Data Analysis for Public Affairs with R

Jerome Dumortier

09 January 2024

Contents

1 Preface

This book serves as an introduction to data analysis for public affairs with R and RStudio. Most examples are drawn from public affairs and economics. Throughout the book, the following notation is used:

- *variablename*: Variable names which represent the columns in a data frame
- `dataset`: Name of the data set in the file DataAnalysisPAData.RData
- ***Exercise Name***: Title of the exercise in the various chapters
- package: Link to the package documentation. The example link is associated with "Linear Models for Panel Data."

The variable definitions and the data sets in DataAnalysisPAData.RData are described in the section Data Sources.

2 Introduction

The purpose of this chapter is to provide an introduction and overview about the various aspects of uncertainty:

- Usefulness of learning about probability, statistics, and regression models
- Installing R and RStudio

There is also a YouTube video and slides associated with this chapter:

- Introduction to Probability and Statistics - Video
- Introduction to Probability and Statistics - Slides

You are also encouraged to review a video on some Basic Mathematical Tools.

There are a few books that may serve as a reference:

- *A Modern Introduction to Probability and Statistics - Understanding Why and How* by Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester.

- *Introductory Econometrics - A Modern Approach* by Jeffrey M. Wooldridge. This book is the standard literature in many regression classes.
- *Basic Econometrics* by Damodar N. Gujarati. An excellent book that can serve as a reference for many years.

2.1 Overview

We are surrounded by probability and statistics on a daily basis because the world around us is uncertain. The purpose of probability theory and statistics is to explain and model stochastic processes such that predictions can be made. Probability and the application of statistics occur basically everywhere. For example, if you order something online, other products are suggested to you. Those suggestions are not random but are based on how you compare to other shoppers interested in similar items. Consider the following examples:

- **Grades**: If you take a university class then the grade you receive in the class is uncertain at the beginning of the semester. You may attach different probabilities associated with the various grades based on your knowledge about the material.
- **911 calls**: While getting my graduate degree at Iowa State University, I was standing at a red light one morning which had a fire station down the road. Two fire trucks with their sirens on arrived at the red light and departed in opposite directions. Thus, two 911 calls must have come in at the same time requiring two trucks from the same station. The fire station has three trucks and as a public safety manager, you may be interested in the probability of more than three trucks being requested.

- **Basketball free throws**: Just because there are two outcomes does not mean that the probability is 50%/50%. Stephen Curry is the career leader in terms of free throw percentage (90.56%). Either he scores or misses, and his success rate is far from 50%.
- **Polls**: Especially before elections, polls are very popular to determine which candidate is favored. The polling results usually include a so-called margin of error which is an indicator of confidence in the results. The chapter on confidence intervals explains how the margin of error is calculated.
- **Hurricanes**: Projected pathways of Hurricanes, e.g., Sandy in 2012, produced by the National Hurricane Center (NHC) include so-called cones of uncertainty. The NHC defines the cone of uncertainty as follows: *The cone represents the probable track of the center of a tropical cyclone, and is formed by enclosing the area swept out by a set of circles along the forecast track (at 12, 24, 36 hours, etc.). The size of each circle is set so that two-thirds of historical official forecast errors over a 5-year sample fall within the circle. Based on forecasts over the previous 5 years, the entire track of a tropical cyclone can be expected to remain within the cone roughly 60-70% of the time. It is important to note that the area affected by a tropical cyclone can extend well beyond the confines of the cone enclosing the most likely track area of the center.*
- **COVID-19**: A recent example is the COVID-19 risk assessment chart developed by the Texas Medical Association. The risk categories can be thought of as probabilities of contracting COVID-19 for the activities listed. They also updated their chart to account for COVID Risks of Various Holiday Activities.

If you are working and receiving retirement benefits, you are likely investing those in mutual fund. The saying "do not put all your eggs in one basket" applies in this context. The figure below shows the evolution of Vanguard 500 Index Fund Investor Shares (VFINX) and the Fidelity Select Retailing Portfolio (FSRPX). Although not perfectly, the funds generally move in the same direction.

Both graphs indicate a certain degree of positive association between the returns, i.e., if one of the mutual funds increases, the other tends to increase as well (and vice versa).

This course can be subdivided into three large topics: probability, statistics, and regression. The basics of probability provide means for modeling populations, experiments, and any other random phenomena. You will be introduced to probability distributions that allow you to model random outcomes. Probability theory

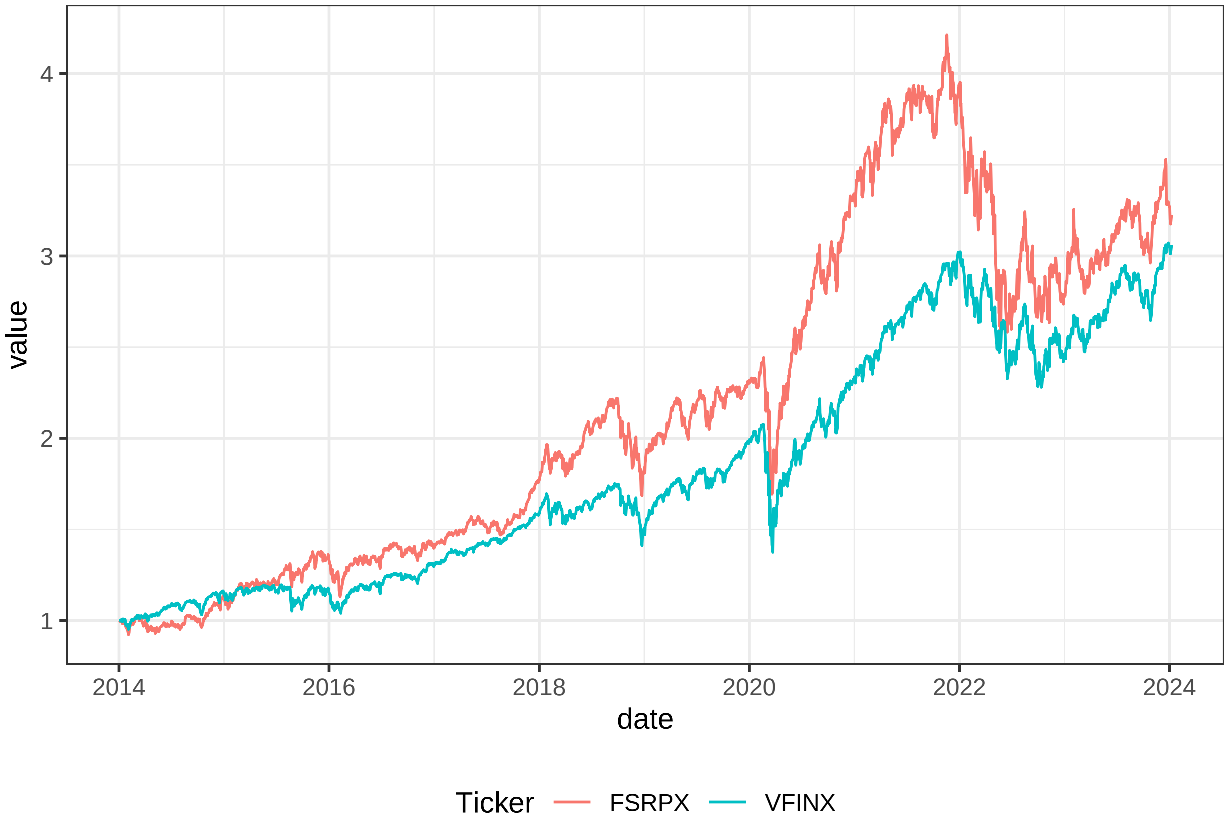


Figure 1: Evolution of the Vanguard 500 Index Fund Investor Shares (VFINX) and Fidelity Select Retailing Portfolio (FSRPX)

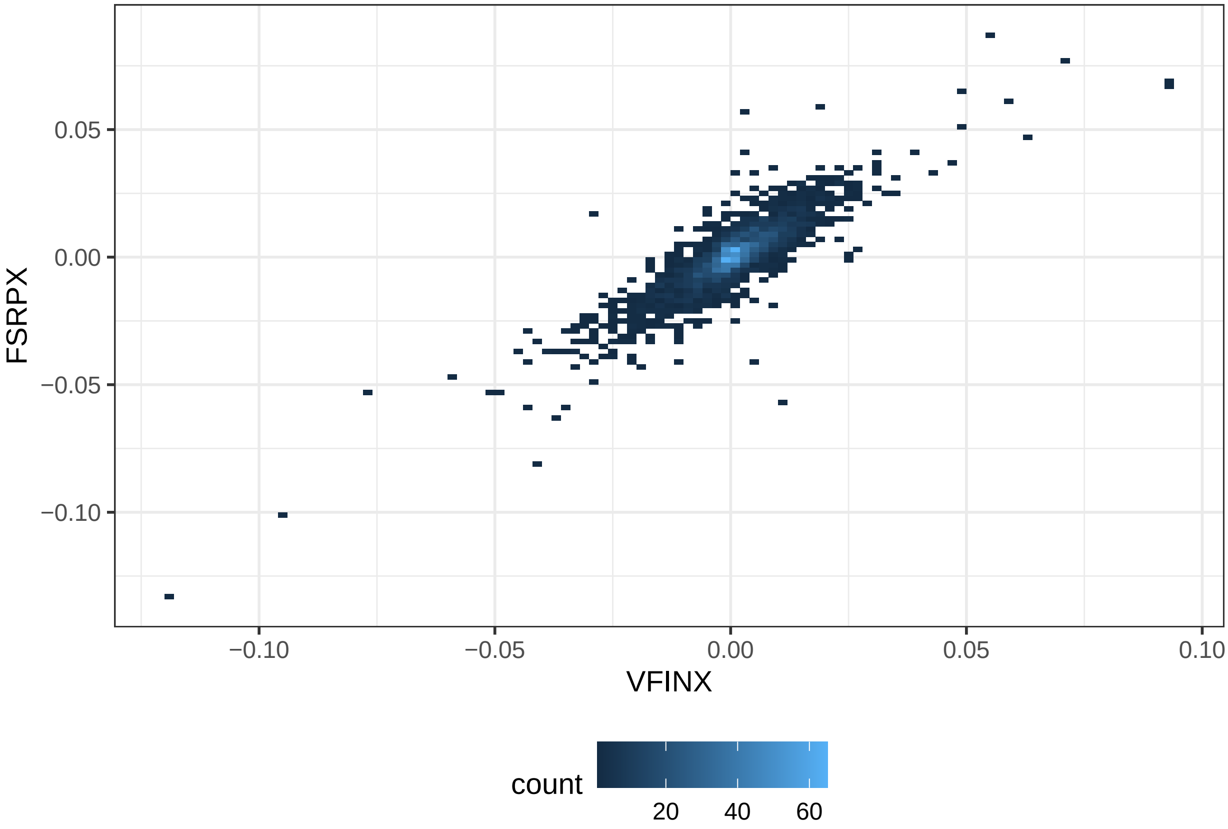


Figure 2: Scatter plot and indication of frequency of the daily returns of VFINX and FSRPX

is also the foundation for statistics. Statistics allows us to learn something about the population based on a sample. Sampling distributions, confidence intervals, and hypothesis testing will be important concepts. The last part will cover regression analysis which states mathematical relationships among variables. For example, the price of a used car can be expressed as a function of model, year, mileage, and cylinders.

To illustrate the difference between probability and statistics let us consider two buckets. The first bucket illustrates the concept of probability and the second bucket illustrates the concept of statistics. Suppose that in the first bucket, you have a bunch of balls of different colors: Green, black, and red. And you also know how many of each color are in the bucket. Probability theory tells you the expected number of green, red, and black balls in your hand after pulling out a bunch of balls from the bucket. It will tell you the likely distribution of colors in your hand.

Statistics is different. Again, you have a bucket but you do not know what is in the bucket. Think about the content of the bucket as your population with unknown characteristics. To learn about the characteristics of the population, you pull out a sample from the bucket. Based on the distribution of colored balls in your hand, you can use statistics to say something of the characteristics of your population, that is, the content of the bucket in this case.

2.2 R and RStudio

Throughout this book, we are going to use the software R to conduct statistical analysis. We will use RStudio to interact with R and you can think of RStudio as a graphical user interface for R. R/RStudio has a similar data setup to Excel but instead of seeing the spreadsheet all the time, the spreadsheet with your data is in the background. R/RStudio requires that your columns be the variables and that the rows are your observations. For convenience, we are going to use Excel for data setup. For R/RStudio, there are great online resources:

- ETH Zurich: If you are looking for documentation about all the various functions in R, this is the website to check them. Note that in most cases, the result of a Google search leads to the ETHZ page.
- StatMethods: This webpage contains a lot of tutorials and introduces you to the basic functionality of R.
- Statistical tools for high-throughput data analysis (STHDA): The site was great to learn how to do data visualization with the R package ggplot2.

R/RStudio has several advantages over Excel:

- R/RStudio is set up to do statistical analysis. Excel is easy to use but has very limited capabilities.
- It is important for your future employer to know that you have been exposed to a modern statistical software besides Excel. You might be the one that introduces a specialized statistical software to your workplace. The advantage of R/RStudio is that it is a free and very powerful statistical software. R is a software that requires some programming and understanding of computer languages but there are almost no limits of what you can do.
- Getting a graduate degree should go beyond simple Excel job training and should expose you to something new.

2.2.1 Preparation for R/RStudio

The next lecture will introduce you to the use of R and RStudio. We will use RStudio to interact with R and you can think of RStudio as a graphical user interface for R. To focus on the use of R and RStudio during the lecture, some easy preparatory steps are necessary for you to perform before class. Those are mostly related to installing R and RStudio and to load sample data into the software. With this document, you should have downloaded the small dataset honda.csv. In preparation for the lecture, you will load honda.csv into R and RStudio.

2.2.2 Installing R and RStudio

You must first install R on your computer by doing the following:

- Go to The R Project for Statistical Computing and download the R version that is appropriate for your computer. This is either the “base” version if you have a Windows computer or the “Latest release” .pkg file if you are using Mac OS. Once you have downloaded the file, install R on your computer.
- Go to RStudio and download the RStudio version that is appropriate for your computer. Note that the various “Installers for Supported Platforms” are at the bottom of the page. Once you have downloaded the file, install RStudio on your computer.

Note that we will only be using RStudio which runs R in the background. You cannot use RStudio without having R installed first. Throughout the lecture, I will be referring to R/RStudio.

2.2.3 Locating Files on your Computer

To import data into R/RStudio, you must know (1) where files are located on your computer and (2) what the current working directory of R/RStudio is. On a windows computer, the directory where files are stored is like `C:\Users\Jerome\Documents\R Lecture` and for Mac OS it is similar to `/Users/Jerome/R Lecture`.

Think of the working directory as the folder on your computer in which R/RStudio is looking for files by default. After opening R/RStudio, you can type `getwd()` in the console window and R/RStudio will return the current working directory. Usually, you have project specific working directories. For this class, create a directory on your computer in which you are going to store all the files associated with this class. You should download the dataset `honda.csv` into the directory you have created for this class. You can use the command `setwd()` to change the R/RStudio to the new working directory. Note - and this is an oddity with R/RStudio - you have to replace the backslash with a forward slash if you are a Windows user, i.e., use `setwd("C:/Users/Jerome/Documents/R Lecture")`. Assuming the `honda.csv` file is in the directory you have set above, use `honda= read.csv("honda.csv")` to load the file into R/RStudio. The data should appear in the Environment tab on the right side in R/RStudio. It is important that you can do the steps described above before the lecture to alleviate any issues at the beginning. A good video explaining the concept of file path can be found here.

3 Introduction to R

Topics covered in this lecture

- Introduction to R and RStudio
- Data Management
- Plotting and Graphs with R
- Basic Statistics

There is also a YouTube video and slides associated with this chapter:

- Introduction to R and RStudio - Video
- Introduction to R and RStudio - Slides

3.1 R Resources and Help

Very large user community for R. Google search for “Some topic R” usually leads quickly to the desired help. Here are the links to a few online tutorials

- UCLA Institute for Digital Research and Education
- StatMethods

Two online resources will provide you the solution to the vast majority of your R questions. Getting on those websites is usually the result of a Google search.

- Statistical Data Analysis R: This resource contains the function manual for R/RStudio including all packages. Example for a function boxplot. The most helpful part are the examples at the bottom of the page.

- Stack Overflow: Resources for developers. For example, a Google search for “r ggplot two y axis” may give you the following result. Note that all questions on Stack Overflow have to be accompanied by a re-creatable dataset.

There are also many R books on GitHub:

- Principles of Econometrics with R
- Introduction to Econometrics with R
- Geocomputation with R
- Introduction to Data Science
- Forecasting: Principles and Practice

Besides many online resources, there are also two useful textbooks:

- Applied Econometrics with R by Christian Kleiber and Achim Zeileis
- Introductory Statistics with R by Peter Dalgaard
- An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

An additional online tutorial is Using R for Introductory Statistics by John Verzani. If you prefer a video, the following Introduction to R and RStudio has been proven useful for people learning R/RStudio.

3.2 Opening RStudio

Work in RStudio is done in four windows:

1. Script Window
    - This is were you type your R Script (.R) and where you execute commands.
    - Comparable to do-file/editor in Stata.
    - This window needs to be opened by File ⇒ New File ⇒ R Script.
2. Console window
    - Use of R interactively. Should only be used for quick calculations and not part of an analysis.
3. Environment
    - Lists all the variables, data frames, and user-created functions.
    - It is tempting to use the “Import Dataset” function . . . Don’t.
4. Plots/Packages/Help

There is a base version of R that allows doing many calculations but the power of R comes through its packages. To use functions associated with a particular package, click “Install’ ’ in the packages window of RStudio and type in the name of the package. Or alternatively, use

- `install.packages("ggplot2")`

To use a package, you have to activate it by either checking the box in the window “Packages” or by including `library(packagename)`. Those packages are updated on a regular basis by users.

The # allows you to include comments in your script file that are not read by R. It is good practice to start any new script with clearing the memory using the command `rm(list=ls())`. Use the command `get()` to determine the current working directory or set the new working directory with the command `setwd()`, e.g., `setwd("E:/")`. For file paths, replace \ with /. Next, you want to load all libraries necessary for your entire script file with the command `library()`. It is also good practice to save your R-script on a regular basis. The frontmatter, i.e., the top of a R-script file, could look as follows

```r
rm(list=ls())
load("DataAnalysisPAData.RData")
library(openxlsx)
```

3.2.1 In-class Exercise 1

Create a R-script file with the following components:

1. Two lines for the title and the date (use #)
2. Clearing all current contents
3. Setting the correct working directory
    - This should be a folder to which you have downloaded all materials.
4. Installing and loading the package `openxlsx`.

3.3 Functions

At the core of R are functions that “do things” based on your input. The basic structure is

- `object = functionname(argument1=value,argument2=value,...)`

The structure has the following components

- `object`: Output of the function will be assigned to object.
- `functionname`: Name of the system function. You can also create and use your own functions. More about this later.
- `argument`: Arguments are function specific.
- `value`: The value you want a particular argument to take.

If a function is executed without an specific assignment, the output will be displayed in the console window. Before using a function, read the documentation. Many functions have default settings. Be aware of default values. In most cases, those defaults are set to values that satisfy most uses. For example, consider the help file for the function t.test.

- `t.test(x,y=NULL,...,mu=0,conf.level=0.95,...)`

For this function we have the following default values

- `y=NULL`
- `mu=0`
- `conf.level=0.95`

3.4 Data in R

The main data types which can appear in the Environment window of R are:

- Vectors
    - `preselection = seq(1788,2016,4)`
    - `midterm = seq(by=4,to=2018,from=1790)`
- Matrix
    - `somematrix = matrix(8,10,4)`
    - Only numerical values are allowed.
- Data frames
    - By far, the most common data type in R.
    - Comparable to an Excel sheet.
    - More on this later.
- Lists
    - Collection of objects from of various types.
    - `myfirstlist = list(preselection,midterm,somematrix)`

3.4.1 Using R as a Calculator

Entering heights of people and storing it in a vector named **`height`**:

```
height = c(71,77,70,73,66,69,73,73,75,76)
```

Calculating the sum, product, natural log, mean, and (element-wise) squaring is done with the following commands:

- `sum(height)`
- `prod(height)`
- `log(height)` # Default is the natural log
- `meanheight = mean(height)`
- `heightsq = height^2`

Removing (i.e., deleting) unused elements: `rm(heightsq,meanheight)`

3.4.2 Creating a Data Frame from Scratch

Data frames are the most commonly used tables in R/RStudio. They are similar to an Excel sheet.

- Column names represent the variables and rows represent observations.
- Column names must be unique and without spaces.

Suggestion: Use only lower-case variable names and objects.

```
studentid       = 1:10
studentnames    = c("Andrew","Linda","William","Daniel","Gina",
                    "Mick","Sonny","Wilbur","Elisabeth","James")
students        = data.frame(studentid,studentnames,height)
rm(studentid,height,studentnames)
```

3.4.3 In-class Exercise 2

Create a data frame called `students` containing the following information:

Name	Economics	English
Mindy	80.0	52.5
Gregory	60.0	60.0
Shubra	95.0	77.5
Keith	77.5	30.0
Louisa	97.5	95.0

Notes:

- Use *name* as the column header for the students' names.
- Once you have created the data frame, remove the unused vectors.

3.4.4 Indexing

Indexing refers to identifying elements in your data:

- For most objects: `students[row number,coloumn number]`
  - `students[3,2]` returns 95. What does `students[3,]` return?
- If you want to select certain columns: `students[c("name")]`
  - Other example: `students[c("name","english")]`
- Selecting results based on certain conditions: `students[which(students$economics>80),]`

Referring to a particular column in a data frame is done through the dollar symbol:

- `students$english`
- You will use this functionality very often.

Creating a new column: `students$average = rowMeans(students[c("economics","english")])`

3.4.5 Importing Data into R

In almost all cases, the data is imported into R from an external data set. The data has to be "machine-readable" which means that the first row must contain the variable names and the actual data starts in the second row. Machine-readable data can be imported as follows:

- `read.csv("filename.csv")`: If you have a comma separated value (.csv) file then this is the easiest and preferred way to import data.
- `readWorkbook(file="filename.xlsx",sheet="sheet name")`: Requires the package openxlsx. Note that there are many packages reading Excel and this is one of the most reliable and user-friendly.
- Importing data from other software packages (e.g., SAS, Stata, Minitab, SPSS) or .dbf (database) files can be achieved using the package foreign. The package works also for Stata data up to version 12. To important data from Stata version 13 and above, the package readstata13.

3.4.6 Sub-setting a Data Frame

To extract variables or observations based on certain criteria, the command `subset()` must be used. Consider the data `vehicles`. Extracting vehicle information only for the year 2015 is done as follows:

- `cars2015 = subset(vehicles,year==2015)`

Note that the double equal sign conducts a logical test. This is similar to Stata. Using a single equal sign does not extract any data and simply returns the original data without (!) an error message.

To list all the distinct values in a column, the command `unique()` can be used. This command only makes sense in the case of categorical data in a particular column. For example, listing all EPA vehicle size classes (*VClass*) can be accomplished as follows:

- `unique(cars2015$VClass)`

Suppose you are only interested in the variables *ghgScore* and *VClass* for the model year 2015.

- `cars2015 = subset(vehicles,year==2015,select=c("ghgScore","VClass"))`

Suppose you are only interested in "Compact Cars" and "Large Cars" in the column *VClass* for the year 2015. There the notation is a bit odd (note that the many line breaks are not necessary to include in R):

```
cars = subset(vehicles,
              year==2015 & vclass %in% c("Compact Cars", "Large Cars"),
              select=c("make","co2tailpipegpm","vclass"))
```

3.4.7 In-class Exercises 3

From the vehicles data set, extract the GHG Score and the vehicle class from the 2014 model year for the following manufacturers: Toyota, Ford, and Audi. Your new data set should contain the following columns: *ghgScore*, *make*, and *VClass*. Is the resulting data frame sensible or do you see a problem?

3.4.8 Aggregating Data

To aggregate data based on a function, e.g., sum or mean:

```
cars = aggregate(cars$co2tailpipegpm,FUN=mean,by=list(cars$make,cars$vclass))
```

3.4.9 Writing Data Frame to .csv-File

To write data to the current working directory:

- `write.csv(cars2014,"cars2014.csv",row.names=FALSE)`

Using the option `row.names=FALSE` avoids an index column in the output file.

3.4.10 Reshaping Data from Long to Wide and Viceversa

R has the ability to reshape data from long to wide format and back. For this demonstration, we use the data in `compactcars` and the command `reshape()` from the package reshape2:

```
cars = melt(compactcars,id=c("year","make","model","displ","drive"))
```

Reshaping the data is generally very useful but also tricky. For detailed information see the section How can I reshape my data in R.

3.4.11 Extending the Basic `table()` Function

The required package to extend the basic `table()` function is called gmodels. Compare the outputs of the functions `table()` and `CrossTable()`. The commands below do the following:

- Standard `table()` function.
- The function `CrossTable()` includes the proportions along the two dimensions of gun ownership and gender. Note that a description of the cell content is at the top of the results page.

```
table(gss$owngun,gss$sex)
```

```
##
##            female male
##   no          741  505
##   refused      20   30
##   yes         291  302
```

```
CrossTable(gss$owngun,gss$sex,prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1889
##
##
##              | gss$sex
##   gss$owngun |    female |      male | Row Total |
## -------------|-----------|-----------|-----------|
##           no |       741 |       505 |      1246 |
##              |     0.595 |     0.405 |     0.660 |
##              |     0.704 |     0.603 |           |
##              |     0.392 |     0.267 |           |
## -------------|-----------|-----------|-----------|
##      refused |        20 |        30 |        50 |
##              |     0.400 |     0.600 |     0.026 |
##              |     0.019 |     0.036 |           |
##              |     0.011 |     0.016 |           |
## -------------|-----------|-----------|-----------|
##          yes |       291 |       302 |       593 |
##              |     0.491 |     0.509 |     0.314 |
##              |     0.277 |     0.361 |           |
##              |     0.154 |     0.160 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1052 |       837 |      1889 |
##              |     0.557 |     0.443 |           |
## -------------|-----------|-----------|-----------|
##
##
```

Note that for almost any R command, you can store the output by assigning it to an object:

- `somename = CrossTable(gssgun$owngun,gssgun$sex,prop.chisq=FALSE)`

3.4.12 Merging Datasets

Consider two data sets from school districts in Ohio:

- `ohioscore` contains an identifier column IRN and a score that indicates the quality of the school.
- `ohioincome` contains the same identifier than the previous sheet in addition to median household income and enrollment.

To merge the two data frames based on the column IRN, the function `merge()` must be used:

```
ohioschool = merge(ohioscore,ohioincome,by=c("irn"))
```

4 Summarizing Data

The purpose of this section is the description of data using numerical and graphical methods. Numerical methods include measures of central tendency and dispersion. Graphical methods presented are histograms, box plots, and empirical cumulative distribution functions. The last part presents the concepts of correlation and covariance.

There is also a YouTube video and slides associated with this chapter:

- Summarizing Data - Video
- Summarizing Data - Slides

4.1 Measures of Central Tendency

The three main measures of central tendency are mean, median, and mode. The mean reports the average value of a data set. Sometimes, it is also called arithmetic mean or average and can be expressed as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{N} x_i$$

The mean does not always explain data very well and there are some other measures to explore. Among them are the median and mode. The median divides the data set into two equal parts, such that half of all the values are greater than the median and half are smaller than the median. The middle term of the data set is the $(n+1)/2$-th term. The median has the advantage that a very high or very low values do not influence the median. For example, the average taxable income in Switzerland is $56,805. The town Vaux-sur-Morges has an average taxable income of $670,046. It turns out that one very wealthy person (Andre Hoffmann from Roche Pharmaceutical) drives up the average in the village with 178 inhabitants. The mode is the value that occurs the most frequently. Consider the table below on the income distribution of ten citizens in three states. Note that the mean income in all three states is 10. Although, there is considerable variation among the citizens. The median is 10, 9.5, and 2 for state A, B, and C, respectively. The mode is 10, 10, and 2 for state A, B, and C, respectively.

Table 1: Income distribution among citizens in three states.

A	B	C
10	2	2
10	3	2
10	7	2
10	8	2
10	9	2
10	10	2
10	10	2
10	14	2
10	17	2
10	20	82

4.2 Measures of Dispersion

The easiest measure of dispersion is called the range which is the largest value minus the smallest value in the data set. A measure used far more often is the variance. Think of the variance as a measure describing how far the data is spread around the mean. Recall from Chapter 1 that there we distinguish between a population and a sample. The population is characterized by unknown parameters and we use statistics based on sample to say something about the unknown population parameters. The distinction between population and sample is important regarding the variance. The population variance is calculated as follows:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

The sample variance is calculated as follows:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

The population variance is used to calculate the variance of the entire population. For example, if a professor is interested in calculating the variance associated with the final exam, they would use population variance equation because the class is not a sample of a larger population. The second equation dividing by $N-1$ is used if we have a sample and want to estimate (!) the variance associated with the population. Consider the figure below which illustrates the difference. A randomly generated data set contains 100,000 observations with $\mu = 50$ and $\sigma = 20$. An simulation procedure takes 1000 sample with sample sizes varying from 2 to 50 (horizontal axis). For each of the sample, the population variance is estimated by dividing by either N or $N-1$. The estimates dividing by *N-1* are closer to the population standard deviation than for the samples divided by *N*.

The coefficient of variation standardizes the standard deviation σ by the mean, i.e., $CV = \sigma/\mu$. Because the magnitude of the standard deviation depends on the mean, it is sometimes necessary to calculate the coefficient of variation to make two or more standard deviations comparable. For example, suppose you are comparing residential home values in California and Indiana. You calculate the mean and standard deviation for California as $2,000,000 and $400,000, respectively. The mean and standard deviation for Indiana are $125,000 and $50,000, respectively. Calculating the coefficient of variation (CV) for both states leads to $CV_{CA} = 0.2$ and $CV_{IN} = 0.4$. Hence, there is much more variation in the home values for Indiana than there is in California.

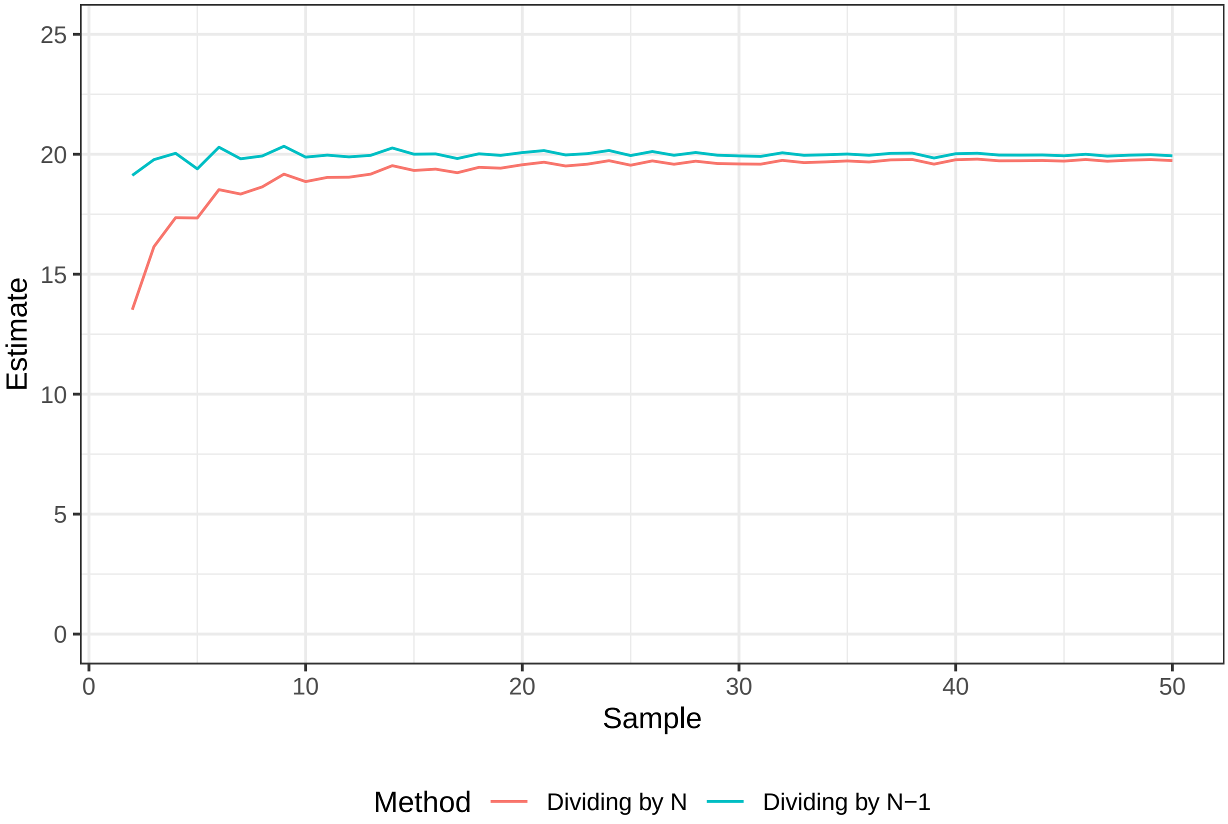


Figure 3: Difference between dividing by N and N-1 to estimate the variance.

4.3 Histograms

In almost all cases, a visual inspection of the data is appropriate. Although numerical statistics such as mean and/or standard deviation exist, many fail to easily identify patterns or anomalies. Histograms are probably the most basic method to graphically summarize data and are also good approximations of the probability distributions. Suppose you have a data set $x_1, x_2, \ldots, x_n$. You divide the range of values into bins. For example, if a data set ranges from 75 to 155, then a possible bin size could be 10, i.e., 75, 85,..., 155. The height of the bar for each bin is determined by the number of observations in the bin range. Those bins usually have the same size but it is not necessary. Consider the eruption and waiting times of Old Faithful geyser in Yellowstone National Park. Panel (a) and (b) represent the histograms of eruption and waiting time, respectively.

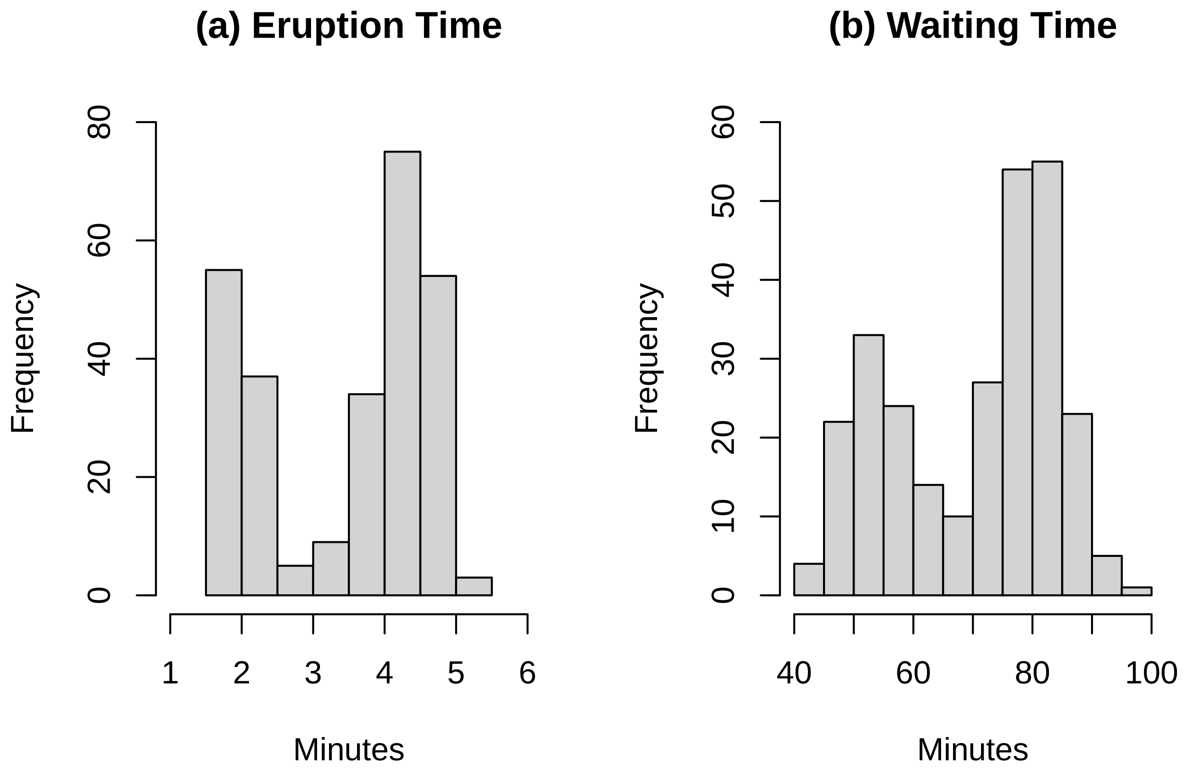


Figure 4: Histogram of eruption and waiting time of Old Faithful geyser

To draw a histogram of the data as shown in the above figure, use the folling command:

- `hist(faithful$eruptions)`

Note that you can control the appearance of the histogram by using options:

- `hist(faithful$eruptions,main="Eruption",xlab="Minutes",xlim=c(1,6),ylim=c(0,80))`

4.4 Empirical Cumulative Distribution Function

The empirical cumulative distribution function can be written as

$$F_n(x) = \frac{\text{number of elements} \leq x}{n}$$

Given the previously mentioned Old Faithful data, the empirical cumulative distribution function is shown shown below. An important measure of empirical cumulative distribution functions are quantiles. Quantiles are the values that cut x% of the distributional area. Below the quantile is a certain share of all values in a set of ordered observations. If the distribution is divided in *n* equal shares, there are *n-1* quantiles. Commonly used quantiles are quartiles, quintiles, deciles, and percentiles. Quartiles divide the sample into 4 equal shares, i.e., 3 quartiles: 25%-, 50%-, 75%- quartile. Quintiles divide the sample into 5 equal shares, deciles into 10 equal shares, and percentiles divide the sample into 100 equal shares. Closely related to quartiles is the *interquartile range (IQR)*.

```r
faithful = data.frame(faithful)
par(mfrow=c(1,2))
     plot(ecdf(faithful$eruptions),main="Eruption Time",xlab="Minutes",xlim=c(1,6),lty=1)
     plot(ecdf(faithful$waiting),main="Waiting Time",xlab="Minutes",xlim=c(40,100),lty=1)
```

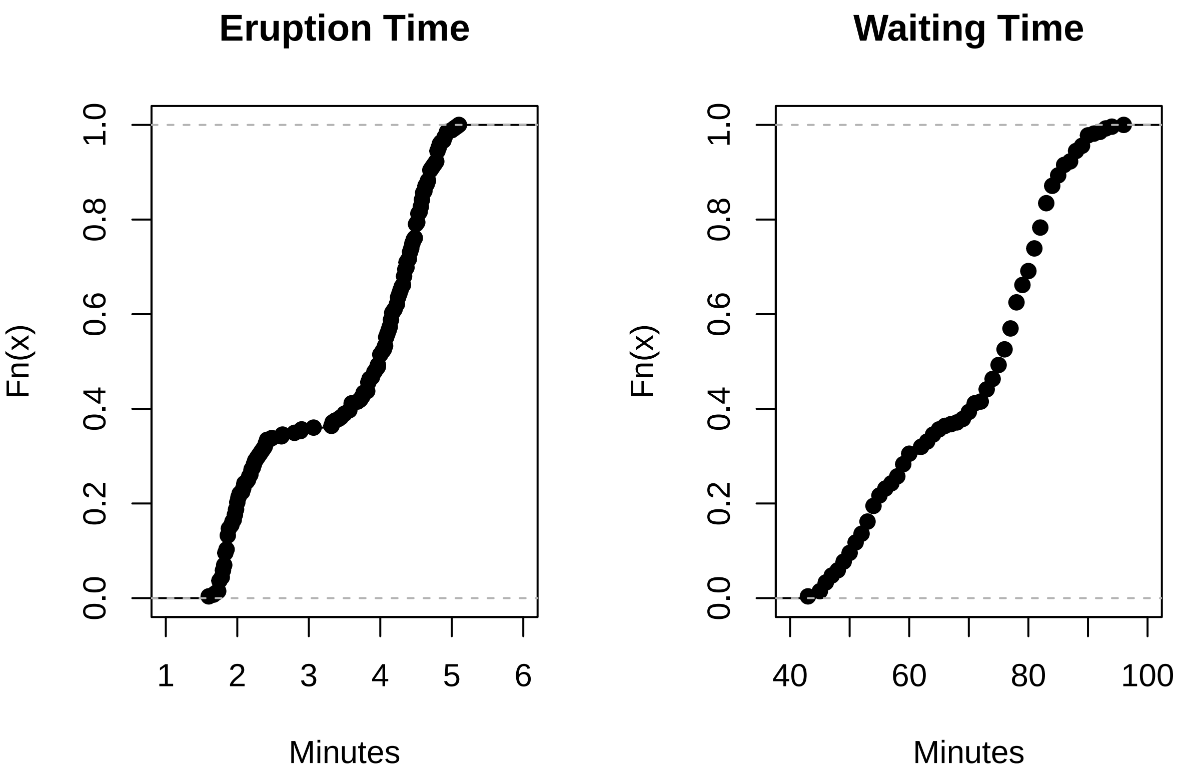


Figure 5: Empirical cumulative distribution functions of eruption and waiting time of Old Faithful geyser

4.5 Boxplots

Outliers are values that extremely deviate from the mean, i.e., extremely high or low values. Those values limit the explanatory power of the mean. Reasons for outliers can be error in measurement, error of judgement, miscalculation, typo in data matrix, real extraordinary values. To treat outliers, we must know whether they are true values or a mistake. Outliers can be neglected under certain conditions, but it is important to bear them in mind when interpreting results!

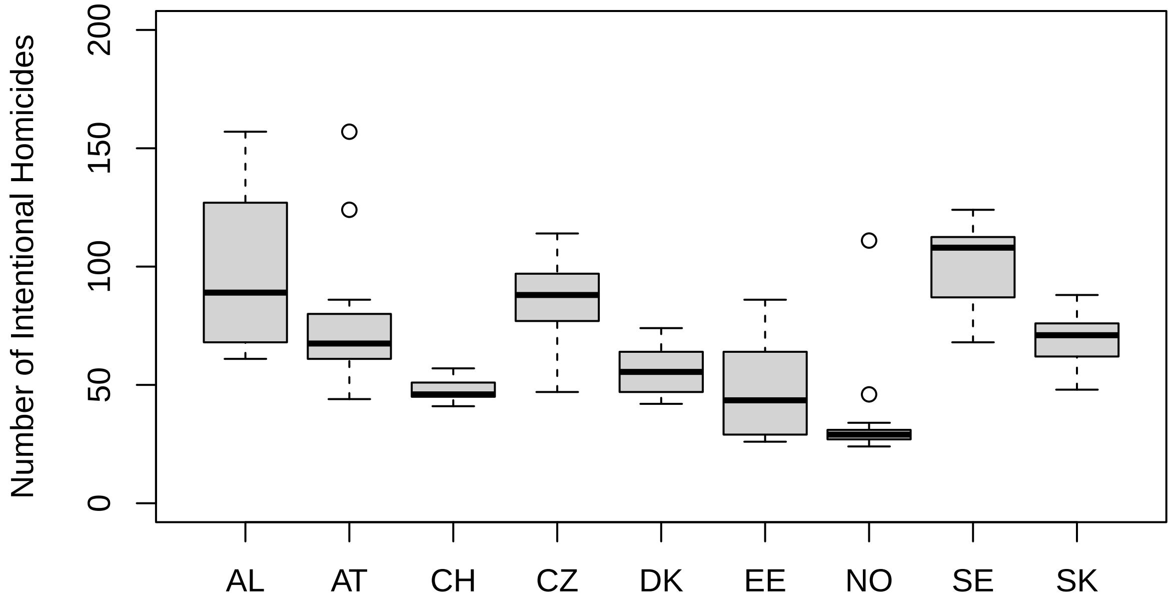


Figure 6: 2008-2018 Intentional homicides in select European countries, i.e., Albania (AL), Austria (AT), Switzerland (CH), Czech Republic (CZ), Denmark (DK), Estonia (EE), Norway (NO), Sweden (SE), and Slovakia (SK) Source: Eurostat

4.6 Covariance and Correlation Coefficicent

The previous sections focused on one random variable at a time. Very often, we have more two or more random variable and we are interested in their relationship. We will analyze the relationship between variables from a causal standpoint in the section on regression analysis. In this chapter, we focus on two variables and how they behave jointly. For now, we will not make any statements about causality. The important part here: Correlation is not causation! As for the variance, there are two definitions/equations to calculate the covariance between two variables:

$$Cov(x,y) = E[(x - E(x)) \cdot (y - E(y))] = E(x \cdot y) - E(x)E(y)$$

If the sign of the covariance is positive, then x and y tend to move in the same direction, i.e., if one variable increases, the other variable increases as well. If the sign of the covariance is negative, then x and y tend to move in opposite directions, i.e., if one variable decreases, the other increases. If X and Y are independent, then $Cov(X,Y) = 0$. The covariance has several properties:

- Property 1: $Var(X+Y) = Var(X) + Var(Y) + 2 \cdot Cov(X,Y)$
- Property 2 (Transformation of the covariance): $Cov(r \cdot X + s, t \cdot Y + u) = r \cdot t \cdot Cov(X,Y)$

The most important aspect about correlation (and statistics in general): Correlation does not mean causation. Causation requires a strong theoretical believe that one variable is the cause of another variable, e.g., influence of education on income. The correlation coefficient (sometimes called Pearson's r) is defined as

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

The correlation coefficient varies between -1 and 1. The Sign provides the direction of the relationship between two variables and the Value provides the magnitude of the relationship. Note that the correlation coefficient has no dimensions!

A second example uses the data `mh2` and the resulting scatter plot is shown below.

