Statistics for Data Analytics

Data Analytics I

Jun.-Prof. Dr. Sven Otto

Last updated: November 5, 2025

Table of contents

Organization of the Course

Statistics for Data Analytics is a graduate-level introductory course in econometrics, focusing on estimation and inference in linear models, with practical illustrations in `R`.

Timetable

See KLIPS for a detailed schedule.

Note: In the first session on 16 October 2025, there will be a lecture instead of exercises. The final lecture will take place on 21 November 2025.

Lecture Material

- This online script and its pdf version
- eWhiteboard lecture and eWhiteboard exercises
- Problemsets and Rscripts
- ILIAS course

Literature

The script is self-contained. To prepare well for the exam, it's a good idea to study this script.

The course is based on James H. Stock and Mark W. Watson's **Introduction to Econometrics (Fourth Edition)**. The Stock and Watson textbook is available for download: PDF by chapter (Uni Köln VPN connection required).

Further recommended textbooks are:

Day	Time	Lecture Hall	Session Type
Thursday	10:00-11:30	XII (Main Building)	Exercises
Thursday	12:00-13:30	XII (Main Building)	Lecture
Friday	10:00-11:30	XII (Main Building)	Lecture

- *Econometric Theory and Methods*, by Russell Davidson and James G. MacKinnon. PDF.
- *Probability and Statistics for Economists*, by Bruce E. Hansen
- *Econometrics*, by Bruce E. Hansen

Printed versions of the books are available from the university library.

Assessment

The course will be graded by a 90-minute exam. For detailed information please visit the ILIAS course.

Communication

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

Important Dates

Registration deadline exam 1	November 13, 2025
Exam 1	November 27, 2025
Registration deadline exam 2	January 27, 2026
Exam 2 (alternate date)	February 10, 2026

Please register for the exam on time. If you miss the registration deadline, you will not be able to take the exam.

R-Packages

The best way to learn statistical methods is to program and apply them yourself. Throughout this lecture script, we will use the `R` programming language to illustrate how econometric methods are applied in practice.

For those of you who are new to `R` and want to learn more about it, here's an introductory tutorial that contains many valuable resources: rintro.svenotto.com. I also recommend the interactive R package SWIRL, which offers an excellent way to learn directly within the R environment.

To run the R code of the lecture script, you will need to install some additional packages via the command `install.packages()`:

```
install.packages(c("AER", "fixest", "moments", "dynlm", "modelsummary", "scatterplot3d", "re
```

Some further datasets are contained in my package TeachData, which is available in a GitHub repository. It can be installed using the following command:

```
remotes::install_github("ottosven/TeachData")
```

1 Data

1.1 Data Structures

Univariate Datasets

A **univariate dataset** consists of a sequence of observations:

$$Y_1, \dots, Y_n.$$

These n observations form a **data vector**:

$$\boldsymbol{Y} = (Y_1, \dots, Y_n)'.$$

Example: Survey of six individuals on their hourly earnings. Data vector:

$$\boldsymbol{Y} = \begin{pmatrix} 10.40 \\ 18.68 \\ 12.44 \\ 54.73 \\ 24.27 \\ 24.41 \end{pmatrix}.$$

Multivariate Datasets

Typically, we have data on more than one variable, such as years of education and gender. Categorical variables are often encoded as **dummy variables** (also called indicator variables), which are binary variables. The female dummy variable is defined as:

$$D_i = \begin{cases} 1 & \text{if person } i \text{ is female,} \\ 0 & \text{otherwise.} \end{cases}$$

person	wage	education	female
1	10.40	12	0

person	wage	education	female
2	18.68	16	0
3	12.44	14	1
4	54.73	18	0
5	24.27	14	0
6	24.41	12	1

A **k-variate dataset** (or multivariate dataset) is a collection of n observations on k variables (i.e., n observation vectors of length k):

$$\boldsymbol{X}_1, \dots, \boldsymbol{X}_n.$$

The i-th vector contains the data on all k variables for individual i:

$$\boldsymbol{X}_i = (X_{i1}, \dots, X_{ik})'.$$

Thus, X_{ij} represents the value for the j-th variable of individual i. The full k-variate dataset is structured in the $n \times k$ **data matrix** $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \begin{pmatrix} X_{11} & \dots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{nk} \end{pmatrix}$$

The i-th row in $\boldsymbol{X}$ corresponds to the values from $\boldsymbol{X}_i$. Since $\boldsymbol{X}_i$ is a column vector, we write rows of the data matrix as $\boldsymbol{X}_i'$ (its transpose), which is a row vector. Note that $\boldsymbol{X} \in \mathbb{R}^{n \times k}$, $\boldsymbol{X}_i \in \mathbb{R}^{k \times 1}$, and $\boldsymbol{X}_i' \in \mathbb{R}^{1 \times k}$.

The data matrix for our example is:

$$\boldsymbol{X} = \begin{pmatrix} 10.40 & 12 & 0 \\ 18.68 & 16 & 0 \\ 12.44 & 14 & 1 \\ 54.73 & 18 & 0 \\ 24.27 & 14 & 0 \\ 24.41 & 12 & 1 \end{pmatrix}$$

with data vectors:

$$\boldsymbol{X}_1 = \begin{pmatrix} 10.40 \\ 12 \\ 0 \end{pmatrix}$$

$$\boldsymbol{X}_2 = \begin{pmatrix} 18.68 \\ 16 \\ 0 \end{pmatrix}$$

$$\boldsymbol{X}_3 = \begin{pmatrix} 12.44 \\ 14 \\ 1 \end{pmatrix}$$

$$\vdots$$

Matrix Algebra

Vector and matrix algebra provide a compact mathematical representation of multivariate data and an efficient framework for analyzing and implementing statistical methods. We will use matrix algebra frequently throughout this course.

To refresh or enhance your knowledge of matrix algebra, consult the following resources:

Crash Course on Matrix Algebra:

matrix.svenotto.com (in particular Sections 1-3)
Section 19.1 of the Stock and Watson textbook also provides a brief overview of matrix algebra concepts.

1.2 Datasets in R

R is a vector-based statistical programming language and is therefore particularly suitable for handling data in tabular or matrix form. Matrix algebra is particularly useful when working with real data in R.

R's most common data structure for tabular data is the data frame (`data.frame`). Like the data matrix $\boldsymbol{X}$ we defined earlier, it organizes data with variables as columns and observations as rows.

CA Schools Data

Let's load the `CASchools` dataset from the `AER` package ("Applied Econometrics with R"). You can install the package with the command `install.packages("AER")`.

```
data(CASchools, package = "AER")
```

The dataset is used throughout Sections 4–8 of Stock and Watson's textbook *Introduction to Econometrics*. It was collected in 1998 and captures California school characteristics including test scores, teacher salaries, student demographics, and district-level metrics.

Variable	Description	Variable	Description
district	District identifier	lunch	% receiving free meals
school	School name	computer	Number of computers
county	County name	expenditure	Spending per student ($)
grades	Through 6th or 8th	income	District avg income ($000s)
students	Total enrollment	english	Non-native English (%)
teachers	Teaching staff	read	Average reading score
calworks	% CalWorks aid	math	Average math score

The Environment pane in RStudio's top-right corner displays all objects currently in your workspace, including the `CASchools` dataset. You can click on it to explore its contents.

The `head()` function displays the first few rows of a dataset, giving you a quick preview of its content.

```
head(CASchools)
```

```
  district                          school  county grades students teachers
1    75119              Sunol Glen Unified Alameda  KK-08      195    10.90
2    61499            Manzanita Elementary   Butte  KK-08      240    11.15
3    61549     Thermalito Union Elementary   Butte  KK-08     1550    82.90
4    61457 Golden Feather Union Elementary   Butte  KK-08      243    14.00
5    61523        Palermo Union Elementary   Butte  KK-08     1335    71.50
6    62042         Burrel Union Elementary  Fresno  KK-08      137     6.40
  calworks   lunch computer expenditure    income   english  read  math
1   0.5102  2.0408       67    6384.911 22.690001  0.000000 691.6 690.0
2  15.4167 47.9167      101    5099.381  9.824000  4.583333 660.5 661.9
```

```
3   55.0323 76.3226      169    5501.955  8.978000 30.000002 636.3 650.9
4   36.4754 77.0492       85    7101.831  8.978000  0.000000 651.9 643.5
5   33.1086 78.4270      171    5235.988  9.080333 13.857677 641.8 639.9
6   12.3188 86.9565       25    5580.147 10.415000 12.408759 605.7 605.4
```

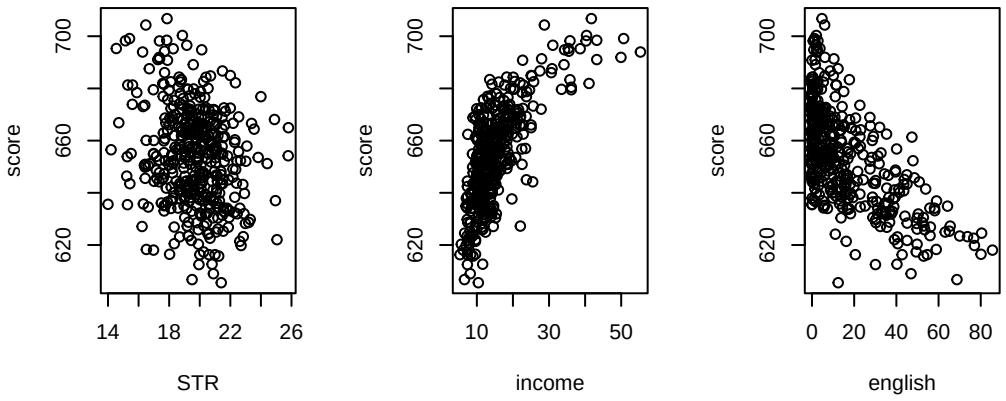
The variable `students` contains the total number of students enrolled in a school. It is the fifth variable in the dataset. To access the variable as a vector, you can type `CASchools[,5]` (the fifth column in your data matrix), `CASchools["students"]`, or simply `CASchools$students`.

We can easily add new variables to our data frame, for instance, the student-teacher ratio (the total number of students per teacher) and the average test score (average of the math and reading scores):

```
# compute student-teacher ratio and append it to CASchools
CASchools$STR = CASchools$students/CASchools$teachers
# compute test score and append it to CASchools
CASchools$score = (CASchools$read + CASchools$math)/2
```

Scatterplots provide further insights:

```
par(mfrow = c(1,3))
plot(score~STR, data = CASchools)
plot(score~income, data = CASchools)
plot(score~english, data = CASchools)
```



The option `par(mfrow = c(1,3))` allows you to display multiple plots side by side. Try what happens if you replace `c(1,3)` with `c(3,1)`.

CPS Data

Another dataset we will use in this course is the CPS dataset from Bruce Hansen's textbook *Econometrics*.

The Current Population Survey (CPS) is a monthly survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics, primarily used to measure the labor force status of the U.S. population.

- Dataset: cps09mar.txt
- Codebook: cps09mar_description.pdf

The dataset is available as a whitespace-separated text file, which can be loaded using `read.table()`.

```
url = "https://users.ssc.wisc.edu/~bhansen/econometrics/cps09mar.txt"
varnames = c("age", "female", "hisp", "education", "earnings", "hours",
             "week", "union", "uncov", "region", "race", "marital")
cps = read.table(url, col.names = varnames)
```

Let's create additional variables:

```
# wage per hour
cps$wage = cps$earnings/(cps$week * cps$hours)
# work experience (years since graduation)
cps$experience = pmax(cps$age - cps$education - 6,0)
# married dummy (see codebook for the categories)
cps$married = (cps$marital %in% c(1, 2, 3)) |> as.numeric()
# Black dummy (see codebook)
cps$Black = (cps$race %in% c(2, 6, 10, 11, 12, 15, 16, 19)) |> as.numeric()
# Asian dummy (see codebook)
cps$Asian = (cps$race %in% c(4, 8, 11, 13, 14, 16, 17, 18, 19)) |> as.numeric()
```

A person is considered married if the `marital` variable takes one of the following categories: 1, 2, or 3 (see the codebook above for more information). Note that `cps$marital %in% c(1, 2, 3)` is a logical expression with either `TRUE` or `FALSE` values. The command `as.numeric()` creates a dummy variable by translating `TRUE` to `1` and `FALSE` to `0`.

The pipe operator `|>` efficiently chains commands. It passes the output of one function as the input to another. For example, `cps$marital %in% c(1, 2, 3) |> as.numeric()` gives the same output as `as.numeric(cps$marital %in% c(1, 2, 3))`.

We will need the CPS dataset later, so it is a good idea to save the dataset to your computer:

```r
write.csv(cps, "cps.csv", row.names = FALSE)
```

This command saves the dataset to a file named `cps.csv` in your current working directory. It's best practice to use an R Project for your course work so that all files (data, scripts, outputs) are stored in a consistent and organized folder structure.

To read the data back into `R` later, just type `cps = read.csv("cps.csv")`.

1.3 Statistical Framework

Data are usually the result of a random experiment. The gender of the next person you meet, the daily fluctuation of a stock price, the monthly music streams of your favorite artist, the annual number of pizzas consumed - all of this information involves a certain amount of randomness.

We distinguish between:

- **Cross-sectional data**: observations on many units at (approximately) one point in time.
- **Time series data**: observations on one unit recorded over multiple time periods.
- **Panel data**: observations on many units recorded over multiple time periods.

In statistical sciences, we interpret a univariate dataset $Y_1, \dots, Y_n$ as a sequence of random variables. Similarly, a multivariate dataset $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is viewed as a sequence of random vectors.

Sampling refers to the process of obtaining data by drawing observations from a population, which is often considered infinite in statistical theory. An infinite population is a conceptual device representing all potential outcomes that could arise under the same conditions, not just the currently existing individuals.

For example, when modeling coin flips, the population includes every possible toss that could ever occur. When analyzing stock returns, the population includes all possible future price movements. When studying human height, the infinite population includes all current humans as well as all hypothetical humans who could exist under similar biological conditions. Formally, the infinite population corresponds to a probability distribution F and the sample is n i.i.d. draws from F.

Random sampling

Econometric methods require specific assumptions about sampling processes. The ideal approach for a cross-sectional study is simple random sampling, where each individual from the population has an equal chance of being selected independently.

This produces observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ that are both identically distributed (drawn from the same population) and independently drawn (as if drawn from an urn with replacement). We call these data **independent and identically distributed (i.i.d.)** or simply a **random sample**.

For example, when conducting a representative survey, the answers of the second randomly selected individual should not depend on the answers of the first randomly selected individual if the individuals are truly randomly selected from the population. A violation of the i.i.d. property is often a matter of data collection quality.

Clustered sampling

While i.i.d. sampling provides a clean theoretical foundation, real-world data sometimes exhibit clustering - where observations are naturally grouped or nested within larger units. This clustering leads to dependencies that violate the i.i.d. assumption:

In cross-sectional studies, clustering occurs when we collect data on individual units that belong to distinct groups. Consider a study on student achievement where researchers randomly select schools, then collect data from all students within those schools:

- Although schools might be selected independently, observations at the student level are dependent
- Students within the same school share common environments (facilities, resources, administration)
- They experience similar teaching quality and educational policies and they influence each other through peer effects and social interactions

For instance, if School A has an exceptional mathematics department, all students from that school may perform better in math tests compared to students with similar abilities in other schools.

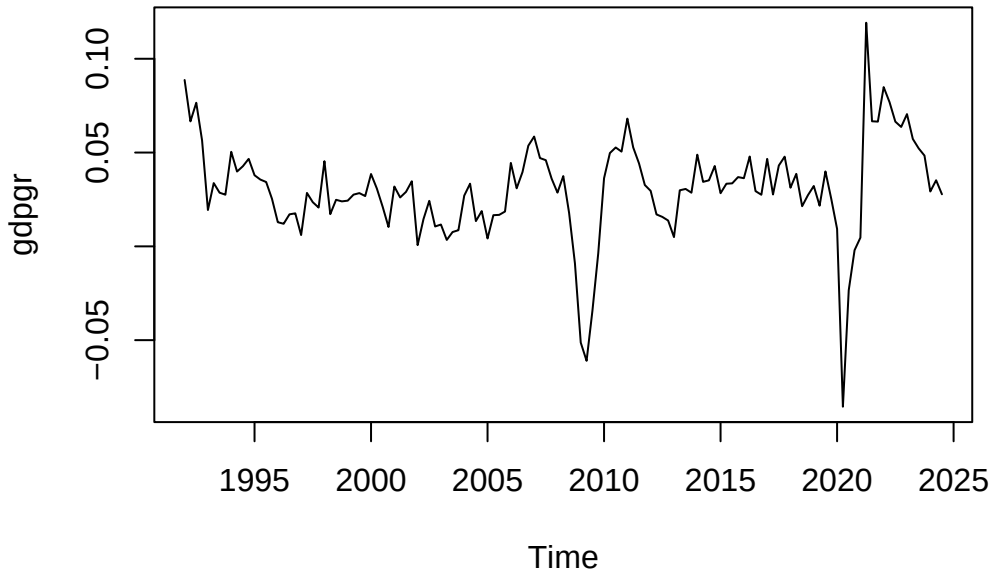
Panel data, by its very nature, introduces clustering across both cross-sectional units and time. If many randomly selected individuals are interviewed over many years, then the observations of two different individuals are independent but, for each individual, observations across different years are dependent due to persistent personal factors.

Time dependence

Time series and panel data are intrinsically not independent due to the sequential nature of the observations. We usually expect observations close in time to be strongly dependent and observations at greater temporal distances to be less dependent.

Consider the quarterly GDP growth rates for Germany in the dataset `gdpgr`. Unlike cross-sectional data where the ordering of observations is arbitrary, the chronological ordering in time series carries crucial information about the dependency structure.

```
library(TeachData)
plot(gdpgr)
```



1.4 R Code

statistics-sec01.R

2 Distribution

Probability Distribution

An event is a collection of different outcomes, typically in form of open, half-open, or closed intervals, or unions of multiple intervals.

The probability distribution F_Y assigns probabilities to all possible events of Y. The cumulative distribution function (CDF) fully characterizes the probability distribution:

Cumulative Distribution Function (CDF)

The CDF of a random variable Y is

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

Any nondecreasing right-continuous function $F_Y(a)$ with $\lim_{a \to -\infty} F_Y(a) = 0$ and $\lim_{a \to +\infty} F_Y(a) = 1$ defines a valid CDF.

For a more detailed introduction to probability theory, see my tutorial at probability.svenotto.com.

2.1 Discrete Random Variables

Coin Toss

Consider the coin toss binary random variable

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

The CDF for a fair coin is

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

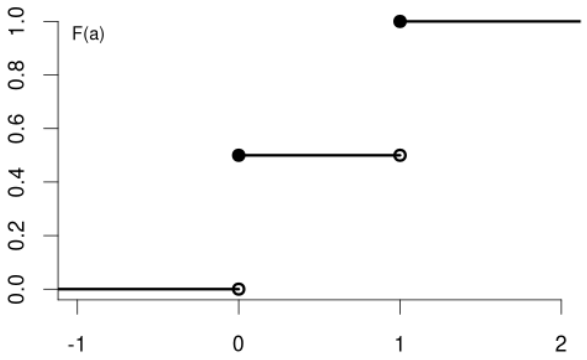


Figure 2.1: CDF of coin (discrete random variable)

with the following CDF plot:

A random variable with a CDF that has jumps and is flat between these jumps is called a **discrete random variable**.

Let $F_Y(a^-) = \lim_{\varepsilon \to 0, \varepsilon > 0} F_Y(a - \varepsilon)$ denote the left limit of F_Y at a.

That means, for the coin toss, we have point values $F_Y(0) = 0.5$ and $F_Y(1) = 1$ and the left-hand limits $F_Y(0^-) = 0$ and $F_Y(1^-) = 0.5$.

The **point probability** $P(Y = a)$ represents the size of the jump at a in the CDF $F_Y(a)$:

$$P(Y = a) = F_Y(a) - F_Y(a^-).$$

Because CDFs are right-continuous, jumps can only be seen when approaching a point a from the left.

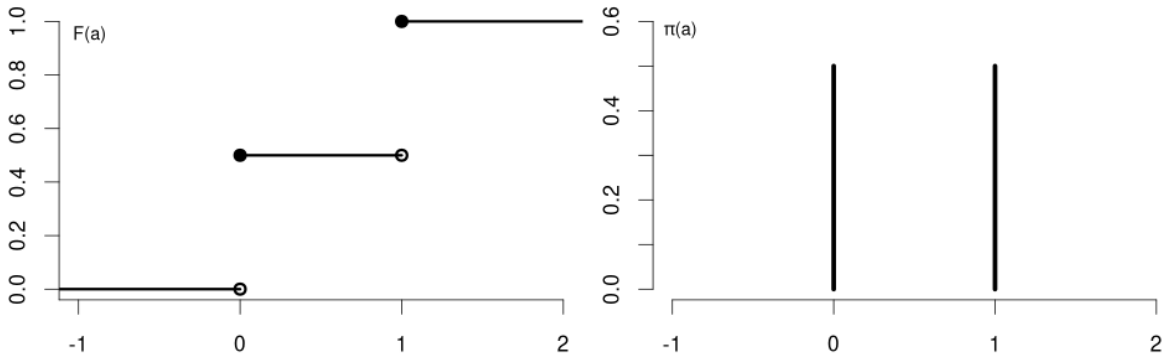
Probability Mass Function (PMF)

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

$$\pi_Y(a) := P(Y = a) = F_Y(a) - F_Y(a^-), \quad a \in \mathbb{R}.$$

The PMF of the *coin* variable is

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



(a) CDF of coin (b) PMF of coin

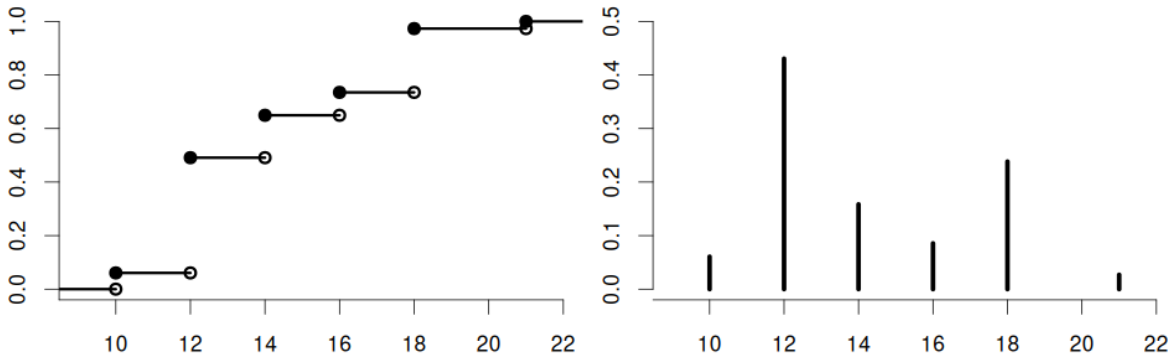
Figure 2.2: Coin variable: CDF (left) and PMF (right)

Years of Education

Suppose you conduct a survey where you ask a randomly selected person about their years of education, with the following answer options:

$$Y \in \{10, 12, 14, 16, 18, 21\}.$$

The *education* variable is a discrete random variable. It may have the following CDF and PMF:



(a) CDF of education (b) PMF of education

Figure 2.3: Education variable: CDF (left) and PMF (right)

Here, the PMF is

$$\pi_Y(a) = \begin{cases} 0.06 & \text{if } a = 10 \\ 0.43 & \text{if } a = 12 \\ 0.16 & \text{if } a = 14 \\ 0.08 & \text{if } a = 16 \\ 0.24 & \text{if } a = 18 \\ 0.03 & \text{if } a = 21 \\ 0 & \text{otherwise} \end{cases}$$

The **support** $\mathcal{Y}$ is the set of all values that Y can take with non-zero probability: $\mathcal{Y} = \{a \in \mathbb{R} : \pi_Y(a) > 0\}$.

For the coin variable, the support is $\mathcal{Y} = \{0, 1\}$, while for the education variable, the support is $\mathcal{Y} = \{10, 12, 14, 16, 18, 21\}$.

The sum of $\pi_Y(a)$ over the support is 1: $\sum_{a \in \mathcal{Y}} \pi_Y(a) = 1$.

For the tail probabilities we have the following rules:

- $P(Y \leq a) = F_Y(a)$
- $P(Y < a) = F_Y(a^-) = F_Y(a) - \pi_Y(a)$
- $P(Y > a) = 1 - F_Y(a)$
- $P(Y \geq a) = 1 - F_Y(a^-) = 1 - F_Y(a) + \pi_Y(a)$

For intervals (with $a < b$):

- $P(a < Y \leq b) = F_Y(b) - F_Y(a)$
- $P(a < Y < b) = F_Y(b^-) - F_Y(a)$
- $P(a \leq Y \leq b) = F_Y(b) - F_Y(a^-)$
- $P(a \leq Y < b) = F_Y(b^-) - F_Y(a^-)$

2.2 Continuous Random Variables

For a continuous random variable Y, the CDF $F_Y(a)$ has no jumps and is continuous. The left limits $F_Y(a^-)$ equal the point values $F_Y(a)$ for all a, which means that the point probabilities are zero:

$$P(Y = a) = F_Y(a) - F_Y(a^-) = 0.$$

Probability is distributed continuously over intervals. Unlike discrete random variables, which are characterized by both the PMF and the CDF, continuous variables have $\pi_Y(a) = P(Y = a) = 0$ for every point, so the PMF is not a useful concept here.

Instead, they are described by the probability density function (PDF), which serves as the continuous analogue of the PMF. If the CDF is differentiable, the PDF is given by its derivative:

Probability Density Function (PDF)

The **probability density function (PDF)** or simply **density function** of a continuous random variable Y is the derivative of its CDF:

$$f_Y(a) = \frac{d}{da} F_Y(a).$$

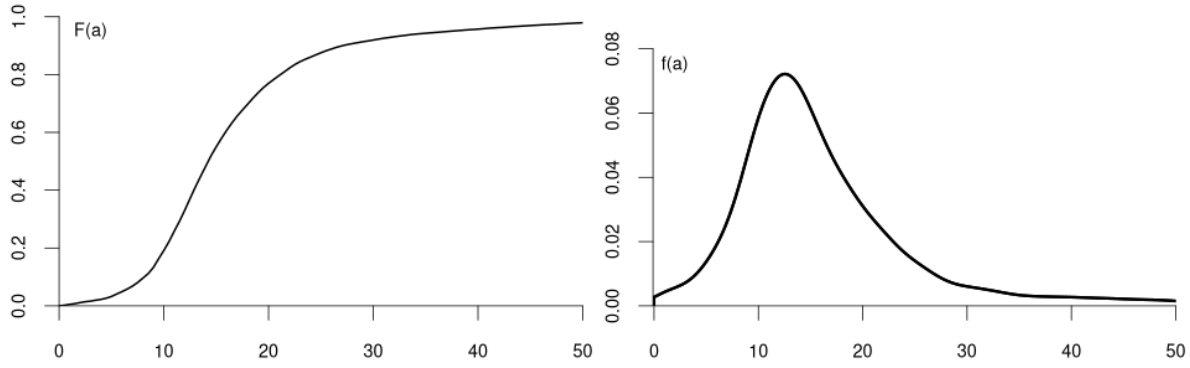
Conversely, the CDF can be obtained from the PDF by integration:

$$F_Y(a) = \int_{-\infty}^{a} f_Y(u)\, du$$

Wage per hour

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 set of possible results of such a random variable is a continuum of different wage levels.

The CDF and PDF of *wage* may have the following form:



(a) CDF of wage (b) PDF of wage

Figure 2.4: Wage variable: CDF (left) and PDF (right)

Basic Rules for Continuous Random Variables (with $a \leq b$):

- $P(Y = a) = \int_a^a f_Y(u)\, du = 0$

- $P(Y \leq a) = P(Y < a) = F_Y(a) = \int_{-\infty}^{a} f_Y(u)\,du$
- $P(Y > a) = P(Y \geq a) = 1 - F_Y(a) = \int_{a}^{\infty} f_Y(u)\,du$
- $P(a < Y < b) = F_Y(b) - F_Y(a) = \int_{a}^{b} f_Y(u)\,du$
- $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b) = P(a \leq Y < b)$

Unlike the PMF, which directly gives probabilities, the PDF does not represent probability directly. Instead, the probability is given by the area under the PDF curve over an interval.

It is important to note that for continuous random variables, the probability of any single point is zero. This is why, as shown in the last rule above, the inequalities (strict or non-strict) don't affect the probability calculations for intervals. This stands in contrast to discrete random variables, where the inclusion of endpoints can change the probability value.

2.3 Conditional Distribution

The distribution of *wage* may differ between men and women. Similarly, the distribution of *education* may vary between married and unmarried individuals. In contrast, the distribution of a *coin flip* should remain the same regardless of whether the person tossing the coin earns 15 or 20 EUR per hour.

The **conditional cumulative distribution function** (conditional CDF),

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

represents the distribution of a random variable Y given that another random variable Z takes a specific value b. It answers the question: "If we know that $Z = b$, what is the distribution of Y?"

For example, suppose that Y represents *wage* and Z represents *education*:

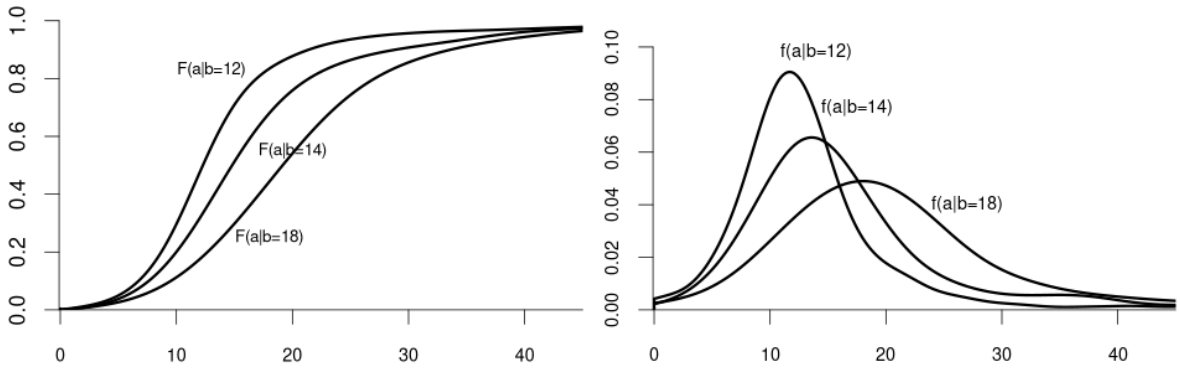
- $F_{Y|Z=12}(a)$ is the CDF of wages among individuals with 12 years of education.
- $F_{Y|Z=14}(a)$ is the CDF of wages among individuals with 14 years of education.
- $F_{Y|Z=18}(a)$ is the CDF of wages among individuals with 18 years of education.

Since *wage* is a continuous variable, its conditional distribution given any specific value of another variable is usually also continuous. The conditional density of Y given $Z = b$ is defined as the derivative of the conditional CDF:

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

We observe that the distribution of wage varies across different levels of education. For example, individuals with fewer years of education are more likely to earn less than 20 EUR per hour:

$$P(Y \leq 20|Z = 12) = F_{Y|Z=12}(20) > F_{Y|Z=18}(20) = P(Y \leq 20|Z = 18).$$



(a) Conditional CDFs of wage given education (b) Conditional PDFs of wage given education

Figure 2.5: Wage distributions conditional on education level

Because the conditional distribution of Y given $Z = b$ depends on the value b of Z, we say that the random variables Y and Z are **dependent random variables**.

Note that the conditional CDF $F_{Y|Z=b}(a)$ can only be defined for values of b in the support of Z.

We can also condition on more than one variable. Let Z_1 represent the labor market *experience* in years and Z_2 be the *female* dummy variable. The conditional CDF of Y given $Z_1 = b$ and $Z_2 = c$ is:

$$F_{Y|Z_1=b,Z_2=c}(a) = F_{Y|Z_1,Z_2}(a|b,c) = P(Y \leq a|Z_1 = b, Z_2 = c).$$

For example:

- $F_{Y|Z_1=10,Z_2=1}(a)$ is the CDF of wages among women with 10 years of experience.
- $F_{Y|Z_1=10,Z_2=0}(a)$ is the CDF of wages among men with 10 years of experience.

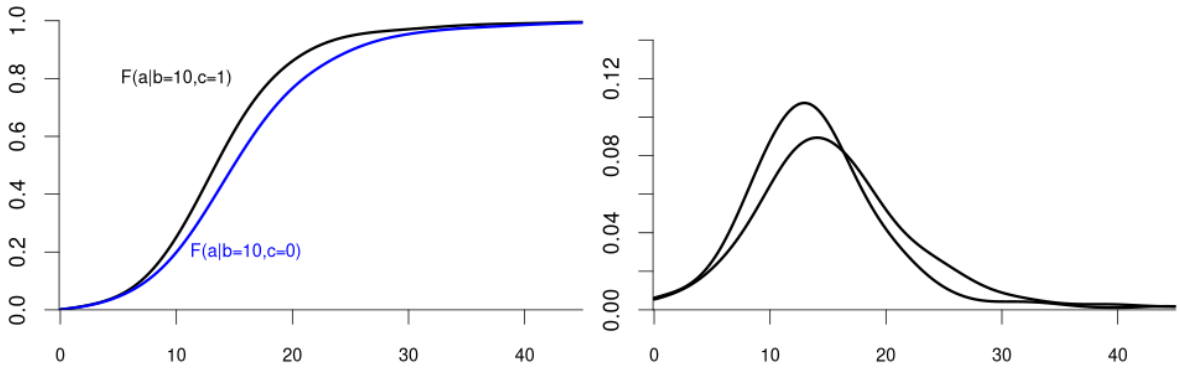
Clearly, the random variable Y and the random vector (Z_1, Z_2) are dependent.

More generally, we can condition on the event that a k-variate random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$ takes the value $\{\boldsymbol{Z} = \boldsymbol{b}\}$, i.e., $\{Z_1 = b_1, \ldots, Z_k = b_k\}$. The conditional CDF of Y given $\{\boldsymbol{Z} = \boldsymbol{b}\}$ is

$$F_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a) = F_{Y|Z_1=b_1,\ldots,Z_k=b_k}(a).$$

The variable of interest, Y, can also be discrete. Then, any conditional CDF of Y is also discrete. Below is the conditional CDF of *education* given the *married* dummy variable:

- $F_{Y|Z=0}(a)$ is the CDF of education among unmarried individuals.
- $F_{Y|Z=1}(a)$ is the CDF of education among married individuals.



(a) Conditional CDFs (b) Conditional PDFs

Figure 2.6: Wage distributions conditional on 10 years of experience and gender

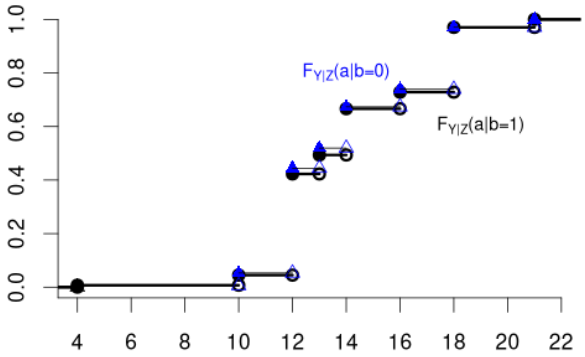
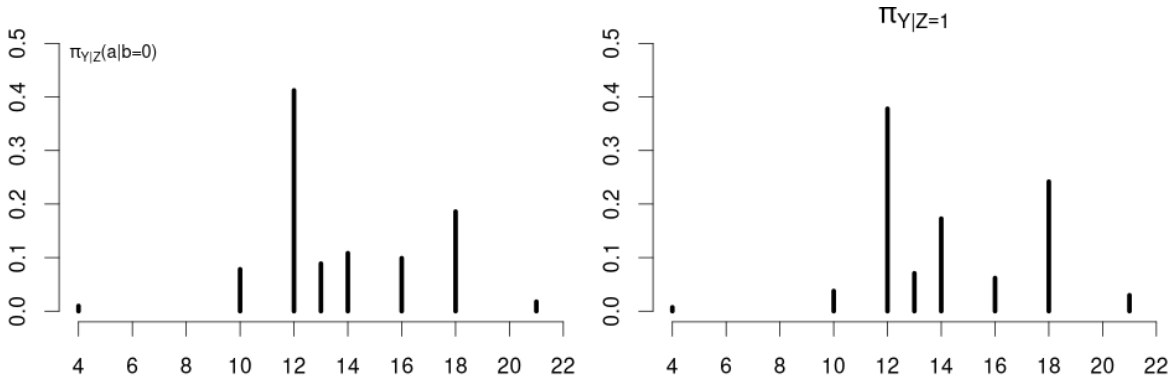


Figure 2.7: Conditional CDFs of education given married



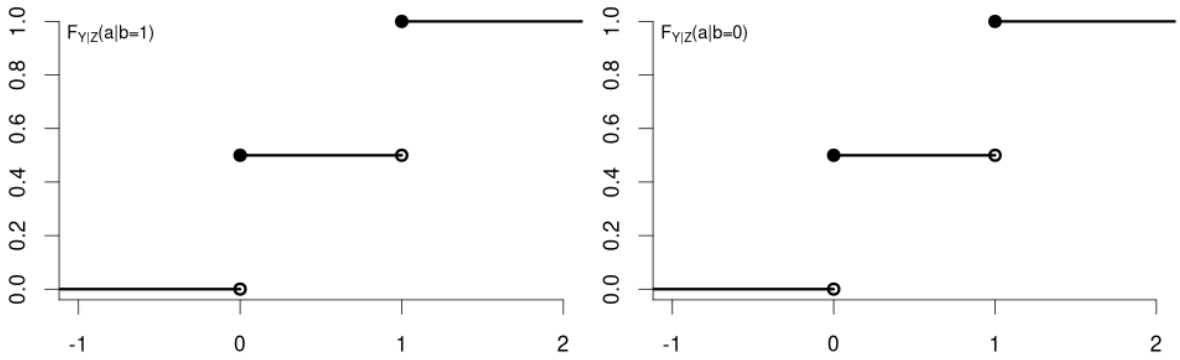
(a) Education given unmarried (b) Education given married

Figure 2.8: Conditional PMFs of education for unmarried (left) and married (right) individuals

The conditional PMFs $\pi_{Y|Z=0}(a) = P(Y = a|Z = 0)$ and $\pi_{Y|Z=1}(a) = P(Y = a|Z = 1)$ indicate the jump heights of $F_{Y|Z=0}(a)$ and $F_{Y|Z=1}(a)$ at a.

