Statistics for Data Analytics

Jun.-Prof. Dr. Sven Otto

Last updated: September 25, 2025

Table of contents

Organization of the Course **3**
Timetable . . . 3

1 Data **6**
1.1 Datasets . . . 6
1.2 R programming language . . . 7
1.3 Datasets in R . . . 8
1.4 Importing data . . . 12
1.5 R-codes . . . 13

2 Sample distribution **14**
2.1 Empirical distribution function . . . 14
2.2 Histogram . . . 17
2.3 Empirical quantiles . . . 19
2.4 Empirical moments . . . 21
2.4.1 Sample moments . . . 21
2.4.2 Central sample moments . . . 22
2.4.3 Degree of freedom corrections . . . 22
2.4.4 Standardized sample moments . . . 23
2.5 Sample covariance . . . 26
2.6 R-codes . . . 28

3 Least squares **29**
3.1 Regression function . . . 29
3.2 Ordinary least squares (OLS) . . . 30
3.3 Simple linear regression (k=2) . . . 30
3.4 Regression plots . . . 32
3.5 Matrix notation . . . 34
3.6 R-squared . . . 36
3.7 Adjusted R-squared . . . 37
3.8 Too many regressors . . . 38
3.9 Perfect multicollinearity . . . 40
3.10 Dummy variable trap . . . 40
3.11 R-codes . . . 41

4 Probability **42**
4.1 Random sampling . . . 42
4.2 Random variables . . . 43
4.3 Events and probabilities . . . 44
4.4 Probability function . . . 46
4.5 Distribution function . . . 47
4.6 Point probabilities . . . 49
4.7 Bivariate distributions . . . 50
4.8 Independence . . . 55
4.9 Multivariate distributions . . . 56
4.10 IID sampling . . . 56
4.11 R-codes . . . 58

5 Expectation **59**
5.1 Discrete random variables . . . 59
5.2 Continuous random variables . . . 60
5.3 Unified definition of the expected value . . . 61
5.4 Transformed variables . . . 62
5.5 Linearity of the expected value . . . 64
5.6 Parameters and estimators . . . 64
5.7 Estimation of the mean . . . 64
5.8 Consistency . . . 65
5.9 Law of large numbers . . . 67
5.10 Heavy tails . . . 67
5.11 Estimation of the variance . . . 68
5.12 Bias-variance tradeoff . . . 69
5.13 R-codes . . . 70

6 Covariance **71**
6.1 Expectation of bivariate random variables . . . 71
6.2 Covariance and correlation . . . 73
6.3 Expectations for random vectors . . . 73
6.4 Population regression . . . 74
6.5 R-codes . . . 75

7 Conditional expectation **76**
7.1 Conditional distribution . . . 76
7.1.1 Conditioning on discrete variables . . . 78
7.1.2 Conditioning on continuous variables . . . 79
7.2 Conditional mean . . . 79
7.3 Rules of calculation . . . 81
7.4 Best predictor property . . . 83

7.5 Linear regression model . . . 84
7.5.1 Conditional mean independence (A1) . . . 84
7.5.2 Random sampling (A2) . . . 86
7.5.3 Finite moments and invertibility (A3 + A4) . . . 88
7.6 R-codes . . . 89

8 Simulations **90**
8.1 Consistent estimation . . . 90
8.2 Set up . . . 91
8.3 Monte Carlo algorithm . . . 91
8.4 Sample mean of coin flips . . . 92
8.5 Linear and nonlinear regression . . . 93
8.5.1 Simulation of the linear case . . . 96
8.5.2 Simulation of the nonlinear case . . . 97
8.6 R-codes . . . 99

9 Marginal effects **100**
9.1 Marginal Effects . . . 100
9.2 Control Variables . . . 102
9.3 CASchools: class size effect . . . 103
9.4 Polynomials . . . 106
9.5 Interactions . . . 107
9.6 Logarithms . . . 108
9.7 CASchools: nonlinear specifications . . . 110
9.8 R-codes . . . 113

10 Confidence intervals **114**
10.1 Estimation uncertainty . . . 114
10.2 Gaussian distribution . . . 115
10.2.1 Multivariate Gaussian distribution . . . 116
10.2.2 Chi-squared distribution . . . 117
10.2.3 Student t-distribution . . . 118
10.3 Classical Gaussian regression model . . . 119
10.4 Confidence interval: known variance . . . 120
10.5 Classical standard errors . . . 123
10.6 Confidence intervals: heteroskedasticity . . . 127
10.7 Confidence interval with non-normal errors . . . 130
10.8 Central limit theorem . . . 131
10.9 CASchools data . . . 133
10.10R-codes . . . 134

11 Hypothesis testing **135**
11.1 Statistical hypotheses . . . 135

11.2 t-Tests . . . 137
11.3 The p-value . . . 139
11.4 Multiple testing problem . . . 141
11.5 Joint Hypotheses . . . 141
11.6 Wald Test . . . 142
11.7 F-Test . . . 145
11.8 Diagnostics tests . . . 147
11.8.1 Breusch-Pagan Test (Koenker's version) . . . 148
11.8.2 Jarque-Bera Test . . . 148
11.9 Nonliearities in test score regressions . . . 149
11.10R-codes . . . 156

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 applications in `R`.

Timetable

See KLIPS for a detailed schedule.

Lecture Material

- This online script: webpage and pdf version
- eWhiteboard: lecture and exercises
- Problemsets and Rscripts: sciebo folder
- More info on exam/exercises/assignments: ILIAS course

Literature

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

The course is based on Stock and Watson, *Introduction to Econometrics (Fourth Edition)*, Chapters 1–9, 15, 18, and 19. The Stock and Watson textbook is available here (Uni Köln VPN connection required).

Further textbooks I can recommend:

- *Probability and Statistics for Economists*, by Bruce E. Hansen
- *Econometrics*, by Bruce E. Hansen

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	XVIII (Main Building)	Lecture

- *Econometric Theory and Methods*, by R. Davidson and J.G. MacKinnon (link)

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

Assessment

The course will be graded by a 90-minute written exam. There will be two optional bonus assignments during the lecture period. These assignments will allow you to earn bonus points that will be added to your overall exam score, but they are optional and not required to achieve the maximum score on the exam. 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. Of course, you can reach me via e-mail: sven.otto@uni-koeln.de

Important Dates

Bonus assignment 1	Oct 22 - Nov 05
Bonus assignment 2	Nov 06 - Nov 20
Registration deadline exam 1	Nov 14, 2024
Exam 1	Nov 28, 2024
Registration deadline exam 2	Mar 03, 2025
Exam 2 (alternate date)	Mar 17, 2025

Please register for the exam on time. If you miss the registration deadline, you will not be able to take the exam (the Examinations Office is very strict about this). You only need to take one of the two exams to complete the course. The second exam will serve as a make-up exam for those who fail the first exam or do not take the first exam.

R-Packages

To run the R code of the lecture script, you will need to install some additional packages.

```
install.packages(
  c("AER", "fixest", "plm", "dynlm",
    "glmnet", "moments", "urca",
    "tidyverse", "stargazer", "BVAR",
    "kableExtra", "scatterplot3d", "tinytex")
  )
```

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

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

1 Data

1.1 Datasets

A **univariate dataset** is a sequence of observations $Y_1, \dots, Y_n$. These n observations can be organized into the **data vector** $\boldsymbol{Y}$, represented as $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$. For example, if you conduct a survey and ask five individuals about their hourly earnings, your data vector might look like

$$\boldsymbol{Y} = \begin{pmatrix} 18.22 \\ 23.85 \\ 10.00 \\ 6.39 \\ 7.42 \end{pmatrix}.$$

Typically we have data on more than one variable, such as years of education and the gender. Categorical variables are often encoded as **dummy variables**, which are binary variables. The female dummy variable is defined as 1 if the gender of the person is female and 0 otherwise.

person	wage	education	female
1	18.22	16	1
2	23.85	18	0
3	10.00	16	1
4	6.39	13	0
5	7.42	14	0

A k**-variate dataset** (or multivariate dataset) is a collection of n vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ containing data on k variables. The i-th vector $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ik})'$ contains the data on all k variables for individual i. 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 use the transpose notation $\boldsymbol{X}_i'$, which is a row vector. The data matrix and vectors for our example

are:

$$\boldsymbol{X} = \begin{pmatrix} 18.22 & 16 & 1 \\ 23.85 & 18 & 0 \\ 10.00 & 16 & 1 \\ 6.39 & 13 & 0 \\ 7.42 & 14 & 0 \end{pmatrix}, \quad \boldsymbol{X}_1 = \begin{pmatrix} 18.22 \\ 16 \\ 1 \end{pmatrix}, \boldsymbol{X}_2 = \begin{pmatrix} 23.85 \\ 18 \\ 0 \end{pmatrix}, \ldots$$

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, please consult the following resources:

Crash Course on Matrix Algebra:

matrix.svenotto.com
Section 19.1 of the Stock and Watson book also provides a brief overview of matrix algebra concepts.

1.2 R programming language

The best way to learn statistical methods is to program and apply them yourself. Throughout this course, we will use the `R` programming language for implementing empirical methods and analyzing real-world datasets.

If you are just starting with `R`, it is crucial to familiarize yourself with its basics. Here's an introductory tutorial, which contains a lot of valuable resources:

Getting Started with R:

rintro.svenotto.com

For those new to R, I also recommend the interactive R package SWIRL, which offers an excellent way to learn directly within the R environment. Additionally, a highly recommended online book to learn R programming is Hands-On Programming with R.

One of the best features of R is its extensive ecosystem of packages contributed by the statistical community. You find R packages for almost any statistical method out there and many statisticians provide R packages to accompany their research.

One of the most frequently used packages in applied econometrics is the `AER` package ("Applied Econometrics with R"), which provides a comprehensive collection of inferential methods for

linear models. You can install the package with the command `install.packages("AER")` and you can load it with

```
library(AER)
```

at the beginning of your code. We will explore several additional packages in the course of the lecture.

1.3 Datasets in R

R includes many built-in datasets and packages of datasets that can be loaded directly into your R environment. For illustration, we consider the `CASchools` dataset available in the `AER` package. This dataset is used in the Stock and Watson textbook in sections 4-8. It contains information on various characteristics of schools in California, such as test scores, teacher salaries, and student demographics.

To load this dataset into your R session, simply use:

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

To get a description of the dataset, use the `?CASchools` command.

```
class(CASchools)
```

```
[1] "data.frame"
```

The `CASchools` dataset is stored as a `data.frame`, R's most common data storage class for tabular data as in $\boldsymbol{X}$. It organizes data in the form of a table, with variables as columns and observations as rows.

To inspect the structure of your dataset, you can use `str()`:

```
str(CASchools)
```

```
'data.frame':   420 obs. of  14 variables:
 $ district   : chr  "75119" "61499" "61549" "61457" ...
 $ school     : chr  "Sunol Glen Unified" "Manzanita Elementary" "Thermalito Union Elementary
 $ county     : Factor w/ 45 levels "Alameda","Butte",..: 1 2 2 2 2 6 29 11 6 25 ...
 $ grades     : Factor w/ 2 levels "KK-06","KK-08": 2 2 2 2 2 2 2 2 2 1 ...
 $ students   : num  195 240 1550 243 1335 ...
```

```
 $ teachers   : num  10.9 11.1 82.9 14 71.5 ...
 $ calworks   : num  0.51 15.42 55.03 36.48 33.11 ...
 $ lunch      : num  2.04 47.92 76.32 77.05 78.43 ...
 $ computer   : num  67 101 169 85 171 25 28 66 35 0 ...
 $ expenditure: num  6385 5099 5502 7102 5236 ...
 $ income     : num  22.69 9.82 8.98 8.98 9.08 ...
 $ english    : num  0 4.58 30 0 13.86 ...
 $ read       : num  692 660 636 652 642 ...
 $ math       : num  690 662 651 644 640 ...
```

The dataset contains variables of different types: `chr` for character/text data, `Factor` for categorical data, and `num` for numeric data. The `head()` function displays its first few rows:

```
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 pipe operator `|>` efficiently chains commands. It passes the output of one function as the input to another. For example:

```
CASchools[,c("school", "county", "income")] |> summary()
```

```
   school                  county        income      
 Length:420         Sonoma     : 29   Min.   : 5.335  
 Class :character   Kern       : 27   1st Qu.:10.639  
 Mode  :character   Los Angeles: 27   Median :13.728  
                    Tulare     : 24   Mean   :15.317  
                    San Diego  : 21   3rd Qu.:17.629  
```

```
                Santa Clara: 20   Max.   :55.328
                (Other)    :272
```

The `summary()` function presents a concise overview, showing absolute frequencies for categorical variables and descriptive statistics for numerical variables.

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

We can easily add new variables to a dataframe, 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
```

The variable `english` indicates the proportion of students whose first language is not English and who may need additional support. We might be interested in the dummy variable `HiEL`, which indicates whether the proportion of English learners is above 10 percent or not:

```
# append HiEL to CASchools
CASchools$HiEL = (CASchools$english >= 10) |> as.numeric()
```

Note that `CASchools$english >= 10` 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`.

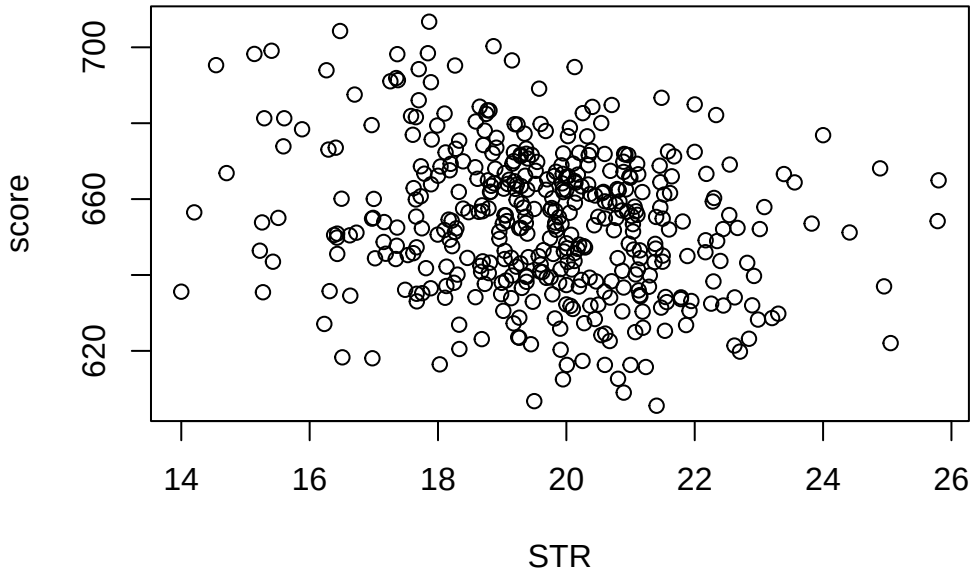
The first few values of some selected variables look like this:

```
CASchools[,c("STR", "score", "english", "HiEL", "income")] |> head()
```

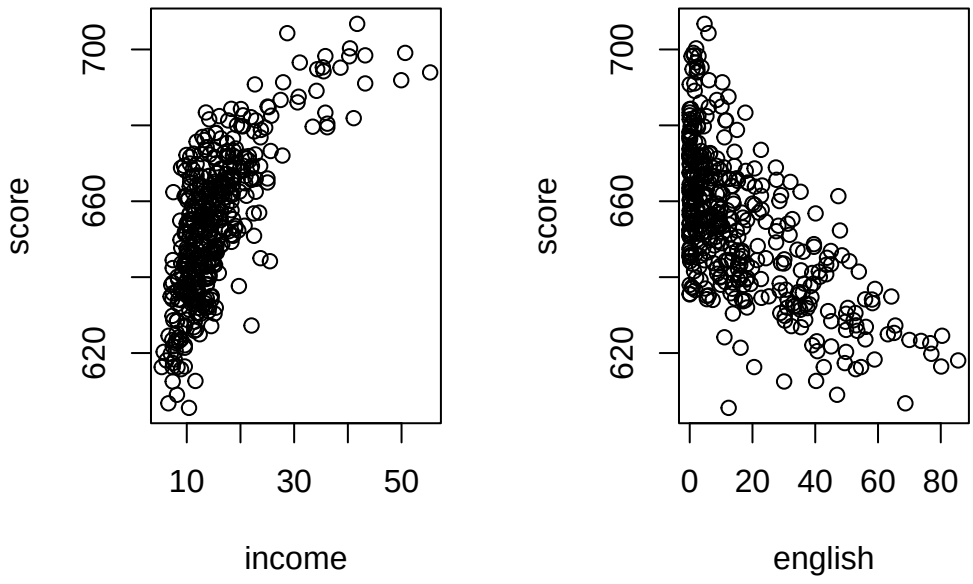
```
       STR  score   english HiEL    income
1 17.88991 690.80  0.000000    0 22.690001
2 21.52466 661.20  4.583333    0  9.824000
3 18.69723 643.60 30.000002    1  8.978000
4 17.35714 647.70  0.000000    0  8.978000
5 18.67133 640.85 13.857677    1  9.080333
6 21.40625 605.55 12.408759    1 10.415000
```

Scatterplots provide further insights:

```
plot(score~STR, data = CASchools)
```




```
par(mfrow = c(1,2))
plot(score~income, data = CASchools)
plot(score~english, data = CASchools)
```



The option `par(mfrow = c(1,2))` allows to display multiple plots side by side. Try what happens if you replace `c(1,2)` with `c(2,1)`.

1.4 Importing data

The internet serves as a vast repository for data in various formats, with `csv` (comma-separated values), `xlsx` (Microsoft Excel spreadsheets), and `txt` (text files) being the most commonly used.

R supports various functions for different data formats:

- `read.csv()` for reading comma-separated values
- `read.csv2()` for semicolon-separated values (adopting the German data convention of using the comma as the decimal mark)
- `read.table()` for whitespace-separated files
- `read_excel()` for Microsoft Excel files (requires the `readxl` package)
- `read_stata()` for STATA files (requires the `haven` package)

Let's import the CPS dataset from Bruce Hansen's textbook. 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
- Description: cps09mar_description.pdf

```
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 further variables:

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

We will be using the cps data in the next sections, so it is a good idea to save the dataset to your computer:

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

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

1.5 R-codes

statistics-sec01.R

2 Sample distribution

In statistics, a univariate dataset $Y_1, \dots, Y_n$ or a multivariate dataset $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is often called a **sample** because it typically represents observations selected from a larger population. The **sample distribution** indicates how the sample values are distributed across possible outcomes. **Summary statistics**, such as the sample mean and sample variance, provide a concise representation of key characteristics of the sample distribution.

2.1 Empirical distribution function

The sample distribution of a univariate sample $Y_1, \dots, Y_n$ is represented by the **empirical cumulative distribution function (ECDF)**, which shows the proportion of observations in the sample that are less than or equal to a certain value a. There are two equivalent ways to define the ECDF: using the **indicator function** and using **order statistics**.

Indicator function

The **indicator function** $I(\cdot)$ is defined as:

$$I(Y_i \leq a) = \begin{cases} 1 & \text{if } Y_i \leq a, \\ 0 & \text{if } Y_i > a. \end{cases}$$

The ECDF is defined as:

$$\widehat{F}(a) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq a).$$

This formula calculates the proportion of sample observations that are less than or equal to the value a.

Order statistics

Equivalently, the ECDF can be defined using **order statistics**. Order statistics are the sample data arranged in ascending order:

$$Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}.$$

In R, you can compute the order statistics of a univariate data vector `Y` using the command `sort(Y)`. The ECDF is then defined as:

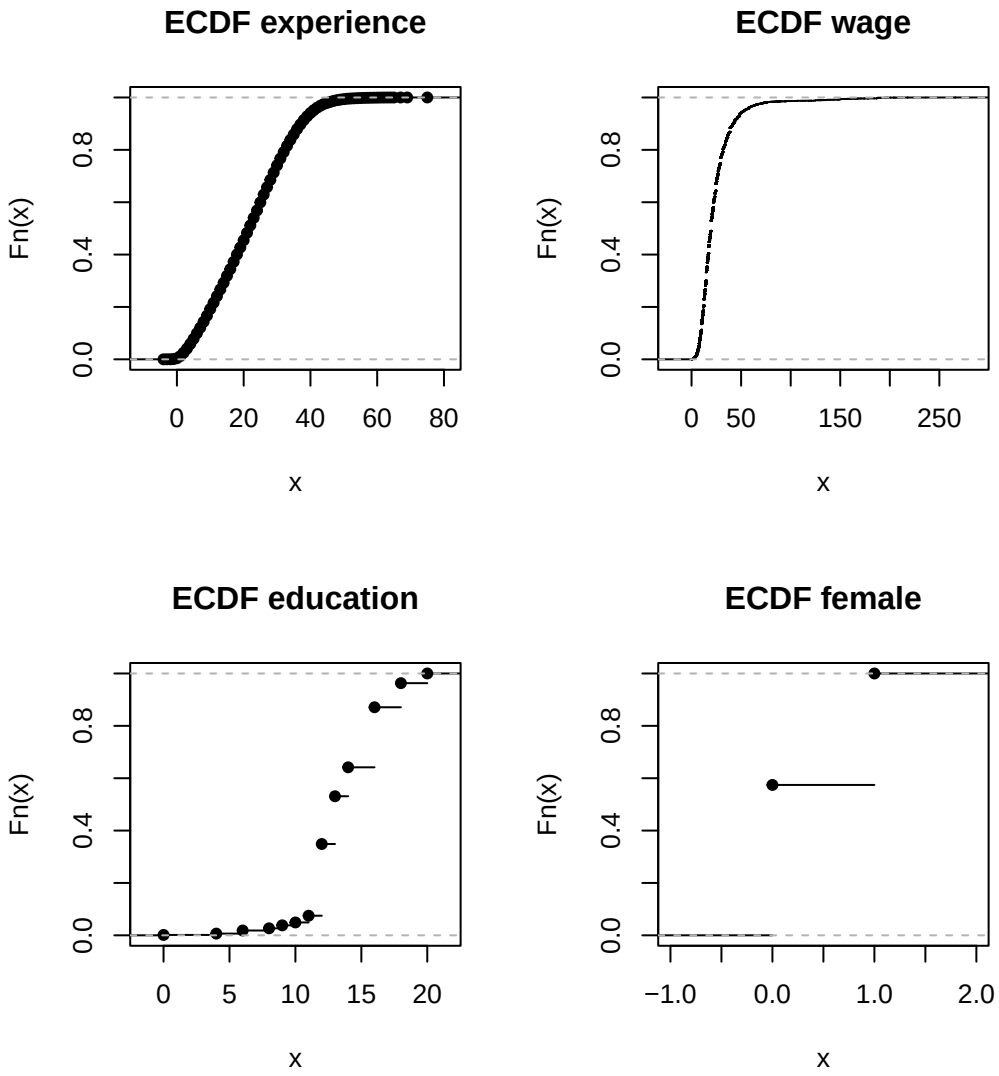
$$\widehat{F}(a) = \begin{cases} 0 & \text{if } a < Y_{(1)}, \\ \frac{k}{n} & \text{if } Y_{(k)} \le a < Y_{(k+1)}, \quad k = 1, 2, \dots, n-1, \\ 1 & \text{if } a \ge Y_{(n)}. \end{cases}$$

The ECDF is a step function that increases by $1/n$ at each data point $Y_{(k)}$. The function remains constant between data points and jumps at each observed value in the sample.

Some ECDFs of the CPS data

```
cps = read.csv("cps.csv")
exper = cps$experience
wage = cps$wage
edu = cps$education
fem = cps$female
```