The function `summary()` gives you mean, median, and quartiles for the eruption and waiting times. The functions `var()` and `cor()` calculate the variance, covariance, and correlation coefficient associated with the data sets. We will see in subsequent chapters how to interpret the covariance and correlation coefficients.

```r
faithful = data.frame(faithful)
summary(faithful)
```

```
##    eruptions        waiting    
##  Min.   :1.600   Min.   :43.0  
##  1st Qu.:2.163   1st Qu.:58.0  
##  Median :4.000   Median :76.0  
##  Mean   :3.488   Mean   :70.9  
##  3rd Qu.:4.454   3rd Qu.:82.0  
##  Max.   :5.100   Max.   :96.0
```

```r
var(faithful)
```

```
##           eruptions   waiting
## eruptions  1.302728  13.97781
## waiting   13.977808 184.82331
```

```r
cor(faithful$eruptions,faithful$waiting)
```

```
## [1] 0.9008112
```

Note that sometimes, we deal with qualitative data, i.e., data that is not expressed as a number but as an expression. Example are gender (male/female), owning a car (yes/no), modes of commute (car, bike, train, bus, etc.). Consider the data in `gssgun`. One way to count the responses for firearm ownership is to use the following command:

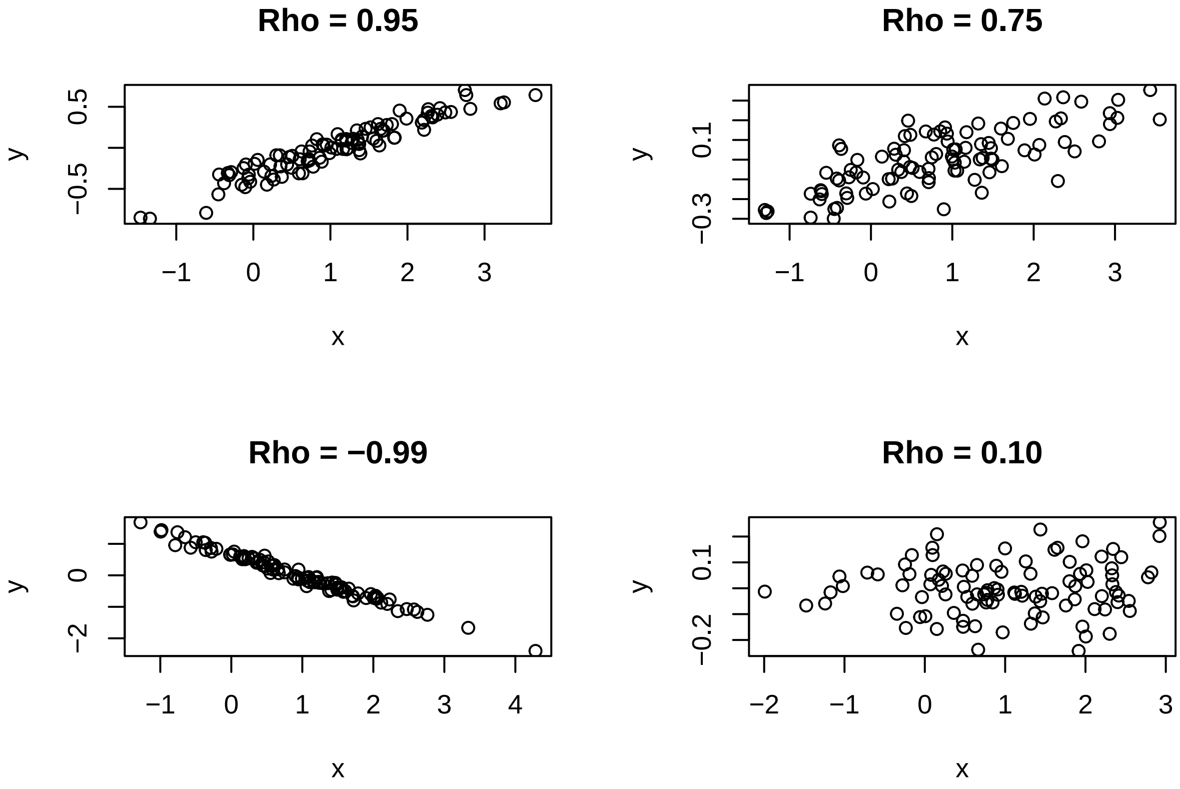


Figure 7: Examples of various correlation coefficients

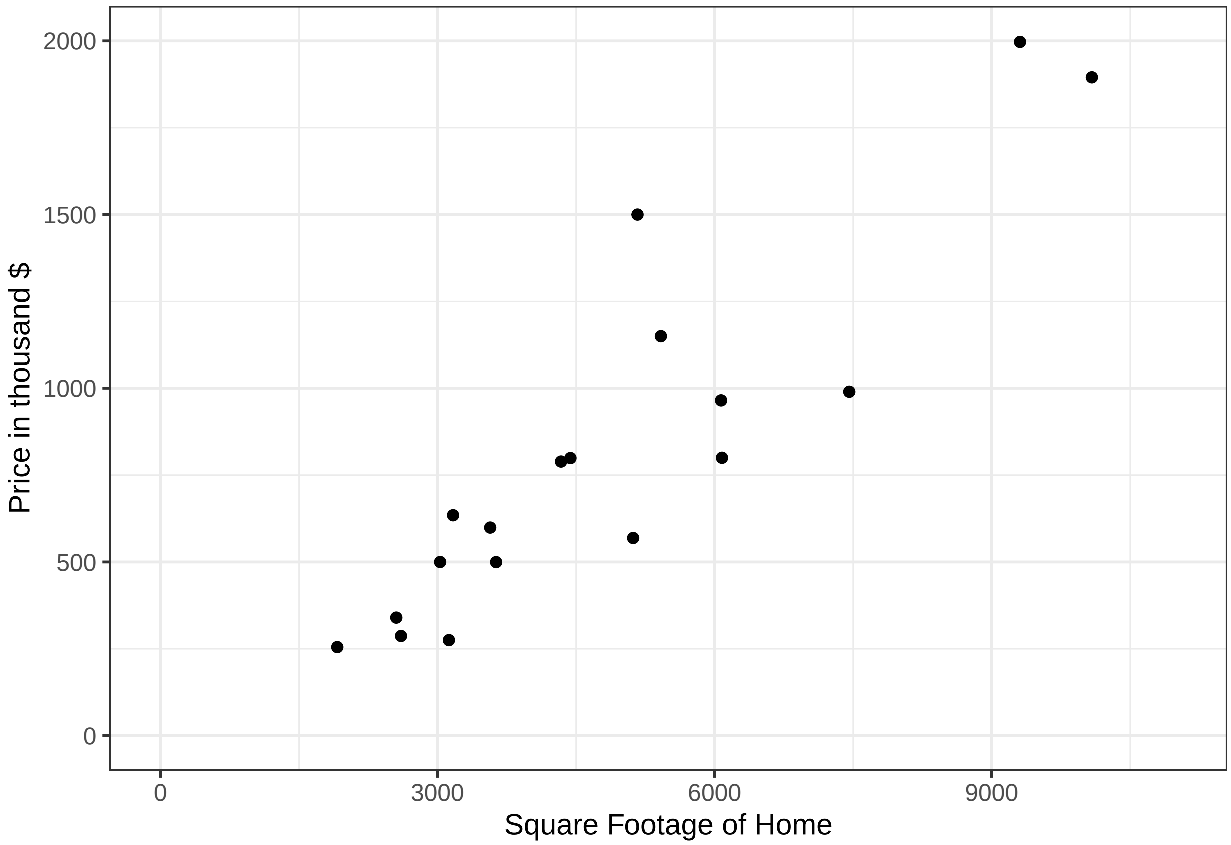


Figure 8: Correlation between the square footage of a home and the price of the home in the Meridian Hills neighborhood in Indianapolis.

- `table(gss$owngun)`

Suppose you are only interested in people who answer *yes* or *no* for the questions concerning arrests and firearms ownership. You will have to subset the data.

```r
CrossTable(gss$owngun,gss$sex,prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1889
##
##
##              | gss$sex
##   gss$owngun |    female |      male | Row Total |
## -------------|-----------|-----------|-----------|
##           no |       741 |       505 |      1246 |
##              |     0.595 |     0.405 |     0.660 |
##              |     0.704 |     0.603 |           |
##              |     0.392 |     0.267 |           |
## -------------|-----------|-----------|-----------|
##      refused |        20 |        30 |        50 |
##              |     0.400 |     0.600 |     0.026 |
##              |     0.019 |     0.036 |           |
##              |     0.011 |     0.016 |           |
## -------------|-----------|-----------|-----------|
##          yes |       291 |       302 |       593 |
##              |     0.491 |     0.509 |     0.314 |
##              |     0.277 |     0.361 |           |
##              |     0.154 |     0.160 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1052 |       837 |      1889 |
##              |     0.557 |     0.443 |           |
## -------------|-----------|-----------|-----------|
##
##
```

4.7 Exercises

1. ***Natural Gas Usage and Temperature*** (**): Use the data `indyheating` for this exercise. Construct a scatter plot with temperature on the horizontal axis and natural gas usage on the vertical axis. What do you observe? How would you expect the graph to look like if electricity consumption was on the vertical axis?

2. ***Exempt Organizations*** (***): Consider the data in `exemptorgs`. You are going to calculate the average revenue across the so-called NTEE codes. Proceed as follows:

   a. Subset the data such that you are only left with the columns *revenue* and *ntee*.

b. Use the function `na.omit()` on the data set you created in (a). What is the function used for?
c. Use the function `aggregate()` and calculate the mean assets, revenue, and income by NTEE code. The table created in your answer should contain the NTEE codes in plain English, that is, not the alphabetical codes.

3. ***GSS Two-Way Table*** (**): In the data set `gss`, pick two of the following variables: *fulltime*, *government*, *married*, *vote*, *gun*, *deathpenalty*. Construct a cross-table similar to the one about gender and guns in the section *Introduction to R.* using the function `CrossTable()`. Ignore the Chi-square statistic but explain if you see any pattern that is of interest.

4. ***Airport Delays Boxplot*** (***): Pick an airport (not IND) and year of your choice in the data set `airlines`. You should be using the function `subset()` to pick year and airport. Next, add a column called *delay* which is the share of delays from all the arriving flights. Next, construct a boxplot with all the airlines on the x-axis, i.e., one boxplot, and the variable *delay* on the vertical axis. Interpret the boxplot. Are there airlines which are particularly on time or always late?

5. ***Housing Price Index*** (***): Pick one of 49 U.S. States (not Indiana). Go to the FRED webpage and download two data series: (1) All-Transactions House Price Index for the United States (USSTHPI) and (2) All-Transactions House Price Index for the state of your choice. Answer the following questions.

   - What do the two data series describe?
   - Plot the two data series over time. Your result should look similar to Panel (b) in Figure **??**.
   - How do the housing prices evolve in your state compared to the United States. Do homes in your state get more expensive than the general trend? Less? How has the housing market evolved during and after the 2008 recession?

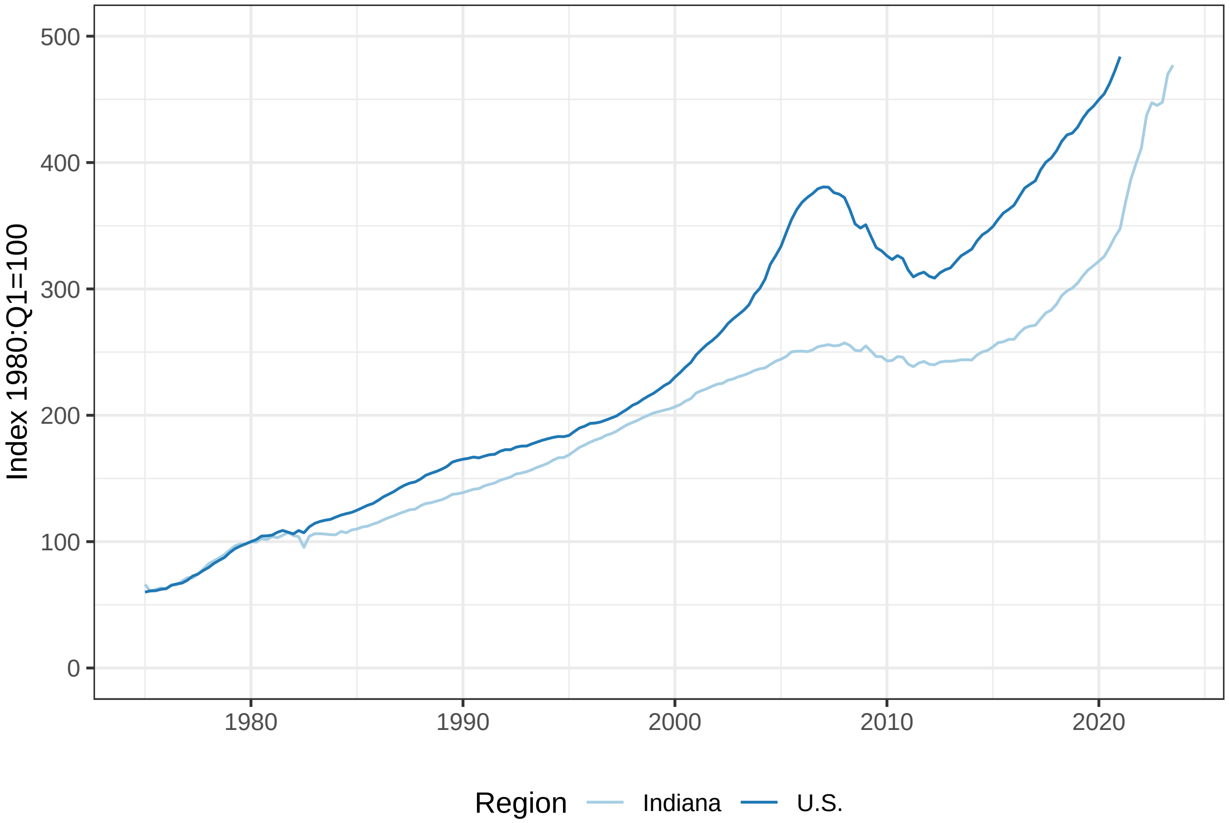


Figure 9: Evolution of the All Transaction House Price Index for the U.S. and Indiana (Source: FRED St. Louis)

6. ***BMW Boxplot*** (**): Consider the used car data set `bmw` for this exercise which contains the prices and miles of a particular BMW model in the Indianapolis area. The column *allwheeldrive* indicates whether the car has all wheel drive (1) or not (0). You must use the R/RStudio commands and not just look at the data. Answer the following questions:

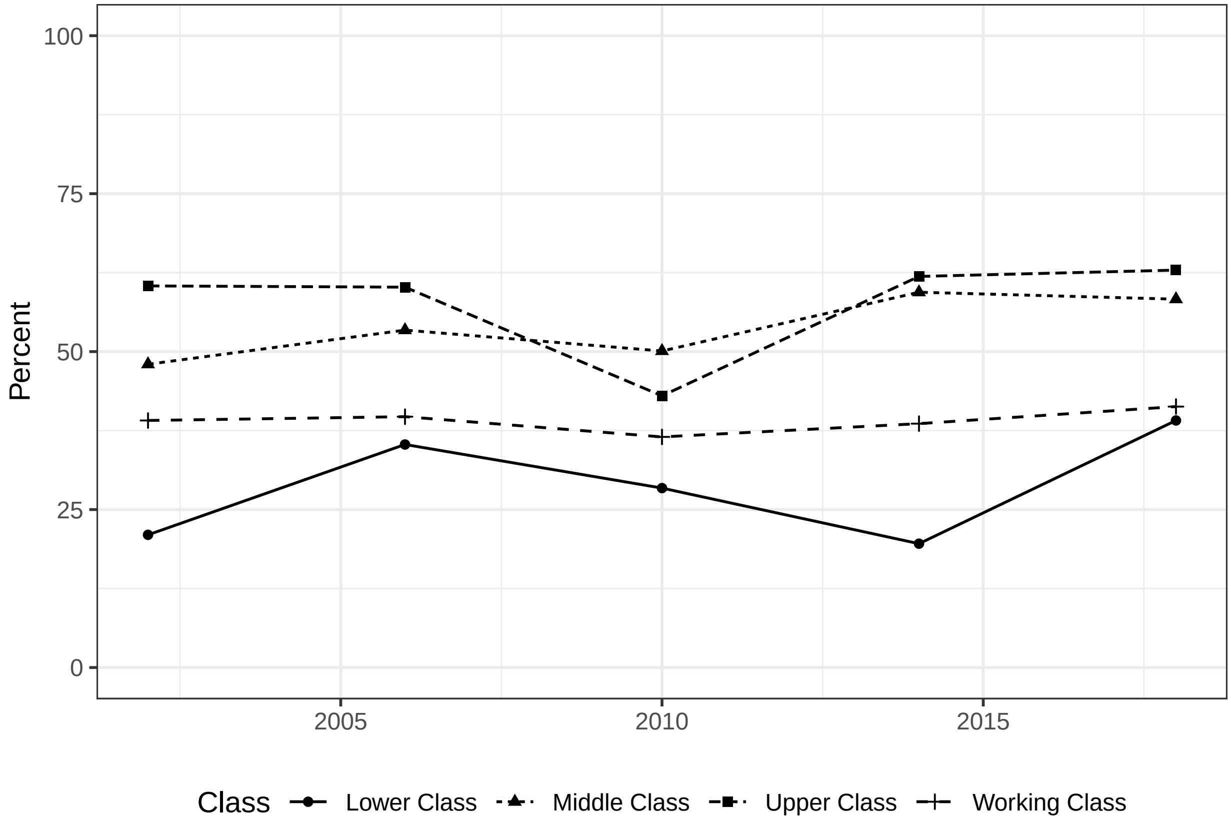
   a. Calculate the following statistics for price and miles: Minimum, maximum, median, and mean.
   b. Calculate the same statistics as in the previous part but separate the data into two groups: (1) with all-wheel drive and (2) without all-wheel drive.
   c. Use the data on prices only. Create a box-and-whisker plot grouped by all-wheel drive. This must be one graph.

7. ***State Capitals*** (**): The data `apartments` contains data about the monthly rent and size of furnished apartments in Berlin and Potsdam, which are not only adjacent cities but are also capitals of two states in Germany. Add a column that shows the rent per square meter. Then, construct a box plot of the rent per square meter on the y-axis and the location on the x-axis. Based on the boxplot, how does the rent per square meter differ between the two cities. Any other observation, which you are able to make?

8. ***EPA Fuel Economy*** (**): Use the 2018 data contained in `vehicles` for this problem.

   a. Generate a scatter plot of the variables `displ` and `comb08U`. What can you say about the shape of the scatter plot and the relationship between engine displacement and fuel economy.
   b. Transform both variables into their natural logarithm and plot the scatter plot. What changes?
   c. Create a table summarizing the average fuel economy by vehicle class (*vclass*) for the following four manufacturers: (1) Ford, (2) Chevrolet, (3) Toyota and (4) Honda. You will have to use the function `aggregate()` for this.

9. ***GSS Trends*** (**): The General Social Survey is an annual survey tracking societal trends in the United States. For a more detailed description, you can go to here. In this exercise, I want you pick a particular issue and one breakdown (e.g., age, sex, political affiliation) and plot a graph with R/RStudio on how attitudes about the issue have evolved over time. I suggest, you go to trends and pick a topic you are interested in. You will see that the website already provides you with a trend graph but I want you to use R/RStudio for this exercise and re-create the graph (i.e., you get zero points if you simply copy the graph provided on the website). For example, suppose you are interested in the variable *rincblls* which asks the respondent "Do you feel that the income from your job alone is enough to meet your family's usual monthly expenses and bills?' ' One of many possibility to plot this with R/RStudio would look like Panel (b) in Figure **??**. Make sure to label the x,y-axis correctly as well as provide a legend for the graph.



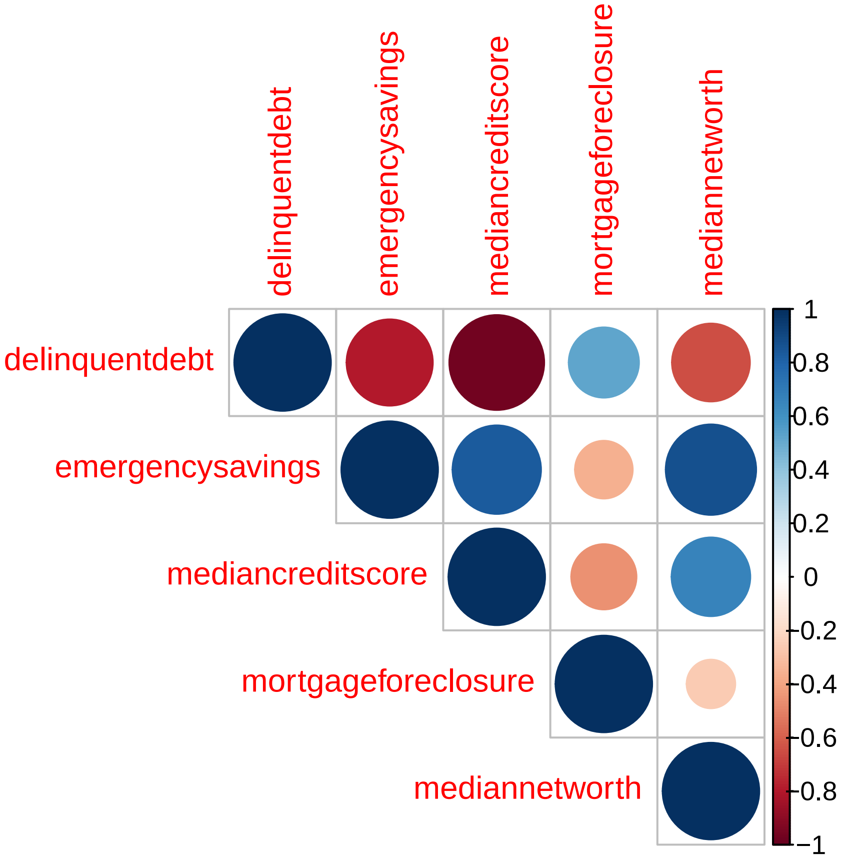
10. ***WDI Boxplot*** (**): The World Bank data `wdi` contains development indicators and also information how the World Bank classifies those countries by region and income. Focus on the years 1975, 1985, 1995, 2005, and 2015 for this question. Use the command `subset()` to extract those years. Next, you have to install and load the package `gglpot2`. The package `ggplot2` allows you to do some great data visualization. Execute the commands below. The result is a boxplot by region and by year. Based on the resulting plot, explain differences between regions in terms of life expectancy and also in terms of evolution over time.

```
wdiofinterest = subset(wdi,year %in% c(1975,1985,1995,2005,2015))
ggplot(wdiofinterest,aes(x=region,y=lifeexp,fill=as.character(year)))+
    geom_boxplot(position=position_dodge(1))
```

11. ***Faithful*** (**): The data set `faithful` which is included with R contains data about the eruption and waiting times in minutes between eruptions from Old Faithful geyser in Yellowstone National Park. Use a scatter plot to visualize the relationship between eruption and waiting time. That is, generate graph with *eruption* on the vertical axis and *waiting* on the horizontal axis. What do you observe? What is the correlation coefficient? Is there anything odd in the resulting scatter plot? Next, use R to construct a empirical cumulative distribution function.

12. ***Financial Health*** (***): Consider the data in `statefinhealth`. Execute the following commands and include the resulting figure in the homework. What does the plot display? Explain the relationship between the components. Are some of the results surprising?

```
library(corrplot)
dffin     = statefinhealth[,c(2:6)]
dffin     = cor(dffin)
corrplot(dffin,type="upper")
```



5 Probability

Probability assigns a number between 0 and 1 to a given event. The number represents the likelihood of this event occurring. To assign a probability to an event, we need to conduct an experiment which is a procedure to obtain outcomes, i.e., the results of an experiment. Suppose you want to know the probability of obtaining exactly seven heads from flipping a coin ten times. We will see later that you can easily calculate this probability but if you do not know how, you have to resort to an experiment. You have to flip a coin ten times and write down the number of heads you obtain. You then have to repeat this experiment multiple times. Each time you get one of eleven possible outcomes. If you conduct the experiment 1,000 times and 117 times the outcome is a 7, then you know the probability of obtaining seven heads from flipping a coin 10 times is around 11.7%.

There is also a YouTube video associated with this lecture:

- Probability

5.1 Sample Spaces, Outcomes, Events, and Set Operations

A sample space is a list of all possible outcomes of an experiment and is denoted Ω. Examples of sample spaces are:

- Rolling a single die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Tossing a coin: $\Omega = \{H, T\}$
- Grades: $\Omega = \{A+, A, A-, \ldots, F\}$
- Number of calls to a fire station in a 24-hour period: $\Omega = \{0, 1, 2 \ldots\}$

An event is a subset of the sample space. Examples are

- Rolling an even number: $E = \{2, 4, 6\}$
- Rolling a number less or equal to 4: $S = \{1, 2, 3, 4\}$
- More than 5 calls to the fire station: $F = \{6, 7, \dots\}$

There are several set operations and we can use Venn diagrams to illustrate them:

- **Intersection**: The intersection W of two sets X and Y is the set of elements that are in both X and Y. We write $W = X \cap Y$.
- **Empty or Null Sets**: The empty set or the null set ($\emptyset$) is the set with no elements. For example, if the sets A and B contain no common elements then these two sets are said to be disjoint, e.g., odd and even numbers: $A \cap B = \emptyset$.
- **Unions**: The union of two sets A and B is the set of all elements in one or the other of the sets. We write $C = A \cup B$.
- **Complements**: The complement of a set X is the set of elements of the universal set U that are not elements of X, and is written X^c.

For a discrete sample space Ω, the probability of an event is a non-negative number, i.e., $Pr(A) \geq 0$, for any subset A of Ω. In addition, we have $Pr(\Omega) = 1$, i.e., all the probabilities of the outcomes in the sample space sum up to 1.

For example, if we flip a coin, the sample space is $\Omega = \{H, T\}$. The corresponding probabilities are $Pr(H) = 0.5$ and $Pr(T) = 0.5$. The sum of both is equal to 1. Keep in mind that $Pr(A^c) = 1 - Pr(A)$. It is sometimes easier to calculate the complement of an event than the event itself. If $A, B, C, \dots$ is a finite or infinite sequence of mutually exclusive events — that is events which cannot happen at the same time — from the sample space Ω, then

$$P(A \cup B \cup C \cup \cdots) = P(A) + P(B) + P(C) + \cdots$$

To illustrate this, suppose you are flipping a coin three times. The eight events in Ω are

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

The probability of each event is equally likely, i.e.,

$$Pr(E_i) = 1/8$$

for $i = 1, 2, 3, \dots, 8$. If your event of interest A is that exactly two of the three tosses results are heads, then $A = \{HHT, HTH, THH\}$. By summing up the probabilities associated with those events, we find that $Pr(A)$:

$$Pr(A) = Pr(HHT) + Pr(HTH) + Pr(THH) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Note that the same result can be obtained by using a probability tree. A probability tree has the advantage that it represents the possible outcome and probabilities in a graphical manner.

5.2 Probability of a Union

For any two events A and B, we have the probability of a union (also known as the addition rule) which states that the probability of A or B occurring can be written as:

$$Pr(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If $Pr(A \cap B) = 0$, then we call them disjoint events.

For example, suppose that researchers are interested in the substances found in people's blood. Let $A = \{Alcohol\}$, $B = \{Cocaine\}$, and $A \cap B = \{Both\}$. Assume that the probabilities are as follows:

$$Pr(A) = 0.86 Pr(B) = 0.35 Pr(A \cap B) = 0.29$$

Hence, $Pr(A \cup B)$ is calculated as follows:

$$Pr(A \cup B) = 0.86 + 0.35 - 0.29 = 0.92$$

If you are intrigued by this example on how to measure drug consumption, check out this EU Project.

For a second example, assume the following events taken from the police by controlling semi-trucks: $A = \{\text{faulty breaks}\}$, $B = \{\text{bad tires}\}$, and $A \cup B = \{\text{faulty breaks and/or bad tires}\}$. Let the probabilities be as follows:

$$Pr(A) = 0.23 Pr(B) = 0.24 Pr(A \cap B) = 0.09$$

Using the above equation, we can determine that $P(A \cup B) = 0.23 + 0.24 - 0.09 = 0.38$. Note that if the events are mutually exclusive, the term $Pr(A \cap B)$ is equal to 0.

Lastly, consider the data below on the gate arrival of airplanes during a week at a mid-sized airport. Everything not within +/- 10 minutes is considered not on time. *Flights* represents the total number of arriving flights which is 275.

Arrival	Event	Flights	Probability
Less than 10 minutes early	A	55	0.20
Within +/- 10 minutes	B	121	0.44
More than 10 minutes late	C	99	0.36

The probability of a flight not begin on time is calculated as follows:

$$Pr(A \cup C) = 0.2 + 0.36 = 0.56$$

5.3 Probability of an Intersection

To find the probability that events A and B occur, we have to use the multiplication rule (i.e., probability of the intersection) which is written as follows:

$$Pr(A \cap B) = P(A) \cdot P(B)$$

For the multiplication rule to hold, the two events must be independent! The multiplication rule for dependent events will be introduced in more detail later but can be written as follows:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B|A)$$

where $Pr(B|A)$ is the probability of A given that even B occurred.

Suppose you have 16 polo shirts in your closet with your company's logo. Nine of them are green and seven are blue. In the morning, you get dressed in the dark and randomly grab a shirt two days in a row (without doing laundry). What is the probability that both shirts are blue? Define the events B_i and G_i as grabbing a blue and green shirt, respectively on day i. In this case, the probabilities are as follows:

$$Pr(B_1) = 7/16 Pr(B_2|B_1) = 6/15$$

Thus,

$$Pr(B_1 \cap B_2) = Pr(B_1) \cdot Pr(B_2|B_1) = 7/16 \cdot 6/15 = 0.175$$

If you are interested in the probability to get a 6 on roll 1 (event A) and a 6 on roll 2 (event B). This is written as $Pr(A) \cdot Pr(B) = 1/6 \cdot 1/6 = 1/36$. This can also be expressed as a joint probability. The joint probability calculates the likelihood of two or more events happening at the same time.

Let $A = \{Hearts\}$ and $B = \{Queen\}$. Then the joint probability is the likelihood of drawing the Queen of Hearts from a deck of cards. This joint probability can be calculated as follows:

$$Pr(A) = \frac{1}{4} Pr(B) = \frac{4}{52} Pr(A \cap B) = Pr(A) \cdot Pr(B) = \frac{1}{52}$$

Let event A be that the first child is a girl and let event B be the second child being a girl. What is the sample space for the gender of the two kids? Given the sample space of $\Omega = \{FF, FM, MF, MM\}$, we have $Pr(A) = 0.5$, $Pr(B) = 0.5$, $Pr(A \cup B) = 0.5 + 0.5 - 0.25$, and $Pr(A \cap B) = 0.25$.

5.4 Conditional Probability

In the previous section, we have seen how to determine the probability of simple events. In this section, we will learn how to calculate the probability of an event knowing that some other event happened before. We call this *conditional probability*, i.e., what can we say about an event if the sample space changes? Consider the following examples:

- What is the probability of a person earning more than $150,000 given graduation from Harvard Law School?
- What is the probability of a person getting arrested given a prior arrest?
- What is the probability of getting an A in graduate statistics given an undergraduate degree in mathematics?
- What is the probability of receiving a grant from a funding agency given prior funding from the same agency?

The conditional probability of event A given that event B happened is expressed as $Pr(A|B)$. Given event B such that $Pr(B) > 0$ and any other event A, we define the conditional probability of A given B as:

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

Rearrangement of the terms from the conditional probability definition leads to the multiplication rule for dependent events:

$$Pr(A \cap B) = Pr(A|B) \cdot Pr(B)$$

Let us consider a couple of examples to illustrate this concept. Consider the following table relating the quality of the service (from an electrician) received to the years of experience.

	Good service	Bad Service	Total
Over 10 years	16	4	50
Below 10 years	10	20	30
Total	26	24	50

Define the event "good service" as G and "more than 10 years of experience" as E. What is $Pr(G)$, $Pr(G|E)$?, $Pr(E)$?

Next, consider rolling a die and the interest in the probability of a 1 given that an odd number was obtained. Let event A be "observe a 1" and event B be "observe an odd number". The probability of interest is $Pr(A|B)$. The event $A \cap B$ requires the observance of both a 1 and an odd number. In this instance, $A \subset B$ so $A \cap B = A$ and $Pr(A \cap B) = Pr(A) = 1/6$. Also, $Pr(B) = 1/2$ and, using the definition,

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

In the last example, a box with red and black balls is considered. The box contains r red balls labeled $1, 2, 3, \ldots, r$ and b black balls labeled $1, 2, 3, \ldots, b$. If a ball from the box is known to be red, what is the probability it is the red ball labeled 1, i.e., $Pr(B|A)$? Let event A be "observe a red ball" and event B be "observe a 1". The probability of A is $Pr(A) = r/(r + b)$ and the probability of a red ball with the number 1 on it is $Pr(A \cap B) = 1/(r + b)$. Then the probability that the ball is red and labeled 1 given that it is red is given by

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{1/(r + b)}{r/(r + b)} = \frac{1}{r}$$

This differs from the probability of B (a 1 on the ball) which is given by $Pr(B) = 2/(r + b)$.

5.5 Independence

Two events are said to be independent if $Pr(A \cap B) = Pr(A) \cdot Pr(B)$. The events A and B are independent if knowledge of B does not affect the probability of A. This can be written in terms of conditional probability as

$$Pr(A|B) = Pr(A)Pr(B|A) = Pr(B)$$

The probability of both events should be positive because division by zero is meaningless. Remember that $Pr(A|B) = Pr(A \cap B)/Pr(B)$ and $Pr(B|A) = Pr(A \cap B)/Pr(B)$.

Suppose you are rolling a red die and a green die. Let event A be "4 on the red die" and event B is "sum of the dice is odd". What is $Pr(A)$, $Pr(B)$, and $Pr(A \cap B)$? Are A and B independent? The table below displays all possible outcomes and can help to calculate the probabilities.

	Green					
Red	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

The probabilities are as follows:

$$Pr(A) = 6/36 = 1/6 Pr(B) = 18/36 = 1/2 Pr(A \cap B) = 3/36 = 1/12$$

To check for independence, multiply $Pr(A)$ and $Pr(B)$ as follows

$$Pr(A) \cdot Pr(B) = \left(\frac{1}{6}\right)\left(\frac{1}{2}\right) = \frac{1}{12}$$

Next consider a different scenario. Assume event C be "at least three dots on a die" and event D as "sum of dice equals 7". Are those two events independent?

$$Pr(C) = \frac{32}{36} Pr(D) = \frac{1}{6}$$

$$Pr(C|D) = \frac{Pr(C \cap D)}{Pr(D)} = \frac{6/36}{6/36} = 1$$

Thus, the two events are dependent.

5.5.1 Birthday Problem

Usually, people underestimate the probability of finding two matching birthdays in a group of people. Here the issue is that people think in terms of matching with their own birthday. That probability is indeed small (1 out of 365). But the matching birthday considers all combinations. Define the following two events: Event B_2 is that "two people have different birthdays" and event B_3 is that "three different birthdays of three people." The probability that two people have different birthdays is

$$Pr(B_2) = 1 - \frac{1}{365}$$

and that probability that three different people have different birthdays is given by

$$Pr(B_3) = Pr(A_3|B_2) \cdot Pr(B_2)$$

$$Pr(A_3|B_2) = 1 - \frac{2}{365}$$

$$Pr(B_3) = \left(1 - \frac{2}{365}\right)\left(1 - \frac{1}{365}\right)$$

In general, we have

$$Pr(B_n) = Pr(A_n|B_{n-1}) \cdot Pr(B_{n-1}) = \left(1 - \frac{n-1}{365}\right) \cdot Pr(B_{n-1}) = \left(1 - \frac{n-1}{365}\right) \cdots \left(1 - \frac{2}{365}\right)\left(1 - \frac{1}{365}\right)$$

This last probability does to zero very quickly as n increases. There are also other examples of paradoxes of probability and other statistical strangeness.

5.6 Law of Total Probability and Bayes Rule

Suppose that you are testing a cow for mad cow disease or bovine spongiform encephalopathy (BSE). Like many medical tests, there is a chance of a false-positive. Define the event B as "cow has BSE" and event T as "cow tests positive." Assume that we have the following probabilities: $Pr(T|B) = 0.7$, $Pr(T|B^C) = 0.1$, $Pr(B) = 0.02$, and $Pr(B^C) = 0.98$. What is $Pr(T) = Pr(T|B) \cdot Pr(B) + Pr(T|B^C) \cdot Pr(B^C)$? Remember from conditional probability

$$Pr(T|B) = \frac{Pr(T \cap B)}{Pr(B)} Pr(T|B^C) = \frac{Pr(T \cap B^C)}{Pr(B^C)}$$

Using the provided probabilities, this can be expressed as $P(T) = 0.7 \cdot 0.02 + 0.1 \cdot 0.95 = 0.112$. What is the probability that a cow has BSE if it tests positive, i.e., $P(B|T)$? The solution to this can be written as

$$Pr(B|T) = \frac{Pr(T \cap B)}{Pr(T)} = \frac{Pr(T|B) \cdot Pr(B)}{Pr(T|B) \cdot Pr(B) + Pr(T|B^C) \cdot Pr(B^C)}$$

Using the provided probabilities, this can be expressed as $Pr(B|T) = 0.7 \cdot 0.02/0.112 = 0.125$. This can also be explained using a probability tree.

5.7 Combinatorial Methods

5.7.1 Permutations

An ordered arrangement of k distinct objects from a total of n objects is called a permutation. The number of ways of ordering n distinct objects taken k at a time is distinguished by the symbol P_k^n:

$$P_k^n = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

where the factorial $n!$ is defined as $n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdots (2) \cdot (1)$ with $0! = 1$. With replacement of the objects, the number of possibilities is n^k.

For example, consider a bowl containing six balls with the letters A, B, C, D,E , and F on the respective balls. Now consider an experiment where I draw one ball from the bowl and write down its letter and then draw a second ball and write down its letter. The outcome is then an ordered pair, i.e., $BA \neq AB$. The number of distinct ways of doing this is given by

$$P_2^6 = \frac{6!}{4!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} = 6 \cdot 5 = 30$$

What is number of ways to arrange objects if $n = 4$ and $k = 2$? Without replacement, we have to calculate P_2^4 and with replacement, we have to calculate n^k.

5.7.2 Combinations

A arrangement of k distinct objects where ordering does not matter is called a combinations. The number of unordered subsets of size r chosen (without replacement) from n available objects is

$$\binom{n}{k} = C_k^n = \frac{P_k^n}{k!} = \frac{n!}{k! \cdot (n-k)!}$$

What is C_6^6? Consider a bowl containing six balls with the letters A, B, C, D, E, F on the respective balls. Now consider an experiment where I draw two balls from the bowl and write down the letter on each of them, not paying any attention to the order in which I draw the balls so that AB is the same as BA. The number of distinct ways of doing this is given by

$$C_2^6 = \frac{6!}{2! \cdot 4!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{6 \cdot 5}{2} = 15$$

The equation for a combination with repetition is as follows:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k! \cdot (n-1)!} = \frac{(n+k-1)!}{k! \cdot (n-k)!}$$

5.8 Exercises

1. ***Simple Probabilities*** (**): Calculate the following probabilities:

   a. Drawing an ace from a deck of cards?
   b. Getting a number divisible by 3 after rolling a die?
   c. Sum of two dice being equal to 7?
   d. Getting at least one head after flipping a coin twice?

2. ***Probability of a Union I*** (*): Consider the two events A and B which are mutually exclusive. The probabilities associated with those two events are *Pr(A)=0.23* and *Pr(B)=0.47*. What is the probability of A or B occurring? What is the probability that neither occurs?

3. ***Probability of a Union II*** (*): Consider the probability associated with the two events A and B, i.e., *Pr(A)=0.45* and *Pr(B)=0.3*. The probability that both of those events occur at the same time, i.e., $Pr(A \cap B)$ is 0.2. What is the probability of the union?

4. ***Party*** (**): Assume that 52% of the population are Republicans and 48% of the population are Democrats. On a particular issue, 64% of Republicans are in favor and 52% of Democrats are in favor. If you randomly pick a person who is in favor of the issue, what is the probability that the person is a Democrat?

5. ***Smoke Detector*** (**): Smoke and fire detectors are essential to save lives. Unfortunately, there are many false fire alarms. Suppose that the probability of an actual fire happening is very low at 5%. Smoke detectors are extremely good at detecting an actual fire, i.e., given a fire, there is a 99% probability that the smoke alarm detects it. If there is no fire, there is a 10% probability that the fire alarm sounds. Suppose that you hear a fire alarm, what is the probability that there is a fire?

6. ***Independence*** (**): Suppose you are rolling a die. Consider the events A ("rolling an even number") and B ("rolling a number less than four"). Are A and B independent?

6 Probability Distributions

There is a YouTube video and slides associated with this chapter:

- Probability Distributions - Video
- Probability Distributions - Slides

In order to associated more complex probabilities to a sample space, we need to employ probability distributions. In a first part, the concept of random variables is presented including expected value and variance of a random variable. The two remaining parts of this chapter cover probability distributions for discrete and continuous random variables, respectively. For each probability distribution, the respective names and commands in R are presented. Many probability distributions in R have a "name" (presented in parenthesis in the lists below). The following discrete probability distributions are presented:

- Bernoulli Distribution
- Binomial Distribution ("binom")
- Poisson Distribution ("pois")

Continuous distributions included in this chapter are:

- Uniform Distribution ("unif")
- Normal Distribution ("norm")
- Student Distribution or *t*-Distribution ("t")

There is a particular set of R commands to find various aspects of the distributions:

- `dname()` for the density or probability function
- `pname()` for the cumulative density function
- `rname()` for the random numbers
- `qname()` for the inverse

6.1 Random Variables

A random variable X is a variable that can take different values and there is a probability associated for each value or range of values. We differentiate between discrete and continuous random variables. The important aspect to keep in mind is that the sum of the probabilities for all values has to be equal to one!

6.1.1 Expected Value and Variance

Think of the expected value as a weighted average. Let X be a discrete random variable taking the values $x_1, x_2, \ldots$ and with probability mass function p. Then the expected value of X, $E(X)$, is defined to be

$$E(X) = \sum_i x_i \cdot P(X = x_i) = \sum_i x_i \cdot p(x_i)$$

Let X be a continuous random variable taking the values with probability density function $f(x)$. Then the expected value of X, $E(X)$, is defined to be

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

The variance can be calculated in with to different equations:

$$Var(X) = E(X - E(X))^2 = E(X^2) - E(X)^2$$

Both equations give you the variance. Sometimes one of the equations is more convenient to use. Note that $E(X^2) \neq E(X)^2$.

Suppose you roll a die and observe the number that comes up. The probability mass or frequency function is given by

$$p(x_i) = P(X = x_i) = \frac{1}{6} \quad \text{for i=1,...,6}$$

Thus, the expected value can be calculated as follows:

$$E(X) = \sum_{i=1}^{6} x_i \cdot \left(\frac{1}{6}\right) = 21/6 = 3.5$$

6.2 Discrete Probability Distributions

A random variable X is said to be discrete if it can assume only a finite or countable infinite number of distinct values. A discrete random variable can be defined on both a countable or uncountable sample space. The probability that X takes on the value k, i.e., $P(X = k)$, is defined as the sum of the probabilities of all sample points that are assigned the value k. For each value within its domain, we have $P(X = k) \geq 0$ and that the sum of all probabilities is equal to one. For example, suppose you are flipping a coin three times and associate the outcome "heads" with the number one. The results of this experiment are illustrated in the table below.

Outcome	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Pr	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
X	3	2	2	1	2	1	1	0

This experiment can also be represented with a histogram which can be turned into a frequency histogram. This is illustrated for tossing a coin ten times and repeating the experiment 10,000 times.

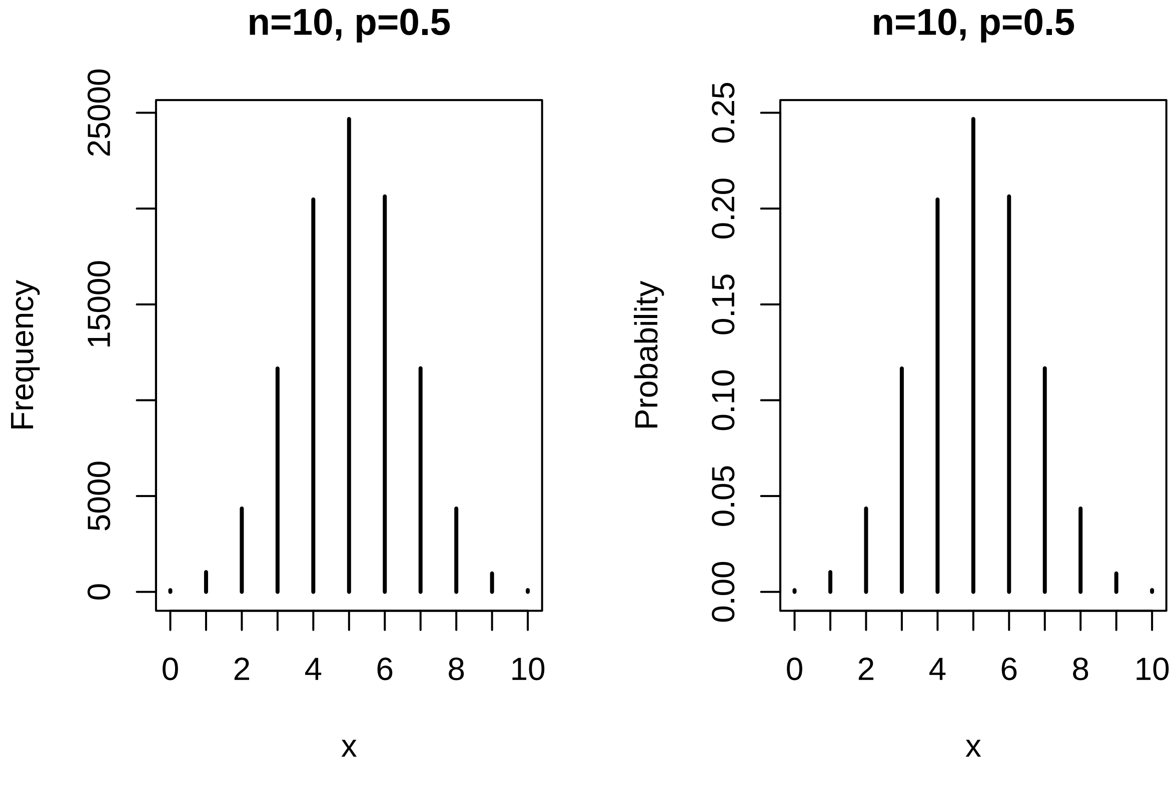


Figure 10: Tossing a coin ten times and repeating the experiment 1000 times.

6.2.1 Bernoulli distribution

The Bernoulli (named after Jacob Bernoulli, 1654-1705) distribution is the simplest probability distribution. There are only two outcomes: *Success* and *Failure*. The distribution is characterized by one parameter, i.e., p. The probability mass function is written as $P(X = 1) = p$ and, correspondingly, $P(X = 0) = 1 - p$. The expected value of the binomial is written as follows:

$$E(x) = 1 \cdot p + 0 \cdot (1 - p) = p$$

To calculate the variance, we have $E(x^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$ and $E(x)^2 = p^2$. Thus, we have

$$Var(x) = E(x^2) - E(x)^2 = p \cdot (1 - p)$$

6.2.2 Binomial distribution

The binomial distribution is closely related to the Bernoulli distribution because it represents *repeated* Bernoulli outcomes. The two parameters are n and p. The number of successes is represented by k. The probability mass function is written as

$$Pr(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

The expected value of X is $E(X) = n \cdot p$ and the variance is $Var = n \cdot p \cdot (1 - p)$. A situation must meet the following conditions for a random variable X to have a binomial distribution:

- You have a fixed number of trials involving a random process; let n be the number of trials.
- You can classify the outcome of each trial into one of two groups: success or failure.
- The probability of success is the same for each trial. Let p be the probability of success, which means $1 - p$ is the probability of failure.
- The trials are independent, meaning the outcome of one trial does not influence the outcome of any other trial.

Imagine a multiple choice test with three questions and five choices for each question. First, draw a probability tree. What is the probability of two correct responses? Next, the binomial distribution is used to calculate the probability.

$$P(X = 2) = \binom{3}{2} \cdot 0.2^2 \cdot (1 - 0.2)^{3-2} = 0.096$$

and

$$P(X = 3) = \binom{3}{3} \cdot 0.2^3 \cdot (1 - 0.2)^{3-3} = 0.008$$

Summing up both probabilities gives us $P(X \geq 2) = 0.104$.

The binomial distribution can be used to analyze the issue of overbooking. Assume that an airline as a plane with a seating capacity of 115. The ticket price for each traveler is \$400. The airline can overbook the flight, i.e., selling more than 115 tickets, but has to pay \$700 in case a person has a valid ticket but needs to be re-booked to another flight. There is a probability of 10% that a booked passenger does not show up. The results for overbooking between 1 and 31 seats are shown in the figure below.

6.2.3 Poisson Distribution

The Poisson distribution is used for count data (i.e., $0, 1, 2, ...$). The probability mass function for the Poisson distribution is given by the following equation:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

An example of the Poisson distribution (named after Simeon Denis Poisson, 1781-1840) for different parameter values is shown below.