Clearly, *education* and *married* are dependent random variables. For example, $\pi_{Y|Z=0}(12) > \pi_{Y|Z=1}(12)$ and $\pi_{Y|Z=0}(18) < \pi_{Y|Z=1}(18)$.

In contrast, consider Y = coin flip and Z = married dummy variable. The CDF of a coin flip should be the same for married or unmarried individuals:



(a) Coin flip given unmarried

(b) Coin flip given married

Figure 2.9: Conditional CDFs of a coin flip for unmarried (left) and married (right) individuals

Because

$$F_Y(a) = F_{Y|Z=0}(a) = F_{Y|Z=1}(a) \quad \text{for all } a$$

we say that Y and Z are **independent random variables**.

2.4 Joint Distribution

When we have two random variables Y and Z, we need to understand three related concepts:

1) **Joint distribution**: How Y and Z behave together
2) **Marginal distributions**: How Y and Z behave individually
3) **Conditional distributions**: How Y behaves given information about Z (and vice versa)

Marginal and Joint Distributions

The joint CDF describes the probability that both variables simultaneously fall below specified values:

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

The marginal CDFs are obtained by "sending the other coordinate to $+\infty$":

$$F_Y(a) = \lim_{b\to\infty} F_{Y,Z}(a,b) = P(Y \le a),$$
$$F_Z(b) = \lim_{a\to\infty} F_{Y,Z}(a,b) = P(Z \le b).$$

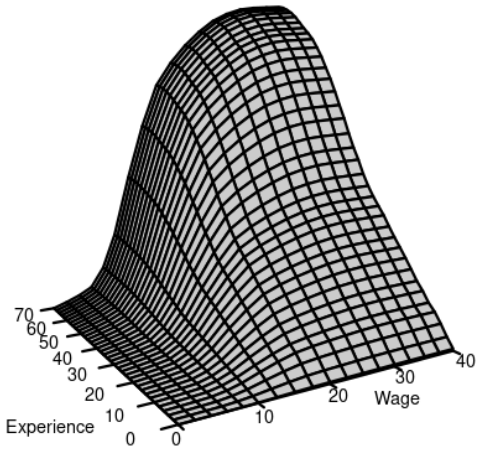
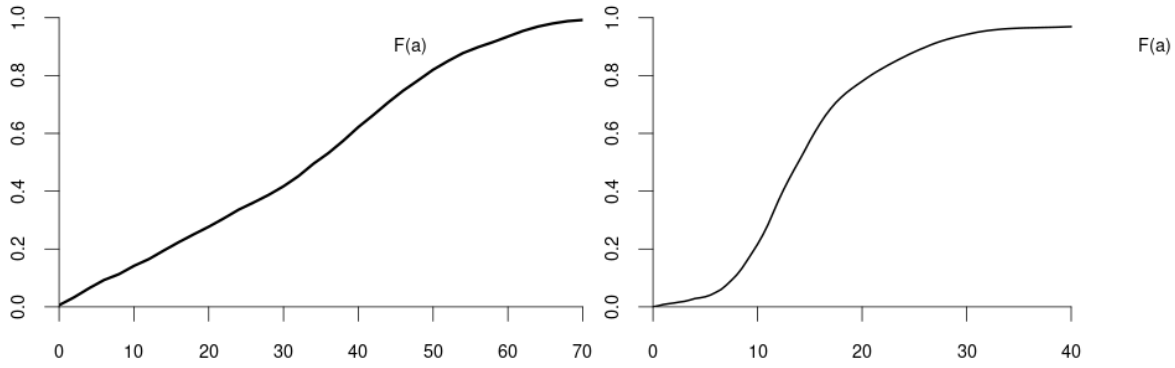


Figure 2.10: Example: Joint CDF of wage and experience



(a) Marginal CDF of experience

(b) Marginal CDF of wage

Figure 2.11: Marginal CDFs of experience (left) and wage (right)

When both variables are continuous, the joint PDF is

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

and the marginal PDFs are

$$f_Y(a) = \int_{-\infty}^{\infty} f_{Y,Z}(a,v)\,dv, \quad f_Z(b) = \int_{-\infty}^{\infty} f_{Y,Z}(u,b)\,du.$$

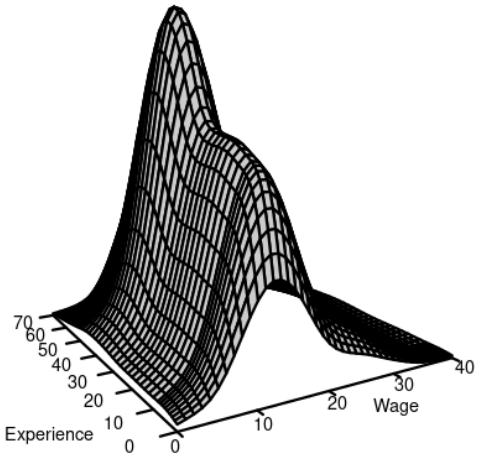


Figure 2.12: Example: Joint PDF of wage and experience

When both variables are discrete, the joint PMF is

$$\pi_{Y,Z}(a,b) = P(Y = a, Z = b),$$

and the marginal PMFs are

$$\pi_Y(a) = \sum_{v \in \mathcal{Z}} \pi_{Y,Z}(a,v), \quad \pi_Z(b) = \sum_{u \in \mathcal{Y}} \pi_{Y,Z}(u,b),$$

where $\mathcal{Y}$ and $\mathcal{Z}$ are the supports of Y and Z.

Conditional and Joint Distributions

If Z is continuous (with $f_Z(b) > 0$):

$$F_{Y|Z=b}(a) = \int_{-\infty}^{a} f_{Y|Z=b}(u)\,du = \frac{\frac{\partial}{\partial b} F_{Y,Z}(a,b)}{f_Z(b)}.$$

If Z is discrete (with $b \in \mathcal{Z}$):

$$F_{Y|Z=b}(a) = \sum_{u \in \mathcal{Y}, u \le a} \pi_{Y|Z=b}(u) = \frac{F_{Y,Z}(a,b) - F_{Y,Z}(a,b^-)}{F_Z(b) - F_Z(b^-)}.$$

In general or mixed cases, the conditional CDF can be defined through limits:

$$F_{Y|Z=b}(a) = \lim_{\epsilon \to 0, \epsilon > 0} \frac{F_{Y,Z}(a, b+\epsilon) - F_{Y,Z}(a, b-\epsilon)}{F_Z(b+\epsilon) - F_Z(b-\epsilon)}.$$

Recovering the Joint from Conditionals

At the CDF level (Riemann-Stieltjes form), the joint can be built from a conditional and the other variable's marginal:

$$F_{Y,Z}(a,b) = \int_{-\infty}^{b} F_{Y|Z=v}(a)\, dF_Z(v) = \int_{-\infty}^{a} F_{Z|Y=u}(b)\, dF_Y(u).$$

Special cases:

- If Z is continuous:
$$F_{Y,Z}(a,b) = \int_{-\infty}^{b} F_{Y|Z=v}(a) f_Z(v)\, dv,$$
- If Z is discrete:
$$F_{Y,Z}(a,b) = \sum_{v \in \mathcal{Z}, v \le b} F_{Y|Z=v}(a) \pi_Z(v).$$
- If Y is continuous:
$$F_{Y,Z}(a,b) = \int_{-\infty}^{a} F_{Z|Y=u}(b) f_Y(u)\, du.$$
- If Y is discrete:
$$F_{Y,Z}(a,b) = \sum_{u \in \mathcal{Y}, u \le a} F_{Z|Y=u}(b) \pi_Y(u).$$

For PDF/PMF, the product rules are:

$$f_{Y,Z}(a,b) = f_{Z|Y=a}(b) f_Y(a) = f_{Y|Z=b}(a) f_Z(b),$$
$$\pi_{Y,Z}(a,b) = \pi_{Z|Y=a}(b) \pi_Y(a) = \pi_{Y|Z=b}(a) \pi_Z(b).$$

2.5 Independence of Random Variables

In the previous section, we saw that the distribution of a coin flip remains the same regardless of a person's marital status, illustrating the concept of independence. Let's now formalize this important concept.

Independence

Y and Z are **independent** if and only if

$$F_{Y|Z=b}(a) = F_Y(a) \quad \text{for all } a \text{ and } b.$$

Note that if $F_{Y|Z=b}(a) = F_Y(a)$ for all b, then automatically $F_{Z|Y=a}(b) = F_Z(b)$ for all a. Due to this symmetry we can equivalently define independence through the property $F_{Z|Y=a}(b) = F_Z(b)$.

> **Technical Note**: Mathematically, the condition is required to hold only for almost every b. That is, for all b except on a set with probability zero under Z. Intuitively, it only needs to hold on the values that Z can actually take. For example, if Z is wage and wages can't be negative, the condition need not hold for negative b.

The definition naturally generalizes to Z_1, Z_2, Z_3 in a sequential chain form. They are **mutually independent** if

(i) $F_{Z_2|Z_1=b_1}(a) = F_{Z_2}(a)$,
(ii) $F_{Z_3|Z_1=b_1,Z_2=b_2}(a) = F_{Z_3}(a)$,

for all a and for (almost) all (b_1, b_2).

Mutual Independence

The random variables $Z_1, \ldots, Z_n$ are **mutually independent** if and only if, for each $i = 2, \ldots, n$,

$$F_{Z_i|Z_1=b_1,\ldots,Z_{i-1}=b_{i-1}}(a) = F_{Z_i}(a)$$

for all a and (almost) all $(b_1, \ldots, b_{i-1})$.

An equivalent viewpoint uses the **joint CDF** of the vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$:

$$F_{\boldsymbol{Z}}(\boldsymbol{a}) = F_{Z_1,\ldots,Z_n}(a_1, \ldots, a_n) = P(Z_1 \leq a_1, \ldots, Z_n \leq a_n).$$

Then, $Z_1, \ldots, Z_n$ are mutually independent if and only if

$$F_{\boldsymbol{Z}}(\boldsymbol{a}) = F_{Z_1}(a_1) \cdots F_{Z_n}(a_n) \quad \text{for all } a_1, \ldots, a_n.$$

2.6 Independent and Identically Distributed

An important concept in statistics is that of an independent and identically distributed (i.i.d.) sample. This arises naturally when we consider multiple random variables that share the same distribution and do not influence each other.

i.i.d. Sample / Random Sample

A collection of random variables $Y_1, \dots, Y_n$ is **i.i.d.** (independent and identically distributed) if:

1. They are mutually independent: for each $i = 2, \dots, n$,

$$F_{Y_i|Y_1=b_1,\dots,Y_{i-1}=b_{i-1}}(a) = F_{Y_i}(a)$$

   for all a and (almost) all $(b_1, \dots, b_{i-1})$.

2. They have the same distribution function: $F_{Y_i}(a) = F(a)$ for all $i = 1, \dots, n$ and all a.

For example, consider n coin flips, where each Y_i represents the outcome of the i-th flip (with $Y_i = 1$ for heads and $Y_i = 0$ for tails). If the coin is fair and the flips are performed independently, then $Y_1, \dots, Y_n$ form an i.i.d. sample with

$$F(a) = F_{Y_i}(a) = \begin{cases} 0 & a < 0 \\ 0.5 & 0 \le a < 1 \\ 1 & a \ge 1 \end{cases} \qquad \text{for all } i = 1, \dots, n.$$

Similarly, if we randomly select n individuals from a large population and measure their wages, the resulting measurements $Y_1, \dots, Y_n$ can be treated as an i.i.d. sample. Each Y_i follows the same distribution (the wage distribution in the population), and knowledge of one person's wage doesn't affect the distribution of another's. The function F is called the **population distribution** or the data-generating process (DGP).

2.7 Independence of Random Vectors

Often in practice, we work with multiple variables recorded for different individuals or time points. For example, consider two random vectors:

$$\boldsymbol{X}_1 = (X_{11}, \dots, X_{1k})', \quad \boldsymbol{X}_2 = (X_{21}, \dots, X_{2k})'.$$

The conditional distribution function of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2$ takes the value $\boldsymbol{b} = (b_1, \ldots, b_k)'$ is

$$F_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{b}}(\boldsymbol{a}) = P(\boldsymbol{X}_1 \leq \boldsymbol{a}|\boldsymbol{X}_2 = \boldsymbol{b}),$$

where the vector inequality $\boldsymbol{X}_1 \leq \boldsymbol{a}$ means the componentwise inequalities $X_{1j} \leq a_j$ for all $j = 1, \ldots, k$ hold.

For instance, if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ represent the survey answers of two different, randomly chosen people, then $F_{\boldsymbol{X}_2|\boldsymbol{X}_1=\boldsymbol{b}}(\boldsymbol{a})$ describes the distribution of the second person's answers, given that the first person's answers are $\boldsymbol{b}$.

If the two people are truly randomly selected and unrelated to one another, we would not expect $\boldsymbol{X}_2$ to depend on whether $\boldsymbol{X}_1$ equals $\boldsymbol{b}$ or some other value $\boldsymbol{c}$. In other words, knowing $\boldsymbol{X}_1$ provides no information that changes the distribution of $\boldsymbol{X}_2$.

Independence of Random Vectors

Two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are **independent** if and only if

$$F_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{b}}(\boldsymbol{a}) = F_{\boldsymbol{X}_1}(\boldsymbol{a}) \quad \text{for all } \boldsymbol{a} \text{ and (almost) all } \boldsymbol{b}.$$

This definition extends naturally to mutual independence of n random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$. They are called **mutually independent** if, for each $i = 2, \ldots, n$,

$$F_{\boldsymbol{X}_i|\boldsymbol{X}_1=\boldsymbol{b}_1,\ldots,\boldsymbol{X}_{i-1}=\boldsymbol{b}_{i-1}}(\boldsymbol{a}) = F_{\boldsymbol{X}_i}(\boldsymbol{a})$$

for all $\boldsymbol{a}$ and (almost) all $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{i-1})$.

Hence, in an independent sample, what the i-th randomly chosen person answers does not depend on anyone else's answers.

i.i.d. Sample of Random Vectors

The concept of i.i.d. samples naturally extends to random vectors. A collection of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is **i.i.d.** if they are mutually independent and have the same distribution function F. Formally,

$$F_{\boldsymbol{X}_i|\boldsymbol{X}_1=\boldsymbol{b}_1,\ldots,\boldsymbol{X}_{i-1}=\boldsymbol{b}_{i-1}}(\boldsymbol{a}) = F(\boldsymbol{a})$$

for all $i = 1, \ldots, n$, for all $\boldsymbol{a}$, and (almost) all $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{i-1})$.

An **i.i.d. dataset** (or **random sample**) is one where each multivariate observation not only comes from the same population distribution F but is independent of the others.

2.8 PMF and PDF Estimation

PMF estimation

With i.i.d. data $Y_1, \ldots, Y_n$ from a discrete random variable Y with support $\mathcal{Y}$, the PMF $\pi_Y(a)$ can be estimated by the empirical PMF

$$\hat{\pi}_Y(a) = \frac{n(a)}{n}, \quad a \in \mathcal{Y}$$

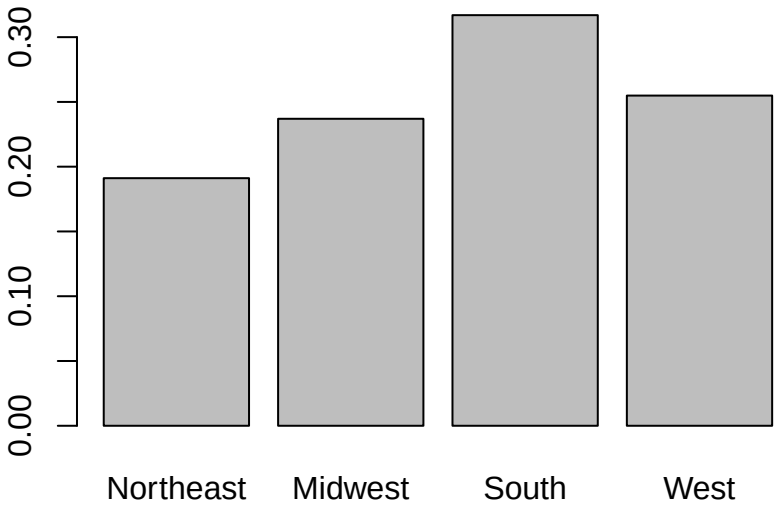
where $n(a)$ is the count of observations equal to a.

Let's load the CPS data from Section 1 and estimate the PMF for *region* (1 = Northeast, 2 = Midwest, 3 = South, 4 = West):

```
cps = read.csv("cps.csv")
n = length(cps$region)  #sample size
pmf = table(cps$region)/n #relative frequencies
pmf
```

```
        1         2         3         4
0.1911434 0.2369832 0.3169761 0.2548973
```

```
barplot(pmf,
        names.arg = c("Northeast", "Midwest", "South", "West"))
```



PDF estimation

Histogram

For continuous data, a histogram provides an intuitive estimate of the PDF.

A histogram divides the data range into B bins, each of equal width h, and counts the number of observations n_j within each bin.

The histogram density estimator is

$$\hat{f}_Y(a) = \frac{n_j}{nh} \quad \text{for } a \text{ in bin } j,$$

so the total area of the rectangles sums to 1.

Kernel density estimator

Suppose we want to estimate the experience density at $a = 21$ and consider the histogram density estimate with $h = 5$. It is based on the frequency of observations in the interval $[20, 25)$ which is a skewed window about $a = 21$.

It seems more sensible to center the window at 21, for example $[18.5, 23.5)$ instead of $[20, 25)$. It also seems sensible to give more weight to observations close to 21 and less to those at the edge of the window.

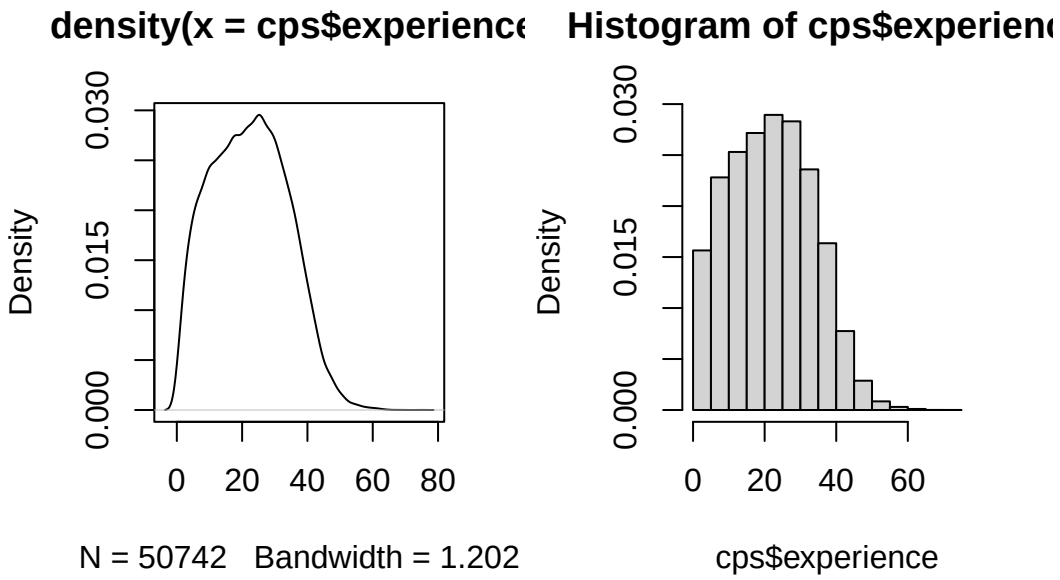
This idea leads to the **kernel density estimator** of $f_Y(a)$, which is a smooth version of the histogram:

$$\hat{f}_Y(a) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{a - Y_i}{h}\Big).$$

Here, $K(u)$ represents a weighting function known as a kernel function, and $h > 0$ is the **bandwidth**. A common choice for $K(u)$ is the Gaussian kernel:

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

```
par(mfrow = c(1,2))
plot(density(cps$experience))
hist(cps$experience, probability=TRUE)
```



The `hist()` and `density()` functions in R automatically choose default values for the number of bins B and the bandwidth h.

2.9 R Code

statistics-sec02.R

3 Moments

In practice, we are interested in characteristics (parameters) of a population distribution, such as the mean or variance of a variable, or correlations between multiple variables. These characteristics are related to the population moments of a distribution.

While the population distribution and its population moments are unobserved, we can learn about these characteristics using the sample moments of a univariate dataset $Y_1, \dots, Y_n$ or a multivariate dataset $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. The ideal scenario is that the dataset forms an i.i.d. sample from the underlying population distribution of interest.

3.1 Sample Moments

The r**-th sample moment** about the origin (also called the r-th empirical moment) of a univariate sample $Y_1, \dots, Y_n$ is defined as

$$\overline{Y^r} = \frac{1}{n} \sum_{i=1}^{n} Y_i^r.$$

For example, the first sample moment ($r = 1$) is the **sample mean** (arithmetic mean):

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

The sample mean is the most common measure of central tendency in a sample.

3.2 Population Moments

The population counterpart of the sample mean is the **expected value**.

Discrete Random Variables

The **expected value**, also called expectation or (population) mean, of a discrete random variable Y with PMF $\pi_Y(\cdot)$ and support $\mathcal{Y}$ is defined as

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

The **r-th moment** (or r-th raw moment) is

$$E[Y^r] = \sum_{u \in \mathcal{Y}} u^r \, \pi_Y(u).$$

Suppose the *education* variable has the following PMF:

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

Then, the expected value of *education* is calculated by summing over all possible values:

$$\begin{aligned} E[Y] = {} & 4 \cdot \pi_Y(4) + 10 \cdot \pi_Y(10) + 12 \cdot \pi_Y(12) \\ & + 13 \cdot \pi_Y(13) + 14 \cdot \pi_Y(14) + 16 \cdot \pi_Y(16) \\ & + 18 \cdot \pi_Y(18) + 21 \cdot \pi_Y(21) = 14.117 \end{aligned}$$

A binary random variable Y has support $\mathcal{Y} = \{0, 1\}$. The probabilities are

- $\pi_Y(1) = P(Y = 1) = p$
- $\pi_Y(0) = P(Y = 0) = 1 - p$

for some $p \in [0, 1]$. The expected value of Y is:

$$\begin{aligned} E[Y] &= 0 \cdot \pi_Y(0) + 1 \cdot \pi_Y(1) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p. \end{aligned}$$

For the variable *coin*, the probability of heads is $p = 0.5$ and the expected value is $E[Y] = p = 0.5$.

Continuous Random Variables

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

$$E[Y] = \int_{-\infty}^{\infty} u\, f_Y(u)\, du.$$

The **r-th moment** is

$$E[Y^r] = \int_{-\infty}^{\infty} u^r\, f_Y(u)\, du.$$

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

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

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

$$E[Y] = \int_{-\infty}^{\infty} u\, f_Y(u)\, du = \int_0^1 u\, du = \frac{1}{2}u^2 \Big|_0^1 = \frac{1}{2}.$$

General Cases

Not every distribution has a PMF or a PDF, but every distribution has a CDF. You can define expectation also directly via the CDF using the Riemann-Stieltjes integral:

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

Similarly, the expected value of a function of multiple random variables can be defined via their joint CDF:

$$E[g(Z_1, \ldots, Z_k)] = \int_{\mathbb{R}^k} g(u_1, \ldots, u_k)\, dF_{Z_1,\ldots,Z_k}(u_1, \ldots, u_k).$$

For more details, see: probability.svenotto.com/part2_expectation.html#general-case

Exceptional Cases

Not every distribution has a well-defined expected value. The simple Pareto distribution with parameter $\alpha = 1$ has the PDF:

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

The expected value is:

$$E[Y] = \int_{-\infty}^{\infty} u f_Y(u)\ \mathrm{d}u = \int_1^{\infty} \frac{u}{u^2}\ \mathrm{d}u = \int_1^{\infty} \frac{1}{u}\ \mathrm{d}u = \log(u)|_1^{\infty} = \infty,$$

where "$= \infty$" means it diverges.

The game of chance from the St. Petersburg paradox is a discrete example with infinite expectation. In this game, a fair coin is tossed until a tail appears; if the first tail is on the n-th toss, the payoff is 2^n dollars. The probability of "first tail on the n-th toss" is 2^{-n}. The expected payoff is:

$$E[Y] = \sum_{n=1}^{\infty} 2^n \cdot \frac{1}{2^n} = \sum_{n=1}^{\infty} 1 = \infty$$

For a t-distributed random variable Y with m degrees of freedom we have $E[|Y|^k] < \infty$ for $k < m$ and $E[|Y|^k] = \infty$ for $k \geq m$. In particular $E[Y] = 0$ for $m > 1$ and $E[Y^2] = m/(m-2)$ for $m > 2$.

Many statistical procedures require conditions such as $E[Y^4] < \infty$ (finite fourth moments). This excludes distributions with very heavy tails and ensures that large outliers (like extreme payoffs as in the St. Petersburg game) are rare enough.

3.3 Convergence in Probability

The sample mean $\overline{Y}$ is a function of the sample $Y_1, \ldots, Y_n$ and the expected value $E[Y]$ is a function of the population distribution F_Y.

$E[Y]$ is a **parameter** of the population distribution F_Y. In general, a parameter θ is a characteristic or feature of a population distribution. Parameters are typically fixed but unknown quantities that we aim to learn about through sampling and estimation.

$\overline{Y}$ is an **estimator** for the parameter $E[Y]$. In general, an estimator $\hat{\theta}_n$ is a function of sample data intended to approximate the unknown parameter θ. Since an estimator is a function of random variables (the sample), it is itself a random variable. When we actually compute the estimator from a specific realized sample, we call the resulting value an estimate.

A desired property for any estimator $\hat{\theta}_n$ is that it gets closer and closer to the true parameter θ as the sample size n increases. It eventually converges to the true parameter value in a hypothetically infinitely large sample.

Because $\hat{\theta}_n$ is a function of the sample and therefore random, we use convergence in probability rather than a purely deterministic limit.

Convergence in Probability

A sequence of random variables $\{W_n\}_{n=1}^{\infty}$ **converges in probability** to a constant c if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|W_n - c| > \epsilon) = 0$$

Equivalently, this can be expressed as:

$$\lim_{n\to\infty} P(|W_n - c| \leq \epsilon) = 1$$

This is denoted as $W_n \xrightarrow{p} c$.

Intuitively, convergence in probability means that as the sample size n increases, the probability that W_n deviates from c by more than any fixed positive amount ϵ becomes arbitrarily small.

For example, if $W_n \xrightarrow{p} c$, then for any small $\epsilon > 0$ (say, $\epsilon = 0.01$), we can make $P(|W_n - c| > 0.01)$ as small as we want by choosing a sufficiently large sample size n. This doesn't mean that W_n will exactly equal c for large n, but rather that the probability of W_n being close to c approaches 1 as n grows.

Applying the concept of convergence in probability to an estimator $\hat{\theta}_n$ for a parameter θ leads to the important property of **consistency**.

Consistency

An estimator $\hat{\theta}_n$ is **consistent** for the parameter θ if:

$$\hat{\theta}_n \xrightarrow{p} \theta \quad \text{as} \quad n \to \infty$$

That is, if for any $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Consistency is a minimal requirement for a good estimator. It ensures that with a large enough sample, the estimator will be arbitrarily close to the true parameter with high probability.

If an estimator $\hat{\theta}_n$ is a continuous random variable, it will almost never equal exactly the true parameter value because for continuous distributions, point probabilities are zero: $P(\hat{\theta}_n = \theta) = 0$.

However, the larger the sample size, the higher the probability that $\hat{\theta}_n$ falls within a small neighborhood around the true value θ. Consistency means that, if we fix some small precision value $\epsilon > 0$, then,

$$P(|\hat{\theta}_n - \theta| \leq \epsilon) = P(\theta - \epsilon \leq \hat{\theta}_n \leq \theta + \epsilon)$$

should increase as the sample size n grows, approaching 1 in the limit.

This property aligns with our intuition that more data should lead to better estimates.

3.4 Law of Large Numbers

The **Law of Large Numbers (LLN)** is one of the fundamental results in probability theory that establishes the consistency of the sample mean for its population mean.

Law of Large Numbers (LLN)

Let $Y_1, Y_2, \ldots, Y_n$ be a univariate i.i.d. sample with $\mu = E[Y_i]$. If $|\mu| < \infty$, then

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu \quad \text{as} \quad n \to \infty.$$

The LLN essentially states that if we take a large enough sample from a population with finite mean, the sample mean will be close to the population mean with high probability.

Below is an interactive Shiny app to visualize the law of large numbers using simulated data for different sample sizes and different distributions.

SHINY APP: LLN

The LLN is not confined to i.i.d. data, but for other sampling schemes additional conditions must be satisfied.

Clustered Data

Clustered data has the form Y_{gj} for $g = 1, \dots, G$ and $j = 1, \dots, n_g$. Here, G is the number of clusters and n_g the number of observations in cluster g.

For instance, G may be the number of classrooms, and n_g the number of students in classroom g that take part in a survey (clustered cross section). Or G may be the number of firms and n_g the number of years where the data for firm g is observed (panel data)

The LLN holds for clustered data if

- Clusters are independent across $g = 1, \dots, G$ (clusters are independently drawn) while dependence inside a cluster is allowed. That is, Y_{gj} and Y_{hk} are independent for $g \neq h$ but Y_{gj} and Y_{gk} may be dependent.
- Cluster observations have a common finite mean: $E[Y_{gj}] = \mu < \infty$ for all g and j
- No single cluster dominates: $\max_{g=1,\dots,G} n_g/N \to 0$ with $N = \sum_{h=1}^{G} n_h$, as $N \to \infty$.

Then,

$$\frac{1}{N} \sum_{g=1}^{G} \sum_{j=1}^{n_g} Y_{gj} \xrightarrow{p} \mu.$$

Time Series Data

A time series Y_t for $t = 1, \dots, n$ is called (strictly) stationary if the vector $(Y_t, Y_{t+1}, \dots, Y_{t+h})'$ has the same probability distribution as the vector $(Y_{t-j}, Y_{t-j+1}, \dots, Y_{t-j+h})'$ for any t, h, and j. That is, the distribution is invariant to time shifts.

For time series data, the LLN

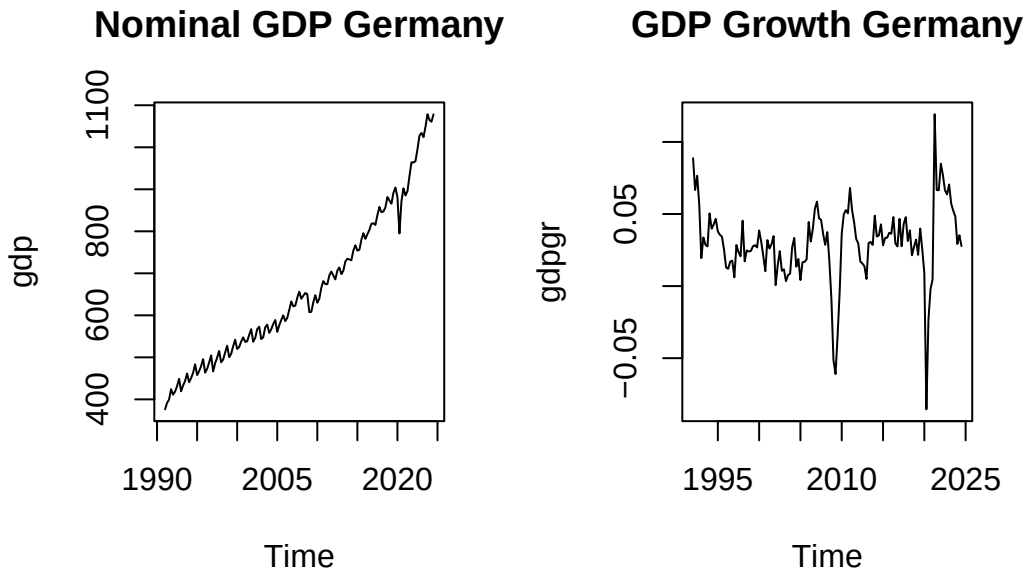
$$\frac{1}{n} \sum_{t=1}^{n} Y_t \xrightarrow{p} \mu$$

holds, if

- Y_t is stationary with a finite mean $E[Y_t] = \mu < \infty$;
- Observations Y_t and Y_{t-j} "become independent as j gets large" (strong mixing).

Nominal GDP is typically nonstationary, while year-on-year GDP growth is often (approximately) stationary:

```
library(TeachData)
par(mfrow = c(1,2))
plot(gdp, main = "Nominal GDP Germany")
plot(gdpgr, main = "GDP Growth Germany")
```



3.5 Central Moments

The r**-th central sample moment** is the average of the r-th powers of the deviations from the sample mean:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^r$$

For example, the second central moment ($r = 2$) is the **sample variance**:

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

The sample variance measures the spread or dispersion of the data around the sample mean.

The population variance of a random variable Y is defined as

$$\text{Var}(Y) = E[(Y - E[Y])^2] = E[Y^2] - E[Y]^2.$$

If $Y_1, \ldots, Y_n$ are i.i.d. draws from the distribution of Y with $E[Y^2] < \infty$, then

$$\hat{\sigma}_Y^2 \xrightarrow{p} \text{Var}(Y).$$

The **sample standard deviation** is the square root of the sample variance:

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

It quantifies the typical deviation of data points from the sample mean in the original units of measurement.

The population standard deviation is

$$\mathrm{sd}(Y) = \sqrt{\mathrm{Var}(Y)},$$

and $\hat{\sigma}_Y \xrightarrow{p} \mathrm{sd}(Y)$ under the same conditions as for the sample variance.

3.6 Cross Moments

For a bivariate sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, we can compute cross moments that describe the relationship between the two variables. The **sample cross moment** is:

$$\overline{YZ} = \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i.$$

Under i.i.d. sampling from the distribution of the bivariate random variable (Y, Z), it converges in probability to the population cross moment $E[YZ]$.

The central sample cross moment is also known as the **sample covariance** and is defined as:

$$\hat{\sigma}_{YZ} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z}) = \overline{YZ} - \overline{Y}\,\overline{Z}.$$

It converges under i.i.d. sampling in probability to the **population covariance**

$$\mathrm{Cov}(Y, Z) = E[(Y - E[Y])(Z - E[Z])] = E[YZ] - E[Y]E[Z].$$

The **sample correlation coefficient** is the standardized sample covariance:

$$r_{YZ} = \frac{\hat{\sigma}_{YZ}}{\hat{\sigma}_Y \hat{\sigma}_Z} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\sqrt{\sum_{i=1}^{n}(Z_i - \overline{Z})^2}}.$$

Its population counterpart is the **population correlation**

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

3.7 Rules of Calculation

For any random variables X, Y, Z, and real numbers $a, b \in \mathbb{R}$, we have the following rules:

- The expected value is linear:

$$E[a + bY] = a + bE[Y].$$

- Expectation of the sum of two random variables:

$$E[Y + Z] = E[Y] + E[Z].$$

- If Y and Z are independent, then

$$E[YZ] = E[Y]E[Z].$$

- The variance has a quadratic scaling property:

$$\mathrm{Var}(a + bY) = b^2\mathrm{Var}(Y)$$

- Variance of the sum of random variables:

$$\mathrm{Var}(Y + Z) = \mathrm{Var}(Y) + 2\mathrm{Cov}(Y, Z) + \mathrm{Var}(Z)$$

- For the sum of k random variables:

$$\mathrm{Var}\Big(\sum_{i=1}^{k} W_i\Big) = \sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{Cov}(W_i, W_j)$$

- If Y and Z are independent, then $\mathrm{Cov}(Y, Z) = 0$. The converse need not hold.
- If Y and Z are independent,

$$\mathrm{Var}(Y + Z) = \mathrm{Var}(Y) + \mathrm{Var}(Z)$$

- The covariance is bilinear:

$$\mathrm{Cov}(aY + bZ, X) = a\mathrm{Cov}(Y, X) + b\mathrm{Cov}(Z, X)$$

3.8 Standardized Moments

The **r-th standardized sample moment** is the central moment normalized by the sample standard deviation raised to the power of r. It is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\widehat{\sigma}_Y}\right)^r,$$

provided that $\hat{\sigma}_Y > 0$.

Its population counterpart is the r-th standardized population moment

$$E\left[\left(\frac{Y - E[Y]}{\text{sd}(Y)}\right)^r\right],$$

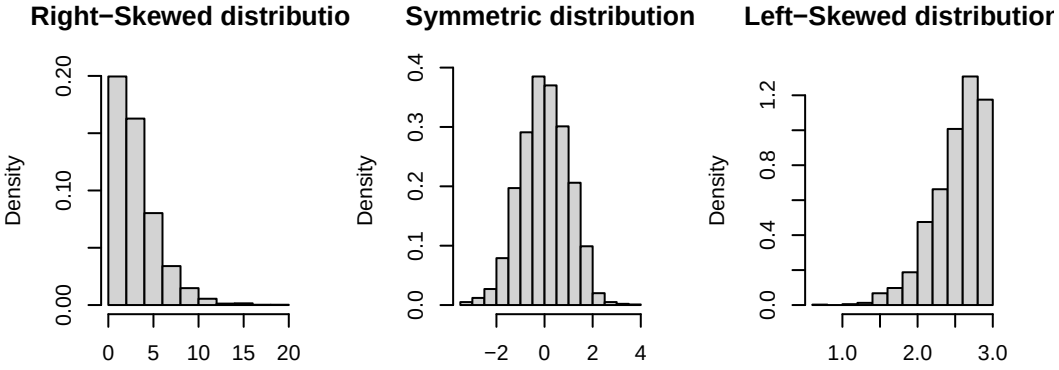
provided that $\text{sd}(Y) > 0$.

Skewness

The sample skewness is the third standardized sample moment:

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

The skewness is a measure of asymmetry around the mean. A positive skewness indicates that the distribution has a longer or heavier tail on the right side (right-skewed), while a negative skewness indicates a longer or heavier tail on the left side (left-skewed). A perfectly symmetric distribution, such as the normal distribution, has a skewness of 0.