```
par(mfrow = c(2,2))
plot.ecdf(exper, main = "ECDF experience")
plot.ecdf(wage, main = "ECDF wage")
plot.ecdf(edu, main = "ECDF education")
plot.ecdf(fem, main = "ECDF female")
```



A variable is **discrete** if it has a countable number of possible outcomes. It is **continuous** if it can take any value within a range or continuum of possible outcomes. The ECDF is always a step function with steps becoming arbitrarily small for continuous distributions as n increases.

The plots show that `edu` and `fem` are discrete variables. The variable `exper`, although measured in years and technically discrete, has a large number of possible values, which makes it effectively "almost" continuous. On the other hand, the variable `wage` is clearly continuous, as it can take on a wide range of values.

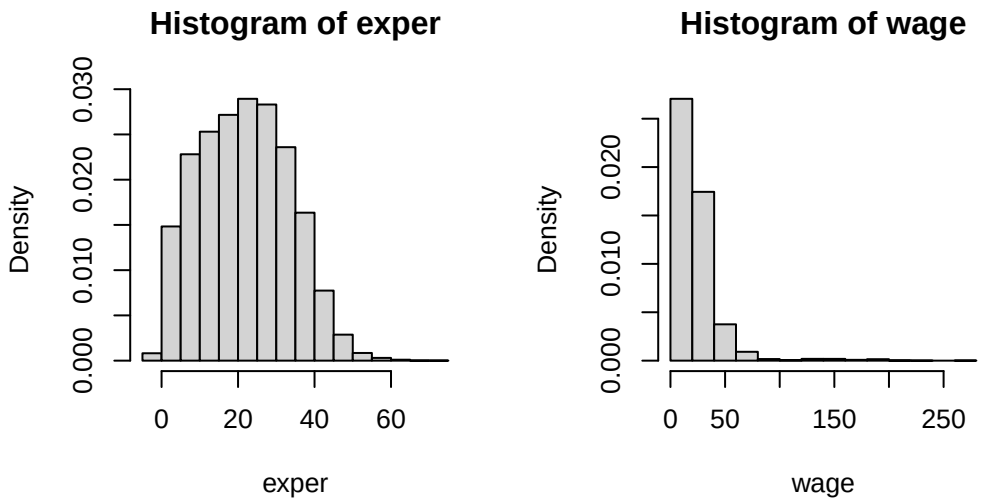
2.2 Histogram

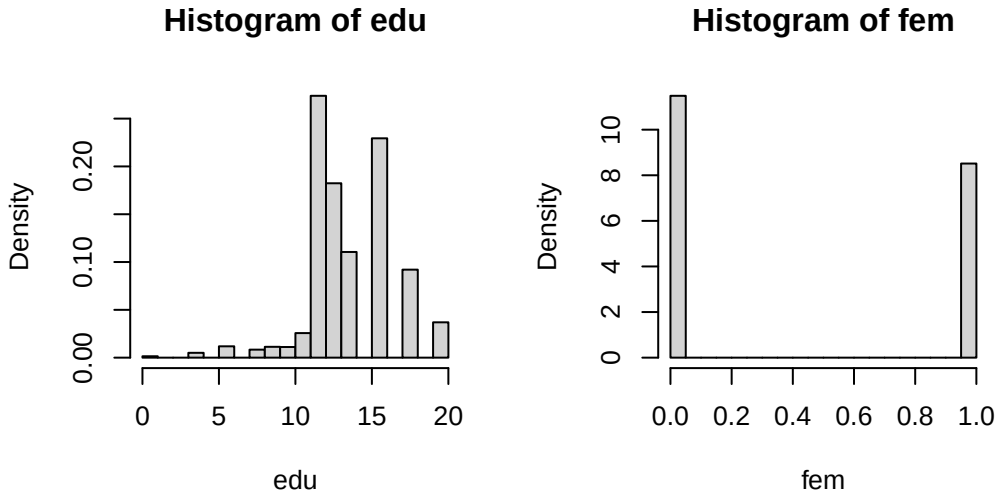
Histograms offer a more intuitive visual representation of the sample distribution compared to the ECDF. A histogram divides the data range into B bins each of equal width h and counts the number of observations n_j within each bin. The height of the histogram at a in the j-th bin is

$$\hat{f}(a) = \frac{n_j}{nh}.$$

The histogram is the plot of these heights, displayed as rectangles, with their area normalized so that the total area equals 1.

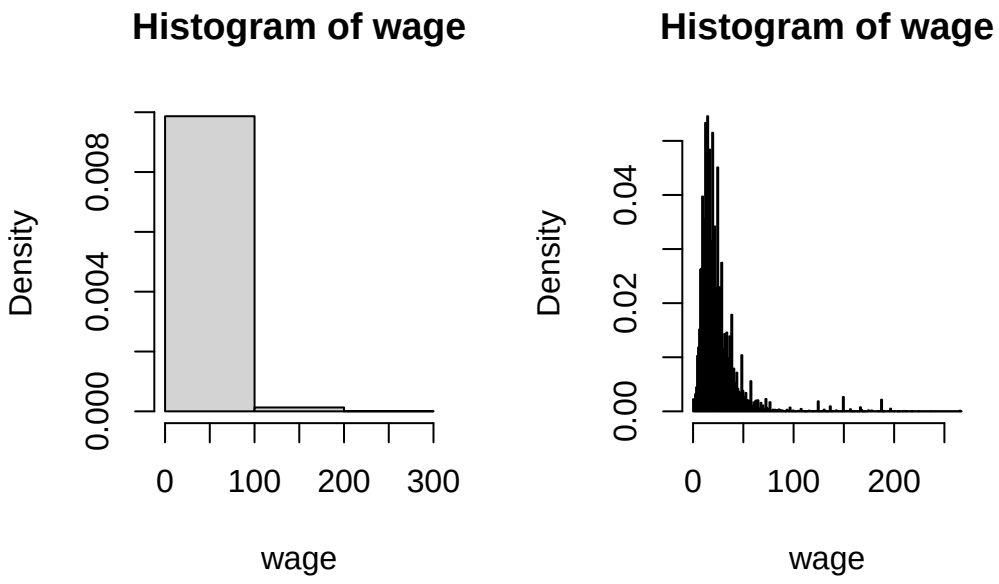
```
par(mfrow = c(2,2))
hist(exper, probability = TRUE)
hist(wage, probability = TRUE)
hist(edu, probability = TRUE)
hist(fem, probability = TRUE)
```





Running `hist(wage, probability=TRUE)` automatically selects a suitable number of bins B. Note that `hist(wage)` will plot absolute frequencies instead of relative ones. The shape of a histogram depends on the choice of B. You can experiment with different values using the `breaks` option:

```
par(mfrow = c(1,2))
hist(wage, probability = TRUE, breaks = 3)
hist(wage, probability = TRUE, breaks = 300)
```



2.3 Empirical quantiles

Another way of characterizing the sample distribution is to use empirical quantiles.

Median

The median is a central value that splits the distribution into two equal parts. The empirical median of a sorted dataset is found at the point where the ECDF reaches 0.5. For an even-sized dataset, the median is the average of the two central observations:

$$\widehat{med} = \begin{cases} Y_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2}(Y_{(\frac{n}{2})} + Y_{(\frac{n}{2}+1)}) & \text{if } n \text{ is even} \end{cases}$$

The median corresponds to the 0.5-quantile of the distribution.

Quantile

The empirical p-quantile $\hat{q}_p$ is a value at which p percent of the data falls below it. It is found at the point where the ECDF reaches p.

Since the ECDF is flat between its jumps, the empirical p-quantile may not be unique. It can be computed as the linear interpolation at $h = (n-1)p + 1$ between $Y_{(\lfloor h \rfloor)}$ and $Y_{(\lceil h \rceil)}$:

$$\hat{q}_p = Y_{(\lfloor h \rfloor)} + (h - \lfloor h \rfloor)(Y_{(\lceil h \rceil)} - Y_{(\lfloor h \rfloor)}).$$

Note that $\lfloor h \rfloor$ and $\lceil h \rceil$ denotes rounding down and rounding up to the next integer. This interpolation scheme is standard in R, although multiple approaches exist to define empirical quantiles (see here).

To calculate the 0.05 quantile, the median and the 0.95 quantile of the data, we can use the following command:

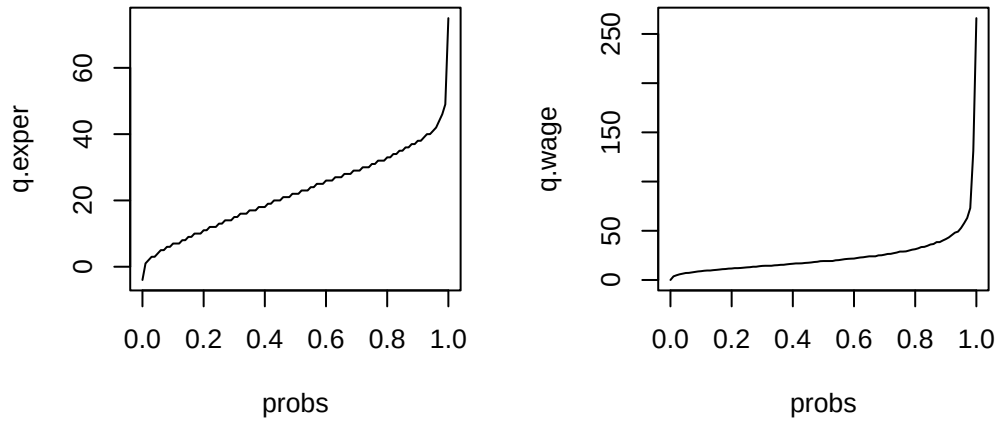
```
quantile(exper, probs = c(0.05, 0.5, 0.95))
```

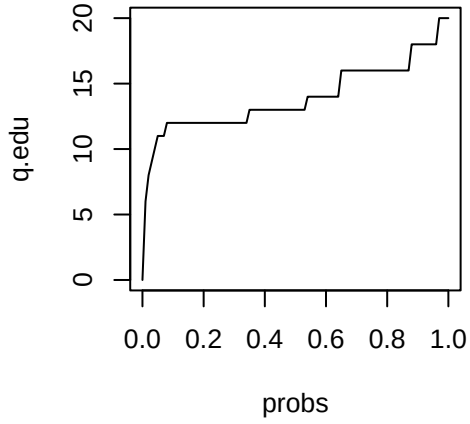
```
 5% 50% 95%
  4  22  41
```

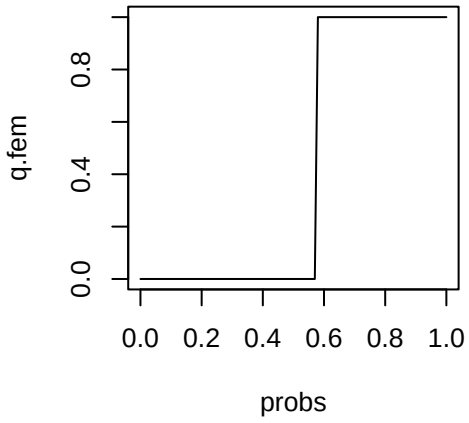
Let's plot all quantiles as a function on a fine grid of probabilities between 0 and 1:

```
# Define a fine grid of probabilities
probs = seq(0, 1, by = 0.01)
# Compute the quantiles
q.exper = quantile(exper, probs)
q.wage = quantile(wage, probs)
q.edu = quantile(edu, probs)
q.fem = quantile(fem, probs)
```

```
par(mfrow = c(2,2))
plot(probs, q.exper, type="l")
plot(probs, q.wage, type="l")
plot(probs, q.edu, type="l")
plot(probs, q.fem, type="l")
```







Check that these are indeed the correct quantiles using the ECDF plots from above.

2.4 Empirical moments

Many stylized features and characteristics of a sample distribution can be computed from sample moments.

2.4.1 Sample moments

The r**-th sample moment** about the origin (also called the raw moment) 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.

To compute the sample mean of a vector `Y` in R, use `mean(Y)` or alternatively `sum(Y)/length(Y)`. The r-th sample moment can be calculated with `mean(Y^r)`.

2.4.2 Central sample 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 **sample standard deviation**, 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.

2.4.3 Degree of freedom corrections

When computing the sample mean $\overline{Y}$, we have n degrees of freedom because each data point Y_i can vary freely. However, when calculating the deviations $(Y_i - \overline{Y})$, these deviations are subject to the constraint:

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

This means that the deviations are not all free to vary; they are connected by this equation. 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:

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

The adjusted sample standard deviation is:

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

To compute the sample variance and sample standard deviaion of a vector `Y` in R, use `mean(Y^2)-mean(Y)^2` and `sqrt(mean(Y^2)-mean(Y)^2)`, respectively. The built-in functions `var(Y)` and `sd(Y)` compute their adjusted versions.

2.4.4 Standardized sample 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:

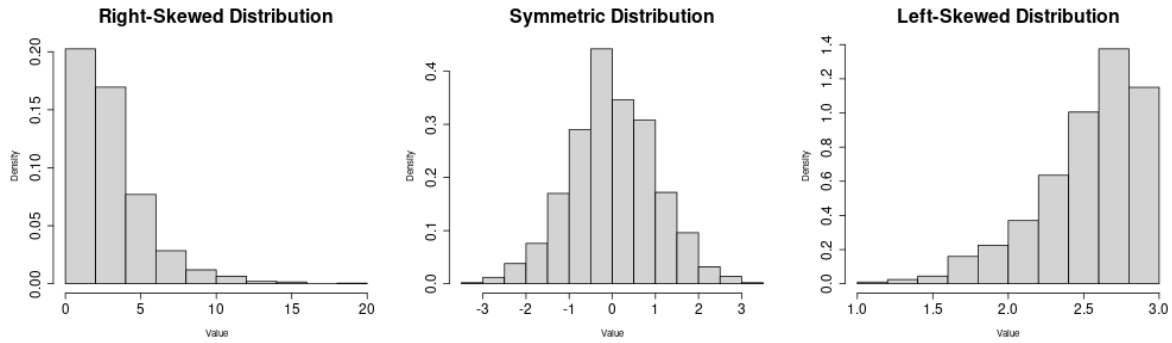
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\hat{\sigma}_Y}\right)^r$$

Skewness

For example, the third standardized sample moment ($r=3$) is the **sample skewness**:

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

The skewness is a measure of asymmetry around the mean. A non-zero skewness indicates an asymmetric distribution, with positive values indicating a right tail and negative values a left tail.



To compute the sample skewness in R, use:

```
mean((Y-mean(Y))^3)/(mean(Y^2)-mean(Y)^2)^(3/2)
```

For convenience, you can use the `skewness(Y)` function from the `moments` package, which performs the same calculation.

```
library(moments)
c(skewness(exper), skewness(wage), skewness(edu), skewness(fem))
```

```
[1]  0.1862605  4.3201570 -0.2253251  0.3004446
```

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

Kurtosis

The **sample kurtosis** is the fourth standardized sample moment ($r = 4$):

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

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

To compute the sample kurtosis in R, use:

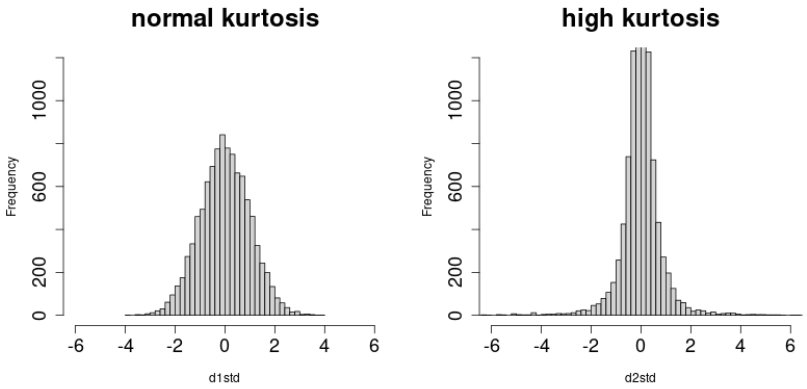
```
mean((Y-mean(Y))^4)/(mean((Y-mean(Y))^2))^2
```

For convenience, you can use the `kurtosis(Y)` function from the `moments` package, which performs the same calculation.

```
c(kurtosis(exper), kurtosis(wage), kurtosis(edu), kurtosis(fem))
```

```
[1]  2.374758 30.370331  4.498264  1.090267
```

The variable `wage` exhibits heavy tails due to a few super-rich outliers in the sample. In contrast, `fem` has light tails because there are approximately equal numbers of women and men.



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.

The plot shows histrograms of two standardized univariate datasets (i.e., their sample mean is 0 and their sample variance is 1). The dataset from the left plot has a normal sample kurtosis (around 3) and the dataset from the right plot has a high sample kurtosis with more obervarions in the tails.

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(wage, probability = TRUE)
hist(log(wage), probability = TRUE, xlim = c(-3, 6))
```


```
c(skewness(log(wage)), kurtosis(log(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.

2.5 Sample covariance

Consider a multivariate dataset $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, such as the following subset of the `cps` dataset:

```
dat = data.frame(wage, edu, fem)
```

Sample mean vector

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

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i.$$

```
colMeans(dat)
```

```
      wage        edu        fem
23.9026619 13.9246187  0.4257223
```

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 **adjusted sample covariance matrix** S is defined as

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

Its elements $s_{h,l}$ are the **adjusted sample covariances**, with main diagonal elements $s_h^2 = s_{h,h}$ being the adjusted sample variances:

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

```
cov(dat)
```

```
           wage         edu         fem
wage 428.948332 21.82614057 -1.66314777
edu   21.826141  7.53198925  0.06037303
fem   -1.663148  0.06037303  0.24448764
```

Sample correlation matrix

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

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

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

$$C = D^{-1} S D^{-1},$$

where D is the diagonal matrix of adjusted sample standard deviations:

$$D = diag(s_1, \ldots, s_k) = \begin{pmatrix} s_1 & 0 & \ldots & 0 \\ 0 & s_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & s_k \end{pmatrix}$$

The matrices $\widehat{\Sigma}$, S, and C are symmetric.

```
cor(dat)
```

```
           wage        edu         fem
wage  1.0000000 0.38398973 -0.16240519
edu   0.3839897 1.00000000  0.04448972
fem  -0.1624052 0.04448972  1.00000000
```

We find a strong positive correlation between `wage` and `edu`, a substantial negative correlation between `wage` and `fem`, and a negligible correlation between `edu` and `fem`.

2.6 R-codes

statistics-sec02.R

3 Least squares

3.1 Regression function

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:

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

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 continuous and categorical variables), and logarithmic transformations.

3.2 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:

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

3.3 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)	Education	Education^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

$$\hat{\beta}_2 = \frac{\hat{\sigma}_{YZ}}{\hat{\sigma}_Z^2}, \quad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{Z}, \tag{3.1}$$

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

3.4 Regression plots

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

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

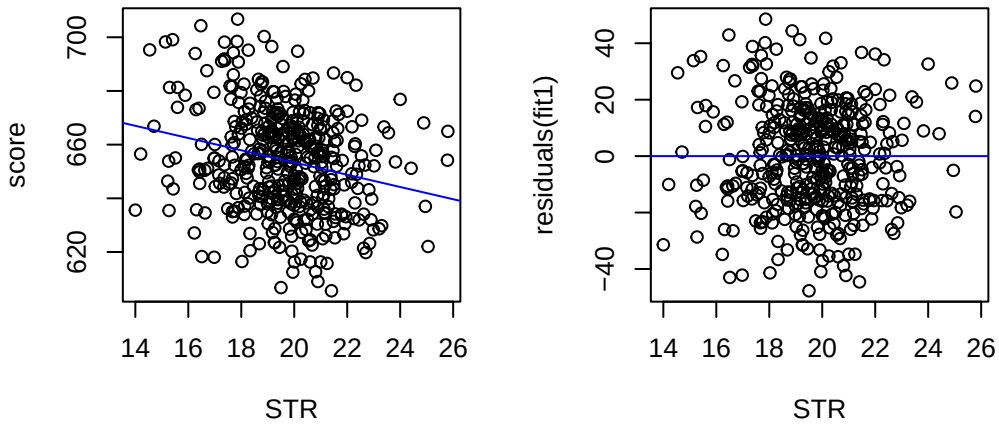
```
(Intercept)         STR
 698.932949   -2.279808
```

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)
abline(fit1, col="blue")
plot(STR, residuals(fit1))
abline(0,0,col="blue")
```

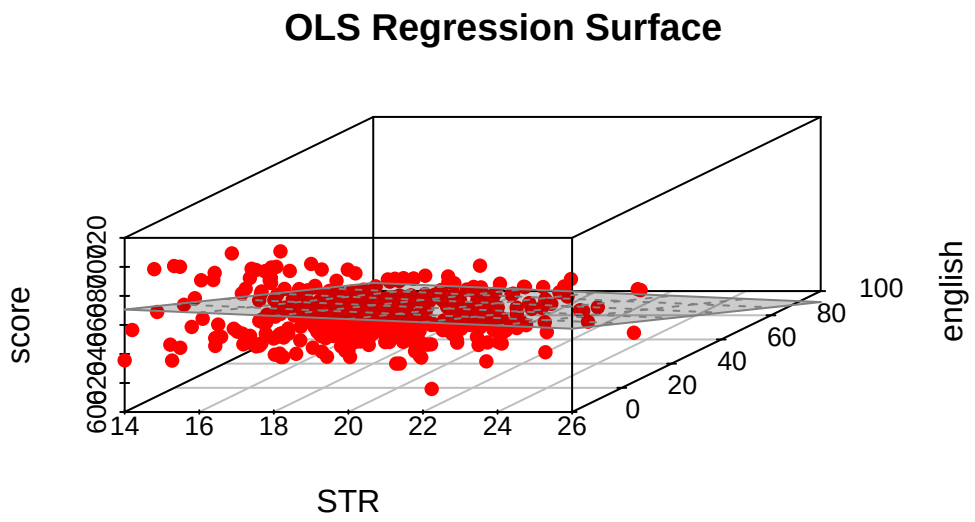


Let's include the percentage of english learners as an additional regressor:

```
english = CASchools$english
fit2= lm(score ~ STR + english)
fit2$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:

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

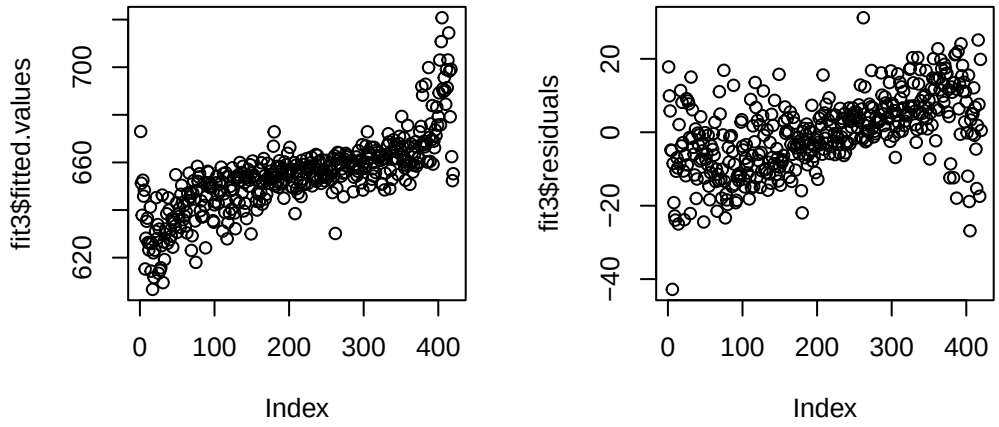
```
 (Intercept)          STR      english       income
640.31549821  -0.06877542  -0.48826683   1.49451661
```

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

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

For specifications with multiple regressors, fitted values and residuals can still be visualized:

```
par(mfrow = c(1,2), cex=0.8)
plot(fit3$fitted.values)
plot(fit3$residuals)
```



The pattern of fitted values arises because the observations in the `CASchools` dataset are sorted in ascending order by test score.

3.5 Matrix notation

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:

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

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

The vector of residuals is given by

$$\hat{\boldsymbol{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}.$$

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

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

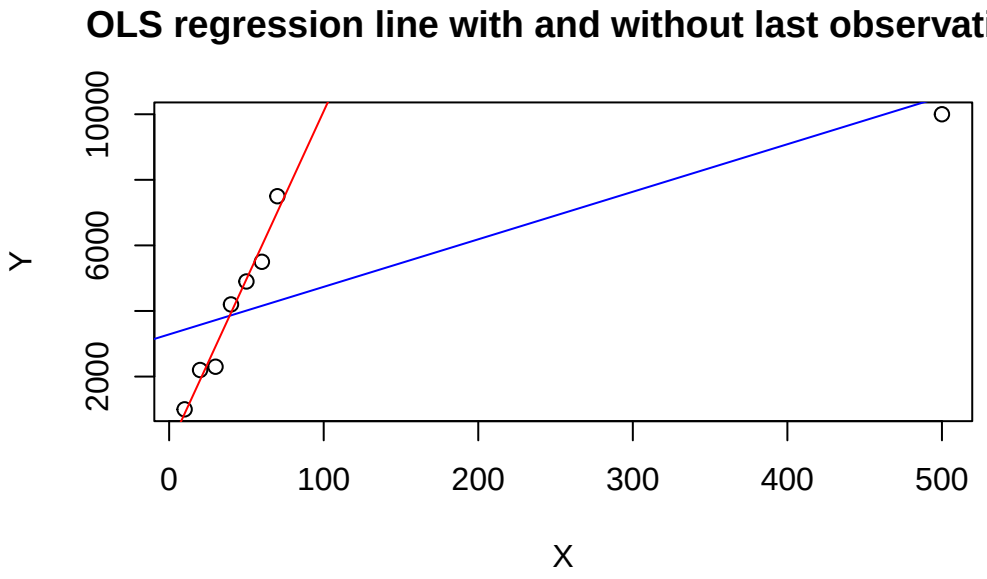
are called **leverage values** or *hat values* and measure how far away the regressor values of the i-th observation X_i are from those of the other observations.

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
```

3.6 R-squared

Consider the following sample variances:

Dependent variable	$\hat{\sigma}^2_Y = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$
Fitted values	$\hat{\sigma}^2_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2$
Residuals	$\hat{\sigma}^2_{\widehat{u}} = \frac{1}{n}\sum_{i=1}^n \hat{u}_i^2$

An important property of the residual vector is that it is orthogonal to the columns of $\boldsymbol{X}$, i.e.

$$\boldsymbol{X}'\hat{\boldsymbol{u}} = \begin{pmatrix} \sum_{i=1}^n \hat{u}_i \\ \sum_{i=1}^n X_{i2}\hat{u}_i \\ \vdots \\ \sum_{i=1}^n X_{ik}\hat{u}_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.2}$$

In particular, the sample mean of the residuals is zero, which is why it does not appear in the residual sample variance $\hat{\sigma}^2_{\widehat{u}}$.