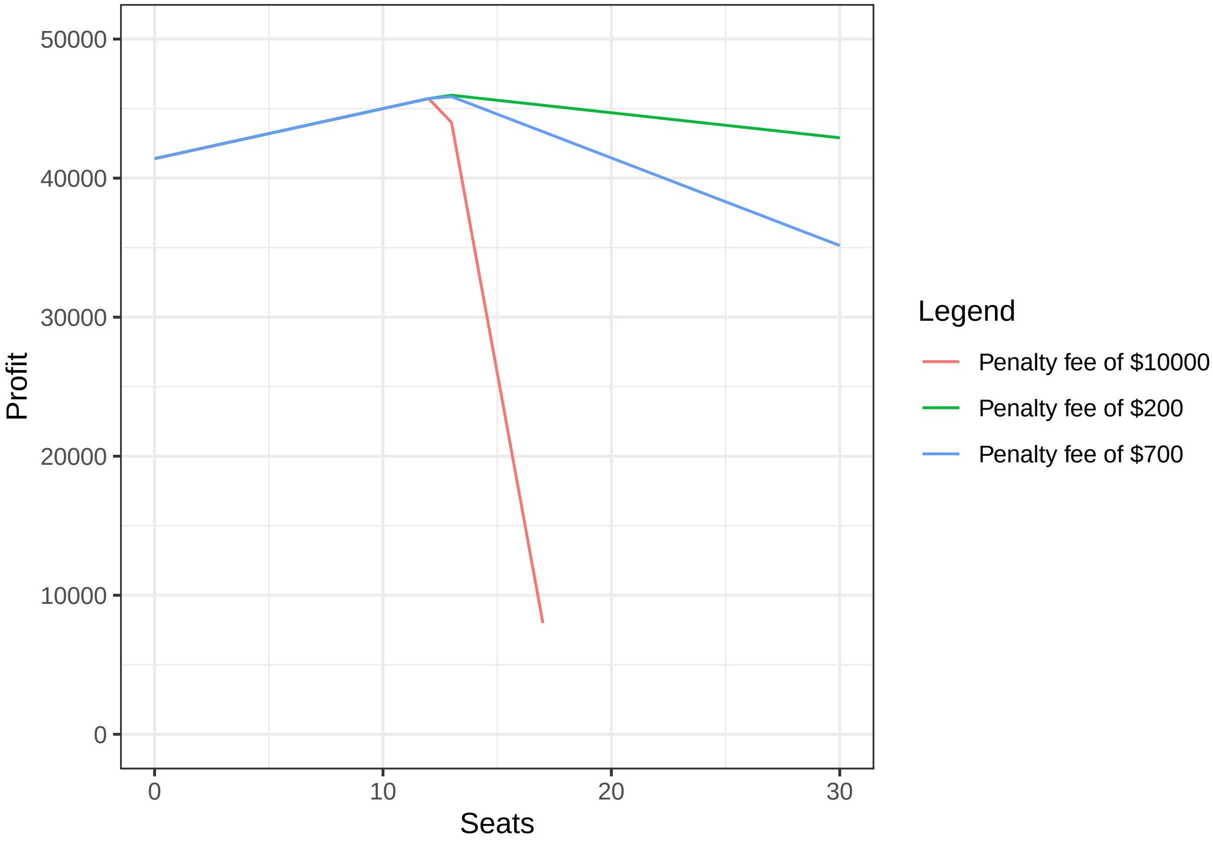


Figure 11: Example of using the binomial distribution to determine how many seats to sell over capacity.

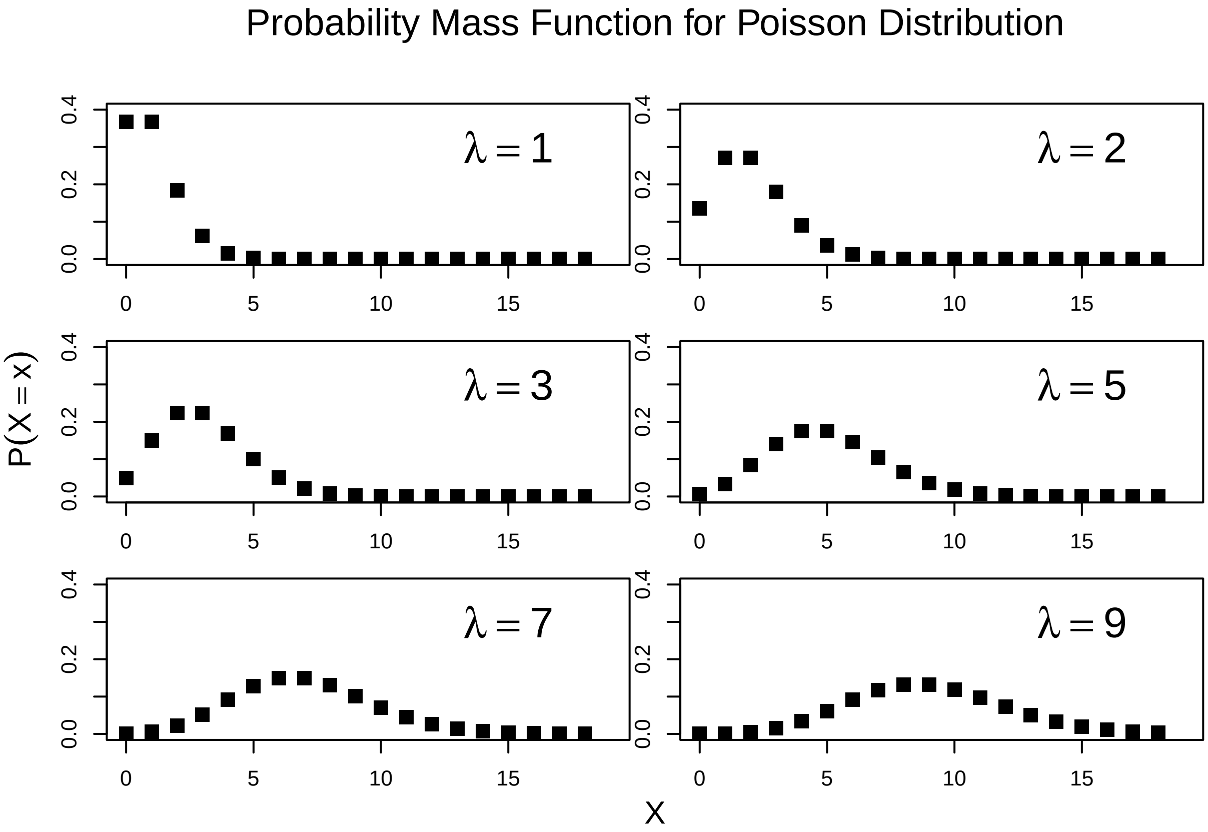


Figure 12: Poisson Distribution

6.3 Continuous distributions

We define a random variable to be continuous if $F_X(x)$ is a continuous function of x and for every real number x. The probability density function, $f_X(x)$, of a continuous random variable X is the function $f(\cdot)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

Properties are that $f_X(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

It is important to note that for a continuous distribution, the probability of a particular event is zero.

6.3.1 Uniform distribution

The simplest continuous distribution is the uniform distribution. If $a < b$, a random variable X is said to have a uniform probability distribution on the interval (a, b) if and only if the density function of X is

$$f(x) = \frac{1}{b-a}$$

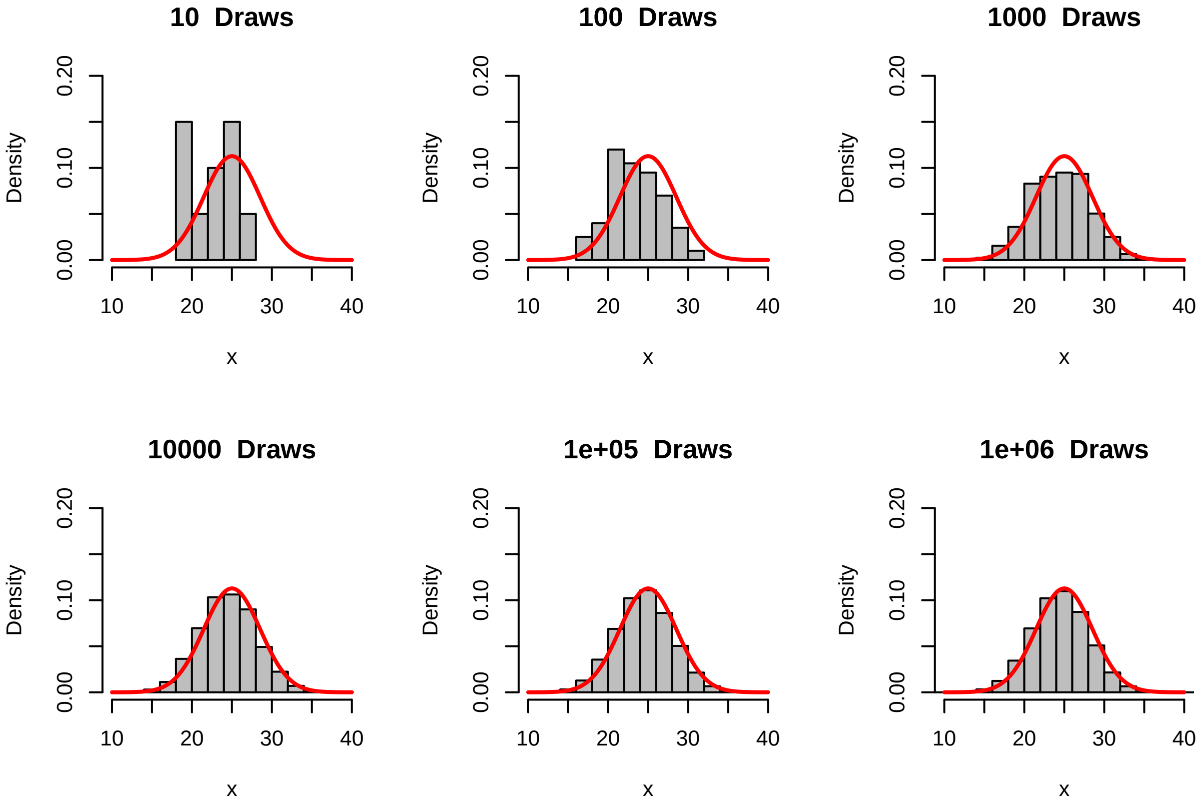
Parameters defining a uniform distribution: a, b. Suppose that a package is scheduled to be delivered between 9:00 (9am) and 20:00 (8pm). You are out for lunch between 12:00 (noon) and 13:30 (1:30pm). The probability that the package arrives during the lunch break is $P(12 < x < 13) = \frac{1}{20-9} \cdot 1.5 \approx 13.64\%$.

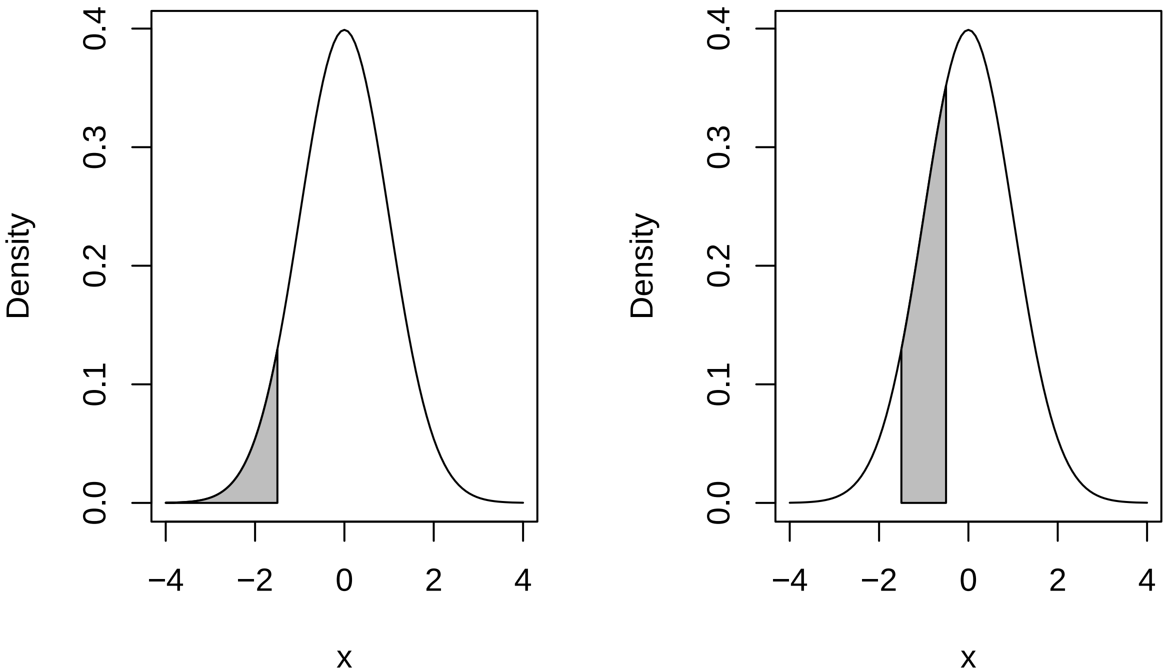
6.3.2 Normal distribution

The random variable X is said to be normally distributed with mean μ and variance σ^2 (abbreviated by $x \sim N[\mu, \sigma^2)$ if the density function of x is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Normal distribution can be derived from the Binomial Distribution.


10 Draws
100 Draws
1000 Draws
10000 Draws
1e+05 Draws
1e+06 Draws
Density
0.00
0.10
0.20
10
20
30
40
x



The normal probability density function is bell-shaped and symmetric as shown in the figure. It is usually necessary to standardize the normal distribution to make it N(0,1):

$$z = \frac{X - \mu}{\sigma}$$

where z is the distance from the mean of a Normal distribution expressed in units of the standard deviation. Suppose that we have $\mu = 75$ and $\sigma = 10$ and we want to find $P(x < 60)$ and $P(60 < x < 70)$.

$$Pr(x < 60) \Rightarrow \frac{60 - 75}{10} = -1.5 P(60 < x < 70) \Rightarrow z_1 = -1.5, z_2 = -0.5$$

The normal distribution can be derived from the binomial distribution (Galton board). Assume that the average height of males and females in the U.S. is $\mu_M = 70$ and $\mu_F = 65$ inches, respectively. The standard deviations are $\sigma_M = 4$ and $\sigma_F = 3.5$. Calculate the probabilities associated with being σ, $2 \cdot \sigma$, and $3 \cdot \sigma$ away from the mean. For more information on how the Normal Distribution evolved, see The Evolution of the Normal Distribution.

6.3.3 *t*-Distribution

Hence, we have to replace it with an estimate, i.e., with the value of the sample standard deviation. To calculate the Student t distribution, we need the degrees of freedom as well: $t_{\alpha,\nu}$. This distribution was published by William Gosset in 1908. His employer, Guinness Breweries, required him to publish under a pseudonym, so he chose “Student”.

6.4 Distribution Fitting

There is a YouTube video and slides associated with this chapter:

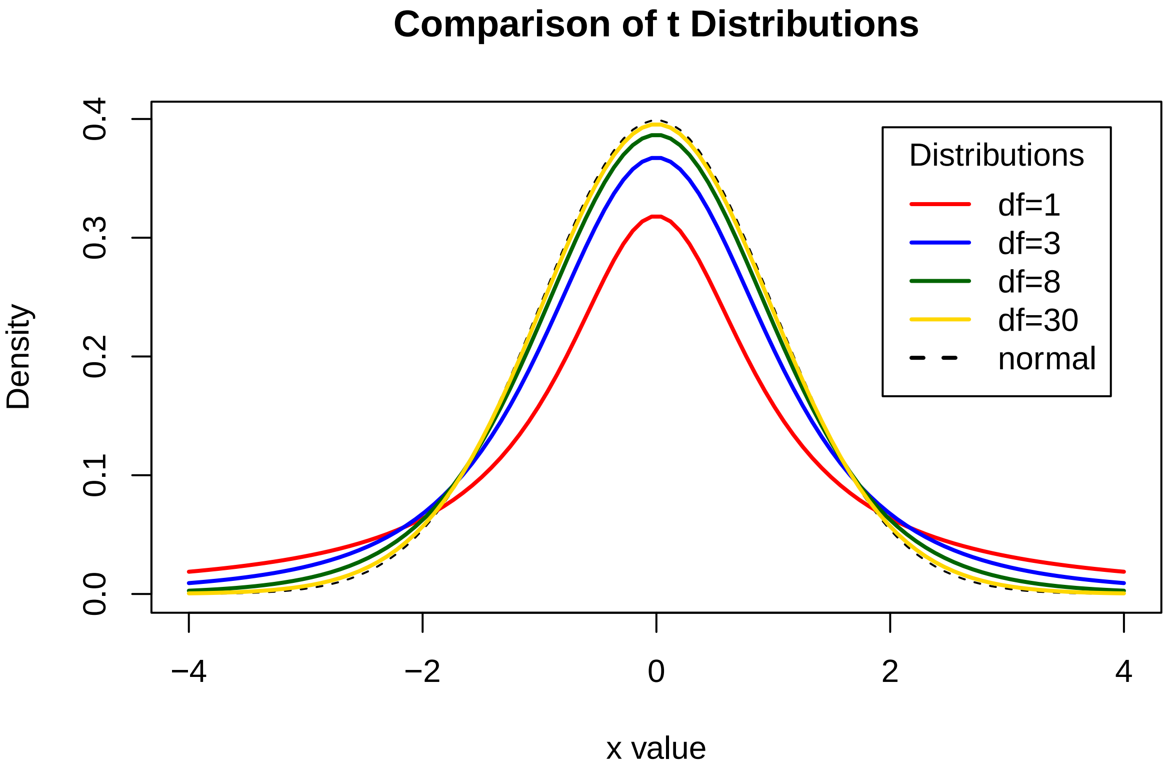


Figure 13: Student Distribution for various levels of degrees of freedom and Normal distribution

- Probability Distributions - Video
- Probability Distributions - Slides

6.5 Exercises

1. ***Fair Game*** (***): A friend of yours has a coin and proposes the following game. You toss the coin 10 times and count the number of heads. The amount you win or lose on k heads is given by $4 \cdot \sqrt{k} - 2 \cdot k$.

   (a) Plot the payoff function.
   (b) Make an exact computation using R to decide if this is a good bet.
   (c) Use the function `rbinom()` and generate 100,000 random outcomes of the game. Calculated the expected winning and compare it to your calculation in part (b).

2. ***O'Neill Air*** (**): The probability of any O'Neill Air flight being delayed more than 15 minutes is 0.1. We randomly select four different O'Neill Air flights.

   (a) Calculate the probability that all four flights arrived within 15 minutes of the scheduled time?
   (b) Calculate the probability that none of the selected flights arrived within 15 minutes of the scheduled time?
   (c) Calculate the probability that at least one of the selected flights arrived within 15 minutes of the scheduled time?

3. ***Apple Juice Machine*** (**): An apple juice company wants to purchase a new bottling machines to fill 16 ounce cans. Two manufactures indicate the following performances for their machines. The first machine fills the cans with 16.5 ounces and a standard deviation of 0.3. The second machine fills the cans with 16.2 ounces and standard deviation 0.1. The filling quantities for both machines are normally distributed. Anything below 16 ounces cannot be sold since it does not meet the advertised 16 ounces. Which machine does the apple juice producer need to buy in order the minimize the quantity of cans which cannot be sold.

4. ***ACT*** (**): Assume ACT scores are normally distributes with mean 19 and standard deviation of 4. A university accepts applicants which score in the top 20%. What is the minimum ACT score that gets you accepted? Calculate the probability of 0, 1, . . . , 10 students meetings the requirements out of 10 students.

5. ***LED Light Bulbs*** (**): On average, a LED light bulb manufacturer produces 250 defective light bulbs out of 5,000. You randomly pick 8 light bulbs. What is the probability of finding more than 1 defective bulbs?

6. ***Vaccine Trials*** (**): A pharmaceutical company estimates that a new vaccine that is undergoing test trials has a 1/10000 chance of causing a serious side effect in a human being. If the company administers the vaccine to 10,000 volunteers, what is the probability that at least 1 volunteers develops those serious side effects? The company wants to make sure that the probability of at least one person developing the side effect is 95%. How many volunteers do they need?

7. ***Quillayute*** (***): Airline magazines are often not very interesting. However, a magazine once covered an interesting story. A hotel in Washington state charges the temperature in Dollars per night. For example, if it is 72F then a room costs $72. Of course, this is too boost demand in the Winter months. The closest weather station to the hotel is at Quillayute Airport. You find the temperature data associated with that location in `quillayute` and for this exercise, you only consider the data for January (i.e., month equals 01). The hotel makes a financial loss if the temperature drops below 30F. a. Using the function `fitdist()` from the package fitdistrplus, fit three probability distributions to the data: Gamma, Weibull, and Log-normal

   b. In your answers, include three plots comparing (1) the histogram to the theoretical densities, (2) the Q-Q plots of the estimates, and (3) empirical and theoretical cumulative distribution functions. Based on those plots, which probability distribution fits best.
   c. Based on the probability distribution you identified in part (b), what is the probability that the average temperature drops below 30F?

7 Basic Statistics and Sampling

Slides: Basic Statistics and Sampling.pdf

In the previous sections, we assumed that we know the parameters associated with probability distributions. From now on, we are interested in finding those parameters by sampling from a population. It is important to differentiate between the population (whose parameters remain unknown to the researcher) and a sample (i.e., a subset) taken from the population. The sample can tell us something about the population parameters. The sampling distribution will be the probability distribution associated with the statistic, e.g., mean or variance, from the sample. Put differently, when we take a sample and calculate the mean, how would that estimate differ and which values would a different sample produce. Suppose you have a set of random variables $X_1, X_2, X_3, \ldots, X_n$ which represent the results of repeating an experiment. The random variables are independent and identically distributed (i.i.d.). The expectation of the average is written as:

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The sampling variance is expressed as:

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

The standard error of the mean is written as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This is different from the sample variance! The sampling variance represents the variation of a particular statistic, e.g., mean. Before we get into the details, let us introduce two very important concepts: (1) the Law of Large Numbers and (2) the Central Limit Theorem.

7.1 Law of Large Numbers

The unemployment rate in the United States is measured via the Current Population Survey (CPS) which samples 60,000 households on a monthly basis. People are classified into *employed*, *unemployed*, *not in the labor force*. The reason for this large sample is that the larger the sample, the smaller the margin of error (more on that concept below). This is a result of the law of large numbers. Any feature of a distribution can be recovered from repeated sampling. In particular, the law of large number can be spelled out as

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

For example, if you flip a coin, there are two possible outcomes: heads or tails. The expected value of heads (or tails) is $E(H) = E(T) = 0.5$. The variance of n coin tosses is

$$Var(n) = \frac{p \cdot (1-p)}{n}$$

Note that $p \cdot (1-p)$ is the variance associated with a Bernoulli. So the variance of n coin tosses is $Var(1) = 0.5$, $Var(10) = 0.025$, $Var(1000) = 0.00025$, etc. It is difficult to predict the share of heads from a single coin toss but high prediction precision from several thousand tosses.

There are very important implications, e.g., insurance business. If there is risk aversion for individuals as well as for firms, why do insurance companies exist? Let us illustrate how the law of large numbers can help us answer this question. Assume an insurance company that sells home insurance policies. The probability of a fire is $P(fire) = 1/250 = 0.004$. Each home is valued at \$250,000 and the value of the home after a fire is \$0. The insurance premium is equal to the expected loss, i.e., \$1000. In a simulation exercise, the damage to n homeowners (where n is the number of homes insured) and the share of homes burned down is calculated. The exercise is repeated a 1000 times and then a histogram is generated.

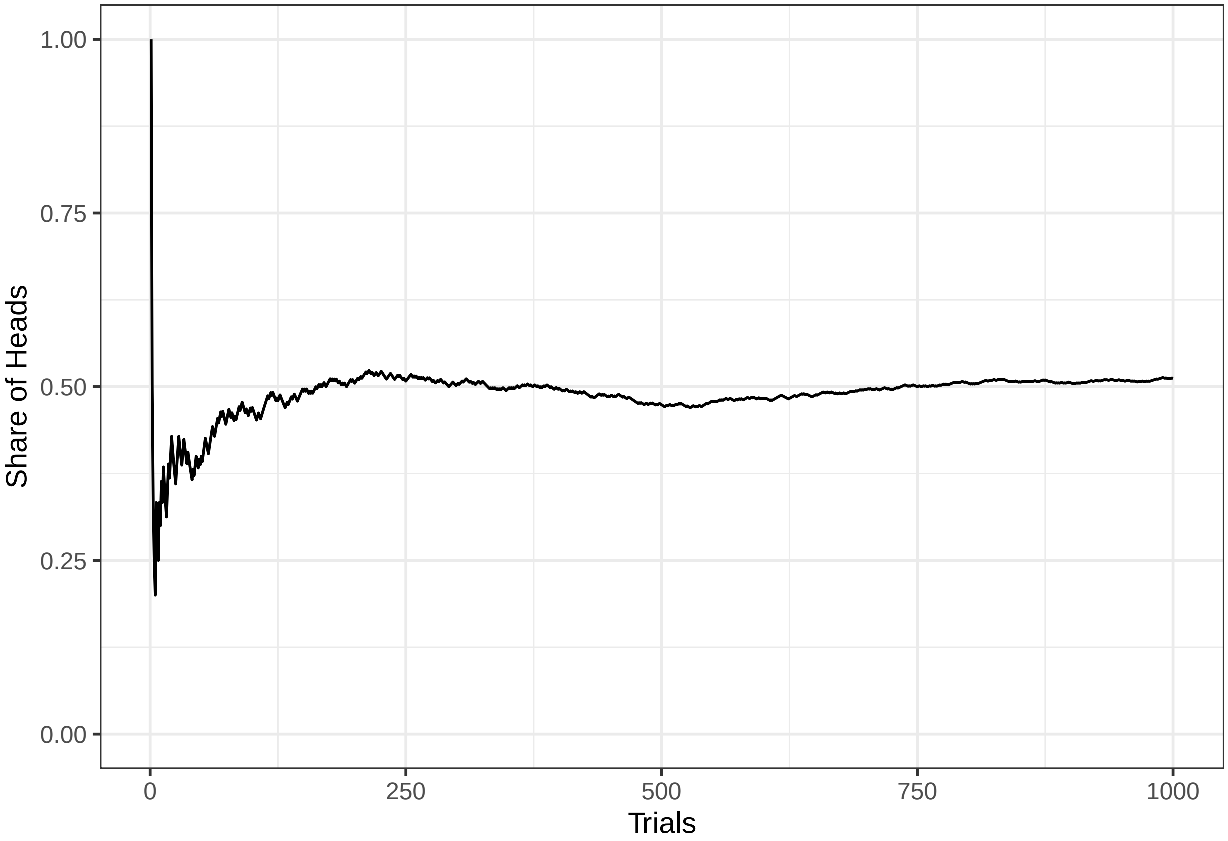


Figure 14: Law of Large Numbers illustrated by flipping a coin up to 50,000 times.

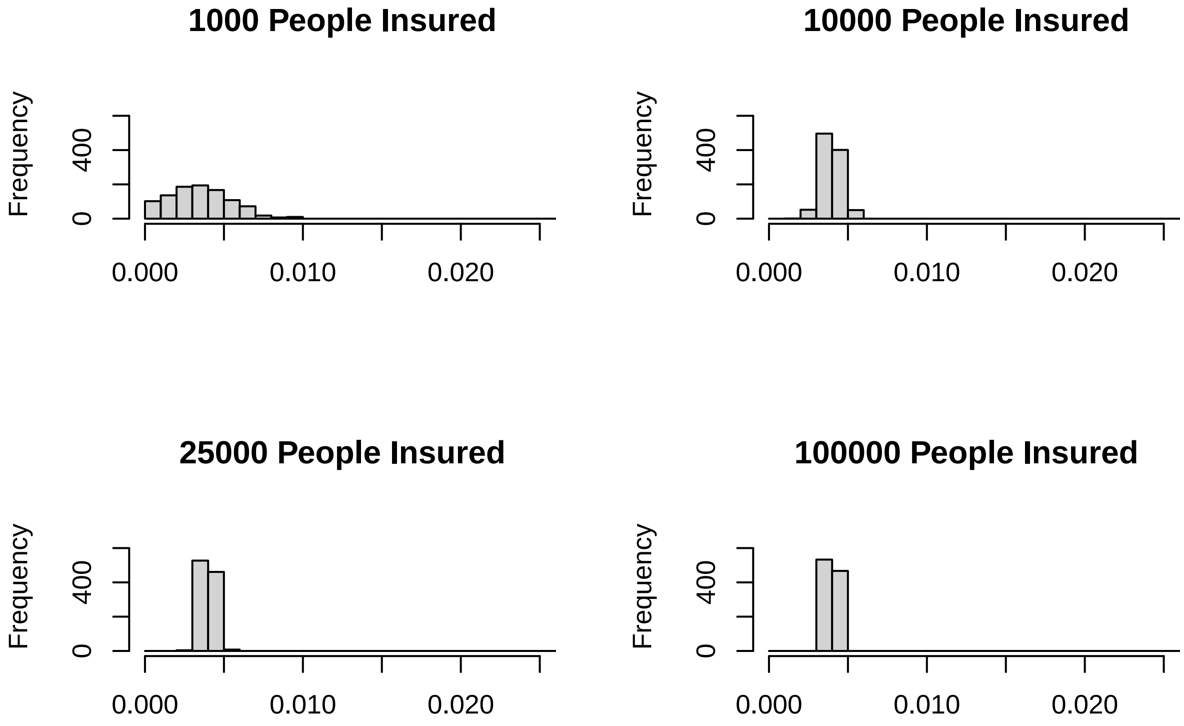


Figure 15: Risk pooling by an insurance company.

7.2 Central Limit Theorem

The Central Limit Theorem tells us that the average of an i.i.d. sample of a random variable X will converge in distribution to the Normal

$$z_n = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \Rightarrow F_{z_n}(a) = \Phi(a)$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The Central Limit Theorem is key for what we are going to see in the remaining sections. What it states is that no matter the underlying distribution (discrete or continuous), if we sample repeatedly and write down the mean of the sample i, i.e., $\bar{x}_i$, those means $\bar{x}_i$ will be normally distributed.

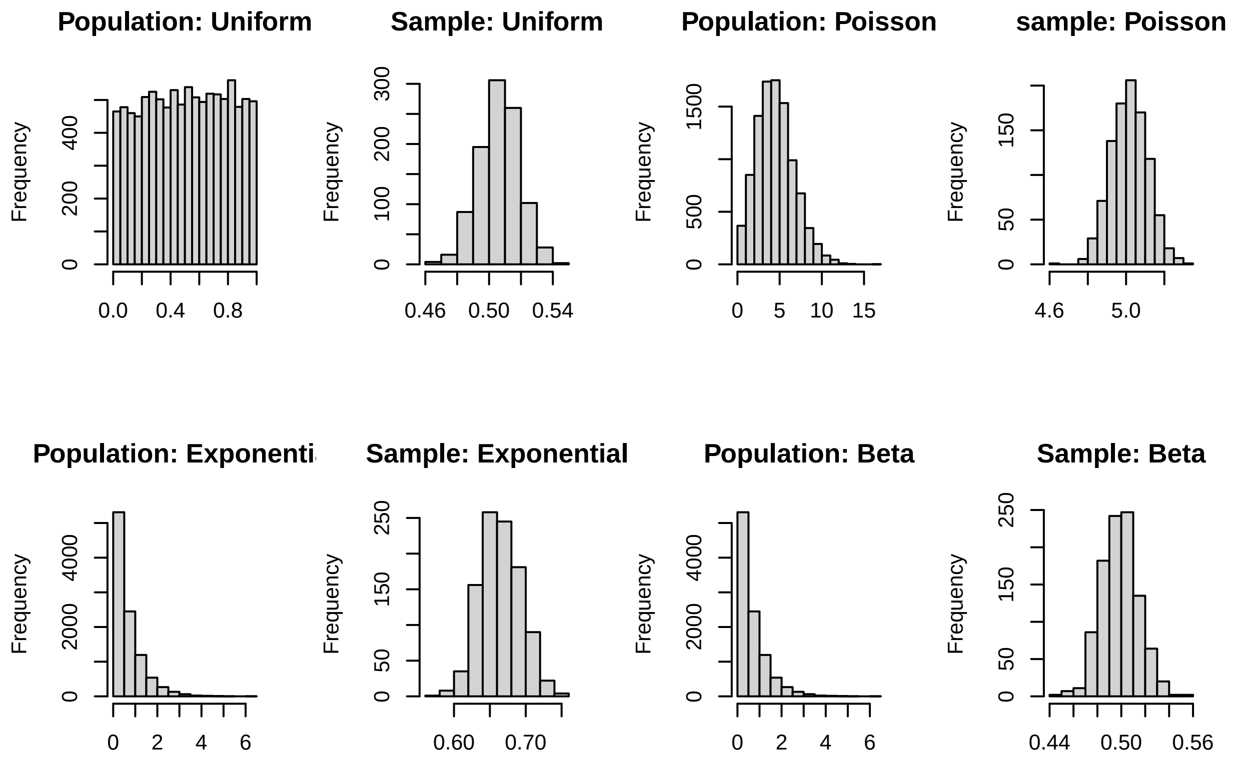


Figure 16: Illustration of the Central Limit Theorem

7.3 Estimation and Estimators

Assume the simple setting of where the random variable X represents a population with probability density function $f(x;\theta)$. This probability density function depends on the parameter θ which is unknown. It is assumed that the probability density function is known except for the parameter θ. So we have to engage in sampling from the population to get information about the parameter of interest. To gather information about θ, we perform random sampling and collect $X_1, X_2, \ldots, X_n$ which represents an identically and identically distributed (i.i.d.) random sample drawn from the probability density function $f(x;\theta)$. In the most abstract formulation, we can think of W as an estimator for the parameter θ in as $W = h(X_1, X_2, \ldots X_n)$.

The goal of a good estimation procedure is to determine an unbiased point estimator $\hat{\theta}$ of the population parameter θ. To evaluate our estimation procedure, we are interested in analyzing the sampling distribution associated with $\hat{\theta}$. From a theoretical perspective, the bias b of an estimator can be expressed as

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta$$

The mean squared error of a point estimator W is defined as

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

Note, that unbiased does not refer to the fact that $\hat{\theta} = \theta$ from one sample but if we were able to repeat the sampling infinite times, the average of the $\hat{\theta}$ would be equal to θ. Some commonly used statistics are the sample mean, the sample variance, the sample standard deviation, and the sample mid-range. The sample mean is the arithmetic average of the values in a random sample. It is denoted

$$\bar{X}(X_1, X_2, \cdot\cdot\cdot, X_n) = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The observed value of $\bar{X}$ in any sample is denoted by the lower case letter, i.e., $\bar{x}$. The sample variance is the statistic defined by

$$S^2(X_1, X_2, \cdot\cdot\cdot, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The observed value of S^2 in any sample is denoted by the lower case letter, i.e., s^2. The sample standard deviation is the statistic defined by $S = \sqrt{S^2}$. The sample mid-range is the statistic defined by

$$\frac{max(X_1, X_2, \cdot\cdot\cdot, X_n) - min(X_1, X_2, \cdot\cdot\cdot, X_n)}{2}$$

Imagine that you have two estimation methods for the estimator $\hat{\theta}$. Which one do you chose? The answer is that you would chose the most efficient estimator based on the sampling variance.

7.4 Exercises

1. ***Random Numbers*** (**): R allows you to generate random numbers from a variety of distributions. For example, you can generate 100 normally distributed random numbers with $\mu = 70$ and $\sigma = 10$ using the following command: `rnorm(100,70,10)`. Generate three datasets ($\mu = 70$ and $\sigma = 10$) with the command `rnorm()` and name them $x1$, $x2$, and $x3$. The data sets $x1$ and x have 50 random numbers each and $x3$ has 1000 random numbers. Calculate and report the mean of $x1$, $x2$, and $x3$. Plot a histogram of all three datasets.
   a. Why do the histograms of $x1$ and $x2$ look different despite the fact that they were generated using the same command? Do they look normally distributed?
   b. The histogram of $x3$ will look much more like a normal distribution. Why is that the case?
   c. Compare the three means and explain which statistical law should make the mean of $x3$ always closer to 70 than for $x1$ and $x2$.

8 Confidence Intervals

This chapter introduces the concept of confidence intervals and interval estimation. An interval estimation is often more useful than a point estimation of the unknown population parameter θ because it gives an interval of numbers within which the parameter value could fall. There are slides and a YouTube Video associated with this chapter:

- Confidence Intervals - Slides
- Confidence Intervals - Video

This chapter on confidence intervals as well as the subsequent chapter on hypothesis testing makes significant use of the t-Distribution. The aforementioned interval estimation allows us to put lower and upper boundaries on the parameter, i.e., $\hat{\theta}_l < \theta < \hat{\theta}_u$. Usually a probability value of 95% is used.

$$Pr(\hat{\theta}_l < \theta < \hat{\theta}_u) = 1 - \alpha$$

The 95% confidence interval for a parameter is an interval calculated from sample data by a method that has a 95% probability of producing an interval containing the true parameter value. Suppose a population has a known mean of 65, e.g., the height of women in the United States. Taking 100 samples of women from the population and calculating a confidence interval given the methods described below, the true mean of 65 will be included (on average) in 95 of those confidence intervals. Note that the **following statement is not correct**:

> The probability that the unknown parameter is contained within a 95% confidence interval is 95%.

Note that there is no way of knowing if the confidence interval actually covers the true parameter. Remember that we have $E(\bar{x}) = \mu$ and $Var(\bar{x}) = \sigma^2/n$. The mean ±1.96 standard deviations includes 95% of the normal distribution. Because the sampling distribution is approximately normal (recall this fact from the Central Limit Theorem and the law of large numbers), the distance of 1.96 standard deviations is the margin of error. The margin of error measures how accurate the point estimate is likely to be in estimating a parameter.

This section is designed to illustrate the concept of confidence interval with R. A population of 1 million voters is generated with 55% of those voters favoring candidate A in an upcoming election. For this exercise, a 95% confidence interval is simulated. The script below proceeds as follows:

```
voters = rbinom(1000000,1,0.55)
output = data.frame(lb=numeric(),ub=numeric(),inside=numeric())
meanA  = mean(voters)
for(i in 1:100){
     poll = sample(voters,1000,replace=FALSE)
     CI   = t.test(poll)
     temp = data.frame(ub=CI$conf.int[1],lb=CI$conf.int[2],inside=0)
     if(CI$conf.int[1]<=meanA & CI$conf.int[2]>=meanA){temp$inside=1}
     output = rbind(output,temp)}
mean(output$inside)
rm(voters,output,meanA,poll,CI,temp)
```

In a first step, a population of 1 million voters is created. The subsequent steps take 100 samples of 1,000 voters. For each sample, the confidence interval (i.e., lower and upper bounds denoted as lb and ub, respectively) is calculated. The next step checks whether the true mean is contained in the confidence interval. If the above code is executed, the output of `mean()` should be around 0.95.

8.1 Confidence Interval for a Proportion

Recall from the Bernoulli distribution that $p + q = 1$ and $\sigma = \sqrt{(p \cdot q)}$. The standard error for the mean is

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Thus, the 95% confidence interval is constructed as follows:

$$\hat{p} \pm 1.96 \cdot \sigma_{\bar{x}} \Leftrightarrow \hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Consider the voting data in `gss2018`. The mean is 0.67 and the standard error is

$$\sigma_{\bar{x}} = \sqrt{\frac{0.67 \cdot (1 - 0.67)}{772}} = 0.0169$$

In this case, the margin of error is $1.96 \cdot 0.0169 = 0.033$. To calculate a confidence interval for a proportion in R, the function `t.test()` is used. Before using the function, the manual calculations are presented.

```
voting         = subset(gss,vote12 %in% c("voted","did not vote"))
voting         = ifelse(voting$vote12=="voted",1,0)
nobs           = nrow(voting)
meandata       = mean(voting)
z              = qnorm(0.975)
stderror       = sqrt(meandata*(1-meandata)/nobs)
CI_lower       = meandata-z*stderror
CI_upper       = meandata+z*stderror
```

Using the function `t.test()` is simpler:

```
t.test(voting)
```

```
##
##  One Sample t-test
##
## data:  voting
## t = 76.794, df = 2608, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.6756645 0.7110737
## sample estimates:
## mean of x
## 0.6933691
```

8.2 Confidence Interval for the Mean

To construct a confidence interval for a population mean, the sample standard deviation must be estimated. The sample variance is denoted with s^2 and the standard error (S.E.) is denoted with

$$S.E. = \frac{s}{\sqrt{n}}$$

In most cases, σ^2 is unknown. If σ^2 was known, the normal distribution could be used. Because we rely on an estimate of the variance, we have to correct for the errors associated with this estimation and we need to use the t-distribution. The t-score is similar to the z-score in that it comes from a bell-shaped curve but the tails are thicker.

Consider the data in `eggweights`. To construct the confidence interval, the sample mean and sample standard deviation are required. Calculating those values leads to $\bar{x} = 61.05$ and $s = 4.46$. So the standard error is

$$S.E. = \frac{4.46}{\sqrt{61}} = 0.57$$

For $n = 61$ and $df = n - 1 = 60$, we find the confidence interval of $61.05 \pm 2.0003 \cdot 0.57$.

```
nobs           = nrow(mh2)
meandata       = mean(mh2$price)
stdev          = sd(mh2$price)
t_alpha_df     = qt(0.975,nobs-1)
CI_lower       = meandata-t_alpha_df*stdev/sqrt(nobs)
CI_upper       = meandata+t_alpha_df*stdev/sqrt(nobs)
t.test(mh2$price)
```

```
##
```

```
##  One Sample t-test
##
## data:  mh2$price
## t = 6.6926, df = 17, p-value = 3.783e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   564678.8 1084609.0
## sample estimates:
## mean of x
##  824643.9
```

8.3 Sample Size Calculation for a Proportion

Recall the confidence interval for a proportion:

$$\hat{p} \pm \underbrace{z \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}}_{\text{Margin of Error (ME)}}$$

If a maximum margin of error is desired, e.g., ±2%, then the above expression must be solved for n. This leads to the following sample size formula:

$$n > \frac{1.96^2 \cdot p \cdot (1-p)}{ME^2}$$

In order to use the formula, the value of p must be known. The problem is that the value is only known after the survey or poll is conducted and not before. To ensure that the margin of error is within the desired limits, a value of $p = 0.5$ is chosen. If the survey or poll is conducted and results in a proportion of people being in favor of an issue or a candidate, then the margin of error will exactly correspond chosen one. If the resulting proportion is different from 0.5, the margin of error is smaller. Thus, the value of $p = 0.5$ can be considered a worst-case scenario because it maximizes the variance. If there is knowledge about the resulting proportion before then that value can be used. For example, using a value of 0.5 to determine the sample size for the U.S. unemployment rate would be prohibitively expensive and unnecessary.

To illustrate the concept of sample size calculation, let us consider a survey that is interested in the proportion of people in support of a property tax reform. You do not have any knowledge about the population parameters but want the estimate to be within 2%. For this reason, you adopt an initial estimate of $p = 0.5$. This results in a worst case scenario.

$$n = \frac{1.96^2 \cdot 0.5 \cdot (1-0.5)}{0.02^2} = 2401$$

If the desired margin of error is reduced in half, then the sample size does not double but quadruples. This is due to squared terms in the above equations.

In some (rare) cases, the sample size necessary depends on the population size. Suppose you are interested in how many students support a privatization of parking. The sample size calculation for a finite population is written as follows:

$$n_f = \frac{n_\infty \cdot N}{n_\infty + (N-1)}$$

The term n_∞ represents the sample size for an infinite population. Consider a college of 10,000 students. The sample size calculation proceeds as follows:

$$\frac{2401 \cdot 10000}{2401 + (10000-1)} = 1937$$

8.4 Exercises

1. ***Meridian Hills I*** (*): The data set `mh1` contains home values for homes in the Meridian Hills area in Indianapolis. Construct a 90%, 95%, and 99% confidence interval around the mean using R.

2. ***GSS Guns and Death Penalty*** (*): Consider the data in `gss`. The data is taken from the 2018 General Social Survey. Note that only observation which have complete responses for all variables are included in the data. The question associated with the variable `gun` is "Do you happen to have in your home (IF HOUSE: or garage) any guns or revolvers?" and the question associated with the variable `deathpenalty` is "Do you favor or oppose the death penalty for persons convicted of murder?" Construct a 90% confidence interval around the proportion for both variables.

3. ***Soda Cans I*** (*): Consider a machine filling soda cans with a reported average of 360 milliliters (mL). The amounts filled into the cans follow a normal distribution with (unknown) mean μ and standard deviation σ. You take a sample of soda cans and measure the volume. Your data (in mL) is found in file `soda`. Construct a 99% confidence interval around the mean.

4. ***Paper Mill I*** (*): The local paper mill claims that it does not discharge more than 1000 gallons of waste water into the White River. An environmental interest group measures the discharge over one week and the data is reported to you in file `discharge`. Construct a 95% confidence interval around the mean discharge.

5. ***Sanders vs. Biden*** (*): A 2020 article about the race of Sanders vs. Biden in Florida states that *"Biden is lapping Sanders in voter support, with support from 66 percent of likely Democratic primary voters to 22 percent for Sanders, according to a University of North Florida poll taken March 5-10."* The article mentions a margin of error ±2.5%. For this exercise, use the number of 66% and calculate the sample size that was used to conduct the poll.

6. ***Privatizing Parking*** (**): You are interested in how many students at IUPUI are supporting privatizing parking services. You have information from a different university who found that 20% of students support privatization. In fall 2018, IUPUI had an enrollment of 29,579 students. Calculate the necessary sample size based on the (1) information you have from the other university and (2) the worst-case scenario in terms of variance You want the margin of error to be with ± 3%.

7. ***VMT by State*** (**): Consider the data in `vehpub`. Pick any state other than Indiana and construct a 95% confidence interval around the average annual miles traveled based on the odometer reading and the age of the vehicles. Keep in mind that the survey was conducted in 2017. Pay attention to eliminate missing or otherwise irrelevant values.

9 Hypothesis Testing

There are slides and a YouTube Video associated with this chapter:

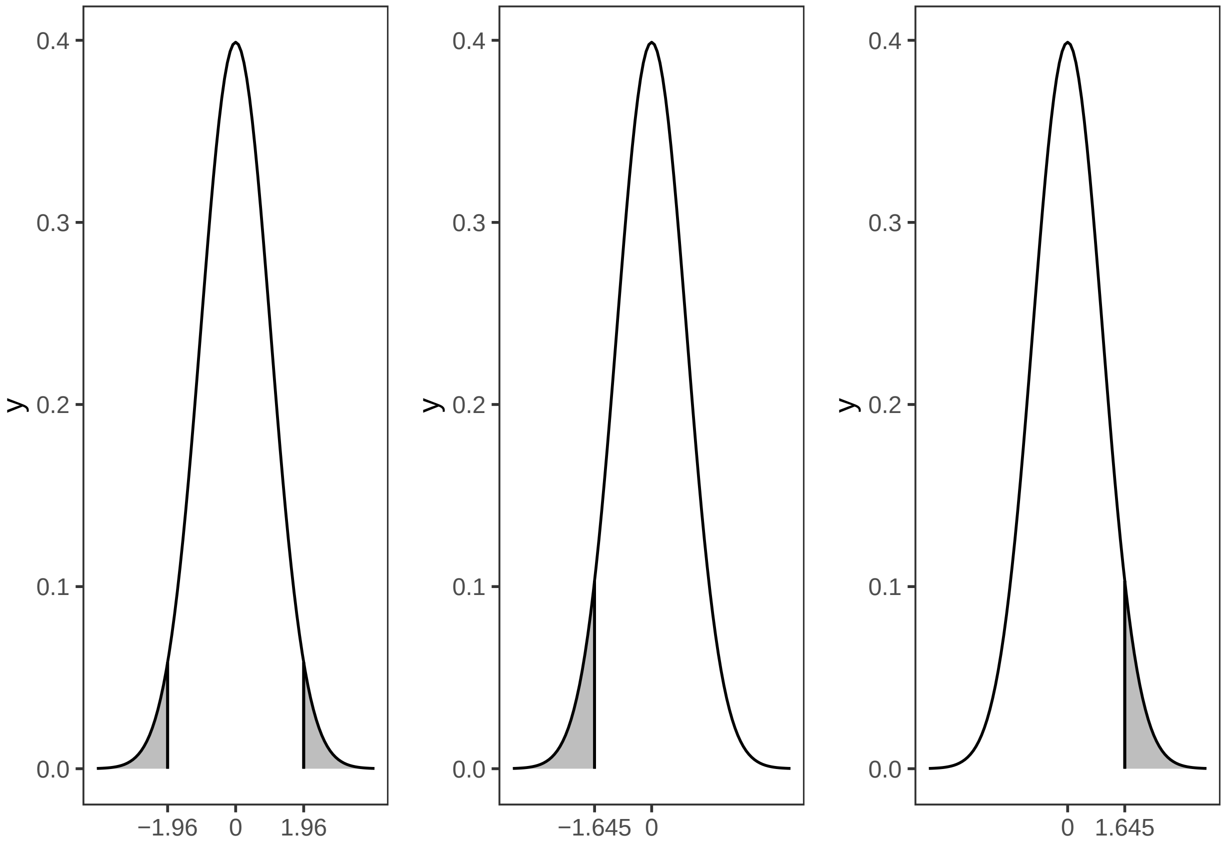
- Hypothesis Testing - Slides
- Hypothesis Testing - Video

A hypothesis is a statement about a population claiming that a parameters takes on a particular value or falls within a certain range of values. The steps of a hypothesis test are:

1. Formulating the null hypothesis H_0 stating that the parameter takes a particular value:
    - One-sided test: H_0: $\mu \geq \mu_0$ or $\mu \leq \mu_0$
    - Two-sided test: H_0: $\mu = \mu_0$
2. Setting the significance level α, e.g., 1%, 5%, or 10%.
3. Test statistic: Value based on the sample used to **reject** or **fail to reject** the null hypothesis.
4. Critical value and p-value:
    - Critical value represents the border point between rejecting and failing to reject H_0.
    - p-Value: Probability of observing the parameter given the null hypothesis. Small p-values represent evidence against H_0.

