Statistics for Data Analytics

Sven Otto

February 6, 2024

Table of contents

Welcome to the course! 3
Course Materials . . . 3
Literature . . . 3
Preparation . . . 4
Assessment . . . 4
Communication . . . 4
Important Dates . . . 4
Timetable . . . 5
R-Packages . . . 5

1 Introduction 7

2 Probability 9
2.1 Random experiments . . . 9
2.2 Random variables . . . 9
2.3 Probability function . . . 11
2.4 Distribution . . . 13
2.5 Cumulative distribution function . . . 14
2.6 Probability density function . . . 19
2.7 Expected value . . . 20
2.7.1 Expectation of a discrete random variable . . . 20
2.7.2 Expectation of a continuous random variable . . . 21
2.7.3 Expectation for general random variables . . . 21
2.7.4 Properties of the expected value . . . 22
2.8 Descriptive features of a distribution . . . 23
2.8.1 Heavy-tailed distributions . . . 25
2.9 The normal distribution . . . 26
2.10 Additional reading . . . 26
2.11 R-codes . . . 27

3 Dependence 28
3.1 Multivariate random variables . . . 28
3.2 Bivariate random variables . . . 28
3.3 Bivariate distributions . . . 30
3.4 Correlation . . . 33

3.5 Independence . . . 34
3.6 Random vectors . . . 35
3.7 Conditional distributions . . . 35
3.8 Conditional expectation . . . 37
3.9 Law of iterated expectations . . . 40
3.10 Conditional variance . . . 40
3.11 Best predictor . . . 41
3.12 Combining normal variables . . . 42
3.12.1 χ^2-distribution . . . 42
3.12.2 F-distribution . . . 43
3.12.3 Student t-distribution . . . 43
3.12.4 Multivariate normal distribution . . . 44
3.12.5 R-commands for parametric distributions . . . 45
3.13 Additional reading . . . 45

4 Sampling **46**
4.1 Data . . . 46
4.2 Random sampling . . . 48
4.3 Dependent sampling . . . 50
4.4 Time series data . . . 51
4.5 Additional reading . . . 54
4.6 R-codes . . . 55

5 Estimation **56**
5.1 Parameters and estimators . . . 56
5.2 Population moments and sample moments . . . 56
5.3 Moment estimators . . . 57
5.4 Consistency . . . 63
5.5 The sample mean . . . 65
5.6 The sample variance . . . 67
5.7 Additional reading . . . 68
5.8 R-codes . . . 68

6 Confidence Intervals **69**
6.1 Estimation uncertainty . . . 69
6.2 Interval estimates . . . 70
6.3 Central limit theorem . . . 72
6.4 Standard errors . . . 74
6.5 Exact confidence intervals under normality . . . 76
6.6 Additional reading . . . 79
6.7 R-codes . . . 79

7 Hypothesis Testing **80**
7.1 Statistical hypotheses . . . 80
7.2 t-Test for the mean . . . 82
7.2.1 The normal i.i.d. case . . . 82
7.2.2 The non-normal i.i.d. case . . . 83
7.2.3 The time series case . . . 84
7.3 The p-value . . . 85
7.4 Power function . . . 86
7.5 One-sided t-test . . . 88
7.6 Testing for autocorrelation . . . 89
7.7 Additional reading . . . 90
7.8 R-codes . . . 90

8 Simulations **91**
8.1 The Monte Carlo principle . . . 91
8.2 Set up . . . 91
8.3 Monte Carlo algorithm . . . 92
8.4 Evaluate an estimator . . . 92
8.5 Evaluate a confidence interval . . . 95
8.6 Evaluate a hypothesis test . . . 96
8.7 Additional reading . . . 97
8.8 R-codes . . . 97

9 Least Squares **98**
9.1 The OLS principle . . . 98
9.2 Simple linear regression (k=2) . . . 100
9.3 Linear regression with k=3 . . . 103
9.4 Matrix notation . . . 105
9.5 Projection matrices . . . 106
9.6 Analysis of Variance . . . 109
9.7 Coefficients of determination . . . 109
9.8 Too many regressors . . . 112
9.9 Multicollinearity . . . 114
9.10 The dummy variable trap . . . 114
9.11 OLS without intercept . . . 118
9.12 Additional reading . . . 118
9.13 R-codes . . . 118

10 Regression Models **119**
10.1 The linear model . . . 119
10.2 Conditional mean independence . . . 120
1) Zero unconditional mean . . . 120
2) Linear best predictor . . . 120

3) Marginal effect interpretation . . . 121
4) Weak exogeneity . . . 121
10.3 Correlation and causation . . . 122
10.4 Nonlinearities . . . 124
10.5 The moment estimator . . . 126
10.6 Random sampling assumption . . . 127
10.6.1 Exogeneity and time series regression . . . 127
10.6.2 Heteroskedasticity . . . 128
10.6.3 Detecting heteroskedasticity . . . 129
10.7 Sampling mean and variance . . . 131
10.8 Consistency . . . 133
10.9 Efficiency . . . 133
10.10 Normality . . . 135
10.11 Asymptotic normality . . . 136
10.12 Additional reading . . . 137
10.13 R-codes . . . 137

11 Classical Inference **138**
11.1 Standardized OLS coefficients . . . 139
11.2 Exact confidence intervals . . . 139
11.3 Exact t-tests . . . 140
11.4 Regression outputs . . . 141
The `lm`-summary output . . . 141
Regression outputs in economic journals . . . 142
11.5 Multiple testing . . . 144
11.6 Exact F-tests . . . 145
11.7 Additional reading . . . 148
11.8 R-codes . . . 148

12 Robust Inference **149**
12.1 Heteroskedasticity-robust standard errors . . . 149
12.2 Robust confidence intervals . . . 152
12.3 Robust t-tests . . . 153
12.4 Robust F-tests . . . 155
12.5 Autocorrelation-robust standard errors . . . 155
12.6 Additional reading . . . 156
12.7 R-codes . . . 156

Welcome to the course!

Statistics for Data Analytics is an introductory graduate-level course in econometrics and statistical inference. We cover basic concepts of mathematical statistics, including estimation and inferential methods in linear models. The goal is to provide the theoretical foundation for data analysis and applied empirical work. Practical applications using the `R` programming language are also integrated into the course.

Course Materials

- This webpage and its pdf version: the online script
- eWhiteboard and eWhiteboard exercises: the whiteboard notes
- ILIAS: further course material
- Problemsets: for the exercises
- R-scripts: codes from the lecture

Literature

The course is based on the following textbooks:

- Stock, J.H. and Watson, M.W. (2019). **Introduction to Econometrics (Fourth Edition, Global Edition)**. Pearson.
- Hansen, B.E. (2022a). **Probability and Statistics for Economists**. Princeton.
- Hansen, B.E. (2022b). **Econometrics**. Princeton.
- Davidson, R., and MacKinnon, J.G. (2004). **Econometric Theory and Methods**. Oxford University Press.

Stock and Watson (2019) is available here. To view the book, please activate your Uni Köln VPN connection. For more information on Hansen (2022a, 2022b), please see the ILIAS course. Davidson and MacKinnon (2004) is available for free on the author's webpage: LINK. Printed versions of the books are available from the university library.

Preparation

You should also be familiar with the basic concepts of **matrix algebra**. Please consider this refresher:

Crash Course in Matrix Algebra

We will be using the statistical programming language R. Please make sure you have **R** and **RStudio** installed before the class. Here you find the installation instructions for the software. If you are a beginner, please consider this short introduction, which contains many valuable resources:

Getting Started with R

Assessment

The course will be graded by a 90-minute written exam. There will be two optional bonus assignments during the lecture period. These assignments will allow you to earn bonus points that will be added to your overall exam score, but they are optional and not required to achieve the maximum score on the exam. More information about the assessment can be found on ILIAS.

Communication

Feel free to use the ILIAS statistics forum to discuss lecture topics and ask questions. Please also let me know if you find any typos. Of course, you can also reach me via e-mail: sven.otto@uni-koeln.de

Important Dates

Bonus assignment 1	Nov 04, 2023 - Nov 17, 2023
Bonus assignment 2	Nov 18, 2023 - Dec 01, 2023
Registration deadline exam 1	Nov 25, 2023
Exam 1	Dec 09, 2023
Registration deadline exam 2	Mar 14, 2024
Exam 2 (alternate date)	Mar 28, 2024

Please register for the exam on time. If you miss the registration deadline, you will not be able to take the exam (the Examinations Office is very strict about this). You only need to

take one of the two exams to complete the course. The second exam will serve as a make-up exam for those who fail the first exam or do not take the first exam.

Timetable

The course is held on Thursdays from 10:00 to 13:30 and on Fridays from 10:00 to 11:30 in **Seminar Room BI** on the fourth floor of building 107b (Universitäts- und Stadtbibliothek).

Day	Time	Lecture/Exercise
Thu, Oct 12	10:00-11:30	Lecture
	12:00-13:30	Lecture
Fri, Oct 13	10:00-11:30	Lecture
Thu, Oct 19	10:00-11:30	Exercises
	12:00-13:30	Lecture
Fri, Oct 20	10:00-11:30	Lecture
Thu, Oct 26	10:00-11:30	Exercises
	12:00-13:30	Lecture
Fri, Oct 27	10:00-11:30	Lecture
Thu, Nov 02	10:00-11:30	Exercises
	12:00-13:30	Lecture
Fri, Nov 03	10:00-11:30	Lecture
Thu, Nov 09	10:00-11:30	Exercises
	12:00-13:30	Lecture
Fri, Nov 10	10:00-11:30	Lecture
Thu, Nov 16	10:00-11:30	Exercises
	12:00-13:30	Lecture
Fri, Nov 17	10:00-11:30	Lecture
Thu, Nov 23	10:00-11:30	Exercises
	12:00-13:30	Lecture
Fri, Nov 24	10:00-11:30	Lecture
Thu, Nov 30	10:00-13:30	Lecture/Q&A

R-Packages

To run the R code of the lecture script, you will need to install some additional packages (only a few, since we will mostly be using base R).

```
install.packages(c("sandwich", "lmtest", "tidyverse", "moments"))
```

To apply inferential methods that are not available in base R packages, we will use `sandwich`, `lmtest`, and `moments`. The `tidyverse` will be useful for data management and visualization. To install the R package that contains the datasets for the lecture please follow the instructions in the ILIAS course.

Some further datasets are contained in my package `teachingdata`, which is available in a GitHub repository:

```
install.packages("remotes")
remotes::install_github("ottosven/teachingdata")
```

1 Introduction

Data is usually the result of a random experiment. The gender of the next person you meet, today's share price of Biontech, the number of Taylor Swift's Spotify streams this month, the sales prices of houses in Cologne, the number of pizzas you eat this year, the delay time of the KVB tram, your grade in your exam - all of this information involves a certain amount of randomness.

Suppose you are conducting a survey and ask 10 random people about their gender, years of education, hourly wage, and years of work experience. It is convenient to work with numerical values only, so we write 1 if the person is female and 0 otherwise. Your data table might look like this:

The random selection of a particular person to be interviewed and to fill out the spreadsheet is a random experiment. Therefore, in order to make statistical inferences about the dependence of the collected variables, we must first understand randomness and uncertainty from a mathematical perspective. Probability is the mathematical language for situations where the outcome is unknown. Probability theory is the basis of mathematical statistics and econometric theory.

We interpret the entries in Table 1.1 to be the outcomes of random variables. For example, the gender of the first person interviewed is a random variable. The value 1 is the result of this random experiment. If the second person had been interviewed before the first, this value would be zero.

Let X_1 be the 4×1 vector of the values of the first person interviewed, X_2 the vector of the second person interviewed, and so on. These are random vectors. It's realizations are

$$X_1 = \begin{pmatrix} 1 \\ 12 \\ 9.2 \\ 41 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 \\ 18 \\ 14.55 \\ 15 \end{pmatrix}, \quad \dots$$

The full data set can be collected in the 10×4 matrix

Table 1.1: Survey data

Person	Female	Education	Wage	Experience
1	1	12	9.20	41
2	0	18	14.55	15
3	1	12	25.29	46
4	1	13	12.18	15
5	0	12	15.33	7
6	1	18	10.95	15
7	1	12	5.18	36
8	0	12	0.00	3
9	0	18	13.14	2
10	0	21	11.03	6

$$\boldsymbol{X} = \begin{pmatrix} 1 & 12 & 9.2 & 41 \\ 0 & 18 & 14.55 & 15 \\ 1 & 12 & 25.29 & 46 \\ 1 & 13 & 12.18 & 15 \\ 0 & 12 & 15.33 & 7 \\ 1 & 18 & 10.95 & 15 \\ 1 & 12 & 5.18 & 36 \\ 0 & 12 & 0 & 3 \\ 0 & 18 & 13.14 & 2 \\ 0 & 21 & 11.03 & 6 \end{pmatrix},$$

where each column corresponds to one of the variables in Table 1.1, and the i-th row correspond to the values of individual i, i.e., X_1 is the first row of $\boldsymbol{X}$.

Matrix algebra provides a compact representation of multivariate data and an efficient framework for analyzing and implementing statistical methods. We will use matrix algebra frequently throughout this course. Please use this refresher below to review the most important concepts (in particular the first three sections):

Crash Course on Matrix Algebra

The best way to learn statistical methods is to program them yourself. We will use the statistical programming language `R` to implement statistical methods and apply them to real data sets. If you are new to `R`, please take a look at this short introduction, which contains a lot of valuable resources:

Getting Started with R

2 Probability

2.1 Random experiments

A random experiment is a procedure or situation where the result is uncertain and determined by a probabilistic mechanism. An **outcome** is a specific result of a random experiment. The **sample space** S is the set/collection of all potential outcomes.

Let's consider some examples:

- *Coin toss*: The outcome of a coin toss can be 'heads' or 'tails'. This random experiment has a two-element sample space: $S = \{heads, tails\}$.
- *Gender*: If you conduct a survey and interview a random person to ask them about their gender, the answer may be 'female', 'male', or 'diverse'. It is a random experiment since the person to be interviewed is selected randomly. The sample space has three elements: $S = \{female, male, diverse\}$.
- *Education level*: If you ask a random person about their education level according to the ISCED-2011 framework, the outcome may be one of the eight ISCED-2011 levels. We have an eight-element sample space:

$$S = \{Level\ 1, Level\ 2, Level\ 3, Level\ 4, Level\ 5, Level\ 6, Level\ 7, Level\ 8\}.$$

- *Wage*: If you ask a random person about their income per working hour in EUR, there are infinitely many potential answers. Any (non-negative) real number may be an outcome. The sample space is a continuum of different wage levels.

2.2 Random variables

A **random variable** is a numerical summary of a random experiment. In econometrics and applied statistics, we always express random experiments in terms of random variables. Let's define some random variables based on the random experiments above:

- *Coin*: A two-element sample space random experiment can be transformed to a binary random variable, i.e., a random variable that takes either 0 or 1. We define the *coin* random variable as

$$Y = \begin{cases} 1 & \text{if outcome is heads,} \\ 0 & \text{if outcome is tails.} \end{cases}$$

  A binary random variable is also called **Bernoulli random variable**.

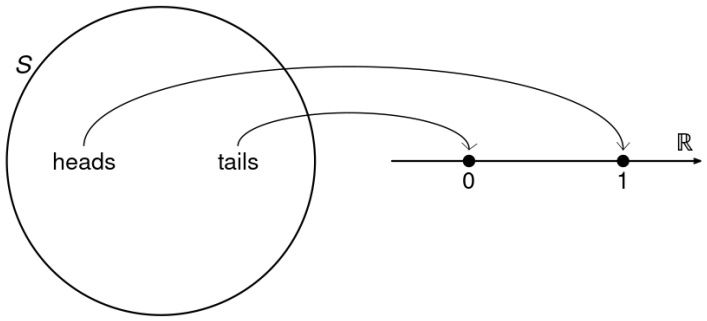


Figure 2.1: Bernoulli random variable

- *Female dummy*: The three-element sample space of the gender random experiment does not provide any natural ordering. A useful way to transform it into random variables are **dummy variables**. The *female* dummy variable is a Bernoulli random variable with

$$Y = \begin{cases} 1 & \text{if the person is female,} \\ 0 & \text{if the person is not female.} \end{cases}$$

  Similarly, dummy variables for *male* and *diverse* can be defined.

- *Education*: The eight-element sample space of the education-level random experiment provides a natural ordering. We define the random variable *education* as the number of years of schooling of the interviewed person:

$$Y = \text{number of years of schooling} \in \{4, 10, 12, 13, 14, 16, 18, 21\}.$$

- *Wage*: The wage level of the interviewed is already numerical. The random variable is

$$Y = \text{income per working hour in EUR.}$$

Table 2.1: ISCED 2011 levels

ISCED level	Education level	Years of schooling
1	Primary	4
2	Lower Secondary	10
3	Upper secondary	12
4	Post-Secondary	13
5	Short-Cycle Tertiary	14
6	Bachelor's	16
7	Master's	18
8	Doctoral	21

2.3 Probability function

In the case of a fair coin, it is natural to assign the following probabilities to the coin variable: $P(Y = 0) = 0.5$ and $P(Y = 1) = 0.5$. By definition, the coin variable will never take the value 2.5, so the corresponding probability is $P(Y = 2.5) = 0$. We may also consider intervals, e.g., $P(Y \geq 0) = 1$ and $P(-1 \leq Y < 1) = 0.5$

The **probability function** P assigns values between 0 and 1 to **events**. Specific subsets of the real line define events. Any real number defines an event, and any open, half-open, or closed interval represents an event as well, e.g.,

$$A_1 = \{Y = 0\}, \quad A_2 = \{Y = 1\}, \quad A_3 = \{Y = 2.5\}$$

and

$$A_4 = \{Y \geq 0\}, \quad A_5 = \{-1 \leq Y < 1\}.$$

We may take **complements**

$$A_6 := A_4^c = \{Y \geq 0\}^c = \{Y < 0\},$$

as well as **unions** and **intersections**:

$$\begin{aligned} A_7 &:= A_1 \cup A_6 = \{Y = 0\} \cup \{Y < 0\} = \{Y \leq 0\}, \\ A_8 &:= A_4 \cap A_5 = \{Y \geq 0\} \cap \{-1 \leq Y < 1\} = \{0 \leq Y < 1\}. \end{aligned}$$

Unions and intersections can also applied iteratively,

$$A_9 := A_1 \cup A_2 \cup A_3 \cup A_5 \cup A_6 \cup A_7 \cup A_8 = \{Y \in (-\infty, 1] \cup \{2.5\}\},$$

and by taking complements, we obtain the full real line and the empty set:

$$\begin{aligned} A_{10} &:= A_9 \cup A_9^c = \{Y \in \mathbb{R}\}, \\ A_{11} &:= A_{10}^c = \{\}. \end{aligned}$$

You may verify that $P(A_1) = 0.5$, $P(A_2) = 0.5$, $P(A_3) = 0$, $P(A_4) = 1$ $P(A_5) = 0.5$, $P(A_6) = 0$, $P(A_7) = 0.5$, $P(A_8) = 0.5$, $P(A_9) = 1$, $P(A_{10}) = 1$, $P(A_{11}) = 0$. If you take the variables *education* or *wage*, the probabilities of these events may be completely different.

To make probabilities a mathematically sound concept, we have to define to which events probabilities are assigned and how these probabilities are assigned. We consider the concept of a **sigma algebra** to collect all events.

Sigma algebra

A collection $\mathcal{B}$ of sets is called sigma algebra if it satisfies the following three properties:

1. $\{\} \in \mathcal{B}$ (empty set)
2. If $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$
3. If $A_1, A_2, ... \in \mathcal{B}$, then $A_1 \cup A_2 \cup ... \in \mathcal{B}$.

If you take all events of the form $\{Y \in (a, b)\}$, where $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, and if you add all unions, intersections, and complements of these events, and again all unions, intersections, and complements of those events, and so on, you will obtain the so-called **Borel sigma algebra**. The Borel sigma algebra contains all events we assign probabilities to, the **Borel sets**.

Probabilities must follow certain conditions. The following axioms ensure that these conditions are fulfilled:

Probability function

A probability function P is a function $P : \mathcal{B} \to [0, 1]$ that satisfies the Axioms of Probability:

1. $P(A) \geq 0$ for every $A \in \mathcal{B}$
2. $P(Y \in \mathbb{R}) = 1$
3. If $A_1, A_2, A_3 ...$ are disjoint then

$$A_1 \cup A_2 \cup A_3 \cup ... = P(A_1) + P(A_2) + P(A_3) + ...$$

Recall that two events A and B are **disjoint** if they have no outcomes in common, i.e., if $A \cap B = \{\}$. For instance, A_1 and A_2 are $A_1 = \{Y = 0\}$ and $A_2 = \{Y = 1\}$ are disjoint, but A_1 and $A_4 = \{Y \geq 0\}$ are not disjoint, since $A_1 \cap A_4 = \{Y = 0\}$ is nonempty.

Probabilities are a well-defined concept if we use the Borel sigma algebra and the axioms of probability. The mathematical details are developed in the field of measure theory.

The axioms of probability imply the following rules of calculation:

Basic rules of probability

- $0 \leq P(A) \leq 1$ for any event A
- $P(A^c) = 1 - P(A)$ for the complement event of A
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events A, B (inclusion-exclusion principle)
- $P(A) \leq P(B)$ if $A \subset B$
- $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint

2.4 Distribution

The **distribution** of a random variable Y is characterized by the probabilities of all events of Y in the Borel sigma algebra. The distribution of the *coin* variable is fully characterized by the probabilities $P(Y = 1) = 0.5$ and $P(Y = 0) = 0.5$. We can compute the probabilities of all other events using the basic rules of probability. The probability mass function summarizes these probabilities:

Probability mass function (PMF)

The probability mass function (PMF) of a random variable Y is

$$\pi(a) := P(Y = a), \quad a \in \mathbb{R}$$

The PMF of the *coin* variable is

$$\pi(a) = P(Y = a) = \begin{cases} 0.5 & \text{if } a \in \{0, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

The *education* variable may have the following PMF:

$$\pi(a) = P(Y = a) = \begin{cases} 0.008 & \text{if } a = 4 \\ 0.048 & \text{if } a = 10 \\ 0.392 & \text{if } a = 12 \\ 0.072 & \text{if } a = 13 \\ 0.155 & \text{if } a = 14 \\ 0.071 & \text{if } a = 16 \\ 0.225 & \text{if } a = 18 \\ 0.029 & \text{if } a = 21 \\ 0 & \text{otherwise} \end{cases}$$

The PMF is useful for distributions where the sum of the PMF values over a discrete (finite or countably infinite) number of domain points equals 1, as in the examples above. These distributions are called **discrete distributions**.

Another example of a discrete distribution is the **Poisson distribution** with parameter $\lambda > 0$, which has the PMF

$$\pi(a) = \begin{cases} \frac{e^{-\lambda}\lambda^a}{a!} & \text{if } a = 0, 1, 2, 3, ... \\ 0 & \text{otherwise.} \end{cases}$$

It has a countably infinite number of domain points with nonzero PMF values, and its probabilities sum to 1, i.e., $\sum_{a=0}^{\infty} \pi(a) = e^{-\lambda} \sum_{a=0}^{\infty} \frac{\lambda^a}{a!} = 1$ since the exponential function has the power series representation $e^{\lambda} = \sum_{a=0}^{\infty} \frac{\lambda^a}{a!}$.

Not all random variables are discrete, e.g., the *wage* variable takes values on a continuum. The cumulative distribution function is a unifying concept summarizing the distribution of any random variable.

2.5 Cumulative distribution function

Cumulative distribution function (CDF)

The cumulative distribution function (CDF) of a random variable Y is

$$F(a) := P(Y \leq a), \quad a \in \mathbb{R},$$

The CDF of the variable *coin* is

$$F(a) = \begin{cases} 0 & a < 0, \\ 0.5 & 0 \leq a < 1, \\ 1 & a \geq 1, \end{cases}$$

with the following CDF plot:

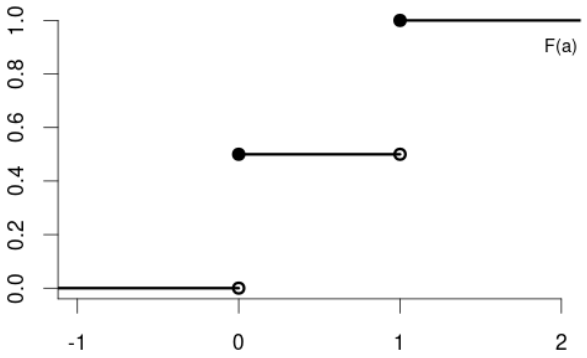


Figure 2.2: CDF of coin

The CDF of the variables *education* is

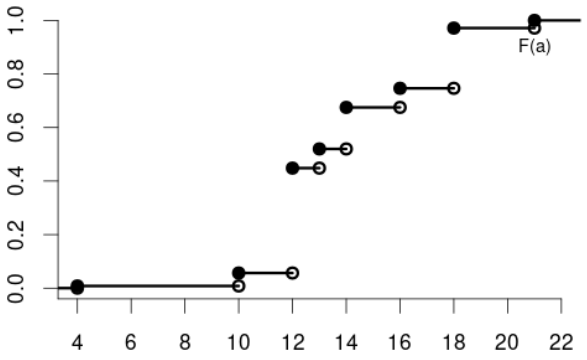


Figure 2.3: CDF of education

and the CDF of the variable *wage* may have the following form:

By the basic rules of probability, we can compute the probability of any event if we know the probabilities of all events of the form $\{Y \leq a\}$.

Some basic rules for the CDF (for $a < b$):

- $P(Y \leq a) = F(a)$
- $P(Y > a) = 1 - F(a)$

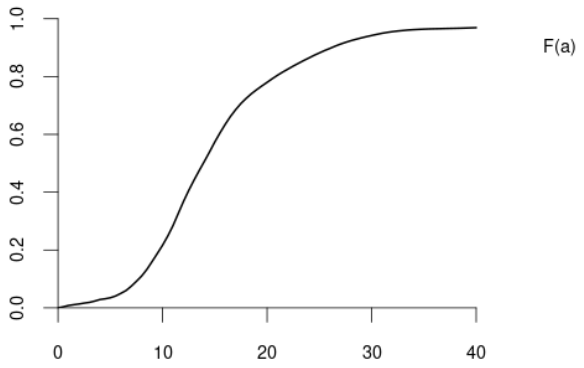


Figure 2.4: CDF of wage

- $P(Y < a) = F(a) - \pi(a)$
- $P(Y \geq a) = 1 - P(Y < a)$
- $P(a < Y \leq b) = F(b) - F(a)$
- $P(a < Y < b) = F(b) - F(a) - \pi(b)$
- $P(a \leq Y \leq b) = F(b) - F(a) + \pi(a)$
- $P(a \leq Y < b) = P(a \leq Y \leq b) - \pi(b)$

Some CDFs have jumps/steps, and some CDFs are smooth/continuous. If F has a jump at domain point a, then the PMF at a is

$$\pi(a) = P(Y = a) = F(a) - \lim_{\epsilon \to 0} F(a - \epsilon) = \text{“jump height at } a\text{”}. \tag{2.1}$$

If F is continuous at domain point a, we have $\lim_{\epsilon \to 0} F(a - \epsilon) = F(a)$, which implies that $\pi(a) = P(Y = a) = 0$.

We call the random variable a **discrete random variable** if the CDF contains jumps and is flat between the jumps. A discrete random variable has only a finite (or countably infinite) number of potential outcomes. The values of the PMF correspond to the jump heights in the CDF as defined in Equation 2.1. The **support** $\mathcal{Y}$ of a discrete random variable Y is the set of all points $a \in \mathbb{R}$ with nonzero probability mass, i.e. $\mathcal{Y} = \{a \in \mathbb{R} : \pi(a) > 0\}$. The probabilities of a discrete random variable sum to 1, i.e., $\sum_{a \in \mathcal{Y}} \pi(a) = 1$.

The Bernoulli variables *coin* and *female* are discrete random variables with support $\mathcal{Y} = \{0, 1\}$. The variable *eduaction* has support $\mathcal{Y} = \{4, 10, 12, 13, 14, 16, 18, 21\}$. A Poisson random variable has thr support $\mathcal{Y} = \mathbb{N} \cup \{0\}$.

We call a random variable a **continuous random variable** if the CDF is continuous at every point $a \in \mathbb{R}$. A continuous random variable has $\pi(a) = P(Y = a) = 0$ for all $a \in \mathbb{R}$. The basic rules for the CDF become simpler in the case of a continuous random variable:

Rules for the CDF of a continuous random variable (for $a < b$):

- $P(Y \leq a) = P(Y < a) = F(a)$
- $P(Y \geq a) = P(Y > a) = 1 - F(a)$
- $P(a < Y \leq b) = P(a \leq Y < b) = F(b) - F(a)$
- $P(a < Y < b) = P(a \leq Y \leq b) = F(b) - F(a)$

Single-outcome events are null sets and occur with probability zero. Therefore, the PMF is not suitable to describe the distribution of a continuous random variable. We use the CDF to compute probabilities of interval events as well as their unions, intersections, and complements.

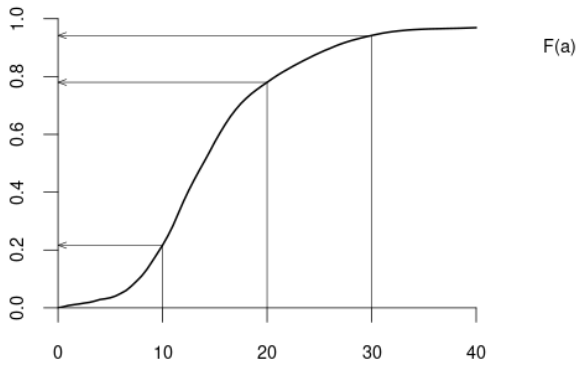


Figure 2.5: CDF of wage evaluated at some points

For instance, $P(Y \leq 30) = 0.942$, $P(Y \leq 20) = 0.779$, $P(Y \leq 10) = 0.217$, and $P(10 \leq Y \leq 20) = 0.779 - 0.217 = 0.562$.

Quantiles

For a continuous random variable Y the α-quantile $q(\alpha)$ is defined as the solution to the equation $\alpha = F(q(\alpha))$, or, equivalently, as the inverse of the distribution function:

$$q(\alpha) = F^{-1}(\alpha)$$

- $q(\cdot)$ is a function from $(0, 1)$ to $\mathbb{R}$.
- Some quantiles have special names:
  - The median is the 0.5 quantile.
  - The quartiles are the 0.25, 0.5 and 0.75 quantiles.

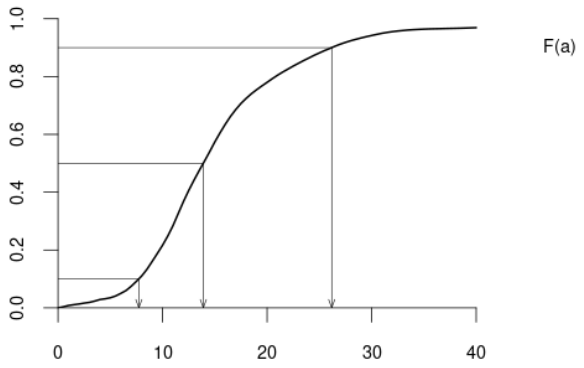


Figure 2.6: Quantiles of variable wage

– The deciles are the 0.1, 0.2,… , 0.9 quantiles.

From the quantile plot, we find that $q(0.1) = 7.73$, $q(0.5) = 13.90$, $q(0.9) = 26.18$. Under this wage distribution, the median wage is 13.90 EUR, the poorest 10% have a wage of less than 7.33 EUR, and the richest 10% have a wage of more than 26.18 EUR.

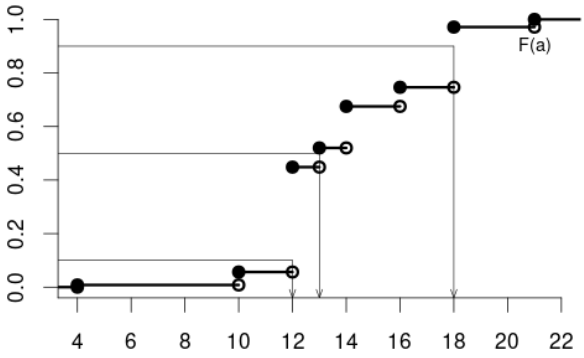


Figure 2.7: Quantiles of variable education

The median of *education* is 13, the 0.1-quantile is 12, and the 0.9-quantile is 18.

A CDF has the following properties:

(i) it is *non-decreasing*,
(ii) it is *right-continuous* (jumps may occur only when the limit point is approached from the left)
(iii) the left limit is zero: $\lim_{a\to-\infty} F(a) = 0$
(iv) the right limit is one: $\lim_{a\to\infty} F(a) = 1$.

Any function F that satisfies these four properties defines a probability distribution. Typically, distributions are divided into discrete and continuous distributions. Still, it may be the case

that a distribution does not fall into either of these categories (for instance, if a CDF has jumps on some domain points and is continuously increasing on other domain intervals). In any case, the CDF characterizes the entire distribution of any random variable.

2.6 Probability density function

For discrete random variables, both the PMF and the CDF characterize the distribution. In the case of a continuous random variable, the PMF does not yield any information about the distribution since it is zero. The continuous counterpart of the PMF is the density function:

Probability density function

The probability density function (PDF) or simply density function of a continuous random variable Y is a function $f(a)$ that satisfies

$$F(a) = \int_{-\infty}^{a} f(u)\ \mathrm{d}u$$

The density $f(a)$ is the derivative of the CDF $F(a)$ if it is differentiable:

$$f(a) = \frac{d}{da}F(a).$$

Properties of a PDF:

(i) $f(a) \geq 0$ for all $a \in \mathbb{R}$

(ii) $\int_{-\infty}^{\infty} f(u)\ \mathrm{d}u = 1$

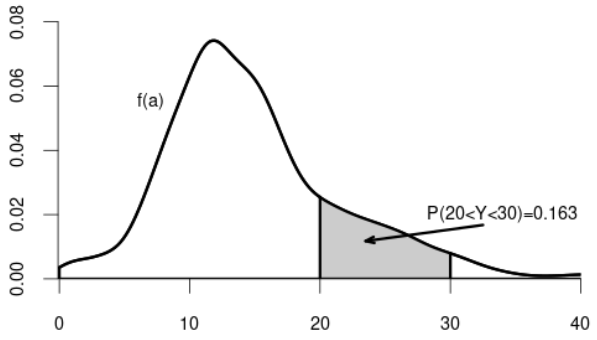


Figure 2.8: PDF of the variable wage

Probability rule for the PDF:

$$P(a < Y < b) = \int_a^b f(u)\ \mathrm{d}u = F(b) - F(a)$$

2.7 Expected value

The **expectation** or **expected value** is the most important measure of the central tendency of a distribution. It gives you the average value you can expect to get if you repeat the random experiment multiple times. We define the expectation first for discrete random variables, then continuous random variables, and finally give a unified definition for all random variables.

2.7.1 Expectation of a discrete random variable

The **expectation** or **expected value** of a discrete random variable Y with PMF $\pi(\cdot)$ and support $\mathcal{Y}$ is defined as

$$E[Y] = \sum_{u \in \mathcal{Y}} u\pi(u).$$

For the *coin* variable, we have $\mathcal{Y} = \{0, 1\}$ and therefore

$$E[Y] = 0 \cdot \pi(0) + 1 \cdot \pi(1) = 0.5.$$

For the variable *education* we get

$$\begin{aligned} E[Y] = {} & 4 \cdot \pi(4) + 10 \cdot \pi(10) + 12 \cdot \pi(12) \\ & + 13 \cdot \pi(13) + 14 \cdot \pi(14) + 16 \cdot \pi(16) \\ & + 18 \cdot \pi(18) + 21 * \pi(21) = 13.557 \end{aligned}$$

The expectation of a Poisson distributed random variable Y with parameter λ is

$$E[Y] = 0 + \sum_{a=1}^{\infty} a \cdot e^{-\lambda} \frac{\lambda^a}{a!} = e^{-\lambda} \sum_{a=1}^{\infty} \frac{\lambda^a}{(a-1)!} = e^{-\lambda} \sum_{a=0}^{\infty} \frac{\lambda^{a+1}}{a!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

2.7.2 Expectation of a continuous random variable

The **expectation** or **expected value** of a of a continuous random variable Y with PDF $f(\cdot)$ is

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \, \mathrm{d}u.$$

Using numerical integration for the density of Figure 2.8 yields the expected value of 16.45 EUR for the wage variable, which is larger than the median value of 13.90 EUR. If the mean is larger than the median, we have a positively skewed distribution, meaning that a few people have high salaries, and many people have medium and low wages.

