot \frac{\ln(n)}{n}, \tag{10.6}$$

$$AIC = -\frac{2}{n}\left(\underbrace{-\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2}_{=\text{Log-likelihood function}}\right) + k \cdot \frac{2}{n} \tag{10.7}$$

$$HQ = -\frac{2}{n}\left(-\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2\right) + k \cdot 2\frac{\ln(\ln(n))}{n} \tag{10.8}$$

$$SC = -\frac{2}{n}\left(-\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2\right) + k \cdot \frac{\ln(n)}{n} \tag{10.9}$$

In `R` the command `AIC()` calculates model selection criteria that differ from the above calculations in two respects:

- It is not divided by n.
- In addition to the estimated parameters in $\boldsymbol{\beta}$, the variance is also added as an estimated parameter.

For a comparison of different models for **given** n this is irrelevant.

$$\begin{array}{ccc} \text{Fit measure} & \text{Number of par.} & \text{Penalty function}(n) \end{array}$$

$$AIC = -2\left(-\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2\right) + (k+1) \cdot 2 \tag{10.10}$$

$$HQ = -2\left(-\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2\right) + (k+1) \cdot 2\ln(\ln(n)) \tag{10.11}$$

$$SC = -2\left(-\frac{n}{2}\left(1 + \ln(2\pi)\right) - \frac{n}{2}\ln\hat{\sigma}^2\right) + (k+1) \cdot \ln(n) \tag{10.12}$$

There are also definitions in which the model selection criteria are maximised, e. g. in Davidson & MacKinnon (2004, Section 15.4). So always note the exact definitions in the software used!

Formula	Software - Command
Akaike Information Criterion (AIC)	
(10.4)	R: `extractAIC()`
(10.7)	EViews, R: own program `SelectCritEViews()`, see section B.2
(10.10)	R: `AIC()`
Hannan-Quinn (HQ)	
(10.5)	R: `extractAIC(,k = log(log(n))))`
(10.8)	EViews, R: own program `SelectCritEViews()`, see section B.2
(10.11)	R: `AIC(,k = log(log(n))))`
Bayesian Information Criterion (BIC)/Schwarz Criterion (SC)	
(10.6)	R: `extractAIC(,k = log(n))`
(10.9)	EViews, R: own program `SelectCritEViews()`, see section B.2
(10.12)	R: `AIC(,k = log(n)))`

- Alternative to the use of model selection criteria: Sequential testing. This requires t-tests or F-tests, which are discussed in chapter 11.
- ♯ The comparison of two models using a model selection criterion can also be interpreted as a test, whereby the significance level is determined by the penalty term.

10.2. Tests for non-nested models

See section 9.3.1 in course material for bachelor course **Introduction to Econometrics** or Wooldridge (2009, Chapter 9) or Davidson & MacKinnon (2004, Section 15.3).

The following is dealt with there:

- **Encompassing test**, R command: `encomptest(model_1,model_2)`
  (requires R package `lmtest`)
- ***J*-Test**, R command: `jtest(model_1,model_2)`
  (requires R package `lmtest`)

10.3. Empirical analysis of trade flows: Part 2

Continuation of **Empirical analysis of trade flows: Part 1** in section 6.3.

To step II.3: **Specifying, estimating and selecting an econometric model**

- **Specifying and estimating** different models:
  Five different models are now specified and estimated:

R commands:

Model 1

```
mod_1_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o)
```

Model 2

```
mod_2_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)
```

Model 3a

```
mod_3a_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)
                    +  ebrd_tfes_o
```

Model 3b

```
mod_3b_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)
                    + log(cepii_area_o)
```

Model 4

```
mod_4_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)
                    +  ebrd_tfes_o + log(cepii_area_o)
```

Calculate the models via

R code (Extract from R program in section A.4)

```
# Apply the function "SelectCritEviews" to four different models

mod_1_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o))
summary(mod_1_kq)
deviance(mod_1_kq)                  # Calculates SSR
SelectCritEviews(mod_1_kq)      # Calculates AIC, HQ, SC

mod_2_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))
summary(mod_2_kq)
deviance(mod_2_kq)                  # Calculates SSR
SelectCritEviews(mod_2_kq)      # Calculates AIC, HQ, SC

mod_3a_kq     <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                      ebrd_tfes_o)
summary(mod_3a_kq)
deviance(mod_3a_kq)           # Calculates SSR
SelectCritEviews(mod_3a_kq)     # Calculates AIC, HQ, SC

mod_3b_kq     <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                                    log(cepii_area_o))
summary(mod_3b_kq)
deviance(mod_3b_kq)                 # Calculates SSR
SelectCritEviews(mod_3b_kq)     # Calculates AIC, HQ, SC

mod_4_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                      ebrd_tfes_o + log(cepii_area_o))
summary(mod_4_kq)
deviance(mod_4_kq)                  # Calculates SSR
SelectCritEviews(mod_4_kq)      # Calculates AIC, HQ, SC
```

Listing 10.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides output from which the following table can be generated:

Dependent variable: ln(*Imports to Germany*)					
Independent variable/model	(1)	(2)	(3a)	(3b)	(4)
Constant	-5.77	4.676	2.741	3.409	2.427
	(2.184)	(2.178)	(2.175)	(2.098)	(2.132)
ln(*GDP*)	1.077	0.975	0.940	1.080	1.025
	(0.087)	(0.063)	(0.0613)	(0.071)	(0.076)
ln(*distance*)	—	-1.074	-0.970	-915	-0.888
		(0.156)	(0.152)	(0.159)	(0.156)
Openness	—	—	0.507	—	0.353
			(0.191)		(0.206)
ln(*area*)	—	—	—	-0.213	-0.151
				(0.089)	(0.085)
Sample size	49	49	49	49	49
R^2	0.765	0.883	0.900	0.900	0.906
Standard error of the regression	1.304	0.928	0.873	0.871	0.853
Residual sum of squares	80.027	39.644	34.302	34.148	32.017
AIC	3.4100	2.7484	2.6445	2.6400	2.6164
HQ	3.4393	2.7924	2.7031	2.6986	2.6896
SC	3.4872	2.8642	2.7989	2.7945	2.8094

- **Selection of model**: The table shows that model 4 must be selected if the Akaike criterion (AIC) or the Hannan-Quinn (HQ) criterion is chosen, but model 3b if the Schwarz (SC) criterion is chosen.

Continuation of **Empirical analysis of trade flows: Part 3** in section 11.7.

11. (Asymptotic) distribution of the OLS estimator and testing in the multiple linear regression model

11.1. Exact distribution of the OLS estimator

- With previous assumptions, the following applies to the OLS estimator,

$$\hat{\boldsymbol{\beta}}_n \overset{(B3)}{=} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \overset{(B1)}{=} \boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}.$$

- Without a distribution assumption for the error vector $\mathbf{u}$, obviously nothing more can be said about the distribution of $\hat{\boldsymbol{\beta}}_n$, even if the $\mathbf{X}$ are given.

  We make the assumption (cf. for the notation Davidson (2000, Section 2.4.1))

  (B4) Multivariate normally distributed errors given X

$$\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

  where the error variance of the DGP is $\sigma^2 = \sigma_0^2$.

  The joint density (conditional on $\mathbf{X}$) is (cf. (2.32))

$$f(u_1, u_2, \dots, u_n|\mathbf{X}; \sigma^2) = f(\mathbf{u}|\mathbf{X}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{u}^T\mathbf{u}\right). \tag{11.1}$$

- If we apply (2.31) to $\hat{\boldsymbol{\beta}}_n$, we obtain on the basis of assumption **(B4)** and the previous assumptions **(B2a)**, **(B2b)** that **for each(!) sample size** n

$$\hat{\boldsymbol{\beta}}_n|\mathbf{X} \sim N\left(\boldsymbol{\beta}_0, \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}\right), \tag{11.2}$$

  i. e. the **OLS estimator given X is exactly multivariate normally distributed**.

- If we apply (2.31) to $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}$, we obtain

$$\mathbf{y}|\mathbf{X} \sim N\left(\mathbf{X}\boldsymbol{\beta}_0, \sigma_0^2\mathbf{I}\right) \quad \Longleftrightarrow \quad y_t|\mathbf{X} \sim NID\left(\mathbf{X}_t\boldsymbol{\beta}_0, \sigma_0^2\right), t = 1, \dots, n. \tag{11.3}$$

  For arbitrary parameters, the **normal multiple linear regression model**

$$y_t|\mathbf{X}_t \sim NID(x_{t1}\beta_1 + x_{t2}\beta_2 + \dots + x_{tk}\beta_k, \sigma^2), \quad \beta_1, \dots, \beta_k \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+, \tag{11.4}$$

  is obtained.

- Note that a simple exact distribution such as (11.2) is only possible under the multivariate normal distribution assumption. Why?
- **Summary of the assumptions of the normal multiple linear regression model**
  - **(B1)** Correctly specified model: The DGP is contained in the multiple linear regression model for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.
  - **(B3)** $\mathbf{X}$ has full column rank and
  - **(B4)** $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

  Note that the assumption **(B4)** contains the assumption **(B2)**.
- If there is a conditional distribution for the error vector $\mathbf{u}$ that differs from the normal distribution, the exact distribution of the OLS estimator can generally only be determined using simulation methods.
- If nothing is known about the type of conditional distribution of the errors, then the exact distribution for finite n is unknown, i. e. $\hat{\boldsymbol{\beta}}_n|\mathbf{X} \sim$ *unknown distribution*. However, as shown below, it is possible to determine the asymptotic distribution under certain conditions.

11.2. Asymptotic distribution of the OLS estimator

- **Derivation**
  - As in the case of the expected value estimator, the OLS estimator must also be multiplied by $\sqrt{n}$ in order to obtain a non-singular asymptotic variance-covariance matrix. Under assumptions **(B1)** and **(B3)**, we obtain

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = \sqrt{n}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u} = \underbrace{\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1}}_{:=\mathbf{A}_n} \underbrace{\frac{\mathbf{X}^T\mathbf{u}}{\sqrt{n}}}_{:=\mathbf{a}_n}.$$

  - Now we have to apply Slutzky's theorem (3.4) from section 3.5: If

    i) $\mathbf{A}_n \xrightarrow{P} \mathbf{A}$ and

    ii) $\mathbf{a}_n \xrightarrow{d} \mathbf{a}$ hold,

    then $\mathbf{A}_n\mathbf{a}_n \xrightarrow{d} \mathbf{A}\mathbf{a}$.
  - For i) to apply, **(A1)** must also apply, so that

$$\operatorname*{plim}_{n\to\infty}\left(\mathbf{X}^T\mathbf{X}/n\right)^{-1} = \mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}$$

    applies.

- For ii) to hold, assumption (A2) must be "'reinforced". A central limit theorem must now apply for $\mathbf{X}^T\mathbf{u}/\sqrt{n}$:

  (A3) $\frac{1}{\sqrt{n}}\mathbf{X}^T\mathbf{u} \xrightarrow{d} \mathbf{w}_\infty \sim N\left(\mathbf{0}, \sigma_0^2 \mathbf{S}_{\mathbf{X}^T\mathbf{X}}\right)$

- **Asymptotic distribution of the OLS estimator**
  The assumptions **(B1)**,**(B3)**, as well as the assumptions **(A1)** and **(A3)** hold for the multiple linear regression model. Then it holds that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) &= \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{X}^T\mathbf{u} \\ &\xrightarrow{d} \mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}\mathbf{w}_\infty \sim N\left(\mathbf{0}, \sigma_0^2 \mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}\right). \qquad (11.5)\end{aligned}$$

- In practice, the assumptions **(A1)** and **(A3)** (high level assumptions) cannot be checked directly. Therefore, these assumptions are generally replaced by assumptions that are more descriptive and easier to check. More on this soon.

- **Application of the asymptotic distribution in practice**:

  - In heuristic notation, the asymptotic distribution can also be written as

$$\hat{\boldsymbol{\beta}}_n \overset{approximately}{\sim} N\left(\boldsymbol{\beta}_0, \frac{\sigma_0^2}{n}\mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}\right),$$

    since for a given sample size n is cancelled out.

  - Since $\mathbf{S}_{\mathbf{X}^T\mathbf{X}}$ and σ_0^2 are unknown, the asymptotic distribution is not applicable. The error variance σ_0^2 can be estimated with s^2 and $\mathbf{S}_{\mathbf{X}^T\mathbf{X}}$ by

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^T\mathbf{X}_t. \qquad (11.6)$$

    This gives the following heuristic notation

$$\hat{\boldsymbol{\beta}}_n \overset{approximately}{\sim} N\left(\boldsymbol{\beta}_0, s^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

    The **main difference** to the exact distribution is that the normal distribution only applies approximately, but the approximation becomes more and more accurate as the sample size n increases.

  - If you want to analyse how good the approximation of the asymptotic normal distribution is, you generally have to do this with the help of computer simulations, so-called **Monte Carlo simulations**.

- **When is assumption (A3) fulfilled?**

  For example, if there is a **random sample** and assumption **(B2)** applies. These assumptions can be weakened, see section 13.4.

Proof sketch:

- It holds that $\mathbf{X}^T\mathbf{u} = \sum_{t=1}^n \underbrace{\mathbf{X}_t^T u_t}_{:=\mathbf{v}_t}$. First, $E[\mathbf{v}_t]$ and $Var(\mathbf{v}_t)$ are determined.

- From assumption **(B2a)** $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ (strict exogeneity) it follows that

$$\begin{aligned} E[u_t|\mathbf{X}] &= 0 \quad \text{for all } t = 1, 2, \dots, n. \\ E\left[E[u_t|\mathbf{X}]|\mathbf{X}_t\right] &= E\left[u_t|\mathbf{X}_t\right] = 0. \\ E[\mathbf{X}_t^T u_t] &= E\left[E[\mathbf{X}_t^T u_t|\mathbf{X}_t]\right] = \mathbf{0}. \end{aligned}$$

  Thus, the expected value of $\mathbf{v}_t = \mathbf{X}_t^T u_t$ is a zero vector.

- Due to the assumption **(B2b)**, it holds that $Var(\mathbf{u}|\mathbf{X}) = \sigma_0^2\mathbf{I}$, and

$$\begin{aligned} Var(u_t|\mathbf{X}) &= E\left[u_t^2|\mathbf{X}\right] = \sigma_0^2 \quad \text{for all } t = 1, 2, \dots, n. \\ E\left[E[u_t^2|\mathbf{X}]|\mathbf{X}_t\right] &= E\left[u_t^2|\mathbf{X}_t\right] = Var(u_t|\mathbf{X}_t) = \sigma_0^2. \\ Var(\mathbf{v}_t) = Var\left(\mathbf{X}_t^T u_t\right) &= E\left[\mathbf{X}_t^T u_t^2 \mathbf{X}_t\right] = E\left[E[u_t^2\mathbf{X}_t^T\mathbf{X}_t|\mathbf{X}_t]\right] = \sigma_0^2 E\left[\mathbf{X}_t^T\mathbf{X}_t\right]. \end{aligned}$$

  Since $\mathbf{v}_t \sim (\mathbf{0}, Var(\mathbf{v}_t))$ and thus $\mathbf{X}_t^T u_t \sim (\mathbf{0}, Var(\mathbf{X}_t^T u_t))$ holds and a random sample was assumed, the multivariate central limit theorem (5.73) can be applied to the estimator of the expected value

$$\hat{\boldsymbol{\mu}}_{\mathbf{v},n} = \frac{1}{n}\mathbf{X}^T\mathbf{u} = \frac{1}{n}\sum_{t=1}^n \mathbf{X}_t^T u_t.$$

  One obtains

$$\sqrt{n}\hat{\boldsymbol{\mu}}_{\mathbf{v},n} \xrightarrow{d} N\left(\mathbf{0}, \sigma_0^2 \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n E\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right).$$

  It can be shown that the following applies based on assumption **(A1)**:

$$\mathbf{S}_{\mathbf{X}^T\mathbf{X}} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n E\left[\mathbf{X}_t^T\mathbf{X}_t\right].$$

  Thus one obtains

$$\frac{1}{\sqrt{n}}\sum_{n=1}^n \mathbf{X}_t^T u_t \xrightarrow{d} N\left(\mathbf{0}, \sigma_0^2\mathbf{S}_{\mathbf{X}^T\mathbf{X}}\right). \qquad (11.7)$$

- ♯ Use of the Cramér-Wold device: Choose arbitrary $(k \times 1)$ vector $\boldsymbol{\lambda}$. With the previous results, the following applies,

$$\boldsymbol{\lambda}^T\mathbf{X}_t^T u_t \sim \left(0, \sigma_0^2\boldsymbol{\lambda}^T E\left[\mathbf{X}_t^T\mathbf{X}_t\right]\boldsymbol{\lambda}\right).$$

The asymptotic properties of the estimator of the expected value are then considered

$$\hat{\nu}_n = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\lambda}^T \mathbf{X}_t^T u_t$$

(= (scalar) random sequence). Under the **additional condition** that the **sample observations are stochastically independent** and the usual regularity conditions, the central limit theorem for heterogeneous but independent random variables (5.72) can be applied and the following holds

$$\sqrt{n}\hat{\nu}_n \xrightarrow{d} N\left(0, \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \sigma_0^2 \boldsymbol{\lambda}^T E\left[\mathbf{X}_t^T \mathbf{X}_t\right] \boldsymbol{\lambda}\right).$$

Since this applies to all $\boldsymbol{\lambda}$ with $||\boldsymbol{\lambda}|| > 0$, you can omit $\boldsymbol{\lambda}$ due to the Cramér-Wold device and one obtains

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \mathbf{X}_t^T u_t \xrightarrow{d} N\left(\mathbf{0}, \sigma_0^2 \lim_{t\to\infty} \frac{1}{n}\sum_{t=1}^{n} E\left[\mathbf{X}_t^T \mathbf{X}_t\right]\right)$$

or again

$$\frac{1}{\sqrt{n}}\sum_{n=1}^{n} \mathbf{X}_t^T u_t \xrightarrow{d} N\left(\mathbf{0}, \sigma_0^2 \mathbf{S}_{\mathbf{X}^T\mathbf{X}}\right). \tag{11.7}$$

□

R commands

Calculate the variance-covariance matrix of two variables with `cov()`. Convert the variance-covariance matrix into a **correlation matrix** with `cov2cor()`.

11.3. Exact tests

Applications of exact tests:

- **Specification of the normal linear regression model and checking of assumptions**, cf. section 11.1
  - **(B1)** and $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ (**(B2a)**): $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ contains DGP
    - t-tests, see section 11.3.1; F-tests, see section 11.3.2.
    - Testing the correct functional form, e. g. with **RESET test**, see section 15.3.
    - Testing for parameter stability, e. g. with **Chow test**, see (11.34) in section 11.3.2.
  - **(B3)**: $\mathbf{X}^T\mathbf{X}$ has rank k: Violation leads to error message "‘singular matrix"’.
  - **(B4)**: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$:

  - ∗ Assumes $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$, see above.
  - ∗ Assumes: $Var(\mathbf{u}|\mathbf{X}_t) = \sigma^2$ (**Homoscedasticity**): **Tests for heteroscedasticity**, see section 15.2.
  - ∗ Requires normally distributed errors: **Lomnicki-Jarque-Bera test**, see section 15.4.

- **Checking economic hypotheses**

11.3.1. t-tests: Testing a single restriction

- The parameter to be tested is called β_2. The normal multiple linear regression model is then:
$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \mathbf{u}, \quad \mathbf{u}|\mathbf{X}_1, \mathbf{x}_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{11.8}$$

- Pair of hypotheses: $H_0 : \beta_2 = \beta_{2,H_0}$ versus $H_1 : \beta_2 \neq \beta_{2,H_0}$

- **t-test with known error variance σ_0^2:**
  - **Test statistic**:
$$z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{\sigma_{\hat{\beta}_2}}. \tag{11.9}$$
  - **Exact distribution**: Under the assumptions **(B1)**, **(B3)**, **(B4)** and known error variance, the following applies **under** H_0:
$$z_{\beta_2}|\mathbf{X} \sim N(0, 1). \tag{11.10}$$

    The test distribution is completely known under H_0.

Proof:

Overview of the procedure (The procedure is analogous to the derivation of the test regarding the expected value (5.80))

1. With the help of the Frisch-Waugh-Lovell theorem, page 162, the test statistic z_{β_2} can be written as a linear combination of normally distributed errors.
2. Since a linear combination of multivariate normally distributed random variables is normally distributed again, the test statistic z_{β_2} is normally distributed.
3. The standardisation in (11.9) was chosen so that under H_0 (11.10) applies.

The steps in detail

1. Calculation of the test statistic: Applying the FWL theorem to β_2 in $\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{x}_2\beta_2 + \mathbf{M}_1\mathbf{u}$ results in
$$\hat{\beta}_2 = \frac{\mathbf{x}_2^T\mathbf{M}_1\mathbf{y}}{\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2}, \quad \sigma_{\hat{\beta}_2}^2 = \sigma_0^2(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{-1}$$

Substituting into (11.9) gives

$$z_{\beta_2} = \frac{\frac{\mathbf{x}_2^T\mathbf{M}_1\mathbf{y}}{\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2} - \beta_{2,H_0}}{\sigma_0(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{-1/2}}.$$

Substituting (11.8) with $\beta_2 = \beta_{2,H_0}$, since H_0 is assumed, provides a linear combination of $\mathbf{u}$ for $z_{\hat{\beta}_2}$

$$z_{\beta_2} = \frac{\mathbf{x}_2^T\mathbf{M}_1\mathbf{u}}{\sigma_0(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{1/2}} = \mathbf{B}\mathbf{u}.$$

2. and 3. Derivation of the distribution: Based on (2.33) one obtains

$$z_{\beta_2}|\mathbf{X} \sim N(0,1)$$

since $E[z_{\beta_2}|\mathbf{X}] = E[\mathbf{B}\mathbf{u}|\mathbf{X}] = 0$ and

$$Var(z_{\beta_2}|\mathbf{X}) = Var\left(\mathbf{B}\,Var(\mathbf{u}|\mathbf{X})\mathbf{B}^T|\mathbf{X}\right) = \frac{E(\mathbf{x}_2^T\mathbf{M}_1\mathbf{u}\mathbf{u}^T\mathbf{M}_1\mathbf{x}_2|\mathbf{X}_1,\mathbf{x}_2)}{\sigma_0^2(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)} = \frac{\sigma_0^2(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)}{\sigma_0^2(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)} = 1.$$

□

- If H_1 applies, the test statistic is also normally distributed, but with a mean value that is different from zero. Cf. calculation of the power function in section 5.6.

- ***t*-test with estimated error variance $\hat{\sigma}^2$**:
  - **Idea**: (Cf. derivation of (5.74) in section 5.6) One replaces σ with s in test statistic (11.9). This results in the following estimator for $\sigma_{\hat{\beta}_2}$

$$s^2_{\hat{\beta}_2} = s^2(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{-1} = \frac{\mathbf{y}^T\mathbf{M}_\mathbf{X}\mathbf{y}}{n-k}(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{-1}.$$

  - **Test statistic**:

$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{s_{\hat{\beta}_2}}. \tag{11.11}$$

  - **Exact distribution**: Under the assumptions **(B1)**, **(B3)**, **(B4)** and known error variance, the following applies **under** H_0:

$$t_{\beta_2}|\mathbf{X} \sim t_{n-k}. \tag{11.12}$$

  The test distribution is completely known under H_0.

Proof:

Overview of procedure

1. Reformulate the test statistic as the quotient (11.13) of the test statistic z_{β_2} and a random variable for which the χ^2-distribution is shown in step 2.

2. Show that the denominator in (11.13) is χ^2-distributed.

3. Show that in (11.13) the normally distributed random variable in the numerator and the χ^2-distributed random variable in the denominator are stochastically independent.

4. According to (2.36), the t-distribution then applies.

The steps in detail:

1. Calculation: $\hat{\beta}_2$ remains the same and the variance of the parameter estimator $\sigma^2_{\hat{\beta}_2}$ is estimated by $s^2_{\hat{\beta}_2}$, so that under H_0 we obtain:

$$\begin{aligned} t_{\beta_2} &= \left(\underbrace{\frac{\mathbf{y}^T\mathbf{M_X}\mathbf{y}}{(n-k)}}_{s^2}\right)^{-1/2} \frac{\mathbf{x}_2^T\mathbf{M}_1\mathbf{u}}{(\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{1/2}} \\ &= \left(\frac{\mathbf{y}^T\mathbf{M_X}\mathbf{y}}{\sigma_0^2(n-k)}\right)^{-1/2} \frac{\mathbf{x}_2^T\mathbf{M}_1\mathbf{u}}{(\sigma_0^2\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{1/2}} = \frac{z_{\beta_2}}{\left(\frac{s^2}{\sigma_0^2}\right)^{1/2}}. \end{aligned} \tag{11.13}$$

2. Derivation of the distribution of the random variables in the denominator:

It holds that $\frac{\mathbf{y}^T}{\sigma_0}\mathbf{M_X}\frac{\mathbf{y}}{\sigma_0} = \frac{\mathbf{u}^T}{\sigma_0}\mathbf{M_X}\frac{\mathbf{u}}{\sigma_0} = \frac{(n-k)s^2}{\sigma_0^2} \sim \chi^2(n-k)$, since $\mathbf{u}/\sigma_0 \sim N(\mathbf{0},\mathbf{I})$ and in the term $\frac{\mathbf{u}^T}{\sigma_0}\mathbf{M_X}\frac{\mathbf{u}}{\sigma_0}$ the projection matrix $\mathbf{M_X}$ has just rank $n-k$. This results in a χ^2-distribution with $n-k$ degrees of freedom due to (2.35).

3. Stochastic independence of numerator and denominator

- Numerator:

$$\mathbf{x}_2^T\mathbf{M}_1\mathbf{y} = \mathbf{x}_2^T\mathbf{P_X}\mathbf{M}_1\mathbf{y} = \mathbf{x}_2^T\mathbf{M}_1\mathbf{P_X}\mathbf{y}$$

  since $\mathbf{x}_2$ is already in the subspace of $\mathbf{P_X}$ and

$$\mathbf{P_X}\underbrace{(\mathbf{I}-\mathbf{P}_1)}_{\mathbf{M}_1} = \mathbf{P_X} - \mathbf{P_X}\mathbf{P}_1 = \mathbf{P_X} - \mathbf{P}_1\mathbf{P_X} = \mathbf{M}_1\mathbf{P_X}.$$

  Taking $\mathbf{P_X}\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P_X}\mathbf{u}$ into account, the numerator

$$\mathbf{x}_2^T\mathbf{M}_1\mathbf{y} = \mathbf{x}_2^T\mathbf{M}_1\mathbf{X}\boldsymbol{\beta} + \mathbf{x}_2^T\mathbf{M}_1\mathbf{P_X}\mathbf{u}$$

  given $\mathbf{X}$ depends only on the random vector $\mathbf{P_X}\mathbf{u}$.

- Denominator: is based on the square root of the quadratic form of $\mathbf{M_X}\mathbf{u}/\sigma_0$

- Given $\mathbf{X}$ the random vectors are $\mathbf{P_X u}$ in the numerator and $\mathbf{M_X u}$ in the denominator. Their covariance is zero, since

$$E\left(\mathbf{P_X u u}^T\mathbf{M_X}|\mathbf{X}_1,\mathbf{x}_2\right) = \mathbf{P_X}\sigma_0^2\mathbf{I}\mathbf{M_X} = \sigma_0^2\mathbf{P_X M_X} = \mathbf{0},$$

  since the respective subspaces are orthogonal to each other.

- Since $\mathbf{P_X u}$ and $\mathbf{M_X u}$ are both multivariate normally distributed on the basis of *the same* vector $\mathbf{u}$, the uncorrelatedness results in independence (cf. Davidson (2000, Theorem C.4.1, S. 466)).

4. Validity of the t-distribution:

This means that the **t-statistic** (11.11) according to (2.36) is **under** H_0 **exactly t-distributed with $n-k$ degrees of freedom**, since numerator and denominator are stochastically independent, the numerator is standard normally distributed, and in the denominator $\frac{\mathbf{y}^T}{\sigma_0}\mathbf{M_X}\frac{\mathbf{y}}{\sigma_0}$ is just $\chi^2(n-k)$ distributed and after division by the number of degrees of freedom results in s^2/σ_0^2:

$$t_{\beta_2}|\mathbf{X} = \frac{\hat\beta_2 - \beta_{2,H_0}}{s_{\hat\beta_2}}|\mathbf{X} \sim t_{n-k}. \tag{11.14}$$

□

- The t-test can also be used to test more complicated single restrictions.

Scale elasticity of a Cobb-Douglas production function:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + u$$

where Y, K and L denote output, capital and labour respectively. The null or alternative hypothesis of linear scale elasticity

$$H_0 : \beta_2 + \beta_3 = 1 \quad \text{versus} \quad H_1 : \beta_2 + \beta_3 \neq 1$$

can be written with $\theta = \beta_2 + \beta_3$ as

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta \neq 1,$$

where then with $\beta_3 = \theta - \beta_2$

$$\log Y = \beta_1 + \beta_2(\log K - \log L) + \theta \log L + u$$

is estimated. Alternatively, an F-test can also be carried out.

11.3.2. F-tests: Testing multiple restrictions

An (economic) theory often implies several restrictions regarding the parameters of a regression model.

- Examples of possible linear restrictions:
  1. $H_0 : \beta_2 = \beta_k$
  2. $H_0 : \beta_1 = 1,\ \beta_k = 0$
  3. $H_0 : \beta_1 = \beta_3,\ \beta_2 = \beta_3$
  4. $H_0 : \beta_j = 0,\ j = 2, \ldots, k$
  5. $H_0 : \beta_j + 2\beta_{j+1} = 1,\ \beta_k = 2.$
- All $q \leq k$ linear restrictions can be represented in the following form:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r} \quad \text{vs.} \quad H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r} \tag{11.15}$$

  where the $(q \times k)$ matrix $\mathbf{R}$ and the $(q \times 1)$ vector $\mathbf{r}$ are given and fixed. When formulating, it must of course be ensured that all restrictions in (11.15) are free of contradictions and not redundant.

  Illustrations of the examples:

  1. $H_0 : \beta_2 = \beta_k \Leftrightarrow \beta_2 - \beta_k = 0$:

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{k-1} \\ \beta_k \end{pmatrix} = 0.$$

  2. $H_0 : \beta_1 = 1,\ \beta_k = 0$:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

  3. $H_0 : \beta_1 = \beta_3,\ \beta_2 = \beta_3$:

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

4. $H_0 : \beta_j = 0,\ j = 2, \dots, k$:

$$\underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}}_{\begin{pmatrix} \mathbf{0} & \mathbf{I}_{k-1} \end{pmatrix}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{((k-1)\times 1)} .$$

5. $H_0 : \beta_j + 2\beta_{j+1} = 1,\ \beta_k = 2$:

$$\begin{pmatrix} 0 & \cdots & 1 & 2 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \beta_{j+1} \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} .$$

Continuation Trade flows: Consider regression model

$$\ln(Imports_i) = \beta_1 + \beta_2 \ln(GDP_i) + \beta_3 \ln(Distance_i) \\ + \beta_4\ Openness_i + \beta_5 \ln(Area) + u_i.$$

Question: Do the variables `Openness` and `Area` play a role in combination? In other words: Are both parameters **jointly** statistically significant? The **pair of hypotheses** is:

$$\begin{aligned} \mathrm{H}_0 &: \beta_4 = 0 \quad \text{and} \quad \beta_5 = 0 \quad \text{versus} \\ \mathrm{H}_1 &: \beta_4 \neq 0 \quad \text{and/or} \quad \beta_5 \neq 0. \end{aligned}$$

Writing the null hypothesis in matrix form $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$

$$H_0 : \quad \underbrace{\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{R}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{\mathbf{r}} . \tag{11.16}$$

F-**test: overview and summary**: (cf. section 5.6):

1. **Pair of hypotheses with disjoint null and alternative hypothesis**: $q \leq k$ linear restrictions can be tested, which can be represented in the following form:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r} \quad \text{vs.} \quad H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r} \tag{11.15}$$

where the $(q \times k)$ matrix $\mathbf{R}$ and the $(q \times 1)$ vector $\mathbf{r}$ are given and fixed.

2. **Test statistic**: The F-test statistic is:

$$F = \frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)/q}{s^2} \sim F_{q,n-k} \tag{11.17}$$

   The F-statistic (11.28) is F-distributed with q and $n-k$ degrees of freedom.

3. **Decision rule** for F-test: Reject $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ if

$$F > F_{q,n-k,1-\alpha}.$$

   Alternatively: Reject H_0 if p-value is lower than the significance level. (If σ_0^2 is known, in F s^2 is replaced by σ_0^2 and the $1-\alpha$ quantile of the χ^2-distribution with $n-k$ degrees of freedom is used, see (11.21).)

Derivation of the F-test statistic (11.17):

- How can you form a scalar test statistic for multiple hypotheses?

  Basic idea: By summing the squared deviations

$$\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right) > \text{critical value}.$$

  Is it possible to determine the probability distribution for the squared deviations?

- **Distribution of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$**:
  - If the assumptions **(B1)** and **(B3)** are fulfilled, the following applies,

$$\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{R}\boldsymbol{\beta}_0 + \mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{u}$$

    or

$$\mathbf{R}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{u},$$

  - If the errors are also **multivariate normally distributed given X**, i. e. assumption **(B4)** applies, $\mathbf{R}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ is also multivariate normally distributed due to (2.33):

$$\mathbf{R}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)|\mathbf{X} \sim N\left(\mathbf{0}, \sigma_0^2\,\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right), \tag{11.18}$$

    where $\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T$ has rank q, since $\text{rk}(\mathbf{AB}) = \text{rk}(\mathbf{A})$ if $\mathbf{B}$ is not singular (vgl. Schmidt & Trenkler 2006, Rule 3.2.7).

    Proof: Derivation of the variance:

$$\begin{aligned} Var\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}_0|\mathbf{X}\right) &= Var\left(\mathbf{R}\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = \mathbf{R}Var\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right)\mathbf{R}^T \\ &= \sigma_0^2\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T \end{aligned}$$

    □

  - Adding and subtracting $\mathbf{r}$ in $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}_0 = \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} + \mathbf{r} - \mathbf{R}\boldsymbol{\beta}_0$ yields:

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}|\mathbf{X} \sim N\left(\mathbf{R}\boldsymbol{\beta}_0 - \mathbf{r}, \sigma_0^2\,\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right), \tag{11.19}$$

- Under $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ is contained, (11.19) simplifies to

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}|\mathbf{X} \sim N\left(\mathbf{0}, \sigma_0^2\,\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right), \tag{11.20}$$

- **Distribution of the weighted sum of squares**:
  - **Error variance σ_0^2 known**: Due to the properties of the χ^2-distribution (2.34), under H_0 for the weighted sum of squares of the $(q \times 1)$ normally distributed vector $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ (11.20), it holds that

$$\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[\sigma_0^2\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right) \sim \chi_q^2. \tag{11.21}$$

A weighted instead of an unweighted sum of the squared deviations of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should therefore be used as the test statistic, as the distribution under H_0 is known for this if the error variance σ_0^2 is known.

  - **Error variance σ_0^2 unknown**: In the test statistic (11.21), the error variance σ_0^2 is in the denominator. If σ_0^2 is replaced by the estimator s^2, the denominator now also shows a random variable. The following statistic is therefore a candidate for the F-distribution

$$\frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)}{s^2}. \tag{11.22}$$

Check the requirements for the F-distribution, cf. (2.37): For a random variable to be F-distributed, the numerator and denominator must be χ^2-distributed.

1. Numerator: Since (11.21) is χ^2-distributed, one divides the numerator and denominator of (11.22) by σ_0^2 so that the new numerator corresponds exactly to (11.21) and is therefore χ^2-distributed:

$$\frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[\sigma_0^2\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)}{s^2/\sigma_0^2}. \tag{11.23}$$

2. Denominator: The denominator s^2/σ_0^2 must still be multiplied by $n-k$, because according to the previous section 11.3.1 for the t-test, the following applies,

$$(n-k)s^2/\sigma_0^2 = \mathbf{y}^T\mathbf{M}_{\mathbf{X}}\mathbf{y}/\sigma_0^2 = \mathbf{u}^T\mathbf{M}_{\mathbf{X}}^T\mathbf{M}_{\mathbf{X}}\mathbf{u}/\sigma_0^2 \sim \chi_{n-k}^2. \tag{11.24}$$

One obtains:

$$\frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[\sigma_0^2\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)}{(n-k)s^2/\sigma_0^2}. \tag{11.25}$$

3. Are the numerator and denominator in (11.25) stochastically independent? Yes.

Proof:

- The numerator and denominator can each be written as a quadratic form (cf. section 9.3) $\mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i$, $i = Z, N$:
- Numerator:

$$\begin{aligned} \mathbf{x}_Z &= \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} = \mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{u}, \\ \mathbf{A}_Z &= \left[\sigma_0^2 \mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1} \end{aligned}$$

- Denominator: The term in (11.24) is also a quadratic form with

$$\mathbf{x}_N = \mathbf{M}_{\mathbf{X}}\mathbf{u}, \tag{11.26}$$

$$\mathbf{A}_N = \frac{1}{\sigma_0^2}\mathbf{I}. \tag{11.27}$$

- Since $\mathbf{A}_Z$ and $\mathbf{A}_N$ are known given $\mathbf{X}$, the distribution properties of $\mathbf{x}_Z$ and $\mathbf{x}_N$ are crucial. Due to assumption **(B4)**, $\mathbf{u}$ given $\mathbf{X}$ is multivariate normally distributed with expected value of zero. Therefore, $\mathbf{x}_Z$ and $\mathbf{x}_N$ given $\mathbf{X}$ are also multivariate normally distributed with expected value of zero. Since $\mathbf{x}_Z$ and $\mathbf{x}_N$ depend on the same multivariate normally distributed error vector $\mathbf{u}$, they are stochastically independent if they are uncorrelated, i. e. $Cov(\mathbf{x}_Z, \mathbf{x}_N|\mathbf{X}) = E\left[\mathbf{x}_Z\mathbf{x}_N^T|\mathbf{X}\right] = \mathbf{0}$ applies. Substitution gives:

$$\begin{aligned} E\left[\mathbf{x}_Z\mathbf{x}_N^T|\mathbf{X}\right] &= E\left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{u}(\mathbf{M}_{\mathbf{X}}\mathbf{u})^T|\mathbf{X}\right] \\ &= \mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E\left[\mathbf{u}\mathbf{u}^T|\mathbf{X}\right]\mathbf{M}_{\mathbf{X}} = \sigma_0^2\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\underbrace{\mathbf{X}^T\mathbf{M}_{\mathbf{X}}}_{=\mathbf{0}} = \mathbf{0} \end{aligned}$$

  Numerator and denominator are χ^2-distributed because $\mathbf{x}_Z$ and $\mathbf{x}_N$ are multivariate normally distributed. If the latter are stochastically independent, this must also hold for functions that depend on them, such as the quadratic forms here. Therefore, the numerator and denominator in (11.25) are stochastically independent.

□

4. Division by the correct number of degrees of freedom? Must still be done. I. e. the numerator in (11.25) must be divided by the number of restrictions q. The denominator in (11.25) must be divided by the correct number of degrees of freedom $n-k$. One obtains:

$$F = \frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T\left[\sigma_0^2\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)/q}{\left[(n-k)s^2/\sigma_0^2\right]/(n-k)}.$$

Note: σ_0^2 and $n-k$ are cancelled out so that (11.17) is obtained. This means that F under H_0 is an F-distributed test statistic

$$F \sim F_{q,n-k}. \tag{11.28}$$

The F-statistic (11.28) is therefore F-distributed with q and $n-k$ degrees of freedom.

The test statistic F is called F**-statistic**.

Alternative notations of the F-statistic:

$$\begin{aligned} F &= \frac{\left(\mathbf{R}\hat{\boldsymbol\beta} - \mathbf{r}\right)^T \left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol\beta} - \mathbf{r}\right)/q}{s^2} && (11.17) \\ &= \frac{\left(\hat{\boldsymbol\beta} - \boldsymbol\beta_0\right)^T \mathbf{R}^T\left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\mathbf{R}\left(\hat{\boldsymbol\beta} - \boldsymbol\beta_0\right)/q}{\mathbf{y}^T\mathbf{M}_\mathbf{X}\mathbf{y}/(n-k)} && (11.29) \end{aligned}$$

Power of F-tests:

- It can be shown that under the alternative hypothesis, all quantiles of the distribution of the F-statistic lie to the right of those of the F-distribution under H_0. The further to the right the quantiles under H_1 are compared to the quantiles under H_0 (e.g. due to an increasing sample size n), the higher the power of the F-test.

Joint exclusion restrictions: further ways of calculating the F-statistic

- You can always rearrange the variables in a multiple regression model so that all **exclusion restrictions** with regard to $\boldsymbol\beta$ in the model

  $$\mathbf{y} = \underbrace{\mathbf{X}_1}_{(n\times k_1)} \boldsymbol\beta_1 + \underbrace{\mathbf{X}_2}_{(n\times k_2)} \boldsymbol\beta_2 + \mathbf{u},$$

  $k = k_1 + k_2$, can be summarised in $\boldsymbol\beta_2$.

  The pair of hypotheses is then

  $$\begin{aligned} &H_0 : \beta_j = 0, j = k_1+1, \ldots, k_1+k_2 \Leftrightarrow \boldsymbol\beta_2 = \mathbf{0} \quad \text{versus} \\ &H_1 : \beta_{k_1+1} \neq 0 \text{ and/or } \ldots \text{ and/or } \beta_{k_1+k_2} \neq 0 \Leftrightarrow \boldsymbol\beta_2 \neq \mathbf{0}. \end{aligned}$$

  The exclusion restrictions can then be written as

  $$\underbrace{\begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}}_{\left(\mathbf{0}_{k_2\times k_1} \quad \mathbf{I}_{k_2}\right)} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{k_1-1} \\ \beta_{k_1} \\ \beta_{k_1+1} \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

  $$\left(\mathbf{0}_{k_2\times k_1} \quad \mathbf{I}_{k_2}\right)\boldsymbol\beta = \mathbf{0}_{k_2\times 1}.$$

In this case, there are further ways of calculating the F-statistic (see also bachelor course **Introduction to Econometrics**) by estimating the restricted and unrestricted model separately:

1. **Restricted regression**: Regress $\mathbf{y}$ exclusively on $\mathbf{X}_1$ and store the residual sum of squares $SSR_1 = \tilde{\mathbf{u}}^T\tilde{\mathbf{u}}$ or, in the case of a constant contained in $\mathbf{X}_1$, also R_1^2.
2. **Unrestricted regression**: Regress $\mathbf{y}$ on $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix}$ and store $SSR = \hat{\mathbf{u}}^T\hat{\mathbf{u}}$ bzw. R^2.

The **further calculation options** are (note $q = k_2$):

$$\begin{aligned} F &= \frac{(SSR_1 - SSR)/k_2}{SSR/(n-k)} && (11.30)\\ &= \frac{\left(\tilde{\mathbf{u}}^T\tilde{\mathbf{u}} - \hat{\mathbf{u}}^T\hat{\mathbf{u}}\right)/k_2}{\hat{\mathbf{u}}^T\hat{\mathbf{u}}/(n-k)} \\ &= \frac{(R^2 - R_1^2)/k_2}{(1-R^2)/(n-k)} && (11.31)\\ &\sim F_{k_2,n-k}. \end{aligned}$$

Continuation Trade flows: Testing the null hypothesis (11.16):

- Determine the critical value c: Calculate $1-\alpha$-quantile of the $F_{2,44}$-distribution with R command `qf(1-alpha,2,44)`. For $\alpha = 0.05$ one obtains 3.209278.
- **Calculating the F-statistic** and the **p-value** is most easily done with the R command `linearHypothesis` (requires R package `car`):

R code (Extract from R program in section A.4)

```
################################################################################
################################################################################
# Section 11.3 Exact Tests
################################################################################

alpha                 <- 0.05         # Significance level
# Estimating model 4
mod_4_kq              <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) +
                         log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))
summary(mod_4_kq)
```

Listing 11.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

and yields

```
                Hypothesis:
                ebrd_tfes_o = 0
                log(cepii_area_o) = 0
                Model 1: restricted model
                Model 2: log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)
                + ebrd_tfes_o +  log(cepii_area_o)

                Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

```
1     46 39.645
2     44 32.018  2    7.6272 5.2408 0.009088 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

– The null hypothesis is rejected. At least one of the two variables Openness or log(Area) is significantly different from zero at the 5% level.