Note that equality is always part of H_0, i.e., $=$, $\leq$, or $\geq$.

There are two types of errors associated with hypothesis testing. A Type I error occurs if H_0 is true but you reject H_0. A Type II error occurs if H_a is true but you fail to reject H_0. The significance level is the largest acceptable probability of committing a Type I error. This is denoted with α. There are two-sided and one-sided hypothesis tests:



9.1 One-Group: Proportions

To execute a hypothesis test for a population proportion, we have to assume that the data is categorical with the population proportion p defined in the context. Assuming that the sample size is above 30, the test statistic is written as

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)/n}}$$

Recall that the sampling distribution of a sample proportion has mean p and standard deviation $\sqrt{p \cdot (1-p)/n}$. This z-score measures the number of standard errors between the sample proportion $\hat{p}$ and the null hypothesis p_0. The significance level shows us how strong the evidence must be. For example, assume we have a sample size of $n = 100$ and that $\hat{p} = 0.48$. He hypothesize that $p_0 = 0.5$, i.e., H_0: $p_0 = 0.5$. The standard error is

$$S.E. = \sqrt{\frac{0.5 \cdot 0.5}{100}} = 0.05$$

Thus, the z-score is

$$z = \frac{0.48 - 0.5}{0.05} = \frac{0.02}{0.05} = -0.4$$

For a two-sided hypothesis test at the $\alpha = 0.05$ level, we fail to reject the hypothesis because $-0.4 > -1.96$.

Consider the data in `gsssocialmedia`. Suppose that Instagram claims that 1/3 of Americans use their service. This can be verified with a two-sided hypothesis test:

```r
socialmedia              = gss[c("instagrm")]
socialmedia              = na.omit(socialmedia)
socialmedia$instagrm     = ifelse(socialmedia$instagrm=="yes",1,0)
t.test(socialmedia$instagrm,mu=1/3,alternative=c("two.sided"))
```

```
##
##  One Sample t-test
##
## data:  socialmedia$instagrm
## t = -2.0065, df = 1371, p-value = 0.045
## alternative hypothesis: true mean is not equal to 0.3333333
## 95 percent confidence interval:
##  0.2838431 0.3327750
## sample estimates:
## mean of x
##  0.308309
```

The p-value is above 5% and thus, we fail to reject the hypothesis. Now suppose that Instagram claims that more than 1/3 of Americans use their service. The term “more than” suggests a one-sided hypothesis. The hypothesis is formulated as follows:

$$H_0 : p \geq 1/3 H_a : p < 1/3$$

It is very important to correctly **state the alternative hypothesis in R**.

```r
t.test(socialmedia$instagrm,mu=1/3,alternative=c("less"))
```

```
##
##  One Sample t-test
##
## data:  socialmedia$instagrm
## t = -2.0065, df = 1371, p-value = 0.0225
## alternative hypothesis: true mean is less than 0.3333333
## 95 percent confidence interval:
##       -Inf 0.3288373
## sample estimates:
## mean of x
##  0.308309
```

In this case, the p-value is below 5% and thus, the hypothesis is rejected.

9.2 One-Group: Mean

The requirements to use the t-test are that (1) the variable is quantitative, (2) the data production employed randomization, and the population distribution is approximately normal. In this section, hypothesis tests with unknown population variance are considered. The test statistic is

$$t = \frac{\hat{x} - \mu_0}{s/\sqrt{n}}$$

For example, consider the scores from a graduate MPA class which has eighteen students in `mpa`. The mean of the data is 69 and the sample standard deviation is 21.15. The hypothesis test is formulated as follows:

$$H_0 : \mu = 80 H_a : \mu \neq 80$$

The t-statistic can be calculated as follows:

$$t = \frac{69 - 80}{21.15/\sqrt{18}} = -2.207$$

The critical value for this two-sided test is -2.11 and thus, the hypothesis is rejected. The hypothesis can aslo be implemented manually in R:

```
n                    = nrow(mpa)
xbar                 = mean(mpa$scores)
stdev                = sd(mpa$scores)
tstatistic           = (xbar-80)/(stdev/sqrt(n))
criticalvalue        = qt(0.025,df=n-1)
pvalue               = pt(tstatistic,n-1)
```

Or the function `t.test()` can be used.

```
t.test(mpa$scores,mu=80)
```

Similar to the hypothesis tests on proportions, a one-sided test can be implemented with the function `t.test()` as well. Consider the data in `eggweights`. Consider the following one-sided hypothesis test:

$$H_0 : \mu \geq 63 H_a : \mu < 63$$

This hypothesis test is implemented as follows:

```
t.test(eggweights$weight,mu=63,alternative="less")
```

```
##
##  One Sample t-test
##
## data:  eggweights$weight
## t = -3.4172, df = 60, p-value = 0.0005709
## alternative hypothesis: true mean is less than 63
## 95 percent confidence interval:
##      -Inf 62.00294
## sample estimates:
## mean of x
##  61.04918
```

In this case, the null hypothesis is rejected because the p-value is below 5%.

9.3 Two-Groups: Proportion

The hypothesis test for comparing the proportion between two groups is written as follows:

$$H_0 : \hat{p}_1 - \hat{p}_2 = 0 H_a : \hat{p}_1 - \hat{p}_2 \neq 0$$

We first need to calculate the pooled estimate, i.e., the total number of success over the total number of observations. You can think of this as a weighted average. Let this weighted average be $\hat{p}^*$.:

$$\hat{p}^* = \frac{p_1 + p_2}{n_1 + n_2}$$

Then, the standard error can be calculated as follows:

$$S.E. = \sqrt{\hat{p}^* \cdot (1 - \hat{p}^*) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

The resulting test statistic is written as

$$z = \frac{\hat{p}_1 - \hat{p}_2}{S.E.}$$

Consider the hypothesis about the proportion of gun owners differentiated by male vs. female. The data is in `gssgun`. A summary of the data is in the following table:

owngun	male	female
yes	232	208
no	335	501

The pooled proportion is calculated as follows:

$$\hat{p}^* = \frac{232 + 208}{440 + 836} = 0.345$$

Given those numbers, the standard error is calculated as:

$$S.E. = \sqrt{0.345 \cdot (1 - 0.345) \cdot \left(\frac{1}{440} + \frac{1}{836}\right)} = 0.028$$

The test statistic is:

$$z = \frac{0.41 - 0.29}{0.028} = 4.286$$

Thus, we reject the hypothesis that the proportion of gun owners among males and females is the same.

9.4 Two-Groups: Mean

To conduct a hypothesis test about the difference between two means in R, the command `t.test()` must be used. Consider the schools in the data set `ohioschool` and `ohioscore`. Suppose you are interested in whether large schools perform as well as small schools. The school enrollment cut-off values chosen are 1000 and 3000 for small and large schools:

```
ohio        = merge(ohioincome,ohioscore,by=c("irn"))
ohio_small  = subset(ohio,enrollment<1000)
ohio_large  = subset(ohio,enrollment>=3000)
t.test(ohio_small$score,ohio_large$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  ohio_small$score and ohio_large$score
## t = 1.5676, df = 240.24, p-value = 0.1183
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5216299  4.5874560
## sample estimates:
## mean of x mean of y
##  89.21070  87.17778
```

Since the p-value is above 5%, we fail to reject the null hypothesis.

9.5 Exercises

1. ***Milk Containers*** (*): A bottling machine fills one-gallon containers with 128 fluid ounces of milk. You suspect that there is some variation in the amount filled and you take measurements from 50

containers. The measurements are in the data set `milk`. Test the null hypothesis that the machine fills the containers with more than 128 fluid ounces.

2. ***Soda Cans II*** (*): Consider a machine filling soda cans with a reported average of 360 milliliters (mL). The amounts filled into the cans follow a normal distribution with (unknown) mean μ and standard deviation σ. You take a sample of soda cans and measure the volume. Your data (in mL) is found in data set `soda`. Test the hypothesis (at the 5% significance level) that the machine fills cans with more than 360 mL.

3. ***Paper Mill II*** (*): The local paper mill claims that it does not discharge more than 1000 gallons of waste water into the White River. An environmental interest group measures the discharge over one week and the data is reported to you in the data set `discharge`. Formulate and test the hypothesis with regard to the claims of the paper mill.

4. ***Meridian Hills II*** (*): The data set *mh1* contains home values of 101 homes in the Meridian Hills area in Indianapolis. Test the hypothesis that the home values are greater than $500,000.

5. ***HDI*** (***): The United Nations Development Programme (UNDP) creates an annual Human Development Report (HRD) including a Human Development Index (HDI). It attempts to measures quality of life in various countries. According to UNDP: *"Human development – or the human development approach – is about expanding the richness of human life, rather than simply the richness of the economy in which human beings live. It is an approach that is focused on people and their opportunities and choices."* Go to the UNDP data webpage and download the 2019 HDR tables. You can either click on "Download 2019 Human Development Data All Tables and Dashboards" on the data webpage or you can here. For this question, you only need the data contained in sheet "Table 1":

   - The second to last column is named *GNI per capita rank minus HDI rank*. Interpret the meaning of the column. What does a negative/positive value mean?
   - Construct a scatter plot with *Gross national income (GNI) per capita* on the horizontal axis and *Human development index (HDI)* on the vertical axis. What do you observe and what can be concluded?
   - Subset the original data into two groups. The first group contains the top 10 countries in terms of income. The second group contains the countries ranked 11-20 in terms of income. You can do this separation in Excel. Is there a statistically significant difference in HDI between those two groups?
   - Subset the original data into two groups. The first group contains the top 20 countries in terms of income. The second group contains the countries ranked 21-40 in terms of income. Is there a statistically significant difference in HDI between those two groups?
   - Compare your answers from parts (3) and (4). What do you conclude?

6. ***Airlines*** (***): You will analyze airline delay data from the Bureau of Transportation Statistics. The data is contained in the data set `airlines`. Pick a random airport (except Indianapolis). Answer the following questions:

   - We are going to focus on the three major carriers: American Airlines, Delta Air Lines, and United Air Lines. Note that United Air Lines is the result of a merger from United and Continental in 2011. The records for Continental Air Lines in the data set stops in that year. To make the data for United comparable over the entire time frame, add the *arr_flights* and $arr_del15* numbers for United and Continental between 2003 and 2011. That is, we are looking at the merged company over the entire time horizon.
   - Create a column called *delay* which represents the share of flights delayed by airline, month, and year. Use the columns $arr\_flights$ and arr_del15 for this calculations. Graph the share of delayed arrivals (i.e., delay) for the three carriers over time. Is there a pattern? For example, is it upward trending or downward trending. Is one airline consistently worse than others? Is an airline improving over time compared to others?
   - Using the data from January 2014 to today, do a boxplot using the *delay* column grouped by the three airlines.
   - Do three two-sample hypothesis tests using the *delay* data from January 2014 to today: (1) United

vs. Delta, (2) Delta vs. American, and (3) American vs. United. The null hypothesis for all three tests is that there is no difference in delays. Report and interpret your results.

7. ***Compact Cars*** (**): Consider the data in `compactcars`. For a long time, cars with a manual transmission were more fuel efficient than cars with an automatic transmission. This has changed in recent years due to improvements for automatic transmissions. In this exercise, you will conduct two paired hypothesis tests: one for compact cars of the 1995 model year and one for the 2015 model year. The data set contains only vehicles and models of the EPA category *Compact Cars* for which the identical model was available with either automatic or manual transmission. Conduct a paired hypothesis test for 1995 and 2015 with the null hypothesis that there is no difference in fuel efficiency. Based on your calculations, what do you conclude? Note that you are not conducting a hypothesis test to compare the 1995 and 2015 fuel efficiency. It is fairly intuitive and clear that the fuel efficiency has improved over that time period.

8. ***Automatic vs. Manual Transmission*** (**): This question is based on the same motivation than the question "Compact Cars". Consider the data in `fetransmission`. Pick a vehicle class of your choice as well as one year in the 1980s and one year in the 2010s. Conduct a paired hypothesis test (individually for each year) with the null hypothesis that there is no difference in fuel economy. Based on your calculations, what do you conclude?

9. ***Green Laws*** (**): Go to the data repository of the General Social Survey (GSS). Read through page to familiarize you with the GSS. This data goes beyond the homework but could be useful to you in the future either for work or if you are interested in a particular question about public opinions. If you are interested in a particular topic, go to *Browse Variables*. For this question, search for the variable `GRNLAWS`.
   - What is the question associated with this variable and which years are covered?
   - Construct the 95% confidence interval for the years covered by this question. Interpret in context. Can you conclude whether or not a majority or minority of the population would answer yes?
   - How has this variable evolved over the years? Make sure to report the share of of respondents in favor. Include a graph with time on the horizontal axis.

10. ***Ohio Schools I*** (***): The data set `ohioincome` and `ohioscore` contain information about the school districts in Ohio with regard to enrollment, overall school performance (think of that as a measure of how good a school is), and median income. First, merge the two data sets based on IRN (serving as an identifier in the two data sets) using the R command `merge()`. Test the hypothesis that there is no difference in performance for the top 25% and bottom 25% of schools in terms of median income. That is, you are testing the hypothesis that low median income and high median income school districts are performing equally well.

10 Additional Topics in Statistics

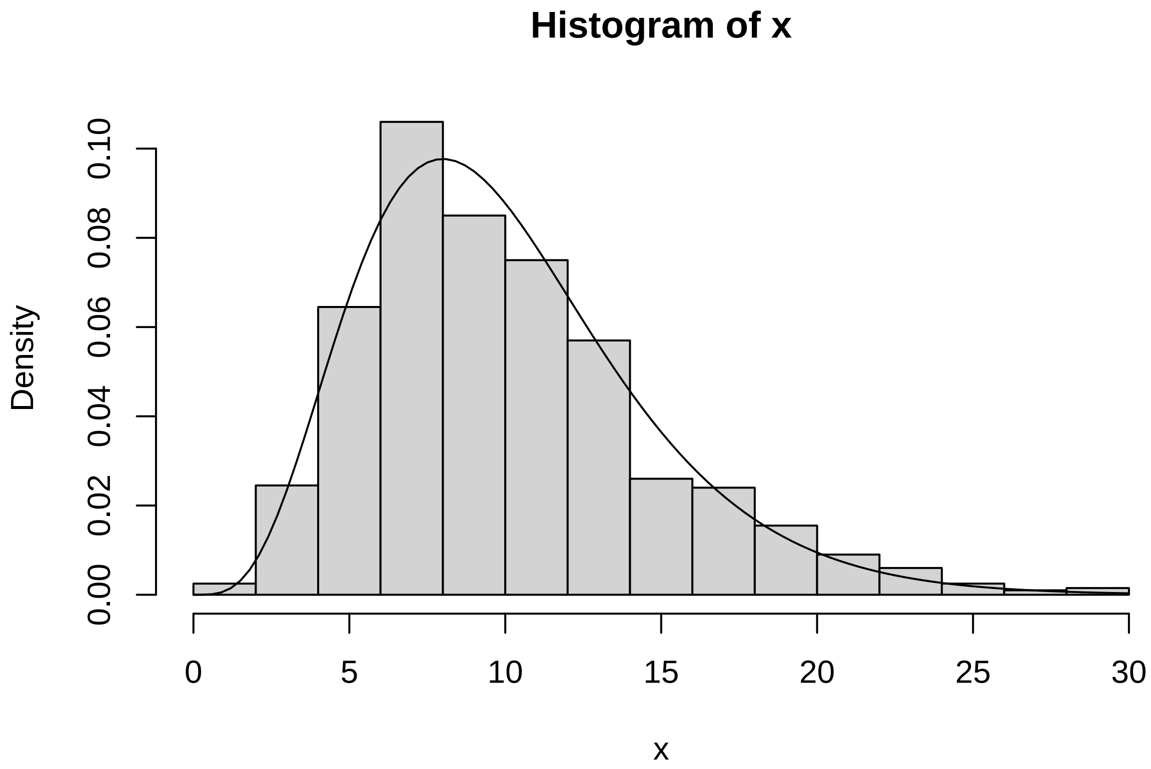
10.1 Chi-Square Test (χ^2-Test)

The χ^2-test is used to conduct a hypothesis test on qualitative variables. Before introducing the procedure, a presentation of the χ^2-distribution is necessary.

Given the values $Z_1, Z_2, \ldots, Z_k$ independently drawn from standard normal distribution, then the squared sum of those values follows a χ^2-distribution:

$$\sum_{i=1}^{k} Z_i^2 = Z \sim \chi_k^2$$

where k specifies the degrees of freedom.



To conduct the hypothesis test, consider the data on voting and education in `gss`. Using the function `CrossTable` associated with the package gmodels, the following table can be constructed.

```
CrossTable(gss$degree,gss$vote12,prop.r = FALSE,prop.c = FALSE,prop.chisq = FALSE,prop.t=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  2806
##
##
##                | gss$vote12
##     gss$degree | did not vote |   ineligible |        voted |    Row Total |
## ---------------|--------------|--------------|--------------|--------------|
##       bachelor |           84 |           17 |          425 |          526 |
## ---------------|--------------|--------------|--------------|--------------|
##       graduate |           32 |           17 |          266 |          315 |
## ---------------|--------------|--------------|--------------|--------------|
##    high school |          447 |          122 |          870 |         1439 |
## ---------------|--------------|--------------|--------------|--------------|
## junior college |           67 |            7 |          137 |          211 |
## ---------------|--------------|--------------|--------------|--------------|
## lt high school |           168 |            37 |           110 |           315 |
## ---------------|--------------|--------------|--------------|--------------|
##   Column Total |           798 |           200 |          1808 |          2806 |
## ---------------|--------------|--------------|--------------|--------------|
##
##
```

This table is similar to what has been presented previously except that all proportions have been removed and only the counts are presented. This is a 5-by-2 contingency table (the total columns are not considered part of the table). The variable education is less than high school (0), high school (1), junior college (2), bachelor (3), and graduate (4). Assuming independence, the expected value E for each cell is

$$E = \frac{(\text{total of row}) \cdot (\text{total of column})}{\text{total count}}$$

For example, consider the cell "high school" and "voting" which contains 214 counts. The expected value under the null hypothesis that voting behavior is independent of education leads to the following:

$$E_{1,1} = \frac{367 \cdot 518}{772} = 246.2513$$

Those calculations can be conducted for each cell. Then, then the χ^2-test statistic is calcualted as follows:

$$\chi^2 = \sum_{i=1}^{r \cdot c} \frac{(O_i - E_i)^2}{E_i}$$

where $r \cdot c$ is number of rows multiplied by the number of colunms, O_i and E_i are the observed and exepected count in a cell, respectively. The degrees of freedom are calcualted as $(r-1) \cdot (c-1)$. Of course, instead of doing it manually, the following command can be used:

```
chisq.test(gss$degree,gss$vote12)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gss$degree and gss$vote12
## X-squared = 256.2, df = 8, p-value < 2.2e-16
```

Due to the p-value being very small, we reject the hypothesis of independence. The χ^2 hypothesis test should only be used for qualitative data. For example, do not categorize income into quartiles to conduct a hypothesis test on whether voting depends on income.

The example of a 2-by-2 table is a special case

```
votegun   = subset(gss,
                   vote12 %in% c("voted","did not vote") &
                        owngun %in% c("yes","no"),
                   select = c("vote12","owngun"))
votegun$vote12 = ifelse(votegun$vote12=="voted",1,0)
votegun$owngun = ifelse(votegun$owngun=="yes",1,0)
t.test(votegun$vote12~votegun$owngun)
```

```
##
##  Welch Two Sample t-test
##
## data:  votegun$vote12 by votegun$owngun
## t = -2.583, df = 1187.1, p-value = 0.009913
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.10556727 -0.01442532
## sample estimates:
## mean in group 0 mean in group 1
##       0.6758193       0.7358156
```

```
chisq.test(votegun$vote12,votegun$owngun)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  votegun$vote12 and votegun$owngun
## X-squared = 6.1159, df = 1, p-value = 0.0134
```

```
CrossTable(votegun$vote12,votegun$owngun,prop.r = FALSE,prop.c = FALSE,prop.chisq = FALSE,prop.t=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  1693
##
##
##                | votegun$owngun
## votegun$vote12 |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##              0 |       366 |       149 |       515 |
## ---------------|-----------|-----------|-----------|
##              1 |       763 |       415 |      1178 |
## ---------------|-----------|-----------|-----------|
##   Column Total |      1129 |       564 |      1693 |
## ---------------|-----------|-----------|-----------|
##
##
```

10.2 Binomial Test

Hypothesis test about the probability of success.

Many government forms indicate the estimated time to fill out a form. This information can be used to calculate the (financial) burden associated with providing information to government agencies. For example, the instructions of the 2020 IRS Form 1040 state that *the estimated average time burden for all taxpayers filing a Form 1040 or 1040-SR is 12 hours.* Consider the data in `irs` which contains 40 observations of taxpayers filling out the form. Conducting a regular t-test leads to the rejection of the hypothesis:

The hypothesis is also rejected

In the sample, 16 observations are below the estimated 12 hours and 24 observations are above 12 hours.

```
binom.test(16,40)
```

```
##
##  Exact binomial test
##
## data:  16 and 40
## number of successes = 16, number of trials = 40, p-value = 0.2682
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2486500 0.5667329
## sample estimates:
## probability of success
##                    0.4
```

checkout = c(3.8,5.3,3.5,4.5,7.2,5.1) t.test(checkout,mu=4,alternative=c("greater"))

10.3 Wilcoxon Signed-Rank Test

10.4 Exercises

1. ***Chi-Squared Test*** (**): Use the `gss` data for this exercise. Select one of the following variables: $fulltime$, *government*, *married*, *gun*, or *deathpenalty*. Conduct a Chi-Squared hypothesis test between the variable chosen and education. Remember that the null hypothesis is independence between the variables. Report your results and interpret them as well.

11 Bivariate Regression

At the end of this chapter, the reader should be able to understand the following concepts:

- Identify the dependent and the independent variables in a linear regression model.
- Calculate a linear regression model finding the intercept and slope coefficients.
- Evaluate the statistical significance of a regression coefficient based on hypothesis testing.

There are slides and a YouTube Video associated with this chapter:

- Bivariate Regression - Slides
- Bivariate Regression - Video

The goal of a regression model is to establish causality between the dependent variable (y) and the independent variable(s) (x_k). It is assumed that the direction of influence is clear, i.e., x influences y and not vice versa. Each observation y_i is a function of x_i plus some random term ϵ_i. The bivariate regression model is adequate to explain the mechanics of regression models. Every regression equation can be decomposed into four parts:

1. y as the dependent variable
2. x_k as the independent variable(s)
3. β_0 as the intercept,
4. β_k as the slope coefficient(s) associated with the independent variable(s) x_k.

The bivariate model can be written as follows:

$$y = \beta_0 + \beta_1 \cdot x + \epsilon$$

Any regression model aims to minimize the sum of the squared residuals which is why it is also called ordinary least square (OLS) model. Consider the above model for a particular observation i:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \Rightarrow e_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$$

where ϵ is the disturbance term with $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$ for all i. This is equivalent to stating that the population from which y_i is drawn has a mean of $\beta_1 + \beta_2 x_i$ and a variance of σ^2. Now if these estimated errors e_i are squared and summed we obtain

$$\sum_{i=1}^{N} e_i^2 = \sum_{t=1}^{N} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \right)^2$$

The estimated errors e_i is the vertical distance between y_i and the predicted $\hat{y}_i$ on the sample regression line. Different values for the parameters β_0 and β_1 give different values for the sum of squared errors. Equation **??** must be minimized with respect to β_0 and β_1. Using calculus, it can be shown that β_0 and β_1 that minimize equation **??** can be determined as follows:

- Mean of x:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Mean of y

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- Slope coefficients

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Intercept

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

The last equation implies that the regression line goes through the point ($\bar{y}$, $\bar{x}$). Independent of the number of observations, there will always be only one β_0 and one β_1. The linear function with the intercept β_0 and the slope coefficient β_1 does not exactly provide y given value of x. Rather it provides the expected value of y, i.e., $E(y|x)$. Panel (a) of the figure provides an example with the price of a home determined by the square footage.

The direction of the relationship is clear in the sense that the square footage (independent variable) influences the home value (dependent variable) and not vice versa. The red dashed lines are the vertical error terms. In the graph, the solid regression line is such that the sum of the squared error terms is minimized.

For the OLS Model to provide unbiased estimates, assumptions associated with the model are necessary. Later sections cover how to test if those assumptions are satisfied and how to correct the model if needed. The assumptions are:

- Linear in parameters, i.e., $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. This does not assume a linear relationship between x and y. Later chapters cover functional forms and how to model non-linear relationships with a linear model.
- The error terms have a zero mean, i.e., $E(\epsilon_i) = 0$.
- Homoscedasticity, i.e., $Var(\epsilon_i) = \sigma^2$ (Figure **??**).
- No autoregression, i.e., $Cov(\epsilon_i, \epsilon_j) = 0$.
- Exogeneity of the independent variable, i.e., $E(\epsilon_i|x_i) = 0$.

11.1 Measuring the Strength of the Relationship

To measure the strength of the hypothesized statistical relationship between the dependent and independent variables of the regression equation, we calculate a value called R^2. The value of R^2 can also be thought of as an indicator of goodness of fit, or how well the sample regression line fits the sample data. To see how this statistic is used, we decompose the variation of y in the sample into two components, i.e., the *unexplained variation* and the *explained variation.* Let the total sum of squares (TSS) be

$$TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$$

Let the explained sum of squares (ESS) be

$$ESS = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

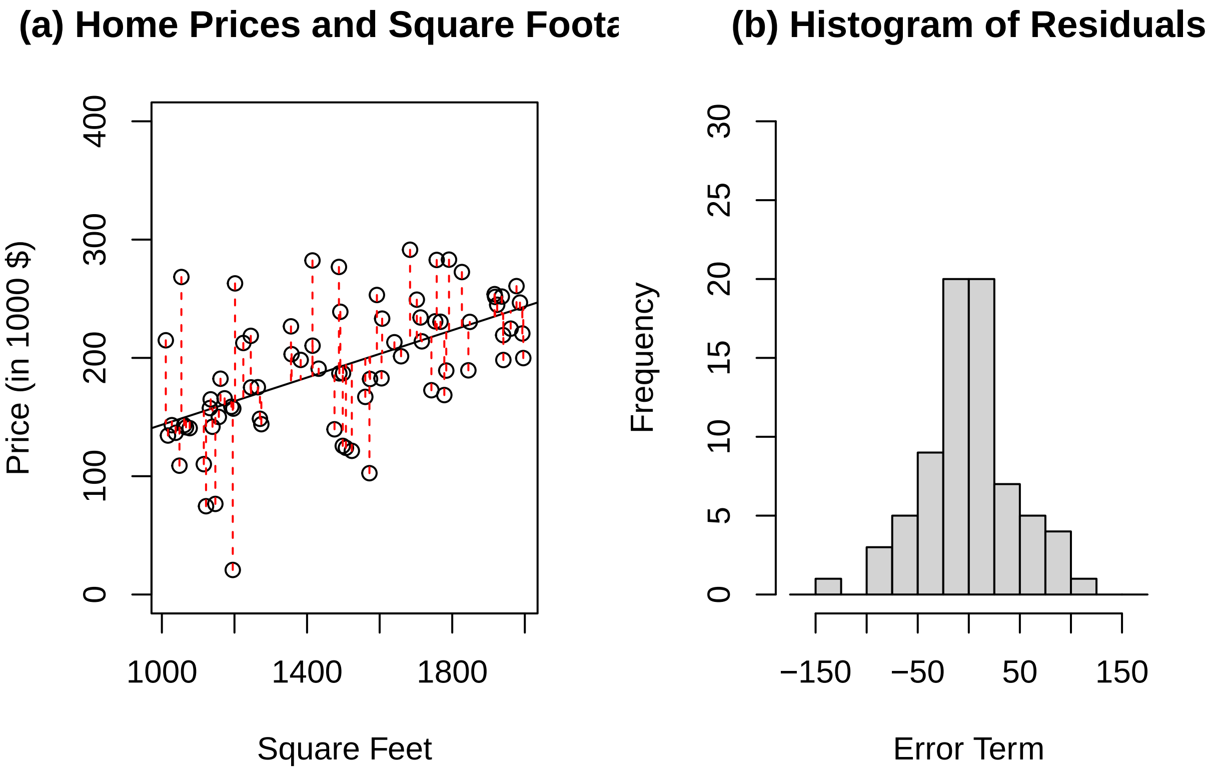


Figure 17: Example of regression line to model home values as a function of square footage. The red dashed lines in panel (a) represent the error terms associated with each observation. Panel (b) is the histogram associated with the error terms. The expected value of the error terms is zero and by assumption, the error terms are normally distributed.

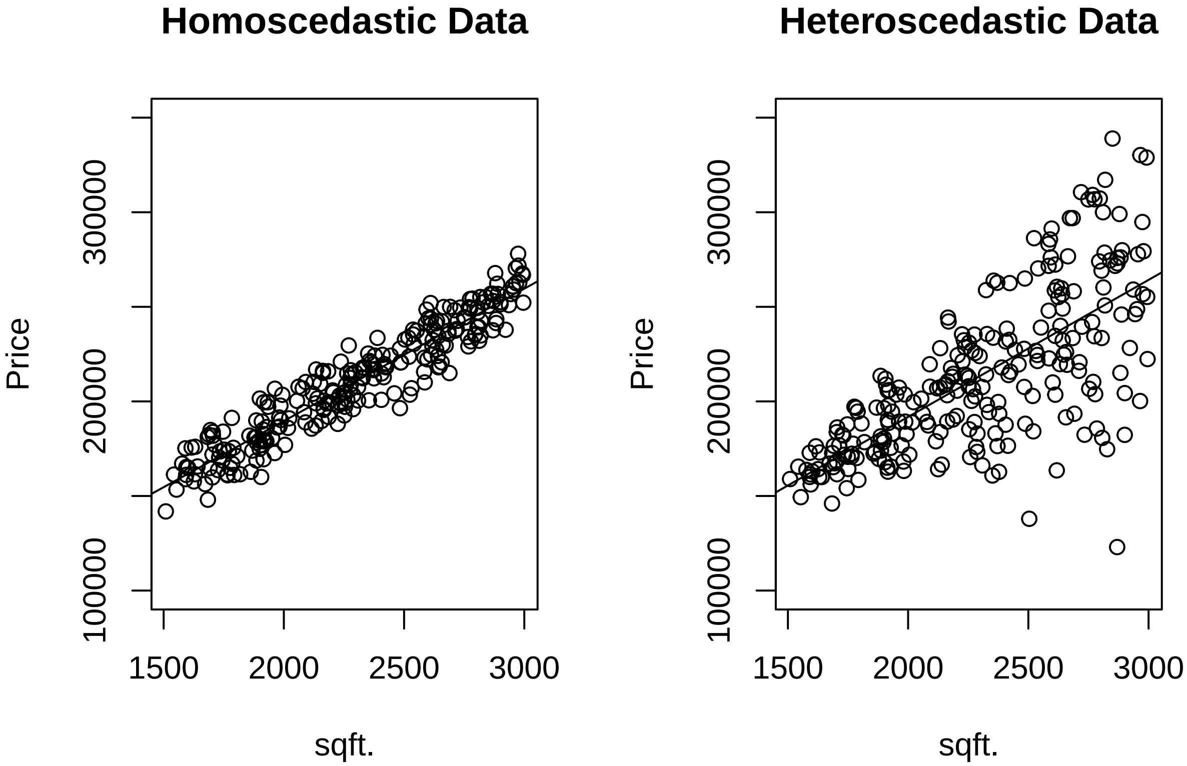


Figure 18: Panel (a) illustrates homoscedastic data whereas Panel (b) illustrates heteroscedastic data. The coefficient estimates will be unbiased by the standard error are larger for the model suffering from heteroscedasticity.

And let the unexplained (residual) sum of square (RSS) be

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Thus, the total sum of squares is equal to the explained sum of squares plus the unexplained sum of squares, i.e., TSS=RSS+ESS. The RSS represents the “unexplained’ ’ variation, since it indicates the amount of error (or the residual) in the prediction of Y; i.e., the difference between the actual value of y and its predicted value. The SSE represents the variation of the predicted values of y around $\bar{y}$, and indicates the gain in predictive power achieved by using $\hat{y}$ as a predictor of y instead of $\bar{y}$. Hence, the ESS is the amount of total variation in y which is accounted for (or explained) by the regression line. So R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

11.2 Hypothesis Testing

To determine whether there is a statistically significant relationship between the variables, a hypothesis test with respect to the coefficients β_0 and β_1 must be conducted. Every statistical software package provides this hypothesis test and no additional calculations are necessary. For the hypothesis test to be valid, the error terms must be normally distributed. Given this assumption, we are using the following t-statistic with $n-2$ degrees of freedom. Note that the degrees of freedom decrease with every additional β. This becomes relevant in the case of multivariate regression. The test statistic is

$$\frac{\hat{\beta} - \beta}{se_{\hat{\beta}}} \sim t_{n-2}$$

The test statistic for β_1 can be written as

$$\frac{\hat{\beta}_1 - \beta_1}{se_{\hat{\beta}_1}} \sim t_{n-2}$$

where

$$se_{\hat{\beta}_1} = \sqrt{\frac{\sum (y_i - \hat{y})^2/(n-2)}{\sum (x_i - \bar{x})^2}}$$

The existence of a linear relationship between X and Y can be tested with the above t-statistic by specifying H$_0$: $\beta_1 = 0$. Tests of hypotheses concerning β_0 are much less frequent then tests concerning β_1.

11.2.1 Numeric Example using Used Car Data

We are going to use a used car data set relating the price to the mileage of the car. Note that the direction of the relationship is clear, i.e., mileage affects price and not vice versa. The data can be found in the file `honda.csv`. In R/RStudio, the regression can performed simply by the command `bhat = lm(price miles,data=honda)`.

In general, there are two problems with interpreting the intercept. First, if the range of x and y does not include the intercept then it is difficult to attach any meaning to the intercept. Second, the intercept can be negative although in reality, this could not be possible. In almost all regression models, we do not care about the intercept.

For the purpose of this example, we are using 25 observations and have divided the price and miles by 1000. Note that scaling the variables does not affect your statistical model in terms of significance.

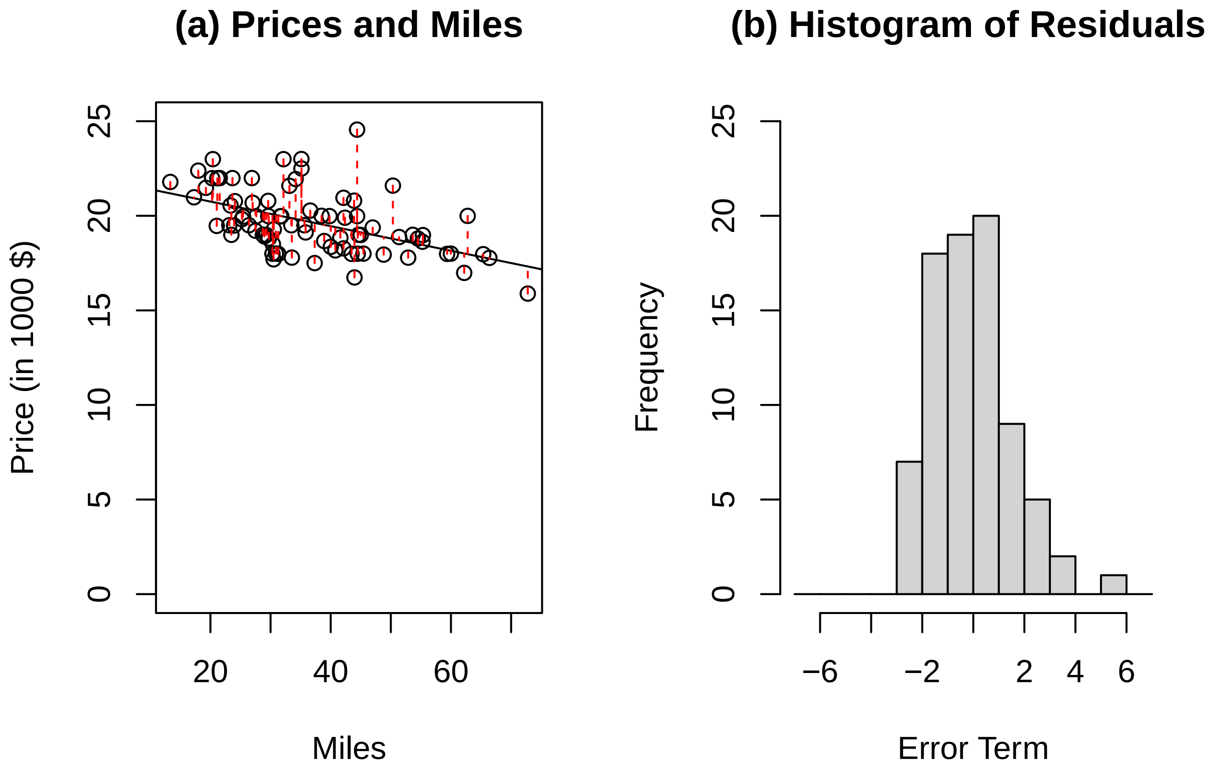


Figure 19: Example of regression line to model Honda prices as a function of miles. The red dashed lines in panel (a) represent the error terms associated with each observation. Panel (b) is the histogram associated with the error terms. The expected value of the error terms is zero and by assumption, the error terms are normally distributed.

Note that the average price is 18.963 and the average mileage is 38.922. Given the values above, we find that $\beta_1 = -246.07/3844.44 = -0.064$ and $\beta_0 = 18.963 - (-0.064) \cdot 38.922 = 21.45$. For the goodness of fit measure R^2, we have

$$R^2 = 1 - \frac{53.41}{69.16} = 0.2278$$

And for the standard error of β_1, we have

$$se_{\hat{\beta}_1} = \sqrt{\frac{53.41/23}{3844.44}} = 0.02458$$

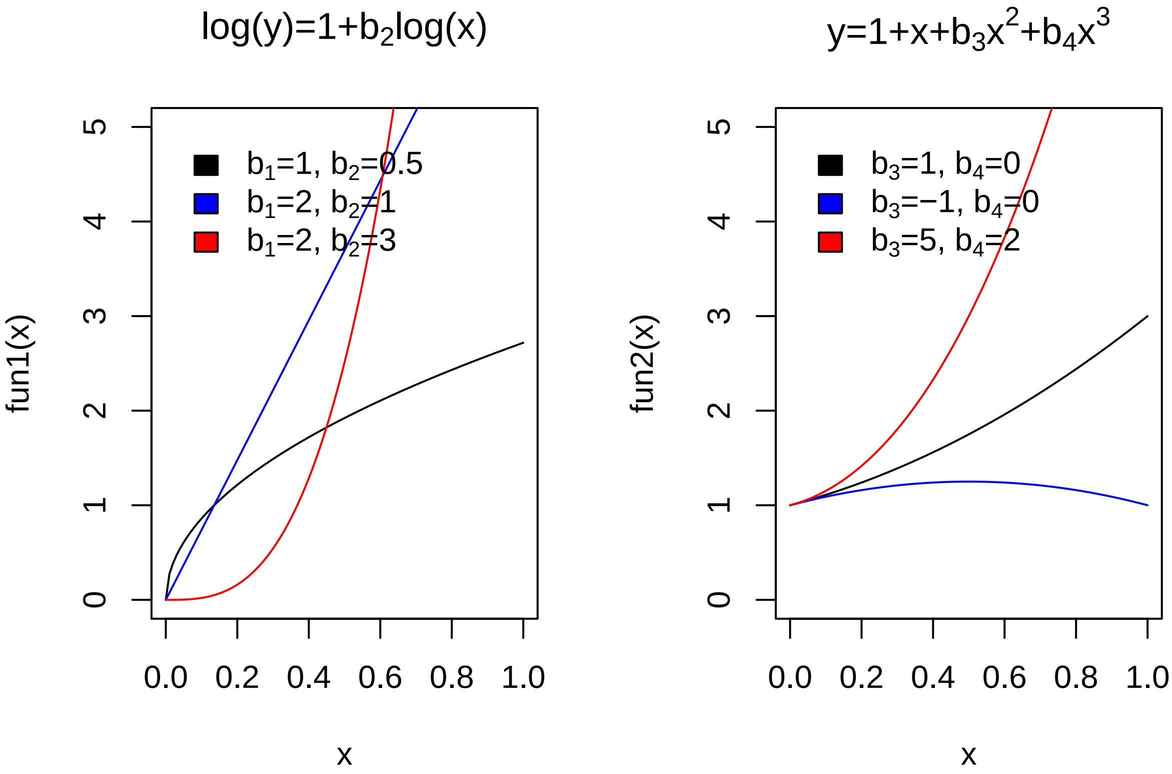
Since we have the intercept and slope coefficient, the regression line can be written as $price = 21.45 - 0.064 \cdot miles$. For example, having a car with 37.329 miles (in thousand), leads to $21.45 - 0.064(37.329) = 19.07$.

11.3 Functional Forms

Despite the fact that the regression model is linear, non-linear relationships can be measured. For example, the relation between consumption and income might be non linear since a change in consumption due to extra income may decrease with income. Or the relationship between income and education can exhibit a non-linear form because a change in income due to more education may decrease with more education. Consider the following relationship between y and x:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

If a nonlinear relation can be expressed as a linear relation by re-defining variables we can estimate that relation using ordinary least square. For the aforementioned equation, we can define new variables x_1 and x_2: $x_1 = x$ and $x_2 = x^2$. Note that this will lead to a multivariate relationship.



11.4 About the Importance of the Assumptions

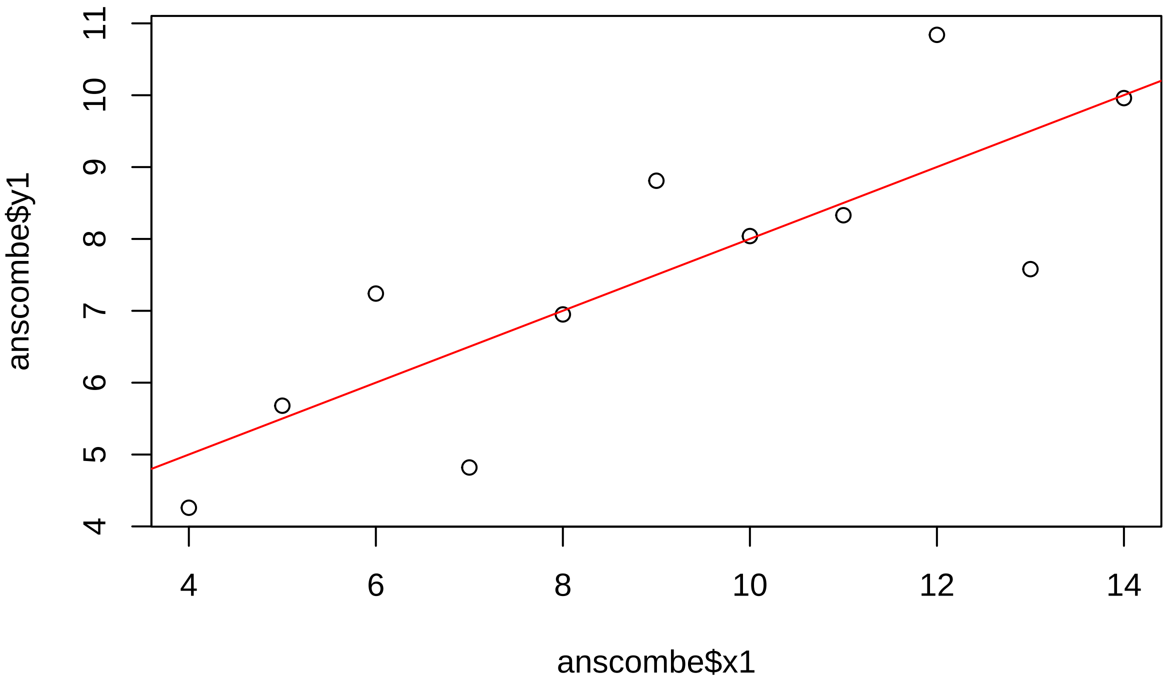
The data in `anscombe` illustrates the danger of simply relying on the regression output. The so-called Anscombe's Quartet includes $i = 1, \ldots, 4$ data series denoted y_i (dependent variable) and x_i (independent variable). Estimate the four regression models and compare the results and the conclusions you draw from the output. Next, plot the observations and include the fitted line. The regression output for the first set:

```r
bhat1 = lm(y1~x1,data=anscombe)
summary(bhat1)
```

```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

And the associated scatter plot with the regression equation:

```r
plot(anscombe$x1,anscombe$y1)
abline(bhat1,col="red")
```



11.5 Exercises

1. ***Accidents*** (***): Researchers at IUPUI attempt to predict the number of auto accidents in the city depending on temperature. They randomly select 30 days during the year and run a regression to determine if temperature significantly affected the number of accidents. Using the data `accidents`, I want you to manually re-create the table we have seen in class to calculate the slope and intercept coefficient and then use R to confirm your result. Note that it is best to copy the `accident` data and convert it into a regular Excel file for the first part of the exercise.
   a. With temperature as the independent variable and accidents as the dependent variable, create four new columns in Excel: (1) $x_i - \bar{x}$, (2) $y_i - \bar{y}$, (3) $(x_i - \bar{x})(y_i - \bar{y})$, and (4) $(x_i - \bar{x})^2$. From there, use the OLS equations provided in the slides to calculate slope and intercept.
   b. Run a simple bivariate regression using the command `lm()` in R and report the results. The results from the calculation with Excel and R must match.
2. ***Ohio Schools II*** (***): Consider the data sets `ohioincome` and `ohioscore`. In the section on hypothesis testing, the school districts were divided by median income into the top 25% and bottom 25%. In this exercise, two linear regression models are fitted to the data.
   a. In a first step, merge the data sets `ohioincome` and `ohioscore` by IRN.
   b. The first regression model is written as follows:

$$score = \beta_0 + \beta_1 \cdot medianincome$$

   Estimate the above equation using R and report the output. Interpret the coefficient β_1. Is it statistically significant?
   c. Do a scatter plot and include the regression line estimated above in the plot. Is the model a good fit for the data. Compare your answer to the one in the previous part which was based on the numerical output.

d. Estimate a second model written as:

$$score = \beta_0 + \beta_1 \cdot medianincome + \beta_2 \cdot medianincome^2$$

For this model, make sure to include the squared term by using the function `I()` in R. If you do not include it, R simply drops the last term. Report and interpret the output.

e. Do a scatter plot and include the (nonlinear) regression line estimated above in the plot. Is the model a good fit for the data. Compare your answer to the previous parts.

3. ***Indy Home Heating*** (***): Consider the data set `heating` which shows the consumption of natural gas and average temperature. Run a regression with *usage* as the dependent variable and *temperature* as the independent variable. Interpret the coefficients. Is the variable *usage* indicative of your total energy consumption over the time period covered in the data set?

12 Basic Multivariate Regression

This chapter extends the bivariate model to a multivariate model, i.e., the case with more than one independent variable. There is also a YouTube Video associated with this section:

- Multivariate Regression with R - Video
- Multivariate Regression with R - Slides

The following topics associated with multivariate regression models are covered in this chapter:

- Dummy variables
- Natural logarithm
- Functional forms
- Interaction Terms
- Multicollinearity

12.1 Introduction

Recall the bivariate regression model with one independent and one dependent variable:

$$y = \beta_0 + \beta_1 \cdot x_1 + \epsilon$$

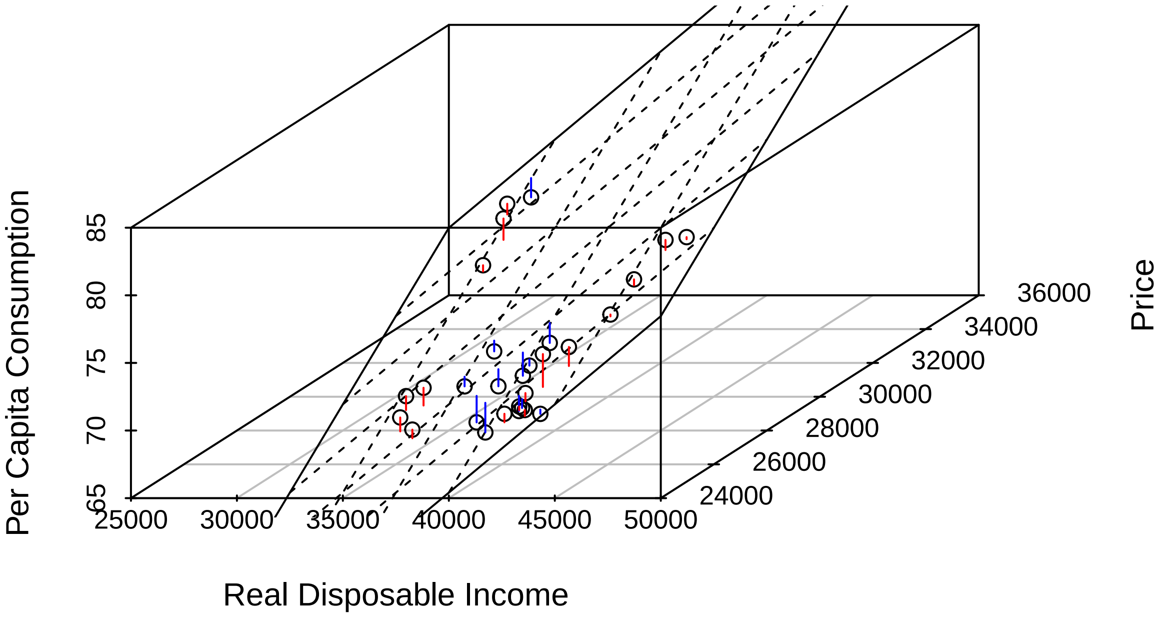
The multivariate linear regression model includes more than one independent variable and is simply an extension of the bivariate regression model:

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

Whether we consider the bivariate or multivariate model, the objective is always to minimize the sum of squared errors which has led to the name ordinary least square (OLS) model. The equation of a line can be determined using slope (β_0) and the intercept ($\beta 1$), i.e.:

$$E(y|x_1) = \beta_0 + \beta_1 \cdot x_1$$

The case of a regression model with two independent variable can still be represented in a 3-dimensional graph as depicted below



The purpose of the multivariate regression model is to measure the effect of independent variables on the dependent variable. It is crucial to control for everything else that could influence the dependent variable. For example, measuring the weekly grocery bill as a function of years of education might give you a statistically significant effect for education but if income is included, the effect for education might (most likely) disappear.

The first example involves estimating home values based on square footage and number of garage spots of a house in the 46268 ZIP code in Indianapolis. The data is contained in `indyhomes`.