Moreover, the following relationship holds (analysis of variance formula):

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

Hence, the larger the proportion of the explained sample variance, the better the fit of the OLS regression. This motivates the definition of the **R-squared coefficient**:

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

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.

3.7 Adjusted R-squared

Recall that the deviations $(Y_i - \overline{Y})$ cannot vary freely because they are subject to the constraint $\sum_{i=1}^n (Y_i - \overline{Y})$, which is why we loose 1 degree of freedom in the sample variance of $\boldsymbol{Y}$.

For the sample variance of $\hat{\boldsymbol{u}}$, we loose k degrees of freedom because the residuals are subject to the constraints from Equation 3.2. The adjusted sample variance of the residuals is therefore defined as:

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

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

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

The squareroot 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 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.

The commands `summary(fit)$r.squared` and `summary(fit)$adj.r.squared` return the R-squared and adjusted R-squared values, respectively. The SER can be returned by `summary(fit)$sigma`.

The `stargazer()` function can be used to produce nice regression outputs:

```
library(stargazer)
```

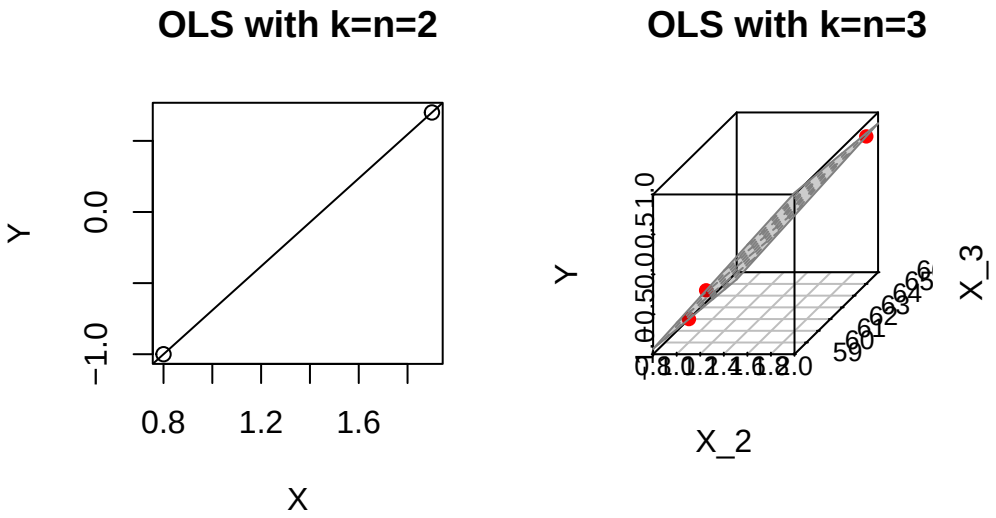
```
stargazer(fit1, fit2, fit3, type="latex", report="vc*", omit.stat = "f",
          star.cutoffs = NA, df=FALSE, omit.table.layout = "n",
          digits = 4, header = FALSE)
```

3.8 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.

Table 3.2

	Dependent variable:		
	score		
	(1)	(2)	(3)
STR	−2.2798	−1.1013	−0.0688
english		−0.6498	−0.4883
income			1.4945
Constant	698.9329	686.0322	640.3155
Observations	420	420	420
R^2	0.0512	0.4264	0.7072
Adjusted R^2	0.0490	0.4237	0.7051
Residual Std. Error	18.5810	14.4645	10.3474



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

3.9 Perfect multicollinearity

The only requirement for computing the OLS coefficients is the invertibility of the matrix $\boldsymbol{X}'\boldsymbol{X}$. 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 $\hat{\boldsymbol{\beta}}$ 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.

Multicollinearity means that at least one regressor is redundant and can be dropped.

3.10 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
  19.338695    6.997155          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 $\boldsymbol{X}'\boldsymbol{X}$.

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:

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

```
   married nonmarried
  26.33585   19.33869
```

3.11 R-codes

statistics-sec03.R

4 Probability

4.1 Random sampling

From the perspective of empirical analysis, a dataset $Y_1, \dots, Y_n$ or $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is simply an array of fixed numbers presented to a researcher. The summary statistics we compute – such as sample means, sample correlations, and OLS coefficients – are functions of this given dataset.

While these statistics provide a snapshot of the data at hand, they do not automatically offer insights into the broader world from which the data originated. To add deeper meaning to these numbers and draw conclusions about underlying dependencies and causalities, we need to consider how the data were obtained.

In statistical theory, a dataset is viewed as the result of a **random experiment**. The gender of the next person you meet, daily fluctuations in stock prices, monthly music streams of your favorite artist, or the annual number of pizzas consumed – all involve a certain amount of randomness.

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 theoretical construct, representing not just the existing physical population but all possible future or hypothetical individuals.

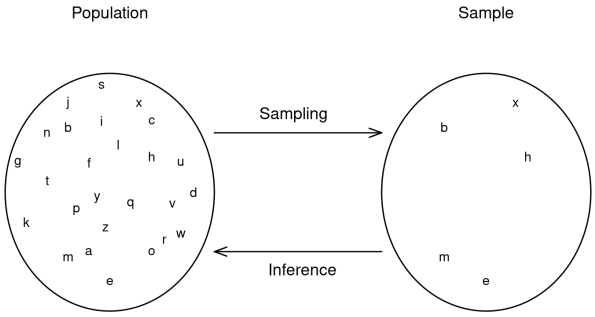


Figure 4.1: Sampling illustration

This figure demonstrates the concept of sampling. The left side displays the full set of letters from "a" to "z", representing the entire (infinite) population. From this population, five letters are randomly chosen, forming the sample shown on the right side.

The goal of **statistical inference** is to learn about the underlying population distribution by analyzing the observed sample. To do so, we need to make assumptions about how the data were sampled.

The simplest and ideal case is **random sampling**, where observations are randomly drawn from this infinite distribution with replacement – like randomly drawing balls from an urn, or randomly selecting individuals for a representative survey. This principle is also known as **i.i.d. sampling** (independent and identically distributed sampling). To define these concepts rigorously, we require probability theory.

4.2 Random variables

A **random variable** is a numerical summary of a random experiment. An **outcome** is a specific result of a random experiment. The **sample space** S is the set/collection of all potential outcomes.

Let's consider some examples:

- *Coin toss:* The outcome of a coin toss can be "heads" or "tails". This random experiment has a two-element sample space: $S = \{heads, tails\}$. We can express the experiment as a binary random variable:

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

- *Gender:* If you conduct a survey and interview a random person to ask them about their gender, the answer may be "female", "male", or "diverse". It is a random experiment since the person to be interviewed is selected randomly. The sample space has three elements: $S = \{female, male, diverse\}$. To focus on female vs. non-female, we can define the female dummy variable:

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

  Similarly, dummy variables for *male* and *diverse* can be defined.
- *Education level:* If you ask a random person about their education level according to the ISCED-2011 framework, the outcome may be one of the eight ISCED-2011 levels. We have an eight-element sample space:

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

Table 4.1: ISCED 2011 levels

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

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

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

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

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

These random variables have in common that they take values on the real line $\mathbb{R}$ but their outcome is uncertain before conducting the random experiment (i.e. flipping the coin or selecting a random person to be interviewed).

4.3 Events and probabilities

An **event** of a random variable Y is a specific subset of the real line. Any real number defines an event (elementary event), and any open, half-open, or closed interval represents an event as well.

Let's define some specific events:

- Elementary events:

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

- Half-open events:

$$A_4 = \{Y \geq 0\} = \{Y \in [0, \infty)\}$$
$$A_5 = \{-1 \leq Y < 1\} = \{Y \in [-1, 1)\}.$$

The **probability function** P assigns values between 0 and 1 to events. It is natural to assign the following probabilities for a fair coin toss:

$$P(A_1) = P(Y = 0) = 0.5, \quad P(A_2) = P(Y = 1) = 0.5$$

By definition, the coin variable will never take the value 2.5, so we assign

$$P(A_3) = P(Y = 2.5) = 0.$$

For each intervals, we check whether the events $\{Y = 0\}$ and/or $\{Y = 1\}$ are subsets of the event of interest. If both $\{Y = 0\}$ and $\{Y = 1\}$ are contained in the event, the probability is 1. If only one of them is contained, the probability is 0.5. If neither is contained, the probability is 0.

$$P(A_4) = P(Y \geq 0) = 1, \quad P(A_5) = P(-1 \leq Y < 1) = 0.5.$$

Every event has a **complementary event**, and for any pair of events we can take the **union** and **intersection**. Let's define further events:

- Complements:

$$A_6 = A_4^c = \{Y \geq 0\}^c = \{Y < 0\} = \{Y \in (-\infty, 0)\},$$

- Unions:

$$A_7 = A_1 \cup A_6 = \{Y = 0\} \cup \{Y < 0\} = \{Y \leq 0\}$$

- Intersections:

$$A_8 = A_4 \cap A_5 = \{Y \geq 0\} \cap \{-1 \leq Y < 1\} = \{0 \leq Y < 1\}$$

- Iterations of it:

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

- Certain event:

$$A_{10} = A_9 \cup A_9^c = \{Y \in (-\infty, \infty)\} = \{Y \in \mathbb{R}\}$$

- Empty event:

$$A_{11} = A_{10}^c = \{Y \notin \mathbb{R}\} = \{\}$$

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

4.4 Probability function

The Borel sigma algebra $\mathcal{B}$ is the collection of all events to which we assign probabilities. The events $A_1, ..., A_{11}$ mentioned earlier are elements of $\mathcal{B}$. Any event of the form $\{Y \in (a, b)\}$, where $a, b \in \mathbb{R}$, is also an element of $\mathcal{B}$. Furthermore, all possible unions, intersections, and complements of these events are contained in $\mathcal{B}$. In essence, $\mathcal{B}$ can be thought of as the comprehensive collection of all events for which we would ever compute probabilities in practice.

The following mathematical axioms ensure that the concept of probability is well-defined and possesses the desired properties:

Probability function

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

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

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

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

The axioms of probability imply the following rules of calculation:

Basic rules of probability

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

4.5 Distribution function

Assigning probabilities to events is straightforward for binary variables, like coin tosses. For instance, knowing that $P(Y = 1) = 0.5$ allows us to derive the probabilities for all events in $\mathcal{B}$. However, for more complex variables, such as *education* or *wage*, defining probabilities for all possible events becomes more challenging due to the vast number of potential set operations involved.

Fortunately, it turns out that knowing the probabilities of events of the form $\{Y \leq a\}$ is enough to determine the probabilities of all other events. These probabilities are summarized in the cumulative distribution function.

Cumulative distribution function (CDF)

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

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

The CDF is sometimes referred to as the **distribution function**, or simply the **distribution**. The distribution defines the probabilities for all possible events in $\mathcal{B}$.

The CDF of the variable *coin* is

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

with the following CDF plot:

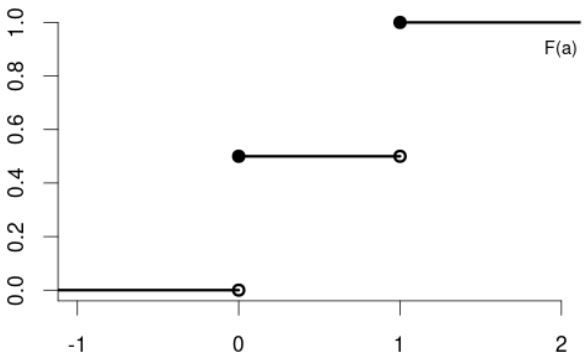


Figure 4.2: CDF of coin

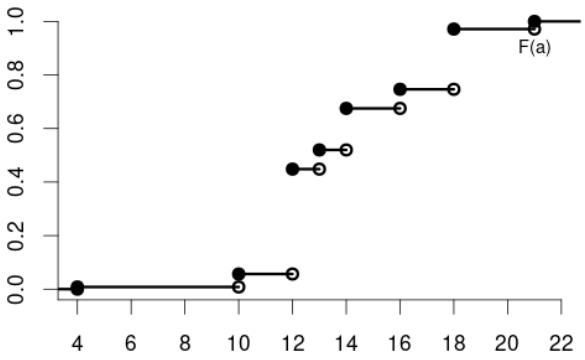


Figure 4.3: CDF of education

The CDF of the variable *education* may be

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

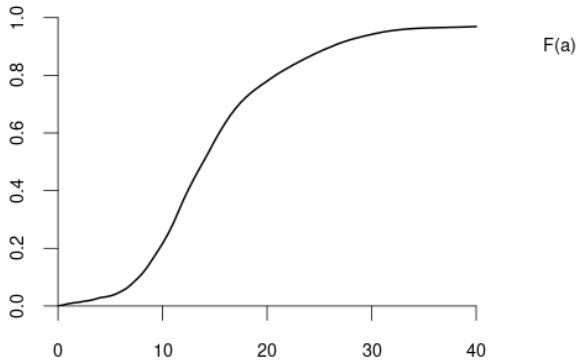


Figure 4.4: CDF of wage

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

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

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

A probability of the form $P(Y = a)$, which involves only an elementary event, is called a **point probability**.

4.6 Point probabilities

The CDF of a **continuous random variable** is smooth, while the CDF of a **discrete random variable** contains jumps and is flat between jumps. For example, variables like *coin* and *education* are discrete, whereas *wage* is continuous.

The **point probability** $P(Y = a)$ represents the size of the jump at $a \in \mathbb{R}$ in the CDF $F(a)$:

$$P(Y = a) = F(a) - \lim_{\epsilon \to 0} F(a - \epsilon),$$

which is the jump height at a. Since continuous variables have no jumps in their CDF, all point probabilities for such variables are zero. The total probability of continuous random variables is spread continuously over an interval, so the probability of the variable being exactly equal to any specific value is zero. Positive probabilities are assigned to intervals.

Basic rules for **continuous random variables** (with $a < b$):

- $P(Y = a) = 0$
- $P(Y \leq a) = P(Y < a) = F(a)$
- $P(Y > a) = P(Y \geq a) = 1 - F(a)$
- $P(a < Y \leq b) = P(a < Y < b) = F(b) - F(a)$
- $P(a \leq Y \leq b) = P(a \leq Y < b) = F(b) - F(a)$

Discrete random variables, unlike continuous ones, have non-zero probabilities at individual points. We summarize the CDF jump heights or point probabilities in the probability mass function:

Probability mass function (PMF)

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

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

The PMF of the *coin* variable is

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

The *education* variable may have the following PMF:

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

4.7 Bivariate distributions

A **bivariate random variable** is a vector of two univariate random variables, e.g., (Y, Z), where Y is *wage* and Z is *experience*.

Bivariate distribution

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

$$\begin{aligned} F_{YZ}(a, b) &= P(Y \leq a, Z \leq b) \\ &= P(\{Y \leq a\} \cap \{Z \leq b\}) \end{aligned}$$

Probabilities can be calculated using a bivariate distribution function in the following way:

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

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

Marginal distributions

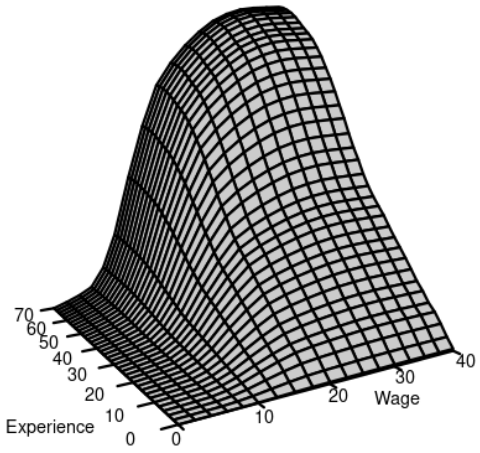


Figure 4.5: Joint CDF of wage and experience

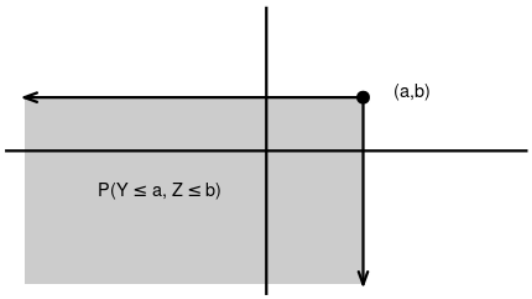


Figure 4.6: Calculate probabilities using the joint CDF

The **marginal distributions** of Y and Z are

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

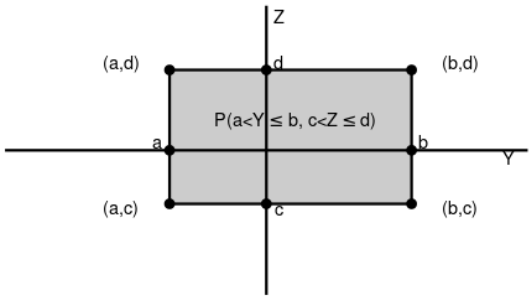


Figure 4.7: Calculate probabilities using the joint CDF

and

$$\begin{aligned} F_Z(b) &= P(Z \leq b) \\ &= P(Y < \infty, Z \leq b) \\ &= \lim_{a \to \infty} F_{YZ}(a, b). \end{aligned}$$

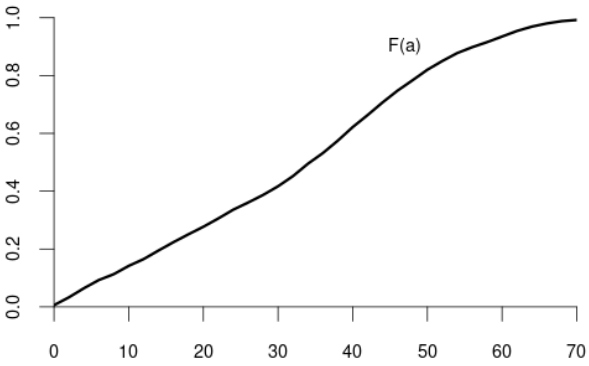


Figure 4.8: Marginal CDF of experience

While the above example shows a bivariate random variable containing two continuous random variables, we can also study discrete variables: Consider, for instance, the coin toss variable Y with $P(Y = 1) = 0.5$ and $P(Y = 0) = 0.5$, and let Z be a second coin toss with the same probabilities. $X = (Y, Z)$ is a bivariate random variable where both entries are discrete random variables.

Since the two coin tosses are performed separately from each other, it is reasonable to assume that the probability that the first and second coin tosses show "heads" is 0.25, i.e., $P(\{Y =$

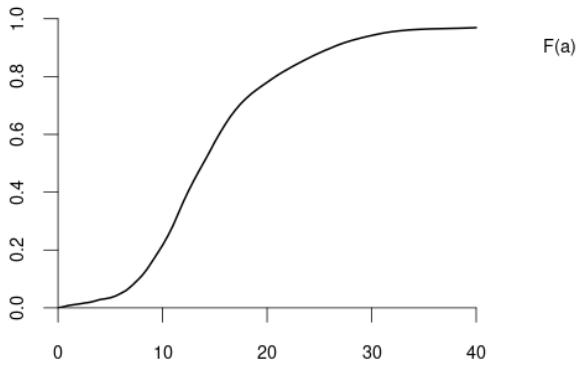


Figure 4.9: Marginal CDF of wage

$1\} \cap \{Z = 1\}) = 0.25$. We would expect the following joint probabilities:

Table 4.2: Joint probabilities of coin tosses

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

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

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

The joint CDF is:

$$F_{YZ}(a,b) = \begin{cases} 0 & \text{if } a < 0 \text{ or } b < 0, \\ 0.25 & \text{if } 0 \le a < 1 \text{ and } 0 \le b < 1, \\ 0.5 & \text{if } 0 \le a < 1 \text{ and } b \ge 1, \\ 0.5 & \text{if } a \ge 1 \text{ and } 0 \le b < 1, \\ 1 & \text{if } a \ge 1 \text{ and } b \ge 1. \end{cases}$$

The marginal CDF of Y is:

$$F_Y(a) = \begin{cases} 0 & \text{if } a < 0, \\ 0.5 & \text{if } 0 \le a < 1, \\ 1 & \text{if } a \ge 1. \end{cases}$$

The marginal CDF of Z is:

$$F_Z(b) = \begin{cases} 0 & \text{if } b < 0, \\ 0.5 & \text{if } 0 \leq b < 1, \\ 1 & \text{if } b \geq 1. \end{cases}$$

Another example are the random variables Y, a dummy variable for the event that the person has a high wage (more than 25 USD/hour), and Z, a dummy variable for the event that the same person has a university degree.

Similarly, $X = (Y, Z)$ is a bivariate random variable consisting of two univariate Bernoulli variables. The joint probabilities might be as follows:

Table 4.3: Joint probabilities of wage and education dummies

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

The joint probability mass function is

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

The joint CDF is:

$$F_{YZ}(a, b) = \begin{cases} 0 & \text{if } a < 0 \text{ or } b < 0, \\ 0.52 & \text{if } 0 \leq a < 1 \text{ and } 0 \leq b < 1, \\ 0.69 & \text{if } 0 \leq a < 1 \text{ and } b \geq 1, \\ 0.64 & \text{if } a \geq 1 \text{ and } 0 \leq b < 1, \\ 1 & \text{if } a \geq 1 \text{ and } b \geq 1. \end{cases}$$

The marginal CDF of Y is:

$$F_Y(a) = \begin{cases} 0 & \text{if } a < 0, \\ 0.69 & \text{if } 0 \leq a < 1, \\ 1 & \text{if } a \geq 1. \end{cases}$$

The marginal CDF of Z is:

$$F_Z(b) = \begin{cases} 0 & \text{if } b < 0, \\ 0.64 & \text{if } 0 \leq b < 1, \\ 1 & \text{if } b \geq 1. \end{cases}$$

4.8 Independence

Two events A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

For instance, in the bivariate random variable of Table 4.2 (two coin tosses), we have

$$\begin{aligned} P(Y = 1, Z = 1) &= 0.25 \\ &= 0.5 \cdot 0.5 \\ &= P(Y = 1)P(Z = 1). \end{aligned}$$

Hence, $\{Y = 1\}$ and $\{Z = 1\}$ are independent events. In the bivariate random variable of Table 4.3 (wage/education), we find

$$\begin{aligned} P(Y = 1, Z = 1) &= 0.19 \\ &\neq P(Y = 1)P(Z = 1) \\ &= 0.31 \cdot 0.36 \\ &= 0.1116. \end{aligned}$$

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

Independence

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

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

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

The random variables Y and Z of Table 4.2 are independent, and those of Table 4.3 are dependent.

4.9 Multivariate distributions

In statistics, we typically study multiple random variables simultaneously. We can collect n random variables $Z_1, \dots, Z_n$ in a $n \times 1$ **random vector**

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = (Z_1, \dots, Z_n)'.$$

We also call $\boldsymbol{Z}$ a n**-variate random variable**.

For example, $Z_1, \dots, Z_n$ could represent n repeated coin tosses or the wage levels of the first n individuals interviewed.

Since $\boldsymbol{Z}$ is a random vector, its outcome is also a vector, e.g., $\{\boldsymbol{Z} = \boldsymbol{b}\}$ with $\boldsymbol{b} = (b_1, \dots, b_n)' \in \mathbb{R}^n$. Events of the form $\{\boldsymbol{Z} \leq \boldsymbol{b}\}$ mean that each component of the random vector $\boldsymbol{Z}$ is smaller than the corresponding values of the vector $\boldsymbol{b}$, i.e.

$$\{\boldsymbol{Z} \leq \boldsymbol{b}\} = \{Z_1 \leq b_1, \dots, Z_n \leq b_n\}.$$

The concepts of the CDF and independence can be generalized to any n-variate random vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)'$. The joint CDF of $\boldsymbol{Z}$ is

$$\begin{aligned} F_{\boldsymbol{Z}}(\boldsymbol{b}) &= P(Z_1 \leq b_1, \dots, Z_n \leq b_n) \\ &= P(\{Z_1 \leq b_1\} \cap \dots \cap \{Z_n \leq b_n\}). \end{aligned}$$

$\boldsymbol{Z}$ has mutually independent entries if

$$F_{\boldsymbol{Z}}(\boldsymbol{b}) = \prod_{i=1}^{n} F_{Z_i}(b_i).$$

That is,

$$P(Z_1 \leq b_1, \dots, Z_n \leq b_n) = P(Z_1 \leq b_1) \cdot \dots \cdot P(Z_n \leq b_n).$$

4.10 IID sampling

In statistical analysis, a dataset $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ that is drawn from some population F is called **sample**.