Proof: ♯ **Possibilities of proof** for (11.30) and (11.31)

– **1st possibility of proof**: using the formula for the inversion of partitioned matrices

– **2nd possibility of proof**: with the help of the Frisch-Waugh-Lovell theorem:

1. Note that the residual sum of squares of the unrestricted model

$$SSR = \mathbf{y}^T \mathbf{M}_{\mathbf{X}} \mathbf{y}$$

using the decomposition of the residual sum of squares and the **Frisch-Waugh-Lovell theorem**, see section 7.1, based on the regression

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \boldsymbol{\beta}_2 + Residuals$$

can also be written as

$$\begin{aligned} SSR &= TSS - ESS \\ &= \mathbf{y}^T \mathbf{M}_1 \mathbf{y} - \mathbf{y}^T \mathbf{M}_1 \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y} \\ &= \mathbf{y}^T \mathbf{M}_1 \mathbf{y} - \mathbf{y}^T \mathbf{M}_1 \underbrace{\mathbf{M}_1 \mathbf{X}_2 \left(\mathbf{X}_2^T \mathbf{M}_1 \mathbf{M}_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2^T \mathbf{M}_1}_{\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2}} \mathbf{M}_1 \mathbf{y} \\ &= \mathbf{y}^T \mathbf{M}_1 \mathbf{y} - \mathbf{y}^T \mathbf{M}_1 \mathbf{X}_2 \left(\mathbf{X}_2^T \mathbf{M}_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2^T \mathbf{M}_1 \mathbf{y}. \end{aligned}$$

2. The numerator in the F-statistic (11.30) is then

$$\begin{aligned} SSR_1 - SSR &= \mathbf{y}^T \mathbf{M}_1 \mathbf{y} - \left[\mathbf{y}^T \mathbf{M}_1 \mathbf{y} - \mathbf{y}^T \mathbf{M}_1 \mathbf{X}_2 \left(\mathbf{X}_2^T \mathbf{M}_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2^T \mathbf{M}_1 \mathbf{y} \right] \\ &= \mathbf{y}^T \mathbf{M}_1 \mathbf{X}_2 \left(\mathbf{X}_2^T \mathbf{M}_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2^T \mathbf{M}_1 \mathbf{y} \qquad (11.32) \\ &= \mathbf{u}^T \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{u}. \end{aligned}$$

The last equal sign holds, since under H_0 $\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{u}$ (verify!).

3. Since $\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2}$ is a projection matrix with rank k_2 it follows from the property (2.35) of the χ^2-distribution that for normally distributed errors under H_0

$$SSR_1 - SSR = \frac{\mathbf{u}^T}{\sigma} \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \frac{\mathbf{u}}{\sigma} \sim \chi^2(k_2).$$

The following applies to the denominator,

$$SSR = \frac{\mathbf{u}^T}{\sigma} \mathbf{M_X} \frac{\mathbf{u}}{\sigma} \sim \chi^2(n-k).$$

Numerator and denominator are therefore both χ^2-distributed.

The random vectors in the numerator $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{u}$ and denominator $\mathbf{M_X u}$ have covariance zero, since

$$\mathbf{M_X M}_1 = \mathbf{M}_1 \mathbf{M_X}$$

and consequently

$$\mathbf{M_X M}_1 \mathbf{X}_2 = \mathbf{M}_1 \mathbf{M_X X}_2 = \mathbf{0}$$

(the columns of $\mathbf{X}_2$ are contained in the orthogonal space of $\mathbf{M_X}$) and thus $E\left[\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{u}\mathbf{u}^T\mathbf{M_X}\right] = \mathbf{0}$. Due to the multivariate normal distribution assumption, the random vectors are therefore also stochastically independent.

Thus, based on the definition of the F-distribution

$$F = \frac{(SSR_1 - SSR)/k_2}{SSR/(n-k)} \sim F_{k_2, n-k}.$$

□

- Due to (11.32), there is another notation of the F-statistic (11.30)

$$F = \frac{\mathbf{y}^T\mathbf{M}_1\mathbf{X}_2 \left(\mathbf{X}_2^T\mathbf{M}_1\mathbf{X}_2\right)^{-1} \mathbf{X}_2^T\mathbf{M}_1\mathbf{y}/k_2}{\mathbf{y}^T\mathbf{M_X y}/(n-k)} \tag{11.33}$$

- The F-statistic (11.30) can also be used for general linear restrictions. For this, however, the model under H_0 must be suitably transformed, see bachelor course **Introduction to Econometrics**.

Other well-known F-tests:

- **Single hypothesis**: F-statistic is square of the t-statistic and corresponds to a two-sided t-test.

- **Chow test for structural breaks**: Test for constancy of all/some parameters across 2 subsamples, each indexed by I and II. If they are not constant, a separate estimate must be made for each subsample

$$\mathbf{y}_I = \mathbf{X}_I \boldsymbol{\beta}_I + \mathbf{u}_I \tag{11.34a}$$
$$\mathbf{y}_{II} = \mathbf{X}_{II} \boldsymbol{\beta}_{II} + \mathbf{u}_{II}. \tag{11.34b}$$

  The null hypothesis (parameter constancy) is

$$H_0 : \boldsymbol{\beta}_I = \boldsymbol{\beta}_{II}.$$

Thus, under H_0 the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

is to be estimated.

Under H_1, on the other hand, elements of $\boldsymbol{\beta}_{II}$ and $\boldsymbol{\beta}_I$ can differ and one estimates in matrix notation with

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_I \\ \mathbf{y}_{II} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_I \\ \mathbf{X}_{II} \end{pmatrix}$$

the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \begin{pmatrix} \mathbf{O} \\ \mathbf{X}_{II} \end{pmatrix} \boldsymbol{\gamma} + \mathbf{u}.$$

The pair of hypotheses is

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \quad \text{versus} \quad H_1 : \gamma_j \neq 0 \text{ for at least one } j.$$

If, in addition to **(B1)**, **(B3)**, **(B4)**, the subsamples are stochastically independent, the **Chow test** is exact.

Pivotal tests

- **Definition**: A random variable, e. g. a test statistic under H_0, with the property that its distribution is the same for all DGPs in a model $\mathbb{M}$ is called **pivotal** for the model $\mathbb{M}$.
- The null hypothesis rarely specifies the complete DGP. If this is the case, it is referred to as a **simple hypothesis**.
- In general, the model contains several different DGPs under the null hypothesis: **compound hypothesis**. If the **exact** distribution of a test of a compound null hypothesis depends on the DSP that generated the sample data, the test statistic is not pivotal, as the test distribution changes depending on the specific DSP for the same null hypothesis. An **exception** are **exact tests**.
- Possible solutions for all other cases:
  - Without knowledge of the DGP: **asymptotically pivotal tests**, see next section 11.4, i. e. the asymptotic distribution of the test statistic is pivotal.
  - With knowledge of the DGP: **Monte Carlo tests**, see section 11.5.1.
  - Without knowledge of the DGP: **bootstrap tests**, see section 11.5.2.

11.4. Asymptotic tests

The normal multiple linear regression model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

If the assumption **(B2)** of (strictly) exogenous regressors is not fulfilled, for example due to lagged endogenous variables as regressors or the assumption **(B4)** of normally distributed errors, then the exact distribution of the t-statistic and the F-statistic from section 11.3 cannot generally be determined analytically. Even if this were possible, the distribution of the t-statistic would generally not be pivotal.

- Under assumptions **(B1)**, **(B2)**, **(B3)**, **(A1)** and **(A3)**, the introduced t-tests and F-tests are **asymptotically** valid, since the OLS estimator is asymptotically normally distributed.
- The results also remain valid under the assumptions of the dynamic linear regression model **(C1)**, **(C2)**, **(C3)** and **(C4a)** or **(C4b)**, see section 13.4.

11.4.1. Asymptotic t-test

Here: The parameter to be tested in the linear regression model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \mathbf{u}, \quad u_t|\mathbf{X}_t \sim (0, \sigma^2), t = 1, \dots, n.$$

is β_2.

Asymptotic t-test: Overview

1. The **pair of hypotheses** is: $H_0 : \beta_2 = \beta_{2,H_0}$ versus $H_1 : \beta_2 \neq \beta_{2,H_0}$.
2. **Test statistic** and **test distribution**: Under H_0, it holds that
$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{s_{\hat{\beta}_2}} = \frac{z_{\beta_2}}{(s^2/\sigma_0^2)^{1/2}} \xrightarrow{d} N(0,1).$$
3. **Decision rule: analogue to decision rule for two-tailed or one-tailed tests**.

Note: **In practice**, the **t-distribution with $n-k$ degrees of freedom** is usually used, as this often provides a better approximation of the (unknown) exact distribution than the standard normal distribution.

Derivation of the asymptotic distribution

1. Denominator: Under H_0, it holds that
$$\text{plim}_{n\to\infty} \left(s^2/\sigma_0^2\right)^{1/2} = 1.$$
2. Numerator: The numerator in (11.13) is expanded with $n^{-1/2}$ to
$$z_{\beta_2} = \frac{n^{-1/2}\mathbf{x}_2^T\mathbf{M}_1\mathbf{u}}{\sigma_0(n^{-1}\mathbf{x}_2^T\mathbf{M}_1\mathbf{x}_2)^{1/2}}$$
and obviously has an expected value of 0 and a variance of 1, as the variance of the numerator is equal to the square of the denominator (verify both!).

3. Assuming that the regularity conditions for a multivariate central limit theorem for $n^{-1/2}\mathbf{x}_2^T\mathbf{M}_1\mathbf{u}$ (i. e for **(A1)**, **(A3)**) are fulfilled, we obtain

$$z_{\beta_2} \xrightarrow{d} N(0,1).$$

4. Using the rule, see section 3.4, "‘If $\mathbf{a}_n \xrightarrow{d} \mathbf{a}$ and $\operatorname{plim}\mathbf{A}_n = \mathbf{A}$, then $\mathbf{A}_n\mathbf{a}_n \xrightarrow{d} \mathbf{A}\mathbf{a}$"’ it follows that

$$t_{\beta_2} = \underbrace{\left(s^2/\sigma_0^2\right)^{-1/2}}_{\xrightarrow{P}1} \underbrace{z_{\beta_2}}_{\xrightarrow{d}N(0,1)} \xrightarrow{d} N(0,1) \tag{11.35}$$

Then the standard normal distribution is obtained again asymptotically under H_0 and all properties of the t-test remain asymptotically valid.

11.4.2. Asymptotic F-test

Asymptotic F-test: Overview

1. **Pair of hypotheses with disjoint null and alternative hypothesis**: as for exact F-test.

2. **Test statistic** and **test distribution**: Under H_0, it holds that

$$qF_n = \left(\mathbf{R}(\hat{\boldsymbol\beta} - \boldsymbol\beta_0)\right)^T \left[s^2\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T\right]^{-1} \left(\mathbf{R}(\hat{\boldsymbol\beta} - \boldsymbol\beta_0)\right) \xrightarrow{d} \chi^2(q) \tag{11.36}$$

In small samples, the F-statistic is often used directly instead together with an (approximate) F-distribution with q and $n-k$ degrees of freedom, as this often provides a better approximation of the (unknown) exact distribution than the χ^2-distribution.

3. **Decision rule**:

$$qF > \chi^2_{q,1-\alpha} \tag{11.37}$$

$$F > F_{q,n-k,1-\alpha} \tag{11.38}$$

Alternatively: Reject H_0 if p-value (based on the asymptotic distribution) is lower than the significance level.

Derivation of asymptotic F-test

- If the relevant assumptions, cf. beginning of the section, are fulfilled, so that

$$\sqrt{n}\left(\hat{\boldsymbol\beta} - \boldsymbol\beta_0\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_0^2\mathbf{S}^{-1}_{\mathbf{X}^T\mathbf{X}}\right)$$

applies, an asymptotic χ^2-distribution follows from the theorem about continuous mappings (3.3) and the properties of the χ^2-distribution (2.34):

$$n\left(\hat{\boldsymbol\beta} - \boldsymbol\beta_0\right)^T \left[\sigma_0^2\mathbf{S}^{-1}_{\mathbf{X}^T\mathbf{X}}\right]^{-1} \left(\hat{\boldsymbol\beta} - \boldsymbol\beta_0\right) \xrightarrow{d} \chi^2(k). \tag{11.39}$$

- Asymptotic distribution of the F-statistic (11.28): Using $\mathbf{A}_n \mathbf{a}_n \xrightarrow{d} \mathbf{Aa}$, cf. section 3.4, we obtain again (11.39) from (11.28) together with $\text{plim}_{n\to\infty} s^2 = \sigma_0^2$ and **(A1)** (or **(C3)**) after multiplying by q. This means that under H_0

$$qF_n = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)^T \left[s^2\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T\right]^{-1} \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right) \xrightarrow{d} \chi^2(q). \tag{11.36}$$

- In the case of exclusion restrictions, the F-statistic (11.28) can alternatively be written as (11.33) under H_0. In this case, of course, the following also applies

$$qF_n = q\frac{\mathbf{y}^T\mathbf{M}_1\mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{M}_1\mathbf{y}/q}{\mathbf{y}^T\mathbf{M}_{\mathbf{X}}\mathbf{y}/(n-k)} \xrightarrow{d} \chi^2(q). \tag{11.40}$$

- Since it holds (cf. section 2.9) that for $n \to \infty$ a sequence of F-distributed random variables $X_n \sim F(q, n-k)$ converges to a χ^2-distribution,

$$qX_n \xrightarrow{d} \chi^2(q), \tag{11.41}$$

  F_n can also be approximated by an $F(q, n-k)$ distribution, which in small samples often provides a better approximation than the χ^2-distribution.

Power: It can be shown that the following applies under H_1,

$$qF \overset{n\to\infty}{\longrightarrow} \infty. \tag{11.42}$$

This means that the power for $n \to \infty$ approaches 1 asymptotically, since $\lim_{n\to\infty} P(qF > c) = 1$. In finite samples, the power is typically less than 1.

Actual versus nominal size

- **Actual size**: Size of a test (5.78), which results from the exact but possibly unknown distribution.
- **Nominal size**: Size of a test that results on the basis of the asymptotic distribution.
- Since the exact distribution for each DGP and sample size is known for exact tests, the nominal and actual significance levels match.
- In asymptotic tests, the closer the asymptotic distribution approximates the actual distribution (which generally depends on the DGP and the number of observations), the closer the nominal and actual significance levels are. For predetermined DGPs, the degree of similarity can be determined using Monte Carlo simulations.
- For asymptotic tests, the actual size is unknown. Therefore, the critical value is chosen so that the nominal size corresponds to the chosen significance level.
- A test is called “oversized” if the actual size (e. g. determined by simulations) is greater than the significance level.

11.5. Monte Carlo tests and bootstrap tests

11.5.1. Monte Carlo tests

- **Empirical distribution function** of the observed sample elements x_t, $t = 1, \ldots, n$:

$$\hat{F}(x) = \frac{1}{n} \sum_{t=1}^{n} 1(x_t \leq x), \tag{11.43}$$

where $1(\cdot)$ denotes the **indicator function**

$$1(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false.} \end{cases} \tag{11.44}$$

Fundamental Theorem of Statistics The empirical distribution function is consistent in the case of i.i.d. random variables

$$\text{plim } \hat{F}(x) = F(x). \tag{11.45}$$

The i.i.d. assumption can be weakened.

- Notation in this section: τ denotes an arbitrary test statistic and $\hat{\tau} = \hat{\tau}(\mathbf{X}, \mathbf{y})$ a value of the test statistic τ calculated on the basis of sample observations.

- The **exact** p**-value** of a calculated test statistic $\hat{\tau}$ **with right-tailed critical value** results (cf. (5.85)) from

$$p(\hat{\tau}) := P(\tau > \hat{\tau}|\boldsymbol{\theta}_{H_0}) = 1 - P(\tau \leq \hat{\tau}|\boldsymbol{\theta}_{H_0}) = 1 - F(\hat{\tau}|\boldsymbol{\theta}_{H_0}), \tag{11.46}$$

where $F(\cdot|\boldsymbol{\theta}_{H_0})$ denotes the exact distribution of the calculated test statistic $\hat{\tau}$ under H_0.

As a reminder: Reject H_0 if $p(\hat{\tau}) < \alpha$ or $\hat{\tau} > c_\alpha$.

If $F(\cdot|\boldsymbol{\theta}_{H_0})$ is unknown, the test distribution can be approximated as accurately as required by the empirical distribution function, **if the DGP is completely known** or the test is **pivot**. The greater the number of replications (Monte Carlo simulations) B, the more accurate the approximation. The **computer-simulated** p value is

$$\hat{p}(\hat{\tau}) = 1 - \hat{F}(\hat{\tau}|\boldsymbol{\theta}_{H_0}) = 1 - \frac{1}{B} \sum_{j=1}^{B} 1(\tau_j^* \leq \hat{\tau}), \tag{11.47}$$

where τ_j^* is the value of the test statistic in the j-th simulation under H_0.

- Performing a Monte Carlo test requires the generation of random numbers using a random number generator, see e. g. Davidson & MacKinnon (2004, S. 157-159).

11.5.2. Bootstrap tests

- The **idea** of a bootstrap test is to estimate the unknown DGP and then apply the Monte Carlo test technique.

- Necessary condition: All necessary properties of the DGP can be estimated consistently with a suitable convergence rate.
- Example: Multiple regression model

$$y_t = \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \quad u_t|\mathbf{X} \sim IID(0, \sigma^2)$$

  We want to test

$$H_0 : \beta_k = 0 \quad \text{versus} \quad H_1 : \beta_k \neq 0.$$

  The bootstrap test works if, among other things,

  - $\boldsymbol{\beta}$ and σ^2 can be estimated consistently with rate $\sqrt{n}$ and
  - the distribution of $\mathbf{u}$ given $\mathbf{X}$ is known or can be estimated accordingly.

Generate bootstrap samples

- Estimate $\boldsymbol{\beta}$ with a consistent and as efficient as possible estimator and calculate the desired test statistic $\hat{\tau}$.
  - One can estimate $\boldsymbol{\beta}$ under H_1 and gets the OLS residual vector $\hat{\mathbf{u}}$.
  - One can estimate $\begin{pmatrix}\beta_1 & \dots & \beta_{k-1}\end{pmatrix}^T$ under H_0 and gets the OLS residual vector $\tilde{\mathbf{u}}$.

  In general, better results are obtained if the estimation is carried out under H_0.

- **Assumption i.i.d. normally distributed errors** (Assumption (B4)): **Parametric bootstrap** One can then estimate σ^2 and generate the n error variables in $\mathbf{u}_j^*$ from $N(\mathbf{0}, s^2\mathbf{I})$ for each bootstrap sample $(\mathbf{y}_j^*, \mathbf{X}_j)$.

  1. Then, for the jth bootstrap sample, the vector of dependent variables under H_0 can be iteratively generated by

$$y_{jt}^* = \tilde{\beta}_1 x_{t1} + \cdots + \tilde{\beta}_{k-1} x_{t,k-1} + u_{jt}^*, \quad t = 1, 2, \ldots, n.$$

  2. For the j-sample $(\mathbf{y}_j^*, \mathbf{X}_j)$, the test statistic, here the squared t-test (=F-test), can then be calculated by estimating the unrestricted model

$$\tau_j^* = \left(t_j^*\right)^2, \quad t_j^* = \hat{\beta}_{jk}^* / \hat{\sigma}_{j,\hat{\beta}_{jk}^*}^*.$$

  After B replications, the empirical distribution function is then calculated and the bootstrap p-value is obtained according to (11.47) from

$$\hat{p}(\hat{\tau}) = 1 - B^{-1} \sum_{j=1}^{B} 1\left(\tau_j^* \leq \hat{\tau}\right).$$

- **Assumption i.i.d. errors** (Assumption (B2)): **Nonparametric / Semiparametric bootstrap**

1. Under H_0 the OLS parameter estimators are consistent and so are the estimated errors

$$\begin{aligned}\operatorname*{plim}_{n\to\infty} \tilde{u}_t &= \operatorname*{plim}_{n\to\infty} \left(y_t - \tilde{\beta}_{n1}x_{t1} - \cdots - \tilde{\beta}_{n,k-1}x_{t,k-1}\right) \\ &= y_t - x_{t1} \operatorname*{plim}_{n\to\infty} \tilde{\beta}_{n1} - \cdots - x_{t,k-1} \operatorname*{plim}_{n\to\infty} \tilde{\beta}_{n,k-1} \\ &= y_t - \beta_1 x_{t1} - \cdots - \beta_{k-1}x_{t,k-1} = u_t.\end{aligned}$$

2. 'Asymptotically', one can also draw from the errors with replacement (**resampling**), because due to the Fundamental Theorem of Statistics, the empirical distribution of the u_t's approximates the true error distribution.

3. Due to the consistency of the residual estimator, the residuals can also be used instead of the unknown errors.

4. Refinements:

   – **rescaled residuals**

   $$\tilde{u}_t^+ = \tilde{u}_t \left(\frac{n}{n - k_{H_0}}\right)^{1/2}.$$

   This corrects the variance of the residuals, which is smaller than the variance of the errors (cf. section 9.5), so that it corresponds to the estimated variance of the errors s^2.

   – **centred and rescaled residuals**

   $$\tilde{u}_t^+ = \left(\tilde{u}_t - \bar{\tilde{u}}\right) \left(\frac{n}{n - k_{H_0}}\right)^{1/2}.$$

   This is necessary if, for example, the regression model does not contain a constant, because then the mean value of the residuals is not equal to zero and thus the bootstrap test is biased.

- **Wild Bootstrap** and **Block Bootstrap**: In the case of heteroscedastic and autocorrelated errors, the above methods do not work. More complicated methods are required here.

- **Number of bootstrap replications**: Choose B so that the quantile, cf. (2.8) in section 2.5.1, can be determined exactly for type I errors:

  – In total, there are $B + 1$ rank positions r for the test statistic $\hat{\tau}$. Example: $B = 2$, where the ranks are arranged in descending order (cf. Davidson & MacKinnon (2004), p. 164):

  $$r = 2 : \hat{\tau} < \min_j(\tau_j^*), \quad r = 1 : \min_j(\tau_j^*) < \hat{\tau} < \max_j(\tau_j^*), \quad r = 0 : \max_j(\tau_j^*) < \hat{\tau}$$

  – If you divide the rank position r by the number of bootstrap replications B, you get the p-value for $\hat{\tau}$, because $0 = \frac{0}{B} \leq \frac{r}{B} \leq \frac{B}{B} = 1$.

  – This means that the bootstrap test rejects under H_0 if $r/B < \alpha$, where α denotes the chosen significance level. Therefore $r < B\alpha$ applies.

  - Let $\lfloor x \rfloor$ denote the largest integer number that is smaller than x. Then, for a given $B\alpha$, the number of rank positions for which H_0 is rejected can be expressed as $\lfloor B\alpha \rfloor + 1$. Example: $B = 9$ and $\alpha = 0.5$. This means that the null hypothesis is rejected for $r = 0, 1, 2, 3, 4$. There are $\lfloor B\alpha \rfloor + 1 = \lfloor 4.5 \rfloor + 1 = 5$ rank positions with rejection.

  - Since there are a total of $B + 1$ rank positions,

$$\frac{\lfloor B\alpha \rfloor + 1}{B + 1}$$

  must be equal to α. Given α, B is therefore determined from

$$\alpha(B + 1) = \lfloor \alpha B \rfloor + 1.$$

  For $\alpha = 0.05$, for example, $B = 99$ makes sense.

Remarks

- **Bootstrap test instead of asymptotic test**?

  If

  - the distribution of the test statistic is asymptotically pivotal and

  - the errors of the model are i.i.d. (otherwise more complicated bootstrap methods must be used, e.g. block bootstrap for correlated errors),

  then the distribution of the bootstrap test converges faster with increasing sample size to the (unknown) exact distribution of the test statistic than the asymptotic distribution, more precisely with n^{-1} instead of $n^{-1/2}$. This explains the widespread use of bootstrap.

- **Caution**: If the test statistic is not asymptotically pivotal, then the bootstrap test and the asymptotic test have the same convergence rate, so bootstrap is useless.

- Bootstrap methods can also be used for dynamic regression models under certain conditions. In this case, $\mathbf{X}_j^*$ is also generated for the jth sample $(y_j^*, \mathbf{X}_j^*)$. For the implementation in a simple example, see Davidson & MacKinnon (2004, p. 160).

- Further reading: e. g. Horowitz (2001), Horowitz (2003).

11.6. Confidence intervals and ellipsoids

11.6.1. Confidence intervals

- Definition: **Confidence interval**:

  - A random interval that can be calculated on the basis of sample information $(\mathbf{y}, \mathbf{X})$ and contains the true parameter value θ_0 with probability $1 - \alpha$ is called a confidence interval.

(It follows that for a large number of samples all generated by the same DGP, the true parameter value should be approximately contained in $1-\alpha$ of all calculated confidence intervals).

- Davidson & MacKinnon (2004, Chapter 5) choose an alternative definition: If all null hypotheses (regarding a parameter)

  $$H_0 : \theta = \theta_{H_0},$$

  that are not rejected at a given significance level of α are summarised in an interval, a confidence interval with a confidence level of

  $$1-\alpha$$

  is obtained.

- Formal: Given a non-negative test statistic $\tau(\mathbf{y}, \mathbf{X}, \theta_{H_0})$ and a significance level α, a confidence interval contains all θ_{H_0} for which the following applies,

  $$KI = \left\{\theta_{H_0} | P_{\theta_{H_0}} \left(\tau(\mathbf{y}, \mathbf{X}, \theta_{H_0}) \leq c_\alpha\right) = 1-\alpha\right\}, \qquad (11.48)$$

  where $P_{\theta_{H_0}}(\cdot)$ means that the probability is calculated under the respective null hypothesis H_0 and c_α is the critical value at the significance level α.

- The bounds $[\theta_l, \theta_u]$ of the confidence interval are obtained by solving

  $$\tau(\mathbf{y}, \mathbf{X}, \theta) = c_\alpha$$

  for θ and result, so to speak, from "‘inverting"’ the test statistic $\tau(\mathbf{y}, \mathbf{X}, \theta_{H_0})$.

- The length and bounds of confidence intervals are random, as they depend on the sample $\mathbf{y}, \mathbf{X}$.

- The **coverage probability** indicates the probability of drawing a sample and calculating a confidence interval based on it, which contains the true parameter θ_0.

- **If a sample is already available**, then the true parameter θ_0 is either contained in the confidence interval calculated on the basis of the observed sample or not. In other words, it makes no sense to speak of a coverage probability with regard to the sample in question in the case of a **already available sample**.

- **Exact confidence intervals** cover the true parameter θ_0 with a coverage probability of $1-\alpha$. This is the case if the test statistic in (11.48) is pivotal.

- If the test statistic in (11.48) is not pivotal, but **asymptotically pivotal**, i. e. its **asymptotic distribution** is known for all DGPs under the null hypothesis and independent of the respective DGP in the model under consideration $\mathbb{M}$, then a **asymptotic confidence interval** is obtained.

- With **asymptotic** confidence intervals, the actual and nominal (chosen) coverage probabilities generally do not coincide. If several methods are available for calculating approximate confidence intervals, one should choose the one for which the difference between the actual and nominal coverage probability is as small as possible.

- If a parameter vector is considered instead of a parameter, **multidimensional confidence ellipsoids** are obtained, see section 11.6.2.
- **Asymptotic confidence interval for β_j in the multiple linear regression model** based on the χ^2-statistic
$$\tau(\mathbf{y},\mathbf{X},\beta_{j,H_0}) = \left(\frac{\hat\beta_j - \beta_{j,H_0}}{s_{\hat\beta_j}}\right)^2$$
with
$$s_{\hat\beta_j} = s(\mathbf{x}_j^T\mathbf{M}_{-j}\mathbf{x}_j)^{-1/2},$$
where $\mathbf{M}_{-j} = \mathbf{I} - \mathbf{X}_{-j}\left(\mathbf{X}_{-j}^T\mathbf{X}_{-j}\right)^{-1}\mathbf{X}_{-j}^T$ and $\mathbf{X}_{-j}$ contains all regressors except the j-th regressor.
  - The bounds of the confidence interval result from
$$\left(\frac{\hat\beta_j - \beta_{j,H_0}}{s_{\hat\beta_j}}\right)^2 = c_\alpha = q_{1-\alpha}$$
(as above by solving for β_{j,H_0}) as
$$[\hat\beta_j - s_{\hat\beta_j}c_\alpha^{1/2}\,,\,\hat\beta_j + s_{\hat\beta_j}c_\alpha^{1/2}].$$
  - For $\alpha = 0.05$, we get $c_\alpha^{1/2} = \sqrt{3.84} = 1.96 = z_{1-\alpha/2}$ for the $(1-\alpha)$ quantile $c_\alpha = q_{1-\alpha}$ of the χ^2-distribution, where z_β denotes the β quantile of the standard normal distribution.
  - This interval is identical to the interval obtained on the basis of the t-statistic if its asymptotic standard normal distribution is taken into consideration.
  - Asymmetric confidence intervals are possible on the basis of the t-statistic, for example. When do you want an asymmetric confidence interval?
- An **exact confidence interval for β_j in the normal linear model** is determined on the basis of the t-statistic and the t-distribution with $n-k$ degrees of freedom:
$$P\left(t_{\alpha/2}(n-k) \le \frac{\hat\beta_j - \beta_{j,H_0}}{s_{\hat\beta_j}} \le t_{1-\alpha/2}(n-k)\right) = 1-\alpha$$
provides
$$[\hat\beta_j - s_{\hat\beta_j}t_{1-\alpha/2}(n-k), \hat\beta_j - s_{\hat\beta_j}t_{\alpha/2}(n-k)]$$
resp.
$$[\hat\beta_j - s_{\hat\beta_j}t_{1-\alpha/2}(n-k), \hat\beta_j + s_{\hat\beta_j}t_{1-\alpha/2}(n-k)].$$

- **Relationship t-test and confidence interval**: Since a two-tailed t-test corresponds to an F-test if the t-statistic is squared, it follows from the construction of confidence intervals carried out here that the null hypothesis of a two-tailed t-test with significance level α cannot be rejected if and only if the null hypothesis lies within the confidence interval with confidence level $1 - \alpha$.

- **Bootstrap confidence intervals**
  - Calculation of the critical values by bootstrap, see section 11.5.2.
  - Important: Compared to an asymptotic confidence interval, a bootstrap confidence interval can only converge faster towards the exact confidence interval if the associated asymptotic distribution of the test statistic is pivotal!
  - There are various methods for carrying out the bootstrap.

    There are differences with regard to

    * the estimation method for the parameters $(\boldsymbol{\beta}, \sigma_0)$ of the DGP,
    * the bootstrap procedure for drawing the errors,
    * the choice of the t-statistic or the F-statistic as the basis.
  - If the t-statistic is used, the boostrap distribution is often asymmetric and the bounds of the confidence interval must be determined carefully, see Davidson & MacKinnon (2004, Section 5.3).
  - Confidence intervals based on the t-statistic are often referred to as **studentized bootstrap** confidence interval or as **percentile-t** or **bootstrap-t** confidence interval.

11.6.2. Confidence ellipsoids

- If (11.39) holds and $\mathbf{R} = \mathbf{I}_k$ is chosen, the boundary of the approximate confidence ellipsoid results from

$$\tau(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}_0) = kF_n = c_\alpha = q_{1-\alpha}.$$

- If the normal distribution applies exactly to the OLS estimators, then exact confidence ellipsoids can also be determined on the basis of the F-statistic and the corresponding critical value from the F-distribution with q and $n - k$ degrees of freedom.

- It can happen that a parameter vector $\boldsymbol{\beta}$ lies in a confidence ellipsoid, but not in the individual confidence intervals for the individual elements of $\boldsymbol{\beta}$ and vice versa (please verify graphically!). The reason for this is generally a strong collinearity between the individual parameter estimators. Cf. discussion in bachelor course **Introduction to Econometrics**.

- **Confidence ellipse**: two-dimensional confidence ellipsoid, example in section 11.7.

- Confidence ellipsoids can be calculated using the bootstrap method as in the one-dimensional case.

11.7. Empirical analysis of trade flows: Part 3

Continuation of **Empirical analysis of trade flows: Part 2** in section 10.3.

Repeat the estimation of the model 4 selected in step II.3 (based on the AIC),

$$\begin{aligned} \ln(Imports_i) = \beta_1 + \beta_2 \ln(GDP_i) + \beta_3 \ln(Distance_i) \\ + \beta_4 \; Openness_i + \beta_5 \ln(Area) + u_i. \end{aligned} \tag{11.49}$$

`R` code see section 10.3

Output:

```
Call:
lm(formula = mod_formula)

Residuals:
Min      1Q  Median      3Q     Max
-2.1825 -0.6344  0.1613  0.6301  1.5243

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.42778    2.13258   1.138   0.2611
log(wdi_gdpusdcr_o)  1.02502    0.07654  13.392  < 2e-16 ***
log(cepii_dist)     -0.88865    0.15614  -5.691 9.57e-07 ***
ebrd_tfes_o          0.35315    0.20642   1.711   0.0942 .
log(cepii_area_o)   -0.15103    0.08523  -1.772   0.0833 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.853 on 44 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.9062,     Adjusted R-squared: 0.8976
F-statistic: 106.2 on 4 and 44 DF,  p-value: < 2.2e-16
```

4. **Check the selected model** (**Part 1**):
   - Test the underlying model assumptions: Either
     - **(B1)**, **(B3)**, **(B4)** (cf. section 11.3), so that **exact** tests can be performed, or
     - **(B1)**, **(B2)**, **(B3)**, **(A1)** and **(A3)**, so that **asymptotic** tests can be performed.
   - Example of an assumption check: Does the assumption of **homoscedastically distributed errors** **(B2b)**, which is also an assumption for **(B4)**, hold?

     Plot of the residuals against the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ with

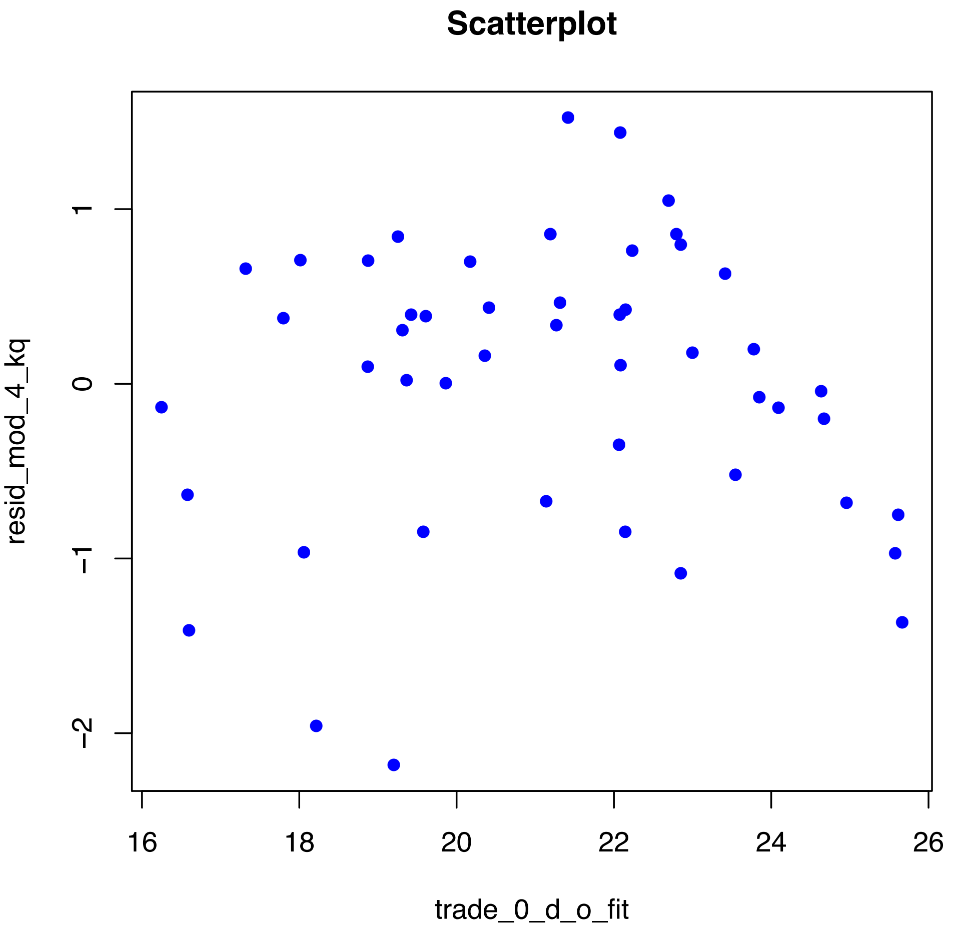
     R code (Extract from R program in section A.4)

```
                                      test=c("Chisq"))
F_stat

################################################################################
# Section 11.7 Empirical analysis of trade flows
################################################################################

# Model 4 was calculated in section 10.3
```

Listing 11.2: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

or with

```
                    plot(mod_4_kq,which=1)

```



Note:

- Under the assumptions mentioned above, residuals are consistent estimators of the errors, i. e.

$$\operatorname*{plim}_{n\to\infty} \hat{u}_t = u_t,$$

so that in large samples, consideration of the residuals comes close to consideration of the unknown errors.

- $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}_{\mathbf{X}}\mathbf{y}$ lies in the subspace of $\mathbf{X}$.

Therefore, a dependence of the dispersion of the residuals $\hat{u}_t$ on $\hat{y}_t$ indicates that the distribution of the errors depends on one or more regressors.

The cause may be

- a violation of the assumption of homoscedastic errors or
- a misspecified regression function.

Visually difficult to say in this case, therefore better: use heteroscedasticity tests, see section 15.2, or tests for correct specification of the functional form, see section 15.3.

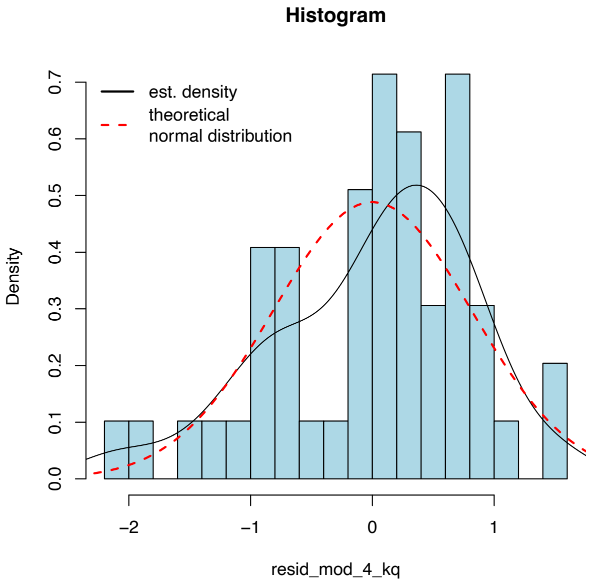
- Check for a possible violation of the **assumption of normally distributed errors (B4)**.
  - Plot of a **histogram** and an estimated density of the residuals as well as a normal distribution density with corresponding variance and calculation of various key figures with

    R code (Extract from R program in section A.4)

```
trade_0_d_o_fit <- mod_4_kq$fitted      # Fitted values of model 4

# Plot of residuals vs. fitted values
if (save.pdf) pdf("plot_fits_vs_resids_mod_4.pdf", 6, 6)
plot(trade_0_d_o_fit, resid_mod_4_kq, col = "blue", pch = 16, main = "Scatterplot")
if (save.pdf) dev.off()

# Plot of the histogram of the residuals
if (save.pdf) pdf("plot_hist_resids_mod_4.pdf", 6, 6)
hist(resid_mod_4_kq, breaks = 20, col = "lightblue", prob = T, main = "Histogram")
    # Estimated density of the residuals
lines(density(resid_mod_4_kq),col = "black", prob = T, add="T")
    # Plot the corresponding theoretical normal distribution
curve(dnorm(x, mean = mean(resid_mod_4_kq), sd = sd(resid_mod_4_kq)),
```

Listing 11.3: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

  - Calculation of statistical measures of the residuals, including a **normality test** ((Lomnicki-)Jarque-Bera test, see section 15.4). The (Lomnicki-)Jarque-Bera test can be calculated with `R` command `jarque.test(model_kq)`(requires `R` package `moments`).



Mean	-1.304583e-17
Median	1.612610e-01
Maximum	1.524291e+00
Minimum	-2.182553e+00
Std. Dev.	8.167224e-01
Skewness	-6.341491e-01
Kurtosis	3.084715e+00
Jarque Bera	3.298837e+00
Probability	1.921616e-01

The smaller the p-value of the normality test, the more likely one can expect the approximation error of the asymptotic normal distribution to be small for strictly exogenous regressors.

p-value of the (Lomnicki-)Jarque-Bera test contradicts visual impression: Assumption of normally distributed errors is not rejected because p-value is too large.

- even better than histogram: Plot of an estimated **density** and comparison with the density of the normal distribution with the estimated error variance.

See section 15.7 for continuation of the model checking **Checking of the selected model (part 2)**.

5. **Using the checked model: confidence intervals and performing tests**:

   - **Confidence intervals**

     – Choice of a **confidence level** $1 - \alpha$, in the following 95%.

     – Calculation of the confidence intervals of all estimated regression parameters with

     R code (Extract from R program in section A.4)

```
legend("topleft", c("est. density","theoretical\nnormal distribution"),
       col = c("black","red"), lwd = 2, lty = c(1,2), bty = "n")
```

Listing 11.4: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides:

```
2.5 %      97.5 %
(Intercept)        -1.87014867  6.72570228
log(wdi_gdpusdcr_o)  0.87076944  1.17927579
log(cepii_dist)     -1.20331827 -0.57397436
ebrd_tfes_o         -0.06286079  0.76916951
```

```
                        log(cepii_area_o)   -0.32280233  0.02074077

```

- **Two-tailed test**
  - Statistical **pair of hypotheses**:

    H_0 : The GDP elasticity of imports is 1. versus H_1 : The elasticity is not equal to 1.

    $$\mathrm{H}_0 : \beta_2 = 1 \quad \text{versus} \quad \mathrm{H}_1 : \beta_2 \neq 1.$$

  - Choose **significance level**, e. g. $\alpha = 0.05$.

    Calculation of the (approximate) **critical values**: $n - k = 49 - 5 = 44$ degrees of freedom. Since the t-statistic is exactly t-distributed under strict assumptions, but under weaker assumptions the t-distribution is a good approximation, the (approximate) critical values are determined on the basis of the t-distribution:

    R code (Extract from R program in section A.4)

```

#### Confidence intervals
confint(mod_4_kq)
```

Listing 11.5: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

$$-c = -2.015368, \quad c = 2.015368$$

  - Calculate ***t*-statistic** using the appropriate line of the output

```
                        Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
                        log(wdi_gdpusdcr_o)  1.02502    0.07654  13.392  < 2e-16 ***

```

    with R command

    R code (Extract from R program in section A.4)

```
# Two-tailed test
    # Determining the critical values
```

Listing 11.6: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{s_{\hat{\beta}_2}} = \frac{1.02502 - 1}{0.7654} = 0.3269286$$

  - **Decision**: Since test statistic

$$\begin{aligned} -c < & t_{\beta_2} < c \\ -2.015368 < & 0.3269286 < 2.015368 \end{aligned}$$

    is outside the (approximate) critical range, do not reject the null hypothesis.

– (Approximate) p-**value** is 0.7452378, calculated with

R code (Extract from R program in section A.4)

```
qt(alpha/2,mod_4_kq$df)
qt(1-alpha/2,mod_4_kq$df)
```

Listing 11.7: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

Since p-value is greater than the chosen significance level, H_0 cannot be rejected (even if a significance level of 10% is chosen).

The p-value states that under H_0 in about 75 out of 100 samples a t-statistic would be obtained whose absolute value is at least 0.33.

– If there is already an (approximate) **confidence interval** for β_2 with confidence level $1 - \alpha$: If the value of the null hypothesis is within the confidence interval, H_0 is not rejected.

An alternative that is faster: Using the `R` command `linearHypothesis` (requires `R` package `car`).

Note: it calculates $F = t^2$, p-values based on the $F_{1,n-k}$-distribution

R code (Extract from R program in section A.4)

```
    # p-value
2*pt(-abs(t),mod_4_kq$df)
```

Listing 11.8: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

```
Linear hypothesis test

Hypothesis:
log(wdi_gdpusdcr_o) = 1

Model 1: restricted model
Model 2: log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes_o +
log(cepii_area_o)

Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 32.095
2     44 32.018  1  0.077776 0.1069 0.7453
```

- **One-tailed test**

  – With regard to the model (11.49), a hypothesis can also be made regarding a negative impact of *distance* on *imports*. Since evidence for $\beta_3 < 0$ is obtained by statistically rejecting $\beta_3 \geq 0$, the **pair of hypotheses** is

$$\mathrm{H}_0 : \beta_3 \geq 0 \quad \text{versus} \quad \mathrm{H}_1 : \beta_3 < 0.$$

  – Choose a **significance level** of $\alpha = 0.05$ and calculate the (approximate) **critical value**. Note that only the **left** critical value is required, as the parameter range of the alternative hypothesis is to the left of the parameter range of the null hypothesis and thus the critical range is also to the left of the non-rejection range:

R code (Extract from R program in section A.4)

```
    # download.packages("car", destdir="C:/Program Files/R/R-2.15.1/library")
    # install.packages("car")
```

Listing 11.9: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

returns -1.68023.