The uniform distribution on the unit interval $[0, 1]$ has the PDF

$$f(u) = \begin{cases} 1 & \text{if } u \in [0, 1], \\ 0 & \text{otherwise,} \end{cases}$$

and the expected value of a uniformly distributed random variable Y is

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \, \mathrm{d}u = \int_{0}^{1} u \, \mathrm{d}u = \frac{1}{2}.$$

2.7.3 Expectation for general random variables

We can also define the expected value in a unified way for any random variable so we do not have to distinguish between discrete and continuous random variables. Let $F(\cdot)$ be the CDF of the random variable of interest and consider the differential $\mathrm{d}F(u)$, which corresponds to an infinitesimal change in $F(\cdot)$ at u. For a discrete random variable, $F(u)$ changes only if there is a step/jump at u and zero otherwise because it is flat. Thus, for a discrete distribution,

$$\mathrm{d}F(u) = \begin{cases} \pi(u) & \text{if } u \in \mathcal{Y} \\ 0 & \text{if } u \notin \mathcal{Y}. \end{cases}$$

In the case of a continuous random variable with differentiable CDF $F(\cdot)$, we have

$$\mathrm{d}F(u) = f(u) \, \mathrm{d}u,$$

where $f(\cdot)$ is the PDF of the random variable. This gives rise to the following unified definition of the expected value:

The **expectation** or **expected value** of any random variable with CDF $F(\cdot)$ is defined as

$$E[Y] = \int_{-\infty}^{\infty} u \, \mathrm{d}F(u). \tag{2.2}$$

Note that Equation 2.2 is the Riemann-Stieltjes integral of a with respect to the function $F(\cdot)$. Recall that the Riemann integral of u with respect to u over the interval $[-1, 1]$ is

$$\int_{-1}^{1} u \, \mathrm{d}u := \lim_{N\to\infty} \sum_{j=1}^{2N} \Big(\frac{j}{N} - 1\Big)\Big((\tfrac{j}{N} - 1) - (\tfrac{j-1}{N} - 1)\Big) = \lim_{N\to\infty} \sum_{j=1}^{2N} \Big(\frac{j}{N} - 1\Big)\frac{1}{N},$$

for the interval $[-z, z]$ we have

$$\int_{-z}^{z} u \, \mathrm{d}u := \lim_{N\to\infty} \sum_{j=1}^{2N} z\Big(\frac{j}{N} - 1\Big)\frac{z}{N},$$

and we obtain $\int_{-\infty}^{\infty} u \, \mathrm{d}u := \lim_{z\to\infty} \int_{-z}^{z} u \, \mathrm{d}u$ for the integral over the entire real line. Note that $z/N = z(\frac{j}{N} - 1) - z(\frac{j-1}{N} - 1)$ corresponds to a change in u on $[-z, z]$ so we approximate

$$\mathrm{d}u \approx z(\tfrac{j}{N} - 1) - z(\tfrac{j-1}{N} - 1) = \tfrac{z}{N}$$

and let N tend to infinity. In the case of the Riemann-Stieltjes integral, where we integrate with respect to changes in a function $F(\cdot)$, i.e., $\mathrm{d}F(u)$. In an interval $[-z, z]$, we have

$$\mathrm{d}F(u) \approx F\Big(z(\tfrac{j}{N} - 1)\Big) - F\Big(z(\tfrac{j-1}{N} - 1)\Big),$$

and we define

$$\int_{-z}^{z} u \, \mathrm{d}F(u) := \lim_{N\to\infty} \sum_{j=1}^{2N} z\Big(\tfrac{j}{N} - 1\Big)F\Big(z(\tfrac{j}{N} - 1)\Big) - F\Big(z(\tfrac{j-1}{N} - 1)\Big)$$

$$\int_{-\infty}^{\infty} u \, \mathrm{d}F(u) := \lim_{z\to\infty} \int_{-z}^{z} u \, \mathrm{d}F(u)$$

2.7.4 Properties of the expected value

The expected value is a measure of central tendency. It is a **linear** function. For any two random variables Y and Z and any $a, b \in \mathbb{R}$, we have

$$E[aY + bZ] = aE[Y] + bE[Z].$$

The expected value has some optimality properties in terms of prediction. The best predictor of a random variable Y in the mean square error sense is the value g^* that minimizes $E[(Y-g)^2]$ over g. We have

$$E[(Y - g)^2] = E[Y^2] - 2gE[Y] + g^2,$$

and minimizing over g yields

$$\frac{\mathrm{d}E[(Y-g)^2]}{\mathrm{d}g} = -2E[Y] + 2g,$$

which is zero if $g = E[Y]$. The second derivative is positive. Therefore, the expected value is the **best predictor** for a random variable if you do not have any further information available.

We often transform random variables by taking, for instance, squares Y^2 or logs $\log(Y)$. For any transformation function $g(\cdot)$, the expectation of the transformed random variable $g(Y)$ is

$$E[g(Y)] = \int_{-\infty}^{\infty} g(u)\ \mathrm{d}F(u),$$

where $\mathrm{d}F(u)$ can be replaced by the PMF or the PDF as discussed in Section 2.7.3 for the different cases. For instance, if we take the *coin* variable Y and consider the transformed random variable $\log(Y+1)$, the expected value is

$$E[\log(Y+1)] = \log(1)\cdot\frac{1}{2} + \log(2)\cdot\frac{1}{2} = \frac{\log(2)}{2}$$

Moments

The r-th moment of a random variable Y is defined as

$$E[Y^r] = \int_{-\infty}^{\infty} u^r\ \mathrm{d}F(u) = \begin{cases} \sum_{u\in y} u^r \pi(u) & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} u^r f(u)\mathrm{d}u & \text{if } Y \text{ is continuous.} \end{cases}$$

2.8 Descriptive features of a distribution

Table 2.2: Some important features of the distribution of Y

$E[Y^r]$	r-th moment of Y
$E[(Y-E[Y])^r]$	r-th central moment of Y
$Var[Y] = E[(Y-E[Y])^2]$	variance of Y
$sd(Y) = \sqrt{Var[Y]}$	standard deviation of Y
$E[((Y-E[Y])/sd(Y))^r]$	r-th standardized moment of Y
$skew = E[((Y-E[Y])/sd(Y))^3]$	skewness of Y
$kurt = E[((Y-E[Y])/sd(Y))^4]$	kurtosis of Y

The mean is a measure of central tendency and equals the expected value. The variance and standard deviation are measures of dispersion. We have

$$Var[Y] = E[(Y-E[Y])^2] = E[Y^2] - E[Y]^2$$

and

$$Var[a + bY] = b^2 Var[Y]$$

for any $a, b \in \mathbb{R}$. The skewness

$$skew = \frac{E[(Y - E[Y])^3]}{sd(Y)^3} = \frac{E[Y^3] - 3E[Y^2]E[Y] + 2E[Y]^3}{(E[Y^2] - E[Y]^2)^{3/2}}$$

is a measure of asymmetry

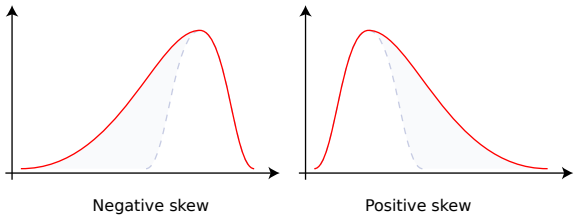


Figure 2.9: Positive and negative skewness

A random variable Y has a **symmetric distribution** about 0 if $F(u) = 1 - F(-u)$. If Y has a density, it is symmetric if $f(x) = f(-x)$. If Y is symmetric about 0, then the skewness is 0. The skewness of the variable *wage* (see Figure 2.8) is positive, i.e., the distribution is positively skewed. The **standard normal distribution** $\mathcal{N}(0, 1)$, which has the density

$$f(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

Below you find a plot of the PDFs of $N(0, 1)$ together with the t_5-distribution, which is the t-distribution with 5 degrees of freedom:

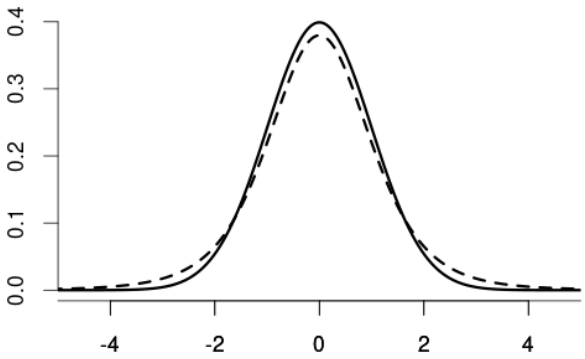


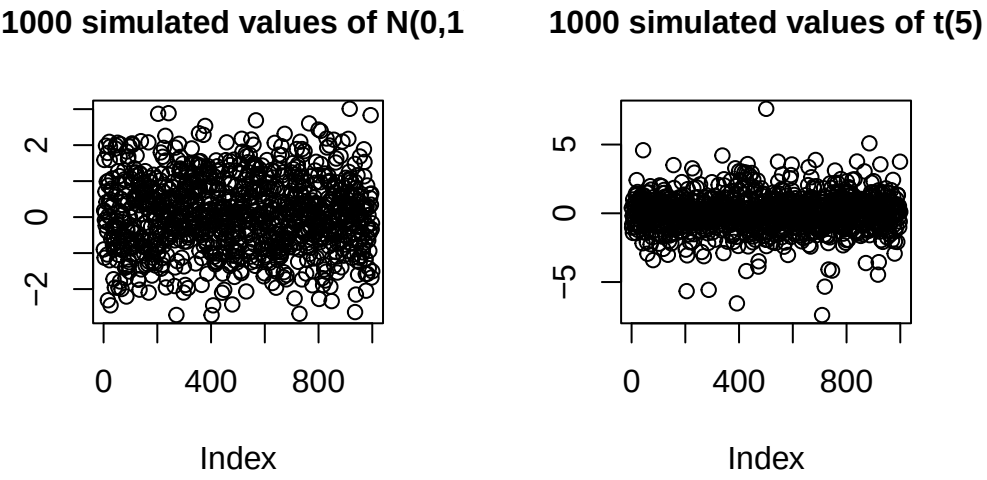
Figure 2.10: PDFs of the standard normal distribution (solid) and the t_5-distribution (dashed)

The standard normal distribution and the t(5) distribution have skewness 0. The kurtosis

$$kurt = \frac{E[(Y - E[Y])^4]}{sd(Y)^4} = \frac{E[Y^4] - 4E[Y^3]E[Y] + 6E[Y^2]E[Y]^2 - 3E[Y]^4}{(E[Y^2] - E[Y]^2)^2}$$

is a measure of how likely extreme outliers are. The standard normal distribution has kurtosis 3 and the t(5) distribution has kurtosis 9 so that outliers in $t(5)$ are more likely than in $\mathcal{N}(0,1)$:

```
par(mfrow=c(1,2), cex.main=1)
plot(rnorm(1000), main = "1000 simulated values of N(0,1)", ylab = "")
plot(rt(1000,5), main = "1000 simulated values of t(5)", ylab = "")
```



The kurtosis of the variable *wage* is also larger than 3, meaning outliers are much more likely than in the standard normal distribution. In this case, the positive skewness means that more people have a wage less than the average, and the large kurtosis means that there are very few people with exceptionally high salaries (outliers).

All features discussed above are functions of the first four moments $E[Y]$, $E[Y^2]$, $E[Y^3]$ and $E[Y^4]$.

2.8.1 Heavy-tailed distributions

Expectations might be infinity. For instance, the simple Pareto distribution has the PDF

$$f(a) = \begin{cases} \frac{1}{a^2} & \text{if } a > 1, \\ 0 & \text{if } a \leq 1, \end{cases}$$

and the expected value is

$$E[X] = \int_{-\infty}^{\infty} a f(a)\, \mathrm{d}a = \int_{1}^{\infty} \frac{1}{a}\, \mathrm{d}a = \log(a)|_1^{\infty} = \infty.$$

The game of chance from the St. Petersburg paradox (see https://en.wikipedia.org/wiki/St._Petersburg_paradox) is an example of a discrete random variable with infinite expectation.

There are distributions with finite mean with some higher moments that are infinite. For instance, the first $m-1$ moments of the t_m distribution (Student's-t distribution with m degrees of freedom) are finite, but the m-th moment and all higher order moments are infinite. Random variables with infinite first four moments have a so-called **heavy-tailed distribution** and may produce huge outliers. Many statistical procedures are only valid if the underlying distribution is not heavy-tailed.

2.9 The normal distribution

A random variable X is normally distributed with parameters (μ, σ^2) if it has the density

$$f(a \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(a-\mu)^2}{2\sigma^2} \Big).$$

We write $Y \sim \mathcal{N}(\mu, \sigma^2)$. Mean and variance are

$$E[Y] = \mu, \quad var[Y] = \sigma^2.$$

Special case: standard normal distribution $\mathcal{N}(0,1)$ with density

$$\phi(a) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{a^2}{2} \Big)$$

and CDF

$$\Phi(a) = \int_{-\infty}^{a} \phi(u) \mathrm{d}u.$$

$\mathcal{N}(0,1)$ is symmetric around zero:

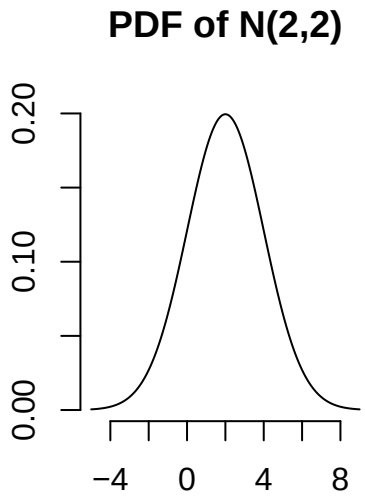
$$\phi(a) = \phi(-a), \quad \Phi(a) = 1 - \Phi(-a)$$

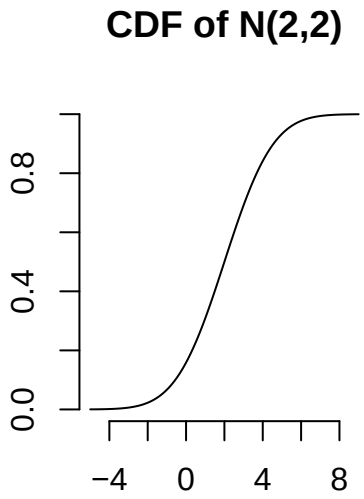
```
par(mfrow=c(1,2), bty="n", lwd=1)
x <- seq(-5,9,by=0.01)
plot(x,dnorm(x,2,2),ylab="",xlab="", type="l", main= "PDF of N(2,2)")
plot(x,pnorm(x,2,2),ylab="",xlab="", type="l", main = "CDF of N(2,2)")
```

If $Y_1, \ldots, Y_n$ are normally distributed and $c_1, \ldots, c_n \in \mathbb{R}$, then $\sum_{j=1}^{n} c_j Y_j$ is normally distributed.

2.10 Additional reading

- Stock and Watson (2019), Section 2
- Hansen (2022a), Section 1-2
- Davidson and MacKinnon (2004), Section 1





2.11 R-codes

statistics-sec2.R

3 Dependence

3.1 Multivariate random variables

In statistics, we typically study multiple random variables simultaneously. We can collect k random variable $X_1, \dots, X_k$ in a **random vector**

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} = (X_1, \dots, X_k)'.$$

We also call X a **k-variate random variable**.

Since X is a random vector, its outcome is also vector-valued, e.g. $X = x \in \mathbb{R}^k$ with $x = (x_1, \dots, x_k)'$. Events of the form $\{X \leq x\}$ mean that each component of the random vector X is smaller than the corresponding values of the vector x, i.e.

$$\{X \leq x\} = \{X_1 \leq x_1, \dots, X_k \leq x_k\}.$$

3.2 Bivariate random variables

If $k = 2$, we call X a **bivariate random variable**. Consider, for instance, the coin toss Bernoulli variable Y with $P(Y = 1) = 0.5$ and $P(Y = 0) = 0.5$, and let Z be a second coin toss with the same probabilities. $X = (Y, Z)$ is a bivariate random variable where both entries are discrete random variables. Since the two coin tosses are performed separately from each other, it is reasonable to assume that the probability that the first and second coin tosses show 'heads' is 0.25, i.e., $P(\{Y = 1\} \cap \{Z = 1\}) = 0.25$. We would expect the following joint probabilities:

Table 3.1: Joint probabilities of coin tosses

	$Z = 1$	$Z = 0$	any result
$Y = 1$	0.25	0.25	0.5
$Y = 0$	0.25	0.25	0.5
any result	0.5	0.5	1

The probabilities in the above table characterize the **joint distribution** of Y and Z. The table shows the values of the **joint probability mass function**:

$$\pi_{YZ}(a,b) = \begin{cases} 0.25 & \text{if } a \in \{0,1\} \text{ and } b \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}$$

Another example are the random variables Y, a dummy variable for the event that the person has a high wage (more than 25 USD/hour), and Z, a dummy variable for the event that the same person has a university degree. Similarly, $X = (Y, Z)$ is a bivariate random variable consisting of two univariate Bernoulli variables. The joint probabilities might be as follows:

Table 3.2: Joint probabilities of wage and education dummies

	Z=1	Z=0	any education
Y=1	0.19	0.12	0.31
Y=0	0.17	0.52	0.69
any wage	0.36	0.64	1

The joint probability mass function is

$$\pi_{YZ}(a,b) = \begin{cases} 0.19 & \text{if } a = 1, b = 1, \\ 0.12 & \text{if } a = 1, b = 0, \\ 0.17 & \text{if } a = 0, b = 1, \\ 0.52 & \text{if } a = 0, b = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The **marginal probability mass function** of Y is

$$\begin{aligned} \pi_Y(a) = P(Y = a) &= \pi_{YZ}(a,0) + \pi_{YZ}(a,1) \\ &= \begin{cases} 0.19 + 0.12 = 0.31 & \text{if } a = 1, \\ 0.17 + 0.52 = 0.69 & \text{if } a = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

and the **marginal probability mass function** of Z is

$$\begin{aligned} \pi_Z(b) = P(Z = b) &= \pi_{YZ}(0,b) + \pi_{YZ}(1,b) \\ &= \begin{cases} 0.19 + 0.17 = 0.36 & \text{if } b = 1, \\ 0.12 + 0.52 = 0.64 & \text{if } b = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

An example of a continuous bivariate random variable is $X = (Y, Z)$, where Y is the wage level in EUR/hour and Z is the labor market experience of the same person measured in years.

3.3 Bivariate distributions

Bivariate distribution

The joint distribution function of a bivariate random variable (Y, Z) is

$F_{YZ}(a, b) = P(Y \leq a, Z \leq b) = P(\{Y \leq a\} \cap \{Z \leq b\})$.

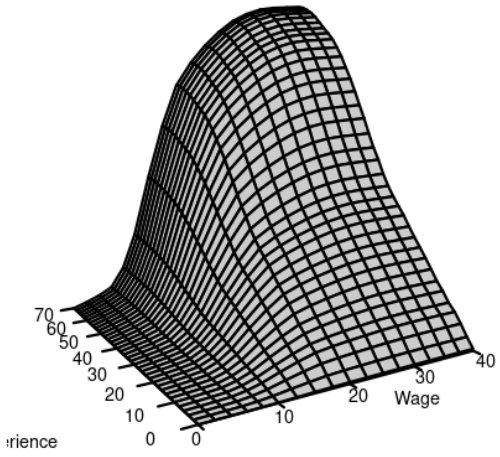


Figure 3.1: Joint CDF of wage and experience

Calculation of probabilities using a bivariate distribution function:

$$P(Y \leq a, Z \leq b) = F_{YZ}(a, b)$$
$$P(a < Y \leq b, c < Z \leq d) = F_{YZ}(b, d) - F_{YZ}(b, c) - F_{YZ}(a, d) + F_{YZ}(a, c)$$

Marginal distributions

The marginal distributions of Y and Z are

$$F_Y(a) = P(Y \leq a) = P(Y \leq a, Z < \infty) = \lim_{b \to \infty} F_{YZ}(a, b),$$
$$F_Z(b) = P(Z \leq b) = P(Y < \infty, Z \leq b) = \lim_{a \to \infty} F_{YZ}(a, b)$$

Bivariate density function

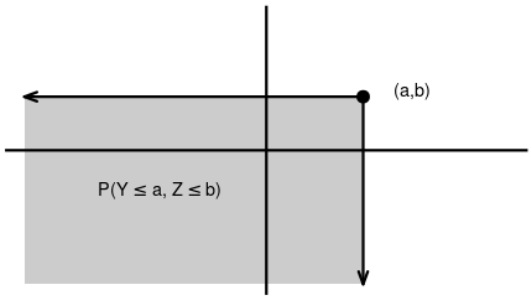


Figure 3.2: Calculate probabilities using the joint CDF

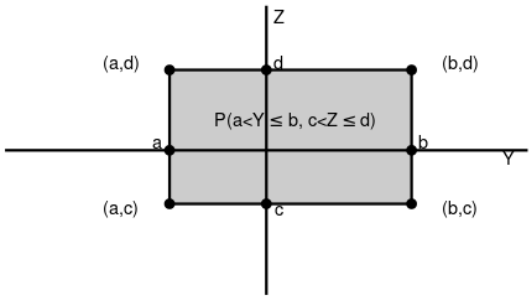


Figure 3.3: Calculate probabilities using the joint CDF

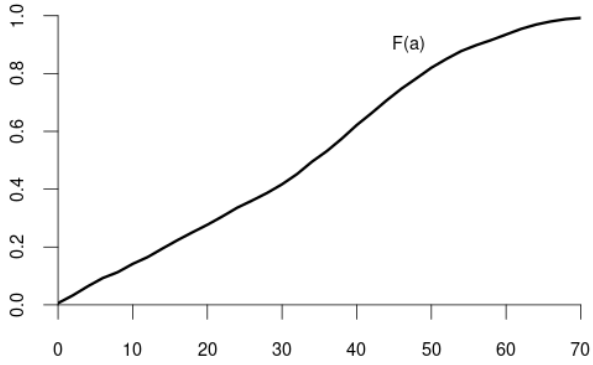


Figure 3.4: Marginal CDF of experience

The joint density function of a bivariate continuous random variable (Y, Z) with differentiable joint CDF $F_{YZ}(a, b)$ equals

$$f_{YZ}(a, b) = \frac{\partial^2}{\partial a \partial b} F_{YZ}(a, b).$$

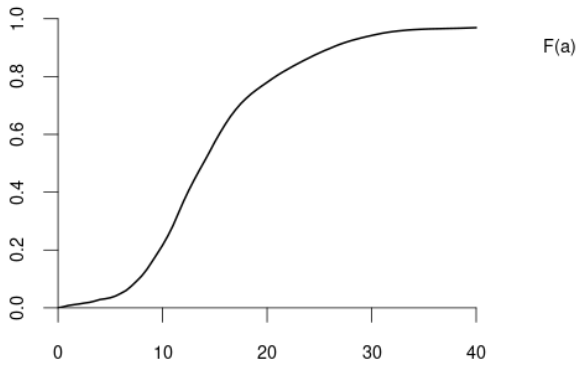


Figure 3.5: Marginal CDF of wage

The marginal densities of Y and Z are

$$f_Y(a) = \frac{d}{da}F_Y(a) = \int_{-\infty}^{\infty} f_{YZ}(a,b)\mathrm{d}b,$$
$$f_Z(b) = \frac{d}{db}F_Z(b) = \int_{-\infty}^{\infty} f_{YZ}(a,b)\mathrm{d}a.$$

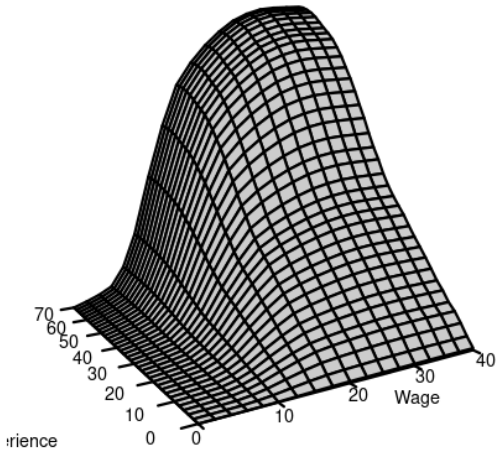


Figure 3.6: Joint CDF of wage and experience

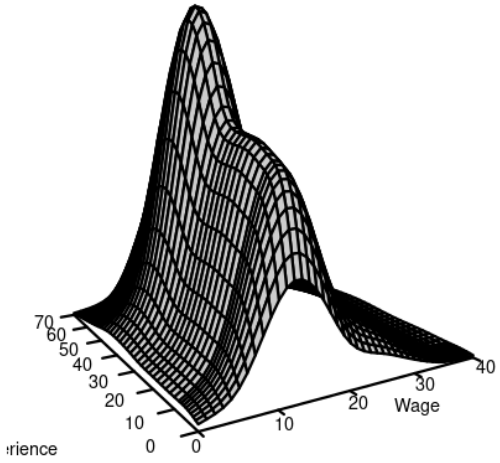


Figure 3.7: Joint PDF of wage and experience

3.4 Correlation

Consider the bivariate continuous random variable (Y, Z) with joint density $f_{YZ}(a, b)$. The expected value of $g(Y, Z)$, where $g(\cdot, \cdot)$ is any real-valued function, is given by

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(a, b) f_{YZ}(a, b) \, \mathrm{d}a \, \mathrm{d}b.$$

The first **cross moment** of Y and Z is $E[YZ]$. We have $E[YZ] = E[g(Y, Z)]$ for the function $g(Y, Z) = Y \cdot Z$. Therefore,

$$E[YZ] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab f_{YZ}(a, b) \, \mathrm{d}a \, \mathrm{d}b.$$

The **covariance** of Y and Z is defined as

$$Cov(Y, Z) = E[(Y - E[Y])(Z - E[Z])] = E[YZ] - E[Y]E[Z].$$

The covariance of Y and Y is the variance:

$$Cov(Y, Y) = Var[Y].$$

The variance of the sum of two random variables depends on the covariance:

$$Var[Y + Z] = Var[Y] + 2Cov(Y, Z) + Var[Z]$$

The **correlation** of Y and Z is

$$Corr(Y,Z) = \frac{Cov(Y,Z)}{sd(Y)sd(Z)}$$

Uncorrelated

Y and Z are **uncorrelated** if $Corr(Y,Z) = 0$, or, equivalently, if $Cov(Y,Z) = 0$.

If Y and Z are uncorrelated, we have

$$E[YZ] = E[Y]E[Z]$$
$$var[Y+Z] = var[Y] + var[Z]$$

3.5 Independence

Two events A and B are independent if

$$P[A \cap B] = P[A]P[B].$$

For instance, in the bivariate random variable of Table 3.1 (two coin tosses), we have

$$P(Y=1, Z=1) = 0.25 = 0.5 \cdot 0.5 = P(Y=1)P(Z=1).$$

Hence, $\{Y=1\}$ and $\{Z=1\}$ are independent events. In the bivariate random variable of Table 3.2 (wage/education), we find

$$P(Y=1, Z=1) = 0.19 \neq P(Y=1)P(Z=1) = 0.31 \cdot 0.36 = 0.1116.$$

Therefore, the two events are not independent. In this case, the two random variables are dependent.

Independence

Y and Z are **independent** random variables if, for all a and b, the bivariate distribution function is the product of the marginal distribution functions:

$$F_{YZ}(a,b) = F_Y(a)F_Z(b).$$

If this property is not satisfied, we say that X and Y are **dependent**.

The random variables Y and Z of Table 3.1 are independent, and those of Table 3.2 are dependent.

If Y and Z are independent and have finite second moments, then Y and Z are uncorrelated. The reverse is not true!

3.6 Random vectors

The above concepts can be generalized to any k-variate random vector $X = (X_1, \dots, X_k)$. The joint CDF of X is

$$F_X(x) = P(X_1 \leq x_1, \dots, X_k \leq x_k).$$

X has independent entries if

$$F_X(x) = \prod_{i=1}^{k} P(X_i \leq x_i) = \prod_{i=1}^{k} F_{X_i}(x_i)$$

If $F_X(x)$ is a continuous CDF, the joint k-dimensional density is

$$f_X(x) = f_X(x_1, \dots, x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_X(x_1, \dots, x_k).$$

The expectation vector of X is

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_k] \end{pmatrix},$$

and the covariance matrix of X is

$$\begin{aligned} Var[X] &= E[(X - E[X])(X - E[X])'] \\ &= \begin{pmatrix} Var[X_1] & Cov(X_1, X_2) & \dots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var[X_2] & \dots & Cov(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \dots & Var[X_k] \end{pmatrix} \end{aligned}$$

For any random vector X, the covariance matrix $Var[X]$ is symmetric and positive semi-definite.

3.7 Conditional distributions

Conditional probability

The conditional probability of an event A given an event B with $P(B) > 0$ is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Let's revisit the wage and schooling example from Table 3.2:

$$P(Y = 1 \mid Z = 1) = \frac{P(\{Y = 1\} \cap \{Z = 1\})}{P(Z = 1)} = \frac{0.19}{0.36} = 0.53$$

$$P(Y = 1 \mid Z = 0) = \frac{P(\{Y = 1\} \cap \{Z = 0\})}{P(Z = 0)} = \frac{0.12}{0.64} = 0.19$$

Note that

$$P(Y = 1 \mid Z = 1) = 0.53 > 0.31 = P(Y = 1)$$

implies

$$P(\{Y = 1\} \cap \{Z = 1\}) > P(Y = 1) \cdot P(Z = 1).$$

If $P(A \mid B) = P(A)$, then the events A and B are independent. If $P(A \mid B) \neq P(A)$, they are dependent.

Conditional distribution of continuous variables

Consider the density $f_{YZ}(a, b)$ of two continuous random variables Y and Z. The **conditional density** of Y given $Z = b$ is

$$f_{Y|Z}(a \mid b) = \frac{f_{YZ}(a, b)}{f_Z(b)}.$$

The **conditional distribution** of Y given $Z = b$ is

$$F_{Y|Z}(a \mid b) = \int_0^a f_{Y|Z}(u \mid b)\ \mathrm{d}u.$$

If Y is continuous and Z is discrete, the **conditional distribution function** of Y given $\{Z = b\}$ with $P(Z = b) > 0$ is

$$F_{Y|Z}(a \mid b) = P(Y \leq a \mid Z = b) = \frac{P(Y \leq a, Z = b)}{P(Z = b)}.$$

If $F_{Y|Z}(a \mid b)$ is differentiable with respect to b, the **conditional density** of Y given $Z = b$ is

$$f_{Y|Z}(a \mid b) = \frac{\partial}{\partial a} F_{Y|Z}(a \mid b).$$

We often are interested in conditioning on multiple variables, such as the wage given a particular education and experience level. Let $f(y, x) = f(y, x_1, \ldots, x_k)$ be the joint density of the composite random vector $(Y, X_1, \ldots, X_k)$ with $X = (X_1, \ldots, X_k)$. The conditional density of a random variable Y given $X = x = (x_1, \ldots, x_k)'$ is

$$f_{Y|X}(y \mid x) = f(y \mid x_1, \ldots, x_k) = \frac{f(y, x_1, \ldots, x_k)}{f_X(x_1, \ldots, x_k)} = \frac{f(y, x)}{f_X(x)}$$

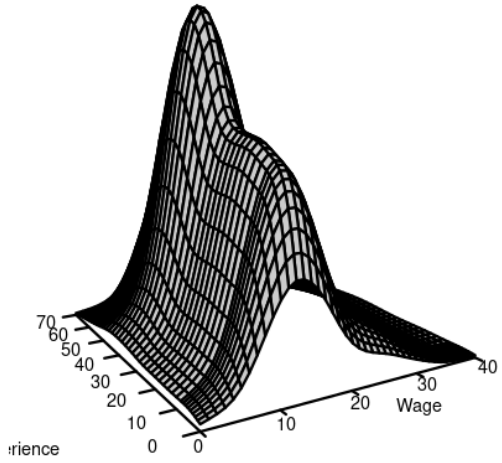


Figure 3.8: Joint PDF of wage and experience

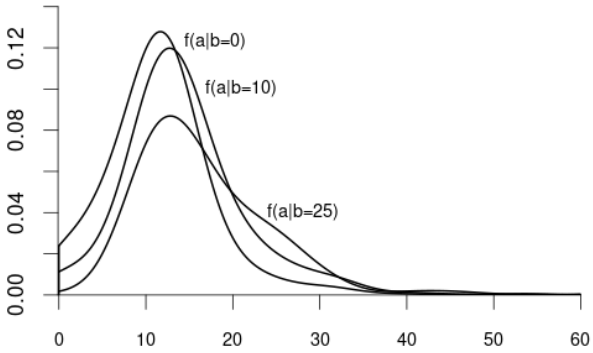


Figure 3.9: Conditional PDFs of wage given experience

The conditional distribution of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \int_0^y f(u \mid x) \, \mathrm{d}u.$$

3.8 Conditional expectation

Conditional expectation function

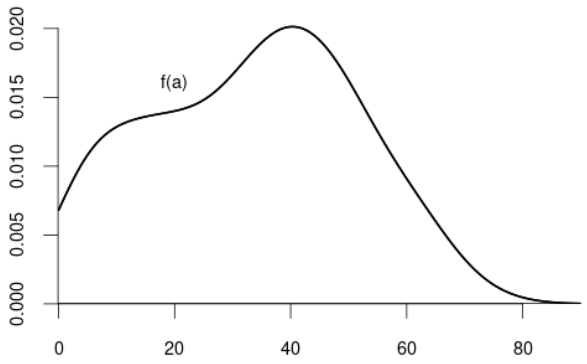


Figure 3.10: PDF of variable experience

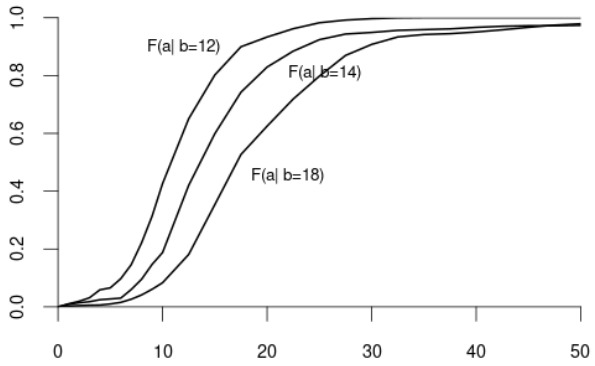


Figure 3.11: Conditional CDFs of wage given education

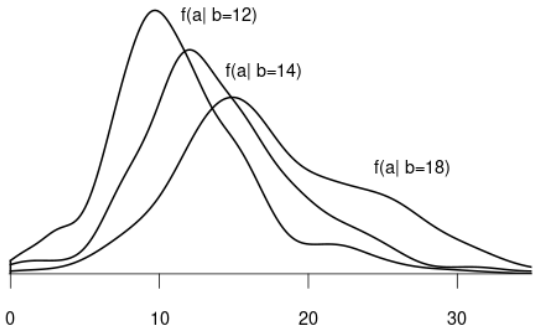
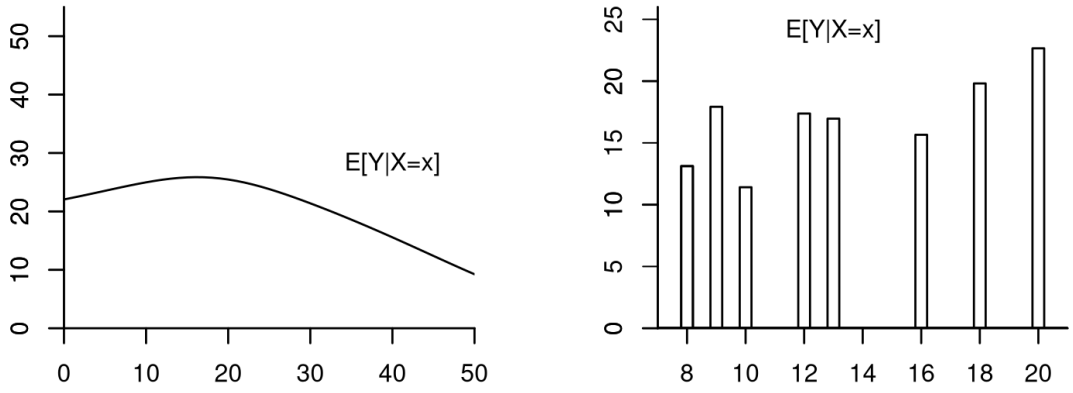


Figure 3.12: Conditional PDFs of wage given education

The **conditional expectation** of Y given $X = x$ is the expected value of the distribution $F_{Y|X}(y \mid x)$. For continuous Y with conditional density $f_{Y|X}(y \mid x)$, the conditional expectation is

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) \, \mathrm{d}y.$$



(a) CEF wage given experience

(b) CEF wage given education

Figure 3.13: Conditional expectation functions

Consider again the wage and experience example. Suppose that the conditional expectation has the functional form

$$E[wage \mid experience = x] = m(x) = 14.5 + 0.9x - 0.017x^2.$$

E.g., for $x = 10$ we have $E[wage \mid experience = 10] = m(10) = 21.8$.

Note that $m(x) = E[wage \mid experience = x]$ is not random. It is a feature of the joint distribution.

Sometimes, it is useful not to fix the experience level to a certain value but to treat it as random:

$$\begin{aligned} E[wage \mid experience] &= m(experience) \\ &= 14.5 + 0.9experience - 0.017experience^2 \end{aligned}$$

$m(experience) = E[wage \mid experience]$ is a function of the random variable experience and, therefore, itself a random variable.

The conditional expectation function (CEF) of Y given the specific event $\{X = x\}$ is

$$m(x) = E[Y \mid X = x].$$

$m(x)$ is deterministic (non-random) and a feature of the joint distribution.

The conditional expectation function (CEF) of Y given the random vector X is

$$m(X) = E[Y \mid X].$$

$m(X)$ is a function of the random vector X and therefore itself a random variable.

3.9 Law of iterated expectations

Rules of calculation for the conditional expectation

Let Y be a random variable and X a random vector.

(i) Law of the iterated expectations (LIE):

$$E[E[Y \mid X]] = E[Y].$$

A more general LIE: For any two random vectors X and $\widetilde{X}$,

$$E[E[Y \mid X, \widetilde{X}] \mid X] = E[Y \mid X].$$

(ii) Conditioning theorem (CT): For any function $g(\cdot)$,

$$E[g(X)Y \mid X] = g(X)E[Y \mid X].$$

(iii) If Y and X are independent then $E[Y \mid X] = E[Y]$.

3.10 Conditional variance

Conditional variance

If $E[Y^2] < \infty$, the **conditional variance** of Y given the event $\{X = x\}$ is

$$Var[Y \mid X = x] = E[(Y - E[Y \mid X = x])^2 \mid X = x].$$

The conditional variance of Y given the random vector X is

$$Var[Y \mid X] = E[(Y - E[Y \mid X])^2 \mid X].$$

3.11 Best predictor

A typical application is to find a good prediction for the outcome of a random variable Y. Recall that the expected value $E[Y]$ is the best predictor for Y in the sense that $g^* = E[Y]$ minimizes $E[(Y-g)^2]$.

With the knowledge of an additional random vector X, we can use the joint distribution of Y and X to improve the prediction of Y.

It turns out that the CEF $m(X) = E[Y \mid X]$ is the best predictor for Y given the information contained in the random vector X:

Best predictor

If $E[Y^2] < \infty$, then the CEF $m(X) = E[Y \mid X]$ minimizes the expected squared error $E[(Y - g(X))^2]$ among all predictor functions $g(X)$.