The population skewness is

$$\text{skew}(Y) = E\left[\left(\frac{Y - E[Y]}{\text{sd}(Y)}\right)^3\right] = \frac{E[(Y - E[Y])^3]}{\text{sd}(Y)^3}$$

Kurtosis

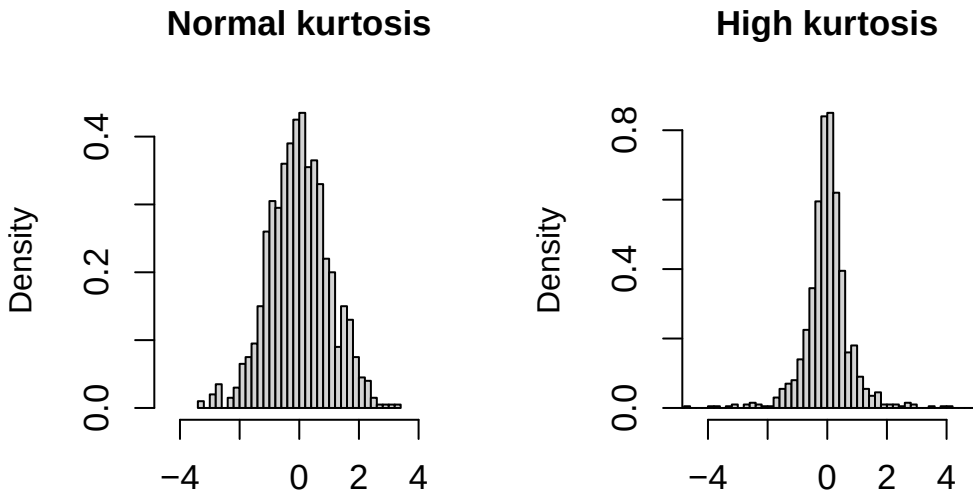
The sample kurtosis is the fourth standardized sample moment:

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

Kurtosis measures the "tailedness" or heaviness of the tails of a distribution and can indicate the presence of extreme outliers. The reference value of kurtosis is 3, which corresponds to the kurtosis of a normal distribution. Values greater than 3 suggest heavier tails, while values less than 3 indicate lighter tails.

The population kurtosis is

$$\text{kurt}(Y) = E\left[\left(\frac{Y - E[Y]}{\text{sd}(Y)}\right)^4\right] = \frac{E[(Y - E[Y])^4]}{\text{Var}(Y)^2}.$$

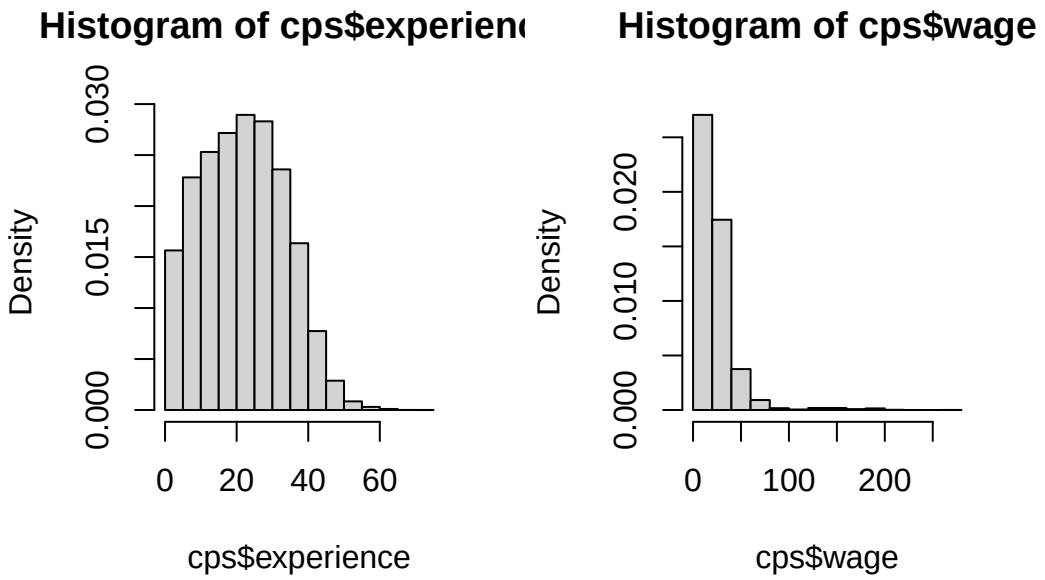


The plots display histograms of two standardized datasets (both have a sample mean of 0 and a sample variance of 1). The left dataset has a normal sample kurtosis (around 3), while the right dataset has a high sample kurtosis with heavier tails.

Log-transformations

Let's load the CPS dataset from Section 1:

```
cps = read.csv("cps.csv")
par(mfrow = c(1,2))
hist(cps$experience, probability = TRUE)
hist(cps$wage, probability = TRUE)
```



To compute the sample skewness and kurtosis we can use the `moments` package

```
library(moments)
c(
  skewness(cps$experience),
  skewness(cps$wage)
)
```

```
[1] 0.1872222 4.3201570
```

Wages are right-skewed because a few very rich individuals earn much more than the many with low to medium incomes. Experience does not indicate any pronounced skewness.

```
c(
  kurtosis(cps$experience),
  kurtosis(cps$wage)
)
```

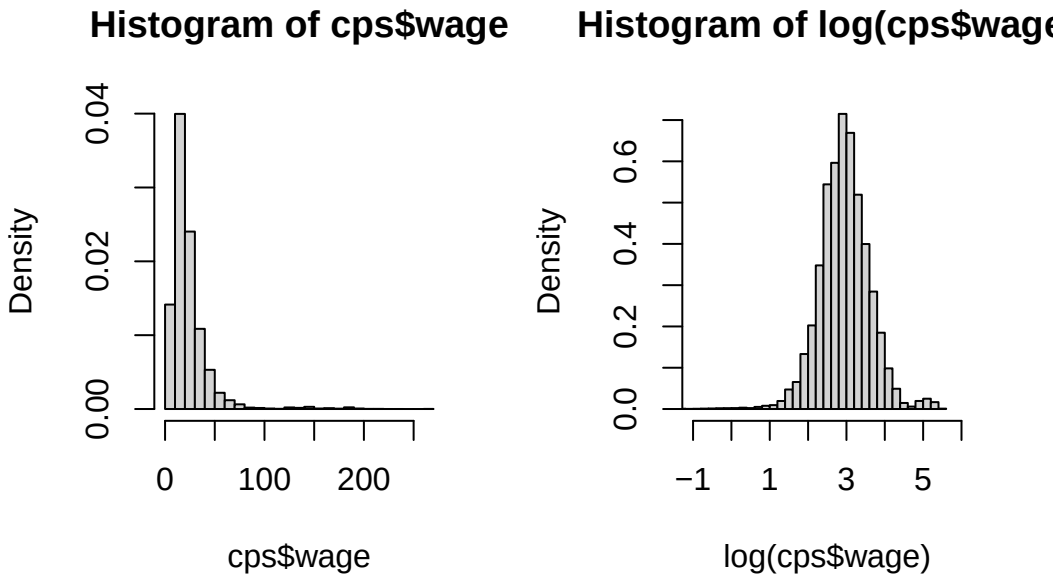
```
[1]  2.373496 30.370331
```

Wages have a large kurtosis due to a few super-rich individuals in the sample. The kurtosis of experience is close to 3 and thus similar to a normal distribution.

Right-skewed, heavy-tailed variables are common in real-world datasets, such as income levels, wealth accumulation, property values, insurance claims, and social media follower counts. A

common transformation to reduce skewness and kurtosis in data is to use the natural logarithm:

```
par(mfrow = c(1,2))
hist(cps$wage, probability = TRUE, breaks = 20)
hist(log(cps$wage), probability = TRUE, breaks = 50, xlim = c(-1, 6))
```




```
c(
  skewness(log(cps$wage)),
  kurtosis(log(cps$wage))
  )
```

```
[1] -0.6990539 11.8566367
```

In econometrics, statistics, and many programming languages including R, $\log(\cdot)$ is commonly used to denote the natural logarithm (base e).

Note: On a pocket calculator, use **LN** to calculate the natural logarithm $\log(\cdot) = \log_e(\cdot)$. If you use **LOG**, you will calculate the logarithm with base 10, i.e., $\log_{10}(\cdot)$, which will give you a different result. The relationship between these logarithms is $\log_{10}(x) = \log_e(x)/\log_e(10)$.

3.9 Multivariate Moments

Consider a multivariate i.i.d. dataset $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$, such as the following subset of the `cps` dataset:

```r
dat = data.frame(wage = cps$wage, education = cps$education, female = cps$female)
```

The **sample mean vector** $\overline{\boldsymbol{X}}$ contains the sample means of the k variables and is defined as

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} X_{i1} \\ \vdots \\ X_{ik} \end{pmatrix}.$$

Its population counterpart is the population mean vector

$$E[\boldsymbol{X}_i] = E[(X_{i1}, \dots, X_{ik})']$$

Cross Moment Matrix

The **sample cross-moment matrix** is

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i' = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} X_{i1}^2 & X_{i1}X_{i2} & \dots & X_{i1}X_{ik} \\ X_{i1}X_{i2} & X_{i2}^2 & \dots & X_{i2}X_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i1}X_{ik} & X_{i2}X_{ik} & \dots & X_{ik}^2 \end{pmatrix}$$

Its population counterpart is $E[\boldsymbol{X}_i\boldsymbol{X}_i']$.

Sample covariance matrix

The **sample covariance matrix** $\widehat{\Sigma}$ is the $k \times k$ matrix given by

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})'.$$

Its elements $\hat{\sigma}_{h,l}$ represent the pairwise **sample covariance** between variables h and l:

$$\hat{\sigma}_{h,l} = \frac{1}{n}\sum_{i=1}^{n}(X_{ih} - \overline{X_h})(X_{il} - \overline{X_l}), \quad \overline{X_h} = \frac{1}{n}\sum_{i=1}^{n}X_{ih}.$$

The population counterpart is the **population covariance matrix**:

$$\begin{aligned} \text{Var}(\boldsymbol{X}_i) &= E[(\boldsymbol{X}_i - E[\boldsymbol{X}_i])(\boldsymbol{X}_i - E[\boldsymbol{X}_i])'] \\ &= E[\boldsymbol{X}_i\boldsymbol{X}_i'] - E[\boldsymbol{X}_i]E[\boldsymbol{X}_i]'. \end{aligned}$$

Sample correlation matrix

The **sample correlation coefficient** between the variables h and l is the standardized sample covariance:

$$r_{h,l} = \frac{\hat{\sigma}_{h,l}}{\hat{\sigma}_h \hat{\sigma}_l} = \frac{\sum_{i=1}^{n}(X_{ih} - \overline{X_h})(X_{il} - \overline{X_l})}{\sqrt{\sum_{i=1}^{n}(X_{ih} - \overline{X_h})^2}\sqrt{\sum_{i=1}^{n}(X_{il} - \overline{X_l})^2}}.$$

These coefficients form the **sample correlation matrix** R, expressed as:

$$R = D^{-1}\widehat{\Sigma}D^{-1},$$

where D is the diagonal matrix of sample standard deviations:

$$D = \text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_k) = \begin{pmatrix} \hat{\sigma}_1 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_k \end{pmatrix}$$

The covariance and correlation matrices are symmetric and positive semidefinite.

```
cor(dat)
```

```
                wage  education      female
wage       1.0000000 0.38398973 -0.16240519
education  0.3839897 1.00000000  0.04448972
female    -0.1624052 0.04448972  1.00000000
```

We find a strong positive correlation between `wage` and `education`, a substantial negative correlation between `wage` and `female`, and a negligible correlation between `education` and `female`.

3.10 R Code

statistics-sec03.R

4 Least squares

4.1 Regression Fundamentals

Regression Problem

The idea of regression analysis is to approximate a univariate dependent variable Y_i (also known as the regressand or response variable) as a function of the k-variate vector of the independent variables $\boldsymbol{X}_i$ (also known as regressors or predictor variables). The relationship is formulated as

$$Y_i \approx f(\boldsymbol{X}_i), \quad i = 1, \dots, n,$$

where $Y_1, \dots, Y_n$ is a univariate dataset for the dependent variable and $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ a k-variate dataset for the regressor variables.

The goal of the least squares method is to find the regression function that minimizes the squared difference between actual and fitted values of Y_i:

$$\min_{f(\cdot)} \sum_{i=1}^{n} (Y_i - f(\boldsymbol{X}_i))^2.$$

Linear Regression

If the regression function $f(\boldsymbol{X}_i)$ is linear in $\boldsymbol{X}_i$, i.e.,

$$f(\boldsymbol{X}_i) = b_1 + b_2 X_{i2} + \dots + b_k X_{ik} = \boldsymbol{X}_i' \boldsymbol{b}, \quad \boldsymbol{b} \in \mathbb{R}^k,$$

the minimization problem is known as the **ordinary least squares (OLS)** problem. The coefficient vector has k entries:

$$\boldsymbol{b} = (b_1, b_2, \dots, b_k)'.$$

To avoid the unrealistic constraint of the regression line passing through the origin, a constant term (intercept) is always included in $\boldsymbol{X}_i$, typically as the first regressor:

$$\boldsymbol{X}_i = (1, X_{i2}, \dots, X_{ik})'.$$

Despite its linear framework, linear regressions can be quite adaptable to nonlinear relationships by incorporating nonlinear transformations of the original regressors. Examples include polynomial terms (e.g., squared, cubic), interaction terms (combining different variables), and logarithmic transformations.

4.2 Ordinary least squares (OLS)

The **sum of squared errors** for a given coefficient vector $\boldsymbol{b} \in \mathbb{R}^k$ is defined as

$$S_n(\boldsymbol{b}) = \sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2 = \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2.$$

It is minimized by the **least squares coefficient vector**

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{b}\in\mathbb{R}^k} \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2.$$

Least squares coefficients

If the $k \times k$ matrix $(\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i')$ is invertible, the solution for the ordinary least squares problem is uniquely determined by

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i.$$

The **fitted values** or predicted values are

$$\widehat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik} = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}, \quad i = 1, ..., n.$$

The **residuals** are the difference between observed and fitted values:

$$\hat{u}_i = Y_i - \widehat{Y}_i = Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}, \quad i = 1, ..., n.$$

4.3 Simple linear regression (k=2)

A simple linear regression is a linear regression of a dependent variable Y on a constant and a single independent variable Z. I.e., we are interested in a regression function of the form

$$\boldsymbol{X}_i'\boldsymbol{b} = b_1 + b_2 Z_i.$$

The regressor vector is $\boldsymbol{X}_i = (1, Z_i)'$. Let's consider $Y = \log(\text{wage})$ and $Z = \text{education}$ from the following dataset with $n = 20$ observations:

Person	log(Wage)	Edu	Edu^2	Edu x log(Wage)
1	2.56	18	324	46.08
2	2.44	14	196	34.16
3	2.32	14	196	32.48
4	2.44	16	256	39.04
5	2.22	16	256	35.52
6	2.7	14	196	37.8
7	2.46	16	256	39.36
8	2.71	16	256	43.36
9	3.18	18	324	57.24
10	2.15	12	144	25.8
11	3.24	18	324	58.32
12	2.76	14	196	38.64
13	1.64	12	144	19.68
14	3.36	21	441	70.56
15	1.86	14	196	26.04
16	2.56	12	144	30.72
17	2.22	13	169	28.86
18	2.61	21	441	54.81
19	2.54	12	144	30.48
20	2.9	21	441	60.9
sum	**50.87**	**312**	**5044**	**809.85**

The OLS coefficients are

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} &= \Big(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_iY_i\\ &= \begin{pmatrix} n & \sum_{i=1}^{n}Z_i\\ \sum_{i=1}^{n}Z_i & \sum_{i=1}^{n}Z_i^2\end{pmatrix}^{-1}\begin{pmatrix}\sum_{i=1}^{n}Y_i\\ \sum_{i=1}^{n}Z_iY_i\end{pmatrix}\end{aligned}$$

Evaluate sums:

$$\sum_{i=1}^{n}\boldsymbol{X}_iY_i = \begin{pmatrix}50.87\\ 809.85\end{pmatrix}, \quad \sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i' = \begin{pmatrix}20 & 312\\ 312 & 5044\end{pmatrix}$$

OLS coefficients:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}20 & 312\\ 312 & 5044\end{pmatrix}^{-1}\begin{pmatrix}50.87\\ 809.85\end{pmatrix} = \begin{pmatrix}1.107\\ 0.092\end{pmatrix}$$

The fitted regression line is

$$1.107 + 0.092 \text{ education}$$

There is another, simpler formula for $\hat{\beta}_1$ and $\hat{\beta}_2$ in the simple linear regression. It can be expressed in terms of sample means and covariances:

Simple linear regression

The least squares coefficients in a simple linear regression can be written as

$$\hat{\beta}_2 = \frac{\hat{\sigma}_{YZ}}{\hat{\sigma}_Z^2}, \quad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{Z}, \tag{4.1}$$

where $\hat{\sigma}_{YZ}$ is the sample covariance between Y and Z, and $\hat{\sigma}_Z^2$ is the sample variance of Z.

4.4 Regression Plots

Line Fitting

Let's examine the linear relationship between average test scores and the student-teacher ratio:

```
data(CASchools, package = "AER")
CASchools$STR = CASchools$students/CASchools$teachers
CASchools$score = (CASchools$read+CASchools$math)/2
fit1 = lm(score ~ STR, data = CASchools)
fit1$coefficients
```

```
(Intercept)         STR
 698.932949   -2.279808
```

We have

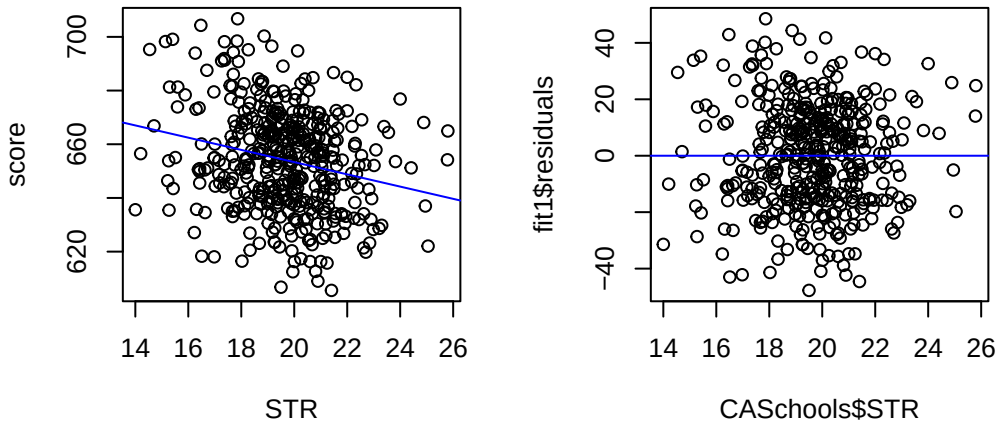
$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 698.9 \\ -2.28 \end{pmatrix}.$$

The fitted regression line is

$$698.9 - 2.28 \text{ STR}.$$

We can plot the regression line over a scatter plot of the data:

```
par(mfrow = c(1,2), cex=0.8)
plot(score ~ STR, data = CASchools)
abline(fit1, col="blue")
plot(CASchools$STR, fit1$residuals)
abline(0, 0, col="blue")
```



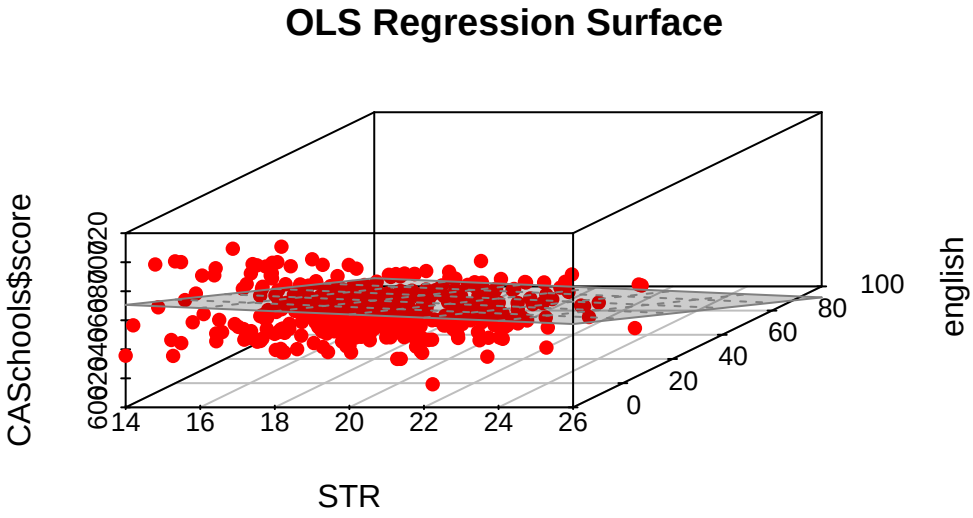
Multidimensional Visualizations

Let's include the percentage of english learners as an additional regressor:

```
fit2= lm(score ~ STR + english, data = CASchools)
fit2$coefficients
```

```
(Intercept)         STR     english
686.0322445  -1.1012956  -0.6497768
```

A 3D plot provides a visual representation of the resulting regression line (surface):



Adding the additional predictor `income` gives a regression specification with dimensions beyond visual representation:

```
fit3 = lm(score ~ STR + english + income, data = CASchools)
fit3$coefficients
```

```
 (Intercept)          STR      english       income
640.31549821  -0.06877542  -0.48826683   1.49451661
```

The fitted regression line now includes three predictors and four coefficients:

$$640.3 - 0.07 \text{ STR} - 0.49 \text{ english} + 1.49 \text{ income}$$

4.5 Matrix notation

OLS Formula

Matrix notation is convenient because it eliminates the need for summation symbols and indices. We define the response vector $\boldsymbol{Y}$ and the regressor matrix (design matrix) $\boldsymbol{X}$ as follows:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{12} & \dots & X_{1k} \\ \vdots & & & \vdots \\ 1 & X_{n2} & \dots & X_{nk} \end{pmatrix}$$

Note that $\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i' = \boldsymbol{X}'\boldsymbol{X}$ and $\sum_{i=1}^n \boldsymbol{X}_i Y_i = \boldsymbol{X}'\boldsymbol{Y}$.

The least squares coefficient vector becomes

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\Big)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

The vector of fitted values can be computed as follows:

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Projection Matrix

The vector of fitted values can be computed as follows:

$$\widehat{\boldsymbol{Y}} = \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{=\boldsymbol{P}}\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}.$$

The **projection matrix** $\boldsymbol{P}$ is also known as the *influence matrix* or *hat matrix* and maps observed values to fitted values.

The diagonal entries of $\boldsymbol{P}$, given by

$$h_{ii} = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i,$$

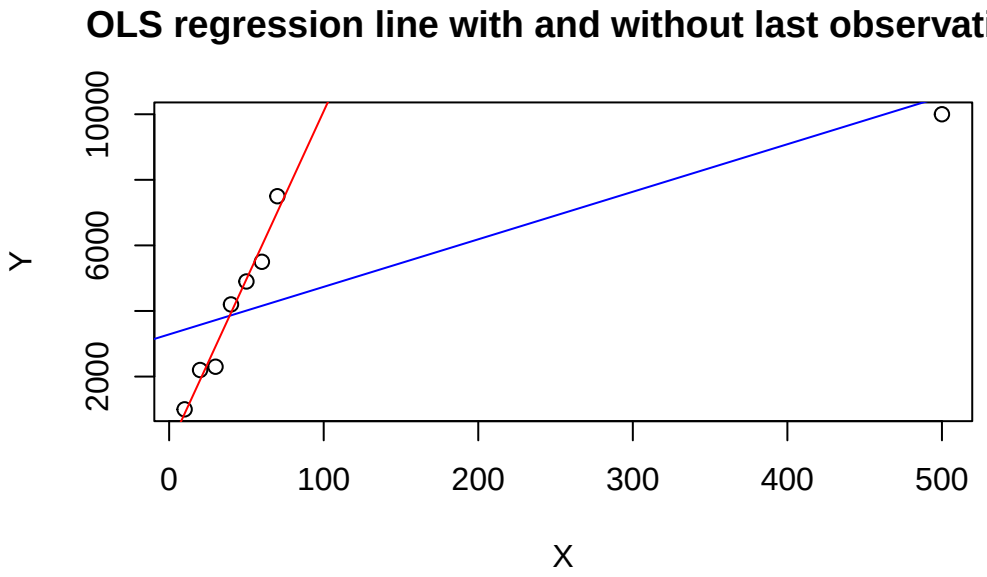
are called **leverage values** or *hat values* and measure how far away the regressor values of the i-th observation X_i are from those of the other observations.

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:

```
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
```

Residuals

The vector of residuals is given by

$$\hat{\boldsymbol{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

An important property of the residual vector is: $\boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{0}$. To see that this property holds, let's rearrange the OLS formula:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad \Leftrightarrow \quad \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}.$$

The dependent variable vector can be decomposed into the vector of fitted values and the residual vector:

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}.$$

Substituting this into the OLS formula from above gives:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}) \quad \Leftrightarrow \quad \boldsymbol{0} = \boldsymbol{X}'\hat{\boldsymbol{u}}.$$

This property has a geometric interpretation: it means the residuals are orthogonal to all regressors. This makes sense because if there were any linear relationship left between the residuals and the regressors, we could have captured it in our model to improve the fit.

4.6 Goodness of Fit

Analysis of Variance

The orthogonality property of the residual vector can be written in a more detailed way as follows:

$$\boldsymbol{X}'\hat{\boldsymbol{u}} = \begin{pmatrix} \sum_{i=1}^n \hat{u}_i \\ \sum_{i=1}^n X_{i2}\hat{u}_i \\ \vdots \\ \sum_{i=1}^n X_{ik}\hat{u}_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.2}$$

In particular, the sample mean of the residuals is zero:

$$\frac{1}{n}\sum_{i=1}^n \hat{u}_i = 0.$$

Therefore, the sample variance of the residuals is simply the sample mean of squared residuals:

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

The sample variance of the dependent variable is

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

and the sample variance of the fitted values is

$$\hat{\sigma}^2_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2.$$

The three sample variances are connected through the analysis of variance formula:

$$\hat{\sigma}^2_{Y} = \hat{\sigma}^2_{\widehat{Y}} + \hat{\sigma}^2_{\hat{u}}.$$

Hence, the larger the proportion of the explained sample variance, the better the fit of the OLS regression.

R-squared

The analysis of variance formula motivates the definition of the **R-squared coefficient**:

$$R^2 = 1 - \frac{\hat{\sigma}^2_{\hat{u}}}{\hat{\sigma}^2_{Y}} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}.$$

The R-squared describes the proportion of sample variation in $\boldsymbol{Y}$ explained by $\widehat{\boldsymbol{Y}}$. We have $0 \leq R^2 \leq 1$.

In a regression of Y_i on a single regressor Z_i with intercept (simple linear regression), the R-squared is equal to the squared sample correlation coefficient of Y_i and Z_i.

An R-squared of 0 indicates no sample variation in $\widehat{\boldsymbol{Y}}$ (a flat regression line/surface), whereas a value of 1 indicates no variation in $\hat{\boldsymbol{u}}$, indicating a perfect fit. The higher the R-squared, the better the OLS regression fits the data.

However, a low R-squared does not necessarily mean the regression specification is bad. It just implies that there is a high share of unobserved heterogeneity in $\boldsymbol{Y}$ that is not captured by the regressors $\boldsymbol{X}$ linearly.

Conversely, a high R-squared 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.

Degree of Freedom Corrections

Adjusted Sample Variance

When computing the sample mean $\overline{Y}$, we have n degrees of freedom because all data points $Y_1, \dots, Y_n$ can vary freely.

When computing variances, we take the sample mean of the squared deviations

$$(Y_1 - \overline{Y})^2, \dots, (Y_n - \overline{Y})^2.$$

These elements cannot vary freely because $\overline{Y}$ is computed from the same sample and implies the constraint

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0.$$

This means that the deviations are connected by this equation and are not all free to vary. Knowing the first $n-1$ of the deviations determines the last one:

$$(Y_n - \overline{Y}) = -\sum_{i=1}^{n-1}(Y_i - \overline{Y}).$$

Therefore, only $n-1$ deviations can vary freely, which results in $n-1$ degrees of freedom for the sample variance.

Because $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ effectively contains only $n-1$ freely varying summands, it is common to account for this fact. The **adjusted sample variance** uses $n-1$ in the denominator:

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

The adjusted sample variance relates to the unadjusted sample variance as:

$$s_Y^2 = \frac{n}{n-1}\hat{\sigma}_Y^2.$$

Its square root, $s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$, is the adjusted sample standard deviation. Note: the built-in `R` functions `var(Y)` and `sd(Y)` compute the adjusted versions of the sample variance and standard deviations.

Adjusted Residual Variance

For the sample variance of $\hat{\boldsymbol{u}}$, we lose k degrees of freedom because the residuals are subject to the constraints from Equation 4.2. The adjusted sample variance of the residuals is therefore defined as:

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

The square root of the adjusted sample variance of the residuals is called the **standard error of the regression (SER)** or **residual standard error**:

$$SER := s_{\hat{u}} = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2}.$$

The square root of the unadjusted sample variance of the residuals is also called the **Root Mean Squared Error (RMSE)**:

$$RMSE(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_{\hat{u}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}.$$

Adjusted R-squared

By incorporating adjusted versions of the sample variances in the R-squared definition, we penalize regression specifications with large k. The **adjusted R-squared** is

$$\overline{R}^2 = 1 - \frac{\frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}.$$

	(1)	(2)	(3)
(Intercept)	698.933	686.032	640.315
STR	−2.280	−1.101	−0.069
english		−0.650	−0.488
income			1.495
Num.Obs.	420	420	420
R2	0.051	0.426	0.707
R2 Adj.	0.049	0.424	0.705
RMSE	18.54	14.41	10.30

The R-squared should be used for interpreting the share of variation explained by the fitted regression line. The adjusted R-squared should be used for comparing different OLS regression specifications.

4.7 Regression Table

The `modelsummary()` function can be used to produce comparison tables of regression outputs:

```
library(modelsummary)
mymodels = list(fit1, fit2, fit3)
modelsummary(mymodels,
             statistic = NULL,
             gof_map = c("nobs", "r.squared", "adj.r.squared", "rmse"))
```

Model (3) explains the most variation in test scores and provides the best fit to the data, as indicated by the highest R^2 and the lowest residual standard error.

In model (1), schools with one more student per class are predicted to have a 2.28-point lower test score. This effect decreases to 1.1 points in model (2), after accounting for the percentage of English learners, and drops further to just 0.07 points in model (3), once income is also included.

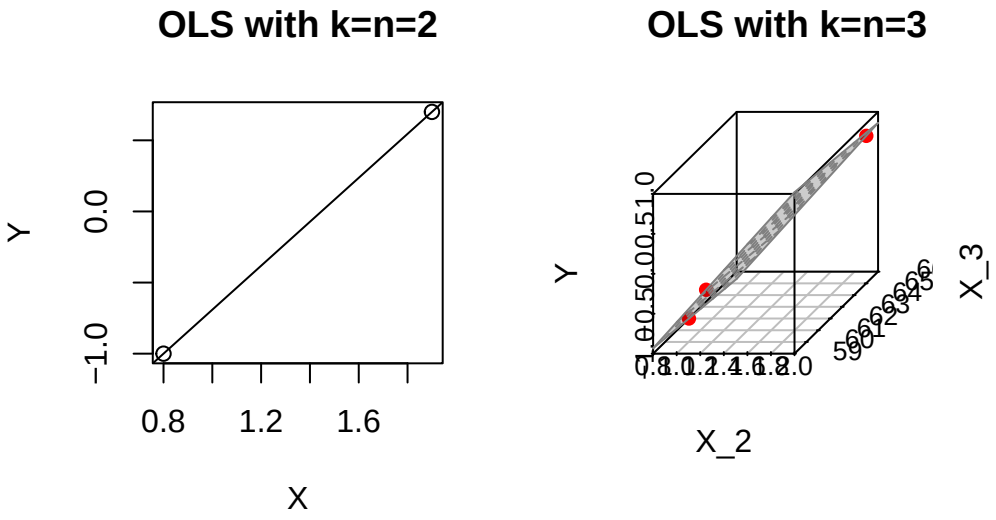
While the R-squared increases in the number of regressors, the RMSE decreases.

To give deeper meaning to these results and understand their interpretation within a broader context, we turn to a formal probabilistic model framework in the next section.

4.8 When OLS Fails

Too many regressors

OLS should be considered for regression problems with $k << n$ (small k and large n). When the number of predictors k approaches or equals the number of observations n, we run into the problem of overfitting. Specifically, at $k = n$, the regression line will perfectly fit the data.



If $k = n \geq 4$, we can no longer visualize the OLS regression line in the 3D space, but the problem of a perfect fit is still present. If $k > n$, there exists no unique OLS solution because $\boldsymbol{X'X}$ is not invertible. Regression problems with $k \approx n$ or $k > n$ are called **high-dimensional regressions**.

Perfect multicollinearity

The only requirement for computing the OLS coefficients is the invertibility of the matrix $\boldsymbol{X'X}$. As discussed above, a necessary condition is that $k \leq n$.

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. Multicollinearity arises if one variable is a linear combination of the other variables.

Common causes are duplicating a regressor or using the same variable in different units (e.g., GDP in both EUR and USD).

Perfect multicollinearity (or strict multicollinearity) arises 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 $\widehat{\boldsymbol{\beta}}$ cannot be computed.

Near multicollinearity 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.

If $k \leq n$ and multicollinearity is present, it means that at least one regressor is redundant and can be dropped.

Dummy variable trap

A common cause of strict multicollinearity is the inclusion of too many dummy variables. Let's consider the `cps` data and add a dummy variable for non-married individuals:

```
cps = read.csv("cps.csv")
cps$nonmarried = 1-cps$married
fit4 = lm(wage ~ married + nonmarried, data = cps)
fit4$coefficients
```

```
(Intercept)     married  nonmarried
  19.305329    6.920139          NA
```

The coefficient for `nonmarried` is `NA`. We fell into the dummy variable trap!

The dummy variables `married` and `nonmarried` are collinear with the intercept variable because $married + nonmarried = 1$, which leads to a singular matrix $\boldsymbol{X}'\boldsymbol{X}$ and therefore to perfect multicollinearity.

The solution is to use one dummy variable less than factor levels, as R automatically does by omitting the last dummy variable. Another solution would be to remove the intercept from the model, which can be done by adding `-1` to the model formula:

```
fit5 = lm(wage ~ married + nonmarried - 1, data = cps)
fit5$coefficients
```

```
   married nonmarried
  26.22547   19.30533
```

4.9 R Code

statistics-sec04.R

5 Regression

5.1 Conditional Expectation

In econometrics, we often analyze how a variable of interest (like wages) varies systematically with other variables (like education or experience). The **conditional expectation function (CEF)** provides a powerful framework for describing these relationships.

The conditional expectation of a random variable Y given a random vector $\boldsymbol{X}$ is the expected value of Y given any possible value of $\boldsymbol{X}$. Using the conditional CDF, the conditional expectation (or conditional mean) is

$$E[Y|\boldsymbol{X}=\boldsymbol{x}] = \int_{-\infty}^{\infty} y \, dF_{Y|\boldsymbol{X}=\boldsymbol{x}}(y).$$

For a continuous random variable Y we have

$$E[Y|\boldsymbol{X}=\boldsymbol{x}] = \int_{-\infty}^{\infty} y \, f_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) \, dy,$$

where $f_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ is the conditional density of Y given $\boldsymbol{X}=\boldsymbol{x}$.

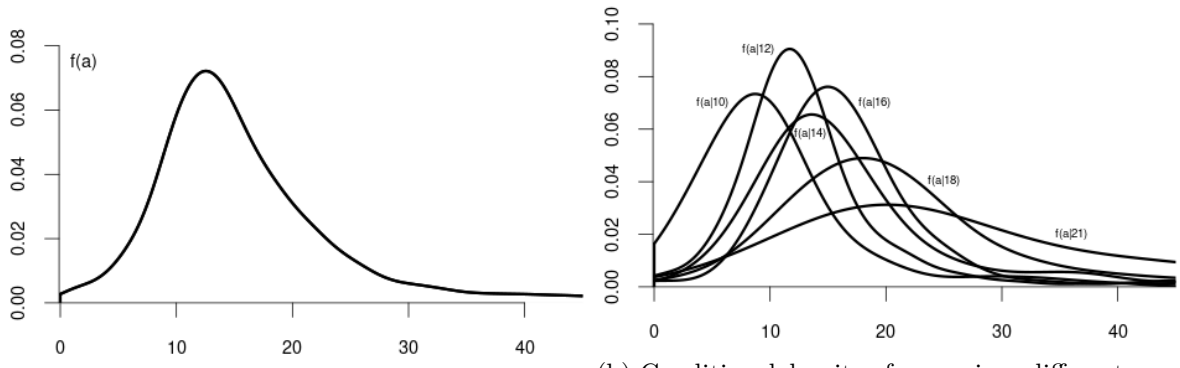
When Y is discrete with support $\mathcal{Y}$, we have

$$E[Y|\boldsymbol{X}=\boldsymbol{x}] = \sum_{y\in\mathcal{Y}} y \, \pi_{Y|\boldsymbol{X}=\boldsymbol{x}}(y).$$

The CEF maps values of $\boldsymbol{X}$ to corresponding conditional means of Y. As a function of the random vector $\boldsymbol{X}$, the CEF itself is a random variable:

$$E[Y|\boldsymbol{X}] = m(\boldsymbol{X}), \quad \text{where } m(\boldsymbol{x}) = E[Y|\boldsymbol{X}=\boldsymbol{x}]$$

For a comprehensive treatment of conditional expectations see Probability Tutorial Part 2



(a) Unconditional density of wage

(b) Conditional density of wage given different years of education

Figure 5.1: Unconditional density $f_Y(y)$ and conditional densities $f_{Y|X=x}(y)$ of wage given x years of education

Examples

Let's examine this concept using wage and education as examples. When X is univariate and discrete (such as years of education), we can analyze how wage distributions change across education levels by comparing their **conditional distributions**:

Notice how the conditional distributions tend to shift rightward as education increases, indicating higher average wages with higher education.