```
indyhomes46268 = subset(indyhomes,zip==46268)
bhat = lm(price~sqft+garage,data=indyhomes46268)
summary(bhat)
```

```
##
## Call:
## lm(formula = price ~ sqft + garage, data = indyhomes46268)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -58780  -7817   1582   7886  51803
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81733.141  15896.004   5.142 5.20e-06 ***
## sqft           40.897      4.383   9.331 2.85e-12 ***
## garage      16580.964   7136.866   2.323   0.0245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Residual standard error: 20710 on 47 degrees of freedom
## Multiple R-squared:  0.675,  Adjusted R-squared:  0.6611 
## F-statistic:  48.8 on 2 and 47 DF,  p-value: 3.388e-12
```

Depending on the nature of the variables, it might be necessary to scale your variables for ease of interpretation. This might be necessary if coefficients are very large or very small. A rescaling, e.g., dividing income by 1000, does affect the coefficients and the standard errors but has no effect on the t-statistics.

12.2 Dummy Variables

So far, independent variables were quantitative such as price, income, square footage, miles, and so on. But very often, a qualitative characteristic such as religion or gender must be modeled. For this purpose, dummy variables that can be either 0 or 1 are used. Dummy variables represent a single qualitative characteristic. For example, consider the price (y_i) of a car depending on miles (x_i) and whether the car has all-wheel drive (AWD) or rear-wheel drive (RWD). This characteristic can be modeled using a dummy variable (d_i). If $d_i = 1$, the car has AWD and if $d_1 = 0$, the car has RWD. The regression equation can be written as follows:

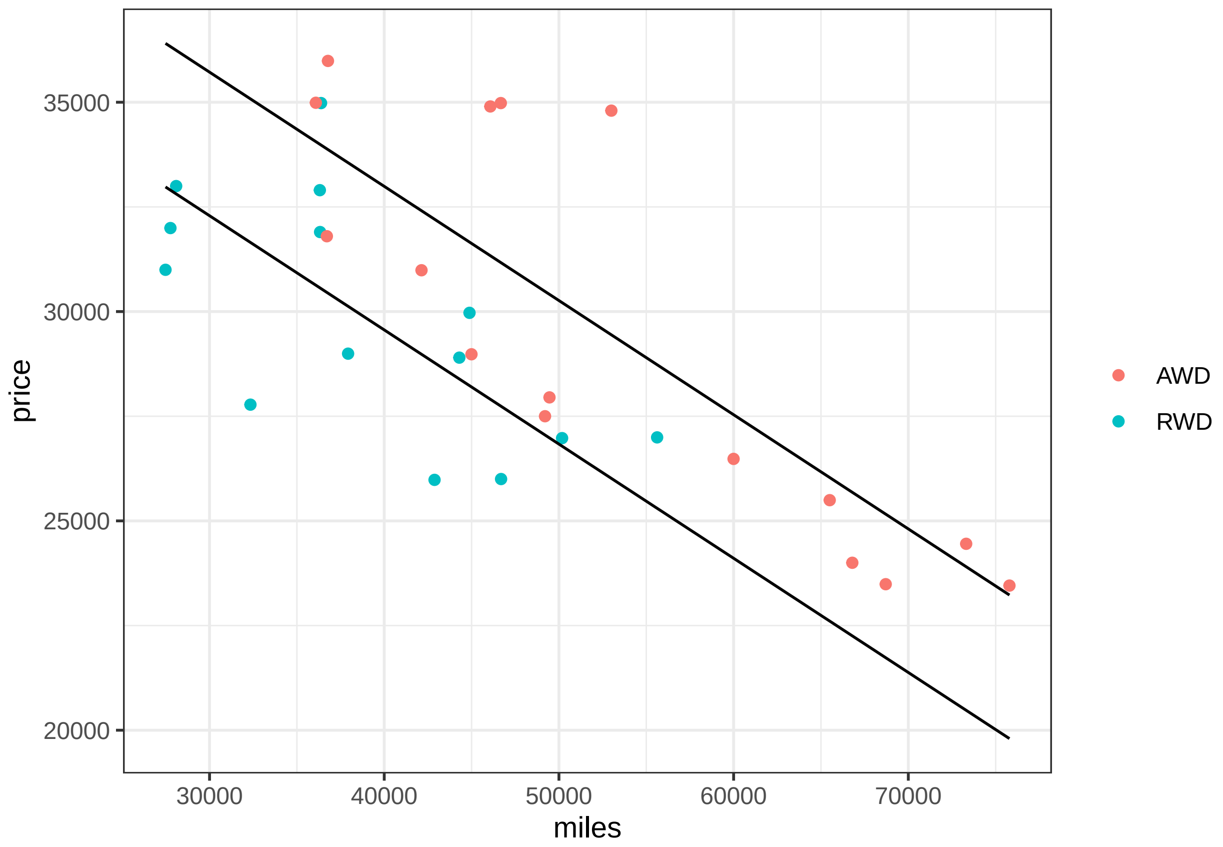
$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot d_i + \epsilon_i$$

his regression can theoretically be separated into two single equations:

- RWD: $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$
- AWD: $y_i = (\beta_0 + \beta_2) + \beta_1 \cdot x_i + \epsilon_i$

To interpret the dummy variables, it is necessary to know how it is coded. In the above case, if the coefficient β_2 is positive, then the dummay variable adds to the price. That is, the coefficient β_2 represents the value of all-wheel drive.

```
## 
## Call:
## lm(formula = price ~ miles + allwheeldrive, data = bmw)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3874.1 -1724.0  -176.5  1604.5  5355.0 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    4.047e+04  1.711e+03  23.660  < 2e-16 ***
## miles         -2.728e-01  4.044e-02  -6.745 3.05e-07 ***
## allwheeldrive  3.429e+03  1.063e+03   3.227  0.00327 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2449 on 27 degrees of freedom
## Multiple R-squared:  0.6287, Adjusted R-squared:  0.6012 
## F-statistic: 22.86 on 2 and 27 DF,  p-value: 1.553e-06
```



12.3 Natural Logarithm

Transforming the dependent and/or independent variables using the natural logarithm has some important and useful interpretations. Consider the following simple consumption equation in which both variables are in logarithmic form:

$$ln(consumption) = \beta_0 + \beta_1 \cdot \ln(income) + \epsilon$$

In this case, β_1 is the elasticity of consumption with respect to income, i.e., a 1% increase in income leads to a $\beta_1 \cdot 1\%$ increase in consumption. For example, if $\beta_1 = 0.4$, then a 1% increase in income will rise consumption by 0.4%. Note that the percentage increase is only an approximation for small changes.

Dep. Var.	Indep. Var	Interpretation
y	x	1 dollar change in x changes y by $\hat{\beta}$ dollars
$\ln(y)$	x	1 dollar change in x changes y by 100 $\times\hat{\beta}$ percent
$\ln(y)$	$\ln(x)$	1 percent change in x changes y by $\hat{\beta}$ percent
y	$\ln(x)$	1 percent change in x changes y by $\hat{\beta}/100$ dollars

```
bhat = lm(log(total)~yards+att+exp+draft1+veteran+changeteam+pbowlever,data=nfl)
summary(bhat)
```

```
##
## Call:
## lm(formula = log(total) ~ yards + att + exp + draft1 + veteran +
##     changeteam + pbowlever, data = nfl)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4899 -0.4998 -0.0801  0.4554  3.1959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.9289322  0.0767846 -12.098  < 2e-16 ***
## yards        0.0003566  0.0001411   2.527 0.011783 *
## att          0.0003927  0.0009657   0.407 0.684408
## exp          0.0108812  0.0160213   0.679 0.497312
## draft1       0.8876564  0.1132374   7.839 2.30e-14 ***
## veteran      0.6735244  0.1144567   5.885 6.88e-09 ***
## changeteam  -0.3095919  0.0893125  -3.466 0.000568 ***
## pbowlever    0.4093324  0.0936078   4.373 1.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7909 on 560 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:   0.55,  Adjusted R-squared:  0.5444
## F-statistic:  97.8 on 7 and 560 DF,  p-value: < 2.2e-16
```

12.4 Functional Form

To model non-linear relationships, an independent variable can be transformed by squaring it. For example, consider the relationship between income and food expenditure. The regular OLS assumes a linear relationship in the sense that an increase in income always leads to a proportional increase in food expenditure. In realty, there is likely a flattening out of food expenditure for high incomes because only so much money can be spent on food.

```
bhat = lm(log(total)~yards+att+exp+exp2+draft1+veteran+changeteam+pbowlever,data=nfl)
summary(bhat)
```

```
##
## Call:
## lm(formula = log(total) ~ yards + att + exp + exp2 + draft1 +
##     veteran + changeteam + pbowlever, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4477 -0.5010 -0.0807  0.4452  3.1638
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0270821  0.1228294  -8.362 4.93e-16 ***
## yards        0.0003353  0.0001426   2.351 0.019087 *
## att          0.0005137  0.0009728   0.528 0.597691
## exp          0.0702541  0.0601679   1.168 0.243452
## exp2        -0.0037576  0.0036704  -1.024 0.306398
## draft1       0.8910570  0.1132812   7.866 1.90e-14 ***
## veteran      0.5752784  0.1493617   3.852 0.000131 ***
## changeteam  -0.3221519  0.0901474  -3.574 0.000383 ***
## pbowlever    0.4158535  0.0938203   4.432 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7909 on 559 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.5509, Adjusted R-squared:  0.5445
## F-statistic: 85.71 on 8 and 559 DF,  p-value: < 2.2e-16
```

12.5 Interaction Effects

Interaction effects are used when the influence of one independent variable depends on the level of another independent variable. Suppose that you want to measure time spent volunteering (y) and you think that it depends on the marital status (x_1), the number of children (x_2), and some other independent variables (X). So you could have the following regression equation

$$y = \beta_1 \cdot x_1 \cdot x_2 + \beta \cdot X + \epsilon$$

$$\frac{dy}{dx} = \beta_1 x_2 + \beta_2$$

12.6 Exercises

1. ***Ohio Schools III*** (***): Consider the data sets `ohioincome` and `ohioscore`. In a first step, merge the two data sets by IRN. For all questions below, interpret the coefficients in terms of direction, magnitude, and statistical significance.

   a. Estimate the following equation and report the output.

   $$score = \beta_0 + \beta_1 \cdot medianincome + \beta_2 \cdot enrollment$$

   b. Estimate the following equation. Compare your answer to the previous part. Do the coefficients change in magnitude? How do you interpret the squared term?

   $$score = \beta_0 + \beta_1 \cdot medianincome + \beta_2 \cdot enrollment + \beta_3 \cdot medianincome^2$$

2. ***Honda vs. BMW*** (***): The data sets `honda` and `bmw` contain prices and mileage of used Honda and BMW cars in the Indianapolis area. For BMW, you have a dummy variable which indicates all-wheel drive ($allwheeldrive = 1$) or rear-wheel drive ($allwheeldrive = 0$).

   a. Run a regression with price as the dependent variable and miles as the independent variable for both cars (separately). Report the intercept and slope coefficients. Interpret your results, e.g., how does an increase in miles affect the price of the cars.
   b. Generate two scatter plots of the data and the fitted lines. For each car, I want the scatter plot and the fitted line in the same graph. What can you say about the difference in depreciation of the two cars.

3. ***WDI*** (**): Using the data in `wdi`, estimate the equation below for the year 2018. Report and interpret the results:

   $$fertrate = \beta_0 + \beta_1 \cdot gdp + \beta_2 \cdot litrate$$

4. ***Retail*** (***): This exercise will demonstrate the use of dummy variables to model so-called seasonality in the data `retail`. Note that time series analysis is a fairly complex topic and this question only serves as an introduction. Using the data in `retail`, estimate the following regression model:

   $$retail = \beta_0 + \beta_1 \cdot t + \sum_{m=1}^{11} \beta_m \cdot D_m$$

   where t represents a simple time trend and D_m are monthly dummy variables. Make sure to only include 11(!) monthly dummy variables. Is there seasonality in the data? Interpret.

5. ***Indy Homes I*** (***): The data `indyhomes` contains home values of two ZIP codes in Indianapolis. In this exercise, you will estimate the value of homes (dependent variable) based on a set of independent variables. The variables are mostly self-explanatory. The variables *levels* and garage refers to the number of *stories* and the garage parking spots, respectively.

   a. Create a dummy variable called *northwest* for the 46268 ZIP code.
   b. Report the results of the following regression equation

   $$\ln(price) = \beta_0+\beta_1\cdot\ln(sqft)+\beta_2\cdot northwest+\beta_3\cdot\ln(lot)+\beta_4\cdot bed+\beta_5\cdot garage+\beta_6\cdot levels+\beta_7\cdot northwest\cdot levels$$

   c. Interpret each coefficient from the previous part and how it affects $ln(price)$. How do you interpret the interaction term?
   d. What is the expected home value of a house in the 46228 ZIP code area with the following characteristics: 1800 sqft, 0.54 acres lot, 4 bedrooms, 3 bathrooms, 2 garage spots, and 1 story.

6. ***Pork Demand*** (***): In this exercise, you will estimate the per-capita pork demand as a function of pork prices and the prices of substitutes (beef and chicken) as well as real disposable income. Use the date `meatdemand` for this exercise. Estimate the following equation and interpret the coefficients. Are the signs of the coefficients what you would expect?

   $$\ln(q_{pork}) = \beta_0 + \beta_1 \cdot \ln(p_{pork}) + \beta_2 \cdot \ln(p_{chicken}) + \beta_3 \cdot \ln(p_{beef}) + \beta_4 \cdot \ln(rdi)$$

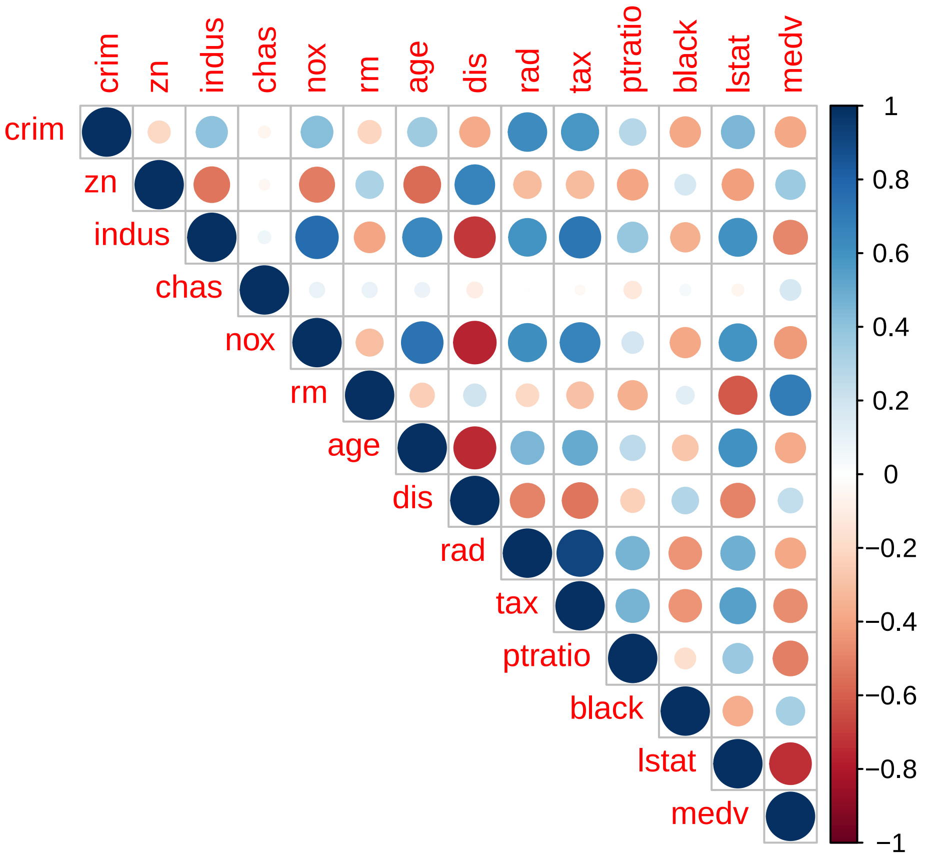
7. ***NFL I*** (***): This question will have you create a similar analysis to the one found in Berri et al. (2011). The corresponding data is in `nfl`:

   $$\ln(total) = \beta_0+\beta_1\cdot yards+\beta_2\cdot att+\beta_3\cdot exp+\beta_4\cdot exp^2+\beta_5\cdot draft1+\beta_6\cdot draft2+\beta_7\cdot veteran+\beta_8\cdot changeteam+\beta_9\cdot pbowlever+$$

   Report the output and interpret the coefficients in terms of statistical significance and direction (i.e., sign).

8. ***Boston*** (***): For this exercise, use the data set `boston`. In a first step, execute the following code:

```r
library(corrplot)
corr_matrix = cor(boston)
corrplot(corr_matrix,type="upper")
```



What does the resulting plot represent? In a second step, estimate the following model:

$$medv = \beta_0 + \beta_1 \cdot lstat + \beta_2 \cdot crim + \beta_3 \cdot age$$

Explain what exactly you estimated and what hypotheses are underlying the model. Lastly, estimate the model including all remaining independent variables. Are any of the results surprising?

9. ***BLM I*** (***): The following question is based on the article Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity. Note that we use a simplified version of the data set for this question. The dependent variable for this exercise is protest frequency ($totprotests$) and the independent variables are city population (pop), population density ($popdensity$), percent Black ($percentblack$), black poverty rate ($blackpovertyrate$), percent of population with at least a bachelor ($percentbachelor$), college enrollment ($collegeenrollpc$), share of democrats ($demshare$), and black police-caused deaths per 10,000 people ($deathsblackpc$). Interpret the output.

10. ***Furnished Apartments*** (***): Long-term furnished apartments are usually managed by companies. Some of those companies are more expensive than others. The dependent variable is $rent$. Estimate a regression model using all other columns as independent variables. Which companies are more expensive than others and by how much? Is there a difference between living in Berlin (city) or one of its close suburbs (Potsdam). On a side note, people in Potsdam prefer the city not to be called a suburb since it is fairly sizable state capital.

13 ANOVA

Analysis of Variance (ANOVA) models (also know as Dummy Variable Regression models) are regressions with only dummy variables. An ANOVA model with two independent variables can be written as follows:

$$y_i = \beta_0 + \beta_1 \cdot d_1 + \beta_2 \cdot d_2$$

where d_1 and d_2 are dummy variables. Consider the following model using the `nfl` data for the year 2005:

$$total = \beta_0 + \beta_1 \cdot draft1 + \beta_2 \cdot veteran$$

where *draft1* and *veteran* are dummy variables. That is, if $draft1 = 1$, then the player was selected in the first draft round. If $veteran = 1$, then the player has played multiple seasons in the NFL. To distinguish *j* categories only *j-1* dummy variables are needed. Otherwise, we have perfect multicollinearity. The category without a dummy variable is the base category.

```r
bhat = lm(total~draft1+veteran,data=subset(nfl,year=2005))
```

```
## Warning: In subset.data.frame(nfl, year = 2005) :
##  extra argument 'year' will be disregarded
```

```r
summary(bhat)
```

```
##
## Call:
## lm(formula = total ~ draft1 + veteran, data = subset(nfl, year = 2005))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.340 -1.865 -0.702  0.792 32.429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9999     0.2534   3.945 8.63e-05 ***
## draft1        2.6422     0.4262   6.200 8.81e-10 ***
## veteran       1.6083     0.2820   5.703 1.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.083 on 848 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared:  0.05191,    Adjusted R-squared:  0.04968
## F-statistic: 23.22 on 2 and 848 DF,  p-value: 1.526e-10
```

For a player who was not drafted in the first round and is not a veteran, the income is close to $1 million. Note that both dummy variables are statistically significant. Note that the R-squared is very low.

13.1 Exercises

1. ***Indy Homes II*** (**)Consider the data in `indyhomes`. A real estate agent once mentioned that homes built in the eighties are of lower quality and thus, are cheaper. One possibility to asses this claim is with an ANOVA model. Add three new dummy variables to the data frame: eighties (equal to 1 if the home was built in any year between 1980 to 1989), northwest (equal to 1 if the home is in the 46268 ZIP code), and singlestory (equal to 1 if the home has one story). Estimate the following ANOVA model:

   $$price = \beta_0 + \beta_1 \cdot d_{80s} + \beta_2 \cdot d_{northwest} + \beta_3 \cdot d_{single}$$

   What do you conclude? Next, add square footage to the model. What has changed in terms of coefficients? How do you interpret that change?

14 Violating Assumptions

This chapter introduces the detection and correction of problems with the estimation procedure due to the violation of the key assumptions necessary for the OLS model to work. The following R packages are needed for this chapter: `car`, `lmtest`, `orcutt`, and `sandwich`.

There are also slides associated with this chapter:

- Violating Assumptions - Slides

14.1 Heteroscedasticity

A key assumption of the OLS model is homoscedasticity error terms. That is, the error variance is constant:

$$Var(\epsilon_i) = \sigma^2$$

With heteroscedasticity, the variance of the error term is not constant:

$$Var(\epsilon_i) = \sigma_i^2$$

For a bivariate regression model with heteroscedastic data, it can be shown that

$$Var(\hat{\beta}_1) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

This is different from the variance of the coefficient estimate under homoscedasticity:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}$$

Unbiasedness of the OLS estimator is not affected but the variance of β_1 will be larger compared to other estimators. Note that the measure of R^2 is unaffected by heteroscedasticity. Homoscedasticity is needed to justify the t-test, F-test, and confidence intervals. The F-statistic does no longer have an F-distribution. In short, hypothesis tests on the β-coefficients are no longer valid.

If σ_i^2 was known, the use of a Generalized Least Squares (GLS) model would be appropriate:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Dividing both sides by the known variance:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{\epsilon_i}{\sigma_i}$$

If $\epsilon_i^* = \epsilon_i / \sigma_i$, then it can be shown that $Var(\epsilon_i^*) = 1$, i.e., constant. Under the usual OLS model:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i \right)^2$$

Under GLS model:

$$\sum_{i=1}^{N} w_i e_i^2 = \sum_{i=1}^{N} w_i \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i \right)^2$$

That is, GLS minimizes the weighted sum of the residual squares. Since in reality, the variance of σ^2 is not known, other techniques have to be employed to obtain so-called heteroscedasticity-consistent (HC) standard errors. But first, two tests are introduced to detect heteroscedasticity.

14.1.1 Detecting Heteroscedasticity

Two test are presented to detect heteroscedasticity:

- Goldfeld-Quandt Test (1965)
- Breusch-Pagan-Godfrey Test (1979)

The steps necessary for the **Goldfeld-Quandt Test** are as follows:

1. Sort observations by ascending order of the dependent variable.
2. Pick C as the number of central observations to drop in the middle of the dependent variable.
3. Run two separate regression equations, i.e., with the "lower" and "upper" part.
4. Compute
$$\lambda = \frac{RSS_2/df}{RSS_1/df}$$
5. λ follows an F-distribution.

The Goldfeld-Quandt Test can be illustrated with `gqtestdata`. In a first step, the data is separated into two groups with $C = 6$. In a second step, both groups are used to run a regression. And lastly, λ is calculated.

```
gqtestdata1         = gqtestdata[1:22,]
gqtestdata2         = gqtestdata[29:50,]
bhat1               = lm(price~sqft,data=gqtestdata1)
bhat2               = lm(price~sqft,data=gqtestdata2)
lambda              = sum(bhat2$residuals^2)/sum(bhat1$residuals^2)
```

Of course, there is also a function in R called gqtest which simplifies the procedure.

```
library(lmtest)
bhat                = lm(price~sqft,data=gqtestdata)
gqtest(bhat,fraction = 6)
```

```
##
##  Goldfeld-Quandt test
##
## data:  bhat
## GQ = 8.0391, df1 = 20, df2 = 20, p-value = 9.811e-06
## alternative hypothesis: variance increases from segment 1 to 2
```

In any case, the hypothesis of homoscedasticity is rejected for `gqtestdata`.

The **Breusch-Pagan-Godfrey Test** is an alternative and does not rely on choosing C as the number of central observations to be dropped. The steps include the following:

1. Run a regular OLS model and obtain the residuals.
2. Calculate
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N}$$
3. Construct the variable p_i as follows: $p_i = e_i^2/\hat{\sigma}^2$
4. Regress p_i on the X's as follows
$$p_i = \alpha_0 + \alpha_1 \cdot x_{i1} + \alpha_2 \cdot x_{i2} + \dots$$
5. Obtain the explained sum of squares (ESS) and define $\Theta = 0.5 \cdot ESS$. Then $\Theta \sim \chi^2_{m-1}$.

The much simpler procedure is to use the function `bptest()` in R.

```
library(lmtest)
bhat                = lm(price~sqft,data=gqtestdata)
bptest(bhat)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bhat
## BP = 14.579, df = 1, p-value = 0.0001344
```

14.1.2 Correcting Heteroscedasticity

To correct for heteroscedasticity, robust standard errors must be obtained.

```
bhat = lm(price~sqft,data=gqtestdata)
summary(bhat)
vcov = vcovHC(bhat)
coeftest(bhat,vcov.=vcov)
```

Note that there are multiple variations to calculate the standard error and thus, it is possible for slight variations among the results from different packages.

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 e_i^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The square root of the following equation is called heteroscedastic robust standard error:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{N} \hat{r}_{ij}^2 e_i^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Standard errors can be either larger or smaller. Note that in this example, we do not know whether heteroscedasticity is present or not.

14.2 Multicollinearity

Multicollinearity describes the situation in which two or more independent variables are linearly related. Under perfect multicollinearity:

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k = 0$$

where λ_i are constants that are not all zero simultaneously. For example, consider $x_1 = \{8, 12, 15, 17\}$, $x_2 = \{24, 36, 45, 51\}$, and $x_3 = \{2, 3, 3.75, 4.25\}$. In this case, $\lambda_1 = 1$, $\lambda_2 = -1/5$, and $\lambda_3 = 2$. Note, multicollinearity refers to linear relationships! Including a squared or cubed term is not an issue of multicollinearity. It can be shown that the variance of the estimator increases in the presence of multicollinearity. There are various indications that the data suffers from multicollinearity:

- High R^2 but few significant variables
- Fail to reject the hypothesis for H$_0$: $\beta_i = 0$ based on t-values but rejection all slopes being simultaneously zero based on F-test.
- High correlation among explanatory variables
- Variation of statistically significant variables between models.

14.2.1 Variance Inflated Factors (VIF)

Identifies possible correlation among multiple independent variables and not just two as in the case of a simple correlation coefficient. Consider the model:

$$y_i = \beta_0 + \beta_k x_{ik} + \epsilon_i$$

The estimated variances of the coefficient β_k is written as

$$Var(\beta_k)^* = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2}$$

Without any multicollinearity, this variance is minimized. If some some independent variables are correlated with the independent variable k, then

$$Var(\beta_k) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2} \cdot \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 if variable x_k is taken as the dependent variable. The VIF can be written as

$$\frac{Var(\beta_k)}{Var(\beta_k)^*} = \frac{1}{1-R_k^2}$$

If $VIF = 1$, then there is no relationship between the variable x_k and the remaining independent variables. Otherwise, $VIF > 1$. In general, the interpretation is as follows:

- VIF of 4 warrants attention
- VIF of 10 indicates a serious problem.

14.2.2 Examples

To illustrate the concept of multicollinearity, the data set from `nfl` is used (Berri et al. (2011)). The first model includes total salary as the dependent variable and the following independent variables: prior season passing yards, pass attempts, experience (squared) in the league, draft round pick, veteran (more than 3 years in the league), pro bowl appearance, and facial symmetry.

```
bhat = lm(log(total)~yards+att+exp+exp2+draft1+draft2+veteran+changeteam+pbowlever+symm,data=nfl)
summary(bhat)
```

After estimating the results, the function `vif()` from the package `car` is used:

```
library(car)
vif(bhat)
```

```
##      yards        att        exp       exp2     draft1     draft2    veteran changeteam  pbowlever
##  32.547700  30.920282  39.889877  26.715342   1.621048   1.228091   5.253525   1.194254   1.581753
```

The results indicate multicollinearity for *yards*, *att*, and experience. Passings yards and attempts may be correlated and thus, one of them (*att*) is dropped.

```
bhat = lm(log(total)~yards+exp+exp2+draft1+draft2+veteran+changeteam+pbowlever+symm,data=nfl)
summary(bhat)
```

```
vif(bhat)
```

```
##      yards        exp       exp2     draft1     draft2    veteran changeteam  pbowlever       symm
##   1.460849  39.339639  26.162804   1.616171   1.227479   5.253502   1.141435   1.569621   1.052906
```

This improves the estimation but experience (and its squared term) are still problematic. The last estimation removes experience and the VIF terms are now in the acceptable range.

```
bhat = lm(log(total)~yards+draft1+draft2+veteran+changeteam+pbowlever+symm,data=nfl)
summary(bhat)
```

```
vif(bhat)
```

```
##      yards     draft1     draft2    veteran changeteam  pbowlever       symm
##   1.406241   1.653634   1.229459   1.976506   1.101988   1.406095   1.010855
```

The important part is that the conclusion of the paper has not changed with regard to facial symmetry.

14.3 Other Issues and Problems with Data

More serious problems than heteroscedasticity:

- Functional form mis-specification
- Measurement error
- Missing data: Estimating a standard regression model is not possible with missing values. All statistical software packages ignore missing data. Missing data is a minor problem if it is due to random error.

Missing data can be problematic if it is systematically missing, e.g., missing education data for people with lower education
- Non-random samples
- Outliers

14.4 Autocorrelation

The correlation of error terms is called autocorrelation. The issue usually arises if there is a time component in the data. Recall the main types of data available for research:

- Cross-sectional data (multiple observations at same time point)
- Time series data (one variable observed over time)
- Pooled data (multiple observations at different time points)
- Panel data (same observations at different time points)

There is a distinction between serial correlation and autocorrelation:

- Serial correlation: Correlation between two series
- Autocorrelation: Correlation with lagged variables

The OLS estimator is still unbiased but there is no longer minimum variance since $E(\epsilon_i \epsilon_j) \neq 0$. Autocorrelation is unlikely for cross-sectional data except in the case of spatial auto-correlation. One cause of autocorrelation could be inertia in economic variables. For example, variables such as income, production, or employment increase after a recession. But there are a number of other reasons for autocorrelation.

Autocorrelation could be caused by specification bias due to excluded variables or incorrect functional forms. For example, assume that the correct equation is

$$q_{beef} = \beta_0 + \beta_1 \cdot p_{beef} + \beta_2 \cdot p_{income} + \beta_3 \cdot p_{pork} + \epsilon_t$$

The estimated equation is:

$$q_{beef} = \beta_0 + \beta_1 \cdot p_{beef} + \beta_2 \cdot p_{income} + \upsilon_t$$

The error terms in both equations are denoted ϵ_t and υ_t, respectively. This results in a systematic patters of υ_t:

$$\upsilon_t = \beta_3 \cdot p_{pork} + \epsilon_t$$

Correlation between the error terms can also be caused by specifying an incorrect functional form. Assume that the correct equation is written as follows:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \epsilon_i$$

But the estimated equation is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Serial correlation is caused by lagged terms in the regression equation:

$$consumption_t = \beta_0 + \beta_1 \cdot income_t + \beta_3 \cdot consumption_{t-1} + \epsilon_t$$

The issues of lagged terms will be covered in the part on dynamic regression and time series and this section serves only as an introduction to first-order autoregressive schemes. Consider the model:

$$y_t = \beta_0 + \beta_1 \cdot x_t + \upsilon_t$$

Assume the following form of υ:

$$\upsilon_t = \rho \cdot \upsilon_{t-1} + \epsilon_t$$

his is called a first-order autoregressive AR(1) scheme. An AR(2) would be written as

$$\upsilon_t = \rho_1 \cdot \upsilon_{t-1} + \rho_2 \cdot \upsilon_{t-2} + \epsilon_t$$

This can be illustrated with simulated data. Consider the following model:

$$y_t = 1 + 0.8 \cdot x_t + \upsilon_t$$

and assume the following form of υ:

$$\upsilon_t = 0.7 \cdot \upsilon_{t-1} + \epsilon_t$$

1. Simulate the above model 100 times 2. Compare variance of coefficients under different two different methods: (1) OLS and (2) Cochrane-Orcutt

14.4.1 Durbin Watson d-Test

The test statistic of the Durbin-Watson test is written as:

$$d = \frac{\sum_{t=2}^{N} (e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2}$$

Assumptions underlying the test are

- No intercept
- AR(1) process, i.e., $\upsilon_t = \rho \upsilon_{t-1} + \epsilon_t$
- No lagged independent variables

Original papers derive lower (d_L) and upper (d_U) bounds, i.e., critical values, that depend on N and k only.

- $d \approx 2 \cdot (1 - \rho)$ and since $-1 \leq \rho \leq 1$, we have $0 \leq d \leq 4$.

Rule of thumb indicates that $d = 2$ signals no problems.

14.4.2 Breusch-Godfrey Test

Consider the following model $y_t = \beta_0 + \beta_1 x_t + \upsilon_t$ with the following error term structure:

$$\upsilon_t = \rho_1 \upsilon_{t-1} + \rho_2 \upsilon_{t-2} + \cdots + \rho_p \upsilon_{t-p} + \epsilon_t$$

The null hypothesis for the test is expressed as follows:

- H_0: $\rho_1 = \rho_2 = \cdots = \rho_p = 0$

When the following regression is executed:

$$\hat{\upsilon}_t = \alpha_0 + \alpha_1 \cdot x_t + \hat{\rho}_1 \cdot \hat{\upsilon}_{t-1} + \hat{\rho}_2 \cdot \hat{\upsilon}_{t-2} + \cdots + \hat{\rho}_p \cdot \hat{\upsilon}_{t-p} + \epsilon_t$$

Then

$$(n - p) \cdot R^2 \sim \chi_p^2$$

Wages and Productivity in the United States 1959-1998 Consider the data in *business.csv* and do the following:

1. Plot the data in a scatter plot
2. Run the regression in level form as well as log format
3. Plot the diagnostic plots.
4. Run the Durbin-Watson test and the Breusch-Godfrey test. What do you conclude?
5. Run the regression by (1) including a trend variable and (2) a squared term but no trend.

14.5 Exercises

1. ***WDI and Heteroscedasticity*** (***): Using the data in `wdi`, estimate the following equation for the year 2018 and report the results:

   $$fertrate = \beta_0 + \beta_1 \cdot gdp + \beta_2 \cdot litrate$$

   Conduct a test for heteroscedasticity of your choice. Is there heteroscedasticity present in the model? If yes, execute additional calculations to correct for it and report the results.

2. ***Indy Homes III*** (***): The data `indyhomes` contains home values of two ZIP codes in Indianapolis. The model estimates the home value (dependent variable) based on a set of independent variables. The variables *levels* and *garage* refers to the number of stories and garage spots, respectively. The remaining variables are self-explanatory.

   a. Create a dummy variable called *northwest* for the 46268 ZIP code.
   b. Report and interpret the results of the regression equation that uses ln(*price*) as the dependent variables and the folloing as independent variables: ln($sqft$), *northwest*, ln(*lot*), *bed*, *garage*, and $levels$. In addtion, include an interaction between northwest and levels.
   c. What is the expected home value of a house in the 46228 ZIP code area with the following characteristics: 1900 sqft, 0.65 acres lot, 3 bedrooms, 3 bathrooms, 2 garage spots, and 2 story.
   d. Conduct a Breusch-Pagan-Godfrey test for heteroscedasticity. What do you conclude?
   e. Estimate the above model with heteroscedasticity-consistent (HC) standard errors. What changes compared to the model from Part b?

3. ***WDI and Multicollinearity*** (***): Use the command `subset()` on the WDI data and to select the variables *fertrate*, *gdp*, *litrate*, *lifeexp*, and *mortrate* for the year 2015. Estimate the following model

   $$fertrate = \beta_0 + \beta_1 \cdot gdp + \beta_2 \cdot litrate + \beta_3 \cdot lifeexp + \beta_4 \cdot mortrate$$

   Interpret the results. What do you conclude in terms of statistical significance and the value of R^2? Use the function `vif` from the package `car`. What can you say about the issue of multicollinearity in this case? Correct the issue of multicollinearity by adjusting your model.

15 Binary Choice

There are slides and a YouTube Video associated with this chapter:

- Binary Choice - Slides
- Binary Choice - Video

Binary choice models are part of a large class of so-called qualitative choice models which are used for qualitative dependent variables. Consider the following outcomes of interest:

- Is a person in the labor force?
- Will an individual vote yes on a particular issue?
- Did a person watch the last Super Bowl?
- Have you purchased a new car in the past year
- Did you do any charitable contributions in the past year?
- Did you vote during the last election?
- Does an individual recidivate after being released from prison?

For those questions, the dependent variable is either 0 ("no") or 1 ("yes"). For binary choice models, the outcome is interpreted as a probability, i.e., what is the probability of a person to answer "yes" to those questions.

In the next chapter, the model is expanded to consider more than binary outcomes. Those models include categorical dependent variable that are either naturally ordered or have no ordering. Examples of naturally ordered categorical variables are:

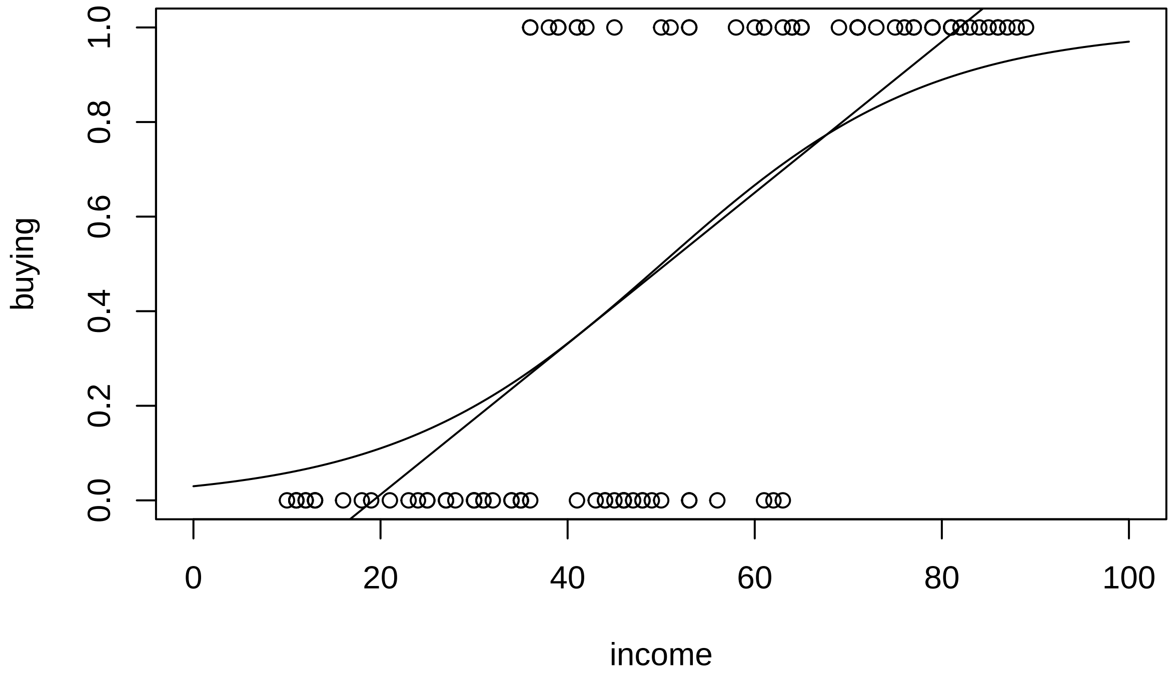
- Level of happiness: Very happy, happy, ok, or sad.
- Intention to get a COIVID-19 vaccine: Definitely yes, probably yes, probably no, or definitely no

Two examples about categorical dependent variable which have no ordering are:

- Commute to campus: Bike, car, walk, or bus
- Voter registration: Democrat, Republican, or independent

For all those models, the outcome of interest is the probability to fall into a particular category. For binary choice models which are considered in this chapter, the outcome of interest is the probability to fall into the 1 ("yes") category. For binary choice models, y takes one of two values: 0 or 1. And the model will specify $Pr(y = 1|x)$ where x are the independent variables.

Consider the decision to purchase organic food. Assume that you have data about the income of respondents as well as information if they purchase organic food. The purchase decision ("yes" or "no") is on the vertical axis and the income is on the horizontal axis.



Remember that the probability has be bounded between 0 and 1. Hence, we need to find a function $G(z)$ such that $0 \leq G(z) \leq 1$ for all values of z and $P(y = 1|x) = G(z)$. Popular choices for $G(z)$ are the cumulative normal distribution function ("Probit Model") and the logistic function ("Logit Model"). For what follows, let $z = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$. For the probit model, $G(z)$ is written as

$$Pr(y = 1) = G(z) = \Phi(z)$$

where Φ represents the cumulative normal. And for the logit model, $G(z)$ is written as

$$Pr(y = 1) = G(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

The interpretation of the logit and probit estimates is not as straightforward as in the multivariate regression case. In general, we care about the effect of x on $P(y = 1|x)$. The sign of the coefficient shows the direction of the change in probability. The approximation to the marginal effect if x is roughly continuous:

$$\Delta P(y = 1|x) \approx g(\hat{\beta}_0 + x \cdot \hat{\beta}) \cdot \beta_j \cdot \Delta x_j$$

To obtain the marginal effects in R, an additional step is necessary. Let us illustrate the binary choice model using the data set `organic` and a logit model. The results of interest for the binary choice model are the (1) coefficient estimates, (2) marginal effects, and (3) predicted probabilities.

15.1 Binary Choice Estimation in R

There are (at least) two possibilities to obtain the coefficient estimates in R. The first is using the built in R command `glm()`:

```
bhat_glm_logit = glm(buying~income,family=binomial(link="logit"),data=organic)
summary(bhat_glm_logit)
```

```
##
## Call:
## glm(formula = buying ~ income, family = binomial(link = "logit"),
##     data = organic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.87557    1.13842  -5.161 2.45e-07 ***
## income       0.11709    0.02247   5.211 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  70.931  on 98  degrees of freedom
## AIC: 74.931
##
## Number of Fisher Scoring iterations: 6
```

Note that interpretation of the coefficients is slightly different from the regular linear model. The sign of the coefficient estimate for income is interpreted as the direction in which the probability changes. In this case, the coefficient is positive and thus, an increase (decrease) in income leads to an increase (decrease) in the probability of purchasing organic food. In addition, the coefficients are statistically significant. As aforementioned, the coefficients do not indicate the marginal effects though. To calculate the marginal effects, a slightly different approach is necessary. Let us first look at the second approach of obtaining the coefficient estimates and the R package mfx is required to do so.

```
bhat = logitmfx(buying~income,data=organic)
summary(bhat$fit)
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = data,
##     start = start, control = control, x = T)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.87557    1.13842  -5.161 2.45e-07 ***
## income       0.11709    0.02247   5.211 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  70.931  on 98  degrees of freedom
## AIC: 74.931
##
## Number of Fisher Scoring iterations: 6
```

The results are identical as before using the `glm()` but the command `logitmfx()` allows for the calculation of the marginal effects as well. This is done with the command the `bhat\$mfxest`.