Let us find the function $g(\cdot)$ that minimizes $E[(Y-g(X))^2]$:

$$\begin{aligned} &E[(Y-g(X))^2] = E[(Y-m(X)+m(X)-g(X))^2] \\ &= \underbrace{E[(Y-m(X))^2]}_{=(i)} + 2\underbrace{E[(Y-m(X))(m(X)-g(X))]}_{=(ii)} + \underbrace{E[(m(X)-g(X))^2]}_{(iii)} \end{aligned}$$

The first term (i) does not depend on $g(\cdot)$ and is finite if $E[Y^2] < \infty$.

For the second term (ii), we use the LIE and CT:

$$\begin{aligned} &E[(Y-m(X))(m(X)-g(X))] \\ &= E[E[(Y-m(X))(m(X)-g(X)) \mid X]] \\ &= E[E[Y-m(X) \mid X](m(X)-g(X))] \\ &= E[(\underbrace{E[Y \mid X]}_{=m(X)} - m(X))(m(X)-g(X))] = 0 \end{aligned}$$

The third term (iii) $E[(m(X)-g(X))^2]$ is minimal if $m(\cdot) = g(\cdot)$

Therefore, $m(X) = E[Y \mid X]$ minimizes $E[(Y-g(X))^2]$.

The best predictor for Y given X is $m(X) = E[Y \mid X]$, but Y can typically only partially be predicted. We have a prediction error (CEF error)

$$e = Y - E[Y \mid X].$$

The conditional expectation of the CEF error does not depend on X and is zero:

$$\begin{aligned} E[e \mid X] &= E[(Y - m(X)) \mid X] \\ &= E[Y \mid X] - E[m(X) \mid X] \\ &= m(X) - m(X) = 0 \end{aligned}$$

We say that Y is **conditional mean independent** of Z if $E[Y \mid Z]$ does not depend on Z.

If Y and Z are independent, they are also conditional mean independent, but not necessarily vice versa. If Y and Z are conditional mean independent, they are also uncorrelated, but not necessarily vice versa.

Since the CEF is the best predictor of Y, it is of great interest to study the CEF in practice. Much of the statistical and econometric research deals with methods to approximate and estimate the CEF. This field of statistics is called **regression analysis**.

Consider the following model for Y and X:

$$Y = m(X) + e, \quad E[e \mid X] = 0. \tag{3.1}$$

We call $m(\cdot)$ **regression function** and e **error term**.

From equation Equation 3.1 it follows that

$$E[Y \mid X] = E[m(X) + e \mid X] = E[m(X) \mid X] + E[e \mid X] = m(X).$$

I.e., the nonparametric regression model is a model for the CEF.

If $m(\cdot)$ is a linear function, then Equation 3.1 is a **linear regression model**. We will study this model in detail in the next sections.

3.12 Combining normal variables

Some of the distributions commonly encountered in econometrics are combinations of univariate normal distributions, such as the multivariate normal, chi-squared, Student t, and F distributions.

3.12.1 χ^2-distribution

Let $Z_1, \ldots, Z_m$ be independent $\mathcal{N}(0,1)$ random variables. Then, the random variable

$$Y = \sum_{i=1}^{m} Z_i^2$$

is **chi-square distributed** with parameter m, written $Y \sim \chi^2_m$.

The parameter m is called the degrees of freedom.

Expectation and variance:

$$E[Y] = m, \quad var[Y] = 2m$$

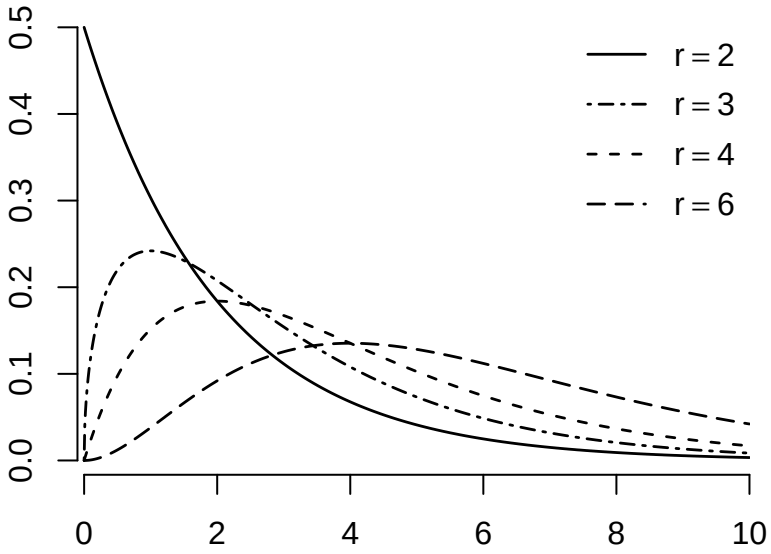


Figure 3.14: χ^2 -distribution

3.12.2 F-distribution

If $Q_1 \sim \chi^2_m$ and $Q_2 \sim \chi^2_r$, and if Q_1 and Q_2 are independent, then

$$Y = \frac{Q_1/m}{Q_2/r}$$

is **F-distributed** with parameters m and r, written $Y \sim F_{m,r}$.

The parameter m is called the degrees of freedom in the numerator; r is the degree of freedom in the denominator.

If $r \to \infty$ then the distribution of mY approaches χ^2_m

3.12.3 Student t-distribution

If $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2_m$, and Z and Q are independent, then

$$Y = \frac{Z}{\sqrt{Q/m}}$$

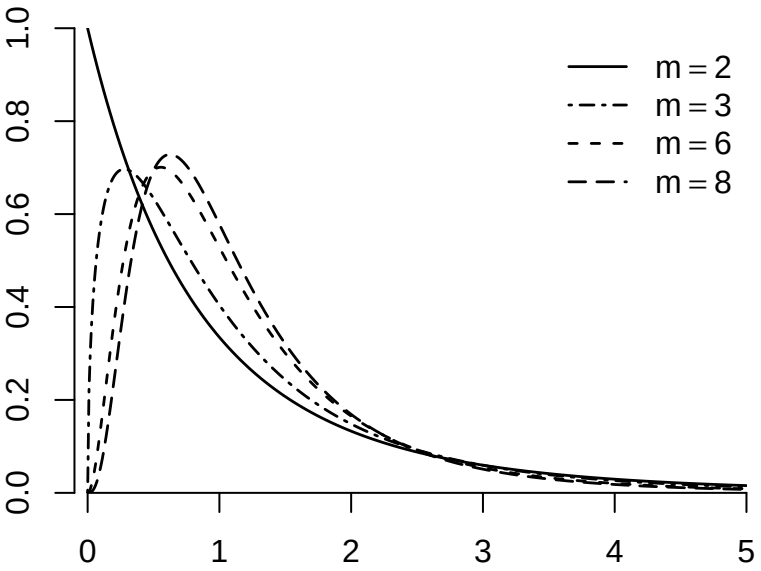


Figure 3.15: F-distribution

is t**-distributed** with parameter m degrees of freedom, written $Y \sim t_m$.

Expectation, variance, and moments:

$$E[Y] = 0 \quad (\text{if } m \geq 2),$$
$$var[Y] = \frac{m}{m-2} \quad (\text{if } m \geq 3)$$

The first $m-1$ moments are finite: $E[|Y|^r] < \infty$ for $r \leq m-1$ and $E[|Y|^r] = \infty$ for $r \geq m$.

The t-distribution with $m = 1$ is also called **Cauchy distribution**. The t-distributions with 1, 2, 3, and 4 degrees of freedom are heavy-tailed distributions. If $m \to \infty$ then $t_m \to \mathcal{N}(0,1)$

3.12.4 Multivariate normal distribution

Let $X_1, \ldots, X_k$ be independent $\mathcal{N}(0,1)$ random variables. Then, the k-vector $X = (X_1, \ldots, X_k)'$ has the **multivariate standard normal distribution**, written $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$. Its joint density is

$$f(x) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{x'x}{2}\right).$$

If $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$ and $\widetilde{X} = \mu + \boldsymbol{B}X$ for a $q \times 1$ vector μ and a $q \times k$ matrix $\boldsymbol{B}$, then $\widetilde{X}$ has a **multivariate normal distribution** with parameters μ and $\Sigma = \boldsymbol{B}\boldsymbol{B}'$, written $\widetilde{X} \sim \mathcal{N}(\mu, \Sigma)$. Its joint density is

$$f(x) = \frac{1}{(2\pi)^{k/2}(\det(\Sigma))^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big).$$

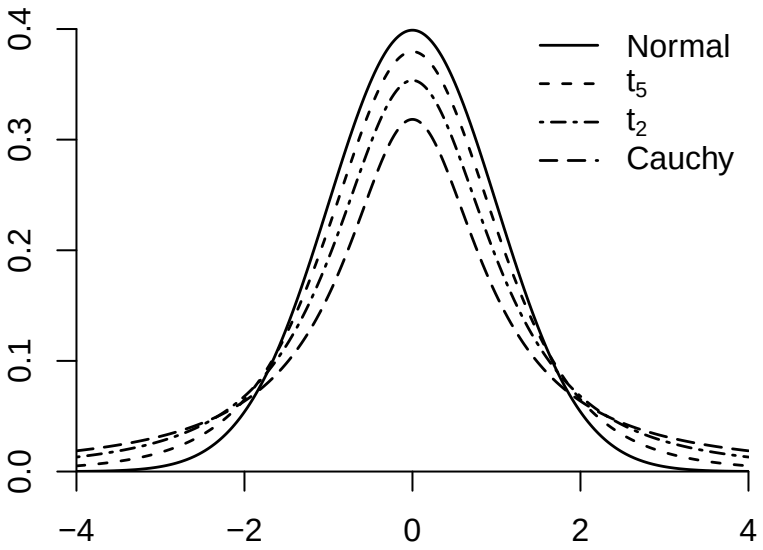


Figure 3.16: Student t-distribution

The expectation vector and covariance matrix are

$$E[\widetilde{X}] = \mu, \quad var[\widetilde{X}] = \Sigma.$$

3.12.5 R-commands for parametric distributions

	get CDF $F(a)$	quantile function $q(p)$	generate n independent random numbers
$\mathcal{N}(0,1)$	`pnorm(a)`	`qnorm(p)`	`rnorm(n)`
χ^2_r	`pchisq(a,r)`	`qchisq(p,r)`	`rchisq(n,r)`
t_r	`pt(a,r)`	`qt(p,r)`	`rt(n,r)`
$F_{r,k}$	`pf(a,r,k)`	`qf(p,r,k)`	`rf(n,r,k)`

3.13 Additional reading

- Stock and Watson (2019), Section 2
- Hansen (2022a), Section 4
- Hansen (2022b), Section 2
- Davidson and MacKinnon (2004), Section 1

4 Sampling

4.1 Data

Consider a **dataset** $\{X_1, \dots, X_n\}$ with n **observations** or individuals $i = 1, \dots, n$. Each observation X_i consists of k **variables** or measurements, i.e., X_i is a $k \times 1$ vector $X_i = (X_{i1}, X_{i2}, \dots, X_{ik})' \in \mathbb{R}^k$.

For instance, let's consider a sub-sample of $n = 100$ observations of the 2021 German General Social Survey (ALLBUS) of the variables *wage* and *education*.

We have a dataset of $n = 100$ bivariate observation vectors

$$X_i = \begin{pmatrix} W_i \\ S_i \end{pmatrix}, \quad i = 1, \dots, 100,$$

where W_i and S_i are the wage level and years of schooling of individual i.

From the perspective of empirical analysis, a dataset is simply an array of numbers that are fixed and presented to a researcher.

From the perspective of statistical theory, a dataset consists of n repeated realizations of a k-variate random variable X with distribution F. In the above case, we have $X = (W, S)'$, where W and S are random variables for wage and education.

The distribution F is called **population distribution**. The population can be thought of as an infinitely large group of hypothetical individuals from which we draw our data rather than a fixed group of a physical population. A finite physical population would be the 83 million inhabitants of Germany. An infinite population is a theoretical auxiliary construct. For example, one can imagine the infinite group of potentially born persons with all their possible characteristics.

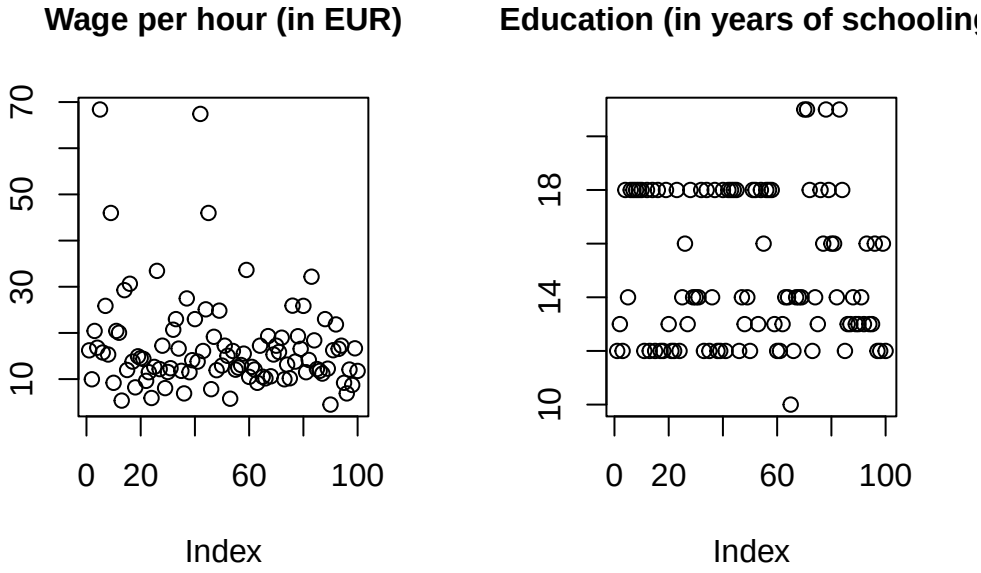
A **parameter** is a feature (function) of the population distribution F, such as expectations, variances, or correlations. Statistical analysis is concerned with inference on parameters of the population distribution F.

Table 4.1: ALLBUS data sub-sample

Person	Wage	Education
1	16.23	12
2	9.96	13
3	20.43	12
4	16.80	18
5	68.39	14
6	15.76	18
7	25.86	18
8	15.33	18
9	45.98	18
10	9.20	18
11	20.43	12
12	20.04	18
13	5.35	12
14	29.26	18
15	11.97	12
16	30.65	18
17	13.79	12
18	8.21	12
19	14.94	18
20	14.55	13
21	14.37	12
22	9.62	12
23	11.49	18
24	5.91	12
25	12.64	14
26	33.44	16
27	12.17	13
28	17.24	18
29	8.05	14
30	11.54	14
31	12.36	14
32	20.69	18
33	22.99	12
34	16.61	18
35	11.79	12
36	6.90	14
37	27.47	18
38	11.49	12
39	14.10	12
40	22.99	18
41	13.79	12
42	67.43	18
43	16.09	18
44	25.11	18
45	45.98	18
46	7.82	12
47	19.16	14
48	11.93	13

4.2 Random sampling

In statistical analysis, a dataset $\{X_1, \dots, X_n\}$ that is drawn from some population F is also called **sample**.



The ALLBUS data are **cross-sectional** data, where n individuals are randomly selected from the German population and independently interviewed on k variables. The ALLBUS data consists of n independently replicated random experiments.

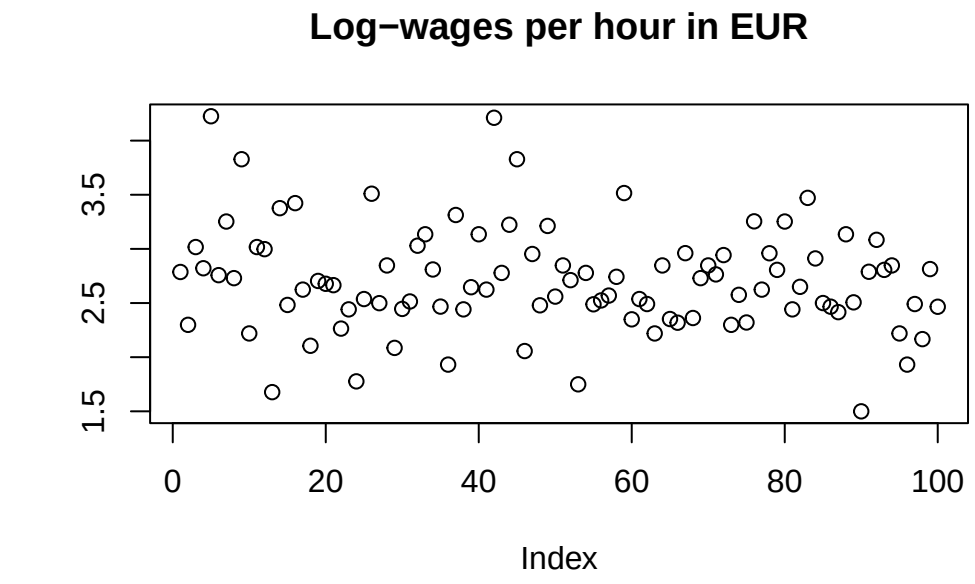
i.i.d. sample / random sample

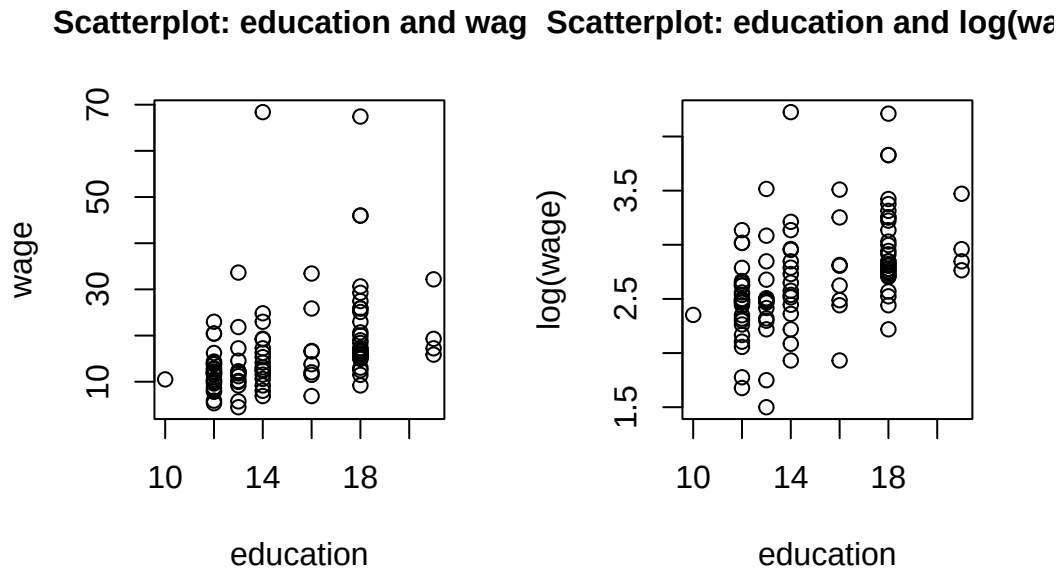
A collection of random vectors $\{X_1, \dots, X_n\}$ is **i.i.d. (independent and identically distributed)** if X_i and X_j are mutually independent and have the same distribution F for all $i \neq j$.

An i.i.d. dataset or i.i.d. sample is also called a **random sample**. F is called **population distribution** or **data-generating process (DGP)**.

Any transformed sample $\{g(X_1), \dots, g(X_n)\}$ of an i.i.d. sample $\{X_1, \dots, X_n\}$ is also an i.i.d. sample (g may be any function). For instance, we may consider the log-transformed data, which produces a less skewed and fat-tailed distribution:

Below, you find a simulated random sample from a Poisson distribution with parameter $\lambda = 14$:

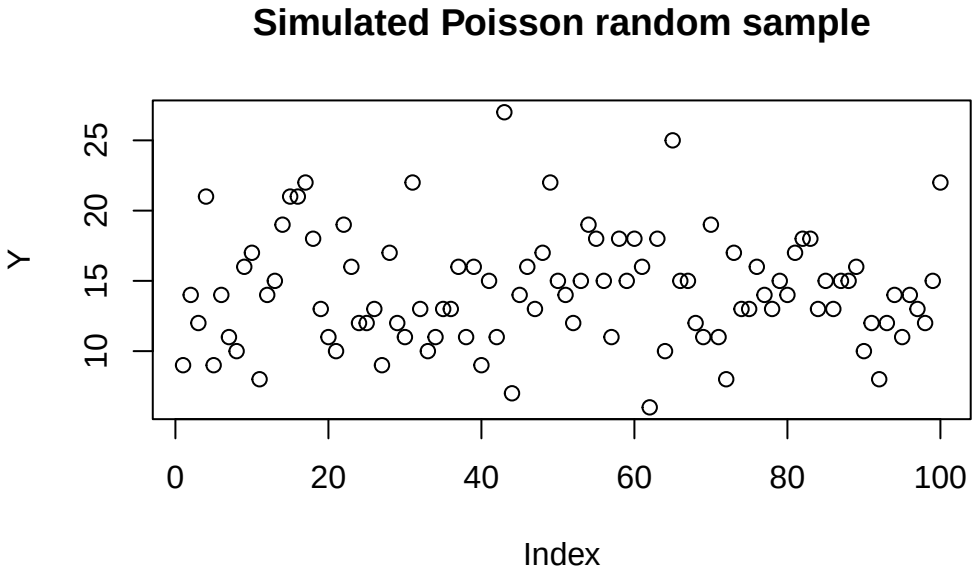





```r
Y = rpois(100, 14)
plot(Y, main="Simulated Poisson random sample")
```

Statistical analysis is concerned with inference on parameters of the population distribution F from which the data is sampled.

Sampling methods of obtaining economic datasets that may be considered as random sampling are:



- **Survey sampling**
  Examples: representative survey of randomly selected households from a list of residential addresses; online questionnaire to a random sample of recent customers
- **Administrative records**
  Examples: data from a government agency database, Statistisches Bundesamt, ECB, etc.
- **Direct observation**
  Collected data without experimental control and interactions with the subject. Example: monitoring customer behavior in a retail store
- **Web scraping**
  Examples: collected house prices on real estate sites or hotel/electronics prices on booking.com/amazon, etc.
- **Field experiment**
  To study the impact of a treatment or intervention on a treatment group compared with a control group. Example: testing the effectiveness of a new teaching method by implementing it in a selected group of schools and comparing results to other schools with traditional methods
- **Laboratory experiment**
  Example: a controlled medical trial for a new drug

4.3 Dependent sampling

Examples of cross-sectional data sampling that may produce some dependence across observations are:

- **Stratified sampling**
  The population is first divided into homogenous subpopulations (strata), and a random sample is obtained from each stratum independently. Examples: divide companies into industry strata (manufacturing, technology, agriculture, etc.) and sample from each stratum; divide the population into income strata (low-income, middle-income, high-income).
  The sample is independent within each stratum, but it is not between different strata. The strata are defined based on specific characteristics that may be correlated with the variables collected in the sample.

- **Clustered sampling**
  Entire subpopulations are drawn. Example: new teaching methods are compared to traditional ones on the student level, where only certain classrooms are randomly selected, and all students in the selected classes are evaluated.
  Within each cluster (classroom), the sample is dependent because of the shared environment and teacher's performance, but between classrooms, it is independent.

Other types of data we often encounter in econometrics are time series data, panel data, or spatial data:

- **Time series data** consists of observations collected at different points in time, such as stock prices, daily temperature measurements, or GDP figures. These observations are ordered and typically show temporal trends, seasonality, and autocorrelation.
- **Panel data** involves observations collected on multiple entities (e.g., individuals, firms, countries) over multiple time periods.
- **Spatial data** includes observations taken at different geographic locations, where values at nearby locations are often correlated.

Time series, panel, and spatial data cannot be considered a random sample given their temporal or geographic dependence.

4.4 Time series data

Time series data $\{Y_1, \dots, Y_n\}$ is a sequence of real-valued observations arranged chronologically and indexed by time. In contrast to cross-sectional data, we typically use the index t to indicate the observation index, i.e., we write Y_t for the time t observation.

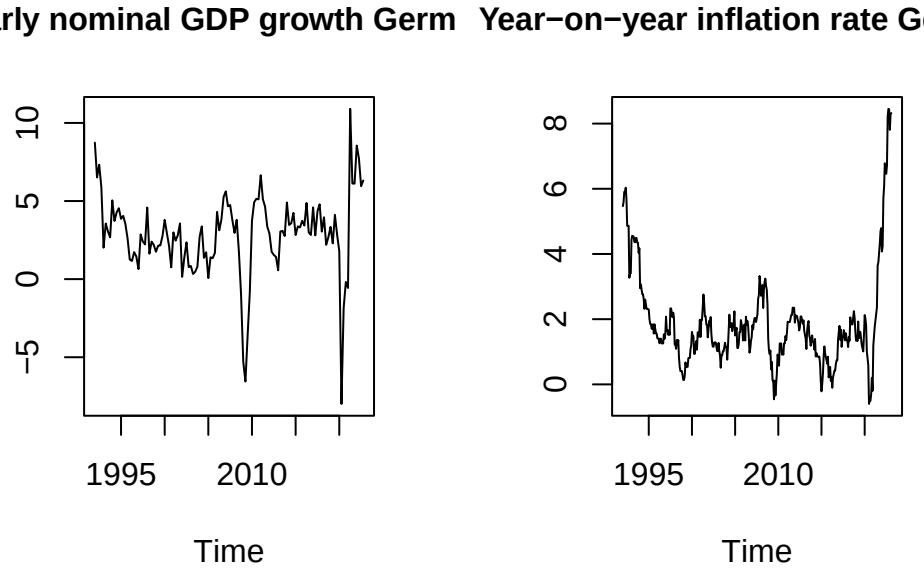
The order defined by the time indices is essential to time series data. We usually expect observations close in time to be strongly dependent and observations at greater distances to be less dependent. It is a fundamental difference to cross-sectional data, where each index represents an individual, and the ordering of the indices is interchangeable.

The **time series process** is the underlying doubly infinite sequence of random variables

$$\{Y_t\}_{t\in\mathbb{Z}} = \{\dots, Y_{-1}, Y_0, \underbrace{Y_1, \dots, Y_n}_{\text{observed part}}, Y_{n+1}, \dots\}.$$

where the time series sample $\{Y_1, \dots, Y_n\}$ is only the observed part of the process.

Examples of **time series data** are the nominal GDP growth and the inflation rate of Germany.



Stationary time series

A time series Y_t is called **stationary** if the mean μ and the autocovariances $\gamma(\tau)$ do not depend on the time point t. That is,

$$\mu := E[Y_t] < \infty, \quad \text{for all } t,$$

and

$$\gamma(\tau) := Cov(Y_t, Y_{t-\tau}) < \infty \quad \text{for all } t \text{ and } \tau.$$

The function $\gamma(\cdot)$ of a stationary time series Y_t is called **autocovariance function**, and $\gamma(\tau)$ is the autocovariance of order τ. The **autocorrelation of order** τ is

$$\rho(\tau) = \frac{Cov(Y_t, Y_{t-\tau})}{Var[Y_t]}, \quad \tau \in \mathbb{Z}.$$

The autocorrelations of stationary time series typically decay to zero quite quickly as τ increases. If $\rho(\tau) \to 0$ as $\tau \to \infty$ quickly enough so that $\sum_{\tau=1}^{\infty} \rho(\tau) < \infty$, the time series is called a **short memory time series**. For short memory time series, observations close in time may be highly correlated, but observations farther apart have little dependence.

One of the most commonly studied time series processes is the **autoregressive process of order one** with parameter ϕ. It is defined as

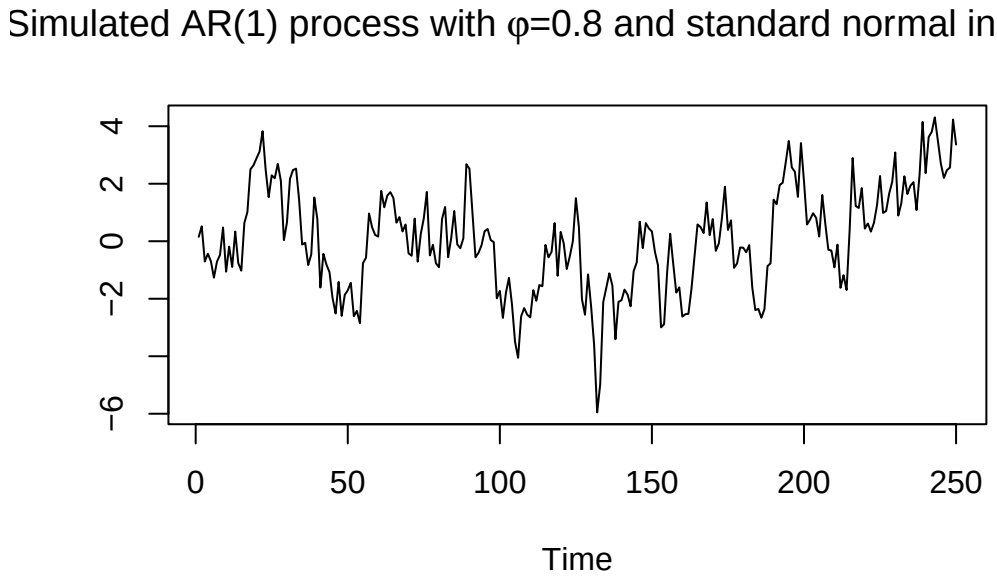
$$Y_t = c + \phi Y_{t-1} + u_t,$$

where $\{u_t\}$ is an i.i.d. process of increments with $E[u_t] = 0$ and $Var[u_t] = \sigma_u^2$, and c is a constant. If $|\phi| < 1$, the AR(1) process is stationary with

$$\mu = \frac{c}{1-\phi}, \quad \gamma(\tau) = \frac{\phi^\tau \sigma_u^2}{1-\phi^2}, \quad \rho(\tau) = \phi^\tau, \quad \tau \geq 0.$$

Its autocorrelations $\rho(\tau) = \phi^\tau$ decay exponentially in the lag order τ.

```
u = rnorm(250)
AR1 = stats::filter(u, 0.8, "recursive")
plot(AR1, main=expression(paste("Simulated AR(1) process with ",varphi,"=0.8 and standard no
```



More complex time dependencies can be modeled using an AR(p) process

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + u_t.$$

Many macroeconomic time series, such as *GDP growth* or *inflation rates* can be modeled using autoregressive processes.

A time series Y_t is **nonstationary** if the mean $E[Y_t]$ or the autocovariances $Cov(Y_t, Y_{t-\tau})$ change with t, i.e., if there exist time points $s \neq t$ with

$$E[Y_t] \neq E[Y_s] \quad \text{or} \quad Cov(Y_t, Y_{t-\tau}) \neq Cov(Y_s, Y_{s-\tau})$$

for some τ.

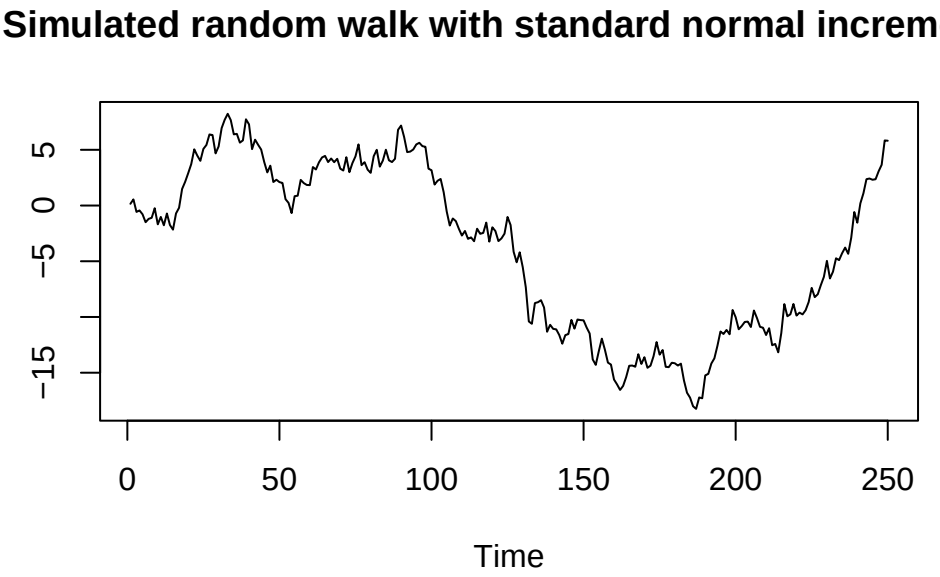
The **simple random walk** is an example of a nonstationary time series process. It is an AR(1) process with $\phi = 1$, $c = 0$, and starting value $Y_0 = 0$, i.e.,

$$Y_t = Y_{t-1} + u_t, \quad t \geq 1.$$

By backward substitution, it can be expressed as the cumulative sum $Y_t = \sum_{j=1}^{t} u_j$.

It is nonstationary since $Cov(Y_t, Y_{t-\tau}) = (t-\tau)\sigma_u^2$, which depends on t and becomes larger as t gets larger.

```
RW = stats::filter(u, 1, "recursive")
plot(RW, main= "Simulated random walk with standard normal increments", ylab = "")
```



4.5 Additional reading

- Stock and Watson (2019), Sections 1-2, 15
- Hansen (2022a), Section 6
- Hansen (2022b), Section 3
- Davidson and MacKinnon (2004), Section 1

4.6 R-codes

statistics-sec4.R

5 Estimation

5.1 Parameters and estimators

A **parameter** θ is a feature (function) of the population distribution F. We often use Greek letters for parameters. The expectation, variance, correlation, autocorrelation, and regression coefficients are parameters.

A **statistic** is a function of a sample $\{X_i,\ i = 1, \ldots, n\}$. An **estimator** $\widehat{\theta}$ for θ is a statistic intended as a guess about θ. It is a function of the random vectors $X_1, \ldots, X_n$ and, therefore, a random variable. When an estimator $\widehat{\theta}$ is calculated in a specific realized sample, we call $\widehat{\theta}$ an **estimate**.

5.2 Population moments and sample moments

Consider the **moments** of some bivariate random variable (Y, Z). The **sample moments** are the corresponding average values in the sample $\{(Y_i, Z_i),\ i = 1, \ldots, n\}$:

population moment	sample moment
$E[Y]$	$\frac{1}{n}\sum_{i=1}^{n} Y_i$
$E[Z]$	$\frac{1}{n}\sum_{i=1}^{n} Z_i$
$E[Y^2]$	$\frac{1}{n}\sum_{i=1}^{n} Y_i^2$
$E[Z^2]$	$\frac{1}{n}\sum_{i=1}^{n} Z_i^2$
$E[YZ]$	$\frac{1}{n}\sum_{i=1}^{n} Y_i Z_i$

5.3 Moment estimators

Many parameters of interest can be expressed as a function of the population moments. A common estimation approach is the **moment estimator**, where the population moments are replaced by their corresponding sample moments.

The vectors `wg`, `edu`, `gdp`, and `infl` contain the datasets on wage, education, GDP growth, and inflation from the previous section, and `AR1` and `RW` are the simulated AR(1) and random walk series.