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

i.i.d. sample / random sample

A collection of random vectors $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ is **i.i.d. (independent and identically distributed)** if they are mutually independent and have the same distribution function F for all $i \neq j$.

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

Any transformed sample $\{g(\boldsymbol{X}_1), \ldots, g(\boldsymbol{X}_n)\}$ of an i.i.d. sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ is also an i.i.d. sample (g may be any function). For instance, if the wages of n interviewed individuals are i.i.d., then the log-wages are also i.i.d.

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

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

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

- **Stratified sampling**
  The population is first divided into homogenous subpopulations (strata), and a random sample is obtained from each stratum independently. Examples: divide companies into industry strata (manufacturing, technology, agriculture, etc.) and sample from each

stratum; divide the population into income strata (low-income, middle-income, high-income).
The sample is independent within each stratum, but it is not between different strata. The strata are defined based on specific characteristics that may be correlated with the variables collected in the sample.

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

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

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

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

4.11 R-codes

statistics-sec04.R

5 Expectation

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

5.1 Discrete random variables

Recall that a discrete random variable Y is a variable that can take on a countable number of distinct values. Each possible value a has an associated probability $\pi(a) = P(Y = a)$, known as the probability mass function (PMF).

The support $\mathcal{Y}$ of Y is the set of all values that Y can take with non-zero probability:

$$\mathcal{Y} = \{a \in \mathbb{R} : \pi(a) > 0\}.$$

The total probability sums to 1: $\sum_{a \in \mathcal{Y}} \pi(a) = 1$.

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

$$E[Y] = \sum_{u \in \mathcal{Y}} u\pi(u). \tag{5.1}$$

The expected value of the variable *education* from the previous section is calculated by summing over all possible values:

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

A **binary** or **Bernoulli** random variable Y takes on only two possible values: 0 and 1. The support is $\mathcal{Y} = \{0, 1\}$. The probabilities are

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

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

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

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

5.2 Continuous random variables

For discrete random variables, both the PMF and the CDF characterize the distribution. For continuous random variables, the PMF concept does not apply because the probability of any specific point is zero. The continuous counterpart of the PMF is the density function:

Probability density function

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

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

If the CDF is differentiable, the density $f(a)$ is its derivative:

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

Properties of a PDF:

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

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

Probability rule for the PDF:

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

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

$$E[Y] = \int_{-\infty}^{\infty} u f(u) \, \mathrm{d}u. \tag{5.2}$$

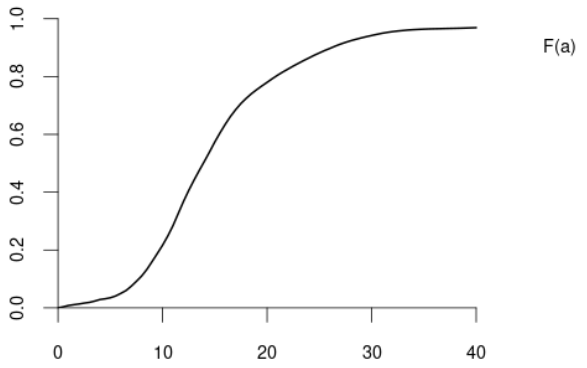


Figure 5.1: CDF of wage

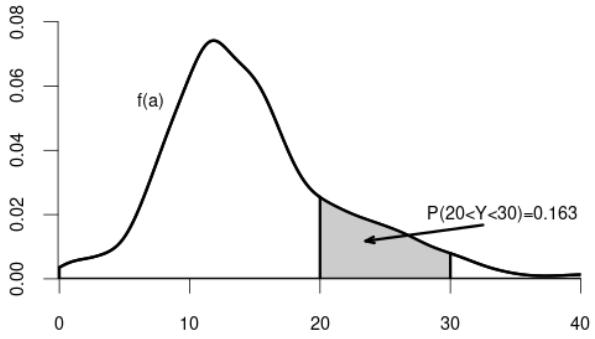


Figure 5.2: PDF of wage

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

$$f(u) = \begin{cases} 1 & \text{if } u \in [0, 1], \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

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

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

5.3 Unified definition of the expected value

The expected value of a random variable Y can be defined in a unified way that applies to both discrete and continuous cases by using its CDF $F(u)$:

$$E[Y] = \int_{-\infty}^{\infty} u\ \mathrm{d}F(u). \tag{5.4}$$

This integral, known as the **Riemann-Stieltjes integral**, generalizes the concept of integration to include functions that may not be smooth or differentiable everywhere.

For a continuous random variable with PDF $f(u)$, the CDF $F(u)$ is smooth and differentiable. The relationship between the CDF and the PDF is:

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

Substituting this into our unified definition gives:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} u\ \mathrm{d}F(u) \\ &= \int_{-\infty}^{\infty} uf(u)\ \mathrm{d}u, \end{aligned}$$

which matches the standard definition of the expected value for continuous random variables as in Equation 5.2.

For a discrete random variable, the CDF $F(u)$ is a step function that increases in jumps at the possible values $u \in \mathcal{Y}$ that Y can take. The "change" or jump in the CDF at each $u \in \mathcal{Y}$ is:

$$\Delta F(u) = F(u) - F(u^-) = P(Y = u) = \pi(u),$$

where $F(u^-)$ is the value of $F(u)$ just before u, and $\pi(u)$ is the PMF of Y.

Integrating with respect to $F(u)$ simplifies to summing over these jumps:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} u\ \mathrm{d}F(u) \\ &= \sum_{u \in \mathcal{Y}} u\ \Delta F(u) \\ &= \sum_{u \in \mathcal{Y}} u\pi(u), \end{aligned}$$

which aligns with the standard definition of the expected value for discrete random variables as in Equation 5.1.

The unified definition $E[Y] = \int_{-\infty}^{\infty} u\ \mathrm{d}F(u)$ allows us to treat all types of random variables consistently, whether the variable is discrete, continuous, or a mixture of both. It can also handle non-standard cases such as distributions with CDFs that are not differentiable everywhere.

5.4 Transformed variables

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

is

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

where $F(u)$ is the CDF of Y. As discussed in Section 5.3 for the different cases, $\mathrm{d}F(u)$ can be replaced by the PMF or the PDF, i.e.,

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

For instance, if we take the *coin* variable Y and consider the transformed random variable $\log(Y+1)$, the expected value is

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

We can define the population counterparts of the sample moments and their centralized and standardized versions:

- **r-th moment** of Y:
$$E[Y^r] = \int_{-\infty}^{\infty} u^r\ \mathrm{d}F(u)$$
- **r-th central moment**:
$$E[(Y - E[Y])^r] = \int_{-\infty}^{\infty} (u - E[Y])^r\ \mathrm{d}F(u)$$
- **Variance** (2nd central moment):
$$Var[Y] = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty} (u - E[Y])^2\ \mathrm{d}F(u)$$
- **Standard deviation**:
$$sd(Y) = \sqrt{Var[Y]}$$
- **r-th standardized moment**:
$$E\left[\left(\frac{Y - E[Y]}{sd(Y)}\right)^r\right] = \int_{-\infty}^{\infty} \left(\frac{u - E[Y]}{sd(Y)}\right)^r\ \mathrm{d}F(u)$$
- **Skewness** (3rd standardized moment):
$$skew(Y) = E\left[\left(\frac{Y - E[Y]}{sd(Y)}\right)^3\right]$$
- **Kurtosis** (4th standardized moment):
$$kurt(Y) = E\left[\left(\frac{Y - E[Y]}{sd(Y)}\right)^4\right]$$

5.5 Linearity of the expected value

The expected value is a **linear** function. For any $a, b \in \mathbb{R}$, we have

$$E[aY + b] = aE[Y] + b.$$

For the variance, the following rule applies:

$$Var[aY + b] = a^2 Var[Y].$$

For any two random variables Y and Z, we have

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

A similar result for the variance does not hold in general. However, if Y and Z are independent random variables, we have

$$Var[aY + bZ] = a^2 Var[Y] + b^2 Var[Z]. \tag{5.5}$$

5.6 Parameters and estimators

A **parameter** θ is a feature (function) of the population distribution F of some random variable Y. The expectation, variance, skewness, and kurtosis are parameters.

A **statistic** is a function of a sample $Y_1, ..., Y_n$. An **estimator** $\hat{\theta}$ for θ is a statistic intended as a guess about θ. It is a function of the random variables $Y_1, ..., Y_n$ and, therefore, a random variable as well. The sample mean, sample variance, sample skewness and sample kurtosis are estimators. When an estimator $\hat{\theta}$ is calculated in a specific realized sample, we call $\hat{\theta}$ an **estimate**.

5.7 Estimation of the mean

The expected value $E[Y]$ is also called **population mean** because it is the population counterpart of the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, where the sample $Y_1, ..., Y_n$ is identically distributed and has the same distribution as Y. In particular, we have:

$$E[Y_1] = ... = E[Y_n] = E[Y].$$

The true population mean $E[Y]$ is unknown in practice, but we can use the sample mean $\overline{Y}$ to estimate it. The sample mean is an unbiased estimator for the population mean because

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

The **bias** of an estimator is the expected value of the estimator minus the parameter to be estimated. The bias of the sample mean is zero:

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

When repeating random experiments and computing sample means, we can expect the sample means to be distributed around the true population mean, with the population mean at the center of this distribution.

To assess how large the spread around the true population mean is, we can compute the variance:

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

To simplify this term further, let's assume that the sample is i.i.d. (independent and identically distributed), i.e. the observations are randomly sampled from the population. Then, we can apply Equation 5.5:

$$Var\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} Var[Y_i].$$

By the identical distribution of the sample, we have

$$Var[Y_1] = ... = Var[Y_n] = Var[Y].$$

Therefore, the variance of the sample mean becomes:

$$Var[\overline{Y}] = \frac{1}{n^2} \sum_{i=1}^{n} Var[Y_i] = \frac{1}{n^2} \sum_{i=1}^{n} Var[Y] = \frac{Var[Y]}{n}.$$

The spread of sample means around the true mean becomes smaller, the larger the sample size n is. The more observations we have, the more precisely the sample mean can estimate the true population mean.

5.8 Consistency

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

Consistency

An estimator $\hat{\theta}$ is **consistent** for a true parameter θ if, for any $\epsilon > 0$,

$$P(|\hat{\theta} - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

Equivalently, consistency can be defined by the complementary event:

$$P(|\hat{\theta} - \theta| \leq \epsilon) \to 1 \qquad \text{as } n \to \infty.$$

If $\hat{\theta}$ is consistent, we say it **converges in probability** to θ, denoted by

$$\hat{\theta} \xrightarrow{p} \theta \qquad \text{as } n \to \infty.$$

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

However, the larger the sample size, the higher should be the probability that $\hat{\theta}$ is close to the true value θ. Consistency means that, if we fix some small precision value $\epsilon > 0$, then,

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

should increase in the sample size n and eventually reach 1.

An estimator is called **inconsistent** if it is not consistent. An inconsistent estimator is practically useless and leads to false inference. Therefore, it is important to verify that your estimator is consistent.

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

Sufficient condition for consistency

Let $\hat{\theta}$ be an estimator for some parameter θ. The **bias** of $\hat{\theta}$ is

$$Bias[\hat{\theta}] = E[\hat{\theta}] - \theta.$$

If the **bias** and the **variance** of $\hat{\theta}$ tends to zero for large sample sizes, i.e., if

i) $Bias[\hat{\theta}] \to 0$ (as $n \to \infty$),
ii) $Var[\hat{\theta}] \to 0$ (as $n \to \infty$),

then $\hat{\theta}$ is consistent for θ.

The reason for this sufficient condition is the fact that

$$P(|\hat{\theta} - \theta| > \epsilon) \leq Var[\hat{\theta}] + Bias[\hat{\theta}]^2,$$

which follows from Markov's inequality.

5.9 Law of large numbers

The sample mean $\overline{Y}$ of an i.i.d. sample is consistent for the population mean $E[Y]$ because

i) $Bias[\overline{Y}] = 0$ for all n;
ii) $Var[\overline{Y}] = Var[Y]/n \to 0$, as $n \to \infty$, provided $Var[Y] < \infty$.

The consistency result of the sample mean is also known as the **law of large numbers (LLN)**:

$$\overline{Y} \xrightarrow{p} E[Y] \qquad \text{as } n \to \infty.$$

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

5.10 Heavy tails

The sample mean of i.i.d. samples from most distributions is consistent. However, there are some exceptional cases where consistency fails. For instance, the simple Pareto distribution has the PDF

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

and the expected value is

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

The population mean is infinity, so the sample mean cannot converge and is inconsistent. The game of chance from the St. Petersburg paradox (see https://en.wikipedia.org/wiki/St._Petersburg_paradox) is an example of a discrete random variable with infinite expectation.

Another example is the t-distribution with 1 degree of freedom, also denoted as t_1 or Cauchy distribution, which has the PDF

$$f(u) = \frac{1}{\pi(1+u^2)}.$$

The lack of consistency of the sample mean from a t_1 distribution is visualized in the shiny application above.

The Pareto, St. Petersburg, and Cauchy distributions have infinite population mean, and the sample mean of observations from these distributions is inconsistent. These are distributions that produce huge outliers.

There are other distributions that have a finite mean but an infinite variance, skewness, or kurtosis.

For instance, the t_2 distribution has a finite mean but an infinite variance. The t_3 distribution has a finite variance but an infinite skewness. The t_4 distribution has a finite skewness but an infinite kurtosis.

If Y is t_m-distributed (t-distribution with m degrees of freedom), then

$$E[Y], E[Y^2], ..., E[Y^{m-1}] < \infty$$

but

$$E[Y^m] = E[Y^{m+1}] = ... = \infty.$$

Random variables with infinite first four moments have a so-called **heavy-tailed distribution** and may produce huge outliers. Many statistical procedures are only valid if the underlying distribution is not heavy-tailed.

5.11 Estimation of the variance

Consider an i.i.d. sample $Y_1, ..., Y_n$ from some population distribution with population mean $\mu = E[Y]$ and population variance $\sigma^2 = Var[Y] < \infty$.

We introduced two sample cointerparts of σ^2: the sample variance

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

and the adjusted sample variance

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

The sample variance can be decomposed as

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

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

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

where we used the fact that $Var[\overline{Y}] = \sigma^2/n$.

The sample variance is **downward biased**:

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

On the other hand, the adjusted sample variance is **unbiased**:

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

The variance of the sample variance can be computed as

$$Var[\hat{\sigma}_Y^2] = \frac{\sigma^4}{n}\Big(kurt - \frac{n-3}{n-1}\Big)\frac{(n-1)^2}{n^2},$$

while the variance of the adjusted sample variance is

$$Var[s_Y^2] = \frac{\sigma^4}{n}\Big(kurt - \frac{n-3}{n-1}\Big).$$

As long as the kurtosis of the underlying distribution is finite, the sufficient conditions for consistency are satisfied as the bias and variance tend to zero as $n \to \infty$. The adjusted sample variance is unbiased for any n. The sample variance is biased for fixed n but **asymptotically unbiased** as the bias tends to zero for large n. The sample variance and the adjusted sample variance are consistent for the variance if the sample is i.i.d. and the distribution is not heavy-tailed.

5.12 Bias-variance tradeoff

From a bias perspective, adjusted sample variance s_Y^2 is preferred over $\hat{\sigma}_Y^2$ because s_Y^2 is unbiased. However, from a variance perspective, $\hat{\sigma}_Y^2$ is preferred due to its smaller variance. Traditionally, the emphasis on unbiasedness has led to a preference for $\hat{\sigma}_Y^2$, even at the cost of a higher variance.

A more modern approach balances bias and variance, known as the **bias-variance tradeoff**, by selecting an estimator that minimizes the **mean squared error (MSE)**:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var[\hat{\theta}] + Bias[\hat{\theta}]^2.$$

For the variance estimators, the MSEs are

$$MSE[\hat{\sigma}_Y^2] = Var[\hat{\sigma}_Y^2] + Bias[\hat{\sigma}_Y^2]^2 = \frac{\sigma^4}{n}\left[\left(kurt - \frac{n-3}{n-1}\right)\frac{(n-1)^2}{n^2} + \frac{1}{n}\right]$$

and

$$MSE[s_Y^2] = Var[s_Y^2] = \frac{\sigma^4}{n}\left(kurt - \frac{n-3}{n-1}\right).$$

Since s_Y^2 is unbiased, its MSE equals its variance.

It is not possible to universally determine which estimator has a lower MSE because this depends on the population kurtosis ($kurt$) of the underlying distribution. However, it can be shown that for all distributions with $kurt \geq 1.5$, the relation $MSE[s_Y^2] > MSE[\hat{\sigma}_Y^2]$ holds, which implies that $\hat{\sigma}_Y^2$ is preferred based on the bias-variance tradeoff for all moderately tailed distributions.

To give an indication of typical kurtosis values:

- Symmetric Bernoulli distribution with $P(Y = 0) = P(Y = 1) = 0.5$: kurtosis of 1 (light-tailed).
- Uniform distribution (see Equation 5.3): kurtosis of 1.8 (moderately light-tailed).
- Normal distribution: kurtosis of 3 (moderately tailed).
- t_5 distribution: kurtosis of 9 (moderately heavy-tailed).
- t_4 distribution: infinite kurtosis (heavy-tailed).

Therefore, according to the bias-variance tradeoff, the adjusted sample variance s_Y^2 is preferred only for extremely light-tailed distributions, while $\hat{\sigma}_Y^2$ is preferred in cases with moderate or higher kurtosis.

In practice, especially with larger samples, the difference between s_Y^2 and $\hat{\sigma}_Y^2$ becomes negligible, and either estimator is generally acceptable. Therefore, the discussion about a better variance estimator is a bit nitpicky and not of much practical relevance.

However, for instance in high-dimensional regression problems with near multicollinearity ($k \approx n$), the bias-variance tradeoff is crucial. In such cases, biased but low-variance estimators like ridge or lasso (shrinkage estimators) are often preferred over ordinary least squares (OLS).

5.13 R-codes

statistics-sec05.R

6 Covariance

6.1 Expectation of bivariate random variables

We often are interested in expected values of functions involving two random variables, such as the **cross-moment** $E[YZ]$ for variables Y and Z.

If $F(a, b)$ is the joint CDF of (Y, Z), then the cross-moment is defined as:

$$E[YZ] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab \ \mathrm{d}F(a, b). \tag{6.1}$$

If Y and Z are continuous and $F(a, b)$ is differentiable, the joint probability density function (PDF) of (Y, Z):

$$f(a, b) = \frac{\partial^2}{\partial a \partial b} F(a, b).$$

This allows us to write the differential of the CDF as

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

and the cross-moment becomes:

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

In the *wage* and *experience* example, we have the following joint CDF and joint PDF:

If Y and Z are discrete with joint PMF $\pi(a, b)$ and support $\mathcal{Y}$, the cross moment is

$$E[YZ] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ab \ \mathrm{d}F(a, b) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{Y}} ab \ \pi(a, b).$$

If one variable is discrete and the other is continuous, the expectation involves a mixture of summation and integration.

In general, the expected value of any real valued function $g(Y, Z)$ is given by

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

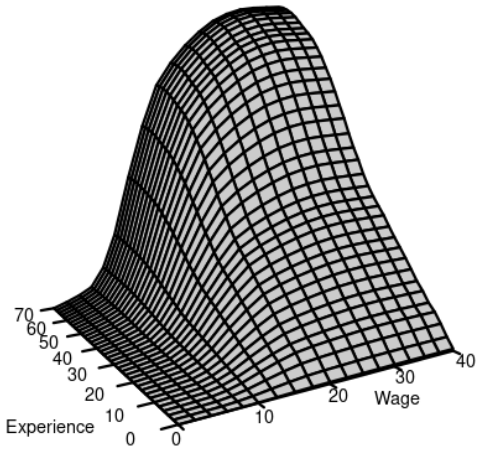


Figure 6.1: Joint CDF of wage and experience

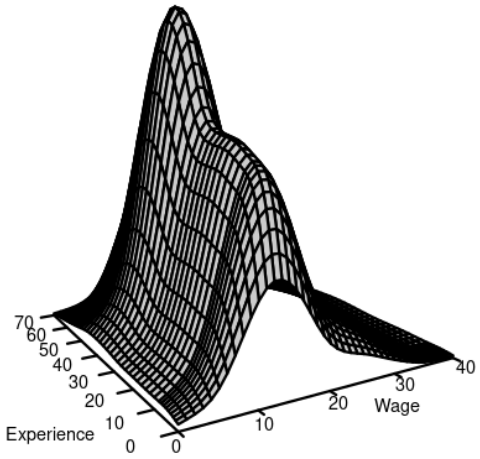


Figure 6.2: Joint PDF of wage and experience

6.2 Covariance and correlation

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

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

The covariance of Y with itself is the variance:

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

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

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

The **correlation** of Y and Z is

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

where $sd(Y)$ and $sd(Z)$ are the standard deviations of Y and Z, respectively.

Uncorrelated

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

If Y and Z are uncorrelated, then:

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

If Y and Z are independent and have finite second moments, they are uncorrelated. However, the reverse is not necessarily true; uncorrelated variables are not always independent.

6.3 Expectations for random vectors

These concepts generalize to any k-dimensional random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$.

The expectation vector of $\boldsymbol{Z}$ is:

$$E[\boldsymbol{Z}] = \begin{pmatrix} E[Z_1] \\ \vdots \\ E[Z_k] \end{pmatrix}.$$

The covariance matrix of $\boldsymbol{Z}$ is:

$$\begin{aligned} Var[\boldsymbol{Z}] &= E[(\boldsymbol{Z} - E[\boldsymbol{Z}])(\boldsymbol{Z} - E[\boldsymbol{Z}])'] \\ &= \begin{pmatrix} Var[Z_1] & Cov(Z_1, Z_2) & ... & Cov(X_1, Z_k) \\ Cov(Z_2, Z_1) & Var[Z_2] & ... & Cov(Z_2, Z_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Z_k, Z_1) & Cov(Z_k, Z_2) & ... & Var[Z_k] \end{pmatrix} \end{aligned}$$

For any random vector $\boldsymbol{Z}$, the covariance matrix $Var[\boldsymbol{Z}]$ is symmetric and positive semi-definite.

6.4 Population regression

Consider the dependent variable Y_i and the regressor vector $\boldsymbol{X}_i = (1, X_{i2}, ..., X_{ik})'$ for a representative individual i from the population. Assume the linear relationship:

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

where $\boldsymbol{\beta}$ is the vector of population regression coefficients, and u_i is an error term satisfying $E[\boldsymbol{X}_i u_i] = \boldsymbol{0}$.

The error term u_i accounts for factors affecting Y_i that are not included in the model, such as measurement errors, omitted variables, or unobserved/unmeasured variables. We assume all variables have finite second moments, ensuring that all covariances and cross-moments are finite.

To express $\boldsymbol{\beta}$ in terms of population moments, compute:

$$\begin{aligned} E[\boldsymbol{X}_i Y_i] &= E[\boldsymbol{X}_i(\boldsymbol{X}_i'\beta + u_i)] \\ &= E[\boldsymbol{X}_i\boldsymbol{X}_i']\boldsymbol{\beta} + E[\boldsymbol{X}_i u_i]. \end{aligned}$$

Since $E[\boldsymbol{X}_i u_i] = \boldsymbol{0}$, it follows that

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

Assuming $E[\boldsymbol{X}_i\boldsymbol{X}_i']$ is invertible, we solve for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = E[\boldsymbol{X}_i\boldsymbol{X}_i']^{-1}E[\boldsymbol{X}_i Y_i].$$

Applying the method of moments, we estimate $\boldsymbol{\beta}$ by replacing the population moments with their sample counterparts:

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

This estimator $\hat{\boldsymbol{\beta}}$ coincides with the OLS coefficient vector and is known as the OLS estimator or the method of moments estimator for $\boldsymbol{\beta}$.

6.5 R-codes

statistics-sec06.R

7 Conditional expectation

7.1 Conditional distribution

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.

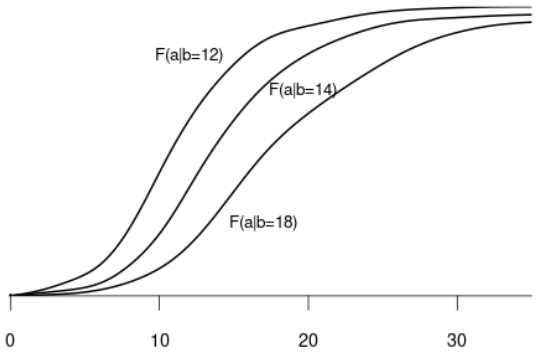


Figure 7.1: Conditional CDFs of wage given education

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

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

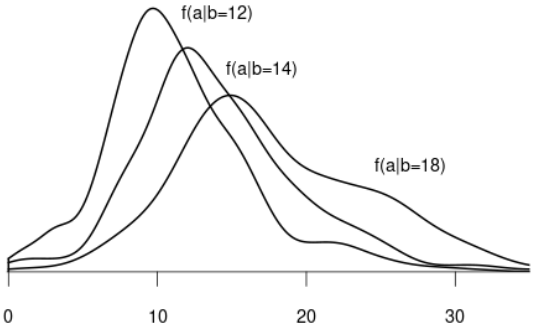


Figure 7.2: Conditional PDFs of wage given education

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

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

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.