– The ***t*-statistic** is contained in the R output:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
log(cepii_dist)     -0.88865    0.15614  -5.691 9.57e-07 ***
```

or in this case results as follows:

R-Code (Extract from R program in section A.4)

```
(F_stat              <- linearHypothesis(mod_4_kq,c("log(wdi_gdpusdcr_o)=1")))
```

Listing 11.10: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

$$t_{\beta_3} = -5.691.$$

– **Decision**: Because of

$$t_{\beta_3} = -5.691 < -1.68023 = c,$$

rejection of the null hypothesis

– ***p*-value**:

R code (Extract from R program in section A.4)

```
    # Critical values
alpha            <-0.05
```

Listing 11.11: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

results in 4.783876e-07. The distance therefore has a negative impact even at the 1% significance level.

– **Interpretation**: If the distance increases by 1%, then ceteris paribus the expected imports to Germany fall by approx. 0.9%.

- **F-test: testing joint hypotheses**:

– **Question**: In the model (11.49), the parameters of the variables openness and area are not statistically significant for the chosen significance level of 5%. However, is it possible that both parameters are statistically **jointly** significant? The **pair of hypotheses** is:

$$\begin{aligned} &H_0 : \beta_4 = 0 \quad \text{and} \quad \beta_5 = 0 \quad \text{versus} \\ &H_1 : \beta_4 \neq 0 \quad \text{and/or} \quad \beta_5 \neq 0. \end{aligned}$$

- Choice of the **significance level**: $\alpha = 0.05$ and the (approximate) **critical values**. The **critical range** lies to the right of the critical value.

  Using the F-statistic, determining the (approximate) critical value based on the $F_{2,44}$-distribution gives 3.209278 with

  R-Code (Extract from R program in section A.4)

```

    # p-value
```

  Listing 11.12: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

  Using the χ^2-statistic, determining the (approximate) critical value based on the $\chi^2(2)$-distribution gives 5.991465 with

  R code (Extract from R program in section A.4)

```
(qf(1-alpha,2,mod_4_kq$df))
```

  Listing 11.13: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

- **Calculating the F-statistic** and the **p-value** is most easily done with the R command `linearHypothesis` (requires R package `car`):

  R code (Extract from R program in section A.4)

```

#### F-test, correlation matrix and confidence ellipses
```

  Listing 11.14: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

  and returns

```
                 Linear hypothesis test

                 Hypothesis:
                 ebrd_tfes_o = 0
                 log(cepii_area_o) = 0

                 Model 1: restricted model
                 Model 2: log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes
_o +

                 log(cepii_area_o)

                 Res.Df    RSS Df Sum of Sq      F   Pr(>F)
                 1     46 39.645
                 2     44 32.018  2    7.6272 5.2408 0.009088 **
                 ---
                 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Calculating the χ^2-statistic and the ***p*-value** is most easily done with the R command `linearHypothesis` (requires R package `car`):

R code (Extract from R program in section A.4)

```
(F2_stat       <- linearHypothesis(mod_4_kq,c("ebrd_tfes_o=0","log(cepii_area_o)=0"),
             test=c("F")))
```

Listing 11.15: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

and returns

```
                           Linear hypothesis test

                           Hypothesis:
                           ebrd_tfes_o = 0
                           log(cepii_area_o) = 0

                           Model 1: restricted model
                           Model 2: log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes
      _o +

                           log(cepii_area_o)

                           Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
                           1     46 39.645
                           2     44 32.018  2    7.6272 10.482   0.005296 **
                           ---
                           Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In both cases, the respective test statistic is in the critical range, so that the null hypothesis is rejected at the 5% significance level. Based on both *p*-values, it can be seen that the null hypothesis is also rejected at the 1% significance level.

– **Interpretation**: At least one of the two variables openness or logarithmised area has an effect on exports to Germany. One possible reason for the different test results of the individual tests and the joint tests is the correlation of 0.42 between the parameter estimators, see below.

- **Correlation matrix of the parameter estimates**

R code (Extract from R program in section A.4)

```
   # chi^2-statistic
(Chisq_stat     <- linearHypothesis(mod_4_kq,c("ebrd_tfes_o=0","log(cepii_area_o)=0"),
                           test=c("Chisq")))
```

Listing 11.16: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides

```
                      (Intercept) log(wdi_gdpusdcr_o) log(cepii_dist) ebrd_tfes_o log(cepii_area_o)
                      (Intercept)          1.00000000           -0.6077120      -0.71380469 -0.26879925
      0.08289662
                      log(wdi_gdpusdcr_o) -0.60771198            1.0000000       0.30644626 -0.41648145
     -0.62198317
```

```
                   log(cepii_dist)     -0.71380469          0.3064463      1.00000000  0.09807673
   -0.29518939
                   ebrd_tfes_o         -0.26879925         -0.4164814      0.09807673  1.00000000
    0.42127548
                   log(cepii_area_o)    0.08289662         -0.6219832     -0.29518939  0.42127548
    1.00000000
```

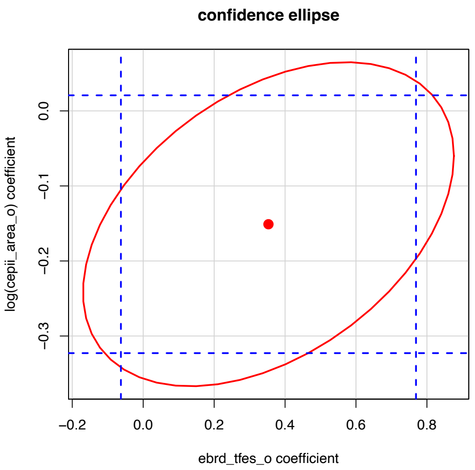
- **Confidence ellipse** for β_4 and β_5:
  - Choose confidence level, here 95%.
  - Choose two parameters, here β_4 and β_5
  - Use `R` command `confidenceEllipse()` (requires `R` package `car`):

    R code (Extract from R program in section A.4)

```
    # Covariance matrix
(cov_par          <- vcov(mod_4_kq))
    # Correlation matrix
(corr_par         <- cov2cor(cov_par))

#### Confidence ellipsoids

# Confidence ellipse
if (save.pdf) pdf("plot_conf_ellipse.pdf", 6, 6)
confidenceEllipse(mod_4_kq, which.coef = c(4, 5), levels = 0.95,
```

Listing 11.17: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides



12. Univariate time series models

This chapter mainly deals with models for univariate time series. Time series are samples in which the observations are available over time. A time series is univariate if there is exactly one variable over time.

> **Example:** The ifo Business Climate: Figure 12.1 shows the time series of the monthly indices of ifo business outlook, ifo business situation and ifo business climate for the period January 1991 to January 2023.

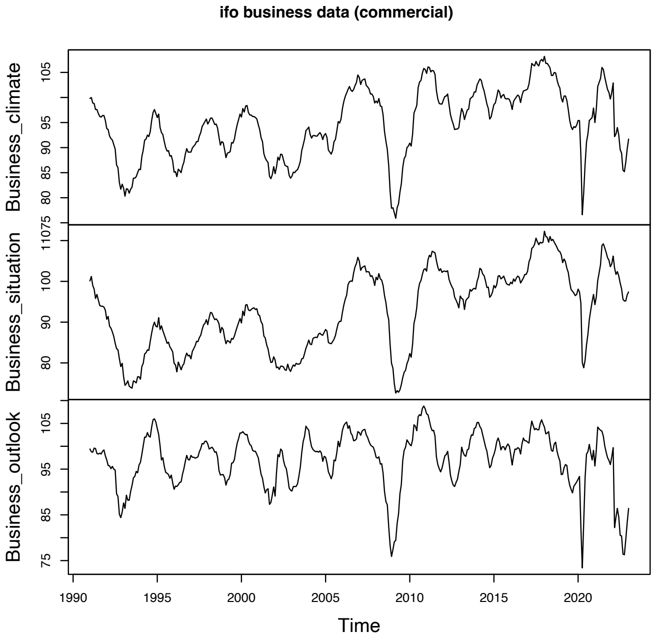


Figure 12.1.: The ifo business outlook, the ifo business situation and the ifo business climate for trade and industry (January 1991 to January 2023) (**R program**, see section A.9, page 351)

Time series are realisations of DGPs, which are then referred to as stochastic processes. The latter are considered in more detail below. Models for univariate time series are sets that contain stochastic processes.

12.1. Stochastic processes

Stochastic process

A stochastic process $\{y_t\}_{t\in\mathbb{T}}$ is a sequence of random variables

$$\{y_t|t \in \mathbb{T}\} \equiv \{y_t(\omega)|t \in \mathbb{T}, \omega \in \Omega\} \equiv \{y(t,\omega)|t \in \mathbb{T}, \omega \in \Omega\} : \quad \Omega \times \mathbb{T} \to \mathbb{R}^{\mathbb{T}}, \tag{12.1}$$

which are defined on a sample space Ω and a given index set $\mathbb{T}$ (Hassler (2007, Section 2.3), Mikosch (1998, Section 1.2)).

Remarks:

- Other terms: random process, random sequence (Davidson 2000, Section 4.1).
- Other common notations are: $\{y_t\}_{t\in\mathbb{T}}$ or without specifying the index set $\{y_t\}$.
- If the index t represents time, a stochastic process is also referred to as a **time series process**:
  - **Continuous-time processes**: $\mathbb{T}$ is an interval in $\mathbb{R}$.
  - **Discrete-time processes**: $\mathbb{T}$ is a finite or countably infinite set, typically $\mathbb{T} = \mathbb{Z}$ or $\mathbb{T} = \mathbb{N}$.

    In the case of discrete-time processes, a distinction is made between discrete-time stochastic processes with

    - **regular observation frequency**:

      Examples: monthly observations of the ifo business climate, annual GDP growth rates, weekly observations of the DAX.

    - **irregular observation frequency**:

      Example: Reuter's ticker data.
- Univariate and multivariate stochastic processes:
  - **Univariate stochastic process**: y_t is a scalar random variable

    Example: Observations of the ifo Business Climate Index.
  - **Multivariate stochastic process**: $\mathbf{y}_t$ is a random vector.

    Example: $\mathbf{y}_t = \begin{pmatrix} \text{ifo business climate}_t \\ \text{ifo business situation}_t \\ \text{ifo business outlook}_t \end{pmatrix}$.

  In this chapter, we consider almost exclusively discrete-time univariate stochastic processes with a regular observation frequency. Chapter 13 takes a closer look at models for multivariate stochastic processes.

- Literature on general **existence conditions** for stochastic processes is given in Hassler (2007, Section 2.3, footnote 9).
- Important: A stochastic process is a function of 2 variables:
  - For a given time period t_0,

$$y_{t_0} = y(t_0, \omega), \quad \omega \in \Omega,$$

is a random variable. The expected values $E[y_t] = \mu_t, t \in \mathbb{T}$, are called **ensemble averages**. Figure 12.2 shows different realisations for each time period t.

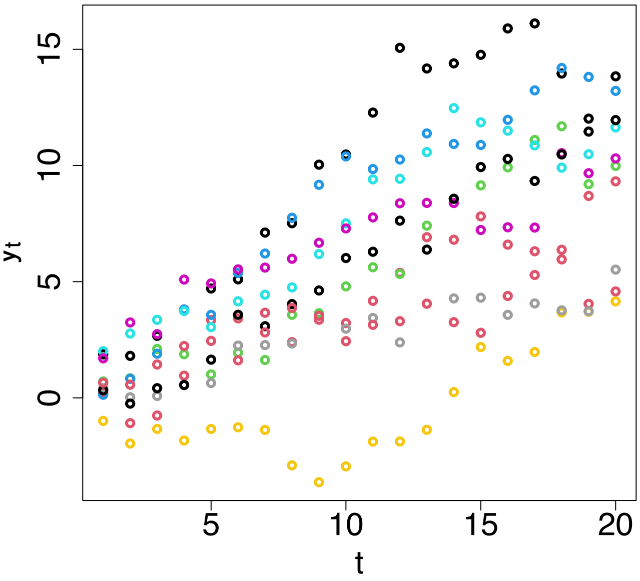


Figure 12.2.: Ten different realisations for each time period t of a stochastic process (**R program**, see section A.10, 352)

  - For a given elementary event ω_0,

$$y_t = y(t, \omega_0), \quad t \in \mathbb{T},$$

is a function of time.

The function is then called a **realisation**, a **trajectory** or a **path** of the stochastic process $\{y_t\}$. Figure 12.3 shows different trajectories of a stochastic process. Some authors refer exclusively to the realisation of a stochastic process as a time series (Hassler 2007, Section 2.3).

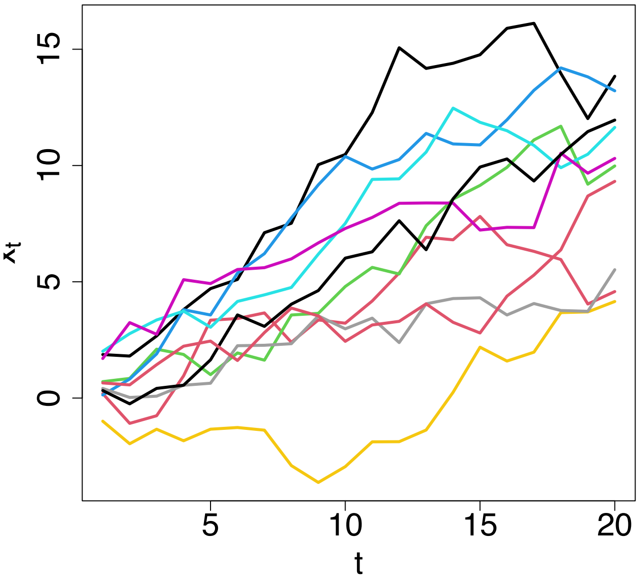


Figure 12.3.: Ten different trajectories of a stochastic process (**R program**, see section A.10, 352)

Summary: realisation

- of a random variable: number.
- of a stochastic process: trajectory, path: function of time t or a sequence of real numbers.

DGPs, joint and conditional densities for stochastic processes

- **Univariate stochastic processes**

  The DGP of a univariate stochastic process $\{y_t | t \in \mathbb{T}\}$, $\mathbb{T} = \{1, 2, \dots, T\}$ for T possible sample observations $(y_1, y_2, \dots, y_T)$ is fully determined by the joint density $f_{Y_1,Y_2,\dots,Y_T}(y_1, y_2, \dots, y_T)$, which in turn can be represented as a product of conditional densities (cf. for multivariate stochastic processes (5.2) in section 5.1):

$$f_{Y_1,Y_2,\dots,Y_n}(y_1, y_2, \dots, y_n) = \prod_{t=1}^{T} f_{Y_t|Y_{t-1},\dots,Y_1}(y_t | y_{t-1}, \dots, y_1). \tag{12.2}$$

- **Multivariate stochastic processes**

  (Cf. (5.2) in section 5.1):

$$f_{\mathbf{Y}_1,\mathbf{Y}_2,\dots,\mathbf{Y}_T}(\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_T)=\prod_{t=1}^{T} f_{\mathbf{Y}_t|\mathbf{Y}_{t-1},\dots,\mathbf{Y}_1}(\mathbf{y}_t|\mathbf{y}_{t-1},\dots,\mathbf{y}_1) \tag{5.2}$$

Fully and partially specified time series models

- If the conditional densities $f_{Y_t|Y_{t-1},\dots,Y_1}(y_t|y_{t-1},\dots,y_1)$ or $f_{\mathbf{Y}_t|\mathbf{Y}_{t-1},\dots,\mathbf{Y}_1}(\mathbf{y}_t|\mathbf{y}_{t-1},\dots,\mathbf{y}_1)$, $t=1,\dots,T$, are known, the DGP of the stochastic process is known.
- **Fully specified models**

  Time series models in which the conditional densities are modelled are **fully specified**.
- **Partially specified models**

  Often one is only interested in individual characteristics of the conditional densities, typically the conditional expected value or the conditional variance. In this case, it is generally sufficient to use models with partially specified stochastic processes:

  – in the univariate case $E[y_t|y_{t-1},\dots,y_1]$:

    * LLinear stochastic processes ⟶ section 12.2
    * Moving average processes ⟶ section 12.2
    * Autoregressive processes ⟶ section 12.3
    * Autoregressive integrated moving average processes ⟶ section 12.3
    * Nonlinear autoregressive processes ⟶ Examples in **Advanced Econometrics**
    * ...

  – in the multivariate case $E[\mathbf{y}_t|\mathbf{y}_{t-1},\dots,\mathbf{y}_1]$:

    * Vector autoregressive processes ⟶ **Quantitative Economic Research II**
    * Vector error correction processes ⟶ **Quantitative Economic Research II**
    * ...

  – in the univariate case $Var(y_t|y_{t-1},\dots,y_1)$:

    * Autoregressive Conditional Heteroskedastic Processes (ARCH processes) ⟶ **Applied Financial Econometrics**

∗ Generalized Autoregressive Conditional Heteroskedastic Processes (GARCH processes) ⟶ **Applied Financial Econometrics**

Motivation of the following definitions:

Can ensemble average μ_t be estimated?

- If R realisations $y_t^{(r)}$ were available for y_t, we could use the estimator of the expected value

$$\hat{\mu}_t = \frac{1}{R}\sum_{r=1}^{R} y_t^{(r)}.$$

  Problem: In practice, typically $R = 1$.

- **Central question**: Under what conditions can the ensemble average μ_t be estimated by the time average

$$\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t? \tag{12.3}$$

  Note: here $R = 1$.

- Response requires measures to quantify the stochastic dependencies between observations at different points in time as well as concepts that define a constancy of relevant properties of stochastic processes over time (including the ensemble average $\mu_t = \mu$). These are introduced below. The answers themselves can be found in section 12.4.1.

Measurement of the temporal stochastic dependencies

The following measures are typically used to measure the dependency structure characterised by the conditional densities (representation for univariate stochastic processes):

- Autocovariance function
- Autocorrelation function
- Partial autocorrelation function

Concepts that define a constancy of relevant properties of stochastic processes over time

- Mean stationarity
- Weak stationarity
- Strict stationarity

Autocovariance function

The autocovariance function of a stochastic process $\{y_t | t \in \mathbb{T}\}$ is defined for all $t, t-k \in \mathbb{T}$, k integer, as

$$\begin{aligned} Cov(y_t, y_{t-k}) &\equiv E\left[(y_t - E[y_t])(y_{t-k} - E[y_{t-k}])\right] = E[y_t y_{t-k}] - E[y_t]E[y_{t-k}] \\ &= E[y_t y_{t-k}] - \mu_t \mu_{t-k}. \end{aligned} \tag{12.4}$$

Note: from $E(y_t | y_{t-k}) = E(y_t)$ follows $Cov(y_t, y_{t-k}) = 0$.

Autocorrelation function

The autocorrelation function of a stochastic process $\{y_t | t \in \mathbb{T}\}$ is defined for all $t, t-k \in \mathbb{T}$, k integer, as

$$Corr\left(y_t, y_{t-k}\right) \equiv \frac{Cov\left(y_t, y_{t-k}\right)}{\sqrt{Var\left(y_t\right) Var\left(y_{t-k}\right)}}.$$

Partial autocorrelation function

The partial autocorrelation function specifies the conditional autocorrelation between y_t and y_{t-k}, whereby the condition is composed of all observations that lie between the periods t and $t-k$, i.e. $y_{t-1}, \ldots, y_{t-k+1}$,

$$Corr\left(y_t, y_{t-k} | y_{t-1}, \ldots, y_{t-k+1}\right).$$

An example of a partial autocorrelation function is (12.29) in section 12.3.3.

Stationarity concepts

Mean stationary process

$\{y_t\}$ is **mean stationary** if the following applies

$$\mu_t = \mu \quad \text{for all } t \in \mathbb{T}. \tag{12.5}$$

Nothing is assumed about the autocovariances.

Weakly stationary or covariance stationary process

A univariate stochastic process $\{y_t | t \in \mathbb{T}\}$ is referred to as **(weakly) stationary** or **covariance stationary** if the following properties are fulfilled with regard to the first two moments:

- $E[y_t] = \mu$ for all $t \in \mathbb{T}$,
- $Cov(y_t, y_{t-k}) = \gamma_k$, for all $t, t-k \in \mathbb{T}$,

i. e. the mean value does not depend on the time period t and the autocovariance function depends exclusively on lag k, but not on the time period t.

Inferences:

- Weakly stationary processes are homoscedastic, since for $k = 0$ it holds that $Var(y_t) = \gamma_0$.
- For weakly stationary processes, the following applies to the autocorrelation function $\rho_k \equiv Corr(y_t, y_{t-k})$:
$$\rho_k = \gamma_k/\gamma_0.$$

Strict stationarity

The definition is provided here for multivariate stochastic processes: A multivariate stochastic process $\{\mathbf{y}_t\}$ is called **strictly stationary**, if for any set of time indices $t_1 < t_2 < \cdots < t_m$ the joint probability distribution for $(\mathbf{y}_{t_1}, \mathbf{y}_{t_2}, \ldots, \mathbf{y}_{t_m})$ and the joint probability distribution for $(\mathbf{y}_{t_1+k}, \mathbf{y}_{t_2+k}, \ldots, \mathbf{y}_{t_m+k})$ are equal for any integer k.

Examples for strictly stationary univariate processes

- $$\begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix} \sim N\left(\underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{\boldsymbol{\mu}}, \underbrace{\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}}_{\boldsymbol{\Sigma}} \right)$$

  and

  $$\begin{pmatrix} y_{t+k} \\ y_{t+k-1} \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  have the same bivariate normal distribution for any $t, t-1, t+k, t+k-1 \in \mathbb{T}$.

- $\sharp$ Given the random variable z_1 with an arbitrary distribution. Then

  $$z_t = z_1, \quad t = 2, 3, \ldots,$$

  is a strictly stationary process, where all autocorrelations for $k \neq 0$ are one (Hayashi 2000, Example 2.2).

Stochastic processes without autocorrelations

White Noise

$u_t \sim$ WN(0,σ^2) means for all $t \in \mathbb{T}$:

- $E[u_t] = 0$,
- $Var(u_t) = E[u_t^2] = \sigma^2$,
- $Cov(u_t, u_{t-k}) = 0$ für $k \neq 0$.

The conditions mean that the unconditional mean value of u_t is zero for each period and there is no heteroscedasticity.

Note: *No* assumption is made about the distribution of u_t's, but only the first two moments are specified.

Independent white noise

A sequence of IID random variables is called an **IID process** or **independent white noise**:

$$u_t \sim IID(0, \sigma^2), \quad t \in \mathbb{T}.$$

I. e., it does not help to observe u_{t-k} in order to specify the probability that a realisation of u_t occurs in a certain interval.

Note: No assumption is made about the distribution of u_t's.

Gaussian White Noise

If you add a normal distribution assumption to the independent white noise, you get Gaussian white noise:

$$u_t \sim NID(0, \sigma^2), \quad t \in \mathbb{T},$$

resp.

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

Note: Independent white noise and Gaussian white noise are both strictly stationary.

`R` commands

Generating Gaussian white noise with `rnorm()`.

♯ Examples of white noise that are not independent white noise

- (Hayashi 2000, Example 2.4): Let the random variable w be uniformly distributed on $[0, 2\pi]$ and $u_t = cos(tw), t = 1, 2, \ldots$. This means that $E[u_t|u_{t-k}] \neq 0, t-k, k > 0$, since all u_t are affected by ω. There can be no independent white noise. Nevertheless, $E(u_t) = 0$, $Var(u_t) = 1/4$, $Cov(u_t, u_{t-k}) = 0, k \neq 0$, so that white noise is present.

- A simple process with conditional heteroscedasticity, namely an ARCH(1) process ⟶ **Applied Financial Econometrics**.

In the following, (partially specified) univariate linear stochastic processes that allow non-zero autocorrelations are covered.

Only a short introduction is given. Very good, detailed textbooks are Hamilton (1994), Kirchgässner et al. (2013), Neusser (2009) and the application-orientated description in Lütkepohl & Kraetzig (2008).

12.2. Linear stochastic processes and moving average processes

Linear process

A stochastic process $\{y_t\}_{t\in\mathbb{Z}}$ is called **linear process** if it has the following representation (Neusser (2006, Definition 2.4),Brockwell & Davis (1991, Section 11.1, p. 404))

$$y_t = \sum_{j=-\infty}^{\infty} \psi_j u_{t-j} \tag{12.6}$$

with the parameters $\psi_j \in \mathbb{R}$, $j \in \mathbb{Z}$, and

$$u_t \sim WN(0, \sigma^2), \tag{12.7a}$$

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \tag{12.7b}$$

Example: $y_t = \psi_0 u_t + \psi_{-1} u_{t+1} + \psi u_{t-1}$

Remarks:

- Without specifying the distribution of u_t, a linear stochastic process is only partially specified (cf. white noise).
- **Note: For infinite sums of random variables, swapping the expected value and summation is generally not possible**.

  A swapping of expected value and infinite sum is **only possible if** a suitably defined limit value exists for the infinite sum.

  The above condition (12.7b) is sufficient for the infinite sum of random variables $\sum_{j=-\infty}^{\infty} \psi_j u_{t-j}$ to converge to a well-defined limit value, which is denoted by y_t. The convergence occurs **with probability one** (Brockwell & Davis 1991, Proposition 3.3.1, p. 83).

  If (12.7b) is valid, the expected value and summation can be swapped. This condition can be weakened somewhat; see Appendix A.2 in **Advanced Econometrics**.
- Under all conditions mentioned, the linear process is weakly stationary.

- Note that in this general definition y_t may also be affected by future errors u_{t-j}, $j < 0$. If this is excluded, the result is a moving average process, see (12.8) below.

Lag operator

The lag operator defines an operation on an ordered set (e. g. a discrete stochastic process) on which it maps each element to the previous element

$$Ly_t \equiv y_{t-1}$$

with the following properties:

$$\begin{aligned} L^0 &= 1 \\ L^2 y_t &= L(Ly_t) = Ly_{t-1} = y_{t-2} \\ L^{-1} &= y_{t+1} \\ Lc &= c \\ L^m L^n y_t &= y_{t-m-n}. \end{aligned}$$

Lag polynomial, filter

The lag polynomial is a linear combination of different powers of lag operators with integer exponents

$$\Psi(L) = \ldots + \psi_{-2}L^{-2} + \psi_{-1}L^{-1} + \psi_0 + \psi_1 L + \psi_2 L^2 + \ldots$$

and is called a **linear filter** (Neusser 2006, Definition 2.4).

Moving Average process of order ∞ (MA(∞) process)

A moving average process $\{y_t\}_{t\in\mathbb{Z}}$ of order ∞ is a linear process with $\psi_j = 0$ for all negative j and $\sum_{j=0}^{\infty} |\psi_j| < \infty$

$$\begin{aligned} y_t &= \sum_{j=0}^{\infty} \psi_j u_{t-j}, \quad \psi_0 = 1, \\ &= (1 + \psi_1 L + \psi_2 L^2 + \cdots)u_t = \Psi(L)u_t. \end{aligned} \tag{12.8}$$

u_t is often referred to as **shock**, **innovation** or **error**.
The term $\Psi(L) \equiv (1 + \psi_1 L + \psi_2 L^2 + \cdots)$ is called **MA(∞) polynomial**.

Remarks: In a MA(∞) process, the future has no impact on the present.

Properties

- Mean value: $E[y_t] = E[\sum_{j=0}^{\infty} \psi_j u_{t-j}] = \sum_{j=0}^{\infty} \psi_j E[u_{t-j}] = 0$. It is possible to swap the expected value and infinite sum, since $\sum_{j=0}^{\infty} |\psi_j| < \infty$, so that (12.7b) holds.

- (Auto)covariance function::

$$\begin{aligned}
Cov(y_t, y_{t-k}) &= E[y_t y_{t-k}] \\
&= E\left[\left(\sum_{j=0}^{\infty}\psi_j u_{t-j}\right)\left(\sum_{l=0}^{\infty}\psi_l u_{t-k-l}\right)\right] \\
&= \sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\psi_j\psi_l E[u_{t-j}u_{t-k-l}] \\
&= \sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\psi_j\psi_l \begin{cases}\sigma^2 & \text{if } t-j=t-k-l \\ 0 & \text{otherwise}\end{cases} \qquad \text{with } j=k+l \\
&= \sigma^2\sum_{l=0}^{\infty}\psi_{k+l}\psi_l = \gamma_k. \qquad (12.9)
\end{aligned}$$

- Variance

$$\gamma_0 = Var(y_t) = \sigma^2\sum_{j=0}^{\infty}\psi_j^2 < \infty. \qquad (12.10)$$

- ♯ Technical remark on the derivation of (12.9) and (12.10): It can be shown that under (12.7b) the limit value $\left(\sum_{j=0}^{\infty}\psi_j u_{t-j}\right)\left(\sum_{l=0}^{\infty}\psi_l u_{t-k-l}\right)$ is well-defined as $y_t y_{t-k}$, so that the swapping of expected value and summation is possible. Note that from $\sum_{j=0}^{\infty}|\psi_j| < \infty$ follows: $\sum_{j=0}^{\infty}\psi^2 < \infty$. The latter condition is a necessary condition for the existence of the variance in (12.10). It is referred to as quadratic summability.

Special cases for the practice:

MA(q) processes

$$y_t = \sum_{l=0}^{q}\psi_j u_{t-j} \qquad (12.11)$$

Properties as for MA(∞) processes except:

$$\gamma_k = 0, \quad |k| > q. \qquad (12.12)$$

`R` commands

- Generating MA(q) processes: using the command `filter` with the specification `sides=1,method="convolution"`.
- Calculating the theoretical autocorrelation function: with the command `ARMAacf`

Example: Generating a MA(2) process and theoretical autocorrelation function

- DGP: $y_t = u_t + \psi_1 u_{t-1} + 0.6u_{t-2}, \quad u_t \sim NID(0,4)$.
- Sample size (or length): $n = 1000$ **R program**, see section A.11, page 353 provides figure 12.4 and autocorrelations $\rho_0 = 1, \rho_1 = 0.64, \rho_2 = 0.3$ und $\rho_k = 0$ für $k = 3, \ldots$.

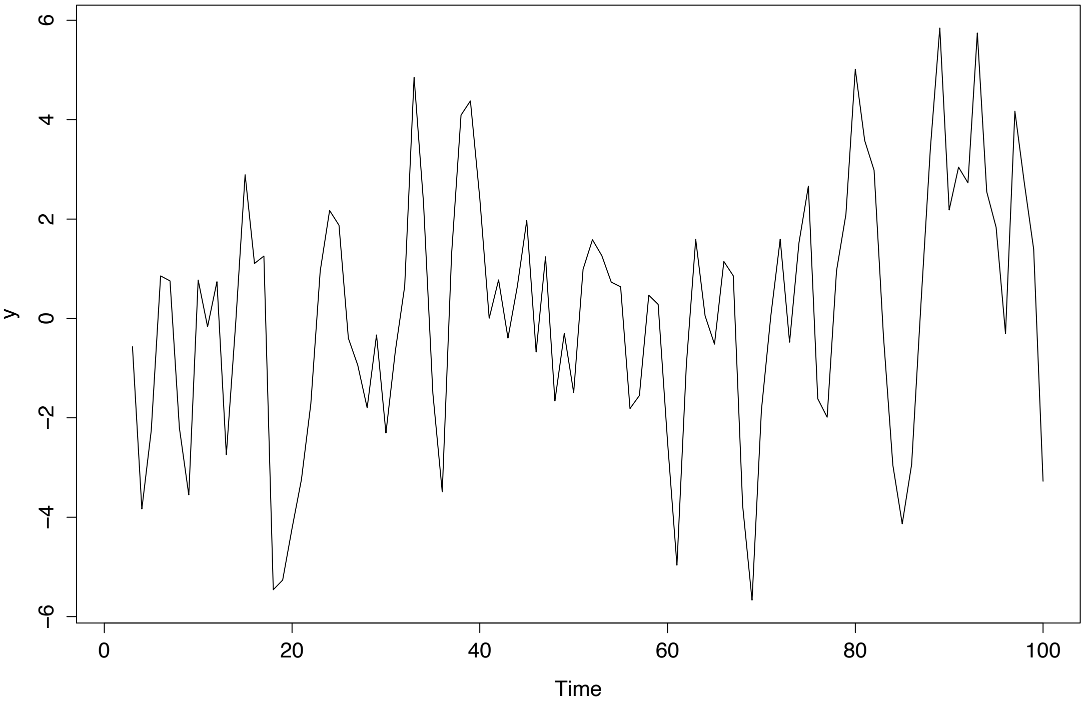


Figure 12.4.: A realisation for $n = 100$ of a MA(2) process with $\psi_1 = 0.8$ and $\psi_2 = 0.6$ and $\sigma^2 = 4$ (**R program**, see section A.11, page 353)

Problem with MA(q) processes: They cannot be estimated with OLS, but require maximum likelihood estimation methods, see **Advanced econometrics**. In contrast, the following autoregressive processes can be estimated with OLS.

12.3. Autoregressive processes

Autoregressive process of order p (AR(p) process)

A stochastic process $\{y_t\}$ is called **autoregressive process of order p (AR(p) process)** if it fulfils the following stochastic difference equation

$$y_t = \nu + \alpha_1 y_{t-1} + \cdots \alpha_p y_{t-p} + u_t, \quad t \in \mathbb{T}, \tag{12.13}$$

or

$$
\begin{aligned}
(1-\alpha_1 L - \cdots - \alpha_p L^p) y_t &= \nu + u_t, \\
\alpha(L) y_t &= \nu + u_t
\end{aligned}
$$

and $\{u_t\}$ is white noise. The term $\alpha(L) \equiv (1-\alpha_1 L - \cdots - \alpha_p L^p)$ is called an AR(p) polynomial.

Conditional expected value of AR(p) processes

The conditional expected value $E[y_t|y_{t-1},\ldots,y_1]$ can be easily calculated for AR(p) processes:

$$E[y_t|y_{t-1},\ldots,y_1] = \nu + \alpha_1 y_{t-1} + \cdots \alpha_p y_{t-p}. \tag{12.14}$$

This means that

$$E[y_t|y_{t-1},\ldots,y_1] = E[y_t|y_{t-1},\ldots,y_{t-p}]. \tag{12.15}$$

`R` commands

Generate AR(p) processes: using the command `filter` with the specification `method="recursive"`.

The derivation of the properties of AR(p) processes is more complex than for MA processes and is therefore omitted. However, the essential properties can be easily analysed using AR(1) processes.

12.3.1. AR(1) processes

Stochastic properties of an AR(1) process

$$y_t = \nu + \alpha_1 y_{t-1} + u_t, \quad u_t \sim WN(0,\sigma^2), \quad t \in \mathbb{T}, \tag{12.16}$$

- **Solution**: k times substitution returns:

$$
\begin{aligned}
y_t &= \nu + \nu\alpha_1 + \alpha_1\left(\alpha_1 y_{t-2} + u_{t-1}\right) + u_t = \cdots \\
&= \nu \sum_{k=0}^{j-1} \alpha_1^k + \alpha_1^j y_{t-j} + \sum_{k=0}^{j-1} \alpha_1^k u_{t-k}.
\end{aligned}
$$

For $j = t$

$$y_t = \nu \sum_{k=0}^{t-1} \alpha_1^k + \alpha_1^t y_0 + \sum_{k=0}^{t-1} \alpha_1^k u_{t-k}. \tag{12.17}$$

- **Stability properties**

  For (12.17), the following applies **to arbitrary** α_1 and $\nu = 0$, $\mathbb{T} = \mathbb{N}$, initial value y_0 and $j = t$:

  - **AR process explosive**, if $|\alpha_1| > 1$.
  - **Random walk with/without drift**, if $\alpha_1 = 1$:
    - **Random walk with drift**, if $\alpha_1 = 1, \nu \neq 0$:

$$y_t = \nu\, t + y_0 + \sum_{j=0}^{t-1} u_{t-j}. \tag{12.18}$$

    - **Random walk without drift**, if $\alpha_1 = 1, \nu = 0$:

$$y_t = y_0 + \sum_{j=0}^{t-1} u_{t-j}. \tag{12.19}$$

      Example: Figure 12.3 shows different realisations of a random walk.

  - **AR process stable**, if $|\alpha_1| < 1$.

  Stationary (and stable) AR(1) process: If $t \in \mathbb{Z}$ and $j \to \infty$ (process has been running indefinitely), the following applies for $|\alpha_1| < 1$

$$y_t = \nu/(1 - \alpha_1) + \sum_{j=0}^{\infty} \alpha_1^j u_{t-j}. \tag{12.20}$$

- **Unconditional expected value** $\mu_t \equiv E[y_t]$ for period t:

$$\mu_t = \begin{cases} \nu(1 + \alpha_1 + \ldots + \alpha_1^{t-1}) + \alpha_1^t E[y_0] & \text{if } t = 1, 2, \ldots \text{ — dependent on } t, \\ \nu/(1 - \alpha_1) & \text{if } |\alpha_1| < 1 \text{ and } t \in \mathbb{Z} \text{ — independent of } t, \\ t\nu + E[y_0] & \text{if } \alpha_1 = 1 \text{ and } t = 1, 2, \ldots \text{ — dependent on } t. \end{cases}$$

  For a given μ_t, the mean-adjusted autoregressive process is obtained

$$y_t - \mu_t = \alpha_1(y_{t-1} - \mu_{t-1}) + u_t.$$

- **Unconditional variance**:

$$\begin{aligned} Var(y_t) &= \alpha_1^2 Var(y_{t-1}) + \sigma^2 \\ &= \begin{cases} \sigma^2 \sum_{j=0}^{t-1} \alpha_1^{2j} & \text{if } Var(y_0) = 0 \text{ and } t = 1, 2, \ldots \text{ — dependent on } t, \\ \sigma^2/(1 - \alpha_1^2) & \text{if } |\alpha_1| < 1 \text{ and } t \in \mathbb{Z} \text{ — independent of } t, \\ \sigma^2 t & \text{if } \alpha_1 = 1,\ Var(y_0) = 0 \text{ and } t = 1, 2, \ldots \text{ — dependent on } t. \end{cases} \end{aligned}$$

- **Autocovariance function** $Cov(y_t, y_s) \equiv E[(y_t - \mu_t)(y_s - \mu_s)]$:

$$
\begin{aligned}
Cov(y_t, y_{t-k}) &= \alpha_1^k Var(y_{t-k}) \\
&= \begin{cases} \alpha_1^k \sigma^2 \sum_{j=0}^{t-1-k} \alpha_1^{2j} & \text{if } Var(y_0) = 0 \text{ and } t = 1, 2, \dots \text{ — dependent on } t, \\ \alpha_1^k \sigma^2/(1 - \alpha_1^2) & \text{if } |\alpha_1| < 1 \text{ and } t \in \mathbb{Z} \text{ — independent of } t, \\ (t-k)\sigma^2 & \text{if } \alpha_1 = 1,\ Var(y_0) = 0,\ t = 1, 2, \dots \text{ — dependent on } t. \end{cases}
\end{aligned}
$$

- **Weakly stationary AR(1) process** If $|\alpha_1| < 1$ and $t \in \mathbb{Z}$, an AR(1) process is weakly stationary, as the first two moments are independent of t:

$$
\begin{aligned}
E[y_t] &= \mu = \nu/(1 - \alpha_1) \\
Var(y_t) &= \gamma_0 = \sigma^2/(1 - \alpha_1^2) \\
Cov(y_t, y_{t-k} &= \gamma_k = \alpha_1^k \gamma_0
\end{aligned}
$$

- **Autocorrelation function** In the case of a weakly stationary AR(1) process, the following applies

$$
\rho_k = Corr(y_t, y_{t-k}) = \alpha_1^k \tag{12.21}
$$

- **Properties of a (weakly) stationary AR(1) process**: If $|\alpha_1| \neq 0$:
  - $\gamma_k \neq 0$ for all $k \in \mathbb{Z}$,
  - Autocovariances and autocorrelations converge exponentially fast to zero:

$$
\begin{aligned}
\gamma_k &= \alpha_1^k \gamma_0, \\
\rho_k &= \alpha_1^k
\end{aligned}
$$

  This means that the effect of shocks is 'forgotten' relatively quickly. We therefore also speak of **models with short memory**. In extreme contrast to this is the random walk. Here, there is a perfect memory, as the effect of a shock is never forgotten. Random walks are an example of **models with long memory**. See remark after equation (12.33).

  Example: Plot of the autocorrelation function (12.21) of an AR(1) process for $\alpha_1 = 0.8$ and $k = 1, \dots, 20$