Mean and sample mean

$$\mu_Y = E[Y]$$
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

If `Y` is your sample stored in a vector, the R command for the sample mean is: `mean(Y)` or `sum(Y)/length(Y)`

```
mean(wg)
```

```
[1] 17.0556
```

```
mean(edu)
```

```
[1] 14.95
```

Variance and sample variance

$$\sigma_Y^2 = Var[Y] = E[(Y - E[Y]^2)] = E[Y^2] - E[Y]^2$$
$$\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2$$

Sample variance in R: `mean((Y-mean(Y))^2)`.
Note that the command `var(Y)` returns the bias-corrected sample variance, which slightly deviates from the sample variance:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{n}{n-1}\hat{\sigma}_Y^2. \tag{5.1}$$

It is common to report the bias-corrected version s_Y^2 instead of $\hat{\sigma}_Y^2$, as it has slightly better properties (see Section 5.6 below). However, this does not matter for large samples since $n/(n-1) \to 1$ as $n \to \infty$.

```r
mean((wg - mean(wg))^2)   ## sample variance
```

```
[1] 108.9961
```

```r
mean((edu - mean(edu))^2)
```

```
[1] 7.4875
```

```r
var(wg)   ## bias-corrected sample variance
```

```
[1] 110.0971
```

```r
var(edu)
```

```
[1] 7.563131
```

Standard deviation and sample standard deviation

$$\sigma_Y = sd(Y) = \sqrt{Var[Y]}$$
$$\hat{\sigma}_Y = \sqrt{\hat{\sigma}_Y^2}$$

Sample standard deviation in R: `sqrt(mean((Y-mean(Y))^2))`.
Note that the command `sd(Y)` returns the bias-corrected sample standard deviation, which is the square root of Equation 5.1, and slightly deviates from $\hat{\sigma}_Y$. Similarly to the bias-corrected variance, the bias-corrected standard deviation s_Y is typically reported instead of $\hat{\sigma}_Y$.

```r
sqrt(mean((wg - mean(wg))^2)) ## sample standard deviation
```

```
[1] 10.44012
```

```r
sd(wg) ## bias-corrected sample standard deviation
```

```
[1] 10.49272
```

Skewness and sample skewness

$$skew(Y) = \frac{E[(Y - E[Y])^3}{sd(Y)^3}$$
$$\widehat{skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^3}{\hat{\sigma}_Y^3}$$

Sample skewness in R: `mean((Y-mean(Y))^3)/sqrt(mean((Y-mean(Y))^2))^3`. Alternatively, the command `skewness(Y)` of the package `moments` can be used.

```
library(moments)
skewness(wg)
```

```
[1] 2.767215
```

```
skewness(edu)
```

```
[1] 0.4184468
```

Kurtosis and sample kurtosis

$$kurt(Y) = \frac{E[(Y - E[Y])^4}{sd(Y)^4}$$
$$\widehat{kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^4}{\hat{\sigma}_Y^4}$$

Sample skewness in R: `mean((Y-mean(Y))^4)/mean((Y-mean(Y))^2)^2`. Alternatively, the command `kurtosis(Y)` of the package `moments` can be used.

```
kurtosis(wg)
```

```
[1] 12.83567
```

```
kurtosis(edu)
```

```
[1] 1.906658
```

Covariance and sample covariance

$$\sigma_{YZ} = Cov(Y, Z) = E[(Y - E[Y])(Z - E[Z])] = E[YZ] - E[Y]E[Z]$$
$$\hat{\sigma}_{YZ} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z}) = \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i - \overline{Y} \cdot \overline{Z}$$

Sample covariance in R: `mean((Y-mean(Y))*(Z-mean(Z)))`. Note that `cov(Y,Z)` returns the bias-corrected sample covariance $s_{YZ} = n/(n-1) \cdot \hat{\sigma}_{YZ}$.

```
mean((wg - mean(wg))*(edu - mean(edu))) ## sample covariance
```

```
[1] 10.43228
```

```
cov(wg, edu) ## bias-corrected sample covariance
```

```
[1] 10.53766
```

Correlation and sample correlation

$$\rho_{YZ} = \frac{Cov(Y, Z)}{sd(Y)sd(Z)}$$
$$\hat{\rho}_{YZ} = \frac{\hat{\sigma}_{YZ}}{\hat{\sigma}_Y \hat{\sigma}_Z}$$

Sample correlation in R: `cor(Y,Z)`.

```
cor(wg, edu)
```

```
[1] 0.3651786
```

Autocovariance and sample autocovariance

$$\gamma(\tau) = Cov(Y_t, Y_{t-\tau}) = E[(Y_t - \mu_Y)(Y_{t-\tau} - \mu_Y)]$$
$$\hat{\gamma}(\tau) = \frac{1}{n}\sum_{i=\tau+1}^{n}(Y_i - \overline{Y})(Y_{i-\tau} - \overline{Y})$$

Sample autocovariances in R: `acf(Y, type = "covariance", plot = FALSE)`. Note that `acf(Y, type = "covariance")` returns a plot of the sample autocovariance function.

```r
acf(gdp, type = "covariance", plot = FALSE)
```

```
Autocovariances of series 'gdp', by lag

   0.00    0.25    0.50    0.75    1.00    1.25    1.50    1.75    2.00    2.25
 6.6348  4.2772  2.9628  1.5468 -0.7084 -0.9210 -1.3339 -1.7877 -1.7673 -1.2969
   2.50    2.75    3.00    3.25    3.50    3.75    4.00    4.25    4.50    4.75
-0.9022 -0.3446  0.2861  0.3126  0.6607  0.8516  0.4438  0.5634  0.2310 -0.1504
   5.00
-0.0926
```

```r
acf(infl, type = "covariance", plot = FALSE)
```

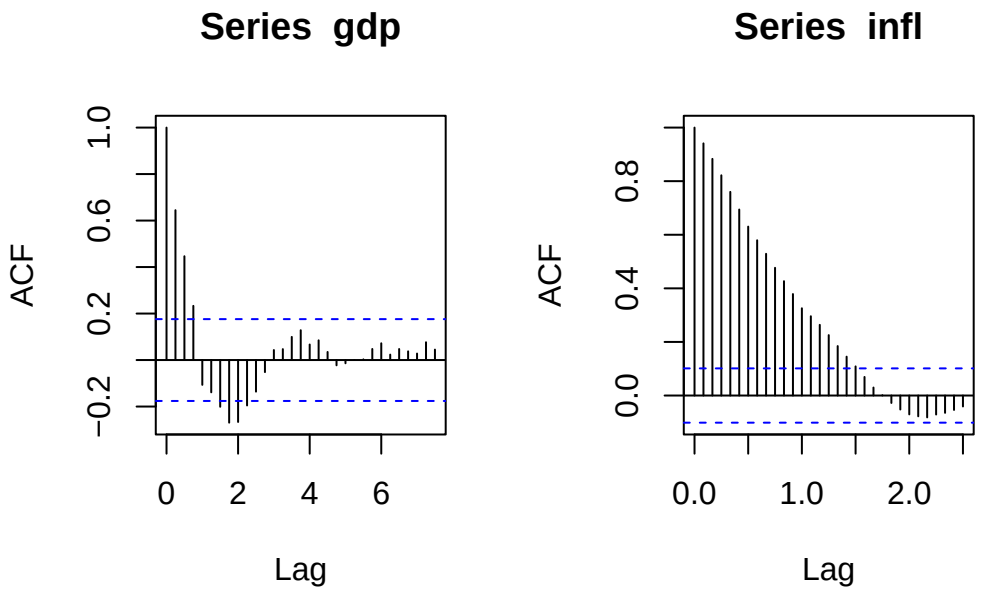
```
Autocovariances of series 'infl', by lag

  0.0000   0.0833   0.1667   0.2500   0.3333   0.4167   0.5000   0.5833
 2.25940  2.12625  1.99568  1.85704  1.71710  1.56860  1.42454  1.30920
  0.6667   0.7500   0.8333   0.9167   1.0000   1.0833   1.1667   1.2500
 1.19518  1.07702  0.96450  0.85573  0.73584  0.66869  0.59613  0.50944
  1.3333   1.4167   1.5000   1.5833   1.6667   1.7500   1.8333   1.9167
 0.41708  0.32785  0.24573  0.15681  0.06751  0.00289 -0.06278 -0.11801
  2.0000   2.0833
-0.15911 -0.17548
```

Autocorrelation and sample autocorrelation

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \frac{Cov(Y_t, Y_{t-\tau})}{Var[Y_t]}$$
$$\hat{\rho}(\tau) = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)}$$

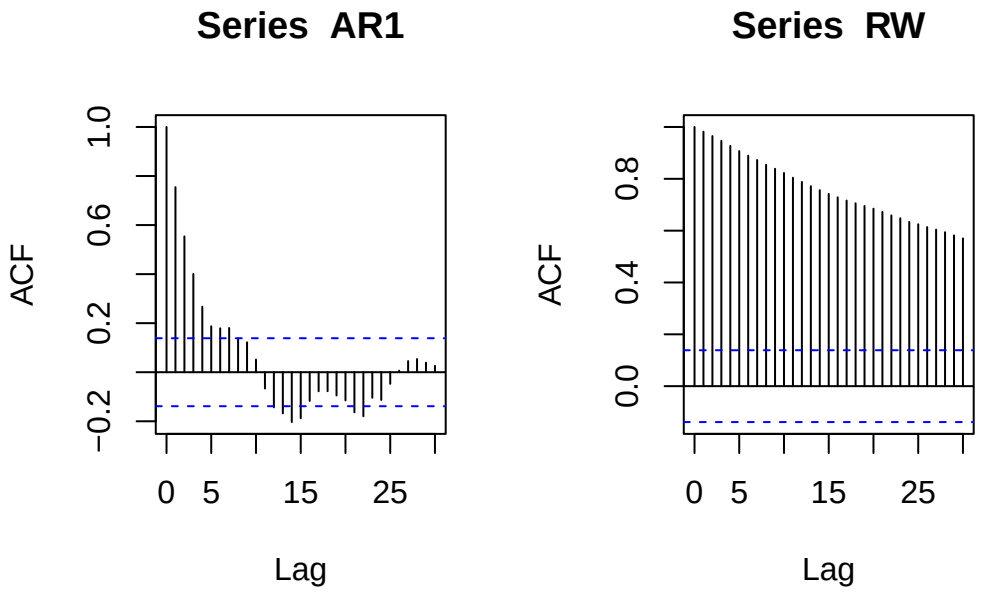
Sample autocorrelations in R: `acf(Y, plot = FALSE)`. Note that `acf(Y)` returns a plot of the sample autocorrelation function.

```r
par(mfrow=c(1,2))
acf(gdp, lag.max=30)
acf(infl, lag.max = 30)
```



Note that the lags on the x-axis are measured in yeas, where `gdp` is quarterly data and `infl` is monthly data.

```
par(mfrow=c(1,2))
acf(AR1, lag.max = 30)
acf(RW, lag.max = 30)
```



The ACF plots indicate the dynamic structure of the time series and whether they can be

regarded as a stationary and short-memory time series. The ACF of `gdp` tends to zero quickly, similar to the ACF of the `AR1`. They can be treated as stationary and short-memory time series. The ACF of `infl` tends to zero slowly, indicating a high persistence, so the short-memory condition may not be satisfied. The ACF of `RW` does not tend to zero. It is a nonstationary time series.

5.4 Consistency

Good estimators get closer and closer to the true parameter being estimated as the sample size n increases, eventually returning the true parameter value in a hypothetically infinitely large sample. This property is called **consistency**.

Consistency

An estimator $\hat{\theta}$ is **consistent** for a true parameter value θ if

$$\lim_{n\to\infty} P(|\hat{\theta}-\theta|>\epsilon)=0 \qquad \text{for any } \epsilon>0,$$

or, equivalently, if $P(|\hat{\theta}-\theta|\leq\epsilon)\to 1$ as $n\to\infty$.

If $\hat{\theta}$ is consistent for θ, we say that $\hat{\theta}$ **converges in probability** to θ. A common notation for convergence in probability is the probability limit:

$$\text{plim}_{n\to\infty}\,\hat{\theta}=\theta, \quad \text{or} \quad \hat{\theta}\xrightarrow{p}\theta.$$

Since an estimator $\hat{\theta}$ is usually a continuous random variable, it will almost never reach exactly the true parameter value: $P(\hat{\theta}=\theta)=0$. However, the larger the sample size, the higher should be the probability that $\hat{\theta}$ is close to the true value θ. Consistency means that, if we fix some small precision value $\epsilon>0$, then, $P(|\hat{\theta}-\theta|\leq\epsilon)=P(\theta-\epsilon\leq\hat{\theta}\leq\theta+\epsilon)$ should increase in the sample size n and eventually reach 1.

An estimator is called **inconsistent** if it is not consistent. An inconsistent estimator is practically useless.

To show whether an estimator is consistent, we can check the sufficient condition for consistency:

Sufficient condition for consistency

The **mean squared error (MSE)** of an estimator $\hat{\theta}$ for some parameter θ is defined as

$$mse(\hat{\theta})=E[(\hat{\theta}-\theta)^2].$$

If $mse(\hat{\theta})$ tends to 0 as $n\to\infty$, then $\hat{\theta}$ is consistent for θ.

The sufficient condition is a direct application of Markov's inequality, which states that, for any random variable Z, any $\epsilon > 0$, and any $r \in \mathbb{N}$, we have

$$P(|Z| \geq \epsilon) \leq \frac{E[|Z|^r]}{\epsilon^r}.$$

Consequently, with $Z = \hat{\theta} - \theta$, and $r = 2$, we get

$$P(|\hat{\theta} - \theta| > \epsilon) \leq \frac{E[|\hat{\theta} - \theta|^2]}{\epsilon^2} = \frac{mse(\hat{\theta})}{\epsilon^2}.$$

Then, $mse(\hat{\theta}) \to 0$ is a sufficient condition for consistency since it implies $P(|\hat{\theta} - \theta| > \epsilon) \to 0$ due to the inequality in the formula above.

The mean square error of any estimator $\hat{\theta}$ can be decomposed into two additive terms,

$$mse(\hat{\theta}) = var[\hat{\theta}] + bias[\hat{\theta}]^2,$$

where

$$var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

is the **sampling variance** of $\hat{\theta}$, and

$$bias[\hat{\theta}] = E[\hat{\theta}] - \theta$$

is the **bias** of $\hat{\theta}$. The bias of an estimator measures how closely it approximates the true parameter on average, while the sampling variance quantifies how much the estimator's values typically fluctuate around that average. The MSE measures the precision of an estimator $\hat{\theta}$ for a given sample size n.

A converging MSE is a sufficient but not a necessary condition. It may be the case that an estimator is consistent with an MSE that does not converge, but these are some quite exceptional cases that we do not cover in this lecture (for instance, cases where the variance is infinite).

The MSE formula can be verified by some algebra:

$$\begin{aligned}
mse(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}])^2] + 2E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] + (E[\hat{\theta}] - \theta)^2 \\
&= var[\hat{\theta}] + 2(\underbrace{E[\hat{\theta}] - E[\hat{\theta}]}_{=0})(E[\hat{\theta}] - \theta) + bias[\hat{\theta}]^2 \\
&= var[\hat{\theta}] + bias[\hat{\theta}]^2
\end{aligned}$$

The bias and the sampling variance must tend to 0 as the sample size increases to have a consistent estimator. The estimator may have some bias for a fixed sample size n, but this

bias must tend to zero as n tends to infinity. We say that $\hat{\theta}$ is **asymptotically unbiased** if $bias[\hat{\theta}] \neq 0$ but $\lim_{n\to\infty} bias[\hat{\theta}] = 0$.

A particular class of estimators are **unbiased estimators**. An estimator $\hat{\theta}$ is unbiased if $bias[\hat{\theta}] = 0$ for any sample size n. In most cases, unbiased estimators should be preferred to biased/asymptotically unbiased estimators, but in some cases, we have a large variance for unbiased estimators and a small variance for asymptotically unbiased estimators. Balancing out the bias and sampling variance to obtain the smallest possible MSE is the so-called **bias-variance tradeoff**.

5.5 The sample mean

For any i.i.d. sample or stationary time series $\{Y_1, \dots, Y_n\}$ with $E[Y_i] = \mu$ and $Var[Y_i] = \sigma^2 < \infty$, the sample mean $\overline{Y}$ satisfies

$$E[\overline{Y}] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

It is an unbiased estimator for the mean μ since

$$bias[\overline{Y}] = E[\overline{Y}] - \mu = \mu - \mu = 0.$$

If the sample is i.i.d., the sampling variance is

$$Var[\overline{Y}] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[Y_i] = \frac{\sigma^2}{n}, \tag{5.2}$$

where the second equality follows by the fact that the variance of a sum of uncorrelated random variables is the sum of its variances. The variance declines in the sample size n, and the MSE converges to 0 as $n \to \infty$:

$$mse[\overline{Y}] = var[\overline{Y}] + bias[\overline{Y}]^2 = var[\overline{Y}] = \frac{\sigma^2}{n} \to 0.$$

Therefore, the sample mean is a consistent and unbiased estimator.

If the data is a stationary time series, the second equation in Equation 5.2 does not hold, and

we have

$$\begin{aligned} n \cdot Var[\overline{Y}] &= \frac{1}{n} Var\left[\sum_{t=1}^{n} Y_t\right] \\ &= \frac{1}{n}\sum_{t=1}^{n} Var[Y_t] + \frac{2}{n}\sum_{\tau=1}^{n-1}\sum_{t=\tau+1}^{n} Cov(Y_t, Y_{t-\tau}) \\ &= \gamma(0) + 2\sum_{\tau=1}^{n-1}\frac{n-\tau}{n}\gamma(\tau) \\ &\to \gamma(0) + 2\sum_{\tau=1}^{\infty}\gamma(\tau) =: \omega^2, \end{aligned}$$

where $\omega^2 < \infty$ if the time series is a short-memory time series. The parameter ω^2 is called **long-run variance**. Then, $n \cdot mse[\overline{Y}] \to \omega^2$, which implies that

$$mse[\overline{Y}] \to 0.$$

Hence, the sample mean is also an unbiased and consistent estimator for stationary short-memory time series.

The consistency of the sample mean to the population mean is also known as the **Law of Large Numbers (LLN)**.

Law of Large Numbers (LLN)

If $\{Y_1, \dots, Y_n\}$ is

(i) an i.i.d. sample from a population with $E[|Y_i|] < \infty$, or
(ii) a stationary time series with $\gamma(\tau) \to 0$ as $\tau \to \infty$,

then the sample mean is consistent for the population mean $\mu = E[Y_i]$, i.e.

$$\overline{Y} \xrightarrow{p} \mu \qquad (\text{as } n \to \infty)$$

Below is an interactive Shiny app to visualize the sample mean of simulated data for different sample sizes.

SHINY APP: LLN

The sample mean converges quickly to the population mean in the IID standard normal, Bernoulli, t(2), and stationary AR(1) case. The t(1) is a distribution with extremely heavy tails and infinite expectation. The LLN does not hold for this distribution. In the random walk case, the sample mean also does not converge. The random walk $Y_t = \sum_{j=1}^{t} u_j$ is a nonstationary time series. The sample mean is unbiased for a random walk but has a diverging sampling variance due to the strong persistence.

5.6 The sample variance

Consider an i.i.d. sample $\{Y_1, \dots, Y_n\}$ from some population distribution with mean $E[Y_i] = \mu$ and variance $Var[Y_i] = \sigma^2 < \infty$. The sample variance can be decomposed as

$$\begin{aligned}
\hat{\sigma}_Y^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu + \mu - \overline{Y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 + \frac{2}{n}\sum_{i=1}^{n}(Y_i - \mu)(\mu - \overline{Y}) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \overline{Y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - 2(\overline{Y} - \mu)^2 + (\overline{Y} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - (\overline{Y} - \mu)^2
\end{aligned}$$

The sampling mean of $\hat{\sigma}_Y^2$ is

$$\begin{aligned}
E[\hat{\sigma}_Y^2] &= \frac{1}{n}\sum_{i=1}^{n}E[(Y_i - \mu)^2] - E[(\overline{Y} - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n}Var[Y_i] - Var[\overline{Y}] \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2,
\end{aligned}$$

where we used the fact that $Var[\overline{Y}] = \sigma^2/n$ for i.i.d. data. The sample variance is **downward biased**:

$$bias[\hat{\sigma}_Y^2] = E[\hat{\sigma}_Y^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

By rescaling by the bias factor, we define the bias-corrected sample variance:

$$s_Y^2 = \frac{n}{n-1}\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

It is unbiased for σ^2:

$$bias[s_Y^2] = E[s_Y^2] - \sigma^2 = \frac{n}{n-1}E[\hat{\sigma}_Y^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

The bias correction is only valid for uncorrelated data. In the case of an autocorrelated stationary time series, s_Y^2 would still yield a bias. In any case, the sample variance is asymptotically unbiased. Suppose the underlying distribution is not heavy-tailed (i.e., fourth moments are bounded). In that case, the sampling variance also converges to zero so that the sample variance is consistent for the true variance for i.i.d. and stationary short-memory time series data.

5.7 Additional reading

- Stock and Watson (2019), Sections 2-3
- Hansen (2022a), Sections 6-7
- Hansen (2022b), Sections 3, 6
- Davidson and MacKinnon (2004), Section 1

5.8 R-codes

statistics-sec5.R

6 Confidence Intervals

6.1 Estimation uncertainty

An estimator approximates an unknown population parameter in the form of a single estimate, which is a real number. We call it a **point estimate**. The sample mean of the variable wage of the dataset of the previous section is 17.06. It is a point estimate for the population mean of the variable wage. It is unbiased and consistent, but it does not provide any information about how accurate the estimate is for a given sample size n.

A point estimate does not indicate the uncertainty inherent in the estimation process. Consistency results, such as the law of large numbers, state that as the sample size n approaches infinity, the estimate becomes increasingly accurate, i.e., the uncertainty of the estimate diminishes for large n. In practice, however, we are always faced with finite sample sizes and must understand the estimation uncertainty for a fixed sample size n.

We already learned that the MSE for the sample mean is

$$mse(\overline{Y}) = \frac{\sigma^2}{n},$$

where $0 < Var[Y] = \sigma^2 < \infty$. A quantity with better interpretability than the MSE is the square root of the MSE, similar to the variance and standard deviation. The **root mean squared error (RMSE)** of an estimator $\hat{\theta}$ for θ is

$$rmse(\hat{\theta}) = \sqrt{mse(\hat{\theta})} = \sqrt{E[(\hat{\theta} - \theta)^2]}.$$

The RMSE measures how much an estimate differs on average from its true parameter value for a given sample size n. The RMSE of the sample mean is

$$rmse(\overline{Y}) = \frac{\sigma}{\sqrt{n}}.$$

Since the RMSE is a linear function of $1/\sqrt{n}$, we say that the sample mean has the **rate of convergence** $\sqrt{n}$. We have $\lim_{n\to\infty} \sqrt{n} \cdot rmse(\hat{\theta}) = \sigma$.

Rate of convergence

An estimator $\hat{\theta}$ with $\lim_{n\to\infty} mse(\hat{\theta}) = 0$ has convergence rate $\sqrt{n}$ if

$$0 < \lim_{n\to\infty} \left(\sqrt{n} \cdot rmse(\hat{\theta}) \right) < \infty$$

More generally, the rate of convergence is $g(n)$ if

$$0 < \lim_{n\to\infty} \left(g(n) \cdot rmse(\hat{\theta}) \right) < \infty.$$

The rate $\sqrt{n}$ is the standard convergence rate for estimators and valid for most estimators we use in practice under mild conditions. If the rate of convergence is $\sqrt{n}$, we say that the estimator has a **parametric convergence rate**. There are exceptions where estimators have slower or faster convergence rates (nonparametric estimators, bootstrap, cointegration, long-memory time series).

The rate of convergence gives a first indication of how fast the uncertainty decreases as we get more observations. Consider the case of a $\sqrt{n}$ case as in the sample mean case. To halve the average deviation of the estimate from the true parameter value by a factor of 2, we need to increase the sample size by a factor of 4 since $\sqrt{4} = 2$. To halve the rmse by a factor of 4, we already need to increase the sample size by a factor of 16.

6.2 Interval estimates

The convergence rate indicates the relative estimation uncertainty, i.e., how much more accurate an estimate gets if we increase the sample size by a certain factor. However, it does not offer a way to quantify the uncertainty precisely.

One of the most common methods of incorporating estimation uncertainty into estimation results is through **interval estimates**, often referred to as **confidence intervals**. A confidence interval defines a range of values within which the true parameter is expected to fall, with a specified **coverage probability**, denoted as $1 - \alpha$.

More precisely, if θ is the parameter of interest and $\hat{\theta}$ is a point estimator for θ, a symmetric confidence interval I with coverage probability $1 - \alpha$ can be expressed as

$$I_{1-\alpha} = [\hat{\theta} - c_{1-\alpha};\ \hat{\theta} + c_{1-\alpha}]$$

with the property that

$$P(\theta \in I_{1-\alpha}) = 1 - \alpha. \tag{6.1}$$

Common coverage probabilities are 0.95, 0.99, and 0.90.

To derive the value $c_{1-\alpha}$ for a given sample size, we need to solve Equation 6.1 for $c_{1-\alpha}$. Note that the value $c_{1-\alpha}$ will depend on the distribution of $\hat{\theta}$. Also, note that $\hat{\theta}$ is a consistent estimator with a sampling variance that converges to 0. It is useful to consider the standardized estimator

$$Z = \frac{\hat{\theta} - E[\hat{\theta}]}{sd(\hat{\theta})},$$

which satisfies $E[Z] = 0$ and $Var[Z] = 1$ for any sample size n.

Let's reformulate Equation 6.1 with respect to Z_n. For simplicity, let's focus on the case where $\hat{\theta}$ is unbiased, i.e., $E[\hat{\theta}] = \theta$. The interval event can be rearranged as

$$\begin{aligned}
& \theta \in I_{1-\alpha} \\
\Leftrightarrow \quad & \hat{\theta} - c_{1-\alpha} \leq \theta \leq \hat{\theta} + c_{1-\alpha} \\
\Leftrightarrow \quad & -c_{1-\alpha} \leq \theta - \hat{\theta} \leq c_{1-\alpha} \\
\Leftrightarrow \quad & c_{1-\alpha} \geq \hat{\theta} - \theta \geq -c_{1-\alpha} \\
\Leftrightarrow \quad & \frac{c_{1-\alpha}}{sd(\hat{\theta})} \geq Z \geq -\frac{c_{1-\alpha}}{sd(\hat{\theta})}
\end{aligned}$$

Hence, Equation 6.1 becomes

$$P\left(\frac{-c_{1-\alpha}}{sd(\hat{\theta})} \leq Z \leq \frac{c_{1-\alpha}}{sd(\hat{\theta})}\right) = 1 - \alpha. \tag{6.2}$$

The next step would be to apply the CDF of Z to solve for $c_{1-\alpha}$. Suppose, for instance, $\hat{\theta}$ has a normal distribution. Then, Z is standard normal and has the CDF Φ and quantile function Φ^{-1}, which implies that the equation above becomes

$$\begin{aligned}
1 - \alpha &= \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\theta})}\right) - \Phi\left(\frac{-c_{1-\alpha}}{sd(\hat{\theta})}\right) \\
&= \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\theta})}\right) - \left(1 - \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\theta})}\right)\right) \\
&= 2\Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\theta})}\right) - 1,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
& \frac{2 - \alpha}{2} = \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\theta})}\right) \\
\Leftrightarrow \quad & \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \frac{c_{1-\alpha}}{sd(\hat{\theta})} \\
\Leftrightarrow \quad & z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\theta}) = c_{1-\alpha},
\end{aligned}$$

where $z_{(1-\frac{\alpha}{2})}$ is the $1 - \alpha/2$-quantile of $\mathcal{N}(0, 1)$. The confidence interval is

$$I_{1-\alpha} = [\hat{\theta} - z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\theta});\ \hat{\theta} + z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\theta})].$$

Standard normal quantiles can be obtained using the R command `qnorm` or by using statistical tables:

Table 6.1: Quantiles of the standard normal distribution

0.9	0.95	0.975	0.99	0.995
1.28	1.64	1.96	2.33	2.58

Therefore, 90%, 95%, and 99% confidence intervals for θ are given by

$$I_{0.9} = [\hat{\theta} - 1.64 \cdot sd(\hat{\theta});\ \hat{\theta} + 1.64 \cdot sd(\hat{\theta})]$$
$$I_{0.95} = [\hat{\theta} - 1.96 \cdot sd(\hat{\theta});\ \hat{\theta} + 1.96 \cdot sd(\hat{\theta})]$$
$$I_{0.99} = [\hat{\theta} - 2.58 \cdot sd(\hat{\theta});\ \hat{\theta} + 2.58 \cdot sd(\hat{\theta})]$$

In the case of the sample mean $\hat{\theta} = \overline{Y}$ as an estimator for the population mean $\theta = \mu$, we have $sd(\hat{\theta}) = \sigma/\sqrt{n}$ in the i.i.d. sampling case and $sd(\hat{\theta}) = \omega/\sqrt{n}$ in case of a stationary short-memory time series.

In any case, Equation 6.1 is satisfied, so the true parameter lies inside the confidence interval with probability $1 - \alpha$. With probability α the parameter is not in the interval. The more we want to be sure that the true parameter is in the interval, the smaller we have to choose α and the larger the interval becomes. If we choose $\alpha = 0$, the interval would be infinite, which does not help much. A certain amount of uncertainty always remains. We can control this by choosing the value for α.

Notice that we made two restrictive assumptions in the derivations above. First, we assumed that the estimator is unbiased. The assumption is, in fact, unproblematic if the estimator is asymptotically unbiased. Then, instead of Equation 6.1, the confidence interval is only asymptotically valid such that

$$\lim_{n \to \infty} P(\theta \in I_{1-\alpha}) = 1 - \alpha. \tag{6.3}$$

If Equation 6.3 is satisfied, we say that $I_{1-\alpha}$ is an **asymptotic confidence interval** for θ.

Our second restrictive assumption is that $\hat{\theta}$ follows a normal distribution for any given sample size n. As we will see in the next section, this assumption is less restrictive than one might think. Many estimators are asymptotically normal under general conditions (for instance, maximum likelihood estimators) so that the distribution of $\hat{\theta}$ comes closer and closer to a normal distribution as the sample size increases. Therefore Equation 6.3 is satisfied also under non-normality in many cases.

6.3 Central limit theorem

Consider the sample mean $\hat{\theta} = \overline{Y}$ as an estimator for the population mean $\theta = \mu$, which is unbiased with $E[\overline{Y}] = \mu$.

If the sample is i.i.d., $Var[\overline{Y}] = \sigma^2/n$. Under the additional assumption that the sample $\{Y_1, \dots, Y_n\}$ is $\mathcal{N}(\mu, \sigma^2)$ distributed, it follows that the sample mean is also normal since it is a linear combination of the sample values. Therefore,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big)$$

For this case, the $1-\alpha$ confidence interval is

$$I_{1-\alpha} = \Big[\overline{Y} - z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}};\ \overline{Y} + z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}\Big].$$

For stationary short memory time series the distribution is $\mathcal{N}(\mu, \omega^2/n)$, and we have

$$I_{1-\alpha} = \Big[\overline{Y} - z_{(1-\frac{\alpha}{2})} \cdot \frac{\omega}{\sqrt{n}};\ \overline{Y} + z_{(1-\frac{\alpha}{2})} \cdot \frac{\omega}{\sqrt{n}}\Big] \tag{6.4}$$

instead.

Fortunately, the central limit theorem tells us that we can drop the normality assumption and still obtain a valid asymptotic confidence interval in the sense of Equation 6.3.

Convergence in distribution

A statistic S_n **converges in distribution** to the random variable S if

$$\lim_{n\to\infty} P(S_n \le c) = P(S \le c)$$

for all $c \in \mathbb{R}$ at which the distribution function $F(c) = P(S \le c)$ is continuous. We write $S_n \xrightarrow{D} S$.

If S has the distribution $\mathcal{N}(\mu, \sigma^2)$, we write $S_n \xrightarrow{D} \mathcal{N}(\mu, \sigma^2)$.

To formulate the central limit theorem, consider the standardized sample mean in the i.i.d. case,

$$Z_n = \frac{\overline{Y} - E[\overline{Y}]}{sd(\overline{Y})} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma}$$

where σ can be replaced by ω in the stationary short-memory time series case.

Central Limit Theorem (CLT)

i) Let $\{Y_1, \dots, Y_n\}$ be an i.i.d. sample with $E[Y_i] = \mu$ and $0 < Var[Y_i] = \sigma^2 < \infty$. Then, the sample mean satisfies

$$\frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} \xrightarrow{D} \mathcal{N}(0, 1),$$

or, equivalently, $\sqrt{n}(\overline{Y} - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$.

ii) Let $\{Y_1, \ldots, Y_n\}$ be a stationary short-memory time series with with mean μ and long-run variance $0 < \omega^2 < \infty$. Moreover, let $E[Y_t^4] < \infty$ and let Y_t and $Y_{t-\tau}$ become independent as τ gets large. Then,

$$\frac{\sqrt{n}(\overline{Y} - \mu)}{\omega} \xrightarrow{D} \mathcal{N}(0, 1).$$

or, equivalently, $\sqrt{n}(\overline{Y} - \mu) \xrightarrow{D} \mathcal{N}(0, \omega^2)$.

Below, you will find an interactive shiny app for the central limit theorem:

SHINY APP: CLT

Note that for an asymptotic confidence interval (Equation 6.3), Equation 6.2 becomes

$$\lim_{n\to\infty} P\left(\frac{-c_{1-\alpha}}{sd(\hat{\theta})} \leq Z_n \leq \frac{c_{1-\alpha}}{sd(\hat{\theta})}\right) = 1 - \alpha,$$

and the CLT implies that $Z_n \xrightarrow{D} \mathcal{N}(0, 1)$. Therefore, all derivations from the previous subsection to obtain $c_{1-\alpha}$ are still valid, and

$$I_{1-\alpha} = \left[\overline{Y} - z_{(1-\frac{\alpha}{2})} \cdot sd(\overline{Y});\ \overline{Y} + z_{(1-\frac{\alpha}{2})} \cdot sd(\overline{Y})\right] \tag{6.5}$$

is an asymptotic confidence interval for μ.

The condition of "becoming independent as τ gets large" in the CLT is a weak dependence condition. Note that the random variables Y_t and $Y_{t-\tau}$ are independent if

$$P(Y_t \leq a, Y_{t-\tau} \leq b) - P(Y_t \leq a)P(Y_{t-\tau} \leq b) = 0 \tag{6.6}$$

for all a and b. Intuitively, weak dependence means that the left-hand side of Equation 6.6 might be nonzero but must converge to 0 as τ tends to infinity. I.e., the amount of dependence must decline for large τ.

6.4 Standard errors

The standard deviation of the sample mean and the confidence intervals depend on the unknown parameter σ^2 (i.i.d. case) or ω^2 (short-memory time series case).

To obtain feasible confidence intervals, we have to estimate the unknown parameters. I.e., we need an estimator for the standard deviation of the sample mean.

Standard error

A standard error $se(\hat{\theta})$ for an estimator $\hat{\theta}$ is an estimator for the estimators' standard deviation $sd(\hat{\theta}) = \sqrt{Var[\hat{\theta}]}$. The standard error is called consistent, if

$$\frac{se(\hat{\theta})}{sd(\hat{\theta})} \xrightarrow{p} 1.$$

For the i.i.d. case., we can replace σ^2 by the sample variance $\hat{\sigma}_Y^2$ or the bias-corrected sample variance s_Y^2. The classical standard error for the sample mean is

$$se(\overline{Y}) = \frac{s_Y}{\sqrt{n}}.$$

```
## Classical standard error
se = sd(wg)/sqrt(length(wg))
se
```

```
[1] 1.049272
```

```
t.test(wg)$stderr
```

```
[1] 1.049272
```

```
## 95% confidence interval
I = mean(wg) + c(-qnorm(0.975)*se, +qnorm(0.975)*se)
I
```

```
[1] 14.99907 19.11213
```

For short-memory stationary time series, classical standard errors are not valid. We need an estimator for the long-run variance $\omega^2 = \gamma(0) + 2\sum_{\tau=1}^{\infty} \gamma(\tau)$. A commonly applied estimator developed by Whitney K. Newey and Kenneth D. West in 1987 is

$$\hat{\omega}_{nw}^2 = \hat{\gamma}(0) + 2\sum_{\tau=1}^{\ell_n - 1} \frac{\ell_n - \tau}{\ell_n} \hat{\gamma}(\tau),$$

where $\hat{\gamma}(\tau)$ is the sample autocovariance function, and ℓ_n is a data-dependent truncation parameter. The corresponding autocorrelation robust standard error is

$$se_{nw}(\overline{Y}) = \frac{\hat{\omega}_{nw}}{\sqrt{n}}.$$

In R, you can get the robust standard error by the following command

```
library(sandwich)
## Robust standard error
seNW = sqrt(NeweyWest(lm(gdp~1)))
seNW
```

```
            (Intercept)
(Intercept)   0.4988509
```

```
## Robust confidence interval
mean(gdp) + c(-qnorm(0.975)*seNW, +qnorm(0.975)*seNW)
```

```
[1] 1.899616 3.855075
```

```
## Classical confidence intervals are too small:
se = sd(gdp)/sqrt(length(gdp))
mean(gdp) + c(-qnorm(0.975)*se, +qnorm(0.975)*se)
```

```
[1] 2.422138 3.332553
```

Confidence interval

Let $\hat{\theta}$ be a consistent estimator for the parameter θ with $mse[\hat{\theta}] \to 0$, and let the standardized estimator satisfy

$$\frac{\hat{\theta} - E[\hat{\theta}]}{sd(\hat{\theta})} \xrightarrow{D} \mathcal{N}(0,1).$$

Moreover, let $se(\hat{\theta})$ be a consistent standard error for $\hat{\theta}$. Then,

$$I_{1-\alpha} = \left[\hat{\theta} - z_{(1-\frac{\alpha}{2})} \cdot se(\hat{\theta});\ \hat{\theta} + z_{(1-\frac{\alpha}{2})} \cdot se(\hat{\theta})\right]$$

is an asymptotic $1-\alpha$ confidence interval for θ.

6.5 Exact confidence intervals under normality

Consider again the sample mean $\overline{Y}$ of an i.i.d. sample $\{Y_1, \dots, Y_n\}$ from some distribution with mean μ and variance $0 < \sigma^2 < \infty$.

The CLT implies that the sample mean is asymptotically normal, which theoretically justifies that Equation 6.5 is an asymptotic confidence interval for μ. In practice, Equation 6.5 is not feasible because σ and thus $sd(\overline{Y})$ are unknown.

We can use classical standard errors, replacing σ with s_Y, and still get an asymptotically valid confidence interval:

$$\lim_{n\to\infty} P\left(\mu \in \left[\overline{Y} - z_{(1-\frac{\alpha}{2})} \cdot \frac{s_Y}{\sqrt{n}};\ \overline{Y} + z_{(1-\frac{\alpha}{2})} \cdot \frac{s_Y}{\sqrt{n}}\right]\right) = 1 - \alpha. \tag{6.7}$$

The approximation is quite accurate for large samples, but for small samples the confidence interval may be imprecise. Therefore, it may be helpful to see if we can derive an exact confidence interval $I^*_{1-\alpha}$ such that $P(\mu \in I^*_{1-\alpha}) = 1 - \alpha$ for any small sample size n.

Under the restrictive assumption that the population distribution is normal, i.e., $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ for all $i = 1, \dots, n$, the sample mean is also normal since it is a weighted average of normal variables:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \tag{6.8}$$

In this case, the infeasible confidence interval in Equation 6.5 is indeed an exact confidence interval for μ. However, the feasible counterpart in Equation 6.7 is still only asymptotically valid, even if the underlying sample is normal.

Fortunately, the additional layer of uncertainty introduced by replacing σ with its estimator s_Y can be precisely quantified. Under the same conditions as in Equation 6.8, the bias-corrected sample variance as a χ^2-distribution and is independent of $\overline{Y}$:

$$\frac{(n-1)s_Y^2}{\sigma^2} \sim \chi^2_{n-1}$$

Consequently, standardizing with $s_Y/\sqrt{n}$ instead of $\sigma/\sqrt{n}$ yields a t-distributed statistic:

$$\frac{\overline{Y} - \mu}{s_Y/n} = \frac{\overline{Y} - \mu}{\sigma/n} \cdot \frac{\sigma}{s_Y} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1}.$$

Therefore, to obtain a feasible exact confidence interval for μ, we replace the standard normal quantile $z_{(1-\frac{\alpha}{2})}$ by the t-quantile $t_{(n-1;1-\frac{\alpha}{2})}$ with $n-1$ degrees of freedom:

$$P\left(\mu \in \left[\overline{Y} - t_{(n-1;1-\frac{\alpha}{2})} \cdot \frac{s_Y}{\sqrt{n}};\ \overline{Y} + t_{(n-1;1-\frac{\alpha}{2})} \cdot \frac{s_Y}{\sqrt{n}}\right]\right) = 1 - \alpha. \tag{6.9}$$

Table 6.2: Student's t-distribution quantiles

df	0.9	0.95	0.975	0.99	0.995
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60

df	0.9	0.95	0.975	0.99	0.995
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
8	1.40	1.86	2.31	2.90	3.36
10	1.37	1.81	2.23	2.76	3.17
15	1.34	1.75	2.13	2.60	2.95
20	1.33	1.72	2.09	2.53	2.85
25	1.32	1.71	2.06	2.49	2.79
30	1.31	1.70	2.04	2.46	2.75
40	1.30	1.68	2.02	2.42	2.70
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
80	1.29	1.66	1.99	2.37	2.64
100	1.29	1.66	1.98	2.36	2.63
$\to \infty$	1.28	1.64	1.96	2.33	2.58

Note that Equation 6.9 is only valid if $\{Y_1, \dots, Y_n\}$ is normally distributed. In any case, if $\{Y_1, \dots, Y_n\}$ is non-normal, we have

$$\lim_{n\to\infty} P\left(\mu \in \left[\overline{Y} - t_{(n-1;1-\frac{\alpha}{2})} \cdot \frac{s_Y}{n};\ \overline{Y} + t_{(n-1;1-\frac{\alpha}{2})} \cdot \frac{s_Y}{n}\right]\right) = 1 - \alpha. \tag{6.10}$$

Hence, Equation 6.7 and Equation 6.10 yield asymptotical confidence intervals for the mean for any distribution with finite variance, where only the latter is an exact confidence interval if the sample is normal.