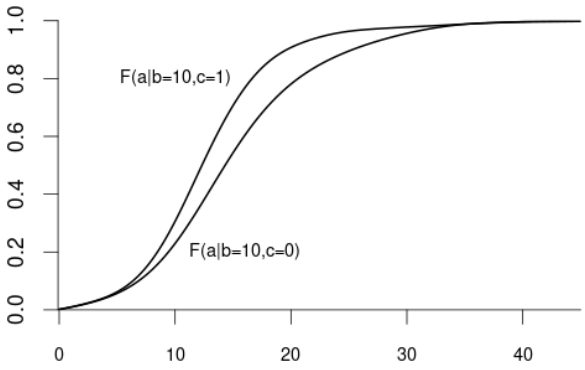


Figure 7.3: Conditional CDFs of wage given experience and gender

Similarly, we can take the derivative to get the conditional density $f_{Y|Z_1=b,Z_2=c}(a)$:

More generally, we can condition on the event that a 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:

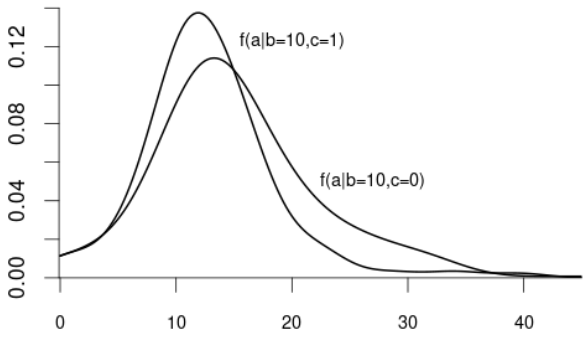


Figure 7.4: Conditional CDFs of wage given experience and gender

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

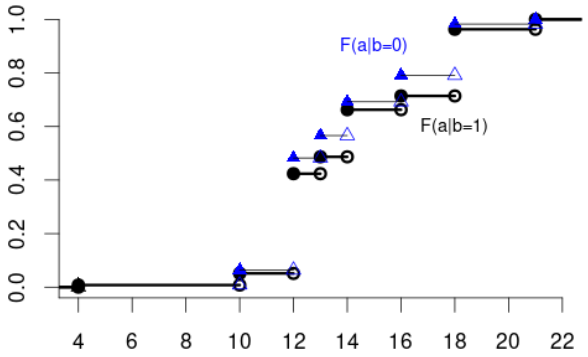


Figure 7.5: Conditional CDFs of education given married

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.

7.1.1 Conditioning on discrete variables

If Z is a discrete random variable, then the conditional CDF can be expressed in terms of conditional probabilities.

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

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

Let's revisit the wage and schooling example from Table 4.3:

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

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

Therefore, the conditional CDF of Y given $\{Z=b\}$ with $P(Z=b)>0$ is:

$$F_{Y|Z=b}(a) = P(Y\le a|Z=b) = \frac{P(Y\le a, Z=b)}{P(Z=b)} = \sum_{u\in\mathcal{Y},u\le a}\frac{\pi_{YZ}(u,b)}{\pi_Z(b)}.$$

7.1.2 Conditioning on continuous variables

If Z is a continuous variable, we have $P(Z=b)=0$ for all b, and $P(Y\le a|Z=b)$ cannot be defined in the same way as for discrete variables.

If $f_{YZ}(a,b)$ is the joint PDF of Y and Z and $f_Z(b)$ is the marginal PDF of Z, the relation of the conditional CDF and the PDFs is as follows:

$$F_{Y|Z=b}(a) = P(Y\le a|Z=b) = \int_{-\infty}^{a}\frac{f_{YZ}(u,b)}{f_Z(b)}\,\mathrm{d}u.$$

7.2 Conditional mean

Conditional expectation

The **conditional expectation** or **conditional mean** of Y given $\boldsymbol{Z}=\boldsymbol{b}$ is the expected value of the distribution $F_{Y|\boldsymbol{Z}=\boldsymbol{b}}$:

$$E[Y|\boldsymbol{Z}=\boldsymbol{b}] = \int_{-\infty}^{\infty} a\,\mathrm{d}F_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a).$$

For continuous Y with conditional density $f_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a)$, we have $\mathrm{d}F_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a) = f_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a)\,\mathrm{d}a$, and the conditional expectation is

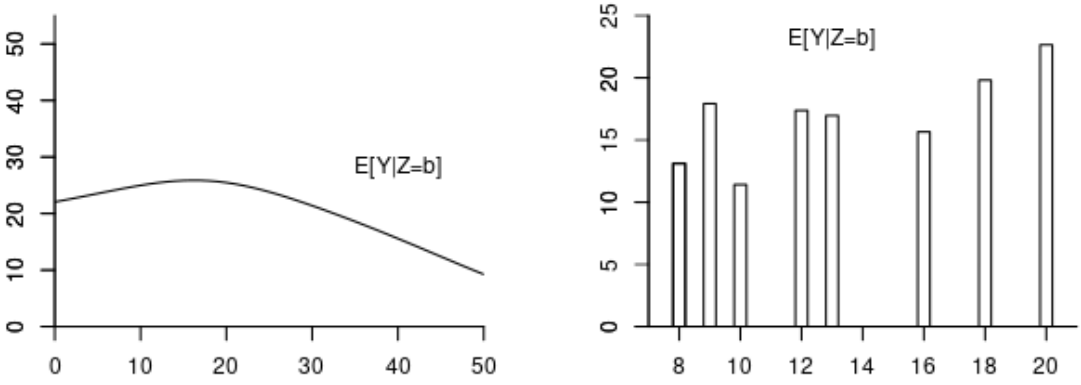
$$E[Y|Z=\boldsymbol{b}] = \int_{-\infty}^{\infty} a f_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a)\,\mathrm{d}a.$$

Similarly, for discrete Y with support $\mathcal{Y}$ and conditional PMF $\pi_{Y|\boldsymbol{Z}=\boldsymbol{b}}(a)$, we have

$$E[Y|Z=\boldsymbol{b}] = \sum_{u\in\mathcal{Y}} u\pi_{Y|\boldsymbol{Z}=\boldsymbol{b}}(u).$$

The conditional expectation is a function of $\boldsymbol{b}$, which is a specific value of $\boldsymbol{Z}$ that we condition on. Therefore, we call it the **conditional expectation function**:

$$m(\boldsymbol{b}) = E[Y|Z=\boldsymbol{b}].$$



(a) CEF wage given experience

(b) CEF wage given education

Figure 7.6: Conditional expectation functions. The x-axis represents b.

Suppose the conditional expectation of wage given experience level b is:

$$m(b) = E[wage|exper = b] = 14.5 + 0.9b - 0.017b^2.$$

For example, with 10 years of experience:

$$m(10) = E[wage|exper = 10] = 21.8.$$

Here, $m(b)$ assigns a specific real number to each fixed value of b; it is a deterministic function derived from the joint distribution of wage and experience.

However, if we treat experience as a random variable, the conditional expectation becomes:

$$m(exper) = E[wage|exper] = 14.5 + 0.9exper - 0.017exper^2.$$

Now, $m(exper)$ is a function of the random variable experexper and is itself a random variable.

In general:

- The conditional expectation given a specific value b is:

$$m(\boldsymbol{b}) = E[Y|\boldsymbol{Z} = \boldsymbol{b}],$$

  which is deterministic.
- The conditional expectation given the random variable Z is:

$$m(\boldsymbol{Z}) = E[Y|\boldsymbol{Z}],$$

  which is a random variable because it depends on the random vector $\boldsymbol{Z}$.

This distinction highlights that the conditional expectation can be either a specific number, i.e. $E[Y|\boldsymbol{Z} = \boldsymbol{b}]$, or a random variable, i.e., $E[Y|\boldsymbol{Z}]$, depending on whether the condition is fixed or random.

7.3 Rules of calculation

Rules of Calculation for Conditional Expectation

Let Y be a random variable and $\boldsymbol{Z}$ a random vector. The rules of calculation rules below are fundamental tools for working with conditional expectations:

(i) Law of Iterated Expectations (LIE):

$$E[E[Y|\boldsymbol{Z}]] = E[Y].$$

Intuition: The LIE tells us that if we first compute the expected value of Y given each possible outcome of $\boldsymbol{Z}$, and then average those expected values over all possible values of $\boldsymbol{Z}$, we end up with the overall expected value of Y. It's like calculating the average outcome across all scenarios by considering each scenario's average separately.

More generally, for any two random vectors $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$:

$$E[E[Y|\boldsymbol{Z}, \boldsymbol{Z}^*]|\boldsymbol{Z}] = E[Y|\boldsymbol{Z}].$$

Intuition: Even if we condition on additional information $\boldsymbol{Z}^*$, averaging over $\boldsymbol{Z}^*$ while keeping $\boldsymbol{Z}$ fixed brings us back to the conditional expectation given $\boldsymbol{Z}$ alone.

(ii) Conditioning Theorem (CT):

For any function $g(\boldsymbol{Z})$:

$$E[g(\boldsymbol{Z})\,Y|\boldsymbol{Z}] = g(\boldsymbol{Z})\,E[Y|\boldsymbol{Z}].$$

Intuition: Once we know $\boldsymbol{Z}$, the function $g(\boldsymbol{Z})$ becomes a known quantity. Therefore, when computing the conditional expectation given $\boldsymbol{Z}$, we can treat $g(\boldsymbol{Z})$ as a constant and factor it out.

(iii) Independence Rule (IR):

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

$$E[Y|\boldsymbol{Z}] = E[Y].$$

Intuition: Independence means that Y and $\boldsymbol{Z}$ do not influence each other. Knowing the value of $\boldsymbol{Z}$ gives us no additional information about Y. Therefore, the expected value of Y remains the same regardless of the value of $\boldsymbol{Z}$, so the conditional expectation equals the unconditional expectation.

Another way to see this is the fact that, if Y and Z are independent, then

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

7.4 Best predictor property

It turns out that the CEF $m(\boldsymbol{Z}) = E[Y|\boldsymbol{Z}]$ is the best predictor for Y given the information contained in the random vector $\boldsymbol{Z}$:

Best predictor

The CEF $m(\boldsymbol{Z}) = E[Y|\boldsymbol{Z}]$ minimizes the expected squared error $E[(Y - g(\boldsymbol{Z}))^2]$ among all predictor functions $g(\boldsymbol{Z})$:

$$m(\boldsymbol{Z}) = \text{argmin}_{g(\boldsymbol{Z})} E[(Y - g(\boldsymbol{Z}))^2]$$

Proof: Let us find the function $g(\cdot)$ that minimizes $E[(Y - g(\boldsymbol{Z}))^2]$:

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

- The first term (i) does not depend on $g(\cdot)$ and is finite if $E[Y^2] < \infty$.
- For the second term (ii), we use the LIE and CT:

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

- The third term (iii) $E[(m(\boldsymbol{Z}) - g(\boldsymbol{Z}))^2]$ is minimal if $g(\cdot) = m(\cdot)$.

Therefore, $m(\boldsymbol{Z}) = E[Y|\boldsymbol{Z}]$ minimizes $E[(Y - g(\boldsymbol{Z}))^2]$.

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

$$u = Y - E[Y|\boldsymbol{Z}].$$

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

$$\begin{aligned} E[u|\boldsymbol{Z}] &= E[(Y - m(\boldsymbol{Z}))|\boldsymbol{Z}] \\ &= E[Y|\boldsymbol{Z}] - E[m(\boldsymbol{Z})|\boldsymbol{Z}] \\ &= m(\boldsymbol{Z}) - m(\boldsymbol{Z}) = 0. \end{aligned}$$

7.5 Linear regression model

Consider again the linear regression framework with dependent variable Y_i and regressor vector $\boldsymbol{X}_i$. The previous section shows that we can always write

$$Y_i = m(\boldsymbol{X}_i) + u_i, \quad E[u_i|\boldsymbol{X}_i] = 0,$$

where $m(\boldsymbol{X}_i)$ is the CEF of Y_i given $\boldsymbol{X}_i$, and u_i is the CEF error.

In the **linear regression model**, we assume that the CEF is linear in $\boldsymbol{X}_i$, i.e.

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

From this equation, by the CT, it becomes clear that

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

Therefore, $\boldsymbol{X}_i'\boldsymbol{\beta}$ is the best predictor for Y_i given $\boldsymbol{X}_i$.

Linear regression model

We assume that $(Y_i, \boldsymbol{X}_i')$ satisfies

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \dots, n, \tag{7.1}$$

with

- (A1) **conditional mean independence**: $E[u_i|\boldsymbol{X}_i] = 0$
- (A2) **random sampling**: $(Y_i, \boldsymbol{X}_i')$ are i.i.d. draws from their joint population distribution
- (A3) **large outliers unlikely**: $0 < E[Y_i^4] < \infty$, $0 < E[X_{il}^4] < \infty$ for all $l = 1, \dots, k$
- (A4) **no perfect multicollinearity**: $\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i'$ is invertible

In matrix notation, the model equation can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{u} = (u_1, \dots, u_n)'$ is the error term vector, $\boldsymbol{Y}$ is the dependent variable vector, and $\boldsymbol{X}$ is the $n \times k$ regressor matrix.

(A1) and (A2) define the structure of the regression model, while (A3) and (A4) ensure that OLS estimation is feasible and reliable.

7.5.1 Conditional mean independence (A1)

Assumption (A1) is fundamental to the regression model and has several key implications:

1) Zero unconditional mean

Using the Law of Iterated Expectations (LIE):

$$E[u_i] \overset{(LIE)}{=} E[E[u_i|\boldsymbol{X}_i]] = E[0] = 0$$

The error term u_i has a zero unconditional mean.

2) Linear best predictor

The conditional mean of Y_i given X_i is:

$$\begin{aligned} E[Y_i|\boldsymbol{X}_i] &= E[\boldsymbol{X}_i'\boldsymbol{\beta} + u_i|\boldsymbol{X}_i] \\ &\overset{(CT)}{=} \boldsymbol{X}_i'\boldsymbol{\beta} + E[u_i|\boldsymbol{X}_i] \\ &= \boldsymbol{X}_i'\boldsymbol{\beta} \end{aligned}$$

The regression function $\boldsymbol{X}_i'\boldsymbol{\beta}$ represents the best linear predictor of Y_i given $\boldsymbol{X}_i$. This means the expected value of Y_i is a linear function of the regressors.

3) Marginal effect interpretation

From the linearity of the conditional expectation:

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

The partial derivative with respect to X_{ij} is:

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

The coefficient β_j represents the marginal effect of a one-unit increase in X_{ij} on the expected value of Y_i, holding all other variables constant.

Note: This marginal effect is not necessarily causal. Unobserved factors correlated with X_{ij} may influence Y_i, so β_j captures both the direct effect of X_{ij} and the indirect effect through these unobserved variables.

4) Weak exogeneity

Using the definition of covariance:

$$Cov(u_i, X_{il}) = E[u_i X_{il}] - E[u_i]E[X_{il}].$$

Since $E[u_i] = 0$:

$$Cov(u_i, X_{il}) = E[u_i X_{il}].$$

Applying the LIE and the CT:

$$\begin{aligned} E[u_i X_{il}] &= E[E[u_i X_{il} | \boldsymbol{X}_i]] \\ &= E[X_{il} E[u_i | \boldsymbol{X}_i]] \\ &= E[X_{il} \cdot 0] = 0 \end{aligned}$$

The error term u_i is uncorrelated with each regressor X_{il}. This property is known as **weak exogeneity**. It indicates that $u_i i$ captures unobserved factors that do not systematically vary with the observed regressors.

Note: Weak exogeneity does not rule out the presence of unobserved variables that affect both Y_i and $\boldsymbol{X}_i$. The coefficient β_j reflects the average relationship between $\boldsymbol{X}_i$ and Y_i, including any indirect effects from unobserved factors that are correlated with $\boldsymbol{X}_i$.

7.5.2 Random sampling (A2)

1) Strict exogeneity

The i.i.d. assumption (A2) implies that $\{(Y_i, \boldsymbol{X}_i', u_i), i = 1, \dots, n\}$ is an i.i.d. collection since $u_i = Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}$ is a function of a random sample, and functions of independent variables are independent as well.

Therefore, u_i and $\boldsymbol{X}_j$ are independent for $i \neq j$. The independence rule (IR) implies $E[u_i | \boldsymbol{X}_1, \dots, \boldsymbol{X}_n] = E[u_i | \boldsymbol{X}_i]$.

The weak exogeneity condition (A1) turns into a **strict exogeneity** property:

$$E[u_i | \boldsymbol{X}] = E[u_i | \boldsymbol{X}_1, \dots, \boldsymbol{X}_n] \overset{(A2)}{=} E[u_i | \boldsymbol{X}_i] \overset{(A1)}{=} 0.$$

Additionally,

$$Cov(u_j, X_{il}) = \underbrace{E[u_j X_{il}]}_{=0} - \underbrace{E[u_j]}_{=0} E[X_{il}] = 0.$$

Weak exogeneity means that the regressors of individual i are uncorrelated with the error term of the same individual i. Strict exogeneity means that the regressors of individual i are uncorrelated with the error terms of any individual j in the sample.

2) Heteroskedasticity

The i.i.d. assumption (A2) is not as restrictive as it may seem at first sight. It allows for dependence between u_i and $\boldsymbol{X}_i = (1, X_{i2}, ..., X_{ik})'$. The error term u_i can have a conditional distribution that depends on $\boldsymbol{X}_i$.

The exogeneity assumption (A1) requires that the conditional mean of u_i is independent of $\boldsymbol{X}_i$. Besides this, dependencies between u_i and $X_{i2}, ..., X_{ik}$ are allowed. For instance, the variance of u_i can be a function of $X_{i2}, ..., X_{ik}$. If this is the case, u_i is said to be **heteroskedastic**.

The **conditional variance** is defined analogously to the unconditional variance:

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

The conditional variance of the error is:

$$Var[u_i|\boldsymbol{X}] = E[u_i^2|\boldsymbol{X}] \overset{(A2)}{=} E[u_i^2|\boldsymbol{X}_i] =: \sigma_i^2 = \sigma^2(\boldsymbol{X}_i).$$

An additional restrictive assumption is **homoskedasticity**, which means that the variance of u_i is not allowed to vary for different values of $\boldsymbol{X}_i$:

$$Var[u_i|\boldsymbol{X}] = \sigma^2.$$

Homoskedastic errors are a restrictive assumption sometimes made for convenience in addition to (A1)+(A2). Homoskedasticity is often unrealistic in practice, so we stick with the heteroskedastic errors framework.

3) No autocorrelation

(A2) implies that u_i is independent of u_j for $i \neq j$, and therefore $E[u_i|u_j, \boldsymbol{X}] = E[u_i|\boldsymbol{X}] = 0$ by the IR. This implies

$$E[u_i u_j|\boldsymbol{X}] \overset{(LIE)}{=} E\big[E[u_i u_j|u_j, \boldsymbol{X}]|\boldsymbol{X}\big] \overset{(CT)}{=} E\big[u_j \underbrace{E[u_i|u_j, \boldsymbol{X}]}_{=0}|\boldsymbol{X}\big] = 0,$$

and, therefore,

$$Cov(u_i, u_j) = E[u_i u_j] \overset{(LIE)}{=} E[E[u_i u_j|\boldsymbol{X}]] = 0.$$

The conditional covariance matrix of the error term vector $\boldsymbol{u}$ is

$$\boldsymbol{D} := Var[\boldsymbol{u}|\boldsymbol{X}] = E[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}] = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix}.$$

It is a diagonal matrix with conditional variances on the main diagonal. We also write $\boldsymbol{D} = diag(\sigma_1^2, ..., \sigma_n^2)$.

7.5.3 Finite moments and invertibility (A3 + A4)

Assuming (A3) excludes frequently occurring large outliers as it rules out heavy-tailed distributions. Hence, we should be careful if we use variables with large kurtosis. Assuming (A4) ensures that the OLS estimator $\hat{\boldsymbol{\beta}}$ can be computed.

7.5.3.1 Unbiasedness

(A4) ensures that $\hat{\boldsymbol{\beta}}$ is well defined. The following decomposition is useful:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}.\end{aligned}$$

The strict exogeneity implies $E[\boldsymbol{u}|\boldsymbol{X}] = \boldsymbol{0}$, and

$$E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X}] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}] \overset{(CT)}{=} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{E[\boldsymbol{u}|\boldsymbol{X}]}_{=\boldsymbol{0}} = \boldsymbol{0}.$$

By the (LIE), $E[\hat{\boldsymbol{\beta}}] = E[E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]] = E[\boldsymbol{\beta}] = \boldsymbol{\beta}$.

Hence, the **OLS estimator is unbiased**: $Bias[\hat{\boldsymbol{\beta}}] = 0$.

7.5.3.2 Conditional variance

Recall the matrix rule $Var[\boldsymbol{A}\boldsymbol{Z}] = \boldsymbol{A}Var[\boldsymbol{Z}]\boldsymbol{A}'$ if $\boldsymbol{Z}$ is a random vector and $\boldsymbol{A}$ is a matrix. Then,

$$\begin{aligned}Var[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= Var[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}] \\ &= Var[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Var[\boldsymbol{u}|\boldsymbol{X}]((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

7.5.3.3 Consistency

The conditional variance can be written as

$$\begin{aligned}Var[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \frac{1}{n}\Big(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\Big)^{-1}\Big(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\Big)\Big(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\Big)^{-1} \\ &= \frac{1}{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\end{aligned}$$

It can be shown, by the multivariate law of large numbers, that $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i' \xrightarrow{p} E[\boldsymbol{X}_i\boldsymbol{X}_i']$ and $\sum_{i=1}^{n} \sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i \xrightarrow{p} E[\sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i']$. For this to hold we need bounded fourth moments, i.e. (A3). In total, we have

$$\begin{aligned}
&\Big(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} \sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i'\Big)^{-1} \\
&\xrightarrow{p} E[\boldsymbol{X}_i\boldsymbol{X}_i']^{-1}E[\sigma_i^2\boldsymbol{X}_i\boldsymbol{X}_i']E[\boldsymbol{X}_i\boldsymbol{X}_i']^{-1}.
\end{aligned}$$

Note that the conditional variance $Var[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ has an additional factor $1/n$, which converges to zero for large n. Therefore, we have

$$Var[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] \xrightarrow{p} \mathbf{0},$$

which also holds for the unconditional variance, i.e. $Var[\hat{\boldsymbol{\beta}}] \to \mathbf{0}$.

Therefore, since the bias is zero and the variance converges to zero, the sufficient conditions for consistency are fulfilled. The OLS estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}$ under (A1)–(A4).

7.6 R-codes

statistics-sec07.R

8 Simulations

8.1 Consistent estimation

Recall the definitions of the bias, variance, and mean squared error (MSE) of an estimator $\hat{\theta}$ for a parameter θ:

- **Bias**: $Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$
- **Variance**: $Var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$
- **MSE**: $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

These quantities are related by the equation:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2.$$

This relationship holds for any estimator and can be derived as follows:

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2] + 2E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] + (E[\hat{\theta}] - \theta)^2 \\ &= Var(\hat{\theta}) + 2(\underbrace{E[\hat{\theta}] - E[\hat{\theta}]}_{=0})(E[\hat{\theta}] - \theta) + Bias(\hat{\theta})^2 \end{aligned}$$

Recall that an estimator is consistent if it gets closer to the true parameter value as we collect more data. In mathematical terms, $\hat{\theta}$ is **consistent** for θ if its MSE tends to zero as the sample size $n \to \infty$. This means both the bias and variance of $\hat{\theta}$ approach zero.

To understand the consistency properties of an estimator $\hat{\theta}$, an alternative to mathematical proofs is to conduct a **Monte Carlo simulation**. These simulations are useful for studying the sampling distribution of a statistic in a controlled environment where the true data-generating population distribution is known. They allow us to compare the biases and MSEs of different estimators for different sample sizes.

While mathematical proofs establish theoretical properties of estimators, Monte Carlo simulations show us how these estimators actually behave with real, finite samples. These simulations let us examine an estimator's performance under different conditions and sample sizes, and help us develop statistical intuition.

