

Introductory Econometrics

Slides

Rolf Tschernig & Harry Haupt

University of Regensburg

University of Passau

August 2020 ¹

¹These slides were originally designed for the course "Intensive Course in Econometrics" that Rolf Tschernig and Harry Haupt created for the TEMPUS Project "New Curricula in Trade Theory and Econometrics" in 2009. Florian Brezina produced the empirical example for data from Germany. Kathrin Kagerer, Joachim Schnurbus und Roland Jucknewitz, whose unmarried name was Weigand, helped us enormously to improve and correct this course material. Patrick Kratzer wrote most of the R program for the empirical examples using functions written by Roland Jucknewitz. We are greatly indebted to all of them. The version of August 2020 is synchronized with the German version "Kursmaterial für Einführung in die Ökonometrie (Bachelor) — August 2020" in terms of slide numbers and the empirical example. Of course, the usual disclaimer applies. Please send possible errors to rolf.tschernig@ur.de.

© These slides may be printed and reproduced for individual or instructional use but not for commercial purposes.

Please cite as: Rolf Tschernig and Harry Haupt, Introductory Econometrics, Slides, Universität Regensburg, August 2020. Downloaded on [Day Month Year].

Contents

- 1 Introduction: What is Econometrics? 4**
 - 1.1 A Trade Example: What Determines Trade Flows? 4
 - 1.2 Economic Models and the Need for Econometrics 14
 - 1.3 Causality and Experiments 22
 - 1.4 Types of Economic Data 25

- 2 The Simple Regression Model 31**
 - 2.1 The Population Regression Model 32
 - 2.2 The Sample Regression Model 47

2.3	The OLS Estimator	50
2.4	Best Linear Prediction, Correlation, and Causality	67
2.5	Algebraic Properties of the OLS Estimator	74
2.6	Parameter Interpretation and Functional Form	78
2.7	Statistical Properties: Expected Value and Variance	90
2.8	Estimation of the Error Variance	97
3	Multiple Regression Analysis: Estimation	100
3.1	Motivation: The Trade Example Continued	100
3.2	The Multiple Regression Model of the Population	105
3.3	The OLS Estimator: Derivation and Algebraic Properties ..	119
3.4	The OLS Estimator: Statistical Properties	132
3.5	Model Specification I: Model Selection Criteria	166

4	Multiple Regression Analysis: Hypothesis Testing	178
4.1	Basics of Statistical Tests	178
4.2	Probability Distribution of the OLS Estimator	209
4.3	The t Test in the Multiple Regression Model	216
4.4	Empirical Analysis of a Simplified Gravity Equation	224
4.5	Confidence Intervals	235
4.6	Testing a Single Linear Combination of Parameters	247
4.7	The F Test	253
4.8	Reporting Regression Results	279
5	Multiple Regression Analysis: Asymptotics	282
5.1	Large Sample Distribution of the Mean Estimator	283
5.2	Large Sample Inference for the OLS Estimator	298

6	Multiple Regression Analysis: Interpretation	303
6.1	Level and Log Models	303
6.2	Data Scaling	304
6.3	Dealing with Nonlinear or Transformed Regressors	311
6.4	Regressors with Qualitative Data	322
7	Multiple Regression Analysis: Prediction	340
7.1	Prediction and Prediction Error	340
7.2	Statistical Properties of Linear Predictions	347
8	Multiple Regression Analysis: Heteroskedasticity	348
8.1	Consequences of Heteroskedasticity for OLS	351
8.2	Heteroskedasticity-Robust Inference after OLS	354
8.3	The General Least Squares (GLS) Estimator	357
8.4	Feasible Generalized Least Squares (FGLS)	366

9	Multiple Regression Analysis: Model Diagnostics	388
9.1	The RESET Test	388
9.2	Heteroskedasticity Tests	391
9.3	Model Specification II: Useful Tests	410
10	Appendix	I
10.1	A Condensed Introduction to Probability	I
10.2	Important Rules of Matrix Algebra	XXIII
10.3	Rules for Matrix Differentiation	XXX
10.4	Data for Estimating Gravity Equations	XXXII
10.5	R Program for Empirical Examples	XXXVIII

Organisation

Contact

Prof. Dr. Rolf Tschernig

Building RW(L), 5th floor, room 514

Universitätsstr. 31, 93040 Regensburg, Germany

Tel. (+49) 941/943 2737, Fax (+49) 941/943 4917

Email: rolf.tschernig@wiwi.uni-regensburg.de

<https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/>

[index.html](#)

Schedule and Location

see LSF or corresponding homepage

<https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/lehre/bachelor/einfuehrung-in-die-oekonometrie/index.html>

Exam

see corresponding homepage

<https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/lehre/bachelor/einfuehrung-in-die-oekonometrie/index.html>

Required Text

Wooldridge, J.M. (2009). *Introductory Econometrics. A Modern Approach*, 4th ed., Thomson South-Western. Or newer edition.

Additional Reading

Stock, J.H. and Watson, M.W. (2007). *Introduction to Econometrics*, 2nd ed., Pearson, Addison-Wesley. (Or newer edition)

Software

All empirical examples are computed with R (<https://www.r-project.org>). The appendix 10.5 contains all R programs.

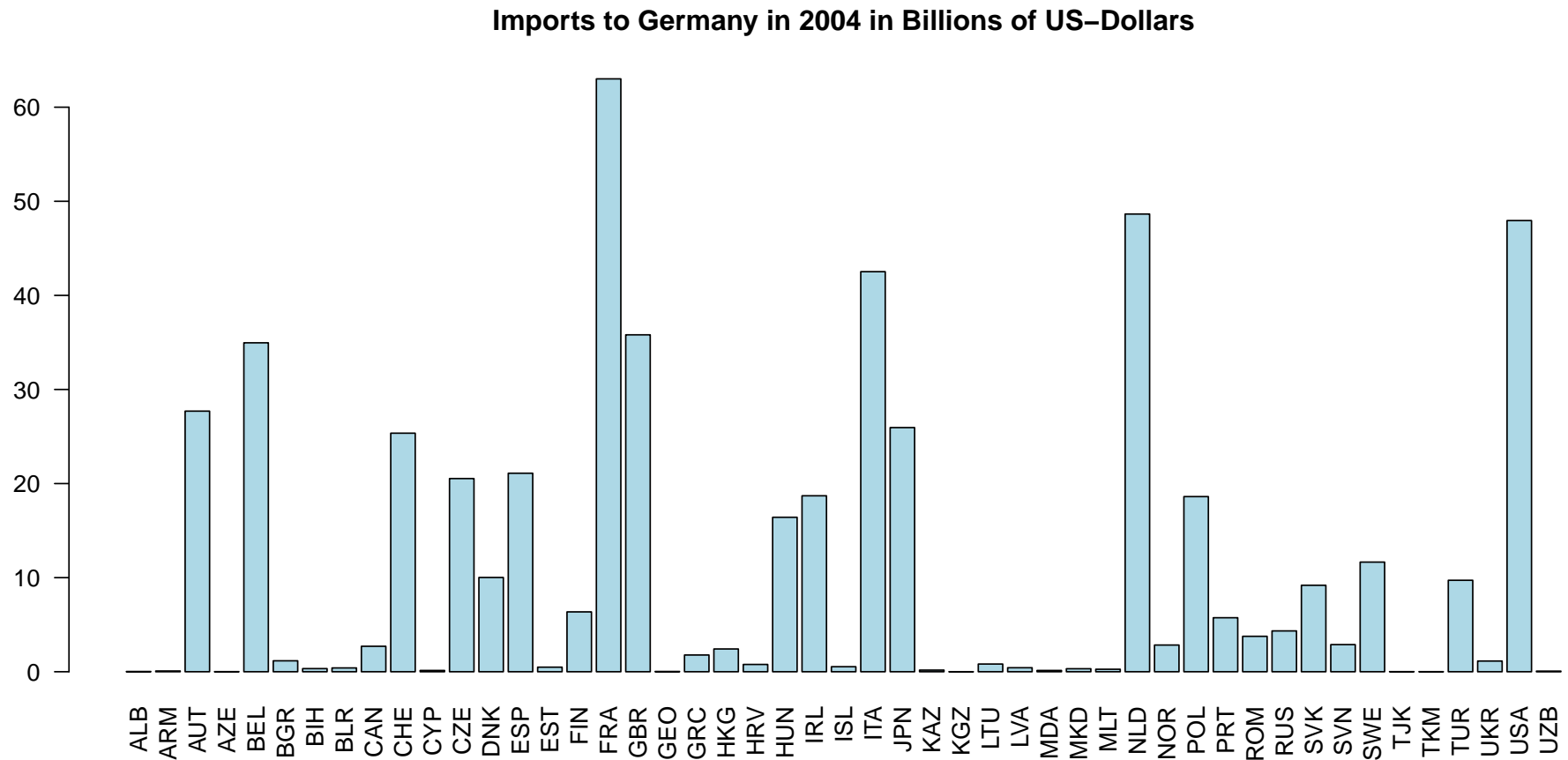
1 Introduction: What is Econometrics?

1.1 A Trade Example: What Determines Trade Flows?

Goal/Research Question: Identify the factors that influence imports to Germany and quantify their impact.

- Three **basic questions** that have to be answered during the analysis:
 1. Which (economic) relationships could be / are “known” to be relevant for this question?
 2. Which data can be useful for checking the possibly relevant economic conjectures/theories?
 3. How to decide about which economic conjecture to reject or to follow?
- Let’s have a first look at some data of interest: the imports (in current US dollars) to Germany from 54 originating countries in 2004.

Imports to Germany in 2004 in current US dollars



The original data are from **UN Commodity Trade Statistics Database (UN COMTRADE)**

- See section 10.4 in the Appendix for detailed data descriptions.

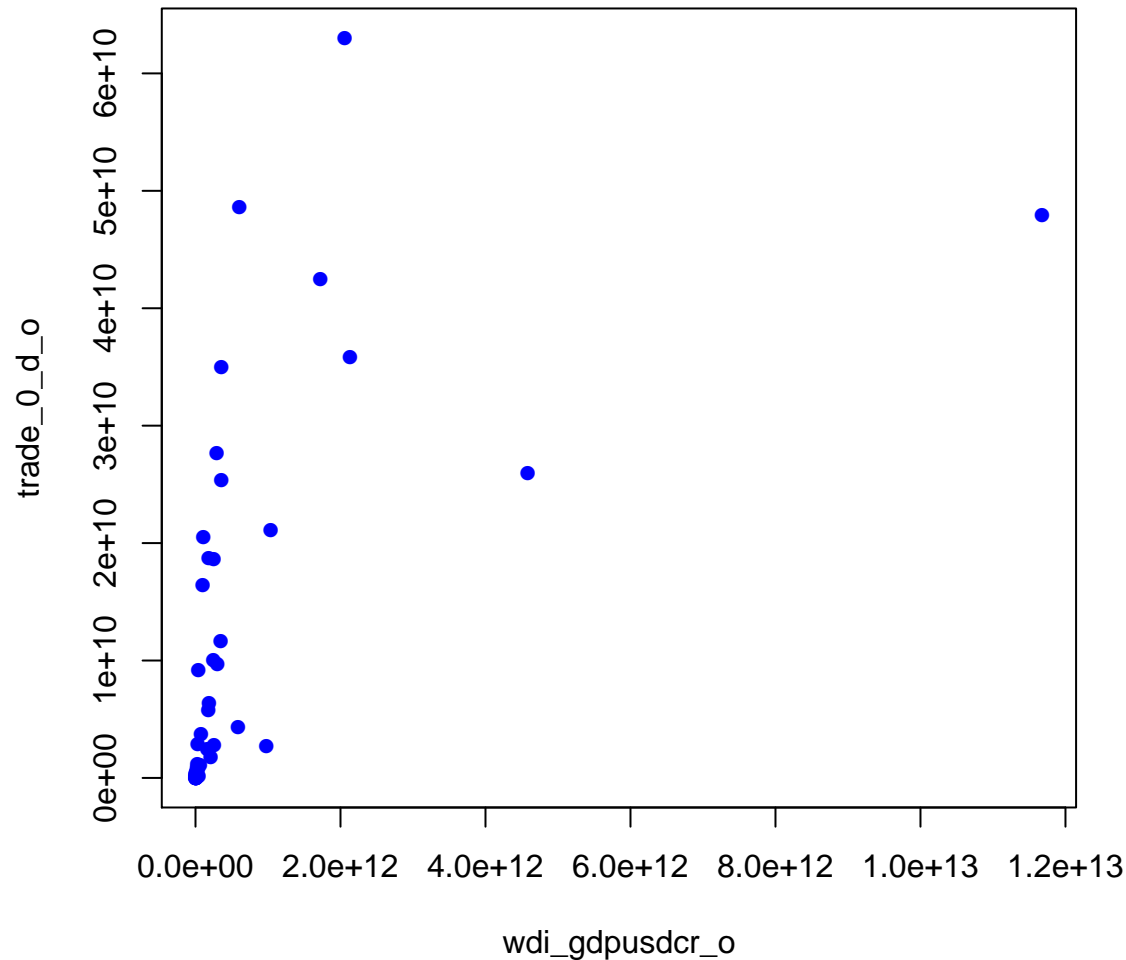
Data are provided in the text file `importe_ger_2004.txt`.

We thank Richard Frensch, Osteuropa-Institut, Regensburg, Germany, who provided all data throughout this course for analyzing trade flows.

- A first attempt to answer the three basic questions:
 1. Ignore for the moment all existing economic theory and simply hypothesize that observed imports depend somehow on the GDP of the exporting country.
 2. Collect GDP data for the countries of origin, e.g. from the [International Monetary Fund \(IMF\) – World Economic Outlook Database](#)
 3. Plot the data, e.g. by using a scatter plot.

Can you decide whether there is a **relationship between trade flows from and the GDP of exporting countries?**

A scatter plot



Some questions:

- What do you see?
- Is there a relationship?
- If so, how to quantify it?
- Is there a causal relationship - what determines what?
- By how much do the imports from the US change if the GDP in Germany changes by 1%?

- Are there other relevant factors determining imports, e.g. distance?
- Is it possible to forecast future trade flows?
- What have we done?
 - We tried to simplify reality
 - by building some kind of (economic) model.

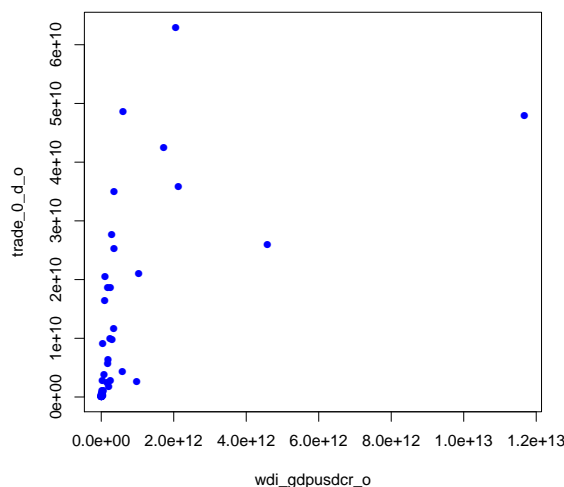
- An **(economic) model**

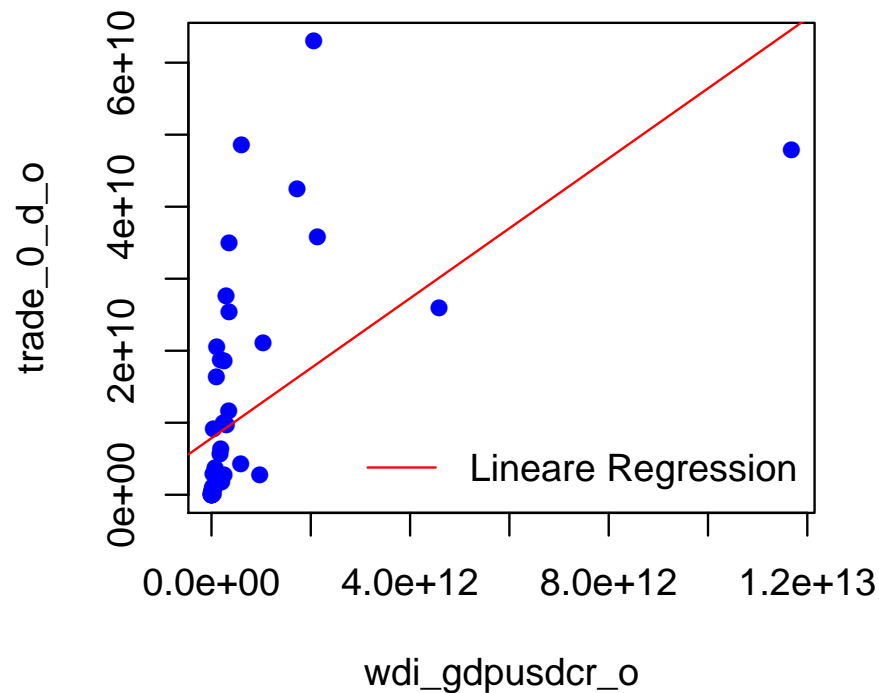
- has to reduce the complexity of reality such that it is useful for answering the question of interest;
- is a collection of cleverly chosen assumptions from which implications can be inferred (using logic) — Example: Heckscher-Ohlin model;
- should be as simple as possible and as complex as necessary;
- cannot be refuted or “validated” without empirical data of some kind.

- Let us consider a simple formal model for the relationship between imports and GDP of the originating countries

$$imports_i = \beta_0 + \beta_1 gdp_i, \quad i = 1, \dots, 49.$$

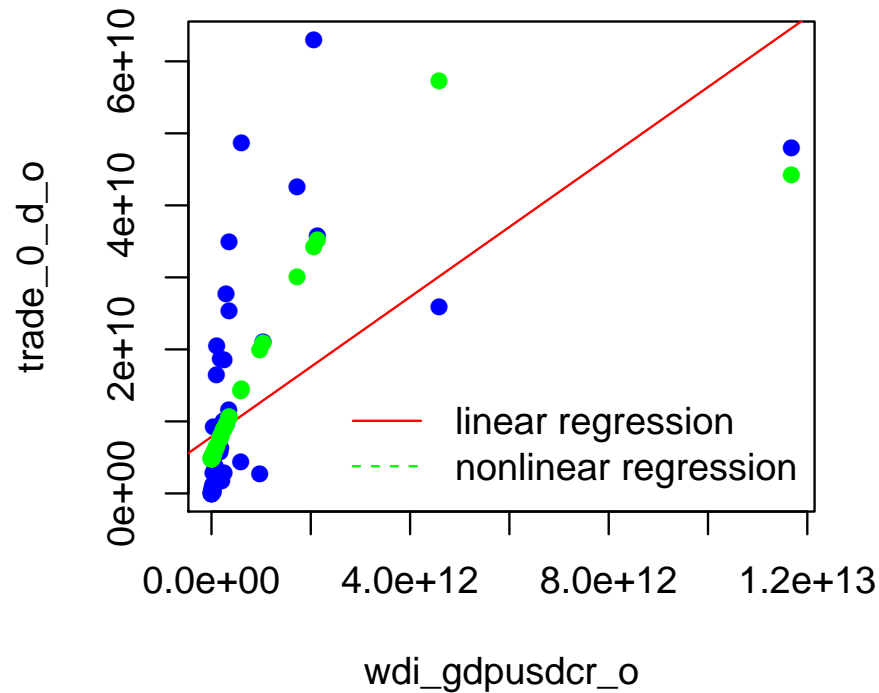
- Does this make sense?
- How to determine the values of the so called parameters β_0 and β_1 ?
- Fit a straight line through the cloud!





More questions:

- How to fit a line through the cloud of points?
- Which properties does the fitted line have?
- What to do with other relevant factors that are currently neglected in the analysis?
- Which criteria to choose for identifying a potential relationship?



Further questions:

- Is the potential relationship really linear? Compare it to the green points of a nonlinear relationship.
- And: how much may results change with a different sample, e.g. for 2003?

1.2 Economic Models and the Need for Econometrics

- **Standard problems of economic models:**

- The conjectured economic model is likely to neglect some factors.
- *Numeric results* to the numerical questions posed depend in general on the choice of a data set. A different data set leads to different numerical results.

⇒ Numeric answers always have some **uncertainty**.

• Econometrics

- offers solutions for dealing with unobserved factors in economic models,
- provides “both a numerical answer to the question and a measure how precise the answer is (Stock and Watson, 2007, p. 7)”,
- as will be seen later, provides tools that allow to refute economic hypotheses using statistical techniques by confronting theory with data *and* to quantify the probability of such decisions to be wrong,
- as will be seen later as well, allows to quantify **risks** of forecasts, decisions, and even of its own analysis.

- Therefore:

Econometrics can also be useful for providing answers to questions like:

- How reliable are predicted growth rates or returns?
- How likely is it that the value realizing in the future will be close to the predicted value? In other words, how precise are the predictions?

- **Main tool:** Multiple regression model

It allows to quantify the effect of a change in one variable on another variable, holding other things constant (**ceteris paribus analysis**).

- **Steps of an econometric analysis:**

1. Careful formulation of question/problem/task of interest.
2. Specification of an economic model.
3. Careful selection of a class of econometric models.
4. Collecting data.
5. Selection and estimation of an econometric model.
6. Diagnostics of correct model specification.
7. Usage of the model.

Note that there exists a large variety of econometric models and model choice depends very much on the research question, the underlying economic theory, availability of data, and the structure of the problem.

- **Goals** of this course:

providing you with basic econometric tools such that you can

- successfully carry out simple empirical econometric analyzes and provide quantitative answers to quantitative questions,
- recognize ill-conducted econometric studies and their consequences,
- recognize when to ask for help of an expert econometrician,
- attend courses for advanced econometrics / empirical economics,
- study more advanced econometric techniques.

Some Definitions of Econometrics

- “... discover empirical relation between economic variables, provide forecast of various economic quantities of interest ... (First issue of volume 1, *Econometrica*, 1933).”
- “The science of model building consists of a set of quantitative tools which are used to construct and then test *mathematical representations of the real world*. The development and use of these tools are subsumed under the subject heading of econometrics **Pindyck and Rubinfeld (1998)**.”

- “At a broad level, econometrics is the science and art of using economic theory and statistical techniques to analyze economic data. Econometric methods are used in many branches of economics, including finance, labor economics, macroeconomics, microeconomics, marketing, and economic policy. Econometric methods are also commonly used in other social sciences, including political science and sociology (Stock and Watson, 2007, p. 3).”

So, some may also say: “Alchemy or Science?”, “Economic-tricks”, “Econo-mystiques” .

- “Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy (Wooldridge, 2009, p. 1).”

- **Summary of tasks for econometric methods**

- **In brief: econometrics can be useful whenever you encounter (economic) data and you want to make sense out of them.**

- **In detail:**

- * **Providing a formal framework for falsifying postulated economic relationships** by confronting economic theory with economic data using statistical methods: Economic hypotheses are formulated and statistically tested on basis of adequately (and repeatedly) collected data such that test results may falsify the postulated hypotheses.

- * **Analyzing the effects of policy measures.**

- * **Forecasting.**

1.3 Causality and Experiments

- Common understanding: “causality means that a specific action” (touching a hot stove) “leads to a specific, measurable consequence” (get burned) ([Stock and Watson, 2007](#), p. 8).
- How to identify causality? Observe **repeatedly** an action and its consequence! However, this approach only allows to draw conclusions on average causality since for one specific action one cannot simultaneously observe outcomes of taking and not taking this action (hand burned, hand not burned).
- Thus, in science one aims at repeating an action and its consequences under **identical** conditions. How to generate repetitions of actions?

- **Randomized controlled experiments:**

- there is a **control group** that receives no treatment (e.g. fertilizer) and a **treatment group** that receives treatment, and
- where **treatment is assigned randomly** in order to eliminate any possible systematic relationship between the treatment and other possible influences.

- **Causal effect:**

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 and Watson, 2007, p. 9).”

- In economics randomized controlled experiments are very often difficult or impossible to conduct. Then a randomized controlled experiment provides a theoretical benchmark and econometric analysis

aims at mimicking as closely as possible the conditions of a randomized controlled experiment using actual data.

- Note that for **forecasting** knowledge of causal effects is not necessary.
- Warning: in general multiple regression models do not allow conclusions about causality!
- A well readable introduction to methods of causality analysis is [Angrist and Pischke \(2015\)](#).

1.4 Types of Economic Data

1. Cross-Sectional Data

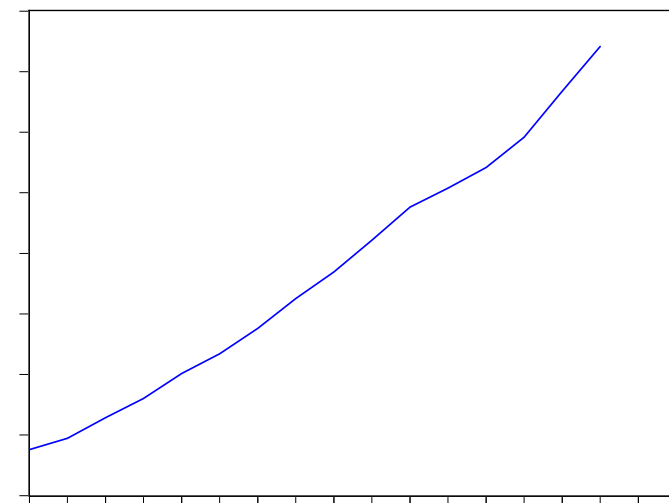
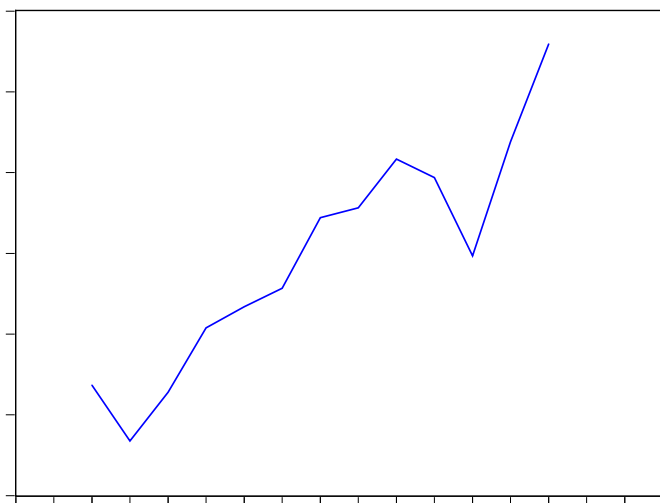
- are collected across several units at a single point or period of time.
- Units: “economic agents”, e.g. individuals, households, investors, firms, economic sectors, cities, countries.
- In general: the order of observations has no meaning.
- Popular to use index i .
- Optimal: the data are a *random sample* of the underlying *population*, see Section 2.1 for details.
- Cross-Sectional data allow to explain differences between individual units.

- Example: sample of countries that export to Germany in 2004 of Section 1.1.

2. Time Series Data (BA: Time Series Econometrics, Quantitative Economic Research I, MA: Methods of Econometrics, Applied Time Series Econometrics, Quantitative Economic Research II)

- are sampled across differing points/periods of time.
- Popular to use index t .
- Sampling frequency is important:
 - variable versus fixed;
 - fixed: annually, quarterly, monthly, weekly, daily, intradaily;
 - variable: ticker data, duration data (e.g. unemployment spells).

- Time series data allow the analysis of dynamic effects.
- Univariate versus multivariate time series data.
- Example: Trade flow from US to Germany and GDP in USA (in current US dollars), 1990 - 2007, $T = 18$.



3. Panel data (BA: Advanced Issues in Econometrics)

- are a collection of cross-sectional data for at least two different points/periods of time.
- Individual units remain identical in each cross-sectional sample (except if units vanish).
- Use of double index: it where $i = 1, \dots, N$ and $t = 1, \dots, T$.
- Typical problem: missing values - for some units and periods there are no data.
- Example: growth rate of imports from 54 different countries to Germany from 1991 to 2008 where all 54 countries were chosen for the sample 1990 and **kept fixed** for all subsequent years ($T = 18, N = 54$).

4. Pooled Cross Sections (BA: Advanced Issues in Econometrics)

- also a collection of cross-sectional data, however, allowing for changing units across time.
- Example: in 1995 countries of origin are the Netherlands, France, Russia and in 1996 countries of origin are Poland, US, Italy.

In this course: focus on the analysis of cross-sectional data and specific types of time series data:

- **simple regression model** → Chapter 2,
- **multiple regression model** → Chapters 3 to 9.
- Time series analysis requires advanced econometric techniques that are beyond the scope of this course (given the time constraints).

Recall the **arithmetic quality of data**:

- quantitative variables,
- qualitative or categorical variables.

Reading: Sections 1.1-1.3 in [Wooldridge \(2009\)](#).

2 The Simple Regression Model

Distinguish between the

- population regression model and the
- sample regression model.

2.1 The Population Regression Model

- **In general:**

y and x are two variables that describe properties of the population under consideration for which one wants “to explain y in terms of x ” or “to study how y varies with changes in x ” or “to predict y for given values of x ”.

Example: By how much changes the hourly wage for an additional year of schooling keeping all other influences fixed?

- **If we knew everything**, then the relationship between y and x may formally be expressed as

$$y = m(x, z^1, \dots, z^s) \quad (2.1)$$

where z^1, \dots, z^s denote s additional variables that in addition to years of schooling x influence the hourly wage y .

- **For practical application** it is possible
 - that relationship (2.1) is too complicated to be useful,
 - that there does not exist an exact relationship, or
 - that there exists an exact relationship for which, however, not all s influential variables z^1, \dots, z^s can be observed, or
 - one has no idea about the structure of the function $m(\cdot)$.
- **Our solution:**
 - **build a useful model**, cf. Section 1.1,
 - which focuses on a relationship that holds **on “average”**. What do we mean by “average”?

- **Crucial building blocks for our model:**

- **Consider the variable y as random.** You may think of y denoting the value of the variable of a random choice out of all units in the population. Furthermore, in case of discrete values of the **random variable** y , a probability is assigned to each value of y . (If the random variable y is continuous, a density value is assigned.)

In other words: apply **probability theory**. See Appendices B and C in [Wooldridge \(2009\)](#).

Examples:

- * The population consists of all apartments in Regensburg. The variable y denotes the rent of a single apartment randomly chosen from all apartments in Regensburg.

- * The population consists of all possible values of imports to Germany from a specific country and period.
 - * For a dice the population consists of all numbers that are written on each side although in this case statisticians prefer to talk about a sample space.
- In terms of probability theory the “average” of a variable y is given by the **expected value** of this variable. In case of discrete y one has

$$E[y] = \sum_{j \in \text{all different } y_j \text{ in population}} y_j \text{Prob}(y = y_j)$$

- Sometimes one may only look at a **subset of the population**, namely all y that have the same value for another variable x .

Example: one only considers the rents of all apartments in Regensburg of size $x = 75 m^2$.

- If the “average” is conditioned on specific values of another variable x , then one considers the **conditional expected value** of y for a given x : $E[y|x]$. For discrete random variables y one has

$$E[y|x] = \sum_{j \in \text{all different } y_j \text{ in population}} y_j \text{Prob}(y = y_j|x)$$

(See Appendix 10.1 for a brief introduction to probability theory and corresponding definitions for continuous random variables.)

Example continued: the conditional expectation $E[y|x = 75]$ corresponds to the average rent of all apartments in Regensburg of size $x = 75 m^2$.

- Note that the variable x can be random, too. Then, the conditional expectation $E[y|x]$ is a function of the (random) variable x

$$E[y|x] = g(x)$$

and therefore a random variable itself.

- From the **identity**

$$y = E[y|x] + (y - E[y|x]) \quad (2.2)$$

one defines the **error term** or **disturbance term** as

$$u \equiv y - E[y|x]$$

so that one obtains a **simple regression model of the population**

$$y = E[y|x] + u \quad (2.3)$$

- **Interpretation:**

- The random variable y varies randomly around the conditional expectation $E[y|x]$:

$$y = E[y|x] + u.$$

- The conditional expectation $E[y|x]$ is called the **systematic part** of the regression.
- The error term u is called the **unsystematic part** of the regression.

- So instead of trying the impossible, namely specifying $m(x, \dots)$ given by (2.1), one focuses the analysis on the “average” $E[y|x]$.

- **How to determine the conditional expectation?**

- This step requires assumptions!

- To keep things simple we make **Assumption (A)** given by

$$E[y|x] = \beta_0 + \beta_1 x. \quad (2.4)$$

- **Discussion** of Assumption (A):

- * **It restricts the flexibility of** $g(x) = E[y|x]$ such that $g(x) = \beta_0 + \beta_1 x$ has to be linear in x . So if $E[y|x] = \delta_0 + \delta_1 \log x$, Assumption A is wrong.

- * It can be fulfilled if there are other variables influencing y linearly. For example, consider

$$E[y|x, z] = \delta_0 + \delta_1 x + \delta_2 z.$$

Then, by the law of iterated expectations one obtains

$$E[y|x] = \delta_0 + \delta_1 x + \delta_2 E[z|x]$$

If $E[z|x]$ is linear in x , one obtains

$$\begin{aligned} E[y|x] &= \delta_0 + \delta_1 x + \delta_2(\alpha_0 + \alpha_1 x) \\ &= \gamma_0 + \gamma_1 x \end{aligned} \tag{2.5}$$

with $\gamma_0 = \delta_0 + \delta_2\alpha_0$ und $\gamma_1 = \delta_1 + \delta_2\alpha_1$. Note, however, that in this case $E[y|x, z] \neq E[y|x]$ in general. Then model choice depends on the goal of the analysis: the smaller model can sometimes be preferable for prediction, the larger model is needed if controlling for z is important \Leftrightarrow controlled random experiments, see Section 1.3.

* In general, Assumption (A) is violated if (2.5) does not hold e.g. if $E[z|x]$ is nonlinear in x . Then the linear population model is called **misspecified**. More on that in Section 3.4.

- **Properties of the error term u : From Assumption (A)**

1. $E[u|x] = 0$,

2. $E[u] = 0$,

3. $Cov(x, u) = 0$.

- An **alternative set of assumptions**:

The above result $E[u|x] = 0$ together with the identity (2.3) allows to rewrite **Assumption (A)** in terms of the following two assumptions:

1. **Assumption SLR.1**
(Linearity in the Parameters)

$$y = \beta_0 + \beta_1 x + u, \quad (2.6)$$

2. **Assumption SLR.4**
(Zero Conditional Mean)

$$E[u|x] = 0.$$

- **Linear Population Regression Model:**

The simple linear population regression model is given by equation (2.6)

$$y = \beta_0 + \beta_1 x + u$$

and obtained by specifying the conditional expectation in the regression model (2.3) by a linear function (linear in the parameters).

The parameters β_0 and β_1 are called the **intercept parameter** and **slope parameter**, respectively.

- **Some terminology for regressions**

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor
	Covariate

- A simple **example**: a game of dice

Let the random numbers x and u denote the throws of two fair dices with $x, u = \{-2.5, -1.5, -0.5, 0.5, 1.5, 2.5\}$. Based on both throws the random number y denotes the following sum

$$y = \underbrace{2}_{\beta_0} + \underbrace{3}_{\beta_1} x + u.$$

This completely describes the population regression model.

- Derive the systematic relationship between y and x holding x fixed.
- Interpret the systematic relationship.
- How can you obtain the values of the parameters $\beta_0 = 2$ and $\beta_1 = 3$ if those values are unknown?

Next section: How can you determine/estimate β_0 and β_1 ?

2.2 The Sample Regression Model

Estimators and Estimates

- In practice one has to estimate the unknown parameters β_0 and β_1 of the population regression model using a sample of observations.
- The sample has to be **representative** and has to be collected/-drawn from the population.
- A sample of the random numbers x and y of size n is given by $\{(x_i, y_i) : i = 1, \dots, n\}$.
- Now we require an **estimator** that allows us — given the sample observations $\{(x_i, y_i) : i = 1, \dots, n\}$ — to compute **estimates** for the unknown parameters β_0 and β_1 of the population.

- **Note:**

- If we want to construct an estimator for the unknown parameters, we have not yet observed a sample. An **estimator** is a function that contains the sample values as arguments.
- Once we have an estimator *and* observe a sample, we can compute **estimates** (=numerical values) for the unknown quantities.

- For estimating the unknown parameters there exist many different estimators that differ with respect to their statistical properties (statistical quality)!

Example: Two different estimators for estimating the mean:

$$\frac{1}{n} \sum_{i=1}^n y_i \text{ and } \frac{1}{2} (y_1 + y_n).$$

- If you denote estimators of the parameters β_0 and β_1 in the population regression model

$$y = \beta_0 + \beta_1 x + u$$

by $\tilde{\beta}_0$ and $\tilde{\beta}_1$, then the sample regression model is given by

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{u}_i, \quad i = 1, \dots, n.$$

It consists of

- the **sample regression function** or **regression line**

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i,$$

- the **fitted values** \tilde{y}_i , and
- the **residuals** $\tilde{u}_i = y_i - \tilde{y}_i$, $i = 1, \dots, n$.

With which method can we estimate?

2.3 The Ordinary Least Squares Estimator (OLS) Estimator

- The ordinary least squares estimator is frequently abbreviated as OLS estimator. The OLS estimator goes back to C.F. Gauss (1777-1855).
- It is derived by choosing the values $\tilde{\beta}_0$ and $\tilde{\beta}_1$ such that the **sum of squared residuals (SSR)**

$$\sum_{i=1}^n \tilde{u}_i^2 = \sum_{i=1}^n \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right)^2$$

is **minimized**.

- One computes the first partial derivatives with respect to $\tilde{\beta}_0$ and $\tilde{\beta}_1$ and sets them equal to zero:

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0, \quad (2.7)$$

$$\sum_{i=1}^n x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0. \quad (2.8)$$

The equations (2.7) and (2.8) are called **normal equations**.

It is important to understand the interpretation of the normal equations.

From (2.7) one obtains

$$\begin{aligned}\hat{\beta}_0 &= n^{-1} \sum_{i=1}^n y_i - \hat{\beta}_1 n^{-1} \sum_{i=1}^n x_i \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x},\end{aligned}\tag{2.9}$$

where $\bar{z} = n^{-1} \sum_{i=1}^n z_i$ denotes the estimated mean of z_i , $i = 1, \dots, n$.

Inserting (2.9) into the normal equation (2.8) delivers

$$\sum_{i=1}^n x_i \left(y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) - \hat{\beta}_1 x_i \right) = 0.$$

Moving terms leads to

$$\sum_{i=1}^n x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}).$$

Note that

$$\sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$
$$\sum_{i=1}^n x_i(x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2,$$

such that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (2.10)$$

- **Terminology:**

- The sample functions (2.9) and (2.10)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

are called the **ordinary least squares (OLS) estimators** for β_0 and β_1 .

- For a given sample, the quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **OLS estimates** for β_0 and β_1 .

- The **OLS sample regression function** or **OLS regression line** for the simple regression model is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (2.11)$$

with **residuals** $\hat{u}_i = y_i - \hat{y}_i$.

- The **OLS sample regression model** is denoted by

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i \quad (2.12)$$

Note:

- The OLS estimator $\hat{\beta}_1$ only exists if the sample observations x_i , $i = 1, \dots, n$ exhibit variation.

Assumption SLR.3**(Sample Variation in the Explanatory Variable):**

In the sample the outcomes of the independent variable x_i , $i = 1, 2, \dots, n$ are not all the same.

- The derivation of the OLS estimator only requires assumption SLR.3 but not the population Assumptions SLR.1 and SLR.4.

- In order to investigate the statistical properties of the OLS estimator one needs further assumptions, see Sections 2.7, 3.4, 4.2.
- One also can derive the OLS estimator from the assumptions about the population, see below.
- The OLS estimator as a **Moment Estimator**:
 - Note that from Assumption SLR.4 $E[u|x] = 0$ one obtains two conditions on moments: $E[u] = 0$ and $Cov(x, u) = 0$. Inserting Assumption SLR.1 $u = y - \beta_0 - \beta_1 x$ defines **moment conditions** for the model parameters

$$E(y - \beta_0 - \beta_1 x) = 0 \quad (2.13)$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0 \quad (2.14)$$

- How to estimate the moment conditions using sample functions?
- **Assumption SLR.2 (Random Sampling):**
The sample of size n is obtained by random sampling that is, the pairs (x_i, y_i) and (x_j, y_j) , $i \neq j$, $i, j = 1, \dots, n$, are pairwise identically and independently distributed following the population model.
- An important result in statistics, see Section 5.1, says:
If Assumption SLR.2 holds, then the expected value can well be estimated by the **sample average**. (Assumption SLR.2 can be weakened, see e.g. Chapter 11 in [Wooldridge \(2009\)](#).)

- If one replaces the expected values in (2.13) and (2.14) by their sample averages, one obtains

$$n^{-1} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0, \quad (2.15)$$

$$n^{-1} \sum_{i=1}^n x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0. \quad (2.16)$$

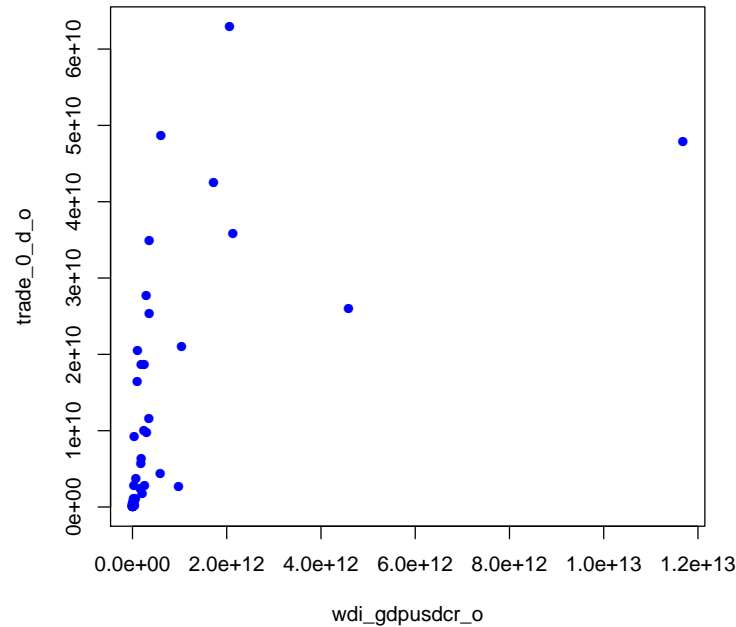
By multiplying (2.15) (2.16) by n one obtains the normal equations (2.7) and (2.8).

The Trade Example Continued

Question:

Do imports to Germany increase if the exporting country experiences an increase in GDP?