In statistical software packages, Equation 6.10 is typically implemented. It is also a little more conservative than Equation 6.7 since $t_{(n-1;1-\frac{\alpha}{2})} > z_{(1-\frac{\alpha}{2})}$.

```
## confidence interval with normal quantiles
n=length(wg)
mean(wg) + c(-qnorm(0.975),+qnorm(0.975))*sd(wg)/sqrt(n)
```

```
[1] 14.99907 19.11213
```

```
## confidence interval with t-quantiles
mean(wg) + c(-qt(0.975,n-1),+qt(0.975,n-1))*sd(wg)/sqrt(n)
```

```
[1] 14.97362 19.13758
```

```r
## built-in confidence interval using the t-test function
t.test(wg, conf.level = 0.95)$conf.int
```

```
[1] 14.97362 19.13758
attr(,"conf.level")
[1] 0.95
```

Since the CDF of a t-distribution with $n-1$ degrees of freedom converges to the CDF of $\mathcal{N}(0,1)$, the confidence intervals in Equation 6.7 and Equation 6.10 are close to each other for large samples.

6.6 Additional reading

- Stock and Watson (2019), Sections 3
- Hansen (2022a), Section 14

6.7 R-codes

statistics-sec6.R

7 Hypothesis Testing

7.1 Statistical hypotheses

A statistical hypothesis is a statement about the population distribution. For instance, we might be interested in the hypothesis that the mean $\mu = E[Y]$ of a random variable Y is equal to some value μ_0 or whether it is unequal to that value.

In hypothesis testing, we divide the parameter space of interest into a null hypothesis and an alternative hypothesis, for instance

$$\underbrace{H_0 : \mu = \mu_0}_{\text{null hypothesis}} \quad \text{vs.} \quad \underbrace{H_1 : \mu \neq \mu_0,}_{\text{alternative hypothesis}} \tag{7.1}$$

or, more generally, $H_0 : \theta = \theta_0$. In practice, two-sided alternatives are more common, i.e. $H_1 : \theta \neq \theta_0$, but one-sided alternatives are also possible, i.e. $H_1 : \theta > \theta_0$ (right-sided) or $H_1 : \theta < \theta_0$ (left-sided).

We are interested in testing H_0 against H_1. The idea of hypothesis testing is to construct a statistic T_0 (**test statistic**) for which the sampling distribution under H_0 (**null distribution**) is known, and for which the distribution under H_1 differs from the null distribution (i.e., the null distribution is informative about H_1).

If the observed value of T_0 takes a value that is likely to occur under the null distribution, we deduce that there is no evidence against H_0, and consequently we do not reject H_0 (we accept H_0). If the observed value of T_0 takes a value that is unlikely to occur under the null distribution, we deduce that there is evidence against H_0, and consequently, we reject H_0 in favor of H_1. "Unlikely" means that its occurrence has only a small probability α. The value α is called the **significance level** and must be selected by the researcher. It is conventional to use the values $\alpha = 0.1$, $\alpha = 0.05$, or $\alpha = 0.01$, but it is not a hard rule.

A hypothesis test with significance level α is a decision rule defined by a rejection region I_1 and an acceptance region $I_0 = I_1^c$ so that we

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } T_0 \in I_0, \\ \text{reject } H_0 \quad & \text{if } T_0 \in I_1. \end{aligned}$$

The rejection region is defined such that a false rejection occurs with probability α, i.e.

$$P(\underbrace{T_0 \in I_1}_{\text{reject}} \mid H_0 \text{ is true}) = \alpha, \tag{7.2}$$

where $P(\cdot \mid H_0 \text{ is true})$ denotes the probability function of the null distribution.

A test that satisfies Equation 7.2 is called a **size-α-test**. The **type I error** is the probability of falsely rejecting H_0 and equals α for a size-α-test. The **type II error** is the probability of falsely accepting H_0 and depends on the sample size n and the unknown parameter value θ under H_1. Typically, the further θ is from θ_0, and the larger the sample size n, the smaller the type II error.

The probability of a type I error is also called the **size of a test**:

$$P(\text{reject } H_0 \mid H_0 \text{ is true}).$$

The **power of a test** is the complementary probability of a type II error:

$$P(\text{reject } H_0 \mid H_1 \text{ is true}) = 1 - P(\text{accept } H_0 \mid H_1 \text{ is true}).$$

A hypothesis test is **consistent for** H_1 if the power tends to 1 as n tends to infinity for any parameter value under the alternative.

Table 7.1: Testing Decisions

	Accept H_0	Reject H_0
H_0 is true	correct decision	type I error
H_1 is true	type II error	correct decision

In many cases, the probability distribution of T_0 under H_0 is known only asymptotically. Then, the rejection region must be defined such that

$$\lim_{n\to\infty} P(T_0 \in I_1 \mid H_0 \text{ is true}) = \alpha.$$

We call this test an asymptotic size-α-test.

The decision "accept H_0" does not mean that H_0 is true. Since the probability of a type II error is unknown in practice, it is more accurate to say that we "fail to reject H_0" instead of "accept H_0". The power of a consistent test tends to 1 as n increases, so type II errors typically occur if the sample size is too small. Therefore, to interpret a "fail to reject H_0", we have to consider whether our sample size is relatively small or rather large.

7.2 t-Test for the mean

For tests concerning the population mean, we can use the standardized sample mean, where we replace the sampling standard deviation with a standard error. The classical **t-statistic** or **t-ratio** for $H_0 : \mu = \mu_0$ is

$$T_0 := \frac{\overline{Y} - \mu_0}{se(\overline{Y})} = \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}}.$$

To construct a test of size α based on T_0, we first have to study the null distribution of T_0.

7.2.1 The normal i.i.d. case

Let's start with the restrictive case that the sample $\{Y_1, \dots, Y_n\}$ is i.i.d. and normally distributed with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, and recall from the previous section that, in this case,

$$\frac{\overline{Y} - \mu}{s_Y/\sqrt{n}} \sim t_{n-1} \tag{7.3}$$

for any fixed sample size n.

If the null hypothesis $H_0 : \mu = \mu_0$ is true, the t-ratio satisfies

$$T_0 := \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}} = \frac{\overline{Y} - \mu}{s_Y/\sqrt{n}} \sim t_{n-1},$$

i.e., the null distribution of T_0 is t_{n-1}.

Under the alternative hypothesis $H_1 : \mu \neq \mu_0$, we have

$$T_0 = \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}} = \underbrace{\frac{\overline{Y} - \mu}{s_Y/\sqrt{n}}}_{\sim t_{n-1}} + \underbrace{\frac{\mu - \mu_0}{s_Y/\sqrt{n}}}_{\neq 0} \nsim t_{n-1}.$$

The second term $(\mu - \mu_0)/(s_Y/\sqrt{n})$ diverges as n increases. It becomes to $+\infty$ or $-\infty$ depending on whether $\mu - \mu_0$ is positive or negative. The distribution under H_1 differs from the null distribution. The null distribution is therefore informative about H_1.

The idea is to compute T_0 for the given sample and check whether its value fits better to the null distribution or the alternative distribution of T_0. More precisely, we define a decision rule so that Equation 7.2 is satisfied. Since the null distribution has mean zero and some finite variance and the alternative distribution diverges to $+\infty$ or $-\infty$, it makes sense to accept H_0 if T_0 is close to zero and reject if it takes large positive or negative values.

We define the acceptance and rejection regions I_0 and I_1 by fixing a **critical value** c with the following decision rule:

$$\begin{aligned} \text{do not reject } H_0 \text{ if } |T_0| &\leq c, \\ \text{reject } H_0 \text{ if } |T_0| &> c. \end{aligned}$$

Let $F_0(a) = P(T_0 \leq a \mid H_0 \text{ is true})$ be the CDF of the null distribution, i.e., F_0 is the CDF of t_{n-1}. To determine a critical value c that satisfies Equation 7.2, first note that c must be positive (otherwise, we would always reject). Moreover,

$$\begin{aligned} P(\text{reject } H_0 \mid H_0 \text{ is true}) &= P(|T_0| > c \mid H_0 \text{ is true}) \\ &= P(\{T_0 < -c\} \cup \{T_0 > c\} \mid H_0 \text{ is true}) \\ &= P(T_0 < -c \mid H_0 \text{ is true}) + P(T_0 > c \mid H_0 \text{ is true}) \\ &= F_0(-c) + (1 - F_0(c)) \\ &= 2(1 - F_0(c)), \end{aligned}$$

where the last step follows from $F_0(-c) = 1 - F_0(c)$ due to the fact that t_{n-1} is a symmetric distribution. Finally, setting equal to α and solving for c yields

$$\begin{aligned} & & 2(1 - F_0(c)) &= \alpha \\ \Leftrightarrow & & F_0(c) &= 1 - \frac{\alpha}{2} \\ \Leftrightarrow & & c &= F_0^{-1}(1 - \tfrac{\alpha}{2}) \\ \Leftrightarrow & & c &= t_{(n-1;1-\frac{\alpha}{2})}, \end{aligned}$$

which is the $1 - \alpha/2$ quantile of the t-distribution with $n - 1$ degrees of freedom.

Hence, we reject H_0 if $|T_0|$ exceeds $t_{(n-1;1-\frac{\alpha}{2})}$. Equivalently, H_0 is rejected if μ_0 is not element of the $(1 - \alpha)$ confidence interval Equation 6.9.

7.2.2 The non-normal i.i.d. case

If $\{Y_1, \ldots, Y_n\}$ is i.i.d. and non-normal, the CLT implies that Equation 7.3 is replaced by the asymptotic result

$$\frac{\overline{Y} - \mu}{s_Y/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$$

as $n \to \infty$. Therefore, an asymptotic size-α-test for the hypotheses in Equation 7.1 is defined by the following decision rule:

$$\begin{aligned} \text{do not reject } H_0 \quad &\text{if } |T_0| \leq z_{(1-\frac{\alpha}{2})}, \\ \text{reject } H_0 \quad &\text{if } |T_0| > z_{(1-\frac{\alpha}{2})}. \end{aligned}$$

Alternatively, since $t_{n-1} \xrightarrow{D} \mathcal{N}(0,1)$ and therefore $\lim_{n\to\infty} t_{(n-1;1-\frac{\alpha}{2})} = z_{(1-\frac{\alpha}{2})}$, we may also use t-quantiles and apply the decision rule

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } |T_0| \leq t_{(n-1;1-\frac{\alpha}{2})}, \\ \text{reject } H_0 \quad & \text{if } |T_0| > t_{(n-1;1-\frac{\alpha}{2})}. \end{aligned}$$

Let's test the hypothesis that $H_0 : E(wage) = 16$ against $H_1 : E(wage) \neq 16$.

```
tratio = (mean(wg)-16)/(sd(wg)/sqrt(length(wg)))
crit = qt(0.975, length(wg)-1)
c(tratio, crit)
```

```
[1] 1.006031 1.984217
```

The test statistic $T_0 = 1.01$ is in absolute value smaller than the critical value $c = 1.98$ at the 5% significance level. Therefore, we do not reject H_0. Alternatively, we can use the `t.test(wg, mu=16)` function (see below). If we use the normal quantile, we have the critical value $c = 1.96$, which yields the same test decision.

7.2.3 The time series case

If $\{Y_1, \ldots, Y_n\}$ is a stationary short-memory time series, classical standard errors are not valid since

$$\frac{\overline{Y} - \mu}{s_Y/\sqrt{n}} \xrightarrow{D} \mathcal{N}\Big(0, \frac{\omega^2}{\sigma^2}\Big).$$

We may use Newey-West standard errors $se_{nw}(\overline{Y})$, which yield

$$\frac{\overline{Y} - \mu}{\widehat{\omega}_{nw}/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

The autocorrelation robust t-ratio for $H_0 : \mu = \mu_0$ is given by

$$T_{0,AC} = \frac{\overline{Y} - \mu_0}{\widehat{\omega}_{nw}/\sqrt{n}},$$

and an asymptotic test of size α is given by the decision rule

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } |T_{0,AC}| \leq z_{(1-\frac{\alpha}{2})}, \\ \text{reject } H_0 \quad & \text{if } |T_{0,AC}| > z_{(1-\frac{\alpha}{2})}. \end{aligned}$$

Alternatively, t-quantiles may be used.

Let's test the hypothesis that $H_0 : E(gdpgrowth) = 4$ against $H_1 : E(gdpgrowth) \neq 4$.

```r
library(sandwich)
seNW = sqrt(NeweyWest(lm(gdp~1)))
tratio = (mean(gdp)-4)/seNW
crit = qnorm(0.975)
c(tratio, crit)
```

```
[1] -2.250481  1.959964
```

We reject H_0 since $|T_{0,AC}| = 2.25 > 1.96$. There is no statistical evidence that the mean GDP growth is 4.

7.3 The p-value

The **p-value** is a criterion to reach the t-test decision conveniently:

$$\begin{aligned} \text{reject } H_0 \quad & \text{if p-value} < \alpha \\ \text{do not reject } H_0 \quad & \text{if p-value} \geq \alpha \end{aligned}$$

Formally, the p-value of a two-sided t-test is defined as

$$p\text{-value} = P(|\widetilde{T}| > |T_0| \mid H_0 \text{ is true}),$$

where $\widetilde{T}$ is a random variable with the null distribution, i.e. $\widetilde{T} \sim t_{n-1}$. The p-value is the probability that a null-distributed random variable produces values at least as extreme as the test statistic T_0 produced for your sample. We can express the p-value also using the CDF F_0 of the null distribution:

$$\begin{aligned} p\text{-value} &= P(|\widetilde{T}| > |T_0| \mid H_0 \text{ is true}) \\ &= 1 - P(|\widetilde{T}| \leq |T_0| \mid H_0 \text{ is true}) \\ &= 1 - F_0(|T_0|) + F_0(-|T_0|) \\ &= 2(1 - F_0(|T_0|)). \end{aligned}$$

Make no mistake, the p-value is not the probability that H_0 is true! It is a measure of how likely it is that the observed test statistic comes from a sample that has been drawn from the null distribution.

To compute the p-value by hand, we have to insert the t-statistic into the CDF of the null distribution:

```r
tratio = (mean(wg)-16)/(sd(wg)/sqrt(length(wg)))
df = length(wg)-1
2*(1-pt(abs(tratio), df)) ## p-value
```

```
[1] 0.3168532
```

The `t.test()` function provides a summary of all relevant statistics for inference on the mean: the sample mean, a 95%-confidence interval, the t-statistic for a specified null hypothesis, and the corresponding p-value:

```r
t.test(wg, mu=16)
```

```

    One Sample t-test

data:  wg
t = 1.006, df = 99, p-value = 0.3169
alternative hypothesis: true mean is not equal to 16
95 percent confidence interval:
 14.97362 19.13758
sample estimates:
mean of x
  17.0556
```

7.4 Power function

Consider an i.i.d. sample $\{Y_1, \dots, Y_n\}$ from a distribution with mean μ and variance σ^2. The power function of the two-sided t-test for $H_0 : \mu = \mu_0$ is defined as

$$\pi(\mu, \sigma^2, n) = P(\text{reject } H_0) = P(|T_0| > c),$$

where $c = z_{(1-\frac{\alpha}{2})}$ (or a corresponding t-quantile).

For $\mu = \mu_0$ (i.e., H_0 is true), $\pi(\mu_0, \sigma^2, n)$ corresponds to the size and the probability of a type I error. For $\mu \neq \mu_0$ (i.e., H_1 is true), we have $\pi(\mu, \sigma^2, n) = (1 - P(\text{type II error}))$.

Consider for simplicity the test for a zero mean, i.e., $\mu_0 = 0$, and suppose the true mean is different, i.e., $\mu \neq 0$. The test statistic is

$$T_0 = \frac{\overline{Y} - 0}{se(\overline{Y})} = \frac{\overline{Y}}{s_Y/\sqrt{n}},$$

and we have

$$T_0 - \frac{\sqrt{n}\mu}{s_Y} = \frac{\overline{Y} - \mu}{s_Y/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1),$$

which implies that

$$P(T_0 \leq c) = P\left(\frac{\overline{Y} - \mu}{s_Y/\sqrt{n}} \leq c - \frac{\sqrt{n}\mu}{s_Y}\right) \approx \Phi(c - \sqrt{n}\mu/\sigma).$$

In the last step, we also used that s_Y^2 is consistent for σ^2. The power function then becomes

$$\begin{aligned}\pi(\mu, \sigma^2, n) &= P(|T_0| > c) \\ &= 1 - P(-c \leq T_0 \leq c) \\ &1 - P(T_0 \leq c) + P(T_0 \leq -c) \\ &\approx 1 - \Phi(c - \sqrt{n}\mu/\sigma) + \Phi(-c - \sqrt{n}\mu/\sigma).\end{aligned}$$

Below you find a plot of the power curve for different values of $\sqrt{n}\mu/\sigma$ and significance levels:

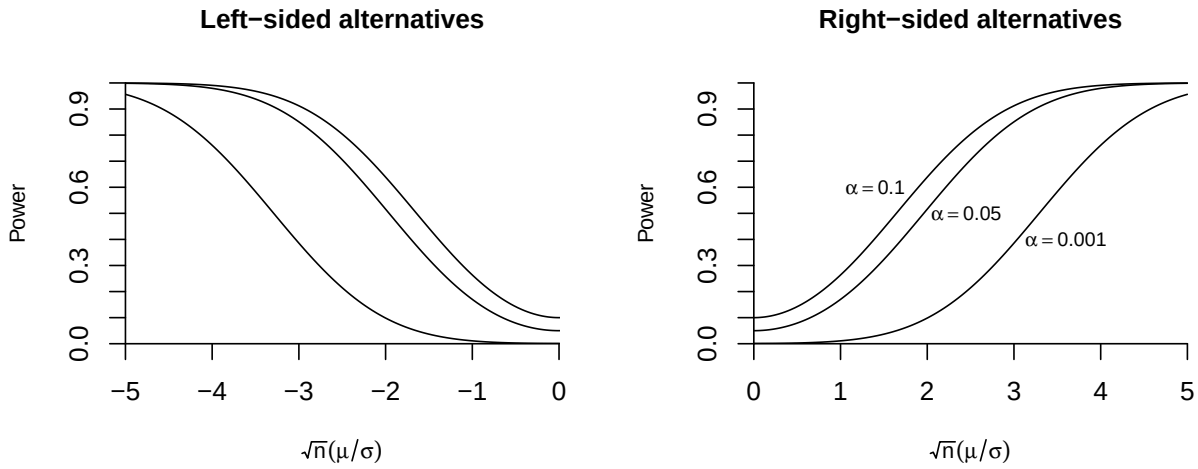


Figure 7.1: Power curves two-sided t-tests for the null hypothesis of a zero mean

Since $(\pm c - \sqrt{n}\mu/\sigma) \to \infty$ as $n \to \infty$, we have, for $\mu > 0$,

$$1 - \underbrace{\Phi(c - \sqrt{n}\mu/\sigma)}_{\to 0} + \underbrace{\Phi(-c - \sqrt{n}\mu/\sigma)}_{\to 0} \to 1.$$

Similarly, for alternatives $\mu < 0$, we have

$$1 - \underbrace{\Phi(c - \sqrt{n}\mu/\sigma)}_{\to 1} + \underbrace{\Phi(-c - \sqrt{n}\mu/\sigma)}_{\to 1} \to 1.$$

Hence, the test is consistent for any alternative $\mu \neq 0$.

7.5 One-sided t-test

The two-sided t-test is a powerful test for any alternative $\mu \neq \mu_0$, but it is not the uniformly most powerful test. For a right-sided alternative $\mu > 0$, the right-sided t-test has a higher power. The following decision rule defines it:

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } T_0 \leq z_{(1-\alpha)}, \\ \text{reject } H_0 \quad & \text{if } T_0 > z_{(1-\alpha)}. \end{aligned}$$

The power function is

$$\pi(\mu, \sigma^2, n) = P(T_0 > z_{(1-\alpha)}) \approx 1 - \Phi(z_{(1-\alpha)} - \sqrt{n}\mu/\sigma),$$

which is larger than the power function of the two-sided test for right-sided alternatives. For left-sided alternatives, the right-sided test has no power. The size is $\pi(0, \sigma^2, n) \approx 1 - \Phi(z_{(1-\alpha)}) = \alpha$.

Similarly, the left-sided t-test is defined by the decision rule

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } T_0 \geq -z_{(1-\alpha)}, \\ \text{reject } H_0 \quad & \text{if } T_0 < -z_{(1-\alpha)}. \end{aligned}$$

Its power function is

$$\pi(\mu, \sigma^2, n) = P(T_0 < -z_{(1-\alpha)}) \approx \Phi(-z_{(1-\alpha)} - \sqrt{n}\mu/\sigma),$$

where the size is $\pi(\mu, \sigma^2, n) \approx \Phi(-z_{(1-\alpha)}) = \alpha$. It has a higher power for left-sided alternatives than the two-sided test but has no power for right-sided alternatives.

It can be shown that, for the fixed simple alternative $\mu > \mu_0$, the right-sided t-test is the test with the highest possible power, and for an alternative of the form $\mu < \mu_0$, the left-sided t-test has the greatest power. It is a result of the Neyman-Pearson lemma, which states that no test exists for a simple hypothesis with greater power than a likelihood-ratio test. It can also be shown that the one-sided t-test is equivalent to the likelihood ratio test for the alternative of interest.

Note that the direction of testing must be specified in advance, which is not always practical. The two-sided t-test has slightly less power but is more common in practice because it has power against both right- and left-sided alternatives.

Below, you will find an interactive shiny app for the right-sided z-test. A z-test is a t-test with known variance, where s_Y is replaced by σ in the test statistic (i.e., the two tests are asymptotically equivalent).

SHINY APP: TTEST

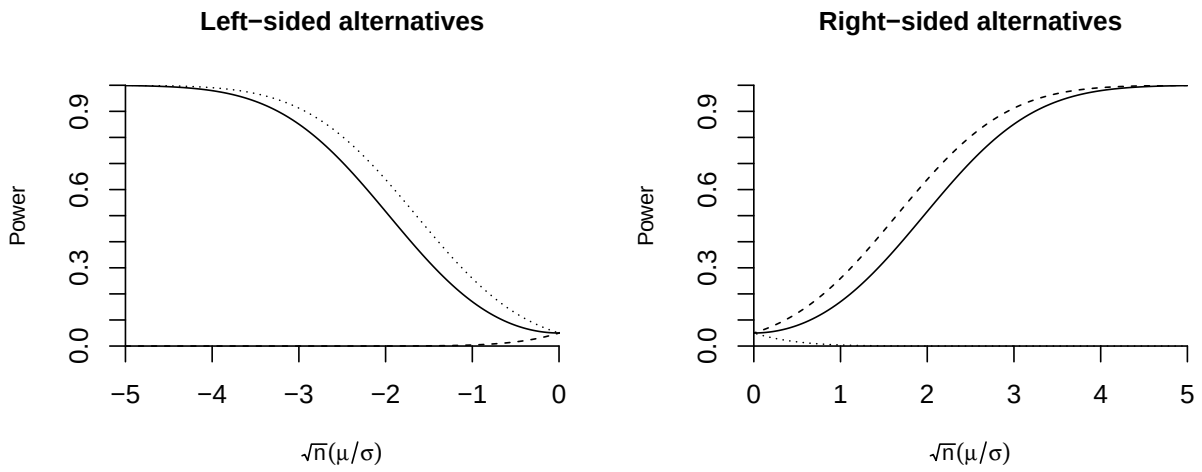


Figure 7.2: Power of two-sided (solid), right-sided (dashed), and left-sided (dotted) t-tests (5% significance level) for the null hypothesis of a zero mean

7.6 Testing for autocorrelation

Consider a stationary time series $\{Y_1, \dots, Y_n\}$ with order τ autocorrelation function $\rho(\tau)$ and finite fourth moments ($E[Y_i^4] < \infty$). We have the following limit theorem for the sample autocorrelation function under $H_0 : \rho(\tau) = 0$:

$$\sqrt{n}(\hat{\rho}(\tau) - \rho(\tau)) \xrightarrow{D} \mathcal{N}(0, 1).$$

To test the null hypothesis $H_0 : \rho(\tau) = 0$ for some fixed τ, consider the test statistic

$$T_0 = \frac{\hat{\rho}(\tau)}{1/\sqrt{n}},$$

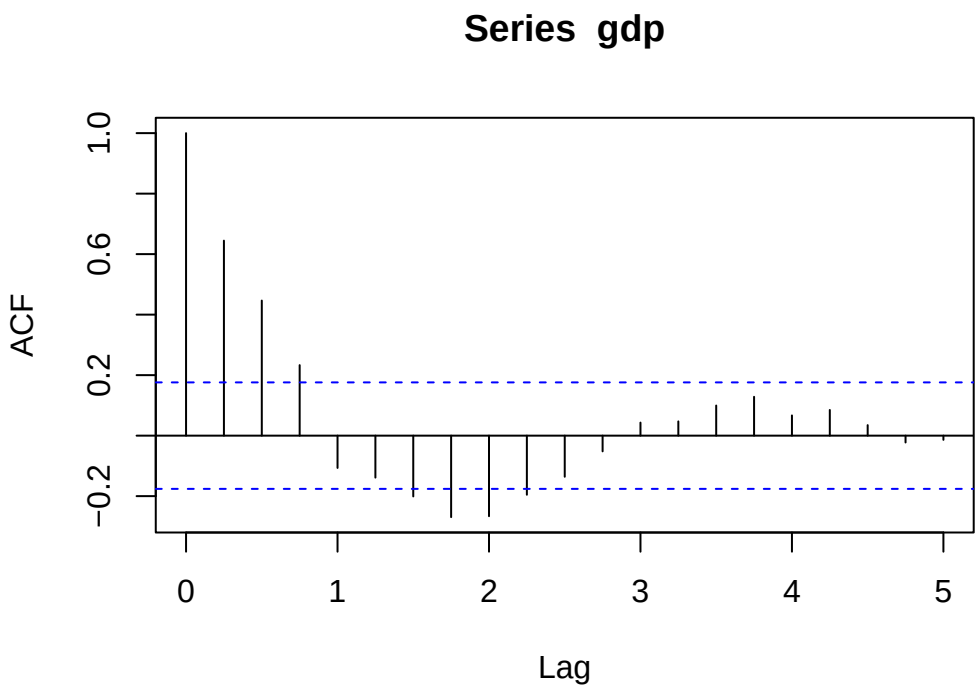
which is standard normal in the limit under H_0. Therefore, a simple test for zero order τ autocorrelation is given by the decision rule

$$\begin{aligned} \text{do not reject } H_0 \quad &\text{if } |T_0| \leq z_{(1-\frac{\alpha}{2})}, \\ \text{reject } H_0 \quad &\text{if } |T_0| > z_{(1-\frac{\alpha}{2})}. \end{aligned}$$

Equivalently, we reject H_0 if $\hat{\rho}(\tau)$ lies outside $\pm z_{(1-\frac{\alpha}{2})}/\sqrt{n}$.

The interval $[-z_{(1-\frac{\alpha}{2})}/\sqrt{n};\ z_{(1-\frac{\alpha}{2})}/\sqrt{n}]$ for $\alpha = 0.05$ is indicated in the ACF plot in R by blue dashed lines. Hence, GDP growth has significant autocorrelation at lag 3 but not at lag 4:

```
acf(gdp)
```



7.7 Additional reading

- Stock and Watson (2019), Sections 3
- Hansen (2022a), Section 13

7.8 R-codes

statistics-sec7.R

8 Simulations

8.1 The Monte Carlo principle

Monte Carlo simulations are useful for studying properties of the sampling distribution of a statistic in a particular controlled environment, where we know the true distribution from which the data are sampled. The statistic of interest could be an estimator, a confidence interval, or a test statistic. Some properties of the statistic that we might be interested in are listed below:

- bias of an estimator
- variance of an estimator
- MSE of an estimator
- moments of the distribution of an estimator
- quantiles of the distribution of an estimator
- coverage rates of a confidence interval
- size of a hypothesis test
- power function of a hypothesis test

Monte Carlo simulations indicate whether an estimator is consistent, a confidence interval has the correct coverage, or a hypothesis test has the right size (although they do not replace formal proof). We can also use Monte Carlo simulations to compare the biases and MSEs of different estimators or the power curves of various tests.

The idea is that we use computer-generated pseudorandom numbers to create an artificial data set of sample size n to which we apply the statistic of interest. This procedure generates a random draw from the sampling distribution of the statistic.

By repeating the procedure B times independently, we obtain an i.i.d. sample of length B from the sampling distribution of our statistic of interest, which we call a **Monte Carlo sample**. From the Monte Carlo sample, we can compute empirical counterparts of the features of interest (e.g., bias, MSE, coverage, power, etc.).

8.2 Set up

To set up the Monte Carlo simulation, we need to specify

(i) a sample size n;
(ii) a specific parametric distribution F from which we sample our data;
(iii) the sampling scheme (i.i.d. or time series process)
(iv) the number of Monte Carlo repetitions B.
(v) the Monte Carlo statistic $\hat{\theta}$.

For example, if we are interested in the MSE of the sample mean of 100 i.i.d. coin flips, we set $n = 100$, F the Bernoulli distribution with probability 0.5, an i.i.d. sampling scheme, a large number of repetitions (e.g., $B = 10000$), and the sample mean as a Monte Carlo statistic.

If we are interested in the size of a t-test applied to an AR(1) process with parameter 0.8, length 200, and standard normal increments, we set $n = 200$, F the standard normal distribution, the AR(1) process sampling scheme, a large number of replications (e.g., $B = 10000$), and a Monte Carlo statistic $\hat{\theta}$ containing both the sample mean and the bias-corrected sample variance, since both are needed to compute the t-statistic.

8.3 Monte Carlo algorithm

The Monte Carlo simulation is performed as follows:

1. Using the specified sampling scheme, draw a sample $\{X_1^*, \dots, X_n^*\}$ of size n from F using the computer's random number generator.
2. Evaluate the statistic $\hat{\theta}$ from $\{X_1^*, \dots, X_n^*\}$.
3. Repeat steps 1 and 2 of the experiment B times and collect the estimates from step 2 in the Monte Carlo sample $\hat{\theta}_{mc} = \{\hat{\theta}_1, \dots, \hat{\theta}_B\}$.
4. Evaluate the features of interest from the Monte Carlo sample.

8.4 Evaluate an estimator

If `MCsample` is the Monte Carlo sample $\hat{\theta}_{mc}$ of an estimator $\hat{\theta}$ for some parameter θ (`theta`), we can evaluate

- the MC-estimated sampling mean `mean(MCsample)`:

$$\hat{\mu}_{mc} = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*$$

- the MC-estimated bias `bias = mean(MCsample) - theta`:

$$\widehat{bias}[\hat{\theta}_{mc}] = \hat{\mu}_{mc} - \theta$$

- the MC-estimated sampling variance `var(MCsample)`:

$$\widehat{var}[\hat{\theta}_{mc}] = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\mu}_{mc})^2$$

- the MC-estimated MSE `mse = var(MCsample) + bias^2`:

$$\widehat{mse}[\hat{\theta}_{mc}] = \widehat{var}[\hat{\theta}_{mc}] + \widehat{bias}[\hat{\theta}_{mc}]^2$$

- The r-th MC-estimated moment: `mean(MCsample^r)`
- The MC-estimated p-quantile: `quantile(MCsample, p)`

We may evaluate whether the estimator is consistent by checking whether the MSE tends to 0 as n gets larger. Let's try it for the sample mean of an i.i.d. standard normal distributed sample of different sample sizes:

```
#MC sample mean
B = 1000 #Monte Carlo replications

getMCsample = function(n){
  X = rnorm(n) #standard normal sample
  mean(X) #sample mean (statistic of interest)
}
# n=10 (small sample size)
MCsample1 = sapply(rep(10,B), getMCsample)
mse1 = var(MCsample1)+(mean(MCsample1)-0)^2
mse1
```

```
[1] 0.1045475
```

```
# n=100 (moderate sample size)
MCsample2 = sapply(rep(100,B), getMCsample)
mse2 = var(MCsample2)+(mean(MCsample2)-0)^2
mse2
```

```
[1] 0.009382021
```

```
# n=1000 (large sample size)
MCsample3 = sapply(rep(1000,B), getMCsample)
mse3 = var(MCsample3)+(mean(MCsample3)-0)^2
mse3
```

```
[1] 0.000980639
```

The simulated MSE tends to 0 as n increases. We already derived the theoretical MSE, which is σ^2/n. With $\sigma^2 = 1$, the simulated numbers fit the theoretical ones.

Another example would be checking whether classical standard errors are consistent. That is, $se(\overline{Y})/sd(\overline{Y}) \xrightarrow{p} 1$ as $n \to \infty$. We have $sd(\overline{Y}) = \sigma^2/\sqrt{n}$ and $se(\overline{Y}) = s_Y/\sqrt{n}$.

```
#MC standard errors
B = 1000 #Monte Carlo replications

getMCsample = function(n){
  X = rnorm(n) #standard normal sample
  se = sd(X)/sqrt(n) #classical standard errors
  sdt = 1/sqrt(n) #true sampling standard deviation
  se/sdt # return ratio
}
# n=10 (small sample size)
MCsample1 = sapply(rep(10,B), getMCsample)
mse1 = var(MCsample1)+(mean(MCsample1)-1)^2
mse1
```

```
[1] 0.05600309
```

```
# n=100 (moderate sample size)
MCsample2 = sapply(rep(100,B), getMCsample)
mse2 = var(MCsample2)+(mean(MCsample2)-1)^2
mse2
```

```
[1] 0.005137395
```

```
# n=1000 (large sample size)
MCsample3 = sapply(rep(1000,B), getMCsample)
mse3 = var(MCsample3)+(mean(MCsample3)-1)^2
mse3
```

```
[1] 0.0004965606
```

The MC-estimated MSE of $se(\overline{Y})/sd(\overline{Y})$ tends to 0, which indicates that the standard error is consistent if the data is sampled from a standard normal distribution.

8.5 Evaluate a confidence interval

An asymptotic $1-\alpha$-confidence interval must have a coverage rate that tends to $1-\alpha$ as the sample size increases.

Let's check how well a confidence interval with normal quantiles performs when t-quantiles would yield exact confidence intervals. Suppose the data is standard normal with zero mean. The interval has correct coverage if the logical statements `lowerbound < 0` and `0 < upperbound` are both true. The command `(lowerbound < 0) && (0 < upperbound)` returns `TRUE` if both statements are true and `FALSE` otherwise.

```
#MC standard errors
B = 1000 #Monte Carlo replications

getMCsample = function(n){
  X = rnorm(n) #standard normal sample
  ## confidence interval bounds:
  lowerbound =  mean(X) - qnorm(0.975)*sd(X)/sqrt(n)
  upperbound = mean(X) + qnorm(0.975)*sd(X)/sqrt(n)
  ## check if the true mean 0 is inside the bounds
  ## The command returns TRUE or FALSE
  (lowerbound < 0) && (0 < upperbound)
}
# n=10 (small sample size)
MCsample1 = sapply(rep(10,B), getMCsample)
sum(MCsample1)/B ## relative frequency of TRUEs
```

```
[1] 0.923
```

```
# n=100 (moderate sample size)
MCsample2 = sapply(rep(100,B), getMCsample)
sum(MCsample2)/B
```

```
[1] 0.949
```

```
# n=1000 (large sample size)
MCsample3 = sapply(rep(1000,B), getMCsample)
sum(MCsample3)/B
```

```
[1] 0.95
```

8.6 Evaluate a hypothesis test

A hypothesis test must have asymptotic size α and a power that tends to 1 as the sample size increases.

Let's evaluate whether the t-test for the mean with $\alpha = 0.05$ works as expected in finite samples if the underlying data has an exponential distribution with parameter λ. We consider the two-sided t-test for $H_0 : \mu = 1$. The mean of an exponentially distributed random variable is $1/\lambda$, so $\mu = 1$ if $\lambda = 1$. If $\lambda = 0.8$ we have $\mu = 1.25$.

We consider different sample sizes and evaluate the MC-estimated size for the test under H_0 (i.e., $\lambda = 1$) and the MC-estimated power of the test for $\lambda = 0.8$, which is a setting under H_1.