The idea is to use computer-generated pseudorandom numbers to create artificial datasets of sample size n. We apply the estimator of interest to each dataset, which generates random draws from the distribution of the estimator. By repeating this procedure independently B times, we obtain an i.i.d. sample of size B from the distribution of the estimator, known as a **Monte Carlo sample**. From this sample, we can compute empirical estimates of quantities like bias, variance, and MSE.

8.2 Set up

To set up the Monte Carlo simulation for $\hat{\theta}$, we need to specify

1. **Estimator ($\hat{\theta}$)**: The estimator of interest.
2. **Population distribution (F)**: The specific distribution from which we sample our data.
3. **Parameter value (θ)**: The particular value of the parameter of F that we aim to estimate.
4. **Sample size (n)**: The number of observations in each simulated dataset.
5. **Sampling scheme**: Typically independent and identically distributed (i.i.d.), but it could also involve dependence (e.g., in time series data).
6. **Number of repetitions (B)**: The number of times the simulation is repeated to generate a Monte Carlo sample.

For example, if we are interested in the MSE of the sample mean of 100 i.i.d. coin flips, we set:

- $\hat{\theta} = \overline{Y}$ (the sample mean),
- F as the Bernoulli distribution with $P(Y = 1) = 0.5$,
- $\theta = E[Y] = 0.5$ (the population mean),
- $n = 100$,
- an i.i.d. sampling scheme,
- a large number of repetitions, such as $B = 10000$.

8.3 Monte Carlo algorithm

The Monte Carlo simulation is performed as follows:

1. Using the specified sampling scheme, draw a sample $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ of size n from F using the computer's random number generator. Evaluate the estimator $\hat{\theta}$ from $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$.
2. Repeat step 1 of the experiment B times and collect the estimates in the Monte Carlo sample
$$\hat{\theta}_{mc} = \{\hat{\theta}_1, \dots, \hat{\theta}_B\}.$$

3. Estimate the features of interest from the Monte Carlo sample:

   - Mean:
   $$\hat{\mu}_{mc} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i.$$
   - Bias:
   $$\widehat{Bias}(\hat{\theta}_{mc}) = \hat{\mu}_{mc} - \theta$$
   - Variance:
   $$\widehat{Var}(\hat{\theta}_{mc}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \hat{\mu}_{mc})^2$$
   - MSE:
   $$\widehat{MSE}(\hat{\theta}_{mc}) = \widehat{Var}(\hat{\theta}_{mc}) + \widehat{Bias}(\hat{\theta}_{mc})^2$$

8.4 Sample mean of coin flips

Let's conduct a Monte Carlo simulation for the sample mean of coin flips.

```
set.seed(1)  # Set seed for reproducibility

# Function to generate a random sample and compute its sample mean
getMCsample = function(n) {
  # Generate an i.i.d. Bernoulli sample of size n with probability 0.5
  X = rbinom(n, size = 1, prob = 0.5)
  # Compute and return the sample mean of X
  mean(X)
}

# True parameter value (population mean) of the Bernoulli distribution
theta = 0.5

# Number of Monte Carlo repetitions
B = 1000

# Function to perform Monte Carlo simulation and calculate Bias, Variance, and MSE for a give
simulate_bias_variance_mse = function(n) {
  # Generate a Monte Carlo sample of B sample means
  MCsample = replicate(B, getMCsample(n))
  # Calculate Bias, Variance, and MSE
  Bias = mean(MCsample) - theta
  Variance = var(MCsample)
```

```
  MSE = Variance + Bias^2
  # Return the results as a vector
  c(Bias, Variance, MSE)
}

# Run the simulation for different sample sizes and store results
result10 = simulate_bias_variance_mse(10)
result20 = simulate_bias_variance_mse(20)
result50 = simulate_bias_variance_mse(50)
results = cbind(result10, result20, result50)

# Assign names to columns and rows for clarity in the output
colnames(results) = c("n=10", "n=20", "n=50")
rownames(results) = c("Bias", "Variance", "MSE")

# Display the results
results
```

```
                n=10        n=20        n=50
Bias     -0.00470000 -0.00370000 0.004740000
Variance  0.02605396  0.01272403 0.004631364
MSE       0.02607605  0.01273772 0.004653831
```

This output shows how the bias, variance, and MSE decrease as the sample size increases, which illustrates the consistency of the estimator.

8.5 Linear and nonlinear regression

Let's use Monte Carlo simulations to study the consistency properties of the OLS estimator in a simple linear regression model. We expect $\hat{\beta}_2$ to be a consistent estimator for β_2 in the following regression model:

$$Y_i = \beta_1 + \beta_2 Z_i + u_i, \quad E[u_i|Z_i] = 0, \tag{8.1}$$

provided (A2)–(A4) hold true. In this case, $\hat{\beta}_2$ is

$$\hat{\beta}_2 = \frac{\hat{\sigma}_{YZ}}{\hat{\sigma}_Z^2}. \tag{8.2}$$

However, the true relationship between Y and Z might be nonlinear such that the true model has the form

$$Y_i = \beta_1 + \beta_2 Z_i + \beta_3 Z_i^2 + \beta_4 Z_i^3 + v_i, \quad E[v_i|Z_i] = 0. \tag{8.3}$$

Note that $u_i = \beta_3 Z_i^2 + \beta_4 Z_i^3 + v_i$. Hence, if $\beta_3 \neq 0$, then

$$\begin{aligned} E[u_i|Z_i] &= E[\beta_3 Z_i^2 + \beta_4 Z_i^3 + v_i|Z_i] \\ &= \beta_3 Z_i^2 + \beta_4 Z_i^3 + E[v_i|Z_i] \\ &= \beta_3 Z_i^2 + \beta_4 Z_i^3 \neq 0, \end{aligned}$$

and the simple model from Equation 8.1 cannot be true. This means the error term contains systematic patterns related to Z_i, which violates a key assumption (A1) of linear regression.

In this case, using $\hat{\beta}_2$ from Equation 8.2 to estimate β_2 from Equation 8.3 will lead to a biased estimate.

Let's simulate data from models Equation 8.1 and Equation 8.3 where:

- Z_i, u_i, v_i are i.i.d. and $\mathcal{N}(0, 1)$ (standard normal distribution)
- $n = 100$
- $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = -3$, $\beta_4 = -1$

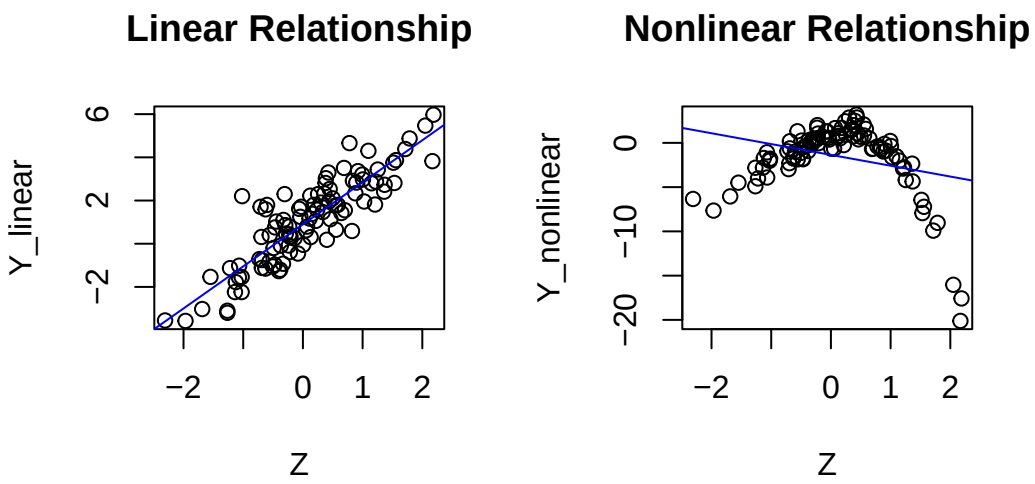
```
set.seed(123)  # For reproducibility

# Parameters
beta1 = 1
beta2 = 2
beta3 = -3
beta4 = -1
n = 100

# Data generation
Z = rnorm(n)
Y_linear = beta1 + beta2 * Z + rnorm(n)
Y_nonlinear = beta1 + beta2 * Z + beta3 * Z^2 + beta4 * Z^3 + rnorm(n)

# Linear Case Plot with Regression Line
par(mfrow = c(1, 2))
plot(Z, Y_linear, main = "Linear Relationship")
fit1 = lm(Y_linear ~ Z) # fit simple linear model
abline(fit1, col = "blue")  # Add linear regression line

# Nonlinear Case Plot with Regression Line
plot(Z, Y_nonlinear, main = "Nonlinear Relationship")
fit2 = lm(Y_nonlinear ~ Z) # fit simple linear model without Z^2
abline(fit2, col = "blue")  # Add linear regression line
```

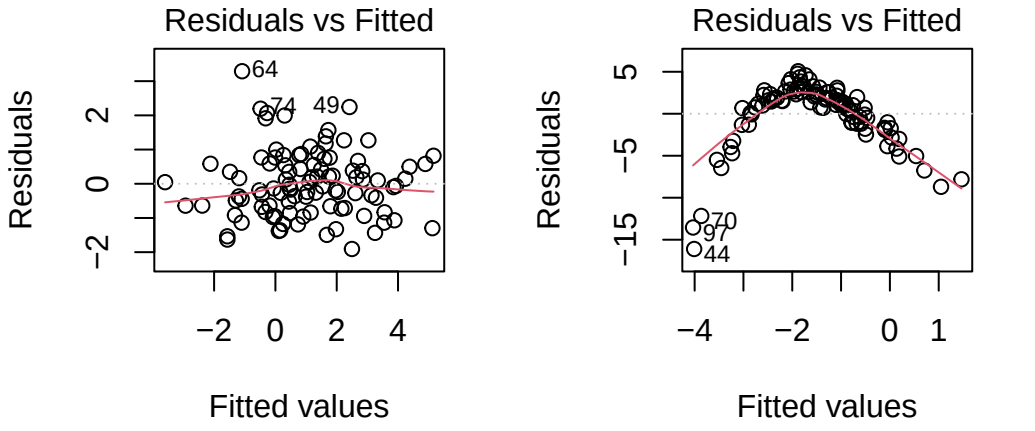


In the left plot, the model is correctly specified, i.e., $E[u_i|Z_i] = 0$ holds. In the right plot, the model is misspecified, i.e., $E[u_i|Z_i] \neq 0$.

This becomes also evident in the residuals versus fitted values plots. The residuals serve as proxies for the unknown error terms, while the fitted values $\widehat{Y}_i = \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ provide a one-dimensional summary of all regressors.

Residuals that are equally spread around a horizontal line without distinct patterns, as shown in the left plot below, indicate a correctly specified linear model. When the size or sign of the residuals systematically depends on the fitted values, as in the right plot below, this suggests hidden nonlinear relationships between the response and predictors that the model fails to capture.

```
## Diagnostics plot
par(mfrow = c(1, 2))
plot(fit1, which = 1)
plot(fit2, which = 1)
```



The red solid line indicates a local scatterplot smoother, which is a smooth locally weighted line through the points on the scatterplot to visualize the general pattern of the data.

8.5.1 Simulation of the linear case

To assess the statistical properties of our estimator, we examine how accurately $\hat{\beta}_2$ from Equation 8.2 estimates the true parameter β_2 in the correctly specified model Equation 8.1.

```
set.seed(1)  # Set seed for reproducibility

# True parameter values
beta1 = 1
beta2 = 2

# Generate a random sample and compute OLS coefficient beta2-hat
getMCsample = function(n) {
  # Data generation
  Z = rnorm(n)
  Y_linear = beta1 + beta2 * Z + rnorm(n)
  fit1 = lm(Y_linear ~ Z) # fit simple linear model
  # Compute and return beta2-hat
  fit1$coefficients[2]
}

# Number of Monte Carlo repetitions
B = 1000

# Function to perform Monte Carlo simulation and calculate Bias, Variance, and MSE for a give
simulate_bias_variance_mse = function(n) {
  # Generate a Monte Carlo sample of B sample means
  MCsample = replicate(B, getMCsample(n))
  # Calculate Bias, Variance, and MSE
  Bias = mean(MCsample) - beta2
  Variance = var(MCsample)
  MSE = Variance + Bias^2
  # Return the results as a vector
  c(Bias, Variance, MSE)
}

# Run the simulation for different sample sizes and store results
result10 = simulate_bias_variance_mse(10)
result20 = simulate_bias_variance_mse(20)
```

```r
result50 = simulate_bias_variance_mse(50)
results = cbind(result10, result20, result50)

# Assign names to columns and rows for clarity in the output
colnames(results) = c("n=10", "n=20", "n=50")
rownames(results) = c("Bias", "Variance", "MSE")

# Display the results
results
```

```
              n=10         n=20         n=50
Bias     0.0155187 -0.003998293 -0.001679989
Variance 0.1468236  0.056849539  0.021480276
MSE      0.1470645  0.056865525  0.021483098
```

- The bias of $\hat{\beta}_2$ is close to zero for all sample sizes.
- The variance decreases as n increases.
- The MSE decreases with larger n, which indicates that $\hat{\beta}_2$ is a consistent estimator when the model is correctly specified.

8.5.2 Simulation of the nonlinear case

We now examine how the OLS estimator $\hat{\beta}_2$ from the linear model Equation 8.2 performs when the true data generating process contains nonlinear terms, as specified in Equation 8.3. This allows us to quantify the bias that arises from omitting the nonlinear terms.

```r
set.seed(1)  # Set seed for reproducibility

# True parameter values
beta1 = 1
beta2 = 2
beta3 = -3
beta4 = -1

# Generate a random sample and compute OLS coefficient beta2-hat
getMCsample = function(n) {
  # Data generation
  Z = rnorm(n)
  Y_nonlinear = beta1 + beta2 * Z + beta3 * Z^2 + beta4 * Z^3 + rnorm(n)
  fit2 = lm(Y_nonlinear ~ Z) # fit simple linear model without Z^2
  # Compute and return beta2-hat
```

```r
  fit2$coefficients[2]
}

# Number of Monte Carlo repetitions
B = 1000

# Function to perform Monte Carlo simulation and calculate Bias, Variance, and MSE for a give
simulate_bias_variance_mse = function(n) {
  # Generate a Monte Carlo sample of B sample means
  MCsample = replicate(B, getMCsample(n))
  # Calculate Bias, Variance, and MSE
  Bias = mean(MCsample) - beta2
  Variance = var(MCsample)
  MSE = Variance + Bias^2
  # Return the results as a vector
  c(Bias, Variance, MSE)
}

# Run the simulation for different sample sizes and store results
result10 = simulate_bias_variance_mse(10)
result20 = simulate_bias_variance_mse(20)
result50 = simulate_bias_variance_mse(50)
results = cbind(result10, result20, result50)

# Assign names to columns and rows for clarity in the output
colnames(results) = c("n=10", "n=20", "n=50")
rownames(results) = c("Bias", "Variance", "MSE")

# Display the results
results
```

```
              n=10      n=20      n=50
Bias     -2.514799 -2.653668 -2.844885
Variance  8.606104  5.340871  2.118467
MSE      14.930317 12.382827 10.211839
```

- The bias of $\hat{\beta}_2$ is substantial and does not decrease with larger n.
- The variance decreases with larger n, but the MSE remains high due to the large bias.
- This demonstrates that omitting the relevant nonlinear terms (Z_i^2 and Z_i^3) leads to a biased and inconsistent estimator of β_2 when the true model is nonlinear.

8.6 R-codes

statistics-sec08.R

9 Marginal effects

9.1 Marginal Effects

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

$$wage_i = \beta_1 + \beta_2 \, edu_i + u_i, \quad i = 1, \dots, n, \tag{9.1}$$

where (A1) holds, i.e.:

$$E[u_i|edu_i] = 0.$$

Population regression function:

$$\begin{aligned} m(edu_i) &= E[wage_i|edu_i] \\ &= \beta_1 + \beta_2 edu_i + E[u_i|edu_i] \\ &= \beta_1 + \beta_2 edu_i \end{aligned}$$

$$m(edu_i) = E[wage_i|edu_i] = \underbrace{\beta_1 + \beta_2 edu_i}_{=m(edu_i)} + \underbrace{E[u_i|edu_i]}_{=0}.$$

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

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

Marginal effect of education:

$$\frac{\partial E[wage_i|edu_i]}{\partial edu_i} = \beta_2.$$

```
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 than people with one year less of education.

The coefficient β_2 describes the **correlative relationship** between education and wages.

To see this, consider the covariance of the two variables:

$$\begin{aligned} Cov(wage_i, edu_i) &= Cov(\beta_1 + \beta_2\ edu_i, edu_i) + \underbrace{Cov(u_i, edu_i)}_{=0} \\ &= \beta_2 Var(edu_i) \end{aligned}$$

Therefore, the coefficient β_2 is proportional to the population correlation coefficient:

$$\beta_2 = \frac{Cov(wage_i, edu_i)}{Var[edu_i]} = Corr(wage_i, edu_i) \cdot \frac{sd(wage_i)}{sd(edu_i)}.$$

The marginal effect is a correlative effect and does not say where exactly a higher wage level for people with more education comes from. **Regression relationships do not necessarily imply a causal relationship**.

People with more education may earn more for a number of reasons. Maybe they are generally smarter or come from wealthier families, which leads to better paying jobs. Or maybe more education actually leads to higher earnings.

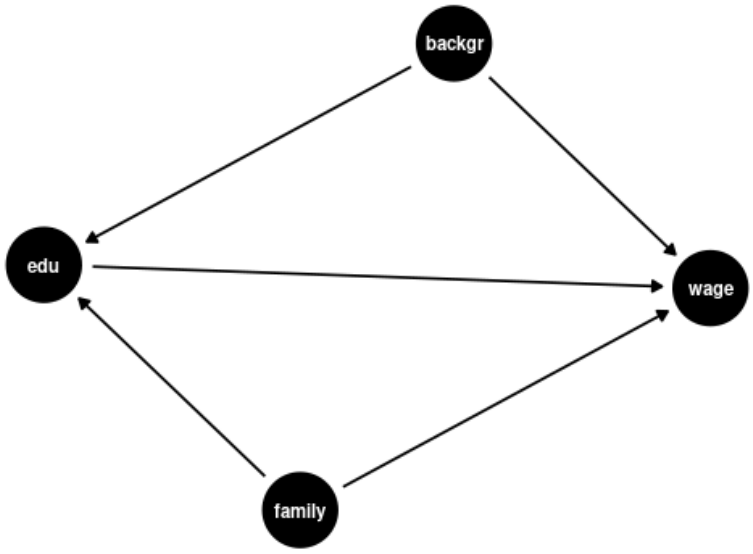


Figure 9.1: A DAG (directed acyclic graph) for the correlative and causal effects of edu on wage

The coefficient β_2 is a measure of how strongly education and earnings are correlated.

This association could be due to other factors that correlate with both wages and education, such as family background (parental education, family income, ethnicity, structural racism) or personal background (gender, intelligence).

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

To disentangle the causal effect of education on wages from other correlative effects, we can include control variables.

9.2 Control Variables

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, in this scenario),
(ii) is correlated with the regressor of interest (education),
(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 Equation 9.1 measures the correlative or marginal effect, not the causal effect. This must always be kept in mind when interpreting regression coefficients.

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 racial background and gender dummy variables for Black and female:

$$wage_i = \beta_1 + \beta_2 edu_i + \beta_3 exper_i + \beta_4 Black_i + \beta_5 fem_i + u_i.$$

In this case,

$$\beta_2 = \frac{\partial E[wage_i | edu_i, exper_i, Black_i, fem_i]}{\partial edu_i}$$

is the marginal effect of education on expected wages, holding experience, race, 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
   -21.7095       3.1350       0.2443      -2.8554      -7.4363
```

Interpretation: Given the same experience, racial background, 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.

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

Good control variables are variables that are determined before the level of education is determined. Control variables should not be the cause of the dependent variable of interest.

Examples of **good controls** for education are parental education level, region of residence, or educational industry/field of study.

A problematic situation is when the control variable is the cause of education. Bad controls are typically highly correlated with the independent variable of interest and irrelevant to the causal effect of that variable on the dependent variable.

Examples of **bad controls** for education are current job position, number of professional certifications obtained, or number of job offers.

A high correlation of the bad control with the variable education also causes a high variance of the OLS coefficient for education and leads to an imprecise coefficient estimate. This problem is called **imperfect multicollinearity**.

Bad controls make it difficult to interpret causal relationships. They may control away the effect you want to measure, or they may introduce additional reverse causal effects hidden in the regression coefficients.

9.3 CASchools: class size effect

Recall the `CASchools` dataset used in the Stock and Watson textbook in sections 4-8.

```
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` conditional on different control variables such as:

- `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.

```
cor(CASchools[,c("STR", "score", "english", "lunch", "expenditure")])
```

```
                  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 sample correlation matrix indicates that `english`, `lunch` and `expenditure` are correlated with `STR` and `score`, which implies these variables could confound the relationship of `STR` on `score` (omitted variable bias).

```
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(stargazer)
```

```
stargazer(fit1, fit2, fit3, fit4, type="latex", report="vc*", omit.stat = "f",
          star.cutoffs = NA, header = FALSE)
```

Interpretations:

- Model (1): Between two classes that differ by one student, the class with more students scores on average 2.280 points lower.
- 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.
- Model (3): Between two classes that differ by one student but have the same share of English learners and students with reduced meals, the larger class scores on average 0.998 points lower.

Table 9.1

	Dependent variable:			
	score			
	(1)	(2)	(3)	(4)
STR	−2.280	−1.101	−0.998	−0.235
english		−0.650	−0.122	−0.128
lunch			−0.547	−0.546
expenditure				0.004
Constant	698.933	686.032	700.150	665.988
Observations	420	420	420	420
R^2	0.051	0.426	0.775	0.783
Adjusted R^2	0.049	0.424	0.773	0.781
Residual Std. Error	18.581 (df = 418)	14.464 (df = 417)	9.080 (df = 416)	8.910 (df = 415)
Note:				NA

- 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.

The variables `english` and `lunch` are **good controls** because they are likely determined before class size decisions and capture important student background characteristics. These pre-existing factors can influence both class size assignments (as schools might create smaller classes for disadvantaged students) and test scores.

Per-pupil `expenditure`, however, is a **bad control** because it is likely determined simultaneously with or after class size decisions. Smaller classes mechanically increase per-pupil expenditure through higher teacher salary costs per student. Including expenditure therefore "controls away" part of the class size effect we aim to measure, which leads to potential underestimation of the true effect.

9.4 Polynomials

A linear dependence on 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.