```
bhat$mfxest
```

```
##             dF/dx   Std. Err.        z        P>|z|
## income 0.02919553 0.005634262 5.181785 2.197728e-07
```

It is important to note that the marginal effects are taken at the mean of the independent variables. To calculate the marginal effects at specific points, the command `margins()` must be used. Before, we used the command `glm()` to calculate the logit coefficients. The reason for using the `glm()` is that it allows us to calculate the predicted probabilities. Consider the example to purchase organic food and assume that there are three new respondents with income levels $25,000, $50,000, and $75,000. To predict the probability of those individuals purchasing organic food, the following functions can be used:

```
datablock = data.frame(income=c(25,50,75))
predict(bhat_glm_logit,newdata=datablock,type="response")
```

```
##         1         2         3
## 0.0498116 0.4946870 0.9481377
```

15.2 Exercises

1. ***Hybrid Cars*** (***): Consider the problem of choosing whether or not to purchase a hybrid vehicle (e.g., the Toyota Prius, Honda Civic Hybrid, Ford Escape). As an analyst, you assume that whether or not an individual purchases a hybrid depends upon the current price of gasoline (*gas*), the difference in purchase price of a hybrid vehicle compared to a comparably equipped vehicle (*increment*), college education which is represented by a dummy variable that equals 1 if individual has completed college and equals 0 otherwise (*college*), and a dummy variable that equals 1 if the individual is a member of an environmental organization, e.g., Nature Conservancy, National Audubon Society, (*env*). Answer the following questions contained in `hybrid`:

   a. Provide summary statistics (i.e., mean, minimum, and maximum) for each variables.

   b. Estimate a regression model that allows to calculate the probability that a person would buy a hybrid.

   c. Using your parameter estimates, compute the probability that the following “types’’ of individuals will buy a hybrid:

      - Type I: gasoline price = 2.50; difference in purchase price = 1,500; college = 0; and member of an environmental organization = 0
      - Type II: gasoline price = 3.50; difference in purchase price = 500; college = 1; and member of an environmental organization = 1
      - Type III: gasoline price = 3.00; difference in purchase price = 1,000; college = 1; and member of an environmental organization = 0

   d. Given the above probabilities, calculate the marginal effect that gasoline prices have on the probability that each of the three “types’’ of individuals will purchase a hybrid vehicle.

2. ***EV Data*** (***): Using the `evdata`, estimate a logit model which calculates the probability that a consumer would purchase a hybrid (choice=2), plug-in hybrid (choice=3), or electric vehicle (choice=4) as opposed to a gasoline car (choice=1). In a first step, you should create a new variable that is equal to 0 if the person purchases a gasoline vehicle and is equal to 1 for everything else. Include the following independent variables in your model: *age*, *numcars*, *politics*, $female$, *edu*, and *income*.

3. ***Voting Behavior*** (***): We are interested in the determinants of voting based on data from the General Social Survey (GSS). The following variables are hypothesized to be a determinant of voting behavior: *age*, *gunownership* (yes=1), *stanceondeathpenalty* (in favor=1), and *educationlevel* (higher number is associated with a higher education level). Run a probit and logit model and report the results. Calculate the marginal probabilities.

4. ***Fatal Car Accidents*** (***): Research suggests that black motorists are driving slower during the day than at night because police officers could engage in racial profiling more easily during the day than at night (Kalinowki et al., 2021). Assuming that lower speeds lead to fewer accidents, the aforementioned hypothesis can be tested via data from the Fatality Analysis Reporting System (FARS). Use the data in `fatalcity` and `fatalstate` to run a logit model with *black* as the dependent variable and *rain*, *daylight*, and *year* as the independent variables. In addition, use $fips$ and *state* as independent variables as well but use the function `factor()` to include them, e.g., `+factor(state)`. Based on the two separate regression models, what do you conclude?

5. ***Biking*** (***): Consider the data in `hhpub`. Your dependent variable will be called *biking* and indicated whether respondents in the survey use a bike or never at all.

6. ***Organic Fruit Purchases*** (***): The data in `fpdata` contains survey results on how many organic tomoatoes and organic strawberries a household purchases. For the

16 Qualitative Choice Models

The last chapter introduced binary choice models that answer questions such as:

- Did you vote during the last election?
- Does an individual recidivate after being released from prison?

In this chapter, cases of dependent variables with more than two categories are considered. The categorical dependent variables can be with or without natural ordering. Here are some examples of natural ordering:

- Level of happiness: Very happy, happy, okay, or sad
- Intent to get vaccinated: Definitely yes, probably yes, probably no, definitely no

Examples without natural ordering:

- Type of car to purchase: Passenger car, Pick-up truck, van, convertible

Those models are also known as Discrete Choice Models. The packages required for this section are mlogit, erer, MASS, AER, glmmML, and nnet. The data required are `fpdata` and `evdata` in addition to `Fishing` and `TravelMode` which are part of the packages mlogit and AER. You can load them into R with the following commands:

- `data("Fishing",package="mlogit")`
- `data("TravelMode",package="AER")`

There is also a YouTube Video associated with this chapter.

16.1 Ordered Logit Model

Suppose that respondents are asked if they are going to get a COVID-19 vaccine. The choices are

- Definitely yes
- Probably yes

- Probably no
- Definitely no

The ordered logit model assumes the presence of a latent (unobserved by the researcher) variable y^*:

$$y_i^* = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

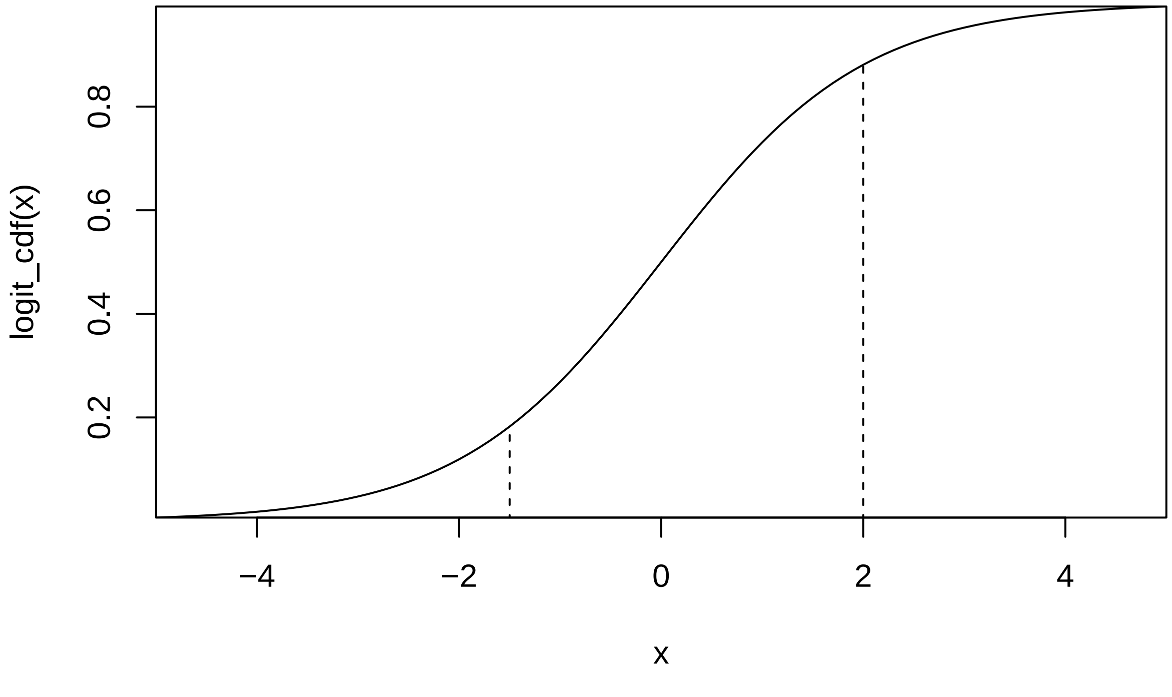
In the case of the vaccine model, this may be a measure of trust in vaccines. What the researcher does measure is an m-alternative ordered model:

$$y_i = j \quad \text{if} \quad \alpha_{j-1} < y_i^* \leq \alpha_j \quad \text{for} \quad j = 1, \ldots, m$$

where $\alpha_0 = -\infty$ and $\alpha_m = \infty$. In this case, we have

$$\begin{aligned} Pr(y = j) &= Pr(\alpha_{j-1} < y_i^* \leq \alpha_j) \\ &= Pr(\alpha_{j-1} < \beta_0 + \beta_1 \cdot x_i + \epsilon_i \leq \alpha_j) \\ &= Pr(\alpha_{j-1} - \beta_0 - \beta_1 \cdot x_i < \epsilon_i \leq \alpha_j - \beta_0 - \beta_1 \cdot x_i) \\ &= F(\alpha_j - \beta_0 - \beta_1 \cdot x_i) - F(\alpha_{j-1} - \beta_0 - \beta_1 \cdot x_i) \end{aligned}$$

For the ordered logit: $F(z) = exp(z)/(1 + exp(z))$. The cut-off example points for the graph below are $\alpha_0 = -\infty$, $\alpha_2 = -1.5$, $\alpha_3 = 2$, and $\alpha_4 = \infty$



16.1.1 Ordered Logit Example: Organic Food Purchase

The order logit model is illustrated with a survey on the purchase frequency of organic tomatoes and organic strawberries `fpdata`:

- Never (1), rarely (2), once per month (3), every 2 weeks (4), 1-2 times a week (5), almost daily (6)

The independent variables included in the model are

- Age and female
- Education: High school (1), some college (2), bachelor (3), master (4), technical school diploma (5), doctorate (6)

```
fpdata$strawberriesorg = as.factor(fpdata$strawberriesorg)
strawdata = fpdata[c("strawberriesorg","age","education","female","kidsunder12")]
strawdata = na.omit(strawdata)
bhat = polr(strawberriesorg~age+education+female+kidsunder12,data=strawdata,Hess=TRUE)
summary(bhat)
```

```
## Call:
## polr(formula = strawberriesorg ~ age + education + female + kidsunder12,
##     data = strawdata, Hess = TRUE)
##
## Coefficients:
##                Value Std. Error t value
## age         -0.02034   0.009838 -2.0676
## education    0.01596   0.112028  0.1425
## female      -0.41533   0.280485 -1.4808
## kidsunder12  0.28560   0.321778  0.8876
##
## Intercepts:
##     Value   Std. Error t value
## 0|1 -1.4958  0.6497    -2.3022
## 1|2 -0.4381  0.6434    -0.6810
## 2|3  0.2084  0.6394     0.3259
## 3|4  0.8352  0.6442     1.2964
## 4|5  1.6314  0.6699     2.4353
##
## Residual Deviance: 526.4547
## AIC: 544.4547
```

For the organic purchases data, the cut-off points are under “Intercepts” and thus, we have (rounded coefficients):

$$z = -0.020 \cdot age + 0.0160 \cdot education - 0.415 \cdot female + 0.286 \cdot kidsunder12$$

The cut-off points can be interpreted as follows:

$$\begin{aligned}
Pr(y = 1) &= P(z + \epsilon_i \leq -1.4958)\\
Pr(y = 2) &= P(-1.4958 < z + \epsilon_i \leq -0.4381)\\
Pr(y = 3) &= P(-0.4381 < z + \epsilon_i \leq 0.2084)\\
Pr(y = 4) &= P(0.2084 < z + \epsilon_i \leq 0.8352)\\
Pr(y = 4) &= P(0.8352 < z + \epsilon_i \leq 1.6314)\\
Pr(y = 6) &= P(1.6314 \leq z + \epsilon_i)
\end{aligned}$$

In order to get the p-values displayed in the output, you have to execute two additional steps:

```
bhat.coef = data.frame(coef(summary(bhat)))
bhat.coef$pval = round((pnorm(abs(bhat.coef$t.value),lower.tail=FALSE)*2),2)
bhat.coef
```

```
##                   Value  Std..Error    t.value pval
## age         -0.02034185 0.009838336 -2.0676110 0.04
## education    0.01595914 0.112027882  0.1424569 0.89
## female      -0.41533145 0.280485372 -1.4807598 0.14
## kidsunder12  0.28560319 0.321777883  0.8875787 0.37
## 0|1         -1.49575271 0.649708891 -2.3021891 0.02
## 1|2         -0.43813109 0.643390358 -0.6809724 0.50
## 2|3          0.20839939 0.639383205  0.3259382 0.74
## 3|4          0.83515493 0.644195302  1.2964313 0.19
## 4|5          1.63135404 0.669870994  2.4353257 0.01
```

16.1.2 Predicted Probability and Marginal Effects

The predicted probability for each observation can be obtained as well assuming that the output of the `polr` command is stored in *bhat*.

```
bhat.pred = predict(bhat,type="probs")
x = ocME(bhat)
x$out$ME.all
```

```
##             effect.0 effect.1 effect.2 effect.3 effect.4 effect.5
## age            0.005    0.000   -0.001   -0.001   -0.001   -0.001
## education     -0.004    0.000    0.001    0.001    0.001    0.001
## female         0.098   -0.003   -0.023   -0.024   -0.023   -0.026
## kidsunder12   -0.067    0.001    0.015    0.017    0.016    0.018
```

16.2 Multinomial Logit and Multinomial Probit Models

Revealed preferences:

- Observed choices of individuals

Stated preference

- Hypothetical choice situations

Economists' modeling of choice

- Utility/happiness/satisfaction associated with multiple choice situations

16.2.1 Theoretical Aspects

Travel choice model dependent on cost (x) and time (z):

$$V_j = \alpha_j + \beta_1 \cdot x_j + \beta_2 \cdot z_j$$

Probability of choosing alternative j (assuming three choices):

$$P(1) = \frac{e^{V_1}}{e^{V_1} + e^{V_2} + e^{V_3}}$$
$$P(2) = \frac{e^{V_2}}{e^{V_1} + e^{V_2} + e^{V_3}}$$
$$P(3) = \frac{e^{V_3}}{e^{V_1} + e^{V_2} + e^{V_3}}$$

Note that $P(1) + P(2) + P(3) = 1$

16.2.2 Data Managment

Long shape

- One row for each alternative

Wide shape

- One row for each choice situation

There are some very good resources on data management and the package in general:

- Estimation of Random Utility Models in R: The mlogit Package

Travel Mode (long format)

- Travel Mode Choice Data
- `mlogit.data(travelmode,choice="choice",shape="long",alt.levels=c("air","train","bus","car"))`

16.2.3 Fishing Data

The data is in wide format:

- Fishing modes: beach, pier, private, and charter
- Alternative-specific variables: price and catch
- Individual-specific variables: income
- Suitability of the "wide" format to store individual-specific variables
- The R parameter `varying` designates alternative specific variables

In a first step, the model is only estimated using the individual-specific variable *income*:

```
data("Fishing",package="mlogit")
fishing = mlogit.data(Fishing,shape="wide",varying=2:9,choice="mode")
bhat = mlogit(mode~0|income,fishing)
summary(bhat)
```

```
##
## Call:
## mlogit(formula = mode ~ 0 | income, data = fishing, method = "nr")
##
## Frequencies of alternatives:choice
##   beach    boat charter    pier
## 0.11337 0.35364 0.38240 0.15059
##
## nr method
## 4 iterations, 0h:0m:0s
## g'(-H)^-1g = 8.32E-07
## gradient close to zero
##
## Coefficients :
##                        Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept):boat     7.3892e-01  1.9673e-01  3.7560 0.0001727 ***
## (Intercept):charter  1.3413e+00  1.9452e-01  6.8955 5.367e-12 ***
## (Intercept):pier     8.1415e-01  2.2863e-01  3.5610 0.0003695 ***
## income:boat          9.1906e-05  4.0664e-05  2.2602 0.0238116 *
## income:charter      -3.1640e-05  4.1846e-05 -0.7561 0.4495908
## income:pier         -1.4340e-04  5.3288e-05 -2.6911 0.0071223 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1477.2
## McFadden R^2:  0.013736
## Likelihood ratio test : chisq = 41.145 (p.value = 6.0931e-09)
```

```
fishing.fitted      = fitted(bhat,outcome=FALSE)
effects(bhat,covariate ="income")
```

```
##         beach          boat       charter          pier
##  7.496226e-08  3.259851e-05 -1.201366e-05 -2.065981e-05
```

In a second step, alternative-specific variables are added:

```
bhat                = mlogit(mode~catch+price|income,data=fishing)
summary(bhat)
```

```
##
## Call:
## mlogit(formula = mode ~ catch + price | income, data = fishing,
##     method = "nr")
##
## Frequencies of alternatives:choice
##   beach    boat charter    pier
## 0.11337 0.35364 0.38240 0.15059
##
## nr method
## 7 iterations, 0h:0m:0s
## g'(-H)^-1g = 1.37E-05
## successive function values within tolerance limits
##
## Coefficients :
##                        Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept):boat     5.2728e-01  2.2279e-01   2.3667 0.0179485 *
## (Intercept):charter  1.6944e+00  2.2405e-01   7.5624 3.952e-14 ***
## (Intercept):pier     7.7796e-01  2.2049e-01   3.5283 0.0004183 ***
## catch                3.5778e-01  1.0977e-01   3.2593 0.0011170 **
## price               -2.5117e-02  1.7317e-03 -14.5042 < 2.2e-16 ***
## income:boat          8.9440e-05  5.0067e-05   1.7864 0.0740345 .
## income:charter      -3.3292e-05  5.0341e-05  -0.6613 0.5084031
## income:pier         -1.2758e-04  5.0640e-05  -2.5193 0.0117582 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1215.1
## McFadden R^2:  0.18868
## Likelihood ratio test : chisq = 565.17 (p.value = < 2.22e-16)
```

```
fishing.fitted      = fitted(bhat,outcome=FALSE)
effects(bhat,covariate="income")
```

```
##         beach          boat       charter          pier
## -7.214167e-07  3.176132e-05 -2.173392e-05 -9.305980e-06
```

```
rm(bhat,Fishing,fishing.fitted)
```

The mlogit package also allows for the estimation of a multinomial probit model.

```
bhat                = mlogit(mode~catch+price|income,data=fishing,probit=FALSE)
summary(bhat)
```

```
##
## Call:
```

```
## mlogit(formula = mode ~ catch + price | income, data = fishing,
##     probit = FALSE, method = "nr")
##
## Frequencies of alternatives:choice
##   beach    boat charter    pier
## 0.11337 0.35364 0.38240 0.15059
##
## nr method
## 7 iterations, 0h:0m:0s
## g'(-H)^-1g = 1.37E-05
## successive function values within tolerance limits
##
## Coefficients :
##                         Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept):boat      5.2728e-01  2.2279e-01   2.3667 0.0179485 *
## (Intercept):charter   1.6944e+00  2.2405e-01   7.5624 3.952e-14 ***
## (Intercept):pier      7.7796e-01  2.2049e-01   3.5283 0.0004183 ***
## catch                 3.5778e-01  1.0977e-01   3.2593 0.0011170 **
## price                -2.5117e-02  1.7317e-03 -14.5042 < 2.2e-16 ***
## income:boat           8.9440e-05  5.0067e-05   1.7864 0.0740345 .
## income:charter       -3.3292e-05  5.0341e-05  -0.6613 0.5084031
## income:pier          -1.2758e-04  5.0640e-05  -2.5193 0.0117582 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1215.1
## McFadden R^2:  0.18868
## Likelihood ratio test : chisq = 565.17 (p.value = < 2.22e-16)
```

```
fishing.fitted      = fitted(bhat,outcome=FALSE)
effects(bhat,covariate="income")
```

```
##         beach          boat       charter          pier
## -7.214167e-07  3.176132e-05 -2.173392e-05 -9.305980e-06
```

```
rm(bhat,Fishing,fishing,fishing.fitted)
```

```
## Warning in rm(bhat, Fishing, fishing, fishing.fitted): object 'Fishing' not found
```

16.2.4 Travel Data

```
data("TravelMode",package="AER")
travelmode = mlogit.data(TravelMode,choice="choice",shape="long",alt.var="mode")
bhat = mlogit(choice~gcost+wait|income+size,data=travelmode,reflevel="car")
summary(bhat)
```

```
##
## Call:
## mlogit(formula = choice ~ gcost + wait | income + size, data = travelmode,
##     reflevel = "car", method = "nr")
##
## Frequencies of alternatives:choice
##     car     air   train     bus
## 0.28095 0.27619 0.30000 0.14286
##
## nr method
## 5 iterations, 0h:0m:0s
## g'(-H)^-1g = 1.66E-07
## gradient close to zero
##
## Coefficients :
##                     Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):air    7.8736084  0.9868475  7.9785 1.554e-15 ***
## (Intercept):train  5.5592051  0.6991387  7.9515 1.776e-15 ***
## (Intercept):bus    4.4331916  0.7783339  5.6957 1.228e-08 ***
## gcost             -0.0196850  0.0054015 -3.6444 0.0002680 ***
## wait              -0.1015659  0.0112306 -9.0436 < 2.2e-16 ***
## income:air         0.0040710  0.0127247  0.3199 0.7490196
## income:train      -0.0551849  0.0144824 -3.8105 0.0001387 ***
## income:bus        -0.0233237  0.0162973 -1.4311 0.1523914
## size:air          -1.0274229  0.2656569 -3.8675 0.0001100 ***
## size:train         0.3023954  0.2256155  1.3403 0.1801437
## size:bus          -0.0300096  0.3339774 -0.0899 0.9284023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -177.45
## McFadden R^2:  0.37463
## Likelihood ratio test : chisq = 212.61 (p.value = < 2.22e-16)
```

```
tavel.fitted = fitted(bhat,outcome=FALSE)
```

16.2.5 Electric Vehicle Data

```
evdata = mlogit.data(evdata,shape="wide",choice="choice")
bhat = mlogit(choice~0|age+female+level2+numcars+edu+income+politics,data=evdata)
summary(bhat)
```

```
##
## Call:
## mlogit(formula = choice ~ 0 | age + female + level2 + numcars +
##     edu + income + politics, data = evdata, method = "nr")
##
## Frequencies of alternatives:choice
##        1        2        3        4
## 0.419355 0.265233 0.225806 0.089606
##
## nr method
## 5 iterations, 0h:0m:0s
## g'(-H)^-1g = 0.000603
## successive function values within tolerance limits
##
## Coefficients :
##                  Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept):2  0.11810109  0.60622429  0.1948 0.8455384
## (Intercept):3  0.90653782  0.63856959  1.4196 0.1557130
## (Intercept):4  0.34528552  0.84542368  0.4084 0.6829675
## age:2         -0.01298148  0.00731409 -1.7749 0.0759210 .
## age:3         -0.04216613  0.00867889 -4.8585 1.183e-06 ***
## age:4         -0.02314160  0.01139425 -2.0310 0.0422561 *
## female:2       0.17935742  0.21591603  0.8307 0.4061537
## female:3      -0.02781447  0.23732560 -0.1172 0.9067019
## female:4       0.00019196  0.32364088  0.0006 0.9995267
## level2:2       0.13808856  0.25338230  0.5450 0.5857665
## level2:3       0.87829370  0.25462065  3.4494 0.0005618 ***
## level2:4       1.14103750  0.33905869  3.3653 0.0007646 ***
## numcars:2      0.08340570  0.10845229  0.7691 0.4418611
## numcars:3      0.02412224  0.11385000  0.2119 0.8322027
## numcars:4     -0.35470532  0.15922101 -2.2278 0.0258969 *
## edu:2          0.09033371  0.08476675  1.0657 0.2865711
## edu:3          0.21765011  0.09487551  2.2941 0.0217871 *
## edu:4         -0.00475892  0.12980749 -0.0367 0.9707550
## income:2      -0.01070840  0.07000961 -0.1530 0.8784328
## income:3      -0.04329121  0.07768626 -0.5573 0.5773519
## income:4      -0.01112286  0.10371253 -0.1072 0.9145930
## politics:2    -0.15253774  0.05211226 -2.9271 0.0034214 **
## politics:3    -0.19656259  0.05687906 -3.4558 0.0005487 ***
## politics:4    -0.07567912  0.07405617 -1.0219 0.3068211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -663.51
## McFadden R^2:  0.062686
## Likelihood ratio test : chisq = 88.75 (p.value = 2.6494e-10)
```

```r
evdata.fitted = fitted(bhat,outcome=FALSE)
rm(bhat,evdata,evdata.fitted)
```

16.3 Exercises

1. ***Happiness*** (***): Using the data set `happy`, generate an ordered logit regression model that regresses the dependent variable *happiness* on those variables that have the strongest potential causal relationship. For your model, interpret the R output and indicate why each independent variable that is included in the model would contribute to higher or lower happiness. Speak to the possibility of multicollinearity in the independent variables.

2. ***AFV Choice*** (***): The data `evdata` contains data about the choice of consumers with respect to alternative fuel vehicles. For each consumer, you have the following variables: *age*, *suv* (whether they are interested in buying a SUV), *level*2 (indicating whether people have a fast charger for electric cars in their community), *own*… (indicating whether the respondent currently has a gas, hybrid, plug- in hybrid, or battery electric vehicle), *gender* (1=female) and *numcars* (number of cars). Estimate a multinomial logit model that estimates the probability of a consumer to purchase a gasoline, hybrid, plug-in hybrid, or battery electric vehicle. Calculate the marginal probabilities as well.

3. ***NHTS*** (***): Consider the data set `hhpub` from the 2017 National Household Travel Survey (NHTS). The data contains information about household characteristics and some of their travel means. For this question, you will focus on the following variables: $BIKE$, $HBPPOPDN$, $HHFAMINC$, $HHVEHCNT$, $HOMEOWN$, and $URBRUR$. You must read the codebook for this question and learn how the variables are coded. Go to the codebook and pick "Household" as the dataset (drop down menu). Before conducting the analysis, delete all entries that are not complete (i.e., all the negative values). Once you have final data set, estimate an ordered logit model with $BIKE$ as the dependent variable and the other variables as the independent variables. What do you conclude from the model?

4. ***Home Heating*** (***): Consider the dataset `Heating` from the You can load the data set into R by typing: `data("Heating",package="mlogit")`. The data contains the choice of heating systems in California homes.

Estimate a multinomial logit model with installation and operating cost as the alternative-specific variables and income, age, and number of rooms as the individual-specific variables.

17 Limited Dependent Variable Models

There are slides associated with this chapter:

- Limited Dependent Variable Models - Slides

This chapter covers three regression models in which the dependent variable is somehow limited:

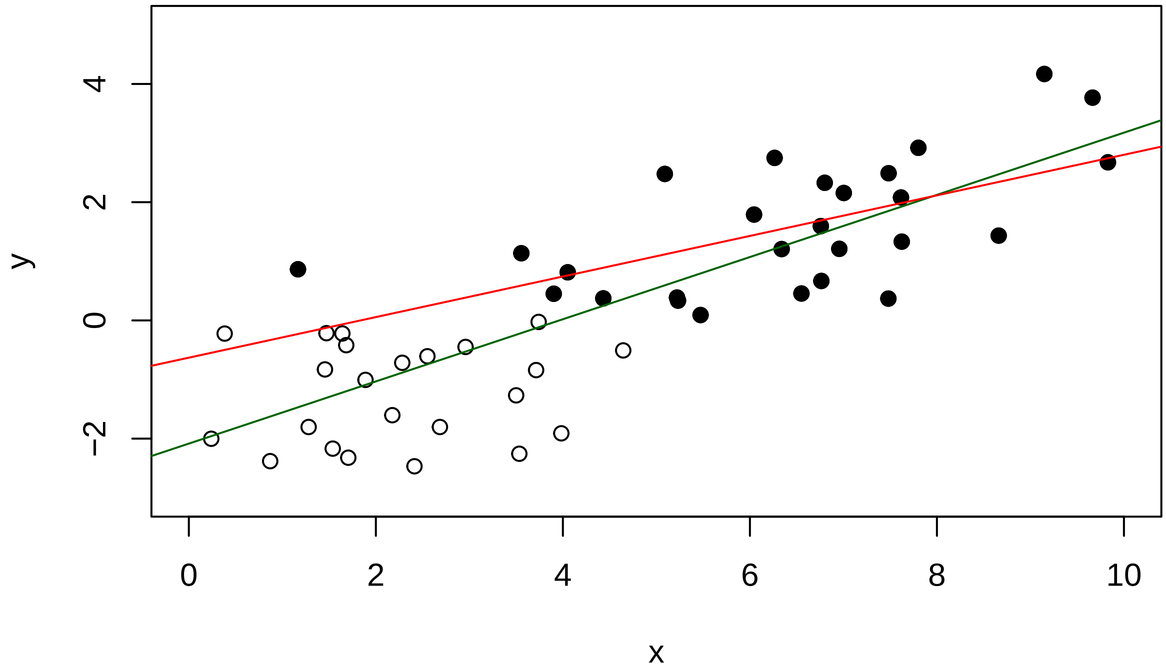
1. Truncation: With truncated data, the researcher does not observe values past a particular point and those values are also not reported. Examples of truncation are low-income household studies, on-site visitation data, or time-of-use pricing experiments (excludes low-usage households).
2. Censoring: In the case of censoring, values that are above or below a certain value are replaced by that value. For example, the demand for a particular class is not fully observed (absence of a waiting list). This is also called a Tobit model.
3. Count dependent variable

The following packages are necessary for this section: AER truncreg, censReg, and pscl.

Truncation and censoring lead to a bias in the estimates. It is not always clear why or if the data is limited in its range.

17.1 Truncation

In the case of truncation, a certain part of the data is not observed. In the graph below, the true parameters are $\beta_0 = -2$ and $\beta_1 = 0.5$. Values $y < 0$ are not reported in the data. The green regression line is "correct" whereas the "red" is the line obtained from a regression model which ignores the truncation.



If all the data was observed, the correct regression model would give the following results:

```
summary(bhatreal)
```

```
##
## Call:
## lm(formula = yreal ~ x, data = truncation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03049 -0.74295  0.08173  0.74879  2.33519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.0842     0.2839  -7.341 2.21e-09 ***
## x             0.5260     0.0545   9.652 7.98e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 48 degrees of freedom
## Multiple R-squared:  0.6599, Adjusted R-squared:  0.6529
## F-statistic: 93.15 on 1 and 48 DF,  p-value: 7.984e-13
```

The estimates are biased if truncation is ignored:

```
summary(bhattruncated)
```

```
##
## Call:
```

```
## lm(formula = yobs ~ x, data = truncation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56809 -0.71414 -0.06843  0.58821  1.65911
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.63071    0.59825  -1.054 0.301851
## x            0.34327    0.08946   3.837 0.000752 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8949 on 25 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.3706, Adjusted R-squared:  0.3455
## F-statistic: 14.72 on 1 and 25 DF,  p-value: 0.000752
```

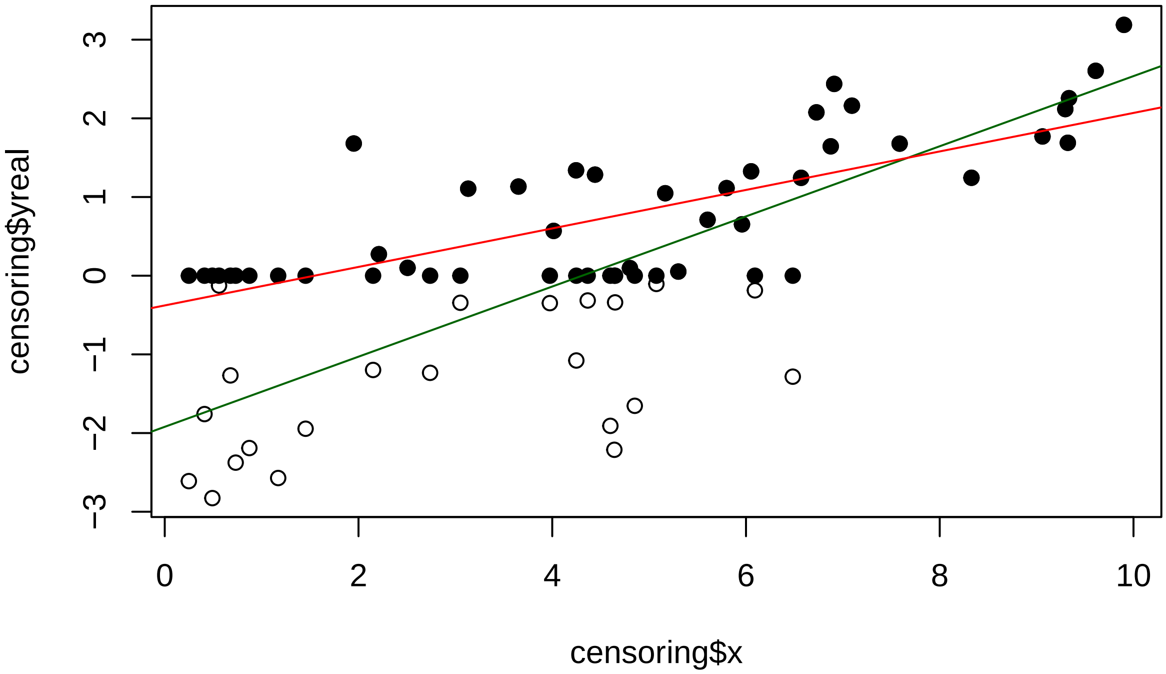
To correct for the truncation, use the functions from the package truncreg which allows to reduce the bias of the coefficients:

```
bhatcorrect = truncreg(yobs~x,data=truncation)
summary(bhatcorrect)
```

```
##
## Call:
## truncreg(formula = yobs ~ x, data = truncation)
##
## BFGS maximization method
## 34 iterations, 0h:0m:0s
## g'(-H)^-1g = 8.53E-11
##
##
##
## Coefficients :
##             Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) -2.47203    1.35511 -1.8242  0.068117 .
## x            0.56968    0.17531  3.2496  0.001156 **
## sigma        1.03294    0.21090  4.8978 9.691e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -29.992 on 3 Df
```

17.2 Censoring

In the case of censoring, the values of the dependent variable are reported at a certain point if they are above or below a certain value.



If all data was reported at the correct value, the following following regression model could be executed:

```
summary(bhat_real)
```

```
##
## Call:
## lm(formula = yreal ~ x, data = censoring)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36087 -0.54654 -0.05193  0.68759  2.73136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.92076    0.29763  -6.453 5.07e-08 ***
## x            0.44580    0.05482   8.132 1.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.054 on 48 degrees of freedom
## Multiple R-squared:  0.5794, Adjusted R-squared:  0.5706
## F-statistic: 66.12 on 1 and 48 DF,  p-value: 1.394e-10
```

Ignoring censoring leads to biased results:

```
summary(bhat_censored)
```

```
##
```

```
## Call:
## lm(formula = y ~ x, data = censoring)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20830 -0.42404  0.08197  0.30730  1.58097
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.37803    0.17476  -2.163   0.0355 *
## x            0.24470    0.03219   7.602 8.85e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6186 on 48 degrees of freedom
## Multiple R-squared:  0.5463, Adjusted R-squared:  0.5368
## F-statistic: 57.79 on 1 and 48 DF,  p-value: 8.846e-10
```

Using the R package censReg) allows for the reduction of the bias:

```
b_correct = censReg(y~x,data=censoring)
summary(b_correct)
```

```
##
## Call:
## censReg(formula = y ~ x, data = censoring)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##             50             22             28              0
##
## Coefficients:
##             Estimate Std. error t value  Pr(> t)
## (Intercept) -1.44613    0.36608  -3.950 7.80e-05 ***
## x            0.38638    0.05924   6.522 6.92e-11 ***
## logSigma    -0.14172    0.14238  -0.995     0.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 6 iterations
## Return code 8: successive function values within relative tolerance limit (reltol)
## Log-likelihood: -48.15507 on 3 Df
```

17.3 Count Regression Models

Count regression models apply to cases in which the dependent variable represents discrete, integer count data. Here are some examples of count outcomes:

- What are the number of arrests for a person?
- What determines the number of credit cards a person owns?

This section on count regression presents three models:

1. Poisson Regression Model: The condition to use this model is the absence of overdispersion, i.e., the expected value of the dependent variable is equal to the variance.
2. Quasi-Poisson Regression Model: Overdispersion occurs if the variance of the dependent variable is larger than its mean. In this case, the Poisson Regression Model leads to unreliable hypothesis tests

regarding the coefficients. The Quasi-Poisson Model solves this issue.

3. Negative Binomial Regression Model: The second possibility to deal with overdispersion is to use a Negative Binomial Regression Model.

The main package used is pscl. There is also an additional resource with more theoretical details on the topic: Regression Models for Count Data in R. A more up-to-date version of the document may be found with the pscl package documentation.

17.3.1 Poisson Regression Model

Recall that the Poisson distribution is written as:

$$Pr(Y = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

The important characteristics of the distribution is that the mean and variance are equal to λ, i.e., $E(Y) = \lambda$ and $Var(Y) = \lambda$, this is also known as the equidispersion property. The mean parameter is written as $\lambda = exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)$.

The first examples uses the 2017 NHTS data contained in `hhpub`. In a first step, the data is prepared for the regression model, i.e., missing or unknown values are eliminated and income is measured in (thousand) dollar terms:

```
hhpubdata = subset(hhpub,hhfaminc %in% c(1:11) & homeown %in% c(1,2) & urbrur %in% c(1,2) & hhvehcnt %ir
hhfaminc  = c(1:11)
income    = c(10,12.5,20,30,42.5,57.5,82.5,112.5,137.5,175,200)
income    = data.frame(hhfaminc,income)
hhpubdata = merge(hhpubdata,income)
hhpubdata$rural = hhpubdata$urbrur-1
hhpubdata$rent = hhpubdata$homeown-1
```

The outcome of interest is the number of vehicles based on household income, home ownership, and urban/rural household location. Before executing the Poisson regression model, calculate the mean and variance of the outcome variable.

```
## [1] 1.981142
```

```
## [1] 1.386027
```

The mean and the variance are similar and thus, estimating a Poisson Regression Model is an appropriate first step. The package AER also contains the function `dispersiontest()` which conducts a hypothesis test assuming no overdispersion. This test will be used after execution of the main regression. To estimate the model, the `glm()` function can be used by specifying `family=poisson`.

```
bhat_pois = glm(hhvehcnt~income+rent+rural,data=hhpubdata,family=poisson)
summary(bhat_pois)
```

```
##
## Call:
## glm(formula = hhvehcnt ~ income + rent + rural, family = poisson,
##     data = hhpubdata)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.654e-01  4.292e-03  108.43   <2e-16 ***
## income       2.986e-03  3.601e-05   82.93   <2e-16 ***
## rent        -3.733e-01  5.797e-03  -64.39   <2e-16 ***
## rural        2.224e-01  4.616e-03   48.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 86505  on 124400  degrees of freedom
## Residual deviance: 68533  on 124397  degrees of freedom
## AIC: 370161
##
## Number of Fisher Scoring iterations: 5
```

All coefficients are statistically significant. The signs associated with the coefficients indicates the direction of influence on the outcome variable, i.e., the number of cars. As expected, higher income is associated with a higher number of cars and so is living in a rural environment. Renting is associated with a lower number of vehicles. Note that income and renting is possibly correlated. In general, the coefficient estimates are interpreted using $\exp(\beta)$. That is, with every unit increase in X, the predictor variable has a multiplicative effect of $\exp(\beta)$ on the mean of Y, i.e., λ:

- If $\beta = 0$, then $\exp(\beta) = 1$, and the expected count $\mu = E(y) = \exp(\alpha)$, and Y and X are not related.
- If $\beta > 0$, then $\exp(\beta) > 1$, and the expected count $E(y)$ is $\exp(\beta)$ times larger than when $X = 0$
- If $\beta < 0$, then $\exp(\beta) < 1$, and the expected count $E(y)$ is $\exp(\beta)$ times smaller than when $X = 0$

The function `overdispersion` tests the null hypothesis of equidispersion:

```
dispersiontest(bhat_pois)
```

```
##
##  Overdispersion test
##
## data:  bhat_pois
## z = -115.75, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.5670593
```

Given the p-value, the null hypothesis cannot be rejected. If the data suggests overdispersion, two alternative regression models can be used: (1) Quasi-Poisson and (2) Negative Binomial.

17.3.2 Quasi-Poisson Regression Model

The dataset `blm` used in this section as well as the one describing the negative binomial model is associated with the article Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity. Note that the paper includes a significant number of supplementary materials which allows for the replication of the results and much more.

The dependent variable is the total number of protests in a city. In a first step, the mean and variance of the variable *totalprotests* are calclated:

```
mean(blm$totprotests)
```

```
## [1] 0.4959529
```

```
var(blm$totprotests)
```

```
## [1] 6.35326
```

The variance is significantly higher than the mean which suggests overdispersion. In a first step, a regular Poisson model is estimated.

```r
eq1 = "totprotests~log(pop)+log(popdensity)+percentblack+blackpovertyrate+I(blackpovertyrate^2)+percentb
bhat1 = glm(eq1,data=blm,family=poisson)
summary(bhat1)
```

```
##
## Call:
## glm(formula = eq1, family = poisson, data = blm)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.001e+01  6.327e-01 -31.625  < 2e-16 ***
## log(pop)               1.129e+00  4.007e-02  28.170  < 2e-16 ***
## log(popdensity)       -1.831e-01  8.654e-02  -2.116   0.0343 *
## percentblack           1.697e-02  3.104e-03   5.467 4.59e-08 ***
## blackpovertyrate       1.461e-01  2.636e-02   5.541 3.02e-08 ***
## I(blackpovertyrate^2) -1.552e-03  3.985e-04  -3.895 9.82e-05 ***
## percentbachelor        3.893e-02  3.918e-03   9.935  < 2e-16 ***
## collegeenrollpc        9.305e-03  2.377e-03   3.914 9.06e-05 ***
## demshare               4.301e-02  5.293e-03   8.126 4.43e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3204.6  on 1225  degrees of freedom
## Residual deviance:  787.4  on 1217  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: 1242.9
##
## Number of Fisher Scoring iterations: 6
```

Note that the signs and statistical significance correspond to Model 1 (Table 3) in the original paper. The coefficients are different because the paper only include negative binomial regression models. The reason for not using a Poisson regression model is the presence of overdispersion:

```r
dispersiontest(bhat1)
```

```
##
##  Overdispersion test
##
## data:  bhat1
## z = 1.4052, p-value = 0.07998
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.212733
```

Note that the null hypothesis is rejected at the 10% but not the 5% significance level. The Quasi-Poisson Regression Model handles overdispersion by adjusting the stand-errors but leaving the coefficicent estimates the same.

```r
bhat2 = glm(eq1,data=blm,family=quasipoisson)
summary(bhat2)
```

```
##
## Call:
```

```
## glm(formula = eq1, family = quasipoisson, data = blm)
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.001e+01  9.841e-01 -20.332  < 2e-16 ***
## log(pop)               1.129e+00  6.232e-02  18.111  < 2e-16 ***
## log(popdensity)       -1.831e-01  1.346e-01  -1.360 0.173942
## percentblack           1.697e-02  4.828e-03   3.515 0.000457 ***
## blackpovertyrate       1.461e-01  4.100e-02   3.562 0.000382 ***
## I(blackpovertyrate^2) -1.552e-03  6.198e-04  -2.504 0.012403 *
## percentbachelor        3.893e-02  6.094e-03   6.387 2.40e-10 ***
## collegeenrollpc        9.305e-03  3.697e-03   2.517 0.011975 *
## demshare               4.301e-02  8.233e-03   5.225 2.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.419275)
##
##     Null deviance: 3204.6  on 1225  degrees of freedom
## Residual deviance:  787.4  on 1217  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Note that in this case, the population density is not statistically significant anymore.

17.3.3 Negative Binomial Regression Model

The Negative Binomial Regression Model can be used in the presence of count data and overdispersion. Below, the results from the article Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity are recreated using the negative binomial models presented in the paper. The required package is MASS

The first model is a basic model of resource mobilization and opportunity structure.

```r
bhat3 = glm.nb(eq1,data=blm,link=log)
summary(bhat3)
```

```
##
## Call:
## glm.nb(formula = eq1, data = blm, link = log, init.theta = 1.559078735)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.090e+01  1.117e+00 -18.719  < 2e-16 ***
## log(pop)               1.292e+00  7.159e-02  18.047  < 2e-16 ***
## log(popdensity)       -3.130e-01  1.328e-01  -2.356  0.01848 *
## percentblack           2.249e-02  4.709e-03   4.777 1.78e-06 ***
## blackpovertyrate       1.319e-01  3.121e-02   4.227 2.37e-05 ***
## I(blackpovertyrate^2) -1.340e-03  4.768e-04  -2.810  0.00496 **
## percentbachelor        4.462e-02  5.465e-03   8.163 3.26e-16 ***
## collegeenrollpc        1.051e-02  4.118e-03   2.553  0.01068 *
## demshare               4.077e-02  7.292e-03   5.591 2.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(1.5591) family taken to be 1)
##
##     Null deviance: 1899.84  on 1225  degrees of freedom
## Residual deviance:  501.15  on 1217  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: 1120.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.559
##           Std. Err.:  0.351
##
##  2 x log-likelihood:  -1100.187
```

In a second model, black deaths are added:

```
bhat4 = glm.nb(totprotests~log(pop)+log(popdensity)+percentblack+blackpovertyrate+I(blackpovertyrate^2)
summary(bhat4)
```

```
##
## Call:
## glm.nb(formula = totprotests ~ log(pop) + log(popdensity) + percentblack +
##     blackpovertyrate + I(blackpovertyrate^2) + percentbachelor +
##     collegeenrollpc + demshare + deathsblackpc, data = blm, link = log,
##     init.theta = 1.685551835)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.073e+01  1.101e+00 -18.824  < 2e-16 ***
## log(pop)               1.281e+00  7.046e-02  18.178  < 2e-16 ***
## log(popdensity)       -3.054e-01  1.315e-01  -2.323 0.020201 *
## percentblack           1.801e-02  4.864e-03   3.704 0.000212 ***
## blackpovertyrate       1.283e-01  3.109e-02   4.127 3.67e-05 ***
## I(blackpovertyrate^2) -1.301e-03  4.743e-04  -2.744 0.006071 **
## percentbachelor        4.372e-02  5.381e-03   8.125 4.47e-16 ***
## collegeenrollpc        1.005e-02  4.073e-03   2.466 0.013657 *
## demshare               4.069e-02  7.249e-03   5.613 1.98e-08 ***
## deathsblackpc          2.825e+00  9.312e-01   3.034 0.002414 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6856) family taken to be 1)
##
##     Null deviance: 1943.21  on 1225  degrees of freedom
## Residual deviance:  500.18  on 1216  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: 1113.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.686
##           Std. Err.:  0.404
```

```
##
##  2 x log-likelihood:  -1091.353
```

And in the third model, the authors use all police-caused deaths instead (victims of any race):

```
bhat5 = glm.nb(totprotests~log(pop)+log(popdensity)+percentblack+blackpovertyrate+I(blackpovertyrate^2)
summary(bhat5)
```

```
##
## Call:
## glm.nb(formula = totprotests ~ log(pop) + log(popdensity) + percentblack +
##     blackpovertyrate + I(blackpovertyrate^2) + percentbachelor +
##     collegeenrollpc + demshare + deathspc, data = blm, link = log,
##     init.theta = 1.621799986)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.080e+01  1.108e+00 -18.773  < 2e-16 ***
## log(pop)               1.277e+00  7.122e-02  17.928  < 2e-16 ***
## log(popdensity)       -3.121e-01  1.321e-01  -2.363  0.01812 *
## percentblack           2.163e-02  4.705e-03   4.596 4.31e-06 ***
## blackpovertyrate       1.295e-01  3.110e-02   4.164 3.12e-05 ***
## I(blackpovertyrate^2) -1.315e-03  4.744e-04  -2.772  0.00558 **
## percentbachelor        4.507e-02  5.472e-03   8.237  < 2e-16 ***
## collegeenrollpc        1.040e-02  4.094e-03   2.540  0.01108 *
## demshare               4.121e-02  7.257e-03   5.678 1.36e-08 ***
## deathspc               9.564e-01  6.330e-01   1.511  0.13085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6218) family taken to be 1)
##
##     Null deviance: 1921.82  on 1225  degrees of freedom
## Residual deviance:  502.82  on 1216  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: 1119.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.622
##           Std. Err.:  0.374
##
##  2 x log-likelihood:  -1097.839
```

17.4 Hurdle and Zero-Inflation Models

Count data often includes many observations at 0 which can lead to problems using a Poisson or a Negative-Binomial Regression Model. The application of both models is first illustrated with the NMES1988 data from the package AER and then with the BLM protest data.