From these conditional densities, we can compute the expected wage for each education level. Plotting these conditional expectations gives the CEF:

$$m(x) = E[\text{wage} \mid \text{edu} = x]$$

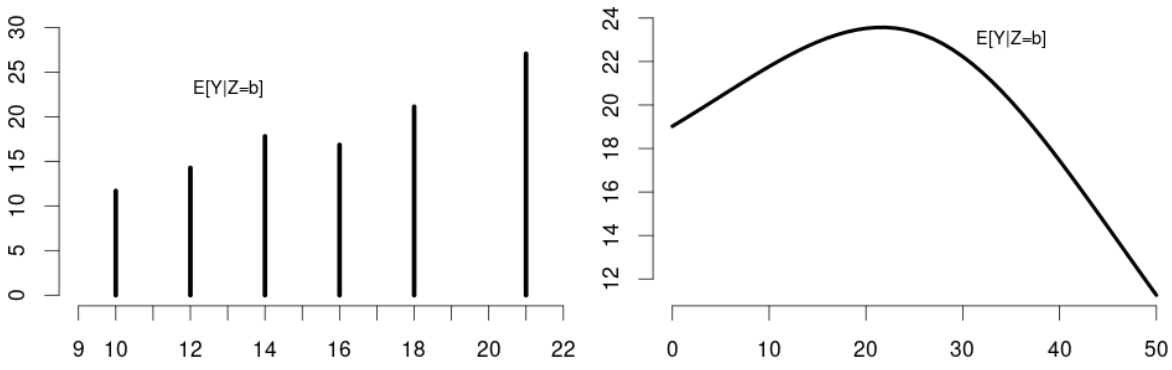
Since education is discrete, the CEF is defined only at specific values, as shown in the left plot below:

When X is continuous (like years of experience), the CEF is often a smooth function (right plot). The shape of $E[\text{wage}|\text{experience}]$ reflects real-world patterns: wages rise quickly early in careers, then plateau, and may eventually decline near retirement.

The CEF as a Random Variable

It's important to distinguish between:

- $E[Y|\boldsymbol{X} = \boldsymbol{x}]$: a **number** (the conditional mean at a specific value)
- $E[Y|\boldsymbol{X}]$: a **function of $\boldsymbol{X}$**, which is itself a **random variable**



(a) CEF of wage given education (b) CEF of wage given experience

Figure 5.2: Conditional expectations of wage given education (left) and experience (right)

For instance, if $X =$ education has the probability mass function:

$$P(X = x) = \begin{cases} 0.06 & \text{if } x = 10 \\ 0.43 & \text{if } x = 12 \\ 0.16 & \text{if } x = 14 \\ 0.08 & \text{if } x = 16 \\ 0.24 & \text{if } x = 18 \\ 0.03 & \text{if } x = 21 \\ 0 & \text{otherwise} \end{cases}$$

Then $E[Y|X]$ as a random variable has the probability mass function:

$$P(E[Y|X] = y) = \begin{cases} 0.06 & \text{if } y = 11.68 \text{ (when } X = 10) \\ 0.43 & \text{if } y = 14.26 \text{ (when } X = 12) \\ 0.16 & \text{if } y = 17.80 \text{ (when } X = 14) \\ 0.08 & \text{if } y = 16.84 \text{ (when } X = 16) \\ 0.24 & \text{if } y = 21.12 \text{ (when } X = 18) \\ 0.03 & \text{if } y = 27.05 \text{ (when } X = 21) \\ 0 & \text{otherwise,} \end{cases}$$

where the values for y are taken from Figure 5.2a.

The CEF assigns to each value of $\boldsymbol{X}$ the expected value of Y given that information.

5.2 CEF Properties

The conditional expectation function has several important properties that make it a fundamental tool in econometric analysis.

Law of Iterated Expectations (LIE)

The **law of iterated expectations** connects conditional and unconditional expectations:

$$E[Y] = E[E[Y|\boldsymbol{X}]]$$

This means that to compute the overall average of Y, we can first compute the average of Y within each group defined by $\boldsymbol{X}$, then average those conditional means using the distribution of $\boldsymbol{X}$.

This is analogous to the law of total probability, where we compute marginal probabilities or densities as weighted averages of conditional ones:

For simplicity consider a univariate conditioning random variable X. When X is discrete:

$$P(Y = y) = \sum_x P(Y = y \mid X = x) \cdot P(X = x)$$

When X is continuous:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y) \cdot f_X(x)\, dx$$

Similarly, the LIE states:

When X is discrete:

$$E[Y] = \sum_x E[Y \mid X = x] \cdot P(X = x)$$

When X is continuous:

$$E[Y] = \int_{-\infty}^{\infty} E[Y \mid X = x] \cdot f_X(x)\, dx$$

Let's apply this to our wage and education example. With X = education and Y = wage, we have:

$$
\begin{aligned}
E[Y|X=10] &= 11.68, \qquad P(X=10) = 0.06 \\
E[Y|X=12] &= 14.26, \qquad P(X=12) = 0.43 \\
E[Y|X=14] &= 17.80, \qquad P(X=14) = 0.16 \\
E[Y|X=16] &= 16.84, \qquad P(X=16) = 0.08 \\
E[Y|X=18] &= 21.12, \qquad P(X=18) = 0.24 \\
E[Y|X=21] &= 27.05, \qquad P(X=21) = 0.03
\end{aligned}
$$

The law of iterated expectations gives us:

$$
\begin{aligned}
E[Y] &= \sum_x E[Y|X=x] \cdot P(X=x) \\
&= 11.68 \cdot 0.06 + 14.26 \cdot 0.43 + 17.80 \cdot 0.16 \\
&\quad + 16.84 \cdot 0.08 + 21.12 \cdot 0.24 + 27.05 \cdot 0.03 \\
&= 0.7008 + 6.1318 + 2.848 + 1.3472 + 5.0688 + 0.8115 \\
&= 16.91
\end{aligned}
$$

This unconditional expected wage of 16.91 aligns with what we would calculate from the unconditional density from Figure 5.1a.

The LIE provides us with a powerful way to bridge conditional expectations (within education groups) and the overall unconditional expectation (averaging across all education levels).

Conditioning Theorem (CT)

The **conditioning theorem** (also called the factorization rule) states:

$$E[g(\boldsymbol{X})Y|\boldsymbol{X}] = g(\boldsymbol{X}) \cdot E[Y|\boldsymbol{X}]$$

This means that when taking the conditional expectation of a product where one factor is a function of the conditioning variable, that factor can be treated as a constant and factored out. Once we condition on $\boldsymbol{X}$, the value of $g(\boldsymbol{X})$ is fixed.

If $Y =$ wage and $X =$ education, then for someone with 16 years of education:

$$E[16 \cdot \text{wage} \mid \text{edu} = 16] = 16 \cdot E[\text{wage} \mid \text{edu} = 16]$$

More generally, if we want to find the expected product of education and wage, conditional on education:

$$E[\text{edu} \cdot \text{wage} \mid \text{edu}] = \text{edu} \cdot E[\text{wage} \mid \text{edu}]$$

Best Predictor Property

If $E[Y^2] < \infty$, the conditional expectation $E[Y|\boldsymbol{X}]$ is the **best predictor** of Y given $\boldsymbol{X}$ in terms of mean squared error, i.e.:

$$E[Y|\boldsymbol{X}] = \arg\min_{g(\cdot)} E[(Y - g(\boldsymbol{X}))^2]$$

This means that among all possible functions of $\boldsymbol{X}$, the CEF minimizes the expected squared prediction error. In practical terms, if you want to predict wages based only on education, the optimal prediction is exactly the conditional mean wage for each education level.

For example, if someone has 18 years of education, our best prediction of their wage (minimizing expected squared error) is $E[\text{wage}|\text{education} = 18] = 21.12$.

No other function of education, whether linear, quadratic, or more complex, can yield a better prediction in terms of expected squared error than the CEF itself.

Proof sketch: Add and subtract $m(\boldsymbol{X}) = E[Y|\boldsymbol{X}]$:

$$\begin{aligned}
&E[(Y - g(\boldsymbol{X}))^2] \\
&= E[(Y - m(\boldsymbol{X}) + m(\boldsymbol{X}) - g(\boldsymbol{X}))^2] \\
&= E[(Y - m(\boldsymbol{X}))^2] \\
&\quad + 2E[(Y - m(\boldsymbol{X}))(m(\boldsymbol{X}) - g(\boldsymbol{X}))] \\
&\quad + E[(m(\boldsymbol{X}) - g(\boldsymbol{X}))^2]
\end{aligned}$$

- The first term is finite and does not depend on $g(\cdot)$.
- The cross term is zero by the LIE and CT.
- The last term is minimal if $g(\boldsymbol{X}) = m(\boldsymbol{X})$.

Independence Implications

If Y and $\boldsymbol{X}$ are independent, then:

$$E[Y|\boldsymbol{X}] = E[Y]$$

When variables are independent, knowing $\boldsymbol{X}$ provides no information about Y, so the conditional expectation equals the unconditional expectation. The CEF becomes a constant function that doesn't vary with $\boldsymbol{X}$.

In our wage example, if education and wage were completely independent, the CEF would be a horizontal line at the overall average wage of 16.91. Each conditional density $f_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)$ would be identical to the unconditional density $f_Y(y)$, and the conditional means would all equal the unconditional mean.

The fact that our CEF for wage given education has a positive slope indicates that these variables are not independent – higher education is associated with higher expected wages.

5.3 Linear Model Specification

Prediction Error

Consider a sample $(Y_i, \boldsymbol{X}_i')$, $i = 1, \ldots, n$. We have established that the **conditional expectation function (CEF)** $E[Y_i|\boldsymbol{X}_i]$ is the best predictor of Y_i given $\boldsymbol{X}_i$, minimizing the mean squared prediction error.

This leads to the following prediction error:

$$u_i = Y_i - E[Y_i|\boldsymbol{X}_i]$$

By construction, this error has a conditional mean of zero:

$$E[u_i|\boldsymbol{X}_i] = 0$$

This property follows directly from the law of iterated expectations:

$$\begin{aligned} E[u_i|\boldsymbol{X}_i] &= E[Y_i - E[Y_i|\boldsymbol{X}_i] \mid \boldsymbol{X}_i] \\ &= E[Y_i|\boldsymbol{X}_i] - E[E[Y_i|\boldsymbol{X}_i] \mid \boldsymbol{X}_i] \\ &= E[Y_i|\boldsymbol{X}_i] - E[Y_i|\boldsymbol{X}_i] = 0 \end{aligned}$$

We can thus always decompose the outcome as:

$$Y_i = E[Y_i|\boldsymbol{X}_i] + u_i$$

where $E[u_i|\boldsymbol{X}_i] = 0$. This equation is not yet a regression model. It's simply the decomposition of Y_i into its conditional expectation and an unpredictable component.

Linear Regression Model

To move to a regression framework, we impose a structural assumption about the form of the CEF. The key assumption of the **linear regression model** is that the conditional expectation is a **linear function** of the regressors:

$$E[Y_i|\boldsymbol{X}_i] = \boldsymbol{X}_i'\boldsymbol{\beta}$$

Substituting this into our decomposition yields the **linear regression equation**:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i \tag{5.1}$$

with the crucial assumption:

$$E[u_i|\boldsymbol{X}_i] = 0 \tag{5.2}$$

Exogeneity

This assumption (Equation 5.2) is called **exogeneity** or **mean independence**. It ensures that the linear function $\boldsymbol{X}_i'\boldsymbol{\beta}$ correctly captures the conditional mean of Y_i.

Under the linear regression equation (Equation 5.1) we have the following equivalence:

$$E[Y_i|\boldsymbol{X}_i] = \boldsymbol{X}_i'\boldsymbol{\beta} \quad \Leftrightarrow \quad E[u_i|\boldsymbol{X}_i] = 0$$

Therefore, the linear regression model in its most general form is characterized by the two conditions: linear regression equation (Equation 5.1) and exogenous regressors (Equation 5.2).

For example, in a wage regression, exogeneity means that the expected wage conditional on education and experience is exactly captured by the linear combination of these variables. No systematic pattern remains in the error term.

Model Misspecification

If the true conditional expectation function is nonlinear (e.g., if wages increase with education at a diminishing rate), then $E[Y_i|\boldsymbol{X}_i] \neq \boldsymbol{X}_i'\boldsymbol{\beta}$, and the model is **misspecified**. In such cases, the linear model provides the best linear approximation to the true CEF, but systematic patterns remain in the error term.

It's important to note that u_i may still be statistically dependent on $\boldsymbol{X}_i$ in ways other than its mean. For example, the **variance** of u_i may depend on $\boldsymbol{X}_i$ in the case of **heteroskedasticity**. For instance, wage dispersion might increase with education level. The assumption $E[u_i|\boldsymbol{X}_i] = 0$ requires only that the conditional mean of the error is zero, not that the error is completely independent of the regressors.

5.4 Population Regression Coefficient

Under the linear regression model

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, \quad E[u_i|\boldsymbol{X}_i] = 0,$$

we are interested in the **population regression coefficient** $\boldsymbol{\beta}$, which indicates how the conditional mean of Y_i varies **linearly** with the regressors in $\boldsymbol{X}_i$.

Moment Condition

A key implication of the exogeneity condition $E[u_i|\boldsymbol{X}_i] = 0$ is that the regressors are **mean uncorrelated** with the error term:

$$E[\boldsymbol{X}_i u_i] = \mathbf{0}$$

This can be derived from the exogeneity condition using the LIE:

$$E[\boldsymbol{X}_i u_i] = E[E[\boldsymbol{X}_i u_i \mid \boldsymbol{X}_i]] = E[\boldsymbol{X}_i \cdot E[u_i|\boldsymbol{X}_i]] = E[\boldsymbol{X}_i \cdot 0] = \mathbf{0}$$

Substituting the linear model into the mean uncorrelatedness condition gives a moment condition that identifies $\boldsymbol{\beta}$:

$$\mathbf{0} = E[\boldsymbol{X}_i u_i] = E[\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})] = E[\boldsymbol{X}_i Y_i] - E[\boldsymbol{X}_i \boldsymbol{X}_i']\boldsymbol{\beta}$$

Rearranging to solve for $\boldsymbol{\beta}$:

$$E[\boldsymbol{X}_i Y_i] = E[\boldsymbol{X}_i \boldsymbol{X}_i']\boldsymbol{\beta}$$

Assuming that the matrix $E[\boldsymbol{X}_i \boldsymbol{X}_i']$ is invertible, we can express the population regression coefficient as:

$$\boldsymbol{\beta} = \left(E[\boldsymbol{X}_i \boldsymbol{X}_i']\right)^{-1} E[\boldsymbol{X}_i Y_i] \tag{5.3}$$

This expression shows that $\boldsymbol{\beta}$ is entirely determined by the joint distribution of $(Y_i, \boldsymbol{X}_i')$ in the population.

The invertibility of $E[\boldsymbol{X}_i \boldsymbol{X}_i']$ is guaranteed if there is no perfect linear relationship among the regressors. In particular, no pair of regressors should be perfectly correlated, and no regressor should be a perfect linear combination of the other regressors.

OLS Estimation

Recall that we have estimated population moments like $E[Y]$ and $\text{Var}(Y)$ by their sample counterparts, i.e. $\overline{Y}$ and $\widehat{\sigma}_Y^2$. This estimation principle is known as the **method of moments**, where we replace population moments by their corresponding sample moments.

To estimate the population regression coefficient

$$\boldsymbol{\beta} = \left(E[\boldsymbol{X}_i \boldsymbol{X}_i']\right)^{-1} E[\boldsymbol{X}_i Y_i]$$

using a given i.i.d. sample $(Y_i, \boldsymbol{X}_i')$, $i = 1, \dots, n$, we replace all population moments by their sample counterparts, i.e.,

$$\widehat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i Y_i\right).$$

This can be simplified to the familiar form

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i Y_i\right),$$

or $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, which is called the **ordinary least squares (OLS) estimator**.

5.5 Consistency

Recall that the law of large numbers for a univariate i.i.d. dataset $Y_1, \dots, Y_n$ states that the sample average converges in probability to the population mean:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E[Y] \quad \text{as } n \to \infty.$$

The OLS estimator is a function of two sample averages: the sample second moment matrix $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'$ and the sample cross-moment vector $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i Y_i$.

If $(Y_i, \boldsymbol{X}_i')$, $i = 1, \dots, n$, are i.i.d., then the multivariate version of the law of large numbers applies:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \xrightarrow{p} E[\boldsymbol{X}_i \boldsymbol{X}_i'], \quad \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i Y_i \xrightarrow{p} E[\boldsymbol{X}_i Y_i].$$

This means that convergence in probability holds componentwise. Each element of the sample moment matrix and vector converges to its corresponding population counterpart.

The continuous mapping theorem and Slutsky's lemma enable us to extend these convergence results to more complex expressions.

- If $f(\cdot)$ is a continuous function and $V_n \xrightarrow{p} c$, then $f(V_n) \xrightarrow{p} f(c)$ (continuous mapping theorem).
- If $V_n \xrightarrow{p} c$ and $W_n \xrightarrow{p} d$ then $V_n W_n \xrightarrow{p} cd$ (Slutsky's lemma).

Since matrix inversion is a continuous function, the continuous mapping theorem implies:

$$\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \xrightarrow{p} \left(E[\boldsymbol{X}_i \boldsymbol{X}_i']\right)^{-1}.$$

Applying Slutsky's lemma to combine the two convergence results yields:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i Y_i\right) \\ &\xrightarrow{p} \left(E[\boldsymbol{X}_i \boldsymbol{X}_i']\right)^{-1} E[\boldsymbol{X}_i Y_i] = \boldsymbol{\beta}.\end{aligned}$$

This establishes the consistency of the OLS estimator. We used the following regularity conditions:

1) **Random sampling**: $(Y_i, \boldsymbol{X}_i')$ are i.i.d.
2) **Exogeneity (mean independence)**: $E[u_i|\boldsymbol{X}_i] = 0$.
3) **Finite second moments**: $E[X_{ij}^2] < \infty$ and $E[Y_i^2] < \infty$.
4) **Full rank**: $E[\boldsymbol{X}_i\boldsymbol{X}_i']$ is positive definite (hence invertible).

Neither normality nor homoskedasticity is required for consistency. Heteroskedasticity is fully compatible with OLS consistency.

For any two random variables Y and Z, the Cauchy-Schwarz inequality states $|E[YZ]| \leq \sqrt{E[Y^2]E[Z^2]}$. Specifically, $|E[X_{ik}X_{il}]| \leq \sqrt{E[X_{ik}^2]E[X_{il}^2]}$ and $|E[X_{ik}Y_i]| \leq \sqrt{E[X_{ik}^2]E[Y_i^2]}$.

Therefore, the finite second moment condition ensures that $E[\boldsymbol{X}_i\boldsymbol{X}_i']$ and $E[\boldsymbol{X}_iY_i]$ are finite. The full rank condition ensures that $E[\boldsymbol{X}_i\boldsymbol{X}_i']^{-1}$ exists. Thus, the full rank and finite second moments conditions ensure that $\boldsymbol{\beta}$ is well-defined.

The exogeneity condition is crucial for OLS consistency. Without it, the model is misspecified, Equation 5.3 does not hold, and the OLS estimator would converge to the best linear predictor, which is the $\boldsymbol{\beta}^*$ that minimizes $E[(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2]$.

Just as with the univariate law of large numbers, the i.i.d. assumption can be relaxed to accommodate other sampling schemes. Under clustered sampling with independent clusters, OLS consistency holds if the number of clusters grows large relative to cluster size as $n \to \infty$. For time series data, $(Y_i, \boldsymbol{X}_i')$ must be stationary, and observations $(Y_i, \boldsymbol{X}_i')$ and $(Y_{i-j}, \boldsymbol{X}_{i-j}')$ must become independent as j increases (strong mixing / weak dependence).

5.6 R Code

statistics-sec05.R

6 Effects

6.1 Marginal Effects

Consider the regression model of hourly wage on education (years of schooling),

$$\text{wage}_i = \beta_1 + \beta_2 \text{edu}_i + u_i, \quad i = 1, ..., n,$$

where the exogeneity assumption holds:

$$E[u_i|\text{edu}_i] = 0.$$

The population regression function, which gives the conditional expectation of wage given education, can be derived as:

$$\begin{aligned} m(\text{edu}_i) &= E[\text{wage}_i|\text{edu}_i] \\ &= \beta_1 + \beta_2 \cdot \text{edu}_i + E[u_i|\text{edu}_i] \\ &= \beta_1 + \beta_2 \cdot \text{edu}_i \end{aligned}$$

Thus, the average wage level of all individuals with z years of schooling is:

$$m(z) = \beta_1 + \beta_2 \cdot z.$$

Interpretation of Coefficients

In the linear regression model

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i,$$

the coefficient vector $\boldsymbol{\beta}$ captures the way the **conditional mean of** Y_i changes with the regressors $\boldsymbol{X}_i$. Under the exogeneity assumption, we have

$$E[Y_i|\boldsymbol{X}_i] = \boldsymbol{X}_i'\boldsymbol{\beta} = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik}.$$

This linearity allows for a simple interpretation. The coefficient β_j represents the **partial derivative** of the conditional mean with respect to X_{ij}:

$$\frac{\partial E[Y_i|\boldsymbol{X}_i]}{\partial X_{ij}} = \beta_j.$$

This means that β_j measures the **marginal effect** of a one-unit increase in X_{ij} on the expected value of Y_i, holding all other variables constant.

If X_{ij} is a dummy variable (i.e., binary), then β_j measures the discrete change in $E[Y_i|\boldsymbol{X}_i]$ when X_{ij} changes from 0 to 1.

For our wage-education example, the marginal effect of education is:

$$\frac{\partial E[\text{wage}_i|\text{edu}_i]}{\partial \text{edu}_i} = \beta_2.$$

This population marginal effect parameter can be estimated using OLS:

```
cps = read.csv("cps.csv")
lm(wage ~ education, data = cps)
```

```
Call:
lm(formula = wage ~ education, data = cps)

Coefficients:
(Intercept)    education
    -16.448        2.898
```

Interpretation: People with one more year of education are paid on average \$2.90 USD more per hour than people with one year less of education, assuming the exogeneity condition holds.

Correlation vs. Causation

The coefficient β_2 describes the **correlative relationship** between education and wages, not necessarily a causal one. To see this connection to correlation, consider the covariance of the two variables:

$$\begin{aligned} Cov(\text{wage}_i, \text{edu}_i) &= Cov(\beta_1 + \beta_2 \cdot \text{edu}_i + u_i, \text{edu}_i) \\ &= Cov(\beta_1 + \beta_2 \cdot \text{edu}_i, \text{edu}_i) + Cov(u_i, \text{edu}_i) \end{aligned}$$

The term $Cov(u_i, \text{edu}_i)$ equals zero due to the exogeneity assumption. To see this, recall that $E[u_i] = E[E[u_i|\text{edu}_i]] = 0$ by the LIE, and similarly

$$E[u_i \text{edu}_i] = E[E[u_i \text{edu}_i|\text{edu}_i]] = E[E[u_i|\text{edu}_i]\text{edu}_i] = 0,$$

which implies

$$Cov(u_i, \text{edu}_i) = E[u_i \text{edu}_i] - E[u_i] \cdot E[\text{edu}_i] = 0$$

The coefficient β_2 is thus proportional to the population correlation coefficient:

$$\beta_2 = \frac{Cov(\text{wage}_i, \text{edu}_i)}{Var(\text{edu}_i)} = Corr(\text{wage}_i, \text{edu}_i) \cdot \frac{sd(\text{wage}_i)}{sd(\text{edu}_i)}.$$

The marginal effect is a correlative effect and does not necessarily reveal the source of the higher wage levels for people with more education.

Regression relationships do not necessarily imply causal relationships.

People with more education may earn more for various reasons:

- They might be naturally more talented or capable
- They might come from wealthier families with better connections
- They might have access to better resources and opportunities
- Education itself might actually increase productivity and earnings

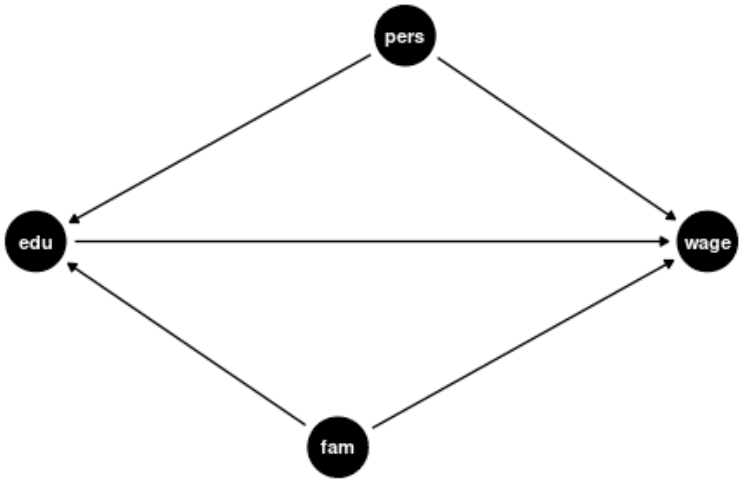


Figure 6.1: A DAG (directed acyclic graph) showing potential confounding factors in the education-wage relationship

The coefficient β_2 measures how strongly education and earnings are correlated, but this association could be due to other factors that correlate with both wages and education, such as:

- Family background (parental education, family income, ethnicity)
- Personal background (gender, intelligence, motivation)

Remember: **Correlation does not imply causation!**

Omitted Variable Bias

To understand the causal effect of an additional year of education on wages, it is crucial to consider the influence of family and personal background. These factors, if not included in our analysis, are known as **omitted variables**. An omitted variable is one that:

(i) is correlated with the dependent variable (wage_i, in this scenario)

(ii) is correlated with the regressor of interest (edu_i)

(iii) is omitted in the regression

The presence of omitted variables means that we cannot be sure that the regression relationship between education and wages is purely causal. We say that we have **omitted variable bias** for the causal effect of the regressor of interest.

The coefficient β_2 in the simple regression model measures the correlative or marginal effect, not the causal effect. This must always be kept in mind when interpreting regression coefficients.

Control Variables

We can include **control variables** in the linear regression model to reduce omitted variable bias so that we can interpret β_2 as a **ceteris paribus marginal effect** (ceteris paribus means holding other variables constant).

For example, let's include years of experience as well as ethnic identity and gender dummy variables for Black and female:

$$\text{wage}_i = \beta_1 + \beta_2 \text{edu}_i + \beta_3 \text{exper}_i + \beta_4 \text{Black}_i + \beta_5 \text{fem}_i + u_i.$$

In this case,

$$\beta_2 = \frac{\partial E[\text{wage}_i | \text{edu}_i, \text{exper}_i, \text{Black}_i, \text{fem}_i]}{\partial \text{edu}_i}$$

is the marginal effect of education on expected wages, holding experience, ethnic identity, and gender fixed.

```
lm(wage ~ education + experience + Black + female, data = cps)
```

```
Call:
lm(formula = wage ~ education + experience + Black + female,
    data = cps)
```

```
Coefficients:
(Intercept)    education   experience        Black       female
   -21.7089       3.1350       0.2443      -2.8554      -7.4363
```

Interpretation of coefficients:

- **Education**: Given the same experience, ethnic identity (whether the individual identifies as Black), and gender, people with one more year of education are paid on average $3.14 USD more than people with one year less of education.
- **Experience**: Each additional year of experience is associated with an average wage increase of $0.24 USD per hour, holding other factors constant.
- **Black**: Black workers earn on average $2.86 USD less per hour than non-Black workers with the same education, experience, and gender.
- **Female**: Women earn on average $7.43 USD less per hour than men with the same education, experience, and ethnic identity.

Note: This regression does not control for other unobservable characteristics (such as ability) or variables not included in the regression (such as quality of education), so omitted variable bias may still be present.

Good vs. Bad Controls

It's important to recognize that control variables are always selected with respect to a particular regressor of interest. A researcher typically focuses on estimating the effect of one specific variable (like education), and control variables must be designed specifically for this relationship.

In causal inference terminology, we can distinguish between different types of variables:

- **Confounders**: Variables that affect both the regressor of interest and the outcome. These are good controls because they help isolate the causal effect of interest.
- **Mediators**: Variables through which the regressor of interest affects the outcome. Controlling for mediators can block part of the causal effect we're trying to estimate.
- **Colliders**: Variables that are affected by both the regressor of interest and the outcome (or by factors that determine the outcome). Controlling for colliders can create spurious associations.

Confounders

Examples of **good controls** (confounders) for education are:

- Parental education level (affects both a person's education and their wage potential)
- Region of residence (geographic factors can influence education access and job markets)

- Family socioeconomic background (affects educational opportunities and wage potential)

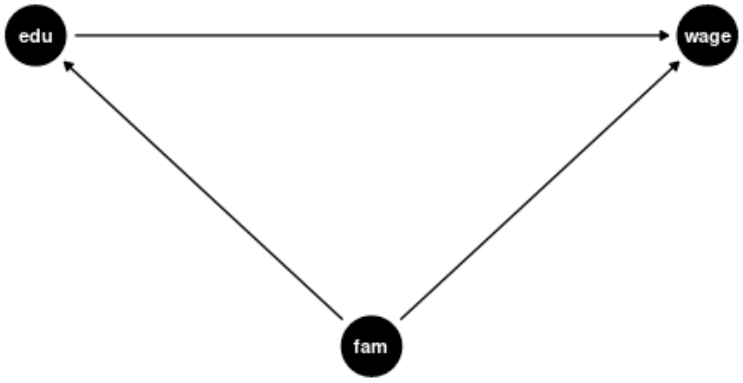


Figure 6.2: A DAG of the education-wage relationship with a family confounder

Mediators and Colliders

Examples of **bad controls** include:

- **Mediators**: Variables that are part of the causal pathway from education to wages
  - Current job position (education → job position → wage)
  - Professional sector (education may determine which sector someone works in)
  - Number of professional certifications (likely a result of education level)

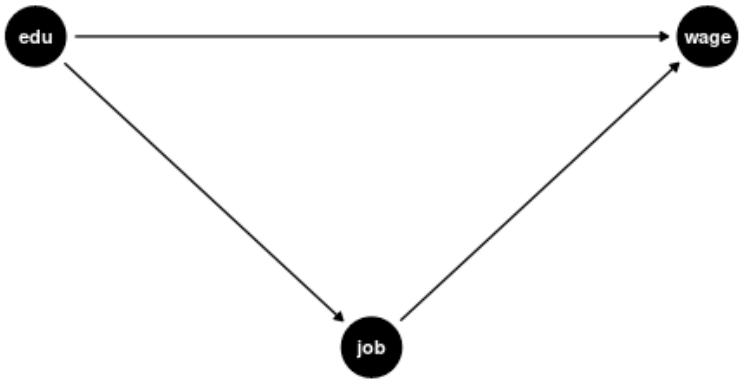


Figure 6.3: A DAG of the education-wage relationship with job position mediator

- **Colliders**: Variables affected by both education and wages (or their determinants)

- Happiness/life satisfaction (might be affected independently by both education and wages)
- Work-life balance (both education and wages might affect this independently)

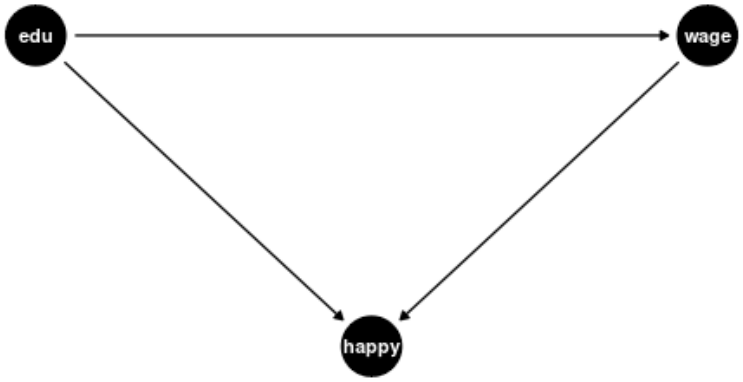


Figure 6.4: A DAG of the education-wage relationship with happiness collider

Bad controls create two problems:

1. **Statistical issue**: High correlation with the variable of interest (like education) causes high variance in the coefficient estimate (high collinearity).
2. **Causal inference issue**: They distort the relationship we're trying to estimate by either blocking part of the causal effect (mediators) or creating artificial associations (colliders).

Good control variables are typically determined before the level of education is determined, while bad controls are often outcomes of the education process itself or are jointly determined with wages.

The appropriate choice of control variables requires not just statistical knowledge but also subject-matter expertise about the causal structure of the relationships being studied.

6.2 Application: Class Size Effect

Let's apply these concepts to a real-world research question: How does class size affect student performance?

Recall the `CASchools` dataset used in the Stock and Watson textbook, which contains information on California school characteristics:

```
data(CASchools, package = "AER")
CASchools$STR = CASchools$students/CASchools$teachers
CASchools$score = (CASchools$read+CASchools$math)/2
```

We are interested in the effect of the student-teacher ratio `STR` (class size) on the average test score `score`. Following our previous discussion on causal inference, we need to consider potential confounding factors that might affect both class sizes and test scores.

Control Strategy

Let's examine several control variables:

- `english`: proportion of students whose primary language is not English.
- `lunch`: proportion of students eligible for free/reduced-price meals.
- `expenditure`: total expenditure per pupil.

First, we should check whether these variables are correlated with both our regressor of interest (`STR`) and the outcome (`score`):

```
library(dplyr)
CASchools |> select(STR, score, english, lunch, expenditure) |> cor()
```

```
                  STR      score     english       lunch expenditure
STR          1.0000000 -0.2263627  0.18764237  0.13520340 -0.61998216
score       -0.2263627  1.0000000 -0.64412381 -0.86877199  0.19127276
english      0.1876424 -0.6441238  1.00000000  0.65306072 -0.07139604
lunch        0.1352034 -0.8687720  0.65306072  1.00000000 -0.06103871
expenditure -0.6199822  0.1912728 -0.07139604 -0.06103871  1.00000000
```

The correlation matrix reveals that `english`, `lunch`, and `expenditure` are indeed correlated with both `STR` and `score`. This suggests they could be confounders that, if omitted, might bias our estimate of the class size effect.

Let's implement a control strategy, adding potential confounders one by one to see how the estimated marginal effect of class size changes:

```
fit1 = lm(score ~ STR, data = CASchools)
fit2 = lm(score ~ STR + english, data = CASchools)
fit3 = lm(score ~ STR + english + lunch, data = CASchools)
fit4 = lm(score ~ STR + english + lunch + expenditure, data = CASchools)
library(modelsummary)
mymodels = list(fit1, fit2, fit3, fit4)
modelsummary(mymodels,
             statistic = NULL,
             gof_map = c("nobs", "r.squared", "adj.r.squared", "rmse"))
```

	(1)	(2)	(3)	(4)
(Intercept)	698.933	686.032	700.150	665.988
STR	−2.280	−1.101	−0.998	−0.235
english		−0.650	−0.122	−0.128
lunch			−0.547	−0.546
expenditure				0.004
Num.Obs.	420	420	420	420
R2	0.051	0.426	0.775	0.783
R2 Adj.	0.049	0.424	0.773	0.781
RMSE	18.54	14.41	9.04	8.86

Interpretation of Marginal Effects

Let's interpret the coefficients on STR from each model more precisely:

- **Model (1)**: Between two classes that differ by one student, the class with more students scores on average 2.280 points lower. This represents the unadjusted association without controlling for any confounding factors.
- **Model (2)**: Between two classes that differ by one student but have the same share of English learners, the larger class scores on average 1.101 points lower. Controlling for English learner status cuts the estimated effect by more than half.
- **Model (3)**: Between two classes that differ by one student but have the same share of English learners and and the same share of students eligible for reduced-price meals, the larger class scores on average 0.998 points lower. Adding this socioeconomic control further reduces the estimated effect slightly.
- **Model (4)**: Between two classes that differ by one student but have the same share of English learners, students with reduced meals, and per-pupil expenditure, the larger class scores on average 0.235 points lower. This represents a dramatic reduction from the previous model.

The sequential addition of controls demonstrates how sensitive the estimated marginal effect is to model specification. Each coefficient represents the partial derivative of the expected test score with respect to the student-teacher ratio, holding constant the variables included in that particular model.

Identifying Good and Bad Controls

Based on our causal framework from the previous section, we can evaluate our control variables:

- **Confounders (good controls)**: `english` and `lunch` are likely **good controls** because they represent pre-existing student characteristics that influence both class size assignments and test performance. For instance, schools with a higher share of immigrants or lower-income households may have on average higher class sizes and lower reading scores.

`STR` $\leftarrow$ `english` $\rightarrow$ `score`

- **Mediator (bad control)**: `expenditure` appears to be a **bad control** because it's likely a mediator in the causal pathway from class size to test scores. Smaller classes mechanically increase per-pupil expenditure through higher teacher salary costs per student.

`STR` $\rightarrow$ `expenditure` $\rightarrow$ `score`

When we control for `expenditure`, we block this causal pathway and "control away" part of the effect of `STR` on `score` we actually want to measure. This explains the dramatic drop in the coefficient in Model (4) and suggests this model likely underestimates the true effect of class size.

This application demonstrates the crucial importance of thoughtful control variable selection in regression analysis. The estimated marginal effect of `STR` on `score` varies substantially depending on which variables we control for. Based on causal reasoning, we should prefer Model (3) with the appropriate confounders but without the mediator.

6.3 Polynomials

Experience and wages

A linear dependence of wages and experience is a strong assumption. We can reasonably expect a nonlinear marginal effect of another year of experience on wages. For example, the effect may be higher for workers with 5 years of experience than for those with 40 years of experience.