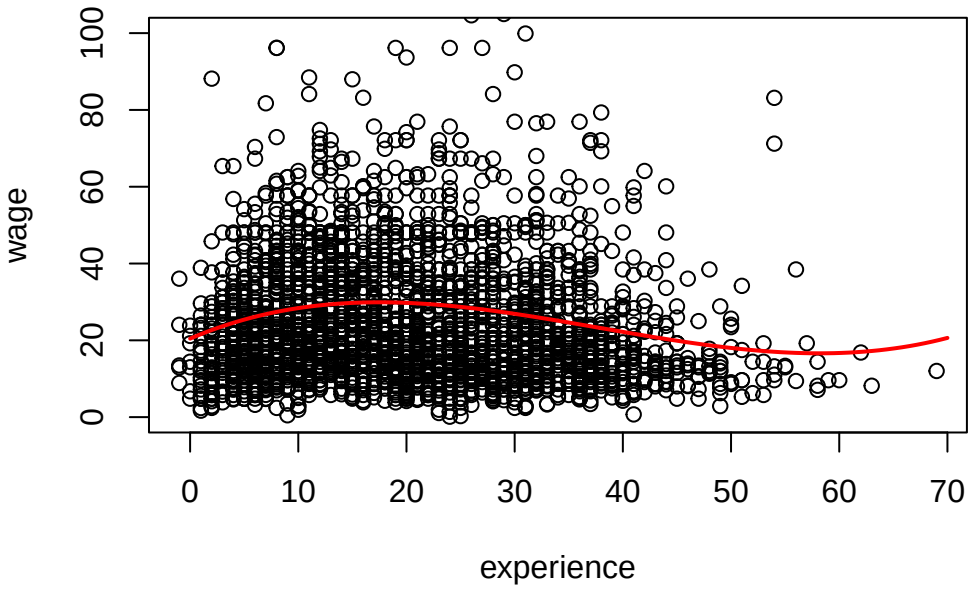
Polynomials can be used to specify a nonlinear regression function:

$$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 exper_i^3 + u_i.$$

```
## 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
```

```
    (Intercept)      experience I(experience^2) I(experience^3)
  20.4547146896    1.2013241316   -0.0446897909    0.0003937551
```

```
## 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:

$$\frac{\partial E[wage_i|exper_i]}{\partial exper_i} = \beta_2 + 2\beta_3 exper_i + 3\beta_4 exper_i^2.$$

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

$$1.43 + 2 \cdot (-0.042) \cdot 10 + 3 \cdot 0.0003 \cdot 10^2 = 0.68.$$

9.5 Interactions

A linear regression with interaction terms:

$$wage_i = \beta_1 + \beta_2 edu_i + \beta_3 fem_i + \beta_4 marr_i + \beta_5 (marr_i \cdot fem_i) + u_i$$

```
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
       -17.886           2.867          -3.266           7.167          -5.767
```

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

$$\frac{\partial E[wage_i|edu_i, female_i, married_i]}{\partial female_i} = \beta_3 + \beta_5 married_i$$

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:

$$\frac{\partial E[wage_i|edu_i, female_i, married_i]}{\partial married_i} = \beta_4 + \beta_5 female_i$$

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.

9.6 Logarithms

When analyzing wage data, we often use logarithmic transformations because they help model proportional relationships and reduce the skewness of the typically right-skewed distribution of wages. A common specification is the log-linear model, where we take the logarithm of wages while keeping education in its original scale:

In the logarithmic specification

$$\log(wage_i) = \beta_1 + \beta_2 edu_i + u_i$$

we have

$$\frac{\partial E[\log(wage_i)|edu_i]}{\partial edu_i} = \beta_2.$$

This implies

$$\underbrace{\partial E[\log(wage_i)|edu_i]}_{\substack{\text{absolute}\\ \text{change}}} = \beta_2 \cdot \underbrace{\partial edu_i}_{\substack{\text{absolute}\\ \text{change}}}.$$

That is, β_2 gives the average absolute change in log-wages when education changes by 1.

Another interpretation can be given in terms of relative changes. Consider the following approximation:

$$E[wage_i|edu_i] \approx \exp(E[\log(wage_i)|edu_i]).$$

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

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

Using the geometric mean approximation from above, we get

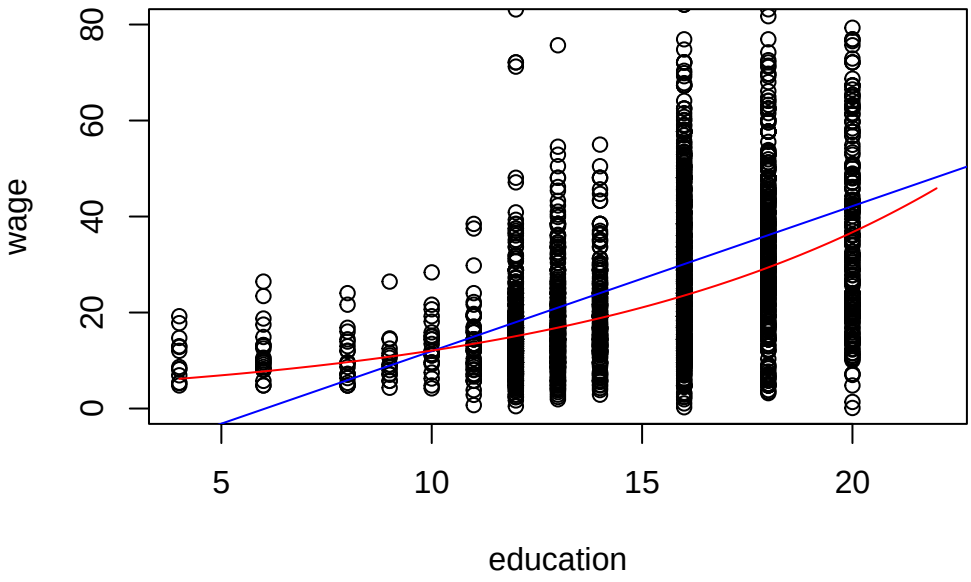
$$\underbrace{\frac{\partial E[wage_i|edu_i]}{E[wage_i|edu_i]}}_{\substack{\text{percentage} \\ \text{change}}} \approx \frac{\partial exp(E[\log(wage_i)|edu_i])}{exp(E[\log(wage_i)|edu_i])} = \beta_2 \cdot \underbrace{\partial edu_i}_{\substack{\text{absolute} \\ \text{change}}}.$$

```
linear_model = lm(wage ~ education, data = cps.as)
log_model = lm(log(wage) ~ education, data = cps.as)
log_model
```

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

Coefficients:
(Intercept)    education
     1.3783       0.1113
```

```
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")
```



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

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

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

and the log-log specification

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

Linear-log interpretation: When X is 1% higher, we observe, on average, a $0.01\beta_2$ higher Y.

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

9.7 CASchools: nonlinear specifications

Let's have a look at an example that explores the relationship between the income of schooling districts and their test scores.

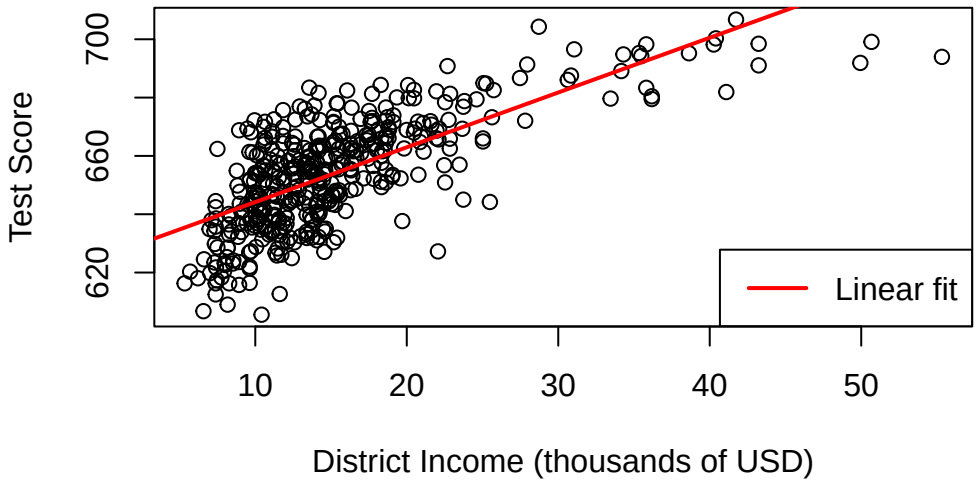
We start our analysis by computing the correlation between both variables.

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

```
[1] 0.7124308
```

Income and test score are positively correlated: school districts with above-average income tend to achieve above-average test scores. But does a linear regression adequately model the data? To investigate this further, let's visualize the data by plotting them and adding a linear regression line.

```
# Create scatterplot
plot(score ~ income, data = CASchools,
    xlab = "District Income (thousands of USD)",
    ylab = "Test Score")
# Fit linear model and add regression line
linear = lm(score ~ income, data = CASchools)
abline(linear, col = "red", lwd = 2)
# Add legend
legend("bottomright", "Linear fit", col = "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 underestimates it for the middle income group. Luckily, OLS isn't limited to linear regressions of the predictors. We have the flexibility to model test scores as a function of income and the square of income.

This leads us to the following regression model:

$$score_i = \beta_1 + \beta_2\, income_i + \beta_3\, income_i^2 + u_i$$

which is a *quadratic regression model*. Here we treat $income^2$ as an additional explanatory variable.

```
# fit the quadratic Model
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
```

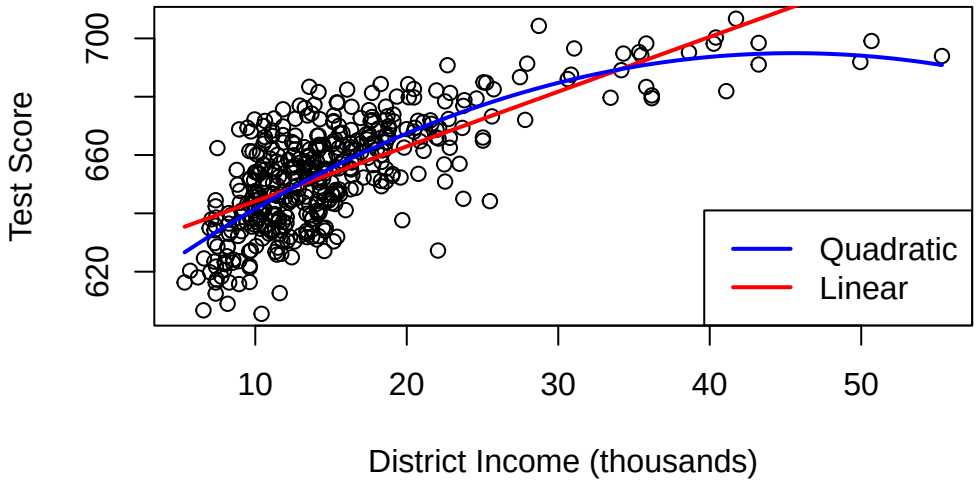
The estimated function is

$$\widehat{score} = 607.3 + 3.85\, income - 0.0423\, income^2$$

```r
# 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)
```



As the plot shows, the quadratic function appears to provide a better fit to the data compared to the linear function.

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

```r
# estimate a level-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
```

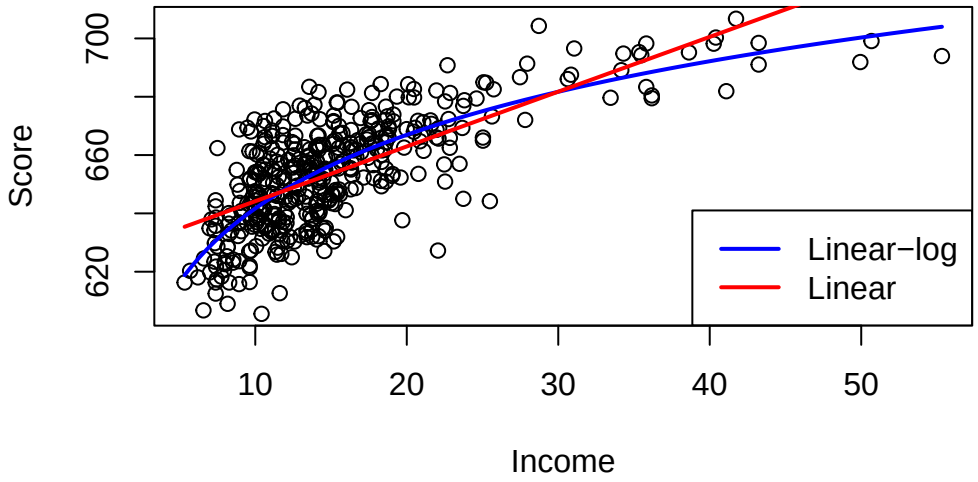
The estimated regression model is

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

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



We can interpret $\hat{\beta}_2$ as follows: a 1% increase in income is associated with an average increase in test scores of $0.01 \cdot 36.42 = 0.36$ points.

9.8 R-codes

statistics-sec09.R

10 Confidence intervals

10.1 Estimation uncertainty

An estimator provides an approximation of an unknown population parameter as a single real number or vector, which we call a **point estimate**. For instance, when we estimate the linear relationship between wage, education, and gender using an OLS, we obtain a specific set of coefficients:

```
cps = read.csv("cps.csv")
lm(wage ~ education + female, data = cps) |> coef()
```

```
(Intercept)   education      female
 -14.081788    2.958174   -7.533067
```

However, the point estimate $\hat{\beta}_j$ alone does not reflect how close or far the estimate might be from the true population parameter β_j. It doesn't capture estimation uncertainty. This inherent uncertainty arises because point estimates are based on a finite sample, which may vary from sample to sample.

Larger samples tend to give more accurate OLS estimates as OLS is unbiased and consistent under assumptions (A1)–(A4). However, we work with fixed, finite samples in practice.

Confidence intervals address this limitation by providing a range of values likely to contain the true population parameter. By constructing an interval around our point estimate that contains the true parameter with a specified probability (e.g., 95% confidence level), we can express the uncertainty more clearly.

In this section, we will introduce **interval estimates**, commonly referred to as **confidence intervals**. To construct a confidence interval for an OLS coefficient $\hat{\beta}_j$, we need two components: a **standard error** (an estimate of the standard deviation of the estimator) and information about the distribution of $\hat{\beta}_j$.

10.2 Gaussian distribution

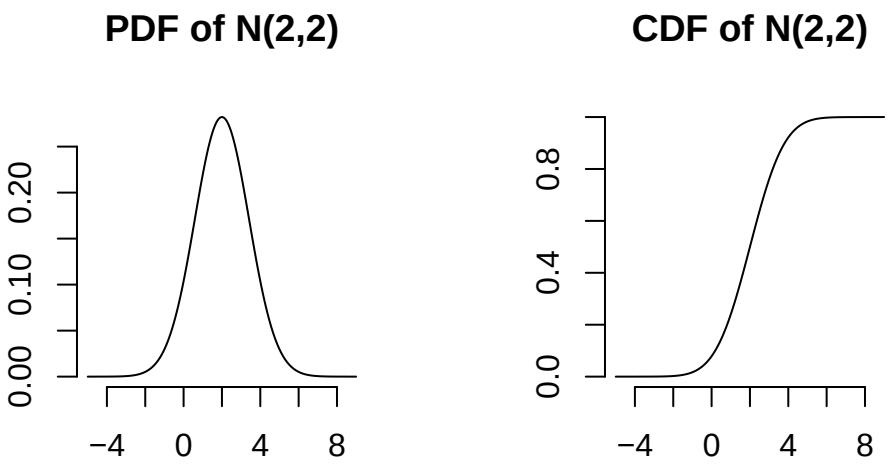
The **Gaussian distribution**, also known as the **normal distribution**, is a fundamental concept in statistics. We often use these terms interchangeably: a random variable Z is said to follow a Gaussian or normal distribution if it has the following probability density function (PDF) with a given mean μ and variance σ^2:

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

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: $Var(Z) = \sigma^2$
- Skewness: $skew(Z) = 0$
- Kurtosis: $kurt(Z) = 3$

```
par(mfrow=c(1,2), bty="n", lwd=1)
x = seq(-5,9,0.01) # define grid for x-axis for the plot
plot(x, dnorm(x, mean = 2, sd = sqrt(2)), type="l", main="PDF of N(2,2)", ylab="", xlab="")
plot(x, pnorm(x, mean = 2, sd = sqrt(2)), type="l", main="CDF of N(2,2)", ylab="", xlab="")
```



Use the R functions `dnorm` to calculate normal PDF values and `pnorm` for normal CDF values.

The Gaussian 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 CDF

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

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

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

Standardizing: If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{Z-\mu}{\sigma} \sim \mathcal{N}(0,1),$$

and the CDF of Z is $\Phi((Z-\mu)/\sigma)$.

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

10.2.1 Multivariate Gaussian 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 density is

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

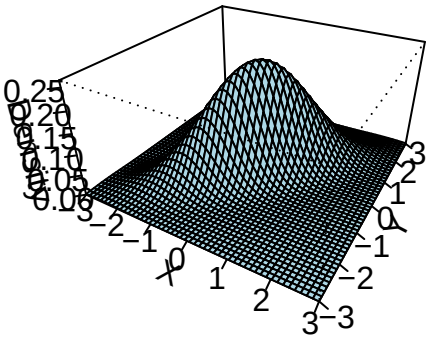
If $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\boldsymbol{Z}^* = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}$ for a $q \times 1$ vector $\boldsymbol{\mu}$ and a $q \times k$ matrix $\boldsymbol{B}$, then $\boldsymbol{Z}^*$ has a **multivariate normal distribution** with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$, written $\boldsymbol{Z}^* \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The k-variate PDF of $\boldsymbol{Z}^*$ is

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

The mean vector and covariance matrix are

$$E[\boldsymbol{Z}^*] = \boldsymbol{\mu}, \quad Var(\boldsymbol{Z}^*) = \boldsymbol{\Sigma}.$$

3D Bivariate Normal Distribution Density



The 3d plot shows the bivariate normal PDF with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

10.2.2 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.

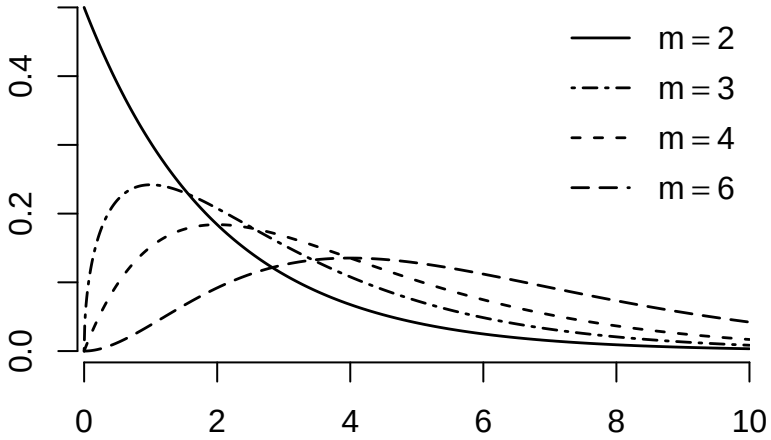


Figure 10.1: PDF of the χ^2-distribution

- Mean: $E[Y] = m$
- Variance: $Var(Y) = 2m$
- Skewness: $skew(Y) = \sqrt{8/m}$
- Kurtosis: $kurt(Y) = 3 + 12/m$

10.2.3 Student t-distribution

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

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

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

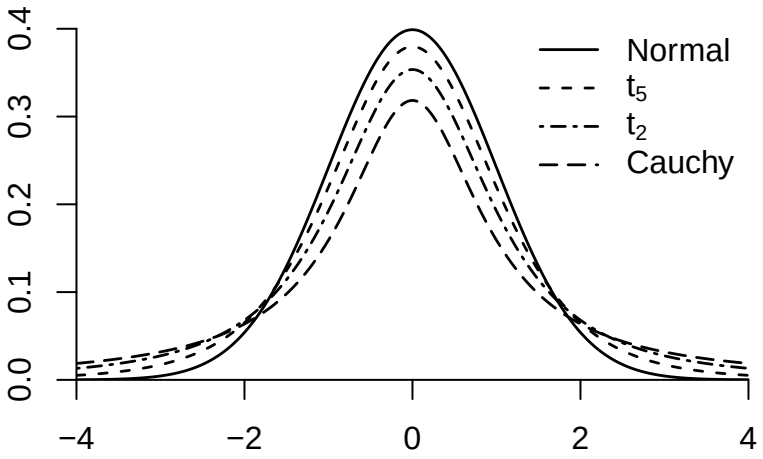


Figure 10.2: PDFs of the Student t-distribution

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

- Mean: $E[Y] = 0$ if $m \geq 2$
- Variance: $Var(Y) = \frac{m}{m-2}$ if $m \geq 3$
- Skewness: $skew(Y) = 0$ if $m \geq 4$
- Kurtosis: $kurt(Y) = 3 + 6/(m - 4)$ if $m \geq 5$

The kurtosis is infinite for $m \leq 4$, the skewness is undefined for $m \leq 3$, the variance is infinite for $m \leq 2$, and the mean is undefined for $m = 1$.

10.3 Classical Gaussian regression model

Let's revisit the linear regression model:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \dots, n. \tag{10.1}$$

Under assumptions (A1)–(A4), the distributional restrictions on the error term are relatively mild:

1) The error terms are i.i.d. but can have different conditional variances depending on the values of the regressors (heteroskedasticity):

$$Var(u_i|\boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i) = \sigma_i^2.$$

   For example, in a regression of `wage` on `female`, the error variances for women may differ from those for men.

2) The error term can follow any distribution, provided that the fourth moment (the kurtosis) is finite. This excludes heavy-tailed distributions.

In standard introductory textbooks, two additional assumptions are often made to further restrict the properties mentioned above. It is beneficial to first study the estimation uncertainty under this simplified setting.

Classical Gaussian regression model

In addition to the linear regression model in Equation 10.1 with assumptions (A1)–(A4), we introduce two more assumptions:

- (A5) **Homoskedasticity**: The error terms have constant variance across all observations, i.e.,

$$Var(u_i|\boldsymbol{X}_i) = \sigma_i^2 = \sigma^2 \quad \text{for all } i = 1, \dots, n.$$

- (A6) **Normality**: The error terms are normally distributed conditional on the regressors, i.e.,

$$u_i|\boldsymbol{X}_i \sim \mathcal{N}(0, \sigma_i^2).$$

(A5)-(A6) combined can be expressed as:

$$u_i|\boldsymbol{X}_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for all } i = 1, \dots, n.$$

The notation $u_i|\boldsymbol{X}_i \sim \mathcal{N}(0, \sigma^2)$ means that the conditional distribution of u_i conditional on $\boldsymbol{X}_i$ is $N(0, \sigma^2)$. The PDF of $u_i|\boldsymbol{X}_i$ is

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{u^2}{2\sigma^2}\Big).$$

Distribution of the OLS coefficients

Conditional on $\boldsymbol{X}$, the OLS coefficient vector is a linear combination of the error term:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}.\end{aligned}$$

Consequently, under (A6), the OLS estimator follows a k-variate normal distribution, conditionally on $\boldsymbol{X}$.

Recall that the mean is $E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta}$ and the covariance matrix is

$$Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Under homoskedasticity (A5), we have $\boldsymbol{D} = \sigma^2\boldsymbol{I}_n$, so the covariance matrix simplifies to

$$Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Therefore,

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

The variance of the j-th OLS coefficient is

$$Var(\hat{\beta}_j|\boldsymbol{X}) = \sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj},$$

where $[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}$ indicates the j-th diagonal element of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The standard deviation is:

$$sd(\hat{\beta}_j|\boldsymbol{X}) = \sqrt{\sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}.$$

Therefore, the standardized OLS coefficient has a standard normal distribution:

$$Z_j := \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j|\boldsymbol{X})} \sim \mathcal{N}(0,1). \tag{10.2}$$

10.4 Confidence interval: known variance

One of the most common methods of incorporating estimation uncertainty into estimation results is through **interval estimates**, often referred to as **confidence intervals**.

A confidence interval 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%). For example, a 95% confidence interval suggests that, across many repeated samples, approximately 95% of the intervals constructed from those samples would contain the true population parameter.

A symmetric confidence interval for β_j with confidence level $1-\alpha$ is an interval

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

with the property that

$$P(\beta_j \in I_{1-\alpha}) = 1-\alpha. \tag{10.3}$$

Common coverage probabilities are 0.95, 0.99, and 0.90.

Note that $I_{1-\alpha}$ is random and β_j is fixed but unknown. Therefore, the coverage probability is the probability that this random interval $I_{1-\alpha}$ contains the true parameter.

A more precise interpretation of a confidence interval is:

If we were to repeat the sampling process and construct confidence intervals for each sample, $1-\alpha$ of those intervals would contain the true population parameter.

It is essential to understand that the confidence interval reflects the reliability of the method, not the probability of the true parameter falling within a particular interval. The interval itself is random – it varies with each sample – but the population parameter is fixed and unknown.

Thus, it is incorrect to interpret a specific confidence interval as having a 95% probability of containing the true value. Instead, the correct interpretation is that the method used to calculate the interval has a 95% success rate across many samples.

The width of the interval

The OLS coefficient $\hat{\beta}_j$ is in the center of $I_{1-\alpha}$. Let's solve for $c_{1-\alpha}$ to get the width of the confidence interval.

The event $\{\beta_j \in I_{1-\alpha}\}$ can be rearranged as

$$\begin{aligned}
& \beta_j \in I_{1-\alpha} \\
\Leftrightarrow \quad & \hat{\beta}_j - c_{1-\alpha} \leq \beta_j \leq \hat{\beta}_j + c_{1-\alpha} \\
\Leftrightarrow \quad & -c_{1-\alpha} \leq \beta_j - \hat{\beta}_j \leq c_{1-\alpha} \\
\Leftrightarrow \quad & c_{1-\alpha} \geq \hat{\beta}_j - \beta_j \geq -c_{1-\alpha} \\
\Leftrightarrow \quad & \frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})} \geq Z_j \geq -\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}
\end{aligned}$$

with Z_j defined in Equation 10.2. Hence, Equation 10.3 becomes

$$P\left(\frac{-c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})} \leq Z_j \leq \frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) = 1-\alpha. \tag{10.4}$$

Since Z_j is standard normal by Equation 10.2, we have

$$\begin{aligned}
&P\left(\frac{-c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})} \leq Z_j \leq \frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) \\
&= \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) - \Phi\left(\frac{-c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) \\
&= \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) - \left(1 - \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right)\right) \\
&= 2\Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) - 1.
\end{aligned}$$

With Equation 10.4, we get

$$1 - \alpha = 2\Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right) - 1.$$

Let's add 1 and divide by 2:

$$1 - \frac{\alpha}{2} = \Phi\left(\frac{c_{1-\alpha}}{sd(\hat{\beta}_j|\boldsymbol{X})}\right), \tag{10.5}$$

where $(2-\alpha)/2 = 1 - \alpha/2$.

The value $z_{(p)}$ is the p**-quantile** of $\mathcal{N}(0,1)$ if $\Phi(z_{(p)}) = p$. We write $\Phi^{-1}(p) = z_{(p)}$, where the quantile function Φ^{-1} is the inverse function of the CDF Φ with $\Phi(\Phi^{-1}(p)) = p$ and $\Phi^{-1}(\Phi^{-1}(z)) = z$.

Then, applying the quantile function Φ^{-1} to Equation 10.5 gives:

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

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

$$I_{1-\alpha} = \left[\hat{\beta}_j - z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\beta}_j|\boldsymbol{X});\ \hat{\beta}_j + z_{(1-\frac{\alpha}{2})} \cdot sd(\hat{\beta}_j|\boldsymbol{X})\right].$$

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

Table 10.1: Some quantiles of the standard normal distribution

0.9	0.95	0.975	0.99	0.995
1.28	1.64	1.96	2.33	2.58

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

$$\begin{aligned} I_{0.9} &= [\hat{\beta}_j - 1.64 \cdot sd(\hat{\beta}_j|\boldsymbol{X});\ \hat{\beta}_j + 1.64 \cdot sd(\hat{\beta}_j|\boldsymbol{X})] \\ I_{0.95} &= [\hat{\beta}_j - 1.96 \cdot sd(\hat{\beta}_j|\boldsymbol{X});\ \hat{\beta}_j + 1.96 \cdot sd(\hat{\beta}_j|\boldsymbol{X})] \\ I_{0.99} &= [\hat{\beta}_j - 2.58 \cdot sd(\hat{\beta}_j|\boldsymbol{X});\ \hat{\beta}_j + 2.58 \cdot sd(\hat{\beta}_j|\boldsymbol{X})] \end{aligned}$$

With probability α, the interval does not cover the true parameter. The smaller we choose α, the more confident we can be that the interval covers the true parameter, but the larger the interval becomes. If we set $\alpha = 0$, the interval would be infinite, providing no useful information.

A certain amount of uncertainty always remains, but we can control it by choosing an appropriate value for α that balances our desired level of confidence with the precision of the estimate. This is why the coverage probability $(1 - \alpha)$ is also called the **confidence level**.

Note that this interval is **infeasible** in practice because the conditional standard deviation is unknown:

$$sd(\hat{\beta}_j|\boldsymbol{X}) = \sqrt{\sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}.$$

It requires knowledge about the true error variance $Var(u_i|\boldsymbol{X}) = \sigma^2$.

10.5 Classical standard errors

A standard error $se(\hat{\beta}_j)$ for an estimator $\hat{\beta}_j$ is an estimator of the standard deviation of the distribution of $\hat{\beta}_j$.

We say that the standard error is consistent if

$$\frac{se(\hat{\beta}_j)}{sd(\hat{\beta}_j|\boldsymbol{X})} \xrightarrow{p} 1. \tag{10.6}$$

This property ensures that, in practice, we can replace the unknown standard deviation with the standard error in confidence intervals.

Under the classical Gaussian regression model, we have

$$sd(\hat{\beta}_j|\boldsymbol{X}) = \sqrt{\sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}.$$

Therefore, it is natural to replace the population error variance σ^2 by the adjusted sample variance of the residuals:

$$s^2_{\widehat{u}} = \frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2 = SER^2.$$

The classical homoskedastic standard errors are:

$$se_{hom}(\hat{\beta}_j) = \sqrt{s^2_{\widehat{u}}[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}.$$

The classical homoskedastic covariance matrix estimator for $Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$ is

$$\widehat{\boldsymbol{V}}_{hom} = s^2_{\widehat{u}}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

