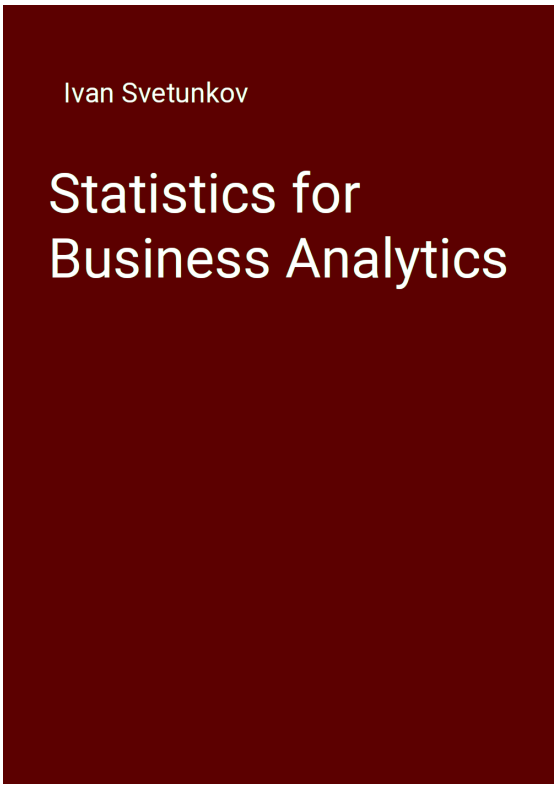
Statistics for Business Analytics

Ivan Svetunkov

2025-03-14

Contents

Preface



Have you encountered the term "Business Analytics" in your life? If not, then you are probably wondering what it means. If yes, then again, you are probably wondering what it means. It is a term that is used nowadays instead of such terms as "Operations Research" and "Management Science". It is a discipline that covers a variety of qualitative and quantitative methods that can be used in practice for real life decisions. It uses methods and approaches from different scientific areas, including statistics, forecasting, optimisation, operations management etc. These lecture notes are focused on the core of quantitative side of the discipline - statistics. While there are many books on statistics, the author failed to find

one that would be focused on the application of statistics both for analysis and forecasting and would rely on modern statistical approaches.

These lecture notes relies heavily on the `greybox` package for R, which focuses on forecasting using regression models. In order to run examples from the lecture notes, you would need to install this package (Svetunkov, 2025):

```
install.packages("greybox")
```

A very important thing to note is that these lecture notes **do not use `tidyverse` packages**. I like base R, and, to be honest, I am sure that `tidyverse` packages are great, but I have never needed them in my research. So, I will not use pipeline operators, `tibble` or `tsibble` objects and `ggplot2`. It is assumed throughout the lecture notes that you can do all those nice tricks on your own if you want to.

If you want to get in touch with me, there are lots of ways to do that: comments section on any page of my website, my Russian website, vk.com, Facebook, Linkedin, Twitter.

You can also find me on ResearchGate, StackExchange and StackOverflow, although I'm not really active there. Finally, I also have GitHub account.

You can use the following to cite the online version of this book:

- Svetunkov, I. (2022) Statistics for Business Analytics: Lancaster, UK. openforecast.org/sba. Accessed on [current date].

If you use LaTeX, the following can be used instead:

```
@MISC{SvetunkovSBA,
    title = {Statistics for Business Analytics},
    author = {Ivan Svetunkov},
    howpublished = {Lecture notes. OpenForecast},
    note = {(version: [current date])},
    url = {https://openforecast.org/sba/},
    year = {2022}
}
```

License

Chapter 1

Introduction

Whenever we talk about statistics, analytics or forecasting, we deal with models that are constructed based on available information. So, it is important to understand what a model is, what types of models exist and how to measure information in order to use it in models afterwards. This is what we will discuss in this chapter.

1.1 What is model?

Reality is complex. Everything is connected with everything, and it is difficult to isolate an object or a phenomenon from the other objects or phenomena and their environment. Furthermore, due to this complexity of reality, we cannot work with itself, we need to simplify it, leave only the most important parts and analyse them. This process of simplification implies that we create a model of reality, work with it and make conclusions about the reality based on it.

Pidd (2010) defines **model** as "*an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality*". Let us analyse this definition.

- It is *external and explicit*, because if you only think about something, it is not a model. The unclear view on an object is not a model, it needs to be formulated.
- It is a *representation of part of reality*, because it is not possible to represent the reality at full - it is too complex, as discussed above.
- Technically speaking, the model without a purpose is still a model, but Pidd (2010) points out in his definition that without it, the model becomes useless (thus "*to understand, to change, to manage and to control*").
- Finally, *as seen by people* is an important element that shows that models are always subjective. One and the same question can be answered with different models based on preferences of analyst.

This definition is wide and covers different types of models, starting from simple graphical ones and ending with complex imitations. In fact, there are four fundamental types of models (they are ordered by the increase of complexity):

1. Textual,
2. Visual,
3. Mathematical,
4. Imitation.

Textual model is just a description of an object or a process. An instruction of how to assemble a chair is an example of a textual model. Any classification will be a textual model as well, so the list of four types of models is a textual model.

Visual model is a graphical or a schematic representation of an object or a process. An example of such a model is provided in Figure 1.1.

Figure 1.1: Chair assembly instruction. Found on Reddit.

Mathematical model is a model that is represented using equations. It is more complex than the previous two, because it requires an understanding of mathematics. At the same time it can be more precise than the previous two models in terms of capturing the structure of reality and making predictions about it.

The mass-energy equivalence equation is an example of such model:

$$E = mc^2.$$

A mathematical model in turn can be either deterministic or stochastic. The former one assumes that there is no randomness in it, while the latter implies that the randomness exists and can be modelled in one way or another. The related classification of models based on the amount of randomness is:

a. White box - the deterministic model. An example of such model is a linear programming, which assumes that there is no randomness in the data;
b. Grey box - the model that assumes some randomness, but for which the structure is known (or assumed). Any statistical model can be considered as a grey box: typically, we have an understanding of how elements in it interact with each other and how the result is obtained;
c. Black box - the model with randomness, for which we do not know what is happening inside. An example of such model is an artificial neural network.

Finally, the *imitation* model is a simplified reproduction of a real object or a process. This can be, for example, a physical model of a building standing in a room of an architect, or a mental arrangement in psychology.

In these lecture notes, we will deal with only first three types of models, focusing on the third one.

When constructing mathematical models, we will inevitably deal with variables, with factors that are potentially related to each other and reflect some aspects of the real object or phenomenon. These variables can be split into two categories:

1. Input, or external, or exogenous, or explanatory variables - those that are provided to us and are assumed to impact the variable (or several variables) of interest;
2. Output, or internal, or endogenous, or response variable(s) - the variable of the main interest, which is assumed to be related by a set of explanatory variables.

The models that have only one response variable are called "*univariate*" models. But in some cases, we might have several response variables (for example, sales of several similar products). We would then deal with a *multivariate* model. In the literature, you might meet a different definition of univariate / multivariate models. For example, some consider a model with several variables multivariate, even if it has only one response and several explanatory ones. But throughout these lecture notes we use the definitions above, focused on response variable.