The MC-estimated size is the relative frequency of rejections under H_0, and the MC-estimated power is the relative frequency of rejections under H_1.

```
#MC standard errors
B = 1000 #Monte Carlo replications

getMCsample = function(n, lambda){
  X = rexp(n, lambda) # i.i.d. exponential sample
  ## p-value of t-test for mean = 1:
  pval = t.test(X, mu=1)$p.value
  ## check if p-value is below 0.05 (return TRUE or FALSE):
  pval < 0.05
}
# n=10 (small sample size)
MCsample1.size = sapply(rep(10,B), getMCsample, lambda=1)
MCsample1.power = sapply(rep(10,B), getMCsample, lambda=0.8)
size = sum(MCsample1.size)/B ## relative frequency of TRUEs
power = sum(MCsample1.power)/B ## relative frequency of TRUEs
c(size,power)
```

```
[1] 0.099 0.067
```

```
# n=100 (moderate sample size)
MCsample1.size = sapply(rep(100,B), getMCsample, lambda=1)
MCsample1.power = sapply(rep(100,B), getMCsample, lambda=0.8)
size = sum(MCsample1.size)/B ## relative frequency of TRUEs
power = sum(MCsample1.power)/B ## relative frequency of TRUEs
c(size,power)
```

```
[1] 0.052 0.480
```

```r
# # n=1000 (large sample size)
MCsample1.size = sapply(rep(300,B), getMCsample, lambda=1)
MCsample1.power = sapply(rep(300,B), getMCsample, lambda=0.8)
size = sum(MCsample1.size)/B ## relative frequency of TRUEs
power = sum(MCsample1.power)/B ## relative frequency of TRUEs
c(size,power)
```

```
[1] 0.050 0.961
```

The size for $n = 10$ is slightly too large (9.9%) but correct for $n = 100$ and $n = 300$ (5%). For the alternative $\mu = 1.25$, the test has almost no power for $n = 10$, a power of 48% for $n = 100$, and for $n = 300$, a power of 96%.

8.7 Additional reading

- Hansen (2022b), Section 9.18

8.8 R-codes

statistics-sec8.R

9 Least Squares

Regression analysis is concerned with the approximation of a dependent variable Y_i (regressand, response variable) by a function $f(X_i)$ of a vector of independent variables X_i (regressors, predictor variables):

$$Y_i \approx f(X_i), \quad i = 1, \dots, n.$$

The least squares method selects the regression function that minimizes the squared approximation error:

$$\min_f \sum_{i=1}^{n} (Y_i - f(X_i))^2. \tag{9.1}$$

In linear regression, the regression function $f(\cdot)$ is a linear function, and the minimization problem of Equation 9.1 is called **ordinary least squares (OLS)** problem.

9.1 The OLS principle

We approximate the dependent variable Y_i by a linear function of k-1 independent variables $X_{i2}, \dots, X_{ik}$ plus an intercept:

$$Y_i \approx b_1 + b_2 X_{i2} + \dots + b_k X_{ik}.$$

The regression function can be written as an inner product:

$$f(X_i) = b_1 + b_2 X_{i2} + \dots + b_k X_{ik} = X_i' b,$$

with

$$X_i = \begin{pmatrix} 1 \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix},$$

where X_i is the vector of k regressors and b is the vector of k regression coefficients.

For a given sample $\{(Y_1, X_1'), \dots, (Y_n, X_n')\}$ and a given coefficient vector $b \in \mathbb{R}^K$ the **sum of squared errors** is

$$S_n(b) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 = \sum_{i=1}^{n} (Y_i - X_i' b)^2.$$

The **least squares coefficient vector** $\hat{\beta}$ is the minimizing argument:

$$\hat{\beta} = \text{argmin}_{b \in \mathbb{R}^k} \sum_{i=1}^{n} (Y_i - X_i' b)^2 \tag{9.2}$$

Least squares coefficients

If the $k \times k$ matrix $(\sum_{i=1}^{n} X_i X_i')$ is invertible, the solution to Equation 9.2 is unique with

$$\hat{\beta} = \Big(\sum_{i=1}^{n} X_i X_i' \Big)^{-1} \sum_{i=1}^{n} X_i Y_i. \tag{9.3}$$

Proof. Write the sum of squared errors as

$$S_n(b) = \sum_{i=1}^{n} (Y_i - X_i' b)^2 = \sum_{i=1}^{n} Y_i^2 - 2b' \sum_{i=1}^{n} X_i Y_i + b' \sum_{i=1}^{n} X_i X_i' b.$$

The first-order condition for a minimum in $b = \hat{\beta}$ is that the gradient is zero, and the second-order condition is that the Hessian matrix is positive definite (see Matrix calculus).

The first-order condition

$$\frac{\partial S_n(b)}{\partial b} = -2 \sum_{i=1}^{n} X_i Y_i + 2 \sum_{i=1}^{n} X_i X_i' b \stackrel{!}{=} \mathbf{0}_k$$

implies

$$\sum_{i=1}^{n} X_i X_i' \hat{\beta} = \sum_{i=1}^{n} X_i Y_i \quad \Leftrightarrow \quad \hat{\beta} = \Big(\sum_{i=1}^{n} X_i X_i' \Big)^{-1} \sum_{i=1}^{n} X_i Y_i.$$

The second-order condition

$$\frac{\partial^2}{\partial b \partial b'} S_n(b) = 2 \sum_{i=1}^{n} X_i X_i' \stackrel{!}{>} 0$$

is satisfied for any b since

$$c' \Big(\sum_{i=1}^{n} X_i X_i' \Big) c = \sum_{i=1}^{n} c' X_i X_i' c = \sum_{i=1}^{n} \underbrace{(c' X_i)^2}_{\geq 0} > 0$$

for any nonzero vector $c \in \mathbb{R}^k$ (see definite matrix). The strict positivity follows from the condition that $\sum_{i=1}^{n} X_i X_i'$ is invertible, i.e., it is nonsingular and does not have any zero eigenvalues (see here). □

□

OLS fitted values and residuals

The **fitted values** or predicted values are

$$\widehat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik} = X_i' \hat{\beta}, \quad i = 1, ..., n.$$

The **residuals** are

$$\hat{u}_i = Y_i - \widehat{Y}_i = Y_i - X_i' \hat{\beta}, \quad i = 1, ..., n.$$

9.2 Simple linear regression (k=2)

A simple linear regression is a linear regression of a dependent variable on a constant and a single independent variable.

Let's revisit the wage and education data from Table 9.1 and consider only the first 20 observations:

We regress $Y_i = \log(wage_i)$ on a constant and $Z_i = education_i$. The regressor vector is $X_i = (1, Z_i)'$.

The OLS coefficients are

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} &= \Big(\sum_{i=1}^n X_i X_i'\Big)^{-1} \sum_{i=1}^n X_i Y_i \\ &= \begin{pmatrix} n & \sum_{i=1}^n Z_i \\ \sum_{i=1}^n Z_i & \sum_{i=1}^n Z_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n Z_i Y_i \end{pmatrix}\end{aligned}$$

Evaluate sums:

$$\sum_{i=1}^n X_i Y_i = \begin{pmatrix} 50.87 \\ 809.85 \end{pmatrix}, \quad \sum_{i=1}^n X_i X_i' = \begin{pmatrix} 20 & 312 \\ 312 & 5044 \end{pmatrix}$$

OLS coefficients:

$$\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix} 20 & 312 \\ 312 & 5044 \end{pmatrix}^{-1} \begin{pmatrix} 50.87 \\ 809.85 \end{pmatrix} = \begin{pmatrix} 1.107 \\ 0.092 \end{pmatrix}$$

In R, we can use the `lm()` function to compute the least squares coefficients:

```
fit1 = lm(logwage~education)
fit1
```

Table 9.1: ALLBUS data sub-sample

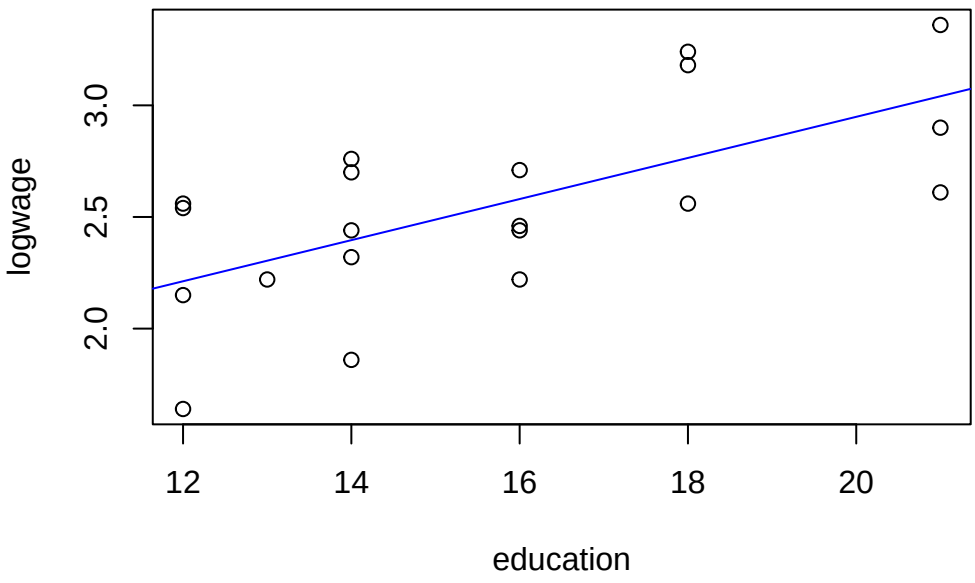
Person	Wage	log(Wage)	Education	Education^2	Edu x log(wage)
1	12.91	2.56	18	324	46.08
2	11.49	2.44	14	196	34.16
3	10.22	2.32	14	196	32.48
4	11.49	2.44	16	256	39.04
5	9.20	2.22	16	256	35.52
6	14.94	2.70	14	196	37.80
7	11.75	2.46	16	256	39.36
8	15.09	2.71	16	256	43.36
9	23.95	3.18	18	324	57.24
10	8.62	2.15	12	144	25.80
11	25.54	3.24	18	324	58.32
12	15.84	2.76	14	196	38.64
13	5.17	1.64	12	144	19.68
14	28.74	3.36	21	441	70.56
15	6.44	1.86	14	196	26.04
16	12.92	2.56	12	144	30.72
17	9.20	2.22	13	169	28.86
18	13.65	2.61	21	441	54.81
19	12.64	2.54	12	144	30.48
20	18.11	2.90	21	441	60.90
sum	**277.91**	**50.87**	**312**	**5044**	**809.85**

```
Call:
lm(formula = logwage ~ education)

Coefficients:
(Intercept)    education
    1.10721      0.09207
```

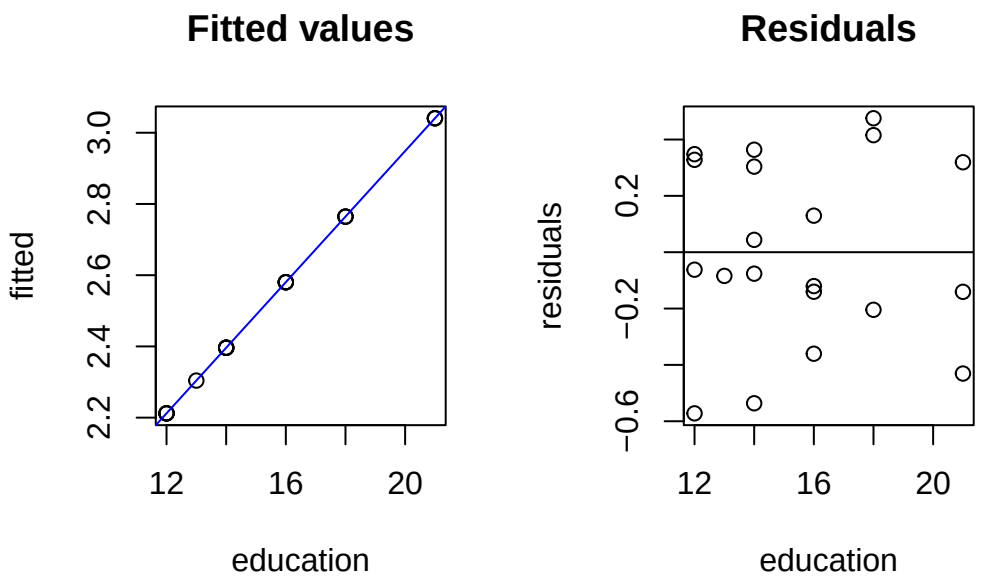
The OLS coefficient vector is $\hat{\beta} = (1.107, 0.092)'$ and the fitted regression line is

$$f(education) = 1.107 + 0.092\ education$$

```
## Plot regression line
plot(education,logwage)
abline(fit1, col="blue")
```




```
par(mfrow = c(1,2))
## Fitted values and residuals
fitted = fit1$fitted.values
residuals = fit1$residuals
plot(education,fitted, main="Fitted values")
abline(fit1, col="blue")
plot(education,residuals, main="Residuals")
abline(h=0)
```



There is another, simpler formula for $\hat{\beta}_1$ and $\hat{\beta}_2$ in the simple linear regression model. It can be expressed in terms of sample moments:

Simple linear regression

The least squares coefficients in a simple linear regression can be written as

$$\hat{\beta}_2 = \frac{\hat{\sigma}_{YZ}}{\hat{\sigma}_Z^2} = \frac{s_{YZ}}{s_Z^2}, \quad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{Z} \tag{9.4}$$

The results coincide with those from above:

```
beta2.hat = cov(logwage,education)/var(education)
beta1.hat = mean(logwage) - beta2.hat*mean(education)
c(beta1.hat, beta2.hat)
```

```
[1] 1.10720588 0.09207014
```

Here is an illustration of fitted values and residuals for an OLS regression line of another dataset:

9.3 Linear regression with k=3

The following code computes $\hat{\beta}$ for a given dataset with $k = 3$ regressors $X_i = (1, X_{i2}, X_{i3})'$:

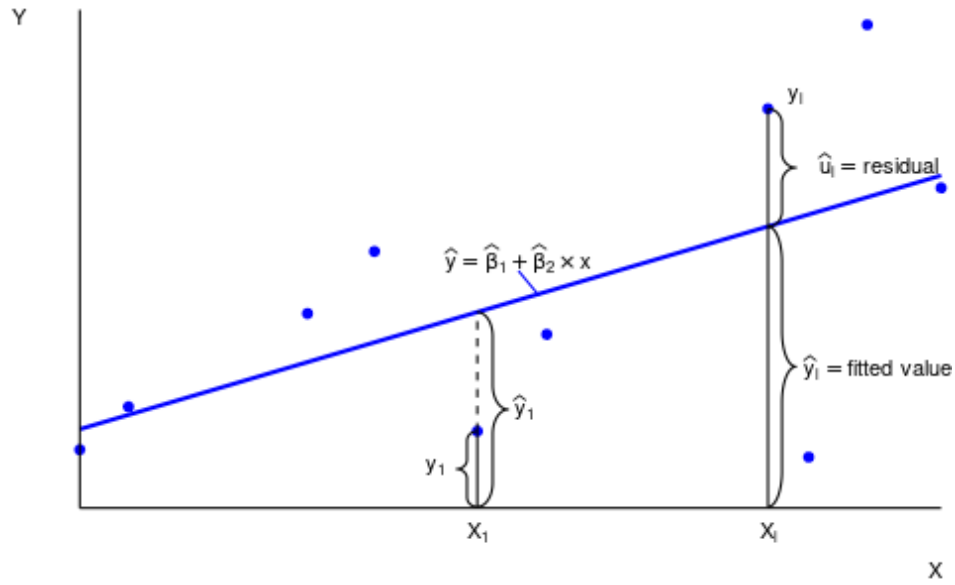
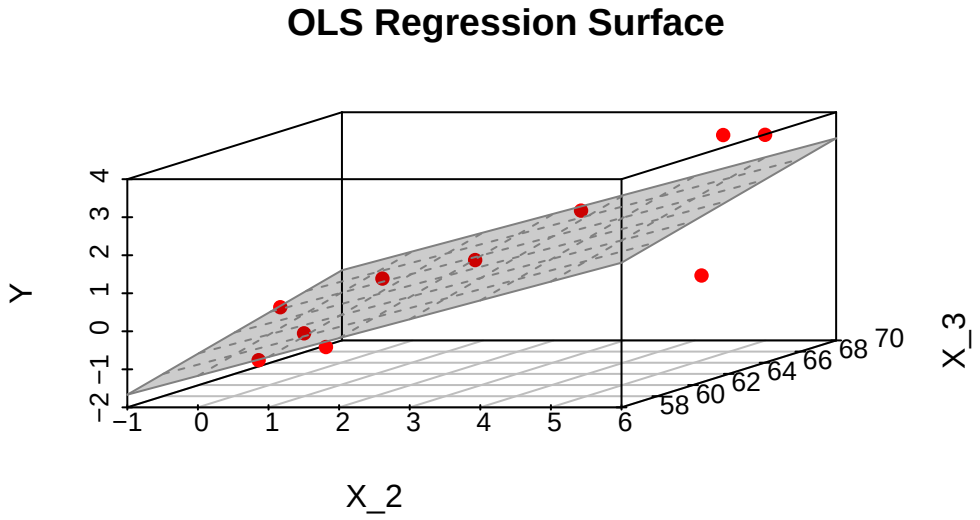


Figure 9.1: Illustration of fitted values and residuals

```
# Some given data
Y = c(0.7,-1.0,-0.2,-1.2,-0.1,3.4,0.0,0.8,3.7,2.0)
X_2 = c(1.9,0.8,1.25,0.1,-0.1,4.4,4.6,1.6,5.5,3.4)
X_3 = c(66, 62, 59, 61, 63, 70, 68, 62, 68, 66)
## Compute the OLS estimation
fit2 = lm(Y ~ X_2 + X_3)
## Plot sample regression surface
library("scatterplot3d") # library for 3d plots
plot3d  <- scatterplot3d(x = X_2, y = X_3, z = Y,
            angle = 33, scale.y = 0.8, pch = 16,
            color ="red",
            xlab = "X_2",
            ylab = "X_3",
            main ="OLS Regression Surface")
plot3d$plane3d(fit2, lty.box = "solid", col=gray(.5), draw_polygon=TRUE)
```

```
fit2$coefficients
```

```
(Intercept)         X_2         X_3
 -8.4780709   0.4955995   0.1259828
```



The fitted regression surface is

$$f(X_i) = -8.478 + 0.496X_{i2} + 0.126X_{i3}.$$

9.4 Matrix notation

To avoid the summation notation in Equation 9.3, we stack the n observations of the dependent variable into a vector and the n regressor vectors as $1 \times k$ row vectors into a $n \times k$ matrix:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{12} & \dots & X_{1k} \\ \vdots & & & \vdots \\ 1 & X_{n2} & \dots & X_{nk} \end{pmatrix}$$

We have $\sum_{i=1}^n X_i X_i' = \boldsymbol{X}'\boldsymbol{X}$ and $\sum_{i=1}^n X_i Y_i = \boldsymbol{X}'\boldsymbol{Y}$.

The least squares coefficients become

$$\hat{\beta} = \Big(\sum_{i=1}^n X_i X_i'\Big)^{-1} \sum_{i=1}^n X_i Y_i = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

By using matrix notation, the OLS coefficient formula can be quickly implemented in R (recall that `%*%` is matrix multiplication, `solve()` computes the inverse, and `t()` the transpose):

```
X = cbind(rep(1,10), X_2, X_3)
solve(t(X) %*% X) %*% t(X) %*% Y
```

```
          [,1]
    -8.4780709
X_2  0.4955995
X_3  0.1259828
```

The vector of fitted values is

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} = \boldsymbol{X}\hat{\beta} = \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{=\boldsymbol{P}}\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}.$$

The vector of residuals is

$$\hat{\boldsymbol{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \underbrace{(\boldsymbol{I}_n - \boldsymbol{P})}_{=\boldsymbol{M}}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{Y}$$

9.5 Projection matrices

The matrix

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

is the $n \times n$ **projection matrix** that projects any vector from $\mathbb{R}^n$ into the column space spanned by the column vectors of $\boldsymbol{X}$. It is also called **influence matrix** or **hat-matrix** because it maps the vector of response values $\boldsymbol{Y}$ on the vector of fitted values (hat values): $\boldsymbol{P}\boldsymbol{Y} = \widehat{\boldsymbol{Y}}$.

Its diagonal entries

$$h_{ii} = X_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}X_i$$

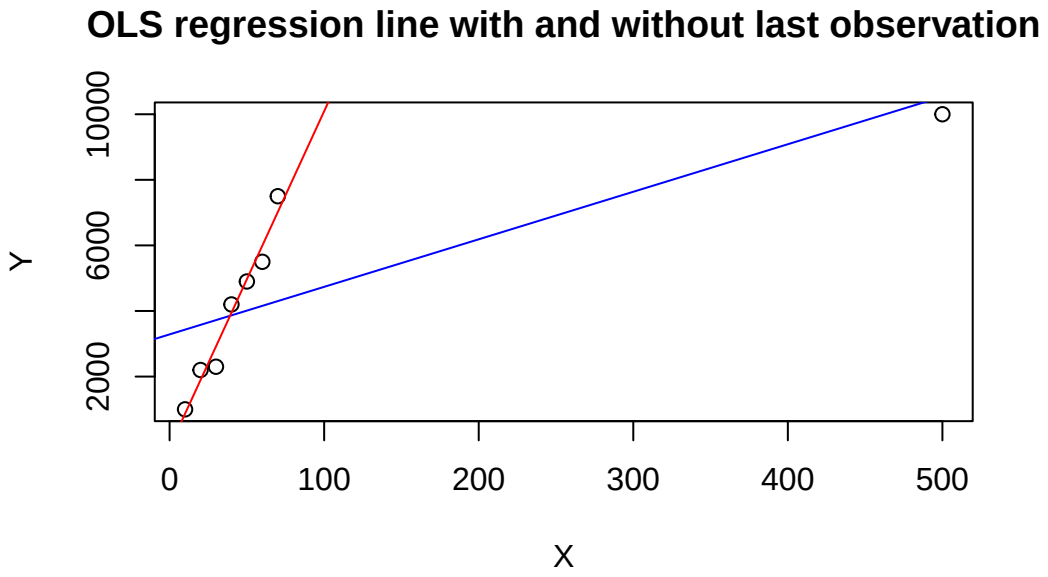
are called **leverage values**, which measure how far away the regressor values of the i-th observation X_i are from those of the other observations.

Properties of leverage values:

$$0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{n} h_{ii} = k.$$

A large h_{ii} occurs when the observation i has a big influence on the regression line, e.g., the last observation in the following dataset:

```r
X=c(10,20,30,40,50,60,70,500)
Y=c(1000,2200,2300,4200,4900,5500,7500,10000)
plot(X,Y, main="OLS regression line with and without last observation")
abline(lm(Y~X), col="blue")
abline(lm(Y[1:7]~X[1:7]), col="red")
```

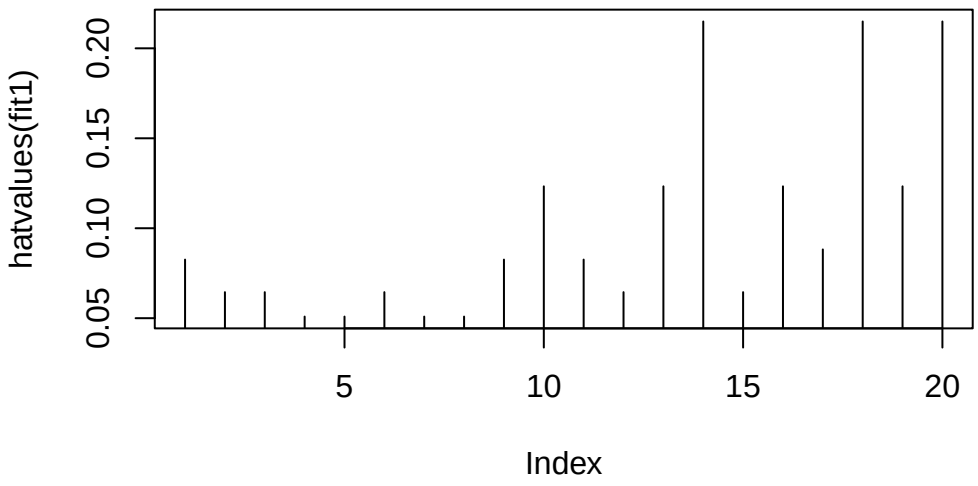



```
hatvalues(lm(Y~X))
```

```
        1         2         3         4         5         6         7         8
0.1657356 0.1569566 0.1492418 0.1425911 0.1370045 0.1324820 0.1290237 0.9869646
```

The wage and education data is quite balanced and has moderate leverage values:

```
plot(hatvalues(fit1), type="h")
```



```
hatvalues(fit1)
```

```
         1          2          3          4          5          6          7
0.08257919 0.06447964 0.06447964 0.05090498 0.05090498 0.06447964 0.05090498
         8          9         10         11         12         13         14
0.05090498 0.08257919 0.12330317 0.08257919 0.06447964 0.12330317 0.21493213
        15         16         17         18         19         20
0.06447964 0.12330317 0.08823529 0.21493213 0.12330317 0.21493213
```

The matrix

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

is the $n \times n$ **orthogonal projection matrix** that projects any vector from $\mathbb{R}^n$ into the vector space orthogonal to that spanned by $\boldsymbol{X}$. It is also called **annihilator matrix** or **residual maker matrix** because it maps the vector of response values $\boldsymbol{Y}$ on the vector of residuals: $\boldsymbol{M}\boldsymbol{Y} = \hat{\boldsymbol{u}}$.

The projection matrices $\boldsymbol{P}$ and $\boldsymbol{M}$ have some nice properties:

(a) $\boldsymbol{P}$ and $\boldsymbol{M}$ are **symmetric**, i.e. $\boldsymbol{P} = \boldsymbol{P}'$ and $\boldsymbol{M} = \boldsymbol{M}'$.
(b) $\boldsymbol{P}$ and $\boldsymbol{M}$ are **idempotent**, i.e. $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$ and $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.
(c) Moreover, we have that $\boldsymbol{X}'\boldsymbol{P} = \boldsymbol{X}'$, $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{M} = \boldsymbol{0}_{k\times n}$, $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}_{n\times k}$, and $\boldsymbol{P}\boldsymbol{M} = \boldsymbol{0}_{n\times n}$.

The properties (a)-(c) follow directly from the definitions of $\boldsymbol{P}$ and $\boldsymbol{M}$ (try it as an exercise). Using these properties one can show that the residual vector $\hat{\boldsymbol{u}}$ is orthogonal to each of the column vectors in $\boldsymbol{X}$:

$$\boldsymbol{X}'\hat{\boldsymbol{u}} = \underbrace{\boldsymbol{X}'\boldsymbol{M}}_{=\boldsymbol{0}_{k\times n}}\boldsymbol{Y} = \boldsymbol{0}_{k\times 1}. \tag{9.5}$$

Equation 9.5 implies that the residual vector $\hat{\boldsymbol{u}}$ is orthogonal to the predicted values vector $\widehat{\boldsymbol{Y}}$ since

$$\boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{0}_{k\times 1} \quad \Rightarrow \quad \underbrace{\hat{\beta}'\boldsymbol{X}'}_{=\widehat{\boldsymbol{Y}}'}\hat{\boldsymbol{u}} = \underbrace{\hat{\beta}'\boldsymbol{0}_{k\times 1}}_{=0}.$$

Hence, OLS produces a decomposition of the dependent variable values into the orthogonal vectors of fitted values and residuals,

$$\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \hat{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{Y} + \boldsymbol{M}\boldsymbol{Y},$$

with

$$\widehat{\boldsymbol{Y}}'\hat{\boldsymbol{u}} = 0. \tag{9.6}$$

9.6 Analysis of Variance

The first and second sample moments of $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ can be expressed in terms of the sample moments of the fitted values $\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \dots, \widehat{Y}_n)'$ and the residuals $\hat{\boldsymbol{u}} = (\hat{u}_1, \dots, \hat{u}_n)'$.

The first sample moment is

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\Bigg(\sum_{i=1}^{n} \widehat{Y}_i + \underbrace{\sum_{i=1}^{n} \hat{u}_i}_{=0}\Bigg) = \overline{\widehat{Y}},$$

which follows from Equation 9.5 and the fact that the linear regression function contains an intercept. The reason is that, due to the intercept, the first column of $\boldsymbol{X}$ consists of 1's, and the first entry of $\boldsymbol{X}'\hat{\boldsymbol{u}}$ is $\sum_{i=1}^{n} \hat{u}_i$.

The second sample moment is

$$\overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i^2 + \frac{2}{n}\underbrace{\sum_{i=1}^{n} \widehat{Y}_i \hat{u}_i}_{=\widehat{\boldsymbol{Y}}'\hat{\boldsymbol{u}}=0} + \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 = \overline{\widehat{Y}^2} + \overline{\hat{u}^2},$$

which follows from Equation 9.6.

The sample variance is

$$\hat{\sigma}_Y^2 = \overline{Y^2} - \overline{Y}^2 = \overline{\widehat{Y}^2} + \overline{\hat{u}^2} - \overline{\widehat{Y}} = \hat{\sigma}_{\widehat{Y}}^2 + \hat{\sigma}_{\hat{u}}^2,$$

where $\hat{\sigma}_{\hat{u}}^2 = \overline{\hat{u}^2} - \overline{\hat{u}}^2 = \overline{\hat{u}^2}$ since $\sum_{i=1}^{n} \hat{u}_i = 0$ due to the intercept in the regression function.

We obtain the well-known variance decomposition result for OLS regressions:

Analysis-of-variance formula

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}_{\text{total sample variance}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2}_{\text{explained sample variance}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}_{\text{unexplained sample variance}}.$$

Alternatively, we can write $\hat{\sigma}_Y^2 = \hat{\sigma}_{\widehat{Y}}^2 + \hat{\sigma}_{\hat{u}}^2$ or $s_Y^2 = s_{\widehat{Y}}^2 + s_{\hat{u}}^2$.

9.7 Coefficients of determination

The larger the proportion of the explained sample variance, the better the fit of the OLS regression. This motivates the definition of the so-called R^2 coefficient of determination:

Coefficient of determination

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{\widehat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

R^2 describes the proportion of sample variation in $\boldsymbol{Y}$ explained by $\widehat{\boldsymbol{Y}}$. Obviously, we have that $0 \leq R^2 \leq 1$.

If $R^2 = 0$, the sample variation in $\widehat{\boldsymbol{Y}}$ is zero (e.g. if the regression line/surface is horizontally flat). The closer R^2 lies to 1, the better the fit of the OLS regression to the data. If $R^2 = 1$, the variation in $\hat{\boldsymbol{u}}$ is zero, so the OLS regression explains the entire variation in $\boldsymbol{Y}$ (perfect fit).

A low R^2 does not necessarily mean the regression specification is bad. It just means that there is a high share of unobserved heterogeneity in $\boldsymbol{Y}$ that is not captured by the regressors $\boldsymbol{X}$. A high R^2 does not necessarily mean a good regression specification. It just means that the regression fits the sample well. Too many unnecessary regressors lead to overfitting. If $k = n$, we have $R^2 = 1$ even if none of the regressors has an actual influence on the dependent variable.

This brings us to the most commonly criticized drawback of R^2: Additional regressors (relevant or not) always increase the R^2. Here is an example of this problem.

```
set.seed(123)
n     <- 100                    # Sample size
X     <- runif(n, 0, 10)        # Relevant X variable
X_ir  <- runif(n, 5, 20)        # Irrelevant X variable
extra <- rnorm(n, 0, 10)        # Unobserved additional variable
Y     <- 1 + 5 * X + extra      # Y variable
fit1  <- lm(Y~X)       # Correct OLS regression
fit2   <- lm(Y~X+X_ir)  # OLS regression with X_ir
c(summary(fit1)$r.squared, summary(fit2)$r.squared)
```

```
[1] 0.6933407 0.6933606
```

So, R^2 increases here even though `X_ir` is a completely irrelevant explanatory variable.

```
## Simulate n-2 additional irrelevant regressors:
X_ir2  <- matrix(runif(n*(n-2), 5, 20),nrow=n)
## use 50 irrelevant regressors:
fit3   <- lm(Y~X+X_ir2[,1:50])
## use all 98 irrelevant regressors:
fit4   <- lm(Y~X+X_ir2)
c(summary(fit3)$r.squared, summary(fit4)$r.squared)
```

```
[1] 0.8353157 1.0000000
```

If we add 50 irrelevant regressors, the R^2 increases even further. If we add 98 irrelevant regressors ($k = 100 = n$), we obtain a perfect fit with $R^2 = 1$.

Possible solutions are given by penalized criteria such as the so-called **adjusted R-squared** or R-bar-squared:

Adjusted R-squared

$$\overline{R}^2 = 1 - \frac{\frac{1}{n-k}\sum_{i=1}^n \widehat{u}_i^2}{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

The adjustment is in terms of the degrees of freedom. We lose k degrees of freedom in the OLS regression since we have k regressors (k linear restrictions). We lose one degree of freedom in computing the sample variance due to the sample mean (one linear restriction).

The squareroot of the adjusted sample variance in the numerator of the adjusted R-squared formula is called the **standard error of the regression (SER)**:

$$SER := \sqrt{\frac{1}{n-k}\sum_{i=1}^n \widehat{u}_i^2} = \sqrt{\frac{n}{n-k}\widehat{\sigma}_{\widehat{u}}^2}.$$

An alternative for comparing different regression specifications is the **leave-one-out R-squared** or R-tilde-squared, which is based on the leave-one-out cross-validation (LOOCV) principle. The i-th leave-one-out OLS coefficient is the OLS coefficient using the sample without the i-th observation:

$$\widehat{\beta}_{(-i)} = \Big(\sum_{j\neq i} X_j X_j'\Big)^{-1} \sum_{j\neq i} X_j Y_j = (\boldsymbol{X}'\boldsymbol{X} - X_i X_i')^{-1}(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i)$$

The i-th leave-one-out predicted value $\widetilde{Y}_i = X_i'\widehat{\beta}_{(-i)}$ is an authentic prediction for Y_i since observation Y_i is not used to construct $\widetilde{Y}_i$. The leave-one-out residual is $\tilde{u}_i = Y_i - \widetilde{Y}_i$ and obeys the relation

$$\tilde{u}_i = \frac{\widehat{u}_i}{1 - h_{ii}},$$

where $h_{ii} = X_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}X_i$ is the i-th leverage value.

The leave-one-out R-squared is the R-squared using leave-one-out residuals:

Leave-one-out R-squared

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^n \tilde{u}_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^n \frac{\widehat{u}_i^2}{(1-h_{ii})^2}}{\sum_{i=1}^n (Y_i - \overline{Y})^2}.$$

Let's compare the three R-squared versions for various numbers of included regressors.

```
library(tidyverse)
R2 = function(fit){
  summary(fit)$r.squared
}
AdjR2 = function(fit){
  summary(fit)$adj.r.squared
}
LoocvR2 = function(fit, Y){
  LoocvResid = residuals(fit)/(1-hatvalues(fit))
  1-sum(LoocvResid^2)/sum((Y - mean(Y))^2)
}

## use 90 irrelevant regressors:
fit5   <- lm(Y~X+X_ir2[,1:90])
## use 95 irrelevant regressors:
fit6   <- lm(Y~X+X_ir2[,1:95])

specification = c(0, 1, 50, 90, 95, 98)
Rsquared = sapply(list(fit1, fit2, fit3, fit5, fit6, fit4), R2)
adjRsquared = sapply(list(fit1, fit2, fit3, fit5, fit6, fit4), AdjR2)
LoocvRsquared = sapply(list(fit1, fit2, fit3, fit5, fit6, fit4), LoocvR2, Y=Y)

round(tibble("#irrelevant regressors" = specification, Rsquared, adjRsquared, LoocvRsquared)
```

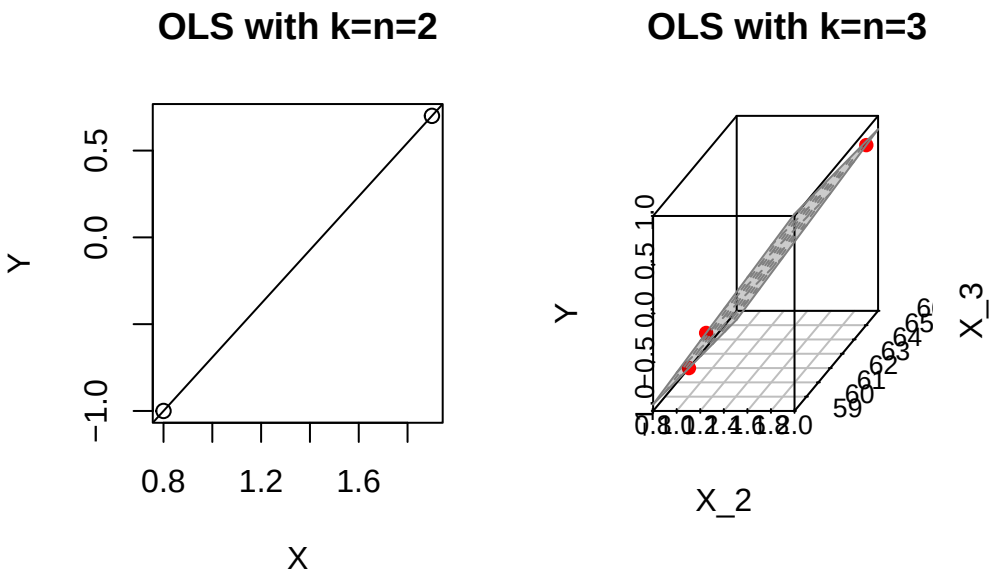
#irrelevant regressors	Rsquared	adjRsquared	LoocvRsquared
0	0.6933	0.6902	0.6799
1	0.6934	0.6870	0.6738
50	0.8353	0.6603	0.2681
90	0.9669	0.5900	-5.7026
95	0.9912	0.7103	-30.7321
98	1.0000	NaN	NaN

For comparing different OLS regression specifications, the leave-one-out R-squared should be preferred.

9.8 Too many regressors

The illustrative simulation above shows that the problem of overfitting increases as the number of regressors increases. If $k = n$ we obtain a perfect fit.

```r
par(mfrow=c(1,2))
## k=n=2
Y = c(0.7,-1.0)
X = c(1.9,0.8)
fit1 = lm(Y~X)
plot(X,Y, main="OLS with k=n=2")
abline(fit1)
## k=n=3
# Some given data
Y = c(0.7,-1.0,-0.2)
X_2 = c(1.9,0.8,1.25)
X_3 = c(66, 62, 59)
fit2 = lm(Y ~ X_2 + X_3)
plot3d  <- scatterplot3d(x = X_2, y = X_3, z = Y,
            angle = 33, scale.y = 0.8, pch = 16,
            color ="red",
            xlab = "X_2",
            ylab = "X_3",
            main ="OLS with k=n=3")
plot3d$plane3d(fit2, lty.box = "solid", col=gray(.5), draw_polygon=TRUE)
```



If the number of regressors and observations $k = n$ are higher, we can no longer visualize the OLS fit, but the problem still exists: we have $Y_i = \widehat{Y}_i$ for all $i = 1, \dots n$.