```
ar1_acf <- ARMAacf(ar=0.8,lag.max=20)
plot(ar1_acf,ylab="Autocorrelations",xlab="Lag",cex=0.8,xlim=c(0,20))
```

  Example: Plot of a realisation of an AR(1) process

  Parameters of the DGP: $\nu = 1, \alpha_1 = 0.8, \sigma^2 = 4$ with $n = 500$. Figure 12.6 shows a realisation generated with **R program**, see section A.12, page 353.

- **(Asymptotic) stationarity**:

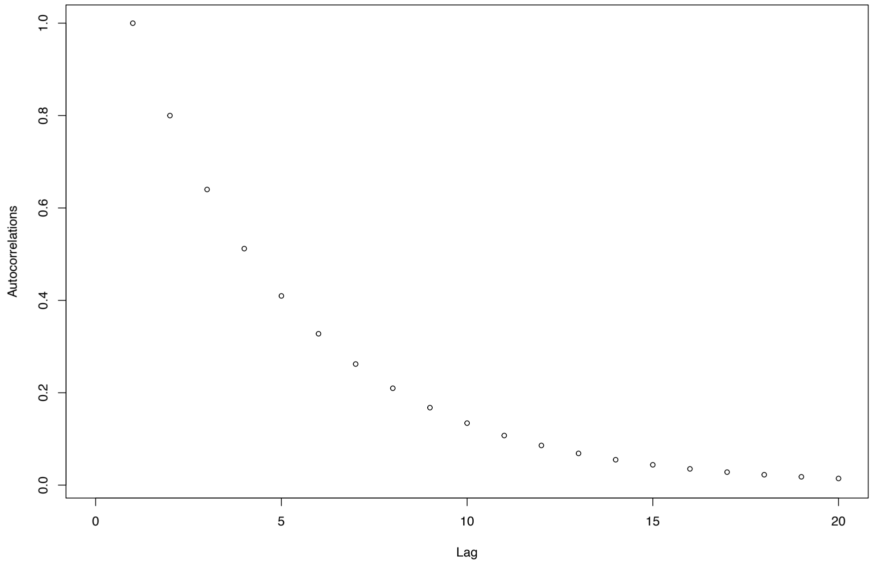


Figure 12.5.: Autocorrelation function of an AR(1) process with $\alpha_1 = 0.8$

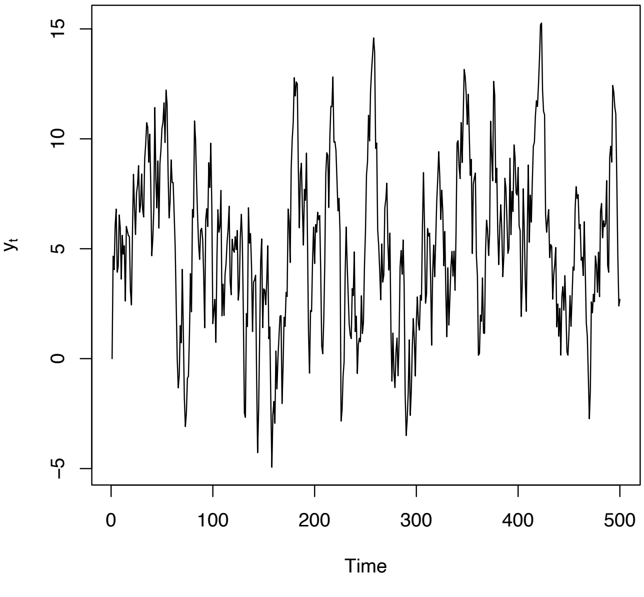


Figure 12.6.: Realisation of an AR(1) process with $\nu = 1, \alpha_1 = 0.8, \sigma^2 = 4$, $y_0 = 0$ and $n = 500$ (**R program** see section A.12, page 353)

- **If** $|\alpha_1| < 1$, the following applies

$$\lim_{t\to\infty} E(y_t) = \mu,$$
$$\lim_{t\to\infty} Cov(y_t, y_{t-k}) = \gamma_k,$$

and the AR(1) process is **asymptotically stationary**.

- Every stationary process is asymptotically stationary.
- If a process is not asymptotically stationary, it is **non-stationary**.
- What conditions are required for strict stationarity?

- **Invertible AR(1) process**: If $\psi_j = \alpha^j$ is defined, then the representation (12.20) with $\psi_0 = 1$ and $\psi_j = 0$, $j < 0$, can also be written as **MA(∞ process)** (12.8)

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j u_{t-j}, \quad t \in \mathbb{Z}. \tag{12.22}$$

  The AR(1) process is then called **invertible**.

- The AR(1) model is partially specified. In order to be able to generate realisations using a Monte Carlo (MC) study, additional assumptions must be made, e. g.:

  - an initial value $y_0 = 0$,
  - a parameter value $\alpha = 0.9$ and
  - a distribution for the errors $u_t \sim NID(0, 2)$. See section 2.9.1 for definition of NID.

  The DGP is now known and the following MC study can be carried out to check the bias of the OLS estimator.

12.3.2. Complex numbers

To motivate the need to study complex numbers, we examine the stability properties of an AR(2) process.

Stability properties of an AR(2) process

- Representation as a combination of two AR(1) processes

$$\begin{aligned}(1-\lambda_1 L)w_t &= u_t,\\ (1-\lambda_2 L)y_t &= w_t.\end{aligned}$$

  Prerequisite for linking:

  - $\{w_t\}$ weakly stationary: $|\lambda_1| < 1$,
  - $\{u_t\}$ White noise.

  Perform:

  1. Invert AR(1) process $\{w_t\}$:

$$w_t = \frac{1}{1-\lambda_1 L} u_t$$

2. and substitute w_t into the y_t equation:

$$\begin{aligned}(1-\lambda_2 L)y_t &= \frac{u_t}{(1-\lambda_1 L)}\\ (1-\lambda_2 L)(1-\lambda_1 L)y_t &= u_t\\ ((1\underbrace{-\lambda_1 L - \lambda_2 L}_{-\alpha_1 L}+\underbrace{\lambda_1\lambda_2 L^2}_{-\alpha_2 L^2})y_t &= u_t\\ y_t - \alpha_1 y_{t-1} - \alpha_2 y_{t-2} &= u_t\\ (1-\alpha_1 L - \alpha_2 L^2)y_t &= u_t\end{aligned}$$

with $\alpha_1 = \lambda_1 + \lambda_2, \quad \alpha_2 = -\lambda_1\lambda_2$.

Result: AR(2) process stationary if $|\lambda_1|, |\lambda_2| < 1$.

Such a decomposition exists for every stationary AR(p) process, but requires knowledge of complex numbers.

Literature on complex numbers: Neusser (2009, Appendix A), Hamilton (1994, Appendix A.2, S. 708-710)

- Motivation: $x^2 + 1 = 0 \iff x^2 = -1 \iff x = \pm\sqrt{-1}$ has **in $\mathbb{R}$ no solution**.
- Idea: Extend $\mathbb{R}$ with an imaginary unit ("'a second dimension"') $i \equiv \sqrt{-1}$ to be able to form roots of negative real numbers, i. e. define the

  $$\textbf{set of complex numbers } \mathbb{C} \equiv \mathbb{R}[i] \equiv \{z \equiv a + ib \mid a, b \in \mathbb{R}\}$$

  as the sum of a real and imaginary number.
  Then for any $a \in \mathbb{R}^+ : \sqrt{-a} = \sqrt{-1 \cdot a} = \sqrt{-1} \cdot \sqrt{a} = i\sqrt{a} \in \mathbb{C}$.
- Important for this course: What is the **modulus of a complex number** $z \in \mathbb{C}$: $||z||_{\mathbb{C}}$?

Conjugate and modulus of a complex number $z \in \mathbb{C}$:

- Complex conjugate of z: $\bar{z} \equiv a - ib$
- “length”, modulus, magnitude or absolute value of z : $||z||_{\mathbb{C}} \equiv r \equiv \sqrt{z\bar{z}} = \sqrt{a^2 + b^2}$
  This is often notated as $||z||_{\mathbb{C}}, ||z||$, or $|z|$ - note the difference to other norms

$$||z|| \equiv |z| \equiv \begin{cases} |z|_{\mathbb{R}} \equiv \begin{cases} z & z \geq 0 \\ -z & z < 0 \end{cases} & \text{if } z \in \mathbb{R} \\ |z|_{\mathbb{R}^n} \equiv \sqrt{z_1^2 + \ldots + z_n^2} & \text{if } z = (z_1, \ldots, z_n)^t \in \mathbb{R}^n \\ |z|_{\mathbb{C}} \equiv \sqrt{z\bar{z}} = \sqrt{a^2 + b^2} & \text{if } z = a + ib \in \mathbb{C} \end{cases}$$

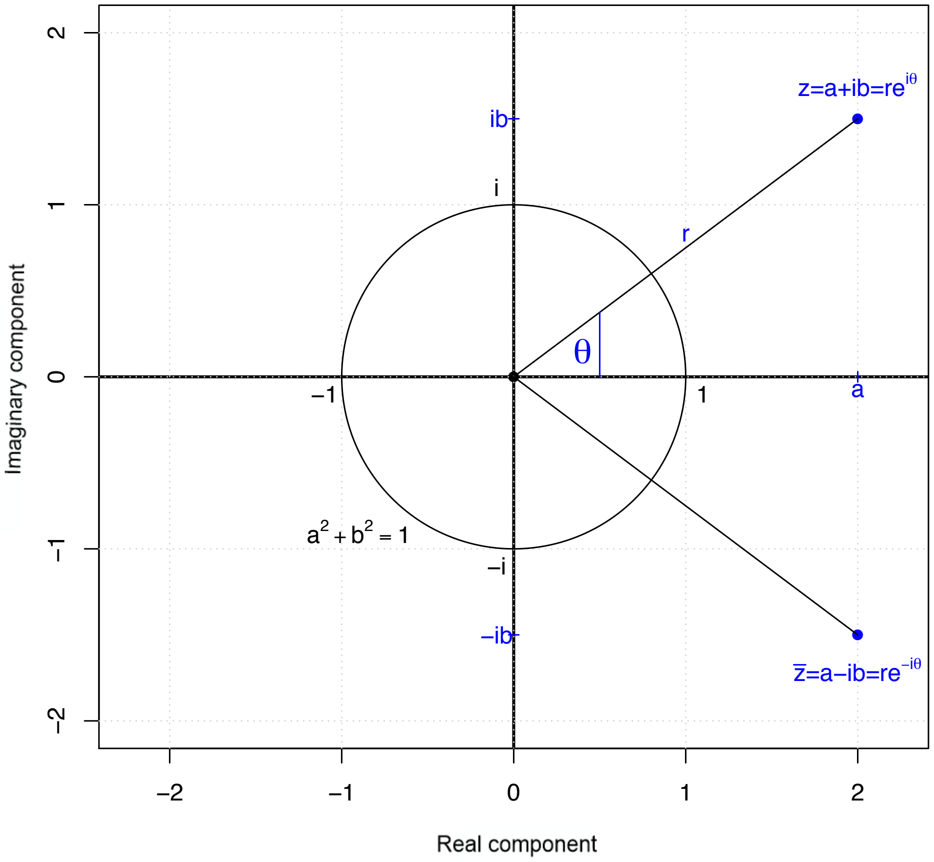
Representation of complex numbers $z \in \mathbb{C}$:

- In Cartesian coordinates: $z = \underbrace{a}_{\text{real}} + i \underbrace{b}_{\text{complex component}} = \mathrm{Re}(z) + i\,\mathrm{Im}(z)$

- In polar coordinates: $z = \underbrace{r}_{\text{length}} \cdot \underbrace{e^{i\theta}}_{\text{direction}} = r(\cos\theta + i\sin\theta)$

  Note: $||e^{i\theta}||_{\mathbb{C}} = \sqrt{e^{i\theta} \cdot e^{-i\theta}} = 1$

The following figure is based on Neusser (2009, Figure A.1, S. 260):



Calculation rules

- **Addition**: $(a + ib) + (c + id) = (a + c) + i(b + d)$
- **Subtraction**: $(a + ib) - (c + id) = (a - c) + i(b - d)$
- **Multiplication**:

$$(a + ib)(c + id) = (ac - bd) + i(ad + bc)$$

- **Division**:

$$\frac{a + ib}{c + id} = \frac{(ac + bd) + i(bc - ad)}{c^2 + d^2}$$

Stability condition for AR(p) processes

An AR(p) process with AR polynomial, $z \in \mathbb{C}$,

$$\begin{aligned}\alpha(z) &= (1 - \alpha_1 z - \cdots - \alpha_p z^p) \\ &= (1 - \lambda_1 z) \cdots (1 - \lambda_p z) \qquad (12.24)\end{aligned}$$

with

- eigenvalues $\lambda_1, \ldots, \lambda_p$ or
- roots $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$

is called stable,

- if all **eigenvalues** are less than one in absolute value

  $$|\lambda_i| < 1, \quad i = 1, 2, \ldots, p, \qquad (12.25)$$

  i. e. lie **within** the unit circle or

- if all **roots/zeros** z_i of the polynomial $\alpha(z)$, i.e. the **characteristic equation** of the AR(p) polynomial

  $$(1 - \alpha_1 z - \cdots - \alpha_p z^p) = 0$$

  lie **outside** the unit circle, i. e.

  $$|z_i| > 1, \quad i = 1, 2, \ldots, p, \qquad (12.26)$$

  applies.

R commands

Calculating the roots of an AR(p) polynomial: with `polyroot()`. Their absolute values can be determined with `abs()`.

Example: AR(2) process: The absolute values of the roots of the AR(2) polynomial

$$\alpha(L) = 1 - 0.1L - 0.9L^2$$

are $z = 1$ and $z = 1.111...$. The polynomial is therefore not stable.