The data NMES1988 contains 4406 observations of people on Medicare who are 66 years or older. The outcome of interest is the number of doctor *visits* as a function of *hospital* (number of hospital visits), *health* (self-indicated health status), *chronic* (number of chronic conditions), *gender*, *school*, and *insurance*.

```
data("NMES1988",package="AER")
eq = visits~hospital+health+chronic+gender+school+insurance
bhat_pois = glm(eq,data=NMES1988,family=poisson)
summary(bhat_pois)
```

```
##
## Call:
## glm(formula = eq, family = poisson, data = NMES1988)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.028874   0.023785  43.258   <2e-16 ***
## hospital          0.164797   0.005997  27.478   <2e-16 ***
## healthpoor        0.248307   0.017845  13.915   <2e-16 ***
## healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
## chronic           0.146639   0.004580  32.020   <2e-16 ***
## gendermale       -0.112320   0.012945  -8.677   <2e-16 ***
## school            0.026143   0.001843  14.182   <2e-16 ***
## insuranceyes      0.201687   0.016860  11.963   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 23168  on 4398  degrees of freedom
## AIC: 35959
##
## Number of Fisher Scoring iterations: 5
```

```
bhat_nb = glm(eq,data=NMES1988)
summary(bhat_nb)
```

```
##
## Call:
## glm(formula = eq, data = NMES1988)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.63203    0.33480   4.875 1.13e-06 ***
## hospital          1.61976    0.13264  12.211  < 2e-16 ***
## healthpoor        1.84532    0.31234   5.908 3.72e-09 ***
## healthexcellent -1.33140    0.36257  -3.672 0.000243 ***
## chronic           0.94440    0.07693  12.276  < 2e-16 ***
## gendermale       -0.63185    0.19454  -3.248 0.001171 **
## school            0.14345    0.02726   5.262 1.49e-07 ***
## insuranceyes      1.10397    0.24362   4.532 6.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 40.02228)
##
##     Null deviance: 201252  on 4405  degrees of freedom
## Residual deviance: 176018  on 4398  degrees of freedom
## AIC: 28769
##
## Number of Fisher Scoring iterations: 2
```

```
bhat_hurdle = hurdle(eq,data=NMES1988,dist="negbin")
summary(bhat_hurdle)
```

```
##
## Call:
## hurdle(formula = eq, data = NMES1988, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.1718 -0.7080 -0.2737  0.3196 18.0092
##
## Count model coefficients (truncated negbin with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.197699   0.058973  20.309  < 2e-16 ***
## hospital         0.211898   0.021396   9.904  < 2e-16 ***
## healthpoor       0.315958   0.048056   6.575 4.87e-11 ***
## healthexcellent -0.331861   0.066093  -5.021 5.14e-07 ***
## chronic          0.126421   0.012452  10.152  < 2e-16 ***
## gendermale      -0.068317   0.032416  -2.108   0.0351 *
## school           0.020693   0.004535   4.563 5.04e-06 ***
## insuranceyes     0.100172   0.042619   2.350   0.0188 *
## Log(theta)       0.333255   0.042754   7.795 6.46e-15 ***
## Zero hurdle model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.043147   0.139852   0.309 0.757688
## hospital         0.312449   0.091437   3.417 0.000633 ***
## healthpoor      -0.008716   0.161024  -0.054 0.956833
## healthexcellent -0.289570   0.142682  -2.029 0.042409 *
## chronic          0.535213   0.045378  11.794  < 2e-16 ***
## gendermale      -0.415658   0.087608  -4.745 2.09e-06 ***
## school           0.058541   0.011989   4.883 1.05e-06 ***
## insuranceyes     0.747120   0.100880   7.406 1.30e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta: count = 1.3955
## Number of iterations in BFGS optimization: 16
## Log-likelihood: -1.209e+04 on 17 Df
```

17.5 Survival Analysis

17.6 Exercises

1. ***Aptitude Tobit Model*** (***): Consider the censored data set in `aptitude` in which the aptitude score is limited at 800. In a first step, estimate a regular OLS model with *apt* as the dependent variable and *read*, *math*, and *factor*(*program*) as the independent variables. In a second step, estimate a model that takes the censored nature of the data into account. Is there a significant difference in estimates?

2. ***Chicago Grocery Stores*** (***): Subdivision of Chicago are called Chicago Community Areas (CCA). The data in `chicagogrocery` includes data about the number of grocery stores (*stores*) in each CCA as well as demographic information. Estimate a Poisson and Negative Binomial Regression Model with *stores* as the dependent variable and the following independent variables: *income*, *pop*, unemployment

rate ($unemployed/laborforce$) and percentage of blacks ($black/pop$). What do you conclude? Are the results what you would expect?

3. ***Extramarital Affairs*** (***): Consider the data set in `fair`. The independent variables, which we are going to use are $male$, $yearsmarried$ (number of years married), $children$, $religious$ (religiousness on a scale of 1-5 with 1 being basically an atheist), and $marriagehappiness$ (self-rating of marriage with 1=very unhappy to 5=very happy). You are going to execute five models: (1) regular OLS, (2) Probit, (3) Poisson, (4) Negative Binomial, and (5) Hurdle Model. For the Probit model, you are running a model with a binary variable of either 0 (no affair) or 1 (at least one affair). Compare the models in terms of statistical significance. What changes from one model to the next? What model is the most appropriate and why?
4. ***Biochemistry Articles*** (***): Publish or perish summarizes life in academia. The dependent variable of interest is $articles$ and the independent variables are $female$, $married$, $kidsbelow6$, and $mentorarticles$ (number of articles by Ph.D. mentor). Estimate a quasi-poisson and a hurdle model. According to the model, what matters in terms of graduate student productivity? Why do those findings matter?
5. ***BLM II*** (***): In a previous exercise, a regular OLS regression model was used to explain the positive number of protests. In this exercise, a zero-inflated and a hurdle model are estimated. The dependent variable for this exercise is protest frequency (*totprotests*) and the independent variables are city population (*pop*), population density (*popdensity*), percent Black (*percentblack*), black poverty rate (*blackpovertyrate*), percent of population with at least a bachelor (*percentbachelor*), college enrollment (*collegeenrollpc*), share of democrats (*demshare*), and Black police-caused deaths per 10,000 people (*deathsblackpc*). Interpret the output of the two models.
6. ***Lung Cancer*** (**): Consider the data in `lung`. Plot the survival curve differentiating by sex. Estimate a survival model with age and $female$ as the independent variables.
7. ***Henning*** (***): Plot the survival curve differentiating by *personal*. Estimate three survival models with the following independent variables: (1) personal, (2) personal and property, and (3) personal, property, and cage. Interpret the output. Is there a big difference in coefficients?

18 Panel Data

This chapter introduces panel data which are observations of the same unit of analysis over time, e.g., people, cities, or countries. So far, only cross-sectional data was analyzed. This chapter on panel data starts to incorporate time aspects into the estimation procedure. There are also a YouTube Video and slides associated with this chapter:

- Panel Data - Video
- Panel Data - Slides

The required packages are plm and lmtest. The plm also contains some excellent documentation on the theoretical aspects of panel data as well as on the use of the package.

18.1 Overview

Two types of data must be distinguished:

- Pooled data: Combination of multiple cross-sectional data over time
  - Two or more different observational units over time
  - Grades in an economics class based on students' concentration combined from multiple semesters
  - American Community Survey (ACS)
- Panel data: Repeated measurement on the same individual i over time t.
  - Individual units can be people, states, firms, counties, countries, etc.
  - National Longitudinal Survey (NLSY79): To access the data: Accessing Data > Investigator > Begin searching as guest.

  - Necessary adjustments of standard error due to correlation across time.

There are some necessary assumptions about linear panel models.

- Regular time intervals
- Errors are correlated
- Parameters may vary across individuals or time
- Intercept: Individual specific effects model (fixed or random)

Note that the General Social Survey (GSS) is not a panel data set because different respondents are questioned every year. Besides the National Longitudinal Survey mentioned above, here are some additional examples of panel data sets:

- Panel Study of Income Dynamics (PSID): Data on approximately 5,000 families on various socioeconomic and demographic variables
- Survey of Income and Program Participation (SIPP): Interviews about economic condition of respondents

Panel models have the advantage that they take into account heterogeneity among observational units, e.g., firms, states, counties. Those models also contribute to the better understanding on the dynamics of change for observational units over time. Panel data combines cross-sectional data with time series data leading to more complete behavioral models.

- Balanced versus unbalance panel: A balanced panel has the same number of time-series observations for each subject or observational unit, whereas an unbalanced panel does not.
- Short versus long panel: A short panel has a larger number of subjects or observational units than there are time periods. A long panel has a greater number of time periods than observational units.

There are three types of regression models presented in this chapter:

- Pooled Ordinary Least Square model
- Fixed Effects Panel Data Model
- Random Effects Panel Data Model

18.2 Pooled Ordinary Least Square model

The first example is data from the General Social Survey (GSS). Recall that the GSS is not a panel data set since the respondents change from one year to the next. The years analyzed for this example are the even years between 1974 and 1984. Note that this data set is accompanying the book *Introductory Econometrics: A Modern Approach* by Jeffrey Wooldridge.

```
##
## Call:
## lm(formula = kids ~ educ + age + I(age^2) + east + northcentral +
##     west + farm + otherrural + town + smallcity + y74 + y76 +
##     y78 + y80 + y82 + y84, data = fertil1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1437 -1.0481 -0.1082  0.9450  5.1055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.785069   3.098715  -2.190 0.028758 *
## educ         -0.129765   0.018653  -6.957 5.94e-12 ***
## age           0.499186   0.140592   3.551 0.000400 ***
## I(age^2)     -0.005436   0.001589  -3.421 0.000648 ***
## east          0.060729   0.132538   0.458 0.646896
## northcentral  0.219568   0.120638   1.820 0.069019 .
## west          0.050807   0.167982   0.302 0.762360
## farm         -0.109598   0.149353  -0.734 0.463217
## otherrural   -0.199208   0.178270  -1.117 0.264043
## town          0.058579   0.126538   0.463 0.643502
## smallcity     0.221122   0.162964   1.357 0.175095
## y74           0.240846   0.175541   1.372 0.170334
## y76          -0.139590   0.181902  -0.767 0.443012
## y78          -0.104546   0.184622  -0.566 0.571324
## y80          -0.086997   0.185803  -0.468 0.639718
## y82          -0.414209   0.174412  -2.375 0.017723 *
## y84          -0.565326   0.177398  -3.187 0.001479 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.581 on 1112 degrees of freedom
## Multiple R-squared:  0.09941,    Adjusted R-squared:  0.08645
## F-statistic: 7.671 on 16 and 1112 DF,  p-value: < 2.2e-16
```

The evolution of fertility rates over time after controlling of other observable factors can be interpreted as follows:

- Base year: 1972
- Negative coefficients indicate a drop in fertility in the early 1980's Coefficient of $y82$ (-0.41) indicates that women had on average 0.41 less children, i.e., 100 women had 41 kids less than 1972.
- This drop is independent from education since we are controlling for education.
- More educated women have fewer children
- Assumes that the effect of each explanatory variable remains constant.

The next example uses `cps7885` and interacts year dummy with key explanatory variables to see if the effect of that variable has changed over time. That is, the following model is estimated:

$$\ln(wage) = \beta_0 + \gamma_0 \cdot y85 + \beta_1 \cdot educ + \gamma_1 \cdot y85 \cdot educ + \beta_2 \cdot exper + \beta_3 \cdot exper^2 + \beta_4 \cdot union + \beta_5 \cdot female + \gamma_5 \cdot y85 \cdot female$$

```
cps7885$y85    = ifelse(cps7885$year==85,1,0)
summary(lm(formula=log(wage)~y85+educ+y85*educ+exper+I(exper^2)+union+female+y85:female,data=cps7885))
```

```
##
## Call:
## lm(formula = log(wage) ~ y85 + educ + y85 * educ + exper + I(exper^2) +
##     union + female + y85:female, data = cps7885)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56098 -0.25828  0.00864  0.26571  2.11669
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.589e-01  9.345e-02   4.911 1.05e-06 ***
## y85          1.178e-01  1.238e-01   0.952   0.3415
## educ         7.472e-02  6.676e-03  11.192  < 2e-16 ***
## exper        2.958e-02  3.567e-03   8.293 3.27e-16 ***
## I(exper^2)  -3.994e-04  7.754e-05  -5.151 3.08e-07 ***
## union        2.021e-01  3.029e-02   6.672 4.03e-11 ***
## female      -3.167e-01  3.662e-02  -8.648  < 2e-16 ***
## y85:educ     1.846e-02  9.354e-03   1.974   0.0487 *
## y85:female   8.505e-02  5.131e-02   1.658   0.0977 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4127 on 1075 degrees of freedom
## Multiple R-squared:  0.4262, Adjusted R-squared:  0.4219
## F-statistic:  99.8 on 8 and 1075 DF,  p-value: < 2.2e-16
```

This model can be interpreted as follows:

- β_0 is the 1978 intercept
- $\beta_0 + \gamma_0$ is the 1985 intercept
- β_1 is the return to education in 1978
- $\beta_1 + \gamma_1$ is the return to education in 1985
- γ_1 measures how the return to education has changed over the seven year period
- 1978 return to education: 7.47%
- 1985 return to education: 7.47%+1.85% = 9.32%
- 1978 gender gap: 31.67%
- 1985 gender gap: 31.67% - 8.51% = 23.16%

The last example regarding pooled data illustrates how misleading a regression model can be if executed incorrectly. The data set is called `kiel` and is on home values near the location of an garbage incinerator. The important aspect of the data set is that there was no knowledge about the proposed incinerator in 1978. In a first step, the data is separated into the two years:

```
kiel1978 = subset(kiel,year==1978)
kiel1981 = subset(kiel,year==1981)
```

Next, two regressions for each of the years are estimated.

```
summary(lm(rprice~nearinc,data=kiel1981))
```

```
##
## Call:
## lm(formula = rprice ~ nearinc, data = kiel1981)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60678 -19832  -2997  21139 136754
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   101308       3093  32.754  < 2e-16 ***
## nearinc       -30688       5828  -5.266 5.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31240 on 140 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1594
## F-statistic: 27.73 on 1 and 140 DF,  p-value: 5.139e-07
```

```
summary(lm(rprice~nearinc,data=kiel1978))
```

```
##
## Call:
## lm(formula = rprice ~ nearinc, data = kiel1978)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56517 -16605  -3193   8683 236307
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82517       2654  31.094  < 2e-16 ***
## nearinc       -18824       4745  -3.968 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29430 on 177 degrees of freedom
## Multiple R-squared:  0.08167,    Adjusted R-squared:  0.07648
## F-statistic: 15.74 on 1 and 177 DF,  p-value: 0.0001054
```

The results can be used for a difference-in-difference estimator: -\$30,688-(-\$18,824)=-\$11,864. Expressed differently:

$$\hat{\delta}_1 = (price_{81,near} - price_{81,far}) - (price_{78,near} - price_{78,far})$$

where $\hat{\delta}_1$ represents the difference over time in average differences in housing prices in the two locations. To determine statistical significance, the following model must be estimated:

$$price = \beta_0 + \gamma_0 \cdot y81 + \beta_1 \cdot nearinc + \delta_1 \cdot y81 \cdot nearinc$$

```
summary(lm(rprice~y81+nearinc+y81:nearinc,data=kiel))
```

```
##
## Call:
## lm(formula = rprice ~ y81 + nearinc + y81:nearinc, data = kiel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60678 -17693  -3031  12483 236307
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82517       2727  30.260  < 2e-16 ***
## y81            18790       4050   4.640 5.12e-06 ***
## nearinc       -18824       4875  -3.861 0.000137 ***
## y81:nearinc   -11864       7457  -1.591 0.112595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30240 on 317 degrees of freedom
## Multiple R-squared:  0.1739, Adjusted R-squared:  0.1661
## F-statistic: 22.25 on 3 and 317 DF,  p-value: 4.224e-13
```

The interpretation of the coefficients is as follows:

- β_0: Average home value which is not near the garbage incinerator
- $\gamma_0 \cdot y81$: Average change in housing values for all homes
- $\beta_1 \cdot nearinc$: Location effect that is not due to the incinerator
- γ_1: Decline in housing values due to incinerator

Include *age* and age^2 in the above equation to take advantage of the information provided in the data leads to the following result:

```
summary(lm(rprice~y81+nearinc+y81:nearinc+age+I(age^2),data=kiel))
```

```
##
```

```
## Call:
## lm(formula = rprice ~ y81 + nearinc + y81:nearinc + age + I(age^2),
##     data = kiel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79349 -14431  -1711  10069 201486
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.912e+04  2.406e+03  37.039  < 2e-16 ***
## y81          2.132e+04  3.444e+03   6.191 1.86e-09 ***
## nearinc      9.398e+03  4.812e+03   1.953 0.051713 .
## age         -1.494e+03  1.319e+02 -11.333  < 2e-16 ***
## I(age^2)     8.691e+00  8.481e-01  10.248  < 2e-16 ***
## y81:nearinc -2.192e+04  6.360e+03  -3.447 0.000644 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25540 on 315 degrees of freedom
## Multiple R-squared:  0.4144, Adjusted R-squared:  0.4052
## F-statistic: 44.59 on 5 and 315 DF,  p-value: < 2.2e-16
```

Other variables such as *cbd*, *rooms*, *area*, *land*, and *baths* can be added as well.

```
summary(lm(rprice~y81+nearinc+y81:nearinc+age+I(age^2)+intst+land+area+rooms+baths,data=kiel))
```

```
##
## Call:
## lm(formula = rprice ~ y81 + nearinc + y81:nearinc + age + I(age^2) +
##     intst + land + area + rooms + baths, data = kiel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -76721  -8885   -252   8433 136649
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.381e+04  1.117e+04   1.237  0.21720
## y81          1.393e+04  2.799e+03   4.977 1.07e-06 ***
## nearinc      3.780e+03  4.453e+03   0.849  0.39661
## age         -7.395e+02  1.311e+02  -5.639 3.85e-08 ***
## I(age^2)     3.453e+00  8.128e-01   4.248 2.86e-05 ***
## intst       -5.386e-01  1.963e-01  -2.743  0.00643 **
## land         1.414e-01  3.108e-02   4.551 7.69e-06 ***
## area         1.809e+01  2.306e+00   7.843 7.16e-14 ***
## rooms        3.304e+03  1.661e+03   1.989  0.04758 *
## baths        6.977e+03  2.581e+03   2.703  0.00725 **
## y81:nearinc -1.418e+04  4.987e+03  -2.843  0.00477 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19620 on 310 degrees of freedom
## Multiple R-squared:   0.66,  Adjusted R-squared:  0.6491
## F-statistic: 60.19 on 10 and 310 DF,  p-value: < 2.2e-16
```

In general, the results show that homes have lost 9.3% in values when including additional independent variables and using the natural logarithm of price.

18.3 Fixed Effects Panel Data Model

The two sections on fixed and random effects panel data model use a very old data set but which is standard in all panel data texts. The data set is called `grunfeld` and is part of the package plm. The data contains the following variables of 10 companies over the period 1935 to 1954:

- *inv*: Investment
- *value*: Value of the firm
- *capital*: Capital stock

The fixed effects model or Least-Squares Dummy Variable (LSDV) regression model assumes constant slope coefficients but varying intercepts over i. The regression equation can be written as:

$$inv_{it} = \beta_{0i} + \beta_1 \cdot value_{it} + \beta_2 \cdot capital_{it}$$

where i and t represent the firms and time, respectively. This model can also be written as

$$inv_{it} = \alpha_0 + \alpha_1 \cdot D_{1i} + \alpha_2 \cdot D_{2i} + \alpha_3 \cdot D_{3i} + \beta_1 \cdot value_{it} + \beta_2 \cdot capital_{it}$$

Individual specific effects:

$$y_{it} = \alpha_i + \beta_i \cdot x_{it} + \epsilon_{it}$$

where α_i can be fixed or random. The companies of interest for this chapter are GM (firm 1), U.S. Steel (firm 2), GE (firm 3), and Westinghouse (firm 8).

```
grunfeld = subset(grunfeld,grunfeld$firm %in% c(1,2,3,8))
```

In a first step, a pooled model is executed, i.e., all cross-sectional and time series observations are combined into a single data set.

$$inv_i = \beta_0 + \beta_1 \cdot value_i + \beta_2 \cdot capital_i$$

The general formulation of the pooled model:

$$y_{it} = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

There are multiple issues associated with a pooled OLS model:

- Ignores heterogeneity among the observations and time.
- Presence of heterogeneity: Correlation between independent variables and error term leads to biased and inconsistent coefficient estimates. Solution: Fixed effects model takes heterogeneity into account. ⇒ Autocorrelation between error terms. Fix: Random effects model

To use the functions from plm, define the data as a panel data set:

```
grunfeld = pdata.frame(grunfeld,index=c("firm","year"))
```

There are two possibilities to execute a pooled OLS Model. Use the regular `lm()` function or use `plm()` specifying the model as “pooling”. The outputs will be names `grunwald.ols` and `grunwald.pooling`.

```
grunfeld.pooling = plm(inv~value+capital,data=grunfeld,model="pooling")
summary(grunfeld.pooling)
```

```
## Pooling Model
##
## Call:
## plm(formula = inv ~ value + capital, data = grunfeld, model = "pooling")
##
## Balanced Panel: n = 4, T = 20, N = 80
```

```
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -319.6766  -99.9523    1.9647   65.9905  336.2072
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) -62.831841  29.725385 -2.1137   0.03778 *
## value         0.110521   0.013776  8.0230 9.186e-12 ***
## capital       0.300463   0.049399  6.0823 4.273e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6410400
## Residual Sum of Squares: 1572700
## R-Squared:      0.75466
## Adj. R-Squared: 0.74829
## F-statistic: 118.424 on 2 and 77 DF, p-value: < 2.22e-16
```

Fixed effects model

- Intercept β_{0i} is firm specific.
- For an individual, this could be education and/or ability, possibly correlated with independent variables
- Intercept is time-invariant.
- Slope coefficients do not vary across individuals (firms) or time

To implement the model in R, the function `plm()` must be used specifying the model as "within". The output will be names `grunwald.fixed`.

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = inv ~ value + capital, data = grunfeld, model = "within")
##
## Balanced Panel: n = 4, T = 20, N = 80
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -184.6581  -48.2612    9.3252   40.5471  197.6681
##
## Coefficients:
##         Estimate Std. Error t-value Pr(>|t|)
## value   0.108400   0.017566  6.1711  3.3e-08 ***
## capital 0.345058   0.026708 12.9195  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2171500
## Residual Sum of Squares: 422220
## R-Squared:      0.80556
## Adj. R-Squared: 0.79242
## F-statistic: 153.291 on 2 and 74 DF, p-value: < 2.22e-16
```

Use `fixef(grunfeld.fixed)` to get the firm specific intercepts. The function `pFtest()` can be used to test whether a fixed effects or OLS is appropriate (H_0: OLS better). If the significance level is set to $\alpha = 0.05$ then the fixed effects model is a better choice if the p-value is below 0.05:

```r
pFtest(grunfeld.fixed,grunfeld.pooling)
```

```
##
##  F test for individual effects
##
## data:  inv ~ value + capital
## F = 67.215, df1 = 3, df2 = 74, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

The fixed effects model can also be implemented using the function `lm()`:

```r
lm(formula=inv~value+capital+factor(firm),data=grunfeld)
```

```
##
## Call:
## lm(formula = inv ~ value + capital + factor(firm), data = grunfeld)
##
## Coefficients:
##   (Intercept)          value        capital  factor(firm)2  factor(firm)3  factor(firm)8
##      -85.5153         0.1084         0.3451       180.5029      -160.7122        26.1296
```

18.4 Random Effects Model

18.5 Exercises

1. ***NFL II*** (***): Consider the data set `nfl` which includes the performance, salary, and facial symmetry of NFL quarterbacks. The data includes the name and the year and thus, it can be estimated as a panel data model. In a first step, convert the data into a panel data set using the function `pdata.frame()`. Next, estimate three models: (a) Regular pooled OLS model, (2) fixed-effects model, and (3) random effects model. The regression equation is the same for each model:

   $$\ln(total) = \beta_0+\beta_1\cdot yards+\beta_2\cdot att+\beta_3\cdot exp+\beta_4\cdot exp^2+\beta_5\cdot draft1+\beta_6\cdot draft2+\beta_7\cdot veteran+\beta_8\cdot changeteam+\beta_9\cdot pbowlever+$$

   Report and interpret the output for all three models. What happens to the variable *symm* in the fixed effects model and why? How does the panel data model compare to the original model which does not incorporate the panel structure?

2. ***Renewable Energy*** (***): This question is based on the paper The effect of the feed-in-system policy on renewable energy investments: Evidence from the EU countries by Alolo et al. (2020).

3. ***Guns and Crime*** (***): There is a fierce debate on the relationship between gun ownership and crime. On the one hand, there is the argument that more guns prevent crime due to a deterrence effect. On the other hand, more guns trickling into society increase the likelihood of crime being committed. This data set is one example on how the issue at hand can be analyzed. The data set which is used for this question is part of the AER and can be imported with the command `data("Guns",package="AER")`. Please read the description of the data set which is available as part of the package. The first regression model to be estimated is written as:

   $$\ln(violent) = \beta_0 + \beta_1 \cdot law + \epsilon$$

   Estimate a fixed effects panel model using the equation above and report the R output. Interpret the results. The second model include year and state fixed effects and is written as follows:

   $$\ln(violent) = \beta_0 + \beta_1 \cdot law + \alpha_i + \lambda_t + \epsilon$$

   In a third and last model, include the variables *prisoners*, *density*, *income*, *population afam*, *cauc*, and *male*. What does the model suggest about the opposing views mentioned at the beginning of the question?

4. ***WDI Panel Data*** (***): Previous questions used the `wdi` data set without incorporating the potential panel structure of the information. For this question, you will have the natural logarithm of *mortrate* as the dependent variables and *gdp* and *litrate* as the independent variables. In a first step, subset the data such that you only have those three variables. Next, estimate two models: (1) Pooled OLS and (2) fixed effects panel model. Compare the results. Does the impact of GDP and the literacy rate change significantly between the models?

19 Time Series

This chapter introduces time as a component in regression models. Times series represent a temporal ordering of the data. In the cross-section models seen so far, the observations could have been in any order. In this chapter, the ordering of the observations matters. It is usually assumed that there is a stochastic process generating the series and the important aspect is that only a single realization of the stochastic process is observed. The following topics are covered:

- Trend and seasonality
- Finite distributed lag models (including past or lagged independent variables)
- Autoregressive model (including past or lagged dependent variables) also known as dynamic models:
- Forecasting

The static regression model is presented before moving into proper time series analysis. The model is used if the data represents various time periods but the independent variables $x_{i,t}$ have an immediate effect on the dependent variable y_t. For example, if the per-capita chicken consumption is a function of the chicken price and real disposable income, then the following model can be estimated using the data in `meatdemand`.

```
summary(lm(qchicken~rdi+pchicken,data=meatdemand))
```

```
##
## Call:
## lm(formula = qchicken ~ rdi + pchicken, data = meatdemand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6335 -2.4765  0.2693  2.2275  6.3905
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.8885914 17.7955903   2.691   0.0107 *
## rdi          0.0014297  0.0002049   6.977 3.52e-08 ***
## pchicken    -0.1143327  0.0454294  -2.517   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.204 on 36 degrees of freedom
## Multiple R-squared:  0.9545, Adjusted R-squared:  0.952
## F-statistic: 377.8 on 2 and 36 DF,  p-value: < 2.2e-16
```

This assumes that there is an immediate effect of income and chicken price on consumption.

19.1 Trend and Seasonality

In general, trends in the data can be linear:

$$y_t = \beta_0 + \beta_1 \cdot t + \epsilon_t$$

or exponential:

$$\ln(y_t) = \beta_0 + \beta_1 \cdot t + \epsilon_t$$

Note that β_1 in the exponential time trend model is the average annual growth rate (assuming t is in years). Often, data can be decomposed into three components:

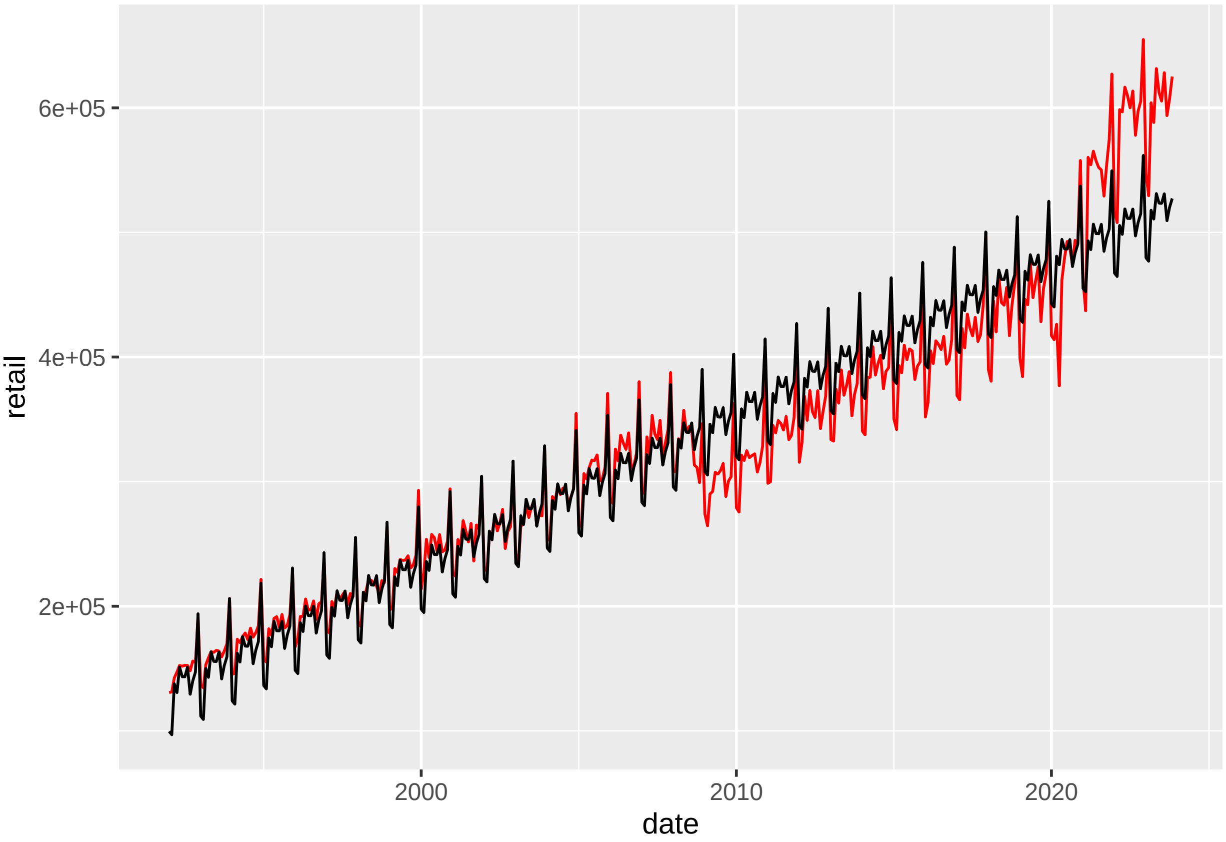
- Trend
- Season
- Random component

The seasonal component can be included via dummy variables. For example, for quarterly data the following model can be used:

$$y_t = \beta_0 + \delta_1 \cdot Q1_t + \delta_2 \cdot Q2_t + \delta_3 \cdot Q3_t + \beta_1 \cdot x_{1,t} + \cdots + \beta_k \cdot x_{k,t} + \epsilon_t$$

One seasonal dummy must be dropped. That is, quarterly and yearly data require three and eleven dummy variables, respectively. Consider the `retail` data.

```
retail$date  = as.Date(retail$date,format="%Y-%m-%d")
retail$month = format(retail$date,"%m")
retail$t     = c(1:nrow(retail))
bhat         = lm(retail~factor(month)+t,data=retail)
retail$fit   = predict.lm(bhat)
ggplot(retail)+
  geom_line(mapping=aes(x=date,y=retail),color="red")+
  geom_line(mapping=aes(x=date,y=fit))
```

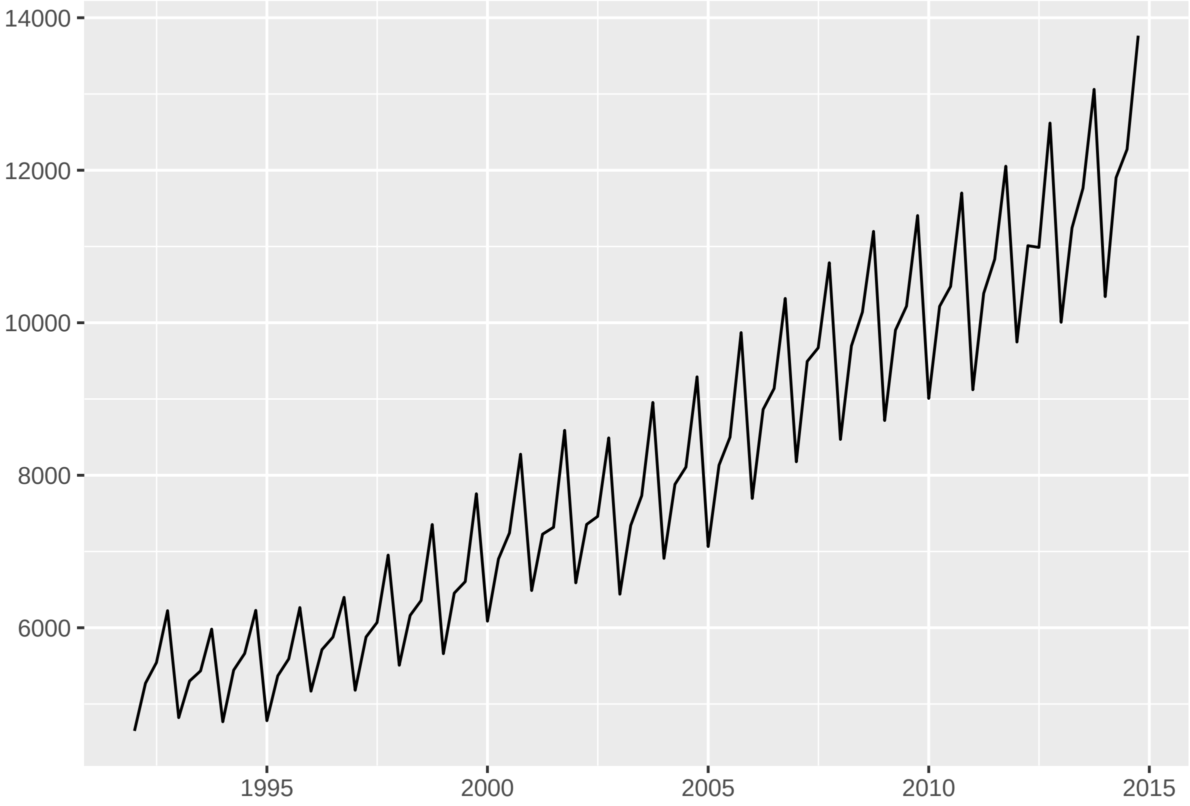


A second example considers quarterly `beer` production in Australia. In a first step, the data is converted into a time series:

```
beer = ts(beer$production,start=c(1992,1),end=c(2014,4),frequency=4)
```

Next, the data is plotted using a function from the package ggfortify. Additional documentation using the package is found under Plotting ts objects.

```
autoplot(beer)
```



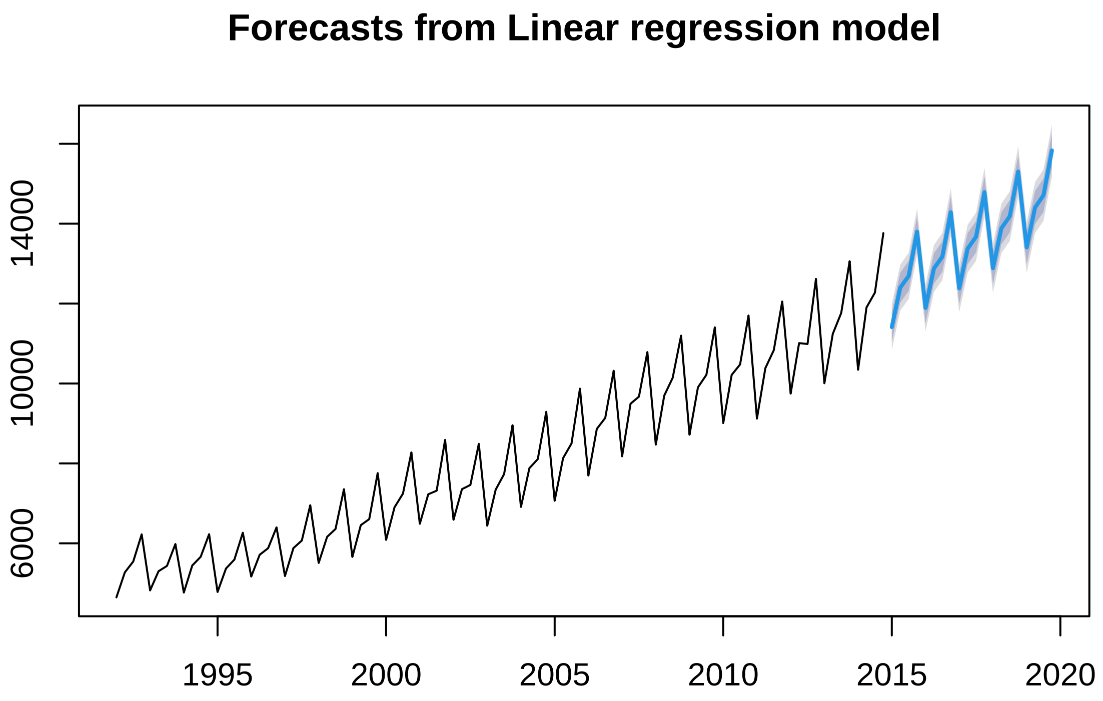
Note that the beer now appears in a different category in the Global Environment, i.e., not under "Data" anymore. The function tslm from the package forecast is used next. The function fits a linear model including seasonality and a trend component (and a trend-squared component if desired).

```
bhat = tslm(beer~trend+I(trend^2)+season)
summary(bhat)
```

```
##
## Call:
## tslm(formula = beer ~ trend + I(trend^2) + season)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -677.24 -163.84    2.12  195.34  529.37
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.082e+03  9.865e+01  41.381  < 2e-16 ***
## trend       3.975e+01  4.298e+00   9.249 1.53e-14 ***
## I(trend^2)  4.200e-01  4.477e-02   9.381 8.26e-15 ***
## season2     8.677e+02  7.987e+01  10.864  < 2e-16 ***
## season3     1.043e+03  7.990e+01  13.061  < 2e-16 ***
## season4     2.031e+03  7.993e+01  25.408  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270.8 on 86 degrees of freedom
## Multiple R-squared:  0.9867, Adjusted R-squared:  0.9859
## F-statistic:  1278 on 5 and 86 DF,  p-value: < 2.2e-16
```

The package forecast can be also be used to forcast:

```
plot(forecast(bhat,h=20))
```



19.1.1 Practice Exercise

1. Consider the data in `ez`. It contains unemployment claims (*uclms*) from Anderson (IN) before and after the establishments of an enterprise zone (EZ). The purpose of an EZ is to provide incentives for business to invest in area that are usually plagued by economic distress. Execute two regression models with ln(*uclms*) as the dependent variable and including a time trend and monthly dummy variables as the independent variables: (1) without EZ and (2) with EZ. What can be concluded in terms of trend, seasonality, and the effectiveness of the EZ.
2. Consider the data in `traffic` which contains information on accidents, traffic laws, and other variables for California. The dependent variable of interest is the natural log of *totacc*. In a first regression, use the time trend and monthly dummy variables as the independent variables. In a second regression, add *wkends*, *unem*, *spdlaw*, and *beltlaw*. What can be concluded in terms of trend, seasonality, and the effectiveness of the laws concerning speed limits and belt usage. Why would weekends and

unemployment be important?

19.2 Finite Distributed Lag Models

Distributed-lag models include past or lagged independent variables:

$$y_t = \alpha + \beta_0 \cdot x_t + \beta_1 \cdot x_{t-1} + \beta_2 \cdot x_{t-2} + \ldots \beta_k \cdot x_{t-k} + \epsilon$$

There are many reasons to include lagged independent variables such as psychological (e.g., it is difficult to break a habit or adjust to a new situation), economic (e.g., contractual obligations), or political (e.g., effectiveness of policy builds up over time) reasons.

The relationship between income and consumption is used to introduce distributed lag models. Assume the following relationship between income and consumption:

$$C_t = \alpha + \beta_0 \cdot I_t + \beta_1 \cdot I_{t-1} + \beta_2 \cdot I_{t-2}$$

Assume that $\alpha_0 = 100$, $\beta_0 = 0.4$, $\beta_1 = 0.3$, and $\beta_2 = 0.2$. For this example, the following questions are of interest:

- What is the long-run consumption with $4,000?
- How does the consumption change if the income increases to $5000?

Note that the sum of the β_i's is 0.9. The long-run multiplier (or long-run propensity) is written as:

$$\sum_{i=1}^{k} \beta_i = \beta_0 + \beta_1 + \beta_2 + \cdots + \beta_k = \beta$$

The question about how may lagged independent variables to include is a difficult to answer question. If the assumption is made that all β_k are of the same sign, then the so-called Koyck transformation can be applied:

$$\beta_k = \beta_0 \cdot \lambda^k \quad \text{for} \quad k = 0, 1, 2, \ldots$$

Characteristics of this assumption:

- $\lambda < 1$ gives less weight to distant values of β
- Long-run multiplier is finite, i.e.,
$$\sum_{k=0}^{\infty} \beta_k = \beta_0 \cdot \left(\frac{1}{1-\lambda}\right)$$

Given the above assumptions, the Koyck transformation can be applied to the regression model. The original model is written as:

$$y_t = \alpha + \beta_0 \cdot x_t + \beta_0 \cdot \lambda x_{t-1} + \beta_0 \cdot \lambda^2 \cdot x_{t-2} + \cdots + \epsilon_t$$

The reformulated equation to be estimated is

$$y_t = \alpha \cdot (1 - \lambda) + \beta_0 \cdot x_t + \lambda \cdot y_{t-1} + \upsilon_t$$

The notation of the error term has changed from ϵ_t to υ_t in order to highlight that the terms will be different.

```
koyck = usdata
koyck$year = substr(koyck$date,1,4)
koyck = aggregate(koyck[c("income","consumption")],FUN=sum,by=list(koyck$year))
colnames(koyck) = c("year","income","consumption")
bhat = lm(formula = consumption ~ income + Lag(consumption), data = koyck)
summary(bhat)
```

```
##
## Call:
## lm(formula = consumption ~ income + Lag(consumption), data = koyck)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14440.4   -849.5    225.1   1291.6   5433.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.071e+03  8.209e+02  -2.523   0.0138 *
## income            6.870e-01  4.395e-02  15.633  < 2e-16 ***
## Lag(consumption)  2.530e-01  4.740e-02   5.338 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2717 on 73 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9961, Adjusted R-squared:  0.996
## F-statistic:  9365 on 2 and 73 DF,  p-value: < 2.2e-16
```

19.3 Basic Theoretical Aspects of Time Series

A collection of random variables ordered in time is called a stochastic process. The observed value in a given time period is usually a particular realization of the stochastic process. An important concept is so-called stationary of a time series or stochastic process. A time series is stationary if mean, variance, and covariance between any lagged observation are constant:

- Constant mean: $E(y_t) = \mu$
- Constant variance: $Var(y_t) = \sigma^2$
- Constant covariance: $Cov(y_t, y_{t-h})$ depends on h but not on t.

A time series with a trend is usually not stationary (or nonstationary). The concepts behind are explained latter but the issue can be illustrated with a simple simulation.

```
spurious             = data.frame(x=matrix(0,500,1),y=matrix(0,500,1))
spurious[1,1]        = 1
spurious[1,2]        = 2
for(i in 2:500){
     spurious[i,1] = spurious[i-1,1]+rnorm(1)
     spurious[i,2] = spurious[i-1,2]+rnorm(1)}
bhat = lm(y~x,data=spurious)
summary(bhat)
```

```
##
## Call:
## lm(formula = y ~ x, data = spurious)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.339  -5.099  -0.836   5.200  15.915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.96742    0.54150 -16.560   <2e-16 ***
## x           -0.08441    0.04111  -2.054   0.0405 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.553 on 498 degrees of freedom
## Multiple R-squared:  0.008397,   Adjusted R-squared:  0.006406
## F-statistic: 4.217 on 1 and 498 DF,  p-value: 0.04054
```

```
rm(spurious,bhat)
```

There is no relationship between y and x yet the regression results indicate statistical significance.