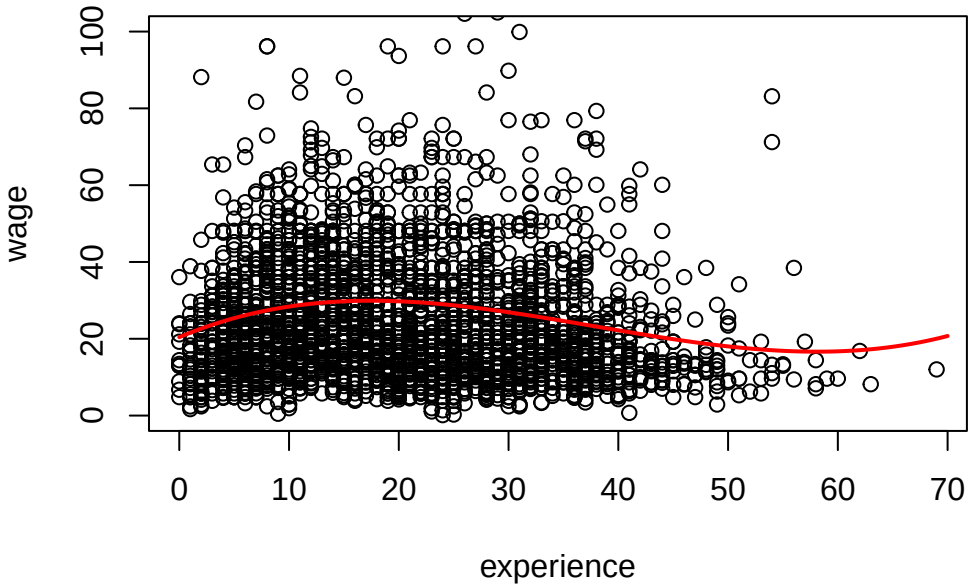
Polynomials can be used to specify a nonlinear regression function:

$$\text{wage}_i = \beta_1 + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{exper}_i^3 + u_i.$$

```r
## we focus on people with Asian background only for illustration
cps.as = cps |> subset(Asian == 1)
fit = lm(wage ~ experience + I(experience^2) + I(experience^3),
         data = cps.as)
beta = fit$coefficients
beta |> round(4)
```

```
    (Intercept)      experience I(experience^2) I(experience^3)
        20.4159          1.2067         -0.0449          0.0004
```

```r
## Scatterplot
plot(wage ~ experience, data = cps.as, ylim = c(0,100))
## plot the cubic function for fitted wages
curve(
  beta[1] + beta[2]*x + beta[3]*x^2 + beta[4]*x^3,
  from = 0, to = 70, add=TRUE, col='red', lwd=2
  )
```



The marginal effect depends on the years of experience:

$$\frac{\partial E[\text{wage}_i|\text{exper}_i]}{\partial \text{exper}_i} = \beta_2 + 2\beta_3 \text{exper}_i + 3\beta_4 \text{exper}_i^2.$$

For instance, the additional wage for a worker with 11 years of experience compared to a worker with 10 years of experience is on average

$$1.2013 + 2 \cdot (-0.0447) \cdot 10 + 3 \cdot 0.0004 \cdot 10^2 = 0.4273.$$

Income and test scores

Another example is the relationship between the income of schooling districts and their test scores.

Income and test score are positively correlated:

```
cor(CASchools$income, CASchools$score)
```

```
[1] 0.7124308
```

School districts with above-average income tend to achieve above-average test scores. But does a linear regression adequately model the data? Let's compare a linear with a quadratic regression specification.

```
linear = lm(score ~ income, data = CASchools)
linear
```

```
Call:
lm(formula = score ~ income, data = CASchools)

Coefficients:
(Intercept)       income
    625.384        1.879
```

Estimated linear regression function:

$$\widehat{\text{score}} = 625.4 + 1.88\,\text{inc.}$$

```
quad = lm(score ~ income + I(income^2), data = CASchools)
quad
```

```
Call:
lm(formula = score ~ income + I(income^2), data = CASchools)

Coefficients:
(Intercept)       income  I(income^2)
  607.30174      3.85099     -0.04231
```

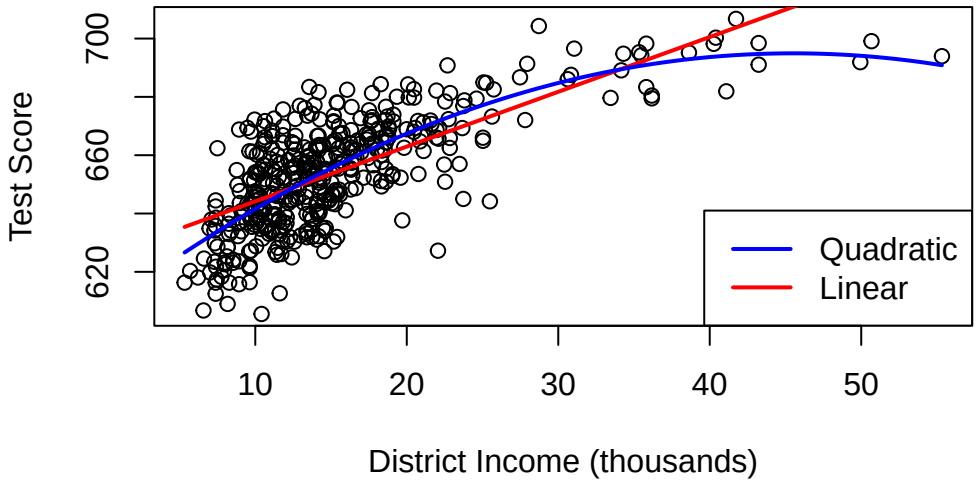
Estimated quadratic regression function:

$$\widehat{\text{score}} = 607.3 + 3.85\,\text{inc} - 0.0423\,\text{inc}^2.$$

```
# Create scatterplot
plot(score ~ income, data = CASchools,
     xlab = "District Income (thousands)",
     ylab = "Test Score")

# Add fitted curves
curve(coef(linear)[1] + coef(linear)[2]*x, add = TRUE, col = "red", lwd=2)
curve(coef(quad)[1] + coef(quad)[2]*x + coef(quad)[3]*x^2, add = TRUE, col = "blue", lwd=2)

# Add legend
legend("bottomright", c("Quadratic", "Linear"), col = c("blue", "red"), lwd = 2)
```



The plot shows that the linear regression line seems to overestimate the true relationship when income is either very high or very low and it tends to underestimate it for the middle income group.

The quadratic function appears to provide a better fit to the data compared to the linear function.

6.4 Logarithms

Log-income and test scores

Another approach to estimate a concave nonlinear regression function involves using a logarithmic regressor.

```
# estimate a linear-log model
linlog = lm(score ~ log(income), data = CASchools)
linlog
```

```
Call:
lm(formula = score ~ log(income), data = CASchools)

Coefficients:
(Intercept)  log(income)
     557.83        36.42
```

The estimated regression function is

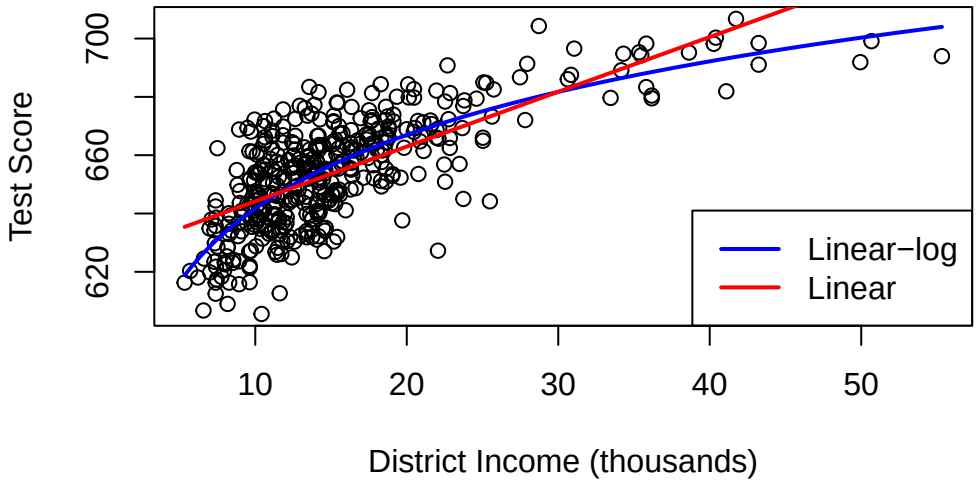
$$\widehat{\text{score}} = 557.8 + 36.42 \log(\text{inc})$$

```
# Create scatterplot
plot(score ~ income, data = CASchools,
    xlab = "District Income (thousands)",
    ylab = "Test Score")

# Add fitted curves
curve(coef(linlog)[1] + coef(linlog)[2]*log(x), add = TRUE, col = "blue", lwd = 2)
curve(coef(linear)[1] + coef(linear)[2]*x, add = TRUE, col = "red", lwd = 2)

# Add legend
legend("bottomright", c("Linear-log", "Linear"), col = c("blue", "red"), lwd = 2)
```

	linear	quad	linlog
(Intercept)	625.384	607.302	557.832
income	1.879	3.851	
I(income^2)		−0.042	
log(income)			36.420
Num.Obs.	420	420	420
R2	0.508	0.556	0.563
R2 Adj.	0.506	0.554	0.561
RMSE	13.35	12.68	12.59




```
library(modelsummary)
modelsummary(list("linear" = linear, "quad" = quad, "linlog" = linlog),
             statistic = NULL,
             gof_map = c("nobs", "r.squared", "adj.r.squared", "rmse"))
```

We observe that the adjusted R-squared is highest for the logarithmic model, indicating that the latter is the most suitable.

The coefficients have a different interpretation.

- Assuming the linear model specification is correct, we have

$$E[\text{score}|\text{inc}] = \beta_1 + \beta_2 \text{inc}.$$

  The marginal effect of `income` on `score` is

$$\frac{\partial E[\text{score}|\text{inc}]}{\partial \text{inc}} = \beta_2.$$

Students from a district with \$1000 higher income have on average 1.879 points higher scores.

- Assuming the quadratic model specification is correct, we have

$$E[\text{score}|\text{inc}] = \beta_1 + \beta_2 \text{inc} + \beta_3 \text{inc}^2.$$

The marginal effect `income` on `score` depends on the income level:

$$\frac{\partial E[\text{score}|\text{inc}]}{\partial \text{inc}} = \beta_2 + 2\beta_3 \text{inc}.$$

When considering a district with x income, students with \$1000 higher income have on average 3.85 - 0.0846x points higher scores.

- Assuming the logarithmic model specification is correct, we have

$$E[\text{score}|\text{inc}] = \beta_1 + \beta_2 \log(\text{inc}).$$

The slope coefficient represents the marginal effect of `log(income)` on `score`:

$$\frac{\partial E[\text{score}|\text{inc}]}{\partial \log(\text{inc})} = \beta_2.$$

Instead, the marginal effect of `income` on `score` is

$$\frac{\partial E[\text{score}|\text{inc}]}{\partial \text{inc}} = \beta_2 \cdot \frac{1}{\text{inc}},$$

so

$$\underbrace{\frac{\partial E[\text{score}|\text{inc}]}{\partial \text{inc}}}_{\text{absolute change}} = \beta_2 \cdot \underbrace{\frac{\partial \text{inc}}{\text{inc}}}_{\text{percentage change}}.$$

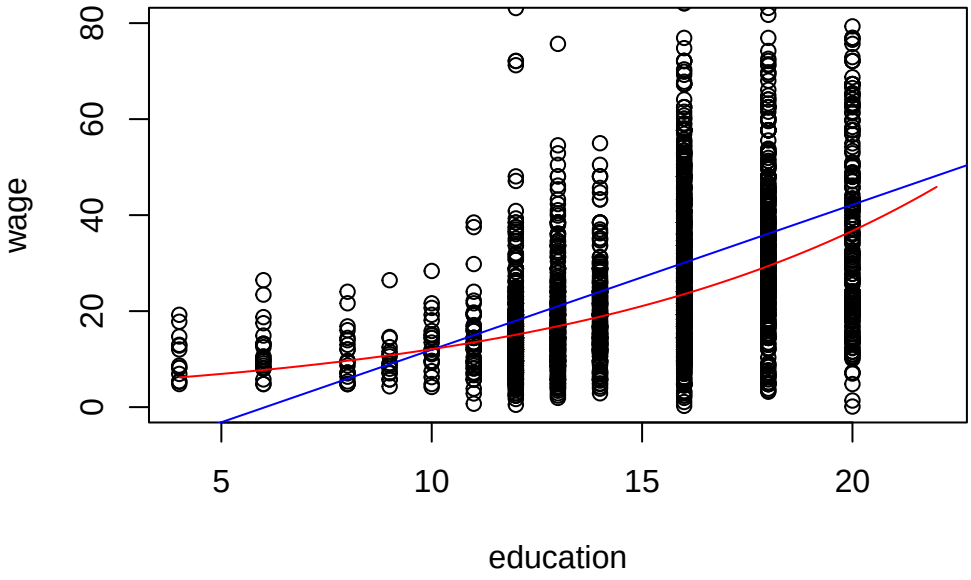
Students from a district with 1% higher income have on average $36.42 \cdot 1\% = 0.3642$ points higher scores.

Education and log-wages

If a convex relationship is expected, we can also use a logarithmic transformation for the dependent variable:

$$\log(\text{wage}_i) = \beta_1 + \beta_2 \text{edu}_i + u_i$$

```
log_model = lm(log(wage) ~ education, data = cps.as)
linear_model = lm(wage ~ education, data = cps.as)
plot(wage ~ education, data = cps.as, ylim = c(0,80), xlim = c(4,22))
abline(linear_model, col="blue")
coef = coefficients(log_model)
curve(exp(coef[1]+coef[2]*x), add=TRUE, col="red")
```



The marginal effect of `education` on `log(wage)` is

$$\frac{\partial E[\log(\text{wage}_i)|edu_i]}{\partial \text{edu}_i} = \beta_2.$$

To interpret β_2 in terms of changes of `wage` instead of `log(wage)`, consider the following approximation:

$$E[\text{wage}_i|\text{edu}_i] \approx \exp(E[\log(\text{wage}_i)|\text{edu}_i]).$$

The left-hand expression is the conventional conditional mean, and the right-hand expression is the geometric mean. The geometric mean is slightly smaller because $E[\log(Y)] < \log(E[Y])$, but this difference is small unless the data is highly skewed.

The marginal effect of a change in *edu* on the geometric mean of *wage* is

$$\frac{\partial exp(E[\log(\text{wage}_i)|\text{edu}_i])}{\partial \text{edu}_i} = \underbrace{exp(E[\log(\text{wage}_i)|\text{edu}_i])}_{\text{outer derivative}} \cdot \beta_2.$$

Using the geometric mean approximation from above, we get

$$\underbrace{\frac{\partial E[\text{wage}_i|\text{edu}_i]}{E[\text{wage}_i|\text{edu}_i]}}_{\substack{\text{percentage}\\ \text{change}}} \approx \frac{\partial exp(E[\log(\text{wage}_i)|\text{edu}_i])}{exp(E[\log(\text{wage}_i)|\text{edu}_i])} = \beta_2 \cdot \underbrace{\partial \text{edu}_i}_{\substack{\text{absolute}\\ \text{change}}}.$$

```
log_model
```

```
Call:
lm(formula = log(wage) ~ education, data = cps.as)

Coefficients:
(Intercept)    education
     1.3783       0.1113
```

Interpretation: A person with one more year of education has a wage that is 11.13% higher on average.

In addition to the linear-log and log-linear specifications, we also have the log-log specification

$$\log(Y) = \beta_1 + \beta_2 \log(X) + u.$$

Log-log interpretation: When X is 1% higher, we observe, on average, a β_2% higher Y.

6.5 Interactions

A linear regression with interaction terms:

$$\text{wage}_i = \beta_1 + \beta_2 \text{edu}_i + \beta_3 \text{fem}_i + \beta_4 \text{marr}_i + \beta_5 (\text{marr}_i \cdot \text{fem}_i) + u_i$$

```
lm(wage ~ education + female + married + married:female, data = cps)
```

```
Call:
lm(formula = wage ~ education + female + married + married:female,
    data = cps)

Coefficients:
   (Intercept)       education          female         married  female:married
       -18.241           2.877          -3.025           7.352          -6.016
```

The marginal effect of gender depends on the person's marital status:

$$\frac{\partial E[\text{wage}_i | \text{edu}_i, \text{fem}_i, \text{marr}_i]}{\partial \text{fem}_i} = \beta_3 + \beta_5 \text{marr}_i$$

Since `female` is a dummy variable, we interpret the marginal effect as a discrete $0 \to 1$ change (ceteris paribus), not literally a derivative.

Interpretation: Given the same education, unmarried women are paid on average 3.27 USD less than unmarried men, and married women are paid on average 3.27+5.77=9.04 USD less than married men.

The marginal effect of the marital status depends on the person's gender:

$$\frac{\partial E[\text{wage}_i|\text{edu}_i, \text{fem}_i, \text{marr}_i]}{\partial \text{marr}_i} = \beta_4 + \beta_5 \text{fem}_i$$

Interpretation: Given the same education, married men are paid on average 7.17 USD more than unmarried men, and married women are paid on average 7.17-5.77=1.40 USD more than unmarried women.

6.6 R Code

statistics-sec06.R

7 Inference

7.1 Strict Exogeneity

Recall the linear regression framework:

- Regression equation $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i$, $i = 1, \dots, n$
- Exogeneity condition $E[u_i|\boldsymbol{X}_i] = 0$
- i.i.d. sample $(Y_i, \boldsymbol{X}_i')$ with finite second moments
- Full rank $E[\boldsymbol{X}_i\boldsymbol{X}_i']$

The **exogeneity condition**

$$E[u_i|\boldsymbol{X}_i] = 0$$

ensures that the regressors are uncorrelated with the error at the individual observation level.

The **i.i.d. condition** implies $(Y_i, \boldsymbol{X}_i')$ is independent of $(Y_j, \boldsymbol{X}_j')$ for all $j \neq i$. Hence, $\boldsymbol{X}_j$ is independent of $u_i = Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}$ for all $j \neq i$, so

$$E[u_i|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n] = E[u_i|\boldsymbol{X}_i].$$

Together, exogeneity and i.i.d. sampling imply **strict exogeneity**:

$$E[u_i|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n] = 0 \quad \text{for all } i.$$

In cross-section with i.i.d. sampling, exogeneity at the unit level implies strict exogeneity.

Equivalently, in matrix form:

$$E[\boldsymbol{u}|\boldsymbol{X}] = \boldsymbol{0}.$$

Strict exogeneity requires the entire vector of errors $\boldsymbol{u}$ to be mean independent of the full regressor matrix $\boldsymbol{X}$. That is, no systematic relationship exists between any regressors and any error term across observations.

7.2 Unbiasedness

Under strict exogeneity, the OLS estimator $\hat{\boldsymbol{\beta}}$ is **unbiased**:

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$

Proof: Recall the model equation in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

Plugging this into the OLS formula:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}.\end{aligned}$$

Taking the conditional expectation:

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

Since $E[\boldsymbol{u}|\boldsymbol{X}] = \boldsymbol{0}$, the conditional mean is

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta}.$$

By the LIE, the unconditional mean becomes

$$E[\hat{\boldsymbol{\beta}}] = E[E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]] = \boldsymbol{\beta}.$$

Thus, each element of the OLS estimator is unbiased:

$$E[\hat{\beta}_j] = \beta_j \quad \text{for } j = 1, \dots, k.$$

7.3 Sampling Variance of OLS

The OLS estimator $\hat{\boldsymbol{\beta}}$ provides a **point estimate** of the unknown population parameter $\boldsymbol{\beta}$. For example, in the regression

$$\text{wage}_i = \beta_1 + \beta_2 \text{education}_i + \beta_3 \text{female}_i + u_i,$$

we obtain specific coefficient estimates:

```r
cps = read.csv("cps.csv")
fit = lm(wage ~ education + female, data = cps)
fit |> coef()
```

```
(Intercept)   education      female
 -14.081788    2.958174   -7.533067
```

The estimate for *education* is $\hat{\beta}_2 = 2.958$. However, this point estimate tells us nothing about how far it might be from the true value β_2.

That is, it does not reflect **estimation uncertainty**, which arises because $\hat{\boldsymbol{\beta}}$ depends on a finite sample that could have turned out differently if a different dataset from the same population had been used.

Larger samples tend to reduce estimation uncertainty, but in practice we only observe one finite sample. To quantify this uncertainty, we study the **sampling variance** of the OLS estimator:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) = \mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}),$$

the conditional variance of $\hat{\boldsymbol{\beta}}$ given the regressors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

Sampling variance of OLS:

Under i.i.d. sampling, the OLS covariance matrix is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{X}_i \boldsymbol{X}_i' \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1},$$

where $\sigma_i^2 = E[u_i^2|\boldsymbol{X}_i]$.

In matrix notation, this can be equivalently written as

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Proof: Recall the general rule that for any matrix $\boldsymbol{A}$,

$$\mathrm{Var}(\boldsymbol{A}\boldsymbol{u}) = \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{u})\,\boldsymbol{A}'.$$

Hence, with $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, by the symmetry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$,

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \mathrm{Var}(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}) \\
&= \mathrm{Var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Under i.i.d. sampling, conditional on $\boldsymbol{X}$, u_i and u_j are independent for $i \neq j$, so

$$E[u_i u_j|\boldsymbol{X}] = E[u_i|\boldsymbol{X}]\, E[u_j|\boldsymbol{X}] = E[u_i|\boldsymbol{X}_i]\, E[u_j|\boldsymbol{X}_j] = 0,$$

and the conditional covariance matrix of $\boldsymbol{u}$ takes a diagonal form:

$$\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}.$$

Also note: $\boldsymbol{X}'\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X})\boldsymbol{X} = \sum_{i=1}^n \sigma_i^2 \boldsymbol{X}_i \boldsymbol{X}_i'$ and $\boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'$.

Homoskedasticity

While u_i is uncorrelated with $\boldsymbol{X}_i$ under the exogeneity assumption, its variance may depend on $\boldsymbol{X}_i$. We say that the errors are **heteroskedastic**:

$$\sigma_i^2 = \mathrm{Var}(u_i|\boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i).$$

In the specific situation where the conditional variance of the error does not depend on $\boldsymbol{X}_i$ and is equal to σ^2 for any value of $\boldsymbol{X}_i$, we say that the errors are **homoskedastic**:

$$\sigma^2 = \mathrm{Var}(u_i) = \mathrm{Var}(u_i|\boldsymbol{X}_i) \quad \text{for all } i.$$

The homoskedastic error covariance matrix has the following simple form:

$$\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X}) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \boldsymbol{I}_n.$$

Because $\boldsymbol{X}'\mathrm{Var}(\boldsymbol{u}|\boldsymbol{X})\boldsymbol{X} = \sigma^2 \boldsymbol{X}'\boldsymbol{X}$, the resulting OLS covariance matrix reduces to

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2 \left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1}.$$

7.4 Gaussian distribution

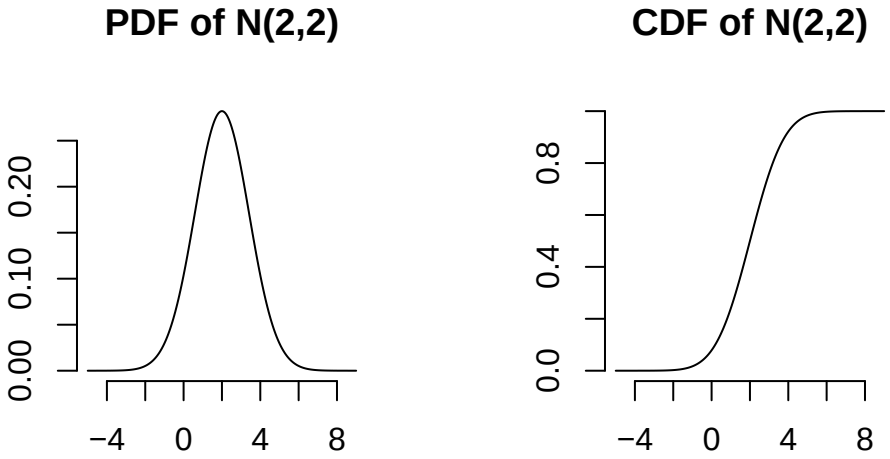
Univariate Normal distribution

The Gaussian distribution, also known as the **normal distribution**, is a fundamental concept in statistics.

A random variable Z is said to follow a normal distribution if it has the following probability density function (PDF):

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(z-\mu)^2}{2\sigma^2} \Big).$$

Formally, we denote this as $Z \sim \mathcal{N}(\mu, \sigma^2)$, meaning that Z is normally distributed with mean μ and variance σ^2.



The normal distribution with mean 0 and variance 1 is called the **standard normal distribution**. It has the PDF

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

and its CDF is

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

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

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

Standardizing: If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{Z-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

The CDF of Z is $P(Z \le a) = \Phi((a-\mu)/\sigma)$.

A normally distributed random variable Z has $\text{skew}(Z) = 0$ and $\text{kurt}(Z) = 3$.

Linear combinations of normally distributed variables are normal: If $Y_1, \dots, Y_n$ are jointly normally distributed and $c_1, \dots, c_n \in \mathbb{R}$, then $\sum_{j=1}^{n} c_j Y_j$ is normally distributed.

Multivariate Normal distribution

Let $Z_1, \ldots, Z_k$ be independent $\mathcal{N}(0,1)$ random variables.

Then, the k-vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$ has the **multivariate standard normal distribution**, written $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$. Its joint PDF is

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

If $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\boldsymbol{Z}^* = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}$ for a $q \times 1$ vector $\boldsymbol{\mu}$ and a $q \times k$ matrix $\boldsymbol{B}$, then $\boldsymbol{Z}^*$ has a **multivariate normal distribution** with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$, written $\boldsymbol{Z}^* \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The q-variate PDF of $\boldsymbol{Z}^*$ is

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

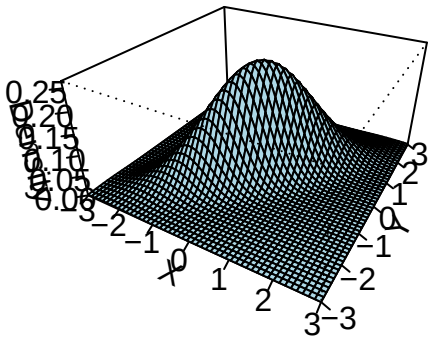
The mean vector and covariance matrix are

$$E[\boldsymbol{Z}^*] = \boldsymbol{\mu}, \quad \mathrm{Var}(\boldsymbol{Z}^*) = \boldsymbol{\Sigma}.$$

The 3D plot below shows the bivariate normal PDF with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

3D Bivariate Normal Distribution Density



7.5 Gaussian Regression Model

The **Gaussian regression model** builds on the linear regression framework by adding a distributional assumption in addition to the i.i.d. and exogeneity assumptions.

It assumes that the error terms are homoskedastic and conditionally normally distributed:

$$u_i|\boldsymbol{X}_i \sim \mathcal{N}(0, \sigma^2) \tag{7.1}$$

That is, conditional on the regressors, the error has mean zero (exogeneity), constant variance (homoskedasticity), and a normal distribution.

Because $u_i = Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}$, Equation 7.1 is equivalent to

$$Y_i|\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i'\boldsymbol{\beta}, \sigma^2).$$

Thus, the normality assumption is a distributional assumption about the dependent variable Y_i, not the regressors $\boldsymbol{X}_i$.

Because $\hat{\boldsymbol{\beta}}$ is a linear combination of the independent and normally distributed variables $Y_1, \dots, Y_n$, the OLS estimator is also normally distributed:

$$\hat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathcal{N}\Big(\boldsymbol{\beta}, \text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})\Big).$$

The variance of the j-th OLS coefficient $\hat{\beta}_j$ is the j-th diagonal element of the covariance matrix. Under homoskedasticity, its conditional standard deviation is

$$\text{sd}(\hat{\beta}_j|\boldsymbol{X}) = \sqrt{\left(\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})\right)_{jj}} = \sigma\sqrt{\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)_{jj}}.$$

Subtracting the mean $E[\hat{\beta}_j] = \beta_j$ and dividing by the $\text{sd}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$ gives the standardized OLS coefficient, which has mean zero and variance 1:

$$Z_j := \frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j|\boldsymbol{X})} \sim \mathcal{N}(0, 1)$$

7.6 Classical Standard Errors

The conditional standard deviation $\text{sd}(\hat{\beta}_j|\boldsymbol{X})$ in the Gaussian regression model is unknown because the population error variance σ^2 is unknown:

$$\text{sd}(\hat{\beta}_j|\boldsymbol{X}) = \sigma\sqrt{\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)_{jj}}.$$

A **standard error** of $\hat{\beta}_j$ is an estimator of the conditional standard deviation. To construct a valid standard error under this setup, we can use the adjusted residual variance to estimate σ^2:

$$s^2_{\widehat{u}} = \frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2.$$

The classical standard error (valid under homoskedasticity) is defined as:

$$\text{se}_{hom}(\hat{\beta}_j) = s_{\widehat{u}}\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}}.$$

To estimate the full sampling covariance matrix $\boldsymbol{V} = \text{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})$, the classical covariance matrix estimator is:

$$\widehat{\boldsymbol{V}}_{hom} = s^2_{\widehat{u}}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

```
## classical homoskedastic covariance matrix estimator:
vcov(fit)
```

```
            (Intercept)     education        female
(Intercept)  0.18825476 -0.0127486354 -0.0089269796
education   -0.01274864  0.0009225111 -0.0002278021
female      -0.00892698 -0.0002278021  0.0284200217
```

Classical standard errors $\text{se}_{hom}(\hat{\beta}_j)$ are the square roots of the diagonal entries:

```
## classical standard errors:
sqrt(diag(vcov(fit)))
```

```
(Intercept)   education      female
 0.43388334  0.03037287  0.16858239
```

They are also displayed in parentheses in a typical regression summary table:

```
library(modelsummary)
modelsummary(fit, gof_map = "none")
```

	(1)
(Intercept)	−14.082
	(0.434)
education	2.958
	(0.030)
female	−7.533
	(0.169)

7.7 Distributions from Normal Samples

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

The parameter m is called the degrees of freedom.

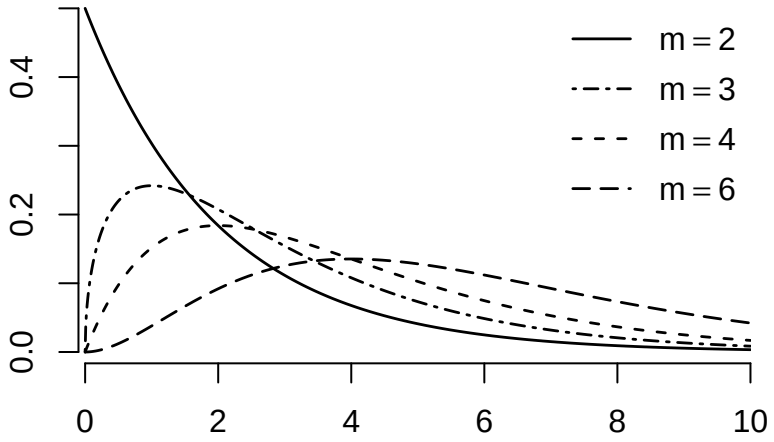


Figure 7.1: PDF of the χ^2-distribution

Under the Gaussian assumption Equation 7.1, $s^2_{\widehat{u}}$ has the following property:

$$\frac{(n-k)s^2_{\widehat{u}}}{\sigma^2} \sim \chi^2_{n-k}. \tag{7.2}$$

Student t-distribution

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

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

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

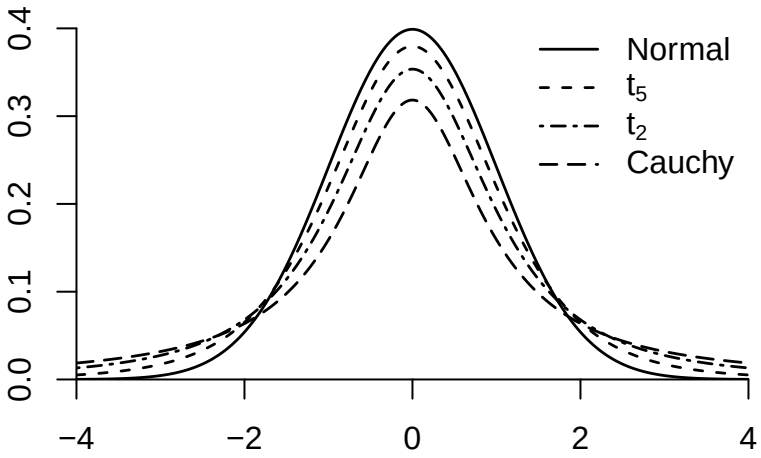


Figure 7.2: PDFs of the Student t-distribution

Under the Gaussian assumption Equation 7.1, the standardized OLS coefficient is standard normal.

When we replace the population standard deviation with its sample estimate (the standard error) then the standardized OLS coefficient has a t-distribution:

$$T_j := \frac{\hat{\beta}_j - \beta_j}{\mathrm{se}_{hom}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\mathrm{sd}(\hat{\beta}_j|\boldsymbol{X})} \cdot \frac{\sigma}{s_{\widehat{u}}} = Z_j \cdot \frac{\sigma}{s_{\widehat{u}}}$$

with

$$T_j \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} = t_{n-k}. \tag{7.3}$$

This means that the OLS coefficient standardized with the homoskedastic standard error instead of the standard deviation follows a t-distribution with $n-k$ degrees of freedom.

Here, we used Equation 7.2 and the fact $\hat{\boldsymbol{\beta}}$ and $s^2_{\widehat{u}}$ are independent.

Like $\mathcal{N}(0,1)$, the t-distribution is symmetric around zero:

$$P(T_j > a) = P(T_j < -a).$$

The t-distribution has heavier tails than the standard normal distribution.

The t_m distribution approaches $\mathcal{N}(0,1)$ as $m \to \infty$.

7.8 Exact Confidence Intervals

A confidence interval is a range of values that is likely to contain the true population parameter with a specified **confidence level** or **coverage probability**, often expressed as a percentage (e.g., 95%).

A $(1-\alpha)$ confidence interval for β_j is an interval $I_{1-\alpha}$ such that

$$P(\beta_j \in I_{1-\alpha}) = 1-\alpha, \tag{7.4}$$

or, equivalently,

$$P(\beta_j \notin I_{1-\alpha}) = \alpha.$$

To construct such an interval, we use the form

$$I_{1-\alpha} = \Big[\hat{\beta}_j - q \cdot \mathrm{se}_{hom}(\hat{\beta}_j); \hat{\beta}_j + q \cdot \mathrm{se}_{hom}(\hat{\beta}_j)\Big].$$

To find the suitable value q, note that, by Equation 7.3,

$$\begin{aligned}
P(\beta_j \in I_{1-\alpha}) &= P\Big(\hat{\beta}_j - q \cdot \mathrm{se}_{hom}(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + q \cdot \mathrm{se}_{hom}(\hat{\beta}_j)\Big) \\
&= P\Big(-q \cdot \mathrm{se}_{hom}(\hat{\beta}_j) \le \beta_j - \hat{\beta}_j \le q \cdot \mathrm{se}_{hom}(\hat{\beta}_j)\Big) \\
&= P\Big(-q \cdot \mathrm{se}_{hom}(\hat{\beta}_j) \le \hat{\beta}_j - \beta_j \le q \cdot \mathrm{se}_{hom}(\hat{\beta}_j)\Big) \\
&= P\Big(-q \le \frac{\hat{\beta}_j - \beta_j}{\mathrm{se}_{hom}(\hat{\beta}_j)} \le q\Big) \\
&= P(|T_j| \le q),
\end{aligned}$$

or, equivalently,

$$P(\beta_j \notin I_{1-\alpha}) = P(|T_j| > q).$$

By the symmetry of the t-distribution,

$$P(|T_j| > q) = P(T_j > q) + P(T_j < -q) = 2P(T_j > q).$$

Therefore, Equation 7.4 is equivalent to

$$\begin{aligned}
& P(\beta_j \notin I_{1-\alpha}) = \alpha \\
\Leftrightarrow \quad & P(|T_j| > q) = \alpha \\
\Leftrightarrow \quad & 2P(T_j > q) = \alpha \\
\Leftrightarrow \quad & P(T_j > q) = \alpha/2 \\
\Leftrightarrow \quad & P(T_j \le q) = 1 - \alpha/2.
\end{aligned}$$

The last condition means that q must be the $1-\alpha/2$ quantile of the distribution of T_j, which is the t_{n-k}-distribution.

We write

$$q = t_{n-k,1-\alpha/2}.$$

Hence, under the Gaussian regression model,

$$P\left(\beta_j \in \left[\hat{\beta}_j - t_{n-k,1-\alpha/2} \cdot \mathrm{se}_{hom}(\hat{\beta}_j); \hat{\beta}_j + t_{n-k,1-\alpha/2} \cdot \mathrm{se}_{hom}(\hat{\beta}_j)\right]\right) = 1-\alpha.$$

Table 7.1: Student's t-distribution quantiles

df	0.95	0.975	0.995	0.9995
1	6.31	12.71	63.66	636.6
2	2.92	4.30	9.92	31.6
3	2.35	3.18	5.84	12.9
5	2.02	2.57	4.03	6.87
10	1.81	2.23	3.17	4.95
20	1.72	2.09	2.85	3.85
50	1.68	2.01	2.68	3.50
100	1.66	1.98	2.63	3.39
$\to \infty$	1.64	1.96	2.58	3.29

The last row (indicated by $\to \infty$) shows the quantiles of the standard normal distribution $\mathcal{N}(0,1)$.

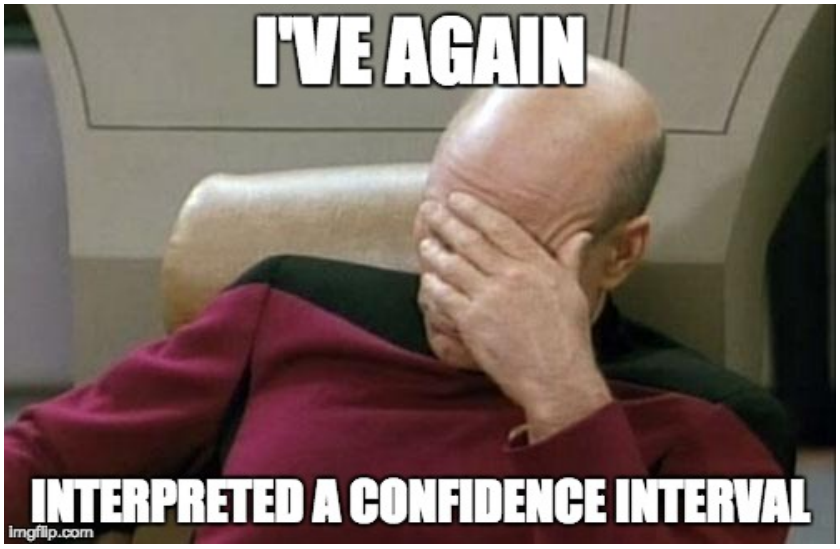
You can display 95% confidence intervals in the `modelsummary` output using the `conf.int` argument:

```
modelsummary(fit, gof_map = "none", statistic = "conf.int")
```

	(1)
(Intercept)	−14.082
	[−14.932, −13.231]
education	2.958
	[2.899, 3.018]
female	−7.533
	[−7.863, −7.203]

7.9 Confidence Interval Interpretation

Note: the confidence interval is **random**, while the parameter β_j is **fixed but unknown**.