For $k > n$, we cannot compute the OLS regression because $\sum_{i=1}^{n} X_i X_i' = \boldsymbol{X}'\boldsymbol{X}$ is not invertible.

Regression problems with $k \approx n$ or $k > n$ are called **high-dimensional regressions**. In this case, regularization techniques such as Ridge, Lasso, Elastic-net, or dimension-reduction techniques such as principal components or partial least squares should be used.

OLS should be considered for regression problems with $k << n$ (small k and large n).

9.9 Multicollinearity

The only condition for computing the OLS coefficients is that $\sum_{i=1}^{n} X_i X_i' = \boldsymbol{X}'\boldsymbol{X}$ is invertible.

A necessary condition is that $k \leq n$, as discussed above.

Another reason the matrix may not be invertible is if two or more regressors are perfectly collinear. Two variables are perfectly collinear if their sample correlation is 1 or -1. We have multicollinearity if one variable is a linear combination of the other variables.

This can happen when the same regressor is included twice or in two different units (e.g., GDP in EUR and USD).

Suppose we include a variable that has the same value for all observations (e.g., a dummy variable where all individuals in the sample belong to the same group). In this case, the variable is collinear with the intercept variable. Including too many dummy variables can lead to the dummy variable trap (see below).

The regressor variables are **strictly multicollinear** (or perfectly multicollinear) if the regressor matrix does not have full column rank: $\text{rank}(\boldsymbol{X}) < k$. It implies $\text{rank}(\boldsymbol{X}'\boldsymbol{X}) < k$, so that the matrix is singular and $\hat{\beta}$ cannot be computed.

A related situation is near multicollinearity (often just called multicollinearity). It occurs when two columns of $\boldsymbol{X}$ have a sample correlation very close to 1 or -1. Then, $(\boldsymbol{X}'\boldsymbol{X})$ is "near singular", its eigenvalues are very small, and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ becomes very large, causing numerical problems.

Multicollinearity means that at least one regressor is redundant and can be dropped.

9.10 The dummy variable trap

In the following, we consider a simple linear regression model that aims to predict wages in 2015 using gender as the only predictor. We use data provided in the accompanying materials of Stock and Watson's *Introduction to Econometrics* textbook. You can download the data stored as an xlsx-file `cps_ch3.xlsx` HERE.

Let us first prepare the dataset:

wage	gender
25.64103	female
16.02564	female
33.65817	male
23.07692	female
12.98077	female
11.53846	male

```
## load the 'tidyverse' package
library("tidyverse")
## load the 'readxl' package
library("readxl")

## import the data into R
cps = read_excel(path = "cps_ch3.xlsx")

## Data wrangling
cps_2015 = cps %>% mutate(
    wage   = ahe15,      # rename "ahe15" as "wage"
    gender = fct_recode( # rename factor "a_sex" as "gender"
      as_factor(a_sex),
                "male" = "1",  # rename factor level "1" to "male"
                "female" = "2" # rename factor level "2" to "female"
              )
  ) %>%
  filter(year == 2015) %>%     # Only data from year 2008
  select(wage, gender)         # Select only the variables "wage" and "gender"
```

The first six lines of the dataset look as follows:

Computing the estimation results:

```
lm_obj = lm(wage ~ gender, data = cps_2015)
coef(lm_obj)
```

gives

- $\hat{\beta}_1 = 28.06$
- $\hat{\beta}_2 = $ -5.02

To compute these estimation results, one must assign a numeric 0/1 coding to the factor levels `male` and `female` of the factor variable `gender`. To see the numeric values used by `R`, one can take a look at `model.matrix(lm_obj)`:

```r
X = model.matrix(lm_obj) # this is the internally used X-matrix
X[1:6,]
```

```
  (Intercept) genderfemale
1           1            1
2           1            1
3           1            0
4           1            1
5           1            1
6           1            0
```

```r
cps_2015$gender[1:6]
```

```
[1] female female male   female female male
Levels: male female
```

Thus, `R` internally codes `female` subjects by 1 and `male` subjects by 0, such that

$$\hat{\beta}_1 + \hat{\beta}_2 X_{i,gender} = \begin{cases} \hat{\beta}_1 & \text{if } X_{i,gender} = \text{male} \\ \hat{\beta}_1 + \hat{\beta}_2 & \text{if } X_{i,gender} = \text{female} \end{cases}$$

Interpretation:

- The average wage of male workers in 2015 was $\hat{\beta}_1 = 28.06$ (USD/Hour).
- The average wage of female workers in 2015 was $\hat{\beta}_1 + \hat{\beta}_2 = 23.04$ (USD/Hour).
- The difference in the earnings between male and female workers in 2015 is $\hat{\beta}_2$ = -5.02 (USD/Hour).

Above, we used `R`'s internal handling of factor variables, which you should always do. However, if you construct a dummy variable for each of the levels of a factor, you may fall into the dummy variable trap:

```r
## Intercept variable
X_1      <- rep(1, times = nrow(cps_2015))

## 1. Dummy variable for 'female'
X_female <- ifelse(cps_2015$gender == "female", 1, 0)

## 2. Dummy variable for 'male'
X_male   <- ifelse(cps_2015$gender == "male", 1, 0)
```

```r
## Construct the model matrix 'X'
X        <- cbind(X_1, X_female, X_male)
```

Computing the OLS coefficients "by hand" yields

```r
## Dependent variable
Y        <- cps_2015$wage

## Computing the estimator
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y
```

```
Error in solve.default(X %*% t(X)) :
  system is exactly singular
```

An error message! We fell into the dummy variable trap!

The estimation result is not computable since $(\boldsymbol{X}'\boldsymbol{X})$ is not invertible due to the perfect multicollinearity between the intercept and the two dummy variables `X_female` and `X_male`

`X_1` = `X_female` + `X_male`

which violates the invertibility condition (no perfect multicollinearity).

Solution: Use one dummy variable less than factor levels. I.e., in this example, you can drop `X_female` or `X_male`.

```r
## New model matrix after dropping X_male:
X        <- cbind(X_1, X_female)

## Computing the estimator
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y
beta_hat
```

```
             [,1]
X_1      28.05536
X_female -5.01638
```

This gives the same result as computed by `R`'s `lm()` function using the factor variable `gender`:

$$f(X_i) = 28.055 - 5.016 X_{i,female}.$$

9.11 OLS without intercept

An intercept is not necessarily required for the OLS method. In this case, the first column of the regressor matrix $\boldsymbol{X}$ does not contain 1's. An alternative to solve the dummy variable trap is to drop the intercept:

```
## Write -1 to drop the intercept
lm_obj2 = lm(Y~X_female+X_male-1)
coefficients(lm_obj2)
```

```
X_female   X_male
23.03898 28.05536
```

The fitted regression function is

$$f(X_i) = 23.039 X_{i,female} + 28.055 X_{i,male}.$$

Notice the different interpretations of the OLS coefficients!

9.12 Additional reading

- Stock and Watson (2019), Sections 4, 6, 19
- Hansen (2022b), Section 4
- Davidson and MacKinnon (2004), Section 2

9.13 R-codes

statistics-sec9.R

10 Regression Models

10.1 The linear model

The previous section discussed OLS regression from a descriptive perspective. A regression model puts the regression problem into a stochastic framework.

Linear Regression Model

Let $\{(Y_i, X_i'), i = 1, ..., n\}$ be a sample from some joint population distribution, where Y_i is individual i's dependent variable, and $X_i = (1, X_{i2}, ..., X_{ik})'$ is the $k \times 1$ vector of individual i's regressor variables.

The linear regression model equation for $i = 1, ..., n$ is

$$Y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + u_i \tag{10.1}$$

where $\beta = (\beta_1, ..., \beta_k)'$ is the $k \times 1$ vector of **regression coefficients** and u_i is the **error term** for individual i.

In vector notation, we write

$$Y_i = X_i'\beta + u_i, \quad i = 1, ..., n.$$

The error term represents further factors that affect the dependent variable and are not included in the model. These factors include measurement errors, omitted variables, or unobserved/unmeasurable variables.

We can use matrix notation to describe the n individual regression equations jointly: The regressor matrix and the response vector are

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \underset{(n\times k)}{\boldsymbol{X}} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{12} & ... & X_{1k} \\ \vdots & & & \vdots \\ 1 & X_{n2} & ... & X_{nk} \end{pmatrix}$$

The vectors of coefficients and error terms are

$$\underset{(k\times 1)}{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \underset{(n\times 1)}{\boldsymbol{u}} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

In matrix notation, the $i = 1, \dots, n$ equations from Equation 10.1 can be jointly written as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{u}.$$

10.2 Conditional mean independence

A1: Mean independence condition

$$E[u_i \mid X_i] = 0 \tag{10.2}$$

Equation 10.2 is also called conditional mean assumption or **weak exogeneity condition**. It is the condition that essentially makes Equation 10.1 a regression model. It has multiple implications.

1) Zero unconditional mean

Equation 10.2 and the LIE imply

$$E[u_i] \overset{(LIE)}{=} E[\underbrace{E[u_i \mid X_i]}_{=0}] = E[0] = 0$$

The error term has a zero unconditional mean: $E[u_i] = 0$.

2) Linear best predictor

The conditional mean of Y_i given X_i is

$$E[Y_i \mid X_i] = \underbrace{E[X_i'\beta \mid X_i]}_{\overset{(CT)}{=} X_i'\beta} + \underbrace{E[u_i \mid X_i]}_{=0} = X_i'\beta.$$

A regression model is a model for the conditional expectation of the response given the regressors. A linear regression model assumes that the conditional expectation is linear.

Recall the best predictor property of the conditional expectation. The regression function $X_i'\beta$ is the best predictor for Y_i given X_i (assuming that $E[Y_i^2] < \infty$).

3) Marginal effect interpretation

$$E[Y_i \mid X_i] = X_i'\beta = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik}$$

implies

$$\frac{\mathrm{d}E[Y_i \mid X_i]}{\mathrm{d}X_{ij}} = \beta_j$$

The coefficient β_j is the marginal effect holding all other regressors fixed (partial derivative).

Note that the marginal effect β_j is not necessarily a causal effect of X_{ij} on Y_i. Unobserved variables that are correlated with X_{ij} might be the actual "cause" of the change in Y_i.

4) Weak exogeneity

For any regressor variable X_{ij}, $j = 1, ..., k$, we have

$$E[u_i \mid X_{ij}] \overset{(LIE)}{=} E[\underbrace{E[u_i \mid X_i]}_{=0} \mid X_{ij}] = 0$$

and

$$E[X_{ij}u_i] \overset{(LIE)}{=} E[E[X_{ij}u_i \mid X_{ij}]] \overset{(CT)}{=} E[X_{ij}\underbrace{E[u_i \mid X_{ij}]}_{=0}] = 0,$$

which implies

$$Cov(X_{ij}, u_i) = \underbrace{E[X_{ij}u_i]}_{=0} - E[X_{ij}]\underbrace{E[u_i]}_{=0} = 0.$$

The error term is uncorrelated with all regressor variables. The regressors are called **exogenous**. Therefore, the condition $E[u_i \mid X_i] = 0$ is also called **weak exogeneity condition**.

It implies that the error term describes only the unobserved variables that are uncorrelated with the regressors. Note that it does not mean that we exclude any effects on the dependent variable of unobserved variables that are correlated with the regressors (which is a quite likely situation).

The exogeneity condition only tells us that the marginal effect β_j is an average effect of a change in X_{ij}, holding all other regressors fixed. We cannot hold fixed the unobserved variables that are correlated with X_{ij}. Any effects of unobserved variables that are correlated with the j-th regressor are implicitly part of the model since they enter the marginal effect β_j through their correlation with X_{ij}.

10.3 Correlation and causation

Let's consider, for example, the wage per hour on education regression model:

$$wage_i = \beta_1 + \beta_2\ edu_i + u_i, \quad E[u_i \mid edu_i] = 0, \quad i = 1, \dots, n.$$

We have

$$E[wage_i \mid edu_i] = \beta_1 + \beta_2 edu_i.$$

The average wage level among all individuals with z years of schooling is $\beta_1 + \beta_2 z$. The marginal effect of education is

$$\frac{\mathrm{d}E[wage_i \mid edu_i]}{\mathrm{d}\ edu_i} = \beta_2.$$

Interpretation: Suppose that $\beta_2 = 1.5$. People with one more year of education are paid on average 1.50 EUR more than people with one year less of education.

$$Cov(wage_i, edu_i) = Cov(\beta_1 + \beta_2 edu_i, edu_i) + \underbrace{Cov(u_i, edu_i)}_{=0} = \beta_2 Var[edu_i]$$

The coefficient β_2 is

$$\beta_2 = \frac{Cov(wage_i, edu_i)}{Var[edu_i]} = Corr(wage_i, edu_i) \cdot \frac{sd(wage_i)}{sd(edu_i)}$$

It describes the **correlative relationship** between education and wages. It makes no statement about where exactly a higher wage level for people with more education comes from. **Regression coefficients do not necessarily yield causal conclusions**.

Maybe people with more education are just smarter on average and therefore earn more on average. Or maybe people with more education have richer parents on average and, therefore, have jobs with higher salaries. Or maybe education really does help to increase wages.

The coefficient β_2 is high when the correlation between education and wage is high.

This could be due to family background (parental education, family income, ethnicity, structural racism) if family background correlates with wage and education.

Or it could be due to personal background (gender, intelligence) if personal background correlates with wage and education.

Of course, it could also be due to an actual high causal effect of education on wages.

Perhaps it is a mixture of different effects. We cannot disentangle these effects unless we include additional control variables.

Notice: Correlation does not imply causation!

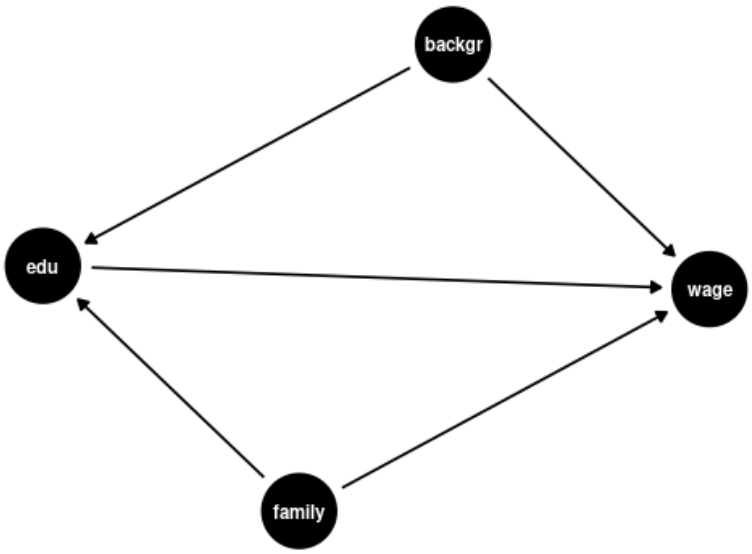


Figure 10.1: A DAG (directed acyclic graph) for the correlative and causal effects of edu on wage

Figure 10.2: Correlation and Causality

Suppose the research question is to understand the causal effect of an additional year of education on wages, with family and personal background held fixed. In that case, family and personal background are so-called **omitted variables**.

A variable is an omitted variable if

(i) it is correlated with the dependent variable (wage in our case),
(ii) correlated with the regressor of interest (education in our case),
(iii) omitted in the regression.

If omitted variables are present, we say that we have an **omitted variable bias** for the causal effect of the regressor of interest. Omitted variables imply that we cannot interpret the coefficient β_2 as a causal effect. It is simply a correlative effect or marginal effect. This must always be kept in mind when interpreting regression coefficients This

We can include **control variables** in the linear regression model to reduce the omitted variables so that we can interpret β_2 as a **ceteris paribus marginal effect** (ceteris paribus means with other variables held constant). For instance, we can include the experience in years and the ethnicity and gender dummy variables for Black and female:

$$wage_i = \beta_1 + \beta_2 edu_i + \beta_3 exp_i + \beta_4 Black_i + \beta_5 female_i + u_i$$

In this case,

$$\beta_2 = \frac{\mathrm{d}E[wage_i \mid edu_i, exp_i, Black_i, female_i]}{\mathrm{d}\ edu_i}$$

is the marginal effect of education on expected wages, holding constant experience, ethnicity, and gender.

But: it does not hold constant further unobservable characteristics (such as ability) nor variables not included in the regression (such as quality of education), so an omitted variaable bias might still be present.

10.4 Nonlinearities

A linear relationship is not always appropriate. E.g., it is not reasonable that an individual with one year of education has the same marginal effect of education on average wages as an individual with 20 years of education. It would make more sense if the marginal effect were a percentage change.

In the level specification

$$wage_i = \beta_1 + \beta_2 edu_i + u_i$$

the parameter β_2 approximates an absolute change in wage when education changes by 1:

$$\frac{\mathrm{d}\ wage_i}{\mathrm{d}\ edu_i} = \beta_2 \quad \Rightarrow \quad \underbrace{\mathrm{d}\ wage_i}_{\approx \text{absolute change}} = \beta_2 \underbrace{\mathrm{d}\ edu_i}_{\approx \text{absolute change}}$$

In the logarithmic specification

$$\log(wage_i) = \beta_1 + \beta_2 edu_i + u_i$$

the parameter β_2 approximates the percentage change in wage when education changes by 1:

$$\begin{aligned} wage_i &= e^{\beta_1 + \beta_2 edu_i + u_i} \\ \Rightarrow \quad \frac{\mathrm{d}\ wage_i}{\mathrm{d}\ edu_i} &= \beta_2 e^{\beta_1 + \beta_2 edu_i + u_i} = \beta_2\ wage_i \\ \Rightarrow \quad \underbrace{\frac{\mathrm{d}\ wage_i}{wage_i}}_{\approx \text{ percentage change}} &= \beta_2 \underbrace{\mathrm{d}\ edu_i}_{\approx \text{ absolute change}} \end{aligned}$$

For instance, if $\beta_2 = 0.05$, then a person with one more year of education has, on average, a 5% higher wage.

The linear regression model is less restrictive than it appears. It can represent nonlinearities very flexibly since we can include as regressors nonlinear transformations of the original regressors.

A linear regression with quadratic and interaction terms

$$\begin{aligned} wage_i = \beta_1 &+ \beta_2 edu_i + \beta_3 exp_i + \beta_4 exp_i^2 \\ &+ \beta_5 female_i + \beta_6 married_i + \beta_7 (married_i \cdot female_i) + u_i \end{aligned}$$

Marginal effects depend on the person's experience level/marital status/gender:

$$\begin{aligned} \frac{\mathrm{d}\ wage_i}{\mathrm{d}\ exp_i} &= \beta_3 + 2\beta_4 exp_i \\ \frac{\mathrm{d}\ wage_i}{\mathrm{d}\ female_i} &= \beta_5 + \beta_7 married_i \\ \frac{\mathrm{d}\ wage_i}{\mathrm{d}\ married_i} &= \beta_6 + \beta_7 female_i \end{aligned}$$

10.5 The moment estimator

Reall that the exogeneity condition $E[u_i \mid X_i] = 0$ implies $E[X_{ij}u_i] = 0$ for all $j = 1, \dots, k$. Thus, the exogeneity condition gives us a system of k linear equations:

$$\left.\begin{array}{c} E[u_i] = 0 \\ E[X_{i2}u_i] = 0 \\ \vdots \\ E[X_{ik}u_i] = 0 \end{array}\right\} \quad \Leftrightarrow \quad \underset{(k\times 1)}{E[X_i u_i]} = \underset{(k\times 1)}{\mathbf{0}_k}$$

It allows us to identify the unknown parameter vector $\beta \in \mathbb{R}^k$ in terms of population moments:

$$\begin{aligned} E[X_i(\underbrace{Y_i - X_i'\beta}_{=u_i})] &= \mathbf{0}_k \\ E[X_iY_i] - E[X_iX_i']\beta &= \mathbf{0}_k \\ E[X_iX_i']\beta &= E[X_iY_i] \\ \beta &= (E[X_iX_i'])^{-1}E[X_iY_i] \end{aligned}$$

The regression coefficent vector β is a function of $E[X_iX_i']$ and $E[X_iY_i]$, which are population moments of the (k+1)-variate random variable (Y_i, X_i'). The corresponding sample moments are $\frac{1}{n}\sum_{i=1}^n X_iX_i'$ and $\frac{1}{n}\sum_{i=1}^n X_iY_i$. The moment estimator for β substitutes population moments by sample moments:

$$\hat{\beta}_{mm} = \left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_iY_i,$$

It can be simplified as follows:

$$\begin{aligned} \hat{\beta}_{mm} &= \left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_iY_i \\ &= \left(\sum_{i=1}^n X_iX_i'\right)^{-1}\sum_{i=1}^n X_iY_i \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \hat{\beta}. \end{aligned}$$

Thus, the method of moments estimator, $\hat{\beta}_{mm}$, coincides with the OLS coefficient vector $\hat{\beta}$. Therefore, we call $\hat{\beta} = \hat{\beta}_{mm}$ the **OLS estimator** for β.

10.6 Random sampling assumption

A2: Random sampling

The (k+1)-variate observations $\{(Y_i, X_i'), i = 1, \dots, n\}$ are an i.i.d. sample from some joint population distribution.

The i.i.d. assumption (A2) implies that $\{(Y_i, X_i', u_i), i = 1, \dots, n\}$ is an i.i.d. collection since $u_i = Y_i - X_i'\beta$ is a function of a random sample, and functions of independent variables are independent as well. Individual i's error term u_i is independent of u_j, X_j, and Y_j for any other individual $j \neq i$.

This implies that the weak exogeneity condition (A1) turns into a **strict exogeneity** property:

$$E[u_i \mid \boldsymbol{X}] = E[u_i \mid X_1, \dots, X_n] \overset{(A2)}{=} E[u_i \mid X_i] \overset{(A1)}{=} 0. \tag{10.3}$$

Weak exogeneity means that individual i's regressors are uncorrelated with individual i's error term. Strict exogeneity means that individual i's regressors are uncorrelated with the error terms of any individual in the sample.

10.6.1 Exogeneity and time series regression

Strict exogeneity is a property in a regression model with randomly sampled data, but it may not hold in dynamic time series regression models with an autocorrelated response variable.

A dynamic time series model is a model where one of the regressors is a lag of the dependent variable. For instance, the model

$$Y_t = \beta_1 + \beta_2 X_t + u_t, \quad E[u_t \mid X_t] = 0,$$

with $X_t = Y_{t-1}$ is called AR(1) dynamic regression model. Since

$$Cov(Y_t, u_t) = \beta_2 \underbrace{Cov(X_t, u_t)}_{=0} + Cov(u_t, u_t) = Var[u_t] \neq 0$$

and $X_{t+1} = Y_t$, we have $Cov(u_t, X_{t+1}) \neq 0$. Hence, weak exogeneity (Equation 10.2) holds in dynamic time series models, but strict exogeneity (Equation 10.3) does not.

For regressions with time series data, the following alternative assumption is made:

A2b: Weak dependence

The (k+1)-variate observations $\{(Y_t, X_t'), t = 1, \dots, n\}$ are a stationary short-memory time series, and (Y_t, X_t') and $(Y_{t-\tau}, X_{t-\tau}')$ become independent as τ gets large.

The precise mathematical statement for A2b is omitted. It essentially requires that the dependent variable and the regressors together have a time-independent autocovariance structure (no structural changes, no non-stationarities) and that the dependence on the time series of τ periods before must decrease with higher τ.

10.6.2 Heteroskedasticity

The i.i.d. assumption (A2) is not as restrictive as it may seem at first sight. It allows for dependence between u_i and $X_i = (X_{i1}, \dots, X_{ik})'$. The error term u_i can have a conditional distribution that depends on X_i.

The exogeneity assumption (A1) requires that the conditional mean of u_i is independent of X_i. Besides this, dependencies between u_i and $X_{i1}, \dots, X_{ik}$ are allowed. For instance, the variance of u_i can be a function of $X_{i1}, \dots, X_{ik}$. If this is the case, u_i is said to be **heteroskedastic**.

Let's have a look at the conditional covariance matrix

$$\boldsymbol{D} := Var[\boldsymbol{u} \mid \boldsymbol{X}] = E[\boldsymbol{u}\boldsymbol{u}' \mid \boldsymbol{X}].$$

(A2) implies $E[u_i \mid u_j, \boldsymbol{X}] = E[u_i \mid \boldsymbol{X}] = 0$ for $j \neq i$ and, therefore,

$$E[u_i u_j \mid \boldsymbol{X}] \overset{(LIE)}{=} E\big[E[u_i u_j \mid u_j, \boldsymbol{X}] \mid \boldsymbol{X}\big] \overset{(CT)}{=} E\big[u_j \underbrace{E[u_i \mid u_j, \boldsymbol{X}]}_{=0} \mid \boldsymbol{X}\big] = 0.$$

Hence, u_i and u_j are conditionally uncorrelated. The off-diagonal elements of $\boldsymbol{D}$ are zero.

The main diagonal elements of $\boldsymbol{D}$ are

$$E[u_i^2 \mid \boldsymbol{X}] \overset{(A2)}{=} E[u_i^2 \mid X_i] =: \sigma_i^2 = \sigma^2(X_i).$$

The conditional variances of u_i may depend on the values of X_i. We have

$$\boldsymbol{D} = E[\boldsymbol{u}\boldsymbol{u}' \mid \boldsymbol{X}] = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} = diag(\sigma_1^2, \dots, \sigma_n^2).$$

Homoskedastic and heteroskedastic errors

An error term is **heteroskedastic** if the conditional variances

$$Var[u_i \mid X_i = x_i] = \sigma^2(X_i)$$

are equal to a non-constant variance function $\sigma^2(x_i) > 0$, which is a function of the value of the regressor $X_i = x_i$.

An error term is **homoskedastic** if the conditional variances

$$Var[u_i \mid X_i = x_i] = \sigma^2 = Var[u_i]$$

are equal to some constant $\sigma^2 > 0$ for every possible regressor realization $X_i = x_i$.

Homoskedastic errors are a restrictive assumption sometimes made for convenience in addition to (A1)+(A2). Homoskedasticity is often unrealistic in practice, so we stick with the heteroskedastic errors framework.

Example (wages and gender):

It may be the case that there is more variation in the wage levels of male workers than in the wage levels of female workers. I.e., $\{(wage_i, female_i),\ i = 1, \dots, n\}$ may be an i.i.d. sample that follows the regression model

$$wage_i = \beta_1 + \beta_2 female_i + u_i, \quad E[u_i \mid female_i] = 0,$$

where

$$Var[u_i \mid female_i] = (1 - 0.4 female_i)\sigma^2 = \sigma_i^2,$$

and $female_i$ is a dummy variable. Male workers have an error variance of σ^2, and female workers have an error variance of $0.6\sigma^2$. The conditional error variance depends on the regressor value, i.e., the errors are heteroskedastic.

If the proportion of female and male workers is 50/50 (i.e. $P(female_i = 1) = 0.5$), the unconditional variance is

$$\begin{aligned} Var[u_i] &= E[u_i^2] \\ &= E[E[u_i^2 \mid female_i]] \\ &= \sigma^2 \cdot 0.5 + 0.6\sigma^2 \cdot 0.5 \\ &= 0.8\sigma^2. \end{aligned}$$

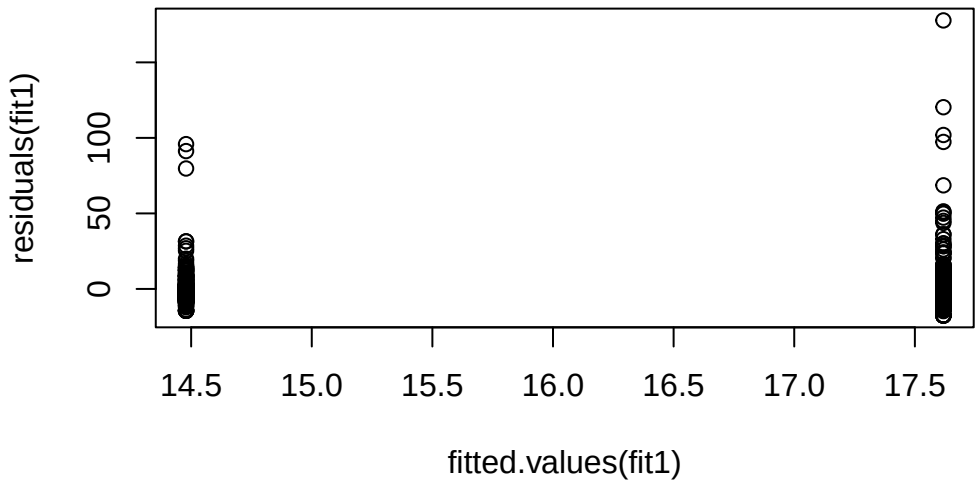
10.6.3 Detecting heteroskedasticity

Let's have a look at what the data tells us. We use data from the Allbus21 survey.

```
library(statisticsdata)
fit1 = lm(wage ~ female, data=allbus21)
fit1
```

```
Call:
lm(formula = wage ~ female, data = allbus21)

Coefficients:
(Intercept)       female
     17.618       -3.139
```

```r
plot(fitted.values(fit1), residuals(fit1))
```



The plot shows the residuals $\hat{u}_i$ plotted against the fitted values $\widehat{Y}_i$. The residuals for male workers have more variation than those for female workers, which indicates heteroskedasticity.

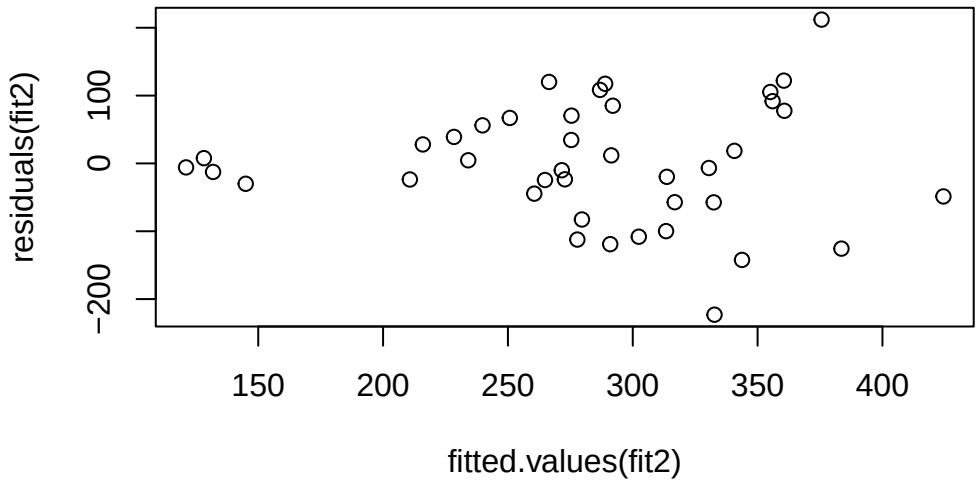
Another example - regressing food expenditure on income:

```r
# install.packages("remotes")
# remotes::install_github("ccolonescu/PoEdata")
library(PoEdata) # for the "food" dataset contained in this package
data("food")     # makes the dataset "food" usable

fit2 = lm(food_exp ~ income, data = food)
## Diagnostic scatter plot of residuals vs fitted values
plot(fitted.values(fit2), residuals(fit2))
```

The diagnostic plot indicates that the variance of unobserved factors for food expenditure is higher for people with high income than those with low income.

Note: Plotting against the fitted values is actually a pretty smart idea since this also works for multiple predictors $X_{i1}, \ldots, X_{ik}$ with $k \geq 3$.

For historical reasons, statistics books often treat homoskedasticity as the standard case and heteroskedasticity as a special case. However, this does not reflect empirical practice since we have to expect heteroskedastic errors in most applications. It turns out that heteroskedasticity is not a problem as long as the correct inference method is chosen (robust standard errors). We will consider heteroskedasticity as the standard case and homoskedasticity as a restrictive special case.



10.7 Sampling mean and variance

Recall that (A1) and (A2) imply the strict exogeneity and heteroskedasticity property:

$$E[\boldsymbol{u} \mid \boldsymbol{X}] = \boldsymbol{0}_n, \quad Var[\boldsymbol{u} \mid \boldsymbol{X}] = \boldsymbol{D} = diag(\sigma_1^2, \dots, \sigma_n^2).$$

To compute the bias and MSE of the OLS estimator $\hat{\beta}$ for β, the following decomposition is useful:

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{u}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}
\end{aligned}$$

The linear regression model is a conditional model, so let's first compute the mean and variance of $\hat{\beta}$ conditional on all regressors of all individuals $\boldsymbol{X}$.

$$\begin{aligned}
E[\hat{\beta} \mid \boldsymbol{X}] &= \beta + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \mid \boldsymbol{X}] \\
&\overset{(CT)}{=} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{E[\boldsymbol{u} \mid \boldsymbol{X}]}_{=\boldsymbol{0}} = \beta
\end{aligned}$$

The OLS estimator is unbiased conditional on $\boldsymbol{X}$, which means that the estimator is unbiased for any realization of the regressors. Unbiased conditional on $\boldsymbol{X}$ is a stronger concept than unconditional unbiasedness since, by the LIE,

$$E[\hat{\beta}] = E[\underbrace{E[\hat{\beta} \mid \boldsymbol{X}]}_{=\beta}] = \beta.$$

Hence, the **OLS estimator is unbiased**: $bias[\hat{\beta}] = 0$.

Recall the matrix rule $Var[\boldsymbol{Az}] = \boldsymbol{A}Var[\boldsymbol{z}]\boldsymbol{A}'$ if $\boldsymbol{z}$ is a random vector and $\boldsymbol{A}$ is a matrix. Then, the conditional sampling variance is

$$\begin{aligned} Var[\hat{\beta} \mid \boldsymbol{X}] &= Var[\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \mid \boldsymbol{X}] \\ &= Var[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \mid \boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Var[\boldsymbol{u} \mid \boldsymbol{X}]((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \sigma_i^2 X_i X_i'\right)\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}. \end{aligned}$$

It turns out that $Var[\hat{\beta}_j \mid \boldsymbol{X}] \xrightarrow{p} 0$ for any j and any realization of the regressors $\boldsymbol{X}$, and $\lim_{n\to\infty} Var[\hat{\beta}_j] = 0$.

Proof. Under the additional condition that $0 < E[X_{ij}^4] < \infty$ and $0 < E[Y_i^4] < \infty$, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} E[X_i X_i'] =: \boldsymbol{Q}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 X_i X_i' \xrightarrow{p} E[(X_i u_i)(X_i u_i)'] =: \boldsymbol{\Omega},$$

where the matrices $\boldsymbol{Q}$ and $\boldsymbol{\Omega}$ are positive definite if there is no strict multicollinearity. Then,

$$\begin{aligned} Var[\hat{\beta} \mid \boldsymbol{X}] &= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} \sigma_i^2 X_i X_i'\right)\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \\ &= \underbrace{\frac{1}{n}}_{\to 0}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i'}_{\xrightarrow{p}\boldsymbol{Q}}\right)^{-1}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 X_i X_i'}_{\xrightarrow{p}\boldsymbol{\Omega}}\right)\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i'}_{\xrightarrow{p}\boldsymbol{Q}}\right)^{-1}, \end{aligned}$$

which implies that $Var[\hat{\beta} \mid \boldsymbol{X}] \xrightarrow{p} \boldsymbol{0}_{k\times k}$, as $n \to \infty$. Some mathematical tools are required to show the steps rigorously (Cauchy-Schwarz inequality, Slutsky's theorem, continuous mapping theorem).

Since $\hat{\beta}_j$ is unbiased and the conditional variance converges to zero, the unconditional variance also converges to zero. □

□

10.8 Consistency

Since, for any $j = 1, \dots, k$, $bias[\hat{\beta}_j] = 0$ and $Var[\hat{\beta}_j] \to 0$ as $n \to \infty$, we have

$$\lim_{n\to\infty} MSE[\hat{\beta}_j] = 0,$$

which implies that the OLS estimator is consistent.

In the above steps, we required that the fourth moments of the population distribution of (Y_i, X_i') are bounded, or equivalently, that the kurtosis of the population from which the data is sampled is finite. That is, fat-tailed population distributions are not permitted.

A3: Large outliers are unlikely

Response and regressor variables have nonzero finite fourth moments:

$$0 < E[Y_i^4] < \infty, \quad 0 < E[X_{ij}^4] < \infty$$

for all $j = 1, \dots, k$.

Another condition that we implicitly used is that the OLS estimator can be computed. I.e., strict multicollinearity is not allowed:

A4: No perfect multicollinearity

The regressor matrix $\boldsymbol{X}$ has full column rank.

The linear regression model Equation 10.1 under assumptions (A1)–(A4) is called **heteroskedastic linear regression model**.

OLS consistency

Under assumptions (A1)–(A4), the OLS estimator $\hat{\beta}$ is consistent for β. It is unbiased, and its conditional variance is

$$\begin{aligned} Var[\hat{\beta} \mid \boldsymbol{X}] &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \sigma_i^2 X_i X_i'\right) \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}. \end{aligned}$$

10.9 Efficiency

It turns out that the OLS estimator is not the most efficient estimator unless the errors are homoskedastic.

A5: Homoskedasticity

The conditional variance of the errors does not depend on the realized regressors:

$$Var[u_i \mid X_i] = \sigma^2, \quad \text{for all } i = 1, \dots, n.$$

The linear regression model Equation 10.1 under assumptions (A1)–(A5) is called **homoskedastic linear regression model**.