Scatter plot (from Section 1.1)

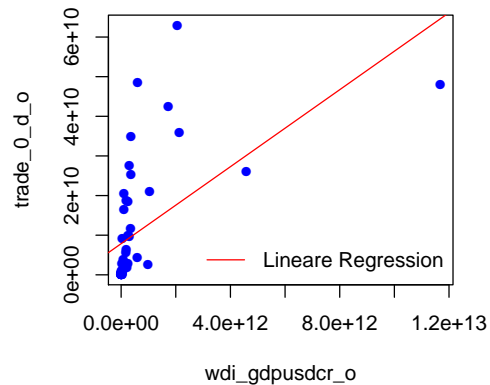


The OLS regression line is given by

$$\widehat{Importe}_i = 7.86 \cdot 10^{09} + 4.857 \cdot 10^{-03} BIP_i, \quad i = 1, \dots, 49,$$

and the sample regression model by

$$Importe_i = 7.86 \cdot 10^{09} + 4.857 \cdot 10^{-03} BIP_i + \hat{u}_i, \quad i = 1, \dots, 49.$$



R-Output

Call:

```
lm(formula = trade_0_d_o ~ wdi_gdpusdcr_o)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.663e+10	-7.736e+09	-6.815e+09	2.094e+09	4.515e+10

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.858e+09	1.976e+09	3.977	0.000239 ***
wdi_gdpusdcr_o	4.857e-03	1.052e-03	4.617	3.03e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.31e+10 on 47 degrees of freedom

Multiple R-squared: 0.3121, Adjusted R-squared: 0.2974

F-statistic: 21.32 on 1 and 47 DF, p-value: 3.027e-05

- For a data description see Appendix [10.4](#):

imports_i (from country *i*) TRADE_0_D_O

gdp_i (in exporting country *i*) WDI_GDPUSDCR_O

- Potential interpretation of estimated slope parameter:

$$\hat{\beta}_1 = \frac{\widehat{\Delta imports}}{\Delta gdp}$$

indicates by how many US dollars average imports in Germany increase if GDP in an exporting country increases by 1 US dollar.

- Does this interpretation really make sense? Aren't there other important influencing factors missing? What about using economic theory as well?
- What about the quality of the estimates?

Example: Wage Regression

Question:

How does education influence the hourly wage of an employee?

- **Data** (Source: Example 2.4 in [Wooldridge \(2009\)](#)): Sample of U.S. employees with $n = 526$ observations. Available data are:
 - wage per hour in dollars and
 - educ years of schooling of each employee.
- The OLS regression line is given by

$$\hat{wage}_i = -0.90 + 0.54 educ_i, \quad i = 1, \dots, 526.$$

The sample regression model is

$$wage_i = -0.90 + 0.54 educ + \hat{u}_i, \quad i = 1, \dots, 526$$

Call:

```
lm(formula = wage ~ educ)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-5.3396 -2.1501 -0.9674  1.1921 16.6085
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.90485    0.68497  -1.321   0.187
educ         0.54136    0.05325  10.167 <2e-16 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.378 on 524 degrees of freedom

Multiple R-squared: 0.1648, Adjusted R-squared: 0.1632

F-statistic: 103.4 on 1 and 524 DF, p-value: < 2.2e-16

- Interpretation of the estimated slope parameter:

$$\hat{\beta}_1 = \frac{\widehat{\Delta wage}}{\Delta educ}$$

indicates by how much the average hourly wage changes if the years of schooling increases by one year:

- An additional year in school or university increases the hourly wage by 54 cent.
- But: Somebody without any education earns an hourly wage of -90 cent? Does this interpretation make sense?
- Is it always sensible to interpret the slope coefficient? Watch out spurious causality, see next section.
- Are these estimates reliable or good in some sense? What do we mean by “good” in econometrics and statistics? To get more insight study
 - the statistical properties of the OLS estimator and the OLS estimates, see Section 2.7 and
 - check the choice of the functional form for the conditional expectation $E[y|x]$, see Section 2.6.

2.4 Best Linear Prediction, Correlation, and Causality

Best Linear Prediction

- What does the OLS estimator estimate if Assumptions SLR.1 and SLR. 4 (alias Assumption (A)) **are not valid** in the population from which the sample is drawn?
- Note that $SSR(\gamma_0, \gamma_1)/n = \sum_{i=1}^n (y_i - \gamma_0 - \gamma_1 x_i)^2 / n$ is a sample average and thus estimates the expected value

$$E \left[(y - \gamma_0 - \gamma_1 x)^2 \right] \quad (2.17)$$

if Assumption SLR.2 (or some weaker form) holds. (For existence of (2.17) it is required that $0 < Var(x) < \infty$ and $Var(y) < \infty$.)

Equation (2.17) is called the **mean squared error** of a **linear predictor**

$$\gamma_0 + \gamma_1 x.$$

- Mimicking minimizing $SSR(\gamma_0, \gamma_1)$, the theoretically best fit of a linear predictor $\gamma_0 + \gamma_1 x$ to y is obtained by minimizing its mean squared error (2.17) with respect to γ_0 and γ_1 . This leads (try to derive it) to

$$\gamma_0^* = E[y] - \gamma_1^* E[x], \quad (2.18)$$

$$\gamma_1^* = \frac{Cov(x, y)}{Var(x)} = Corr(x, y) \sqrt{\frac{Var(y)}{Var(x)}} \quad (2.19)$$

with

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}, \quad -1 \leq Corr(x, y) \leq 1$$

denoting the **correlation** that measures the linear dependence between two variables in a population, here x and y .

The expression

$$\gamma_0^* + \gamma_1^* x \quad (2.20)$$

is called the **best linear predictor** of y where “best” is defined by minimal mean squared error.

- Now observe that for the simple regression model

$$y = \gamma_0^* + \gamma_1^* x + \varepsilon$$

one has $Cov(x, \varepsilon) = 0$, a weaker form of SLR.4, since

$$Cov(x, y) = \frac{Cov(x, y)}{Var(x)} Var(x) + Cov(x, \varepsilon).$$

This indicates that one can show that under Assumption SLR.2 and SLR.3 the **OLS estimator estimates the parameters γ_0^* and γ_1^* of the best linear predictor**. Observe also that the OLS estimator (2.10) for the slope coefficient consists of the sample averages of the moments defining γ_1^*

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Rewriting $\hat{\gamma}_1$ as

$$\hat{\gamma}_1 = \widehat{Corr}(x, y) \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

using the **empirical correlation coefficient**

$$\widehat{Corr}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

shows that the estimated slope coefficient **is non-zero if** there is sample correlation between x and y .

Causality

- Recall Section 1.3.
- **Be aware** that the **slope coefficient of the best linear predictor** γ_1^* **and its OLS estimate** $\hat{\gamma}_1$ **cannot be automatically interpreted in terms of a causal relationship** since estimating the best linear predictor
 - only captures correlation but not direction,
 - may not estimate the model of interest, e.g. if Assumptions SLR.1 and SLR.4 are violated and $\beta_1 \neq \gamma_1^*$,

— may produce garbage if

- * relevant control variables are missing in the simple regression model such that the results cannot represent results of a fictive randomized controlled experiment, see Chapter 3 onwards, or
- * $\widehat{Corr}(x, y)$ estimates spurious correlation ($Corr(x, y) = 0$ and Assumption SLR.2 (or its weaker versions) are violated).

Therefore, before any causal interpretation takes place one has to use specification and diagnostic techniques for regression models. Furthermore, it is important to realize the importance of identification assumptions and to understand the limits of every empirical causality analysis.

2.5 Algebraic Properties of the OLS Estimator

Basic properties:

- $\sum_{i=1}^n \hat{u}_i = 0$, because of normal equation (2.7),
- $\sum_{i=1}^n x_i \hat{u}_i = 0$, because of normal equation (2.8).
- The point (\bar{x}, \bar{y}) lies on the regression line.

Can you provide some intuition for these properties?

- **Total sum of squares (SST)**

$$\text{SST} \equiv \sum_{i=1}^n (y_i - \bar{y})^2$$

- **Explained sum of squares (SSE)**

$$\text{SSE} \equiv \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- **Sum of squared residuals (SSR)**

$$\text{SSR} \equiv \sum_{i=1}^n \hat{u}_i^2$$

- The **decomposition** $\text{SST} = \text{SSE} + \text{SSR}$ holds if the regression model contains an intercept β_0 .

- **Coefficient of Determination R^2 or (R-squared)**

$$R^2 = \frac{\text{SSE}}{\text{SST}}.$$

- Interpretation: share of variation of y_i that is explained by the variation of x_i .
- If the regression model contains an intercept term β_0 , then

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$$

due to the decomposition $\text{SST} = \text{SSE} + \text{SSR}$, and therefore

$$0 \leq R^2 \leq 1.$$

- Later we will see: Choosing regressors with R^2 is in general misleading.

Reading:

- Sections 1.4 and 2.1-2.3 in [Wooldridge \(2009\)](#) and Appendix [10.1](#) if needed.
- 2.4 and 2.5 in [Wooldridge \(2009\)](#).

2.6 Parameter Interpretation and Functional Form and Data Transformation

- The term *linear* in “simple linear regression models” does not imply that the relationship between the explained and the explanatory variable is linear. Instead it refers to the fact that the parameters β_0 and β_1 enter the model linearly.
- Examples for regression models that are *linear* in their parameters:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

$$y_i = \beta_0 + \beta_1 \ln x_i + u_i,$$

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + u_i,$$

$$\ln y_i = \beta_0 + \beta_1 x_i + u_i,$$

$$y_i = \beta_0 + \beta_1 x_i^2 + u_i.$$

The Natural Logarithm in Econometrics

Frequently variables are transformed by taking the natural logarithm \ln . Then the interpretation of the slope coefficient has to be adjusted accordingly.

Taylor approximation of the logarithmic function:

$\ln(1 + z) \approx z$ if z is close to 0.

Using this approximation one can derive a popular approximation of growth rates or returns

$$\begin{aligned} (\Delta x_t)/x_{t-1} &\equiv (x_t - x_{t-1})/x_{t-1} \\ &\approx \ln(1 + (x_t - x_{t-1})/x_{t-1}), \\ (\Delta x_t)/x_{t-1} &\approx \ln(x_t) - \ln(x_{t-1}). \end{aligned}$$

which approximates well if the relative change $\Delta x_t/x_{t-1}$ is close to 0.

One obtains percentages by multiplying with 100:

$$100\Delta \ln(x_t) \approx \% \Delta x_t = 100(x_t - x_{t-1})/x_{t-1}.$$

Thus, the percentage change for small $\Delta x_t/x_{t-1}$ can be well approximated by $100[\ln(x_t) - \ln(x_{t-1})]$.

- Examples of models that are *nonlinear* in the parameters $(\beta_0, \beta_1, \gamma, \lambda, \pi, \delta)$:

$$y_i = \beta_0 + \beta_1 x_i^\gamma + u_i,$$

$$y_i^\gamma = \beta_0 + \beta_1 \ln x_i + u_i,$$

$$y_i = \beta_0 + \beta_1 x_i + \frac{1}{1 + \exp(\lambda(x_i - \pi))} (\gamma + \delta x_i) + u_i.$$

- The last example allows for smooth switching between two linear regimes. The possibilities for formulating nonlinear regression models are huge. However, their estimation requires more advanced

methods such as nonlinear least squares that are beyond the scope of this course.

- Note, however, that linear regression models allow for a wide range of nonlinear relationships between the dependent and independent variables, some of which were listed at the beginning of this section.

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 $\Delta y \rightarrow 0$ and $\Delta x \rightarrow 0$, then it can be shown that $\frac{\Delta y}{\Delta x} \rightarrow \frac{dy}{dx}$.

- If this result is applied to the ratio above, one obtains the **elasticity**

$$\eta(x) = \frac{dy}{dx} \frac{x}{y}.$$

- **Interpretation:** If the relative change of x is 0.01, then the relative change of y given by $0.01\eta(x)$.

In other words: If x changes by 1%, then y changes by $\eta(x)\%$.

- If y, x are **random variables**, then the elasticity is defined with respect to the conditional expectation of y given x :

$$\eta(x) = \frac{dE[y|x]}{dx} \frac{x}{E[y|x]}.$$

This can be derived from

$$\frac{\frac{E[y|x_1=x_0+\Delta x] - E[y|x_0]}{E[y|x_0]}}{\frac{\Delta x}{x_0}} = \frac{E[y|x_1 = x_0 + \Delta x] - E[y|x_0]}{\Delta x} \frac{x_0}{E[y|x_0]}$$

and letting $\Delta x \rightarrow 0$.

Different Models and Interpretations of β_1

For each model it is assumed that SLR.1 and SLR.4 hold.

- **Models that are linear with respect to their variables (level-level models)**

$$y = \beta_0 + \beta_1 x + u.$$

It holds that

$$\frac{dE[y|x]}{dx} = \beta_1$$

and thus

$$\Delta E[y|x] = \beta_1 \Delta x.$$

In words:

The slope coefficient denotes the absolute change in the conditional expectation of the dependent variable y for a one-unit change in the independent variable x .

- **Level-log models**

$$y = \beta_0 + \beta_1 \ln x + u.$$

It holds that

$$\frac{dE[y|x]}{dx} = \beta_1 \frac{1}{x}$$

and thus approximately

$$\Delta E[y|x] \approx \beta_1 \Delta \ln x = \frac{\beta_1}{100} 100 \Delta \ln x \approx \frac{\beta_1}{100} \% \Delta x.$$

In words:

The conditional expectation of y changes by $\beta_1/100$ units if x changes by 1%.

- **Log-level models**

$$\ln y = \beta_0 + \beta_1 x + u$$

or

$$y = e^{\ln y} = e^{\beta_0 + \beta_1 x + u} = e^{\beta_0 + \beta_1 x} e^u.$$

Thus

$$E[y|x] = e^{\beta_0 + \beta_1 x} E[e^u|x].$$

If $E[e^u|x]$ is constant, then

$$\frac{dE[y|x]}{dx} = \beta_1 \underbrace{e^{\beta_0 + \beta_1 x} E[e^u|x]}_{E[y|x]} = \beta_1 E[y|x].$$

One obtains the approximation

$$\frac{\Delta E[y|x]}{E[y|x]} \approx \beta_1 \Delta x, \quad \text{or} \quad \% \Delta E[y|x] \approx 100 \beta_1 \Delta x$$

In words: The conditional expectation of y changes by $100 \beta_1 \%$ if x changes by one unit.

- **Log-log models**

are frequently called **loglinear models** or **constant-elasticity models** and are very popular in empirical work

$$\ln y = \beta_0 + \beta_1 \ln x + u.$$

Similar to above one can show that

$$\frac{dE[y|x]}{dx} = \beta_1 \frac{E[y|x]}{x}, \quad \text{and thus} \quad \beta_1 = \eta(x)$$

if $E[e^u|x]$ is constant.

In these models the slope coefficient is interpreted as the elasticity between the level variables y and x .

In words: The conditional expectation of y changes by $\beta_1\%$ if x changes by 1%.

The Trade Example Continued

R-Output

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

Multiple R-squared: 0.7654, Adjusted R-squared: 0.7604

F-statistic: 153.4 on 1 and 47 DF, p-value: < 2.2e-16

Note the very different interpretation of the estimated slope coefficient $\hat{\beta}_1$:

- Level-level model (Section 2.3): an increase in GDP in the exporting country by 1 billion US dollars corresponds to an average increase of imports to Germany by 4.857 million US dollars.
- Log-log model: an 1%-increase of GDP in the exporting country corresponds to an average increase of imports by 1.077%.

But wait before you take these numbers seriously.

2.7 Statistical Properties of the OLS Estimator: Expected Value and Variance

- Some preparatory transformations (all sums are indexed by $i = 1, \dots, n$):

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{j=1}^n (x_j - \bar{x})^2} \\ &= \sum_{i=1}^n \underbrace{\left[\frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} \right]}_{w_i} y_i = \sum w_i y_i\end{aligned}$$

where it can be shown that (try it):

$$\sum w_i = 0, \quad \sum w_i x_i = 1 \quad \text{and} \quad \sum w_i^2 = \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2}.$$

- **Unbiasedness of the OLS estimator:**

If Assumptions SLR.1 to SLR.4 hold, then

$$E[\hat{\beta}_0] = \beta_0,$$

$$E[\hat{\beta}_1] = \beta_1.$$

Interpretation:

If one keeps repeatedly drawing new samples and estimating the regression parameters, then the average of all obtained OLS parameter estimates roughly corresponds to the population parameters.

The property of unbiasedness is a property of the sample distribution of the OLS estimators for β_0 and β_1 . It **does not** imply that the population parameters are perfectly estimated for a **specific** sample.

Proof for $\hat{\beta}_1$ (clarify where each SLR assumption is needed):

1. $E \left[\hat{\beta}_1 \mid x_1, \dots, x_n \right]$ can be manipulated as follows:

$$\begin{aligned}
 &= E \left[\sum w_i y_i \mid x_1, \dots, x_n \right] \\
 &= E \left[\sum w_i (\beta_0 + \beta_1 x_i + u_i) \mid x_1, \dots, x_n \right] \\
 &= \sum E [w_i (\beta_0 + \beta_1 x_i + u_i) \mid x_1, \dots, x_n] \\
 &= \beta_0 \sum w_i + \beta_1 \sum w_i x_i + \sum E [w_i u_i \mid x_1, \dots, x_n] \\
 &= \beta_1 + \sum w_i E [u_i \mid x_1, \dots, x_n] \\
 &= \beta_1 + \sum w_i E [u_i \mid x_i] \\
 &= \beta_1.
 \end{aligned}$$

2. From $E[\hat{\beta}_1] = E[E[\hat{\beta}_1 \mid x_1, \dots, x_n]]$ one obtains unbiasedness
 $E[\hat{\beta}_1] = \beta_1.$

- **Variance of the OLS estimator**

In order to determine the variance of the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we need another assumption,

Assumption SLR.5 (Homoskedasticity):

$$\text{Var}(u|x) = \sigma^2.$$

- **Variances of parameter estimators conditional on the sample observations**

If Assumptions SLR.1 to SLR.5 hold, then

$$\text{Var} \left(\hat{\beta}_1 \mid x_1, \dots, x_n \right) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\text{Var} \left(\hat{\beta}_0 \mid x_1, \dots, x_n \right) = \sigma^2 \frac{n^{-1} \sum x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Proof (for the conditional variance of $\hat{\beta}_1$):

$$\begin{aligned} & \text{Var} \left(\hat{\beta}_1 \mid x_1, \dots, x_n \right) \\ &= \text{Var} \left(\sum w_i u_i \mid x_1, \dots, x_n \right) \\ &= \sum \text{Var} (w_i u_i \mid x_1, \dots, x_n) \\ &= \sum w_i^2 \text{Var} (u_i \mid x_1, \dots, x_n) \\ &= \sum w_i^2 \text{Var} (u_i \mid x_i) \\ &= \sum w_i^2 \sigma^2 \\ &= \sigma^2 \sum w_i^2 \\ &= \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2}. \end{aligned}$$

• **Covariance between the intercept and the slope estimator:**

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1 | x_1, \dots, x_n) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Proof: $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1 | x_1, \dots, x_n)$ can be manipulated as follows:

$$\begin{aligned} &= \text{Cov} \left(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 \mid x_1, \dots, x_n \right) \\ &= \underbrace{\text{Cov} \left(\bar{u}, \hat{\beta}_1 \mid x_1, \dots, x_n \right)}_{=0 \text{ see below}} - \text{Cov} \left(\hat{\beta}_1 \bar{x}, \hat{\beta}_1 \mid x_1, \dots, x_n \right) \\ &= -\bar{x} \text{Cov} \left(\hat{\beta}_1, \hat{\beta}_1 \mid x_1, \dots, x_n \right) \\ &= -\bar{x} \text{Var} \left(\hat{\beta}_1 \mid x_1, \dots, x_n \right) \\ &= -\sigma^2 \frac{\bar{x}}{\sum (x_i - \bar{x})^2}. \end{aligned}$$

$$\begin{aligned}
& Cov \left(\bar{y}, \hat{\beta}_1 \mid x_1, \dots, x_n \right) \\
&= Cov \left(\beta_0 + \beta_1 \bar{x} + \bar{u}, \sum w_i y_i \mid x_1, \dots, x_n \right) \\
&= Cov \left(\bar{u}, \sum w_i (\beta_0 + \beta_1 x_i + u_i) \mid x_1, \dots, x_n \right) \\
&= Cov \left(\bar{u}, \sum w_i u_i \mid x_1, \dots, x_n \right) \\
&= Cov (\bar{u}, w_1 u_1 \mid x_1, \dots, x_n) + \dots + Cov (\bar{u}, w_n u_n \mid x_1, \dots, x_n) \\
&= w_1 Cov (\bar{u}, u_1 \mid x_1, \dots, x_n) + \dots + w_n Cov (\bar{u}, u_n \mid x_1, \dots, x_n) \\
&= \sum w_i Cov (\bar{u}, u_i \mid x_1, \dots, x_n) \\
&= Cov (\bar{u}, u_1 \mid x_1, \dots, x_n) \sum w_i \\
&= 0.
\end{aligned}$$

2.8 Estimation of the Error Variance

- One possible estimator for the error variance σ^2 is given by

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2,$$

where the \hat{u}_i 's denote the residuals of the OLS estimator.

Disadvantage: The estimator $\tilde{\sigma}^2$ does not take into account that 2 restrictions were imposed on obtaining the OLS residuals, namely:

$$\sum \hat{u}_i = 0, \quad \sum \hat{u}_i x_i = 0.$$

This leads to biased estimates, $E[\tilde{\sigma}^2 | x_1, \dots, x_n] \neq \sigma^2$.

- **Unbiased estimator for the error variance:**

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2.$$

- If Assumptions SLR.1 to SLR.5 hold, then

$$E[\hat{\sigma}^2 | x_1, \dots, x_n] = \sigma^2.$$

- **Standard error of the regression, standard error of the estimate or root mean squared error:**

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

- In the formulas for the variances of and covariance between the parameter estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ the variance estimator $\hat{\sigma}^2$ can be used for estimating the unknown error variance σ .

Example:

$$\widehat{Var}(\hat{\beta}_1 | x_1, \dots, x_n) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}.$$

Denote the standard deviation as

$$sd(\hat{\beta}_1|x_1, \dots, x_n) = \sqrt{Var(\hat{\beta}_1|x_1, \dots, x_n)},$$

then

$$\widehat{sd}(\hat{\beta}_1|x_1, \dots, x_n) = \frac{\hat{\sigma}}{\left(\sum (x_i - \bar{x})^2\right)^{1/2}}$$

is frequently called the **standard error of $\hat{\beta}_1$** and reported in the output of software packages.

Reading: Sections 2.4 and 2.5 in [Wooldridge \(2009\)](#) and Appendix [10.1](#) if needed.

3 Multiple Regression Analysis: Estimation

3.1 Motivation for Multiple Regression: The Trade Example Continued

- In Section 2.6 two simple linear regression models for explaining imports to Germany were estimated (and interpreted): a level-level model and a log-log model.
- It is hardly credible that imports to Germany only depend on the GDP of the exporting country. What about, for example, distance,

borders, and other factors causing trading costs?

- Such quantities have been found to be relevant in the empirical literature on **gravity equations** for explaining intra- and international trade. In general, bi-directional trade flows are considered. Here we consider only one-directional trade flows, namely exports to Germany in 2004. Such a simplified gravity equation reads as

$$\ln(\text{imports}_i) = \beta_0 + \beta_1 \ln(\text{gdp}_i) + \beta_2 \ln(\text{distance}_i) + u_i. \quad (3.1)$$

Standard gravity equations are based on bilateral imports and exports over a number of years and thus require panel data techniques that are treated in the BA module **Advanced Issues in Econometrics**.

- For a brief introduction to gravity equations see e.g. [Fratianne \(2007\)](#). A recent theoretic underpinning of gravity equations was provided by [Anderson and Wincoop \(2003\)](#).
- If relevant variables are neglected, Assumptions SLR.1 and/or SLR.4 could be violated and in this case interpretation of causal effects can be highly misleading, see Section [3.4](#). To avoid this trap, the **multiple regression model** can be useful.
- To get an idea about the change in the elasticity parameter due to a second independent variable, like e.g. distance, inspect the following OLS estimate of the simple import equation ([3.1](#)):

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

Multiple R-squared: 0.8838, Adjusted R-squared: 0.8787

F-statistic: 174.9 on 2 and 46 DF, p-value: < 2.2e-16

Instead of an estimated elasticity of 1.077, see Section 2.6, one obtains a value of 0.976. Furthermore, the R^2 increases from 0.76 to 0.88, indicating a much better statistical fit. Finally, a 1% increase in distance reduces imports by 1.074%. Is this model then better? Or is it (also) misspecified?

To answer these questions we have to study the linear multiple regression model first.

3.2 The Multiple Regression Model of the Population

- **Assumptions:**

The Assumptions SLR.1 and SLR.4 of the simple linear regression model have to be adapted accordingly to the multiple linear regression model (MLR) for the population (see Section 3.3 in [Wooldridge \(2009\)](#)):

- **MLR.1 (Linearity in the Parameters)**

The **multiple regression model** allows for more than one, say k , explanatory variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \quad (3.2)$$

and the model is linear in its parameters.

Example: the import equation (3.1).

– MLR.4 (Zero Conditional Mean)

$$E[u|x_1, \dots, x_k] = 0 \quad \text{for all } x.$$

Observe that all explanatory variables of the multiple regression (3.2) must be included in the conditioning set. Sometimes the conditioning set is called **information set**.

• Remarks:

– To see the need for MLR.4, take the conditional expectation of y in (3.2) given all k regressors

$$E[y|x_1, x_2, \dots, x_k] = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + E[u|x_1, x_2, \dots, x_k].$$

If $E[u|x_1, x_2, \dots, x_k] \neq 0$ for some x , then the systematic part $\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ does not model the conditional expectations $E[y|x_1, \dots, x_k]$ correctly.

– If MLR.1 and MLR.4 are fulfilled, then equation (3.2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

is also called the **linear multiple regression model for the population**. Frequently it is also called the **true model** (even if any model may be far from truth). Alternatively, one may think of equation (3.2) as the **data generating mechanism** (although, strictly speaking, a data generating mechanism also requires specification of the probability distributions of all regressors and the error).

- To guarantee nice properties of the OLS estimator and the sample regression model, we adapt SLR.2 and SLR.3 accordingly:

- **MLR.2 (Random Sampling)**

The sample of size n is obtained by random sampling, that is the observations $\{(x_{i1}, \dots, x_{ik}, y_i) : i = 1, \dots, n\}$ are pairwise independently and identically distributed.

- **MLR.3 (No Perfect Collinearity)**

(more on MLR.3 in Section 3.3)

- **Interpretation:**

- If Assumptions MLR.1 and MLR.4 are correct *and* the population regression model allows for a causal interpretation, then the multiple regression model is a great tool for **ceteris paribus analysis**. It allows to hold the values of all explanatory variables fixed *except one* and check how the conditional expectation of the explained variable changes. This resembles changing one control variable in a randomized control experiment. Let x_j be the

control variable of interest.

- Taking conditional expectations of the multiple regression (3.2) and applying Assumption MLR.4 delivers

$$E[y|x_1, \dots, x_j, \dots, x_k] = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_k x_k.$$

- Consider a change in x_j : $x_j + \Delta x_j$

$$E[y|x_1, \dots, x_j + \Delta x_j, \dots, x_k] = \beta_0 + \beta_1 x_1 + \dots + \beta_j (x_j + \Delta x_j) + \dots + \beta_k x_k.$$

* **Ceteris-paribus effect:**

In (3.2) the absolute change due to a change of x_j by Δx_j is given by

$$\begin{aligned} \Delta E[y|x_1, \dots, x_j, \dots, x_k] &\equiv \\ E[y|x_1, \dots, x_{j-1}, x_j + \Delta x_j, x_{j+1}, \dots, x_k] & \\ - E[y|x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_k] &= \beta_j \Delta x_j, \end{aligned}$$

where β_j corresponds to the first partial derivative

$$\frac{\partial E[y|x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_k]}{\partial x_j} = \beta_j.$$

The parameter β_j gives the partial effect of changing x_j on the conditional expectation of y while all other regressors are held constant.

* **Total effect:**

Of course one can also consider simultaneous changes in the regressors, for example Δx_1 and Δx_k . For this case one obtains

$$\Delta E[y|x_1, \dots, x_k] = \beta_1 \Delta x_1 + \beta_k \Delta x_k.$$

- Note that the **specific interpretation of β_j** depends on how variables enter, e.g. as log variables. In a ceteris paribus analysis the results of Section 2.6 remain valid.

Trade Example Continued

- Considering the log-log model (3.1)

$$\ln(\text{imports}_i) = \beta_0 + \beta_1 \ln(\text{gdp}_i) + \beta_2 \ln(\text{distance}_i) + u_i$$

- a 1% increase in distance leads to a increase of $\beta_2\%$ in imports **keeping GDP fixed**. In other words, one can separate the effect of distance on imports from the effect of economic size. From the output table in Section 3.1 one obtains that a 1% increase in distance decreases imports by about 1.074%.
- Keep in mind that determining distances between countries is a complicated matter and results may change with the choice of the method for computing distances. Our data are from **CEPII**, see also Appendix 10.4.
 - There may still be missing variables, see also Section 4.4.

Wage Example Continued

- In Section 2.3 it was assumed that hourly wage is determined by

$$wage = \beta_0 + \beta_1 educ + u.$$

Instead of a level-level model one may also consider a log-level model

$$\ln(wage) = \beta_0 + \beta_1 educ + u. \quad (3.3)$$

- However, since we expect that experience also matters for hourly wages, we want to include *experience* as well. We obtain

$$\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + v. \quad (3.4)$$

What about the expected log wage given the variables *educ* and *exper*?

$$E[\ln(\text{wage}) | \text{educ}, \text{exper}] = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + E[v | \text{educ}, \text{exper}]$$

$$E[\ln(\text{wage}) | \text{educ}, \text{exper}] = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper},$$

where the second equation only holds if MLR.4 holds, that is if

$$E[v | \text{educ}, \text{exper}] = 0.$$

- Note that if instead of (3.4) one investigates the simple linear log-level model (3.3) **although the population model contains *exper*** one obtains

$$E[\ln(wage)|educ] = \beta_0 + \beta_1 educ + \beta_2 E[exper|educ] + E[v|educ]$$

indicating misspecification of the simple model since it ignores the influence of *exper* via β_2 . Thus, the smaller model suffers from misspecification if

$$E[\ln(wage)|educ] \neq E[\ln(wage)|educ, exper]$$

for some values of *educ* or *exper*.

- **Empirical results:**

See Example 2.10 in [Wooldridge \(2009\)](#), file: wage1.txt, output from R:

- **Simple log-level model**

Call:

```
lm(formula = log(wage) ~ educ)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.21158	-0.36393	-0.07263	0.29712	1.52339

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.583773	0.097336	5.998	3.74e-09 ***
educ	0.082744	0.007567	10.935	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4801 on 524 degrees of freedom

Multiple R-squared: 0.1858, Adjusted R-squared: 0.1843

F-statistic: 119.6 on 1 and 524 DF, p-value: < 2.2e-16

$$\ln(wage_i) = 0.5838 + 0.0827 educ_i + \hat{u}_i, \quad i = 1, \dots, 526,$$
$$R^2 = 0.1858.$$

If SLR.1 to SLR.4 are valid, then each additional year of schooling is estimated to increase hourly wages by 8.3% on average. The sample regression model explains about 18.6% of the variation of the dependent variable $\ln(wage)$.

— Multivariate log-level model:

Call:

```
lm(formula = log(wage) ~ educ + exper)
```

Residuals:

```
Min      1Q   Median      3Q      Max
-2.05800 -0.30136 -0.04539  0.30601  1.44425
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.216854   0.108595   1.997   0.0464 *
educ         0.097936   0.007622  12.848 < 2e-16 ***
exper        0.010347   0.001555   6.653 7.24e-11 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4614 on 523 degrees of freedom

Multiple R-squared: 0.2493, Adjusted R-squared: 0.2465

F-statistic: 86.86 on 2 and 523 DF, p-value: < 2.2e-16

$$\ln(wage_i) = 0.2169 + 0.0979 educ_i + 0.0103 exper_i + \hat{u}_i,$$

$$i = 1, \dots, 526,$$

$$R^2 = 0.2493.$$

- * **Ceteris-paribus interpretation:** If MLR.1 to MLR.4 are correct, then the expected increase in hourly wages due to an additional year of schooling is about 9.8% and thus slightly larger than obtained from the simple regression model.

An additional year of experience corresponds to an increase in expected hourly wages by 1%.

- * **Model fit:**

The model explains 24.9% of the variation of the independent variable. Does this imply that the multivariate model is better than the simple regression model with an R^2 of 18.6%? Be careful with your answer and wait until we investigate model selection criteria.

3.3 The OLS Estimator: Derivation and Algebraic Properties

- For an arbitrary estimator the **sample regression model** for a sample $(y_i, x_{i1}, \dots, x_{ik})$, $i = 1, \dots, n$, is given by

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \dots + \tilde{\beta}_k x_{ik} + \tilde{u}_i, \quad i = 1, \dots, n.$$

- Recall the idea of the OLS estimator: Choose $\tilde{\beta}_0, \dots, \tilde{\beta}_k$ such that the **sum of squared residuals (SSR)**

$$\text{SSR}(\tilde{\beta}_0, \dots, \tilde{\beta}_k) = \sum_{i=1}^n \tilde{u}_i^2 = \sum_{i=1}^n \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \dots - \tilde{\beta}_k x_{ik} \right)^2$$

is minimized. Taking first partial derivatives of $\text{SSR}(\tilde{\beta}_0, \dots, \tilde{\beta}_k)$ with respect to all $k + 1$ parameters and setting them to zero yields

the first order conditions of a minimum:

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \quad (3.5a)$$

$$\sum_{i=1}^n x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \quad (3.5b)$$

$$\sum_{i=1}^n x_{ik} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \quad (3.5c)$$

This system of **normal equations** contains $k + 1$ unknown parameters and $k + 1$ equations. Under some further conditions (see below) it has a unique solution.

Solving this set of equations becomes cumbersome if k is large. This can be circumvented if the normal equations are written in matrix notation.

• The Multiple Regression Model in Matrix Form

Using matrix notation the multiple regression model can be rewritten as (Wooldridge, 2009, Appendix E)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (3.6)$$

where

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1k} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}}_{\mathbf{u}}.$$

The matrix \mathbf{X} is called the regressor matrix and has n rows and $k + 1$ columns. The column vectors \mathbf{y} and \mathbf{u} have n rows each, the column vector $\boldsymbol{\beta}$ has $k + 1$ rows.

• Derivation: The OLS Estimator in Matrix Notation

- One possibility to derive the OLS estimator in matrix notation is to rewrite the normal equations (3.5) in matrix notation. We do this explicitly for the j -th equation

$$\sum_{i=1}^n x_{ij} \left(y_i - \hat{\beta}_0 x_{i0} - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0$$

that is manipulated to

$$\sum_{i=1}^n \left(x_{ij} y_i - \hat{\beta}_0 x_{ij} x_{i0} - \hat{\beta}_1 x_{ij} x_{i1} - \cdots - \hat{\beta}_k x_{ij} x_{ik} \right) = 0$$

and further to

$$\sum_{i=1}^n \left(\hat{\beta}_0 x_{ij} x_{i0} + \hat{\beta}_1 x_{ij} x_{i1} + \cdots + \hat{\beta}_k x_{ij} x_{ik} \right) = \sum_{i=1}^n x_{ij} y_i.$$

By factoring out we have

$$\left(\sum_{i=1}^n x_{ij} x_{i0} \right) \hat{\beta}_0 + \left(\sum_{i=1}^n x_{ij} x_{i1} \right) \hat{\beta}_1 + \cdots + \left(\sum_{i=1}^n x_{ij} x_{ik} \right) \hat{\beta}_k = \sum_{i=1}^n x_{ij} y_i.$$

Similarly, rearranging all other equations and collecting all $k + 1$ equations in a vector delivers

$$\begin{pmatrix} \left(\sum_{i=1}^n x_{i0} x_{i0} \right) \hat{\beta}_0 + \left(\sum_{i=1}^n x_{i0} x_{i1} \right) \hat{\beta}_1 + \cdots + \left(\sum_{i=1}^n x_{i0} x_{ik} \right) \hat{\beta}_k \\ \vdots \\ \left(\sum_{i=1}^n x_{ik} x_{i0} \right) \hat{\beta}_0 + \left(\sum_{i=1}^n x_{ik} x_{i1} \right) \hat{\beta}_1 + \cdots + \left(\sum_{i=1}^n x_{ik} x_{ik} \right) \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{i0} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} y_i \end{pmatrix}.$$

Applying the rules for matrix multiplication yields

$$\underbrace{\begin{pmatrix} (\sum_{i=1}^n x_{i0}x_{i0}) & (\sum_{i=1}^n x_{i0}x_{i1}) & \cdots & (\sum_{i=1}^n x_{i0}x_{ik}) \\ \vdots & \vdots & \ddots & \vdots \\ (\sum_{i=1}^n x_{ik}x_{i0}) & (\sum_{i=1}^n x_{ik}x_{i1}) & \cdots & (\sum_{i=1}^n x_{ik}x_{ik}) \end{pmatrix}}_{\mathbf{X}'\mathbf{X}} \underbrace{\begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}}_{\hat{\boldsymbol{\beta}}} = \underbrace{\begin{pmatrix} \sum_{i=1}^n x_{i0}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{pmatrix}}_{\mathbf{X}'\mathbf{y}}$$

as well as the **normal equations in matrix notation**

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \quad (3.7)$$

– Note: The matrix $\mathbf{X}'\mathbf{X}$ has $k + 1$ columns and rows so that it is a square matrix.

The inverse $(\mathbf{X}'\mathbf{X})^{-1}$ **exists if** all columns (and rows) are linearly independent. This can be shown to be the case if all columns of \mathbf{X} are linearly independent.

This is exactly what the next assumption states.

Assumption MLR.3 (No Perfect Collinearity):

In the sample none of the regressors can be expressed as an exact linear combination of one or more of the other regressors.

Is this a restrictive assumption?

- Finally, multiply the normal equation (3.7) by $(\mathbf{X}'\mathbf{X})^{-1}$ from the left and obtain the **OLS estimator in matrix notation**:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3.8)$$

This is the compact notation for

$$\underbrace{\begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}}_{\hat{\boldsymbol{\beta}}} = \underbrace{\begin{pmatrix} (\sum_{i=1}^n x_{i0}x_{i0}) & (\sum_{i=1}^n x_{i0}x_{i1}) & \cdots & (\sum_{i=1}^n x_{i0}x_{ik}) \\ \vdots & \vdots & \ddots & \vdots \\ (\sum_{i=1}^n x_{ik}x_{i0}) & (\sum_{i=1}^n x_{ik}x_{i1}) & \cdots & (\sum_{i=1}^n x_{ik}x_{ik}) \end{pmatrix}}_{(\mathbf{X}'\mathbf{X})^{-1}}^{-1} \cdot \underbrace{\begin{pmatrix} \sum_{i=1}^n x_{i0}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{pmatrix}}_{\mathbf{X}'\mathbf{y}}.$$

Algebraic Properties of the OLS Estimator

- $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$, that is $\sum_{i=1}^n x_{ij}\hat{u}_i = 0$ for $j = 0, \dots, k$.
Proof: Plugging $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$ into the normal equation yields $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} + \mathbf{X}'\hat{\mathbf{u}}$ and hence $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.
- If $x_{i0} = 1$, $i = 1, \dots, n$, it follows that $\sum_{i=1}^n \hat{u}_i = 0$.
- For the special case $k = 1$, the algebraic properties of the simple linear regression model follow immediately.
- The point $(\bar{y}, \bar{x}_1, \dots, \bar{x}_k)$ is always located on the regression hyperplane if there is a constant in the model.
- The definitions for SST, SSE and SSR are like in the simple regression.
- If a constant term is included in the model, we can decompose

$$\text{SST} = \text{SSE} + \text{SSR}.$$

- The **Coefficient of Determination**:

R^2 is defined as in the SLR case as

$$R^2 = \frac{\text{SSE}}{\text{SST}}$$

or, if there is an intercept in the model,

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}.$$

It can be shown that the R^2 is the squared empirical coefficient of correlation between the observed y_i 's and the explained \hat{y}_i 's, namely

$$\begin{aligned} R^2 &= \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^n (y_i - \bar{y})^2\right) \left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2\right)} \\ &= \left[\widehat{\text{Corr}}(y, \hat{y})\right]^2. \end{aligned}$$

Note that $\left[\widehat{Corr}(y, \hat{y})\right]^2$ can be used even when R^2 is not useful. In this case this expression is called pseudo R^2 .

- **Adjusted R^2 :**

If we rewrite R^2 by expanding the SSR/SST term by n

$$R^2 = 1 - \frac{SSR/n}{SST/n},$$

we can interpret SSR/n and SST/n as estimators for σ^2 and σ_y^2 respectively. They are biased estimators, however.

Using unbiased estimators thereof instead one obtains the “adjusted” R^2

$$\bar{R}^2 = 1 - \frac{SSR/(n - k - 1)}{SST/(n - 1)}.$$

Alternative representations:

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{n-1}{n-k-1} \cdot \frac{\text{SSR}}{\text{SST}} \\ \bar{R}^2 &= 1 - \frac{n-1}{n-k-1} (1 - R^2) \\ &= \frac{-k}{n-k-1} + \frac{n-1}{n-k-1} \cdot R^2\end{aligned}$$

Properties of \bar{R}^2 (see Section 6.3 in [Wooldridge \(2009\)](#)):

- \bar{R}^2 can increase *or* fall when including an additional regressor.
- \bar{R}^2 always increases if an additional regressor reduces the unbiased estimate of the error variance.

Attention: Analogously to R^2 one may not compare \bar{R}^2 of regression models with different y , for example if in one model the regressand is y and in the other one $\ln(y)$.

- The quantities R^2 , \bar{R}^2 , or $\left[\widehat{Corr}(y, \hat{y})\right]^2$ are called **goodness-of-fit measures**.

3.4 The OLS Estimator: Statistical Properties

Assumptions (Recap):

- **MLR.1 (Linearity in the Parameters)**
- **MLR.2 (Random Sampling)**
- **MLR.3 (No Perfect Collinearity)**
- **MLR.4 (Zero Conditional Mean)**

3.4.1 The Unbiasedness of Parameter Estimates

- Let MLR.1 through MLR.4 hold. Then we have $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

Proof:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{MLR.3}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \quad \text{MLR.1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

Taking conditional expectation one obtains

$$E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}]$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}]$$

$$= \boldsymbol{\beta}.$$

MLR.2 and MLR.4

The last equality holds because

$$E[\mathbf{u}|\mathbf{X}] = \begin{pmatrix} E[u_1|\mathbf{X}] \\ E[u_2|\mathbf{X}] \\ \vdots \\ E[u_n|\mathbf{X}] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where the latter follows from

$$\begin{aligned} E[u_i|\mathbf{X}] &= E[u_i|x_{11}, \dots, x_{1k}, \dots, x_{nk}] \\ &= E[u_i|x_{i1}, \dots, x_{ik}] && \text{MLR.2} \\ &= 0 && \text{MLR.4} \end{aligned}$$

for $i = 1, \dots, n$.

• The Danger of Omitted Variable Bias

We partition the $k + 1$ regressors in a $(n \times k)$ matrix \mathbf{X}_A and a $(n \times 1)$ vector \mathbf{x}_a . This yields

$$\mathbf{y} = \mathbf{X}_A \boldsymbol{\beta}_A + \mathbf{x}_a \beta_a + \mathbf{u}. \quad (3.9)$$

In the following it is assumed that the population regression model has the same structure as (3.9).