1.1.1 Models, methods et al.

There are several other definitions that will be useful throughout these lecture notes:

- **Statistical model** (or 'stochastic model', or just 'model' in these lecture notes) is a 'mathematical representation of a real phenomenon with a complete specification of distribution and parameters' (Svetunkov and Boylan, 2019). Very roughly, the statistical model is something that contains a structure (defined by its parameters) and a noise that follows some distribution.
- **True model** is the idealistic statistical model that is correctly specified (has all the necessary components in correct form), and applied to the data in population. By this definition, true model is never reachable in reality, but it is achievable in theory if for some reason we know what components and variables and in what form should be in the model, and have all the data in the world. The notion itself is important when discussing how far the model that we use is from the true one.
- **Estimated model** (aka 'applied model' or 'used model') is the statistical model that was constructed and estimated on the available sample of data. This typically differs from the true model, because the latter is not known. Even if the specification of the true model is known for some reason, the parameters of the estimated model will differ from the true parameters due to sampling randomness, but will hopefully converge to the true ones if the sample size increases.
- **Data generating process** (DGP) is an artificial statistical model, showing how the data could be generated in theory. This notion is utopic and can be used in simulation experiments in order to check, how the selected model with the specific estimator behave in a specific setting. In real life, the data is not generated from any process, but is usually based on complex interactions between different agents in a dynamic environment. Note that I make a distinction between DGP and true model, because I do not think that the idea of something being generated using a mathematical formula is helpful. Many statisticians will not agree with me on this distinction.
- **Forecasting method** is a mathematical procedure that generates point and / or interval forecasts, with or without a statistical model (Svetunkov and Boylan, 2019). Very roughly, forecasting method is just a way of producing forecasts that does not explain how the components of time series interact with each other. It might be needed in order to filter out the noise and extrapolate the structure.

Mathematically in the simplest case the true model can be presented in the form:

$$y_t = \mu_{y,t} + \epsilon_t, \tag{1.1}$$

where y_t is the actual observation, $\mu_{y,t}$ is the structure in the data and ϵ_t is the noise with zero mean, unpredictable element, which arises because of the effect of a lot of small factors and t is the time index. An example would be the daily sales of beer in a pub, which has some seasonality (we see growth in sales every weekends), some other elements of structure plus the white noise (I might go to a different pub, reducing the sales of beer by one pint). So what we typically want to do in forecasting is to capture the structure and also represent the noise

with a distribution with some parameters.

When it comes to applying the chosen model to the data, it can be presented as:

$$y_t = \hat{\mu}_{y,t} + e_t, \tag{1.2}$$

where $\hat{\mu}_{y,t}$ is the estimate of the structure and e_t is the estimate of the white noise (also known as "**residuals**"). As you see even if the structure is correctly captured, the main difference between (1.1) and (1.2) is that the latter is estimated on a sample, so we can only approximate the true structure with some degree of precision.

If we generate the data from the model (1.1), then we can talk about the DGP, keeping in mind that we are talking about an artificial experiment, for which we know the true model and the parameters. This can be useful if we want to see how different models and estimators behave in different conditions.

The simplest forecasting method can be represented with the equation:

$$\hat{y}_t = \hat{\mu}_{y,t}, \tag{1.3}$$

where $\hat{y}_t$ is the point forecast. This equation does not explain where the structure and the noise come from, it just shows the way of producing point forecasts.

In addition, we will discuss in these lecture notes two types of models:

1. Additive, where (most) components are added to one another;
2. Multiplicative, where the components are multiplied.

(1.1) is an example of additive error model. A general example of multiplicative error model is:

$$y_t = \mu_{y,t}\varepsilon_t, \tag{1.4}$$

where ε_t is some noise again, which in the reasonable cases should take only positive values and have mean of one. We will discuss this type of model later in the textbook. We will also see several examples of statistical models, forecasting methods, DGPs and other notions and discuss how they relate to each other.

Remark. Throughout these lecture notes we will use index t to denote the time and index j to denote the cross-sectional elements. So, for example, y_t will mean the response variable changing over time, while y_j will mean the response variable changing from one object to another (for instance, from one person to another).

1.2 Scales of information

Whenever we work with information, we need to understand how to measure it properly. If we cannot do that, then we cannot construct any model and make proper decisions, supported by evidence. For example, if a person feels ill but we cannot say what the temperature of their body is, then we cannot decide,

whether anything needs to be done to reduce it. If we can measure something then we can model it and produce forecasts. Continuing our example, if the temperature is 39°C, then we can conclude that the person is sick and needs to take Paracetamol or some other pills that would reduce the temperature. So, whenever we collect some sort of information about a system's behaviour or about a process, we will inevitably deal with scales of information and it is important to understand what we are dealing with in order to process that information correctly. There are four fundamental scales:

1. Nominal,
2. Ordinal,
3. Interval,
4. Ratio.

The first two form the so called "categorical" scale, while the latter two are typically united in the "numerical" one. Each one of these scales can have one of the following characteristics:

1. Description,
2. Order,
3. Distance,
4. Natural zero,
5. Natural unit.

The last characteristics is typically ignored analytics and forecasting as it does not provide any useful information. But as for the other four, they provide important properties to the scales of information, giving them more flexibility. Here we discuss the scales in detail.

1.2.1 Nominal scale

This is the scale that only has "description" characteristics. It does not have an order, a distance or a natural zero. There is only one operation that can be done in this scale, and it is comparison, whether the value is "equal" or "not equal" to something. An example of data measured in such scale is the following question in a survey:

What is your nationality?

- Russian,
- English,
- Greek,
- Swiss,
- Belgian,
- Lebanese,
- Indonesian,
- Other.

In this case after collecting the data we can only say whether each respondent is

Russian or not, English or not etc. So, the only thing that can be done with the data measured in this scale is to produce a basic summary, showing how many people selected one option or another. Among the statistical instruments, **only the mode** is useful, as it shows which of the options was selected the most. If there are several variables measured in nominal scale, we can calculate some measures of association to see if there are any patterns in respondents behaviour (e.g. those who select "Russian" would prefer Vodka, while those who selected "Belgian" will tend to drink "Beer").

When it comes to constructing models, the nominal scale is typically transformed in a set of dummy variables, which will be discussed later in regression analysis of these lecture notes.

If you are not sure, whether your data is measured in nominal or another scale, you can do a simple test: if changing the places of two values does not break the scale, then this is the nominal one. For example, in the question above, moving "Greek" to the first place will not change anything, so this is indeed the nominal scale. Another example of nominal scale is the number on the T-shirt of football players. They are only descriptive, and if two players change numbers, this will not change anything (although it might confuse football fans).

1.2.2 Ordinal scale

In addition to description, the ordinal scale has the "order". It is possible to say that one value can be placed higher or lower than the other on a scale (thus, permitting operations "greater" and "smaller" in addition to the "equal" and "not equal") However, it is not possible to say how far the elements are placed from each other, so the number of operations in the scale is still limited. Here is an example of a survey question with such scale:

How old are you?

1. Too young,
2. Young,
3. Not too young,
4. Not too old
5. Old,
6. Too old.

In this scale above we have a natural order, and when collecting the data in this scale we can conclude, whether a respondent is older than another one or not. Sometimes ordinal scales look confusing and seem to be of a higher level than they are, here is an example:

How old are you?

1. Younger than 16,
2. 16 - 25,
3. 26 - 40,

4. 41 - 60
5. Older than 60.

This is still an ordinal scale, because it has the natural order, and because we cannot measure the distance between the value: if, for example we subtract "16 - 25" from "26 - 40", we will not get anything meaningful.