An estimator $\hat{\theta}$ is more efficient that an estimator $\widehat{\varphi}$ if $Var[\hat{\theta}] < Var[\widehat{\varphi}]$. If $\hat{\theta}$ and $\widehat{\varphi}$ are vectors, their variances are matrices, and the expression $Var[\hat{\theta}] < Var[\widehat{\varphi}]$ means that $(Var[\widehat{\varphi}] - Var[\hat{\theta}])$ is a positive definite matrix (see matrix tutorial)

Gauss-Markov theorem

In the homoskedastic linear regression model, the OLS estimator $\hat{\beta}$ is the best linear unbiased estimator (BLUE) for β. That is, for any other linear unbiased estimator $\tilde{\beta}$, we have

$$Var[\tilde{\beta} \mid X] \geq Var[\hat{\beta} \mid X]$$

We call the LS estimator $\hat{\beta}$ the efficient estimator in the homoskedastic linear regression model. Homoskedastic errors imply that $\sigma_i^2 = \sigma^2$ for all $i = 1, \dots, n$ so that the conditional variance simplifies in this case to

$$\begin{aligned}
Var[\hat{\beta} \mid \boldsymbol{X}] &= \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} \underbrace{\sigma_i^2}_{=\sigma^2} X_i X_i' \right) \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \\
&= \sigma^2 \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i X_i' \right) \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \\
&= \sigma^2 \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \\
&= \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

In the heteroskedastic linear regression model, OLS is not efficient. We can recover the Gauss-Markov efficiency in the heteroskedastic linear regression model if we use the **generalized least squares estimator (GLS)** instead:

$$\begin{aligned}
\hat{\beta}_{gls} &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{Y} \\
&= \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i Y_i \right),
\end{aligned}$$

where $\boldsymbol{D} = Var[\boldsymbol{u} \mid \boldsymbol{X}] = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$.

The GLS estimator is BLUE in the heteroskedastic linear regression model.

However, GLS is not feasible in practice since $\sigma_1^2, \dots, \sigma_n^2$ are unknown. These quantities can be estimated, but this is another source of uncertainty, and further assumptions are required (feasible GLS - FGLS). Therefore, most practitioners prefer to use the OLS estimator under heteroskedasticity and are willing to accept the loss of efficiency compared to GLS.

10.10 Normality

A6: Normal errors

The conditional distribution of u_i given X_i is $\mathcal{N}(0, \sigma^2)$. That is, the conditional density of u_i given $X_i = x_i$ is

$$f_{u_i|X_i}(z \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2} z^2 \Big)$$

The condition means that the conditional distribution of the errors given the regressors is normal and does not depend on them:

$$f_{u_i|X_i}(z \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2} z^2 \Big)$$

It does not mean that the regressors are normally distributed. It only means that Y_i conditional on X_i must be normal:

$$f_{Y_i|X_i}(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2} (y - x'\beta)^2 \Big)$$

The linear regression model Equation 10.1 under assumptions (A1)–(A6) is called **normal linear regression model**.

The OLS estimator $\widehat{\beta}$ is a linear function of the regressors $\boldsymbol{X}$ and the errors $\boldsymbol{u}$:

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

Conditional on $\boldsymbol{X}$, the OLS estimator is a linear combination of the errors $\boldsymbol{u}$. If $\boldsymbol{u}$ conditional on $\boldsymbol{X}$ is normal, then any linear combination is also normal. Hence, the distribution of $\widehat{\beta}$ conditional on $\boldsymbol{X}$ is normal. Recall that

$$E[\widehat{\beta} \mid \boldsymbol{X}] = \beta, \quad E[\widehat{\beta} \mid \boldsymbol{X}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

This leads to the following result:

Exact normality

Under assumptions (A1)–(A6), the distribution of the OLS estimator conditional on the regressors $\boldsymbol{X}$ is normal for any fixed sample size n with

$$\widehat{\beta} \mid \boldsymbol{X} \sim \mathcal{N}\Big(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\Big).$$

Using this result, we can derive exact inference methods (confidence intervals and t-tests)

10.11 Asymptotic normality

Assumptions (A5) and (A6) are quite restrictive. Dependent variables are rarely exactly normally distributed, and error variances often depend on the values of the regressors.

(A5) and (A6) are quite useful for pedagogical purposes and to understand under what situation the exact distribution of the OLS estimator can be recovered.

Similar to the sample mean case, it turns out that the restrictive assumptions are unnecessary if we are willing to accept that we can only obtain the asymptotic distribution of the estimator. Fortunately, a central limit theorem also holds for the OLS estimator under assumptions (A1)–(A4).

Asymptotic normality of OLS for i.i.d. data

Under assumptions (A1)–(A5), the asymptotic distribution of the OLS estimator is

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}\Big(0, \sigma^2 \boldsymbol{Q}^{-1}\Big),$$

where $\boldsymbol{Q} = E[X_i X_i']$ is the second moment matrix of the regressors.

Under assumptions (A1)–(A4), the asymptotic distribution of the OLS estimator is

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}\Big(0, \boldsymbol{Q}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}^{-1}\Big), \tag{10.4}$$

where $\boldsymbol{\Omega} = E\big[(X_i u_i)(X_i u_i)'\big]$.

Using this result, we can derive asymptotic inference methods (confidence intervals and t-tests)

We can also use the OLS estimator if the variables are time series data. In this case, we have to assume that the time series are weakly dependent in the sense of Assumption (A2b).

The linear regression model Equation 10.1 under assumptions (A1), (A2b), (A3), and (A4) is called **dynamic linear regression model**.

Asymptotic normality of OLS for time series data

Under assumptions (A1), (A2b), (A3), and (A4), the asymptotic distribution of the OLS estimator is

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}\Big(0, \boldsymbol{Q}^{-1} \boldsymbol{\Omega}^* \boldsymbol{Q}^{-1}\Big),$$

where

$$\boldsymbol{\Omega}^* = \boldsymbol{\Omega} + \sum_{\tau=1}^{\infty} \Big(E\big[(X_i u_i)(X_{i-\tau} u_{i-\tau})'\big] + E\big[(X_i u_i)(X_{i-\tau} u_{i-\tau})'\big]' \Big).$$

10.12 Additional reading

- Stock and Watson (2019), Sections 4, 6, 8, 15, 18, 19
- Hansen (2022b), Sections 4, 5, 7
- Davidson and MacKinnon (2004), Section 3

10.13 R-codes

statistics-sec10.R

11 Classical Inference

In this section, we study confidence intervals and hypothesis tests for linear regression coefficients under normality and homoskedasticity.

Of course, normality and homoskedasticity are restrictive cases. It is helpful to consider this special case first because we obtain exact inferential methods, i.e., exact confidence intervals and exact size-α test for any fixed sample size n. For the more general cases, the inferential methods can be easily adapted using large n approximations. We will discuss this in the final section.

Recall that the **normal linear regression model** is defined by the following assumptions:

Normal Linear Regression Assumptions

The variables Y_i and $X_i = (1, X_{i2}, \dots, X_{ik})'$ satisfy the linear regression equation

$$Y_i = X_i'\beta + u_i, \quad i = 1, \dots, n, \tag{11.1}$$

which has the matrix representation $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{u}$.

For each $i = 1, \dots, n$ we assume

- (A1) **conditional mean independence**: $E[u_i \mid X_i] = 0$
- (A2) **random sampling**: (Y_i, X_i') are i.i.d. draws from their joint population distribution
- (A3) **large outliers unlikely**: $0 < E[Y_i^4] < \infty$, $0 < E[X_{ij}^4] < \infty$ for all $j = 1, \dots, k$
- (A4) **no perfect multicollinearity**: $\boldsymbol{X}$ has full column rank
- (A5) **homoskedasticity**: $Var[u_i \mid X_i] = \sigma^2$
- (A6) **normal errors**: $u_i \mid X_i \sim \mathcal{N}(0, \sigma^2)$

Useful results from the previous section:

(i) $\widehat{\beta} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$
(ii) $E[\widehat{\beta} \mid \boldsymbol{X}] = \beta$ under (A1)–(A4)
(iii) $\boldsymbol{D} = Var[\boldsymbol{u} \mid \boldsymbol{X}] = diag(\sigma_1^2, \dots, \sigma_n^2)$ under (A1)–(A4)
(iv) $Var[\widehat{\beta} \mid \boldsymbol{X}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$ under (A1)–(A4)
(v) $\boldsymbol{D} = Var[\boldsymbol{u} \mid \boldsymbol{X}] = \sigma^2\boldsymbol{I}_n$ under (A1)–(A5)

(vi) $Var[\hat{\beta} \mid \boldsymbol{X}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ under (A1)–(A5)
(vii) $\hat{\beta} \mid \boldsymbol{X}$ is normally distributed under (A1)–(A6)

11.1 Standardized OLS coefficients

Under (A1)–(A6) we have

$$\hat{\beta} \mid \boldsymbol{X} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

which is a k-variate random variable. For each component $j = 1, \dots, k$, we have

$$\hat{\beta}_j \mid \boldsymbol{X} \sim \mathcal{N}(\beta_j, \sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}),$$

which is a univariate random variable. Note that $[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}$ is the j-th diagonal element of the $k \times k$ matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

In section Section 6.2 we studied how to obtain a confidence interval for normally distributed estimators.

Analogously to the sample mean case, let's standardize our estimator using the conditional variance:

$$Z_j := \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j \mid \boldsymbol{X})} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 E[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}}.$$

It is standard normal:

$$Z_j \mid \boldsymbol{X} \sim \mathcal{N}(0,1), \quad Z_j \sim \mathcal{N}(0,1).$$

The distribution of Z_j is standard normal conditional on $\boldsymbol{X}$ because $\hat{\beta}_j \mid \boldsymbol{X}$ is normal, and the unconditional distribution of Z_j is standard normal as well because for every $\boldsymbol{X}$ the conditional distribution is the same. This is because we standardize conditionally on $\boldsymbol{X}$.

11.2 Exact confidence intervals

The fact that Z_j is standard normal directly implies that the infeasible confidence interval derived in Section 6.2 can be used (set $\hat{\theta} = \hat{\beta}_j$ and $\theta = \beta_j$):

$$I_{1-\alpha} = [\hat{\beta}_j - z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\beta}_j \mid \boldsymbol{X});\ \hat{\beta}_j + z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\beta}_j \mid \boldsymbol{X})]$$

The interval satisfies $P(\beta_j \in I_{1-\alpha}) = 1 - \alpha$ for any fixed sample size n.

The interval is infeasible since

$$sd(\hat{\beta}_j \mid \boldsymbol{X}) = \sqrt{\sigma^2 E[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}$$

is unknown. The error variance $\sigma^2 = Var[u_i \mid X_i]$ must be estimated.

An unbiased and consistent estimator for σ^2 is

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

We correct by the k degrees of freedom we lose since we have k estimated coefficients that define the residuals $\hat{u}_i = Y_i - X_i'\hat{\beta}$. Similarly to the sample variance case in Section 6.5, it satisfies

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}.$$

It's squareroot s is the **standard error of regression**

$$s = SER = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i}.$$

We also call it **residual standard error**. The classical standard error is

$$se(\hat{\beta}_j) = \sqrt{s^2 E[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}},$$

and the t-ratio for the true coefficient β_j is

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k}.$$

Then, following Section 6.5, a feasible and exact $(1-\alpha)$-confidence interval for β_j is

$$I_{1-\alpha} = [\hat{\beta}_j - t_{(n-k;1-\frac{\alpha}{2})} \cdot se(\hat{\beta}_j);\ \hat{\beta}_j + t_{(n-k;1-\frac{\alpha}{2})} \cdot se(\hat{\beta}_j)]$$

Instead of s^2, we could also use the sample variance $\hat{\sigma}^2_{\hat{u}}$ (divided by n instead of n-k) as an estimate of σ^2. Note that s^2 is unbiased and $\hat{\sigma}^2_{\hat{u}}$ is biased, but $\hat{\sigma}^2_{\hat{u}}$ has a lower variance than s^2 (bias-variance tradeoff). We use s^2 in practice because otherwise the exact t-distribution would not hold.

11.3 Exact t-tests

The t-ratio for the hypothesis $H_0 : \beta_j = \beta_j^0$ is

$$T_0 = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)}.$$

Under H_0, it is t_{n-k}-distributed. Hence, the two-sided t-test for H_0 is given by the test decision

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } |T_0| \leq t_{(n-k;1-\frac{\alpha}{2})}, \\ \text{reject } H_0 \quad & \text{if } |T_0| > t_{(n-k;1-\frac{\alpha}{2})}. \end{aligned}$$

Let F_0 be the CDF of the null-distribution t_{n-k}, i.e. $F_0(t_{(n-k;p)}) = p$. The p-value for the two-sided t-test is

$$p\text{-value} = 2(1 - F_0(|T_0|))$$

The test decision can be equivalently reached by the rule

$$\begin{aligned} \text{reject } H_0 \quad & \text{if p-value} < \alpha \\ \text{do not reject } H_0 \quad & \text{if p-value} \geq \alpha \end{aligned}$$

The most commonly tested hypotheses are $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ for $j = 1, \dots, n$, which is the t-test for individual significance of the j-th variable.

11.4 Regression outputs

Let's consider the dataset `CPSdata` from the `statisticsdata` package and regress log-wages on education, experience, squared experience, and a randomly generated variable.

```
library(statisticsdata)
set.seed(42)
randomdata = rnorm(dim(CPSdata)[1]) ## generate a random regressor
fit = lm(
  log(wage) ~ education + experience + I(experience^2) + randomdata,
  data = CPSdata)
```

The `lm`-summary output

The `summary` output of `R` shows a typical regression output table. It consists of

(i) the **OLS coefficients** $\hat{\beta}_j$ in column 1,
(ii) the **classical standard errors** $se(\hat{\beta}_j)$ in column 2,
(iii) the **t-ratios** for the hypothesis $H_0 : \beta_j = 0$ in column 3, which is $T_0 = \hat{\beta}_j / se(\hat{\beta}_j)$,
(iv) the **p-values** for the two-sided t-tests $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ in column 4.

```
summary(fit)
```

```
Call:
lm(formula = log(wage) ~ education + experience + I(experience^2) +
    randomdata, data = CPSdata)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0782  -0.3134   0.0095   0.3391   2.8193

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.9371641  0.0164574  56.945   <2e-16 ***
education        0.1125763  0.0009752 115.436   <2e-16 ***
experience       0.0362323  0.0008074  44.874   <2e-16 ***
I(experience^2) -0.0005772  0.0000167 -34.556   <2e-16 ***
randomdata      -0.0004240  0.0026086  -0.163    0.871
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5905 on 50737 degrees of freedom
Multiple R-squared:  0.2368,    Adjusted R-squared:  0.2367
F-statistic:  3936 on 4 and 50737 DF,  p-value: < 2.2e-16
```

It also returns the **residual standard error** $s = SER$, the **degrees of freedom** $n - k$, the **R-squared** and the **adjusted R-squared**, and the **overall F-statistic** along with the p-value of the overall F-test (we will discuss the F-test below).

Regression outputs in economic journals

Beautiful LaTeX and HTML output tables can be produced with the `stargazer` package and function:

```
library(stargazer)
stargazer(fit, header=FALSE, type='html')
```

This is the most common style of regression output found in econometric journals. The constant is reported last. T-ratios and p-values are usually not shown because they can be computed from the estimate and standard error if needed ($T_0 = \hat{\beta}_j / se(\hat{\beta}_j)$).

Table 11.1

	Dependent variable:
	log(wage)
education	0.113^{***}
	(0.001)
experience	0.036^{***}
	(0.001)
I(experience^2)	-0.001^{***}
	(0.00002)
randomdata	-0.0004
	(0.003)
Constant	0.937^{***}
	(0.016)
Observations	50,742
R^2	0.237
Adjusted R^2	0.237
Residual Std. Error	0.590 (df = 50737)
F Statistic	$3{,}935.634^{***}$ (df = 4; 50737)
Note:	*p<0.1; **p<0.05; ***p<0.01

The test decision for the two-sided t-test for $H_0 : \beta_j = 0$ (test for individual significance) can be reached by inspecting the number of stars shown. Three stars *** indicate that the t-test rejects at $\alpha = 0.01$, two stars ** indicate a rejection at $\alpha = 0.05$ but not at $\alpha = 0.01$, and one star * indicates a rejection at $\alpha = 0.1$ but not at $\alpha = 0.05$.

As expected, H_0 cannot be rejected for the variable *randomdata*, since it was generated independently of *log(wage)*. All other variables have three stars. We say that the variables are highly significant, i.e. they show a strong correlative relation conditional on the other variables.

Note that `summary` has a different style and uses three stars for $\alpha = 0.001$, one star for $\alpha = 0.05$, and `.` for $\alpha = 0.1$ (always check the table notes).

Stargazer can display different regression outputs in a nice side-by-side view:

```
fit2 = lm(log(wage) ~ education + experience + I(experience^2), data=CPSdata)
fit3 = lm(log(wage) ~ education, data=CPSdata)
fit4 = lm(log(wage) ~ 1, data=CPSdata)
stargazer(fit, fit2, fit3, header=FALSE, type='html')
```

To customize the table to your needs, `stargazer` offers many options (have a look at the documentation by typing `?stargazer`).

11.5 Multiple testing

Consider the usual two-sided t-tests for the hypotheses $H_0 : \beta_1 = 0$ (test1) and $H_0 : \beta_2 = 0$ (test2).

Each test on its own is a valid hypothesis test of size α. However, applying these tests one after the other leads to a **multiple testing problem**. The probability of falsely rejecting the joint hypothesis

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \quad \text{vs.} \quad H_1 : \text{not } H_0$$

is too large. "Not H_0" means "$\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both".

To see this, suppose that, for simplicity, the t-statistics $\hat{\beta}_1/se(\hat{\beta}_1)$ and $\hat{\beta}_2/se(\hat{\beta}_2)$ are independent, which implies that the test decisions of the two tests are independent.

$$\begin{aligned}
&P(\text{both tests do not reject} \mid H_0 \text{ is true}) \\
&= P(\{\text{test1 does not reject}\} \cap \{\text{test2 does not reject}\} \mid H_0 \text{ is true}) \\
&= P(\text{test1 does not reject} \mid H_0) \cdot P(\text{test2 does not reject} \mid H_0) \\
&= (1-\alpha)^2 = \alpha^2 - 2\alpha + 1
\end{aligned}$$

The size of the combined test is larger than α:

$$\begin{aligned}
&P(\text{at least one test rejects} \mid H_0 \text{ is true}) \\
&= 1 - P(\text{both tests do not reject} \mid H_0 \text{ is true}) \\
&= 1 - (\alpha^2 - 2\alpha + 1) = 2\alpha - \alpha^2 = \alpha(2 - \alpha) > \alpha
\end{aligned}$$

If the two test statistics are dependent, then the probability of at least one of the tests falsely rejecting depends on their correlation and will also exceed α.

Therefore, another rejection rule must be found for repeated t-tests.

11.6 Exact F-tests

Suppose we want to test that the last q regression coefficients equal zero. We have q conditions to test:

$$H_0 : \begin{pmatrix} \beta_{k-q+1} \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

For instance, we would like to test in our first regression whether all coefficients except the constant are zero ($q = 4$), i.e.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.$$

$\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = \sum_{i=1}^{n} \hat{u}_i^2$ is the residual sum of squares when using all regressors (unrestricted LS estimation).

```
sum(residuals(fit)^2)
```

```
[1] 17690.81
```

Let $\tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}$ be the residual sum of squares when using only the first $k - q$ regressors (restricted LS estimation). In our case, we use only the constant.

```
sum(residuals(fit4)^2)
```

```
[1] 23179.86
```

The **F-statistic** is

$$F = \frac{(\tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}} - \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})/q}{(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})/(n-k)}$$

```
n = length(fit$fitted.values)
k = fit$rank
q = fit$rank - fit4$rank
## n, k, and q
c(n,k,q)
```

```
[1] 50742     5     4
```

```
numerator = (sum(residuals(fit4)^2) -  sum(residuals(fit)^2))/q
denominator = sum(residuals(fit)^2)/(n-k)
## F-statistic
numerator/denominator
```

```
[1] 3935.634
```

Note that the number coincides with the F-statistic in the regression output. We call this F-statistic the **overall F-statistic** or the F-statistic for overall significance because we compare our model with the intercept-only model.

The null-distribution of the F-statistic in the normal regression model is the F-distribution with q degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator:

$$F \sim \frac{\chi^2_q/q}{\chi^2_{n-k}/(n-k)} = F_{q,n-k}$$

F-test decision rule:

"*Reject H_0 if the F-statistic exceeds the $(1-\alpha)$ quantile of the $F_{q;n-k}$ distribution.*"

Table 11.2: 0.95-quantiles of the $F_{m,r}$-distribution

r / m	1	2	3	4	5	6	7	8
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82

r / m	1	2	3	4	5	6	7	8
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95
$\to \infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94

The F-test also allows for a general hypothesis form

$$H_0 : \boldsymbol{R}'\beta = \boldsymbol{c},$$

where $\boldsymbol{R}$ is a $k \times q$ matrix with $\text{rank}(\boldsymbol{R}) = q$ and $\boldsymbol{c}$ is a $q \times 1$ vector. Then, the F-statistic has the alternative direct representation

$$F = \frac{1}{q}(\boldsymbol{R}'\hat{\beta} - \boldsymbol{c})'(\boldsymbol{R}'\widehat{\boldsymbol{V}}\boldsymbol{R})^{-1}(\boldsymbol{R}'\hat{\beta} - \boldsymbol{c}),$$

where $\widehat{\boldsymbol{V}} = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the classical covariance matrix estimator.

Consider the linear regression with $k = 3$:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i$$

Example with $q = 2$: The hypothesis $H_0 : (\beta_2 = 0 \text{ and } \beta_3 = 0)$ is translated into $H_0 : \boldsymbol{R}'\beta = \boldsymbol{c}$ with

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Example with $q = 1$: The hypothesis $H_0 : \beta_2 + \beta_3 = 1$ is translated into $H_0 : \boldsymbol{R}'\beta = \boldsymbol{c}$ with

$$\boldsymbol{R} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{c} = \begin{pmatrix} 1 \end{pmatrix}$$

The `car` package provides the `linearHypothesis` function

```r
library(car)
linearHypothesis(fit, c("education = 0.11", "randomdata = 0"))
```

```
Linear hypothesis test

Hypothesis:
education = 0.11
randomdata = 0

Model 1: restricted model
Model 2: log(wage) ~ education + experience + I(experience^2) + randomdata

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1  50739 17693
2  50737 17691  2    2.4433 3.5036 0.0301 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hypothesis $H_0 : \beta_2 = 0.11, \beta_5 = 0$ cannot be rejected at the 0.01 significance level, but it can be rejected at the 0.05 significance level.

11.7 Additional reading

- Stock and Watson (2019), Sections 5, 7, 18, 19
- Hansen (2022b), Sections 5
- Davidson and MacKinnon (2004), Section 4, 5

11.8 R-codes

statistics-sec11.R

12 Robust Inference

12.1 Heteroskedasticity-robust standard errors

In the previous section, we discussed that exact inferential methods can be derived under homoskedasticity (A5) and normality (A6). These assumptions are quite restrictive for most practical applications. Let us now focus on the heteroskedastic regression model.

OLS Assumptions for the Heteroskedastic Model

The variables Y_i and $X_i = (1, X_{i2}, \dots, X_{ik})'$ satisfy the linear regression equation

$$Y_i = X_i'\beta + u_i, \quad i = 1, \dots, n,$$

which has the matrix representation $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{u}$.

For each $i = 1, \dots, n$ we assume

- (A1) **conditional mean independence**: $E[u_i \mid X_i] = 0$
- (A2) **random sampling**: (Y_i, X_i') are i.i.d. draws from their joint population distribution
- (A3) **large outliers unlikely**: $0 < E[Y_i^4] < \infty$, $0 < E[X_{ij}^4] < \infty$ for all $j = 1, \dots, k$
- (A4) **no perfect multicollinearity**: $\boldsymbol{X}$ has full column rank

Recall that the conditional covariance matrix is

$$\begin{aligned} Var[\hat{\beta} \mid \boldsymbol{X}] &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\sum_{i=1}^n \sigma_i^2 X_i X_i'\right) \left(\sum_{i=1}^n X_i X_i'\right)^{-1}. \end{aligned}$$

The standard deviation of the j-th coefficient estimate is

$$sd(\hat{\beta}_j \mid \boldsymbol{X}) = \sqrt{\left[\left(\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\sum_{i=1}^n \sigma_i^2 X_i X_i'\right) \left(\sum_{i=1}^n X_i X_i'\right)^{-1}\right]_{jj}}$$

It is not possible to standardize $\hat{\beta}_j$ using $sd(\hat{\beta}_j \mid \boldsymbol{X})$ in practice because the conditional variances $\sigma_i^2 = Var[u_i \mid X_i]$ are unknown.

Estimating σ_i^2 consistently for every $i = 1, \dots, n$ separately is impossible. However, it turns out that that the weighed average $n^{-1} \sum_{i=1}^n \sigma_i^2 X_i X_i'$ can be estimated consistently if we use the squared residuals $\hat{u}_i^2$ as a proxy for the conditional variances σ_i^2:

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 X_i X_i' \xrightarrow{p} \frac{1}{n} \sum_{i=1}^n \sigma_i^2 X_i X_i'$$

Therefore, a consistent covariance matrix estimator under heteroskedasticity is

$$\widehat{Var}_{HC0}[\hat{\beta}] = \left(\sum_{i=1}^n X_i X_i' \right)^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 X_i X_i' \right) \left(\sum_{i=1}^n X_i X_i' \right)^{-1},$$

and the corresponding standard error for the j-th coefficient estimate is

$$se_{HC0}(\hat{\beta}_j) = \sqrt{\left[\left(\sum_{i=1}^n X_i X_i' \right)^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 X_i X_i' \right) \left(\sum_{i=1}^n X_i X_i' \right)^{-1} \right]_{jj}}.$$

It satisfies $se_{HC0}(\hat{\beta}_j) / sd(\hat{\beta}_j \mid \boldsymbol{X}) \xrightarrow{p} 1$, and, together with the asymptotic normality result in Equation 10.4, we obtain

$$Z_{j,HC0} = \frac{\hat{\beta}_j - \beta_j}{se_{HC0}(\hat{\beta}_j)} \xrightarrow{D} \mathcal{N}(0,1),$$

which is the heteroskedasticity-robust t-ratio.

Unfortunately, it is not possible to derive the exact distribution of $Z_{j,HC0}$ even if (A6) is true or if (A5) and (A6) is true. However, the large sample approximation of $Z_{j,HC0}$ by a standard normal distribution is typically quite accurate, even for small samples.

We use the label **HC0** for this version of heteroskedasticity-robust standard errors. There are many other options (HC1–HC5). **HC1** is the version that is implemented as the default standard error in many software packages (e.g., the "r" option in Stata), and is a bias-corrected version of HC0:

$$se_{HC1}(\hat{\beta}_j) = \sqrt{\frac{n}{n-k}} \cdot se_{HC0}(\hat{\beta}_j).$$

Another version is **HC3**, which is based on leave-one-out residuals:

$$se_{HC3}(\hat{\beta}_j) = \sqrt{\left[\left(\sum_{i=1}^n X_i X_i' \right)^{-1} \left(\sum_{i=1}^n \left(\frac{\hat{u}_i}{1 - h_{ii}} \right)^2 X_i X_i' \right) \left(\sum_{i=1}^n X_i X_i' \right)^{-1} \right]_{jj}},$$

where h_{ii} is the i-th leverage value. The idea is similar to the leave-one-out R-squared, where influential observations get much higher weights for standardization to account for their impact on the coefficient estimate. The corresponding full covariance matrix estimate is

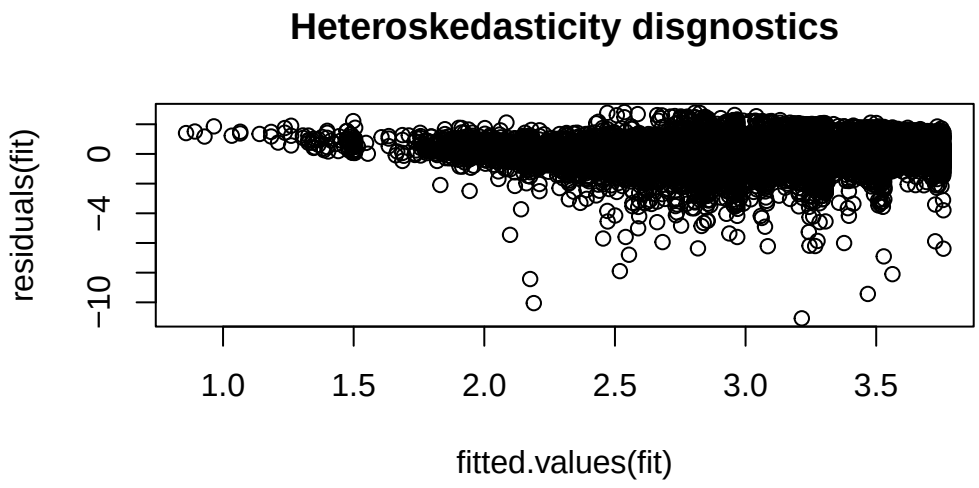
$$\widehat{Var}_{HC3}[\hat{\beta}] = \left(\sum_{i=1}^n X_i X_i' \right)^{-1} \left(\sum_{i=1}^n \left(\frac{\hat{u}_i}{1 - h_{ii}} \right)^2 X_i X_i' \right) \left(\sum_{i=1}^n X_i X_i' \right)^{-1}.$$

Recent research indicates that **HC3** should be the preferred choice for conducting inference, but other HC-versions perform similarly well. Classical standard errors

$$se(\hat{\beta}_j) = \sqrt{s^2 E[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}} = \sqrt{\left(\frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2\right)\left[\left(\sum_{k=1}^{n}X_k X_k'\right)^{-1}\right]_{jj}}$$

should only be used if you have very good reasons that your error terms are homoskedastic.

```
library(statisticsdata)
fit = lm(log(wage) ~ education + experience + I(experience^2), data = CPSdata)
## Heteroscedasticity diagnostics:
plot(fitted.values(fit), residuals(fit), main = "Heteroskedasticity disgnostics")
library(sandwich)
```




```
HOM = sqrt(diag(vcovHC(fit, "const")))
HC0 = sqrt(diag(vcovHC(fit, "HC0")))
HC1 = sqrt(diag(vcovHC(fit, "HC1")))
HC3 = sqrt(diag(vcovHC(fit, "HC3")))
library(tidyverse)
library(kableExtra)
tibble("Variable" = names(coefficients(fit)), "classical SE" = HOM, HC0, HC1, HC3) |> kbl(al
```

Variable	classical SE	HC0	HC1	HC3
(Intercept)	0.0164572	0.0181872	0.0181879	0.0181919
education	0.0009752	0.0010870	0.0010870	0.0010872
experience	0.0008074	0.0008851	0.0008851	0.0008856
I(experience^2)	0.0000167	0.0000194	0.0000194	0.0000194

The HC-robust standard errors are all quite similar. The classical standard errors differ from the robust obes and should not be used since homoskedasticity might not hold. In practice, there is no good reason not to use HC standard errors, even if there is evidence for homoskedasticity.

12.2 Robust confidence intervals

A heteroskedasticity-robust asymptotic $1-\alpha$ confidence interval for β_j is

$$I_{1-\alpha}^{HC} = [\hat{\beta}_j - z_{(1-\frac{\alpha}{2})} se_{HC}(\hat{\beta}_j); \hat{\beta}_j + z_{(1-\frac{\alpha}{2})} se_{HC}(\hat{\beta}_j)],$$

where $se_{HC}(\hat{\beta}_j)$ is any heteroskedasticity-robust standard error.

Instead of $z_{(1-\frac{\alpha}{2})}$ you can also use $t_{(n-k;1-\frac{\alpha}{2})}$ since $\lim_{n\to\infty} t_{(n-k;1-\frac{\alpha}{2})} = z_{(1-\frac{\alpha}{2})}$. However, there is no theoretical justification for preferring t-quantiles over standard normal quantiles.

For our CPSdata regression, we obtain the following robust confidence intervals:

```
cbind(
  coefficients(fit)-HC3*qnorm(0.975),
  coefficients(fit)+HC3*qnorm(0.975)
)
```

```
                       [,1]          [,2]
(Intercept)      0.9015028254  0.9728138849
education        0.1104457789  0.1147075373
experience       0.0344968746  0.0379682407
I(experience^2) -0.0006151078 -0.0005391928
```

Classical confidence intervals are slightly too small:

```
confint(fit)
```

```
                      2.5 %        97.5 %
(Intercept)      0.9049020666  0.9694146436
education        0.1106652300  0.1144880862
experience       0.0346500248  0.0378150905
I(experience^2) -0.0006098851 -0.0005444154
```

12.3 Robust t-tests

The robust version of the two-sided t-test with

$$H_0 : \beta_j = \beta_j^0 \quad vs. \quad H_1 : \beta_j \neq \beta_j^0$$

has the test statistic

$$T_0^{HC} = \frac{\hat{\beta}_j - \beta_j^0}{se_{HC}(\hat{\beta}_j)}.$$

The test decision rule is

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } |T_0^{HC}| \leq z_{(1-\frac{\alpha}{2})}, \\ \text{reject } H_0 \quad & \text{if } |T_0^{HC}| > z_{(1-\frac{\alpha}{2})}. \end{aligned}$$

The null distribution CDF is $F_0 = \Phi$, and the p-value for the two-sided t-test is

$$p\text{-value} = 2(1 - \Phi(|T_0^{HC}|)).$$

Regression outputs with robust standard errors can be created with the following command:

```
library(lmtest)
coeftest(fit, vcov. = vcovHC(fit, type = 'HC3'), df = Inf)
```

```
z test of coefficients:

                   Estimate  Std. Error z value  Pr(>|z|)
(Intercept)      9.3716e-01  1.8192e-02  51.515 < 2.2e-16 ***
education        1.1258e-01  1.0872e-03 103.547 < 2.2e-16 ***
experience       3.6233e-02  8.8557e-04  40.914 < 2.2e-16 ***
I(experience^2) -5.7715e-04  1.9366e-05 -29.802 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can also include robust standard errors in stargazer outputs:

```
library(stargazer)
HC3 = sqrt(diag(vcovHC(fit, "HC3")))
stargazer(fit, header=FALSE, type='html', se = list(HC3), omit.stat = "f")
```

Table 12.1

	Dependent variable:
	log(wage)
education	0.113^{***}
	(0.001)
experience	0.036^{***}
	(0.001)
I(experience^2)	-0.001^{***}
	(0.00002)
Constant	0.937^{***}
	(0.018)
Observations	50,742
R^2	0.237
Adjusted R^2	0.237
Residual Std. Error	0.590 (df = 50738)
Note:	*p<0.1; **p<0.05; ***p<0.01

12.4 Robust F-tests

For a robust F-test of the form

$$H_0 : \boldsymbol{R}'\beta = \boldsymbol{c}, \quad vs. \quad H_1 : \boldsymbol{R}'\beta \neq \boldsymbol{c},$$

the robust F-statistic

$$F_0 = \frac{1}{q}(\boldsymbol{R}'\hat{\beta} - \boldsymbol{c})'(\boldsymbol{R}'(\widehat{Var}_{HC3}[\hat{\beta}])\boldsymbol{R})^{-1}(\boldsymbol{R}'\hat{\beta} - \boldsymbol{c})$$

can be used.

The asymptotic critical value is the $1-\alpha$ quantile of the $F_{q,\infty}$ distribution. Note that the $F_{q,\infty}$ distribution coincides with χ^2_q/q, so you can use the $1-\alpha$ quantile of χ^2_q divided by q.

If c is the $1-\alpha$ critical value, the asymptotic size-α F-test has the following decision rule:

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } F_0 \leq c, \\ \text{reject } H_0 \quad & \text{if } F_0 > c. \end{aligned}$$

12.5 Autocorrelation-robust standard errors

OLS Assumptions for the Dynamic Regression Model

The variables Y_i and $X_i = (1, X_{i2}, \dots, X_{ik})'$ satisfy the linear regression equation

$$Y_i = X_i'\beta + u_i, \quad i = 1, \dots, n,$$

which has the matrix representation $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{u}$.

For each $i = 1, \dots, n$ we assume

- (A1) **conditional mean independence**: $E[u_i \mid X_i] = 0$
- (A2b) **weak dependence**: (Y_i, X_i') is stationary short-memory, and (Y_i, X_i') and $(Y_{i-\tau}, X_{i-\tau}')$ become independent as τ gets large
- (A3) **large outliers unlikely**: $0 < E[Y_i^4] < \infty$, $0 < E[X_{ij}^4] < \infty$ for all $j = 1, \dots, k$
- (A4) **no perfect multicollinearity**: $\boldsymbol{X}$ has full column rank

Since the observations are not independent by (A2b), the covariance matrix $Var[\hat{\beta} \mid \boldsymbol{X}]$ has a more complicated structure.

Similarly, to the long-run variance in the sample mean case, the term $\sum_{i=1}^{n} \sigma_i^2 X_i X_i'$ must be replaced by its long-run counterpart. It turns out that the following matrix is a consistent estimate for $Var[\hat{\beta} \mid \boldsymbol{X}]$ for dynamic linear regressions:

$$\widehat{Var}_{HAC}[\hat{\beta}] = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \boldsymbol{V}_{HAC} \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1}$$

where

$$\boldsymbol{V}_{HAC} = \sum_{i=1}^{n} \hat{u}_i^2 X_i X_i' + \sum_{\tau=1}^{\ell_n - 1} \frac{\ell_n - \tau}{\ell_n} \hat{u}_i \hat{u}_{i-\tau} (X_i X_{i-\tau}' + X_{i-\tau} X_i'),$$

and ℓ_n is some suitable truncation parameter. A common rule of thumb is to select

$$\ell_n = \lfloor 0.75 \cdot n^{1/3} \rfloor$$

We use the label **HAC** because the covariance matrix provides heteroskedasticity and autocorrelation-robust standard errors, which are given by

$$se_{HAC}(\hat{\beta}_j) = \sqrt{[\widehat{Var}_{HAC}[\hat{\beta}]]_{jj}}.$$

For linear regressions with time series data you should use $se_{HAC}(\hat{\beta}_j)$ for confidence intervals and t-tests, and you can use $\widehat{Var}_{HAC}[\hat{\beta}]$ for F-tests.

For a given linear regression model object `lmobj`, the `NeweyWest(lmobj)` command from the `sandwich` package returns the HAC covariance matrix estimate.

12.6 Additional reading

- Stock and Watson (2019), Sections 5, 7, 18, 19
- Hansen (2022b), Sections 7
- Davidson and MacKinnon (2004), Section 4, 5, 6

12.7 R-codes

statistics-sec12.R