Trade Example Continued (from Section 3.2):

Assume that in the population imports depend on gdp, distance, and whether the trading countries share to some extent the same language

$$\begin{aligned} \ln(\text{imports}_i) = & \beta_0 + \beta_1 \ln(\text{gdp}_i) + \beta_2 \ln(\text{distance}_i) \\ & + \beta_3 \ln(\text{openess}_i) + u_i. \end{aligned} \quad (3.10)$$

so that \mathbf{X}_A includes the constant, gdp_i , and distance_i and \mathbf{x}_a

denotes the vector for $openess_i$, each $i = 1, \dots, n$.

Imagine now that you are only interested in the values of β_A (the parameters for the constant, gdp , and $distance$), and that the regressor vector \mathbf{x}_a has to be omitted because, for instance, obtaining data requires too much effort.

Which **effect** has the **omission of the variable** x_a **on the estimation of** $\hat{\beta}_A$ if, for example, the model

$$\mathbf{y} = \mathbf{X}_A \beta_A + \mathbf{w} \quad (3.11)$$

is considered? Model (3.11) is frequently called the smaller model.

Or, stated differently, which estimation properties does the OLS estimator for β_A have on basis of the smaller model (3.11)?

Derivation:

- Denote the OLS estimator for β_A from the small model by $\tilde{\beta}_A$. Following the proof of unbiasedness for the small model but replacing \mathbf{y} with the true population model (3.9) delivers

$$\begin{aligned}\tilde{\beta}_A &= (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{y} \\ &= (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A (\mathbf{X}_A \beta_A + \mathbf{x}_a \beta_a + \mathbf{u}) \\ &= \beta_A + (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{x}_a \beta_a + (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{u}.\end{aligned}$$

- By the law of iterated expectations $E[\mathbf{u}|\mathbf{X}_A] = E[E[\mathbf{u}|\mathbf{X}_A, \mathbf{x}_a]|\mathbf{X}_A]$ and therefore $E[\mathbf{u}|\mathbf{X}_A] = E[0|\mathbf{X}_A] = 0$ by validity of MLR.4 for the population model (3.9).

- Compute the conditional expectation of β_A . Treating the (unobserved) \mathbf{x}_a in the same way as \mathbf{X}_A one obtains

$$E \left[\tilde{\beta}_A | \mathbf{X}_A, \mathbf{x}_a \right] = \beta_A + (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{x}_a \beta_a.$$

Therefore the **estimator** $\tilde{\beta}_A$ is **unbiased only if**

$$(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{x}_a \beta_a = \mathbf{0}. \quad (3.12)$$

Take a closer look at the term on the left hand side of (3.12), i.e.

$$(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{x}_a \beta_a.$$

One observes that

$$\tilde{\delta} = (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{x}_a$$

is the OLS estimator of δ in a regression of \mathbf{x}_a on \mathbf{X}_A

$$\mathbf{x}_a = \mathbf{X}_A \delta + \varepsilon.$$

Condition (3.12) holds (and there is no bias) if

- * $\tilde{\boldsymbol{\delta}} = \mathbf{0}$, so \mathbf{x}_a is uncorrelated with \mathbf{X}_A in the sample or
- * $\beta_a = 0$ holds and the smaller model is the population model.

If neither of these conditions holds, then $\tilde{\boldsymbol{\beta}}_A$ is biased

$$E[\tilde{\boldsymbol{\beta}}_A | \mathbf{X}_A, \mathbf{x}_a] = \boldsymbol{\beta}_A + \tilde{\boldsymbol{\delta}}\beta_a.$$

This means that the OLS estimator $\tilde{\boldsymbol{\beta}}_A$ is in general biased for **every** parameter in the smaller model.

Since these biases are caused by using a regression model in which a variable is omitted that is relevant in the population model, this kind of bias is called **omitted variable bias** and the smaller model is said to be **misspecified**. (See Appendix 3A.4 in [Wooldridge \(2009\)](#).)

- One may also ask about the unconditional bias. Applying the LIE delivers

$$E \left[\tilde{\boldsymbol{\beta}}_A | \mathbf{X}_A \right] = \boldsymbol{\beta}_A + E \left[\tilde{\boldsymbol{\delta}} | \mathbf{X}_A \right] \beta_a,$$
$$E \left[\tilde{\boldsymbol{\beta}}_A \right] = \boldsymbol{\beta}_A + E \left[\tilde{\boldsymbol{\delta}} \right] \beta_a.$$

Interpretation: The second expression delivers the expected value of the OLS estimator if one keeps drawing new samples for \mathbf{y} **and** \mathbf{X}_A . Thus, in repeated sampling there is only bias if there is correlation in the population between the variables in \mathbf{X}_A and \mathbf{x}_a since otherwise $E \left[\tilde{\boldsymbol{\delta}} \right] = \mathbf{0}$, cf. 2.4.

- **Wage Example Continued** (from Section 3.2):

- If the observed regressor $educ$ is correlated with the unobserved variable $ability$, then the regressor $x_a = ability$ is missing in the regression and the OLS estimators, e.g. for the effect of an additional year of schooling, are biased.
- **Interpretation** of the various information sets for computing the expectation of $\hat{\beta}_{educ}$:

- * First consider

$$E[\hat{\beta}_{educ} | educ, exper, ability] = \beta_{educ} + \tilde{\delta} \beta_{ability},$$

where

$$ability = \begin{pmatrix} 1 & educ & exper \end{pmatrix} \delta + \varepsilon.$$

Then the conditional expectation above indicates the average of $\hat{\beta}_{educ}$ computed over many different samples where each

sample of workers is drawn in the following way: You always guarantee that each sample has the same number of workers with e.g. 10 years of schooling, 15 years of experience, and 150 units of ability and the same number of workers with 11 years of schooling, etc., so that for each combination of characteristics there is the same amount of workers although the workers may not be (completely) identical.

* Next consider

$$E[\hat{\beta}_{educ} | educ, exper] = \beta_{educ} + E[\tilde{\delta} | educ, exper] \beta_{ability}.$$

When drawing a new sample you only guarantee that the number of workers with a specific number of years of schooling and experience stay the same. In contrast to above, you do not control ability.

* Finally consider

$$E[\hat{\beta}_{educ}] = \beta_{educ} + E[\tilde{\delta}] \beta_{ability}.$$

Here you simply draw new samples where everything is allowed to vary. If you had, let's say 50 workers with 10 years of schooling in one sample, you may have 73 workers with 10 years of schooling in another sample. This possibility is excluded in the two previous cases.

- **Effect of omitted variables on the conditional mean:**

- General **terminology**:

- * If $E[y|x_A, x_a] \neq E[y|x_A]$,

- then the smaller model omitting x_a is **misspecified** and estimation will suffer from omitted variable bias.

- * If $E[y|x_A, x_a] = E[y|x_A]$,

- then the variable x_a in the larger model is **redundant** and should be eliminated from the regression.

- * **Trade Example Continued:** Assume that the population regression model only contains the variables gdp and $distance$. Then a simple regression model with gdp is misspecified and a multiple regression model with gdp , $distance$, and $openess$ contains the redundant variable $openess$.

- **It can happen that for a misspecified model Assumptions MLR.1 to MLR.4 are fulfilled.**

To see this, consider only one variable in \mathbf{X}_A

$$E[y|x_A, x_a] = \beta_0 + \beta_A x_A + \beta_a x_a.$$

Then, by the law of iterated expectations one obtains

$$E[y|x_A] = \beta_0 + \beta_A x_A + \beta_a E[x_a|x_A].$$

If, in addition, $E[x_a|x_A]$ is linear in x_A

$$x_a = \alpha_0 + \alpha_1 x_A + \varepsilon, \quad E[\varepsilon|x_A] = 0,$$

one obtains

$$\begin{aligned} E[y|x_A] &= \beta_0 + \beta_A x_A + \beta_a(\alpha_0 + \alpha_1 x_A) \\ &= \gamma_0 + \gamma_1 x \end{aligned}$$

with $\gamma_0 = \beta_0 + \beta_a \alpha_0$ und $\gamma_1 = \beta_A + \beta_a \alpha_1$ being the parameters of the best linear predictor, see Section 2.4.

- Note that in this case SLR.1 and SLR.4 are fulfilled for the smaller model although it is not the population model. However

$$E[y|x_A, x_a] \neq E[y|x_A]$$

if $\beta_a \neq 0$ and $\alpha_1 \neq 0$.

- Thus, model choice matters, see Section 3.5. If controlling for x_a is important, then the smaller model is of not much use if the differences between the expected values are large for some values of the regressors.

If one needs a model for prediction, the smaller model may be preferable since it exhibits smaller estimation variance, see Sections 3.4.3 and 3.5.

Reading: Section 3.3 in [Wooldridge \(2009\)](#).

3.4.2 The Variance of Parameter Estimates

- **Assumption MLR.5 (Homoskedasticity):**

$$\text{Var}(u_i | x_{i1}, \dots, x_{ik}) = \sigma^2, \quad i = 1, \dots, n$$

- Assumptions MLR.1 bis MLR.5 are frequently called **Gauss-Markov-Assumptions**.
- Note that by the Random Sampling assumption MLR.2 one has

$$\begin{aligned} \text{Cov}(u_i, u_j | x_{i1}, \dots, x_{ik}, x_{j1}, \dots, x_{jk}) &= 0 \quad \text{for all } i \neq j, 1 \leq i, j \leq n, \\ \text{Cov}(u_i, u_j) &= 0 \quad \text{for all } i \neq j, 1 \leq i, j \leq n, \end{aligned}$$

where for the latter equations the LIE was used. Because of MLR.2 one may also write

$$\text{Var}(u_i | x_{i1}, \dots, x_{ik}) = \text{Var}(u_i | \mathbf{X}), \quad \text{Cov}(u_i, u_j | \mathbf{X}) = 0, \quad i \neq j.$$

One writes all n variances and all covariances in a matrix

$$\text{Var}(\mathbf{u}|\mathbf{X}) \equiv \begin{pmatrix} \text{Var}(u_1|\mathbf{X}) & \text{Cov}(u_1, u_2|\mathbf{X}) & \cdots & \text{Cov}(u_1, u_n|\mathbf{X}) \\ \text{Cov}(u_2, u_1|\mathbf{X}) & \text{Var}(u_2|\mathbf{X}) & \cdots & \text{Cov}(u_2, u_n|\mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(u_n, u_1|\mathbf{X}) & \text{Cov}(u_n, u_2|\mathbf{X}) & \cdots & \text{Var}(u_n|\mathbf{X}) \end{pmatrix} \quad (3.13)$$

$$= \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

or short (MLR.2 and MLR.5 together)

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}. \quad (3.14)$$

• Variance of the OLS Estimator

Under the Gauss-Markov Assumptions MLR.1 to MLR.5 we have

$$\text{Var}(\hat{\beta}_j | \mathbf{X}) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}, \quad x_j \text{ not constant}, \quad (3.15)$$

where SST_j is the **total sample variation** (total sum of squares) of the j -th regressor,

$$\text{SST}_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2,$$

and the coefficient of determination R_j^2 is taken from a regression of the j -th regressor on all other regressors

$$x_{ij} = \delta_0 x_{i0} + \cdots + \delta_{j-1} x_{i,j-1} + \delta_{j+1} x_{i,j+1} + \cdots + \delta_k x_{i,k} + v_i, \\ i = 1, \dots, n. \quad (3.16)$$

(See Appendix 3A.5 in [Wooldridge \(2009\)](#) for the proof.)

Interpretation of the variance of the OLS estimator:

- The larger the **error variance** σ^2 , the larger is the variance of $\hat{\beta}_j$.

Note: This is a property of the population so that this variance component cannot be influenced by sample size. (In analogy to the simple regression model.)

- The larger the **total sample variation** SST_j of the j -th **regressor** x_j is, the smaller is the variance $Var(\hat{\beta}_j|\mathbf{X})$.

Note: The total sample variation can always be increased by increasing sample size since adding another observation increases SST_j .

- If $SST_j = 0$, assumption MLR.3 fails to hold.

- The larger the coefficient of determination R_j^2 from regression (3.16) is, the larger is the variance of $\hat{\beta}_j$.
- The larger R_j^2 , the better the variation in x_j can be explained by variation in the other regressors because in this case there is a high degree of linear dependence between x_j and the other explanatory variables.

Then only a small part of the sample variation in x_j is specific for the j -th regressor (precisely the error variation in (3.16)). The other part of the variation can be explained equally well by the estimated linear combination of all other regressors. This effect is not well attributable by the estimator to either variable x_j or the linear combination of all the remaining variables and thus the estimator suffers from a larger estimation variance.

– **Special cases:**

- * $R_j^2 = 0$: Then x_j and all other explanatory variables are empirically uncorrelated and the parameter estimator $\hat{\beta}_j$ is unaffected by all other regressors.
- * $R_j^2 = 1$: Then MLR.3 fails to hold.
- * R_j^2 near 1: This situation is called **multi- order near collinearity**. In this case $Var(\hat{\beta}_j|\mathbf{X})$ is very large.

– **But:** The multicollinearity problem is reduced in larger samples because SST_j rises and hence variance decreases for a given value of R_j^2 . Therefore multicollinearity is always a problem of small sample sizes, too.

- **Estimation of the error variance σ^2**

- Unbiased **estimation of the error variance σ^2** :

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{n - (k + 1)}.$$

- **Properties** of the OLS estimator (continued):

Call $sd(\hat{\beta}_j | \mathbf{X}) = \sqrt{Var(\hat{\beta}_j | \mathbf{X})}$ the standard deviation, then

$$\hat{sd}(\hat{\beta}_j | \mathbf{X}) = \frac{\hat{\sigma}}{\left(\text{SST}_j (1 - R_j^2) \right)^{1/2}}$$

is the **standard error of $\hat{\beta}_j$** .

- **Variance-covariance-matrix** of the OLS estimator:

Basics: The covariance of jointly estimating β_j and β_l — between the estimators of the j -th and the l -th parameter — is written as

$$\text{Cov}(\hat{\beta}_j, \hat{\beta}_l | \mathbf{X}) = E[(\hat{\beta}_j - \beta_j)(\hat{\beta}_l - \beta_l) | \mathbf{X}], \quad j, l = 0, 1, \dots, k + 1,$$

where unbiasedness of the estimators is assumed. We can write a $((k + 1) \times (k + 1))$ -matrix that contains all variances and covariances (next slide):

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &\equiv \\
&= \begin{pmatrix} \text{Cov}(\hat{\beta}_0, \hat{\beta}_0|\mathbf{X}) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1|\mathbf{X}) & \cdots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_k|\mathbf{X}) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0|\mathbf{X}) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_1|\mathbf{X}) & \cdots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k|\mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_k, \hat{\beta}_0|\mathbf{X}) & \text{Cov}(\hat{\beta}_k, \hat{\beta}_1|\mathbf{X}) & \cdots & \text{Cov}(\hat{\beta}_k, \hat{\beta}_k|\mathbf{X}) \end{pmatrix} \\
&= \begin{pmatrix} E[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_0 - \beta_0)|\mathbf{X}] & \cdots & E[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_k - \beta_k)|\mathbf{X}] \\ E[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0)|\mathbf{X}] & \cdots & E[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_k - \beta_k)|\mathbf{X}] \\ \vdots & \ddots & \vdots \\ E[(\hat{\beta}_k - \beta_k)(\hat{\beta}_0 - \beta_0)|\mathbf{X}] & \cdots & E[(\hat{\beta}_k - \beta_k)(\hat{\beta}_k - \beta_k)|\mathbf{X}] \end{pmatrix} \\
&= E \left[\begin{array}{c} \begin{pmatrix} (\hat{\beta}_0 - \beta_0) \\ \cdots \\ (\hat{\beta}_k - \beta_k) \end{pmatrix} \begin{pmatrix} (\hat{\beta}_0 - \beta_0) & \cdots & (\hat{\beta}_k - \beta_k) \end{pmatrix} \\ \hline \mathbf{X} \end{array} \right] \\
&= E \left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X} \right].
\end{aligned}$$

Next it will be shown that it holds:

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = E \left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X} \right] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Proof:

Remember that correct model specification implies

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},$$

hence $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$. This can be inserted into $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ and obtain

$$\begin{aligned}
E \left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' | \mathbf{X} \right] &= E \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{u} \left((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{u} \right)' | \mathbf{X} \right] \\
&= E \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{u} \mathbf{u}' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X} \right] \\
&= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \underbrace{E[\mathbf{u} \mathbf{u}' | \mathbf{X}]}_{\sigma^2 \mathbf{I}_n} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

From the definition of $Var(\hat{\boldsymbol{\beta}} | \mathbf{X})$ above it can be seen that the diagonal elements are the variances $Var(\hat{\beta}_j | \mathbf{X})$, $j = 0, \dots, k$.

• Efficiency of OLS

Note: The OLS estimator is a linear estimator with respect to the dependent variable because it holds for given \mathbf{X} that

$$\hat{\beta}_j = \sum_{i=1}^n \left(\frac{\hat{v}_i}{\sum_{i=1}^n \hat{v}_i^2} \right) y_i,$$

where \hat{v}_i are the residuals from regression (3.16). Thus, the estimator is a weighted sum of the regressand. The linearity of the *estimator* should not be confused with the linearity of the *parameters in the model*. (For a derivation without matrix algebra see Appendix 3A.2 in [Wooldridge \(2009\)](#).)

Further, OLS is unbiased so that $E[\hat{\beta}_j] = \beta_j$.

Gauss-Markov Theorem: Under assumptions MLR.1 through MLR.5 the OLS estimator is the **best linear unbiased estimator (BLUE)**.

“Best” means that the OLS estimator, that is unbiased since $E[\hat{\beta}_j] = \beta_j$, has minimal variance among linear unbiased estimators.

3.4.3 Trade-off between Bias and Multicollinearity

- **Example:** Let the population model be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

- For a given sample let R_1^2 be close to 1. Then β_1 is estimated with a large variance by (3.15).
- A possible solution? Leaving out the regressor x_2 and estimation of the simple regression. But then, as already shown, the estimator of β_1 is biased.

Hence: If there is correlation between x_1 and x_2 near 1 or -1, then — **for given sample size** — one faces a **trade-off between variance and bias**.

- What we observe is kind of a **statistical uncertainty relation**: The sample does not provide sufficient information to precisely

answer the formulated question.

- **The only good solution:** Increasing sample size.
- **Alternative solution:** Combining highly correlated variables.

- **Variance of parameter estimates in misspecified models:**

Again, there are different possibilities how incorrect regression models might be chosen (cf. Section 3.4.1):

- Too many variables: Parameters are estimated for variables that do not play a role in the “true” data generation mechanism (**redundant variables**).
- Too few variables: One or more variables are missing which are relevant in the population regression model (**omitted variables**).

– Wrong variables: A combination of both.

Effect on the variance of parameter estimators:

– **Case 1 (redundant variables):**

Consider the population model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. Assume that instead the following sample specification is chosen:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\boldsymbol{\alpha} + \mathbf{w},$$

where the vector \mathbf{z} contains all sample observations of the variable z . The variance of the parameter estimator $\hat{\beta}_j$ is

$$\text{Var}(\hat{\beta}_j | \mathbf{X}) = \frac{\sigma^2}{\text{SST}_j(1 - R_{j, \mathbf{X}, \mathbf{z}}^2)},$$

where now $R_{j, \mathbf{X}, \mathbf{z}}^2$ is the coefficient of determination of a regression of x_j on all other variables in \mathbf{X} and on \mathbf{z} . It is easily seen that $R_{j, \mathbf{X}, \mathbf{z}}^2 \geq R_j^2$ because fewer variables are included in the

regression yielding the second R^2 .

Therefore: **Including additional variables in a regression model increases estimation variance or leaves it unchanged.**

– **Case 2 (omitted variables):**

The converse of case 1 holds: **If a variable is omitted, it can be shown that the estimation variance is smaller than when using the true model.**

– **Case 3 (redundant and omitted variables):**

Should really be avoided.

Correct model specification is crucial!

3.5 Model Specification I: Model Selection Criteria

- **Goal of model selection:**

- In principle: find the population model.
- In practice: find the “best” model for the purpose of the analysis.
- More specific: Under the assumption that the population model is a multiple linear regression model find all regressors that are included in the regression and their appropriate transformations (log or level or ...). Avoid omitting variables and including irrelevant variables.

- **Brief theory of model selection:**

- There are two issues:

- a) the **variable (model) choice**,

- b) the **estimation variance**.

- Consider a): Choose a goal function to evaluate different models.

- A popular goal function is the **mean squared error (MSE)**.

- For fixed parameters it is defined as

$$MSE = E \left[(y - \beta_0 x_0 - \beta_1 x_1 - \cdots - \beta_k x_k)^2 \right], \quad (3.17)$$

- see also equation (2.17) for the simple regression case.

- Choose the model for which the MSE is minimal.

Important cases:

- * If x_0, \dots, x_k include **all relevant variables**, the population model is a multiple linear regression, and MSE is minimized with respect to the parameters, then

$$MSE = E \left[u^2 \right] = \sigma^2.$$

- * If **relevant variables are missing**, it can be shown that the MSE decomposes into variance and squared bias. For simplicity, omit all variables except x_1 and fit the simple linear regression

$$y = \gamma_0 + \gamma_1 x_1 + v.$$

Then

$$\begin{aligned} MSE_1 &= E \left[\left\{ (y - E[y|x_1, \dots, x_k]) + (E[y|x_1, \dots, x_k] - E[y|x_1]) \right\}^2 \right] \\ &= \sigma^2 + E \left[(E[y|x_1, \dots, x_k] - E[y|x_1])^2 \right]. \end{aligned}$$

First equation: the first term in parentheses represents the deviation of the observable y from its conditional expectation of the population model (“true” model) and thus u . The second term in parenthesis captures the deviation of the conditional expectation of the “true” model from the conditional expectation of the chosen misspecified model which is the bias of predicting y with a too small model conditional on x_1, \dots, x_k . The second equation can be derived by using the LIE. Since $E \left[(E[y|x_1, \dots, x_k] - E[y|x_1])^2 \right] > 0$ for any misspecified model (see slide/page 145), $MSE < MSE_1$ holds.

- Consider a) and b): If **parameters have to estimated**, a further term enters the mean squared error, namely the variances and covariances for estimating the model parameters. One has

$$\begin{aligned}MSE &= \text{Variance of population error} \\ &+ \text{Bias of chosen model}^2 \\ &+ \text{Estimation variance,}\end{aligned}$$

where the estimation variance in general increases with the number of variables. Now it can happen that for minimizing MSE it is optimal to choose a model that omits variable(s). A typical case is prediction.

- Therefore, reliable methods for estimating the MSE are needed.

- **What does not work:**

- **Selecting the model with the smallest standard error of the regression $\hat{\sigma}$ does not work.**

- * Why? It is always possible to select a model for which every residual is zero, that is $\hat{u}_i = 0$ for all $i = 1, \dots, n$. Then $\hat{\sigma} = 0$ as well although the error variance is $\sigma^2 > 0$ in the true model.
- * How? Simply take $k + 1 = n$ regressors into the sample regression model which fulfil MLR.3 and solve the normal equations (3.5). Then you obtain a perfect fit since you have a linear equation system with n equations and n unknown parameters.
- * Note that you can add any regressors that fulfil MLR.3 even if they have nothing to do with the population regression model.

- * Note also that SSR remains constant or decreases if for a given sample of size n a further regressor variable is added since the linear equation system obtains more flexibility to fit the sample observations. Therefore $\tilde{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n} = \frac{SSR}{n}$ remains constant or decreases as well.
- * For the variance estimator $\hat{\sigma}^2 = \frac{SSR}{n-k-1}$ there are opposing effects: a decrease in SSR maybe offset by the decrease in $n - k - 1$.

In sum, $\tilde{\sigma} = \sqrt{SSR/n}$ is not appropriate for selecting those variables that are part of the population model since $\tilde{\sigma}$ remains the same or decreases if additional regressors are included.

- Selecting the model with the largest R^2 **does not work** either.

Why?

- Although the adjusted R^2 may fall or increase with adding another regressor, it screws up for $k + 1 = n$ since $\bar{R}^2 = 1$ as well in this case.

- **Solution: Use model selection criteria**

- **Basic idea:**

$$\textit{Selection criterion} = \ln \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{n} + (k + 1) \cdot \textit{penalty function}(n)$$

- * **First term:** $\ln \tilde{\sigma}^2$ is based on the **variance estimator** $\tilde{\sigma}^2$ of the chosen model.

Recall that the estimated variance $\tilde{\sigma}^2 = \hat{\mathbf{u}}' \hat{\mathbf{u}} / n$ is reduced or remains constant by every additionally included independent variable.

- * **Second term:** is a **penalty term** punishing the number of parameters to avoid models that include redundant variables.

Because the true error variance is typically underestimated using $\tilde{\sigma}^2$, the penalty term penalizes the inclusion of additional regressors.

The penalty term increases with k and the penalty function must be chosen such that it decreases with n such that a large number of parameters matters less in large samples. Why?

- * This implies a **trade-off**: Regressors are included in the model, if the penalty is smaller than the decrease in the estimated MSE.

By choosing the penalty term (and thus the criterion) one determines how the trade-off is shaped.

- * **Rule**: Choose among all considered candidate models the specification for which the criterion is *minimal*.

– **Popular model selection criteria:**

* the **Akaike Criterion (AIC)**

$$AIC = \ln \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{n} + (k + 1) \frac{2}{n}, \quad (3.18)$$

* the **Hannan-Quinn Criterion (HQ)**

$$HQ = \ln \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{n} + (k + 1) \frac{2 \ln(\ln n)}{n}, \quad (3.19)$$

* the **Schwarz / Bayesian Information Criterion (SC/BIC)**

$$SC = \ln \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{n} + (k + 1) \frac{\ln n}{n}. \quad (3.20)$$

It is advised always to check all criteria although the researcher decides which to use. In nice cases, all criteria deliver the same result. Note that for standard sample sizes SC punishes additional parameters more than HQ, and HQ more than AIC

• Trade Example Continued:

– Modell 1

$$\text{LOG}(\text{TRADE_0_D_0}) = -5.770261 + 1.077624 * \text{LOG}(\text{WDI_GDPUSDCR_0})$$

$$\text{AIC} = 3.410063, \quad \text{HQ} = 3.439359, \quad \text{SC} = 3.487280$$

– Modell 2

$$\text{LOG}(\text{TRADE_0_D_0}) = 4.676112 + 0.975983 * \text{LOG}(\text{WDI_GDPUSDCR_0}) - 1.074076 * \text{LOG}(\text{CEPII_DIST})$$

$$\text{AIC} = 2.748467, \quad \text{HQ} = 2.792411, \quad \text{SC} = 2.864293$$

– Modell 3

$$\text{LOG}(\text{TRADE_0_D_0}) = 2.741040 + 0.940664 * \text{LOG}(\text{WDI_GDPUSDCR_0}) - 0.970318 * \text{LOG}(\text{CEPII_DIST}) \\ + 0.507250 * \text{EBRD_TFES_0}$$

$$\text{AIC} = 2.644544, \quad \text{HQ} = 2.703136, \quad \text{SC} = 2.798979$$

– Modell 4

$$\text{LOG}(\text{TRADE_0_D_0}) = 2.427777 + 1.025023 * \text{LOG}(\text{WDI_GDPUSDCR_0}) - 0.888646 * \text{LOG}(\text{CEPII_DIST}) \\ + 0.353154 * \text{EBRD_TFES_0} - 0.151031 * \text{LOG}(\text{CEPII_AREA_0})$$

$$\text{AIC} = 2.616427, \quad \text{HQ} = 2.689667, \quad \text{SC} = 2.809470$$

– Comparing all four models, SC selects model 3 with regressors *gdp*, *distance* and *openess* while AIC selects model 4 with additional regressor *area*. See Appendix 10.4 for more details on variables. One can nicely see that SC punishes additional variables more than AIC. Statistical tests may provide further information on which model to choose, see Sections 4.3 onwards.

4 Multiple Regression Analysis: Hypothesis Testing and Confidence Intervals

4.1 Basics of Statistical Tests

Foundations of statistical hypothesis testing

- In general: Statistical hypothesis tests allow statistically sound and unambiguous answers to yes-or-no questions:
 - Do men and women earn equal income in Germany?

- Do certain political attempts lead to a decrease in unemployment in 2020?
- Are imports to Germany influenced by the gdp of exporting countries?

- **Elements of a statistical test:**

1. Two disjoint **hypotheses** about one or more value(s) of (a) parameter(s) θ in a population.

That means that one of the two competing hypotheses has to hold in the population:

- **Null hypothesis** H_0

- **Alternative hypothesis** H_1

Were θ known, one immediately can decide whether H_0 holds.

2. A **test statistic** t that is a function of the sample values (\mathbf{X}, \mathbf{y}) .
Prior to observing a sample a test statistic is a random variable, after observing a sample a realization of it. We will denote both as $t(\mathbf{X}, \mathbf{y})$.

3. A **decision rule**, stating for which values of $t(\mathbf{X}, \mathbf{y})$ the **null hypothesis H_0 is rejected** and for which values **the null is not rejected**.

More precisely: Partition the domain of the test statistic T in two disjoint regions:

– **Rejection region, critical region \mathcal{C}**

If the test statistic $t(\mathbf{X}, \mathbf{y})$ is located in the critical region, H_0 is rejected:

$$\text{Reject } H_0 \text{ if } t(\mathbf{X}, \mathbf{y}) \in \mathcal{C}.$$

– **Non-rejection region**

If the test statistic $t(\mathbf{X}, \mathbf{y})$ falls into the non-rejection region, H_0 is *not* rejected:

$$\text{Do not reject } H_0 \text{ if } t(\mathbf{X}, \mathbf{y}) \notin \mathcal{C}.$$

– **Critical value** c : Boundary or boundaries between rejection and non-rejection region.

• **Properties of a test:**

– **Type I error, α error:**

The type I error measures the probability (evaluated before the sample is taken) of rejecting H_0 though H_0 is correct in the population,

$$\alpha(\theta) = P(\text{Reject } H_0 | H_0 \text{ is true}) = P(T \in \mathcal{C} | H_0 \text{ is true}).$$

Note: The type I error may depend on θ .

– **Type II error, β error:**

The type II error gives the probability of not rejecting H_0 though it is wrong,

$$\beta(\theta) = P(\text{Not reject } H_0 | H_1 \text{ is true}).$$

- **Size of a test:** The size of a test denotes the largest type I error that occurs for all admissible parameters θ . To be more precise, it is the supremum of type I errors over all θ that can be considered for the population model.

$$\sup_{\theta} \alpha(\theta)$$

- **Significance level:** The significance level α has to be fixed by the researcher before the test is carried out and specifies how large the type I error is allowed to be:

$$\alpha(\theta) \leq \alpha$$

From this condition one can determine the critical region $\mathcal{C} = \mathcal{C}(\alpha)$.

- **Power of a test:** The **power** of a test gives the probability of rejecting a wrong null hypothesis

$$\begin{aligned}\pi(\theta) &= 1 - \beta(\theta) = 1 - P(\text{Not reject } H_0 | H_1 \text{ is true}) \\ &= P(\text{Reject } H_0 | H_1 \text{ is true}).\end{aligned}$$

To calculate \mathcal{C} for a given α one has to know the **probability distribution** of the test statistic under H_0 .

Deriving Tests about the Sample Mean:

1. Consider two disjoint **hypotheses** about the mean of a sample.
(For example, the mean μ of hourly wages in the US in 1976.)

a) **Null hypothesis**

$$H_0 : \mu = \mu_0$$

(In our example: mean hourly wage is 6 US-\$,
thus $H_0 : \mu = 6$)

b) **Alternative hypothesis**

$$H_1 : \mu \neq \mu_0$$

(In the example: mean hourly wages are not 6 US-\$,
thus $H_1 : \mu \neq 6$)

2. Test statistic:

- a) Choice of an estimator for the unknown mean μ , e.g. the OLS estimator of a regression of hourly wages w on a constant:

Compute the sample mean

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n w_i.$$

out of a sample w_1, \dots, w_n with n observations.

- b) Obtain the probability distribution of the estimator: For simplicity assume that individual wages w_i are jointly normally distributed with expected value μ and variance σ_w^2 , that is

$$w_i \sim N(\mu, \sigma_w^2).$$

From the properties of jointly normally distributed random variables it follows that

$$\hat{\mu} \sim N\left(\mu, \sigma_{\hat{\mu}}^2\right),$$

where $\sigma_{\hat{\mu}}^2 = \text{Var}(\hat{\mu}) = \text{Var}(n^{-1} \sum w_i) = n^{-1} \sigma_w^2$.

- c) In order to obtain a test statistic $t(w_1, \dots, w_n)$ all unknown parameters have to be removed from the distribution. In this simple case this can be achieved by standardizing $\hat{\mu}$

$$t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \sim N(0, 1).$$

- d) The test statistic $t(w_1, \dots, w_n)$ can be calculated *if* we know μ and $\sigma_{\hat{\mu}}$. Assume for the moment that $\sigma_{\hat{\mu}}$ is known.

Which value takes μ under H_0 ?

$$H_0 : \mu = \mu_0.$$

Under H_0 we can compute the test statistic for a given sample as

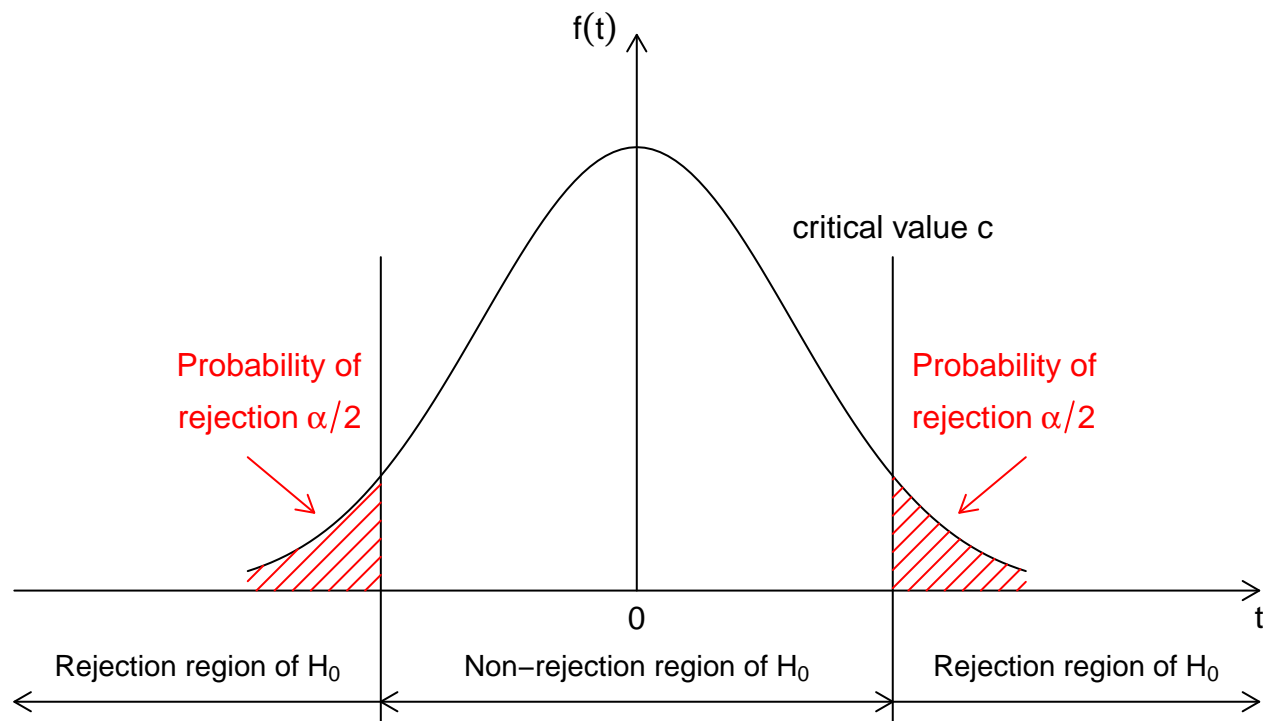
$$t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} \sim N(0, 1).$$

3. Decision rule:

When should we reject H_0 and in which case shouldn't we?

(Now the **significance level** α has to be chosen!)

If the deviation of $\hat{\mu}$ from the null hypothesis value μ_0 is **large enough**, one would reject H_0 .



Intuition: If t is very large (or very small) then

- a) the estimated mean $\hat{\mu}$ is far from μ_0 (under H_0) and / or
- b) the standard deviation $\sigma_{\hat{\mu}}$ of the estimated deviation is small relative to $\hat{\mu} - \mu_0$.

- When is $|t|$ **large enough** (to reject H_0)?
- **Note:** Under H_0 it holds that

$$t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} \sim N(0, 1)$$

and hence for given α the rejection region \mathcal{C} can be determined (see figure).

- Formally:

$$P(T < -c | H_0) + P(T > c | H_0) = \alpha$$

or in this case due to the symmetry of the normal distribution

$$P(T < -c | H_0) = \frac{\alpha}{2} \quad \text{und} \quad P(T > c | H_0) = \frac{\alpha}{2}.$$

The values of $-c$ and c are tabulated — they are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the standard normal distribution.

- Under H_1 it holds that

$$\frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \sim N(0, 1).$$

Expanding yields

$$\frac{\hat{\mu} - \mu + \mu_0 - \mu_0}{\sigma_{\hat{\mu}}} = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} + \frac{\mu_0 - \mu}{\sigma_{\hat{\mu}}} = \underbrace{\frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}}}_{t(w_1, \dots, w_n)} - \underbrace{\frac{\mu - \mu_0}{\sigma_{\hat{\mu}}}}_m$$

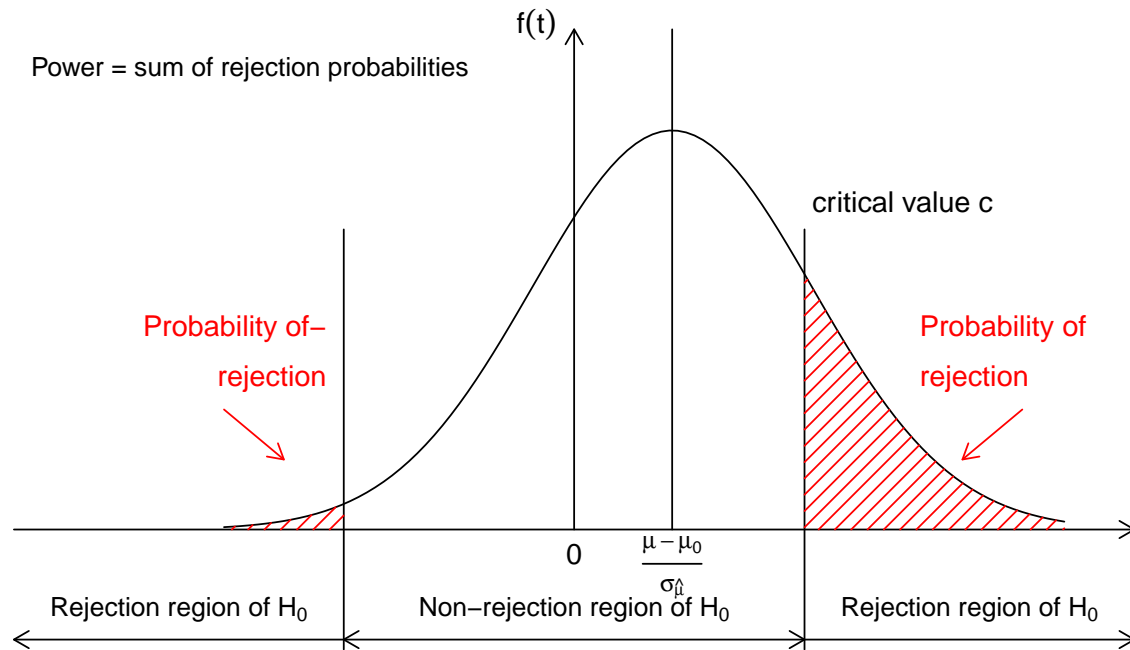
and therefore we have under H_1

$$t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} \sim N\left(\frac{\mu - \mu_0}{\sigma_{\hat{\mu}}}, 1\right)$$

since $X \sim N(m, 1)$ is equivalent to $X - m \sim N(0, 1)$.

- Conclusion: If H_1 is true, then the density of $t(w_1, \dots, w_n)$ is shifted by $(\mu - \mu_0)/\sigma_{\hat{\mu}}$.

- In the figure exhibiting the density under H_1 (for a specific value of $\mu \neq \mu_0$) the power can be seen as the sum of the two shaded areas because $\pi(\mu) = P(t < -c|H_1) + P(t > c|H_1)$.



- For a given $\hat{\sigma}_\mu$, the power of the test increases with the distance between the null hypothesis μ_0 and the true value μ .

- Recall that if H_0 is true, then $(\mu - \mu_0)/\sigma_{\hat{\mu}} = 0$ holds and one obtains the distribution under H_0 .
 - It can further be seen that the type II error — given as $\beta(\mu) = 1 - (1 - \pi(\mu)) = 1 - \pi(\mu)$ — does *not* equal zero!
4. There remains one problem: In real world applications we do not know the standard deviation of the mean estimator $\sigma_{\hat{\mu}} = \sigma_w/\sqrt{n}$.

Remedy: Estimation by

$$\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}_w}{\sqrt{n}}.$$

Then one has the popular t **statistic**

$$t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}},$$

however, watch out!

The test statistic is no longer normally distributed but follows a t distribution with $n - 1$ degrees of freedom (short t_{n-1}). Therefore

$$t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} \sim t_{n-1}.$$

To obtain the critical values

$$P(T < -c | H_0) = \frac{\alpha}{2} \quad \text{und} \quad P(T > c | H_0) = \frac{\alpha}{2},$$

the tables of the t distribution have to be considered (see Appendix G, Table G.2 in [Wooldridge \(2009\)](#)).