The ordinal scale, being more complex than the nominal one, allows using some additional statistical instruments (besides the mode), such as *quantiles of distribution, including median*. Unfortunately, the **arithmetic mean is not applicable to the data in ordinal scale**, because of the absence of distance. Even if you encode every answer in numbers, the resulting average will not be meaningful. Indeed, if in the question above with the five options, we use the numbers ("1" for the first option, "2" for the second one, etc.) and take average, the resulting number of, for example, 3.75 will not mean anything, as there is no element in the scale that would correspond to that number.

When it comes to measuring relations between two variables in ordinal scale, we can use Kendall's τ correlation coefficient, Goodman-Kruskal's γ and Yule's Q. These are discussed in detail in Section 9. As for using the variables in ordinal scale in modelling, the typical thing to do would be to create a set of dummy variables, similarly to how it is done for variables in nominals scale.

As for the identification of scale, if in doubt, you can do any transformation of elements of scale without the loss of its meaning. For example, if we assign numbers from 1 to 5 to the responses above, we can square each one of them and get 1, 4, 9, 16 and 25, which would not change the original scale, but only encode the answers differently (select "16" for the option "41 - 60").

1.2.3 Interval scale

This scale is even more complex than the previous two, as in addition to description and order it also has a distance. This permits doing addition and subtraction to the elements of scale, which are meaningful operations in this case. Arithmetic mean and standard deviation become available in this scale in addition to all those used in lower level scales discussed above. The classical example of a variable measured in this scale is the temperature. Indeed, we can not only say if the temperature of one person is higher than the temperature of the other one, but also by how much: 39°C - 37°C = 2°C, which is a meaningful number in the scale. The only limitation in this scale is that there is no natural zero. 0°C does not mean the absence of temperature, but rather means the point at which water starts freezing. If we switch to Fahrenheit (although why would anyone do that?!), then the 0°F would correspond to the point, where the mixture of ice, water, and ammonium chloride used to stabilise back in 1724, when Fahrenheit proposed the scale.

Almost all descriptive statistics are meaningful in this scale (see Section 5). This includes, but is not limited with **mean, variance, skewness, kurtosis**.

Coefficient of variation and other statistics, where division is done by an actual value or mean, are not meaningful, because the scale does not have a meaningful zero. For example, switching from Celsius to Fahrenheit would change the value of coefficient of variation, although the distribution of the variable will stay the same. Furthermore, some error measures (see Chapter 2 of Svetunkov, 2021) cannot be used for the measurement of accuracy of models for this scale (for example, MAPE cannot be used as it assumes meaningful zero).

The relation between two variables in interval scale can be measured by Pearson's correlation coefficient. The scale can be used in the model as is.

Finally, when it comes to the identification of scale, only linear transformations are permitted for the variables without the loss of its properties. This means that if we measure temperature of two respondents and then do their linear transformations via $y = a + bx$, then the characteristics of scale will not be broken: it will still have description, order and distance with the same meaning as prior to the transformation. In the example of temperature, this is how you switch, for example, from Celsius to Fahrenheit ($y = a + bx$).

1.2.4 Ratio scale

The most complex of the four, this ratio has a natural zero (in addition to all the other characteristics). It permits any operations to the values of scale, including product, division, and non-linear transformations. Coefficient of variation can be used together in addition to all the previous instruments. An example of the information measured in this scale is the height of respondents in meters. You can compare two respondents via their height and not only say that one is higher than the other, but also by how much and how many times. All these operations will be meaningful in this scale.

If you need to check, whether the variables is indeed in ratio scale, note that only the transformation via multiplication would maintain the meaning of the scale. For example, height measured in meters can be transformed into height in feet via the multiplication by approximately 3.28. If you add a constant to the values of scale, it will break it.

All the statistics and error metrics work for the variables measured in this scale. Furthermore, being the most complex, this scale also permits usage of all correlation coefficients.

Finally, the variables measured in this scale can be either integer or continuous. This might cause some confusions, because the integer numbers sometimes look suspiciously similar to the values of ordinal scale, but the tools of identification discussed above might help. If a company needs to buy 7 planes, then this is an integer variable measured in ratio scale: 7 planes is more than 6 planes by one plane, and zero planes means that there are no planes (all the characteristics of ratio scale). Furthermore, squaring the number of planes breaks the distance between them ($7^2 - 6^2 \neq 1^2$), while linear transformation breaks the scale

($7 \times 2 + 3$ has a completely different meaning in the scale than just 7).

1.3 Types of data

After doing measurement (e.g. measuring temperature of patients), an analyst typically obtains data. According to Cambridge dictionary, **data is "information, especially facts or numbers, collected to be examined and considered and used to help decision-making"** (Cam, ????). Data can be *unprocessed* (or raw data), containing an disorganised array of recordings about a process under consideration, or it can be *cleaned*, having a proper structure with correctly encoded variables, without obvious mistakes or missing values. When first dealing with data, an analyst needs to clean it and transform it into a usable format in order to be able to extract useful information from it or apply models.

The process of data cleaning itself depends on a variety of factors, and most importantly on what the analyst wants to do. For example, if an analyst is interested in getting insights about daily A&E attendance, and they have records of every patient arriving to A&E, the first step would be to aggregate those values into daily buckets, obtaining the information about the number of A&E arrivals per day. As another example, if an analyst conducts a survey, aiming to find brand preferences of citizens of Lancaster, then after collecting the survey data, they need to transform some of the questions into appropriate variables to be able to work with them (e.g. if there were multiple choice questions with several options, they need to be transformed into a set of variables equal to the number of that options).

The question related to the data cleaning is what type of data the analyst is working with. There are three fundamental types:

1. Cross-sectional data;
2. Time series;
3. Panel data.

In the example above the **cross-sectional data** is the data collected after conducting a survey in a specific location over a fixed period of time. We end up with answers to the questions (and thus values of variables) of different respondents at the fixed time. Mathematically, we denote these observations with index j, which separates, for example, one respondent from another, and in this textbook we will use letter n to denote the number of elements (respondents) in our sample. So, the variable y_j would mean the value of a variable for the j^{th} respondent.

The **time series data** is typically measured for one and the same object over time. In the example above, the A&E arrival would imply time series data, where we observe a value (number of patients arriving) over time (daily). Mathematically, we denote the observation over time with index t, separating, for example, one

day from another. The last available observation in this case will be denoted with the capital letter T. In our notations, the variable y_t contains the value at a specific moment of time.

Remark. In time series, the observations typically do not happen at random, the number of A&E arrivals will depend on the time of day, day of week and month of year. This is an important characteristic of this specific type of data, and we will come back to it later in this textbook.

Finally, in some situations we might be able to measure data of several objects over time. For example, we could have daily A&E arrival in several hospitals. This type of data would be called **panel data**, and in this situation we would use both indices j and t, ending up with a variable $y_{j,t}$, showing, for example, a specific number of patients arriving to a specific hospital at a specific moment of time.

In this textbook we will focus on cross-sectional data and then will move to the time series one. We will also briefly discuss panel data models, but we do not discuss them in detail, as they become available to analysts less often than the other two types.