```
abs(polyroot(c(1,-0.1,-0.9)))
```

Note the following property of the AR polynomial: $\alpha(1) = 1 - \alpha_1 \cdot 1 - \cdots - \alpha_p \cdot 1^p = 1 - \alpha_1 - \cdots - \alpha_p$.

Moments of a (weakly) stationary AR(p) process

- **Mean value/expected value**:

$$E[y_t] = \mu = \nu/\alpha(1) = \mu/(1 - \alpha_1 - \alpha_2 - \cdots - \alpha_p) \quad \text{for all } t. \tag{12.27}$$

- **Variance** and **autocovariance function**:

  The variance and the autocovariances of a weakly stationary AR(p) process are determined by the following **Yule-Walker equations** (cf. Hamilton (1994, p. 59, Eq. (3.4.36)))

$$\gamma_k = \begin{cases} \alpha_1\gamma_1 + \alpha_2\gamma_2 + \cdots + \alpha_p\gamma_p + \sigma^2 & \text{für } k = 0 \\ \alpha_1\gamma_{k-1} + \alpha_2\gamma_{k-2} + \cdots + \alpha_p\gamma_{k-p} & \text{für } k = 1, 2, \ldots \end{cases} \tag{12.28}$$

  It can be shown that the autocovariances of a stationary AR(p) process converge exponentially fast towards zero (Hamilton (1994, p. 59), Kirchgässner & Wolters (2008, Example 2.4)). See section 12.3.1 for the case of AR(1) processes.

- **Partial autocorrelation function**:

  For a weakly stationary AR(p) process, the following applies,

$$a_k = Corr(y_t, y_{t-k}|y_{t-1}, \ldots, y_{t-k+1}). \tag{12.29}$$

  I. e., all partial autocorrelations for $k > p$ are zero, since $a_k = \alpha_k = 0$ for $k > p$.

Invertibility of a stationary AR(p) process

A stationary AR(p) process can be represented as an MA(∞) process (12.8):

$$\begin{aligned} y_t - \mu &= \psi(L)u_t \qquad (12.30) \\ y_t - \mu &= u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + \ldots + \psi_i u_{t-i} + \ldots, \end{aligned}$$

where the coefficients of the lag polynomial are determined by the following equation:

$$\frac{1}{\alpha(L)} = \psi(L) \tag{12.31}$$

$$\alpha(L)\psi(L) = 1. \tag{12.32}$$

The parameters of $\psi(L)$ can be determined using method of equating the coefficients (Kirchgässner & Wolters 2008, Section 2.1.2)):

$$\phi_j = \sum_{i=1}^{j} \phi_{j-i}\alpha_i, \quad j = 1, 2, \ldots, \quad \nu = 1, \alpha_i = 0 \text{ for } i > p.$$

- **Modulus**: $r = |z| = z \cdot \bar{z} = (a+ib)(a-ib) = a^2 + b^2$
- $\cos\theta = a/r$
- $\sin\theta = b/r$
- **De Moivre's formula**:

$$\begin{aligned} z^n &= (re^{i\theta})^n \\ &= r^n e^{in\theta} \\ &= r^n(\cos n\theta + i \sin n\theta) \end{aligned}$$

Fundamental theorem of algebra

Every polynomial with coefficients $\phi_1, \ldots, \phi_p \in \mathbb{R}$

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \tag{12.23}$$

of degree $p \geq 1$ decomposes in the set (precisely: field) of complex numbers $\mathbb{C}$ exactly into p linear factors (thus has p complex zeros, roots $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$), whereby some roots can occur several times (see Neusser 2009, S. 261):

$$\Phi(z) = (1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_p z).$$

These roots can be real or complex and occur in the complex case as root pairs of conjugate roots. If there are c complex root pairs and r real roots, the following holds: $p = 2c + r$. The $\lambda_1, \ldots, \lambda_p$ are called **eigenvalues** of the polynomial $\Phi(z)$.

Example: The polynomial $\Phi(z) = z^3 - 2z^2 - 23z + 150$ has a root $\{z_1 = -6\}$ in $\mathbb{R}$, but decomposes in $\mathbb{C}$ into all individual components (linear factors) with the complex root pair $\{z_{2,3} = 4 \pm 3i\}$, so it can be written:

$$\Phi(z) = \underbrace{(z+6)\cdot(z^2 - 8z + 25)}_{\text{factorisation over } \mathbb{R}} = \underbrace{(z+6)\cdot(z - 4 - 3i)\cdot(z - 4 + 3i)}_{\text{linear factorisation over } \mathbb{C}}$$

$$= \left(1 - \underbrace{\frac{-1}{6}}_{\lambda_1} z\right) \cdot (1 - \underbrace{(0.16 - 0.12i)}_{\lambda_2} z) \cdot (1 - \underbrace{(0.16 + 0.12i)}_{\lambda_2} z)$$

The fundamental theorem of algebra makes it possible to analyse the stability properties of AR(p) processes.

12.3.3. AR(p) processes and more

Properties of AR(p) processes:

`R` commands

Calculating the MA parameters of an AR(p) process: with `ARMAtoMA()`.

Example: AR(2) process: Realisation, theoretical ACF and PACF as well as MA parameters of the inverted process

The parameters are: $\nu = 1$, $\alpha_1 = -0.5$, $\alpha_2 = -0.8$, $\sigma^2 = 4$ with $n = 500$.

The figure 12.7 is created with the **R program**, see section A.13, page 354.

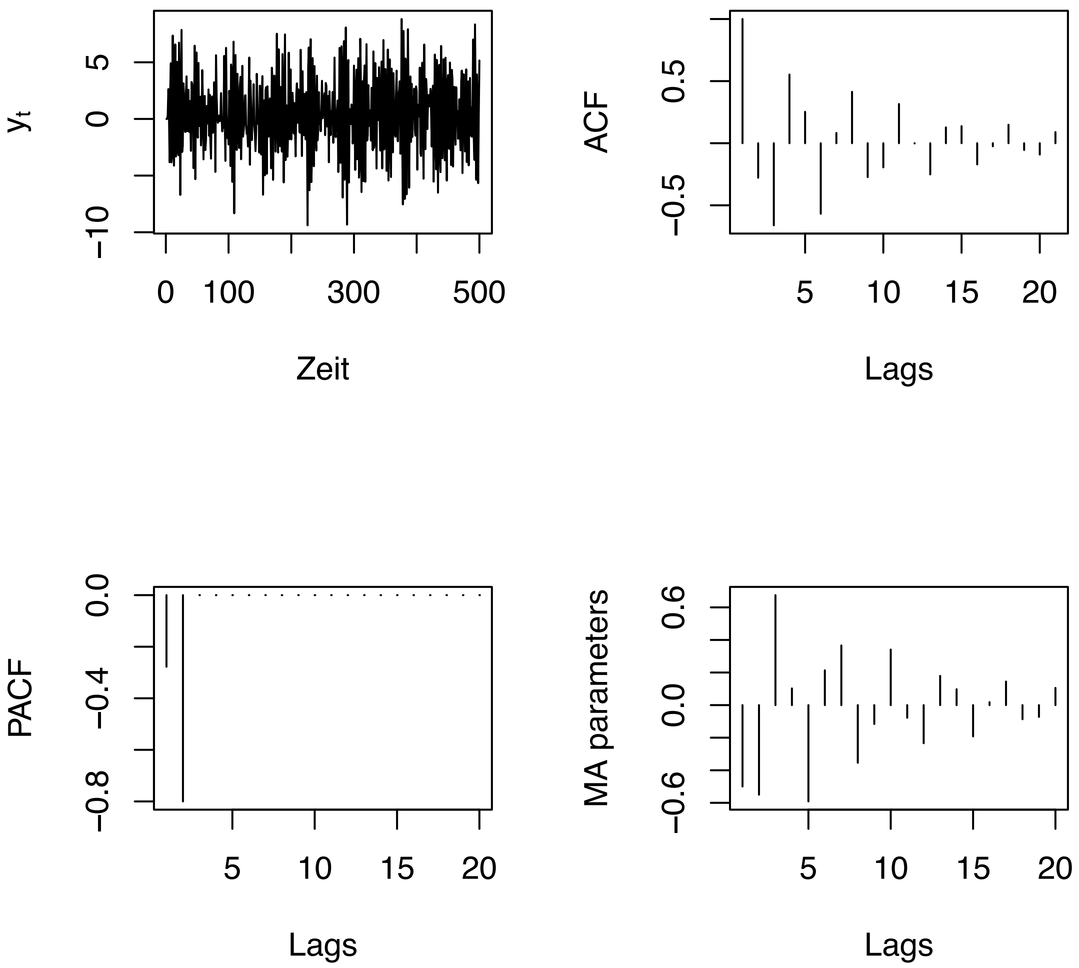


Figure 12.7.: Realisation, ACF, MA parameters, PACF of an AR(2) process with $\nu = 1, \alpha_1 = -0.5, \alpha_2 = -0.8, \sigma^2 = 4$, $y_0 = 0$ and $n = 500$

Non-stationary processes

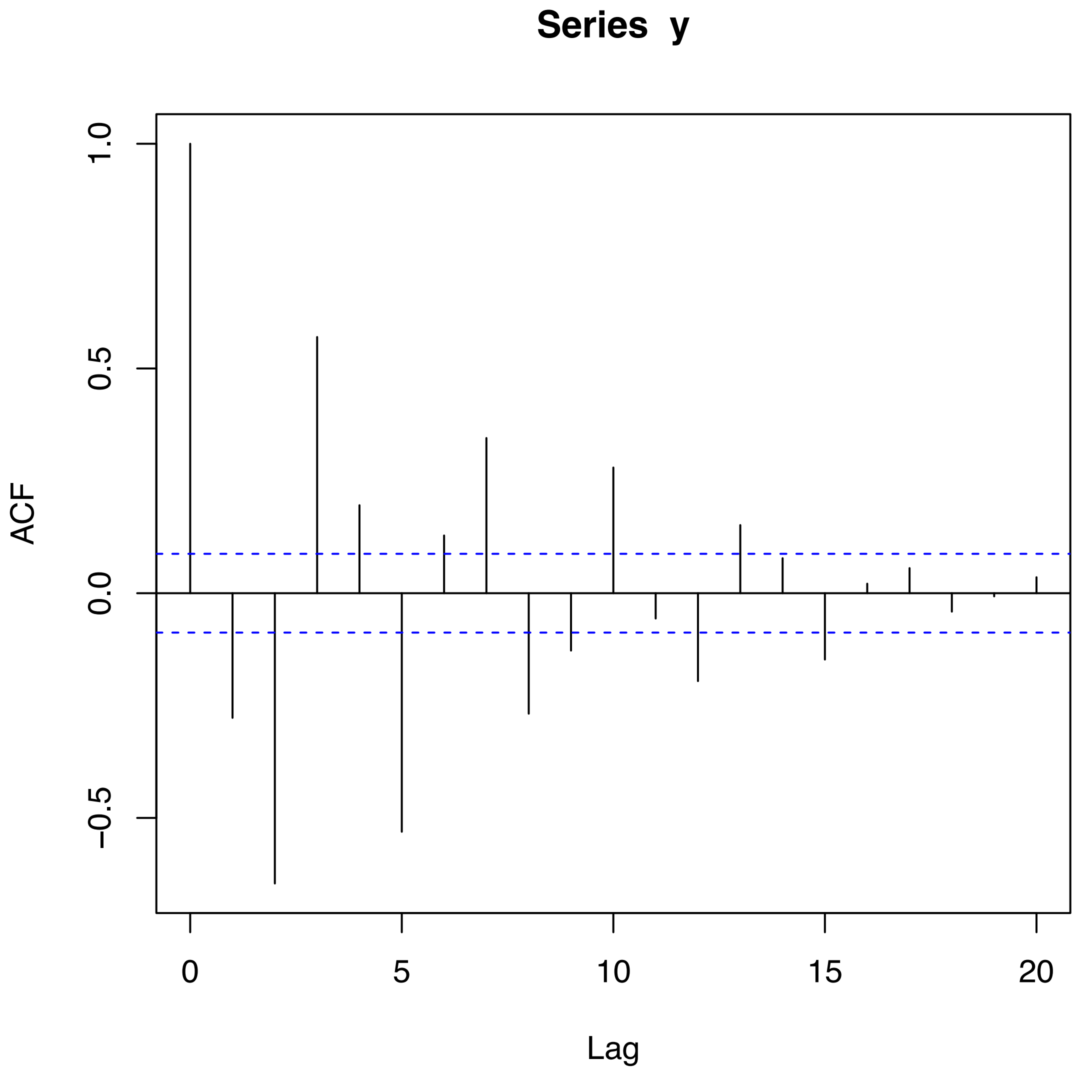


Figure 12.8.: Estimated autocorrelation function of a realisation of an AR(2) process with $\nu = 1, \alpha_1 = -0.5, \alpha_2 = -0.8, \sigma^2 = 4$, $y_0 = 0$ and $n = 500$

- Define the difference operator

$$\Delta y_t \equiv (1 - L)y_t = y_t - y_{t-1}$$

- The stability condition (12.25) is violated, for example, if the AR(p) polynomial $\alpha(z)$ can be decomposed as follows,

$$\begin{aligned} 1 - \alpha_1 z - \cdots - \alpha_p z^p &= (1 - z)\left(1 - \alpha_1^* z - \cdots - \alpha_{p-1}^* z^{p-1}\right) \\ &= (1 - z)\alpha^*(z) = \Delta\alpha^*(z), \end{aligned}$$

  where the AR($p - 1$) polynomial $\alpha^*(L)$ fulfils the stability condition (12.26). In this case, the AR(p) process contains a **random walk component**. This is also referred to as a process that is **integrated with order 1**, in short

$$y_t \sim I(1).$$

After applying the difference operator, a stable process of order 0 is obtained

$$\Delta y_t \sim I(0).$$

The random walk component is often referred to as **stochastic trend**, as it often causes trend-like trajectories, see e. g. figure 12.3.

Example: Figure 12.3 shows different realisations of a random walk.

- General: An AR(p) process $\{y_t\}$ is **integrated with order** d, in short

  $$y_t \sim I(d),$$

  if the following applies,

  $$\alpha(L) = (1-L)^d \alpha^*(L), \tag{12.33}$$

  where d is an integer and $\alpha^*(L)$ fulfils the stability condition. To stabilise an integrated process, the d-times application of the difference operator $(1-L)$ is therefore necessary.

- It is possible that the integration parameter d is a real number → Long memory models. ♯ For a German introduction, see e .g. Tschernig (1994, Chapter 3) and Robinson (2003) with relevant essays on long memory models.

- Autoregressive processes (and stochastic processes in general) can contain a **deterministic trend** or **time trend**. If such processes are stationary after the time trend has been removed, they are referred to as **trend-stationary**.

AR(p) processes are well suited for creating forecasts. These can be calculated as follows:

***h*-step forecast**

$$\begin{aligned}
y_{T+h|T} &\equiv E[y_{t+h}|y_t, \ldots] \\
y_{T+1|T} &= \alpha_1 y_T + \cdots + \alpha_p y_{T-p} \\
y_{T+2|T} &= \alpha_1 y_{T+1|T} + \cdots + \alpha_p y_{T+1-p} \\
&\vdots \\
y_{T+h|T} &= \alpha_1 y_{T+h-1|T} + \alpha_p y_{T+h-p|T} \quad \text{with } y_{T+h-p|T} = y_{T+h-p}, \text{ if } h-p \leq 0.
\end{aligned} \tag{12.34}$$

AR(p) and moving average processes can be combined as follows:

ARIMA(p, d, q) processes

$$\alpha(L)(1-L)^d y_t = \psi(L)u_t, \quad u_t \sim WN(0, \sigma^2) \tag{12.35}$$

where the AR polynomial is stable so that $y_t \sim I(d)$ and $\Delta^d y_t \sim I(0)$.

ARMA(p, q) processes

ARMA(p, q)process is an ARIMA$(p, 0, q)$ process:

$$\alpha(L)y_t = \psi(L)u_t, \quad u_t \sim WN(0, \sigma^2). \tag{12.36}$$

A discussion of the properties of ARMA and ARIMA processes can be found in **Applied Financial Econometrics** or in the textbooks mentioned.

12.3.4. OLS estimator for AR(p) models

An AR(p) model (12.13) can be estimated with the OLS estimator. To determine the estimation properties, the properties of the regressors $y_{t-1}, \ldots, y_{t-p}$ must be checked. To simplify the illustration, this is done for an AR(1) model:

- **Check whether regressor $x_t = y_{t-1}$ is predetermined with respect to u_t:** Since $u_t \sim IID(0, \sigma^2)$ holds for the errors, the following applies (see section 2.7)
$$E[u_t|u_{t-1}, u_{t-2}, \ldots] = E[u_t] = 0.$$
Because of (12.20), y_{t-1} only depends on past u_{t-1-j}, $j \geq 0$. This means that y_{t-1} is determined if the past errors are determined. However, since the expected value of u_t is independent of past errors and therefore independent of the condition on past errors, it is also independent of the condition y_{t-1}. Therefore, (9.4)
$$E[u_t|y_{t-1}] = 0$$
holds and y_{t-1} is predetermined with respect to the errors u_t.

- **Verifying strict exogeneity**: For $x_t = y_{t-1}$ to be strictly exogenous, $Cov(u_t, x_{t+1}) = Cov(u_t, y_t) = 0$ must also apply due to (9.2). This is not the case, since
$$Cov(u_t, y_t) = Cov(u_t, \alpha y_{t-1} + u_t) = \alpha Cov(u_t, y_{t-1}) + Var(u_t) = \sigma^2 > 0.$$
Because of (2.29c), $E[u_t|y_t] \neq 0$ also follows from this. Thus, $x_t = y_{t-1}$ is not strictly exogenous and the **OLS estimator for α is not unbiased**. This generally applies to models with lagged dependent variables.

Since AR(p) models are a special case of dynamic linear regression models, the estimation properties are discussed in more detail in section 13.5.

12.4. Estimation of first and second moments in the case of stationary processes

The ensemble average and variance and autocovariances can also be estimated without specifying a time series model.

12.4.1. Estimating the mean

Consistency of the mean estimator

Let $\{y_t\}$ be a weakly stationary process with mean μ and autocovariance function γ_h. Then the following applies to the mean value estimator,

$$\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$$

for $T \to \infty$:

- If $\gamma_h \overset{h\to\infty}{\longrightarrow} 0$, the following applies (as in the IID case):
$$\lim_{T\to\infty} Var(\bar{y}_T) = \lim_{T\to\infty} E\left[(\bar{y}_T - \mu)^2\right] = 0. \tag{12.37}$$
- If $\sum_{h=-\infty}^{\infty} |\gamma_h| < \infty$, the following applies:
$$\lim_{T\to\infty} T\,Var(\bar{y}_T) = \lim_{T\to\infty} T\,E\left[(\bar{y}_T - \mu)^2\right] = \sum_{h=-\infty}^{\infty} \gamma_h. \tag{12.38}$$

(Brockwell & Davis (1991, Theorem 7.1.1). See there for a proof.)

The mean value estimator $\bar{y}_T$

- is consistent under the weak condition $\gamma_h \overset{h\to\infty}{\longrightarrow} 0$ according to (12.37),

  Example: DGP is stationary AR(p) process.

- converges with $\sqrt{T}$ to the true mean μ according to (12.38) if the autocovariance function is absolutely summable.
- Cf. IID case: $nVar(\bar{y}) = \sigma$ and $Var(\bar{y}) = \sigma/n$. **In the time series case, all autocovariances must be taken into account when calculating the estimation variance**
$$Var(\bar{y}_T) \approx \frac{\gamma_0 + 2\sum_{h=1}^{\infty}\gamma_h}{T} \neq \underbrace{\frac{\gamma_0}{T}}_{IIDcase}$$

Asymptotic distribution of the mean value estimator

Theorem (Brockwell & Davis 1991, Theorem 7.1.2)
If $\{y_t\}$ is a stationary linear process (cf. (12.6)) with mean value $\mu = E(y_t)$ and independent white noise

$$y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j u_{t-j}, \quad u_t \sim IID(0, \sigma^2),$$

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$ apply, then it holds that

$$\sqrt{T}\,(\bar{y}_T - \mu) \xrightarrow{d} N(0, v) \tag{12.39}$$

with

$$v = \sum_{h=-\infty}^{\infty} \gamma_h = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j \right)^2, \tag{12.40}$$

where γ_h denotes the autocovariance function of $\{y_t\}$.

Proof (several pages) see Brockwell & Davis (1991, Section 7.3).

Remarks:

- Application of (12.39) in practice: v is estimated by estimating and summing up only $2p+1$ autocovariances, whereby (12.43) is usually used to estimate γ_h. One obtains

  $$\bar{y}_T \approx N\left(\mu, \sum_{h=-p}^{p} \hat{\gamma}_h \right), \tag{12.41}$$

  where p is chosen with rules of thumb that fulfil $p \sim cT^{1/4}$.

- ♯ If a linear process can be represented as an ARMA(p,q) process (12.36), the mean value can be calculated using the BLUE estimator (GLS estimator, see section 14.1)

  $$\hat{\mu}_T = \left(\boldsymbol{\iota}' \Gamma_T^{-1} \boldsymbol{\iota} \right)^{-1} \boldsymbol{\iota}' \Gamma_T^{-1} \mathbf{y}_T, \quad \boldsymbol{\iota} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}',$$

  whereby the covariance matrix

  $$\Gamma_T = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & \gamma_0 \end{pmatrix} \tag{12.42}$$

  can be determined from the parameters $\alpha_1, \ldots, \alpha_p, m_1, \ldots, m_q$. However, the asymptotic variance is the same (Brockwell & Davis 1991, S. 220, 236),

  $$\lim_{T\to\infty} T Var(\bar{y}_T) = \lim_{T\to\infty} T Var(\hat{\mu}).$$

Exact normal distribution of the mean estimator

If in the above theorem in (12.39) the IID condition is strengthened to Gaussian white noise, $u_t \sim NID(0, \sigma^2)$, then the mean estimator is exactly normally distributed,

$$\bar{y}_T \sim N\left(\mu, \frac{1}{T} \sum_{|h|<T} \left(1 - \frac{|h|}{T} \right) \gamma_h \right).$$

Proof: possible via simple regression model with autocorrelated errors $Cov(u_t, u_{t-h}) = \gamma_h$ by choosing the constant (=strictly exogenous) as regressor,

$$y_t = \mu \cdot 1 + u_t, \quad .$$

□

How can autocovariances be estimated?

A) Estimation of parametric time series models.

Example: AR(p) models, see section 12.3.1.

B) Direct (non-parametric) estimation of the autocovariances, see section 12.4.2

How can convergence or absolute summability of the autocovariances be checked?

Only easily possible with A): Specifying and estimating parametric linear time series models.

♯ Optional: An even more general result for estimating the mean is the following ergodic theorem.

Ergodic stochastic process

A stationary stochastic process is called **ergodic** if an event that affects all random variables y_t, $t \in \mathbb{T}$, has either probability 1 or 0 (Davidson 2000, Section 4.4.3).

Example of a stationary but non-ergodic process $\{y_t\}$ Let $u_t \sim WN(0, \sigma^2)$ and for a continuous random variable z: $z \sim (0, Var(z))$.

$$\begin{aligned} y_t &= z + u_t, \quad t \in \mathbb{Z} \\ Cov(y_t, y_{t-j}) &= Var(z) \qquad \Longrightarrow \qquad Cov(y_t, y_{t-j}) \not\longrightarrow 0 \text{ for } j \to \infty \\ E[y_t] &= E[u_t] + E[z] = E[z] = 0 \end{aligned}$$

The process $\{y_t\}$ is weakly stationary, as neither the mean value nor the variance or the autocovariances depend on the time index. However, the (linear) dependence between two elements of the stochastic process does not disappear with increasing time interval j. Therefore, $\{y_t\}$ is not ergodic. If $P(z = z_0) = 1$, z is in fact a constant. Then $\{y_t\}$ is also ergodic because $P(z \neq z_0) = 0$.

Ergodic Theorem

(Davidson 2000, Theorem 4.4.1)
If $\{y_t\}$ is stationary and ergodic and $E[y_1]$ exists, then

$$\bar{y}_T \xrightarrow{\text{a.s.}} E[y_1].$$

This is also referred to as **mean value ergodicity**. Cf. section 3.3 on almost certain convergence. This means that

$$\bar{y}_T \xrightarrow{\text{P}} E[y_1]$$

also holds.

In general, the following applies: If a stochastic process is ergodic and stationary, the stationary ensemble mean can be estimated by the time mean!

12.4.2. Estimating the autocovariance function

- $Cov(y_t, y_{t-h}) = E\left[(y_t - \mu_t)(y_{t-h} - \mu_{t-h})\right]$ is an expected value.
- Basic idea of estimation: Estimate expected value by averaging. This only works for time series if the underlying DGP is weakly stationary. Then

  $$Cov(y_T, y_{T-k}) = Cov(y_{T-1}, y_{T-1-k}) = \cdots = Cov(y_{1+K}, y_1) = \gamma_k$$

  and one suddenly has $T - k$ observations

  $$(y_T - \bar{y}_T)(y_{T-k} - \bar{y}_T), (y_{T-1} - \bar{y}_T)(y_{T-k-1} - \bar{y}_T), \ldots, (y_{1+k} - \bar{y}_T)(y_1 - \bar{y}_T)$$

  available over which one can average. Possible estimator of the autocovariance function:

  $$\hat{\gamma}_h = \frac{1}{T-h} \sum_{t=h+1}^{T} (y_t - \bar{y})(y_{t-h} - \bar{y}). \tag{12.43}$$

  One problem with this estimator is that for h close to T one again only averages over very few observations, regardless of the sample size.

  Alternative estimator:

  $$\tilde{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y}). \tag{12.44}$$

- Estimation properties (Brockwell & Davis 1991, Section 7.2)
  - biased
  - If the autocovariance matrix (12.42) is estimated with (12.44), the resulting estimated autocovariance matrix

    $$\tilde{\Gamma}_T = \begin{pmatrix} \tilde{\gamma}_0 & \tilde{\gamma}_1 & \cdots & \tilde{\gamma}_{T-1} \\ \tilde{\gamma}_1 & \tilde{\gamma}_0 & \cdots & \tilde{\gamma}_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\gamma}_{T-1} & \tilde{\gamma}_{T-2} & \cdots & \tilde{\gamma}_0 \end{pmatrix} \tag{12.45}$$

    is not negative-definite. However, this does not apply to the estimator $\hat{\gamma}_h$ (12.43).
  - $\tilde{\Gamma}_T$ is positive definite if $\tilde{\gamma}_0 > 0$. See Brockwell & Davis (1991, S. 221) together with Lütkepohl (1996, p. 151).

12.4.3. Estimating the autocorrelation function

- The autocorrelation function ρ_k can also be estimated in two ways:

$$\hat{\rho}_k = \hat{\gamma}_k / \hat{\gamma}_0, \tag{12.46}$$

$$\tilde{\rho}_k = \tilde{\gamma}_k / \tilde{\gamma}_0. \tag{12.47}$$

  - **Estimation properties** **Theorem** (Brockwell & Davis 1991, Theorem 7.2.1) If $\{y_t\}$ is a stationary linear process (12.6) with mean value $\mu = E(y_t)$ and independent white noise

$$y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j u_{t-j}, \quad u_t \sim IID(0, \sigma^2),$$

    where $\sum_{h=-\infty}^{\infty} |\gamma_h| < \infty$ and $E(u_t^4) < \infty$, then for every $h \in \mathbb{N}$, the following applies,

$$\sqrt{T}\,(\tilde{\boldsymbol{\rho}}_h - \boldsymbol{\rho}) \xrightarrow{d} N(\mathbf{0}, \mathbf{W}) \tag{12.48}$$

$$\tilde{\boldsymbol{\rho}}_h' = \begin{pmatrix} \tilde{\rho}_1 & \tilde{\rho}_2 & \cdots \tilde{\rho}_h \end{pmatrix} \tag{12.49}$$

$$\boldsymbol{\rho}_h' = \begin{pmatrix} \rho_1 & \rho_2 & \cdots \rho_h \end{pmatrix} \tag{12.50}$$

    and $\mathbf{W}$ is a covariance matrix with ij-th element

$$w_{ij} = \sum_{k=-\infty}^{\infty} \left(\rho_{k+i}\rho_{k+j} + \rho_{k-i}\rho_{k+j} + 2\rho_i\rho_j\rho_k^2 - 2\rho_i\rho_k\rho_{k+j} - 2\rho_j\rho_k\rho_{k+i} \right). \tag{12.51}$$

  - The condition of the existence of fourth moments in the above theorem can be replaced by (Brockwell & Davis 1991, Theorem 7.2.2)

$$\sum_{j=-\infty}^{\infty} \psi_j^2 |j| < \infty. \tag{12.52}$$

  - If $y_t \sim IID(0, \sigma^2)$, then $\rho_k = 0$ for $|k| > 0$ and $w_{ij} = 1$, if $i = j$ and zero otherwise. This gives an asymptotically pivotal distribution for the estimator

$$\sqrt{T}\,(\tilde{\rho}_h - \mathbf{0}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}). \tag{12.53}$$

    This results in confidence intervals for the estimated autocorrelations of (independent) white noise. For $\alpha = 0.95$ this results in

$$[-1.96/\sqrt{T}, 1.96/\sqrt{T}].$$

- The partial autocorrelation a_k can be easily estimated using the OLS estimator for a_k in the autoregressive model

$$y_t = \nu + a_1 y_{t-1} + \ldots + a_k y_{t-k} + u_t.$$

R commands

Estimating the autocovariance function, the autocorrelation function or the partial autocorrelation function of a time series: with `acf()`.

Example: Estimating the autocorrelation function of a realisation of Gaussian white noise with $\sigma^2 = 4$ and $n = 100$.

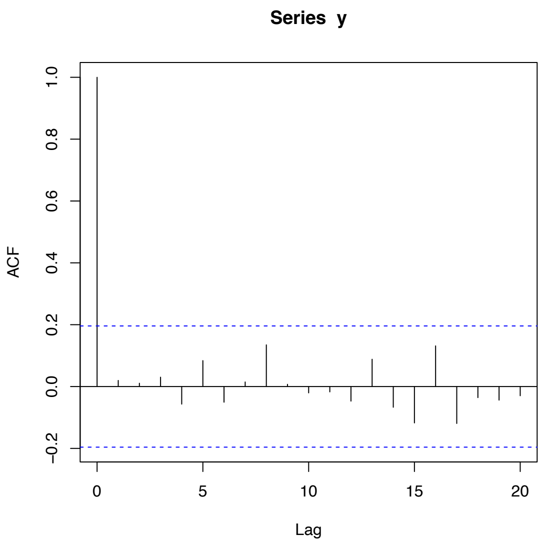


Figure 12.9.: Estimated autocorrelation function with 95% confidence intervals of a white noise realisation with $\sigma^2 = 4$ and $n = 100$ (**R program**, see section A.14, page 355)

In Figure 12.9, no ACF is outside the confidence interval for the lags under consideration. This indicates the presence of white noise.

13. Models for multivariate time series

13.1. Multivariate data generating processes

Supplements to and review of section 5.1:

- Let $\mathbf{s}_t$ denote a $(m \times 1)$ vector of (economic) random variables that are generated in period t and can be related simultaneously and over time.

- Notation: As in section 4.1, the vector $\mathbf{s}_t$ can contain more variables than ultimately needed to be modelled. We will also omit the index for the density functions in the following sections.

- The collection $\{\mathbf{s}_t\}_{t\in\mathbb{T}}$ is a **vector-valued** or **multivariate stochastic process**.

- The **data generating process, DGP** of a m-dimensional multivariate stochastic process is fully described by the conditional density

  $$f_t(\mathbf{s}_t|\mathcal{S}_{t-1})$$

  where $\mathcal{S}_{t-1}$ denotes the information set of all lagged vectors $\mathbf{s}_{t-j}$, $j > 0$,

  $$\mathcal{S}_{t-1} = \{\mathbf{s}_{t-1}, \mathbf{s}_{t-2}, \mathbf{s}_{t-3}, \ldots\}.$$

  This representation is even more general than the conditional densities that occur in (5.2), since presample values, $\mathbf{s}_0, \mathbf{s}_{-1}, \ldots$, are permitted for all conditional densities.

- Of course, one could also use the conditional probability function F instead of the conditional density f. This must even be used if non-continuous random variables are used.

♯ Formally, $\mathcal{S}_{t-1}$ denotes the (smallest) σ-algebra, i.e. the smallest set of subsets that allows to assign probabilities to all possible events based on the considered explanatory (random) vectors $\mathbf{s}_{t-1}$, $\mathbf{s}_{t-2}$, ... See section 2.3 for the definition of a σ-algebra. Instead of $\mathcal{S}_{t-1} = \{\mathbf{s}_{t-1}, \mathbf{s}_{t-2}, \mathbf{s}_{t-3}, \ldots\}$, one should correctly write

$$\mathcal{S}_{t-1} = \sigma(\mathbf{s}_{t-1}, \mathbf{s}_{t-2}, \mathbf{s}_{t-3}, \ldots).$$

- Note that the information set $\mathcal{S}_t$ does not become smaller, since

  $$\mathcal{S}_{t-2} \subseteq \mathcal{S}_{t-1} \subseteq \mathcal{S}_t \subseteq \cdots,$$

  so nothing is forgotten or knowledge is accumulated. The information sets are nested over time (Davidson 2000, Sections 4.1, 5.3.1 and especially 6.2.1).

13.2. Dynamic econometric models

- In the following, we generalise the previous definition of econometric models for random samples from section 5.2 for time series.
- A **dynamic econometric model** $\mathbb{M}$ is a family of functions $M(\cdot)$ depending on the data and a $p \times 1$ parameter vector $\boldsymbol{\psi}$ whose elements are constant over time. The functions describe the entire DGP or parts of it, or at least approximate it (Davidson 2000, Section 4.1.1). The set of possible and allowed parameters is the **parameter space $\boldsymbol{\Psi}$**

$$\mathbb{M} = \{M(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{s}_{t-2}, \dots, \mathbf{s}_2, \mathbf{s}_1, \dots, \mathbf{d}_t; \boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}, \quad \boldsymbol{\Psi} \subseteq \mathbb{R}^p \tag{13.1}$$

- In (13.1), the vector $\mathbf{d}_t$ denotes non-stochastic variables, but possibly time-varying variables, e.g. a constant 1, a time trend t, seasonal dummies, etc.
- Time-dependent parameters are collected via a function $\boldsymbol{\psi}_t = h(\mathbf{d}_t, \boldsymbol{\psi})$.

Example: AR(1) model (12.16):

Parameter vector $\boldsymbol{\psi} = \begin{pmatrix} \nu \\ \alpha_1 \\ \sigma^2 \end{pmatrix}$, parameter space $\boldsymbol{\Psi} = \mathbb{R} \times (-1, 1) \times \mathbb{R}^+$.

Example: Structural vector autoregressive model

The example follows Davidson (2000, Sections 4.5.5 and 4.7.2). Let $\mathbf{s}_t = \begin{pmatrix} y_t & z_t \end{pmatrix}^T$, where the stochastic dynamics of the variables y_t and z_t are determined by the following simultaneous equation system:

$$y_t = \gamma_1 - \alpha_{12} z_t + \beta_{11} y_{t-1} + \beta_{12} z_{t-1} + u_{1t} \tag{13.2a}$$
$$z_t = \gamma_2 - \alpha_{21} y_t + \beta_{21} y_{t-1} + \beta_{22} z_{t-1} + u_{2t}, \tag{13.2b}$$

with

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \sim NID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}}_{\boldsymbol{\Sigma}} \right). \tag{13.2c}$$

The model (13.2) is referred to as a **structural vector autoregressive model** (**SVAR model**). In section 13.3, we analyse which regressors on the right-hand side of (13.2) are exogenous and in which sense. The course **Quantitative Economic Research II** deals with estimation methods and further details. The parameter vector of the model is

$$\boldsymbol{\psi} = \begin{pmatrix} \alpha_{12} & \alpha_{21} & \beta_{11} & \beta_{12} & \beta_{21} & \beta_{22} & \gamma_1 & \gamma_2 & \sigma_{11} & \sigma_{12} & \sigma_{22} \end{pmatrix}^T.$$

- As in the case of models for random samples (cf. section 5.2), we say that the model $\mathbb{M}$ is **fully specified** if a model in reduced form $\mathbb{M}_D$ can be derived from $\mathbb{M}$ that contains conditional densities $f(\mathbf{s}_t|\mathcal{S}_{t-1}, \mathbf{d}_t; \boldsymbol{\theta}(\boldsymbol{\psi}))$ as elements (cf. (5.19))

$$\mathbb{M}_D \equiv \{f(\mathbf{s}_t|\mathcal{S}_{t-1}, \mathbf{d}_t; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}\,.$$

- If a structural dynamic **model** $\mathbb{M}$ **is fully and correctly** specified, there is a parameter vector $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\psi}_0)$ for which the conditional density in $\mathbb{M}_D$ corresponds to the DGP:

$$\begin{aligned} \mathbb{M}_D \supseteq M_D(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{s}_{t-2}, \ldots, \mathbf{s}_2, \mathbf{s}_1, \ldots, \mathbf{d}_t; \boldsymbol{\theta}_0) &\equiv f(\mathbf{s}_t|\mathcal{S}_{t-1}, \mathbf{d}_t; \boldsymbol{\theta}_0) \\ &= \underbrace{f_t(\mathbf{s}_t|\mathcal{S}_{t-1})}_{DGP}. \end{aligned} \tag{13.3}$$

Example: SVAR model – reduced form

The SVAR model (13.2) is a simultaneous equation model. Therefore, the SVAR model $\mathbb{M}$ must be transformed into a model $\mathbb{M}_D$ in reduced form so that the set of DGPs contained in the model becomes visible. For this purpose, it is favourable to write the SVAR model (13.2) in matrix notation

$$\mathbf{B}\mathbf{s}_t = \mathbf{c} + \mathbf{C}\mathbf{s}_{t-1} + \mathbf{u}_t, \quad \mathbf{u}_t \sim NID\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \tag{13.4}$$

with

$$\mathbf{B} = \begin{pmatrix} 1 & \alpha_{12} \\ \alpha_{21} & 1 \end{pmatrix}, \quad \mathbf{x}_t = \begin{pmatrix} y_t \\ z_t \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}, \quad \mathbf{u}_t = \begin{pmatrix} u_{1t} \\ u_{2t}. \end{pmatrix}$$

To obtain the **reduced form** of the SVAR model (13.4), multiply the matrix equation by the inverse of $\boldsymbol{B}$ (assuming

$$\boldsymbol{B}^{-1} = \frac{1}{1 - \alpha_{12}\alpha_{21}} \begin{pmatrix} 1 & -\alpha_{12} \\ -\alpha_{21} & 1 \end{pmatrix}$$

exists, i. e. $\alpha_{12} \neq \alpha_{21}$):

$$\mathbf{s}_t = \underbrace{\mathbf{B}^{-1}\mathbf{c}}_{\mathbf{a}} + \underbrace{\mathbf{B}^{-1}\mathbf{C}}_{\mathbf{A}_1}\mathbf{x}_{t-1} + \underbrace{\mathbf{B}^{-1}\mathbf{u}_t}_{\boldsymbol{\varepsilon}_t},$$

$$\mathbf{s}_t = \mathbf{a} + \mathbf{A}_1\mathbf{s}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{13.5a}$$

$$\boldsymbol{\varepsilon}_t = \mathbf{B}^{-1}\mathbf{u}_t = \frac{1}{1 - \alpha_{21}\alpha_{12}} \begin{pmatrix} u_{1t} - \alpha_{12}u_{2t} \\ u_{2t} - \alpha_{21}u_{1t} \end{pmatrix} \tag{13.5b}$$

$$Var(\boldsymbol{\varepsilon}_t) = \boldsymbol{\Omega} = \mathbf{B}^{-1}\boldsymbol{\Sigma}(\mathbf{B}')^{-1} \tag{13.5c}$$

with conditional density

$$\mathbf{s}_t|\mathcal{S}_{t-1} \sim N\left(\mathbf{B}^{-1}\mathbf{c} + \mathbf{B}^{-1}\mathbf{C}\mathbf{x}_{t-1}, \boldsymbol{\Omega}\right) \tag{13.5d}$$

and with conditional expected value

$$E\left(\mathbf{s}_t|\mathcal{S}_{t-1}\right) = \mathbf{B}^{-1}\mathbf{c} + \mathbf{B}^{-1}\mathbf{C}\mathbf{x}_{t-1}. \tag{13.5e}$$

The covariance matrix of $\boldsymbol{\varepsilon}_t$ is

$$\begin{aligned}\boldsymbol{\Omega} = \begin{pmatrix}\omega_{11} & \omega_{12}\\ \omega_{12} & \omega_{22}\end{pmatrix} = \frac{1}{(1-\alpha_{21}\alpha_{12})^2} \\ \begin{pmatrix}\sigma_{11} - 2\alpha_{12}\sigma_{21} + \alpha_{12}^2\sigma_{22} & (1+\alpha_{12}\alpha_{21})\sigma_{21} - \alpha_{12}\sigma_{22} - \alpha_{21}\sigma_{11}\\ \omega_{12} & \sigma_{22} - 2\alpha_{21}\sigma_{21} + \alpha_{21}^2\sigma_{11}\end{pmatrix}.\end{aligned} \tag{13.5f}$$

The conditional normal distribution for $\mathbf{s}_t$ follows from the linearity of the multivariate normal distribution.

The model (13.5) is the reduced form of an SVAR model and is generally referred to as **VAR model**.

- The elements of a model in structural form, i. e. the functions $M(\cdot)$, typically contain more parameters than are specified by the corresponding model in reduced form. Then there is a function $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\psi})$ that is not one-to-one, so that $\boldsymbol{\psi}$ cannot be uniquely determined from knowledge of $\boldsymbol{\theta}_0$, the true parameter vector of the reduced form.

  Example: SVAR model The parameter vector $\boldsymbol{\psi}$ of the SVAR model (13.2) contains 11 different parameters. The parameter vector of the reduced model (13.5) only 9: $\boldsymbol{\theta} = \begin{pmatrix}a_1 & a_2 & A_{11} & A_{12} & A_{21} & A_{22} & \omega_{11} & \omega_{12} & \omega_{22}\end{pmatrix}$.

  Problem: If this is the case, the information from the data is not sufficient to estimate the parameters of the structural form. In addition, so-called **identification assumptions** are required, which must come from economic theory. See **Quantitative Economic Research II**.

  Because sample information in principle only enables the estimation of the parameters of reduced forms.

- The structural and reduced form of a model can be identical.

  Simple example: SVAR model and VAR model identical If it is assumed a priori in (13.2) that $\alpha_{12} = \alpha_{21} = 0$, then the simultaneous relationship between z_t and y_t disappears and z_t becomes causal for y_t. Then the structural and reduced form correspond.

  The case in the previous example is considered more generally in section 13.3.

- Cf. Davidson (2000, Section 4.1).

13.3. Conditions on exogenous variables in dynamic models

- Even if the correct and complete model $\mathbb{M}_D$ (13.3) were known, it would be impossible – given typical sample sizes – to reliably estimate the correct $(p^* \times 1)$ parameter vector $\boldsymbol{\psi}_0$ if the number of model parameters p^* is extremely large. This is the case if the number of variables under consideration m is very large.

- If you are only interested in the explanation / modelling of selected variables $\mathbf{y}_t$, conditional models can be used as in section 5.2. For time series, however, the definition (5.19) suitable for random samples must be suitably extended.

- The decomposition of the vector $\mathbf{s}_t$ according to (5.5) into irrelevant variables $\mathbf{w}_t$, variables to be explained $\mathbf{y}_t$ and explanatory variables $\mathbf{z}_t$ also applies here.

- Due to the time structure of the data and the DGP, different types of exogeneity can be distinguished, which were introduced by Engle et al. (1983) and are very helpful:

  - **weak exogeneity**: z_t is causal for y_t *within the same time period* in the context of the model under consideration

  - **strong exogeneity**: z_t can be considered as given in multi-step forecasts for y_{t+h}

  - **super-exogeneity**: z_t fulfils the condition to be used as an economic policy control variable.

  At the end of the section, these are related to the previous definitions of strictly exogenous and predetermined variables.

Procedure for defining conditional models for time series

- Partition original vector $\mathbf{s}_t$ and define associated information sets

$$\mathbf{s}_t = \begin{pmatrix} \mathbf{w}_t \\ \mathbf{y}_t \\ \mathbf{z}_t \end{pmatrix}, \qquad \begin{matrix} \mathcal{W}_{t-1} = \{\mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots\} = \sigma(\mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots) \\ \mathcal{Y}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots\} = \sigma(\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots) \\ \mathcal{Z}_{t-1} = \{\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots\} = \sigma(\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots) \end{matrix} \quad ,$$

  so that $\mathbf{w}_t$ contains the irrelevant variables, $\mathbf{y}_t$ the endogenous variables that must be explained within the model, and $\mathbf{z}_t$ the variables that do not have to be explained in the model for the questions at hand, but are relevant for elements of $\mathbf{y}_t$, i. e. are exogenous.

- Combination of information sets. Note the use of the symbol $\vee$ (Davidson 2000, Section B.10):

$$\mathcal{S}_{t-1} = \sigma(\mathbf{w}_{t-1}, \mathbf{y}_{t-1}, \mathbf{z}_{t-1}, \mathbf{w}_{t-2}, \dots) \equiv \mathcal{W}_{t-1} \vee \mathcal{Y}_{t-1} \vee \mathcal{Z}_{t-1} \neq \mathcal{W}_{t-1} \cup \mathcal{Y}_{t-1} \cup \mathcal{Z}_{t-1}.$$

- Define

$$f(\mathbf{s}_t|\mathcal{W}_{t-1} \vee \mathcal{Y}_{t-1} \vee \mathcal{Z}_{t-1}) \equiv f(\mathbf{s}_t|\mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}).$$

- Analogous to (5.6) factorisation of the (parametric) density f for $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\psi})$

$$f(\mathbf{s}_t|\mathcal{S}_{t-1}, \mathbf{d}_t, \boldsymbol{\theta}) = f(\mathbf{w}_t|\mathbf{y}_t, \mathbf{z}_t, \mathcal{S}_{t-1}, \mathbf{d}_t, \boldsymbol{\theta})\ f(\mathbf{y}_t|\mathbf{z}_t, \mathcal{S}_{t-1}, \mathbf{d}_t, \boldsymbol{\theta})\ f(\mathbf{z}_t|\mathcal{S}_{t-1}, \mathbf{d}_t, \boldsymbol{\theta})$$
$$\text{in short:} \qquad f_{\mathbf{w},\mathbf{y},\mathbf{z}} = f_{\mathbf{w}|\mathbf{y},\mathbf{z}}\ f_{\mathbf{y}|\mathbf{z}}\ f_{\mathbf{z}}. \tag{13.6}$$

Assumptions for weak exogeneity

- There exists a partitioning of the parameter vector $\boldsymbol{\psi}$

$$\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\psi}_1 \\ \boldsymbol{\psi}_2 \end{pmatrix}, \quad \boldsymbol{\psi}_1 \in \boldsymbol{\Psi}_1, \boldsymbol{\psi}_2 \in \boldsymbol{\Psi}_2, \quad \boldsymbol{\Psi} = \boldsymbol{\Psi}_1 \times \boldsymbol{\Psi}_2, \tag{13.7}$$

- and the following applies to the conditional densities:

$$f_{\mathbf{w}|\mathbf{y},\mathbf{z}} = f(\mathbf{w}_t|\mathbf{y}_t, \mathbf{z}_t, \mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}, \mathbf{d}_t, \boldsymbol{\theta}_2, \quad) \tag{13.8a}$$
$$f_{\mathbf{y}|\mathbf{z}} = f(\mathbf{y}_t| \quad \mathbf{z}_t, \quad \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}, \mathbf{d}_t, \quad \boldsymbol{\theta}_1) \tag{13.8b}$$
$$f_{\mathbf{z}} = f(\mathbf{z}_t| \quad \mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}, \mathbf{d}_t, \boldsymbol{\theta}_2, \quad) \tag{13.8c}$$

with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1(\boldsymbol{\psi}_1)$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2(\boldsymbol{\psi}_2)$.

- In words:
  - The functions of the conditional densities for the variables $\mathbf{w}_t$, $\mathbf{z}_t$ that are not to be explained do not depend on the parameter vector $\boldsymbol{\psi}_1$.
  - The conditional density function for the variables to be explained $\mathbf{y}_t$ does not depend on $\boldsymbol{\psi}_2$ and does not depend on the past of $\mathbf{w}_t$. The multivariate stochastic process $\{\mathbf{w}_t\}$ is therefore irrelevant for $f_{\mathbf{y}|\mathbf{z}}$.
  - It is not the case that $\boldsymbol{\psi}_1 = \boldsymbol{\psi}_1(\boldsymbol{\psi}_2)$, so that knowledge of $\boldsymbol{\psi}_2$ cannot improve the estimation properties for $\boldsymbol{\psi}_1$. One then denotes $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ as **variation free**.
- If the assumptions (13.7) and (13.8) apply, it does not matter for the complete modelling of $\mathbf{y}_t$ whether $f_{\mathbf{w},\mathbf{y},\mathbf{z}}$ or only $f_{\mathbf{y}|\mathbf{z}}$ is considered.
- If $f_{\mathbf{y}|\mathbf{z}}$ is considered, the model is said to be **conditional on $\mathbf{z}_t$ (conditional model)** and **marginalised with respect to $\mathbf{w}_t$**.
- A **parameterisation of a model is never unique**, since any vector-valued function $\boldsymbol{\phi} = f(\boldsymbol{\psi})$, which is bijective, can be used to generate an alternative parameterisation, but with a different **interpretation**. Example: $\boldsymbol{\phi} = \exp(\boldsymbol{\beta})$.
- For the **existence** of the conditional density $f_{\mathbf{y}|\mathbf{z}}$ in (13.8b) it is therefore important that *any* parameter vector $\boldsymbol{\psi}$ exists that fulfils (13.7) and (13.8).
- It may be possible to further divide $\boldsymbol{\psi}_2$ into parameters for $f_{\mathbf{w}|\mathbf{y},\mathbf{z}}$ and $f_{\mathbf{z}}$. Since both conditional densities are irrelevant for the analysis under assumptions (13.7) and (13.8), this is not necessary.

Weak exogeneity

- If the **goal of the analysis** is limited to the explanation of $\mathbf{y}_t$ (instead of all variables in $\mathbf{s}_t$), you only want to analyse the conditional model for $f_{\mathbf{y}|\mathbf{z}}$ and dispense with the analysis of the marginal model for $f_{\mathbf{z}}$ (cf. section 5.2).

- This is possible precisely when no information is lost for the parameters $\boldsymbol{\psi}_1$ of the conditional model that implies $f_{\mathbf{y}|\mathbf{z}}$ by not analysing the marginal model that implies $f_{\mathbf{z}}$. The conditions for this are (13.7) and (13.8). The variables $\mathbf{z}_t$ of the marginal model are then denoted as **weakly exogenous with respect to $\boldsymbol{\psi}$** for the conditional model, which implies $f_{\mathbf{y}|\mathbf{z}}$. The remaining variables $\mathbf{y}_t$ are referred to as **endogenous**.

- Restricting the analysis to the conditional model (13.8b) makes sense if the conditional density to explain $\mathbf{y}_t$ contains considerably fewer parameters and conditioning variables than the conditional density for $\mathbf{w}_t$.

- Of course, the **interest in the explanation of $\mathbf{y}_t$ by $\mathbf{z}_t$** and $\mathcal{Y}_{t-1}$, $\mathcal{Z}_{t-1}$ based on the conditional density $f_{\mathbf{y}|\mathbf{z}}$ can also refer to any **parameter vector $\boldsymbol{\phi}_1$**, as long as this is determined by $\boldsymbol{\phi}_1 = g(\boldsymbol{\psi}_1)$ (one possibility is $\boldsymbol{\theta}_1$), where $g(\cdot)$ does not have to be invertible and any parameterisation $\boldsymbol{\psi}$ exists for which (13.7) and (13.8) applies.

- Then the variables in $\mathbf{y}_t$ are referred to as **endogenous** and the variables in $\mathbf{z}_t$ as **weakly exogenous for $\boldsymbol{\phi}_1$**. This term was introduced by Engle et al. (1983).

Remarks:

- Splitting the parameter vector $\boldsymbol{\psi}$ into $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$: **'sequential cut of the parameters'** into 'those that have an impact on the analysis' and 'the rest' (Engle et al. 1983).

- The condition (13.7) guarantees that $\boldsymbol{\phi}_1$ is neither directly nor indirectly dependent on $\boldsymbol{\psi}_2$. This means that it is not possible, for example, that knowing the parameters of the marginal model for $\mathbf{z}_t$ would help to determine $\boldsymbol{\phi}_1$ more precisely.

- Note: The decomposition (13.6) can always be different, e.g. $f_{\mathbf{w},\mathbf{y},\mathbf{z}} = f_{\mathbf{w}|\mathbf{y},\mathbf{z}}\, f_{\mathbf{z}|\mathbf{y}}\, f_{\mathbf{y}}$, but possibly without 'sequential cut of the parameters'.

- Frequently searched for: Smallest subset of endogenous variables, i.e. if possible y_t scalar, so that (13.7) and (13.8) just still apply.

- Weak exogeneity is a property that
  - refers to **variable and parameter** within a model,
  - is **not synonymous with causality**,
  - however, excludes simultaneous relationships (cf. section 4.1) and
  - can only be assessed with respect to the 'larger' model $f_{\mathbf{w},\mathbf{y},\mathbf{z}}$.

  The frequent use of the term "$\mathbf{z}_t$ is exogenous for $\mathbf{y}_t$" is therefore imprecise, as the dependence on the model parameters is not clear.

- Cf. Davidson (2000, Section 4.5.3), Hendry (1995, Chapter 5, esp. Sec. 5.3)
- The conditions for weak exogeneity do not rule out the possibility of feedback effects from $\mathbf{y}_t$ to $\mathbf{z}_{t+1}$ (via $\mathcal{Y}_t$ in (13.8c)) and thus to future $\boldsymbol{y}_t$. This is precisely why weak exogeneity is referred to as "weak" . To determine feedback effects, the concept of Granger causality is central, but it is not identical to the concept of causality defined in section 4.1.

For macroeconomic questions, it is generally not possible to conduct controlled random experiments and natural experiments are rare. ⟶ Use of a weaker concept:

Granger causality

Clive Granger (1969) (Nobel Prize winner 2003, together with Robert Engle).

- A variable z_t is **Granger-causal** for y_t if knowledge of z_t somehow helps to improve the prediction for y_{t+h} for at least one $h > 0$. A sufficient condition is that for at least one forecast horizon $h > 0$ the following does *not apply*

$$f(y_{t+h}|\{z_t, z_{t-1}, \dots, \}, \widetilde{\Omega}_t) = f(y_{t+h}|\widetilde{\Omega}_t), \tag{13.9}$$

  where $\widetilde{\Omega}_t$ denotes an information set that can contain all arbitrary variables **except** $\{z_t, z_{t-1}, \dots, \}$.
- If (13.9) applies to all $h > 0$, z_t is **not Granger-causal** for y_t.
- Granger causality $\not\Longleftarrow\not\Longrightarrow$ Existence of a causal mechanism.
- Granger causality refers exclusively to predictive power.
- Cf. Davidson (2000, Section 4.5.4), Lütkepohl (2004, Section 3.7.1).

Strong exogeneity

- Consider $\mathbf{z}_t = \begin{pmatrix} \mathbf{z}_{1t}' & \mathbf{z}_{2t}' \end{pmatrix}'$.
- $\mathbf{y}_t$ is **not Granger-causal** for $\mathbf{z}_{2t}$ if it holds that

$$f(\mathbf{z}_{1t}, \mathbf{z}_{2t}|\mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}) = f(\mathbf{z}_{1t}|\mathbf{z}_{2t}, \mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1})\ f(\mathbf{z}_{2t}|\mathcal{W}_{t-1}, \mathcal{Z}_{t-1}), \tag{13.10}$$

- **$\mathbf{z}_{2t}$ strongly exogenous for ϕ_1**:
  - **$\mathbf{y}_t$ not Granger-causal** for $\mathbf{z}_{2t}$ and
  - **$\mathbf{z}_{2t}$ is weakly exogenous for ϕ_1**
- If $\mathbf{z}_{2t}$ is strongly exogenous (for ϕ_1), **delayed feedback from $\mathbf{y}_t$ is also excluded**. This means that $\mathbf{z}_{2t}$

- can be treated as given in multi-step forecasts.
- can be treated like the non-stochastic variables $\mathbf{d}_t$.

- Cf. Davidson (2000, Section 4.5.4) or Hendry (1995, Section 5.8).

Example: SVAR model

Question: For the explanation of y_t by equation (13.2a), under which parameter restrictions is z_t weakly exogenous?

The prerequisite for weak exogeneity is that the following 'sequential cut' of the parameters

$$\boldsymbol{\psi}_1 = \begin{pmatrix}\alpha_{12} & \gamma_1 & \beta_{11} & \beta_{12} & \sigma_{11}\end{pmatrix}, \quad \boldsymbol{\psi}_2 = \begin{pmatrix}\alpha_{21} & \gamma_2 & \beta_{21} & \beta_{22} & \sigma_{12} & \sigma_{21} & \sigma_{22}\end{pmatrix} \tag{13.11}$$

exists, whereby the parameter vector $\boldsymbol{\psi}_1$ contains the parameters of the structural equation (13.2a) and their error variance.

To check the condition (13.8), first the i) conditional density is derived, ii) then the factorisation is performed and iii) finally the condition is checked.

i) has already been done in (13.5).

ii) Factorise the density f: In this case, there is no $\boldsymbol{w}_t$. If one is interested in the explanation of y_t, one needs the factorisation (13.6) $f_{y,z} = f_{y|z} f_z$.

Procedure:

- 1st step: Derivation of the factorisation of $f_{\varepsilon_{1t},\varepsilon_{2t}} = f_{\varepsilon_{1t}|\varepsilon_{2t}} f_{\varepsilon_{2t}}$.
- 2nd step: Replace ε_{1t} and ε_{2t} with equations in reduced form (13.5b).

1st step: Due to the normal distribution assumption and $E(\boldsymbol{\varepsilon}_t) = 0$ one can write

$$\varepsilon_{1t} = \rho\varepsilon_{2t} + \eta_t, \quad \text{where } E[\eta_t|\varepsilon_{2t}] = 0, \tag{13.12}$$

so that

$$E[\varepsilon_{1t}|\varepsilon_{2t}] = \rho\varepsilon_{2t}. \tag{13.13}$$

Since the ε's are the errors of the reduced form (13.5b), ε_{1t} does not contain z_t and ε_{2t} does not contain y_t (in contrast to the errors of the structural form (13.2)). In the 2nd step, ε_{1t} and ε_{2t} can thus be replaced by the respective equations in $\boldsymbol{\varepsilon}_t = \mathbf{x}_t - \boldsymbol{A}_0 - \boldsymbol{A}_1\mathbf{x}_{t-1}$ and then the conditional density of the conditional model can be determined and its expected value $E[y_t|z_t, \mathcal{X}_{t-1}]$ can be calculated.

The parameter ρ in (13.12) must first be determined. This is done by determining the covariance and variance for equation (13.12):

$$\begin{aligned} Cov(\varepsilon_{1t}, \varepsilon_{2t}) = aVar(\varepsilon_{2t}) = \omega_{12} = \rho\omega_{22} \quad &\Longrightarrow \rho = \frac{\omega_{12}}{\omega_{22}} \\ Var(\varepsilon_{1t}) = \omega_{11} = \rho^2\omega_{22} + Var(\eta_t) \quad &\Longrightarrow Var(\eta_t) = \omega_{11} - \frac{\omega_{12}^2}{\omega_{22}}. \end{aligned}$$

One obtains

$$\underbrace{\varepsilon_{1t}}_{\text{is without } z_t} = \frac{\omega_{12}}{\omega_{22}} \underbrace{\varepsilon_{2t}}_{\text{is without } y_t} + \eta_t. \tag{13.14}$$

2nd step: Now ε_{1t} and ε_{2t} in (13.14) are replaced by the respective equations in $\boldsymbol{\varepsilon}_t = \mathbf{x}_t - \boldsymbol{A}_0 - \boldsymbol{A}_1\mathbf{x}_{t-1}$. After a few transformations (details at the end of the section), one obtains for y_t

$$y_t = \frac{\omega_{12}}{\omega_{22}} z_t + (P\gamma_1 - Q\gamma_2) + (P\beta_{11} - Q\beta_{21})\, y_{t-1} + (P\beta_{12} - Q\beta_{22})\, z_{t-1} + \eta_t \tag{13.15a}$$

mit

$$\frac{\omega_{12}}{\omega_{22}} = \frac{(1 + \alpha_{12}\alpha_{21})\sigma_{21} - \alpha_{12}\sigma_{22} - \alpha_{21}\sigma_{11}}{\sigma_{22} - 2\alpha_{21}\sigma_{21} + \alpha_{21}^2\sigma_{11}}, \tag{13.15b}$$

$$P = \frac{1 + \alpha_{21}\omega_{12}/\omega_{22}}{1 - \alpha_{21}\alpha_{12}}, \quad Q = \frac{\alpha_{12} + \omega_{12}/\omega_{22}}{1 - \alpha_{21}\alpha_{12}}. \tag{13.15c}$$

The **conditional density $f_{y_t|z_t,\mathbf{x}_{t-1}}$ of the conditional model for y_t given z_t** is therefore

$$y_t|z_t, \mathbf{x}_{t-1} \sim \tag{13.16}$$

$$N\left(\frac{\omega_{12}}{\omega_{22}} z_t + (P\gamma_1 - Q\gamma_2) + (P\beta_{11} - Q\beta_{21})\, y_{t-1} + (P\beta_{12} - Q\beta_{22})\, z_{t-1}, \omega_{11} - \frac{\omega_{12}^2}{\omega_{22}}\right).$$

The **conditional expected value for y_t given z_t and lags** is

$$E(y_t|z_t, y_{t-1}, z_{t-1}) = \frac{\omega_{12}}{\omega_{22}} z_t + (P\gamma_1 - Q\gamma_2) + (P\beta_{11} - Q\beta_{21})\, y_{t-1} + (P\beta_{12} - Q\beta_{22})\, z_{t-1} \tag{13.17}$$

Analogously, the expected value for z_t can also be calculated.

iii) Checking the conditions (13.8) for weak exogeneity

- The conditional density (13.16) for the conditional model for y_t includes *all* parameters of the parameter vector $\boldsymbol{\psi}$.

- The conditions (13.8) for weak exogeneity can only be fulfilled if a sequential cut (13.11) exists, so that the parameters of the equation for z_t do not affect the conditional model for y_t. The latter is only possible if

$$\frac{\omega_{12}}{\omega_{22}} = -\alpha_{12} \quad \Longrightarrow \quad P = 1, \quad Q = 0. \tag{13.18}$$

- In order for (13.18) to apply,
  - the equation $\alpha_{12} = \alpha_{21}^{-1}$ must be fulfilled for $\alpha_{21} \neq 0$ and/or $\sigma_{21} \neq 0$, which means that $\mathbf{B}$ is not invertible and no reduced form exists, or
  - $\alpha_{21} = 0$ and $\sigma_{12} = 0$ must hold.
- The **sequential cut** (13.11) is therefore only possible **if** $\alpha_{21} = \sigma_{21} = 0$.
  - Then z_t **is weakly exogenous for** $\boldsymbol{\psi}_1$ or $\boldsymbol{\phi}_1 = g(\boldsymbol{\psi}_1)$ and
  - (13.2) is a **recursive model**.
- Verification of $\alpha_{21} = \sigma_{21} = 0$ not possible with regression. Why is this the case?

Remark: **If z_t is not weakly exogenous with respect to the parameter vector $\boldsymbol{\psi}_1$**, then the OLS estimator does not estimate the parameters of the structural equation (13.2a), but the parameters of the conditional expected value (13.17).

Strong exogeneity z_t **is strongly exogenous for** $\boldsymbol{\psi}_1$ or $\boldsymbol{\phi}_1 = g(\boldsymbol{\psi}_1)$ if the following applies in the reduced form **in addition** to weak exogeneity ($\alpha_{12} = \sigma_{21} = 0$)

$$\mathbf{B}^{-1}\mathbf{C}\mathbf{x}_{t-1} = \frac{1}{1 - \alpha_{12}\alpha_{21}} \begin{pmatrix} \beta_{11} - \alpha_{12}\beta_{21} & \beta_{12} - \alpha_{12}\beta_{22} \\ -\alpha_{21}\beta_{11} + \beta_{21} & -\alpha_{21}\beta_{12} + \beta_{22} \end{pmatrix} \mathbf{x}_{t-1} = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \mathbf{x}_{t-1},$$

thus $\beta_{21} = 0$ (then y_{t-1} has no impact on z_t). The following then applies

$$E(z_t|\mathcal{X}_{t-1}) = E(z_t|z_{t-1}).$$

Comparison of exogeneity concepts

- **Weak exogeneity versus partial independence**
  - Partial independence (9.4) of a regressor can only ever be assessed with regard to an error term.
  - Thus, by construction z_t is partially independent with respect to η_t, since from (13.17) follows: $E[\eta_t|z_t, y_{t-1}, z_{t-1}] = 0$.
  - And so, in general, z_t is *not* partially independent with respect to u_{1t} in (13.2a), because it can be shown that in general $E[u_{1t}|z_t, y_{t-1}, z_{t-1}] \neq 0$.

- However, if z_t is weakly exogenous with respect to $\boldsymbol{\psi}_1$, then z_t is partially independent of u_{1t}, since u_{1t} is then identical to η_t, since $\alpha_{12} = 0$.
- Whether a regressor vector is partially independent depends i) always on the underlying 'error-defining' model, ii) and possibly on the true parameters of the 'larger' model that generates the error - as with weak exogeneity.
- The advantage of the concept of weak exogeneity over the concept of predetermination is that it explicitly makes clear when it is fulfilled by means of parameter restrictions.

- **Strong exogeneity versus strict exogeneity**
  - The regressors $\mathbf{X}_t$ are referred to as **strictly exogenous** if (9.1) is fulfilled.
  - According to the comments on weak exogeneity, the following applies: If a variable is strongly exogenous with regard to a parameter vector, this variable is also strictly exogenous.

Current status in the literature: An issue of the *Journal of Econometrics* (2006) is dedicated to causality and exogeneity, see Bauwens et al. (2006).

♯ **Super-exogeneity**

- exists if the conditional distribution $f_{\mathbf{y}|\mathbf{z}}$ depending on $\boldsymbol{\phi}_1$ is invariant to changes in the marginal/joint distribution $f_{\mathbf{z}}$.
- **Formally**: The vector of non-stochastic variables $\mathbf{d}_t$ can be decomposed into $\mathbf{d}_{1t}$ and $\mathbf{d}_{2t}$, e.g. $\mathbf{d}_t = (1, d_t)'$.

  At least one element of $\mathbf{z}_t$ **is super-exogenous for** $\boldsymbol{\phi}_1 = g(\boldsymbol{\psi}_1)$:
  - $\mathbf{d}_{2t}$ varies over the observation period and is a non-trivial argument of the marginal/joint density $f_{\mathbf{z}}$
  - The conditional density $f_{\mathbf{y}|\mathbf{z}}$ does not depend on $\mathbf{d}_{2t}$.
  - $\mathbf{z}_t$ **is weakly exogenous for** $\boldsymbol{\phi}_1$.
- In order for $\mathbf{z}_t$ **to be super-exogenous for** $\boldsymbol{\phi}_1$, (13.8) with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1(\boldsymbol{\psi}_1)$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2(\boldsymbol{\psi}_2)$ must be further restricted:

$$f_{\mathbf{w}|\mathbf{y},\mathbf{z}} = f(\mathbf{w}_t|\mathbf{y}_t, \mathbf{z}_t, \mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}, \mathbf{d}_{1t}, \mathbf{d}_{2t}, \boldsymbol{\theta}_2) \tag{13.19a}$$

$$f_{\mathbf{y}|\mathbf{z}} = f(\mathbf{y}_t| \quad \mathbf{z}_t, \quad \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}, \mathbf{d}_{1t}, \quad \boldsymbol{\theta}_1) \tag{13.19b}$$

$$f_{\mathbf{z}} = f(\mathbf{z}_t| \quad \mathcal{W}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}, \quad \mathbf{d}_{2t}, \boldsymbol{\theta}_2) \tag{13.19c}$$

  In order for $\mathbf{z}_t$ **to be strongly exogenous for** $\boldsymbol{\phi}_1 = g(\boldsymbol{\psi}_1)$, $\mathcal{Y}_{t-1}$ must not occur in (13.19c).
- Super-exogeneity eliminates symmetry between $\mathbf{y}_t$ and $\mathbf{z}_t$, so that causality can be

determined despite contemporaneous correlation (identification). In the following example, only one of the two conditional expected values depends on $\mathbf{d}_{2t}$.

- allows identification of relationships,
  - which most likely allow structural interpretation,
  - which are invariant to (economic) policy and
  - fulfil a prerequisite for immunity to the Lucas critique of econometric models.
- In essence, the **Lucas critique** reads:

  > 'Given that the structure of an econometric model consists of optimal decision rules for economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models.' (Lucas 1976, p. 41)

  (Quoted from Hendry (1995, Section 14.14). See there for tests regarding the Lucas critique).

  The Lucas critique also applies, for example, when regression parameters are composed of parameters from behavioural equations and expectations, such as in rational expectations models, see e.g. Davidson (2000, Section 5.5).

Example: SVAR model

The second structural equation (13.2b) also contains the term $\delta_2 d_{2t}$, the first structural equation remains unchanged. Then z_t is super-exogenous for y_t.

13.4. Dynamic linear regression models

Dynamic linear regression model

- A **dynamic linear regression model** is a dynamic econometric model (13.1) whose explanatory variable y_t is determined by a linear combination of explanatory variables and an error term. The explanatory variables can contain lagged endogenous variables $y_{t-j}, j > 0$.
- Dynamic linear regression models generally model the conditional expected value of the conditional density (13.6)

  $$f_{y|\mathbf{z}} = f(y_t|\mathbf{z}_t, \mathcal{S}_{t-1}, \mathbf{d}_t; \boldsymbol{\theta}) \tag{13.20}$$

  of the conditional model, where y_t is scalar.
- **Notation**: In the following we assume that

- $\boldsymbol{w}_t$ has already been classified as not relevant.
- We also only consider **one** variable to be explained, namely y_{1t}, which is notated as y_t in the following. There are no further endogenous y_{jt}, $j \geq 2$. The z_k explanatory variables $\mathbf{z}_t$ are summarised as

$$\boldsymbol{Z}_t = \begin{pmatrix} z_{1t} & \cdots & z_{k_z,t} \end{pmatrix} = \mathbf{z}_t^T$$

  so that a sample observation can be written as

$$\begin{pmatrix} y_t \\ \mathbf{Z}_t^T \end{pmatrix}. \tag{13.21}$$

The conditional density (13.20) of the conditional model can thus be specified as

$$f_{Y_t|\boldsymbol{Z}_t,\boldsymbol{Z}_{t-1},\ldots,\boldsymbol{Z}_1,Y_{t-1},\ldots,Y_1}(y_t|\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1, y_{t-1}, \ldots, y_1, \boldsymbol{d}_t).$$

The conditional expected value to be modelled is then

$$E[y_t|\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1, y_{t-1}, \ldots, y_1, \boldsymbol{d}_t].$$

If we now assume that the conditional expected value is linear in the parameters, we obtain the dynamic *linear* regression model, which is discussed in more detail below.

- The **substantively (economically) relevant parameters** of dynamic linear regression models can be consistently estimated with the OLS estimator if certain conditions are met, see section 13.5. These include the weak exogeneity of the regressors (or their predetermination). See the extensive discussion in section 13.3. This prerequisite is already taken into account in the following when defining feasible explanatory variables.

  If the condition of weak exogeneity is not met, the conditional expected value of the reduced form can still be estimated consistently if the latter is linear. Cf. (13.17) in the example in the previous section. However, the parameters are then not interpretable. For forecasting purposes, however, this may be irrelevant.

Dynamic linear regression models

- All regressor variables that can be used to specify a dynamic linear regression model for the endogenous variable y_t form the information set Ω_t of all *potentially* explanatory variables. The information set of the regressor variables actually used in a model is denoted by $\mathcal{I}_t \subset \Omega_t$. See section 5.2.
- Possible regressor variables in $\mathcal{I}_t$ are:
  - deterministic variables, summarised in the row vector $\mathbf{d}_t$: Constant, time trend, seasonal dummies, etc.,

- lagged dependent variables y_{t-j}, $j > 0$,
- predetermined (contemporaneous) variables $\mathbf{Z}_t$ with respect to the error term u_t, i. e. $\boldsymbol{Z}_t \in \Omega_t$, where $E[u_t|\Omega_t] = 0$,
- lagged $\mathbf{Z}_t$, i. e. $\mathbf{Z}_{t-j}, j > 0$,
- (almost) every function of the variables mentioned.

- A **dynamic linear regression model** with information set $\mathcal{I}_t = \{\boldsymbol{d}_t, \boldsymbol{Z}_t, \dots, \boldsymbol{Z}_{t-m}, y_{t-1}, \dots, y_{t-p}\}$ is given by

$$y_t = \mathbf{d}_t\boldsymbol{\nu} + \mathbf{Z}_t\boldsymbol{\delta}_0 + \mathbf{Z}_{t-1}\boldsymbol{\delta}_1 + \cdots + \mathbf{Z}_{t-m}\boldsymbol{\delta}_m + y_{t-1}\alpha_1 + \ldots + y_{t-p}\alpha_p + u_t, \quad t \in \mathbb{T}. \tag{13.22}$$

- With

$$\mathbf{X}_t = \begin{pmatrix} \mathbf{d}_t & \mathbf{Z}_t & \mathbf{Z}_{t-1} & \cdots & \mathbf{Z}_{t-m} & y_{t-1} & \cdots y_{t-p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\delta}_0 \\ \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_m \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}, \tag{13.23}$$

the dynamic linear regression model (13.22) can be written again in the familiar compact form

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t. \tag{13.24}$$

Dynamically correctly specified model

- A dynamic linear regression model is **dynamically correctly specified** if the following applies for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{X}_t \in \mathcal{I}_t$:

$$E[y_t|\Omega_t] = E[y_t|\boldsymbol{d}_t, \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \dots, \boldsymbol{Z}_{t-m}, y_{t-1}, \dots, y_{t-p}] = E[y_t|\boldsymbol{X}_t, \boldsymbol{\beta}_0] = \boldsymbol{X}_t\boldsymbol{\beta}_0 \tag{13.25}$$

where $\boldsymbol{\beta}_0$ is the true parameter vector.

13.5. OLS estimation of dynamic linear regression models

Since AR(p) models are a special case of dynamic linear regression models, it is sufficient to examine the estimation properties for the latter.

- **Assumptions** for asymptotic estimation properties of the OLS estimator of (13.24):

- **(C1)** $\Longleftrightarrow$ Assumption **(B1)**: The DGP is contained in (13.24) for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

- **(C2)**: $u_t|\Omega_t \sim (0, \sigma^2)$ $\Longleftrightarrow$

  (C2a) Regressors predetermined

  $$E(u_t|\Omega_t) = 0,$$

  (C2b) Conditional homoscedasticity of errors

  $$E(u_t^2|\Omega_t) = \sigma^2 := E(u_t^2),$$

  where $\sigma^2 = \sigma_0^2$ applies to the error variance of the DGP.

- **(C3)** $\Longleftrightarrow$ Assumption **(A1)**

  $$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t^T\mathbf{X}_t = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} E(\mathbf{X}_t^T\mathbf{X}_t) = \mathbf{S}_{\mathbf{X}^T\mathbf{X}} < \infty, \quad \mathbf{S}_{\mathbf{X}^T\mathbf{X}} \text{ invertible.}$$

- **(C4a) Strict stationarity** of $\{\mathbf{s}_t\} = \begin{pmatrix} y_t & \mathbf{Z}_t \end{pmatrix}^T$,

- **(C4b)** $E|\boldsymbol{\lambda}^T\mathbf{X}_t u_t|^{2+\delta} \leq B < \infty, \quad \delta > 0$, for all fixed $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}^T\boldsymbol{\lambda} = 1$.

- **Asymptotic estimation properties of the OLS estimator**

  - **Consistency**: Unter den Annahmen **(C1)**, **(C2)**, **(C3)**, the OLS estimator is consistent, i. e.

  $$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 \tag{13.26}$$

  - **Asymptotic normal distribution**: Under assumptions **(C1)**, **(C2)**, **(C3)** and **(C4a)** or **(C4b)**, the OLS estimator is asymptotically normally distributed,

  $$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \xrightarrow{d} N(\mathbf{0}, \sigma_0^2\mathbf{S}_{\mathbf{X}^T\mathbf{X}}^{-1}). \tag{13.27}$$

  - Here without proofs. The (elaborate) proofs can be found in the slides for the MA course **Advanced Dynamic Econometrics** or in Davidson (2000).

- **Notes on the assumptions**:

  - The assumption **(C2a)** assumes that all regressors are predetermined, i. e. (9.4) applies, **and** that the **model is dynamically correctly specified**, i. e.

  $$E[y_t|\mathbf{X}_t, y_{t-1}, \mathbf{X}_{t-1}, y_{t-2}, \mathbf{X}_{t-2}, \ldots] = E[y_t|\mathbf{X}_t] \tag{13.28}$$

  Then the errors $\{u_t\}$ are uncorrelated.

- The assumption **(C2a)** is weaker than strict exogeneity **(B2a)**, therefore the OLS estimator in the dynamic linear regression model is generally biased.
- In order for assumption **(C3)** to apply, for example, in the case of an AR(1) process (12.16), it must hold that
  - $|\alpha| < 1$ (stability condition) holds and
  - $E|u_t|^{2+\delta} \leq B < \infty, \delta > 0$, $t = 1, \ldots, n$, i. e. moments exist for the error distribution beyond the variance.

  For AR(p) processes, the corresponding stability condition must be fulfilled (see e. g. BA course **Time series econometrics** or MA courses mentioned below).

  If all regressors are **weakly stationary**, i. e.

  - $E[\mathbf{X}_t] = E[\mathbf{X}_s]$ and
  - $Cov(\mathbf{X}_s, \mathbf{X}_t) = Cov(\mathbf{X}_{s+k}, \mathbf{X}_{t+k})$ independent of $s, t = 1, \ldots$ and k,

  then assumption **(C3)** is also fulfilled (without proof).
- Assumption **(C4b)** requires that moments beyond the variance exist for the conditional error distribution. (Example: conditional normal distribution, t-distribution with at least 4 degrees of freedom)
- The assumptions correspond to the assumptions in Davidson (2000): Cf. to **(C2a)** (Davidson 2000, Assumption 7.1.1), to **(C2b)** (Davidson 2000, Assumption 7.1.2), to **(C3)** Davidson (2000, 7.1.3), to **(C4b)** (Davidson 2000, Eq. (7.1.12)).

Example: Stationary AR(1) process

R code

```
# ============================== 13_5_KQ_AR1_eng.R ===================================
# Program for generating and OLS estimation of an AR(1) model
# created by : RT, 2011_01_19

graphics.off()      # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation
set.seed(42)        # Randomseed
N     <- 50         # Sample size

beta  <- c(2,0.1)   # Parameter vector
sigma <- 2          # Standard deviation of the error
y0    <- 0          # Start value of the AR(1) process


# Generate a realisation of an AR(1) process
u  <- rnorm(N,mean=0,sd=sigma)       # Draw u
y <- rep(1,N)*y0
for (t in (2:N))
{
  y[t] <- beta[1] + y[t-1] * beta[2] + u[t] # Calculate y_t
}