A correct interpretation of a 95% confidence interval is:

- If we were to repeatedly draw samples and construct a 95% confidence interval from each sample, about 95% of these intervals would contain the true parameter.

Common misinterpretations to avoid:

- "There is a 95% probability that the true value lies in this interval."
- "We are 95% confident this interval contains the true parameter."

These mistakes incorrectly treat the parameter as random and the interval as fixed. In reality, it's the other way around.

A 95% confidence interval should be understood as a coverage probability: Before observing the data, there is a 95% probability that the random interval will cover the true parameter.

A helpful visualization:

https://rpsychologist.com/d3/ci/

7.10 Limitations of the Gaussian Approach

The Gaussian regression framework assumes:

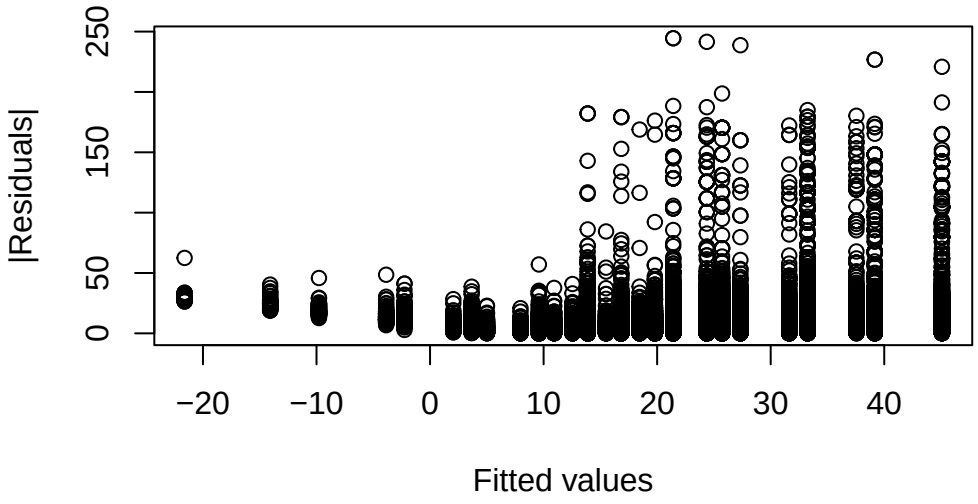
- Exogeneity: $E[u_i \mid \boldsymbol{X}_i] = 0$
- i.i.d. sample: $(Y_i, \boldsymbol{X}_i')$, $i = 1, \dots, n$
- Homoskedastic, normally distributed errors: $u_i | \boldsymbol{X}_i \sim \mathcal{N}(0, \sigma^2)$
- Full rank $E[\boldsymbol{X}_i \boldsymbol{X}_i']$

While mathematically convenient, these assumptions are often violated in practice. In particular, the normality assumption implies homoskedasticity and that the conditional distribution of Y_i given $\boldsymbol{X}_i$ is normal, which is an unrealistic scenario in many economic applications.

Historically, homoskedasticity has been treated as the "default" assumption and heteroskedasticity as a special case. But in empirical work, **heteroskedasticity is the norm**.

A plot of the absolute value of the residuals against the fitted values shows that individuals with predicted wages around 10 USD exhibit residuals with lower variance compared to those with higher predicted wage levels. Hence, the homoskedasticity assumption is implausible:

```
# Plot of absolute residuals against fitted values
plot(abs(residuals(fit)) ~ fitted(fit), xlab="Fitted values", ylab="|Residuals|")
```

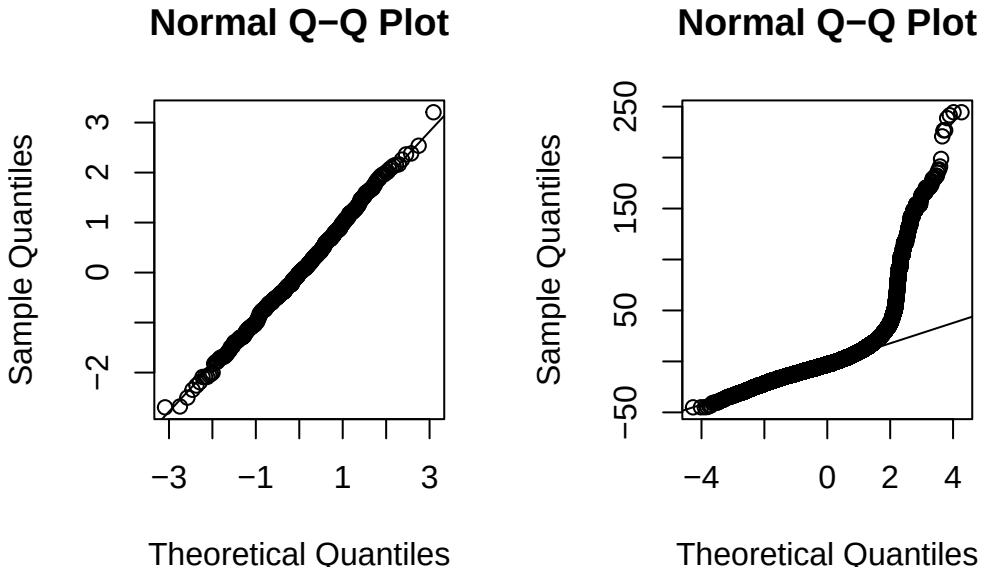


The Q-Q-plot is a graphical tool to help us assess if the errors are conditionally normally distributed.

Let $\hat{u}_{(i)}$ be the sorted residuals (i.e. $\hat{u}_{(1)} \leq ... \leq \hat{u}_{(n)}$). The Q-Q-plot plots the sorted residuals $\hat{u}_{(i)}$ against the $((i-0.5)/n)$-quantiles of the standard normal distribution.

If the residuals line up well on the straight dashed line, there is an indication that the distribution of the residuals is close to a normal distribution.

```
set.seed(123)
par(mfrow = c(1,2))
## auxiliary regression with simulated normal errors:
fit.aux = lm(rnorm(500) ~ 1)
## Q-Q-plot of the residuals of the auxiliary regression:
qqnorm(residuals(fit.aux))
qqline(residuals(fit.aux))
## Q-Q-plot of the residuals of the wage regression:
qqnorm(residuals(fit))
qqline(residuals(fit))
```



In the left plot you see the Q-Q-plot for an example with simulated normally distributed errors, where the Gaussian regression assumption is satisfied.

The right plot indicates that, in our regression of `wage` on `education` and `female`, the normality assumption is implausible.

7.11 Central Limit Theorem

Normality is a strong assumption and fails in many practical applications.

Without normality, it is not possible to construct exact confidence intervals for regression coefficients in general.

Instead, we typically rely on asymptotic arguments. The theoretical justification for these arguments is built upon the central limit theorem.

Convergence in distribution

Let $\boldsymbol{V}_n$ be a sequence of k-variate random variables and let $\boldsymbol{V}$ be a k-variate random variable

$\boldsymbol{V}_n$ **converges in distribution** to $\boldsymbol{V}$, written $\boldsymbol{V}_n \xrightarrow{d} \boldsymbol{V}$, if

$$\lim_{n\to\infty} P(\boldsymbol{V}_n \leq \boldsymbol{a}) = P(\boldsymbol{V} \leq \boldsymbol{a})$$

for all $\boldsymbol{a}$ at which the CDF of $\boldsymbol{V}$ is continuous, where "$\leq$" is componentwise.

If $\boldsymbol{V}$ has the distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we write $\boldsymbol{V}_n \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

By the univariate central limit theorem, the sample mean converges to a normal distribution:

Central Limit Theorem (CLT)

Let $W_1, \ldots, W_n$ be an i.i.d. sample with $E[W_i] = \mu$ and $\text{Var}(W_i) = \sigma^2 < \infty$. Then, the sample mean $\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$ satisfies

$$\sqrt{n}(\overline{W} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

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

SHINY APP: CLT

The same result can be extended to random vectors.

Multivariate Central Limit Theorem (MCLT)

If $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ is a multivariate i.i.d. sample with $E[\boldsymbol{W}_i] = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{W}_i) = \boldsymbol{\Sigma} < \infty$. Then, the sample mean vector $\overline{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i$ satisfies

$$\sqrt{n}(\overline{\boldsymbol{W}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

(see, e.g., Stock and Watson Section 19.2).

7.12 Asymptotic Normality of OLS

Let's apply the MCLT to the OLS vector. Consider $\boldsymbol{W}_i = \boldsymbol{X}_i u_i$, which satisfies

$$E[\boldsymbol{X}_i u_i] = \mathbf{0}, \quad \text{Var}(\boldsymbol{X}_i u_i) = E[u_i^2 \boldsymbol{X}_i \boldsymbol{X}_i'] = \boldsymbol{\Omega}.$$

Therefore, by the MCLT,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i u_i\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}).$$

By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \xrightarrow{p} \boldsymbol{Q} = E[\boldsymbol{X}_i \boldsymbol{X}_i'].$$

Combining these two results:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}}_{\xrightarrow{p} \boldsymbol{Q}^{-1}} \underbrace{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i u_i\right)}_{\xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})}.$$

If $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, then $\boldsymbol{Q}^{-1}\boldsymbol{Z}$ has variance

$$\text{Var}(\boldsymbol{Q}^{-1}\boldsymbol{Z}) = \boldsymbol{Q}^{-1}\text{Var}(\boldsymbol{Z})\boldsymbol{Q}^{-1} = \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}.$$

Hence,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}).$$

Central Limit Theorem for OLS in the heteroskedastic linear model

Consider the linear model $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i$ such that

1) **Random sampling**: $(Y_i, \boldsymbol{X}_i')$ are i.i.d.
2) **Exogeneity (mean independence)**: $E[u_i|\boldsymbol{X}_i] = 0$.
3) **Finite fourth moments**: $E[X_{ij}^4] < \infty$ and $E[u_i^4] < \infty$.
4) **Full rank**: $E[\boldsymbol{X}_i\boldsymbol{X}_i']$ is positive definite (hence invertible).

Then, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}).$$

The only additional assumption compared to the consistency of OLS is the finite fourth moments condition instead of the finite second moments condition. This technical assumption ensures that the variance of $\boldsymbol{X}_i u_i$ is finite.

Specifically, the Cauchy-Schwarz inequality implies that

$$E[X_{ij}^2 u_i^2] \leq \sqrt{E[X_{ij}^4]E[u_i^4]} < \infty,$$

so that the elements of $\boldsymbol{\Omega}$ are finite.

If homoskedasticity holds, then $\boldsymbol{\Omega} = \sigma^2\boldsymbol{Q}$, and the asymptotic variance simplifies to $\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1} = \sigma^2\boldsymbol{Q}^{-1}$.

7.13 Robust standard errors

Unlike in the Gaussian case, the standardized OLS coefficient does **not** follow a standard normal distribution in finite samples:

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j \mid \boldsymbol{X})} \nsim \mathcal{N}(0, 1).$$

However, for large samples, the central limit theorem guarantees that the OLS estimator is **asymptotically normal**.

Asymptotic standard deviation:

$$\begin{aligned}
&\sqrt{n}\,\mathrm{sd}(\hat{\beta}_j|\boldsymbol{X}) \\
&= \sqrt{\left[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\right]_{jj}} \\
&\xrightarrow{p} \sqrt{[\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}]_{jj}}
\end{aligned}$$

Asymptotic distribution:

$$\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{d} \mathcal{N}(0, [\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}]_{jj}).$$

So, the standardized coefficients satisfy

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{sd}(\hat{\beta}_j|\boldsymbol{X})} = \frac{\sqrt{n}(\hat{\beta}_j - \beta_j)}{\sqrt{n}\,\mathrm{sd}(\hat{\beta}_j|\boldsymbol{X})} \xrightarrow{d} \mathcal{N}(0,1).$$

As in the Gaussian case, we can replace the unknown conditional standard deviation by a suitable standard error.

The critical terms in the conditional standard deviation are the unobserved conditional error variances

$$\sigma_i^2 = \mathrm{Var}(u_i|\boldsymbol{X}_i).$$

We replace the unobserved σ_i^2 with the squared OLS residuals:

$$\hat{u}_i^2 = (Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}})^2.$$

This yields a consistent estimator of $\boldsymbol{\Omega}$:

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{X}_i\boldsymbol{X}_i'.$$

Substituting into the asymptotic variance formula, we obtain the **heteroskedasticity-consistent covariance matrix estimator**, also known as the **White estimator** (White, 1980):

White (HC0) Estimator

$$\widehat{\boldsymbol{V}}_{hc0} = \Big(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\Big(\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\Big)\Big(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}$$

The HC0 standard error for the j-th coefficient is the square root of the j-th diagonal entry:

$$\mathrm{se}_{hc0}(\hat{\beta}_j) = \sqrt{[\widehat{\boldsymbol{V}}_{hc0}]_{jj}}$$

This estimator remains consistent for $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$ even if the errors are heteroskedastic. However, it can be biased downward in small samples.

HC1 Correction

To reduce small-sample bias, MacKinnon and White (1985) proposed the **HC1 correction**, which rescales the estimator using a degrees-of-freedom adjustment:

$$\widehat{\boldsymbol{V}}_{hc1} = \frac{n}{n-k} \cdot \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{X}_i \boldsymbol{X}_i' \right) \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1}.$$

The **HC1 standard error** for the j-th coefficient is then:

$$\text{se}_{hc1}(\hat{\beta}_j) = \sqrt{[\widehat{\boldsymbol{V}}_{hc1}]_{jj}}.$$

These standard errors are widely used in applied work because they are valid under general forms of heteroskedasticity and easy to compute. Most statistical software (including R and Stata) uses HC1 by default when robust inference is requested.

HC3 Correction

Recall that an observation i with a high leverage value h_{ii} can distort the estimation of a linear model. Their presence might have a particularly large influence on the estimation of $\boldsymbol{\Omega}$.

An alternative way to construct robust standard errors is to weight the observations by the leverage values:

$$\widehat{\boldsymbol{\Omega}}_{\text{jack}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{u}_i^2}{(1-h_{ii})^2} \boldsymbol{X}_i \boldsymbol{X}_i'.$$

Observations with high leverage values have a small denominator $(1-h_{ii})^2$. Dividing by that amplifies their residuals in $\widehat{\boldsymbol{\Omega}}$, which tends to produce larger standard errors to prevent underestimation of variance driven by leverage.

and are therefore downweighted, which makes this estimator more robust to the influence of leverage points.

The full HC3 covariance matrix estimator is:

$$\widehat{\boldsymbol{V}}_{\text{jack}} = \widehat{\boldsymbol{V}}_{\text{hc3}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \widehat{\boldsymbol{\Omega}}_{\text{jack}} (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

There is also the HC2 estimator, which uses $\hat{u}_i^2/(1-h_{ii})$ instead of $\hat{u}_i^2/(1-h_{ii})^2$, but this is less common.

The HC3 standard errors are:

$$se_{hc3}(\hat{\beta}_j) = \sqrt{[\widehat{\boldsymbol{V}}_{hc3}]_{jj}}.$$

If you have a small sample size and you are worried about high leverage points, you should use the HC3 standard errors instead of the HC1 standard errors.

The HC3 standard error is also called **jackknife standard error** because it is based on the leave-one-out principle, similar to the way a jackknife is used to cut something. The idea is to "cut" the data by removing one observation at a time and then re-estimating the model.

Let $\hat{\boldsymbol{\beta}}_{-i}$ be the OLS estimator when using all observations except those from individual i. The difference of the full sample and the jackknife estimator is

$$\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{n} \boldsymbol{X}_j \boldsymbol{X}_j'\right)^{-1} \boldsymbol{X}_i \frac{\hat{u}_i}{1 - h_{ii}}.$$

The impact of cutting the i-th observation is proportional to $\hat{u}_i/(1 - h_{ii})$. Then, the HC3 covariance matrix can also be defined as:

$$\sum_{i=1}^{n} (\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})' = \left(\sum_{j=1}^{n} \boldsymbol{X}_j \boldsymbol{X}_j'\right)^{-1} \sum_{i=1}^{n} \frac{\hat{u}_i^2}{(1 - h_{ii})^2} \boldsymbol{X}_i \boldsymbol{X}_i' \left(\sum_{j=1}^{n} \boldsymbol{X}_j \boldsymbol{X}_j'\right)^{-1}.$$

7.14 Robust Confidence Intervals

Using heteroskedasticity-robust standard errors, we can construct confidence intervals that remain valid under heteroskedasticity.

For large samples, a $(1 - \alpha)$ confidence interval for β_j is:

$$I_{1-\alpha} = \left[\hat{\beta}_j \pm z_{1-\alpha/2} \cdot se_{hc1}(\hat{\beta}_j)\right],$$

where $z_{1-\alpha/2}$ is the standard normal critical value (e.g., $z_{0.975} = 1.96$ for a 95% interval).

For moderate sample sizes, using a t-distribution with $n - k$ degrees of freedom gives better finite-sample performance:

$$I_{1-\alpha} = \left[\hat{\beta}_j \pm t_{n-k,1-\alpha/2} \cdot se_{hc1}(\hat{\beta}_j)\right].$$

These robust intervals satisfy the asymptotic coverage property:

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

	(1)	(2)	(3)
(Intercept)	−14.082	−14.082	−14.082
	(0.434)	(0.500)	(0.500)
education	2.958	2.958	2.958
	(0.030)	(0.040)	(0.040)
female	−7.533	−7.533	−7.533
	(0.169)	(0.162)	(0.162)
Num.Obs.	50 742	50 742	50 742
R2	0.180	0.180	0.180
RMSE	18.76	18.76	18.76
Std.Errors	IID	HC1	HC3

i Why software uses t-quantiles:

There's no exact finite-sample justification under generic heteroskedasticity. Asymptotically, both t_{n-k} quantiles and standard normal quantiles are valid. Most software uses t-quantiles by default to match the homoskedastic case and improve finite-sample performance. For large samples, this makes little difference, as t-quantiles converge to standard normal quantiles as degrees of freedom grow large.

The `vcov` argument of the `modelsummary()` function allows you to specify the type of covariance matrix estimator to use.

```
## Standard error comparison:
fit = lm(wage ~ education + female, data = cps)
modelsummary(fit,
             vcov = list("IID", "HC1", "HC3"),
             gof_map = c("nobs", "r.squared", "rmse", "vcov.type"))
```

The homoskedasticity-only standard errors (called IID in R) differ from the robust standard errors. The HC1 and HC3 standard errors coincide up to 3 digits after the decimal point in this example.

In practice you should always use HC1 or HC3 standard errors unless you have very good reasons to believe that the Gaussianity and homoskedasticity assumption hold.

```
data(CASchools, package = "AER")
CASchools$STR = CASchools$students/CASchools$teachers
CASchools$score = (CASchools$read+CASchools$math)/2
```

	(1)	(2)	(3)
(Intercept)	686.032	686.032	686.032
	(7.411)	(8.728)	(8.812)
STR	−1.101	−1.101	−1.101
	(0.380)	(0.433)	(0.437)
english	−0.650	−0.650	−0.650
	(0.039)	(0.031)	(0.031)
Num.Obs.	420	420	420
R2	0.426	0.426	0.426
RMSE	14.41	14.41	14.41
Std.Errors	IID	HC1	HC3

```
fit.CASchools = lm(score ~ STR + english, data = CASchools)
## Standard error comparison:
modelsummary(fit.CASchools,
             vcov = list("IID", "HC1", "HC3"),
             gof_map = c("nobs", "r.squared", "rmse", "vcov.type"))
```

Here, HC1 and HC3 standard errors differ slightly. You can also display the confidence interval directly:

```
## Confidence interval comparison:
modelsummary(fit.CASchools,
             vcov = list("IID", "HC1", "HC3"),
             statistic = "conf.int",
             gof_map = c("nobs", "r.squared", "rmse", "vcov.type"))
```

7.15 Summary

- Under i.i.d. sampling, exogeneity, finite second moments, and full rank design matrix, OLS is consistent
- In addition, under finite fourth moments, OLS is asymptotically normal
- Under homoskedastic errors, confidence intervals with classical standard errors are asymptotically valid
- Under homoskedastic and normally distributed errors, confidence intervals with classical standard errors are exactly valid if t-quantiles are used

	(1)	(2)	(3)
(Intercept)	686.032	686.032	686.032
	[671.464, 700.600]	[668.875, 703.189]	[668.710, 703.354]
STR	−1.101	−1.101	−1.101
	[−1.849, −0.354]	[−1.952, −0.250]	[−1.960, −0.242]
english	−0.650	−0.650	−0.650
	[−0.727, −0.572]	[−0.711, −0.589]	[−0.711, −0.588]
Num.Obs.	420	420	420
R2	0.426	0.426	0.426
RMSE	14.41	14.41	14.41
Std.Errors	IID	HC1	HC3

- Without homoskedasticity, confidence intervals with HC1/HC3 standard errors are asymptotically valid.
- If i.i.d. sampling does not hold, other standard errors must be used. Under clustered sampling, use cluster-robust standard errors. For stationary time series, use HAC (heteroskedasticity and autocorrelation consistent) standard errors

7.16 R Code

statistics-sec07.R

8 Testing

In applied regression analysis, we often want to assess whether a regressor has a statistically significant relationship with the outcome variable (conditional on other regressors).

8.1 t-Test

The most common hypothesis test evaluates whether a regression coefficient equals zero:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

This corresponds to testing whether the marginal effect of the regressor X_{ij} on the outcome Y_i is zero, holding other regressors constant.

We use the t-statistic:

$$T_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)},$$

where $se(\hat{\beta}_j)$ is a standard error.

You may use the classical standard error if you have strong evidence that the errors are homoskedastic. However, in most economic applications, heteroskedasticity-robust standard errors are more reliable.

Under the null, T_j follows approximately a t_{n-k} distribution. We reject H_0 at the significance level α if:

$$|T_j| > t_{n-k,1-\alpha/2}.$$

This decision rule is equivalent to checking whether the confidence interval for β_j includes 0:

- Reject H_0 if 0 lies **outside** the $1-\alpha$ confidence interval
- Fail to reject (accept) H_0 if 0 lies **inside** the $1-\alpha$ confidence interval

---

8.2 p-Value

The **p-value** is a criterion to reach a hypothesis 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 represents the probability of observing a test statistic as extreme or more extreme than the one we computed, assuming H_0 is true. For the t-test, the p-value is:

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

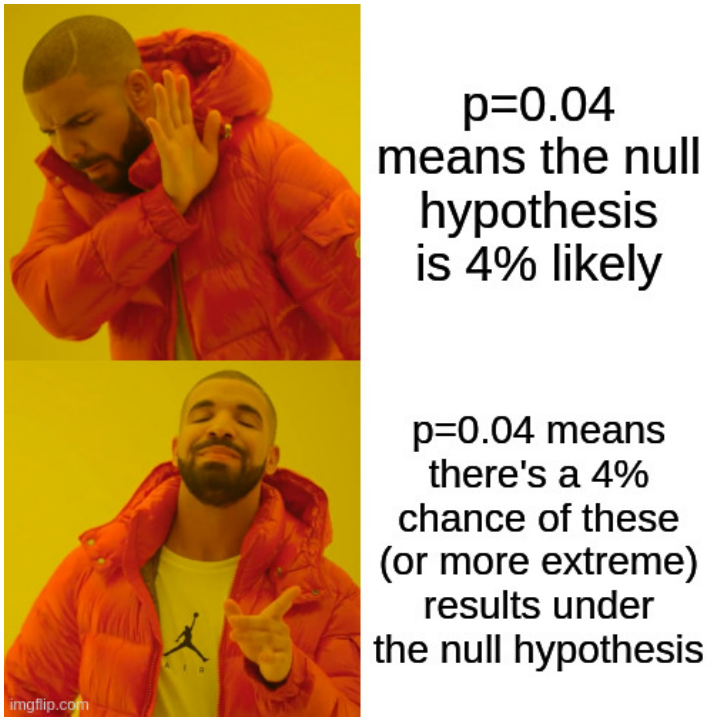
Here, T is a random variable following the null distribution $Z \sim t_{n-k}$, and T_j is the observed value of the test statistic.

Another way of representing the p-values of a t-test is:

$$p\text{-value} = 2(1 - F_{t_{n-k}}(|T_j|)),$$

where $F_{t_{n-k}}$ is the cumulative distribution function (CDF) of the t_{n-k}-distribution.

A common misinterpretation of p-values is treating them as the probability that the null hypothesis is being true. This is incorrect. The p-value is not a statement about the probability of the null hypothesis itself.



The correct interpretation is that the p-value represents the probability of observing a test statistic at least as extreme as the one calculated from our sample, assuming that the null hypothesis is true.

In other words, a p-value of 0.04 means:

- NOT "There's a 4% chance that the null hypothesis is true"
- INSTEAD "If the null hypothesis were true, there would be a 4% chance of observing a test statistic this extreme or more extreme"

Small p-values indicate that the observed data would be unlikely under the null hypothesis, which leads us to reject the null in favor of the alternative. However, they do not tell us the probability that our alternative hypothesis is correct, nor do they directly measure the magnitude or significance of the marginal effect.

> **i** Relation to Confidence Intervals:
>
> Zero lies outside the $(1-\alpha)$ confidence interval for β_j if and only if the p-value for testing $H_0 : \beta_j = 0$ is less than α.

8.3 Significance Stars

Regression tables often use asterisks to indicate levels of statistical significance. Stars summarize statistical significance by comparing the t-statistic to critical values (or equivalently, the p-value or whether 0 is covered by the confidence interval)

The convention within `R` is:

Stars	p-value	t-statistic	Confidence interval
***	$p < 0.001$	$\lvert T_j \rvert > t_{n-k,0.995}$	0 outside $I_{0.999}$
**	$0.001 \leq p < 0.01$	$t_{n-k,0.995} \geq \lvert T_j \rvert > t_{n-k,0.975}$	0 outside $I_{0.99}$, but inside $I_{0.999}$
*	$0.01 \leq p < 0.05$	$t_{n-k,0.975} \geq \lvert T_j \rvert > t_{n-k,0.95}$	0 outside $I_{0.95}$, but inside $I_{0.99}$

	(1)	(2)
(Intercept)	−14.082***	−14.082***
	(0.434)	(0.500)
education	2.958***	2.958***
	(0.030)	(0.040)
female	−7.533***	−7.533***
	(0.169)	(0.162)
Num.Obs.	50 742	50 742
R2	0.180	0.180
R2 Adj.	0.180	0.180
AIC	441 515.9	441 515.9
BIC	441 542.4	441 542.4
RMSE	18.76	18.76
Std.Errors	IID	Heteroskedasticity-robust

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

> **Significance Stars Convention**
>
> Note that most economists use the following significance levels: `***` for 1%, `**` for 5%, and `*` for 10%. In this lecture, we follow the convention of `R`, which uses the significance levels `***` for 0.1%, `**` for 1%, and `*` for 5%.

Regression Tables

Let's revisit the regression of `wage` on `education` and `female`.

```
library(fixest)
library(modelsummary)
cps = read.csv("cps.csv")
fit.hom = feols(wage ~ education + female, data = cps, vcov = "iid")
fit.het = feols(wage ~ education + female, data = cps, vcov = "hc1")
mymodels = list(fit.hom, fit.het)
modelsummary(mymodels, stars = TRUE)
```

To see the exact t-statistics and p-values, you can use the `summary()` function:

```
summary(fit.hom)
```

```
OLS estimation, Dep. Var.: wage
Observations: 50,742
Standard-errors: IID
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -14.08179   0.433883 -32.4552 < 2.2e-16 ***
education     2.95817   0.030373  97.3953 < 2.2e-16 ***
female       -7.53307   0.168582 -44.6848 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.8   Adj. R2: 0.179696
```

```
summary(fit.het)
```

```
OLS estimation, Dep. Var.: wage
Observations: 50,742
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -14.08179   0.500078 -28.1592 < 2.2e-16 ***
education     2.95817   0.040110  73.7512 < 2.2e-16 ***
female       -7.53307   0.161644 -46.6027 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 18.8   Adj. R2: 0.179696
```

All p-values are super small: `2.2e-16` means $2.2 \cdot 10^{-16}$ (15 zeros after the decimal point, followed by 22).

Let's also revisit the `CASchools` dataset and examine four regression models on test scores.

```
library(AER)
data(CASchools, package = "AER")
CASchools$STR = CASchools$students/CASchools$teachers
CASchools$score = (CASchools$read + CASchools$math)/2

fitA = feols(score ~ STR, data = CASchools)
fitB = feols(score ~ STR + english, data = CASchools)
fitC = feols(score ~ STR + english + lunch, data = CASchools)
fitD = feols(score ~ STR + english + lunch + expenditure, data = CASchools)
```

	(1)	(2)	(3)	(4)
(Intercept)	698.933***	686.032***	700.150***	665.988***
	(9.467)	(7.411)	(4.686)	(9.460)
STR	−2.280***	−1.101**	−0.998***	−0.235
	(0.480)	(0.380)	(0.239)	(0.298)
english		−0.650***	−0.122***	−0.128***
		(0.039)	(0.032)	(0.032)
lunch			−0.547***	−0.546***
			(0.022)	(0.021)
expenditure				0.004***
				(0.001)
Num.Obs.	420	420	420	420
R2	0.051	0.426	0.775	0.783
R2 Adj.	0.049	0.424	0.773	0.781
AIC	3648.5	3439.1	3049.0	3034.1
BIC	3656.6	3451.2	3065.2	3054.3
RMSE	18.54	14.41	9.04	8.86
Std.Errors	IID	IID	IID	IID

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Classical (Homoskedastic) Standard Errors

```
mymodels = list(fitA, fitB, fitC, fitD)
modelsummary(mymodels, stars = TRUE, vcov = "iid")
```

Robust (HC1) Standard Errors

```
mymodels = list(fitA, fitB, fitC, fitD)
modelsummary(mymodels, stars = TRUE, vcov = "HC1")
```

Interpretation of `STR` coefficient:

	(1)	(2)	(3)	(4)
(Intercept)	698.933***	686.032***	700.150***	665.988***
	(10.364)	(8.728)	(5.568)	(10.377)
STR	−2.280***	−1.101*	−0.998***	−0.235
	(0.519)	(0.433)	(0.270)	(0.325)
english		−0.650***	−0.122***	−0.128***
		(0.031)	(0.033)	(0.032)
lunch			−0.547***	−0.546***
			(0.024)	(0.023)
expenditure				0.004***
				(0.001)
Num.Obs.	420	420	420	420
R2	0.051	0.426	0.775	0.783
R2 Adj.	0.049	0.424	0.773	0.781
AIC	3648.5	3439.1	3049.0	3034.1
BIC	3656.6	3451.2	3065.2	3054.3
RMSE	18.54	14.41	9.04	8.86
Std.Errors	HC1	HC1	HC1	HC1

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

- Models A–C: The coefficient is negative and statistically significant. However, when using robust standard errors, the coefficient in model B becomes only weakly significant.
- Model D: The coefficient remains negative but becomes insignificant when controlling for `expenditure`.

As discussed earlier, `expenditure` is a **bad control** in this context and should not be used to estimate a ceteris paribus effect of class size on test scores.

8.4 Testing for Heteroskedasticity: Breusch-Pagan Test

Classical standard errors should only be used if you have statistical evidence that the errors are homoskedastic. A statistical test for this is the **Breusch-Pagan Test**.

Under homoskedasticity, the variance of the error term is constant and does not depend on the values of the regressors:

$$Var(u_i \mid \boldsymbol{X}_i) = \sigma^2 \quad \text{(constant)}.$$

To test this assumption, we perform an auxiliary regression of the squared residuals on the original regressors:

$$\hat{u}_i^2 = \boldsymbol{X}_i'\boldsymbol{\gamma} + v_i, \quad i = 1, \dots, n,$$

where:

- $\hat{u}_i$ are the OLS residuals from the original model,
- $\boldsymbol{\gamma}$ are auxiliary coefficients,
- v_i is the error term in the auxiliary regression.

If homoskedasticity holds, the regressors should not explain any variation in $\hat{u}_i^2$, which means the auxiliary regression should have low explanatory power.

Let R^2_{aux} be the R-squared from this auxiliary regression. Then, the **Breusch–Pagan (BP) test statistic** is:

$$BP = n \cdot R^2_{\text{aux}}$$

Under the null hypothesis of homoskedasticity,

$$H_0 : Var(u_i \mid \boldsymbol{X}_i) = \sigma^2,$$

the test statistic follows an asymptotic chi-squared distribution with $k-1$ degrees of freedom:

$$BP \xrightarrow{d} \chi^2_{k-1}$$

We **reject** H_0 at significance level α if:

$$BP > \chi^2_{1-\alpha,\, k-1}.$$

This basic variant of the BP test is Koenker's version of the test. Other variants include further nonlinear transformations of the regressors.

In R, the test is implemented via the `bptest()` function from the **AER** package. Unfortunately, the `bptest()` function does not work directly with `feols` objects, so we need to estimate the model first with `lm()`:

```
fit = lm(wage ~ education + female, data = cps)
bptest(fit)
```

```
    studentized Breusch-Pagan test

data:  fit
BP = 1070.3, df = 2, p-value < 2.2e-16
```

In the wage regression the BP test clearly rejects H_0, which is strong statistical evidence that the errors are heteroskedastic.

Let's apply the test to the `CASchools` model:

```
lm(score ~ STR + english, data = CASchools) |> bptest()
```

```
    studentized Breusch-Pagan test

data:  lm(score ~ STR + english, data = CASchools)
BP = 29.501, df = 2, p-value = 3.926e-07
```

```
lm(score ~ STR + english + lunch, data = CASchools) |> bptest()
```

```
    studentized Breusch-Pagan test

data:  lm(score ~ STR + english + lunch, data = CASchools)
BP = 9.9375, df = 3, p-value = 0.0191
```

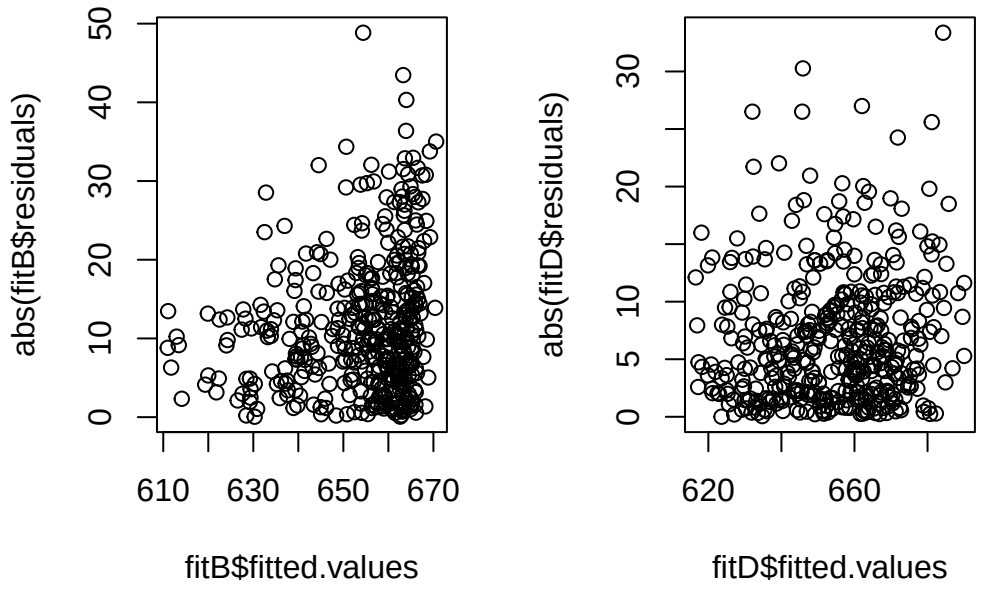
```
lm(score ~ STR + english + lunch + expenditure, data = CASchools) |> bptest()
```

```
    studentized Breusch-Pagan test

data:  lm(score ~ STR + english + lunch + expenditure, data = CASchools)
BP = 5.9649, df = 4, p-value = 0.2018
```

In the regression of `score` on `STR` and `english` there is strong statistical evidence that errors are heteroskedastic, whereas when adding `lunch` and `expenditure` there is no evidence of heteroskedasticity. See the difference in the absolute residuals against fitted values plot:

```
par(mfrow = c(1,2))
plot(abs(fitB$residuals) ~ fitB$fitted.values)
plot(abs(fitD$residuals) ~ fitD$fitted.values)
```



The heteroskedasticity pattern in model (2) likely occurred because of a nonlinear dependence of the omitted variables `lunch` and `expenditure` with the included regressors `STR` and `english`. The inclusion of these variables in model (4) eliminated the heteroskedasticity (apparent heteroskedasticity). Therefore, heteroskedasticity is sometimes a sign of model misspecification.

8.5 Testing for Normality: Jarque–Bera Test

A general property of a normally distributed variable is that it has zero skewness and kurtosis of three. In the Gaussian regression model, this implies:

$$u_i|\boldsymbol{X}_i \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad E[u_i^3] = 0, \quad E[u_i^4] = 3\sigma^4.$$

The **sample skewness** and **sample kurtosis** of the OLS residuals are:

$$\widehat{\text{ske}}(\hat{\boldsymbol{u}}) = \frac{1}{n\hat{\sigma}_{\hat{u}}^3} \sum_{i=1}^{n} \hat{u}_i^3, \quad \widehat{\text{kur}}(\hat{\boldsymbol{u}}) = \frac{1}{n\hat{\sigma}_{\hat{u}}^4} \sum_{i=1}^{n} \hat{u}_i^4$$

A joint test for normality — assessing both skewness and kurtosis — is the **Jarque–Bera (JB) test**, with statistic:

$$JB = n \left(\frac{1}{6} \widehat{\text{ske}}(\hat{\boldsymbol{u}})^2 + \frac{1}{24} (\widehat{\text{kur}}(\hat{\boldsymbol{u}}) - 3)^2 \right)$$

Under the null hypothesis of normal errors, this test statistic is asymptotically chi-squared distributed:

$$JB \xrightarrow{d} \chi_2^2$$

We reject H_0 at level α if:

$$JB > \chi_{1-\alpha,\,2}^2.$$

In R, we can apply the test using the `moments` package:

```
library(moments)
jarque.test(fitD$residuals)
```

```
    Jarque-Bera Normality Test

data:  fitD$residuals
JB = 8.9614, p-value = 0.01133
alternative hypothesis: greater
```

Although the Breusch–Pagan test does not reject homoskedasticity for `fitD` (so classical standard errors are valid asymptotically), the JB rejects the null hypothesis of normal errors at the 5% level and provides statistical evidence that the errors are not normally distributed.