We have already used the term "variable" several times in this chapter, assuming that a reader is familiar with it. In mathematics, **variable is a symbol that represents any of a set of potential values**. In this textbook, we will face several types of variables. We will work with a **response** variable, representing a place holder for a quantity of the main interest of our analysis, something that is formed using an assumed mechanism. This will be denoted with letter y. We will also work with **explanatory** variables, which are supposed to explain how the response variable is formed and are denoted using letter x with potential subscripts, e.g. x_1, x_2 etc, representing a first, a second etc variables. In some cases, we will also use terms "exogenous" and "endogenous" variables, where the former means the variable that is formed on its own and is not impacted by any of variables under consideration, while the latter represents a variable that is created by a combination of variables under consideration. Sometimes, the terms "response" and "endogenous" are used as synonyms. Similarly, the terms "explanatory" and "exogenous" are used as synonyms as well. The basic model with one response variable y_j and one explanatory variable x_j can be written as:

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

where β_0 and β_1 are parameters of the model. This model is discussed in more detail in Chapter 10.

1.4 Sources of uncertainty

When estimating any model on a sample of data, we will inevitably have to deal with uncertainty. Consider an example, when we want to estimate the average height of a person in the room. We could take heights of all the people

in the room, then take average and we would get our answer. But what would happen with that average if another person comes in the room? We would need to do additional measures and re-estimate the average, and inevitably it will be different from the one we had before. This example demonstrates one of the classical sources of uncertainty - the one caused by estimation on a sample of data.

Furthermore, we might be interested in predicting the weight of a person based on their height. The two variables will be related, but would not have a functional relation: with the increase of height we expect that a person will weigh more, but this only holds on average. So, based on a sample of data, we could estimate the relation between the two variables and then having a height of a person, we could predict their expected weight. Their individual weight will inevitably vary from one person to another. This is the second source of uncertainty, appearing because of the individual discrepancies from one person to another.

Finally, the model of weight from height could be wrong for different reasons. For example, there might be plenty of other factors that would impact the weight of person that we have not taken into account. In fact, we never know the true model (see Section 1.1.1), so this is the third source of uncertainty, the one around the model form.

These three sources of uncertainty have been summarised for the first time in Chatfield (1996). Whenever we need to construct any type of model, we will deal with:

1. Uncertainty about the data, e.g. the error term ϵ_t (see Section 6);
2. Uncertainty about estimates of parameters (see Section 12);
3. Uncertainty about the model form (see Section 17).

In these lecture notes we will discuss all of them, slowly moving from (1) to (3), introducing more advanced techniques for model building.

Chapter 2

Probability theory

Before moving to the discussion of statistics, we need to understand the basics of probability theory. In this chapter we will discuss the definition of probability, the definition of random variable then addition and multiplication of probabilities, conditional and independent probabilities and finally the Bayes' Theorem. All these theoretical ideas form the basis of more advanced statistical tools, which is why they are important.

2.1 What is probability?

We start with a classical example: tossing a coin. If you have one, take it in your hands, look at it, and answer a question: what outcome will you have if you toss it? Toss it once and, let's say, it ended up showing heads. Can you predict the outcome of the next toss based on this observation? What if you toss it again and end up with tails? Would that change your prediction for the next toss?

What we could do in this situation to predict future outcomes is to write down the results of tosses as zeroes (for heads) and ones (for tails). We will then have a set of observations of a style:

1 0 1 0 0 1 1 0 1 1

If we then take the mean of this series, we will see that the expected outcome based on our sample is 0.6. We would call this value the **empirical probability**. It shows us that roughly in 50% of the cases in our sample we get tails. But this is based on just 10 experiments. If we continue tossing the coin for many more times, this probability (in the case of a fair coin) will eventually converge to 0.5, meaning that in the 50% of the cases the coin will show heads and in the other 50% it will be tails. In fact, we know that there are only two possible outcomes in this experiment, and that in case of a fair coin, there are no specific forces that could change the outcome and lead to more tails than heads. In this case, we

can say that the **theoretical probability** of having tails is 0.5. Note that this does not tell us anything about each specific outcome, but only demonstrates what happens on average, when we repeat the experiment many times.

Definition 2.1. Probability is the measure of how likely an event is expected to occur if we observe it many times.

This definition implies that we cannot tell what the next outcome of the experiment will be (whether the coin toss will result in heads or tails). Instead, we can say what will happen on average if the experiment is repeated many times. By definition, the probability lies between 0 and 1, where 0 means that the event will not occur and 1 implies that it will always occur.

Remark. When interpreting the probability, we can never say whether the event will happen or not unless the probability equals exactly to 0 or 1. For example, if the probability of rain today is 0.05, this does not mean that we will not have rain today. It only means that if this day repeats many times, in 5% of them it will rain. And there is no guarantee that today we will have one of those 95% sunny days.

We could do other similar experiments, for example rolling a six-sided dice, and calculating the probability of a specific outcome. In the simple cases with coins, cards, dices etc, we can even tell the probability without running the experiments. All we need to do is calculate the number of outcomes of interest and divide it by the sum of all the possible outcomes. For example, the probability of getting 3 on a 6-sided dice is $\frac{1}{6}$, because there are overall six outcomes: 1, 2, 3, 4, 5 and 6, and the probability of getting any one of them is the same for all of them (if the dice is fair). The probability of getting 5 is $\frac{1}{6}$ as well for the same reason: all the six outcomes are considered equally possible and will happen **on average** every sixth roll.

Remark. In some tabletop games, the number of dices and their outcomes are encoded as adb, where a is the number of dices, b is the number of sides and d stands for the word "dice". In our example, the 10-sided dice can be encoded as 1d10, while the classical 6-sided one is 1d6.

Mathematically, we will denote probability as $\mathrm{P}(y)$, where y represents a specific outcome. We can write, for example, that the probability of having 3 in the dice roll experiment is:

$$\mathrm{P}(y = 3) = \frac{1}{6}. \tag{2.1}$$

We can calculate more complicated probabilities. For example, what is the probability of having an odd number when rolling a 1d6? We need to calculate the number of events of interest and divide that number by the number of possible outcomes. In our case, the former is 1, 3, and 5 (three numbers), while the latter is any integer number from 1 to 6 (six numbers). This means that:

$$\mathrm{P}(y \text{ is odd}) = \frac{3}{6} = \frac{1}{2}. \tag{2.2}$$

2.2 What is random variable?

We have already discussed what a variable is in Section 1.3 of this textbook. Just as a reminder, it is a symbol that represents any of a set of potential values. If the value of a variable is known in advance, then it can be considered a deterministic variable. However, if the value depends on random events (and thus is not known in advance) then such variable is called **random variable** (or stochastic variable). In Section 2.1 we discussed the idea of probability and random events with example of coin tossing. If we continue that example then we could encode the outcome of coin tossing as y, expecting it to take value of 0 in case of heads and 1 in case of tails. This variable would be random because the outcome of each coin toss is not known in advance.

Fundamentally speaking, the randomness appears because of the lack of information about the environment. If we knew the initial state of the coin, the power of toss and could take into account all movements of air around it and somehow control all possible uncertainties around the flight of the coin, then we would be able to predict the outcome. In that case, the event would not be random any more, and thus the variable encoding the process would be deterministic. In real life, we do not know all the factors impacting the response variable (the variable of interest) and thus we consider their impact random.

Remark. The randomness disappears as soon as we observe the outcome of the event. For example, if we toss the coin for the first time and obtain tails, then the first value of the variable y is $y_1 = 1$. The variable itself stays random, but the specific outcome for the first trial is not random any more.