# Plot of the time series
```

```
plot(y,xlab="Time",ylab="y",type="l")

# Scatterplot
plot(y[1:(N-1)],y[2:N])

# Calculate the OLS estimator
ols <- lm(y[2:N]~1+y[1:(N-1)])  # Note x=y_{t-1}. Therefore y_t of t=2,...,N
summary(ols)
# ============================== End ========================================
```

Listing 13.1: ./R_code/13_5_KQ_AR1_eng.R

- **Example:** Monte Carlo simulation of the OLS estimation of an AR(1) process with the following

R code

```
# ======================== 13_5_MC_KQ_AR1_eng.R ==================================
# Program for Monte Carlo simulation
# to determine the bias of the OLS estimator in the AR(1) model
# created by : RT, 2010_11_25

graphics.off()      # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation

set.seed(42)        # Randomseed
N     <- 50         # Sample size
R     <- 1000       # Number of replications

beta  <- c(1,0.9)   # Parameter vector
sigma <- 2          # Standard deviation of the error
y0    <- 1          # Start value of the AR(1) process

# Forming a loop
beta_hat_store <- matrix(0,nrow=R,ncol=length(beta))
                    # Initialise matrix to store the OLS estimates
                    # for each realisation
for (r in (1:R))
{
  # Generate a realisation of an AR(1) process
  u <- rnorm(N,mean=0,sd=sigma)      # Draw u
  y <- rep(1,N)*y0
  for (t in (2:N))
  {
      y[t] <- beta[1] + y[t-1] * beta[2] + u[t] # Calculate y_t
  }
  # Calculate the OLS estimator
  ols <- lm(y[2:N]~y[1:(N-1)])     # Note x=y_{t-1}. Therefore y_t of t=2,...,N

  # Store the parameter estimates
  beta_hat_store[r,] <- coef(ols)
}

# Calculate the mean values of the parameter estimates

colMeans(beta_hat_store)

# Create histograms
par(mfrow=c(1,2))     # Draw two plots in a graphic window

hist(beta_hat_store[,1],breaks=sqrt(R))
hist(beta_hat_store[,2],breaks=sqrt(R))

# ========================= End ==================================
```

Listing 13.2: ./R_code/13_5_MC_KQ_AR1_eng.R

14. Generalized least squares estimator and its applications

- The simple linear model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad E(\mathbf{u}|\mathbf{X}) = \mathbf{0}, \quad Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}.$$

  In practice, the assumption of homoscedastic and uncorrelated errors is often violated.

- Generalized linear model with (strictly) exogenous regressors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad E(\mathbf{u}|\mathbf{X}) = \mathbf{0}, \quad Var(\mathbf{u}|\mathbf{X}) = E(\mathbf{u}\mathbf{u}^T|\mathbf{X}) = \boldsymbol{\Omega} \tag{14.1}$$

  where it is assumed that the covariance matrix $\boldsymbol{\Omega}$ is positive definite:

$$\begin{aligned}\boldsymbol{\Omega} &= Var(\mathbf{u}|\mathbf{X}) = E\left[(\mathbf{u} - E[\mathbf{u}|\mathbf{X}])(\mathbf{u} - E[\mathbf{u}|\mathbf{X}])^T\right] \\ &= \begin{pmatrix} Var(u_1|\mathbf{X}) & Cov(u_1,u_2|\mathbf{X}) & \cdots & Cov(u_1,u_n|\mathbf{X}) \\ Cov(u_2,u_1|\mathbf{X}) & Var(u_2|\mathbf{X}) & \cdots & Cov(u_2,u_n|\mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(u_n,u_1|\mathbf{X}) & Cov(u_n,u_2|\mathbf{X}) & \cdots & Var(u_n|\mathbf{X}) \end{pmatrix}.\end{aligned} \tag{14.2}$$

  Special cases:

  - The simple linear model is a special case: $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$.
  - If $\boldsymbol{\Omega}$ is a diagonal matrix with $\omega_t^2 = Var(u_t|\mathbf{X}) \neq \omega_s^2$ for some $s, t, s \neq t$,

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_1^2 & 0 & \cdots & 0 \\ 0 & \omega_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n^2 \end{pmatrix}, \tag{14.3}$$

    the errors are not correlated, but (conditionally) heteroscedastic.

- **(Conditional) heteroscedasticity** is present if the error variance and thus the conditional variance of the dependent variable given the information set Ω_t or parts thereof is not constant, i.e. it holds that

$$Var(u_t|\Omega_t) = \omega_t^2 \neq \sigma^2, \tag{14.4a}$$
$$Var(y_t|\Omega_t) = E\left[(y_t - E[y_t|\Omega_t])^2 |\Omega_t\right] = E[u_t^2|\Omega_t] = \omega_t^2. \tag{14.4b}$$

  - **Heteroscedasticity**: ω_t^2 is a function of deterministic regressors, e. g. time.
  - **Conditional heteroscedasticity**: ω_t^2 is a function of regressors that are random variables.

    Examples:

    - The variance of exports depends on the GDP of the exporting country.
    - The variance of consumption expenditure depends on the level of income.

A rather general model for $E[u_t^2|\Omega_t]$ is:

$$Var(u_t|\Omega_t) = E[u_t^2|\Omega_t] = h(\delta + \mathbf{Z}_t\boldsymbol{\gamma}), \quad \mathbf{Z}_t \in \Omega_t. \tag{14.5}$$

There are three cases:

- The function $h(\cdot)$ is known including all parameter values for $\delta, \boldsymbol{\gamma}$, then use the GLS estimator (14.7), see section 14.1.
- The function $h(\cdot)$ is parametric, but the parameters $\delta, \boldsymbol{\gamma}$ are unknown, then use the FGLS estimator (14.17), see section 14.2.1.
- The function $h(\cdot)$ is completely unknown, then use heteroscedasticity-robust standard errors, see section 14.3.

14.1. Generalized least squares estimator

- **Generalized linear least squares estimator (generalized least squares estimator (GLS estimator))**:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{y}.$$

- **Derivation**:
  - **Cholesky decomposition**: For every symmetric positive definite matrix $\mathbf{A}$ there exists a decomposition $\mathbf{B}\mathbf{B}^T$, where $\mathbf{B}$ is a unique lower triangular matrix with positive elements on the diagonal (Gentle (2007, Section 5.9.2), Lütkepohl (1996, Section 6.2.3 (2))).

- Since $\mathbf{\Omega}$ is symmetrically positive definite, there exists a unique lower triangular matrix $\mathbf{\Psi}$, so that
$$\mathbf{\Omega}^{-1} = \mathbf{\Psi}\mathbf{\Psi}^T.$$
- Multiplying the generalized linear model (14.1) from the left by $\mathbf{\Psi}^T$ yields
$$\underbrace{\mathbf{\Psi}^T\mathbf{y}}_{\mathbf{y}^*} = \underbrace{\mathbf{\Psi}^T\mathbf{X}}_{\mathbf{X}^*}\boldsymbol{\beta} + \underbrace{\mathbf{\Psi}^T\mathbf{u}}_{\mathbf{u}^*}$$
$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*, \tag{14.6}$$
where $\mathbf{\Psi}^T$ was chosen such that $E\left[\mathbf{u}^*\left(\mathbf{u}^*\right)^T|\mathbf{X}\right] = \mathbf{I}$ (verify!).
- This means that the model with the transformed variables fulfils the assumptions of the simple linear model for the covariance matrix of the error vector, so that the OLS estimator can be applied and the GLS estimator follows from this:
$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GLS} &= \left(\left(\mathbf{X}^*\right)^T\mathbf{X}^*\right)^{-1}\left(\mathbf{X}^*\right)^T\mathbf{y}^* &(14.7a)\\ &= \left(\mathbf{X}^T\mathbf{\Psi}\mathbf{\Psi}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Psi}\mathbf{\Psi}^T\mathbf{y} &(14.7b)\\ &= \left(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{y}. &(14.7c)\end{aligned}$$
- The GLS estimator can also be derived directly from the (theoretical) moment conditions
$$\mathbf{X}^T\mathbf{\Psi}\ \left(\mathbf{\Psi}^T\mathbf{y} - \mathbf{\Psi}^T\mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0}$$
$$\mathbf{X}^T\mathbf{\Omega}^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0} \tag{14.8}$$
or from the minimisation of the SSR of the model (14.6).

- **Assumptions for determining the estimation properties**

  (cf. section 11.1 on OLS assumptions)

  - **(B1)** The model is correct, i. e. the DGP is contained in the model (14.1).
  - **(B2')** $\mathbf{u}|\mathbf{X} \sim (\mathbf{0}, \mathbf{\Omega})$.
  - **(B3)** No perfect collinearity in the regressor matrix $\mathbf{X}$.
  - **(B4')** $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Omega})$.

  Note that the assumptions **(B2')** or **(B4')** are weaker than the original assumptions **(B2)** or **(B4)**. Conditional on $\mathbf{X}$, heteroscedasticity as well as autocorrelation for time series data can occur in the errors.

- **Estimation properties of the GLS estimator in finite samples**:

  - Under **(B1)**, **(B2a)** and **(B3)** the GLS estimator is **unbiased**
$$E\left(\hat{\boldsymbol{\beta}}_{GLS}\right) = \boldsymbol{\beta}.$$

- Under **(B1)**, **(B2')** and **(B3)** the GLS estimator has covariance matrix

$$\begin{aligned} Var\left(\hat{\boldsymbol{\beta}}_{GLS}|\mathbf{X}\right) &= \left((\mathbf{X}^*)^T\,\mathbf{X}^*\right)^{-1} \\ &= \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \end{aligned} \qquad (14.9)$$

  and is **BLUE**, i. e. **efficient**. Proof see below for generalized moment estimator.

- **Generalized moment estimator**: For a given sample, let the $(1 \times k)$ vectors of variables $\mathbf{W}_t = \begin{pmatrix} W_{t1} & W_{t2} & \cdots & W_{tk} \end{pmatrix}$, $t = 1, \ldots, n$, be summarised in the matrix $\mathbf{W}^T = \begin{pmatrix} \mathbf{W}_1^T & \mathbf{W}_2^T & \cdots & \mathbf{W}_n^T \end{pmatrix}$. Under the assumption/property

$$E(\mathbf{u}|\mathbf{X}, \mathbf{W}) = \mathbf{0},$$

  a moment estimator is obtained by estimating the theoretical moments $E[\mathbf{W}_t^T u_t] = \mathbf{0}$ based on the resulting moment conditions for a given sample with

$$\mathbf{W}^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0}.$$

  One obtains:

$$\tilde{\boldsymbol{\beta}}_{\mathbf{W}} = \left(\mathbf{W}^T\mathbf{X}\right)^{-1}\mathbf{W}^T\mathbf{y}.$$

  This results in the covariance matrix

$$Var(\tilde{\boldsymbol{\beta}}_{\mathbf{W}}|\mathbf{X}, \mathbf{W}) = \left(\mathbf{W}^T\mathbf{X}\right)^{-1}\mathbf{W}^T\boldsymbol{\Omega}\mathbf{W}\left(\mathbf{X}^T\mathbf{W}\right)^{-1}.$$

  GLS is a special moment estimator (cf. (14.8)) with

$$\mathbf{W} = \boldsymbol{\Omega}^{-1}\mathbf{X}.$$

  The difference between the precision of a generalized moment estimator and the precision of the GLS estimator is positive semidefinite.

  Since every linear unbiased estimator $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$ with $\mathbf{A}\mathbf{X} = \mathbf{I}$, cf. section 9.4, can be represented as a moment estimator (due to $\mathbf{y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{u}}$ follows $\mathbf{A}\tilde{\mathbf{u}} = \mathbf{0}$), the GLS estimator is therefore efficient.

- **Calculating GLS estimators**

  - If n is large, saving and inverting $\boldsymbol{\Omega}$ requires a lot of memory ($n = 10000$ requires 1600 MB, for example). Therefore better: Apply $\boldsymbol{\Psi}$ beforehand without saving $\boldsymbol{\Psi}$ (if possible).

  - **Weighted least squares (WLS) estimator**
    u_t heteroscedastic and uncorrelated (i. e. $\boldsymbol{\Omega}$ diagonal). This means that $\boldsymbol{\Omega}$ is diagonal (14.3) and the approach (14.6) is

$$\frac{y_t}{\omega_t} = \frac{1}{\omega_t}\mathbf{X}_t\boldsymbol{\beta} + \frac{u_t}{\omega_t}$$

    with $Var(u_t/\omega_t|\mathbf{X}) = 1$. **Interpretation, calculation and notes**:

- ∗ Observations with large error variance receive less weight.
- ∗ How to select the weights? Depending on the data structure, e. g. by a linear combination of explanatory variables (example: income level) or averages in different groups.
- ∗ Calculate R^2 for weighted OLS estimation with model (14.6), since the estimated residuals are orthogonal to $\mathbf{\Psi}^T\mathbf{X}$, but not to $\mathbf{X}$.
- ∗ For the unweighted estimation, it is best to use (7.33).

- **Asymptotic estimation properties of the GLS estimator**

$$\hat{\beta}_{GLS} = \left(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{y} = \beta_0 + \left(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{u}.$$

  The assumptions **(A1)**, **(A2)** resp. **(A3)** must be modified accordingly so that an LLN and a CLT apply analogously:

  - **(A1')** $\operatorname{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X} = \mathbf{S}_{\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}}$, $\quad \mathbf{S}_{\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}}$ has full rank.
  - **(A2')** A LLN applies for $\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{u}/n$.
  - **(A3')** $\frac{\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{u}}{\sqrt{n}} \xrightarrow{d} N\left(\mathbf{0}, \mathbf{S}_{\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}}\right).$

  Then it can be shown with the already known procedure that the GLS estimator is

  - **consistent** (**(B1)**, **(B3)** **(A1')**, **(A2')**) and
  - **asymptotically normally distributed** (**(B1)**, **(B3)** **(A1')**, **(A3')**):

$$\sqrt{n}\left(\hat{\beta}_{GLS} - \beta_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{S}^{-1}_{\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}}\right).$$

- Often $\mathbf{\Psi}$ or $\mathbf{\Omega}$ is unknown and must be estimated. Then the GLS estimator is not applicable and must be replaced by the following estimator.
- `R` command: `lm( ,weights=)`, where `weights` must be passed the weights $1/\omega_t^2$.

14.2. Feasible GLS

- If the error covariance matrix $\mathbf{\Omega}$ is unknown, it must be modelled.
- **Asymptotic properties of FGLS**
  - In short, the FGLS estimator

$$\hat{\beta}_{FGLS} = \left(\mathbf{X}^T\widehat{\mathbf{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{\Omega}}^{-1}\mathbf{y} \tag{14.10}$$

    is consistent and asymptotically normally distributed if the error covariance matrix $\mathbf{\Omega}$ is correctly specified and can be consistently estimated.

- The proof is more complex, but the idea is quite simple. It results from the asymptotic properties of the GLS estimator. These are retained for the FGLS estimator if the following applies

$$\text{plim}_{\,n\to\infty}\frac{1}{n}\mathbf{X}^T\widehat{\mathbf{\Omega}}^{-1}\mathbf{X} = \text{plim}_{\,n\to\infty}\frac{1}{n}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}, \tag{14.11a}$$

$$\text{plim}_{\,n\to\infty}\frac{1}{n}\mathbf{X}^T\widehat{\mathbf{\Omega}}^{-1}\mathbf{u} = \text{plim}_{\,n\to\infty}\frac{1}{n}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{u}. \tag{14.11b}$$

  - In general, this requires that the $\boldsymbol{\beta}$ vector is consistently estimated in the first stage, so that from the errors consistently estimated by $\hat{\mathbf{u}}$, e. g. $\boldsymbol{\gamma}$ in (14.5) and thus $\mathbf{\Omega}$ can be consistently estimated.

  - **Attention**: If $\mathbf{\Omega}$ is not diagonal, the OLS estimator for the first stage is generally inconsistent! In this case, other estimation methods are necessary.

14.2.1. Modelling of heteroscedastic errors

- An often suitable model for modelling $Var(u_t|\Omega_t) = \omega_t^2$ is

$$E[u_t^2|\Omega_t] = e^{\delta+\mathbf{Z}_t\boldsymbol{\gamma}} = e^{\delta}e^{\mathbf{Z}_t\boldsymbol{\gamma}}. \tag{14.12}$$

The equation (14.12) specifies the function $h(\cdot)$ in (14.5) as $h(\cdot) = \exp(\cdot)$.

If it is specified for a random variable v_t that

$$E[v_t|\Omega_t] = 0 \quad \text{and} \quad Var(v_t|\Omega_t) = 1,$$

u_t^2 can be written as

$$u_t^2 = e^{\delta+\mathbf{Z}_t\boldsymbol{\gamma}}v_t^2, \tag{14.13}$$

so that (14.12) applies. In order to estimate δ and $\boldsymbol{\gamma}$ with a linear regression, (14.13) is logarithmised:

$$\ln u_t^2 = \delta + \mathbf{Z}_t\boldsymbol{\gamma} + \ln v_t^2. \tag{14.14}$$

Since $E[\ln v_t^2] \neq \ln E[v_t^2] = 0$ (Jensen inequality, see section 2.7), the following trick is used:

$$\ln u_t^2 = \underbrace{\delta + E[\ln v_t^2]}_{\delta'} + \mathbf{Z}_t\boldsymbol{\gamma} + \underbrace{\left(\ln v_t^2 - E[\ln v_t^2]\right)}_{\eta_t \text{ where } E[\eta_t|\Omega_t]=0.}$$

$$\ln u_t^2 = \delta' + \mathbf{Z}_t\boldsymbol{\gamma} + \eta_t. \tag{14.15}$$

- **2-stage estimator:**

1st step: Estimate the model (14.1) with OLS and save the residuals

$$\hat{\mathbf{u}} = \mathbf{M}_{\mathbf{X}}\mathbf{y}.$$

Insert the residuals into (14.15) for OLS estimation of δ' and $\boldsymbol{\gamma}$:

$$\ln \hat{u}_t^2 = \delta' + \mathbf{Z}_t\boldsymbol{\gamma} + errors. \tag{14.16}$$

2nd step: Estimate (14.6) approach with $\hat{\omega}_t^2 = \exp(\mathbf{Z}_t\hat{\boldsymbol{\gamma}})$:

$$\frac{y_t}{\hat{\omega}_t} = \frac{1}{\hat{\omega}_t}\mathbf{X}_t\boldsymbol{\beta} + \frac{u_t}{\hat{\omega}_t}. \tag{14.17}$$

The factor $e^{\hat{\delta}'}$ can be omitted as it is constant for all observations.

- **FGLS or OLS with heteroscedasticity-robust standard errors?** The question is how well $\boldsymbol{\Omega}$ can be estimated. The more imprecise, the more likely it is to use the OLS estimator with heteroscedasticity-robust variance-covariance matrix, see section 14.3.

- It is possible to iterate the FGLS estimator. This has *no* impact on the asymptotic properties, but on the estimation properties in finite samples.

14.2.2. Models with autocorrelated errors

See Davidson & MacKinnon (2004, Sections 7.6-7.9).

14.3. Heteroscedasticity-robust standard errors in OLS estimation

- **Derivation of heteroscedasticity-robust standard errors**

  If heteroscedastic errors are present, then the variance-covariance matrix of the OLS estimator is given by (9.7):

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\mathbf{u}|\mathbf{X})\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} & (9.7)\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}. & (14.18)\end{aligned}$$

  This variance-covariance matrix is often referred to as **sandwich covariance matrix**, where $(\mathbf{X}^T\mathbf{X})^{-1}$ represents the "‘bread slices"’. The variance-covariance matrix of inefficient estimators often have this form.

- An alternative representation of the "‘filling"’ is

$$\mathbf{X}^T\boldsymbol{\Omega}\mathbf{X} = \sum_{t=1}^{n}\omega_t^2\mathbf{X}_t^T\mathbf{X}_t.$$

  Since $E[u_t^2|\mathbf{X}] = \omega_t^2$, ω_t^2 can be estimated by the "‘average based on one observation"’ u_t^2. This is of course not a very good estimator, but for our purpose it does the job. Since u_t is unknown, we take the residual $\hat{u}_t$.

  Accordingly, the covariance matrix (14.18) of the OLS estimator for heteroscedasticity can be estimated using

$$\widehat{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{t=1}^{n}\hat{u}_t^2\mathbf{X}_t^T\mathbf{X}_t\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{14.19}$$

- Remarks:
  - The standard errors obtained from (14.19) are called **heteroskedasticity-robust standard errors** or **White standard errors**. The latter term goes back to Halbert White, an econometrician at the University of California in San Diego.
  - For an individual $\hat{\beta}_j$, the heteroscedasticity-robust standard error can be smaller or larger than the usual OLS standard error.
  - It can be shown that the OLS estimator $\hat{\boldsymbol{\beta}}$ no longer has a known finite sampling distribution if heteroscedasticity-robust standard errors are used. However, it is **asymptotically normally distributed** under quite general conditions. The critical values and the p-values therefore remain approximately valid if (14.19) is used.
  - In Davidson & MacKinnon (2004, Section 5.5) it is explained why (14.19) is a consistent estimator of (14.18).
  - The OLS estimator is **unbiased** and **consistent** regardless of the choice of standard errors (White or non-White), since the assumptions **(B1)**, **(B2a)**, **(B3)** remain unaffected by heteroskedasticity.
  - However, the OLS estimator is (asymptotically) **not efficient** in the case of heteroscedastic errors, as it can be shown that the difference between the (asymptotic) precision of the OLS estimator and the (F)GLS estimator is positive semidefinite. Therefore, if something is known about the functional form of heteroscedasticity and there are sufficient sample observations, the FGLS estimator should be used.
- **Alternative estimators** of (14.18) and their names in Davidson & MacKinnon (2004, Section 5.5) and the `R` packages `car` or `sandwich`.
  - "‘HC0"’: White standard errors (14.19).
  - "‘HC1"’: Multiplies White standard errors (14.19) by $n/(n-k)$. (Default in EViews.)
  - "‘HC2"’: Replaces $\hat{u}_t^2$ White standard errors (14.19) with $\hat{u}_t^2/(1-h_t)$, where h_t is the tth diagonal element of $\mathbf{P_X}$.
  - "‘HC3"’: Replaces $\hat{u}_t^2$ White standard errors (14.19) with $\hat{u}_t^2/(1-h_t)^2$, where h_t is the tth diagonal element of $\mathbf{P_X}$.

  All corrections aim to correct the underestimation of the error variance by using the residuals instead of the errors, see section 9.5. For a more detailed explanation of the respective corrections, see Davidson & MacKinnon (2004, Section 5.5).
- `R` commands for calculating heteroscedasticity-robust variance-covariance matrices:
  - Package `car`: `hcmm(model,type="hc1")`
  - Package `sandwich`: `vcovHC(model,type="HC1")`

- R commands for calculating heteroscedasticity-robust test statistics with package car:
  - coeftest(model,vcov=hccm(model,type="hc1")) provides usual regression output with heteroscedasticity-robust standard errors.
  - linearHypothesis(,vcov=hccm(model,type="hc1")) provides F-test with heteroscedasticity-robust variance-covariance matrix.

14.4. Empirical analysis of trade flows: Part 4

Continuation of the analysis of model (11.49):

$$\begin{aligned} \ln(Imports_i) = \beta_1 + \beta_2 \ln(GDP_i) + \beta_3 \ln(Distance_i) \\ + \beta_4\, Openness_i + \beta_5 \ln(Area) + u_i. \end{aligned} \tag{11.49}$$

- Eliminate **missing values** or **not a number (NAN)** or **not available/not applicable (NA)** (in R): In the data set used, there is a NA for the dependent variable Imports, which leads to the residual vector having fewer rows than the regressor matrix in the further course of the R program. It therefore makes sense to eliminate this observation from the data frame before starting the estimations. If the original data frame is denoted with daten_all, this can be done with the command

```
daten <- daten_all[!is.na(daten$trade_0_d_o),]
```

  Only then use the command attach(daten) so that R searches in the correct data frame!

- **FGLS estimation**

  R code (Extract from R program in section A.4)

```
      col = "blue", lwd = 2)
if (save.pdf) dev.off()

################################################################################
# Section 14.4 FGLS and heteroskedasticity-robust OLS estimation
################################################################################

#### FGLS estimation for model 4
mod_4_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
  ebrd_tfes_o + log(cepii_area_o)

# 1st step
resids        <- residuals(mod_4_kq)
fits          <- fitted(mod_4_kq)
```

Listing 14.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

One obtains:

```
Call:
lm(formula = mod_4_formula, weights = 1/omega)

Weighted Residuals:
```

```
Min     1Q Median     3Q    Max
-4.799 -1.227  0.544  1.174  3.006

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.20213    1.26254   1.744 0.088106 .
log(wdi_gdpusdcr_o)  1.07977    0.05715  18.893  < 2e-16 ***
log(cepii_dist)     -0.90934    0.11505  -7.904 5.54e-10 ***
ebrd_tfes_o          0.25397    0.17561   1.446 0.155201
log(cepii_area_o)   -0.20138    0.05343  -3.769 0.000485 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.897 on 44 degrees of freedom
Multiple R-squared:  0.9345,    Adjusted R-squared:  0.9286
F-statistic: 157.1 on 4 and 44 DF,  p-value: < 2.2e-16
```

- **Heteroskedasticity-robust OLS estimator**

  R code (Extract from R program in section A.4)

```
# 2nd step
omega          <- exp(fitted(lm(mod_formula_ln_u_squared)))
model_gls      <- lm(mod_4_formula, weights=1/omega)
(summary(model_gls))
```

Listing 14.2: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides

```
t test of coefficients:
Estimate Std. Error t value  Pr(>|t|)
(Intercept)          2.427777   1.337400  1.8153  0.076298 .
log(wdi_gdpusdcr_o)  1.025023   0.070679 14.5024 < 2.2e-16 ***
log(cepii_dist)     -0.888646   0.120775 -7.3579 3.428e-09 ***
ebrd_tfes_o          0.353154   0.180896  1.9522  0.057290 .
log(cepii_area_o)   -0.151031   0.050657 -2.9814  0.004662 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Summarising the results in a table**: **Output table for model (11.49) for different estimators**

  Result: Both the parameter estimates themselves and the standard errors do not differ fundamentally. Possible cause: There is no heteroscedasticity in the error variances.

Dependent variable: ln(*Imports to Germany*)		
Independent variables/model	OLS	FGLS
Constant	2.427	2.024
	(2.132)	(1.236)
	[1.337]	
$ln(GDP)$	1.025	1.080
	(0.076)	(0.057)
	[0.070]	
$ln(Distance)$	-0.888	-0.888
	(0.156)	(0.110)
	[0.120]	
$Openness$	0.353	0.263
	(0.206)	(0.179)
	[0.180]	
$ln(Area)$	-0.151	-0.203
	(0.085)	(0.048)
	[0.050]	
Sample size	49	49
R^2	0.906	0.9055
Standard error of the regression	0.853	
Residual sum of squares	32.017	
AIC	2.6164	
HQ	2.6896	
SC	2.8094	

Notes: OLS or FGLS standard errors in round brackets, White standard errors in square brackets.

- Continuation in section 15.7.

15. Model checking

Overview of the modelling process, see section 4.3.

Why is model checking necessary?

Properties of estimation and test procedures only apply under the assumptions made! $\Longrightarrow$ Checking these assumptions is essential by carrying out statistical tests!

Review (see chapter 11.3, p. 219)

Applications of exact tests:

- **Specification of the normal linear regression model and checking the assumptions**, cf. section 11.1

  - **(B1)** and $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ (**(B2a)**): $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ contains DGP
    - t-tests, see section 11.3.1; F-tests, see section 11.3.2.
    - Testing the correct functional form, e. g. with **RESET test**, see section 15.3.
    - Testing for parameter stability, e. g. with **Chow test**, see (11.34) in sections 11.3.2.
  - **(B3)**: $\mathbf{X}^T\mathbf{X}$ has rank k: violation leads to error message "‘singular matrix"’.
  - **(B4)**: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$:
    - Assumes $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$, see above.
    - Assumes: $Var(\mathbf{u}|\mathbf{X}_t) = \sigma^2$ (**Homoscedasticity**): **Tests for heteroscedasticity**, see section 15.2.
    - Requires normally distributed errors: **Lomnicki-Jarque-Bera test**, see section 15.4.

- **Testing economic hypotheses**

15.1. Tests for autocorrelation in the errors

15.2. Tests for heteroscedastic errors

- As already mentioned, it does not make sense to use the FGLS estimator (14.17) "automatically". If the errors are homoscedastic, the OLS estimator with the usual OLS errors should be used.
- You should therefore first test whether there is statistical evidence for heteroscedasticity.
- Two different tests are presented below: The Breusch-Pagan test and the White test. Both have "homoscedastic errors" as null hypothesis.
- In `R` the Breusch-Pagan test is contained in the package `lmtest`.

It is assumed that the assumptions for unbiasedness or consistency of the OLS estimator are fulfilled for the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

The hypothesis test refers to the validity of **(B2b)** or **(C2b)**, i. e. to the presence of homoscedasticity.

The **hypothesis pair** to be tested is

$$\begin{aligned}
&\mathrm{H}_0: Var(u_t|\mathbf{X}_t) = \sigma^2 \quad \text{(Homoscedasticity)},\\
&\mathrm{H}_1: Var(u_t|\mathbf{X}_t) = \omega_t^2 \neq \sigma^2 \quad \text{(Heteroscedasticity)}.
\end{aligned}$$

The basic idea of heteroscedasticity tests is that under the null hypothesis, no regressor should have explanatory power for $Var(u_t|\mathbf{X}_t)$. If the null hypothesis does not hold, the conditional variance $Var(u_t|\mathbf{X}_t)$ can be determined by (almost) any function of the regressors x_{tj}, $(1 \leq j \leq k)$ or other regressors.

Beachte: The Breusch-Pagan test and the White test differ with regard to their alternative hypothesis.

15.2.1. Breusch-Pagan test

- Idea: Let's consider the regression

  $$u_t^2 = \delta_0 + \delta_1 x_{t1} + \cdots + \delta_k x_{tk} + v_t, \quad t = 1, \ldots, n. \tag{15.1}$$

  Under assumptions **(B1)**,**(B2a)**,**(B3)**, the OLS estimator for the δ_j's is unbiased.

  The pair of hypotheses is therefore:

  $$\begin{aligned}
  &\mathrm{H}_0: \delta_1 = \delta_2 = \cdots = \delta_k = 0 \quad \text{versus}\\
  &\mathrm{H}_1: \delta_1 \neq 0 \text{ and/or } \delta_2 \neq 0 \text{ and/or } \ldots,
  \end{aligned}$$

since under H$_0$ it holds that $E[u_t^2|\mathbf{X}] = \delta_0$.

- **Differences to the previous application of the F-test**:
  - The squared errors u_i^2 are not normally distributed under any circumstances because they are squared values and therefore cannot be negative. This means that the v_i cannot be normally distributed either and the F-distribution of the F-statistic is not exactly valid for finite samples.

    An asymptotic F-test must therefore be used. With the results from section 11.4 and appropriate regularity assumptions, it follows that k times the F-statistic is asymptotically $\chi^2(k)$ distributed.
  - The errors u_i are unknown. However, they can be replaced by the residuals $\hat{u}_i$ of the OLS estimation without affecting the asymptotic validity of the F-test. The reason for this is that the errors are consistently estimated by the residuals if the parameters are consistently estimated. (The formal proof is quite complex and is omitted here.)
- One can also use the R^2 version of the test statistic. Note that the R^2 is zero due to SSR = SST if a constant is used as the only regressor (there is then no regressor with variation). We denote the coefficient of determination of the OLS estimation from (15.1) by $R^2_{\hat{u}^2}$ and obtain
$$F = \frac{R^2_{\hat{u}^2}/k}{(1-R^2_{\hat{u}^2})/(n-k)}.$$
  The test statistic of the overall F-test, which tests for the joint significance of all regressors, is output by default by most software programmes.
- H$_0$ is rejected if F or kF exceeds the critical value for a chosen significance level based on the $F_{k,n-k}$ or $\chi^2(k)$ distribution (or if the p-value is smaller than the significance level).
- In `R` the test is performed with the command `bptest()` from the package `lmtest` and returns the kF-statistic, which is asymptotically $\chi^2(k)$-distributed.
- **Note**:
  - If one suspects that the heteroscedasticity is caused by special variables that were not previously considered in the regression, these can be added to the regression (15.1).
  - If H$_0$ is not rejected, this does not automatically mean that the u_i's are homoscedastic. If the specification (15.1) does not contain all relevant variables that could cause heteroscedasticity, it can happen that all δ_j, $j = 1, \dots, k$ are jointly insignificant.
  - A variant of the Breusch-Pagan test is a test for multiplicative heteroscedasticity, i. e. the variance has the form $\sigma_t^2 = h(\delta + \mathbf{X}_t\boldsymbol{\beta})$. If, for example, the case $h(\cdot) = \exp(\cdot)$ is assumed, the test equation
$$\ln(\hat{u}_t^2) = \delta' + \mathbf{X}_t\boldsymbol{\beta} + errors. \tag{14.16}$$
    is obtained.

15.2.2. White test

- **Background**:
  In order to derive the asymptotic distribution of the OLS estimator, the assumption of homoscedastic errors ((**B2b**) or (**C2b**)) is *not* required.

  It is already sufficient that the squared errors u_t^2 are uncorrelated with all regressors, their squares and their cross products.

  This can be tested quite easily with the following regression, where the unknown errors have already been replaced by the residuals:

$$\begin{aligned}\hat{u}_t^2 = \delta_0 &+ \delta_1 x_{t1} + \cdots + \delta_k x_{tk} \\ &+ \delta_{k+1} x_{t1}^2 + \cdots + \delta_{J_1} x_{tk}^2 \\ &+ \delta_{J_1+1} x_{t1} x_{t2} + \cdots + \delta_{J_2} x_{tk-1} x_{tk} \\ &+ v_t, \quad t = 1, \ldots, n. \end{aligned} \tag{15.2}$$

- The pair of hypotheses is:

$$\begin{aligned} \mathrm{H}_0 &: \delta_j = 0 \text{ für } j = 1, 2, \ldots, J_2, \\ \mathrm{H}_1 &: \delta_j \neq 0 \text{ for at least one } j. \end{aligned}$$

  An F-test can be used again, whose distribution is approximately the F-distribution (asymptotic distribution).

- If you have many regressors, it is tedious to carry out the F-test for (15.2) by hand. Most software programmes already provide the White test.

- If k is large, a large number of parameters must be estimated when performing the White test. This can hardly be realised in small samples. In this case, only the squares x_{tj}^2 are included in the regression and all cross products are neglected.

- **Note**: If the null hypothesis is rejected, this may be because
  - the errors are heteroscedastic and/or
  - the model is not specified correctly.

- The White test is not automatically available in `R`. A separate programme `whitetest()` can be found in section B.2.

15.3. Test for correct specification of the functional form: RESET test

RESET test (REgression Specification Error Test)

Idea and implementation:

- The RESET test is used to check whether the present regression model
$$y_t = \mathbf{X}_t \boldsymbol{\beta} + u_t$$
is correctly specified, i. e. whether the assumptions **(B1)** and **(B2a)** or alternatively the assumptions **(C1)** and **(C2a)** are valid and thus
$$E[y_t|\Omega_t] = \mathbf{X}_t \boldsymbol{\beta} \tag{10.1}$$
holds. Cf. chapter 10.

- Every term that is added to the model should therefore be insignificant. Thus, every non-linear function of independent variables should also be insignificant.

- Therefore, the null hypothesis of the RESET test is formulated in such a way that the significance of non-linear functions of the fitted values $\hat{y}_t = \mathbf{X}_t \hat{\boldsymbol{\beta}}$ added to the model can be tested. Note that the fitted values represent a non-linear function of the regressors of the initial model.

- In practice, the second and third power of $\hat{y}_t$ turned out to be sufficient to be able to perform the RESET test:
$$y_t = \mathbf{X}_t \boldsymbol{\beta} + \alpha \hat{y}_t{}^2 + \gamma \hat{y}_t{}^3 + errors. \tag{15.3}$$
The pair of hypotheses is
$$\begin{aligned} &\mathrm{H}_0 : \alpha = 0, \gamma = 0 \quad \text{(linear model is correctly specified)} \\ &\mathrm{H}_1 : \alpha \neq 0 \text{ and/or } \gamma \neq 0. \end{aligned}$$
This null hypothesis is tested using an F-test with 2 degrees of freedom in the numerator and $n-k-2$ in the denominator, whereby the resulting critical value is only asymptotically correct.

- **Note**: If the null hypothesis that the initial model is correctly specified is rejected, this can have a number of causes:
  - The functional form is non-linear.
  - Relevant regressors are missing.
  - Heteroscedasticity is present.

- `R` command: `resettest()`, whereby the second and third power are taken into account without further specifications (requires `R` package `lmtest`).

- See Davidson & MacKinnon (2004, Section 15.2) for more details.

15.4. Normality test: Lomnicki-Jarque-Bera test

- See Davidson & MacKinnon (2004, Section 15.2) for a detailed explanation.

- In `R` the Lomnicki-Jarque-Bera test is performed with the command `jarque.test()` from the package `moments`.

15.5. Stability tests

Chow test

see (11.34) in section 11.3.2.

15.6. Summary of an econometric modelling process

15.7. Empirical analysis of trade flows: Part 5

Continuation of section 14.4.

- **RESET test** of model 4:

  R code (Extract from R program in section A.4)

```
# see section  14.3
(coeftest(mod_4_kq,vcov=hccm(mod_4_kq,type="hc1")))
```

Listing 15.1: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides:

```
        RESET test

data:  mod_4_kq
RESET = 7.14, df1 = 2, df2 = 42, p-value = 0.002142
```

- **Breusch-Pagan test for heteroscedasticity** of model 4:

  R code (Extract from R program in section A.4)

```
################################################################################
# Section 15.7
```

Listing 15.2: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides:

```
        studentized Breusch-Pagan test

data:  mod_4_kq
BP = 4.2779, df = 4, p-value = 0.3697
```

- **White test with cross products for heteroscedasticity** with the OLS residuals:

  R code (Extract from R program in section A.4)

```
#### Perform the RESET test for model 4 with
resettest(mod_4_kq)

#### Performing the Breusch-Pagan test for model 4
bptest(mod_4_kq)


################################################################################
#                   Start function whitetest
################################################################################
# White test for homoskedastic errors with cross products
# RW, 2011_01_26

whitetest <- function(model){

  # Extract data from model
  dat <- model$model
  dat$resid_sq <- model$resid^2

  # Create formula for auxiliary regression
  regr <- attr(model$terms, "term.labels")
  form <- as.formula(paste("resid_sq~(",paste(regr,collapse="+"),")^2+",paste("I(",regr,"^2)",collapse="+"))
     )

  # Estimate auxiliary regression
  test_eq <- lm(form,data=dat)

  # Overall F-test
  fstat <- summary(test_eq)$fstatistic

  # Calculate and display result
  result1 <- c(fstat[1],fstat[2],fstat[3],pf(fstat[1],fstat[2],fstat[3],lower.tail=FALSE))
  names(result1) <- c("F-Statistic","df1","df2","P-Value")
  result <- list(result1,test_eq)
  return(result)
}
```

Listing 15.3: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

provides:

```
[[1]]
F-Statistic          df1          df2      P-Value
  0.5239004   14.0000000   34.0000000    0.9016863

[[2]]

Call:
lm(formula = form, data = dat)

Coefficients:
                          (Intercept)                    log(wdi_gdpusdcr_o)                         log(
    cepii_dist)
                            -77.25103                                3.91926
    4.08526
             I(log(wdi_gdpusdcr_o)^2)                    I(log(cepii_dist)^2)                        I(ebrd_
    tfes_o^2)
                             -0.02898                                0.18986
    0.20200
      log(wdi_gdpusdcr_o):ebrd_tfes_o  log(wdi_gdpusdcr_o):log(cepii_area_o)          log(cepii_dist):ebrd
    _tfes_o
                             -0.71004                                0.08044
    -0.98879
```

```
                   ebrd_tfes_o                      log(cepii_area_o)
                      15.48882                               -2.91347
          I(log(cepii_area_o)^2)  log(wdi_gdpusdcr_o):log(cepii_dist)
                      -0.04813                               -0.06623
log(cepii_dist):log(cepii_area_o)       ebrd_tfes_o:log(cepii_area_o)
                      -0.07797                                0.61849
```

- **Results**:
  - The RESET test provides a rejection of the null hypothesis of correct specification at the 1% significance level. This means that quadratic terms may play an explanatory role, e. g. `I((log(wdi_gdpusdcr_o))`. However, taking this additional regressor into account does not lead to an insignificant RESET test statistic. This may be due to outliers.
  - Both the Breusch-Pagan test and the White test do *not* reject the null hypothesis of homoscedastic OLS residuals at any useful significance level. Thus, the use of heteroscedasticity-robust standard errors or FGLS in section 14.4 was not efficient.
  - Breusch-Pagan test and White test also do not reject the null hypothesis of homoscedastic standardised FGLS errors. The p-values increase again significantly to over 50%.
  - A final model has not yet been found due to the strong rejection of the RESET test, even with a quadratic regressor.