Consider the following so-called autoregressive model. The model depends on lagged terms of the dependent variable:

$$y_t = \alpha + \phi_1 \cdot y_{t-1} + \epsilon_t$$

This is called an AR(1) because the y is lagged by one period. The requirement for a stationary AR(1) is that $|\phi_1| < 1$. The properties of the AR(1) process are:

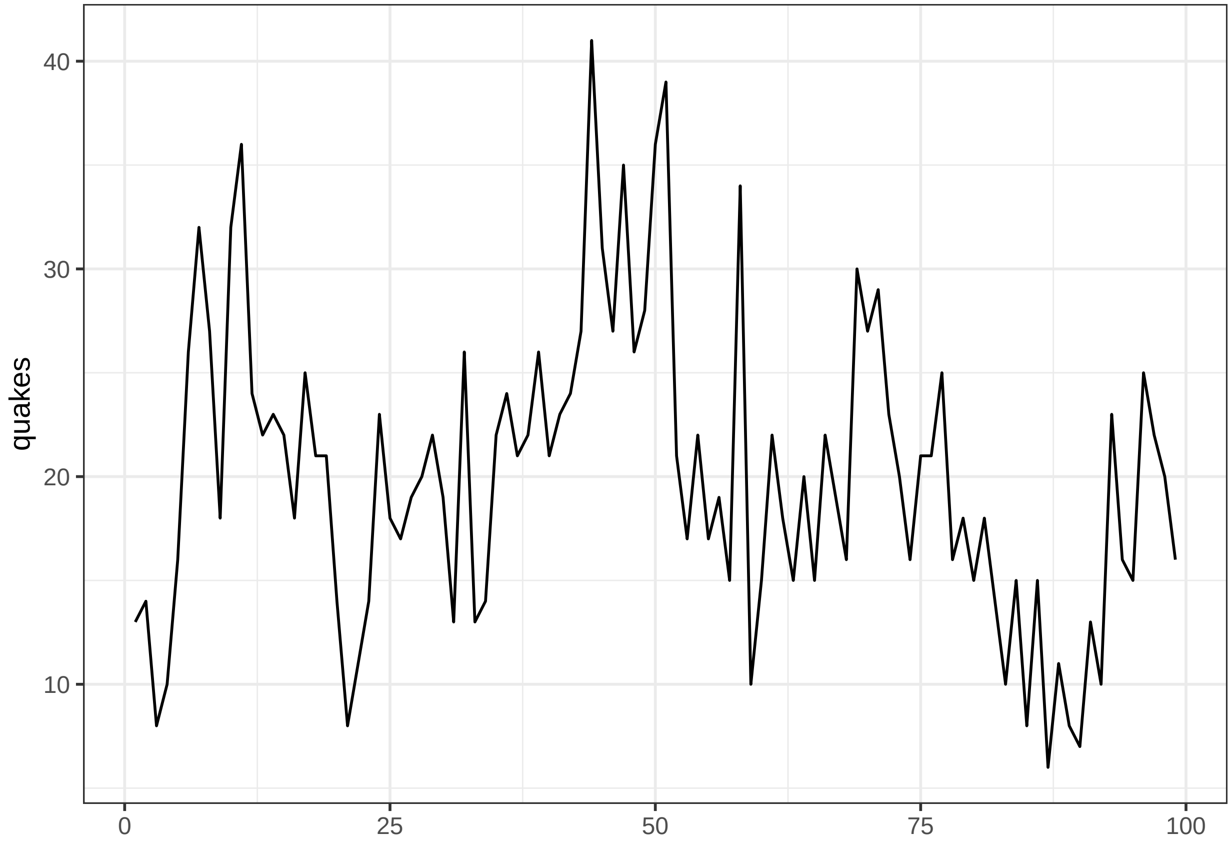
- Mean of y_t: $\mu = \frac{\alpha}{1-\phi_1}$
- Variance: $Var(x_t) = \frac{\sigma_w^2}{1-\phi_1^2}$
- Correlation: $\rho_h = \phi_1^h$

19.4 Autoregressive Model

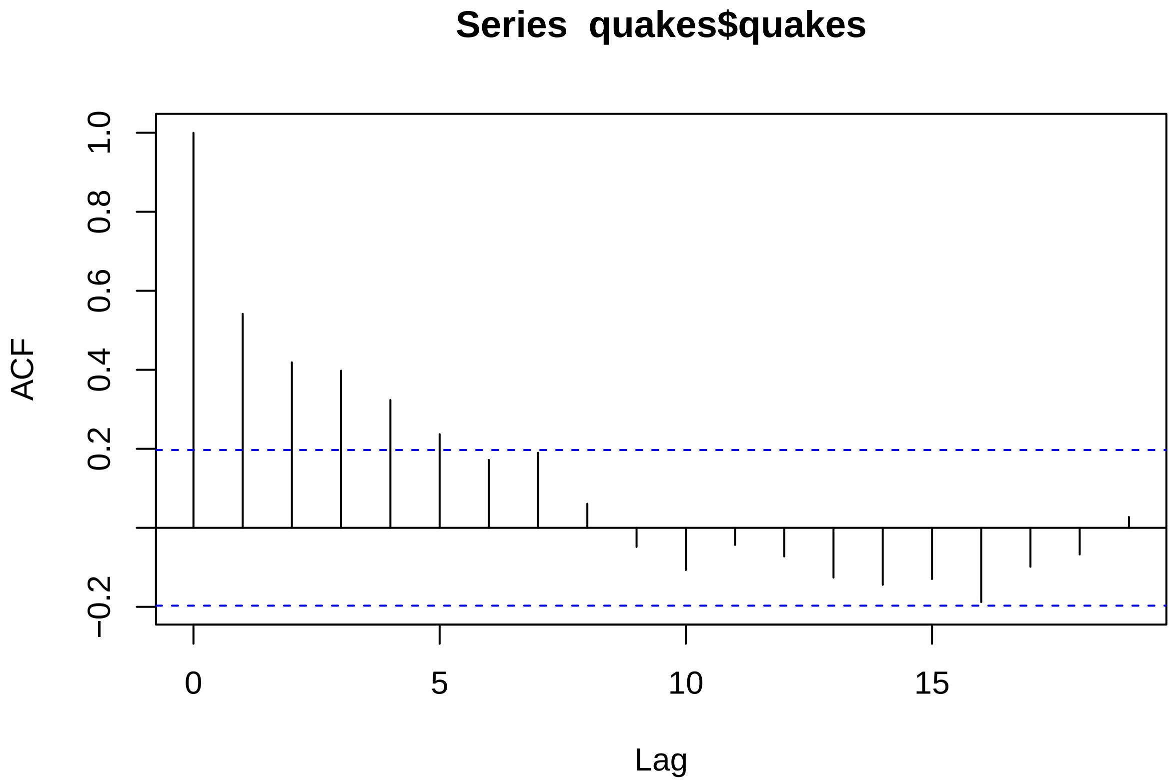
An autoregressive model includes lagged dependent variables. One of the simplest model is an autoregressive model of order 1, i.e., an AR(1) Model

$$y_t = \alpha + \beta \cdot y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2)$. Consider the data of earthquakes over magnitude 7 in the data set `quakes`:



In a first step, a scatter plot is constructed of Y_{t-1} and y_t. The easiest way is to use the function `acf`:



```
## [1] 0.5417329
```

The correlation coefficicent of 0.25 indicates a weak positive correlation between the number of earthquakes in periods t and $t-1$. Remember that correlation is not causation. The AR(1) model can be estimated with the `lm()` function used previously:

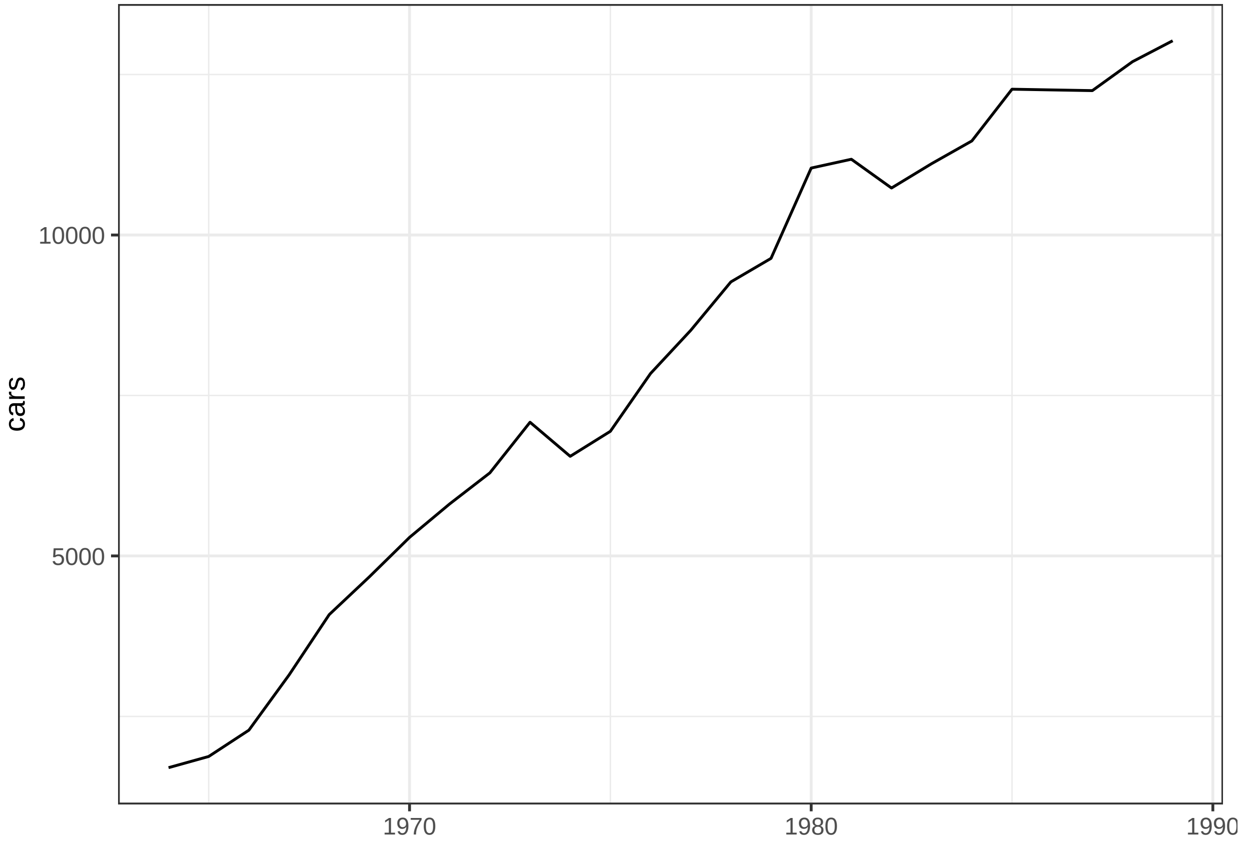
```
summary(lm(quakes~Lag(quakes),data=quakes))
```

```
##
## Call:
## lm(formula = quakes ~ Lag(quakes), data = quakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.666  -3.901  -0.351   3.050  17.138
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.19070    1.81924   5.052 2.08e-06 ***
## Lag(quakes)  0.54339    0.08528   6.372 6.47e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.122 on 96 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2972, Adjusted R-squared:  0.2899
## F-statistic:  40.6 on 1 and 96 DF,  p-value: 6.471e-09
```

Note that although the slope coefficient associated with the lagged term is statistically significant, the R-squared value is very low. The third examples uses data on Japanese car production (`jcars`) to illustrate the concept of autocorrelation. The focus is on car production after 1963.

In a first step, the data is visualized using `ggplot`:

```
ggplot(jcars,aes(x=year,y=cars))+geom_line()+theme_bw()+theme(axis.title.x=element_blank())
```

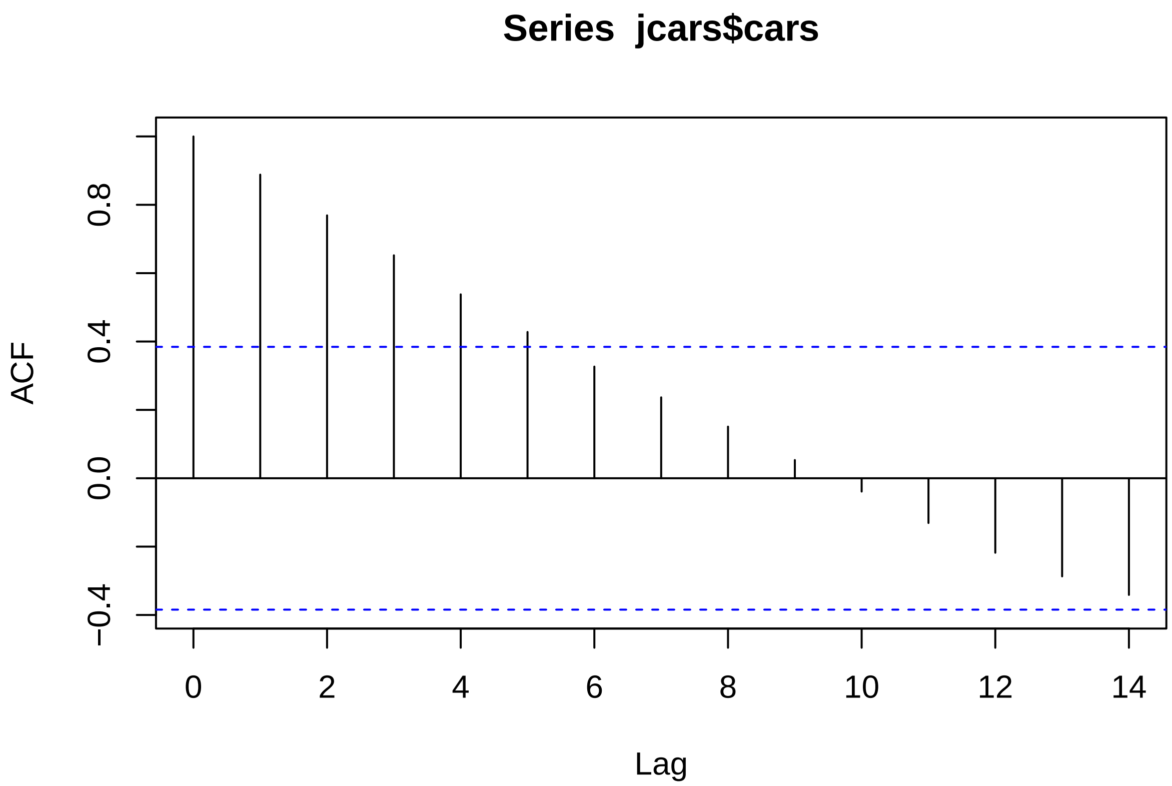


The sample autocorrelation function (ACF) is the correlation between y_t and y_{t-1}, y_{t-2}, y_{t-3}, and so on. It can be written as follows:

$$\rho_j = \frac{Cov(y_t, y_{t-j})}{\sqrt{Var(y_t) \cdot Var(y_{t-j})}}$$

The ACF can be used to identify a possible structure of time series either of the actual time series or the residuals of the regression. The autocorrelation function (ACF) is plotted using the function `acf()` in R.

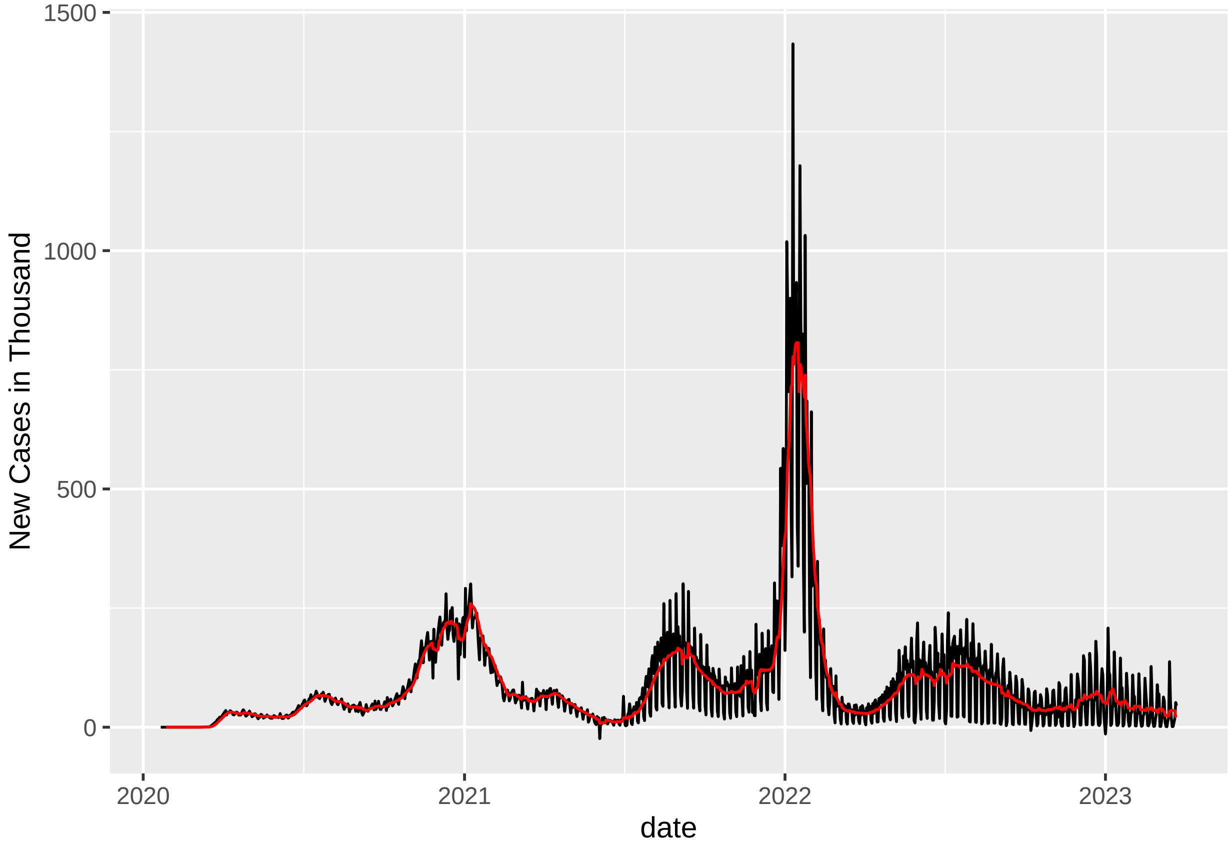
```
acf(jcars$cars)
```



19.5 Moving Average Models

The moving average model presented in this chapter should not be confused with the concept of moving average (also known as rolling mean) which is familiar to many people. To avoid confusion, the term rolling mean is used for the latter. An example of a rolling mean is plotted below:

```
ggplot(covid,aes(x=date))+
     geom_line(aes(y=newcases/1000))+
     geom_line(aes(y=rollmean(newcases/1000,7,na.pad=TRUE,align="right")),
               color="red")+ylab("New Cases in Thousand")
```



A moving average term in a time series model is a past error (multiplied by a coefficient), e.g., MA(1):

$$x_t = \mu + w_t + \theta_1 \cdot w_{t-1}$$

where $w_t \sim N(0, \sigma_w^2)$. The MA(1) model is written as:

$$x_t = \mu + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2}$$

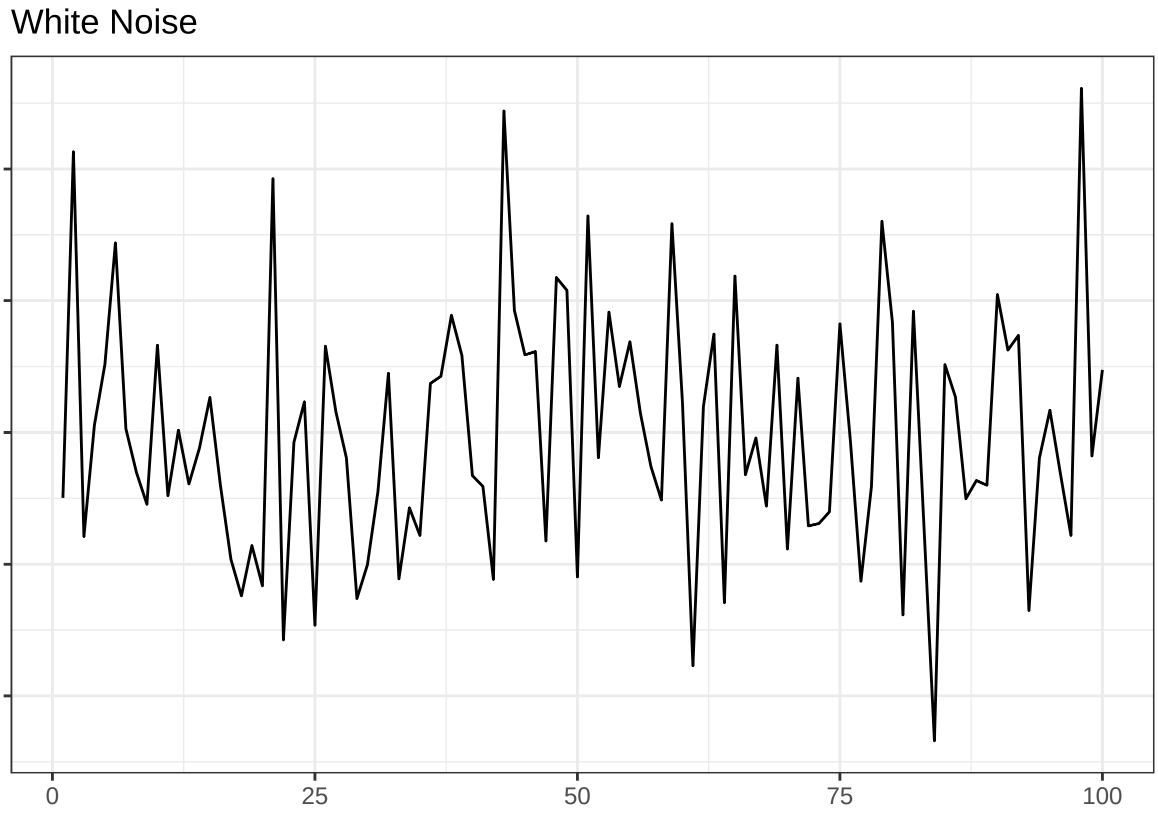
And the properties of an MA(1) model are as follows:

- $E[x_t] = \mu$
- $Var(x_t) = \sigma_w^2(1 + \theta_1^2)$
- ACF is $\rho_1 = \theta_1/(1 + \theta_1^2)$ and $\rho_h = 0$ for $h \geq 2$

19.6 Random Walk

A time series with no autocorrelation is called White Noise. Consider the following plot of White Noise:

```
autoplot(ts(rnorm(100)))+ggtitle("White Noise")+theme_bw()
```



Let ϵ_t be white noise then the random walk without drift is

$$y_t = y_{t-1} + \epsilon_t$$

This is called an autoregressive model of order 1 or AR(1). Example:

$$y_1 = y_0 + \epsilon_1$$

Consider a different model written as

$$y_2 = y_1 + \epsilon_2 = y_0 + \epsilon_1 + \epsilon_2$$

This is not a stationary process and it can be shown that $E(y_t) = y_0$ and $Var(y_t) = t \cdot \sigma^2$. However

$$y_t - y_{t-1} = \Delta y_t = \epsilon_t$$

Let ϵ_t be white noise then the random walk with drift is

$$y_t = c + y_{t-1} + \epsilon_t$$

where c is the drift parameter. It can be shown that $E(y_t) = y_0 + t \cdot c$ and $Var(y_t) = t \cdot \sigma^2$. An autoregressive model AR(p) can be written as

$$y_t = c + \sum_{i=1}^{p} \phi_p y_{t-p} + \epsilon_t$$

19.7 Forecasting Japanense Car Production

Model 1: Regular OLS Model

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

Model 2: Autoregressive Model

$$y_t = \beta_0 + \beta_1 t + n_t \quad \text{where} \quad n_t = \phi_1 n_{t-1} + \epsilon_t$$

```
summary(lm(cars~year,data=jcars))
```

```
##
## Call:
## lm(formula = cars ~ year, data = jcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.62 -406.49   47.09  353.35 1351.64
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -924484.82   30143.45  -30.67   <2e-16 ***
## year            471.81      15.25   30.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.2 on 24 degrees of freedom
## Multiple R-squared:  0.9755, Adjusted R-squared:  0.9745
## F-statistic: 957.1 on 1 and 24 DF,  p-value: < 2.2e-16
```

```
auto.arima(jcars$cars)
```

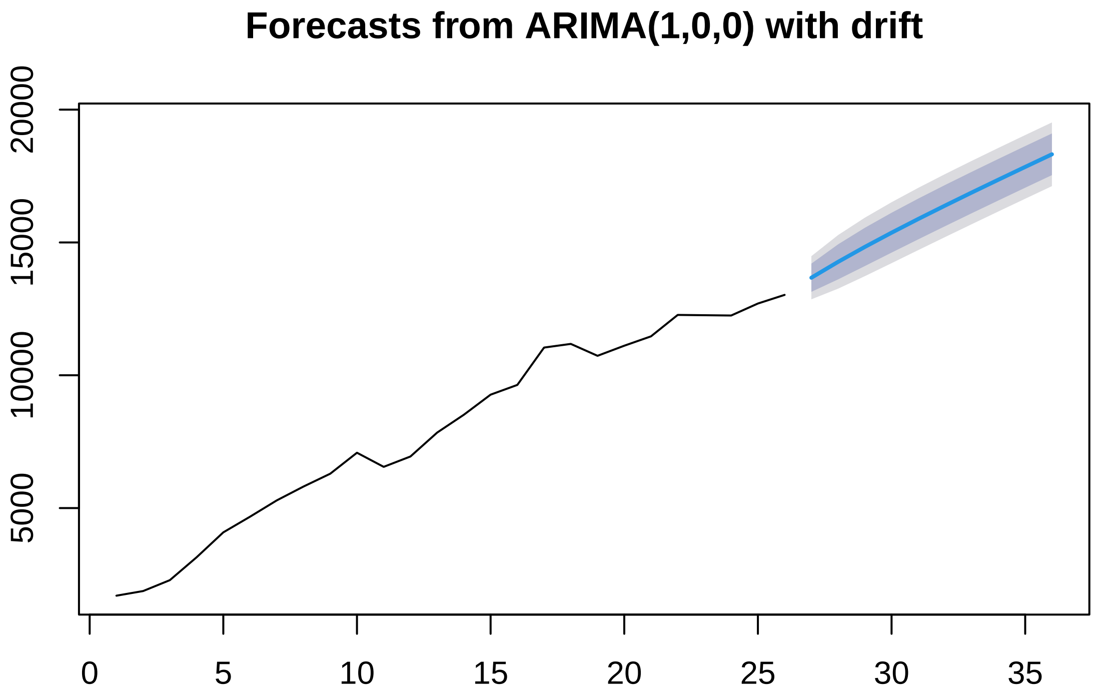
```
## Series: jcars$cars
## ARIMA(0,1,0) with drift
##
## Coefficients:
##          drift
##       452.9600
## s.e.   83.6424
##
## sigma^2 = 182188:  log likelihood = -186.37
## AIC=376.75   AICc=377.29   BIC=379.18
```

```
bhat = Arima(jcars$cars,order=c(1,0,0),include.constant=TRUE,include.drift = TRUE)
summary(bhat)
```

```
## Series: jcars$cars
## ARIMA(1,0,0) with drift
##
## Coefficients:
##          ar1  intercept     drift
##       0.7363  1662.4148  463.5637
## s.e.  0.1347   471.7223   29.2265
##
## sigma^2 = 171700:  log likelihood = -192.38
## AIC=392.77   AICc=394.67   BIC=397.8
##
```

```
## Training set error measures:
##                    ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
## Training set 17.28081 389.7285 311.2957 -0.7522648 5.354775 0.5840007 0.1564618
```

```
plot(forecast(bhat))
```



19.8 Exercises

1. ***Souvenir Shop*** (***): The data in `souvenirs` represents the sales of a souvenir beach shop in Australia. The sales volume in the shop fluctuates with the number of tourists which are in town. A larger number of tourists is observed in December for the holiday season and for the local surfing festival in March. Over the years, the shop has expanded in terms of items offered and also in store area.
   a. Plot the data. What do you observe in terms of trend, seasonality, and magnitude of fluctuations?
   b. Estimate a regular OLS Model with *sales* as the dependent variable and a trend as well as monthly dummy variables as independent variables.
   c. Plot the observed sales and predicted sales over time.
   d. Repeat the estimation and the plot from (b) and (c) but use the natural log of *sales*, i.e., $\ln(sales)$, as the dependent variable instead.
   e. Compare the two graphs you have generated. What is the difference and what is the reason for the difference? Which form of the dependent variable is more appropriate.
2. ***Retail Time Series*** (***): Consider the data `retail` and the following model:

$$\ln(retail_t) = \beta_0 + \beta_1 \cdot t + \delta_1 \cdot Q1_t + \delta_2 \cdot Q2_t + \delta_3 \cdot Q3_t + \epsilon_t$$

   a. Estimate the above regression equation, report, and interpret the results.
   b. How is using $\ln(retail)$ different from *retail* as the dependent variable? How does it change the interpretation of the coefficients?

   c. Re-estimate the model but this time using *retail* as the dependent variable.
   d. Plot (in the same graph) the observed values and both modeled series. What do you observe? Does one model fit better than the other? If yes, which one?
3. ***WDI Per-Capita Consumption*** (**): The data `wdi` includes per-capita income (*gdp*), overall household consumption (*consumption*), and *population*. For a country of your choice, do the following:
   a. Subset the data such that you are only left with the country of your choice.
   b. Create a new column which contains the consumption per capita, i.e., consumption divided by population.
   c. Make sure that there are at least 10 years of observations for the country of your choice.
   d. Estimate a Koyck-Model with consumption as the dependent variable and per-capita consumption as the dependent variable.
4. ***Cardiovascular Mortality*** (***): Consider the data in `cmort` which represents weekly data of cardiovascular mortality in Los Angeles over the period 1970 to 1979.
   a. Estimate an AR(2) model of the following form to the data:

$$y_t = \alpha + \beta_1 \cdot y_{t-1} + \beta_2 \cdot y_{t-2} + \epsilon_t$$

   Make sure to use the function `ar.ols()` with the options `deman=FALSE` and `intercept=TRUE`.
   b. Given the above model, what are the predicted values for the last four weeks of data. Compare the forecast to the observed values. Is it a good fit?

20 Data Sources

This chapter describes the functions and data sets associated with the lecture notes which are contained in the file DataAnalysisPAData.RData.

The following functions are included:

- `bday(k)`: This function calculates the probability of two people having the same birthday in a group of k people.
- `samplesize(me,p=0.5,pop)`: This function requires the desired margin of error as the input, e.g., $me = 0.03$ for ±3%. By default, the function calculates the sample size assuming a variance maximizing probability of $p = 0.5$ and infinite population. Both parameters can be adjusted.

Some data sets are generic, i.e., randomly generated, to highlight a particular concept while other are either based on public sources or are taken from academic papers. In any case, the sources are clearly indicated. Sometimes modifications were made to the data for ease of use, e.g., removing missing values or unnecessary variables. Some of the data sets are also associated with specific R packages which is indicated. All variable names are lowercase. The following data sets are included:

- `accidents` (N=30): Generic data on the number of *accidents*, *temperature* in Fahrenheit, and *precipitation* in inches.
- `airlines` (N=211918): On-time and delay statistics for the 100 largest airports (by number of 2019 arrivals) by airline January 2004 to May 2022. The variable names are mostly self-explanatory. The variables $arrflights$ and $arrdel15$ are the total number of arriving flights and the number of flight delayed at least 15 minutes. The original data is available here.
- `anscombe` (N=11): Anscombe's Quartet where yi and xi represent dependent and independent variable of set i, respectively. This data set is widely available and is also included in R.
- `apartments` (N=147): The data set contains data on furnished apartments for rent. The variables are *rent*, *rooms*, *area* (in square meters), managing *company*, and *location* (Southwest Berlin and Potsdam).
- `aptitude` (N=200): Aptitude (*apt*), reading (*read*), and *math* test scores for students (*id*). The program type is indicated by *prog*. The data set is taken from the UCLA Advanced Research Computing website on Tobit Models.

- `beer` (N=92): Quarterly beer production in Australia. The data is obtained from the lecture notes for STAT 510: Applied Time Series Analysis at Penn State.
- `biochemistry` (N=915): The data set is based on the article The Origins of Sex Differences in Science and assesses the number of articles written by graduate students three years before receiving their Ph.D. in biochemistry.
- `blm` (N=1359): The data is associated with the paper Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity. The variables include city (*geography*), total protests (*totprotests*), population (*pop*), population density (*popdensity*), percent black (*percentblack*), black poverty rate (*blackpovertyrate*), percent college-educated (*percentbachelor*), college students as a percent of the population (*collegeenrollpc*), Democratic vote share (*demshare*), police-caused deaths per 10,000 (*deathspc*), and black police-caused deaths per 10,000 (*deathsblackpc*).
- `bloodpressure` (N=20): The data is taken from STAT 501 Regression Methods at Penn State. It contains blood pressure data for 20 individuals on *bp* (in mm Hg), *age*, *weight* (in kilograms), *bsa* (body surface area in square meters), *dur* (duration of hypertension in years), *pulse* (basal pulse in beats per minutes), and *stress* (index number).
- `bmw` (N=30): Data on the prices and miles of a particular BMW 5-series model in the Indianapolis area. Some of the models have rear-wheel drive ($allwheeldrive = 0$) whereas some have all-wheel drive ($allwheeldrive = 1$).
- `boston` (N=506): Housing values in suburbs of Boston. The data set is part of the R package MASS. An explanation of the various variables can be found here.
- `censoring` (N=50): Generic data in which values of the response variable below 0 are reported as 0. The researcher observes y and x but not $yreal$.
- `chicagogrocery` (N=77): The data set was represents the information from 2013 grocery stores and [Community Data Snapshots Raw Data from March 2014](https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data]. It includes the following variables: *cca* (Name of the Chicago Community Area), *stores* (number of grocery stores), $sqft$ (sum of store square footage), *income* (median income), *homevalue* (median home value), *pop* (population), *whitepop* (white population), $laborforce$ (labor force), *unemployed* (unemployed people), *black* (black population), and *acres* (area of the CCA).
- `childmortality` (N=64): Data on child mortality (*cm*, number of deaths of children under age 5 in a year per 1000 live births), female literacy rate ($flfp$, percent), per-capita GNP in 1980 (*pgnp*), and total fertility rate (tfr 1980–1985, the average number of children born to a woman, using age-specific fertility rates for a given year). The data was taken from Table 6.4 in Basic Econometrics (4th edition) by Damodar N. Gujarati.
- `coffee` (N=29): Historical data about the coffee consumption in the United States. The variable *consumption* is measured in thousand 60-kg bags. The variable *price* represents the retail price of roasted coffee in US Dollars per pound. Both time series are obtained from the International Coffee Association. The variables *rdi* and *cpi* represent real disposable income and the consumer price index, respectively. Those data series are obtained from the St. Louis FRED database. The same is true for the variable *population*. The variables *rprice* and *pcconsumption* represent the real coffee prices and the per capita consumption of coffee.
- `compactcars` (N=78): Fuel efficiency of compact cars in 1995 and 2015. The fuel efficiency is expressed in miles per gallon in the columns *automatic* and *manual*, respectively. The variable *displ* represents the displacement in liters.
- `covid`: Daily time series of COVID-19 *cases* and *death* (cumulative count). The variables *newcases* and *newdeaths* correspond to the additional cases each day. The number of observations is variable since it is updated on a daily basis. The data can be obtained from a New York Times GitHub account.

- `cps7885` (N=1084): Data from the Current Population Survey (CPS) for the year 1978 and 1985. The data accompanies the book *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. It contains the following variables: *educ* (years of schooling), *south* (1 if live in south), *nonwhite* (1 if nonwhite), *female* (1 if female), *married* (1 if married), *exper*, *age*, *union* (1 if belong to union), *wage*, and *year* (78 or 85).
- `crime` (N=99): Crime rates in North Carolina counties.
- `discharge` (N=7): Generic data set of pollutant discharge into a river measured in gallons.
- `eggweights` (N=61): Weight in grams of Large Grade A Brown Eggs from Whole Foods.
- `eucrime`: Intentional homicide data in Europe from Eurostat.
- `evdata` (N=579): Data about the choice of consumers with respect to alternative fuel vehicles. The variable *choice* represents the choice by the consumer for gasoline vehicles (1), conventional hybrids (2), plug-in hybrids (3), and electric vehicles (4). For each consumer, you have the following variables: *age*, *level*2 (indicating whether people have a fast charger for electric cars in their community), *female*, and *numcars* (number of cars). The variable *income* is coded as follows: Under $15,000 (1), $15,000 to $24,999 (2), $25,000 to $34, 999 (3), $35,000 to $49,999 (4), $50,000 to $74,999 (5), $75,000 to $99,999 (6), $100,000 to $149,999 (7), $150,000 to $199,999 (8), $200,000 to $249,000 (9), above $250,000 (10). The variable *edu* is coded as follows: Less than High School (1), High School / GED (2), Some College (3), 2-year College Degree (4), 4-year College Degree (5), Masters Degree (6), Doctoral Degree (7). The variable *politics* is coded as follows: Extremely Liberal (1), Liberal (2), Slightly liberal (3), Moderate (4), Slightly conservative (5), Conservative (6), Extremely conservative (7), Other (8), None (9). See Dumortier et al. (2014) for more details.
- `exemptorgs` (N=42135): Tax exempt organizations in Indiana (Source: IRS Exempt Organizations Business Master File Extract (EO BMF)) updated on 11 December 2023. The variables are *name*, *state*, *deductability* (1 meaning contributions are deductible, 2 meaning contributions are not deductible, and 4 meaning contributions are deductible by treaty, which may refer to foreign organizations), *organization* (1-Corporation, 2-Trust, 3-Cooperative, 4-Partnership, 5-Association). The variables *assets*, *income*, and *revenue* are measured in USD. The variable *ntee* refers to the National Taxonomy of Exempt Entities and may be missing for some older organizations. A full list is available here.
- `ez` (N=107): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. The variable *uclms* represents the unemployment claims in Anderson (Indiana). The variable *ez* indicates the presence of the enterprise zone in the city. An enterprise zone benefits from tax breaks and other public support to spur economic activity.
- `factor` (N=3199): The data is associated with the analysis in Lane et al. (2018). On a scale from 1–5, respondents had to rank the importance of various vehicle attributes.
- `fair` (N=601): The data comes from the analysis presented in the 1978 paper A Theory of Extramarital Affairs by Ray C. Fair. The variable $affairs$ representing the number of extramarital affairs in the past year. Everything of seven and above is coded as 7. The other variables are *male*, *yearsmarried* (number of years married), *children*, *religious* (religiousness on a scale of 1-5 with 1 being basically an atheist), and *marriagehappiness* (self-rating of marriage with 1=very unhappy to 5=very happy).
- `faithful` (N=272): Eruption and waiting in time minutes of Old Faithful Geyser. This data is also included in R.
- `fatalcity` (N=8131): Each row represents a car accident with at least one fatality according to the Fatality Analysis Reporting System (FARS). The accidents occurred in counties (identified by their $fips$ code) comprising a city. Except *year*, all other variables are dummy variables indicating characteristics associated with the accident.
- `fatalstate` (N=26893): Each row represents a car accident with at least one fatality according to the Fatality Analysis Reporting System (FARS). The accidents occurred in counties (identified by their $fips$ code) in various states. The variable *state* identifies the FIPS code of the state.

- `fertil1` (N=1129): Data from the General Social Survey. The data set is accompanying the book *Introductory Econometrics: A Modern Approach* by Jeffrey Wooldridge. The variables are as follows: *year* (72 to 84, even), *educ* (years of schooling), *meduc* and $feduc$ (mother's and father's education), *kids* (number children ever born), *east*, *northcentral*, and *west* (equal to 1 if lived in at 16), $farm$ (equal to 1 if on farm at 16), *otherrural* (equal to 1 if other rural at 16), *town* (equal to 1 if lived in town at 16), and *smallcity* (equal to 1 if in small city at 16).
- `fetransmission` (N=9297): Fuel efficiency of cars from 1984 and 2023. This data is similar to `compactcars` except for all years and all vehicle classes. The fuel efficiency is expressed in miles per gallon in the columns *automatic* and *manual*, respectively. The variable *displ* represents the displacement in liters.
- `fpdata` (N=176):
- `gqtestdata` (N=50): Generic (heteroscedastic) home value data of *price* and living area ($sqft$).
- `grunfeld` (N=200): This is a standard data set to introduce panel data models. It represents investment data from ten firms over the years from 1935 to 1954. The variables are $firm$, *year*, *inv* (gross investment), *value* (value of the firm), and *capital* (stock of plant and equipment).
- `gss` (N=2867): 2016 data from the General Social Survey with the variables that are mostly self-explanatory. The variables *wrkstat* and *cappun* refer to the Work status and the stance on the death penalty, respectively. The variables $facebook$ and *instagrm* indicates if the respondent uses the respective social media platform.
- `happy` (N=631): Data from the 2018 General Social Survey about happiness which includes the following variables: $sexfreq$: Frequency of sex during last year, *gun*: Have gun in home, *sclass*: Subjective class identification, *health*1: Condition of health, *happiness*: General happiness, *party*: Political party affiliation. You may want to combine the "Ind, near democrat" and "Not str democrat" into the same category, e.g., "lean democrat." Do the same for Republicans., *education*: Highest year of school completed, *age*: Age of respondent
- `hdi` (N=189): Human Development Indicator from 2019.
- `heating` (N=900): R package mlogit. There are five types of heating systems: gas central (*gc*), gas room (*gr*), electric central (*ec*), electric room (*er*), and heat pump (*hp*). There are also the following variables: *idcase* gives the observation number (1-900), *depvar* identifies the chosen alternative (gc, gr, ec, er, hp), *ic.alt* is the installation cost for the 5 alternatives, *oc.alt* is the annual operating cost for the 5 alternatives, *income* is the annual income of the household, *agehed* is the age of the household head, *rooms* is the number of rooms in the house, *region* a factor with levels *ncostl* (northern coastal region), *scostl* (southern coastal region), *mountn* (mountain region), *valley* (central valley region).
- `henning` (N=194): The data set `henning` is similar to `rossi` and comes from the paper Cognitive-behavioral treatment of incarcerated offenders: An evaluation of the Vermont Department of Corrections' Cognitive Self-Change Program by Henning and Frueh (1996). The following variables will be used in this analysis: *months* (day of re-arrest measured in months), censor (dummy variable), personal (offense against a person), property (offense against property), and cage (centered age at time of release, i.e., age minus average age).
- `hhpub` (N=129696): The data contains household data from the 2017 National Household Travel Survey (NHTS). Some columns were deleted but the detailed code book can be found here. The variables included are: *homeown* (home ownership), *hhsize* (count of household members), *hhvehcnt* (count of household vehicles), $hhfaminc$ (household income), *bike* (frequency of bicycle use for travel), *price* (price of gasoline affects travel), *place* (travel is financial burden), *ptrans* (public transporation to reduce financial burden of travel), *hhstate* (household state), *urbrur* (household in urban/rural area), and *hbppopdn* (category of population density (persons per square mile) in the census block group of the household's home location.
- `honda` (N=81): Prices and miles of used Honda Accords in the Indianapolis area.

- `housing`: Data from FRED
- `housing1` (N=88): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. The variables are the following: *price* (house price in USD 1000), *assess* (assessed value in USD 1000), *bdrms* (number of bedrooms), *lotsize* (in square feet), $sqrft$ (home size in square feet), and *colonial* (equal to 1 if house is in colonial style).
- `hprice2` (N=506): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. The variables are the following: *price* (median housing price in USD), *crime* (crimes committed per capita), *nox* (nitrogen oxide concentration), *rooms* (number of rooms), *dist* (weighted distance to five employment centers), *radial* (accessibility index highways), *proptax* (property tax per USD 1000), *stratio* (average student-teacher ratio), and *lowstat* (percent of people of lower socioeconomic status).
- `hybrid` (N=1000): This is a generic data set with the following variables: y (equal to 1 if person purchased a hybrid car), *gas* price of gasoline at the time of purchase, *increment* (price difference compared to gasoline car), *college* as dummy variable indicating if person has a college degree, and *env* as a dummy variable indicating whether person is associated with an environmental group.
- `indyheating` (N=24): Contains 24 observations of natural gas usage (in hundreds of cubic feet) and the average outside temperature (in Fahrenheit).
- `indyhomes` (N=102): This data set contains home characteristics from two ZIP codes in Indianapolis.
- `jcars` (N=43): The data is taken from *Forecasting: Methods and Applications* by Makridakis, Wheelwright, and Hyndman (1998). It contains Japanese motor vehicle production in thousand (1947–1989).
- `kiel` (N=321): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. It includes the following variables: *year* (1978 or 1981), *age* (age of house), *nbh* (neighborhood, 1-6), *cbd* (distance to central business district in feet), *intst* (distance to interstate in feet), *price* (selling price), *rooms* (number of rooms in house), *area* (square footage of house), *land* (square footage lot), *baths* (number of bathrooms), *dist* (distance from house to incinerator in feet), *wind* (percent time wind from incinerator to house), $y81$ (equal to 1 if year is 1981), *nearinc* (equal 1 if distance is less than 15840 feet), *rprice* (real price in 1978 dollars).
- `lung` (N=228): The data is associated with the R package survival and includes the survival of patients with lung cancer. The variables included are the following: *inst* (institution code), *time* (survival time in days), *phecog* (ECOG performance score as rated by the physician: 0=asymptomatic, 1=symptomatic but completely ambulatory, 2=in bed less than 50% of the day, 3=in bed more than 50% of the day but not bed bound, 4 = bed bound), *phkarno* (Karnofsky performance score from 0=bad to 100=good as rated by physician), *patkarno* (Karnofsky performance score as rated by the patient), *mealcal* (calories consumed at meals) and *wtloss* (weight loss in last six months in pounds).
- `meatdemand` (N=39): Meet demand quantity (q) and real price (p) data from the U.S. Department of Agriculture. Prices and real disposable income (*rdi*) are in real terms.
- `mh1` (N=101): Home values in the Meridian Hills area in Indianapolis.
- `mh2` (N=18): Prices and characteristics of homes in the Meridian Hills area in Indianapolis.
- `milk` (N=50): Randomly generated milk container fillings in ounces.
- `mpa` (N=18): Randomly generated exam scores in a class of MPA students.
- `nfl` (N=1009): Data from Berri et al. (2011) which includes the following variables: *total* (total salary), *yards* (passing yards from the prior season), *att* (pass attempts), *exp* (total years of experience in the league), $exp2$ (total years of experience in the league squared), $draft1$ (first round draft pick), $draft2$ (second round draft pick), *veteran* (bargaining status changes after a player has completed three years in the NFL), *changeteam* (player has changed team), *pbowlever* (player appeared in the Pro Bowl), and *symm* (facial symmetry).

- `ohioincome` (N=607): Enrollment and median income in Ohio School districts for 2018/2019. *IRN* identifies the school district.
- `ohioscore` (N=608): Performance and achievement scores of Ohio schools for 2018/2019. The data is obtained from the Ohio Department of Education. *IRN* identifies the school district.
- `organic` (N=100): Randomly generated data on the purchase behavior of organic food (binary choice) as a function of income.
- `quakes` (N=99): Time series of the annual number of earthquakes in the world with seismic magnitude over 7.0, for 99 consecutive years. The data is taken from Penn State STAT 510 Applied Time Series Analysis.
- `quillayute` (N=20116): Weather data from Quillayute Airport in Washington State. Weather station USW00094240. The variables are *date*, *temperature* (average between maximum and minimum temperature in degree Fahrenheit), and *month* (month of the year).
- `retail` (N=348): Advance Retail Sales: Retail (Excluding Food Services) obtained from the St. Louis FRED.
- `rossi` (N=432): Experimental recidivism study on 432 male prisoners over a period of one year after release from prison (Rossi et al., 1980). The following variables are included: *week* (week of first arrest after release, or censoring time), *arrest* (the event indicator, equal to 1 for those arrested during the period of the study and 0 for those who were not arrested), fin (a factor, with levels yes if the individual received financial aid after release from prison, and no if he did not; financial aid was a randomly assigned factor manipulated by the researchers), *age* (in years at the time of release), *race* (a factor with levels black and other), *wexp* (a factor with levels yes if the individual had full-time work experience prior to incarceration and no if he did not), *mar* (a factor with levels married if the individual was married at the time of release and not married if he was not), *paro* (a factor coded yes if the individual was released on parole and no if he was not), *prio* (number of prior convictions), *educ* (education, a categorical variable coded numerically, with codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6 (some post-secondary)), and $emp1$ to $emp52$ (factors coded yes if the individual was employed in the corresponding week of the study and no otherwise).
- `skewness` (N=500): Randomly generated data from a beta distribution with various parameters. $beta55$ ($a = 5$ and $b = 5$), $beta28$ ($a = 2$ and $b = 8$), and $beta82$ ($a = 8$ and $b = 2$).
- `soda` (N=25): Randomly generated data of for soda cans (in milliliters).
- `souvenirs` (N=84): The data is taken from *Forecasting: Methods and Applications* by Makridakis, Wheelwright, and Hyndman (1998). It contains monthly sales for a souvenir shop on the wharf at a beach resort town in Queensland, Australia. The data is also contained in the R package fma as `fancy`.
- `statefinhealth` (N=51): Data taken from the Urban Institute's Financial Health and Wealth Dashboard 2022. The data includes the following variables: *state*, *delinquentdebt* (residents with delinquent debt), *emergencysavings* (Households with at least USD 2000 in emergency savings), *mediancreditscore*, $mortgageforeclosure$ (mortgage holders in foreclosure), and *mediannetworth*.
- `states` (N=10): Generic income data for three states.
- `traffic` (N=108): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. It contains the following variables: *year*, *month*, *totacc* (statewide total accidents), t (time trend), *unem* (state unemployment rate), *spdlaw* (equal to 1 after 65 mph in effect), *bltlaw* (equal to 1 after seatbelt law), and *wkends* (number of weekends in month).
- `truncation` (n=50): Generic truncated data for which the researcher only observes *yobs* as the independent variable and x as the dependent variable. *yreal* are the (partially unobserved) values of the dependent variable.
- `usdata` (N=307): Data from the St. Louis FRED data base of real disposable personal income per capita (*income*) and real personal consumption expenditures per capita (*consumption*).

- `vehicles` (N=45885): The data set contains data about the fuel efficiency of vehicles from model year 1984 to 2021 (*year*). It also includes *make* and *model* of the various manufacturers. The variable *displ* refers to engine displacement in liters, *vclass* is the vehicle class, *co*2*tailpipegpm* indicates the greenhouse gas (GHG) emissions, and *comb*08 is combined MPG. The other variables are self-explanatory. The data is obtained from the U.S. Environmental Protection Agency' (EPA) Fuel Economy Data webpage. The full data set contains many more variables which are well documented here.
- `vehpub` (N=256115): The data contains vehicle data from the 2017 National Household Travel Survey (NHTS). Some columns were deleted but the detailed code book can be found here. The variables included are: *houseid* (household identifier), *vehid* (vehicle identifier), *vehyear* (vehicle year), *make* (vehicle make), *fueltype* (fuel type), *vehtype* (vehicle type), *odread* (odometer reading), *hfuel* (type of hybrid vehicle), *hybrid* (hybrid vehicle), *homeown* (home ownership), *hhfaminc* (household income), *hhstate* (household state), *urbrur* (household in urban/rural area).
- `wage` (N=526): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. It contains the following variables: *income* (average hourly earnings), *educ* (years of education), and *exper* (years potential experience).
- `wage2` (N=935): This data is taken from *Introductory Econometrics: A Modern Approach* by Jeffrey M. Wooldridge. It contains the following variables: *wage* (monthly earnings), *hours* (average weekly hours), *kww* knowledge of world work score, *educ* (years of education), *exper* (years of work experience), *tenure* (years with current employer), *age* (age in years), *married* (equal to 1 if married), *black* (equal to 1 if black), *south* (equal to 1 if live in south), *urban* (equal to 1 if live in SMSA), *sibs* (number of siblings), *brthord* (birth order), *meduc* (mother's education), *feduc* (father's education).
- `waterpressure` (N=30): Randomly generated data on the water pressure in a city's water lines.
- `wdi` (N=13671): World Development Indicators (WDI) from the World Bank. The following variables are included in the data set: Country code (*iso*2*c*), *country*, *year*, gross domestic product per capita in constant 2010 USD (*gdp*), life expectancy at birth in years (*lifeexp*), adult literacy rate (percent of people ages 15 and above) (*litrate*), fertility rate (births per woman) (*fertrate*), mortality rate under age of five (per 1,000 live births) (*mortrate*), geographic region (*region*), and income groupings based on gross national income per capita (*incomeLevel*).
- `windsolar` (N=648): The data is taken from the paper The effect of the feed-in-system policy on renewable energy investments: Evidence from the EU countries by Alolo et al. (2020). The data contains the following variables: *country*, *year*, added wind capacity in Megawatts (*agwind*), presence of either a feed-in-premium (FIP) or a feed-in-tariff (FIT) (*fiswind*), presence of a FIT (*fitwind*), presence of a FIP (*fipwind*), existence of a quota trade system (*quota*), existence of tax benefits (*tax*), existence of a *tender* mechanism, existence of an electricity price *cap*, gross domestic product (*gdp*), per capita *electricityconsumption*, percent of electricity from various sources (*coalshare*, *renewableshare*, *oilshare*, *nuclearshare*, and *naturalgasshare*), population growth (*popgrowth*), per capita carbon emissions *co*2, prices of various energy sources (*oilprice*, *naturalgasprice*, and *coalprice*), energy dependence defined as energy imports divided by total energy consumption (*energyimport*).