Furthermore, there are two types of random variables:

1. Discrete;
2. Continuous.

The first type represents the variable that takes count values. For example, variable y for the event "coin tossing" is discrete because it can only take values of 0 and 1. Another classical example is the variable encoding the score on a 1d6, the experiment with dice roll. We cannot get a value of 4.123 in this experiment, so the variable encoding it is discrete.

The second type of random variable represents the case, when it takes non-count value, such as real number over the whole range of values or on a specific interval of values. An example of a continuous variable is the time on a stopwatch, when a runner crosses the finish line.

Remark. The discrete variable can be considered as a continuous or approximated by the models for continuous ones when it has many outcomes. For example, the sales of wine can be measured in bottles, which is a discrete variable. But if the sales are measured in thousands of units then it might be easier to consider the variable to be continuous instead.

Finally, if we want to measure the probability of random variable taking specific

values, then for the discrete variable it can be done by considering the chance of that specific outcome over all possible ones. For example, for the fair dice, the chance of obtaining 3 is $\frac{1}{6}$: it can take values of 1, 2, 3, 4, 5 and 6. However, the probability that a continuous variable takes a specific value is zero, because the number of all possible cases for the continuous variable is infinite. For example, the time of a 100 meter runner can be anything between 9.2 seconds (which comes from the physics of human body) and infinity (if person never finishes). The probability that I will finish a race in 10 seconds is zero not because I am not fit enough, but rather because it is almost impossible to do that precisely on 10.000000 and not, let us say, on 10.000001.

Chapter 3

Discrete distributions

In this chapter we discuss the idea of discrete distributions, their properties and then move to the discussion of specific examples: Bernoulli, Binomial, Poisson, Geometric and Negative Binomial distributions.

3.1 What is discrete distribution?

A random variable discussed in Section 2.2 can take a variety of values. For now we focus on discrete random variable, which means that it can take one of the several possible values with some probabilities. For example, if we roll two six-sided dices, we will have a variety of outcomes from 2 (both having score of one) to 12 (both having score of six), but not all outcomes will have equal probability. For example, there are different ways of obtaining score of 7: 1+6, 2+5, 3+4, 4+3, 5+2 and 6+1 - but there is only one way of obtaining 2: 1+1. In this case, we are dealing with a distribution of values from 2 to 12, where each value has its own probability of occurrence. This situation is shown in Figure 3.1.

As can be seen from Figure 3.1, the distribution of probabilities in this case is symmetric, the chances of having very low and very high scores are lower than the chance of having something closer to the middle. The probability of having 7 is the highest and is $\frac{6}{36} = \frac{1}{6}$, which means that it will occur more often than other values if we repeat the experiment and roll the dices many times.

Any discrete distribution can be characterised using the following functions:

1. Probability Mass Function (PMF);
2. Cumulative Distribution Function (CDF);
3. Moment Generation Function (MMF);
4. Characteristic function (CF).

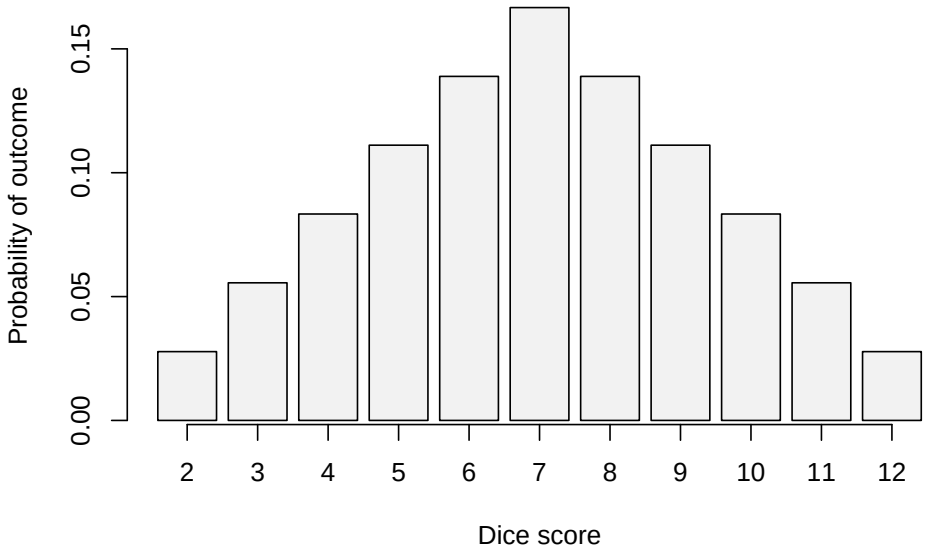


Figure 3.1: Distribution of outcomes for scores based on two dices.

PMF is the function of probability of occurrence from specific values of random variable. An example of PMF is shown in Figure 3.1. Based on it, we can say what the probability of a specific outcome is for the random variable.

CDF shows the probability of the event lower than the specified one. For example, the probability of getting the score lower than 4 is $\frac{1}{36} + \frac{2}{36} = \frac{1}{12}$, which corresponds to the sum of the first two bars in Figure 3.1. The CDF for our example is shown in Figure 3.2.

Any CDF is equal to zero for the values below possible (e.g. it is impossible to get score of 1 rolling two dices) and is equal to one for the values at and above the maximum (if we roll two dices, the score will be below 13). Given that CDF shows probabilities, it can never be greater than one or lower than zero.

Finally, MGF and CF are the functions that allow obtaining the moments of distributions, such as mean, variance, skewness etc. We do not discuss these functions in detail in this textbook, and we will discuss the moments later in the Section 5.1.

Because we considered the discrete random variable, the distribution shown in Figure 3.1 is discrete as well.

3.2 Bernoulli distribution (Tossing a coin)

Consider a case when we track whether it rains at Lancaster or not and want to understand based on this information whether it will rain today or not. In

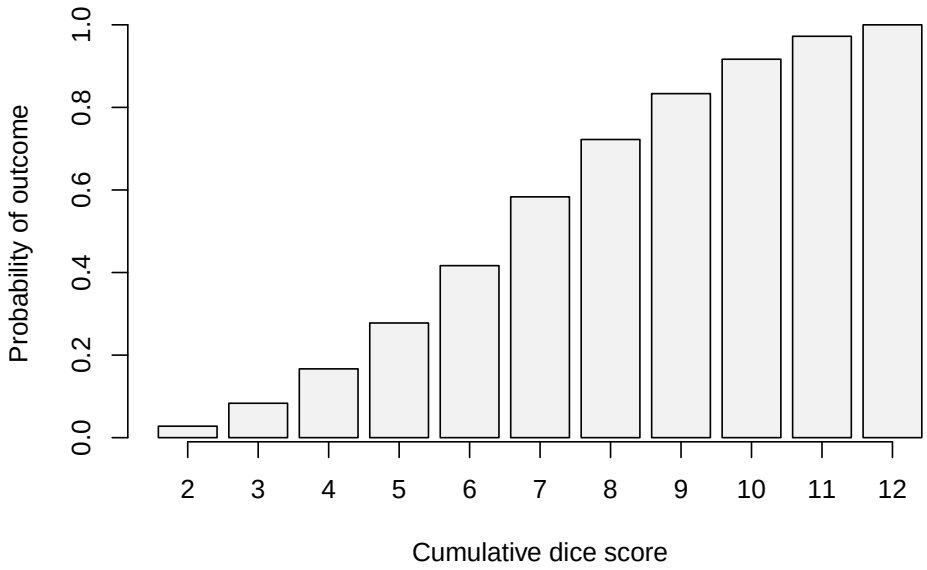


Figure 3.2: Cumulative distribution of outcomes for scores based on two dices.

this example, we have two values for a random variable (1 - it rains, 0 - it does not), and if we do not have any additional information, we assume that the probability that it rains is fixed. This situation can be modelled using the Bernoulli distribution.