Wage Example Continued:

Hourly wages w_i , $i = 1, \dots, 526$ of US employees:

1. Hypotheses:

a) Null hypothesis: $H_0 : \mu = 6$

b) Alternative hypothesis: $H_1 : \mu \neq 6$

2. Estimation and calculation of the t statistic in R:

Call:

```
lm(formula = wage ~ 1)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-5.3661 -2.5661 -1.2461  0.9839 19.0839
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.896      0.161    36.62  <2e-16 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.693 on 525 degrees of freedom

Thus (using rounded values)

$$\hat{\mu} = 5.896, \quad \hat{\sigma}_{\hat{\mu}} = 0.161$$

and

$$t(w_1, \dots, w_{526}) = \frac{5.89 - 6}{0.161} = -0.6459627, \quad \text{exact: } -0.6452201.$$

3. Determination of critical values:

Suppose a significance level of $\alpha = 5\%$. Then the critical value $c = t_{525,0.05}$ can be obtained from the table for the t distribution with $n - 1 = 525$ degrees of freedom: $c = t_{525,0.05} = 1.96$.

4. Test decision: Do not reject $H_0 : \mu = 6$ since

$$-c = -1.96 < t = -0.645 < c = 1.96,$$

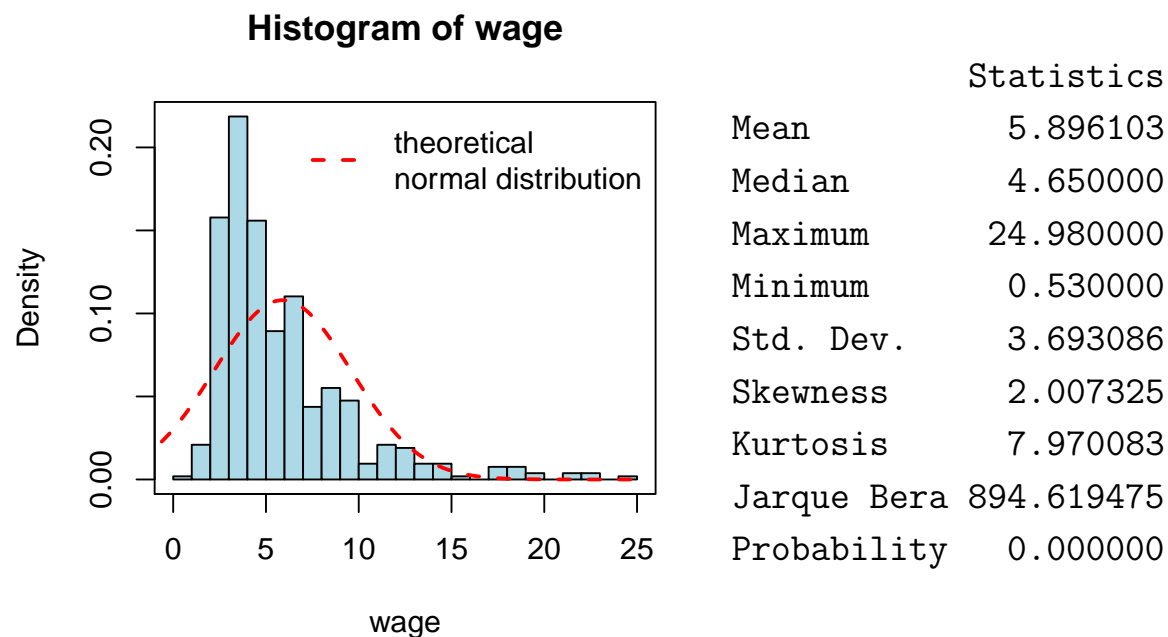
and therefore $t \notin \mathcal{C}$ (the test statistic is not contained in the rejection region).

5. **However:**

Do hourly wages w_i really follow a normal distribution as assumed?

Examine the histogram of the sample observations w_i :

Result:



- The normality condition for our test does not seem to be fulfilled. The test result could be misleading!
- There are also tests that work without the normality assumption, see Section 5.1.

One- and two-sided hypothesis tests

- **Two-sided tests**

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

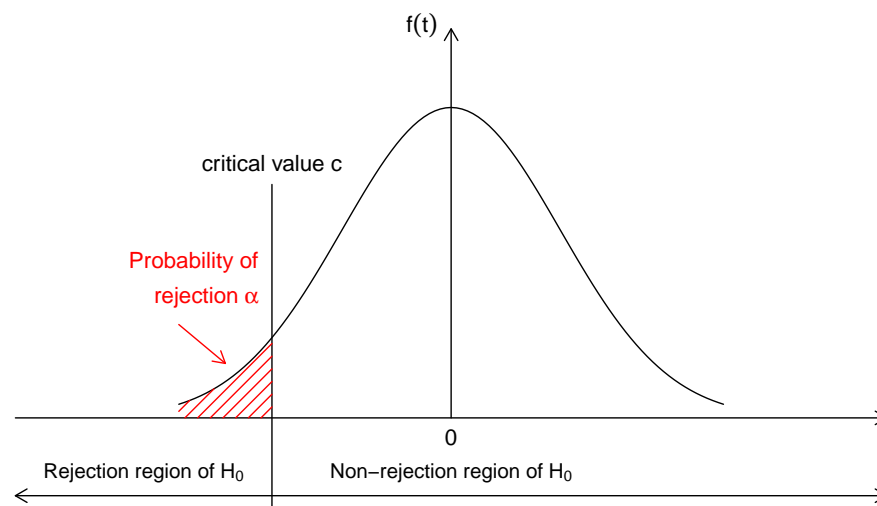
- **One-sided tests**

- **Tests with left-sided alternative hypothesis**

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

Notice: Often, also in [Wooldridge \(2009\)](#), you can read $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$. This notation, however, is somewhat imprecise since either H_0 or H_1 has to be true. This is not made clear by the latter notation.

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$



* **Decision rule:**

$$t < c \quad \Rightarrow \quad \text{Reject } H_0.$$

* You do not need a rejection region on the right hand side since all $\theta > \theta_0$ are elements of H_0 and thus fall into the non-rejection region.

- * The critical value is obtained on basis of the density for $\theta = \theta_0$ since then for a given critical value c the shaded area is larger than for any $\theta > \theta_0$ and one prefers a test for which the maximum of the type I error and thus its size is controlled. That means the size of the test is limited by the given significance level.

Wage Example Continued:

(In the following we ignore that wages are not normally distributed.)

- * The null hypothesis states that mean hourly wages are US-\$ 6 or more (H_1 says it is less than US-\$ 6):

$$H_0 : \mu \geq 6 \quad \text{versus} \quad H_1 : \mu < 6$$

- * Calculation of the test statistic: as in the two-sided case, because again μ_0 is the boundary between null and alternative hypothesis:

$$t(w_1, \dots, w_{526}) = \frac{5.896 - 6}{0.161} = -0.6459627, \quad \text{exact: } -0.6452201.$$

* Calculation of the critical value: For $\alpha = 0.05$ the critical value (note: one-sided test) from the t distribution with 525 degrees of freedom (df) is 1.645. Thus, $c = -1.645$ since the left-sided critical value is needed.

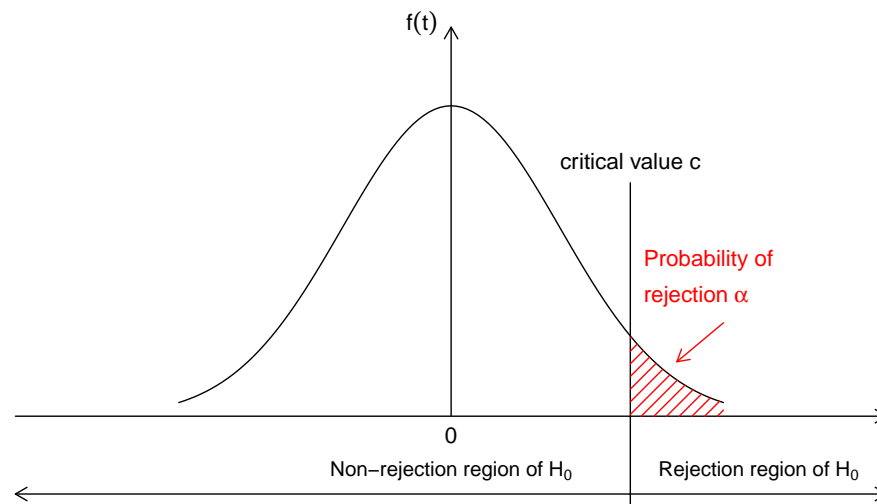
* Decision: Since

$$t = -0.6459627 > c = -1.645$$

the null hypothesis is not rejected.

– Test with right-sided alternative

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$



As with left-sided alternatives, but reversed.

- **Why do we carry out one-sided tests?** Consider the following issue: Provide statistical evidence that the mean wage is above \$ 5.60.
 - Since by using statistical tests we can never confirm but only

reject a hypothesis, we have to choose the alternative hypothesis such that it reflects our conjecture. Here, this is a mean wage larger than \$ 5.60. Rejecting the null hypothesis then provides statistical evidence for the alternative hypothesis. However, there are exceptions to this rule, see e.g. Sections 4.6 and 4.7.

- We thus have to test if the mean wage is *statistically significantly larger* than \$ 5.60.

We therefore need a test with a one-sided alternative. Our pair of hypotheses is

$$H_0 : \mu \leq 5.60 \quad \text{versus} \quad H_1 : \mu > 5.60.$$

- For $\alpha = P(T > c | H_0) = 0.05$ the critical value is $c = 1.645$.
- Decision:

$$t = \frac{5.896 - 5.60}{0.161} = 1.838509 > c = 1.645$$

⇒ Reject H_0 (for size 5%) that means data confirm that the mean wage is **statistically significantly** above \$ 5.60.

- If, on the contrary, we want to examine whether mean wages deviate from \$ 5.60 **in any direction**, the pair of hypotheses is:

$$H_0 : \mu = 5.60 \quad \text{versus} \quad H_1 : \mu \neq 5.60.$$

Given the chosen significance level, $\alpha = 0.05$, the critical values are -1.96 and 1.96, respectively, and hence

$$-1.96 < 1.84 < 1.96.$$

Thus, the null hypothesis cannot be rejected.

- It is therefore easier to reject if one has knowledge about the location of the alternative because then the region of rejection can be made smaller and it is “easier” to reject the null hypothesis if it is false.

p -values

- For every test statistic one can calculate the *largest* significance level for which — given a sample of observations — the computed test statistic would have just *not* led to a rejection of the null. This probability is called **p -value (probability value)**.

In case of a one-sided test with right-hand alternative one has (Wooldridge, 2009, Appendix C.6, p. 776)

$$P(T \leq t(\mathbf{y}) | H_0) \equiv 1 - p$$

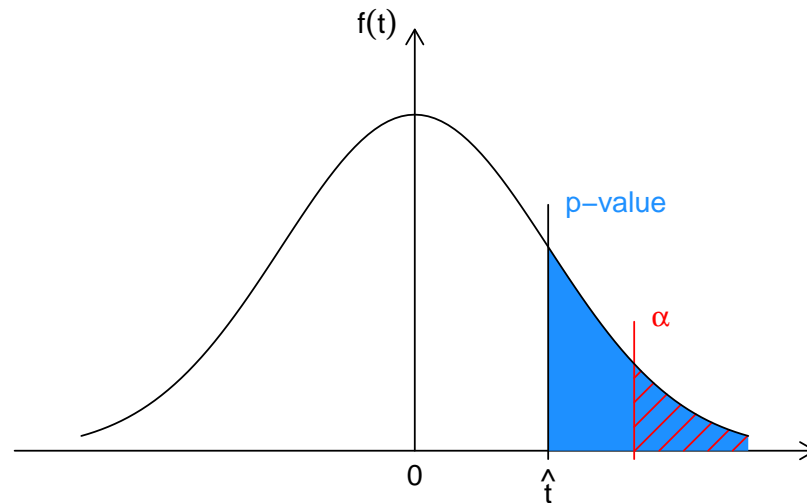
- Since $P(T > t(\mathbf{y}) | H_0) = 1 - P(T \leq t(\mathbf{y}) | H_0)$, one also has

$$P(T > t(\mathbf{y}) | H_0) = p$$

and thus it is common to say that the p -value is the *smallest* significance level at which the null *can* be rejected. Cf. Section 4.2, p. 133 in Wooldridge (2009)

- The **decision rule** of a test can also be stated **in terms of p -values**:

Reject H_0 if the **p -value is smaller** than the significance level α .



Note: In the figure \hat{t} is shorthand for $t(\mathbf{y})$.

Left-sided test: $p = P(T < t(\mathbf{X}, \mathbf{y}))$,

Right-sided test: $p = P(T > t(\mathbf{X}, \mathbf{y}))$

Two-sided test: $p = P(T < -|t(\mathbf{X}, \mathbf{y})|) + P(T > |t(\mathbf{X}, \mathbf{y})|)$

- Most software packages (e.g. R) give p -values for

$$H_0 : \theta = 0 \quad \text{versus} \quad \theta \neq 0.$$

Reading: Appendix C.6 in [Wooldridge \(2009\)](#).

4.2 Probability Distribution of the OLS Estimator

For the multiple regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

we assume MLR.1 to MLR.5, as we did in Sections 3.2 and 3.4.

- Recall from Section 3.4.1 that under MLR.1 the OLS estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

can be written as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{W}}\mathbf{u}. \quad (4.1)$$

- In order to derive the probability distribution of a test statistic one needs the probability distribution of the underlying estimators since the former is a function of the latter. Furthermore, the probability distribution of the OLS estimator is necessary to construct interval estimators, see Section 4.5.

Conditioning on the regressor matrix \mathbf{X} , it follows from (4.1) that the probability distribution of the OLS estimator only depends on the error vector \mathbf{u} . Similarly to the case of testing the mean we make the assumption that the relevant random variables are normally distributed.

- **Assumption MLR.6 (Normality of Errors):**

Conditionally on the regressor matrix \mathbf{X} , the vector of sample errors \mathbf{u} is stochastically independently and identically normally distributed as

$$u_i | x_{i1}, \dots, x_{ik} \sim i.i.d. N(0, \sigma^2), \quad i = 1, \dots, n.$$

Jointly with MLR.2, it can be equivalently written that \mathbf{u} is multivariate normal with mean zero and variance-covariance matrix $\sigma^2 \mathbf{I}$

$$\mathbf{u} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

- Of course, one could assume for the errors \mathbf{u} any other probability distribution. However, assuming normally distributed errors has two advantages:

1. The probability distribution of the OLS estimator and derived test statistics can easily be derived, see the remaining sections.

2. Under certain conditions the resulting probability distribution for the OLS estimator holds even if the errors are not normally distributed. Then it is called **asymptotic distribution**, see Chapter 5.

See Appendix B and D in [Wooldridge \(2009\)](#) for rules and properties of normally distributed random variables and vectors.

- **Properties of the multivariate normal distribution:**

- If $Z \sim N(\mu, \sigma^2)$, then $aZ + b \sim N(a\mu + b, a^2\sigma^2)$.
- If the random numbers Z and V are jointly normally distributed, then Z and V are stochastically independent if and only if $Cov(Z, V) = 0$. (Note that the conditional independence follows from $Cov(Z, V) = 0$ only for the normal distribution.)
- Every linear combination of a vector of identically and independently normally distributed random variables $\mathbf{z} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ is also normally distributed. Let

$$\mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}.$$

Then $\sum_{j=1}^n w_j z_j | \mathbf{w} = \mathbf{w}'\mathbf{z} | \mathbf{w} \sim N(\mathbf{w}'\boldsymbol{\mu}, \sigma^2\mathbf{w}'\mathbf{w})$.

More generally, it holds for $\mathbf{z} = (z_1, \dots, z_n)' \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ and

$$\mathbf{W} = \begin{pmatrix} w_{01} & w_{02} & \cdots & w_{0n} \\ w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{k1} & w_{k2} & \cdots & w_{kn} \end{pmatrix}$$

$$\begin{pmatrix} \sum_{j=1}^n w_{0j} z_j \\ \vdots \\ \sum_{j=1}^n w_{kj} z_j \end{pmatrix} | \mathbf{W} = \mathbf{Wz} | \mathbf{W} \sim N \left(\mathbf{W}\boldsymbol{\mu}, \sigma^2 \mathbf{W}\mathbf{W}' \right).$$

(4.2)

- The property (4.2) for linear combinations of normally distributed random numbers is very helpful for us since the OLS estimator (4.1)

is just such a linear combination.

Thus, one obtains

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{W} = \mathbf{W}\mathbf{u} | \mathbf{W} \sim N \left(\mathbf{0}, \sigma^2 \mathbf{W}\mathbf{W}' \right).$$

Since $\mathbf{W}\mathbf{W}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$, one obtains

$$\hat{\boldsymbol{\beta}} | \mathbf{X} \sim N \left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right).$$

Similarly one can show that

$$\hat{\beta}_j | \mathbf{X} \sim N \left(\beta_j, \sigma_{\hat{\beta}_j}^2 \right) \quad (4.3)$$

with $\sigma_{\hat{\beta}_j}^2 = \frac{\sigma^2}{\text{SST}_j(1-R_j^2)}$ (see (3.15) in Section 3.4).

- Note that (4.3) generalizes the example of Section 4.1 for testing hypotheses on the mean. If \mathbf{X} is a column vector of ones, then $\hat{\beta}_0 = \hat{\mu}$.

4.3 The t Test in the Multiple Regression Model

- **Derivation of the test statistic and its distribution**

- From (4.3) $\hat{\beta}_j | \mathbf{X} \sim N \left(\beta_j, \sigma_{\hat{\beta}_j}^2 \right)$.

- Standardizing leads to

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim N(0, 1), \quad \text{no conditioning since } \mathbf{X} \text{ only contained in } \sigma_{\hat{\beta}_j}.$$

For estimated σ^2 (no proof) the test statistic follows a t distribution with $n - k - 1$ degrees of freedom. Estimating $k + 1$ regression parameters implies $k + 1$ restrictions from the normal equations

$$t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1}.$$

- **Critical region and decision rule**

- **Two-sided test**

- * **Hypotheses:**

$$H_0 : \beta_j = \beta_{j0} \quad \text{versus} \quad H_1 : \beta_j \neq \beta_{j0}.$$

For a given significance level one obtains the critical values from the table of the t distribution such that $P(T < -c|H_0) = \alpha/2$ and $P(T > c|H_0) = \alpha/2$ or equivalently $2 \cdot P(T > c|H_0) = \alpha$.

- * **Decision rule:**

- Reject H_0 if $|t(\mathbf{X}, \mathbf{y})| > c$, otherwise do not reject H_0 .
- Alternatively: Calculate p -value

$$p = P(|T| > |t(\mathbf{X}, \mathbf{y})| | H_0) = 2 \cdot P(T > t(\mathbf{X}, \mathbf{y}) | H_0)$$

and reject H_0 if $p < \alpha$, otherwise do not reject H_0 .

– One-sided test with left-sided alternative

* Hypotheses:

$$H_0 : \beta_j \geq \beta_{j0} \quad \text{versus} \quad H_1 : \beta_j < \beta_{j0}.$$

For a given significance level one obtains the critical value from the table of the t distribution such that

$$P(T < c | H_0) = \alpha.$$

* Decision rule:

- Reject H_0 if $t(\mathbf{X}, \mathbf{y}) < c$, otherwise do not reject H_0 .
- Alternatively: Calculate p -value

$$p = P(T < t(\mathbf{X}, \mathbf{y}) | H_0).$$

and reject H_0 if $p < \alpha$, otherwise do not reject H_0 .

– One-sided test with right-sided alternative

* Hypotheses:

$$H_0 : \beta_j \leq \beta_{j0} \quad \text{versus} \quad H_1 : \beta_j > \beta_{j0}.$$

For a given significance level one obtains the critical value from the table of the t distribution such that

$$P(T > c | H_0) = \alpha.$$

* Decision rule:

- Reject H_0 if $t(\mathbf{X}, \mathbf{y}) > c$, otherwise do not reject H_0 .
- Alternatively: Calculate p -value

$$p = P(T > t(\mathbf{X}, \mathbf{y}) | H_0)$$

and reject H_0 if $p < \alpha$, otherwise do not reject H_0 .

- **Economic versus statistical significance**

- For a given (statistical) significance level α , the power of a test increases with increasing sample size since $\sigma_{\hat{\beta}_j}$ in the denominator of the test statistic decreases with sample size.
- Not being able to reject a null hypothesis may thus be simple caused by a too small sample size (if the null hypothesis is wrong in the population).
- On the other hand, if a variable has only weak influence in the population, its parameter will be significantly different from zero if the sample size is large enough. Thus, even if $\beta_j x_j$ **only has small economic** impact on the dependent variable, the variable is statistically significant.
- **Be careful:** In order to avoid estimation bias due to too small

models, significant variables must be kept in the model, see Section 3.4.1.

- **Choice of significance level**

- Two reasons for decreasing the significance level α with increasing sample size n :

- * Larger sample sizes make tests more powerful. Thus, one can decide whether the benefits of a larger sample size is only attributed to reducing the Type II error $\beta(\theta) = 1 - \pi(\theta)$ or whether one wants also to decrease the Type I error as well. In case of standard significance testing, the type I error represents the probability to include a variable in the model although it is irrelevant in the population model. Thus, it makes sense to reduce this probability as well.

- * In general one selects relevant variables from a large number of possibly relevant variables. Since for each statistically significant variable a significance level α holds, one includes erroneously on average about αK redundant variables where K denotes the total number of variables considered. Since frequently K is allowed to increase with sample size n , the significance level α should fall in order to avoid αK to increase.
- If one uses the Hannan-Quinn (HQ) (3.19) or the Schwarz (SC) (3.20) model selection criterion, then the significance level decreases with sample size. This is not the case for the AIC criterion (3.18).

- **Insignificance, multicollinearity, and sample size**

- Recall: The test statistic $t(\mathbf{X}, \mathbf{y})$ is small since

- * the deviation between the true value and the null hypothesis is small, for example between β_j and β_{j0}

- * or the estimated standard error $\hat{\sigma}_{\hat{\beta}_j}$ of β_j is large.

The latter can also be caused by multicollinearity in \mathbf{X} . Thus: A high degree of multicollinearity makes it more unlikely to reject the null hypothesis (since $|t(\mathbf{X}, \mathbf{y})|$ is small on average).

- For this reason one may keep insignificant variables in the regression. However, corresponding parameter estimates have then to be interpreted with care.

Reading: Appendices C.5, E.3 in [Wooldridge \(2009\)](#) if needed.

4.4 Example of an Empirical Analysis I: A Simplified Gravity Equation

Trade Example Continued (from Section 3.5):

Compare steps of an econometric analysis, see Section 1.2.

1. Question of interest:

Quantify impact of changes of gdp in exporting country and changes in imports to Germany.

2. Economic model:

Under idealized assumptions including complete specialization in production and identical consumption preferences among countries, no trading costs, and focusing exclusively on imports, economic theory implies (see Section II, equation (5) in Fratianni (2007))

$$imports_i = A gdp_i distance_i^{\beta_2}, \quad \beta_2 < 0.$$

This implies a unit elasticity (elasticity of 1) of *gdp* on *imports*. This means that a 1% change in *gdp* in the exporting country increases imports by 1% as well.

This hypothesis can be statistically tested.

3. **Econometric model:**

The simplest econometric model is obtained by taking logs of the economic model and adding an error term. This delivers

$$\ln(\text{imports}_i) = \beta_0 + \beta_1 \ln(\text{gdp}_i) + \beta_2 \ln(\text{distance}_i) + u_i.$$

4. **Collecting data:** see Appendix 10.4.

5. Selection and estimation of an econometric model:

In practice, there may be further variables influencing imports. Thus, further control variables have to be added. Based on the Schwarz criterion the model selection exercise in Section 3.5 suggested to add the control variable *openness*

(Model 3),

$$\ln(\text{imports}_i) = \beta_0 + \beta_1 \ln(\text{gdp}_i) + \beta_2 \ln(\text{distance}_i) + \beta_3 \text{openness}_i + u_i.$$

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

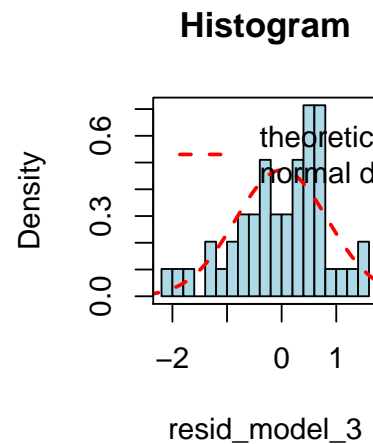
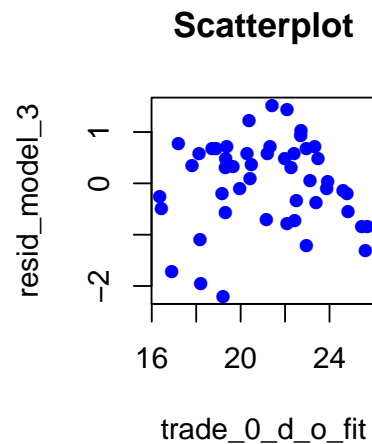
Residual standard error: 0.8731 on 45 degrees of freedom

Multiple R-squared: 0.8995, Adjusted R-squared: 0.8928

F-statistic: 134.2 on 3 and 45 DF, p-value: < 2.2e-16

6. Model diagnostics:

- Check possible violation of MLR.5 (Homoskedasticity) by plotting the residuals against the fitted values.
- Check possible violation of MLR.6 (Normal errors) by plotting a histogram of the residuals.



	Statistics
Mean	7.087363e-17
Median	1.008609e-01
Maximum	1.521959e+00
Minimum	-2.199881e+00
Std. Dev.	8.453628e-01
Skewness	-6.137689e-01
Kurtosis	2.990075e+00
Jarque Bera	3.076685e+00
Probability	2.147368e-01

The scatter plot does not indicate a violation of MLR.5. Why?
 Statistical tests for checking MLR.5 will be presented in section 9.2.

In contrast, the histogram points at an asymmetric distribution. If this is the case, the errors were not normally distributed. The asymmetry of a distribution can be measured by the third moment, the skewness. The symmetric normal distribution has a skewness of zero. Inspecting the box right to the histogram shows that the estimated skewness is about -0.6.

The fourth moment, the kurtosis is estimated close to 3 which is the theoretical value implied by the standard normal distribution.

For specialists: Whether the 3. and/or 4. moment (skewness and kurtosis) contradict the normal distribution, can be checked with the **Lomnicki-Jarque-Bera-Test**. The corresponding p value is the last line in the box. Thus, the null hypothesis of normally distributed errors cannot be rejected given any reasonable significance level.

Thus, we may continue to use this model.

7. Usage of the model: Conduct tests:

A two-sided test

- Now we can formulate the pair of statistical hypotheses:

H_0 : The elasticity of imports to gdp is 1. versus H_1 : The elasticity is unequal to 1.

$$H_0 : \beta_1 = 1 \quad \text{versus} \quad H_1 : \beta_1 \neq 1.$$

- Compute t statistic from the relevant line of the output

```

                Estimate Std. Error t value Pr(>|t|)
log(wdi_gdpusdcr_o)  0.94066    0.06134  15.335 < 2e-16 ***

```

$$t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.94066 - 1}{0.06134} = -0.9673948$$

- Choose a significance level, e.g. $\alpha = 0.05$.

Compute critical values: The degrees of freedom are $n - k - 1 =$

$49 - 3 - 1 = 45$. One may obtain an approximate critical value from Table G.2 in [Wooldridge \(2009\)](#) or a precise critical value e.g. from

– R: `(crit <- qt(0.975, df = 49 - 3 - 1))` in the command window delivering 2.014103 or

– Excel using $c = (\text{TINV}(\alpha; n - k - 1)) = 2.0106$. (Note that the Excel function already assumes a two-sided test.)

- Since

$$-c < t(\mathbf{X}, \mathbf{y}) < c$$

$$-2.014103 < -0.9673948 < 2.014103$$

one cannot reject the null hypothesis.

- p -values can be computed in R using
`pval <- 2 * pt(teststat, df = 49-3-1) = 0.3385174.`
 Thus, one cannot reject H_0 even at the 10% significance level.
 The p -value means that we would observe a t statistic of at least 0.9673948 in absolute value in about 34 samples out of 100 samples drawn given that H_0 is true.

One-sided test

- Now we can formulate the pair of statistical hypotheses with respect to the sign of β_2 , e.g. the impact of *distance* on *imports*.
 To provide evidence for $\beta_2 < 0$, we put this into H_1 :

$$H_0 : \beta_2 \geq 0 \quad \text{versus} \quad H_1 : \beta_2 < 0.$$

- Compute t statistic from the relevant line of the output

Estimate	Std. Error	t value	Pr(> t)
-9.703183e-01	1.526847e-01	-6.355048e+00	9.262691e-08

$$t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_2 - \beta_{20}}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{-0.9703183 - 0}{0.1526847} = -6.355046.$$

- Choosing again $\alpha = 0.05$, we compute the critical value using R function

$$\text{qt}(1-0.05, \text{df}=49-3-1) = 1.679427.$$

- Since

$$t(\mathbf{X}, \mathbf{y}) = -6.3550 < -1.6794 = c,$$

one rejects the null hypothesis. Thus, log distance has a statistically significant negative impact on imports at the given significance level.

- The corresponding p -value using R is $\text{pt}(\text{teststat}, \text{df}=49-3-1) = 4.631369\text{e-}08$. Thus, distance has a negative impact even at the 1% significance level.

Note that we already considered other model specifications in Section 3.5. It might be interesting to check whether these test results are robust if other model specifications are used such as Model 2 or Model 4.

4.5 Confidence Intervals

- How large is the probability that the estimated parameter value corresponds to the true value?
- A parameter estimator — to be more precise, a point estimator — does not allow any conclusions how “close” the estimate is to the true value of the population.
- Following the position of Sir Karl Popper who advocated the **critical rationalism** in the philosophy of science, point estimates are not very useful since it cannot be falsified. Instead, an empirical hypothesis is only scientific if it is **falsifiable**.
- Example: Assume we predicted on basis of an econometric model a price index and obtained a predicted value of 5.12. the realized value, however, will be 5.24. → Then we made a **wrong** prediction

since it did not realize exactly.

This “error” can only have three reasons:

- The random error of the population regression model.
- The estimation error of the sample regression model.
- The regression model is not correct or (more realistic) it is a bad approximation. At least one of our assumptions is not justified.

Problem:

From an subjective point of view one can have different opinions about these “explanations”:

- One believes that the deviation is due to the random error.
- Another claims that the model is wrong.

Solution

One should specify **objective criteria** such that one can make a scientific decision. These criteria should be determined before any predicted value realizes.

Then one cannot escape a potential falsification of a hypothesis afterwards. This makes a hypothesis scientific in the sense of Popper.

- Let's be more precise:

How large is the probability that the estimated value $\hat{\beta}_j$ corresponds exactly to the true value β_j if, as was shown in Section 4.3,

$$\hat{\beta}_j \sim N \left(\beta_j, \sigma_{\hat{\beta}_j}^2 \right)$$

and $(\hat{\beta}_j - \beta_j) / \sigma_{\hat{\beta}_j} \sim N(0, 1)$, or if $\sigma_{\hat{\beta}_j}$ is estimated,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} ?$$

- **Alternative question:**

How large is the probability that *prior to observing a sample* the true value β_j lies in the interval

$$[\hat{\beta}_j - c \cdot \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + c \cdot \hat{\sigma}_{\hat{\beta}_j}]$$

where c is given?

Note that the endpoints of the interval are random *prior to obtaining a sample*. Its **location is random** through $\hat{\beta}_j$ and its **length is random** through $\hat{\sigma}_{\hat{\beta}_j}$

This interval is the most well known example of an **interval estimator**.

- **Answer for given $\sigma_{\hat{\beta}_j}$:**

How large is the probability that the true value β_j is contained in

the interval $[\hat{\beta}_j - c \cdot \sigma_{\hat{\beta}_j}, \hat{\beta}_j + c \cdot \sigma_{\hat{\beta}_j}]$ which is random prior to observing a sample and where the value c is chosen by you?

– It is $2\Phi(c) - 1$ since

$$\begin{aligned}
 P\left(\hat{\beta}_j - c\sigma_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + c\sigma_{\hat{\beta}_j}\right) &= P\left(-c\sigma_{\hat{\beta}_j} \leq \beta_j - \hat{\beta}_j \leq c\sigma_{\hat{\beta}_j}\right) \\
 &= P\left(-c \leq \frac{\beta_j - \hat{\beta}_j}{\sigma_{\hat{\beta}_j}} \leq c\right) \\
 &= P\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \leq c\right) \\
 &= \Phi(c) - \Phi(-c) \\
 &= \Phi(c) - (1 - \Phi(c)) \\
 &= 2\Phi(c) - 1.
 \end{aligned}$$

– Example: For $c = 1.96$ one obtains $\Phi(1.96) - \Phi(-1.96) = 0.975 - 0.025 = 0.95$:

The true value β_j will be with 95% probability within the interval $\hat{\beta}_j \pm c \cdot \sigma_{\hat{\beta}_j}$. One also relates this probability to α by writing $0.95 = 1 - \alpha$. Thus one has $\alpha = 0.05$.

- **Answer for estimated $\sigma_{\hat{\beta}_j}$:** The true value β_j lies in the interval $\hat{\beta}_j \pm c \cdot \hat{\sigma}_{\hat{\beta}_j}$ with probability $1 - \alpha$. Note, however, that for computing the probability one has to use the t_{n-k-1} distribution since

$$P\left(\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + c\hat{\sigma}_{\hat{\beta}_j}\right) = P\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \leq c\right).$$

- The interval

$$[\hat{\beta}_j - c \cdot \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + c \cdot \hat{\sigma}_{\hat{\beta}_j}]$$

is called **confidence interval**. One says that the confidence interval contains the true value with a probability of confidence of $(1 - \alpha)100\%$. The value $(1 - \alpha)$ is also called **confidence level** or **coverage probability** of the confidence interval.

- **In practice** one determines the confidence level $1 - \alpha$ and then computes the value c using the appropriate distribution: either $N(0, 1)$ or t_{n-k-1} .
- **Interpretation:** Would one draw R times new samples from a given population and compute a confidence interval for each sample for given confidence level $1 - \alpha$, then the true value would be contained in the confidence intervals in about $(1 - \alpha)R$ cases.
- **Note:**
 - If a sample was already taken and a confidence interval computed,

then the true parameter is either contained in the confidence interval computed for this sample or not. In other words, it does not make sense to talk about a coverage probability w.r.t. the given sample.

- The constant c corresponds to the (upper) critical value of a two-sided test with significance level α .
- Since the confidence interval is a random interval, its location and length is in general different for each sample.
- The larger $(1 - \alpha)$, the smaller α , the larger is the confidence interval. In other words: the more you want to be on the safe side, the larger the confidence interval becomes. Why?

- A two-sided t test and a confidence interval contain the same amount of information. The null hypothesis of a two-sided t test is rejected if and only if the value of the null hypothesis lies outside the confidence interval. Draw a graph to make this clear.
- A confidence interval for a given sample contains all null hypotheses of a two-sided t test that cannot be rejected for significance level α .
- If keep drawing new samples from a population, how many confidence intervals do not contain the true value on average?

- **Trade Example Continued** (from Section 4.4):

- Compute a 95% confidence interval for the elasticity β_{gdp} of imports with respect to gdp.
- From Section 4.4 it can be justified that MLR.1 to MLR.6 hold and imports are normally distributed.
- Since $\sigma_{\hat{\beta}_{gdp}}$ has to be estimated, one has to use the t distribution with $n - k - 1 = 45$ degrees of freedom. For a confidence level of 0.95 one obtains $\alpha = 0.05$ and thus $c = 2.014103$ (z.B. in R via `qt(1-0.05/2, df = 49 - 3 - 1)`).
- The relevant line of output was, see Section 4.4:

```

                Estimate Std. Error t value Pr(>|t|)
log(wdi_gdpusdcr_o) 0.94066    0.06134  15.335 < 2e-16 ***

```

– Therefore the 95% confidence interval is given by

$$\begin{aligned} & [\hat{\beta}_{BIP} - c \cdot \hat{\sigma}_{\hat{\beta}_{BIP}}, \hat{\beta}_{BIP} + c \cdot \hat{\sigma}_{\hat{\beta}_{BIP}}] \\ & [0.94066 - 2.014103 \cdot 0.06134, 0.94066 + 2.014103 \cdot 0.06134] \\ & [0.81712, 1.06421]. \end{aligned}$$

– All null hypotheses for the elasticity of imports with respect to gdp in the confidence interval $[0.81712, 1.06421]$ cannot be rejected given confidence level 95%. Note that 1 is included in the confidence interval. This reflects the test result in Section 4.4 of not rejecting $H_0 : \beta_{gdp} = 1$.

4.6 Testing a Single Linear Combination of Parameters

- **Example:** Cobb-Douglas production function

$$\log Y = \beta_0 + \beta_1 \log K + \beta_2 \log L + u,$$

where Y denotes output, K and L denote the production factors capital and labor, respectively. Note that β_1 and β_2 are elasticities here.

If the restriction $\beta_1 + \beta_2 = 1$ holds true, the production function has constant returns to scale, e.g. a 1% increase of labor and capital leads to a 1% increase of output on average.

For an empirical test of constant returns to scale, we employ the following pair of hypotheses:

$$H_0 : \beta_1 + \beta_2 = 1 \quad \text{versus} \quad H_1 : \beta_1 + \beta_2 \neq 1.$$

- **How to construct the test statistic:**

1. First, define auxiliary parameters θ and θ_0 , where

$$\theta = \beta_1 + \beta_2, \quad \theta_0 = 1,$$

or, equivalently

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

2. Second, solve θ for one of the parameters β_i , here β_1

$$\beta_1 = \theta - \beta_2$$

and insert it into the initial regression equation and reformulate the latter to

$$\begin{aligned} \log Y &= \beta_0 + (\theta - \beta_2) \log K + \beta_2 \log L + u \\ \log Y &= \beta_0 + \theta \log K + \beta_2 \underbrace{(\log L - \log K)}_{\text{new variable}} + u. \end{aligned} \quad (4.4)$$

Then estimate (4.4) and obtain the test statistic

$$t_\theta = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}}$$

which can be directly calculated from the estimation of (4.4).

Example:

In a classical marketing model we regress (the natural logarithm of) sales (S) of a consumer good on (the natural logarithm of) this good's price (P) as well as on (the natural logarithms of) cross prices (P_{K1} , P_{K2}) of competing goods. The following regression output is calculated from the data:

Call:

```
lm(formula = log(S) ~ log(P) + log(P_K1) + log(P_K2))
```

Residuals:

Min	1Q	Median	3Q	Max
-4.8760	-0.6421	-0.0098	0.6352	3.7577

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.40779	0.07956	55.40	<2e-16 ***
log(P)	-3.95528	0.06809	-58.09	<2e-16 ***
log(P_K1)	0.71027	0.07391	9.61	<2e-16 ***
log(P_K2)	1.15416	0.07982	14.46	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.022 on 6913 degrees of freedom

Multiple R-squared: 0.3323, Adjusted R-squared: 0.332

F-statistic: 1147 on 3 and 6913 DF, p-value: < 2.2e-16

We wish to test the following statement: the cross price elasticities are identical, keeping everything else fixed (*ceteris paribus*) (though the competing goods come from different market segments).

- The initial hypotheses are given by

$$H_0 : \beta_{K1} = \beta_{K2} \quad \text{versus} \quad H_1 : \beta_{K1} \neq \beta_{K2}.$$

We reformulate them by re-parametrization according to

$$\theta = \beta_{K1} - \beta_{K2}, \quad \theta_0 = 0$$

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0.$$

- Thus, due to $\beta_{K1} = \theta + \beta_{K2}$, the initial regression model

$$\ln(S) = \beta_1 + \beta_2 \ln(P) + \beta_{K1} \ln(P_{K1}) + \beta_{K2} \ln(P_{K2}) + u$$

can be rendered to

$$\ln(S) = \beta_1 + \beta_2 \ln(P) + \theta \ln(P_{K1}) + \beta_{K2}(\ln(P_{K2}) + \ln(P_{K1})) + u.$$

- Given the estimates of the last regression

```
lm(formula = log(S) ~ log(P) + log(P_K1) + I(log(P_K1) + log(P_K2)))
```

Residuals:

Min	1Q	Median	3Q	Max
-4.8760	-0.6421	-0.0098	0.6352	3.7577

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.40779	0.07956	55.403	< 2e-16 ***
log(P)	-3.95528	0.06809	-58.085	< 2e-16 ***
log(P_K1)	-0.44389	0.11254	-3.944	8.09e-05 ***
I(log(P_K1) + log(P_K2))	1.15416	0.07982	14.460	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.022 on 6913 degrees of freedom

Multiple R-squared: 0.3323, Adjusted R-squared: 0.332

F-statistic: 1147 on 3 and 6913 DF, p-value: < 2.2e-16

calculate the t statistic as

$$t = \frac{-0.44389 - 0}{0.112544} \approx -3.94, \quad \text{exact value: } -3.944165.$$

For a given significance level of $\alpha = 0.05$, the critical values are -1.96 and 1.96. Thus, we have to reject H_0 .

Reading: Sections 4.3-4.4 in [Wooldridge \(2009\)](#).

4.7 Jointly Testing Several Linear Combinations of Parameters: The F Test

Some examples of possible restrictions within the MLR framework:

1. $H_0 : \beta_1 = 3$
2. $H_0 : \beta_2 = \beta_k$
3. $H_0 : \beta_1 = 1, \beta_k = 0$
4. $H_0 : \beta_1 = \beta_3, \beta_2 = \beta_4$
5. $H_0 : \beta_j = 0, j = 1, \dots, k$
6. $H_0 : \beta_j + 2\beta_l = 1, \beta_k = 2$

We can already check case 1. and case 2. by applying t tests. For all other cases we need the F test.

4.7.1 Testing of Several Exclusion Restrictions

Trade Example Continued (from Section 4.5):

Consider Model 4 in Section 3.5:

```
lm(formula = log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
    ebrd_tfes_o + log(cepii_area_o))
```

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

Multiple R-squared: 0.9062, Adjusted R-squared: 0.8976

F-statistic: 106.2 on 4 and 44 DF, p-value: < 2.2e-16

Are the control variables *openness* (`EBRD_TFES_0`) and *area* (`LOG(CEPII_AREA_0)`) really needed in the specification of Model 4?

To put it more precisely, are both parameters of two variables mentioned jointly significantly different from zero?

$$H_0 : \beta_{openness} = 0 \text{ and } \beta_{area} = 0$$

versus

$$H_1 : \beta_{openness} \neq 0 \text{ and/or } \beta_{area} \neq 0$$

How can one jointly test several hypotheses?

- Note that SSR decreases (or stays constant) with an additional regressor.

⇒ **Idea:** Compare the SSR of a model on which the null hypotheses are imposed (restricted model) with the SSR of another model that does not impose the joint restrictions (unrestricted model).

- The estimation under H_0 is easy: simply exclude all regressors from the regression whose parameters under H_0 are set to zero and re-estimate the restricted model.

In case of Model 4 for the trade example the OLS estimates are for the restricted model (that corresponds to Model 2 in Section 3.5):

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

Multiple R-squared: 0.8838, Adjusted R-squared: 0.8787

F-statistic: 174.9 on 2 and 46 DF, p-value: < 2.2e-16

Results:

- The R^2 of the unrestricted model is 0.9062 while the R^2 of the restricted model is 0.8838.
- Correspondingly, the standard error of regression $\hat{\sigma}$ increases from 0.853 to 0.9284.
- Are these changes large? It looks like that but what does “large” really mean here?
- Note that all three model selection criteria, AIC, HQ, and SC, “prefer” the unrestricted model, see Section 3.5. Will this finding be confirmed by the test?

- In order to be able to use a statistic (a function that can be computed from sample values) as a test statistic, one has to know its probability distribution under the null hypothesis H_0 .

One can show (\rightarrow master course Methods of Econometrics or Section 4.4 in [Davidson and MacKinnon \(2004\)](#)) that the following test statistic follows an F distribution

$$F = \frac{(\text{SSR}_{H_0} - \text{SSR}_{H_1})/q}{\text{SSR}_{H_1}/(n - k - 1)} \sim F_{q, n-k-1}.$$

Therefore this test is called F **test** and the test statistic is abbreviated as F **statistic**.

- Note that the F **distribution** has two different degrees of freedom, q degrees of freedom for the random variable in the numerator, and $n - k - 1$ degrees of freedom for the random variable in denominator.

The value q contains the **number of restrictions that are jointly tested**.

- **Details** of the F statistic:
 - Its minimum is 0 since $SSR_{H_0} \geq SSR_{H_1}$ and $SSR_{H_1} > 0$. (Therefore the F statistic cannot be normally distributed!)
 - There is no upper bound.
- When should the joint null hypothesis be rejected?
 - The larger the absolute difference between the SSRs of the restricted and the unrestricted model, $SSR_{H_0} - SSR_{H_1}$, the more likely is a violation of the exclusion restrictions since then the excluded variables are likely to contribute to a much smaller SSR of the unrestricted model which points at the relevance of the excluded variables.

- However, be aware that absolute differences do not say much. Why?
- It makes much more sense to consider the relative difference between the SSRs. This is exactly what the F statistic does. It scales the difference in SSRs by the SSR of the unrestricted model. If the relative difference is large, then the joint null hypothesis is likely to be violated.
- On the other hand, if the relative difference is small, then it is likely that the excluded variables do not have any relevant impact in the unrestricted model since they can be neglected without any noticeable effect.

- **Decision rule:**

Reject H_0 if the test statistic is larger than the **critical value**:

$$\text{Reject } H_0 \text{ if } F > c.$$

Thus, the **critical region** is (c, ∞) .

Calculation of the critical region:

For a given significance level α , the critical value c is implicitly defined by the probability

$$P(F > c | H_0) = \alpha.$$

The corresponding value for c given α can be found in tables on the F distribution, e.g. Table G.3 in Appendix G in [Wooldridge \(2009\)](#) or be computed in R (`qf(1-alpha, df1=q, df2= n-k-1)`) or Excel (`Finv(0,05;q;n-k-1)` für `alpha=0,05`).

Trade Example Continued (from the beginning of this section):

- The joint null hypothesis contains two exclusion restrictions, thus the degree of freedoms for the numerator are two, $q = 2$. The degrees of freedom for the denominator correspond to the degrees of freedom of Model 4, $n - k - 1 = 49 - 4 - 1 = 44$. Choosing a significance value of $\alpha = 0.05$, we check Table G.3 in Appendix G in [Wooldridge \(2009\)](#) for the appropriate critical value. Listed values are $F_{2,40} = 3.23$ and $F_{2,60} = 3.15$. While the former implies a true significance level smaller than 0.05, the latter implies one above 0.05. If one is interested in an exact critical value, one can obtain it from R, nämlich $\text{qf}(1-0.05, 2, 44) = 3.209278$.
- From the standard errors and degrees of freedom of the regression outputs for Model 4 and Model 2 at the beginning of the section, one can compute the SSRs ($\text{SSR} = (\text{Residual standard error})^2$)

* df) and thus the F statistic as

$$F = \frac{(39,64485 - 32,01770)/2}{32,01770/44} = 5,240768.$$

Since

$$F = 5,240768 > 3,20928 = c,$$

reject H_0 on a significance level of 5%.

- Check that the same decision holds for a significance level of 1%.

The two variables openness ($EBRD_TFES_O$) und area ($LOG(CEPII_AREA_O)$) are statistically significant at the 1% significance level and thus at least one of the two variables has an impact on imports on the 5% as well as on the 1% significance level.

Calculation of p -values for F statistics:

- In empirical work one is frequently interested in the *largest* significance level for which it is not possible to reject the null hypothesis **given** the observed test statistic.

As explained in Section 4.1, this information is provided by the p -value. Alternatively, it is the *smallest* significance level at which the null can be rejected.

Given the significance level that was chosen prior to any calculations, the null hypothesis is rejected if the p -value is *smaller* than the given significance level α .

- **Trade Example Continued:** The p -value can be computed in Excel (`=FVERT(5,24077;2;44)=0,00909`). Der p value can also be calculated in R:

$$1 - \text{pf}(5.24077, \text{df1} = 2, \text{df2} = 44) = 0.00908809.$$

Thus, there is strong statistical evidence against the null hypothesis.

Direct calculation of the F statistik in R:

- For computing the F statistic one uses the R package `car`, which has to be installed when used for the first time with `install.packages("car")`. One always has to load the package with the command `library(car)`.
- To carry out the F test, one applies the command `linearHypothesis(model,...)`. In the given example one uses:


```
linearHypothesis(model_4, c(" ebrd_tfes_o = 0",
"log(cepii_area_o) = 0")). One obtains
```

Linear hypothesis test

Hypothesis:

$\text{ebrd_tfes_o} = 0$

$\log(\text{cepii_area_o}) = 0$

Model 1: restricted model

Model 2: $\log(\text{trade_0_d_o}) \sim \log(\text{wdi_gdpusdcr_o}) + \log(\text{cepii_dist}) + \text{ebrd_tfes_o} + \log(\text{cepii_area_o})$

	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

Remarks:

- One can, of course, test the simple null hypothesis with a two-sided alternative

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

by means of an F test.

It holds that the square of a random variable X that follows a t distribution with $n - k - 1$ degrees of freedom just corresponds to a random variable that follows an F distribution with $(1, n - k - 1)$ degrees of freedom

$$X \sim t_{n-k-1} \quad \Longrightarrow \quad X^2 \sim F_{1,n-k-1}.$$

Therefore, a two-sided t test and an F test lead to exactly the same result for the pair of hypotheses above.

- It may happen that each regressor tested by itself is not statistically significant but if they are jointly tested they are statistically significant (at the same significance level). This is a sign of multicollinearity between the regressors considered. Then, the given sample size is only sufficient for providing statistical significance jointly for both regressors. However, it is not sufficient for providing statistical evidence for each regressor separately. In such cases you may check the covariance between the parameter estimates that are included in the test (in R: `vcov(model)` returns covariance matrix of parameter estimates).
- It may also happen that one variable is statistically significant but if jointly tested with other variables it becomes insignificant. This can happen if the other variables that are included in the joint hypothesis are redundant in the population regression. In this case, the power of

a single hypothesis test is weakened by the other irrelevant variables.

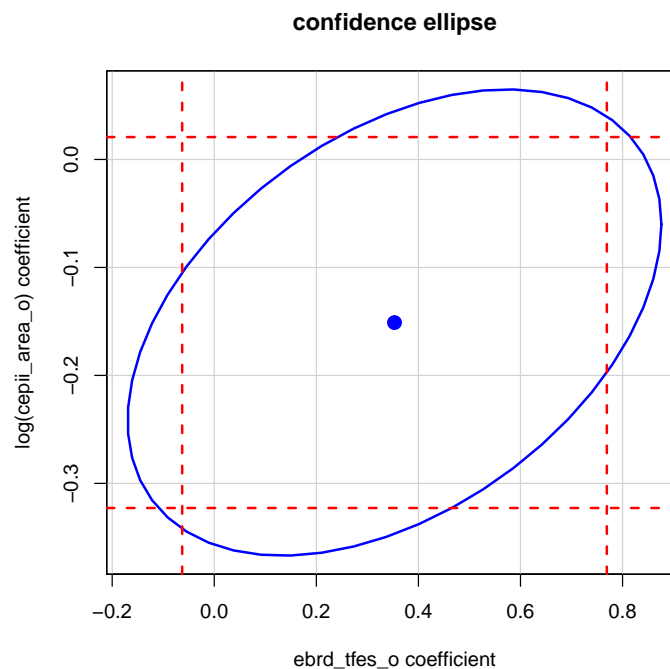
- Thus, there is no general rule on whether to prefer joint or single tests results.
- **Trade Example Continued** (from the middle of this section): Comparing four different model specifications using model selection criteria, see Section 3.5, AIC favors Model 4 (SC favors Model 3). Inspecting its parameter estimates at the beginning of this section, one finds two parameters to be statistically insignificant even at the 5% level: $\beta_{openess}$ and β_{area} .

Why, then, was Model 4 found to be best by AIC but not Model 2 that does not contain both insignificant variables?

Answer:

The parameter estimators for $\beta_{openess}$ and β_{area} might be highly correlated so that only a joint impact is significant. One reason could be that a lot of variation of *openess* can be explained by *area*, among other things. The F test above already showed that both parameters are jointly significant at the 1% level.

The effect of multicollinearity can nicely be seen in



The ellipse is a generalization of confidence intervals to two dimensions. Thus, all points outside the ellipse are joint null hypotheses that are rejected. Note that the origin also lies outside while the zero is included in each one-dimensional confidence interval. ((One obtains the plot with the R command `confidenceEllipse(...)`. See the R program in the appendix [10.5](#), Folie 270, for details.)

- **R^2 version of the F statistic:**

If a regression model contains a constant, then the decomposition $SSR = SST(1 - R^2)$ holds. Inserting each SSR into the F statistic delivers

$$F = \frac{(R_{H_1}^2 - R_{H_0}^2)/q}{(1 - R_{H_1}^2)/(n - (k + 1))} \sim F_{q, n-k-1}.$$

Note:

- SST is canceled if the dependent variable y is the same under H_0 and H_1 as, for example, in case of exclusion restrictions. However, this is not always true if general linear restrictions are tested.
- There can be slight differences between both versions of the F statistic due to rounding errors.

Overall F Test

Standard software packages (such as R) include in their OLS output for the multiple regression model $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$ the F statistic and its p -value for the pair of hypotheses:

“None of the (non-constant) regressors has impact on the dependent variable and thus the corresponding parameters are all zero.”

$$H_0 : \beta_1 = \dots = \beta_k = 0 \quad (\text{and } y = \beta_0 + u)$$

$$H_1 : \beta_j \neq 0 \text{ for at least one } j = 1, \dots, k.$$

If H_0 is not rejected, this possibly indicates that

- all regressors are possible badly/wrongly chosen,
- or at least a substantial number of regressors has no impact on y ,
- or too many regressors were considered for given sample size n .

This test is a first rough check for the validity of the model.

4.7.2 Testing of Several General Linear Restrictions

- Generalization of the F test for exclusion restrictions.
- Works equivalently by computing the relative change in the SSRs.
- R^2 version cannot be used in this case!

Examples of possible pairs of hypotheses:

$$H_0 : \beta_2 = \beta_3 = 1 \quad \text{versus} \quad H_1 : \beta_2 \neq 1 \text{ and/or } \beta_3 \neq 1,$$

$$H_0 : \beta_1 = 1, \beta_j = 2\beta_l \quad \text{versus} \quad H_1 : \beta_1 \neq 1 \text{ and/or } \beta_j \neq 2\beta_l.$$

Trade Example Continued (from previous subsection):

- One may conjecture that due to the multicollinearity between the estimates for *openness* and *area* the impact of *openness* might be underestimated in absolute value (in Model 3 the parameter estimate was 0.507250) while the impact of *area* is zero. Thus, consider the

pair of hypotheses:

$$H_0 : \beta_{openess} = 0.5 \quad \text{and} \quad \beta_{area} = 0$$

$$H_1 : \beta_{openess} \neq 0.5 \quad \text{and/or} \quad \beta_{area} \neq 0$$

In order to compute the SSR under H_0 impose these restrictions on the regression as

$$\log(imports) - (0,5)openess = \beta_0 + \beta_{gdp} \log(gdp) + \beta_{distance} distance + u$$

The R output is:

Call:

```
lm(formula = log(trade_0_d_o) - 0.5 * ebrd_tfes_o ~ log(wdi_gdpusdcr_o) +
    log(cepii_dist))
```

Residuals:

Min	1Q	Median	3Q	Max
-2.1968	-0.5605	0.1032	0.5904	1.5233

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.76870	2.02633	1.366	0.178
log(wdi_gdpusdcr_o)	0.94117	0.05922	15.893	< 2e-16 ***
log(cepii_dist)	-0.97180	0.14596	-6.658	2.97e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8636 on 46 degrees of freedom

Multiple R-squared: 0.8884, Adjusted R-squared: 0.8836

F-statistic: 183.1 on 2 and 46 DF, p-value: < 2.2e-16

This allows to compute the F statistic

$$\begin{aligned}
 F &= \frac{(SSR_{H_0} - SSR_{H_1}) / q}{SSR_{H_1} / (n - k - 1)} \\
 &= \frac{(34.30373 - 32.01770) / 2}{32.01770 / 44} = 1.570776 < c = 3.20928.
 \end{aligned}$$

Direkt in R: `linearHypothesis(model_4, c("ebrd_tfes_o = 0.5", "log(cepii_area_o) = 0"))`:

Linear hypothesis test

Hypothesis:

`ebrd_tfes_o = 0.5`

`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	34.304				
2	44	32.018	2	2.286	1.5708	0.2193

→ *The claim that the “area of a country has no effect and openness an impact of 0.5”, cannot be rejected given any reasonable significance level since the p value is about 22%.*

4.8 Reporting Regression Results

In general, empirical researchers investigate a number of different specifications of regression functions.

In order to make visible how robust the conclusions are with respect to model choice it is good practice to report the results of the most important specifications so that each reader can evaluate the findings in her own manner.

This is most easily achieved by summarizing the relevant results in a table, see the example below.

For each specification a minimum number of results should be:

- OLS parameter estimates $\hat{\beta}_j$ of the regression parameters β_j , $j = 0, 1, \dots, k$ (plus variable names),
- Standard error of $\hat{\beta}_j$, $\hat{\sigma}_{\hat{\beta}_j}$,
- Number of observations n ,
- R^2 and adjusted R^2 ,
- Standard error of regression or estimated variance of the regression error $\hat{\sigma}^2$.

If possible, one should also report

- Model selection criteria such as AIC, HQ or SC,
- Sum of squared residuals (SSR).

Based on the SSRs one can easily compute F tests.

Trade Example Continued:

Dependent Variable: $\ln(\text{Imports by Germany})$				
Independent Variables / Model	(1)	(2)	(3)	(4)
constant	-5.77 (2.184)	4.676 (2.178)	2.741 (2.175)	2.427 (2.132)
$\ln(gdp)$	1.077 (0.087)	0.975 (0.063)	0.940 (0.0613)	1.025 (0.076)
$\ln(distance)$	—	-1.074 (0.156)	-0.970 (0.152)	-0.888 (0.156)
$openess$	—	—	0.507 (0.191)	0.353 (0.206)
$\ln(area)$	—	—	—	-0.151 (0.085)
Number of observations	49	49	49	49
R^2	0.765	0.883	0.899	0.906
Standard error of regression	1.304	0.928	0.873	0.853
Sum of squared residuals	80.027	39.644	34.302	32.017
AIC	3.4100	2.7484	2.6445	2.6164
HQ	3.4393	2.7924	2.7031	2.6896
SC	3.4872	2.8642	2.7989	2.8094

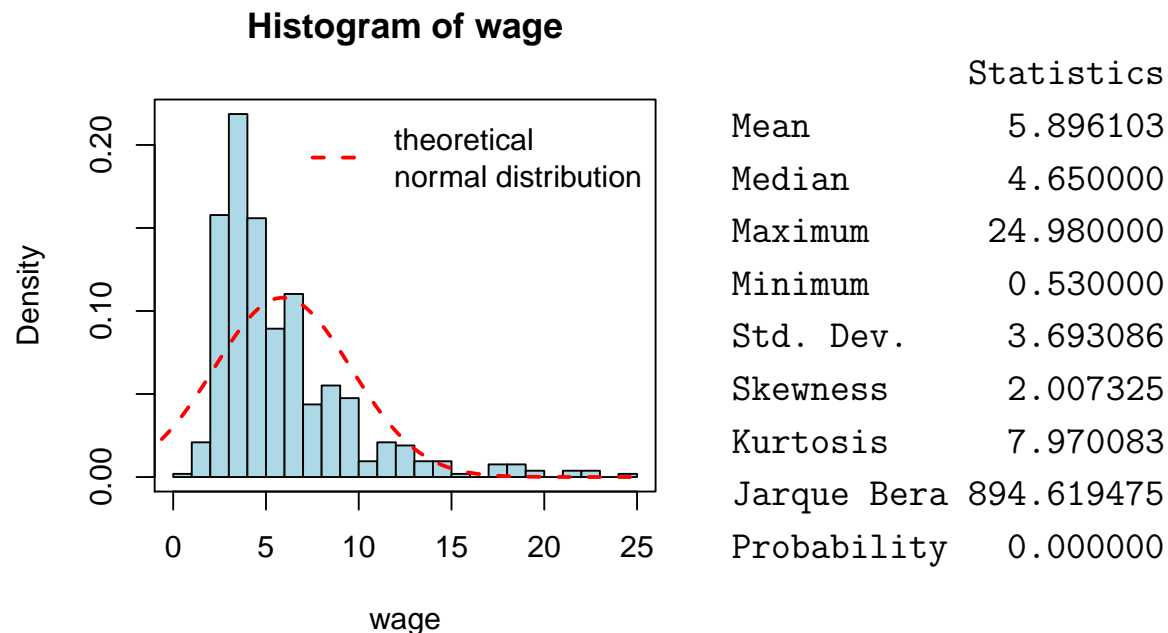
Reading: Sections
4.5-4.6 in **Wooldridge**
(2009).

5 Multiple Regression Analysis: Asymptotics

The assumption of a normal (or gaussian) distribution MLR.6 is frequently violated in empirical practice. How can we then proceed to calculate test statistics or confidence intervals?

5.1 Large Sample Distribution of the Mean Estimator

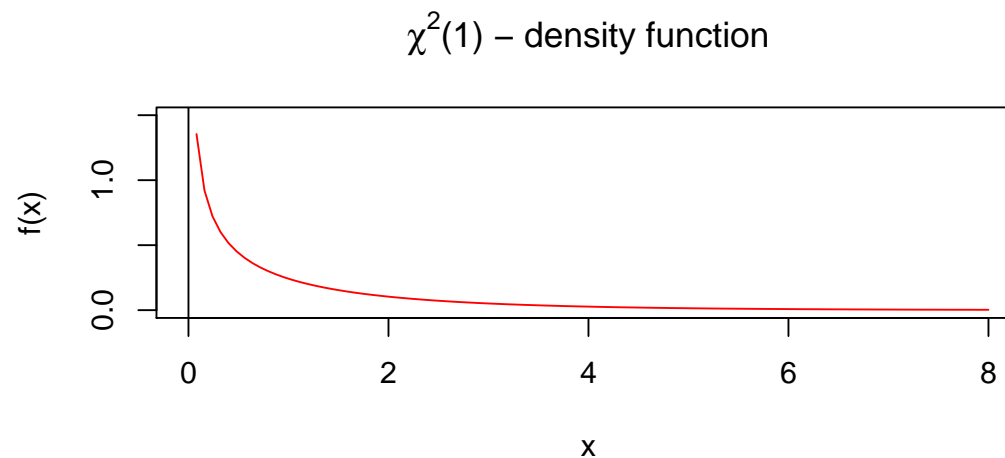
- Example: Testing the mean of hourly wages: the empirical distribution is steep at the left and skewed to the right (as is typical for prices and wages which are not generated additively).



- Examples of random variables with right-skewed distribution:
 - A $\chi^2(m)$ distributed random variable X is defined as the sum of m squared i.i.d. standard normal random variables

$$X = \sum_{j=1}^m u_j^2, \quad u_j \sim i.i.d.N(0, 1).$$

(Details on the χ^2 distribution can be found in Appendix B in [Wooldridge \(2009\)](#).)



Moments of a $\chi^2(1)$ distributed random variable:

$$E[X] = E[u^2] = \text{Var}(u) + E[u]^2 = 1,$$

$$\text{Var}(X) = E[X^2] - E[X]^2 = E[u^4] - 1^2 = 2,$$

$$\frac{u^2 - 1}{\sqrt{2}} = \frac{X - 1}{\sqrt{2}} \sim (0, 1).$$

Note that for a standard normal random variable we have $E[u^4] = 3$ (= kurtosis).

– Linear functions of a $\chi^2(1)$ distributed random variable, e.g.

$$y_i = \nu + \sigma_y \frac{u_i^2 - 1}{\sqrt{2}}, \quad u_i \sim i.i.d.N(0, 1). \quad (5.1)$$

Moments:

$$E[y_i] = \nu,$$

$$Var(y_i) = Var\left(\sigma_y \frac{u_i^2 - 1}{\sqrt{2}}\right) = \sigma_y^2 Var\left(\frac{u_i^2}{\sqrt{2}}\right) = \sigma_y^2.$$

- **Expectation and variance of mean estimators**

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n y_i.$$

$$E[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^n E[y_i] = \nu,$$

$$\text{Var}(\hat{\mu}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(y_i) = \frac{\text{Var}(y_i)}{n} = \frac{\sigma_y^2}{n},$$

$$\text{sd}(\hat{\mu}_n) = \frac{\sigma_y}{\sqrt{n}}.$$

In this example the estimator is unbiased and the variance decreases with rate n as sample size increases.

- **Consistency of an estimator $\hat{\theta}_n$:**

For every $\epsilon > 0$ and $\delta > 0$ there exists an N such that

$$P\left(|\hat{\theta}_n - \theta| < \epsilon\right) > 1 - \delta \quad \text{for all } n > N.$$

Alternatively:

- $\lim_{n \rightarrow \infty} P\left(|\hat{\theta}_n - \theta| < \epsilon\right) = 1,$
- $\text{plim } \hat{\theta}_n = \theta,$
- $\hat{\theta}_n \xrightarrow{p} \theta.$

The “plim” notation stands for **probability limit**. This concept of convergence is usually denoted as convergence in probability or (weak) consistency. Some notes on calculation rules for the “plim” are given in Appendix C.3 in [Wooldridge \(2009\)](#).

A consistent estimator $\hat{\theta}_n$ has the properties

$$- \lim_{n \rightarrow \infty} E \left[\hat{\theta}_n \right] = \theta \text{ and}$$

$$- \lim_{n \rightarrow \infty} \text{Var} \left(\hat{\theta}_n \right) = 0.$$

If one of these conditions fails to hold, the estimator is called **inconsistent**. In general:

- **Weak law of large numbers (WLLN):**

For $y_i \sim i.i.d.$ with $-\infty < E[y_i] = \mu < \infty$, the mean estimator $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n y_i$ is weakly consistent, that is

$$\hat{\mu}_n \xrightarrow{p} \mu.$$

- Then we can consistently estimate the variance of i.i.d. random variables $w_i \sim i.i.d.(\mu_w, \sigma_w^2)$ with $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (w_i - \mu_w)^2$. Why?

- **But how can we derive the asymptotic probability distribution of the mean estimator $\hat{\mu}_n$?**

- **Monte Carlo Simulation (MC):**

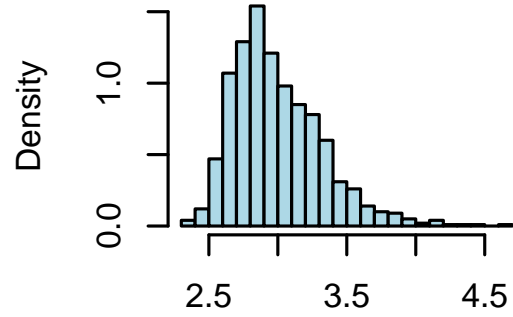
The R program `EOE_ws19_Emp_Beispiele.R`, line 559 following, allows us to iteratively draw $R = 1000$ samples of size n with elements $\{y_1, \dots, y_n\}$, where $y_i \sim i.i.d.(\nu, \sigma_y^2)$ with $\nu = 3$ and $\sigma_y^2 = 1$ and y_i is generated from (5.1). One frequently calls (5.1) the **data generating process (DGP)**. For every sample $\{y_1^r, y_2^r, \dots, y_n^r\}$ generated in this way, where $r = 1, \dots, 1000$, the mean estimator $\hat{\mu}^r = \frac{1}{n} \sum_{i=1}^n y_i^r$ is calculated and stored. After all R iterations, a histogram is calculated based on R estimates $\hat{\mu}^1, \hat{\mu}^2, \dots, \hat{\mu}^R$.

First, the results for the simulated moments:

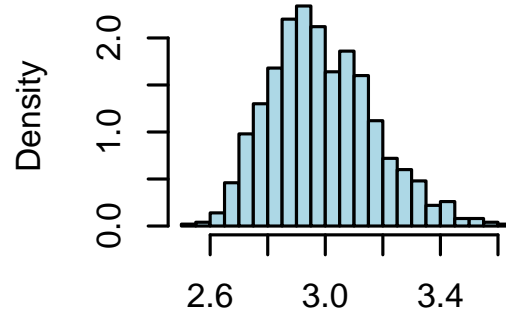
	Average of estimated means	Standard deviation	True standard deviation of MC DGP
$n = 10$	2.999717	0.323645	0.316228
$n = 30$	2.988812	0.180521	0.182574
$n = 50$	3.005385	0.148377	0.141421
$n = 100$	3.001922	0.098153	0.100000
$n = 500$	3.003529	0.045176	0.044721
$n = 1000$	3.000575	0.031675	0.031623

- The true moments are accurately estimated,
- and we can observe how the LLN works.

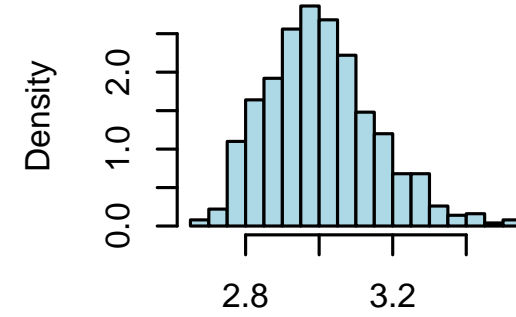
n = 10



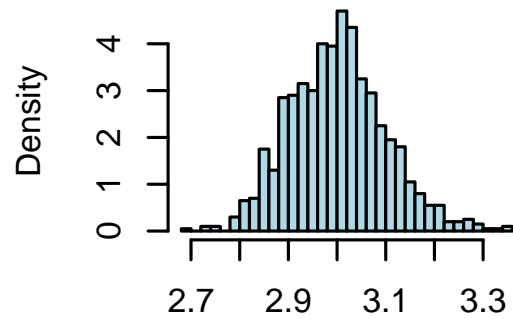
n = 30



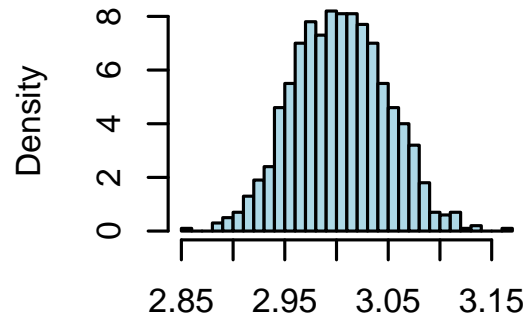
n = 50



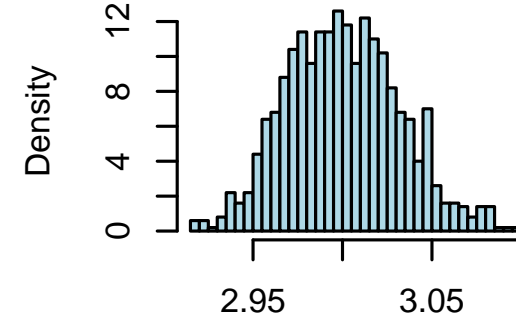
n = 100



n = 500



n = 1000



- Results for simulated distributions:
 - Right-skewness decreases with increase in sample size n .
 - A test for normality (Jarque-Bera-Test): null hypothesis of normal distribution cannot be rejected for large n .

Theoretical explanation of this phenomenon: a central limit theorem holds under certain (rather weak) conditions that is one of the most important tools in statistics!

- **Central limit theorem (CLT):**

For $y_i \sim i.i.d.(\mu, \sigma^2)$ with $0 < \sigma^2 < \infty$, $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n y_i$ is asymptotically normally distributed:

$$\sqrt{n} (\hat{\mu}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

.

- **Interpretation:** the larger the number of sample elements n , the more precise is the approximation of the exact distribution of $\hat{\mu}_n$ (see the MC results) by an exactly specified normal distribution. Hence the label **large sample distribution**.
- But how good is the asymptotic approximation for a given sample size n ?
 - * The CLT is not informative on this question, though we may get an answer by conducting MC simulations for certain cases or by using rather involved finite sample statistics.
 - * Experience: as the distribution of the y_i approaches the normal distribution, smaller and smaller n suffice for a very good approximation. In some cases even $n = 30$ is enough.

- Alternative notations ($\Phi(z)$ is the cumulative distribution function of the standard normal distribution):

$$\sqrt{n} \left(\frac{\hat{\mu}_n - \mu}{\sigma} \right) \xrightarrow{d} N(0, 1) \quad (5.2)$$

$$P \left(\sqrt{n} \left(\frac{\hat{\mu}_n - \mu}{\sigma} \right) \leq z \right) \longrightarrow \Phi(z) \quad (5.3)$$

$$\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} \underset{\sim}{\text{approx}} N(0, 1) \quad (5.4)$$

$$\hat{\mu}_n \underset{\sim}{\text{approx}} N \left(\mu, \frac{\sigma^2}{n} \right) \quad (5.5)$$

Notation: the mean estimator is **asymptotically normally distributed**.

- In large samples the standardized mean estimator is approximated by a standard normal distribution. Then, due to (5.4)

$$w_i \sim i.i.d.N(\mu, \sigma^2) \quad t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \sim N(0, 1)$$

$$w_i \sim i.i.d.(\mu, \sigma^2) \quad t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \stackrel{\text{approx}}{\sim} N(0, 1)$$

and it can be shown that

$$w_i \sim i.i.d.N(\mu, \sigma^2) \quad t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} \sim t_{n-k-1}$$

$$w_i \sim i.i.d.(\mu, \sigma^2) \quad t(w_1, \dots, w_n) = \frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} \stackrel{\text{approx}}{\sim} N(0, 1)$$

and we get the following (very convenient) result: **the (small sample) theory of t tests and confidence intervals for the mean estimator of i.i.d. variables holds approximately in large (enough) samples.**

- Hence the test results in our empirical exercise are still approximately valid!
- How about this concept of validity in a regression context?

5.2 Large Sample Inference for the OLS Estimator

- The OLS-estimator

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + \mathbf{W}\mathbf{u}$$

depends on \mathbf{X} or \mathbf{W} . Hence, for the OLS estimator to be consistent and asymptotically normal, certain conditions must hold for the regressor variables as $n \rightarrow \infty$. One of these conditions is that for all $i, l = 0, 1, \dots, k$ we have $\text{plim} \frac{1}{n} \sum_{i=1}^n x_{ij}x_{il} = E[x_jx_l] = a_{ij}$ or

$$\frac{1}{n} \mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{A}. \quad (5.6)$$

- **Asymptotic normality of the OLS estimator**

All necessary conditions for asymptotic normality are fulfilled if the standard assumptions MLR.1-MLR.5 hold true. Then (see a sketch of proof in Appendix E.4 in [Wooldridge \(2009\)](#)):

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \xrightarrow{d} N \left(0, \sigma^2 \mathbf{A} \right). \quad (5.7)$$

For the **(asymptotic) distributions of the t statistics** we get:

$$\text{MLR.1-MLR.6 } t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim N(0, 1)$$

$$\text{MLR.1-MLR.5 } t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \underset{\text{approx}}{\sim} N(0, 1)$$

and it can be shown that

$$\text{MLR.1-MLR.6 } t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} / (\text{SST}_j (1 - R_j^2))} \sim t_{n-k-1}$$

$$\text{MLR.1-MLR.5 } t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} / (\text{SST}_j (1 - R_j^2))} \underset{\text{approx}}{\sim} N(0, 1)$$

A frequent observation from many Monte Carlo simulations and empirical **practice** is that

- for small n one proceeds as in the case of normally distributed errors and uses the critical values of the t distribution:

$$\text{MLR.1-MLR.5 } t(\mathbf{X}, \mathbf{y}) = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} / (\text{SST}_j(1 - R_j^2))} \underset{\sim}{\text{approx}} t_{n-k-1}$$

- and analogously for the F statistic the critical values are determined from the F distribution.
- Note again: the critical values are valid only approximately, not exactly. Analogously, the p -values (calculated in R) are valid only approximately!

- **Conclusion:**

- For the calculation of test statistics and confidence intervals (exception: forecast intervals) we proceed as hitherto. However, all statistical results hold only as an approximation.
- If the assumption of homoskedasticity is violated, even the asymptotic results do not hold and models for heteroskedastic errors are required (with stronger assumptions for LLN and CLT), see Chapter 8.

Reading: Chapter 5 and Appendix C.3 in [Wooldridge \(2009\)](#).

6 Multiple Regression Analysis: Interpretation

6.1 Level and Log Models

Recall section 2.6 on level-level, level-log, log-level, log-log models. All the results remain valid in the multiple regression model in a ceteris-paribus analysis.

6.2 Data Scaling

- **Scaling the dependent variable:**

- Initial model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

- Variable transformation: $y_i^* = a \cdot y_i$ with scale factor a .

- New, transformed regression equation:

$$\underbrace{a\mathbf{y}}_{\mathbf{y}^*} = \mathbf{X} \underbrace{a\boldsymbol{\beta}}_{\boldsymbol{\beta}^*} + \underbrace{a\mathbf{u}}_{\mathbf{u}^*}$$

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{u}^* \quad (6.1)$$

– OLS-estimator for β^* in (6.1):

$$\begin{aligned}\hat{\beta}^* &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}^* \\ &= a (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = a\hat{\beta}.\end{aligned}$$

– Error variance:

$$Var(\mathbf{u}^*) = Var(a\mathbf{u}) = a^2 Var(\mathbf{u}) = a^2 \sigma^2 \mathbf{I}.$$

– Variance-covariance matrix:

$$Var(\hat{\beta}^*) = \sigma^{*2} (\mathbf{X}'\mathbf{X})^{-1} = a^2 \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = a^2 Var(\hat{\beta})$$

– t statistic:

$$t^* = \frac{\hat{\beta}^*_j - 0}{\sigma_{\hat{\beta}^*_j}} = \frac{a\hat{\beta}_j}{a\sigma_{\hat{\beta}_j}} = t.$$

- **Scaling explanatory variables:**

- Variable transformation: $\mathbf{X}^* = \mathbf{X} \cdot a$. New regression equation:

$$\mathbf{y} = \mathbf{X}a \cdot a^{-1}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{u}. \quad (6.2)$$

- OLS estimator for $\boldsymbol{\beta}^*$ in (6.2):

$$\begin{aligned} \hat{\boldsymbol{\beta}}^* &= \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1} \mathbf{X}^{*\prime}\mathbf{y} = \left(a^2\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'a\mathbf{y} \\ &= a^{-2}a \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} = a^{-1}\hat{\boldsymbol{\beta}}. \end{aligned}$$

- Result: The sole magnitude of β_j is no indicator for the relevance of the impact of the j -th regressor. One always has to take the scale of the variable into account.

- **Example:** In Section 2.3 a simple level-level model was estimated for imports on gdp. The parameter estimate $\hat{\beta}_{BIP} = 4.857 \cdot 10^{-03}$ appears very small. However, taking into account that

gdp is measured in dollars, this estimate is not small. Simply rescale gdp to millions of dollars with $a = 10^{-6}$ and you obtain $\hat{\beta}^*_{BIP} = 10^6 \cdot 4.857 \cdot 10^{-03} = 4857$.

- **Scaling of variables in logarithmic form**

just alters the constant β_0 since $\ln y^* = \ln ay = \ln a + \ln y$.

- **Standardized Coefficients:**

We just saw that it is not possible to deduce the relevance of explanatory variables from the magnitude of the corresponding coefficient. This is possible, however, if the regression is suitably standardized.

Deviation: First, consider the following sample regression model

$$y_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \dots + x_{ik}\hat{\beta}_k + \hat{u}_i, \quad (6.3)$$

and its representation after taking means over all n observations

$$\bar{y} = \hat{\beta}_0 + \bar{x}_1\hat{\beta}_1 + \dots + \bar{x}_k\hat{\beta}_k. \quad (6.4)$$

Then we calculate the difference between (6.4) and (6.3)

$$(y_i - \bar{y}) = (x_{i1} - \bar{x}_1)\hat{\beta}_1 + \dots + (x_{ik} - \bar{x}_k)\hat{\beta}_k + \hat{u}_i. \quad (6.5)$$

Finally, we divide equation (6.5) by the estimated standard deviation of y , say $\hat{\sigma}_y$, and expand every term on the right-hand side by the estimated standard deviations of the corresponding explanatory variables, say $\hat{\sigma}_{x_j}$, $j = 1, \dots, k$,

$$\frac{(y_i - \bar{y})}{\hat{\sigma}_y} = \frac{(x_{i1} - \bar{x}_1)}{\hat{\sigma}_y} \cdot \frac{\hat{\sigma}_{x_1}}{\hat{\sigma}_{x_1}} \hat{\beta}_1 + \dots + \frac{(x_{ik} - \bar{x}_k)}{\hat{\sigma}_y} \cdot \frac{\hat{\sigma}_{x_k}}{\hat{\sigma}_{x_k}} \hat{\beta}_k + \frac{\hat{u}_i}{\hat{\sigma}_y}.$$

Simple algebra gives

$$\underbrace{\frac{(y_i - \bar{y})}{\hat{\sigma}_y}}_{z_{i,y}} = \underbrace{\frac{(x_{i1} - \bar{x}_1)}{\hat{\sigma}_{x_1}}}_{z_{i,x_1}} \underbrace{\frac{\hat{\sigma}_{x_1} \hat{\beta}_1}{\hat{\sigma}_y}}_{\hat{b}_1} + \dots + \underbrace{\frac{(x_{ik} - \bar{x}_k)}{\hat{\sigma}_{x_k}}}_{z_{i,x_k}} \underbrace{\frac{\hat{\sigma}_{x_k} \hat{\beta}_k}{\hat{\sigma}_y}}_{\hat{b}_k} + \underbrace{\frac{\hat{u}_i}{\hat{\sigma}_y}}_{\xi_i}.$$

In the literature the transformed variables $z_{i,y}$ and $z_{i,x_1}, \dots, z_{i,x_k}$ are usually denoted as ***z-scores***.

In **compact notation** we get

$$z_{i,y} = z_{i,x_1} \hat{b}_1 + \dots + z_{i,x_k} \hat{b}_k + \xi_i.$$

where \hat{b}_j are denoted as **standardized coefficients** (or simply **beta coefficients**).

The magnitudes of the standardized coefficients can be compared to each other. Hence, the explanatory variable with the largest parameter $\hat{\beta}_j$ has the relatively largest impact on the dependent variable.

Interpretation: a one standard deviation increase in x_j changes y by \hat{b}_j standard deviations.

Standardized coefficients can be calculated in SPSS (see Example 6.1 in [Wooldridge \(2009\)](#)).

6.3 Dealing with Nonlinear or Transformed Regressors

- **Further details on logarithmic variables:**

Consider the following log-level regression model

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad (6.6)$$

where x_2 is a dummy variable (it is either equal to 0 or 1).

– How can we determine the **exact impact** of x_2 , that is, how should we interpret β_2 ? From (6.6) follows

$$y = e^{\ln y} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + u} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \cdot e^u$$

and for the conditional expectation

$$E[y|x_1, x_2] = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \cdot E[e^u|x_1, x_2]. \quad (6.7)$$

Inserting the two possible values of x_2 into (6.7) delivers

$$\begin{aligned} E[y|x_1, x_2 = 0] &= e^{\beta_0 + \beta_1 x_1} \cdot E[e^u|x_1, x_2] \\ E[y|x_1, x_2 = 1] &= e^{\beta_0 + \beta_1 x_1} \cdot E[e^u|x_1, x_2] \cdot e^{\beta_2} \\ &= E[y|x_1, x_2 = 0] \cdot e^{\beta_2}. \end{aligned}$$

- Thus, if $E[e^u|x_1, x_2]$ is constant (for x_2), the **relative mean change of the dependent variable with respect to a unit change in x_2** is equal to

$$\begin{aligned} \frac{\Delta E[y|x_1, x_2]}{E[y|x_1, x_2 = 0]} &= \frac{E[y|x_1, x_2 = 1] - E[y|x_1, x_2 = 0]}{E[y|x_1, x_2 = 0]} \\ &= \frac{E[y|x_1, x_2 = 0] \cdot e^{\beta_2} - E[y|x_1, x_2 = 0]}{E[y|x_1, x_2 = 0]} \\ &= e^{\beta_2} - 1. \end{aligned}$$

This implies

$$\% \Delta E[y|x_1, x_2] = 100 \left(e^{\beta_2} - 1 \right).$$

– In the general case of k regressors:

$$\% \Delta E[y|x_1, x_2, \dots, x_k] = 100 \left(e^{\beta_j \Delta x_j} - 1 \right). \quad (6.8)$$

Obviously (6.8) represents the **exact partial effect**, whereas the interpretation as an approximate semi-elasticity may be rather crude in some cases.

– **Trade Example Continued** (from Section 4.8 and specifically from Section 4.4):

For Model 3 we obtained the sample regression

$$\begin{aligned} \text{LOG}(\text{TRADE}_0\text{D}_0) = & 2.74104 + 0.9406645 \cdot \text{LOG}(\text{WDI_GDPUSDCR}_0) \\ & - 0.9703183 \cdot \text{LOG}(\text{CEPII_DIST}) + 0.5072497 \cdot \text{EBRD_TFES}_0 + \text{RESIDUAL} \end{aligned}$$

Recall that CEPII_COMCOL_REV denotes a dummy variable.

- * The approximate interpretation of $\hat{\beta}_{openess}$ is that 1 unit change changes *imports* on average by $100\hat{\beta}_{openess} = 50.7\%$.
- * The exact partial effect is $100 \left(e^{\hat{\beta}_{comcol}} - 1 \right) = 66.1\%$ and thus substantially larger.
- * Of course, the difference between the approximate and exact effect are even larger if $\hat{\beta}$ is further away from zero.

- **Models with quadratic regressors:**

- For example, consider the multiple regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + u.$$

The marginal effect of a change in x_2 on the conditional expectation of y is equal to

$$\frac{\partial E[y|x_1, x_2]}{\partial x_2} = \beta_2 + 2\beta_3 x_2.$$

Therefore a change of Δx_2 in x_2 changes ceteris paribus the dependent variable y on average by

$$(\beta_2 + 2\beta_3 x_2)\Delta x_2.$$

Clearly, this effect depends on the level of x_2 (and an interpretation of β_2 alone does not make any sense!).

- In some empirical applications regressor variables are considered using quadratics and logarithms, in order to approximate a **non-linear regression function**.

Example: we can approximate non-constant elasticities using the model

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 \ln x_2 + \beta_3 (\ln x_2)^2 + u.$$

Then the elasticity of y with respect to x_2 equals

$$\beta_2 + 2\beta_3 \ln x_2$$

and is constant if and only if $\beta_3 = 0$.

– Trade Example Continued:

So far we only considered multiple regression models that are log-log or log-level in the original variables.

Now consider a further specification for modeling imports where a log regressors also enters as square.

Model 5:

$$\ln(\textit{imports}) = \beta_0 + \beta_1 \ln(\textit{gdp}) + \beta_2 (\ln(\textit{gdp}))^2 + \beta_3 \ln(\textit{distance}) \\ + \beta_4 \textit{openess} + \beta_5 \textit{area} + u.$$

Using the previous result, the elasticity of *imports* with respect to *gdp* is

$$\beta_1 + 2\beta_2 \ln(\textit{gdp}). \quad (6.9)$$

Estimation of Model 5 delivers:

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

Multiple R-squared: 0.9155, Adjusted R-squared: 0.9056

F-statistic: 93.12 on 5 and 43 DF, p-value: < 2.2e-16

Comparing the AIC, HQ, and SC of Model 5 with those of Models 1 to 4, see Section 4.4, one finds that Model 5 exhibits the lowest values throughout. In addition, the (approximate) p -value of β_2 is 0.03523 and the quadratic term is statistically significant at the 5% significance level.

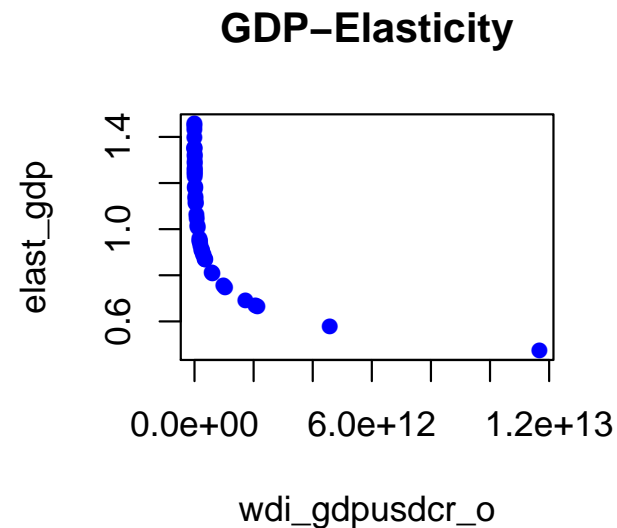
This provides also evidence for a nonlinear elasticity. Inserting the parameter estimates into (6.9) delivers

$$\eta(BIP) = 3.90881 - 0.05711 \ln(BIP).$$

One may plot the elasticity $\eta(gdp)$ versus gdp for each observed value of gdp . In R this can be done by a little program

R-Code

```
# Modell 5:
model_5 <- lm(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))
# Generiere die Elastizitäten für verschiedene BIPs
elast_gdp <- model_5$coef[2] + 2* model_5$coef[3]*log(wdi_gdpusdcr_o)
# Erstelle Scatterplot
plot(wdi_gdpusdcr_o, elast_gdp, pch = 16, col = "blue", main = "GDP-Elasticity")
```



The import elasticity with respect to gdp is much larger for small economies in terms of gdp than for large economies.

Warning: Nonlinearities are sometimes due to missing variables. Can you think of any control variables left out that should be included in Model 5?

- **Interactions:**

Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2 x_1 + u.$$

The marginal effect of a change in x_2 is given by

$$\Delta E[y|x_1, x_2] = (\beta_2 + \beta_3 x_1) \Delta x_2.$$

Hence, in this case the marginal effect also depends on the level of x_1 !

6.4 Regressors with Qualitative Data

Dummy variables or binary variables

A **binary variable** can take exactly **two** different values and allows to describe **two** qualitatively different states.

Examples: female vs. male, employed vs. unemployed, etc.

- In general these values are coded as 0 and 1. This allows for a very easy and straightforward interpretation. Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \delta D + u,$$

where D equals 0 or 1.

- **Interpretation** (well known by now):

$$\begin{aligned} E[y|x_1, \dots, x_{k-1}, D = 1] - E[y|x_1, \dots, x_{k-1}, D = 0] = \\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \delta \\ - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1}) = \delta \end{aligned}$$

The coefficient of a dummy variable is equal to an intercept shift of size δ in the case $D = 1$. All slope parameters β_i , $i = 1, \dots, k - 1$ remain *unchanged*.

- **Wage Example Continued:**

- Question of interest: Do females earn significantly less than males?
- Data: a sample of $n = 526$ U.S. workers obtained in 1976. (Source: Examples 2.4, 7.1 in [Wooldridge \(2009\)](#)).

- * wage in dollars per hour,
- * educ: years of schooling of each worker,
- * exper: years of professional experience,
- * tenure: years of employment in current firm,
- * female: dummy=1 if female, dumm=0 otherwise.