A. R Programs

A.1. R programs for Graphs in Section 2.5 to Distribution and Density Functions

CDF and PDF of the standard normal distribution, see figure 2.1

```r
# Distribution and density function of the standard normal distribution
# KK, 21.10.2010, RT, 29.09.2015 (pdf- instead of eps-graphs)

#######
# Density function of the standard normal distribution
#######


# Open the graph output and specify that save as .eps-file
# (filename, size of the graph)
#postscript("pdf_std_normal.eps", height=4, width=6, horizontal=FALSE)
# or save as .pdf
pdf("pdf_std_normal_eng.pdf", height=4, width=6)


# Parameters for graphs: (optional)
#     las=1: Axis scaling horizontal
#     mai: width of margins (bottom, left, top, right)
#     mgp: position of axes, axis scaling and axis labelling
par(las=1, mai=c(0.6,0.1,0.1,0.1), mgp=c(1.5,1,0))


# Plot the two points (-3.5,0) and (3.5,0.48) (-> proportions of the graph)
#     type="n": empty plot
#     bty="n": no box around the graph
#     xaxt="n", yaxt="n": no x- and y-axis
#     xlab="x", ylab="": x-axis labelling is x, y-axis unlabelled
plot(c(-3.5,3.5), c(0,0.48), type="n", bty="n", xaxt="n", yaxt="n",
     xlab="x", ylab="")


# Complement the axes (1 x-axis, 2 y-axis)
#     pos=0: Axis goes through 0
#     labels: Axis scaling
#     at: Positions of the axis scaling
axis(1, pos=0, labels=-3:3, at=-3:3)
axis(2, pos=0, labels=1:4/10, at=1:4/10)
#     x-axis too short -> extend with line at y=0 (h horizontal line)
#     y-axis too short -> draw a line from 0 to 0.44
abline(h=0)
lines(c(0,0), c(0,0.44))
# y-axis labelling
text(0, 0.472, expression(phi(x)))


# Plot the function
#     dnorm: density of the normal distribution (analogously pnorm: distribution function)
#     from, to: range in which the function is plotted
#     add=TRUE: plots into existing graph window
plot(function(x) dnorm(x), from=-3.5, to=3.5, add=TRUE)
```

```
# Close the current graphs window (e.~g. .eps or .pdf file)
dev.off()


#######
# Distribution function of the standard normal distribution
#######


# postscript("cdf_std_normal.eps", height=4, width=6, horizontal=FALSE)
pdf("cdf_std_normal_eng.pdf", height=4, width=6)

par(las=1, mai=c(0.6,0.1,0.1,0.1), mgp=c(1.5,1,0))

plot(c(-3.5,3.5), c(0,1.2), type="n", bty="n", xaxt="n", yaxt="n",
     xlab="x", ylab="")

axis(1, pos=0, labels=-3:3, at=-3:3)
axis(2, pos=0, labels=c("",0.5,1), at=c(0,0.5,1))
abline(h=0); lines(c(0,0), c(0,1.1))
text(0, 1.18, expression(Phi(x)))

plot(function(x) pnorm(x), from=-3.5, to=3.5, add=TRUE)

dev.off()
```

Listing A.1: ./R_code/2_5_Plot_PDF_CDF_StNormal_eng.R

0.85-**quantile of the standard normal distribution**, see figure 2.2

```
# R program on quantiles, section confidence intervals
# RT,KK, 01.02.2011

alpha   <- 0.85         # Set probability for quantile
dev.off()             # Close all graph windows
split.screen(c(2,1))  # Split a graph window
# Plot density function
screen(1)
plot(function(x) dnorm(x), from=-4, to=4, lwd=2, ylab="Density",
     main="Standard normal distribution")
abline(h=0)
abline(v=qnorm(alpha), col="red")

# Plan: Draw a polygon (which can then be coloured).
# Polygon first needs all x-values, then all y-values
#   these are then connected
# x-values: from -4 to qnorm(alpha) (-> x_tmp)
# y-values: density values for the x-values
x_tmp   <- seq(from=-4, to=qnorm(alpha), length.out=1000)
polygon(c(x_tmp, x_tmp[length(x_tmp)]),     # last double (point q_alpha,0)
        c(dnorm(x_tmp), 0),             # last 0 (Punkt y_alpha,0)
        border=NA,                  # no border
        density=10,                 # dashed, 10%
        col="blue")                 # colour
# Plot probability function
screen(2)
plot(function(x) pnorm(x), from=-4, to=4, lwd=2, ylab="Probability function",
     main="Standard normal distribution")
abline(h=alpha)
abline(v=qnorm(alpha),col="red")
```

Listing A.2: ./R_code/2_5_Plot_Quant_StNormal_eng.R

PDF of the bivariate normal distribution, see figure 2.3

```
# Density function of the bivariate normal distribution
# RT, 21.10.2010, 25.10.2010

rm(list = ls()) # cleans workspace

# install package mnormt if not installed yet
if(!require(mnormt)){
  install.packages("mnormt")
}

library(mnormt)    # load package mnormt

# parameters of bivariate normal distribution
mu_1         <- 0
mu_2         <- 0

sigma_1      <- 1
sigma_2      <- 1
rho          <- 0.0

# determine mean vector

Mean         <- c(mu_1,mu_2)

# compute variance-covariance matrix

sigma2_1     <- sigma_1^2
sigma2_2     <- sigma_2^2
sigma_12     <- sigma_1 * sigma_2 * rho

Sigma        <- matrix(c(sigma2_1,sigma_12,sigma_12,sigma2_2),2)

# determine grid on which density is computed

x1_limit     <- mu_1 + 3*sigma_1
x2_limit     <- mu_2 + 3*sigma_2
ngridpoints  <- 50

x1           <- seq(-x1_limit,x1_limit,2*x1_limit/(ngridpoints-1))
x2           <- seq(-x2_limit,x2_limit,2*x2_limit/(ngridpoints-1))
X            <- expand.grid(x1=x1,x2=x2)

# compute density
Density      <- apply(X,1,dmnorm,mean=Mean,varcov=Sigma)

Density      <- matrix(Density,length(x1),length(x2),byrow=FALSE)

# Colors for surface = estimates
n_col        <- 80
nrDensity    <- nrow(Density)
ncDensity    <- ncol(Density)
Densitylim   <- c(min(Density),max(Density))
#couleurs    <- tail(heat.colors(trunc(1 * n_col)),n_col)
couleurs     <- topo.colors(trunc(1 * n_col))
Densitycol   <- couleurs[trunc((Density-Densitylim[1])/
                                 (Densitylim[2]-Densitylim[1])*(n_col-1))+1]
dim(Densitycol) <- c(nrDensity,ncDensity)
Densitycol      <- Densitycol[-nrDensity,-ncDensity]

# plot surface and contour lines
pdf("Biv_Normal_Surface_col_eng.pdf", height=6, width=6)
#par(mfrow=c(1,1))
#split.screen(c(2,1))
#screen(1)
par(mai=c(0.5,0.5,0.3,0.1))
persp(y=x1, x=x2, z=Density, col=Densitycol,
      main="Density of Bivariate Normal Distribution for (x1,x2)" ,
      theta=35, phi=20 , r=10, shade=0.1,   ticktype="detailed")
```

```
dev.off()

# ?contour
# screen(2)
pdf("Biv_Normal_Surface_con_eng.pdf", height=6, width=6)
contour(x1,x2,Density,nlevels=50,main="Density of Bivariate Normal Distribution  for (x1,x2)" )
dev.off()
# close.screen(all=TRUE)
```

Listing A.3: ./R_code/2_5_Plot_PDF_biv_Normal_eng.R

A.2. R Programs for Monte Carlo Simulation in the Section 5.5.1 concerning the Law of Large Numbers

```
# ======================== 5_4_MC_bar_y_LLN_CLT_eng.R ==============================
#
# Program for Monte Carlo simulation
# to illustrate the LLM and CLTs of the arithmetic mean
# Calculates mean and standard deviation over all replications
# as well as histograms
# DGP: mean + chi-squared distributed error
# Note: Program is written with for-loops for readability
# Status: RT, 2015_10_02

graphics.off()            # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation

set.seed(42)                   # Randomseed
N             <- c(10,50,100,500)   # Sample sizes
R             <- 10000          # Number of replications

mu            <- 1             # Mean value
deg_freedom   <- 1             # Degrees of freedom of the qui-squared distribution
sigma         <- 2             # Standard deviation of the error

save.pdf      <- 1             # 1=create PDFs of graphs, 0=otherwise
# Form two loops:
#   Outer loop on the number of replications
#   Inner loop on the sample size

n_max         <- N[length(N)]  # Maximum sample size
#   Initialise the output matrices
mu_hat_store  <- matrix(0,nrow=R,ncol=length(N))
mu_tilde_store  <- matrix(0,nrow=R,ncol=length(N))

for (r in (1:R))
{
  # Fenerate a realisation of a simple linear regression model
  # for the maximum sample size
  u           <- rchisq(n_max,df=deg_freedom,)       # Drawing u
  u           <- (u-deg_freedom)/sqrt(2*deg_freedom) # Standardising
  y           <- mu+u

  for (i in (1:length(N)))
  {
    # Store the estimates
    mu_hat_store[r,i]   <- mean(y[1:N[i]])   # arithm. mean
    mu_tilde_store[r,i] <- (y[1]+y[N[i]])/2  # alternative estimator
  }
}

# Calculate the arithmetic means of the parameter estimates
mu_hat_mean    <- colMeans(mu_hat_store)
mu_tilde_mean <- colMeans(mu_tilde_store)

# Calculate the variances of the parameter estimates
mu_hat_sd      <- sqrt(diag(var(mu_hat_store)))
mu_tilde_sd    <- sqrt(diag(var(mu_tilde_store)))


# Display on screen
(cbind(N,mu_hat_mean,mu_hat_sd,mu_tilde_mean,mu_tilde_sd))

# Create histograms
if (save.pdf) pdf("plot_MC_mu_hat_Konsistenz_eng.pdf", height=6, width=6)
par(mfrow=c(2,2))      # Draw four plots in a graphic window
for (i in (1:4))
```

```
{
  # Sample size N[i]
  hist(mu_hat_store[,i], breaks=sqrt(R),
       xlab=expression(hat(mu)), main=paste("Histogram for n= ",N[i],sep=""))
}
if (save.pdf) dev.off()

if (save.pdf) pdf("plot_MC_mu_tilde_Konsistenz_eng.pdf", height=6, width=6)
par(mfrow=c(2,2))     # Draw four plots in a graphic window
for (i in (1:4))
{
  # Sample size N[i]
  hist(mu_tilde_store[,i], breaks=sqrt(R),
       xlab=expression(tilde(mu)), main=paste("Histogram for n= ",N[i],sep=""))
}
if (save.pdf) dev.off()
# ========================= End ==================================
```

Listing A.4: ./R_code/5_4_MC_bar_y_LLN_CLT_eng.R

A.3. R Programs for Graphs in the Section 5.6 on Basics of Tests

Test on mean value of the DAX, see page 127

```
# ===================== 5_5_Test_Mean_DAX_eng.R ===============================#
# Program tests expected value of DAX returns.
# Data comes from Yahoo-Finance
# Options for loading data:
# a) directly from Yahoo-Finance
# b) download a csv file from Yahoo-Finance and read it into R
# c) for info: load .xlsx files into R,
#               e.g. if .csv file was converted into .xlsx file
# Status: 2015_10_01, ....
# 2021_11_25, RT substantial revision


# -------------- load packages --------------------------------------------------
    # Package for zoo data frame, which allows exact dates to be specified
if (!require(zoo)) install.packages("zoo")
library(zoo) # RT, 2021_11_24
    # Package with command for direct download of data from Yahoo-Finance
if (!require(tseries)) install.packages("tseries") # requires library zoo
library(tseries)
    # Package for loading xlsx files
if (!require(readxl)) install.packages("readxl")
library(readxl)

# -------------- set parameters ----------------------------------------------
price_to_check  <- "Close"   # "Open", "High", "Low", "Close", "
                            # Adj.Close" (only for .csv or .xlsx files)
csv_file_name   <- "^GDAXI.csv"
excel_file_name <- "DAX_19930401_20211123.xlsx"

    # for Yahoo download
series_yahoo    <- "^GDAXI"         # choose which price to analyse
# the first entry indicates for which index / security the prices are to be
# downloaded (use of Yahoo-Finance designations)

start_yahoo     <- "1993-04-01"     # Start
end_yahoo       <- "2022-11-28"     # End

alpha           <- 0.05             # Significance level

# Set working directory to the directory of the R-file in RStudio via
# Session -> Set Working Directory -> To Source File Location.
# or use the following two lines after uncommenting:
# WD <- ""         # Add directory structure for R-file
# setwd(WD)        # set it as working directory

# ---------------- Load DAX data ----------------------------------------

# To a) Download directly from Yahoo-Finance:
#       Use package "tseries" which contains the command get.hist.quote()
#       which allows direct download from yahoo.finance.
#       The command returns a zoo data frame.
#       A zoo data frame is a special data frame that allows a more detailed
#       control of dates than the normal data frame in R.

data_all_zoo    <- get.hist.quote(instrument = series_yahoo,
                                  start = start_yahoo, end = end_yahoo)
                        # Command in package tseries
head(data_all_zoo)
tail(data_all_zoo)
data_all_zoo[(1:10),]

# To b) Load data from .csv file downloaded earlier from Yahoo-Finance:

data_all        <- read.csv(csv_file_name, header = TRUE, sep = ",",
                            na.strings = "null",
                            colClasses = c("Date", rep("numeric", 6)) )
data_all_zoo<- zoo(x = data_all[,-1], order.by = as.Date(data_all[,1]))
head(data_all_zoo)
tail(data_all_zoo)

# To c):
#    c1) Use package "readxl", which does not require Java.
#       However, it returns a "tibble" instead of a "data frame",
#       which can be converted directly into a zoo data frame.

excel_daten <- read_xlsx(path = excel_file_name, range = "Tabelle1!A1:G7362",
                          col_names = TRUE, col_types = c("date", rep("numeric",6)), na = "null")
excel_daten_zoo <- zoo(x = excel_daten[,-1], order.by = as.Date(excel_daten$Date) )
head(excel_daten_zoo)
tail(excel_daten_zoo)

#    c2) Use package "xlsx"
#      requires older Java version to be present,
#      therefore does not work on every MAC
#      therefore commented out with ## in the following.

## if (!require(xlsx)){
##     install.packages("xlsx")
## }
## library(xlsx)    # library to read files in xls or xlsx format.
##
## excel_daten <- read.xlsx(excel_file_name, sheetIndex = 1, colIndex = (1:7),
##                          startRow = 1, colClasses = c("Date", rep("numeric",6)), header = TRUE)
##                 # create zoo data frame with exact dates
## excel_daten_zoo<- zoo(x = excel_daten[,-1], order.by = excel_daten[,1])
## head(excel_daten_zoo)
## tail(excel_daten_zoo)
##
## data_all_zoo <- excel_daten_zoo

# ------------------------ Plot and calculate returns ----------------------

price_all_zoo       <- data_all_zoo[, price_to_check] # Choose price type

## price_zoo    <- rev(price_zoo)      # only for Excel file for data up to 2015
## sort prices so that oldest value is at the beginning of the price-vector so
## that returns are calculated correctly
```

```
plot(price_all_zoo, ylab = price_to_check, xlab = "Year")  # Plot the data

price_zoo       <- na.omit(price_all_zoo)       # Remove missing values
n               <- length(price_zoo)            # Number of observations

# Derive left critical value for two-sided test
crit_left <- qt(alpha/2, df = n-1)

# A) Calculate the returns and the corresponding t-statistic

r              <- (price_zoo[2:n] - as.numeric(price_zoo[1:(n-1)])) /
                    as.numeric(price_zoo[1:(n-1)])
    # the first value is used as a zoo dataframe to keep the correct date.
    # The delayed values are used without a date by converting the values to
    # "numeric" so as not to create a conflict in the date.

plot(r, xlab = "Year", ylab = "Returns")
tail(r)
mean(r)
sd(r)
(t        <- mean(r)/(sd(r)/sqrt(n))) # Calculate test statistic

plot(density(r))    # Estimate density of returns
plot(function(x) dnorm(x, mean = mean(r), sd = sd(r) ),
     from = min(r), to =max(r), add=TRUE, col = "red")
                    # Add density of the normal distribution
                    # with mean and SD of the returns

# B) Calculate the log returns and the corresponding t-statistic
r_log           <- diff(log(price_zoo), lag=1)
            # the diff() command can handle zoo data frames

plot(r_log, xlab = "Year", ylab = "Log-Returns")
tail(r_log)
mean(r_log)
sd(r_log)
(t_log         <- mean(r_log)/(sd(r_log)/sqrt(n))) # Calculate test statistic

plot(density(r_log))    # Estimate density of log returns
plot(function(x) dnorm(x, mean = mean(r_log), sd = sd(r_log) ),
     from = min(r_log), to =max(r_log), add=TRUE, col = "red")
                    # Add density of the normal distribution
                    # with mean and SD of the log returns

# ==================== End ==============================================
```

Listing A.5: ./R_code/5_5_Test_Mean_DAX_eng.R

Power function of the test on the mean value, see figure 5.3

```
# ======================== 5_5_Plot_Power_Function_eng.R ==========================
# Program to create the graph for plotting the power function
# in slides Methods, Section 4.1
# created by: RT, 2012_12_20, kor. 2015_12_01, 2021_10_21 (sigma_0 removed from X-signature)

# ===== Define function to calculate power
z_power <- function(c, mu_v, sigma_mu, mu_H0)
{
  power_left    <- pnorm(-c, mean = (mu_v - mu_H0)/sigma_mu, sd=1)
  power_right   <- 1 - pnorm(c, mean = (mu_v - mu_H0)/sigma_mu, sd=1)
  return(power_left + power_right)
}
# ==== End Function =============================

# Parameters for plot

graphics.off()            # Close all graphic windows

alpha       <- 0.05     # Significance level
n           <- 50           # Number of observations
mu_H0       <- 0            # Mean value under H_0
sigma       <- 1            # Standard deviation

save.pdf    <- 1      # 1=create PDFs of graphs, 0=otherwise

c           <- qnorm(1 - alpha/2) # Calculate critical value
mu_v        <- seq(mu_H0 - 2, mu_H0 + 2, 0.1) # Grid for density under H_1

# Create plot

if (save.pdf) pdf("plot_power_function_2021_eng.pdf", height=4, width=7)
plot(mu_v, z_power(c, mu_v, sigma/sqrt(n), mu_H0), type="l",
    xlab = expression(mu[0]-mu[H[0]]#/(sigma[0]/sqrt(n)))
                  ), ylab="Power function")
abline(h = 0.05, col = "red")
axis(2, at = 0.05, labels = expression(alpha), tick=FALSE)
lines(mu_v, z_power(c, mu_v, 2 * sigma/sqrt(n), mu_H0), type="l", col="blue")
if (save.pdf) dev.off()
# ========================= End ====================== ========================
```

Listing A.6: ./R_code/5_5_Plot_Power_Function_eng.R

Illustration of the power function on a grid, see figure 5.4

```
# ================== power_function_persp ==================================
# Program for creating the perspective graph for displaying the power
# function in slides Methods, section 4.1
# created by: RT, 2012_12_20, 2022_12_06 Adaptation to macOS with x11()
# requires XQuartz, which needs to be installed additionally, and library(tcltk)
# Note: If library aplpack is not installed, install it first!
# Load the library aplpack, which contains the functions for slider

graphics.off()  # Close all graphic windows
library(aplpack)# Load the library aplpack
library(tcltk)  # 2022_12_06, RT: is necessary for XQuartz

# =================== Define functions ====================================
# ----------------------- z_power_grid------------------------------------------
z_power_grid <- function(mu_d_sigma_mu,mu_0,c)
{
    mu_v        <- mu_d_sigma_mu[1]
    sigma_mu    <- mu_d_sigma_mu[2]
    power_left  <- pnorm(-c,mean=(mu_v-mu_0)/sigma_mu,sd=1)
    power_right <- 1-pnorm(c,mean=(mu_v-mu_0)/sigma_mu,sd=1)
    return(power_left+power_right)
}
# ---------------------- End z_power_grid ---------------------------------------

# ----------------------- col_persp -------------------------------------------
# Function for colouring the surface
col_persp <- function(Z)
  {
  # Colors for surface = estimates
  n_col   <- dim(Z)
  nrZ   <- nrow(Z)
  ncZ   <- ncol(Z)
  Zlim  <- c(min(Z),max(Z))
  couleurs  <- heat.colors(trunc(1 * n_col))
  #  couleurs  <- topo.colors(trunc(1 * n_col))
  Zcol  <- couleurs[trunc((Z-Zlim[1])/(Zlim[2]-Zlim[1])*(n_col-1))+1]
  dim(Zcol) <- c(nrZ,ncZ)
  return(Zcol         <- Zcol[-nrZ,-ncZ])
  }
# ---------------------- End col_persp---------------------------------------

# ----------------------- flexible plot -----------------------------------
beweglicher_plot  <- function(...)
  # Create perspective graph
{
    persp(x=mu_v,y=sigma_mu_v,z=power_grid_mat, ticktype="detailed", col=power_grid_col,
        r=slider(no=3), #5,
        xlab = expression(mu[0]-mu[H[0]]), ylab = expression(sigma/sqrt(n)),
        zlab = "Power function",
        theta=slider(no=1), #35,
        phi=slider(no=2), #20,
        expand=1) -> res  #phi = 30
}
# ----------------------- End flexible plot --------------------------------
# ========================= End functions ====================================

# ========================= Main program ======================================
# Define parameters

alpha       <- 0.05         # significance level
mu_0        <- 0            # mean value under H_0
mu_diff     <- 1

sigma       <- 1            # standard deviation
sigma_min   <- sigma
sigma_max   <- sigma
n_min       <- 20
n_max       <- 1000
```

```
c           <- qnorm(1-alpha/2)     # critical value

# Grid
    # grid for mu
mu_v        <- seq(mu_0-mu_diff,mu_0+mu_diff,0.05)
    # grid for sigma_hat_mu
sigma_mu_step       <- (sigma/sqrt(n_min)-sigma/sqrt(n_max))/(length(mu_v)-1)
sigma_mu_v          <- seq(sigma/sqrt(n_max),sigma/sqrt(n_min),by=sigma_mu_step)

grid                <- expand.grid(mu_v,sigma_mu_v)
power_grid          <- apply(grid,1,z_power_grid,mu_0,c)
power_grid_mat      <- matrix(power_grid,length(mu_v),length(sigma_mu_v),byrow=FALSE)



# Colour hyperplane of subspace with function "col_persp", see above
power_grid_col <- col_persp(power_grid_mat)




# create the 3D graph
# call slider with function beweglicher_plot to create and possibly
# rotate the 3D graph
# windows() # opens a new graphic window,
            # 2022_12_06, RT: Command only works on Windows
x11()       # also works on MacOS
slider(beweglicher_plot,
       sl.names        = c("turn", "tilt", "distance"),
       sl.mins        = c(0, 0, 1),        # minimum values for sliders
       sl.maxs        = c(360, 360, 100),        # maximum values for sliders
       sl.deltas = c(1, 1, 1),        # step size for sliders
       sl.defaults = c(35, 20, 5)        # default values for parametersr
       , prompt = TRUE        # ensures that the effect of a slider movement is
                              # seen immediately on the screen and not only
                              # after releasing the mouse button
)
# End slider
```

Listing A.7: ./R_code/5_5_Plot_Power_Function_Persp_eng.R

A.4. R Program for an Empirical Example about Trade Flows, starting in Section 6.3

```
################### 4_ff_Beispiel_Handelsstroeme_eng.R #############################
#
################################################################################
################################################################################
# Example of trade flows in the Methods of Econometrics script,
# University of Regensburg
# Commented R code
# Status: 01.10.2015, RT 22.12.2021 (library stargazer), RT 2023_01_17 attempt to include AIC etc in atargazer -
      did not work
# Antecedent:
# - aussenhandel_beispiel_hk.r WS 2014/15 for part up to
# END COMPULSORY COURSE PO 2011 - MATERIAL
# - aussenhandel_beispiel_pflichtkurs.r
################################################################################
################################################################################
# In order to be able to run the script, the data for the
# trade flows example "importe_ger_2004_ebrd.txt" is needed.
#
```

```
# Note: First the functions stats and SelectCritEviews are defined.
# Then the main program begins in line ??

################################################################################
#                   Start definition functions
################################################################################

############################ Function stats ####################################
# Useful function that returns statistical key figures when a vector is entered
# analogous to EViews-Output of "Descriptive Statistics"
#

stats <- function(x) {

  n           <- length(x)
  sigma       <- sd(x) * sqrt((n-1)/n)
  skewness    <- 1/n * sum(((x-mean(x))/sigma)^3)
  kurtosis    <- 1/n * sum(((x-mean(x))/sigma)^4)
  jarquebera  <- n/6*((skewness)^2 + 1/4 * ((kurtosis-3))^2)
  pvalue      <- 1- pchisq(jarquebera, df = 2)

  Statistics  <- c(mean(x), median(x), max(x), min(x), sd(x),
                        skewness, kurtosis, jarquebera, pvalue)

  names(Statistics) <- c("Mean", "Median", "Maximum", "Minimum", "Std. Dev.",
                        "Skewness", "Kurtosis", "Jarque Bera", "Probability")

  return(data.frame(Statistics))
}
############################### End ###########################################

####################### Function SelectCritEviews ##############################
# Function for calculating model selection criteria as in EViews
# RT, 2011_01_26

SelectCritEviews <- function(model)
{
  n           <- length(model$residuals)
  k           <- length(model$coefficients)
  fitmeasure  <- -2*logLik(model)/n

  aic         <- fitmeasure + k * 2/n
  hq          <- fitmeasure + k * 2*log(log(n))/n
  sc          <- fitmeasure + k * log(n)/n
  sellist     <- list(aic=aic[1],hq=hq[1],sc=sc[1])
  return(t(sellist))
}
############################### End ###########################################

################################################################################
#                   End definition functions
################################################################################

################################################################################
#                   Start main program
################################################################################
save.pdf      <- 1             # 1=create PDFs of graphs, 0=otherwise

# The following libraries are loaded during the process: car,lmtest

# If these are not installed, they will be installed first:
if (!require(car)){
  install.packages("car")
}
if (!require(lmtest)){
  install.packages("lmtest")
}

# Determination of the working directory
# in which the R program and the data are located
WD            <- getwd() # Determine the directory of the R file and
setwd(WD)                # set it as working directory

# Read the data as data frame
daten_all     <-read.table("importe_ger_2004_ebrd.txt", header = TRUE)
# Assign the variable names and
# eliminate the observation export country: GER, import country: GER.
attach(daten_all[-20,])

# To try out, if importe_ger_2004_ebrd.txt has already been read in
stats(trade_0_d_o)

################################################################################
# Section 6.3
################################################################################

############# Scatterplot with (linear) regression line #####################
# I.1 Aim/scientific issue: first empirical attempt

# Define file name for output in PDF format
if (save.pdf)   pdf("plot_wdi_vs_trade.pdf", height=6, width=6)

# OLS estimation of a simple linear regression model, stored in ols
ols           <- lm(trade_0_d_o ~ wdi_gdpusdcr_o)
# Scatterplot of the two variables
plot(wdi_gdpusdcr_o, trade_0_d_o, col = "blue", pch = 16)
# Plot the linear regression line using abline
abline(ols, col = "red")
# Add a legend
legend("bottomright", "Lineare Regression", col = "red", lty = 1, bty = "n")

# Close device
if (save.pdf) dev.off()

######## Estimate two multiple linear regression models ##############
# II.3 Specifying, estimating and selecting an econometric model
# Note:
# The numbering of the regression models is based on
# the models in the script, section 10.3

# Run a linear regression and save the results as an object
mod_2_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))

# Display of the regression results
summary(mod_2_kq)

# II.4 Validating the estimated model
# Running linear regression with additional regressor and
# using the formula command
mod_3a_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                  ebrd_tfes_o
mod_3a_kq      <- lm(mod_3a_formula)
# Display the regression results of the second linear regression model
summary(mod_3a_kq)

################################################################################
# Section 8.2
################################################################################
# Functional form: level-level, ... , log-log

summary(lm(trade_0_d_o ~ wdi_gdpusdcr_o))           #level - level model
summary(lm(trade_0_d_o ~ log(wdi_gdpusdcr_o)))      #level - log model
summary(lm(log(trade_0_d_o) ~ wdi_gdpusdcr_o))      #log - level model
summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o))) #log - log models
summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o)+log(cepii_dist)))

################################################################################
# Section 8.5
```

```
################################################################################
# Is there a non-linear relationship between imports and GDP?
# Simple modelling option: GDP regressor is also quadratic in the model

# Model 5: Also use log(BIP)^2 as regressor
mod_5_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) +
  I(log(wdi_gdpusdcr_o)^2) + log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o)

mod_5_kq      <- lm(mod_5_formula)
summary(mod_5_kq)

# Generate plot of elasticities for different GDPs
elast_gdp     <- mod_5_kq$coef[2] + 2* mod_5_kq$coef[3]*log(wdi_gdpusdcr_o)
# Create scatterplot
if (save.pdf)  pdf("plot_modell5_elast.pdf.pdf", height=6, width=6)
plot(wdi_gdpusdcr_o, elast_gdp, pch = 16, col = "blue", main = "GDP-Elasticity")
if (save.pdf) dev.off()

################################################################################
# Section 9.5
###############################################################################
# Estimate the variance-covariance matrix of the OLS estimators for model 3a
summary(mod_3a_kq)$cov

# Estimate the correlation matrix of the OLS estimators for model 3a
cov2cor(summary(mod_3a_kq)$cov)

# Estimate the covariance matrix of sample observations for model 3a
cor(data.frame(log_wdi_gdpusdcr_o = log(wdi_gdpusdcr_o),
               log_cepii_dist=log(cepii_dist),ebrd_tfes_o))

################################################################################
# Section 10.3 Information Criteria
################################################################################

# Calculate the values of the table
# Apply the function "SelectCritEviews" to four different models

mod_1_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o))
summary(mod_1_kq)
deviance(mod_1_kq)                  # Calculates SSR
SelectCritEviews(mod_1_kq)      # Calculates AIC, HQ, SC

mod_2_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))
summary(mod_2_kq)
deviance(mod_2_kq)                  # Calculates SSR
SelectCritEviews(mod_2_kq)      # Calculates AIC, HQ, SC

mod_3a_kq     <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                      ebrd_tfes_o)
summary(mod_3a_kq)
deviance(mod_3a_kq)           # Calculates SSR
SelectCritEviews(mod_3a_kq)     # Calculates AIC, HQ, SC

mod_3b_kq     <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                                    log(cepii_area_o))
summary(mod_3b_kq)
deviance(mod_3b_kq)                 # Calculates SSR
SelectCritEviews(mod_3b_kq)     # Calculates AIC, HQ, SC

mod_4_kq      <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
                      ebrd_tfes_o + log(cepii_area_o))
summary(mod_4_kq)
deviance(mod_4_kq)                  # Calculates SSR
SelectCritEviews(mod_4_kq)      # Calculates AIC, HQ, SC

    # RT 2021_12_22: Automatic creation of a table (also for latex)
library(stargazer)
stargazer(mod_1_kq, mod_2_kq, mod_3a_kq, mod_3b_kq, mod_4_kq, type = "text"
          #keep.stat = "aic"
          )
          # RT 2023_01_17: keep.stat = C("all") added


################################################################################
################################################################################
# Section 11.3 Exact Tests
################################################################################

alpha                <- 0.05         # Significance level
# Estimating model 4
mod_4_kq             <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) +
                        log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))
summary(mod_4_kq)
qf(1-alpha,2,44)                   # Critical value
library(car)                         # Library car load for F-test
       # F-test
F_stat           <- linearHypothesis(mod_4_kq,
                                     c("ebrd_tfes_o=0","log(cepii_area_o)=0"),
                                     test=c("F"))
F_stat
       # Chi^2-test
F_stat           <- linearHypothesis(mod_4_kq,
                                     c("ebrd_tfes_o=0","log(cepii_area_o)=0"),
                                     test=c("Chisq"))
F_stat

################################################################################
# Section 11.7 Empirical analysis of trade flows
################################################################################

# Model 4 was calculated in section 10.3
resid_mod_4_kq  <- mod_4_kq$resid       # Residuals of model 4
trade_0_d_o_fit <- mod_4_kq$fitted      # Fitted values of model 4

# Plot of residuals vs. fitted values
if (save.pdf) pdf("plot_fits_vs_resids_mod_4.pdf", 6, 6)
plot(trade_0_d_o_fit, resid_mod_4_kq, col = "blue", pch = 16, main = "Scatterplot")
if (save.pdf) dev.off()

# Plot of the histogram of the residuals
if (save.pdf) pdf("plot_hist_resids_mod_4.pdf", 6, 6)
hist(resid_mod_4_kq, breaks = 20, col = "lightblue", prob = T, main = "Histogram")
    # Estimated density of the residuals
lines(density(resid_mod_4_kq),col = "black", prob = T, add="T")
    # Plot the corresponding theoretical normal distribution
curve(dnorm(x, mean = mean(resid_mod_4_kq), sd = sd(resid_mod_4_kq)),
      from = -3, to = 3, add = T, col = "red", lty = 2, lwd = 2)
legend("topleft", c("est. density","theoretical\nnormal distribution"),
       col = c("black","red"), lwd = 2, lty = c(1,2), bty = "n")
if (save.pdf) dev.off()

# statistical evaluation of the residuals
stats(resid_mod_4_kq)

####  Confidence intervals
confint(mod_4_kq)

####  t-tests, two-tailed and one-tailed

# Two-tailed test
    # Determining the critical values
alpha          <-0.05
qt(alpha/2,mod_4_kq$df)
qt(1-alpha/2,mod_4_kq$df)

      # t-statistic
(t             <- (coefficients(mod_4_kq)["log(wdi_gdpusdcr_o)"]-1)/
  sqrt(vcov(mod_4_kq)["log(wdi_gdpusdcr_o)","log(wdi_gdpusdcr_o)"]))

     # p-value
2*pt(-abs(t),mod_4_kq$df)

  # Alternative via F
  # for this you need the car package
    # download.packages("car", destdir="C:/Program Files/R/R-2.15.1/library")
    # install.packages("car")
library("car")
(F_stat             <- linearHypothesis(mod_4_kq,c("log(wdi_gdpusdcr_o)=1")))

#  One-tailed test with left alternative
    # Critical values
alpha           <-0.05
(qt(alpha,mod_4_kq$df))

    # t-test statistic
(t              <- summary(mod_4_kq)$coefficients["log(cepii_dist)",3])

    # p-value
(pt(t,mod_4_kq$df))

#### F-test, correlation matrix and confidence ellipses

# F-test with 2 restrictions
    # critical value for F-statistic
(qf(1-alpha,2,mod_4_kq$df))

    # F-statistic
(F2_stat        <- linearHypothesis(mod_4_kq,c("ebrd_tfes_o=0","log(cepii_area_o)=0"),
            test=c("F")))

# chi^2-test
    # critical value for chi^2-statistic
(qchisq(1-alpha,2))

    # chi^2-statistic
(Chisq_stat     <- linearHypothesis(mod_4_kq,c("ebrd_tfes_o=0","log(cepii_area_o)=0"),
                                test=c("Chisq")))

#### Covariance and correlation matrix

    # Covariance matrix
(cov_par        <- vcov(mod_4_kq))
    # Correlation matrix
(corr_par       <- cov2cor(cov_par))

#### Confidence ellipsoids

# Confidence ellipse
if (save.pdf) pdf("plot_conf_ellipse.pdf", 6, 6)
confidenceEllipse(mod_4_kq, which.coef = c(4, 5), levels = 0.95,
                  main = "confidence ellipse", col = "red")
# Confidence interval
abline(v = confint(mod_4_kq, "ebrd_tfes_o", level = 0.95), lty = 2,
       col = "blue", lwd = 2)
abline(h = confint(mod_4_kq, "log(cepii_area_o)", level = 0.95), lty = 2,
       col = "blue", lwd = 2)
if (save.pdf) dev.off()

################################################################################
# Section 14.4 FGLS and heteroskedasticity-robust OLS estimation
################################################################################

#### FGLS estimation for model 4
mod_4_formula <- log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
  ebrd_tfes_o + log(cepii_area_o)
```

```
# 1st step
resids        <- residuals(mod_4_kq)
fits          <- fitted(mod_4_kq)
mod_formula_ln_u_squared <- log(resids^2) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)
+  ebrd_tfes_o + log(cepii_area_o)

# 2nd step
omega         <- exp(fitted(lm(mod_formula_ln_u_squared)))
model_gls     <- lm(mod_4_formula, weights=1/omega)
(summary(model_gls))

#### Regression output with heteroskedasticity-robust standard errors

library(lmtest)
# For choices for estimating the heteroskedastic variance-covariance matrix,
# see section  14.3
(coeftest(mod_4_kq,vcov=hccm(mod_4_kq,type="hc1")))

################################################################################
# Section 15.7
################################################################################

#### Perform the RESET test for model 4 with
resettest(mod_4_kq)

#### Performing the Breusch-Pagan test for model 4
bptest(mod_4_kq)


################################################################################
#                   Start function whitetest
################################################################################
# White test for homoskedastic errors with cross products
# RW, 2011_01_26

whitetest <- function(model){

  # Extract data from model
  dat <- model$model
  dat$resid_sq <- model$resid^2

  # Create formula for auxiliary regression
  regr <- attr(model$terms, "term.labels")
  form <- as.formula(paste("resid_sq~(",paste(regr,collapse="+"),")^2+",paste("I(",regr,"^2)",collapse="+")))

  # Estimate auxiliary regression
  test_eq <- lm(form,data=dat)

  # Overall F-test
  fstat <- summary(test_eq)$fstatistic

  # Calculate and display result
  result1 <- c(fstat[1],fstat[2],fstat[3],pf(fstat[1],fstat[2],fstat[3],lower.tail=FALSE))
  names(result1) <- c("F-Statistic","df1","df2","P-Value")
  result <- list(result1,test_eq)
  return(result)
}

################################################################################
#                   End function whitetest
################################################################################

#### Performing the white test for model 4
whitetest(mod_4_kq)

########################## End main program #################################
```

Listing A.8: ./R_code/4_ff_Beispiel_Handelsstroeme_eng.R

A.5. R Program for Graphs in Section 7.1 The Geometry of the LS Estimator

Geometry of the LS estimator, see figures 7.1 and 7.2 ♯ **Derivation of the function** `comp_d3` in the following R program for the calculation of the 3rd coordinate of the hyperplane that is spanned by $\delta(\mathbf{X})$ in the case of $k = 2$ and $n = 3$:

The axes of the 3D graph are orthogonal to each other. According to the directions of the axes, the three unit basis vectors $\mathbf{e}_i$, $i = 1, 2, 3$ are chosen (see leverage effect in section 7.2). For these, therefore $\mathbf{e}_i^T \mathbf{e}_j = 0$, $i \neq j$ holds. For all vectors in the subspace $\delta(\mathbf{X})$

$$\mathbf{X}\mathbf{a} = d_1\mathbf{e}_1 + d_2\mathbf{e}_2 + d_3\mathbf{e}_3 = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} \tag{A.1}$$

holds. To calculate the hyperplane of the subspace inE^3, d_1 and d_2 can each be specified on a grid, $d_1, d_2 = 0, 0.25, 0.5, \ldots, 10$. The problem is now to determine d_3 in such a way that (A.1) is satisfied:

1. To do this, one first determines the (2×1) vector $\mathbf{a}$ depending on d_1, d_2:

$$\underbrace{\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}}_{:=\mathbf{X}_I} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}.$$

   Multiplication by $\mathbf{X}_I^{-1}$ ($\mathbf{X}_I$ is quadratic here) yields

$$\mathbf{a} = \mathbf{X}_I^{-1} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}.$$

2. Calculating d_3 using the 3rd row of (A.1) yields:

$$d_3 = \mathbf{X}_3\mathbf{a} = \mathbf{X}_3\mathbf{X}_I^{-1} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}.$$

```r
# ================== 7_1_Projektion_KQ_n3_eng.R ==================================
# Program for creating the graphs in the Methods of Econometrics script,
# Section 7.1 The Geometry of the OLS estimator
# created by: RT,KK,JS, 2010_11_24, 2015_09_30,
# 2022_12_20, RT (window() replaced by x11() )
#

# If library aplpack is not installed, this will be installed now.

if (!require(aplpack)){
  install.packages("aplpack")
}
graphics.off()  # Close all graphic windows

# Load the library aplpack which contains the functions for slider
library(aplpack)# Load the library aplpack
```

```
# =================== Define functions ====================================

# ----------------------- comp_d3 --------------------------------------------
# Function "comp_d3" that calculates d_3 given d_1 and d_2 and X,
  # cf. script Methods of Econometrics, section A.3
comp_d3 <- function(d,X) X[3,] %*% solve(X[1:2,]) %*% d
# ---------------------- End comp_d3 ---------------------------------------

# ----------------------- col_persp -------------------------------------------
# Function for colouring the surface
col_persp <- function(Z)
  {
  # Colors for surface = estimates
  n_col    <- dim(Z)
  nrZ      <- nrow(Z)
  ncZ      <- ncol(Z)
  Zlim     <- c(min(Z),max(Z))
  couleurs  <- heat.colors(trunc(1 * n_col))
  #  couleurs  <- topo.colors(trunc(1 * n_col))
  Zcol  <- couleurs[trunc((Z-Zlim[1])/(Zlim[2]-Zlim[1])*(n_col-1))+1]
  dim(Zcol) <- c(nrZ,ncZ)
  return(Zcol         <- Zcol[-nrZ,-ncZ])
  }
# ---------------------- End col_persp ---------------------------------------
# ---------------------- beweglicher_plot -------------------------------------
# Define function "beweglicher_plot", which is called by the program "slider"
# and creates the surface/perspective graph.
# Note: the function uses variables from the main program without
# passing them explicitly
beweglicher_plot  <- function(...) {
  # Create perspective graph
  persp(x=d1,y=d2,z=d3_mat, ticktype="detailed", col=d3_col,
        r=slider(no=3),       #5,
        xlab = "e1", ylab = "e2", zlab = "e3",
        theta=slider(no=1),   #35,
        phi=slider(no=2),     #20,
        expand=1) -> res      #phi = 30
    # x_1 vector
  lines(trans3d(x=c(0,x1[1]), y=c(0,x1[1]), z=c(0,x1[3]), pmat=res),
        col="black", lwd=2)
  text(trans3d(x1[1], x1[2]+1, x1[3], pmat=res), expression(x[1]),cex=sym_gr)

    # x_2 vector
  lines(trans3d(x=c(0,x2[1]), y=c(0,x2[2]), z=c(0,x2[3]), pmat=res),
        col="black", lwd=2)
  text(trans3d(x2[1]-1, x2[2]-0.5, x2[3], pmat=res), expression(x[2]),cex=sym_gr)

    # X beta vector
  lines(trans3d(x=c(0,Xbeta[1]), y=c(0,Xbeta[2]), z=c(0,Xbeta[3]), pmat=res),
        col="grey2", lwd=2)
  text(trans3d(Xbeta[1]+1, Xbeta[2], Xbeta[3], pmat=res), expression(X*beta),
       cex=sym_gr)

    # shifted u vector
  lines(trans3d(x=c(Xbeta[1],y[1]), y=c(Xbeta[2],y[2]), z=c(Xbeta[3],y[3]),
                pmat=res), col="brown", lwd=2)

    # y vector
  lines(trans3d(x=c(0,y[1]), y=c(0,y[2]), z=c(0,y[3]), pmat=res),
        col="red", lwd=2)
  text(trans3d(y[1], y[2], y[3]+1, pmat=res), expression(y),cex=sym_gr)

    # X hat beta vector
  lines(trans3d(x=c(0,y_hat[1]), y=c(0,y_hat[2]), z=c(0,y_hat[3]), pmat=res),
        col="purple", lwd=2)
  text(trans3d(y_hat[1]+1, y_hat[2], y_hat[3], pmat=res), expression(X*hat(beta)),
       cex=sym_gr )

    # hat u vector
  lines(trans3d(x=c(0,y[1]-y_hat[1]), y=c(0,y[2]-y_hat[2]), z=c(0,y[3]-y_hat[3]),
                pmat=res), col="green", lwd=2)
  text(trans3d(y[1]-y_hat[1], y[2]-y_hat[2], y[3]-y_hat[3]+1, pmat=res),
       expression(hat(u)),cex=sym_gr)

    # shifted hat u vector
  lines(trans3d(x=c(y[1],y_hat[1]), y=c(y[2],y_hat[2]), z=c(y[3],y_hat[3]),
                pmat=res), col="green", lwd=2)
}
# ----------------------- End beweglicher_plot -------------------------------
# ========================= End functions ====================================

# ========================= Main program =====================================
# Defining parameters
  # Specifying the parameters for the graph
sym_gr  <- 1.5  # Symbol size
step <- .25     # step size for the grid over which X beta is calculated
                # and plotted

# Specifying the three observations and the parameter vector beta
x1  <- c(1,1,1)
x2  <- 2*c(0.5,2,1.3)
beta <- c(5,-1)
u <- c(-3,4,5)

X <- cbind(x1,x2) # X = { x_1 x_2 }
Xbeta <- X%*%beta # X*beta
y <- Xbeta + u

# Calculating the OLS estimator and the fitted values
beta_hat<- solve(t(X) %*% X) %*% t(X) %*% y
y_hat <- X %*% beta_hat

# Calculating the 3D graph
d1_min <- min(Xbeta,0)
d1_max <- max(Xbeta,10)
d1 <- seq(d1_min,d1_max, by=step) # Grid points in first direction
d2 <- seq(d1_min,d1_max, by=step) # Grid points in second direction
d_grid <- expand.grid(d1,d2) #creating the grid over which subspace
                             # delta(X) is to be plotted
  # Apply function "comp_d3" on grid of d_1 and d_2
  # note: arguments passed in "apply" to the function are passed without "="
d3_grid <- apply(d_grid,1,comp_d3,X)
  # "apply" outputs a vector, which is subsequently converted into a matrix
  # so that d_3 matches the correct d_1 and d_2.
d3_mat <- matrix(d3_grid,length(d1),length(d2),byrow=FALSE)

# Colour hyperplane of subspace with function "col_persp", see above
d3_col <- col_persp(d3_mat)

# Create a scatterplot with regression line of the DGP and estimated
# regression line, errors and residuals for 2nd observation.
plot(x2,y,col="red",pch=16,xlab=expression(x[2]),ylab=expression(y)) # Scatterplot
abline(a=beta[1],b=beta[2],col="black") # Regression line of the DGP
points(x2,Xbeta,col="black",pch=16) # X beta on the regression line
abline(a=beta_hat[1],b=beta_hat[2],col="blue")# Estimated regression line
points(x2,y_hat,col="blue",pch=16) # hat y on the estimated regression line
  # plotting the error vector and residual vector for 2. observation
t <- 2
lines(cbind(x2[t],x2[t]),cbind(Xbeta[t],y[t]),col="brown") # Error vector
text(x2[t]-.2,(y[t]-Xbeta[t])*0.75, expression(u[2]),cex=sym_gr)

lines(cbind(x2[t],x2[t]),cbind(y_hat[t],y[t]),col="green") # Residual vector
text(x2[t]-.2,y[t]-(y[t]-y_hat[t])*0.5, expression(hat(u)[2]),cex=sym_gr)
text(x2[t]-.4,y[t],expression((list(x[22],y[2]))),cex=sym_gr)

# Creating the 3D graph
# call slider with function beweglicher_plot to create and possibly rotate
# the 3D graph
# windows() # opens a new graphic window
x11()
slider(beweglicher_plot,
       sl.names        = c("turn", "tilt", "distance"),
       sl.mins        = c(0, 0, 1),         # minimum values for sliders
       sl.maxs        = c(360, 360, 100),   # maximum values for sliders
       sl.deltas = c(1, 1, 1),              # step size for sliders
       sl.defaults = c(35, 20, 5)           # default values for parameters
       , prompt = TRUE        # ensures that the effect of a slider movement
                              # is seen immediately on the screen and not only
                              # after releasing the mouse button
)
# End slider
# ================ End main program =====================================
```