It is one of the simplest distributions. It can be used to characterise the situation, when there are only two outcomes of event, like the classical example with coin tossing. In this special case, according to this distribution, the random variable can only take two values: zero (e.g. for heads) and one (e.g. tails) with a probability of having tails equal to $p = 0.5$. It is a useful distribution not only for the coin experiment, but for any other experiment with two outcomes and some probability p. For example, consumers behaviour when making a choice whether to buy a product or not can be modelled using Bernoulli distribution: we do not know what a consumer will choose and why, so based on some external information we can assign them a probability of purchase p.

In general, the distribution can be characterised with the following PMF:

$$f(y, p) = p^y (1 - p)^{1-y}, \tag{3.1}$$

where y can only take values of 0 and 1. Figure 3.3 demonstrates how the PMF (3.1) looks for several probabilities.

The mean of this distribution equals to p, which is in practice used in the estimation of the probability of occurrence p: collecting a vector of zeroes and ones and then taking the mean will give the empirical probability of occurrence $\hat{p}$. The variance of Bernoulli distribution is $p \times (1 - p)$.

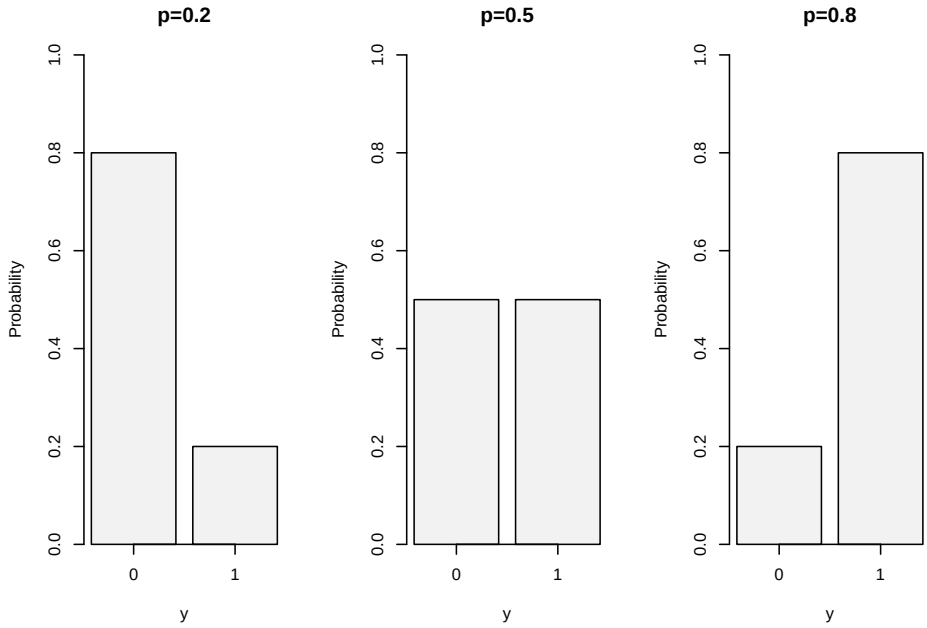


Figure 3.3: Probability Mass Function of Bernoulli distribution with probabilities of 0.2, 0.5 and 0.8

Finally, the CDF of the distribution is:

$$F(y, p) = \begin{cases} 1 - p, \text{ for } y = 0 \\ 1, \text{ for } y = 1. \end{cases} \tag{3.2}$$

which can be plotted as shown in Figure 3.4.

The CDF of Bernoulli distribution is seldom used in practice and is provided here for completeness.

While sitting at home during the COVID pandemic isolation, Vasiliy conducted an experiment: he threw paper balls in a rubbish bin located in the far corner of his room. He counted how many times he missed and how many times he got the balls in the bin. It was 36 to 64.

1. What is the probability distribution that describes this experiment?
2. What is the probability that Vasiliy will miss when he throws the next ball?
3. What is the variance of his throws?

Solution. This is an example of Bernoulli distribution: it has two outcomes and a probability of success.

We will encode the miss as zero and the score as one. Based on that, taking the mean of the outcomes, we can estimate the mean of Bernoulli probability of

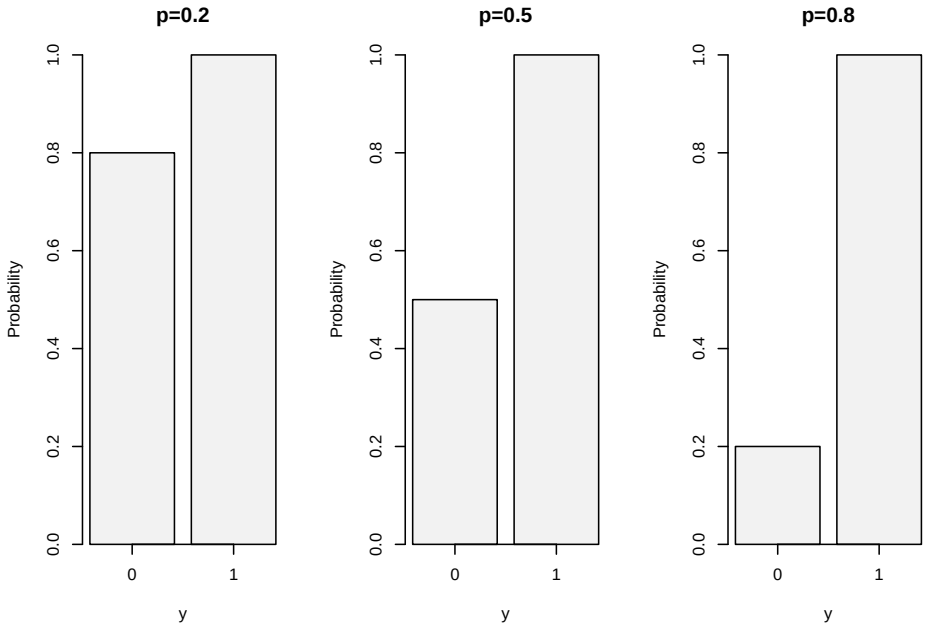


Figure 3.4: Cumulative Distribution Function of Bernoulli distribution with probabilities of 0.2, 0.5 and 0.8

miss:

$$\bar{y} = \hat{p} = \frac{36}{100} = 0.36.$$

So, when Vasiliy throws the next ball in the bin, he will miss with the probability of 0.36.

The variance is $p \times (1 - p) = 0.36 \times 0.64 = 0.2304$.

In R, this distribution is implemented in `stats` package via `dbinom(size=1)`, `pbinom(size=1)`, `qbinom(size=1)` and `rbinom(size=1)` for PDF, CDF, QF and random generator respectively. The important parameter in this case is `size=1`, which will be discussed in Section 3.3.