```
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

- **Note:** In order to be able to interpret the coefficients of dummy variables one has to know the reference group. The reference group is given by the group for which the dummy equals zero.
- **Prediction:** How much earns a woman with 12 years of schooling, 10 years of experience, and 1 year tenure? (Or course, you can insert any other numbers here.)

$$\begin{aligned} E[\ln(wage) | female = 1, educ = 12, exper = 10, tenure = 1] \\ &= 0.4167 - 0.2965 \cdot 1 + 0.0802 \cdot 12 + 0.0294 \cdot 10 \\ &\quad - 0.0006 \cdot (10^2) + 0.0317 \cdot 1 - 0.0006 \cdot (1^2) \\ &= 1.35 \end{aligned}$$

Thus, the expected hourly wage is **approximately** $\exp(1.35) = 3.86$ US dollar.

- We already know that in case of a log-level model the expected value of y given the regressors x_1, x_2 is given by

$$E[y|x_1, x_2] = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \cdot E[e^u|x_1, x_2].$$

The true value of $E[e^u|x_1, x_2]$ depends on the probability distribution of u .

It holds that: If u is normally distributed with variance σ^2 , then

$$E[e^u|x_1, x_2] = e^{E[u|x_1, x_2] + \sigma^2/2}.$$

The precise prediction is therefore

$$E[y|x_1, x_2] = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + E[u|x_1, x_2] + \sigma^2/2}.$$

The exact prediction of the desired hourly wage is

$$\begin{aligned}
 & E[wage | female = 1, educ = 12, exper = 10, tenure = 1] \\
 & = \exp(0.4167 - 0.2965 \cdot 1 + 0.0802 \cdot 12 + 0.02943 \cdot 10 \\
 & \quad - 0.0006 \cdot (10^2) + 0.0317 \cdot 1 - 0.0006 \cdot (1^2) + 0.3998^2/2) \\
 & = 4.18.
 \end{aligned}$$

Thus, the precise value of the mean hourly wage for the specified person is about 4.18\$ and thus 30 Cent larger than the approximate value.

- The parameter δ corresponds to the difference between the log income of female and male workers *keeping everything else constant* (e.g. years of schooling, experience, etc.).

Question: How large is the exact wage difference?

Answer: $100(e^{-0.2965} - 1) = 34.51\%$.

Note that *ceteris paribus* analysis is much more informative than the comparison of the unconditional means of male and female wages. Assuming normal errors one has

$$\frac{E[wage_f] - E[wage_m]}{E[wage_m]} = \frac{e^{E[\ln(wage_f)] + \sigma_f^2/2} - e^{E[\ln(wage_m)] + \sigma_m^2/2}}{e^{E[\ln(wage_m)] + \sigma_m^2/2}}.$$

Inserting estimates one obtains

$$\frac{e^{1.416 + 0.44^2/2} - e^{1.814 + 0.53^2/2}}{e^{1.814 + 0.53^2/2}} = -0.3570,$$

which, by the way, is very similar to inserting estimates for $(E[wage_f] - E[wage_m])/E[wage_m]$ leading to -0.3538.

Females earn 36% less than males if one does not control for other effects.

Several subgroups

- **Example:** A worker is female or male and married or unmarried

⇒ 4 subgroups:

1. female and not married
2. female and married
3. male and not married
4. male and married

How to proceed:

- Choose one subgroup to be the **reference group**, for example: female and not married
- Define dummy variables for the other subgroups. For example, in R:

```
* femmarr <- female * married
* malesing <- (1 - female) * (1 - married)
* malemarr <- (1 - female) * married
```

```
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 for Interpretation:

- Married women earn about 8.8% less than unmarried women. However, this effect is only significant at the 10% significance level (for a two-sided test).
- The wage difference between married men and women is about $32.3 - (-8.8) = 41.1\%$. A t test cannot be directly applied. (Solution: Choose a new reference group with one of the two subgroups as the reference group.)

Remarks:

- Using dummies for all subgroups is not recommended since then differences with respect to the ref. group cannot be tested directly.
- If you use dummies for all subgroups you cannot include a constant. Otherwise MLR.3 is violated. Why?

• Using ordinal information in regression

Example: Ranking of universities

The quality difference between ranks 1 and 2 and ranks 11 and 12, respectively, may be dramatically different. Hence, ranks should **not** be used as regressors. Instead, we have to assign a dummy variable D_j for all but one (the “reference category”) of the universities, inducing several new parameters which have to be estimated. (Therefore we may split in the trade example the variable *openess*

in several dummy variables.)

Note: Then, the coefficient of a dummy variable D_j denotes the intercept shift between university j and the reference university.

Sometimes there are too many ranks and hence too many parameters to be estimated. Then it proves useful to group the data, e.g., ranks 1-10, 11-20, etc.

Interactions and Dummy Variables

- **Interactions between dummy variables:**

- May be used to define sub-groups (e.g., married males).
- Note that a useful interpretation and comparison of sub-group effects crucially depends on a correct setup of dummies. For example, let us include the dummies *male* and *married* and their interaction in a wage equation

$$y = \beta_0 + \delta_1 \text{male} + \delta_2 \text{married} + \delta_3 \text{male} \cdot \text{married} + \dots$$

Then, a comparison between male-married and male-single is given by

$$\begin{aligned} E[y|male = 1, married = 1] - E[y|male = 1, married = 0] \\ = \beta_0 + \delta_1 + \delta_2 + \delta_3 + \dots - (\beta_0 + \delta_1 + \dots) = \delta_2 + \delta_3 \end{aligned}$$

- **Interactions between dummies and quantitative variables:**

- Allows different slope parameters for different groups

$$y = \beta_0 + \beta_1 D + \beta_2 x_1 + \beta_3 (x_1 \cdot D) + u.$$

Note: here β_1 denotes the difference between both groups only for the case $x_1 = 0$.

If $x_1 \neq 0$, then this difference is equal to

$$\begin{aligned} & E[y|D = 1, x_1] - E[y|D = 0, x_1] \\ &= \beta_0 + \beta_1 \cdot 1 + \beta_2 x_1 + \beta_3(x_1 \cdot 1) - (\beta_0 + \beta_2 x_1) \\ &= \beta_1 + \beta_3 x_1 \end{aligned}$$

Even if β_1 is negative, the total effect may be positive!

— Wage Example Continued:

```
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

Are returns to schooling sensitive to gender?

- **Testing for differences between groups**

- Can be done with F tests.
- **Chow Test:** Allows to test whether there is a difference between groups in a sense that there may be group specific intercepts and/or (at least one) slope parameter.

Illustration:

$$y = \beta_0 + \beta_1 D + \beta_2 x_1 + \beta_3 (x_1 \cdot D) + \beta_4 x_2 + \beta_5 (x_2 \cdot D) + u. \quad (6.10)$$

Pair of hypotheses:

$$H_0 : \beta_1 = \beta_3 = \beta_5 = 0 \quad \text{vs.}$$

$$H_1 : \beta_1 \neq 0 \text{ and/or } \beta_3 \neq 0 \text{ and/or } \beta_5 \neq 0$$

Application of F tests:

- * Estimate the regression equation for each group l

$$y = \beta_{0l} + \beta_{2l}x_1 + \beta_{4l}x_2 + u, \quad l = 1, 2,$$

and calculate SSR_1 and SSR_2 .

- * Then estimate this regression for both groups together and calculate SSR.

- * Compute the F statistic

$$F = \frac{SSR - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \frac{n - 2(k + 1)}{(k + 1)}$$

where the degrees of freedom for the F distribution are equal to $k + 1$ and $n - 2(k + 1)$.

Reading: Chapter 6 (without Section 6.4) and Chapter 7 (without Sections 7.5 und 7.6) in [Wooldridge \(2009\)](#).

7 Multiple Regression Analysis: Prediction

7.1 Prediction and Prediction Error

- Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, i.e.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \quad 1 \leq i \leq n.$$

- We search for a **predictor** \hat{y}_0 for y_0 given x_{01}, \dots, x_{0k} .
- Define the **prediction error**

$$y_0 - \hat{y}_0.$$

- We assume that MLR.1 to MLR.5 hold for the **prediction sample** (\mathbf{x}_0, y_0) . Then

$$y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} + u_0 \quad (7.1)$$

and

$$E[u_0 | x_{01}, \dots, x_{0k}] = 0,$$

so that

$$E[y_0 | x_{01}, \dots, x_{0k}] = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} = \mathbf{x}'_0 \boldsymbol{\beta},$$

where $\mathbf{x}'_0 = (1, x_{01}, \dots, x_{0k})$.

MLR.4 guarantees that for *known* parameters the predictions are unbiased. Then, the prediction is, loosely speaking, correct on average (if averaged over many samples).

It can be shown that the conditional expectation is optimal in the sense of minimizing the mean squared prediction error.

- In practice, the true regression coefficients β_j , $j = 0, \dots, k$, are unknown. Inserting the OLS estimators $\hat{\beta}_j$ gives

$$\hat{y}_0 = \hat{E}[y_0 | x_{01}, \dots, x_{0k}] = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_k x_{0k}.$$

Using compact notation the **prediction rule** is:

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}} \quad (7.2)$$

- This prediction rule only makes sense if (y_0, \mathbf{x}'_0) belongs to the population as well. Otherwise the population regression model is not valid for (y_0, \mathbf{x}'_0) and the prediction based on the estimated version possibly strongly misleading.

- General **decomposition** of the **prediction error**

$$\begin{aligned}
 \hat{u}_0 &= y_0 - \hat{y}_0 && (7.3) \\
 &= \underbrace{(y_0 - E[y_0|\mathbf{x}_0])}_{\text{unavoidable error } v_0} \\
 &\quad + \underbrace{(E[y_0|\mathbf{x}_0] - \mathbf{x}'_0\boldsymbol{\beta})}_{\text{possible specification error}} \\
 &\quad + \underbrace{(\mathbf{x}'_0\boldsymbol{\beta} - \mathbf{x}'_0\hat{\boldsymbol{\beta}})}_{\text{estimation error}}
 \end{aligned}$$

- If MLR.1 and MLR.4 are correct for the population and if the prediction sample also belongs to the population, then the specification error is zero. Then $v_0 = u_0$ in (7.1).
- If the estimator is consistent, $\text{plim } \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, then the estimation error becomes negligible in large samples.

– Using the OLS estimator, the estimation error is

$$\begin{aligned}
 \mathbf{x}'_0\boldsymbol{\beta} - \mathbf{x}'_0\hat{\boldsymbol{\beta}} &= \mathbf{x}'_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
 &= \mathbf{x}'_0\boldsymbol{\beta} - \mathbf{x}'_0 \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \right) \\
 &= \mathbf{x}'_0\boldsymbol{\beta} - \mathbf{x}'_0 \left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \right) \\
 &= -\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \tag{7.4}
 \end{aligned}$$

Thus, the estimation error only depends on the estimation sample.

– The OLS prediction error under MLR.1 to MLR.5 is given by (using (7.3) and (7.4)):

$$\begin{aligned}
 \hat{u}_0 &= u_0 + \mathbf{x}'_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
 &= u_0 - \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \tag{7.5}
 \end{aligned}$$

- **Variance of the prediction error:**

- **Extension of Assumption MLR.2** (Random Sampling):

u_0 and \mathbf{u} are uncorrelated.

- Conditional variance of (7.5) given \mathbf{X} and \mathbf{x}_0 :

$$\begin{aligned} Var(\hat{u}_0|\mathbf{X}, \mathbf{x}_0) &= Var(u_0|\mathbf{X}, \mathbf{x}_0) + Var\left(\mathbf{x}'_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|\mathbf{X}, \mathbf{x}_0\right) \\ &= \sigma^2 + \mathbf{x}'_0 Var(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|\mathbf{X})\mathbf{x}_0 \\ &= \sigma^2 + \mathbf{x}'_0\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \end{aligned}$$

or

$$Var(\hat{u}_0|\mathbf{X}, \mathbf{x}_0) = \sigma^2 \left(1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right). \quad (7.6)$$

- Relevant in practice: **Estimated variance of the prediction error**

$$\widehat{Var}(\hat{u}_0|\mathbf{X}, \mathbf{x}_0) = \hat{\sigma}^2 \left(1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right).$$

- **Prediction interval:** A prediction interval is (given an a priori chosen confidence probability $1 - \alpha$) for the multiple regression model given by

$$\left[\hat{y}_0 - t_{n-k-1} \sqrt{\widehat{Var}(\hat{u}_0 | \mathbf{X}, \mathbf{x}_0)}, \hat{y}_0 + t_{n-k-1} \sqrt{\widehat{Var}(\hat{u}_0 | \mathbf{X}, \mathbf{x}_0)} \right].$$

Notes:

- Derivation and structure are analogous to the case of confidence intervals for the parameter estimates.
- Prediction intervals are in contrast to confidence intervals even in large samples only valid if the prediction errors are normally distributed. This is because there is no averaging of the true prediction error u_0 as it occurs for $\hat{\beta} - \beta = \mathbf{W}\mathbf{u}$ due to the central limit theorem.

7.2 Statistical Properties of Linear Predictions

Apparently the prediction rule is linear (in y) since

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}} = \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}.$$

Gauss-Markov property of linear prediction

If $\hat{\boldsymbol{\beta}}$ is the BLU estimator for $\boldsymbol{\beta}$, then

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}$$

is the BLU prediction rule. Among all linear prediction rules with a mean prediction error of zero it exhibits the smallest prediction error variance.

Reading: Section 6.4 in [Wooldridge \(2009\)](#).

8 Multiple Regression Analysis: Heteroskedasticity

- In this chapter Assumptions MLR.1 through MLR.4 continue to hold.
- If MLR.5 fails to hold such that

$$\text{Var}(u_i | x_{i1}, \dots, x_{ik}) = \sigma_i^2 \neq \sigma^2, \quad i = 1, \dots, n,$$

the errors of the regression model exhibit heteroscedasticity. More precisely (instead of MLR.5) we have

– Assumption GLS.5: Heteroskedasticity

$$\begin{aligned} \text{Var}(u_i | x_{i1}, \dots, x_{ik}) &= \sigma_i^2(x_{i1}, \dots, x_{ik}) \\ &= \sigma^2 h(x_{i1}, \dots, x_{ik}) = \sigma^2 h_i, \quad i = 1, \dots, n. \end{aligned}$$

The error variance of the i -th sample observation σ_i^2 is a function $h(\cdot)$ of the regressors.

• Examples:

- The variance of net rents depends on the size of the flat.
- The variance of consumption expenditures depends on the level of income.
- The variance of log hourly wages depends on years of education.

- The **covariance matrix of the errors** of the regression is given by:

$$\text{Var}(\mathbf{u}|\mathbf{X}) = E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \begin{pmatrix} \sigma^2 h_1 & 0 & \cdots & 0 \\ 0 & \sigma^2 h_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 h_n \end{pmatrix} = \sigma^2 \underbrace{\begin{pmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_n \end{pmatrix}}_{\Psi}.$$

Thus, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \boldsymbol{\Psi}, \quad (8.1)$$

which will be referred to as the **original model in matrix notation**.

- When estimating models with heteroskedastic errors three cases have to be distinguished:
 1. Function $h(\cdot)$ is known, see Section 8.3.
 2. Function $h(\cdot)$ is only partially known, see Section 8.4.
 3. Function $h(\cdot)$ is completely unknown, see Section 8.2.

8.1 Consequences of Heteroskedasticity for OLS

- The OLS estimator is **unbiased** and **consistent**.
- **Variance of the OLS estimator** in the presence of heteroskedastic errors (compare Section 3.4.2):

From $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ it can be derived that

$$\begin{aligned}
 \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= E \left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)' | \mathbf{X} \right] \\
 &= E \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{u} \mathbf{u}' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X} \right] \\
 &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \underbrace{E [\mathbf{u} \mathbf{u}' | \mathbf{X}]}_{\sigma^2 \boldsymbol{\Psi}} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\
 &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \sigma^2 \boldsymbol{\Psi} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}. \tag{8.2}
 \end{aligned}$$

- Note that with homoskedastic errors one has $\boldsymbol{\Psi} = \mathbf{I}$. Then (8.2) yields the usual OLS covariance matrix, namely $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- **If heteroskedasticity is present, using the usual covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is misleading** and leads to faulty inference.

- The problem with using (8.2) directly is that Ψ is unknown. The next section introduces an appropriate estimator.
- Even if Ψ is known, OLS is not the best linear unbiased estimator, and thus **not efficient**. One has to use the GLS estimator instead, see Section 8.3.

8.2 Heteroskedasticity-Robust Inference after OLS

- **Derivation of heteroskedasticity-robust standard errors**

Let $\mathbf{x}'_i = (1, x_{i1}, \dots, x_{ik})$. Note that the middle term in the variance-covariance matrix (8.2) with dimension $(k+1) \times (k+1)$ can be written as

$$\mathbf{X}'\sigma^2\Psi\mathbf{X} = \sum_{i=1}^n \sigma^2 h_i \mathbf{x}_i \mathbf{x}'_i.$$

Because $E[u_i^2 | \mathbf{X}] = \sigma^2 h_i$, one can estimate $\sigma^2 h_i$ by the “one observation average” u_i^2 . Of course this is not a good estimator but for the present purpose it is doing well enough. Since u_i is not known, one takes the residual \hat{u}_i .

Hence one can estimate the covariance matrix (8.2) of the OLS estimator in presence of heteroskedasticity by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \right) (\mathbf{X}'\mathbf{X})^{-1}. \quad (8.3)$$

- Comments:

- Standard errors obtained from (8.3) are called **heteroskedasticity-robust standard errors** or also **White standard errors** named after Halbert White, an econometrician at the University of California in San Diego.
- For single $\hat{\beta}_j$ heteroskedasticity-robust standard errors can be smaller or larger than the usual OLS standard errors.

- If heteroskedasticity-robust standard errors are used, it can be shown that the OLS estimator $\hat{\beta}$ has no longer a known finite sample distribution. However, it is **asymptotically normally distributed**. Thus, critical values and p -values remain approximately valid if (8.3) is used.
- The OLS estimator with White standard errors is **unbiased** and **consistent** since MLR.1 to MLR.4 are unaffected by heteroskedasticity.
- However, the OLS estimator is **not efficient**. Efficient estimators will be presented in the next sections.

8.3 The General Least Squares (GLS) Estimator

- Original model (8.1):

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i, \quad (8.4)$$

$$\text{Var}(u_i | x_{i1}, \dots, x_{ik}) = \sigma^2 h(x_{i1}, \dots, x_{ik}) = \sigma^2 h_i.$$

- **Basic idea:** Weighted estimation of (8.4):

Transformation of the initial model to a model that satisfies all assumptions, including MLR.5. This is achieved by kind of standardizing the regression error u_i . This amounts to dividing u_i and thus the whole regression equation (8.4) by the square root of h_i :

$$\underbrace{\frac{y_i}{\sqrt{h_i}}}_{y_i^*} = \beta_0 \underbrace{\frac{1}{\sqrt{h_i}}}_{x_{i0}^*} + \beta_1 \underbrace{\frac{x_{i1}}{\sqrt{h_i}}}_{x_{i1}^*} + \dots + \beta_k \underbrace{\frac{x_{ik}}{\sqrt{h_i}}}_{x_{ik}^*} + \underbrace{\frac{u_i}{\sqrt{h_i}}}_{u_i^*}.$$

The resulting model is

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \dots + \beta_k x_{ik}^* + u_i^*. \quad (8.5)$$

Note: For the transformed error u_i^* we have

$$\begin{aligned} \text{Var}(u_i^* | x_{i1}, \dots, x_{ik}) &= \text{Var} \left(\frac{u_i}{\sqrt{h_i}} \middle| x_{i1}, \dots, x_{ik} \right) \\ &= E \left[\frac{u_i^2}{h_i} \middle| x_{i1}, \dots, x_{ik} \right] \\ &= \frac{1}{h_i} E[u_i^2 | x_{i1}, \dots, x_{ik}] = \frac{1}{h_i} \sigma^2 h_i = \sigma^2. \end{aligned}$$

Result: We have transformed the original regression (8.4) in such a way that the homoskedasticity assumption MLR.5 holds for the resulting regression model (8.5).

- Therefore the OLS estimator based on the transformed model (8.5) has all desirable properties: **BLU** (best linear unbiased).
- The OLS estimator of the transformed model (8.5) is based on the minimization of a *weighted* sum of squared residuals

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2 / h_i.$$

Therefore, it is called a **weighted least squares** (WLS) procedure. Note in its current form it requires that $h(\cdot)$ is known.

- The transformed model does not contain a constant term if $\sqrt{h_i}$ is not identical to one of the regressors in model (8.4).
- Next we derive the transformed model in matrix notation.

- Explicit statement of \mathbf{y}^* , \mathbf{X}^* , and \mathbf{u}^* in matrix notation:

$$\underbrace{\begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{pmatrix}}_{\mathbf{y}^*} = \underbrace{\begin{pmatrix} h_1^{-1/2} & 0 & \cdots & 0 \\ 0 & h_2^{-1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_n^{-1/2} \end{pmatrix}}_{\mathbf{P}} \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}}$$

$$\underbrace{\begin{pmatrix} x_{10}^* & x_{11}^* & \cdots & x_{1k}^* \\ x_{20}^* & x_{21}^* & \cdots & x_{2k}^* \\ \vdots & \vdots & & \vdots \\ x_{n0}^* & x_{n1}^* & \cdots & x_{nk}^* \end{pmatrix}}_{\mathbf{X}^*} = \mathbf{P} \cdot \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{\mathbf{X}}, \quad \underbrace{\begin{pmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_n^* \end{pmatrix}}_{\mathbf{u}^*} = \mathbf{P} \cdot \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}}_{\mathbf{u}}$$

- For the transformation matrix \mathbf{P} it holds that

$$\mathbf{P}'\mathbf{P} = \mathbf{\Psi}^{-1}$$

and hence

$$E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{\Psi} = \sigma^2(\mathbf{P}'\mathbf{P})^{-1}.$$

- Therefore, the **transformed model (8.5) in matrix notation** is given by

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{u},$$

or

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*, \quad E[\mathbf{u}^*(\mathbf{u}^*)'|\mathbf{X}^*] = \sigma^2\mathbf{I}. \quad (8.6)$$

- Obviously (8.6) is obtained by multiplying the original model (8.1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by the transformation matrix \mathbf{P} from the left.
- What is the explicit formula for the OLS estimator in terms of the transformed model (8.6) and the original model (8.1)?

GLS (generalized least squares) estimator

- OLS estimation of (8.6) yields

$$\begin{aligned}\hat{\beta}_{GLS} &= (\mathbf{X}^{*'}\mathbf{X}^*)^{-1} \mathbf{X}^{*'}\mathbf{y}^* \\ &= ((\mathbf{PX})'\mathbf{PX})^{-1} (\mathbf{PX})'\mathbf{Py} \\ &= (\mathbf{X}'\mathbf{P}'\mathbf{PX})^{-1} \mathbf{X}'\mathbf{P}'\mathbf{Py}\end{aligned}$$

and therefore

$$\hat{\beta}_{GLS} = \left(\mathbf{X}'\Psi^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\Psi^{-1}\mathbf{y}. \quad (8.7)$$

$\hat{\beta}_{GLS}$ in (8.7) is called **generalized least squares estimator** or **GLS estimator**.

In case of heteroskedasticity Ψ is a diagonal matrix and each of the n observations is weighted by $1/\sqrt{h_i}$.

- **Properties** for known $h(\cdot)$:

Under MLR.1 to MLR.4 and GLS.5 the GLS-estimator $\hat{\beta}_{GLS}$

- is **unbiased** and **consistent**,

- is **BLUE** (best linear unbiased), and thus **efficient**,

- has variance-covariance matrix $Var(\hat{\beta}_{GLS}|\mathbf{X}) = \sigma^2 (\mathbf{X}'\Psi^{-1}\mathbf{X})^{-1}$,

- is unbiased and consistent even if Ψ is misspecified since Ψ is a function of \mathbf{X} and not of \mathbf{u} and thus

$$E[\hat{\beta}_{GLS} - \beta|\mathbf{X}] = (\mathbf{X}'\Psi^{-1}\mathbf{X})^{-1} \mathbf{X}'\Psi^{-1}E[\mathbf{u}|\mathbf{X}] = \mathbf{0}.$$

As a consequence, OLS is inefficient since OLS and GLS are both linear estimators. OLS variances are larger than or equal to those of the GLS estimator. This can be shown using matrix algebra.

- Analogously to MLR.6 in Section 4.2 above, we assume
 - **Assumption GLS.6: Normal Distribution**

$$u_i | \mathbf{x}_i \sim N(0, \sigma^2 h_i), \quad i = 1, \dots, n,$$

which, together with MLR.2 (Random Sampling) implies the **multivariate normal distribution**

$$\mathbf{u} | \mathbf{X} \sim N \left(\mathbf{0}, \sigma^2 \mathbf{\Psi} \right).$$

Note GLS.6 implies that u_i given \mathbf{x}_i is independently but not identically distributed since the variance changes with i . (The covariances have not changed. They are zero due to MLR.2.)

All **test statistics** based on the transformed model (8.6) and appropriately modified for the original model (8.1) exhibit the **exact distributions** of Chapter 4 (normal, t , F).

- **Frequent problem in practice:** h_i is not known. In this case, the feasible GLS estimator has to be used \longrightarrow Case 2.

8.4 Feasible Generalized Least Squares (FGLS)

- In general, the variance function h_i is not known and has to be estimated. Frequently neither the relevant factors nor the functional relationship are known.
- Hence, one needs a specification that flexibly captures a large range of possibilities, e.g.

$$h_i = h(x_{i1}, \dots, x_{ik}) = \exp(\delta_1 x_{i1} + \dots + \delta_k x_{ik})$$

and thus

$$\text{Var}(u_i | x_{i1}, x_{i2}, \dots, x_{ik}) = \sigma^2 h_i = \sigma^2 \exp(\delta_1 x_{i1} + \dots + \delta_k x_{ik}).$$

Remark: On pp. 282, [Wooldridge \(2009\)](#) considers in h_i additionally the factor $\exp \delta_0$. As this factor is constant, it can also be captured by σ^2 .

- How can one estimate the unknown parameters $\delta_1, \dots, \delta_k$?

Standardizing u_i delivers $v_i = u_i/(\sigma\sqrt{h_i})$ with $E[v_i|\mathbf{X}] = 0$ and $Var(v_i|\mathbf{X}) = 1$. Therefore $u_i = \sigma\sqrt{h_i}v_i$ and

$$u_i^2 = \sigma^2 h_i v_i^2, \quad i = 1, \dots, n.$$

Taking logarithms leads to

$$\begin{aligned} \ln u_i^2 &= \ln \sigma^2 + \ln h_i + \ln v_i^2 \\ &= \ln \sigma^2 + \ln \exp(\delta_1 x_{i1} + \dots + \delta_k x_{ik}) + \ln v_i^2 \\ &= \underbrace{\ln \sigma^2 + E[\ln v_i^2]}_{\alpha_0} + \delta_1 x_{i1} + \dots + \delta_k x_{ik} + \underbrace{\ln v_i^2 - E[\ln v_i^2]}_{e_i} \\ \ln u_i^2 &= \alpha_0 + \delta_1 x_{i1} + \dots + \delta_k x_{ik} + e_i. \end{aligned} \quad (8.8)$$

For the regression equation (8.8) the assumptions MLR.1-MLR.4 are satisfied. Hence, the OLS estimator for δ_j is unbiased and consistent.

In practice, the u_i^2 's in the variance regression (8.8) are replaced by the squared OLS residuals \hat{u}_i^2 's from the sample regression $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$ of (8.1). The resulting $\hat{\delta}_j$'s are used to get the fitted values \hat{h}_i 's which are inserted into the GLS estimator (8.7) in step II.

- Outline of the **FGLS**-method:

Step I

- a) Regress \mathbf{y} on \mathbf{X} and compute the residual vector $\hat{\mathbf{u}}$ by OLS estimation of the original specification (8.1).
- b) Calculate $\ln \hat{u}_i^2$, $i = 1, \dots, n$, that are used as regressand in the variance regression (8.8).
- c) Estimate the variance regression (8.8) by OLS.
- d) Compute $\hat{h}_i = \exp \left(\hat{\delta}_1 x_{i1} + \dots + \hat{\delta}_k x_{ik} \right)$, $i = 1, \dots, n$.

Step II

The FGLS estimator $\hat{\beta}_{FGLS}$ is obtained **analogously** to the GLS procedure. The original regression (8.1) is multiplied from the left with the matrix

$$\hat{\mathbf{P}} = \begin{pmatrix} \hat{h}_1^{-1/2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \hat{h}_n^{-1/2} \end{pmatrix}.$$

This delivers a variant of the transformed regression

$$\mathbf{y}^\# = \mathbf{X}^\# \boldsymbol{\beta} + \mathbf{u}^\#. \quad (8.9)$$

Hence, OLS estimation of (8.9) leads to the **FGLS estimator**

$$\hat{\beta}_{FGLS} = \left(\mathbf{X}' \hat{\boldsymbol{\Psi}}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \hat{\boldsymbol{\Psi}}^{-1} \mathbf{y}, \quad (8.10)$$

with $\hat{\boldsymbol{\Psi}}^{-1} = \hat{\mathbf{P}}' \hat{\mathbf{P}}$.

- **Estimation properties of the FGLS estimators:**

- They are **consistent**, that is, they converge in probability to the true parameters for $n \rightarrow \infty$

$$\text{plim } \hat{\beta}_{FGLS} = \beta.$$

- The FGLS estimator is **asymptotically efficient**: For a correctly specified h_i and a sufficiently large sample, the FGLS estimator is preferable to the OLS estimator as the former one has a lower estimation-variance. (This is plausible, as FGLS also uses information on the functional form of the heteroskedasticity while OLS with heteroskedasticity-robust standard errors does not.)
- If the variance function h_i is **misspecified**, then the FGLS estimator is **inefficient**.

- Be aware that there may be considerable differences between the FGLS estimates and the OLS estimates.
- **Comparing OLS with heteroskedasticity-robust standard errors and FGLS**
 - If you know something about the variance function h_i , then FGLS is preferable. If you have no idea about it, then OLS with heteroskedasticity-robust standard errors may be better.
 - It is always a **good idea to run an OLS regression also with heteroskedasticity-robust standard errors** in order to see whether the significance of parameters depends on the presence of heteroskedasticity.
 - Since any estimator taking into account heteroskedasticity should be avoided if there is no heteroskedasticity, one should **test for**

the presence of heteroskedasticity, see Section 9.2.

• Trade Example Continued

- Consider Model 5 of Section 6.3 and compare OLS estimates, FGLS estimates, and OLS estimates with heteroskedasticity-robust standard errors.
- R program to run OLS, FGLS with both steps, and OLS with White standard errors, and scatter plots of residuals against fitted values for both estimators (part of `EOE_ws19_Emp_Beispiele.R`, lines 67ff.):

```
# definiere log der abhängigen Variable
log_imp <- log(trade_0_d_o)

### Erster Schritt a) KQ-Regression und Berechnung der Residuen

# KQ-Regression
eq_ols_model5 <- lm(log_imp ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +
                    log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))
```

```
# Berechne Residuen
res_ols_model5 <- eq_ols_model5$resid

# Berechne gefittete/angepasste Werte
fit_ols_model5 <- fitted.values(eq_ols_model5)

# Plote die Residuen gegen die gefitteten Werte, um zu untersuchen,
# ob Heteroskedastie vorliegen könnte
dev.off()
plot(fit_ols_model5, res_ols_model5, pch = 16)

### Erster Schritt b) bis d)

# Quadriere die Residuen und logarithmiere sie anschließend
ln_u_hat_sq <- log(res_ols_model5^2)

# Schätze die Varianzgleichung
eq_h_model5 <- lm(ln_u_hat_sq ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +
                 log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))

# Berechne die gefitteten Werte der logarithmierten Residuenanalyse
ln_u_hat_sq_hat <- fitted.values(eq_h_model5)
```

```
# Berechne die h's aus den gefitteten Werten der Varianzregression
h_hat <- exp(ln_u_hat_sq_hat)

### Zweiter Schritt: FGLS-Schätzung

# Schätze FGLS mit den gewichteten weights = 1/h_hat
eq_fgls_model5 <- lm(log_imp ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +
                    log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o),
                    weights = 1/h_hat)
summary(eq_fgls_model5)

# Berechne die gefitteten Werte aus FGLS
fit_fgls_model5 <- fitted.values(eq_fgls_model5)

# Berechne die Residuen aus FGLS
res_fgls_model5 <- resid(eq_fgls_model5)

# Standardisierung der Residuen mittels der Gewichte
res_fgls_model5_star <- res_fgls_model5*h_hat^(-1/2)

# Plote die Residuen gegen die gefitteten Werte
plot(fit_fgls_model5, res_fgls_model5_star, pch = 16)
```

```
### KQ-Regression mit heteroskedastie-robusten Standardfehlern
library(lmtest)
eq_white_model5 <- coeftest(eq_ols_model5, vcov=hccm(eq_ols_model5,type="hc1"))

# Graphiken/Outputs für Skript
summary(eq_ols_model5)
summary(eq_h_model5)
summary(eq_fgl5_model5)
eq_white_model5
```

— OLS output with usual standard errors:

Call:

```
lm(formula = log_imp ~ 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

Multiple R-squared: 0.9155, Adjusted R-squared: 0.9056

F-statistic: 93.12 on 5 and 43 DF, p-value: < 2.2e-16

— FGLS - Step I: estimate variance regression (8.8)

Call:

```
lm(formula = ln_u_hat_sq ~ 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
-5.6970	-0.6885	0.4991	1.4881	2.8326

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	63.57453	48.98487	1.298	0.201
log(wdi_gdpusdcr_o)	-4.79105	3.73067	-1.284	0.206
I((log(wdi_gdpusdcr_o))^2)	0.08839	0.07377	1.198	0.237
log(cepii_dist)	-0.36408	0.45827	-0.794	0.431
ebrd_tfes_o	0.23452	0.56327	0.416	0.679
log(cepii_area_o)	0.03706	0.23109	0.160	0.873

Residual standard error: 2.3 on 43 degrees of freedom

Multiple R-squared: 0.09998, Adjusted R-squared: -0.004677

F-statistic: 0.9553 on 5 and 43 DF, p-value: 0.4557

— Estimate FGLS - Step II: estimate (8.10)

Call:

```
lm(formula = log_imp ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +
    log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o), weights = 1/h_hat)
```

Weighted Residuals:

Min	1Q	Median	3Q	Max
-4.1788	-1.3479	0.2645	1.2478	3.6620

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-30.66686	16.80239	-1.825	0.0749	.
log(wdi_gdpusdcr_o)	3.55935	1.28177	2.777	0.0081	**
I((log(wdi_gdpusdcr_o))^2)	-0.05016	0.02482	-2.021	0.0495	*
log(cepii_dist)	-0.74852	0.11358	-6.590	5.06e-08	***
ebrd_tfes_o	0.39046	0.18441	2.117	0.0401	*
log(cepii_area_o)	-0.13856	0.05551	-2.496	0.0165	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.973 on 43 degrees of freedom

Multiple R-squared: 0.9055, Adjusted R-squared: 0.8945

F-statistic: 82.41 on 5 and 43 DF, p-value: < 2.2e-16

In contrast to EViews, the R command `eq_fgls_modelt <- lm(..., weights=...)` only delivers results for the weighted model ((8.6) or (8.9)). The corresponding residual sum of squares and further statistics for the weighted model, which EViews reports, are obtained with

```
(SSR <- sum(w_scaled*(log_imp_star - regressor_star%%coef(eq_fgls_model5))^2)) # SSR
mean(log_imp * (w_scaled))           # Mean dependent var
sd(log_imp * (w_scaled))             # S.D. dependent var
sqrt(SSR/(n-k-1))                   # S.E. of regression
```

Corresponding statistics for the unweighted model (in EViews “Unweighted Statistics”) are obtained in R with

```
# R-squared
(r_squared <- 1 - sum(residuals(eq_fgls_model5)^2) /
  sum((log_imp - mean(log_imp))^2))
# Adjusted R-squared
-k/(n-k-1) + (n-1)/(n-k-1)*r_squared
# Mean dependent var
mean(log_imp)
# S.D. dependent var
sd(log_imp)
# S.E. of regression
sqrt(sum(residuals(eq_fgls_model5)^2)/(n-k-1))
# Sum squared resid
sum(residuals(eq_fgls_model5)^2)
```

– OLS with heteroskedasticity-robust standard errors

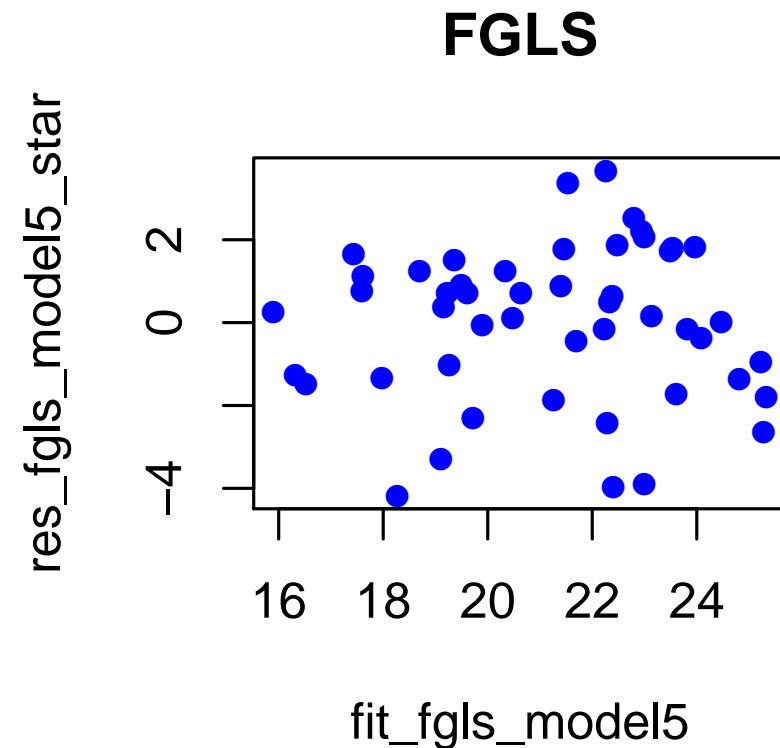
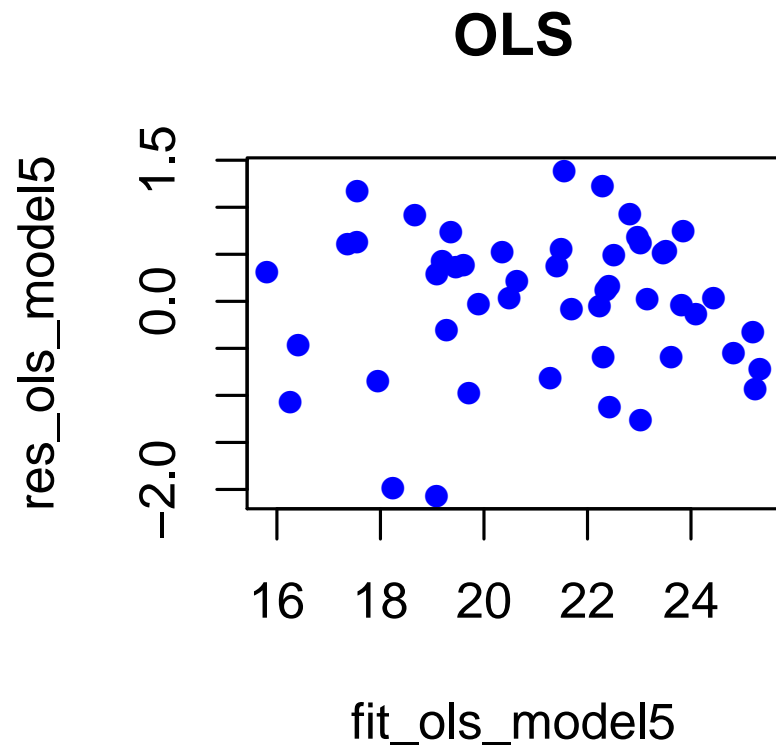
In R they are obtained with the command `coefTest()` from the R package `lmtest`

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-35.233143	16.148517	-2.1818	0.034635	*
log(wdi_gdpusdcr_o)	3.908811	1.244314	3.1413	0.003041	**
I((log(wdi_gdpusdcr_o))^2)	-0.057108	0.024340	-2.3462	0.023644	*
log(cepii_dist)	-0.748559	0.124427	-6.0160	3.465e-07	***
ebrd_tfes_o	0.419883	0.155896	2.6934	0.010045	*
log(cepii_area_o)	-0.132380	0.046455	-2.8496	0.006693	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Diagnostic plots: (standardized) residuals against fitted values



- **Output table for Model 4 and Model 5 using various estimators** (compare Section 4.8):

Dependent Variable: $\ln(\text{imports by Germany})$			
Independent variables / Model	(4)-OLS	(5)-OLS	(5)-FGLS
constant	2.427 (2.132) [1.337]	-35.233 (17.441) [16.148]	-30.666 (16.802)
$\ln(\text{gdp})$	1.025 (0.076) [0.070]	3.908 (1.328) [1.244]	3.559 (1.281)
$(\ln(\text{gdp}))^2$	—	-0.057 (0.026) [0.024]	-0.050 (0.024)
$\ln(\text{distance})$	-0.888 (0.156) [0.120]	-0.748 (0.163) [0.124]	-0.748 (0.113)
openess	0.353 (0.206) [0.180]	0.419 (0.200) [0.155]	0.390 (0.184)
$\ln(\text{area})$	-0.151 (0.085) [0.050]	-0.132 (0.082) [0.046]	-0.138 (0.055)
number of observations	49	49	49
R^2	0.906	0.915	0.905
standard error of the regression	0.853	0.819	0.736
residual sum of squares	32.017	28.846	23.330
AIC	2.6164	2.5529	
HQ	2.6896	2.6408	
SC	2.8094	2.7845	

Notes: OLS or FGLS standard errors in parentheses, White standard errors in brackets

– Results and Interpretation:

- * OLS and FGLS parameter estimates are quite similar for all parameters. The effect of potential heteroskedasticity is only weak. Therefore, one should test whether heteroskedasticity is present at all. If not, the FGLS estimator would not be efficient and we should use the OLS estimator instead.
- * When taking into account heteroskedasticity, based on FGLS there is no insignificant parameter at the 5% significance level. This also holds when using heteroskedasticity-robust OLS standard errors.