This means that exact inference based on t-distributions is not valid in finite samples, and confidence intervals or t-test results give only large sample approximations.

In econometrics, asymptotic large sample approximations have become the convention because exact finite sample inference is rarely feasible.

8.6 Joint Hypothesis Testing

So far, we've tested whether a single coefficient is zero. But often we want to test **multiple restrictions simultaneously**, such as whether a group of variables has a joint effect.

The **joint exclusion** hypothesis formulates the null hypothesis that a set of coefficients or linear combinations of coefficients are equal to zero:

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \mathbf{0}$$

where:

- $\boldsymbol{R}$ is a $q \times k$ restriction matrix,
- $\mathbf{0}$ is the $q \times 1$ vector of zeros,
- q is the number of restrictions.

Consider for example the `score` on `STR` regression with interaction effects:

$$\text{score}_i = \beta_1 + \beta_2 \text{STR}_i + \beta_3 \text{HiEL}_i + \beta_4 \text{STR}_i \cdot \text{HiEL}_i + u_i.$$

```
## Create dummy variable for high proportion of English learners
CASchools$HiEL = (CASchools$english >= 10) |> as.numeric()
fitE = feols(score ~ STR + HiEL + STR:HiEL, data = CASchools, vcov = "hc1")
fitE |> summary()
```

```
OLS estimation, Dep. Var.: score
Observations: 420
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 682.245837  11.867815 57.487065 < 2.2e-16 ***
STR          -0.968460   0.589102 -1.643961   0.10094
HiEL          5.639135  19.514560  0.288971   0.77275
STR:HiEL     -1.276613   0.966920 -1.320289   0.18746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.8   Adj. R2: 0.305368
```

The model output reveals that none of the individual t-tests reject the null hypothesis that the individual coefficients are zero.

However, these results are misleading because the true marginal effects are a mixture of these coefficients:

$$\frac{\partial E[\text{score}_i \mid \boldsymbol{X}_i]}{\partial \text{STR}_i} = \beta_2 + \beta_4 \cdot \text{HiEL}_i.$$

Therefore, to test if `STR` has an effect on `score`, we need to test the joint hypothesis:

$$H_0 : \beta_2 = 0 \quad \text{and} \quad \beta_4 = 0.$$

In terms of the multiple restriction notation $H_0 : \boldsymbol{R\beta} = \mathbf{0}$, we have

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Similarly, the marginal effects of `HiEL` is:

$$\frac{\partial E[\text{score}_i \mid \boldsymbol{X}_i]}{\partial \text{HiEL}_i} = \beta_3 + \beta_4 \cdot \text{STR}_i.$$

We test the joint hypothesis that $\beta_3 = 0$ and $\beta_4 = 0$:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Wald Test

The **Wald test** is based on the **Wald distance**:

$$\boldsymbol{d} = \boldsymbol{R}\hat{\boldsymbol{\beta}},$$

which measures how far the estimated coefficients deviate from the hypothesized restrictions.

The covariance matrix of the Wald distance is: $Var(\boldsymbol{d}|\boldsymbol{X}) = \boldsymbol{R}Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X})\boldsymbol{R}'$, which can be estimated as:

$$\widehat{Var}(\boldsymbol{d} \mid \boldsymbol{X}) = \boldsymbol{R}\widehat{\boldsymbol{V}}\boldsymbol{R}'.$$

The **Wald statistic** is the squared, variance-standardized distance:

$$W = \boldsymbol{d}'(\boldsymbol{R}\widehat{\boldsymbol{V}}\boldsymbol{R}')^{-1}\boldsymbol{d},$$

where $\widehat{\boldsymbol{V}}$ is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ (e.g., HC1 robust: $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_{hc1}$).

Under the null hypothesis, and assuming (A1)–(A4), the Wald statistic has an asymptotic chi-squared distribution:

$$W \xrightarrow{d} \chi^2_q,$$

where q is the number of restrictions.

The null is rejected if $W > \chi^2_{1-\alpha,q}$.

F-test

The Wald test is an asymptotic size-α-test under (A1)–(A4). Even if normality and homoskedasticity hold true as well, the Wald test is still only asymptotically valid, i.e.:

$$\lim_{n\to\infty} P(\text{Wald test rejects } H_0 | H_0 \text{ true}) = \alpha.$$

The F-test is the small sample correction of the Wald test. It is based on the same distance as the Wald test, but it is scaled by the number of restrictions q:

$$F = \frac{W}{q} = \frac{1}{q}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}\widehat{\boldsymbol{V}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}).$$

Under the restrictive assumption that the Gaussian regression model holds, and if $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_{hom}$ is used, it can be shown that

$$F \sim F_{q;n-k}$$

for any finite sample size n. Here, $F_{q;n-k}$ is the F-distribution with q degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator.

The test decision for the **F-test**:

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } F \leq F_{(1-\alpha,q,n-k)}, \\ \text{reject } H_0 \quad & \text{if } F > F_{(1-\alpha,q,n-k)}, \end{aligned}$$

where $F_{(p,m_1,m_2)}$ is the p-quantile of the F distribution with m_1 degrees of freedom in the numerator and m_2 degrees of freedom in the denominator.

> **i** F- and Chi-squared distribution
>
> Similar to how the t-distribution t_{n-k} approaches the standard normal as sample size increases, we have $q \cdot F_{q;n-k} \to \chi^2_q$ as $n \to \infty$. Therefore, the F-test and Wald test become asymptotically equivalent and lead to identical statistical conclusions in large samples. For single constraint (q=1) hypotheses of the form $H_0 : \beta_j = 0$, the F-test is equivalent to a two-sided t-test.
> The F-test can be viewed as a finite-sample correction of the Wald test. It tends to be more conservative than the Wald test in small samples, meaning that rejection by the F-test generally implies rejection by the Wald test, but not necessarily vice versa. Due to this more conservative nature, which helps control false rejections (Type I errors) in small samples, the F-test is often preferred in practice.

F-tests in R

The function `wald()` from the `fixest` package performs an F-test:

```
wald(fitE, keep = "STR")
```

```
Wald test, H0: joint nullity of STR and STR:HiEL
 stat = 5.6381, p-value = 0.003837, on 2 and 416 DoF, VCOV: Heteroskedasticity-robust.
```

```
wald(fitE, keep = "HiEL")
```

```
Wald test, H0: joint nullity of HiEL and STR:HiEL
 stat = 89.9, p-value < 2.2e-16, on 2 and 416 DoF, VCOV: Heteroskedasticity-robust.
```

The hypotheses that `STR` and `HiEL` have no effect on `score` can be clearly rejected.

Another research question is whether the effect of `STR` on `score` is zero only for the subgroup of schools with a high proportion of English learners (`HiEL` $= 1$). In this case, the marginal effect is:

$$\frac{\partial E[\text{score}_i \mid \boldsymbol{X}_i, \text{HiEL}_i = 1]}{\partial \text{STR}_i} = \beta_2 + \beta_4 \cdot 1,$$

and the null hypothesis is:

$$H_0 : \beta_2 + \beta_4 = 0.$$

The corresponding restriction matrix is:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix},$$

where the number of restrictions is $q = 1$.

The function `linearHypothesis()` from the `AER` package is more flexible for these cases:

```
## Define hypothesis matrix:
R = matrix(c(0,1,0,1), ncol = 4)
linearHypothesis(fitE, hypothesis.matrix = R, test = "F", vcov. = vcovHC(fitE, type = "HC1"))
```

```
Linear hypothesis test:
STR  + STR:HiEL = 0

Model 1: restricted model
Model 2: score ~ STR + HiEL + STR:HiEL
```

```
Note: Coefficient covariance matrix supplied.

  Res.Df Df      F   Pr(>F)
1    417
2    416  1 8.5736 0.003598 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarly, this hypothesis can be rejected at the 0.01 level.

8.7 Jackknife Methods

Projection Matrix

Recall the vector of fitted values $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Inserting the model equation gives:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{=\boldsymbol{P}}\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}.$$

The **projection matrix** $\boldsymbol{P}$ is also known as the *influence matrix* or *hat matrix* and maps observed values to fitted values.

Leverage Values

The diagonal entries of $\boldsymbol{P}$, given by

$$h_{ii} = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i,$$

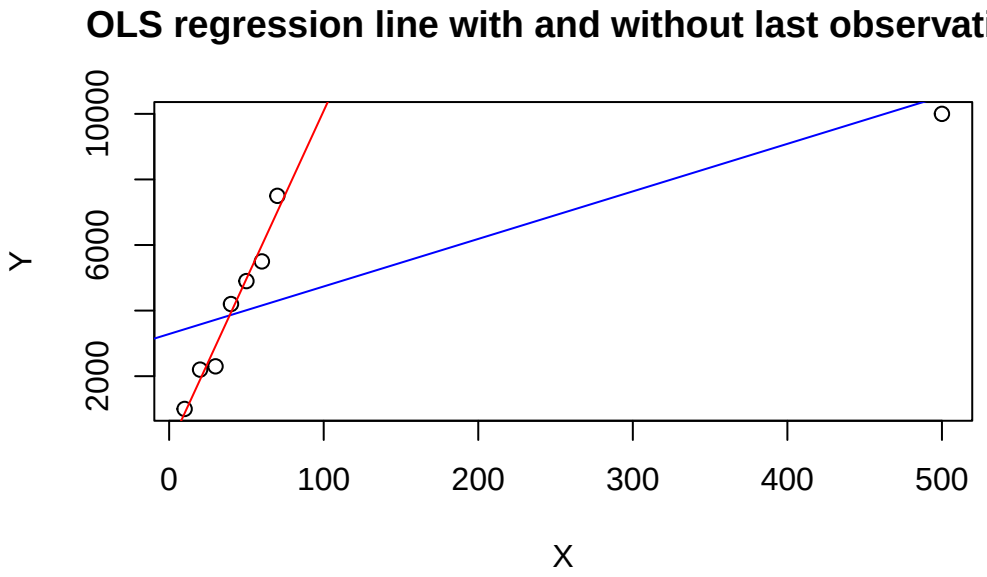
are called **leverage values** or *hat values* and measure how far away the regressor values of the i-th observation X_i are from those of the other observations.

Properties of leverage values:

$$0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{n} h_{ii} = k.$$

Leverage values h_{ii} indicate how much influence an observation $\boldsymbol{X}_i$ has on the regression fit, e.g., the last observation in the following artificial 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")
```




```r
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
```

A low leverage implies the presence of many regressor observations similar to $\boldsymbol{X}_i$ in the sample, while a high leverage indicates a lack of similar observations near $\boldsymbol{X}_i$.

An observation with a high leverage h_{ii} but a response value Y_i that is close to the true regression line $\boldsymbol{X}_i'\boldsymbol{\beta}$ (indicating a small error u_i) is considered a **good leverage point**. Despite being unusual in the regressor space, this point improves estimation precision because it provides valuable information about the regression relationship in regions where data is sparse.

Conversely, a **bad leverage point** occurs when both h_{ii} and the error u_i are large, indicating both unusual regressor and response values. This can misleadingly impact the regression fit.

The actual error term is unknown, but standardized residuals can be used to differentiate between good and bad leverage points.

Standardized Residuals

Many regression diagnostic tools rely on the residuals of the OLS estimation $\hat{u}_i$ because they provide insight into the properties of the unknown error terms u_i.

Under the homoskedastic linear regression model (A1)–(A5), the errors are independent and have the property

$$Var(u_i \mid \boldsymbol{X}) = \sigma^2.$$

Since $\boldsymbol{PX} = \boldsymbol{X}$ and, therefore,

$$\hat{\boldsymbol{u}} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{u},$$

the residuals have a different property:

$$Var(\hat{\boldsymbol{u}} \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{P}).$$

The i-th residual satisfies

$$Var(\hat{u}_i \mid \boldsymbol{X}) = \sigma^2(1 - h_{ii}),$$

where h_{ii} is the i-th leverage value.

Under the assumption of homoskedasticity, the variance of $\hat{u}_i$ depends on $\boldsymbol{X}$, while the variance of u_i does not. Dividing by $\sqrt{1 - h_{ii}}$ removes the dependency:

$$Var\left(\frac{\hat{u}_i}{\sqrt{1 - h_{ii}}} \,\middle|\, \boldsymbol{X} \right) = \sigma^2$$

The **standardized residuals** are defined as follows:

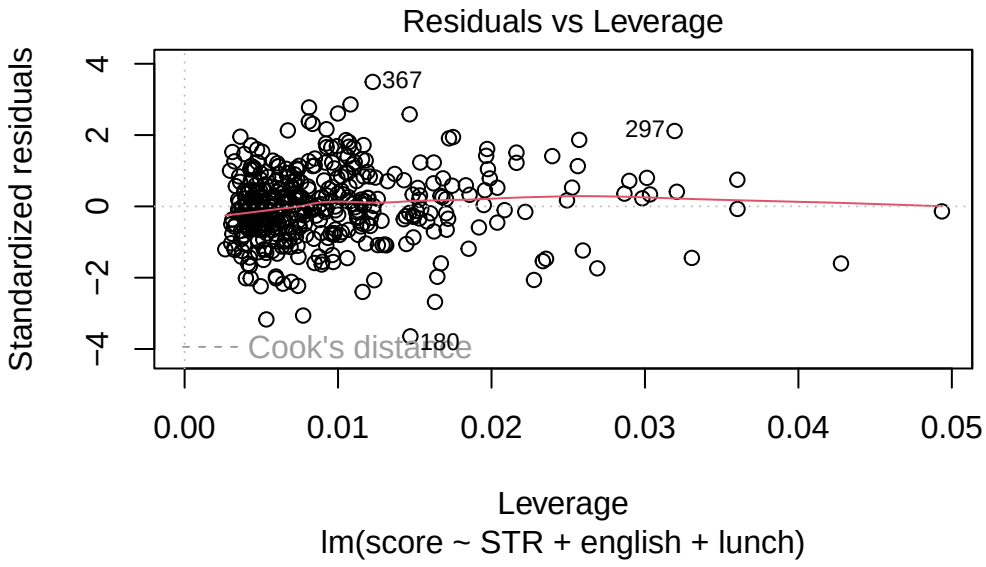
$$r_i := \frac{\hat{u}_i}{\sqrt{s^2_{\hat{u}}(1 - h_{ii})}}.$$

Standardized residuals are available using the R command `rstandard()`.

Residuals vs. Leverage Plot

Plotting standardized residuals against leverage values provides a graphical tool for detecting outliers. High leverage points have a strong influence on the regression fit. High leverage values with standardized residuals close to 0 are good leverage points, and high leverage values with large standardized residuals are bad leverage points.

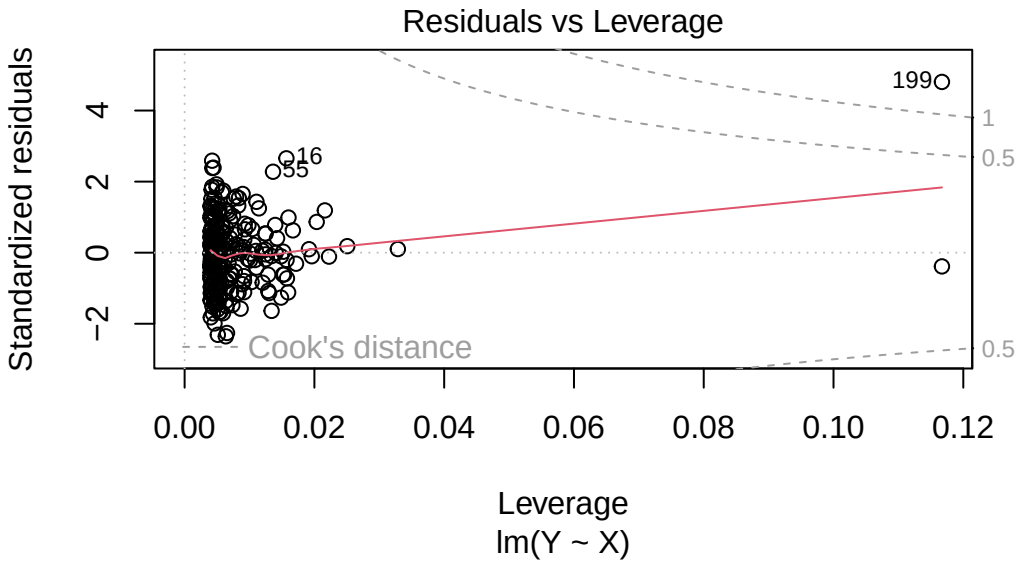
```
fit = lm(score ~ STR + english + lunch, data = CASchools)
plot(fit, which = 5)
```



The plot indicates that some observations have a higher leverage value than others, but none of these have a large standardized residual, so they are not bad leverage points.

Here is an example with two high leverage points. Observation $i = 200$ is a good leverage point and $i = 199$ is a bad leverage point:

```
## simulate regressors and errors
X = rnorm(250)
u = rnorm(250)
## set some unusual observations manually
X[199] = 6
X[200] = 6
u[199] = 5
u[200] = 0
## define dependent variable
Y = X + u
## residuals vs leverage plot
plot(lm(Y ~ X), which = 5)
```



The plot also shows Cook's distance thresholds. Cook's distance for observation i is defined as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})' \boldsymbol{X}' \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})}{k s^2_{\hat{u}}},$$

where

$$\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}_i \frac{\hat{u}_i}{1 - h_{ii}}.$$

Here, $\hat{\boldsymbol{\beta}}_{(-i)}$ is the i-th leave-one-out estimator (the OLS estimator when the i-th observation is left out).

This principle is called **Jackknife** because it is similar to the way a jackknife is used to cut something. The idea is to "cut" the data by removing one observation at a time and then re-estimating the model. The impact of cutting the i-th observation is proportional to $\hat{u}_i/(1 - h_{ii})$.

We should pay special attention to points outside Cook's distance thresholds of 0.5 and 1 and check for measurement errors or other anomalies.

Jackknife Standard Errors

Recall the heteroskedasticity-robust White estimator for the meat matrix $\boldsymbol{\Omega} = E[u_i^2 \boldsymbol{X}_i \boldsymbol{X}_i']$ in the sandwich formula tor the OLS variance:

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{X}_i \boldsymbol{X}_i'.$$

If there are leverage points in the data, their presence might have a large influence on the estimation of $\boldsymbol{\Omega}$.

An alternative way of estimating the covariance matrix is to weight the observations by the leverage values:

$$\widehat{\boldsymbol{\Omega}}_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{u}_i^2}{(1-h_{ii})^2}\boldsymbol{X}_i\boldsymbol{X}_i'.$$

Observations with high leverage values have a small denominator $(1-h_{ii})^2$ and are therefore downweighted, which makes this estimator more robust to the influence of leverage points.

The full jackknife covariance matrix estimator is conventionally labeled as the **HC3** estimator:

$$\widehat{\boldsymbol{V}}_{\text{jack}} = \widehat{\boldsymbol{V}}_{\text{hc3}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\widehat{\boldsymbol{\Omega}}_{\text{jack}}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

There is also the HC2 estimator, which uses $\hat{u}_i^2(1-h_{ii})$ instead of $\hat{u}_i^2/(1-h_{ii})^2$, but this is less common.

The HC3 standard errors are:

$$se_{hc3}(\hat{\beta}_j) = \sqrt{[\widehat{\boldsymbol{V}}_{hc3}]_{jj}}.$$

If you have a small sample size and you are worried about influential observations, you should use the HC3 standard errors instead of the HC1 standard errors.

To display the HC3 standard errors in the regression table, you can use `modelsummary(fit, vcov = "HC3")`.

8.8 Cluster-robust Inference

Recall that in many economic applications, observations are naturally clustered. For instance, students within the same school, workers in the same firm, or households in the same village may share common unobserved factors that induce correlation in their outcomes.

As discussed in Section 5, for clustered observations we can use the notation $(\boldsymbol{X}_{ig}, Y_{ig})$, where the linear regression equation is:

$$Y_{ig} = \boldsymbol{X}_{ig}'\boldsymbol{\beta} + u_{ig}, \quad i = 1, \dots, n_g, \quad g = 1, \dots, G.$$

Under independence across clusters but allowing for arbitrary correlation within clusters, the OLS estimator remains unbiased, but its standard variance formula is no longer valid. As we saw in Section 5, the conditional variance

$$Var(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

satisfies

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{g=1}^{G} E\Big[\Big(\sum_{i=1}^{n_g} \boldsymbol{X}_{ig} u_{ig}\Big)\Big(\sum_{i=1}^{n_g} \boldsymbol{X}_{ig} u_{ig}\Big)' \Big| \boldsymbol{X}\Big].$$

Cluster-robust Standard Errors

When observations within clusters are correlated, using ordinary standard errors (even heteroskedasticity-robust ones) will typically underestimate the true sampling variability of the OLS estimator.

To account for within-cluster correlation, we use **cluster-robust standard errors**. The key insight is to estimate the middle part of the sandwich formula above by allowing for arbitrary within-cluster correlation, while maintaining the independence assumption across clusters.

The **cluster-robust variance estimator** is:

$$\widehat{\boldsymbol{V}}_{CR0} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \sum_{g=1}^{G} \Big(\sum_{i=1}^{n_g} \boldsymbol{X}_{ig} \hat{u}_{ig}\Big)\Big(\sum_{i=1}^{n_g} \boldsymbol{X}_{ig} \hat{u}_{ig}\Big)' (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

This estimator, also known as the **clustered sandwich estimator**, allows for arbitrary correlation of errors within clusters, including both heteroskedasticity and serial correlation. Like the HC estimators, it is consistent under large-sample asymptotics.

Finite Sample Correction

Similar to the HC1 correction for heteroskedasticity, a small-sample correction for the cluster-robust estimator is commonly applied:

$$\widehat{\boldsymbol{V}}_{CR1} = \frac{G}{G-1} \cdot \frac{n-1}{n-k} \cdot \widehat{\boldsymbol{V}}_{CR0},$$

where G is the number of clusters, n is the total sample size, and k is the number of regressors.

The corresponding cluster-robust standard errors are:

$$se_{CR1}(\hat{\beta}_j) = \sqrt{[\widehat{\boldsymbol{V}}_{CR1}]_{jj}}.$$

When to Cluster

You should use cluster-robust standard errors when:

1. There's a clear grouping structure in your data (schools, villages, firms, etc.)
2. You expect errors to be correlated within these groups
3. You have a sufficient number of clusters (generally at least 30-50)

Common examples include: - Student-level data clustered by school or classroom - Firm-level data clustered by industry - Individual-level data clustered by geographic region - Panel data clustered by individual or time period

Implementation in R

The `CASchools` dataset contains information on 420 California Schools from 45 different counties, which can be viewed as clusters.

The `fixest` package makes it easy to implement cluster-robust standard errors:

```
feols(score ~ STR + english, data = CASchools, cluster = "county") |> summary()
```

```
OLS estimation, Dep. Var.: score
Observations: 420
Standard-errors: Clustered (county)
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 686.032245  15.802838  43.41196 < 2.2e-16 ***
STR          -1.101296   0.754387  -1.45986   0.15143
english      -0.649777   0.030230 -21.49427 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 14.4   Adj. R2: 0.423681
```

After accounting for clustering, the coefficient on `STR` is no longer statistically significant.

You can also use the `modelsummary()` function to compare the same regression with different standard errors:

```
fit1 = feols(score ~ STR + english, data = CASchools)
## List of standard errors:
myvcov = list("IID", "HC1", "HC3", ~county)
modelsummary(fit1, stars = TRUE, statistic = "conf.int", vcov = myvcov)
```

	(1)	(2)	(3)	(4)
(Intercept)	686.032***	686.032***	686.032***	686.032***
	[671.464, 700.600]	[668.875, 703.189]	[668.710, 703.354]	[654.969, 717.095]
STR	−1.101**	−1.101*	−1.101*	−1.101
	[−1.849, −0.354]	[−1.952, −0.250]	[−1.960, −0.242]	[−2.584, 0.382]
english	−0.650***	−0.650***	−0.650***	−0.650***
	[−0.727, −0.572]	[−0.711, −0.589]	[−0.711, −0.588]	[−0.709, −0.590]
Num.Obs.	420	420	420	420
R2	0.426	0.426	0.426	0.426
R2 Adj.	0.424	0.424	0.424	0.424
AIC	3439.1	3439.1	3439.1	3439.1
BIC	3451.2	3451.2	3451.2	3451.2
RMSE	14.41	14.41	14.41	14.41
Std.Errors	IID	HC1	HC3	by: county

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Challenges with Cluster-robust Inference

The cluster-robust variance estimator relies on having a large number of clusters. With few clusters (generally $G < 30$), the estimator may be biased downward, leading to confidence intervals that are too narrow and overly frequent rejection of null hypotheses.

To account for high leverage points, the CR3 correction is similar to HC3 and applies a leverage adjustment at the cluster level:

$$\widehat{\boldsymbol{V}}_{CR3} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \sum_{g=1}^{G} \Big(\sum_{i=1}^{n_g} \boldsymbol{X}_{ig} \frac{\hat{u}_{ig}}{1 - h_{ig}} \Big) \Big(\sum_{i=1}^{n_g} \boldsymbol{X}_{ig} \frac{\hat{u}_{ig}}{1 - h_{ig}} \Big)' (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

8.9 R Code

statistics-sec08.R

9 Hypothesis testing

9.1 Statistical hypotheses

A statistical hypothesis is a statement about the population distribution. For instance, we might be interested in the hypothesis that a population regression coefficient β_j of a linear regression model is equal to some value β_j^0 or whether it is unequal to that value.

For instance, in a regression of test scores on the student-teacher ratio, we might be interested in testing whether adding one more student per class has no effect on test scores – that is, whether $\beta_j = \beta_j^0 = 0$.

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

$$\underbrace{H_0 : \beta_j = \beta_j^0}_{\text{null hypothesis}} \quad \text{vs.} \quad \underbrace{H_1 : \beta_j \neq \beta_j^0}_{\text{alternative hypothesis}}. \tag{9.1}$$

This idea is not limited to regression coefficients. For any parameter θ we can test the hypothesis $H_0 : \theta = \theta_0$ against its alternative $H_1 : \theta \neq \theta_0$.

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 distribution of T_0 under the assumption that H_0 holds(**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.

$$P(\underbrace{T_0 \in I_1}_{\text{reject}} \mid H_0 \text{ is true}) = \alpha, \tag{9.2}$$

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

A test that satisfies Equation 9.2 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 9.1: 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.

9.2 t-Tests

The **t-statistic** is the OLS estimator standardized with the standard error. Under (A1)–(A4) we have

$$T = \frac{\hat{\beta}_j - \beta_j}{se_{hc}(\hat{\beta}_j)} \xrightarrow{d} \mathcal{N}(0,1).$$

This result can be used to test the hypothesis $H_0 : \beta_j = \beta_j^0$. The t-statistic for this hypothesis is

$$T_0 = \frac{\hat{\beta}_j - \beta_j^0}{se_{hc}(\hat{\beta}_j)},$$

which satisfies $T_0 = T \xrightarrow{d} \mathcal{N}(0,1)$ under H_0.

Therefore, we can test H_0 by checking whether the presumed value β_j^0 falls into the confidence interval. We do not reject H_0 if

$$\beta_j^0 \in I_{1-\alpha}^{(hc)} = [\hat{\beta}_j - t_{(1-\frac{\alpha}{2},n-k)} se_{hc}(\hat{\beta}_j);\ \hat{\beta}_j + t_{(1-\frac{\alpha}{2},n-k)} se_{hc}(\hat{\beta}_j)].$$

By the definition of T_0, we have $\beta_j^0 \in I_{1-\alpha}^{(hc)}$ if and only if $|T_0| \leq t_{(1-\frac{\alpha}{2},n-k)}$.

Therefore, the **two-sided t-test** for H_0 against $H_1 : \beta_j \neq \beta_j^0$ is given by the test decision

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

The value $t_{(1-\frac{\alpha}{2},n-k)}$ is called the **critical value**.

This test is asymptotically of size α:

$$\lim_{n\to\infty} P(\text{we reject } H_0 | H_0 \text{ is true}) = \alpha.$$

This is because the confidence interval has asymptotically a $1-\alpha$ coverage rate:

$$\begin{aligned} &\lim_{n\to\infty} P(\text{we do not reject } H_0 | H_0 \text{ is true}) \\ &= \lim_{n\to\infty} P(\beta_j^0 \in I_{1-\alpha}^{(hc)} | H_0 \text{ is true}) \\ &= \lim_{n\to\infty} P(\beta_j \in I_{1-\alpha}^{(hc)}) \\ &= 1-\alpha. \end{aligned}$$

If (A5)–(A6) hold, and $se_{hom}(\hat{\beta}_j)$ is used instead of $se_{hc}(\hat{\beta}_j)$, then the t-test is of exact size α. However, as discussed in the previous section, (A5)–(A6) is an unlikely scenario in practice. Therefore $se_{hc}(\hat{\beta}_j)$ is the preferred choice.

```
library(AER)
cps = read.csv("cps.csv")
fit = lm(wage ~ education + female, data = cps)
coefci(fit, vcov = vcovHC, level = 0.99)
```

```
                0.5 %     99.5 %
(Intercept) -15.370102 -12.793475
education     2.854842   3.061506
female       -7.949469  -7.116664
```

The 99% confidence intervals indicate that:

- the null hypothesis $H_0 : \beta_2 = 0$ ("the marginal effect of education on the wage conditional on gender is 0") is rejected at the 1% significance level.
- the null hypothesis $H_0 : \beta_2 = 3$ ("the marginal effect of education on the wage conditional on gender is 3") is not rejected at the 1% significance level.

Let's compute T_0 for the hypothesis $\beta_2 = 3$ by hand:

```
## OLS coefficient
betahat2 = fit$coefficient[2]
## HC standard error
se = sqrt(vcovHC(fit)[2,2])
## presumed value for beta2
beta20 = 3
c(betahat2, beta20, se)
```

```
 education
2.95817398 3.00000000 0.04011445
```

```
## test statistic
T0 = (betahat2 - beta20)/se
T0
```

```
education
-1.042667
```

```
## critical values for 1=%, 5% and 1% levels
n = length(fit$fitted.values)
qt(c(0.95, 0.975, 0.995), df=n-3)
```

```
[1] 1.644884 1.960011 2.575926
```

Since $|T_0| = 1.04$ is smaller that the critical values for all common significance levels, we cannot reject $H_0 : \beta_2 = 3$.

9.3 The p-value

The **p-value** is a criterion to reach a hypothesis 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

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

where T^* is a random variable following the null distribution (in this case, $T^* \sim t_{n-k}$), and T_0 is the observed value of the test statistic.

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_{T_0} of the null distribution (in this case, t_{n-k}):

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

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 a population where the null hypothesis is true.

Let's compute the p-value for the hypothesis $\beta_2 = 3$ in the wage on education and female regression by hand. Here, F_{T_0} is the CDF of the t-distribution with $n - 3$ degrees of freedom. To compute $F_{T_0}(a)$, we can use `pt(a, df=n-3)`.

```
## p-value
2*(1-pt(abs(T0), df = n-3))
```

```
education
0.2971074
```

The p-value is larger than any common significance level. Hence, we do not reject H_0.

For the hypothesis $H_0 : \beta_2 = 0$, we get the following p-value:

```
T0 = (betahat2 - 0)/se
2*(1-pt(abs(T0), df = n-3))
```

```
education
        0
```

The p-value is (almost) 0. Hence, we reject H_0.

More conveniently, the `coeftest` function from the `AER` package provides a full summary of the regression results including the t-statistics and p-values for the hypotheses that $H_0 : \beta_j = 0$ for $j = 1, \dots, k$.

```
coeftest(fit, vcov = vcovHC)
```

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -14.081788   0.500136 -28.156 < 2.2e-16 ***
education     2.958174   0.040114  73.743 < 2.2e-16 ***
female       -7.533067   0.161652 -46.601 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can specify different standard errors: `coeftest(fit, vcov = vcovHC, type = "HC1")`. `coeftest(fit)` returns the t-test results for classical standard errors which is identical to the output of the base-R command `summary(fit)`, which should not be used in applications with heteroskedasticity.

To represent very small numbers where there are, e.g., `16` zero digits before the first nonzero digit after the decimal point, R uses scientific notation in the form `e-16`. For example, `2.2e-16` means `0.00000000000000022`.

9.4 Multiple testing problem

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 random variables, which implies that the test decisions of the two tests are independent.

$$\begin{aligned}
&P(\text{both tests do not reject} \mid H_0 \text{ true}) \\
&= P(\{\text{test1 does not reject}\} \cap \{\text{test2 does not reject}\} \mid H_0 \text{ true}) \\
&= P(\text{test1 does not reject} \mid H_0 \text{ true}) \cdot P(\text{test2 does not reject} \mid H_0 \text{ true}) \\
&= (1-\alpha)^2 = \alpha^2 - 2\alpha + 1
\end{aligned}$$

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.

Each t-test has a probability of falsely rejecting H_0 (type I error) of α, but if multiple t-tests are used on different coefficients, then the probability of falsely rejecting at least once (joint type I error probability) is greater than α (multiple testing problem).

Therefore, when multiple hypotheses are to be tested, repeated t-tests will not yield valid inferences, and another rejection rule must be found for repeated t-tests.

9.5 Joint Hypotheses

Consider the general hypothesis

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r},$$

where $\boldsymbol{R}$ is a $q \times k$ matrix with $\text{rank}(\boldsymbol{R}) = q$ and $\boldsymbol{r}$ is a $q \times 1$ vector.

Let's look at a linear regression with $k = 3$:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i$$

- Example 1: The hypothesis $H_0 : (\beta_2 = 0 \text{ and } \beta_3 = 0)$ implies $q = 2$ constraints and is translated to $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ with

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

- Example 2: The hypothesis $H_0 : \beta_2 + \beta_3 = 1$ implies $q = 1$ constraint and is translated to $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ with

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{r} = \begin{pmatrix} 1 \end{pmatrix}.$$

In practice, the most common multiple hypothesis tests are tests of whether multiple coefficients are equal to zero, which is a test of whether those regressors should be included in the model.

9.6 Wald Test

The Wald distance is the vector $\boldsymbol{d} = \boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}$, and the Wald statistic is the squared standardized Wald distance vector:

$$\begin{aligned} W &= \boldsymbol{d}'(\boldsymbol{R}\widehat{\boldsymbol{V}}\boldsymbol{R}')^{-1}\boldsymbol{d} \\ &= (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}\widehat{\boldsymbol{V}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \end{aligned}$$

Here, $\widehat{\boldsymbol{V}}$ is a suitable estimator for covariance matrix of the OLS coefficient vector, i.e. $\widehat{\boldsymbol{V}}_{hc}$ for robust testing under (A1)–(A4), and $\widehat{\boldsymbol{V}}_{hom}$ for testing under the special case of homoskedasticity.

Under H_0 we have

$$W \xrightarrow{d} \chi^2_q.$$

The test decision for the **Wald test**:

$$\begin{aligned} \text{do not reject } H_0 \quad &\text{if } W \leq \chi^2_{(1-\alpha,q)}, \\ \text{reject } H_0 \quad &\text{if } W > \chi^2_{(1-\alpha,q)}, \end{aligned}$$

where $\chi^2_{(p,q)}$ is the p-quantile of the chi-squared distribution with q degrees of freedom. $\chi^2_{(p,q)}$ can be returned using `qchisq(p,q)`.

To test $H_0 : \beta_2 = \beta_3 = 0$ in the regression of wage on education and female (example 1), we can use the `linearHypothesis()` function from the `AER` package:

```r
## Define r and R
r = c(0,0)
R = rbind(
  c(0,1,0),
  c(0,0,1)
)
R
```

```
     [,1] [,2] [,3]
[1,]    0    1    0
[2,]    0    0    1
```

```r
linearHypothesis(fit,
                 hypothesis.matrix = R,
                 rhs = r,
                 vcov = vcovHC,
                 test = "Chisq")
```

```
Linear hypothesis test:
education = 0
female = 0

Model 1: restricted model
Model 2: wage ~ education + female

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1  50741
2  50739  2 5977.4  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is rejected because the p-value is very small. To confirm this, we see in the output that the Wald statistic is $W = 5977$. The critical value for the common significance levels are:

```r
qchisq(c(0.9, 0.95, 0.99), df=2)
```

```
[1] 4.605170 5.991465 9.210340
```

To compute the Wald statistic W by hand, we need matrix algebra:

```
betahat = fit$coefficients
## Wald distance:
d = R %*% betahat - r
## Wald statistic
W = t(d) %*% solve(R %*% vcovHC(fit) %*% t(R)) %*% d
W
```

```
         [,1]
[1,] 5977.396
```

Instead of definition the matrix `R` and vector `r`, we can also specify our restrictions in `linearHypothesis()` directly:

```
linearHypothesis(fit,
                 c("education = 0", "female = 0"),
                 vcov = vcovHC,
                 test = "Chisq")
```

```
Linear hypothesis test:
education = 0
female = 0

Model 1: restricted model
Model 2: wage ~ education + female

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1  50741
2  50739  2 5977.4  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If `vcov = vcovHC` is omitted, then the homoskedasticity-only covariance matrix $\widehat{\boldsymbol{V}}_{hom}$ is used.
If `test = "Chisq` is omitted, then the F-test is applied, which is introduced below.

9.7 F-Test

The Wald test is an asymptotic size-α-test under (A1)–(A4). Even if (A5) and (A6) hold true as well, the Wald test is still only asymptotically valid, i.e.:

$$\lim_{n\to\infty} P(\text{Wald test rejects } H_0 | H_0 \text{ true}) = \alpha.$$

Similarly to the classical t-test, we can construct a test joint test that is of exact size α under (A1)–(A6).

The F statistic is the Wald statistic scaled by the number of constraints:

$$F = \frac{W}{q} = \frac{1}{q}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}\widehat{\boldsymbol{V}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}).$$

If (A1)–(A6) hold true, and if $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_{hom}$ is used, it can be shown that

$$F \sim F_{q;n-k}$$

for any finite sample size n, where $F_{q;n-k}$ is the F-distribution with q degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator.