Remark. A nice example of a task using Bernoulli distribution is the Saint Petersburg Paradox (Kotz et al., 2005, page 8318). The idea of it is as follows. Imagine that I offer you to play a game. I will toss the coin as many times as needed to get first heads. We will calculate how many tails I had in that tossing and I will pay you an amount of money, depending on that number. If I toss tails once, I will pay you £2. If I toss it twice, I will pay $£2^2 = 4$. In general, I will pay $£2^n$ if I toss consecutively n tails before getting heads. The question of this task, is how much you will be ready to pay to enter such game (i.e. what is the fair price?). Daniel Bernoulli proposed that the fair price can be calculated

via the expectation of prize, which in this case is:

$$\text{E(tails } n \text{ times and heads on } n+1) = \sum_{j=1}^{\infty} 2^j \left(\frac{1}{2}\right)^{j+1}$$

The logic behind this formula is that mathematically, we assume that we can have infinite number of experiments, and each prize has its outcome. For example, the probability to get just £2 is $\frac{1}{2}$, while the probability to get £4 is $\frac{1}{4}$ etc. But the values cancel each other out in this formula leaving us with:

$$\text{E(tails } n \text{ times and heads on } n+1) = \sum_{j=1}^{\infty} \frac{1}{2} = \infty.$$

So, although it is unrealistic to expect in real life that the streak of tails will continue indefinitely, the statistics theory tells us that the fair price for the game is infinity. Practically speaking, the infinite amount of tails will never happen, so we should have a finite number for the price. But mathematics assumes that the experiment can be repeated infinite amount of times, and in this case it is entirely possible that we will observe an infinite streak of tails. This is the Saint Petersburg paradox, which demonstrates how sometimes the asymptotic properties relate to reality. I think that it provides a good demonstration of what statistics typically works with.

Finally, coming back to the rain example, we cannot say for sure whether it will rain tomorrow or not, but based on the collected sample, we can calculate the probability of that event. But remember that even if the probability is low, it does not mean that you do not need to bring an umbrella with you.

3.3 Binomial distribution (Multiple coin tosses)

In the previous example with coin tossing we focused on the experiment that contained only one trial: toss a coin, see what score you got (zero or one). However, we could make the game more complicated and do it in, let us say, 10 trials. In this more complicated experiment we would sum up the scores to see what we get after those 10 trials. In theory, we can have any integer number from zero to ten, but the resulting score will be random and its chance of appearance will vary with the score. For example, we will get 0 only if in all 10 trials we get heads, while we can get 1 in a variety of ways:

1. first trial is one and then the rest are zero;
2. second trial is one and the rest are zero;
3. etc.

This means that the score in this experiment will have a distribution of its own. But can we describe it somehow?

Yes, we can. The distribution that underlies this experiment is called "Binomial". For the coin tossing experiment with 10 trials and $p = 0.5$ it looks like one shown in Figure 3.5.

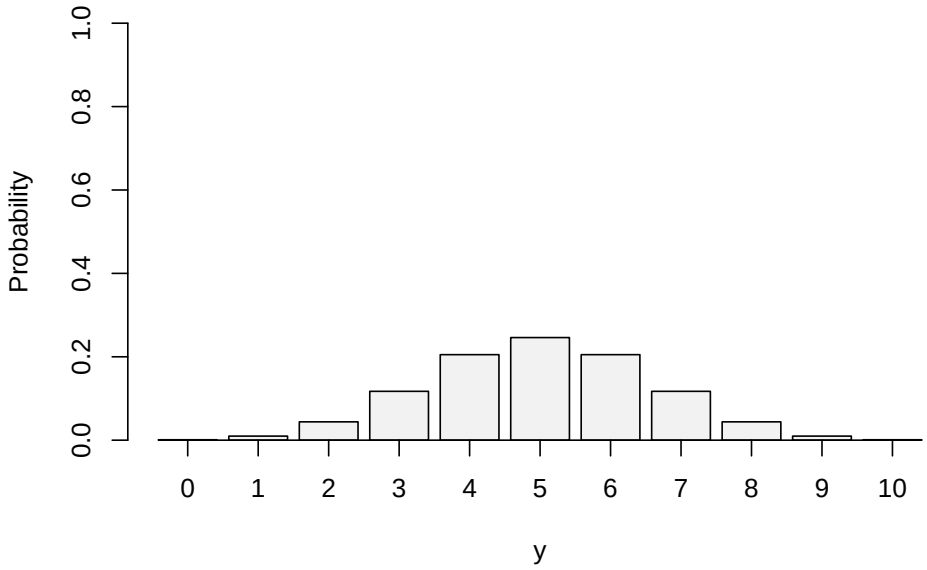


Figure 3.5: Probability Mass Function for Binomial distribution with p=0.5 and n=10.

From the Figure 3.5, we can see that the most probable outcome is the score of 5. This is because there are more ways of getting 5 than getting 4 in our example. In fact the number of ways can be calculated using **Binomial coefficient**, which is defined as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{3.3}$$

where k is the score of interest, n is the number of trials and ! is the symbol for factorial. In our example, $k = 5$ and $n = 10$, meaning that we have:

$$\binom{10}{5} = \frac{10!}{5!(10-5)!},$$

which can be calculated in R as:

```
factorial(10)/(factorial(5)^2)
```

and is equal to 252. Using this formula, we can calculate other scores for comparison:

```
factorial(10)/(factorial(c(0:10))*factorial(10-c(0:10))) |>
    setNames(nm=c(0:10))
```

```
##   0   1   2   3   4   5   6   7   8   9  10
##   1  10  45 120 210 252 210 120  45  10   1
```

Given that we throw the coin 10 times, there are overall $2^{10} = 1024$ theoretical of outcomes of this experiment, which allows calculating the probability of each outcome:

```
round((factorial(10) /
          (factorial(c(0:10))*factorial(10-c(0:10)))) /
         1024,3) |>
    setNames(nm=c(0:10))
```

```
##     0     1     2     3     4     5     6     7     8     9    10
## 0.001 0.010 0.044 0.117 0.205 0.246 0.205 0.117 0.044 0.010 0.001
```

These probabilities can be obtained via the PMF of Binomial distribution:

$$f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{3.4}$$

We will denote this distribution as $\mathcal{B}\rangle(n, p)$. In R, this is implemented in the `dbinom()` function from `stats` package. Note that so far we assumed that $p = 0.5$, however in real life this is not necessarily the case.

Consider a problem of demand forecasting for expensive medical equipment in the UK. These are not typically bought in large quantities, and each hospital might need only one machine for their purposes, and that machine can last for many years. For demonstration purposes, assume that there is a fixed probability $p = 0.1$ that any given hospital decides to bye such machine. It is safe to assume that the probability that machine is needed in a hospital is independent of the probability in the other one. In this case, the distribution of demand for machines in the UK per year will be Binomial. For completeness of our example, assume that there are 20 hospitals in the country that might need such machine, then this will be characterised as $\mathcal{B}\rangle(20, 0.1)$ and will be an asymmetric distribution, as shown in Figure 3.6.

From the Figure 3.6, it is clear that there is a high probability that only a few machines will be bought: 0, 1, 2 or 3. The probability that we will sell 10 is almost zero. We can also say that with the highest probability, the company will sell 2 machines. We could also use mean, saying how much the company will sell on average, which in Binomial distribution is calculated as $n \times p$ and in our case is $20 \times 0.1 = 2$. However, producing expensive medical equipment typically takes time and should be done in advance. So, if we told the company that they should produce only two, we might then face a situation, when the demand was higher (for example, 6) and they lost the sales.