Listing A.9: ./R_code/7_1_Projection_KQ_n3_eng.R

A.6. R Program for Regression Results in Section 8.3 on Qualitative Data as Regressors

```
# ===================== 8_4_Interpretationen_Wage_eng.R ===========================
# Program for wage regressions with dummies and interaction terms,
# see section 8.4 in script Methods of Econometrics
# Status: 2015_10_02
# Predecessor: app_interpretationen_wage.r from WS 2013/14


# Specification of the working directory
# in which the R program and the data are located

WD            <- getwd() # set the directory of the R file and
setwd(WD)                # set it as working directory

# Import the data
# The data file "wage1.txt" must be located in the same directory as the
# R file
wage_data     <- read.table("wage1.txt", header = TRUE)
attach(wage_data)

# Wage regression with dummy variable, see section 8.4.1
wage_mod_1_kq <- lm(log(wage) ~ female +
                      educ + exper + I(exper^2) + tenure + I(tenure^2))
summary(wage_mod_1_kq)

# Relative difference of unconditional mean wages of women and men
(mean(wage[female==1])-mean(wage[female==0]))/mean(wage[female==0])
  # alternative calculation possibility
wage_mean <- lm(wage~0+female+I(1-female))
(wage_mean$coef[1]-wage_mean$coef[2])/wage_mean$coef[2]

# Wage regression with multiple dummy variables: Interaction of dummies,
# see section 8.4.2

  # Define dummy variables for subgroups
femmarr       <- female * married
malesing      <- (1 - female) * (1 - married)
malemarr      <- (1 - female) * married

wage_mod_2_kq <- lm(log(wage) ~ femmarr + malesing + malemarr +
                      educ + exper + I(exper^2) + tenure + I(tenure^2))
summary(wage_mod_2_kq)

# Wage regression with a dummy and a dummy interaction term
wage_mod_3_kq <- lm(log(wage) ~ female +
                      educ + exper + I(exper^2) + tenure + I(tenure^2) +
                      I(female*educ))
summary(wage_mod_3_kq)
```

Listing A.10: ./R_code/8_4_Interpretationen_Wage_eng.R

A.7. R Program for Graphs in Section 9.1 on Unbiasedness of the LS Estimator

Monte-Carlo-Simulation zur Erwartungstreue, see figure 9.1

```
# ======================== 9_1_MC_KQ_einf_lin_Reg_eng.R ============================
#
# Program for Monte Carlo simulation
# to illustrate the unbiasedness of the OLS estimator
# in the simple linear regression model.
# In addition, the covariance between the estimated OLS parameters
# is illustrated with a scatterplot..
# created by : RT, 2010_11_25

graphics.off()      # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation

set.seed(42)                # Randomseed
n             <- 50         # Sample size
R             <- 1000       # Number of replications

beta_0        <- c(1,0.9)   # Parameter vector
sigma_0       <- 2          # Standard deviation of the error

save.pdf      <- 1          # 1=create PDFs of graphs, 0=otherwise

# Form a loop
beta_hat_store <- matrix(0,nrow=R,ncol=length(beta))
                    # Initialise matrix to store the OLS estimates
                    # for each realisation
for (r in (1:R))
{
  # Generate a realisation of a simple linear regression model
  u           <- rnorm(n,mean=0,sd=sigma_0)     # Draw u
  x           <- sample(1:20, n, replace=TRUE)  # Draw x
  y           <- beta_0[1] + x * beta_0[2] + u  # Calculate y

  # Calculate the OLS estimator
  ols         <- lm(y~x)

  # Save the parameter estimate
  beta_hat_store[r,] <- coef(ols)
}

# Calculate the mean values of the parameter estimates
colMeans(beta_hat_store)

# Create histograms
if (save.pdf) pdf("plot_MC_KQ_einf_lin_Reg_hist.pdf", height=6, width=6)
par(mfrow=c(1,2))     # Display two plots in one graphic window
hist(beta_hat_store[,1],breaks=sqrt(R))
hist(beta_hat_store[,2],breaks=sqrt(R))
if (save.pdf) dev.off()

# Variance-covariance matrix of the estimators from the R realisations
(var(beta_hat_store))

# Asymptotic variance-covariance matrix
S_XX          <- matrix(c(1,10.5,10.5,143.5),2,2)
cov_asymp     <- sigma_0^2 * solve(S_XX)
   # Adjustment to sample size
(cov_asymp / n)

# Scatterplot of the R OLS estimates
par(mfrow=c(1,1))
plot(beta_hat_store[,1],beta_hat_store[,2])
```

```
# ========================= End ==============================================
```

Listing A.11: ./R_code/9_1_MC_KQ_einf_lin_Reg_eng.R

A.8. R Program for Monte Carlo Simulation in the Section 9.2 on Consistency of the LS Estimator

Monte Carlo simulation on consistency and the central limit theorem, see figures 9.2 und 9.3

```
# ======================== 9_2_MC_KQ_Konsistenz_einf_lin_Reg_eng.R =================
# Program for Monte Carlo simulation
# to illustrate the consistency and the asymptotic normal distribution
# of the OLS estimator in the simple linear regression model.
# Calculates mean and standard deviation of all replications
# and histograms.
# Note: Program is written with for loops for the sake of readability
# Status: RT, 2015_10_04

graphics.off()            # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation

set.seed(42)                   # Randomseed
N             <- c(50,100,500,1000,10000,100000)   # Sample sizes
R             <- 10000        # Number of replications

beta          <- c(1,0.9)      # Parameter vector
sigma         <- 2             # Standard deviation of the error

save.pdf      <- 1             # 1=create PDFs of graphs, 0=otherwise
# Form two loops:
#  Outer loop on the number of replications
#  Inner loop on the sample size

n_max         <- N[length(N)]  # Maximum sample size
# Initialise the output matrices
beta_1_hat_store <- matrix(0,nrow=R,ncol=length(N))
                    # Initialise matrix to store the OLS estimates
                    # for beta_1 each realisation and each sample size
beta_2_hat_store <- matrix(0,nrow=R,ncol=length(N))
                    # Initialise matrix to store the OLS estimates
                    # for beta_1 each realisation and each sample size


for (r in (1:R))
{
  # Generate a realisation of a simple linear regression model
  # for the maximum sample size
  u           <- rnorm(n_max,mean=0,sd=sigma)       # Draw u
  x           <- sample(1:20, n_max, replace=TRUE)  # Draw x
  y           <- beta[1] + x * beta[2] + u          # Calculate y

  for (i in (1:length(N)))
  {
    # Calculate the OLS estimator for all sample sizes

      # ols   <- lm(y[1:N[i]]~x[1:N[i]]) # Standard command for OLS estimation
          # Fast lm command to save time in the simulation
      ols     <- lm.fit(cbind(rep(1,N[i]),x[1:N[i]]),y[1:N[i]])

  # Save the parameter estimates
  beta_1_hat_store[r,i] <- coef(ols)[1]
  beta_2_hat_store[r,i] <- coef(ols)[2]
  }
}

# Calculate the mean values of the parameter estimates
beta_1_hat_mean  <- colMeans(beta_1_hat_store)
beta_2_hat_mean  <- colMeans(beta_2_hat_store)
```

```
# Calculate the standard deviations of the parameter estimates
beta_1_hat_sd    <- sqrt(diag(var(beta_1_hat_store)))
beta_2_hat_sd    <- sqrt(diag(var(beta_2_hat_store)))

  # Display on the screen
(cbind(N,beta_1_hat_mean,beta_1_hat_sd,beta_2_hat_mean,beta_2_hat_sd))

# Create histograms
if (save.pdf) pdf("plot_MC_KQ_Konsistenz_einf_lin_Reg1_eng.pdf", height=6, width=6)
par(mfrow=c(2,2))     # Display four plots in a graphic window
for (i in (1:2))
{
  # Sample size N[i]
  hist(beta_1_hat_store[,i], breaks=sqrt(R),
     xlab=expression(hat(beta)[1]), main=paste("Histogram for n= ",N[i],sep=""))
  hist(beta_2_hat_store[,i], breaks=sqrt(R),
     xlab=expression(hat(beta)[2]), main=paste("Histogram for n= ",N[i],sep=""))
}
if (save.pdf) dev.off()

if (save.pdf)  pdf("plot_MC_KQ_Konsistenz_einf_lin_Reg2_eng.pdf", height=6, width=6)
par(mfrow=c(2,2))     # Display four plots in a graphic window
for (i in (3:4))
{
  # Sample size N[i]
  hist(beta_1_hat_store[,i], breaks=sqrt(R),
       xlab=expression(hat(beta)[1]), main=paste("Histogram for n= ",N[i],sep=""))
  hist(beta_2_hat_store[,i], breaks=sqrt(R),
       xlab=expression(hat(beta)[2]), main=paste("Histogram for n= ",N[i],sep=""))
}
if (save.pdf) dev.off()

if (save.pdf)  pdf("plot_MC_KQ_Konsistenz_einf_lin_Reg3_eng.pdf", height=6, width=6)
par(mfrow=c(2,2))     # Display four plots in a graphic window
for (i in (5:6))
{
  # Sample size N[i]
  hist(beta_1_hat_store[,i], breaks=sqrt(R),
       xlab=expression(hat(beta)[1]), main=paste("Histogram for n= ",N[i],sep=""))
  hist(beta_2_hat_store[,i], breaks=sqrt(R),
       xlab=expression(hat(beta)[2]), main=paste("Histogram for n= ",N[i],sep=""))
}
if (save.pdf) dev.off()
# ========================= End ==============================================
```

Listing A.12: ./R_code/9_2_MC_KQ_Konsistenz_einf_lin_Reg_eng.R

A.9. R Program for the Representation of ifo Business Climate Time Series in the Section 12 on Univariate Time Series Models

```
# ======================== 12_0_ifo_Geschaeftsklima_1991-2023_eng.R ===========================
#
# generates graphs of the time series on ifo business outlook,
# the ifo business situation and the ifo business climate
# Note: Previous data only referred to the commercial economy
# Current data from: https://www.ifo.de/node/67013
# last change: 2023_02_01, RT

save.pdf      <- 0             # 1=create PDFs of graphs, 0=otherwise

# If these are not installed, they are installed first:
if (!require(dynlm)){
  install.packages("dynlm")
}
if (!require(readxl)) {
    install.packages("readxl")
}

# Set the working directory
# The easiest way is via RStudio

# Read in the data

# with package "readxl", which does not require Java.
#      However, it returns a "tibble" instead of a "dataframe", so that
#      the "data.frame" command is also required to convert a tibble
#      into a dataframe
library(readxl)
library(dynlm)
excel_daten  <- data.frame(read_excel(path="ifo_geschaeftsklima_1991_01-2023_01_gsk-d-202301.xlsx",
                                      sheet = 2, range = "B10:D394", col_names=FALSE))
                    # Note that data from 1991 onwards are only available for the
                    # Gewerbliche Wirtschaft on sheet 2
# Create a time series object with dataframe properties
daten <- zoo( ts((excel_daten),
              start = c(1991, 1), end = c(2023,01), frequency = 12,
              names = c("Business_climate", "Business_situation",
                        "Business_outlook")) )

head(daten)

# Plot time series
if (save.pdf)  pdf("ifo_geschaeftsklima_1991_01-2023_01_eng.pdf", height=6,width=6)
plot(daten, xlab="Time", main="ifo business data (commercial)")
if (save.pdf) dev.off()

# Create scatterplot for business outlook
n <- nrow(daten)
if (save.pdf)  pdf("ifo_geschaeftsklima_scatter_1991_01-2023_01_eng.pdf",
                   height=6, width=6)
plot(Business_outlook[2:n] ~ Business_outlook[1:(n-1)], data=daten)
if (save.pdf) dev.off()

# Estimate AR(1) model for business climate and business outlook
gk_ols <- lm(Business_climate[2:n] ~ Business_climate[1:(n-1)],data=daten)
summary(gk_ols)

ge_ols <- lm(Business_outlook[2:n] ~ Business_outlook[1:(n-1)],
             data=daten)
summary(ge_ols)

# alternatively with dynlm package (allows lag notation as in EViews)
  # simplified regression with time series
gk_dynlm <- dynlm(Business_climate ~ L(Business_climate),data=daten)
```

```
summary(gk_dynlm)
# ========================= End ==============================================
```

Listing A.13: ./R_code/12_0_ifo_Geschaeftsklima_1991–2023_eng.R

A.10. R Program for the Representation of different Realisations of Time Series in Section 12.1 on Stochastic Processes

```
# ======================== 12_1_Traj_RW_eng.R =====================================
#
# generates a graph with ten realisations of a random walk with or without drift
# and plots the time series
# last change: 2015_10_10, RT, 2022_02_02, RT (also with drift)

save.pdf <- 0      # 1=create PDFs of graphs, 0=otherwise

# Parameters of the DGP and the MC
n        <- 20     # Length of the time series

alpha      <- 1          # AR parameters of AR(1) process with mean value 0
                               # 0:      Gaussian white noise
                               # 1:      Random walk
                               # 0 < |alpha| < 1: stationary process
nu        <- 0.5        # allows to draw random walks with drift nu * t
                        # added RT 2022_02_02

R        <- 10     # number of trajectories

var_z      <- 0      # Variance of z to illustrate ergodicity:
                               # 0   => ergodic
                               # >0      => not ergodic
# Parameters for plots
lwd      <- 3
cexmu    <- 2

set.seed(42)       # seed value

# Initialisation of the output matrices
y        <- matrix(nu + rnorm(n*R), n)  # Initialisation of the time series vectors with
                    # standard-normally distributed error process (Gaussian
                    # white noise)
                    # here nu added to allow for drift using filter command, 2022_02_02
z        <- rnorm(R) * var_z    # draw a random number that is the same for all t

# Generate all R trajectories of the AR(1) process

for (i in 1:R) y[,i] <- filter(y[,i], alpha, method="recursive")

# Plotting the time series - Displaying ensemble
if (save.pdf) pdf("Traj_RW_points_eng.pdf")
  # First trajectory
plot(y[,1]+z[1], cex.lab=cexmu, cex.axis=cexmu, lwd=lwd, ylim=c(min(y+min(z)),
                max(y)+max(z)), ylab=expression(y[t]), xlab="t")
    # 2nd to Rth trajectory
for (i in 2:R) points(y[,i]+z[i], col=i, lwd=lwd)
dev.off()

 # Plotting the time series - displaying trajectories
if (save.pdf) pdf("Traj_RW_lines_eng.pdf")
    # First trajectory
plot(y[,1], cex.lab=cexmu, cex.axis=cexmu, lwd=lwd, type="l", ylim=c(min(y),
      max(y)), ylab=expression(x[t]), xlab="t")
    # 2nd to Rth trajectory
for (i in 2:R) lines(y[,i], col=i, lwd=lwd)
```

```
dev.off()
# ========================= End ==============================================
```

Listing A.14: ./R_code/12_1_Traj_RW_eng.R

A.11. R Program for Monte Carlo Simulation in Section 12.2 on Linear Stochastic Processes and MA Processes

```
# ==========================12_2_MA2_Realisation_eng ===============================
# Program for creating a realisation of a MA(2) process
# created by:  RT, 2015_09_11

# Define the sample size and the parameters of a MA(2) process
n      <- 100                   # Sample size
sigma  <- 2                     # Standard deviation of the white noise
psi    <- c(1, 0.8, 0.6)        # MA parameters for y_t = u_t + 0.8 u_{t-1} + 0.6 u_{t-2}
set.seed(1)                     # Set seed value for random generator

save.pdf      <- 1             # 1=create PDFs of graphs, 0=otherwise

# Generate a realisation
u      <- rnorm(n + length(psi) - 1, sd = sigma)
                                # Generate white noise
                                # Generate a realisation of a MA(2) process
y      <- filter(u, filter = psi, sides = 1, method = "convolution")

if (save.pdf) pdf("MA2_Realisation_eng.pdf", height=6, width=6)
plot(y, xlab = "Time", ylab = expression(y[t]))
                                # Plotting a MA(2) time series
if (save.pdf) dev.off()


# Calculate the theoretical autocorrelation function for k=0,1,...,10
ARMAacf(ma=psi[2:3],lag.max=10)
# =============================== End =========================================
```

Listing A.15: ./R_code/12_2_MA2_Realisation_eng.R

A.12. R Program for Monte Carlo Simulation in Section 12.3.1 on AR(1) Processes

```
# ===================== 12_3_AR1_Realisierung_eng.R ===============================
# Program for creating a realisation of an AR(1) process
# created by: RT, 2015_10_10

# AR(1) parameters of the DGP
nu          <- 1
alpha_1     <- 0.8
sigma2      <- 4
y_0         <- 0

# Length of the time series
n           <- 500

set.seed(15)     # Set seed value
u           <- rnorm(n,sd=sqrt(sigma2)) # Generate Gaussian White Noise
y           <- rep(y_0,n)               # Initialise output vector
```

```
save.pdf      <- 0              # 1=create PDFs of graphs, 0=otherwise
# Set the working directory
# in which the R program and the data are located
WD            <- getwd() # Determine the directory of the R file and
setwd(WD)                # set it as working directory

# Generate AR(1) realisation
for (i in (2:n))
{
  y[i]      <- nu + alpha_1 * y[i-1] + u[i]
}
# Plot realisation
if (save.pdf) pdf("AR1_Realisierung_eng.pdf", height=6, width=6)
plot(seq(1:n),y,xlab="Time",ylab=expression(y[t]),type="l")
if (save.pdf) dev.off()
```

Listing A.16: ./R_code/12_3_AR1_Realisierung_eng.R

A.13. R Program for Monte Carlo Simulation in Section 12.3.3 on AR(p) Processes and more

Realisation, ACF, MA paramter, PACF of an AR(2) Process, see figure 12.7

```
# ======================== 12_3_AR2_Realisierung_eng.R ==========================
# Program for creating a realisation of an AR(2) process
# and for calculating the ACF, the MA representation and the roots
# created by: RT, 2015_29_09

# AR(2) parameters of the DGP

alpha_0       <- 1
alpha         <- c(-0.5,-0.8)
sigma2        <- 4

# Start values
y_start       <- c(0,0)

# Length of the time series
n             <- 500

save.pdf      <- 1              # 1=create PDFs of graphs, 0=otherwise
# Set the working directory
# in which the R program and the data are located
WD            <- getwd() # Determine the directory of the R file and
setwd(WD)                # set it as working directory


# Check stability of the AR(2) polynomial
AR2_wurzeln   <- polyroot(c(1,-alpha))
abs(AR2_wurzeln)

set.seed(15)    # Set seed value
u             <- rnorm(n,sd=sqrt(sigma2)) # Generate Gaussian White Noise
y             <- rep(NA,n)                # Initialise output vector
y[1:length(y_start)] <- y_start         # Plug in start values

# Generate AR(2) realisation
for (i in ((length(alpha)+1):n))
{
  y[i]        <- alpha_0 + alpha %*% y[(i-1):(i-2)] + u[i]
}
```

```
# Generate plots
if (save.pdf) pdf("AR2_Realisierung_eng.pdf", height=6, width=6)
split.screen(c(2,2))

# Plot realisation
screen(1)
plot(seq(1:n),y,xlab="Zeit",ylab=expression(y[t]),type="l")

# Plot theoretical ACF
screen(2)
plot(ARMAacf(ar=alpha,lag.max=20),type="h",ylab="ACF",xlab="Lags")

# Plot theoretical PACF
screen(3)
plot(ARMAacf(ar=alpha,lag.max=20,pacf=TRUE),type="h",ylab="PACF",xlab="Lags")

# Plot MA parameters of the inverted AR(2) process
screen(4)
plot((1:20),ARMAtoMA(ar=alpha,lag.max=20),type="h",ylab="MA parameters",xlab="Lags")
if (save.pdf) dev.off()

if (save.pdf) pdf("AR2_Realisierung_ACF_eng.pdf", height=6, width=6)
acf(y,lag.max=20,type="correlation")
if (save.pdf) dev.off()
# ========================= End ==============================================
```

Listing A.17: ./R_code/12_3_AR2_Realisierung_eng.R

A.14. R Program for Estimating the Autocorrelation Function in Section 12.4 on Estimating First and Second Moments in the Case of Stationary Processes

Estimated autocorrelation function of a white noise realisation, see figure 12.9

```
# ============================== 12_4_WN_ACF_Est_eng ==============================
# Program for estimating the autocorrelation function of a realisation of a
# Gaussian white noise process with n=100 observations
# created by: RT, 2015_18_10

# Variance
sigma2   <- 4
# Length of the time series
n        <- 100
save.pdf      <- 1              # 1=create PDFs of graphs, 0=otherwise

# Set seed value
set.seed(15)
# Generate Gaussian white noise
y        <- rnorm(n,sd=sqrt(sigma2))

# Plot the estimated autocorrelation function
# with 95\% confidence intervals
if (save.pdf) pdf("ACF_WN_Est_eng.pdf", height=6, width=6)
acf(y,lag.max=20,type="correlation")
if (save.pdf) dev.off()
```

Listing A.18: ./R_code/12_4_WN_ACF_Est_eng.R

A.15. R Program for the Simulation and Estimation of AR(1) Processes in Section 13.5 on LS Estimation of Dynamic Linear Regression Models

Generation and estimation of a process

```
# =============================== 13_5_KQ_AR1_eng.R ====================================
# Program for generating and OLS estimation of an AR(1) model
# created by : RT, 2011_01_19

graphics.off()      # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation
set.seed(42)        # Randomseed
N     <- 50         # Sample size

beta  <- c(2,0.1)   # Parameter vector
sigma <- 2          # Standard deviation of the error
y0    <- 0          # Start value of the AR(1) process


# Generate a realisation of an AR(1) process
u  <- rnorm(N,mean=0,sd=sigma)       # Draw u
y <- rep(1,N)*y0
for (t in (2:N))
{
  y[t] <- beta[1] + y[t-1] * beta[2] + u[t] # Calculate y_t
}

# Plot of the time series
plot(y,xlab="Time",ylab="y",type="l")

# Scatterplot
plot(y[1:(N-1)],y[2:N])

# Calculate the OLS estimator
ols <- lm(y[2:N]~1+y[1:(N-1)])  # Note x=y_{t-1]. Therefore y_t of t=2,...,N
summary(ols)
# =============================== End ========================================
```

Listing A.19: ./R_code/13_5_KQ_AR1_eng.R

Monte Carlo simulation

```
# ======================== 13_5_MC_KQ_AR1_eng.R ==================================
# Program for Monte Carlo simulation
# to determine the bias of the OLS estimator in the AR(1) model
# created by : RT, 2010_11_25

graphics.off()      # Close all graphic windows

# Set parameters of the model and the Monte Carlo simulation

set.seed(42)        # Randomseed
N     <- 50         # Sample size
R     <- 1000       # Number of replications

beta  <- c(1,0.9)   # Parameter vector
sigma <- 2          # Standard deviation of the error
y0    <- 1          # Start value of the AR(1) process

# Forming a loop
beta_hat_store <- matrix(0,nrow=R,ncol=length(beta))
                    # Initialise matrix to store the OLS estimates
                    # for each realisation
```

```
for (r in (1:R))
{
  # Generate a realisation of an AR(1) process
  u <- rnorm(N,mean=0,sd=sigma)       # Draw u
  y <- rep(1,N)*y0
  for (t in (2:N))
  {
      y[t] <- beta[1] + y[t-1] * beta[2] + u[t] # Calculate y_t
  }
  # Calculate the OLS estimator
  ols <- lm(y[2:N]~y[1:(N-1)])    # Note x=y_{t-1}. Therefore y_t of t=2,...,N

  # Store the parameter estimates
  beta_hat_store[r,] <- coef(ols)
}

# Calculate the mean values of the parameter estimates

colMeans(beta_hat_store)

# Create histograms
par(mfrow=c(1,2))     # Draw two plots in a graphic window

hist(beta_hat_store[,1],breaks=sqrt(R))
hist(beta_hat_store[,2],breaks=sqrt(R))

# ========================= End ==================================
```

Listing A.20: ./R_code/13_5_MC_KQ_AR1_eng.R

B. R Commands for Regression Analysis

B.1. Overview of Available Commands

Required R packages: stats usually loaded, car, lmtest, moments, sandwich.

Performing a lineaer regression:

```
model_kq <- lm()
```

creates a regression object that is the basis for the following commands:

`R` Command	Functional Description	`R` Package
	Output of estimations and forecasts	
`print()`	simple printed display	
`summary()`	standard regression output	
`coef()`	(or `coefficients()`) extracts estimated regression parameters	
`residuals()`	(or `resid()`) extracts residuals	
`fitted()`	(or `fitted.values()`) extracts fitted values	
`anova()`	Comparison of nested models	
`predict()`	Prediction for new regression values	
`confint()`	Confidence intervals for regression coefficients	
`confidenceEllipse()`	Confidence intervals for regression coefficients	`car`
`deviance()`	Sum of squared residuals (SSR)	
`vcov()`	(estimated) variance-covariance matrix of the parameter estimators	
`logLik()`	Log-likelihood (under the assumption of normal distributed errors)	
	Testing	
`hccm()`	Heteroscedasticity corrected variance-covariance matrix of the parameter estimators; with `type="hc0"` White variance-covariance matrix (14.19)	`car`
`coeftest()`	standard regression output, if necessary using heteroscedasticity-robust standard errors	`lmtest`
`linearHypothesis()`	F-test `test=c("F")` or (asymptotical) χ^2 test `test=c("Chisq")`; with `white.adjust=c(FALSE, TRUE, "hc0")` White heteroscedasticity-robust variance-covariance matrix	`car`
`lrtest()`	Likelihoodratio test, see **Advanced Issues in Econometrics** or **Advanced Econometrics**	`lmtest`
`waldtest()`	Wald test, see **Advanced Issues in Econometrics** or **Advanced Econometrics**	`lmtest`

Continued on the next page.

R Command	Functional Description	R Package
Model specification		
`AIC()`	Information criteria including AIC, BIC/SC (assuming normally distributed errors) - Note: In contrast to EViews, the estimated parameter variance is counted as a parameter and not divided by the number of observations, see section 10.1	
`SelectCritEViews()`	Information criteria a la EViews, see section 10.1	own R program, see section ??
`encomptest()`	Encompassing test for testing non-nested regression models, see section 10.2 or Davidson & MacKinnon (2004, Section 15.3)	`lmtest`
`jtest()`	*J*-test for testing non-nested regression models, see section 10.2 or Davidson & MacKinnon (2004, Section 15.3)	`lmtest`
Model diagnostics		
`plot()`	Graphs for model checking	
`resettest()`	RESET test to test the functional form, see section 15.3	`lmtest`
`jarque.test()`	Lomnicki-Jarque-Bera test for checking normally distributed errors, see section 15.4	`moments`
`bptest()`	Breusch-Pagan test to test for the presence of heteroscedastic errors, see section 15.2.1	`lmtest`
`whitetest()`	White test to test for the presence of heteroscedastic errors, see section 15.2.2	own R program, see section B.2

Commands for graphs/plots:

```
# INFORMATION FOR GRAPHS


# save as .eps file (file name, size of the graphic)
# postscript("pdf_std_normal.eps", height=4, width=6, horizontal=FALSE)
# or save as .pdf
# pdf("pdf_std_normal.pdf", height=4, width=6)
# windows()         # opens a new graphic window

# split.screen(c(2,1))  # splits a graphic window
# screen(1)             # selects window 1
# dev.off()             # closes open graphic window
# close.screen(all=TRUE)# closes all windows

# Parameters for graphics: (optional)
#     las=1: Axis scaling horizontal
#     mai: width of margins (bottom, left, top, right)
#     mgp: Position of axes, axis scaling and axis labelling
```

B.2. Own R Packages

```
# --------------- SelectCritEViews --------------------------------------------------
# function to compute model selection criteria for linear regressions as EViews
# RT, 2011_01_26

SelectCritEViews <- function(model)
  {
  n   <- length(model$residuals)
  k   <- length(model$coefficients)
  fitmeasure  <- -2*logLik(model)/n

  aic <- fitmeasure + k * 2/n
  hq  <- fitmeasure + k * 2*log(log(n))/n
  sc  <- fitmeasure + k * log(n)/n
  sellist <- list(aic=aic[1],hq=hq[1],sc=sc[1])
return(sellist)
  }
# -----------------------------------------------------------------------------
```

```
# --------------- whitetest --------------------------------------------------
# function to conduct White test including cross terms
# RW, 2011_01_26

whitetest <- function(model){

# Extract data from model
dat <- model$model
dat$resid_sq <- model$resid^2

# Create formula for auxiliary regression
regr <- attr(model$terms, "term.labels")
form <- as.formula(paste("resid_sq~(",paste(regr,collapse="+"),")^2+",
            paste("I(",regr,"^2)",collapse="+")))

# Estimate auxiliary regression
test_eq <- lm(form,data=dat)

# Overall F-test
fstat <- summary(test_eq)$fstatistic

# Calculate and display result
result1 <- c(fstat[1],fstat[2],fstat[3],pf(fstat[1],fstat[2],
        fstat[3],lower.tail=FALSE))
names(result1) <- c("F-Statistic","df1","df2","P-Value")
result <- list(result1,test_eq)
return(result)
}
# ----------------------------------------------------------------------------
```

More:

- Course **Programming with R**
- Kleiber & Zeileis (2008)
- Overview of available packages in R

  http://cran.r-project.org/web/views/

C. Data for Estimation of the Gravity Equation

Section corresponds to section 10.4 in course material on Introduction to Econometrics.

Legend for the data in `importe_ger_2004_ebrd.txt`

- **Countries and country codes**

1	ALB	Albania
2	ARM	Armenia
3	AUT	Austria
4	AZE	Azerbaijan
5	BEL	Belgium and Luxembourg
6	BGR	Bulgaria
7	BIH	Bosnia and Herzegovina
8	BLR	Belarus
9	CAN	Canada
10	CHE	Switzerland
11	CYP	Cyprus
12	CZE	Czech Republic
13	DNK	Denmark
14	ESP	Spain
15	EST	Estonia
16	FIN	Finland
17	FRA	France
18	GBR	United Kingdom
19	GEO	Georgia
20	GER	Germany
21	GRC	Greece
22	HKG	Hong Kong
23	HRV	Croatia
24	HUN	Hungary
25	IRL	Ireland
26	ISL	Iceland
27	ITA	Italy
28	JPN	Japan
29	KAZ	Kazakhstan
30	KGZ	Kyrgyzstan
31	LTU	Lithuania
32	LVA	Latvia
33	MDA	Republic of Moldova
34	MKD	FYR Macedonia
35	MLT	Malta
36	NLD	Netherlands
37	NOR	Norway
38	POL	Poland
39	PRT	Portugal
40	ROM	Romania
41	RUS	Russia
42	SVK	Slovakia
43	SVN	Slovenia
44	SWE	Sweden
45	TJK	Tajikistan
46	TKM	Turkmenistan
47	TUR	Turkey
48	UKR	Ukraine
49	USA	United States
50	UZB	Uzbekistan

Countries that only appear as countries of origin:

BIH	Bosnia und Herzegovina	CHN	China	KOR	South Korea
TJK	Tajikistan	HKG	Hong Kong	TWN	Taiwan
UZB	Uzbekistan	JPN	Japan	THA	Thailand

R **code**:

```
setwd(’d:/..’)                # set working directory
daten   <- read.table("importe_ger_2004_ebrd.txt", header=TRUE, sep="\t")
```

```
attach(daten)
cbind(matrix(iso_o,50,1),matrix(d_o,50,1))  # Abbreviations of country names
                                            # and trade directions
```

- **Endogenous variable**:
  - TRADE_0_D_O: Imports of country D from country O (i.e. exports of country O to country D) in current US dollars.
  - Product classes: Trade flows are based on the aggregation of trade flows recorded according to the Standard International Trade Classification, Revision 3 (SITC, Rev.3) at the lowest level of aggregation (4 or 5 digits). Source: UN COMTRADE
  - Fuels and lubricants are not included (i.e. specifically fuel and natural gas products). Minimum limit of the underlying split trade flows (at the SITC Rev.3 5-digit level) is 500 US dollars.

- **Explanatory variables:**

Country of origin (O-country)

WDI_GDPUSDCR_O	country of origin GDP data; in current US dollars	World Bank - World Development Indicators
WDI_GDPPCUSDCR_O	country of origin GDP per capita data; in current US dollars	World Bank - World Development Indicators
WEO_GDPCR_O	country of destination and origin GDP data; in current US dollars	IMF - World Economic Outlook database
WEO_GDPPCCR_O	country of destination and origin GDP per capita data; in current US dollars	IMF - World Economic Outlook database
WEO_POP_O	country of origin population data	IMF - World Economic Outlook database
CEPII_AREA_O	area of country of origin in km^2	CEPII
CEPII_COL45	dummy; d- and o-country had a colonial relationship after 1945	CEPII
CEPII_COL45_REV	dummy; revised by “expert knowledge”	
CEPII_COLONY	dummy; d- and o-country ever had a colonial relationship	CEPII
CEPII_COMCOL	dummy; d- and o-country share a common colonial master after 1945	CEPII
CEPII_COMCOL_REV	dummy; revised by “expert knowledge”	
CEPII_COMLANG_ETHNO	dummy; d- and o-country share a common language	CEPII
CEPII_COMLANG_ETHNO_REV	spoken by at least 9% of the population	
CEPII_COMLANG_OFF	dummy; d- and o-country share common official language	CEPII
CEPII_CONTIG	dummy; d- and o-country are neighbouring states	CEPII
CEPII_DISINT_O	domestic distance in country of origin	CEPII
CEPII_DIST	geodetic distance between d- and o-country	CEPII
CEPII_DISTCAP	distance between d- and o-country based on their capitals $0.67\sqrt{Fläche/\pi}$	CEPII
CEPII_DISTW	weighted distances, see CEPII for details	CEPII
CEPII_DISTWCES	weighted distances, see CEPII for details	CEPII
CEPII_LAT_O	latitude of the city	CEPII
CEPII_LON_O	longitude of the city	CEPII
CEPII_SMCTRY_REV	dummy; d- and o-country were/are the same country	CEPII, revised
ISO_O	three-letter ISO code for country of origin	CEPII
EBRD_TFES_O	EBRD measure of the degree of liberalisation of the trade and payment flows of the o-country	EBRD

Country of destination (D-country)

WDI_GDPUSDCR_D	country of destination GDP data; in current US dollars	World Bank - World Development Indicators
WDI_GDPPCUSDCR_D	country of destination GDP per capita data; in current US dollars	World Bank - World Development Indicators
WEO_GDPCR_D	country of destination and origin GDP data; in current US dollars	IMF - World Economic Outlook database
WEO_GDPPCCR_D	country of destination and origin GDP per capita data; in current US dollars	IMF - World Economic Outlook database
WEO_POP_D	country of destination population data	IMF - World Economic Outlook database

Notes: The EBRD measures reform efforts on a scale of 1 to 4+ (=4.33); 1 indicates no or marginal progress; 2 indicates important progress; 3 indicates substantial progress; 4 indicates extensive progress, while 4+ means that the country has reached the standard and performance norms of advanced industrialised countries, i.e., OECD countries. This variable is by construction qualitative and not cardinal.

- **Thanks:** to Richard Frensch, Institute for Eastern Europe, Regensburg, who provided the data.
- **Websites** CEPII

D. Basic concepts of sets

Definition

Fischer (cf. 2014, p. 43) A set G together with an operation $*$ is called **group** if the following axioms are fulfilled:

1. Associative property: $(a * b) * c = a * (b * c)$ for all $a, b, c \in G$
2. Existence of a neutral element $e \in G$ with the following properties:
   a) $e * a = a$ for all $a \in G$.
   b) Existence of an inverse element: For each $a \in G$ there is an $a' \in G$ with $a' * a = e$.

The group is called **abelian** or **commutative** if, in addition $a * b = b * a$ for all $a, b \in G$.

Definition

Fischer (cf. 2014, p. 54) A set R together with two operations

$$\begin{aligned} + :& R \times R \to R, \quad (a, b) \to a + b, \quad \text{and} \\ \cdot :& R \times R \to R, \quad (a, b) \to a \cdot b, \end{aligned}$$

is called **ring** if the following applies:

1. R together with the addition $+$ is an abelian group ($=$ commutative group).
2. The multiplication $\cdot$ is associative.
3. The distributive properties hold, i.e. for all $a, b, c \in R$ it holds that

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \text{and} \quad (a + b) \cdot c = a \cdot c + b \cdot c.$$

Definition

Fischer (cf. 2014, p. 56) A set K together with two operations

$$\begin{aligned} + :& R \times R \to R, \quad (a, b) \to a + b, \quad \text{and} \\ \cdot :& R \times R \to R, \quad (a, b) \to a \cdot b, \end{aligned}$$

is called **field** if the following applies:

1. K together with the addition $+$ is an abelian group.

2. Existence of a subset for which the multiplication is also an abelian group: If $K^* := K\backslash 0$, then for $a, b \in K^*$ it also holds that $a \cdot b \in K^*$, and K^* together with the multiplication thus obtained is an abelian group.

3. The distributive properties hold, i.e. for all $a, b, c \in K$ it holds that

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \text{and} \quad (a + b) \cdot c = a \cdot c + b \cdot c.$$

Bibliography

Anderson, J. E. & Wincoop, E. v. (2003), 'Gravity with gravitas: A solution to the border puzzle', *The American Economic Review* **93**, 170–192. 147

Angrist, J. & Pischke, J. (2009), *Mostly harmless econometrics. An Empiricist's Companion*, Princeton University Press.

Bauwens, L., Boswijk, H. P. & Urbain, J.-P. (2006), 'Causality and exogeneity in econometrics', *Journal of Econometrics* **132**, 305 – 309. 298

Brockwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods*, 2. edn, Springer, New York, NY. 263, 281, 282, 284, 285

Cameron, A. & Trivedi, P. (2005), *Microeconometrics*, Cambridge University Press.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, 2nd edn, Duxbury - Thomson. 38, 70, 83

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press. 119, 123

Davidson, J. (2000), *Econometric Theory*, Blackwell Publishers. 71, 82, 100, 120, 215, 223, 255, 283, 287, 288, 290, 291, 294, 295, 299, 302, 303

Davidson, R. & MacKinnon, J. (1993), *Estimation and Inference in Econometrics.*, Oxford University Press.
URL: *http://www.oup.com/uk/catalogue/?ci=9780195060119*

Davidson, R. & MacKinnon, J. G. (2004), *Econometric Theory and Methods*, Oxford University Press, Oxford. 26, 38, 55, 99, 101, 103, 104, 107, 109, 114, 123, 137, 138, 143, 148, 160, 162, 163, 164, 165, 166, 167, 191, 207, 211, 212, 237, 239, 240, 241, 243, 311, 312, 320, 360

Engle, R., Hendry, D. & Richard, J.-F. (1983), 'Exogeneity', *Econometrica* **51**, 277–304. 291, 293

Fahrmeier, L., Künstler, R., Pigeot, I. & Tutz, G. (2004), *Statistik*, Spinger. 39

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I. & Tutz, G. (2016), *Statistik: Der Weg zur Datenanalyse*, 8 edn, Springer. 38

Fischer, G. (2010), *Lineare Algebra*, 17 edn, Vieweg + Teubner. 6, 26

Fischer, G. (2014), *Lineare Algebra*, 18 edn, Springer Spektrum.
URL: *http://dx.doi.org/10.1007/978-3-658-03945-5* 3, 367

Fratianni, M. (2007), The gravity equation in international trade, Technical report, Dipartimento di Economia, Universita Politecnica delle Marche. 147

Gentle, J. E. (2007), *Matrix Algebra. Theory, Computations, and Applications in Statistics*, Springer Texts in Statistics, Springer.
URL: *http://www.springerlink.com/content/x4rj03/* 3, 8, 10, 11, 26, 306

Granger, C. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica* **37**, 424 – 438. 294

Greene, W. (2008), *Econometric Analysis*, 6 edn, Pearson.
URL: *http://www.pearsonhighered.com/educator/academic/product/0,3110,0135132452,00.html* 165

Greene, W. (2012), *Econometric Analysis*, 7 edn, Pearson.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press. 263, 272, 276

Hansen, B. E. (2015), *Econometrics*.

Hassler, U. (2007), *Stochastische Integration und Zeitreihenmodellierung*, Springer, Berlin, Heidelberg. 255, 256

Hayashi, F. (2000), *Econometrics*, Princeton University Press, Princeton, NJ [u.a.]. 261, 262

Hendry, D. F. (1995), *Dynamic Econometrics*, Oxford University Press. 123, 294, 295, 299

Horowitz, J. (2001), The bootstrap, *in* J. Heckman & E. Leamer, eds, 'Handbook of Econometrics', Vol. 5, North-Holland. 240

Horowitz, J. (2003), 'The boothstrap in econometrics', *Statistical Science* **18**, 211–218. 240

Kirchgässner, G. & Wolters, J. (2008), *Introduction To Modern Time Series Analysis*, Springer, Berlin, [u.a.]. 276

Kirchgässner, G., Wolters, J. & Hassler, U. (2013), *Introduction To Modern Time Series Analysis*, 2nd. ed. edn, Springer, Berlin, [u.a.]. 263

Kleiber, C. & Zeileis, A. (2008), *Applied Econometrics with R*, Springer. 362

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics*, Princeton University Press. 104

Lucas, R. (1976), Econometric policy evaluation: A critique, *in* K. Brunner & A. Meltzer, eds, 'The Phillips Curve and Labor Markets', Vol. Vol. 1 of *Carnegie-Rochester Conferences on Public Policy*, North-Holland, Amsterdam, pp. 19 – 46. 299

Lütkepohl, H. (1996), *Handbook of Matrices*, John Wiley & Sons, Chichester. 3, 284, 306

Lütkepohl, H. (2004), Vector autoregressive and vector error correction models, *in* H. Lütkepohl & M. Krätzig, eds, 'Applied Time Series Econometrics', Cambridge University Press, Cambridge, chapter 3, pp. 86–158. 294

Lütkepohl, H. & Kraetzig, M. (2008), *Applied Time Series Econometrics*, Cambridge University Press. 263

Mikosch, T. (1998), *Elementary Stochastic Calculus*, World Scientific Publishing, Singapore. 255

Neusser, K. (2006), *Zeitreihenanalyse in den Wirtschaftswissenschaften*, Teubner, Wiesbaden. 263, 264

Neusser, K. (2009), *Zeitreihenanalyse in den Wirtschaftswissenschaften*, 2. edn, Teubner, Wiesbaden. 263, 272, 273, 274

Peracchi, F. (2001), *Econometrics*, John Wiley and Sons.
URL: *http://www.wiley-vch.de/publish/dt/books/bySubjectEC00/ISBN0-471-98764-6/?sID=he2l84vhvc6o6e4f1mc7i17k05*

Robinson, P. M., ed. (2003), *Time Series with Long Memory*, Oxford University Press. 279

Ruud, P. (2000), *An Introduction to Classical Econometric Theory*, Oxford University Press.
URL: *http://www.oup.com/uk/catalogue/?ci=9780195111644* 167

Schmidt, K. & Trenkler, G. (2006), *Einführung in die Moderne Matrix-Algebra. Mit Anwendungen in der Statistik*, Springer. 24, 25, 26, 29, 226

Schmidt, K. & Trenkler, G. (2015), *Einführung in die Moderne Matrix-Algebra. Mit Anwendungen in der Statistik*, 3 edn, Springer. 3

Steland, A. (2010), *Basiswissen Statistik : Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 2., komplett überarb. und erw. aufl. edn, Spinger, Berlin ; Heidelberg : Springer. 45

Steland, A. (2013), *Basiswissen Statistik : Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 3., überarb. und erw. aufl. edn, Spinger, Berlin ; Heidelberg : Springer.
URL: *http://link.springer.com/book/10.1007/978-3-642-37201-8* 38

Steland, A. (2016), *Basiswissen Statistik : Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 4 edn, Spinger, Berlin ; Heidelberg : Springer.
URL: *http://link.springer.com/book/10.1007/978-3-642-37201-8*

Stock, J. H. & Watson, M. W. (2007), *Introduction to Econometrics*, 2nd. edn, Pearson, Boston, Mass. 86, 87, 95

Stock, J. H. & Watson, M. W. (2012), *Introduction to Econometrics*, 3rd. edn, Pearson, Boston, Mass.

Tschernig, R. (1994), *Wechselkurse, Unsicherheit und Long Memory*, Physica-Verlag, Heidelberg.
URL: *http://epub.uni-regensburg.de/6928/* 279

Vaart, A. v. d. (1998), *Asymptotic Statistics*, Cambridge series in statistical and probabilistic mathematics, Cambridge University Press. 81

Verbeek, M. (2012), *A guide to modern econometrics*, Wiley, Chichester.

Wooldridge, J. M. (2009), *Introductory Econometrics. A Modern Approach*, 4th edn, Thomson South-Western, Mason. 60, 63, 64, 107, 131, 134, 137, 138, 159, 165, 174, 175, 185, 188, 189, 190, 191, 203, 212

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press. 119