```
fit = lm(wage ~ education + female, data = cps)
## 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
```

The classical standard errors are the square roots of the diagonal elements of this matrix:

```
## classical standard errors:
sqrt(diag(vcov(fit)))
```

```
(Intercept)   education      female
 0.43388334  0.03037287  0.16858239
```

These standard errors are also displayed in the second column of a regression output:

```
summary(fit)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-45.071  -9.035  -2.973   4.472 244.491
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.08179    0.43388  -32.45   <2e-16 ***
education     2.95817    0.03037   97.39   <2e-16 ***
female       -7.53307    0.16858  -44.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.76 on 50739 degrees of freedom
Multiple R-squared:  0.1797,    Adjusted R-squared:  0.1797
F-statistic:  5559 on 2 and 50739 DF,  p-value: < 2.2e-16
```

Because $s_{\widehat{u}}^2/\sigma^2 \xrightarrow{p} 1$, property Equation 10.6 is satisfied and $se_{hom}(\hat{\beta}_j)$ is a consistent standard error under homoskedasticity.

Note that the main result we used to derive the confidence interval is that the standardized OLS coefficient is standard normal:

$$Z_j := \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j|\boldsymbol{X})} \sim \mathcal{N}(0,1).$$

If we replace the unknown standard deviation $sd(\hat{\beta}_j|\boldsymbol{X})$ with the standard error $se_{hom}(\hat{\beta}_j)$, the distribution changes.

The OLS estimator standardized with the standard error is called **t-statistic**:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{se_{hom}(\hat{\beta}_j)} = \frac{sd(\hat{\beta}_j|\boldsymbol{X})}{se_{hom}(\hat{\beta}_j)} \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j|\boldsymbol{X})} = \frac{sd(\hat{\beta}_j|\boldsymbol{X})}{se_{hom}(\hat{\beta}_j)} Z_j.$$

The additional factor satisfies

$$\frac{sd(\hat{\beta}_j|\boldsymbol{X})}{se_{hom}(\hat{\beta}_j)} = \frac{\sigma}{s_{\widehat{u}}} \sim \sqrt{(n-k)/\chi^2_{n-k}},$$

where χ^2_{n-k} is the chi-squared distribution with $n-k$ degrees of freedom, independent of Z_j.

Therefore, the t-statistic is t-distributed:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{se_{hom}(\hat{\beta}_j)} = \frac{\sigma}{s_{\widehat{u}}} Z_j \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} = t_{n-k}. \tag{10.7}$$

Consequently, if we replace the unknown standard deviation $sd(\hat{\beta}_j|\boldsymbol{X})$ with the standard error $se_{hom}(\hat{\beta}_j)$ in the confidence interval formula, we have to replace the standard normal quantiles by t-quantiles:

$$I_{1-\alpha}^{(hom)} = [\hat{\beta}_j - t_{(1-\frac{\alpha}{2},n-k)} se_{hom}(\hat{\beta}_j);\ \hat{\beta}_j + t_{(1-\frac{\alpha}{2},n-k)} se_{hom}(\hat{\beta}_j)]$$

This interval is feasible and satisfies $P(\beta_j \in I_{1-\alpha}^{(hom)}) = 1 - \alpha$ under (A1)–(A6).

Table 10.2: Student's t-distribution quantiles

df	0.9	0.95	0.975	0.99	0.995
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
8	1.40	1.86	2.31	2.90	3.36
10	1.37	1.81	2.23	2.76	3.17
15	1.34	1.75	2.13	2.60	2.95
20	1.33	1.72	2.09	2.53	2.85
25	1.32	1.71	2.06	2.49	2.79
30	1.31	1.70	2.04	2.46	2.75
40	1.30	1.68	2.02	2.42	2.70
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
80	1.29	1.66	1.99	2.37	2.64
100	1.29	1.66	1.98	2.36	2.63
$\to \infty$	1.28	1.64	1.96	2.33	2.58

We can use the `coefci` function from the `AER` package:

```
library(AER)
coefci(fit)
```

```
                2.5 %     97.5 %
(Intercept) -14.932204 -13.231372
education     2.898643   3.017705
female       -7.863490  -7.202643
```

```
coefci(fit, level = 0.99)
```

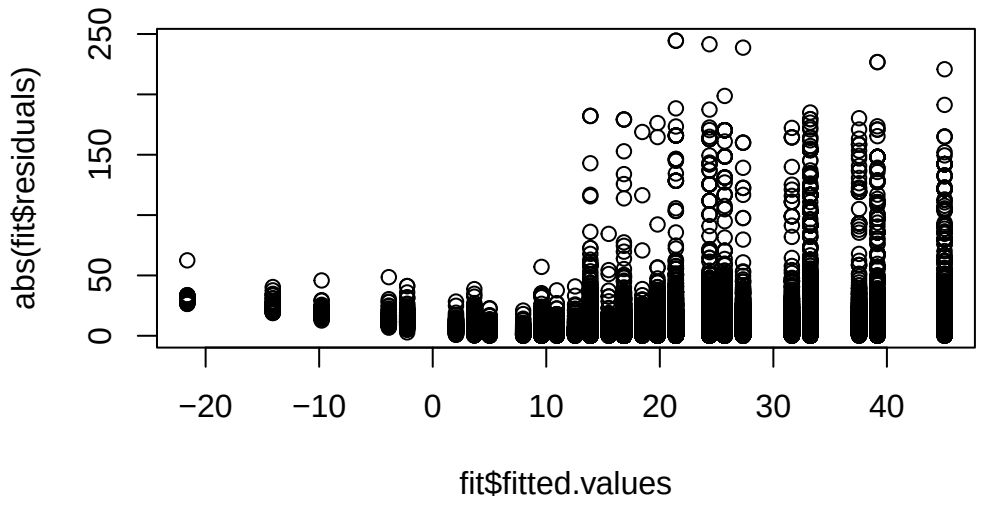
```
                 0.5 %     99.5 %
(Intercept) -15.199440 -12.964137
education     2.879936   3.036412
female       -7.967322  -7.098811
```

10.6 Confidence intervals: heteroskedasticity

The exact confidence interval $I_{1-\alpha}^{(hom)}$ is only valid under the restrictive assumption of homoskedasticity (A5) and normality (A6).

For historical reasons, statistics books often treat homoskedasticity as the standard case and heteroskedasticity as a special case. However, this does not reflect empirical practice since we have to expect heteroskedastic errors in most applications. It turns out that heteroskedasticity is not a problem as long as the robust standard errors are used.

```
plot(abs(fit$residuals)~fit$fitted.values)
```



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 (A5) is implausible.

If (A5) does not hold, then standard deviation is

$$sd(\hat{\beta}_j|\boldsymbol{X}) = \sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}.$$

To estimate $sd(\hat{\beta}_j|\boldsymbol{X})$, we will have to replace the diagonal matrix

$$\boldsymbol{D} = diag(\sigma_1^2, \dots, \sigma_n^2)$$

by some sample counterpart

$$\widehat{\boldsymbol{D}} = diag(\hat{\sigma}_1^2, \dots, \hat{\sigma}_n^2).$$

Various **heteroskedasticity-consistent (HC)** standard errors have been proposed in the literature:

HC type	weights
HC0	$\hat{\sigma}_i^2 = \hat{u}_i^2$
HC1	$\hat{\sigma}_i^2 = \frac{n}{n-k}\hat{u}_i^2$
HC2	$\hat{\sigma}_i^2 = \frac{\hat{u}_i^2}{1-h_{ii}}$
HC3	$\hat{\sigma}_i^2 = \frac{\hat{u}_i^2}{(1-h_{ii})^2}$

HC0 replaces the unknown variances with squared residuals, and HC1 is a bias-corrected version of HC0. HC2 and HC3 use the leverage values h_{ii} (the diagonal entries of the influence matrix $\boldsymbol{P}$) and give less weight to influential observations.

HC1 and HC3 are the most common choices and can be written as

$$se_{hc1}(\hat{\beta}_j) = \sqrt{\Big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\Big(\frac{n}{n-k}\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\Big)(\boldsymbol{X}'\boldsymbol{X})^{-1}\Big]_{jj}},$$

$$se_{hc3}(\hat{\beta}_j) = \sqrt{\Big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\Big(\sum_{i=1}^{n}\frac{\hat{u}_i^2}{(1-h_{ii})^2}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)(\boldsymbol{X}'\boldsymbol{X})^{-1}\Big]_{jj}}.$$

All versions perform similarly well in large samples, but HC3 performs best in small samples and is the preferred choice.

HC standard errors are also known as **heteroskedasticity-robust standard errors** or simply **robust standard errors**.

Estimators for the full covariance matrix of $\hat{\boldsymbol{\beta}}$ have the form

$$\widehat{\boldsymbol{V}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{D}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

The HC3 covariance estimator can be written as

$$\widehat{\boldsymbol{V}}_{hc3} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\Big(\sum_{i=1}^{n}\frac{\hat{u}_i^2}{(1-h_{ii})^2}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Therefore, we can use confidence intervals of the form:

$$I_{1-\alpha}^{(hc)} = [\hat{\beta}_j - t_{(1-\frac{\alpha}{2},n-k)} se_{hc}(\hat{\beta}_j);\ \hat{\beta}_j + t_{(1-\frac{\alpha}{2},n-k)} se_{hc}(\hat{\beta}_j)].$$

In contrast to Equation 10.7, the distribution of the ratio $sd(\hat{\beta}_j|\boldsymbol{X})/se_{hc}(\hat{\beta}_j)$ is unknown in practice, and the t-statistic is not t-distributed.

However, for large n, we have

$$T_j^{(hc)} = \frac{\hat{\beta}_j - \beta_j}{se_{hc}(\hat{\beta}_j)} = \underbrace{\frac{sd(\hat{\beta}_j|\boldsymbol{X})}{se_{hc}(\hat{\beta}_j)}}_{\xrightarrow{p} 1} \underbrace{Z_j}_{\sim\mathcal{N}(0,1)}$$

which implies that

$$\lim_{n\to\infty} P(\beta_j \in I_{1-\alpha}^{(hc)}) = 1 - \alpha. \tag{10.8}$$

Therefore $I_{1-\alpha}^{(hc)}$ is an **asymptotic confidence interval** for β_j.

```
## HC3 covariance matrix estimate Vhat-hc3
vcovHC(fit)
```

```
            (Intercept)    education       female
(Intercept)  0.25013606 -0.019590435  0.013394891
education   -0.01959043  0.001609169 -0.002173848
female       0.01339489 -0.002173848  0.026131235
```

```
## HC3 standard errors
sqrt(diag(vcovHC(fit)))
```

```
(Intercept)   education      female
 0.50013604  0.04011445  0.16165158
```

```
## HC1 standard errors
sqrt(diag(vcovHC(fit, type = "HC1")))
```

```
(Intercept)   education      female
 0.50007811  0.04011017  0.16164436
```

```
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
```

Robust confidence intervals can also be used and hold asymptotically under (A5). Therefore, the exact classical confidence intervals should only be used if there are very good reasons for the error terms to be homoskedastic and normally distributed.

10.7 Confidence interval with non-normal errors

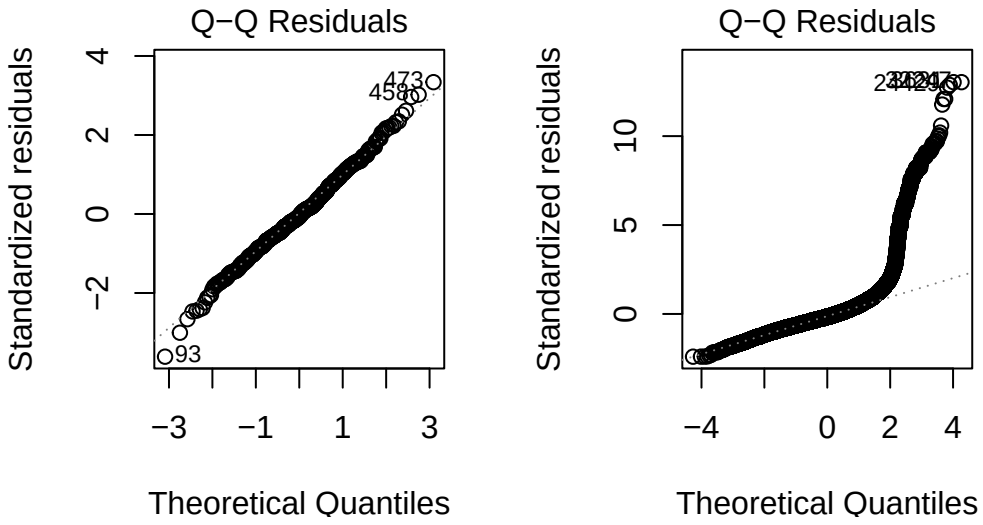
Similar to the homoskedasticity assumption (A5), the normality assumption (A6) is also not satisfied in most applications. A useful diagnostic plot is the Q-Q-plot.

The Q-Q-plot is a graphical tool to help us assess if the errors are conditionally normally distributed, i.e. whether assumption (A6) is satisfied.

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 are lined well on the straight dashed line, there is indication that the distribution of the residuals is close to a normal distribution.

```
par(mfrow = c(1,2))
# Normally distributed response variable
plot(lm(rnorm(500) ~ 1), which = 2)
plot(fit, which=2)
```



In the left plot you see the Q-Q-plot for an example with normally distributed errors. The right plot indicates that, in our regression of `wage` on `education` and `female`, the normality assumption is implausible.

If (A6) does not hold, then Z_j is not normally distributed, and it is unclear whether Equation 10.8 holds. However, by the central limit theorem, we still can establish that

$$\lim_{n\to\infty} P(\beta_j \in I_{1-\alpha}^{(hc)}) = 1 - \alpha.$$

Therefore, the robust confidence interval $I_{1-\alpha}^{(hc)}$ is asymptotically valid if (A1)–(A4) hold.

10.8 Central limit theorem

Convergence in distribution

Let $\boldsymbol{W}_n$ be a sequence of k-variate random variables and let $\boldsymbol{V}$ be a k-variate random variable

$\boldsymbol{W}_n$ converges in distribution to $\boldsymbol{V}$, written $\boldsymbol{W}_n \xrightarrow{d} \boldsymbol{V}$, if

$$\lim_{n\to\infty} P(\boldsymbol{W}_n \leq \boldsymbol{a}) = P(\boldsymbol{V} \leq \boldsymbol{a})$$

for all $\boldsymbol{a}$ at which the CDF of $\boldsymbol{V}$ is continuous.

If $\boldsymbol{V}$ has the distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we write $\boldsymbol{W}_n \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Consider for simplicity the regression on an intercept only. In this case, we have $k = 1$ and $\hat{\beta}_1 = \overline{Y}$ (see the second problem set).

By the univariate central limit theorem, the centered sample mean converges to a normal distribution:

Central Limit Theorem (CLT)

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

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

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

SHINY APP: CLT

The same result can be extended to k-variate random vectors.

Multivatiate Central Limit Theorem (MCLT)

If $\{\boldsymbol{W}_1, \dots, \boldsymbol{W}_n\}$ is an i.i.d. sample with $E[\boldsymbol{W}_i] = \boldsymbol{\mu}$ and $Var(\boldsymbol{W}_i) = \boldsymbol{\Sigma} < \infty$. Then,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

(see, e.g., Stock and Watson Section 19.2).

If we apply the MCLT to the random sequence $\boldsymbol{W}_i = \boldsymbol{X}_i u_i$ with $E[\boldsymbol{X}_i u_i] = \boldsymbol{0}$ and $Var(\boldsymbol{X}_i u_i) = \boldsymbol{\Omega} = E[u_i^2 \boldsymbol{X}_i \boldsymbol{X}_i']$, then we get

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i u_i\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}).$$

Therefore, we get

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i u_i\right) \xrightarrow{d} \boldsymbol{Q}^{-1}\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}),$$

because $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \xrightarrow{p} \boldsymbol{Q} = E[\boldsymbol{X}_i \boldsymbol{X}_i']$. Since $Var[\boldsymbol{Q}^{-1}\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega})] = \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$, we have the following central limit theorem for the OLS estimator:

Central Limit Theorem for OLS

Consider the general linear regression model Equation 10.1 under assumptions (A1)–(A4). Then, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}).$$

A direct consequence is that the robust t-statistic is asymptotically standard normal:

$$T_j^{(hc)} = \frac{\hat{\beta}_j - \beta_j}{se_{hc}(\hat{\beta}_j)} \xrightarrow{d} \mathcal{N}(0,1).$$

Also note that the t-distribution t_{n-k} approaches the standard normal distribution as n grows. Therefore, we have

$$t_{n-k} \xrightarrow{d} \mathcal{N}(0,1)$$

and we can write

$$T_j^{(hc)} = \frac{\hat{\beta}_j - \beta_j}{se_{hc}(\hat{\beta}_j)} \overset{a}{\sim} t_{n-k}.$$

This notation means that $T_j^{(hc)}$ is asymptotically t-distributed. I.e., the distributions of $T_j^{(hc)}$ becomes closer to a t_{n-k} distribution as n grows.

Therefore, it is still reasonable to use t-quantiles in robust confidence intervals instead of standard normal quantiles. It also turns out that for smaller sample sizes, confidence intervals with t-quantiles tend to yield better small sample coverages that using standard normal quantiles.

10.9 CASchools data

Let's revisit the test score application from the previous section and compare HC-robust confidence intervals:

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

```
coefci(fit1, vcov=vcovHC)
```

```
                2.5 %   97.5 %
(Intercept) 678.371140 719.4948
STR          -3.310516  -1.2491
```

```
coefci(fit2, vcov=vcovHC)
```

```
                 2.5 %      97.5 %
(Intercept) 668.7102930 703.3541961
STR          -1.9604231  -0.2421682
english      -0.7112962  -0.5882574
```

```
coefci(fit3, vcov=vcovHC)
```

```
                 2.5 %      97.5 %
(Intercept) 689.0614539 711.2384604
STR          -1.5364346  -0.4601833
english      -0.1869188  -0.0562281
lunch        -0.5951529  -0.4995380
```

The confidence intervals for `STR` in the first three models do not cover 0 and are strictly negative. This gives strong statistical evidence that the marginal effect of `STR` on `score` is negative, holding `english` and `lunch` fixed.

```
coefci(fit4, vcov=vcovHC)
```

```
                   2.5 %       97.5 %
(Intercept) 645.329067184 686.64732942
STR          -0.882408250   0.41163186
english      -0.192981575  -0.06370184
lunch        -0.592410029  -0.50037547
expenditure   0.001738419   0.00550568
```

In the fourth model, the point estimator for the marginal effect of `STR` is negative, but the confidence interval also covers positive values. Therefore, there is no statistical evidence that the marginal effect of `STR` on `score` holding `english`, `lunch`, and `expenditure` fixed.

However, as discussed in the previous section, `expenditure` is a bad control for `STR` and should not be used to estimate the effect of class size on test score.

10.10 R-codes

statistics-sec10.R

11 Hypothesis testing

11.1 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

$$\underbrace{H_0 : \beta_j = \beta_j^0}_{\text{null hypothesis}} \quad \text{vs.} \quad \underbrace{H_1 : \beta_j \neq \beta_j^0}_{\text{alternative hypothesis}}. \tag{11.1}$$

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.

$$P(\underbrace{T_0 \in I_1}_{\text{reject}} \mid H_0 \text{ is true}) = \alpha, \tag{11.2}$$

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

A test that satisfies Equation 11.2 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 11.1: 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.

11.2 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 $H_0 : \beta_2 = 3$.

11.3 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`.

11.4 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.

11.5 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.

11.6 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.

11.7 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**:

$$\begin{aligned} \text{do not reject } H_0 \quad & \text{if } F \leq F_{(1-\alpha,q,n-k)}, \\ \text{reject } H_0 \quad & \text{if } F > F_{(1-\alpha,q,n-k)}, \end{aligned}$$

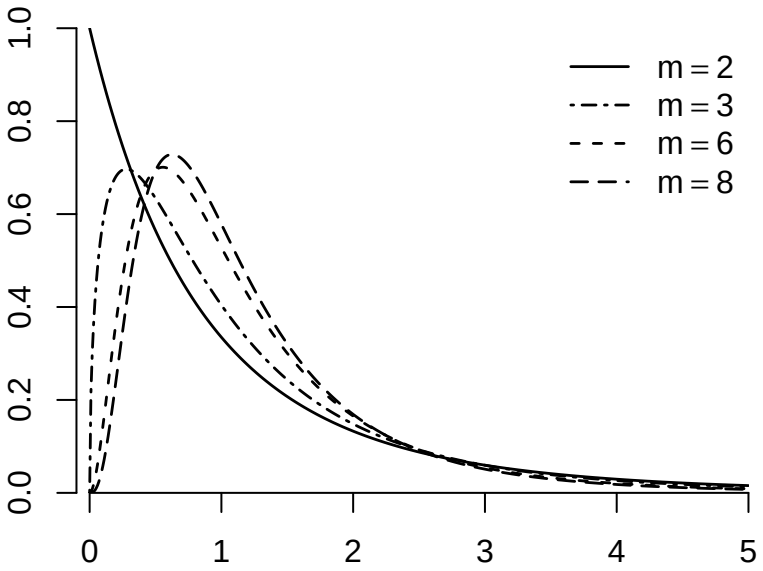


Figure 11.1: F-distribution

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.

11.8 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.

11.8.1 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.

$$\hat{u}_i^2 = \boldsymbol{X}_i'\boldsymbol{\gamma} + v_i, \quad i = 1, \dots, n. \tag{11.3}$$

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 11.3. 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.

11.8.2 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.

11.9 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 11.2

	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.

11.10 R-codes

statistics-sec11.R