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

F-test decision rule

The test decision for the **F-test**:

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } F \leq F_{(1-\alpha,q,n-k)}, \\ \text{reject } H_0 \quad & \text{if } F > F_{(1-\alpha,q,n-k)}, \end{aligned}$$

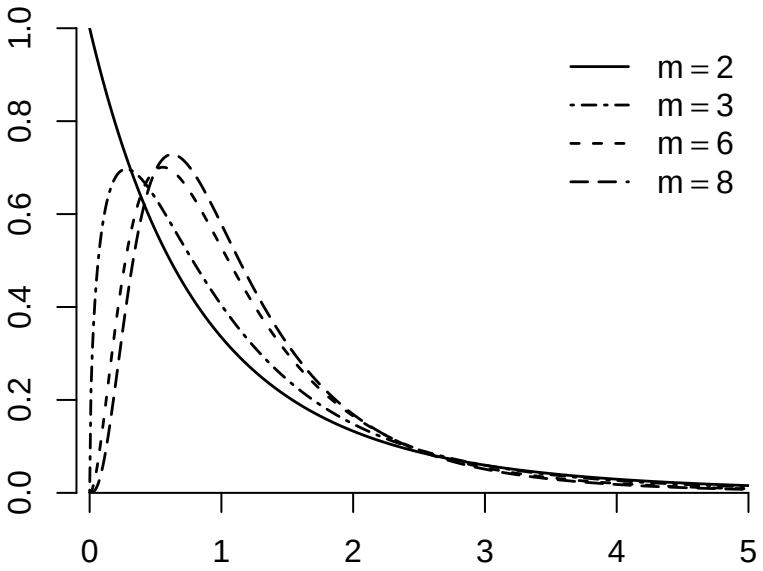


Figure 9.1: F-distribution

where $F_{(p,m_1,m_2)}$ is the p-quantile of the F distribution with m_1 degrees of freedom in the numerator and m_2 degrees of freedom in the denominator. $F_{(p,m_1,m_2)}$ can be returned using `qf(p,m1,m2)`.

For single constraint ($q = 1$) hypotheses of the form $H_0 : \beta_j = \beta_j^0$, the F-test is equivalent to a two-sided t-test.

- If (A1)–(A6) hold true and $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_{hom}$ is used, the F-test has exact size α, similar to the exact t-test for this case.
- If (A1)–(A5) hold true and $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_{hom}$ is used, the F-test and the Wald-test have asymptotic size α.
- If (A1)–(A4) hold true and $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_{hc}$ is used, the F-test and the Wald-test have asymptotic size α.

The F-test tends to be more conservative than the Wald test in small samples, meaning that rejection by the F-test generally implies rejection by the Wald test, but not necessarily vice versa. Due to this more conservative nature, which helps control false rejections (Type I errors) in small samples, the F-test is often preferred in practice.

```
linearHypothesis(fit,
                 c("education = 0", "female = 0"),
                 vcov = vcovHC,
                 test = "F")
```

```
Linear hypothesis test:
education = 0
female = 0

Model 1: restricted model
Model 2: wage ~ education + female

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1  50741
2  50739  2 2988.7 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we have $F = W/2$. The critical values for the common significance level can be obtained as follows:

```
n = length(fit$fitted.values)
k = 3
q = 2
qf(c(0.9, 0.95, 0.99), q, n-k)
```

```
[1] 2.302690 2.995909 4.605588
```

Since $F = 2988.7$, the null hypothesis is rejected at all common significance levels.

9.8 Diagnostics tests

The asymptotic properties of the OLS estimator and inferential methods using HC-type standard errors do not depend on the validity of the homoskedasticity and normality assumptions (A5)–(A6).

However, if you are interested in exact inference, verifying the assumptions (A5)–(A6) becomes crucial, especially in small samples.

9.8.1 Breusch-Pagan Test (Koenker's version)

Under homoskedasticity, the variance of the error term does not depend on the values of the regressors.

To test for heteroskedasticity, we regress the squared residuals on the regressors.

$$\hat{u}_i^2 = \boldsymbol{X}_i'\boldsymbol{\gamma} + v_i, \quad i = 1, \dots, n. \tag{9.3}$$

Here, $\boldsymbol{\gamma}$ are the auxiliary coefficients and v_i are the auxiliary error terms. Under homoskedasticity, the regressors should not be able to explain any variation in the residuals.

Let R^2_{aux} be the R-squared coefficient of the auxiliary regression of Equation 9.3. The test statistic:

$$BP = nR^2_{aux}$$

Under the null hypothesis of homoskedasticity, we have

$$BP \xrightarrow{d} \chi^2_{k-1}$$

Test decision rule: Reject H_0 if BP exceeds $\chi^2_{(1-\alpha,k-1)}$.

In R we can apply the `bptest()` function from the `AER` package to the `lm` object of our regression.

```
bptest(fit)
```

```
    studentized Breusch-Pagan test

data:  fit
BP = 1070.3, df = 2, p-value < 2.2e-16
```

The BP test clearly rejects H_0, which is strong statistical evidence that the errors are heteroskedastic.

9.8.2 Jarque-Bera Test

A general property of any normally distributed random variable is that it has a skewness of 0 and a kurtosis of 3.

Under (A5)–(A6), we have $u_i \sim \mathcal{N}(0, \sigma^2)$, which implies $E[u_i^3] = 0$ and $E[u_i^4] = 3\sigma^4$.

Consider the sample skewness and the sample kurtosis of the residuals from your regression:

$$\widehat{skew}_{\widehat{u}} = \frac{1}{n\hat{\sigma}^3_{\widehat{u}}} \sum_{i=1}^{n} \hat{u}_i^3, \quad \widehat{kurt}_{\widehat{u}} = \frac{1}{n\hat{\sigma}^4_{\widehat{u}}} \sum_{i=1}^{n} \hat{u}_i^4$$

Jarque-Bera test statistic and null distribution if (A5)–(A6) hold:

$$JB = n\left(\frac{1}{6}(\widehat{skew}_{\widehat{u}})^2 + \frac{1}{24}(\widehat{kurt}_{\widehat{u}} - 3)^2\right) \xrightarrow{d} \chi^2_2.$$

Test decision rule: Reject the null hypothesis of normality if JB exceeds $\chi^2_{(1-\alpha,2)}$.

Note that the Jarque-Bera test is sensitive to outliers.

In R we apply use the `jarque.test()` function from the `moments` package to the residual vector from our regression.

```
library(moments)
jarque.test(fit$residuals)
```

```
    Jarque-Bera Normality Test

data:  fit$residuals
JB = 2230900, p-value < 2.2e-16
alternative hypothesis: greater
```

The JB test clearly rejects H_0, which is strong statistical evidence that the errors are not normally distributed.

The results of the BP and the JB test indicate that classical standard errors $se(\beta_j)$ and the classical covariance matrix estimators $\widehat{\boldsymbol{V}}_{hom}$ should not be used. Instead, HC-versions should be applied.

9.9 Nonliearities in test score regressions

Let's use the hypothesis tests from this section to conduct a study on the relationship between test scores and the student-teacher ratio.

```
data(CASchools, package = "AER")
## append student-teacher ratio
CASchools$STR = CASchools$students/CASchools$teachers
## append average test score
CASchools$score = (CASchools$read+CASchools$math)/2
## append high English learner share dummy variable
CASchools$HiEL = (CASchools$english >= 10) |> as.numeric()
```

This section examines three key questions about test scores and the student-teacher ratio.

- First, it explores if reducing the student-teacher ratio affects test scores differently based on the number of English learners, even when considering economic differences across districts.
- Second, it investigates if this effect varies depending on the student-teacher ratio.
- Lastly, it aims to determine the expected impact on test scores when the student-teacher ratio decreases by two students per teacher, considering both economic factors and potential nonlinear relationships.

The logarithm of district `income` is used following our previous empirical analysis, which suggested that this specification captures the nonlinear relationship between scores and income.

We leave out the expenditure per pupil (`expenditure`) from our analysis because including it would suggest that spending changes with the student-teacher ratio (in other words, we would not be holding expenditures per pupil constant: bad control).

We will consider 7 different model specifications:

```
# estimate all models
mod1 = lm(score ~ STR + english + lunch, data = CASchools)
mod2 = lm(score ~ STR + english + lunch + log(income), data = CASchools)
mod3 = lm(score ~ STR + HiEL + HiEL:STR, data = CASchools)
mod4 = lm(score ~ STR + HiEL + HiEL:STR + lunch + log(income), data = CASchools)
mod5 = lm(score ~ STR + I(STR^2) + I(STR^3) + HiEL + lunch + log(income),
              data = CASchools)
mod6 = lm(score ~ STR + I(STR^2) + I(STR^3) + HiEL + HiEL:STR + HiEL:I(STR^2)
          + HiEL:I(STR^3) + lunch + log(income), data = CASchools)
mod7 = lm(score ~ STR + I(STR^2) + I(STR^3) + english + lunch + log(income),
              data = CASchools)

# gather robust standard errors in a list
rob_se = list(sqrt(diag(vcovHC(mod1))),
              sqrt(diag(vcovHC(mod2))),
```

```
               sqrt(diag(vcovHC(mod3))),
               sqrt(diag(vcovHC(mod4))),
               sqrt(diag(vcovHC(mod5))),
               sqrt(diag(vcovHC(mod6))),
               sqrt(diag(vcovHC(mod7))))
```

```
library(stargazer)
stargazer(mod1, mod2, mod3, mod4,
          mod5, mod6, mod7,
          font.size = "footnotesize",
          se = rob_se,
          type="latex",
          omit.stat = "f", df=FALSE, header = FALSE)
```

The stars in the regression output indicate the statistical significance of each coefficient based on a t-test of the hypothesis $H_0 : \beta_j = 0$. No stars indicate that the coefficient is not statistically significant (cannot reject H_0 at conventional significance levels). One star (*) denotes significance at the 10% level (pval < 0.10), two stars (**) indicate significance at the 5% level (pval < 0.05), and three stars (***) indicate significance at the 1% level (pval < 0.01).

What can be concluded from the results presented?

i) First, we find that there is evidence of heteroskedasticity and non-normality, because the Breusch-Pagan test and the Jarque-Bera test reject. Therefore, HC-robust tests should be used.

```
bptest(mod1)
```

```
    studentized Breusch-Pagan test

data:  mod1
BP = 9.9375, df = 3, p-value = 0.0191
```

```
jarque.test(mod1$residuals)
```

```
    Jarque-Bera Normality Test

data:  mod1$residuals
```

Table 9.2

	Dependent variable:						
	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
STR	−0.998***	−0.734***	−0.968	−0.531	64.339**	83.702***	65.285**
	(0.274)	(0.261)	(0.599)	(0.350)	(27.295)	(31.506)	(27.708)
english	−0.122***	−0.176***					−0.166***
	(0.033)	(0.034)					(0.035)
I(STR^2)					−3.424**	−4.381***	−3.466**
					(1.373)	(1.597)	(1.395)
I(STR^3)					0.059***	0.075***	0.060***
					(0.023)	(0.027)	(0.023)
lunch	−0.547***	−0.398***		−0.411***	−0.420***	−0.418***	−0.402***
	(0.024)	(0.034)		(0.029)	(0.029)	(0.029)	(0.034)
log(income)		11.569***		12.124***	11.748***	11.800***	11.509***
		(1.841)		(1.823)	(1.799)	(1.809)	(1.834)
HiEL			5.639	5.498	−5.474***	816.076**	
			(19.889)	(10.012)	(1.046)	(354.100)	
STR:HiEL			−1.277	−0.578		−123.282**	
			(0.986)	(0.507)		(54.290)	
I(STR^2):HiEL						6.121**	
						(2.752)	
I(STR^3):HiEL						−0.101**	
						(0.046)	
Constant	700.150***	658.552***	682.246***	653.666***	252.050	122.353	244.809
	(5.641)	(8.749)	(12.071)	(10.053)	(179.724)	(205.050)	(181.899)
Observations	420	420	420	420	420	420	420
R^2	0.775	0.796	0.310	0.797	0.801	0.803	0.801
Adjusted R^2	0.773	0.794	0.305	0.795	0.798	0.799	0.798
Residual Std. Error	9.080	8.643	15.880	8.629	8.559	8.547	8.568
Note:							*p<0.1; **p<0.05; ***p<0.01

```
JB = 10.626, p-value = 0.004926
alternative hypothesis: greater
```

ii) We see the estimated coefficient of `STR` is highly significant in all models except from specifications (3) and (4).

iii) When we add `log(income)` to model (1) in the second specification, all coefficients remain highly significant while the coefficient on the new regressor is also statistically significant at the 1% level. In addition, the coefficient on `STR` is now 0.27 higher than in model (1), which suggests a possible reduction in omitted variable bias when including `log(income)` as a regressor. For these reasons, it makes sense to keep this variable in other models too.

iv) Models (3) and (4) include the interaction term between `STR` and `HiEL`, first without control variables in the third specification and then controlling for economic factors in the fourth. The estimated coefficient for the interaction term is not significant at any common level in any of these models, nor is the coefficient on the dummy variable `HiEL`. However, this result is misleading and we should not conclude that none of the variables has a non-zero marginal effect because the coefficients cannot be interpreted separately from each other. What we can learn from the fact that the coefficient of `STR:HiEL` alone is not significantly different from zero is that the impact of the student-teacher ratio on test scores remains consistent across districts with high and low proportions of English learning students. Let's test the hypotheses that all coefficients that involve `STR` are zero and all coefficients that involve `HiEL` are zero. We find that H_0 is rejected for both hypotheses and the overall marginal effects are clearly significant:

```
linearHypothesis(mod3, c("STR = 0", "STR:HiEL = 0"), vcov=vcovHC)
```

```
Linear hypothesis test:
STR = 0
STR:HiEL = 0

Model 1: restricted model
Model 2: score ~ STR + HiEL + HiEL:STR

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F   Pr(>F)
1    418
2    416  2 5.4228 0.004732 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mod3, c("HiEL = 0", "STR:HiEL = 0"), vcov=vcovHC)
```

```
Linear hypothesis test:
HiEL = 0
STR:HiEL = 0

Model 1: restricted model
Model 2: score ~ STR + HiEL + HiEL:STR

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    418
2    416  2 88.806 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

v) In regression (5) we have included quadratic and cubic terms for `STR`, while omitting the interaction term between `STR` and `HiEL`, since it was not significant in specification (4). The results indicate high levels of significance for these estimated coefficients and we can therefore assume the presence of a nonlinear effect of the student-teacher ration on test scores. This can be verified with an F-test of $H_0 : \beta_3 = \beta_4 = 0$:

```
linearHypothesis(mod5, c("I(STR^2) = 0", "I(STR^3) = 0"), vcov=vcovHC)
```

```
Linear hypothesis test:
I(STR^2) = 0
I(STR^3) = 0

Model 1: restricted model
Model 2: score ~ STR + I(STR^2) + I(STR^3) + HiEL + lunch + log(income)

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F  Pr(>F)
1    415
2    413  2 5.0205 0.00701 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

vi) Regression (6) further examines whether the proportion of English learners influences the student-teacher ratio, incorporating the interaction terms $HiEL{\cdot}STR$, $HiEL{\cdot}STR^2$ and $HiEL \cdot STR^3$. Each individual t-test confirms significant effects. To validate this, we perform a robust F-test to assess $H_0 : \beta_8 = \beta_9 = \beta_1 0 = 0$.

```
linearHypothesis(mod6, c("STR:HiEL = 0", "I(STR^2):HiEL = 0", "I(STR^3):HiEL = 0"), vcov=vco
```

```
Linear hypothesis test:
STR:HiEL = 0
I(STR^2):HiEL = 0
I(STR^3):HiEL = 0

Model 1: restricted model
Model 2: score ~ STR + I(STR^2) + I(STR^3) + HiEL + HiEL:STR + HiEL:I(STR^2) +
    HiEL:I(STR^3) + lunch + log(income)

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F  Pr(>F)
1    413
2    410  3 2.1885 0.08882 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

vii) With a p-value of `0.08882` we can just reject the null hypothesis at the 10% level. This provides only weak evidence that the regression functions are different for districts with high and low percentages of English learners.

viii) In model (7), we employ a continuous measure for the proportion of English learners instead of a dummy variable (thus omitting interaction terms). We note minimal alterations in the coefficient estimates for the remaining regressors. Consequently, we infer that the findings observed in model (5) are robust and not influenced significantly by the method used to measure the percentage of English learners.

We can now address the initial questions raised in this section:

- First, in the linear models, the impact of the percentage of English learners on changes in test scores due to variations in the student-teacher ratio is minimal, a conclusion that holds true even after accounting for students' economic backgrounds. Although the cubic specification (6) suggests that the relationship between student-teacher ratio and test scores is influenced by the proportion of English learners, the magnitude of this influence is not significant.

- Second, while controlling for students' economic backgrounds, we identify nonlinearities in the association between student-teacher ratio and test scores.
- Lastly, under the **linear specification** (2), a reduction of two students per teacher in the student-teacher ratio is projected to increase test scores by approximately 1.46 points. As this model is linear, this effect remains consistent regardless of class size. For instance, assuming a student-teacher ratio of 20, the **nonlinear model** (5) indicates that the reduction in student-teacher ratio would lead to an increase in test scores by

$$\begin{aligned} & 64.33 \cdot 18 + 18^2 \cdot (-3.42) + 18^3 \cdot (0.059) \\ & \quad - (64.33 \cdot 20 + 20^2 \cdot (-3.42) + 20^3 \cdot (0.059)) \\ & \approx 3.3 \end{aligned}$$

  points. If the ratio was 22, a reduction to 20 leads to a predicted improvement in test scores of

$$\begin{aligned} & 64.33 \cdot 20 + 20^2 \cdot (-3.42) + 20^3 \cdot (0.059) \\ & \quad - (64.33 \cdot 22 + 22^2 \cdot (-3.42) + 22^3 \cdot (0.059)) \\ & \approx 2.4 \end{aligned}$$

  points. This suggests that the effect is more evident in smaller classes.

9.10 R-codes

statistics-sec11.R

10 Estimation Theory

10.1 Bias, Variance, and MSE

Let θ denote a population parameter and $\hat{\theta}_n$ an estimator based on a sample of size n.

Definitions

- **Bias:**
$$\text{Bias}(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta.$$
- **Sampling Variance:**
$$\text{Var}(\hat{\theta}_n) = E[(\hat{\theta}_n - E[\hat{\theta}_n])^2]$$
- **Mean Squared Error:**
$$\text{MSE}(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2] = \text{Var}(\hat{\theta}_n) + \text{Bias}(\hat{\theta}_n)^2.$$

An estimator $\hat{\theta}_n$ is **unbiased** if $\text{Bias}(\hat{\theta}_n) = 0$ for any fixed n.

$\hat{\theta}_n$ is **asymptotically unbiased** if $\lim_{n\to\infty} \text{Bias}(\hat{\theta}_n) = 0$.

Bias–variance trade-off: Unbiasedness is only one dimension of estimator quality. A slightly biased estimator can have much smaller variance and thus lower MSE.

Sample mean

Let $Y_1, \ldots, Y_n$ be an i.i.d. sample with $\mu = E[Y_i]$ and $\sigma^2 = \text{Var}(Y_i) < \infty$.

The sample mean is unbiased:
$$E[\overline{Y}] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \mu.$$

Thus, the MSE equals the variance:
$$\text{MSE}(\overline{Y}) = \text{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(Y_i) = \frac{\sigma^2}{n}.$$

Sample variance

Let $Y_1, \dots, Y_n$ be an i.i.d. sample with $\mu = E[Y_i]$, $\sigma^2 = \text{Var}(Y_i) < \infty$, and $\kappa = \text{kurt}(Y_i) < \infty$. By decomposing $Y_i - \overline{Y} = (Y_i - \mu) - (\overline{Y} - \mu)$, we can rearrange the sample variance as follows:

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

Thus, the mean of the sample variance is

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

The sample variance is a downward biased estimator for the population variance σ^2:

$$\text{Bias}(\hat{\sigma}_Y^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

However, the bias tends to zero as $n \to \infty$, so the sample variance is asymptotically unbiased.

Recall the adjusted sample variance:

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

It is unbiased:

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

The degree of freedom correction can be interpreted as a bias correction.

The sampling variances of the unadjusted and adjusted sample variance are

$$\begin{aligned} \text{Var}(\hat{\sigma}_Y^2) &= \frac{\sigma^4}{n}\Big(\kappa - \frac{n-3}{n-1}\Big)\frac{(n-1)^2}{n^2}, \\ \text{Var}(s_Y^2) &= \frac{\sigma^4}{n}\Big(\kappa - \frac{n-3}{n-1}\Big). \end{aligned}$$

Thus, $\text{Var}(s_Y^2) > \text{Var}(\hat{\sigma}_Y^2)$. s_Y^2 is unbiased for σ^2 but it estimates σ^2 less precise than $\hat{\sigma}_Y^2$.

The MSE of s_Y^2 equals $Var(s_Y^2)$ while the MSE of $\hat{\sigma}_Y^2$ is

$$\begin{aligned}\text{MSE}(\hat{\sigma}_Y^2) &= \text{Var}(\hat{\sigma}_Y^2) + \text{Bias}(\hat{\sigma}_Y^2)^2 \\ &= \frac{\sigma^4}{n}\left[\left(\kappa - \frac{n-3}{n-1}\right)\frac{(n-1)^2}{n^2} + \frac{1}{n}\right].\end{aligned}$$

It is not possible to universally determine which estimator has a lower MSE because this depends on the population kurtosis κ of the underlying distribution.

However, it can be shown that for all distributions with $\kappa \geq 1.5$, the relation $\text{MSE}(s_Y^2) > \text{MSE}(\hat{\sigma}_Y^2)$ holds, which implies that $\hat{\sigma}_Y^2$ is preferred based on the bias-variance tradeoff for most distributions (recall that the normal distribution has $\kappa = 3$).

OLS Coefficient

Bias

Recall the model equation in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

Plugging this into the OLS formula:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}.\end{aligned}$$

Taking the conditional expectation:

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

Under exogeneity and i.i.d. sampling,

$$E[u_i|\boldsymbol{X}] = E[u_i|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n] = E[u_i|\boldsymbol{X}_i] = 0,$$

hence, $E[\boldsymbol{u}|\boldsymbol{X}] = \boldsymbol{0}$. Thus, the conditional mean is

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta},$$

and the unconditional mean becomes

$$E[\hat{\boldsymbol{\beta}}] = E[E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]] = \boldsymbol{\beta}.$$

Thus, **each element** of the OLS estimator is **unbiased**:

$$E[\hat{\beta}_j] = \beta_j \quad \text{for } j = 1, \dots, k.$$

Variance

Recall the general rule that for any matrix $\boldsymbol{A}$,

$$Var(\boldsymbol{Au}) = \boldsymbol{A}\, Var(\boldsymbol{u})\, \boldsymbol{A}'.$$

Hence, with $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, by the symmetry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$,

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \text{Var}(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}) \\
&= \text{Var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' Var(\boldsymbol{u}|\boldsymbol{X})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' Var(\boldsymbol{u}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Under i.i.d. sampling, the conditional covariance matrix of $\boldsymbol{u}$ takes a diagonal form:

$$Var(\boldsymbol{u}|\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix},$$

where $\sigma_i^2 = E[u_i^2|\boldsymbol{X}_i] = \sigma^2(\boldsymbol{X}_i)$.

While u_i is uncorrelated with $\boldsymbol{X}_i$ under the exogeneity assumption, its variance may depend on $\boldsymbol{X}_i$. In this case, we say that the errors are **heteroskedastic**.

Inserting this diagonal structure into the OLS covariance matrix gives

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^{n}\sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\sum_{i=1}^{n}\sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}
\end{aligned}$$

In the specific situation where the conditional variance of the error does not depend on $\boldsymbol{X}_i$ and is equal to σ^2 for any value of $\boldsymbol{X}_i$, we say that the errors are **homoskedastic**.

The homoskedastic error covariance matrix has the following simple form:

$$Var(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}.$$

The resulting OLS covariance matrix is

$$\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}$$

10.2 Convergence

Consistency

Recall that an estimator $\hat{\theta}_n$ is consistent for a parameter θ if for any $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \quad \text{as } n \to \infty.$$

Using Markov's inequality, we can bound this term from above by the MSE of the estimator:

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

So if the MSE of an estimator converges to zero as the sample size n approaches infinity, then $P(|\hat{\theta}_n - \theta| > \epsilon)$ also converges to zero, and the estimator is consistent.

Sufficient condition for consistency

If $\lim_{n\to\infty} \text{MSE}(\hat{\theta}_n) = 0$, then $\hat{\theta}_n$ is consistent for θ.

Rate of Convergence

The consistency property of an estimator indicates that the estimation uncertainty vanishes as the sample size n approaches infinity, but it does not quantify how accurate the estimate is for a given finite sample size n.

Recall that the MSE for the sample mean is

$$\text{MSE}(\overline{Y}) = \frac{\sigma^2}{n}.$$

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 $\hat{\theta}_n$ for θ is

$$\text{RMSE}(\hat{\theta}_n) = \sqrt{\text{MSE}(\hat{\theta}_n)} = \sqrt{E[(\hat{\theta}_n - \theta)^2]}.$$

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

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

Since the RMSE is proportional to $1/\sqrt{n}$, we say that the sample mean has the **rate of convergence** $\sqrt{n}$:

$$\lim_{n\to\infty} \sqrt{n} \cdot \text{RMSE}(\overline{Y}) = \sigma.$$

Rate of convergence

A consistent estimator $\hat{\theta}_n$ has convergence rate $\sqrt{n}$ if

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

More generally, the rate of convergence is r_n if

$$0 < \lim_{n\to\infty} \left(r_n \cdot \text{RMSE}(\hat{\theta}_n) \right) < \infty.$$

The rate $\sqrt{n}$ holds for many common estimators. In this case, we say that the estimator has a **parametric convergence rate**. There are important exceptions where estimators have slower or faster convergence rates (nonparametric estimators, certain machine learning methods, 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 parametric convergence rate $\sqrt{n}$ like in the sample mean case. To halve the RMSE, we need to increase the sample size by a factor of 4 since $\sqrt{4} = 2$. To reduce the RMSE by a factor of 4, we already need to increase the sample size by a factor of 16.

Convergence Rate of OLS

Under i.i.d sampling and the exogeneity condition, OLS is unbiased, so the conditional MSE equals the conditional variance:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\sum_{i=1}^{n} \sigma_i^2 \boldsymbol{X}_i \boldsymbol{X}_i' \right) \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \\
&= \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{X}_i \boldsymbol{X}_i' \right) \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1},
\end{aligned}$$

where $\sigma_i^2 = E[u_i^2|\boldsymbol{X}_i]$. Write

$$\boldsymbol{Q} := E[\boldsymbol{X}_i \boldsymbol{X}_i'], \quad \boldsymbol{\Omega} = E[u_i^2 \boldsymbol{X}_i \boldsymbol{X}_i'].$$

By the law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \xrightarrow{p} \boldsymbol{Q}, \quad \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{X}_i \boldsymbol{X}_i' \xrightarrow{p} \boldsymbol{\Omega},$$

and hence

$$n \, \text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) \xrightarrow{p} \boldsymbol{Q}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}^{-1}.$$

Taking expectations gives the unconditional statement

$$n\,\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) = n\,\mathrm{Var}(\widehat{\boldsymbol{\beta}}) \overset{n\to\infty}{\longrightarrow} \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}.$$

For the j-th coefficient,

$$\lim_{n\to\infty} \sqrt{n} \cdot \mathrm{RMSE}(\widehat{\beta}_j) = \sqrt{[(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1})]_{jj}},$$

where $[\cdot]_{jj}$ indicates the (j,j)-th diagonal element of a matrix.

Thus, each OLS coefficient has a parametric (i.e. $\sqrt{n}$) rate of convergence, and the asymptotic variance of the OLS coefficient vector is $\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$.

10.3 Gaussian distribution

Univariate Normal distribution

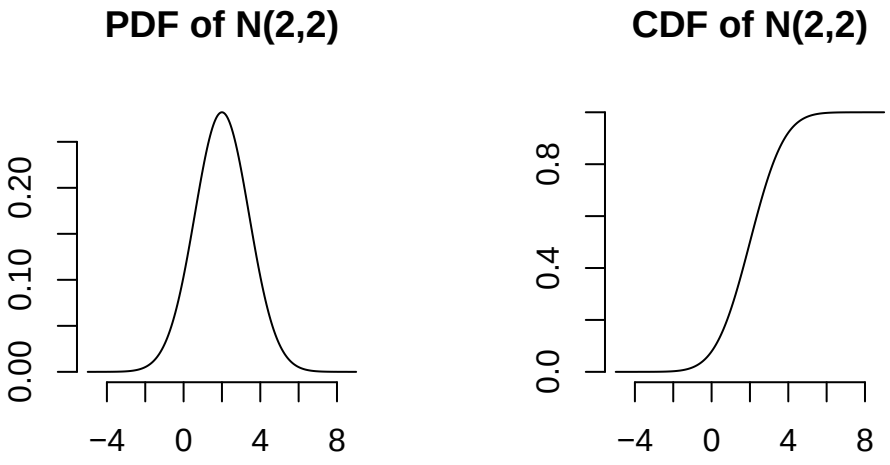
The Gaussian distribution, also known as the **normal distribution**, is a fundamental concept in statistics.

A random variable Z is said to follow a normal distribution if it has the following probability density function (PDF):

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

Formally, we denote this as $Z \sim \mathcal{N}(\mu, \sigma^2)$, meaning that Z is normally distributed with mean μ and variance σ^2.

- Mean: $E[Z] = \mu$
- Variance: $\mathrm{Var}(Z) = \sigma^2$
- Skewness: $\mathrm{skew}(Z) = 0$
- Kurtosis: $\mathrm{kurt}(Z) = 3$



The normal distribution with mean 0 and variance 1 is called the **standard normal distribution**. It has the PDF

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

and its CDF is

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

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

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

Standardizing: If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{Z-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

The CDF of Z is $P(Z \le a) = \Phi((a-\mu)/\sigma)$.

Linear combinations of normally distributed variables are normal: If $Y_1, \dots, Y_n$ are jointly normally distributed and $c_1, \dots, c_n \in \mathbb{R}$, then $\sum_{j=1}^{n} c_j Y_j$ is normally distributed.

Multivariate Normal distribution

Let $Z_1, \dots, Z_k$ be independent $\mathcal{N}(0,1)$ random variables.

Then, the k-vector $\boldsymbol{Z} = (Z_1, \dots, Z_k)'$ has the **multivariate standard normal distribution**, written $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$. Its joint PDF is

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

If $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\boldsymbol{Z}^* = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}$ for a $q \times 1$ vector $\boldsymbol{\mu}$ and a $q \times k$ matrix $\boldsymbol{B}$, then $\boldsymbol{Z}^*$ has a **multivariate normal distribution** with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$, written $\boldsymbol{Z}^* \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

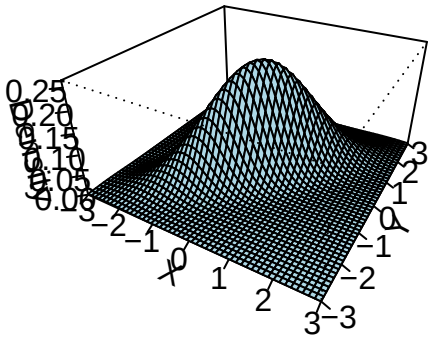
The q-variate PDF of $\boldsymbol{Z}^*$ is

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

The mean vector and covariance matrix are

$$E[\boldsymbol{Z}^*] = \boldsymbol{\mu}, \quad \mathrm{Var}(\boldsymbol{Z}^*) = \boldsymbol{\Sigma}.$$

3D Bivariate Normal Distribution Density



The 3D plot shows the bivariate normal PDF with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

10.4 Central Limit Theorem

Convergence in distribution

Let $\boldsymbol{V}_n$ be a sequence of k-variate random variables and let $\boldsymbol{V}$ be a k-variate random variable

$\boldsymbol{V}_n$ **converges in distribution** to $\boldsymbol{V}$, written $\boldsymbol{V}_n \xrightarrow{d} \boldsymbol{V}$, if

$$\lim_{n\to\infty} P(\boldsymbol{V}_n \leq \boldsymbol{a}) = P(\boldsymbol{V} \leq \boldsymbol{a})$$

for all $\boldsymbol{a}$ at which the CDF of $\boldsymbol{V}$ is continuous, where "$\leq$" is componentwise.

If $\boldsymbol{V}$ has the distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we write $\boldsymbol{V}_n \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

By the univariate central limit theorem, the sample mean converges to a normal distribution:

Central Limit Theorem (CLT)

Let $W_1, \ldots, W_n$ be an i.i.d. sample with $E[W_i] = \mu$ and $\text{Var}(W_i) = \sigma^2 < \infty$. Then, the sample mean $\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$ satisfies

$$\sqrt{n}(\overline{W} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

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

SHINY APP: CLT

The same result can be extended to random vectors.

Multivariate Central Limit Theorem (MCLT)

If $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ is a multivariate i.i.d. sample with $E[\boldsymbol{W}_i] = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{W}_i) = \boldsymbol{\Sigma} < \infty$. Then, the sample mean vector $\overline{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_i$ satisfies

$$\sqrt{n}(\overline{\boldsymbol{W}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

(see, e.g., Stock and Watson Section 19.2).

10.5 Asymptotic Normality

Let's apply the MCLT to the OLS vector. Consider $\boldsymbol{W}_i = \boldsymbol{X}_i u_i$, which satisfies

$$E[\boldsymbol{X}_i u_i] = \boldsymbol{0}, \quad \text{Var}(\boldsymbol{X}_i u_i) = E[u_i^2 \boldsymbol{X}_i \boldsymbol{X}_i'] = \boldsymbol{\Omega}.$$

Therefore, by the MCLT,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i u_i\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}).$$

Thus, because $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i' \xrightarrow{p} \boldsymbol{Q}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i u_i\right) \xrightarrow{d} \boldsymbol{Q}^{-1}\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}).$$

where the right-hand side means the distribution of $\boldsymbol{Q}^{-1}\boldsymbol{Z}$ for $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega})$.

Finally, since the variance of $\boldsymbol{Q}^{-1}\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega})$ is $\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$, we have the following central limit theorem for the OLS estimator:

Central Limit Theorem for OLS

Consider the linear model $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i$ such that

1) **Random sampling**: $(Y_i, \boldsymbol{X}_i')$ are i.i.d.
2) **Exogeneity (mean independence)**: $E[u_i|\boldsymbol{X}_i] = 0$.
3) **Finite fourth moments**: $E[X_{ij}^4] < \infty$ and $E[u_i^4] < \infty$.
4) **Full rank**: $E[\boldsymbol{X}_i \boldsymbol{X}_i']$ is positive definite (hence invertible).

Then, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}).$$

The only additional assumption compared to the consistency of OLS is the finite fourth moments condition instead of the finite second moments condition. This technical assumption ensures that the variance of $\boldsymbol{X}_i u_i$ is finite.

Specifically, the Cauchy-Schwarz inequality implies that

$$E[X_{ij}^2 u_i^2] \leq \sqrt{E[X_{ij}^4]E[u_i^4]} < \infty,$$

so that the elements of $\boldsymbol{\Omega}$ are finite.

If homoskedasticity holds, then $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{Q}$, and the asymptotic variance simplifies to $\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1} = \sigma^2 \boldsymbol{Q}^{-1}$.

10.6 Efficiency

When comparing two unbiased estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ then $\hat{\theta}_n$ is at least as efficient as $\tilde{\theta}_n$ if it has no larger variance:

$$\operatorname{Var}(\hat{\theta}_n) \leq \operatorname{Var}(\tilde{\theta}_n).$$

For vector-valued estimators, we compare covariance matrices in the Loewner order: $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is a positive semidefinite matrix (see matrix tutorial for details).

Then, the estimator $\hat{\boldsymbol{\theta}}_n$ is at least as efficient as $\tilde{\boldsymbol{\theta}}_n$ if the covariance matrices satisfy

$$\operatorname{Var}(\hat{\boldsymbol{\theta}}_n) \preceq \operatorname{Var}(\tilde{\boldsymbol{\theta}}_n).$$

Under homoskedasticity, the OLS coefficient is the efficient estimator for $\boldsymbol{\beta}$.

Gauss–Markov Theorem

Consider the linear model $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i$ with

1. **Random sampling:** $(Y_i, \boldsymbol{X}_i')$ i.i.d.
2. **Exogeneity:** $E[u_i|\boldsymbol{X}_i] = 0$.
3. **Full rank:** $\boldsymbol{X}$ has column rank k.
4. **Homoskedasticity:** $\operatorname{Var}(u_i|\boldsymbol{X}_i) = \sigma^2$.

Then the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is the **Best Linear Unbiased Estimator (BLUE)**:

- **Linear:** it is linear in $\boldsymbol{Y}$;
- **Unbiased:** $E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta}$;
- **Best:** for any other linear unbiased estimator $\tilde{\boldsymbol{\beta}}$,

$$\text{Var}(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) \succeq \text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

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:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{gls} &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{Y} \\ &= \left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\boldsymbol{X}_iY_i\right),\end{aligned}$$

where

$$\boldsymbol{D} = \text{Var}(\boldsymbol{u}|\boldsymbol{X}) = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$$

GLS can be derived from the Method of Moments principle using the following moment condition:

$$E[\sigma_i^{-2}\boldsymbol{X}_iu_i] = E[\sigma_i^{-2}\boldsymbol{X}_iE[u_i|\boldsymbol{X}_i]] = \boldsymbol{0}.$$

Then $E[\sigma_i^{-2}\boldsymbol{X}_iu_i] = E[\sigma_i^{-2}\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})] = \boldsymbol{0}$ can be rearranged as

$$\boldsymbol{\beta} = E[\sigma_i^{-2}\boldsymbol{X}_i\boldsymbol{X}_i']^{-1}E[\sigma_i^{-2}\boldsymbol{X}_iY_i].$$

The GLS estimator is BLUE in the heteroskedastic linear regression model.

However, GLS is generally infeasible because $\boldsymbol{D}$ (or the σ_i^2) is unknown. Unless we can credibly model $\boldsymbol{D}$, the standard approach is to use OLS and account for heteroskedastity in additional inferential steps. When a plausible variance structure is available, one can estimate it and run feasible GLS (FGLS).

10.7 R Code

statistics-sec07.R