- * Inspecting the scatter plots of OLS and standardized FGLS residuals against fitted values does not automatically suggest heteroskedasticity. Thus, heteroskedasticity tests are useful, see Section 9.2.

- **Cigarette Example** (Wooldridge, 2009, Example 8.7) with R:

Step I

1. OLS estimation

```
lm(formula = cigs ~ lincome + lcigpric + educ + age + I(age^2) +
    restaurn, data = smoke_all)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.819	-9.381	-5.975	7.922	70.221

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.639855	24.078660	-0.151	0.87988
lincome	0.880268	0.727783	1.210	0.22682
lcigpric	-0.750855	5.773343	-0.130	0.89656
educ	-0.501498	0.167077	-3.002	0.00277 **
age	0.770694	0.160122	4.813	1.78e-06 ***
I(age^2)	-0.009023	0.001743	-5.176	2.86e-07 ***
restaurn	-2.825085	1.111794	-2.541	0.01124 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 800 degrees of freedom

Multiple R-squared: 0.05274, Adjusted R-squared: 0.04563

F-statistic: 7.423 on 6 and 800 DF, p-value: 9.499e-08

2. Save the residuals using `u_hat_cig <- resid(ols_1)`

3. Taking the logarithm of the squared residuals by using

```
ln_u_sq <- log(u_hat_cig^2)
```

4. Estimation of variance regression (8.8) with OLS yields

```
lm(formula = ln_u_sq ~ lincome + lcigpric + educ + age + I(age^2) +
    restaurn, data = smoke_all)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-11.2186 -0.2237 -0.0227  0.2951  4.9588
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.9207040  2.5630344  -0.749  0.45384
lincome      0.2915405  0.0774683   3.763  0.00018 ***
lcigpric     0.1954209  0.6145390   0.318  0.75057
educ        -0.0797036  0.0177844  -4.482 8.49e-06 ***
age          0.2040054  0.0170441  11.969 < 2e-16 ***
I(age^2)    -0.0023921  0.0001855 -12.893 < 2e-16 ***
restaurn    -0.6270116  0.1183440  -5.298 1.51e-07 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.427 on 800 degrees of freedom
```

```
Multiple R-squared:  0.2474, Adjusted R-squared:  0.2417
```

```
F-statistic: 43.82 on 6 and 800 DF,  p-value: < 2.2e-16
```

– Save the \hat{h}_i with `h_hat_cig <- exp(ln_u_sq - resid(ols_2))`

Step II

Weighted LS estimate with weights $h_hat_cig^{(-1)}$

Call:

```
lm(formula = cigs ~ lincome + lcigpric + educ + age + I(age^2) +
    restaurn, data = smoke_all, weights = h_hat_cig^(-1))
```

Weighted Residuals:

Min	1Q	Median	3Q	Max
-1.9036	-0.9532	-0.8099	0.8415	9.8556

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.6354329	17.8031409	0.317	0.751674	
lincome	1.2952396	0.4370117	2.964	0.003128	**
lcigpric	-2.9403048	4.4601450	-0.659	0.509932	
educ	-0.4634464	0.1201587	-3.857	0.000124	***
age	0.4819480	0.0968082	4.978	7.86e-07	***
I(age^2)	-0.0056272	0.0009395	-5.990	3.17e-09	***
restaurn	-3.4610642	0.7955050	-4.351	1.53e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.579 on 800 degrees of freedom

Multiple R-squared: 0.1134, Adjusted R-squared: 0.1068

F-statistic: 17.06 on 6 and 800 DF, p-value: < 2.2e-16

— Compare them with the OLS estimates based on White standard errors.

9 Multiple Regression Analysis: Model Diagnostics

9.1 The RESET Test

RESET Test (regression specification error test)

Idea and implementation:

- If the original model

$$y = x_0\beta_0 + \dots + x_k\beta_k + u = \mathbf{x}'\boldsymbol{\beta} + u$$

satisfies assumption MLR.4 $E[u|x_0, \dots, x_k] = 0$, it holds that

$$E[y|x_0, \dots, x_k] = x_0\beta_0 + \dots + x_k\beta_k = \mathbf{x}'\boldsymbol{\beta}.$$

- Then, any further term added to the model should not be significant. Thus, any nonlinear function of the independent variables should be insignificant.
- Thus, the null hypothesis of the RESET test is formulated such that one can test the significance of nonlinear functions of the fitted values $\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$ that are added to the model. Note that the fitted values are a linear function of the regressors of the original specification.

- In practice it turned out that for implementing the RESET test it is sufficient to include quadratic and cubic terms of \hat{y} only

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha\hat{y}^2 + \gamma\hat{y}^3 + \varepsilon.$$

The pair of hypotheses is

$$H_0 : \alpha = 0, \gamma = 0 \quad (\text{linear model is correctly specified})$$

$$H_1 : \alpha \neq 0 \text{ and/or } \gamma \neq 0.$$

The null hypothesis is tested using an F test with 2 degrees of freedom in the numerator and $n - k - 3$ in the denominator.

- **Be aware** that the null hypothesis may also be rejected because of omitting relevant regressor variables.
- In R use the command `resettest()` in the R package `lmtest`.

9.2 Heteroskedasticity Tests

- As already noted, it does not make sense to “automatically” use the FGLS estimator. If the errors are homoskedastic, the OLS estimator with OLS standard errors should be used.
- Thus, one should test if there is statistical evidence for heteroskedasticity.
- In the following, two different tests for heteroskedasticity are discussed: the Breusch-Pagan test and the White test. For both, the null hypothesis is “homoskedastic errors”.
- Both tests are implemented in R. The Breusch-Pagan test `bptest` in the R package `lmtest`. The White test `white_lm` in the R package `skedastic`. The latter is also programmed in `EOE_ws19_Emp_Beispiele.R`, lines 848 and following.

It is assumed that for the multiple linear regression

$$y = \beta_0 + x_1\beta_1 + \dots + x_k\beta_k + u$$

assumptions MLR.1 to MLR.4 hold.

The **pair of hypotheses** that has to be tested is

$$H_0 : \text{Var}(u_i|\mathbf{x}_i) = \sigma^2 \quad (\text{homoskedasticity}),$$

$$H_1 : \text{Var}(u_i|\mathbf{x}_i) = \sigma_i \neq \sigma^2 \quad (\text{heteroskedasticity}).$$

The general idea underlying heteroskedasticity tests is that under the null hypothesis no regressor should have any explanatory power for $\text{Var}(u_i|\mathbf{x}_i)$. If the null hypothesis is not true, $\text{Var}(u_i|\mathbf{x}_i)$ can be a (nearly arbitrary) function of the regressors x_j , ($1 \leq j \leq k$).

Note: The Breusch-Pagan test and the White test differ with respect to the specification of their alternative hypothesis.

Breusch-Pagan Test

- Idea: Consider the regression

$$u_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik} + v_i, \quad i = 1, \dots, n. \quad (9.1)$$

Under assumptions MLR.1 to MLR.4 the OLS estimator for the δ_j 's is unbiased.

The pair of hypotheses is:

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0 \quad \text{versus}$$

$$H_1 : \delta_1 \neq 0 \text{ and/or } \delta_2 \neq 0 \text{ and/or } \dots,$$

since under H_0 it holds that $E[u_i^2 | \mathbf{X}] = \delta_0$.

- **Difference from the previous application of the F test:**

- The squares of the errors u_i^2 are by no means normally distributed since they are squared quantities and thus cannot take negative values. Hence, the v_i cannot be normally distributed and the F distribution of the F statistic does not hold exactly in finite samples. However, the central limit theorem (CLT) works here as well, see Section 5.2, and the F statistic follows approximately an F distribution in large samples.
- The errors u_i are unknown. They can be replaced by the OLS residuals \hat{u}_i . In doing so, the F test remains asymptotically valid (proof is formally sophisticated).

- The R^2 version of the test statistic can be used. Note that for a regression including only a constant, it holds that $R^2 = 0$ since $SSR = SST$ (there are no regressors that show a variation). Call the coefficient of variation of the OLS estimation of (9.1) $R_{\hat{u}^2}^2$ then

$$F = \frac{R_{\hat{u}^2}^2/k}{(1 - R_{\hat{u}^2}^2)/(n - k - 1)}.$$

The F statistic for testing the joint significance of all regressors is generally given by the appropriate software.

- H_0 is rejected if F exceeds the critical value for a chosen significance level (or equivalently if the p -value is smaller than the significance level).

• Cigarette Example Continued: (from Section 8.4):

```
lm(formula = u_hat_sq ~ lincome + lcigpric + educ + age + I(age^2) +
    restaurn, data = smoke_all)
```

Residuals:

```
   Min       1Q   Median       3Q      Max
-270.1 -127.5  -94.0  -39.1 4667.8
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-636.30311	652.49456	-0.975	0.3298
lincome	24.63849	19.72180	1.249	0.2119
lcigpric	60.97656	156.44869	0.390	0.6968
educ	-2.38423	4.52753	-0.527	0.5986
age	19.41748	4.33907	4.475	8.75e-06 ***
I(age^2)	-0.21479	0.04723	-4.547	6.27e-06 ***
restaurn	-71.18138	30.12789	-2.363	0.0184 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 363.2 on 800 degrees of freedom

Multiple R-squared: 0.03997, Adjusted R-squared: 0.03277

F-statistic: 5.552 on 6 and 800 DF, p-value: 1.189e-05

The F statistic for the above H_0 is 5.55 and the corresponding p -value is smaller than 1%. The null hypothesis of homoskedastic errors thus is rejected at a level of 1%.

- **Note:**

- If one conjectures that the heteroskedasticity is caused by specific variables that have not been included previously, they can be included in regression (9.1).
- If H_0 is not rejected, this does not mean automatically that the u_i 's are homoskedastic. If the specification (9.1) does not contain all relevant variables causing heteroskedasticity, then it may happen that all δ_j , $j = 1, \dots, k$, are jointly insignificant.
- A variant of the Breusch-Pagan test is a test for multiplicative heteroskedasticity, i.e. the variance is of the form $\sigma_i^2 = \sigma^2 \cdot h(\mathbf{x}_i' \boldsymbol{\beta})$. If, for example, the case $h(\cdot) = \exp(\cdot)$ is assumed, the test equation $\ln(\hat{u}_i^2) = \ln(\sigma^2) + \mathbf{x}_i' \boldsymbol{\beta} + v$ results.

White Test

- **Background:**

For deriving the asymptotic distribution of the OLS estimator the assumption of homoskedastic errors MLR.5 is not necessary.

It is enough that the squared errors u_i^2 are uncorrelated with all regressors and the squares and cross products of the latter.

This can easily be tested using the following regression, where

the errors are already replaced by the residuals:

$$\begin{aligned}\hat{u}_i^2 &= \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik} \\ &\quad + \delta_{k+1} x_{i1}^2 + \cdots + \delta_{J_1} x_{ik}^2 \\ &\quad + \delta_{J_1+1} x_{i1} x_{i2} + \cdots + \delta_{J_2} x_{ik-1} x_{ik} \\ &\quad + v_i, \quad i = 1, \dots, n.\end{aligned}\tag{9.2}$$

- The pair of hypotheses is:

$$H_0 : \delta_j = 0 \text{ for all } j = 1, 2, \dots, J_2,$$

$$H_1 : \delta_j \neq 0 \text{ for at least one } j.$$

Again, an F test can be used whose distribution is approximated by the F distribution (asymptotic distribution).

Better known is the LM version of the test. It is computed as $LM = n R^2$ with R^2 obtained from estimating (9.2). The LM test statistic is asymptotically $\chi^2(J_2)$ distributed.

- With many regressors, it is tedious to implement the F test for (9.2) manually. However, most software packages provide the White test.
- When implementing the White test, a large number of parameters has to be estimated if the original model exhibits large k . This is hardly possible in small samples. Then one only includes the squares x_{ij}^2 into the regression and neglects all cross products.
- **Note:** If the null hypothesis is rejected, this may also be due to violation of MLR.1 or MLR.4. Then, the original regression is **mis-specified!**
- **Cigarette Example Continued:**
Use of R function `whitetest()`, see appendix 10.5, slide LXVII.
Not all result lines reproduced:

F Statistic	df1	df2	p Value
2.159257e+00	2.500000e+01	7.810000e+02	9.047555e-04

LM Statistic	df	p Value
52.172439390	25.000000000	0.001139947

Call:

```
lm(formula = form, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-326.8	-138.2	-81.2	-10.4	4620.0

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.937e+04	2.056e+04	1.429	0.1535
lincome	-1.050e+03	9.634e+02	-1.089	0.2763
lcigpric	-1.034e+04	9.755e+03	-1.060	0.2894
educ	-1.175e+02	2.513e+02	-0.467	0.6403
age	-2.641e+02	2.358e+02	-1.120	0.2629
I(age^2)	3.469e+00	3.195e+00	1.086	0.2779
restaurn	-2.868e+03	2.987e+03	-0.960	0.3372
I(lincome^2)	-3.941e+00	1.707e+01	-0.231	0.8175
I(lcigpric^2)	6.685e+02	1.204e+03	0.555	0.5790
I(educ^2)	-2.903e-01	1.288e+00	-0.225	0.8217
I(I(age^2)^2)	1.178e-04	1.458e-04	0.808	0.4196
I(restaurn^2)	NA	NA	NA	NA
lincome:lcigpric	3.299e+02	2.392e+02	1.379	0.1683
lincome:educ	-9.592e+00	8.047e+00	-1.192	0.2336
lincome:age	-3.355e+00	6.682e+00	-0.502	0.6158
lincome:I(age^2)	2.670e-02	7.302e-02	0.366	0.7147
lincome:restaurn	-5.989e+01	4.969e+01	-1.205	0.2285

```

lcigpric:educ      3.291e+01  5.906e+01  0.557  0.5775
lcigpric:age       6.288e+01  5.529e+01  1.137  0.2558
lcigpric:I(age^2) -6.224e-01  5.947e-01 -1.046  0.2957
lcigpric:restaurn  8.622e+02  7.206e+02  1.196  0.2319
educ:age           3.617e+00  1.725e+00  2.097  0.0363 *
educ:I(age^2)     -3.556e-02  1.766e-02 -2.013  0.0445 *
educ:restaurn     -2.896e+00  1.066e+01 -0.272  0.7859
age:I(age^2)      -1.911e-02  2.866e-02 -0.667  0.5050
age:restaurn      -4.933e+00  1.084e+01 -0.455  0.6492
I(age^2):restaurn  3.845e-02  1.205e-01  0.319  0.7497
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 362.9 on 781 degrees of freedom
Multiple R-squared:  0.06465, Adjusted R-squared:  0.03471
F-statistic: 2.159 on 25 and 781 DF,  p-value: 0.0009048

```

Result: With the White test H_0 is also rejected.

Trade Example Continued

(from Section 8.4):

- Breusch-Pagan test for heteroskedasticity using OLS residuals with R command `bptest()` in the R package `lmtest`

```
studentized Breusch-Pagan test
```

```
data: eq_ols_model5
```

```
BP = 5.3378, df = 5, p-value = 0.3761
```

- White test (without cross terms) for heteroskedasticity using OLS residuals

```
# führe White-Test durch, Funktion whitetest() auf Folie 399 definiert
ols_model5_white <- whitetest(eq_ols_model5, crossterms=0)
```

Ergebnis:

F Statistic	df1	df2	p Value
1.0337453	5.0000000	43.0000000	0.4101294

LM Statistic	df	p Value
5.2579260	5.0000000	0.3852202

Call:

```
lm(formula = form, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.8842	-0.3981	-0.1658	0.1013	3.2860

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.879e+00	4.939e+00	0.988	0.329

I(log(wdi_gdpusdcr_o)^2)	-1.269e-02	1.400e-02	-0.906	0.370
I(I((log(wdi_gdpusdcr_o))^2)^2)	7.637e-06	1.070e-05	0.714	0.479
I(log(cepii_dist)^2)	-4.135e-03	1.213e-02	-0.341	0.735
I(ebrd_tfes_o^2)	3.897e-02	3.575e-02	1.090	0.282
I(log(cepii_area_o)^2)	1.065e-03	3.938e-03	0.270	0.788

Residual standard error: 0.871 on 43 degrees of freedom

Multiple R-squared: 0.1073, Adjusted R-squared: 0.003503

F-statistic: 1.034 on 5 and 43 DF, p-value: 0.4101

- Breusch-Pagan test for heteroskedasticity using standardized FGLS residuals

```
LM-Teststatistik      p-Wert
      2.5984906      0.7615946
```

```
lm(formula = data.frame(cbind(u_star_sq, regressor_star)))
```

Residuals:

```
      Min      1Q  Median      3Q      Max
-0.6089 -0.3920 -0.1971  0.2204  1.9828
```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.069617   0.388161   0.179   0.859
log.wdi_gdpusdcr_o.  0.035974   0.105189   0.342   0.734
I..log.wdi_gdpusdcr_o...2. -0.002430   0.002957  -0.822   0.416
log.cepii_dist.  -0.040875   0.095084  -0.430   0.669
ebrd_tfes_o     0.224651   0.168832   1.331   0.190
log.cepii_area_o.  0.039700   0.049151   0.808   0.424
```

Residual standard error: 0.6394 on 43 degrees of freedom

Multiple R-squared: 0.05303, Adjusted R-squared: -0.05708

F-statistic: 0.4816 on 5 and 43 DF, p-value: 0.788

- White test (without cross terms) for heteroskedasticity using FGLS residuals

```
LM-Teststatistik      p-Wert
      5.5752453      0.4724093
```

Call:

```
lm(formula = cbind(u_star_sq, regressor_white))
```

Residuals:

```
      Min      1Q   Median      3Q      Max
-0.68248 -0.40380 -0.13190  0.07897  1.91210
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.577e-01	4.313e-01	-0.598	0.5533
w_scaled_sq	1.358e+01	8.538e+00	1.590	0.1193
log.wdi_gdpusdcr_o.	-3.678e-02	2.292e-02	-1.605	0.1160
I..log.wdi_gdpusdcr_o...2.	2.384e-05	1.550e-05	1.538	0.1315
log.cepii_dist.	-1.139e-02	7.836e-03	-1.453	0.1536
ebrd_tfes_o	7.024e-02	3.207e-02	2.191	0.0341 *
log.cepii_area_o.	1.983e-03	1.516e-03	1.308	0.1981

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6259 on 42 degrees of freedom
Multiple R-squared: 0.1138, Adjusted R-squared: -0.01282
F-statistic: 0.8987 on 6 and 42 DF, p-value: 0.5049

Results:

- Note that the specification of the White test without cross terms follows EViews 6.0 and does not include level terms (in contrast to (9.2)).
- Both, the Breusch-Pagan and the White test *do not* reject the null hypothesis of homoskedastic errors for the OLS residuals at any reasonable significance level. Thus, using OLS with heteroskedasticity-robust standard errors or FGLS in Section 8.4 was not efficient.
- Both, the Breusch-Pagan and the White test do not reject the null hypothesis of homoskedastic standardized errors in the FGLS

framework. Thus, the variance regression in Section 8.4 does not seem to be misspecified.

- In sum, among all models and estimation procedures considered, the FGLS estimates of Model 5 seem to be the most reliable ones.

Reading: Chapter 8 in [Wooldridge \(2009\)](#) (without Section 8.5 concerning linear probability models).

9.3 Model Specification II: Useful Tests

9.3.1 Comparing Models with Identical Regressand

Starting point: two non-nested models

$$(M1) \quad y = x_0\beta_0 + \dots + x_k\beta_k + u = \mathbf{x}'\boldsymbol{\beta} + u,$$

$$(M2) \quad y = z_0\gamma_0 + \dots + z_m\gamma_m + v = \mathbf{z}'\boldsymbol{\gamma} + v,$$

where $k = m$ does not have to hold.

Decision between (M1) and (M2): using

- information criteria (AIC, SC, HQ, ...),
- encompassing test,
- non-nested F test,
- J test.

All three tests can be constituted on the encompassing principle.

Encompassing Principle

Let two non-nested models be given:

$$(M1) \quad y = \mathbf{x}'\boldsymbol{\beta} + u,$$

$$(M2) \quad y = \mathbf{z}'\boldsymbol{\gamma} + v.$$

For clarifying the non-nested relationship between (M1) and (M2), define

$$\mathbf{x}' = \begin{pmatrix} \mathbf{w}' & \mathbf{x}'_B \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_A & \boldsymbol{\beta}_B \end{pmatrix},$$

$$\mathbf{z}' = \begin{pmatrix} \mathbf{w}' & \mathbf{z}'_B \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_A & \boldsymbol{\gamma}_B \end{pmatrix},$$

such that \mathbf{w} contains all common regressors

$$(M1) \quad y = \mathbf{w}'\boldsymbol{\beta}_A + \mathbf{x}'_B\boldsymbol{\beta}_B + u,$$

$$(M2) \quad y = \mathbf{w}'\boldsymbol{\gamma}_A + \mathbf{z}'_B\boldsymbol{\gamma}_B + v.$$

Idea of the encompassing principle:

- If (M1) is correctly specified, it must be able to explain the results of an estimation of (M2) (and vice versa).
- If not, (M1) has to be rejected (and vice versa).

Derivation:

Consider the “**artificial nesting model**”

$$(ANM) \quad y = \mathbf{w}'\mathbf{a} + \mathbf{x}'_B \mathbf{b}_x + \mathbf{z}'_B \mathbf{b}_z + \varepsilon, \quad E[\varepsilon | \mathbf{w}, \mathbf{x}_B, \mathbf{z}_B] = 0.$$

Different settings:

- (ANM) correctly specified model such that (M1) and (M2) are misspecified. Model (M2) is estimated.
- (M1) correctly specified model. Model (M2) is estimated.
- (M2) correctly specified model. Model (M1) is estimated.

In general an **omitted variable bias** results for all cases.

Details:

- (ANM) correctly specified model such that (M1) and (M2) are misspecified. Model (M2) is estimated. $\Rightarrow \mathbf{x}_B$ omitted.

$$\begin{aligned}
 E[y|\mathbf{w}, \mathbf{z}_B] &= E[\mathbf{w}'\mathbf{a} + \mathbf{x}'_B\mathbf{b}_x + \mathbf{z}'_B\mathbf{b}_z + \varepsilon|\mathbf{w}, \mathbf{z}_B] \\
 &= E[\mathbf{w}'\mathbf{a}|\mathbf{w}, \mathbf{z}_B] + E[\mathbf{x}'_B\mathbf{b}_x|\mathbf{w}, \mathbf{z}_B] \\
 &\quad + E[\mathbf{z}'_B\mathbf{b}_z|\mathbf{w}, \mathbf{z}_B] + E[\varepsilon|\mathbf{w}, \mathbf{z}_B] \\
 &= \mathbf{w}'\mathbf{a} + E[\mathbf{x}'_B|\mathbf{w}, \mathbf{z}_B]\mathbf{b}_x + \mathbf{z}'_B\mathbf{b}_z + E[\varepsilon|\mathbf{w}, \mathbf{z}_B].
 \end{aligned}$$

For simplicity it is assumed that x_B is scalar. Then it holds that

$$\begin{aligned}
 x_B &= \mathbf{w}'\mathbf{q} + \mathbf{z}'_B\mathbf{p} + \nu, \\
 E[x_B|\mathbf{w}, \mathbf{z}_B] &= \mathbf{w}'\mathbf{q} + \mathbf{z}'_B\mathbf{p}.
 \end{aligned}$$

It also holds that

$$E [E[\varepsilon|\mathbf{w}, \mathbf{x}_B, \mathbf{z}_B]|\mathbf{w}, \mathbf{z}_B] = E[\varepsilon|\mathbf{w}, \mathbf{z}_B].$$

Since (ANM) is correct, it holds that $E[\varepsilon|\mathbf{w}, \mathbf{x}_B, \mathbf{z}_B] = 0$ and thus

$$E[0] = 0 = E[\varepsilon|\mathbf{w}, \mathbf{z}_B].$$

When estimating (M2) instead of (ANM), one gets

$$\begin{aligned} E[y|\mathbf{w}, \mathbf{z}_B] &= \mathbf{w}'\mathbf{a} + [\mathbf{w}'\mathbf{q} + \mathbf{z}'_B\mathbf{p}]b_x + \mathbf{z}'_B\mathbf{b}_z \\ &= \mathbf{w}' \underbrace{[\mathbf{a} + \mathbf{q}b_x]}_{\gamma_A} + \mathbf{z}'_B \underbrace{[\mathbf{b}_z + \mathbf{p}b_x]}_{\gamma_B}. \end{aligned} \quad (9.3)$$

Note that the biases $\mathbf{q}b_x$ and $\mathbf{p}b_x$ are caused by omitting the variable \mathbf{x}_B . These effects bias the direct impact of \mathbf{w} via \mathbf{a} and of \mathbf{z}_B via \mathbf{b}_z on y .

- *(M1) correctly specified model. Model (M2) is estimated.*

Then $\mathbf{b}_z = \mathbf{0}$ and from (9.3) the following restriction results:

$$\mathbf{p}b_x = \gamma_B.$$

Now it can be seen that knowing the correctly specified model (M1) is enough for deriving model (M2), thus predicting γ_B or the expectation of the OLS estimator. In other words: Since (M2) is “smaller” than (M1) with respect to the relevant variables, the behavior of (M2) can be predicted with the help of (M1) when an unbiased estimator is used for the latter. Then one says “**(M1) encompasses (M2)**”. (Knowing (M1) is not enough here if (ANM) is the correct model, $\mathbf{b}_z \neq \mathbf{0}$.)

- *(M2) correctly specified model. Model (M1) is estimated.*

Can be derived just as in the above case.

Thus, for the null hypothesis “(M1) encompasses (M2)” two equivalent hypotheses can be tested:

- $H_0 : \mathbf{p}b - \gamma_B = \mathbf{0}$ - more complicated, no details here. (This version is often termed **encompassing test** and sometimes has advantages in more general models.)
- $H_0 : \mathbf{b}_Z = \mathbf{0}$ in (ANM) - easy: by the help of a **non-nested F test**.

Proceeding for more than two alternatives:

- Based on this same principle, the remaining model competes with further alternative models as long as it is not rejected.

- Problem of this principle: it can happen that both null hypotheses have to be rejected.

Non-nested F test

Idea and implementation:

- Hypotheses: “ H_0 : model (M1) is correct” versus “ H_1 : model (M1) incorrect”.
- Again, partition $\mathbf{z}' = (\mathbf{w}', \mathbf{z}'_B)$, where the k_A regressors from \mathbf{w} are contained in \mathbf{x} but the k_B regressors from \mathbf{z}_B are not contained.
- Formulate the **artificial nesting model (ANM)**

$$y = \mathbf{x}'\boldsymbol{\beta} + \mathbf{z}'_B \mathbf{b}_z + \varepsilon.$$

- Based on this ANM test H_0 where

$$H_0 : \mathbf{b}_z = \mathbf{0}$$

using an F test with k_B degrees of freedom in the numerator and $n - m - k_B$ in the denominator.

- For the test of (M2) vs. (M1) proceed analogously with partition $\mathbf{x}' = (\mathbf{w}', \mathbf{x}'_B) \dots$

J test (Davidson-MacKinnon test)

Idea and implementation:

- For the *J* test the ANM is formulated such that both (M1) and (M2) are nested in the ANM:

$$y = (1 - \lambda)\mathbf{x}'\boldsymbol{\beta} + \lambda\mathbf{z}'\boldsymbol{\gamma} + \varepsilon.$$

For the case $\lambda = 0$ the model (M1) results, for $\lambda = 1$ model (M2).

- Problem: λ , $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are not identified in the above approach.
- Solution: replace $\boldsymbol{\gamma}$ by the OLS estimator from (M2) $\hat{\boldsymbol{\gamma}}$.
I.e. test $H_0 : \lambda = 0$ with test equation $y = \mathbf{x}'\boldsymbol{\beta}^* + \lambda\hat{y}_{M2} + \eta$, where $\boldsymbol{\beta}^* = (1 - \lambda)\boldsymbol{\beta}$ and $\hat{y}_{M2} = \mathbf{z}'\hat{\boldsymbol{\gamma}}$ is the fitted value from the OLS estimation of (M2).

- For testing whether (M2) is valid, proceed analogously ...
- Interpretation of the logic of the test:
For testing model (M1) it is enlarged by the fitted values of model (M2); these (i.e. the by the regressors in (M2) explained part of y) are tested for their significance in the test equation.
- Advantages of the J test compared to the non-nested F test:
 - only one single restriction has to be tested,
 - higher power, if k_B or respectively m_B are very large,
 - in case of $k_B = 1$ or respectively $m_B = 1$ the tests are equivalent.

9.3.2 Comparing Models with differing Regressand

Idea and implementation (of the P test):

Example: **linear model versus log-log alternative**

- Step 1: Run an OLS estimation for both models.
- Step 2: Compute the corresponding fitted values \hat{y}_{lin} (linear model) and $\widehat{\ln(y_{log})}$ (log-log model).
- Step 3a: Test the linear approach against the log-log alternative using the ANM

$$y = \sum x_j \beta_{j,lin} + \delta_{lin} [\ln(\hat{y}_{lin}) - \widehat{\ln(y_{log})}] + u,$$

by a t test with the null hypothesis

$$H_0 : \delta_{lin} = 0 \quad (\text{linear model is correct}).$$

- Step 3b: Test the log-log approach against the linear alternative using the ANM

$$\ln(y) = \sum \ln(x_j)\beta_{j,log} + \delta_{log}[\hat{y}_{lin} - \exp(\widehat{\ln y_{log}})] + v,$$

by a t test with the null hypothesis

$$H_0 : \delta_{log} = 0 \quad (\text{log-log model is correct}).$$

Problem: it is possible that both hypotheses are rejected (i.e. another functional form is relevant) or both cannot be rejected (i.e. the problem of lacking power or something else).

Note: in this case a comparison using the information criteria is **not possible**.

Reading: Chapter 9 in [Wooldridge \(2009\)](#).

10 Appendix

10.1 A Condensed Introduction to Probability

Preliminary Statement: The following pages are not considered as deterrence, but as supplement to the illustrations found in introductory textbooks for econometrics. This supplement is intended to explain the intuition underlying the large amount of definitions and concepts in probability theory.

Nevertheless it is not possible to completely avoid formulas, although it may take some time to clarify your mind.

A very detailed introduction to probability theory is e. g. [Casella and Berger \(2002\)](#).

- **Sample space, outcome space:**

The set Ω contains all possible outcomes of a random experiment.

This set can contain (countably) finite or infinite outcomes.

Examples:

- Urn with 4 balls of different color: $\Omega = \{\text{yellow, red, blue, green}\}$
- Monthly income of a household in the future: $\Omega = [0, \infty)$

Remark:

- If there is a finite number of outcomes, they are often denoted as ω_i . For S outcomes, Ω appears as

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_S\}.$$

- If there is an infinite number of outcomes, each one is often denoted as ω .

- **Event:**

- If a particular outcome realizes, an event occurs.
- If an event contains exactly one outcome of the sample space, it is called elementary event.
- An event is a subset of the sample space Ω . Thus every set of possible outcomes = every subset of the set Ω including Ω itself.

Examples:

- Urn-example: possible events are for example $\{\text{yellow, red}\}$ or $\{\text{red, blue, green}\}$.

- Household income: possible events are all possible subintervals and combinations of them, e.g. $(0, 5000]$, $[1000, 1001)$, $(400, \infty)$, 4000 , and so on.

Remark: By using the general point of view with the ω 's, one has

- for the case of S outcomes: $\{\omega_1, \omega_2\}$, $\{\omega_S\}$, $\{\omega_3, \dots, \omega_S\}$, and so on.
- for the case of infinitely many outcomes located inside an interval $\Omega = (-\infty, \infty)$: $(a_1, b_1]$, $[a_2, b_2)$, $(0, \infty)$, and so on, where the lower bound always has to be lower or equal the upper bound ($a_i \leq b_i$).

- **Random variable:**

A random variable is a function that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

Urn example: $X(\omega_1) = 0$, $X(\omega_2) = 3$, $X(\omega_3) = 17$, $X(\omega_4) = 20$.

- **Density function**

- Preliminary statement: As we have already seen, it gets complicated if Ω contains infinitely many outcomes. Consider for example $\Omega = [0, 4]$. If one wants to compute the probability for the number π to appear, this probability is equal to zero. If it were not equal to zero, we had the problem that a sum of all probabilities for all (infinitely many) numbers could not be equal to 1. What to do?

- A back door is the following trick: Consider the probability for the outcome of the random variable X being located in the interval $[0, x]$, with $x < 4$. This probability can be written as $P(X \leq x)$. Now determine how the probability changes by extending the size of this interval $[0, x]$ by h . The solution to this is: $P(X \leq x + h) - P(X \leq x)$. By relating this change in probability to the interval length one gets

$$\frac{P(X \leq x + h) - P(X \leq x)}{h}.$$

For a decreasing interval length h that approaches zero, one obtains the following limit:

$$\lim_{h \rightarrow 0} \frac{P(X \leq x + h) - P(X \leq x)}{h} = f(x).$$

This limit is called **probability density function** or shortly **density function** that belongs to the probability function P .

– How to interpret a density function?

By using the sloppy formulation

$$\frac{P(X \leq x + h) - P(X \leq x)}{h} \approx f(x)$$

and rewriting as

$$P(X \leq x + h) - P(X \leq x) \approx f(x)h,$$

one can see that $f(x)$ determines the rate of change for the probability that X falls into the interval $[0, x]$ if the interval length is extended by h . Hence, the density function is a **rate**.

– As the density function is a derivative, we get conversely for our example

$$\int_0^x f(u)du = P(X \leq x) = F(x).$$

Here, $F(x) = P(X \leq x)$ is called **probability distribution function**. Certainly, in this example we get

$$\int_0^4 f(u)du = P(X \leq 4) = 1.$$

In general, the integral of the density function over the full support of the random variable yields a value of 1. Consider for example $X(\omega) \in \mathbb{R}$:

$$\int_{-\infty}^{\infty} f(u)du = P(X \leq \infty) = 1.$$

- **Conditional probability function**

Let's begin with an **example**:

Let the random variable $X \in [0, \infty)$ be the payoff in a lottery. The probability function (distribution function) $P(X \leq x) = F(x)$ is the probability for a maximum payoff x . Additionally, we know that there are two machines (machine A and B) that determine the payoff.

Question: What is the probability for a maximum payoff of x if machine A is used?

In other words, what is the probability of interest if the condition “Machine A is used” is applied? Hence, the probability under consideration is also called **conditional probability** and written as

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

Accordingly one writes $P(X \leq x|B)$, if the condition “Machine B is used” is applied.

Question: What is the relationship between the **unconditional probability** $P(X \leq x)$ and the **conditional probabilities** $P(X \leq x|A)$ and $P(X \leq x|B)$?

To answer this question one has to clarify what the corresponding probabilities of using machine A or B are. Denoting these probabilities by $P(A)$ and $P(B)$ we have:

$$P(X \leq x) = P(X \leq x|A)P(A) + P(X \leq x|B)P(B)$$
$$F(x) = F(x|A)P(A) + F(x|B)P(B)$$

In this example there are two outcomes. The corresponding relationship can be extended to n discrete outcomes $\Omega = \{A_1, A_2, \dots, A_n\}$:

$$F(x) = F(x|A_1)P(A_1) + F(x|A_2)P(A_2) + \dots + F(x|A_n)P(A_n) \quad (10.1)$$

Until now we defined the conditions in terms of events and not in terms of random variables. An example for the latter one were if the payoff is determined by only one machine, but where the mode of operation for this machine is conditioned upon the payoffs' magnitude Z . In this case, the conditional distribution function is $F(x|Z = z)$, with $Z = z$ meaning that the random variable Z exactly takes the value z . For relating the unconditional and conditional probability we have to replace the sum by an integral, and the probability of the conditioning event by the corresponding

density function, as Z can have infinitely many values. For our example we obtain:

$$F(x) = \int_0^{\infty} F(x|Z = z)f(z)dz = \int_0^{\infty} F(x|z)f(z)dz$$

or generally

$$F(x) = \int F(x|Z = z)f(z)dz = \int F(x|z)f(z)dz \quad (10.2)$$

Another important property:

If the random variables X and Z are stochastically independent, we have

$$F(x|z) = F(x).$$

• Conditional density function

The conditional density function can be heuristically derived from the conditional distribution function in the same way as for the case of the unconditional density function: one simply replaces the unconditional probabilities by conditional probabilities. The **conditional density function** arises from

$$\lim_{h \rightarrow 0} \frac{P(X \leq x + h|A) - P(X \leq x|A)}{h} = f(x|A).$$

For finitely many conditions equation (10.1) becomes

$$f(x) = f(x|A_1)P(A_1) + f(x|A_2)P(A_2) + \cdots + f(x|A_n)P(A_n).$$

The relationship (10.2) turns to

$$f(x) = \int f(x|Z = z)f(z)dz = \int f(x|z)f(z)dz. \quad (10.3)$$

• Expectation

Consider again the payoff example.

Question: Which payoff would you expect “on average”?

Answer: $\int_0^\infty x f(x) dx$. For a payoff paid in n different discrete amounts, one would expect $\sum_{i=1}^n x_i P(X = x_i)$ on average. Each possible payoff is multiplied by its probability of entry and added up. It is not surprising that the result is denoted as expectation.

In general the **expectation** is defined as

$$E[X] = \int x f(x) dx, \quad \text{continuous } X,$$

$$E[X] = \sum x_i P(X = x_i), \quad \text{discrete } X.$$

- **Rules for the expectation** e.g. Appendix B in [Wooldridge \(2009\)](#).

1. For each constant c it holds that

$$E[c] = c.$$

2. For all constants a and b and all random variables X and Y it holds that

$$E[aX + bY] = aE[X] + bE[Y].$$

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

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

- **Conditional expectation**

So far we did not care for the machine that was used to create the payoff. If we are interested in the expected payoff of using machine

A , we have to calculate the **conditional expectation**

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

This is easily achieved by replacing the unconditional density $f(x)$ by the conditional density $f(x|A)$ and stating the condition in the notation of expectations accordingly. Analogously the expected payoff for machine B is determined as

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

In general one has for discrete conditioning events

$$E[X|A] = \int x f(x|A) dx, \quad \text{continuous } X,$$

$$E[X|A] = \sum x_i P(X = x_i|A), \quad \text{discrete } X,$$

and for continuous conditions

$$E[X|Z = z] = \int x f(x|Z = z) dx, \quad \text{continuous } X,$$

$$E[X|Z = z] = \sum x_i P(X = x_i|Z = z), \quad \text{discrete } X.$$

Remark: Frequently, the short versions are used as in [Wooldridge \(2009\)](#).

$$E[X|z] = \int x f(x|z) dx, \quad \text{continuous } X,$$

$$E[X|z] = \sum x_i P(X = x_i|z), \quad \text{discrete } X.$$

In accordance to the relationship of unconditional and conditional probabilities there is a similar relationship for unconditional and conditional expectations. The relationship is

$$E[X] = E[E[X|Z]]$$

which is denoted as **law of iterated expectations (LIE)**.

Sketch of proof:

$$\begin{aligned}
 E[X] &= \int x f(x) dx \\
 &= \int x \left[\int f(x|z) f(z) dz \right] dx \quad (\text{insert (10.3)}) \\
 &= \int \int x f(x|z) f(z) dz dx \\
 &= \int \underbrace{\int x f(x|z) dx}_{E[X|z]} f(z) dz \quad (\text{interchange } dx \text{ and } dz) \\
 &= \int E[X|z] f(z) dz \\
 &= E[E[X|Z]]
 \end{aligned}$$

In our example with 2 machines, the law of iterated expectations

yields

$$E[X] = E[X|A]P[A] + E[X|B]P(B).$$

This example also shows that the conditional expectations $E[X|A]$ and $E[X|B]$ are random variables. If they are weighted by the corresponding probabilities of entry $P(A)$ and $P(B)$, they yield $E[X]$. Suppose that, prior to the lottery, you only know both conditional expectations but not which machine is used. Then the expected payoff is equal to $E[X]$ and both conditional expectations are considered as random variables. After knowing what machine is used, the corresponding conditional expectation is the outcome of the random variable. This is a general property of conditional expectations.

- **Rules for conditional expectations**

e.g. Appendix B in [Wooldridge \(2009\)](#).

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)**

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

5. $E[Y|X] = E[E[Y|X, Z]|X]$.

6. If it holds that $E[Y|X] = E[Y]$, then it also holds that $Cov(X, Y) = 0$.
7. If $E[Y^2] < \infty$ and $E[g(X)^2] < \infty$ for an arbitrary function $g(\cdot)$, then the following inequalities hold:

$$E\{[Y - E[Y|X]]^2|X\} \leq E\{[Y - g(X)]^2|X\}$$
$$E\{[Y - E[Y|X]]^2\} \leq E\{[Y - g(X)]^2\}.$$

10.2 Important Rules of Matrix Algebra

Matrix addition

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{T1} & a_{T2} & \cdots & a_{TK} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1K} \\ c_{21} & c_{22} & \cdots & c_{2K} \\ \vdots & \vdots & & \vdots \\ c_{T1} & c_{T2} & \cdots & c_{TK} \end{pmatrix}.$$

If \mathbf{A} and \mathbf{C} are of the same dimension

$$\mathbf{A} + \mathbf{C} = \begin{pmatrix} a_{11} + c_{11} & a_{12} + c_{12} & \cdots & a_{1K} + c_{1K} \\ a_{21} + c_{21} & a_{22} + c_{22} & \cdots & a_{2K} + c_{2K} \\ \vdots & \vdots & & \vdots \\ a_{T1} + c_{T1} & a_{T2} + c_{T2} & \cdots & a_{TK} + c_{TK} \end{pmatrix}.$$

Matrix multiplication

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{T1} & a_{T2} & \cdots & a_{TK} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1L} \\ b_{21} & b_{22} & \cdots & b_{2L} \\ \vdots & \vdots & & \vdots \\ b_{K1} & b_{K2} & \cdots & b_{KL} \end{pmatrix}.$$

If the number of columns in \mathbf{A} is equal to the number of rows in \mathbf{B} , then the product $\mathbf{C} = \mathbf{AB}$ is defined and the following equality holds for every element in \mathbf{C}

$$c_{ij} = \begin{pmatrix} a_{i1} & \cdots & a_{iK} \end{pmatrix} \begin{pmatrix} b_{1j} \\ \vdots \\ b_{Kj} \end{pmatrix} = a_{i1}b_{1j} + \cdots + a_{iK}b_{Kj} = \sum_{l=1}^K a_{il}b_{lj}.$$

Caution: In general it holds that $\mathbf{AB} \neq \mathbf{BA}$.

Transpose of a matrix

Given the (2×3) -matrix (i.e. 2 rows, 3 columns)

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

the transpose of \mathbf{A} is the (3×2) -matrix

$$\mathbf{A}' = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}.$$

It holds that

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'.$$

Inverse of a matrix

Let \mathbf{A} be the $(K \times K)$ -matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KK} \end{pmatrix},$$

then the inverse of \mathbf{A} is \mathbf{A}^{-1} and is defined by

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_K = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

with \mathbf{I}_K as identity matrix of dimension $(K \times K)$.

The matrix \mathbf{A} is invertible if the rows respectively columns are linearly independent. In other words: No row (column) can be described as linear combination of the other rows (columns). Technically this is satisfied whenever the determinant of \mathbf{A} is unequal to zero.

Frequently, a noninvertible matrix is called singular.

The calculation of an inverse is better left to a computer. Only for matrices of 2 or 3 columns/rows, the calculation is of moderate complexity. Hence a manual calculation can be useful.

Special issue of a (2×2) matrix:

For a square (2×2) matrix

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

the **determinant** is computed as

$$\det(\mathbf{B}) = b_{11}b_{22} - b_{21}b_{12}$$

and the **inverse** as

$$\begin{aligned} \mathbf{B}^{-1} &= \frac{1}{\det(\mathbf{B})} \begin{pmatrix} b_{22} & -b_{12} \\ -b_{21} & b_{11} \end{pmatrix} \\ &= \frac{1}{b_{11}b_{22} - b_{21}b_{12}} \begin{pmatrix} b_{22} & -b_{12} \\ -b_{21} & b_{11} \end{pmatrix}. \end{aligned}$$

Example:

$$\mathbf{C} = \begin{pmatrix} 0 & 2 \\ 1 & -1 \end{pmatrix}, \quad \text{with} \quad \det(\mathbf{C}) = 0 \cdot (-1) - 1 \cdot 2 = -2$$

$$\mathbf{C}^{-1} = \frac{1}{-2} \begin{pmatrix} -1 & -2 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & 0 \end{pmatrix}$$

Check:

$$\mathbf{C}\mathbf{C}^{-1} = \begin{pmatrix} 0 & 2 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Reading: As supplement for matrix algebra and its implementation in the multiple linear regression framework see Appendices D, E.1 in [Wooldridge \(2009\)](#).

10.3 Rules for Matrix Differentiation

•

$$c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_T \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{pmatrix}$$

$$z = c'w = \begin{pmatrix} c_1 & c_2 & \cdots & c_T \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{pmatrix}$$

$$\frac{\partial z}{\partial w} = c$$



$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1T} \\ a_{21} & a_{22} & \cdots & a_{2T} \\ \dots & \dots & \dots & \dots \\ a_{T1} & a_{T2} & \cdots & a_{TT} \end{pmatrix}$$

$$z = w'Aw = \begin{pmatrix} w_1 & w_2 & \cdots & w_T \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1T} \\ a_{21} & a_{22} & \cdots & a_{2T} \\ \dots & \dots & \dots & \dots \\ a_{T1} & a_{T2} & \cdots & a_{TT} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{pmatrix}$$

$$\frac{\partial z}{\partial w} = (A' + A)w$$

10.4 Data for Estimating Gravity Equations

Legend for data in `importe_ger_2004_ebrd.txt`

• Countries and country codes

1	ALB	Albania	17	GBR	United Kingdom	33	NLD	Netherlands
2	ARM	Armenia	18	GEO	Georgia	34	NOR	Norway
3	AUT	Austria	19	GER	Germany	35	POL	Poland
4	AZE	Azerbaijan	20	GRC	Greece	36	PRT	Portugal
5	BEL	Belgium and Luxembourg	21	HRV	Croatia	37	ROM	Romania
6	BGR	Bulgaria	22	HUN	Hungary	38	RUS	Russia
7	BLR	Belarus	23	IRL	Ireland	39	SVK	Slovakia
8	CAN	Canada	24	ISL	Iceland	40	SVN	Slovenia
9	CHE	Switzerland	25	ITA	Italy	41	SWE	Sweden
10	CYP	Cyprus	26	KAZ	Kazakhstan	42	TKM	Turkmenistan
11	CZE	Czech Republic	27	KGZ	Kyrgyzstan	43	TUR	Turkey
12	DNK	Denmark	28	LTU	Lithuania	44	UKR	Ukraine
13	ESP	Spain	29	LVA	Latvia	45	USA	United States
14	EST	Estonia	30	MDA	Moldova	46	YUG	Serbia and Montenegro
15	FIN	Finland	31	MKD	Macedonia			
16	FRA	France	32	MLT	Malta			

Note:

This table is based on Table 1 in `gravity_data.pdf`.

Countries that feature only as origin countries:

BIH	Bosnia and Herzegovina
TJK	Tajikistan
UZB	Uzbekistan
CHN	China
HKG	Hong Kong
JPN	Japan
KOR	South Korea
TWN	Taiwan
THA	Thailand

- **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

- Commodity classifications: Trade flows are based on aggregating disaggregate trade flows according to the Standard International Trade Classification, Revision 3 (SITC, Rev.3) at the lowest aggregation levels (4- or 5-digit). Source: **UN COMTRADE**

- Without fuels and lubricants (i.e., specifically without petrol and natural gas products). Cut-off value for underlying disaggregated trade flows (at SITC Rev.3 5-digit level) is 500 US dollars.

- **Explanatory variables:**

Origin country

WDI_GDPUSDCR_O	Origin country GDP data; in current US dollars	World Bank - World Development Indicators
WDI_GDPPCUSDCR_O	Origin country GDP per capita data; in current US dollars	World Bank - World Development Indicators
WEO_GDPCR_O	Destination and origin country GDP data; in current US dollars	IMF - World Economic Outlook database
WEO_GDPPCCR_O	Destination and origin country GDP per capita data; in current US dollars	IMF - World Economic Outlook database
WEO_POP_O	Origin country population data	IMF - World Economic Outlook database
CEPII_AREA_O	area of origin country in km ²	CEPII
CEPII_COL45	dummy; d and o country have had a colonial relationship after 1945	CEPII
CEPII_COL45_REV	dummy; revised by “expert knowledge”	
CEPII_COLONY	dummy; d and o country have ever had a colonial link	CEPII
CEPII_COMCOL	dummy; d and o country share a common colonizer since 1945	CEPII
CEPII_COMCOL_REV	dummy; revised by “expert knowledge”	
CEPII_COMLANG_ETHNO	dummy; d and o country share a language	CEPII
CEPII_COMLANG_ETHNO_REV	at least spoken by 9% of each population	
CEPII_COMLANG_OFF	dummy; d and o country share common official language	CEPII
CEPII_CONTIG	dummy; d and o country are contiguous (neighboring countries)	CEPII
CEPII_DISINT_O	internal distance in origin country	CEPII
CEPII_DIST	geodesic distance between d and o country	CEPII
CEPII_DISTCAP	distance between d and o country based on capitals $0.67\sqrt{area/\pi}$	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	ISO codes in three characters of origin country	CEPII
EBRD_TFES_O	EBRD measure of foreign trade and payments liberalisation of o country	EBRD

Destination country

WDI_GDPUSDCR_D	Destination country GDP data; in current US dollars	World Bank - World Development Indicators
WDI_GDPPCUSDCR_D	Destination country GDP per capita data; in current US dollars	World Bank - World Development Indicators
WEO_GDPCR_D	Destination and origin country GDP data; in current US dollars	IMF - World Economic Outlook database
WEO_GDPPCCR_D	Destination and origin country GDP per capita data; in current US dollars	IMF - World Economic Outlook database
WEO_POP_D	Destination country population data	IMF - World Economic Outlook database

Notes: The EBRD measures reform on a scale between 1 and 4+ (=4.33); 1 represents no or little progress; 2 indicates important progress; 3 is substantial progress; 4 indicates comprehensive progress, while 4+ indicates countries have reached the standards and performance norms of advanced industrial countries, i.e., of OECD countries. By construction, this variable is ordered qualitative rather than cardinal.

- **Thanks:** to Richard Frensch, **IOS - Leibniz-Institut für Süd- und Südosteuropaforschung, Regensburg** und **Universität Regensburg**, for providing the data set.

- EViews-Commands to extract selected data from main workfile:
 - to select observations of countries that export to Kazakhstan:
in workfile: Proc → Copy/Extract from Current Page
→ By Value to New Page or Workfile:
in Sample - observations to copy: @all if (iso_d="KAZ"). Objects to copy:
select. Page Destination: select.
 - to select observations for one period, e.g. 2004:
as above but: in Sample - observations to copy: 2004 2004
 - to select observations for trade flows from Germany to Kazakhstan for all periods:
as above but: in Sample - observations to copy: @all if (iso_o="KAZ") and
(iso_d="GER")
- **Websites** **CEPII**

10.5 R Program for Empirical Examples

```
##### EOE_ws19_Emp_Beispiele.R #####
#
#####
#####
# R-Programm zum Reproduzieren der empirischen Beispiele in den
# Folien Einführung in die Ökonometrie, Universität Regensburg
# erstellt von Patrick Kratzer, Roland Weigand und Rolf Tschernig
# Stand: 18.10.2019, 25.08.2020

#####
#####
# Hinweise:
# a) Um das Skript ausführen zu können, werden folgende Daten benötigt:
#   - Handelsströme-Beispiele "importe_ger_2004_ebrd.txt",
#   - Löhne-Beispiele "wage1.txt"
#   - Zigaretten-Beispiele "smoke.txt"
# b) Die Daten-Dateien müssen im gleichen Verzeichnis wie das Programm liegen
#   und das working directory muss dem Verzeichnis entsprechen, von dem aus
#   dieses R-Programm aufgerufen wurde. Dazu muss das working directory
#   definiert werden, siehe Hinweise ab Zeile 82.
# c) Zunächst werden die Funktionen stats und SelectCritEviews definiert.
#   Anschließend beginnt das Hauptprogramm in Zeile 75.
# d) Graphiken können als PDF-Datei ausgegeben werden, siehe Hinweis Zeile 81.

#####
#                               Beginn Definition Funktionen
#####
```

```
##### Funktion stats #####
# Nützliche Funktion, die bei Eingabe eines Vektors statistische Kennzahlen liefert
# analog zu EViews-Output von "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))
}

##### Ende #####

##### Funktion SelectCritEviews #####
# Funktion zur Berechnung von Modellselektionskriterien wie 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))
}

##### Ende #####

#####
#               Ende Definition Funktionen
#####

#####
#               Beginn Hauptprogramm
#####

##### Bestimme Parameter für das R-Programm #####

save.pdf <- 1          # 1=Erstelle PDFs von Graphiken, 0=sonst
WD       <- ""

# Working Directory, in dem die R-Datei und die
# Daten liegen
# MUSS INDIVIDUELL ANGEPASST WERDEN
# In RStudio über "Session" -> "Set Working Directory"
# -> "To Source File Location" zu bestimmen

```

```

# Beispiele: WD = "~/EOE/R-code" oder
# WD = "C:/users/r-code"

##### Ende Parameter Eingabe #####

# Folgende Libraries werden im Verlauf geladen: car,lmtest

# Falls diese nicht installiert sind, werden diese zunächst installiert:
if (!require(car)){
  install.packages("car")
}
if (!require(lmtest)){
  install.packages("lmtest") # benötigt ab Folie 194
}
if (!require(xtable)){
  install.packages("xtable") # benötigt ab Folie 290
}

# Festlegung des Arbeitsverzeichnisses (working directory)
# in welchem sich das R-Program und die Daten befinden
setwd(WD)          # setze es als Working Directory

##### Einlesen der Handelsströme-Daten als data frame
daten_all <-read.table("importe_ger_2004_ebrd.txt", header = TRUE)
# Zuweisung der Variablennamen und
# Eliminieren der Beobachtung Exportland: GER, Importland: GER
attach(daten_all[-20,])

# Zum Ausprobieren, falls importe_ger_2004_ebrd.txt schon eingelesen worden ist
stats(trade_0_d_o)

```

```
##### Einlesen der wage-Daten als data frame
attach(read.table("wage1.txt", header = TRUE))

#####

##### Histogram, Folie 6 #####

# Für Ausgabe im PDF Format Dateiname definieren
if (save.pdf) pdf("r_imports_barplot.pdf", 12, 6)
# Histogramm
barplot(trade_0_d_o*10^-9, names.arg = iso_o, las = 2, col = "lightblue",
        main = "Imports to Germany in 2004 in Billions of US-Dollars")
# Device schließen
if (save.pdf) dev.off()

#####

##### Scatterplot, Folien 8, 11, 60 #####

# Für Ausgabe im PDF Format Dateiname definieren
if (save.pdf) pdf("scatter.pdf", height=6, width=6)
# Scatterplot der beiden Variablen
plot(wdi_gdpusdcr_o, trade_0_d_o, col = "blue", pch = 16)
# Device schließen
if (save.pdf) dev.off()

#####
```

```
#####
# Scatterplot mit (linearer) Regressionsgerade,
#           Folien 12, 61

# Für Ausgabe im PDF Format Dateiname definieren
if (save.pdf) pdf("plot_wdi_vs_trade.pdf", height=4, width=4)
# KQ-Schätzung eines einfachen linearen Regressionsmodells, abgespeichert in ols
ols_trade_wdi <- lm(trade_0_d_o ~ wdi_gdpusdcr_o)
# Scatterplot der beiden Variablen
plot(wdi_gdpusdcr_o, trade_0_d_o, col = "blue", pch = 16)
# Einzeichnen der linearen Regressionsgeraden mittels abline
abline(ols_trade_wdi, col = "red")
# Hinzufügen einer Legende
legend("bottomright", "Lineare Regression", col = "red", lty = 1, bty = "n")
# Device schließen
if (save.pdf) dev.off()

#####
#####

# Scatterplot mit linearer Regressionsgerade
#           und nichtlinearer Regressionsgerade
#           in Punkten an Beobachtungen dargestellt, Folie 13

if (save.pdf) pdf("r_imports_scatter_nonlin.pdf", 4, 4)
# Schätzung Regressionsmodell mit quadratischem Regressor
ols_nonlin <- lm((trade_0_d_o) ~ wdi_gdpusdcr_o + I(wdi_gdpusdcr_o^2)) # quadr.
# Definiere quadratische Funktion mit geschätzten Parametern
fx <- function(x){ols_nonlin$coefficients[1] +
```



```

      ols_nonlin$coefficients[2]*x + ols_nonlin$coefficients[3]*x^2}
# Erstelle Scatterplot
plot(wdi_gdpusdcr_o, trade_0_d_o, col = "blue", pch = 16)
# Füge lineare Regressionsgerade dazu
abline(ols_trade_wdi, col = "red")
# Füge Prognosepunkte der quadratischen Regression dazu
lines(wdi_gdpusdcr_o, fx(wdi_gdpusdcr_o),
      col = "green", type="p", pch = 16)
# Erstelle Legende
legend("bottomright",
      c("linear regression", "nonlinear regression"),
      col = c("red", "green"), lty = c(1,2), bty = "n")
if (save.pdf) dev.off()

#####
#####

# Handelsstrom von USA -> Deutschland, Folie 27
# aus anderer Datei %RT1920

#####
#####

# Regressionsoutput Handelsbeispiel Folie 61
# siehe auch Folie 12

# Anzeige der Ergebnisse der einfachen linearen Regression
summary(ols_trade_wdi)

#####
#####

```

```
#####
# Regressionsoutput Folie 65

summary(lm(wage ~ educ))

#####
#####

# Regressionsoutput Außenhandelsbeispiel Folie 88

summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o)))

#####
#####

# Regressionsoutput Außenhandelsbeispiel Folie 103

summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist)))

#####
#####

# Regressionsoutput Lohnbeispiel Folie 115

summary(lm(log(wage) ~ educ))

#####
#####
```

```

# Regressionsoutput Lohnbeispiel Folie 117

summary(lm(log(wage) ~ educ + exper))

#####

#####

# Bestimmung der Informationskriterien, Folie 177
# Anwendung der Funktion "SelectCritEviews" auf vier
# verschiedene Modelle:

model_1 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o))
coef(model_1)
SelectCritEviews(model_1)

model_2 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))
coef(model_2)
SelectCritEviews(model_2)

model_3 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o)
coef(model_3)
SelectCritEviews(model_3)

model_4 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o + log(cepii_area_o))
coef(model_4)
SelectCritEviews(model_4)

#####

```

```
#####
# t-Statistik in Eviews, Folie 195, 205

model_wage_m <- lm(wage ~ 1)
summary(model_wage_m)
# t-Statistik für H_0: mu=6
# mit gerundeten Werten
(5.896 - 6) / 0.161

# mit exakten Werten aus KQ-Schätzung
(coef(summary(model_wage_m))[1] - 6) / coef(summary(model_wage_m))[2]

# mit package car
sqrt(linearHypothesis(model_wage_m, c("(Intercept)=6"))$F[2])

# für Seite 205
# t-Statistik für H_0: mu=5.6
# mit gerundeten Werten
# mit gerundeten Werten
(5.896 - 5.6) / 0.161
# mit exakten Werten
(coef(summary(model_wage_m))[1] - 5.6) / coef(summary(model_wage_m))[2]

#####
#####

# Histogram von "wage", Folie 197, 283

if (save.pdf) pdf("r_wage_hist.pdf", 4, 4)
```

```
hist(wage, breaks = 20, col = "lightblue", prob = T)
curve(dnorm(x, mean = mean(wage), sd = sd(wage)),
      from = -5, to = 25, add = T, col = "red", lty = 2, lwd = 2)
legend("topright", "theoretical\nnormal distribution", col = "red",
      lwd = 2, lty = 2, bty = "n")
box()

if (save.pdf) dev.off()

# Ausgabe der deskriptiven Statistiken und Test auf Normalverteilung
stats(wage)

#####
#####

# Gravitationsgleichung, Folie 227

summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes_o))

#####
#####

# Visualisierung der Residuen, Folie 228

model_3 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o)
resid_model_3 <- model_3$resid
trade_0_d_o_fit <- model_3$fitted
```

```

if (save.pdf) pdf("r_resid_model_3.pdf", 5, 3)

par(mfrow = c(1,2))
plot(trade_0_d_o_fit, resid_model_3, col = "blue", pch = 16, main = "Scatterplot")
hist(resid_model_3, breaks = 20, col = "lightblue", prob = T, main = "Histogram")
curve(dnorm(x, mean = mean(resid_model_3), sd = sd(resid_model_3)),
      from = -3, to = 3, add = T, col = "red", lty = 2, lwd = 2)
legend("topleft", "theoretical\nnormal distribution", col = "red", lwd = 2,
      lty = 2, bty = "n")
box()
if (save.pdf) dev.off()

# statistische Auswertung der Residuen
stats(resid_model_3)

#####
#####

# Outputzeile auf Folie 230

summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes_o))
# Befehlszeile für log(wdi_gdpusdcr_o) reinkopieren

#           Estimate Std. Error t value Pr(>|t|)
# log(wdi_gdpusdcr_o)  0.94066    0.06134  15.335 < 2e-16 ***

# t-Statistik für gerundete Werte im Output
(teststat <- (0.94066 - 1)/0.06134)

#####

```

```
#####
# Befehl für Quantil der t-Verteilung auf Folie 230
(crit <- qt(0.975, df = 49 - 3 - 1))

#####
#####

# Outputzeilen auf Folie 232, 233

(pval <- 2 * pt(teststat, df = 49 - 3 - 1))    # (Hälfte von 9.262691e-08)

(summary(model_3)$coef[3,])

# t-Statistik (basierend auf Output)
(teststat2 <- (-9.703183e-01 - 0) / 1.526847e-01)

# kritischer Wert
(crit <- qt(0.95, df = 49 - 3 - 1))

# p-Wert
(pval <- pt(teststat2, df = 49 - 3 - 1))
#####
#####

# Befehle für Folien 245, 246

(crit <- qt(1-0.05/2, df = 49 - 3 - 1))
```

```

summary(lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes_o))

#           Estimate Std. Error t value Pr(>|t|)
# log(wdi_gdpusdcr_o)  0.94066    0.06134  15.335 < 2e-16 ***

# Konfidenzintervall
(0.94066 - 2.014103* 0.06134)
(0.94066 + 2.014103* 0.06134)
#####
#####

# Regressionsoutput auf Folie 250

marketing_102 <-read.table("marketing_102.txt", header = TRUE)

#summary(lm(labsatz ~ log(preis) + log(preis_qualig) +
#           log(preis_qualimo), data=marketing_102))
S <- marketing_102$absatz
P <- marketing_102$preis
P_K1 <- marketing_102$preis_qualig
P_K2 <- marketing_102$preis_qualimo
summary(lm(log(S) ~ log(P) + log(P_K1) + log(P_K2)))

#####
#####

# Regressionsoutput auf Folie 252
# verwendet Daten von Folie 250
summary( lm( log(S) ~ log(P) + log(P_K1) + I(log(P_K1)+log(P_K2)) ) )

```



```
#####
#####

# Regressionsoutput zu Folie 254, Fortsetzung des Außenhandelsbeispiels

model_4 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o + log(cepii_area_o))
summary(model_4)

#####
#####

# Regressionsoutput auf Folie 257

model_2 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist))
summary(model_2)

#####
#####

# F-Test auf Folie 264

model_2_sum <- summary(model_2)
(SSR_model_2 <- (model_2_sum$sigma)^2 * model_2_sum$df[2])

model_4_sum <- summary(model_4)
(SSR_model_4 <- (model_4_sum$sigma)^2 * model_4_sum$df[2])

# F-Statistik
( (SSR_model_2 - SSR_model_4)/2 ) /
```

```

(SSR_model_4/model_4_sum$df[2])
#####
#####

# F-Test auf Folie 266/267

library(car)

1 - pf(5.24077, df1 = 2, df2 = 44)

model_4 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o + log(cepii_area_o))
linearHypothesis(model_4, c("ebrd_tfes_o = 0", "log(cepii_area_o) = 0"))

#####
#####

# Hinweis zu Folie 269, Ermittlung der Kovarianzmatrix

model_4 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o + log(cepii_area_o))
vcov(model_4)
coef(summary(model_4))[,2]^2

#####
#####

# Konfidenzellipse auf Folie 272

if (save.pdf) pdf("r_conf_ellipse.pdf", 6, 6)

```

```

model_4 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o + log(cepii_area_o))

confidenceEllipse(model_4, which.coef = c(4, 5), levels = 0.95,
                  main = "confidence ellipse", col = "blue")
abline(v = confint(model_4, "ebrd_tfes_o", level = 0.95), lty = 2,
       col = "red", lwd = 2)
abline(h = confint(model_4, "log(cepii_area_o)", level = 0.95), lty = 2,
       col = "red", lwd = 2)

if (save.pdf) dev.off()

#####
#####

# Regressionsoutput auf Folie 277, 278

model_4 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) +
              ebrd_tfes_o + log(cepii_area_o))
model_4_h0 <- lm(log(trade_0_d_o)-0.5*ebrd_tfes_o ~ log(wdi_gdpusdcr_o) +
                 log(cepii_dist))

summary(model_4_h0)

# F-Statistik auf Basis der Outputs
(SSR_model_4 <- (model_4_sum$sigma)^2 * model_4_sum$df[2])
(SSR_model_4_h0 <- (summary(model_4_h0)$sigma)^2 * summary(model_4_h0)$df[2])
( (SSR_model_4_h0 - SSR_model_4)/2 ) /
  (SSR_model_4/model_4_sum$df[2])

```

```

# F-Statistik mit library(car)
linearHypothesis(model_4, c("ebrd_tfes_o = 0.5", "log(cepii_area_o) = 0"))

#####
#####

# Folie 281: siehe Folie 177

#####
#####

# Folie 284, Dichte der Chi-Quadrat(1)-Verteilung
if (save.pdf) pdf("r_chi_2_1_verteilung.pdf", 6, 3)

curve(dchisq(x, df = 1), from = 0, to = 8, col = 2, ylab = "f(x)", ylim = c(0, 1.5),
      main = expression(paste(chi^2, "(1) - density function")))
abline(v=0)
if (save.pdf) dev.off()

# Folie 284, Dichte und Verteilung verschiedener Chi-Quadrat-Verteilungen
# (nicht in Folien)

if (save.pdf) pdf("r_chi_2_verteilung.pdf", 6, 3)

par(mfrow = c(1, 2))

curve(dchisq(x, df = 1), from = 0, to = 8, col = 1, ylab = "f(x)", ylim = c(0, 0.5),
      main = expression(paste(chi^2, " - density function")))
lines(c(-1, 0), c(0, 0), col = 1)

```

```

grid()
curve(dchisq(x, df = 2), from = 0, to = 8, col = 2, add = T)
curve(dchisq(x, df = 3), from = 0, to = 8, col = 3, add = T)
curve(dchisq(x, df = 5), from = 0, to = 8, col = 4, add = T)
curve(dchisq(x, df = 10), from = 0, to = 8, col = 5, add = T)
legend("topright", c("df = 1", "df = 2", "df = 3", "df = 5", "df = 10"),
      col = 1:5, lty = 1, bty = "n")
curve(pchisq(x, df = 1), from = 0, to = 8, ylab = "F(x)", col = 1, ylim = c(0, 1),
      main = expression(paste(chi^2, " - distribution function")))
lines(c(-1, 0), c(0, 0), col = 1)
grid()
curve(pchisq(x, df = 2), from = 0, to = 8, col = 2, add = T)
curve(pchisq(x, df = 3), from = 0, to = 8, col = 3, add = T)
curve(pchisq(x, df = 5), from = 0, to = 8, col = 4, add = T)
curve(pchisq(x, df = 10), from = 0, to = 8, col = 5, add = T)
# legend

if (save.pdf) dev.off()
#####
#####

# Monte Carlo Simulation auf Folien 290, 291, 292
if (save.pdf) pdf("r_mcarlo.pdf", 6, 4)
par(mfrow = c(2, 3))
set.seed(12345) # setze Random seed (für Replizierbarkeit)

reps <- 1000 # Anzahl der Replikationen
n <- c(10, 30, 50, 100, 500, 1000) # Stichprobenumfang für die 6 Auswertungen

means <- matrix(NA, nrow = reps, ncol = 6) # Initialisierung der

```

```

# Matrix mit den simulierten Mittelwerten

for(j in 1:6)
{
  for(i in 1:reps)
  {
    means[i,j] <- mean(3 + (rnorm(n[j])^2-1)*2^-0.5) # Simulation der Mittelwerte
  }
  hist(means[,j], breaks = 30, freq = F, xlab = "", # graphische Ausgabe
       col = "lightblue", main = paste("n = ",n[j])) # der Schätzrealisationen
}
if (save.pdf) dev.off()

# Erstelle Tabelle mit Mittelwerten und Standardabweichungen
fx <- function(x) {c(mean(x), sd(x))}
table_output <- apply(means, 2, fx)
# füge Zeile mit wahren Standardabweichungen des Schätzers hinzu
table_output <- rbind(table_output,sqrt(1/n))
# gebe Spalten und Zeilen Namen
rownames(table_output) <- c("Mittelwerte", "Standardabweichungen",
                           "theoret. Standardabw. geg. DGP")
colnames(table_output) <- paste0("n = ",n)
# erstelle Latex-Code für Tabelle
xtable(t(table_output), digits=6)
# Lösche Matrix means aus Simulation
rm(means)

#####
#####

```

```

# Koeffizienten von Modell 3, Folie 313

model_3 <- lm(log(trade_0_d_o) ~ log(wdi_gdpusdcr_o) + log(cepii_dist) + ebrd_tfes_o)
coef(model_3)

#####
#####

# Modell 5 auf Folie 318
model_5 <- lm(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))
summary(model_5)

#####
#####

# Programm zu Folie 320
if (save.pdf) pdf("r_bib_elasticity.pdf", 3, 3)

# Modell 5:
model_5 <- lm(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))

# Generiere die Elastizitäten für verschiedene BIPs
elast_gdp <- model_5$coef[2] + 2* model_5$coef[3]*log(wdi_gdpusdcr_o)

# Erstelle Scatterplot
plot(wdi_gdpusdcr_o, elast_gdp, pch = 16, col = "blue", main = "GDP-Elasticity")

if (save.pdf) dev.off()

```

```
#####
#####

# Regressionsoutput Folie 324, wage-Beispiel

ols <- lm(log(wage) ~ female + educ + exper + I(exper^2) + tenure + I(tenure^2))
summary(ols)

#####
#####

# Regressionsoutput Folie 330, wage-Beispiel

femmarr <- female * married
malesing <- (1 - female) * (1 - married)
malemarr <- (1 - female) * married

ols <- lm(log(wage) ~ femmarr + malesing + malemarr + educ + exper + I(exper^2) + tenure + I(tenure^2))
summary(ols)

#####
#####

# Fortsetzung des Lohnbeispiels, Folie 337

ols <- lm(log(wage) ~ female + educ + exper + I(exper^2) + tenure + I(tenure^2) + I(female*educ))
summary(ols)

#####
```



```
#####  
  
# Fortsetzung Außenhandelsbeispiel, Folie 372 ff.  
  
# R-Programm zur FGLS-Schätzung, Kapitel 8 Heteroskedastie  
# Florian Brezina, PK, 19.02.2011  
# mit Datei importe_ger_2004_ebrd.txt  
  
# Daten einlesen und 20. Beobachtung (Germany) entfernen  
# daten <- read.table("importe_ger_2004_ebrd.txt", header = TRUE)[-20,]  
# attach(daten)  
  
# definiere Variablen  
  
# definiere log der abhängigen Variable  
log_imp <- log(trade_0_d_o)  
  
### Erster Schritt a) KQ-Regression und Berechnung der Residuen  
  
# KQ-Regression  
eq_ols_model5 <- lm(log_imp ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +  
                    log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))  
  
# Berechne Residuen  
res_ols_model5 <- eq_ols_model5$resid  
  
# Berechne gefittete/angepasste Werte  
fit_ols_model5 <- fitted.values(eq_ols_model5)  
  
# Plote die Residuen gegen die gefitteten Werte, um zu untersuchen,
```

```
# ob Heteroskedastie vorliegen könnte
dev.off()
plot(fit_ols_model5, res_ols_model5, pch = 16)

### Erster Schritt b) bis d)

# Quadriere die Residuen und logarithmiere sie anschließend
ln_u_hat_sq <- log(res_ols_model5^2)

# Schätze die Varianzgleichung
eq_h_model5 <- lm(ln_u_hat_sq ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +
                 log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o))

# Berechne die gefitteten Werte der logarithmierten Residuenanalyse
ln_u_hat_sq_hat <- fitted.values(eq_h_model5)

# Berechne die h's aus den gefitteten Werten der Varianzregression
h_hat <- exp(ln_u_hat_sq_hat)

### Zweiter Schritt: FGLS-Schätzung

# Schätze FGLS mit den gewichteten weights = 1/h_hat
eq_fgls_model5 <- lm(log_imp ~ log(wdi_gdpusdcr_o) + I((log(wdi_gdpusdcr_o))^2) +
                 log(cepii_dist) + ebrd_tfes_o + log(cepii_area_o),
                 weights = 1/h_hat)
summary(eq_fgls_model5)

# Berechne die gefitteten Werte aus FGLS
fit_fgls_model5 <- fitted.values(eq_fgls_model5)
```

```

# Berechne die Residuen aus FGLS
res_fgls_model5 <- resid(eq_fgls_model5)

# Standardisierung der Residuen mittels der Gewichte
res_fgls_model5_star <- res_fgls_model5*h_hat^(-1/2)

# Plote die Residuen gegen die gefitteten Werte
plot(fit_fgls_model5, res_fgls_model5_star, pch = 16)

### KQ-Regression mit heteroskedastie-robusten Standardfehlern
library(lmtest)
eq_white_model5 <- coeftest(eq_ols_model5, vcov=hccm(eq_ols_model5,type="hc1"))

# Graphiken/Outputs für Skript
summary(eq_ols_model5)
summary(eq_h_model5)
summary(eq_fgls_model5)
eq_white_model5

if (save.pdf) pdf("r_model_5_fgls.pdf", 6, 3)
par(mfrow = c(1,2))
plot(fit_ols_model5, res_ols_model5, col = "blue", pch = 16, main = "OLS")
plot(fit_fgls_model5, res_fgls_model5_star, col = "blue", pch = 16, main = "FGLS")
if (save.pdf) dev.off()

##### Einschub #####
# ein paar Anmerkungen:
# R^2 und F-Statistik bei R-Output entsprechen den Ergebnissen
# für weighted statistics im Eviews-Output

```

```
# Nachbau des Eviews-Outputs:
w      <- h_hat^-0.5
w_scaled <- length(residuals(eq_fgls_model5)) / sum(w) * w
sum((w_scaled)) # Probe

log_imp_star <- log_imp * sqrt(w_scaled) # Wurzel!?
regressor_star <- model.matrix(eq_fgls_model5) * sqrt(w_scaled)

k <- ncol(model.matrix(eq_fgls_model5))-1
n <- length(resid(eq_fgls_model5))

# Weighted Statistics

# R-squared
summary(eq_fgls_model5)$r.squared
# Adjusted R-squared
summary(eq_fgls_model5)$adj.r.squared
# SSR
(SSR <- sum(w_scaled*(log_imp_star - regressor_star%%coef(eq_fgls_model5))^2))
# Mean dependent var
mean(log_imp * (w_scaled))
# S.D. dependent var
sd(log_imp * (w_scaled))
# S.E. of regression
sqrt(SSR/(n-k-1))

# Unweighted Statistics

# R-squared
```

```

(r_squared <- 1 - sum(residuals(eq_fgls_model5)^2) /
  sum((log_imp - mean(log_imp))^2))
# Adjusted R-squared
-k/(n-k-1) + (n-1)/(n-k-1)*r_squared
# Mean dependent var
mean(log_imp)
# S.D. dependent var
sd(log_imp)
# S.E. of regression
sqrt(sum(residuals(eq_fgls_model5)^2)/(n-k-1))
# Sum squared resid
sum(residuals(eq_fgls_model5)^2)

##### Ende Einschub #####

#####

#####

# Zigarettenbeispiel ab Folie 385

smoke_all <- read.table("smoke.txt", header = TRUE)

# Erster Schritt

# 1. KQ-Schätzung
ols_1 <- lm(cigs ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
  data=smoke_all)
summary(ols_1)

```

```

# 2. Speichere die Residuen
u_hat_cig <- resid(ols_1)

# 3. Logarithmiere die quadrierten Residuen
ln_u_sq <- log(u_hat_cig^2)

# 4. Schätzung der Varianzregression mittels KQ führt zu
ols_2 <- lm(ln_u_sq ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
            data=smoke_all)
summary(ols_2)
# Speichere die Residuen
h_hat_cig <- exp(ln_u_sq - resid(ols_2))
#
# Zweiter Schritt

# Gewichtete KQ-Schätzung mit den Gewichten h_hat_cig^(-1)
ols_3 <- lm(cigs ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
            weights = h_hat_cig^(-1), data=smoke_all)
summary(ols_3)

# Anmerkung: Im Vergleich zu EViews fehlen einige Statistiken. Siehe Hinweise
#           zu Folie 372 zu deren Berechnung

#####
#####

# Fortsetzung des Zigarettenbeispiels auf Folie 396

ols <- lm(cigs ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
          data=smoke_all)

```

```

u_hat_sq <- resid(ols)^2
summary(lm(u_hat_sq ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
          data=smoke_all))

#####
#####

# Fortsetzung Zigarettenbeispiel mit White Test, Folie 401 ff.

# Definition von Funktion für White Test
##### Beginn Funktion whitetest #####
# Function to conduct White test including and without cross terms
# Specification of test equations as in EViews
# Roland Weigand, 2011_01_26, Rolf Tschernig, 2019_10_18, 2020_08_25 (LM test)
# Input:
#     model_est      lm object with estimated model
#     crossterms     1: include cross terms, 0, do not include them
# Output: a list with the following components
#     ftest_result  a vector containing the F statistic, the
#                   degrees of freedom and the p value
#     lmtest_result a vector containing the LM statistic,
#                   the degrees of freedom and the p value
#     test_eq       an lm object with the results of the White regression

whitetest <- function(model_est, crossterms=1){

  # Daten aus model extrahieren
  dat <- model_est$model           # dat is dataframe
  dat$resid_sq <- model_est$resid^2 # resid_sq is added to dataframe

```

```

# Formel für die Hilfsregression erstellen
regr <- attr(model_est$terms, "term.labels")
if (crossterms){
  form <- as.formula(paste("resid_sq ~ (", paste(regr, collapse=" + "), ")^2 +",
                          , paste("I(",regr,"^2)", collapse=" + ") ) )
} else {
  form <- as.formula(paste("resid_sq ~ ", paste("I(",regr,"^2)",
                                              collapse=" + ") ) )
}

# Hilfsregression schätzen
test_eq <- lm(form, data=dat)

# Overall F-Test
fstat <- summary(test_eq)$fstatistic
# LM statistic
lmstat <- length(summary(test_eq)$residuals) * summary(test_eq)$r.squared

# Ergebnis berechnen und ausgeben
ftest_result <- c(fstat[1], fstat[2], fstat[3],
                 pf(fstat[1], fstat[2], fstat[3], lower.tail = FALSE))
names(ftest_result) <- c("F Statistic","df1","df2"," p Value")
lmtest_result <- c(lmstat, summary(test_eq)$df[1] - 1,
                  pchisq(lmstat, summary(test_eq)$df[1] - 1, lower.tail = FALSE))
names(lmtest_result) <- c("LM Statistic", "df", "p Value")
result <- list(lmtest_result = lmtest_result, ftest_result = ftest_result,
              test_eq = test_eq)
return(result)
}
##### Ende Funktion whitetest #####

```



```

# Anwendung der Funktion

ols <- lm(cigs ~ lincome + lcigpric + educ + age + I(age^2) + restaurn,
         data=smoke_all)
ols_white <- whitetest(ols)
# gebe F-Testergebnis aus
ols_white$ftest_result
# gebe LM-Testergebnis aus
ols_white$lmtest_result
# gebe Testgleichung aus
summary(ols_white$test_eq)

#####
#####

# BP-Test auf Folie 403, Fortsetzung des Außenhandelsbeispiels

bptest(eq_ols_model5)

#####
#####

# White-Test auf Folie 404, 405 (ohne Kreuzprodukte)
# führe White-Test durch, Funktion whitetest() auf Folie 399 definiert
ols_model5_white <- whitetest(eq_ols_model5, crossterms=0)
# gebe F-Testergebnis aus
ols_model5_white$ftest_result
# gebe LM-Testergebnis aus
ols_model5_white$lmtest_result

```

```

# gebe Testgleichung aus
summary(ols_model5_white$test_eq)

#####
#####

# Folie 406

# Breusch Pagan Test für FGLS (funktioniert mit "bptest" leider nicht
# Ergebnisse entsprechen denen von Eviews
log_imp_star <- log_imp * (w_scaled)
regressor_star <- model.matrix(eq_fgls_model5)[,-1] * (w_scaled)
u_star_sq <- (resid(eq_fgls_model5) * (w_scaled))^2

bpg_eq_fgls <- lm(data.frame(cbind(u_star_sq, regressor_star)))

t_bpg_fgls <- summary(bpg_eq_fgls)$r.squared * n
bp_fgls_res <- c(t_bpg_fgls,
                1-pchisq(t_bpg_fgls, df = k))
names(bp_fgls_res) <- c("LM-Teststatistik", "p-Wert")
bp_fgls_res
summary(bpg_eq_fgls)

#####
#####

# Folie 407 und 408

# White-Test manuell, erfordert Variablen definiert für Folie 404, 405

```

```
w_scaled_sq <- w_scaled^2
regressor_white <- data.frame(w_scaled_sq, regressor_star^2)

white_eq_fgls <- lm(cbind(u_star_sq , regressor_white))

t_white_fgls <- summary(white_eq_fgls)$r.squared * n
white_fgls_res <- c(t_white_fgls,
                   1-pchisq(t_white_fgls, df = k+1))
names(white_fgls_res) <- c("LM-Teststatistik", "p-Wert")
white_fgls_res
summary(white_eq_fgls)

#####
#####
#                               ENDE
#####
#####
```

Listing 10.1: ../R_code/EOE_ws19_Emp_Beispiele.R

Bibliography

Anderson, J. E., and E. v. Wincoop (2003), “Gravity with Gravitas: A Solution to the Border Puzzle,” *The American Economic Review*, 93, 170–192. 102

Angrist, J. D., and J.-S. Pischke (2015), *Mastering Metrics. The Path from Cause to Effect*, Princeton University Press, Princeton. 24

Casella, G., and R. L. Berger (2002), *Statistical Inference*, 2nd edn., Duxbury - Thomson. II

Davidson, R., and J. G. MacKinnon (2004), *Econometric Theory and Methods*, Oxford University Press, Oxford. 259

Fратиани, M. (2007), “The gravity equation in international trade,” Tech. rep., Dipartimento di Economia, Universita Politecnica delle Marche. 102, 224

Pindyck, R. S., and D. L. Rubinfeld (1998), *Econometric models and economic forecasts*, Irwin McGraw-Hill. 19

Stock, J. H., and M. W. Watson (2007), *Introduction to Econometrics*, Pearson, Boston, Mass. 15, 20, 22, 23

Wooldridge, J. M. (2009), *Introductory Econometrics. A Modern Approach*, 4th edn., Thomson South-Western. 20, 30, 34, 58, 64, 77, 99, 105, 115, 121, 130, 140, 148, 152, 160, 194, 198, 206, 208, 212, 223, 231, 252, 262, 263, 281, 284, 288, 299, 302, 310, 323, 339, 347,

366, 385, 409, 423, XVI, XVIII, XXI, XXIX