So, while we already have a useful information about the distribution of demand in this situation, it is not helpful for decision making. What we could do is consider a case of satisfying, let us say, 99% of demand on machines. In our example, we should sum up the probabilities and find for which number of cases

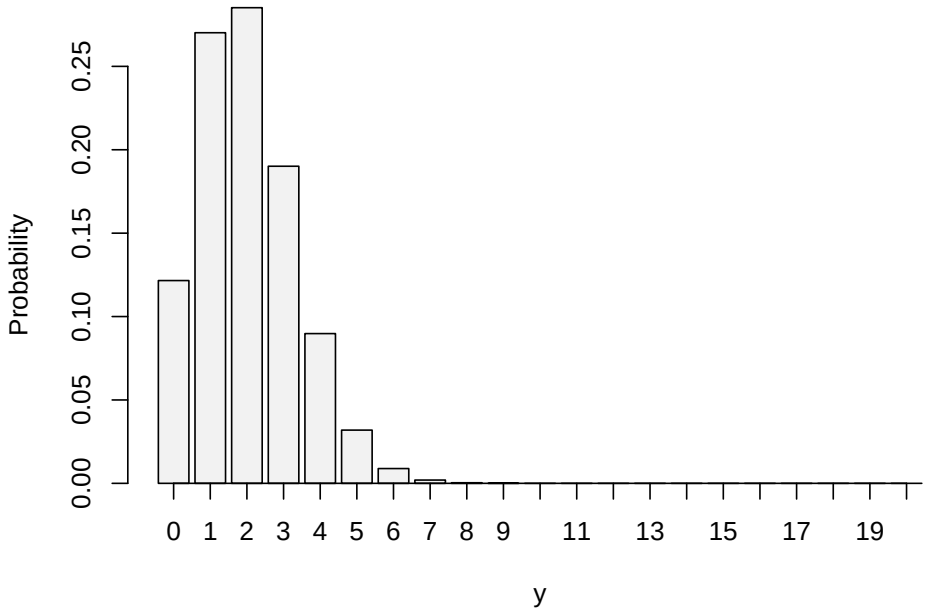


Figure 3.6: Probability Mass Function for Binomial distribution with p=0.1 and n=20.

we get to 99%. Formally, this can be written as $P(y < k) = 0.99$ and we need to find k. One way of doing this is by looking at cumulative distribution function of Binomial distribution, which is mathematically represented as:

$$F(k, n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}. \tag{3.5}$$

This CDF in our example will have the shape shown in Figure 3.7.

We can see from Figure 3.7 that the cumulative probability reaches value close to 1 after the outcome of 6. Numerically, for the $y = 6$, we have probability:

```
pbinom(6, size=20, prob=0.1)
```

```
## [1] 0.9976139
```

while for 5 we will have approximately 0.989. So, for our example, we should recommend the company to produce 6 machines - in this case in 99% of the cases the company will not loose the demand. Yes, in some cases, only 2 or 3 machines will be bought instead of 6, but at least the company will avoid a scenario, when their product is unavailable for clients.

We can get the same result by using Quantile Function, which is implemented in R as `qbinom()`:

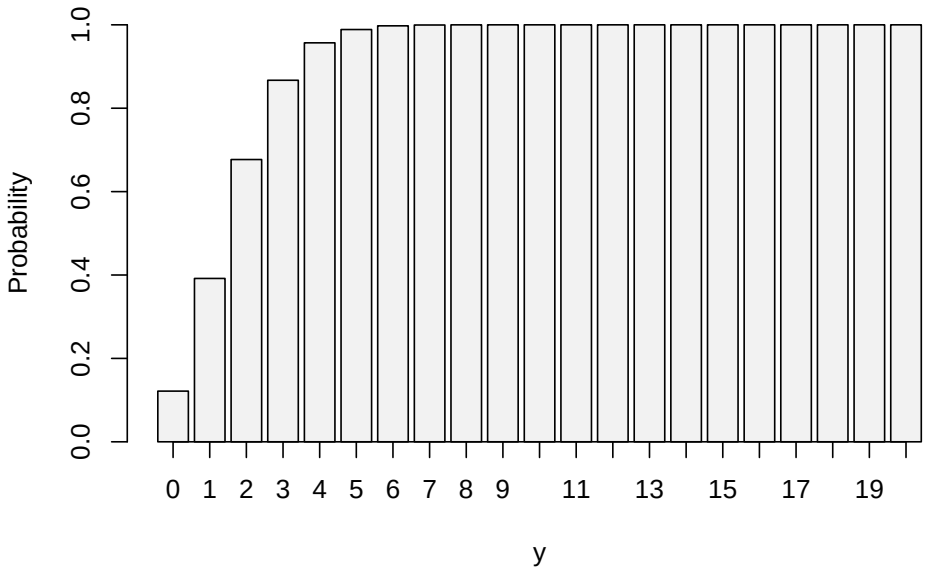


Figure 3.7: Cumulative Distribution Function for Binomial distribution with p=0.1 and n=20.

```
qbinom(0.99, 20, 0.1)
```

In some cases, we might need to know the mean and variance of the distribution. For the Binomial distribution, they can be calculated via the following formulae:

$$\mathrm{E}(y) = p \times n, \tag{3.6}$$

and

$$\mathrm{V}(y) = p \times (1-p) \times n. \tag{3.7}$$

So, in our situation the mean is $20 \times 0.1 = 2$, while the variance is $20 \times 0.1 \times 0.9 = 1.8$.

From the example above, we can take away several important characteristics ofthe Binomial distribution:

1. There are many trials: we have many hospitals, each one of which can be considered as a "trial' ';
2. In each trial, there are only two outcomes: buy or do not buy;
3. The probability of success (someone decides to buy the machine) is fixed between the trials: we assume that the probability that a hospital A decides to buy the machine is the same as for the hospital B;
4. The trials are independent: if one hospital buys a machine, this should not impact the decision of another one;
5. The number of trials n is fixed and known.

Final thing to note about the Binomial distribution is that with the growth of the number of trials it converges to the Normal one, no matter what the probability of success is. We can use this property to get answers about the Binomial distribution in those cases. We will discuss this in Section 4.3.

In R, the Binomial distribution is implemented via `dbinom()`, `pbinom()`, `qbinom()` and `rbinom()` from `stats` package. The functions have parameter `size` which is the number of trials n, and parameter `prob`, which is a probability of success p. If `size=1`, the distribution functions revert to the ones from Bernoulli distribution (Section 3.2).

3.4 Poisson distribution (Modelling arrivals)

Consider a situation, when we want to model an arrival of patients in a hospital for each hour of day. In that case, they arrive for different reasons: some have fractures, others have headaches, some come because they sneezed and the others are there because they are seriously ill. There is potentially a lot of people that could come to the hospital (the whole population living in the area), but typically only few of them will show up in each specific hour. Furthermore, all these people are typically not related (unless there was an event that caused a specific condition to a group of people, such as a big pub fight) and arrive at random. In this example, we could argue that the process of arrival is memoryless, because the arrivals of different patients are not related. For the probability theory, this implies that the probability that a patient arrives in a specific period of time should be independent of when the previous patient arrived. Mathematically, this is represented as:

$$\mathrm{P}(t > \tau_1 + \tau_2) = \mathrm{P}(t > \tau_1)\mathrm{P}(t > \tau_2), \tag{3.8}$$

where t is the waiting time until the next arrival, and τ_1 and τ_2 are waiting times. The formula (3.8) shows that the probability that we will wait more than τ_1 and τ_2 is just equal to the product of probabilities of waiting for more than τ_1 and more than τ_2. The formula relies on the independence of probabilities (discussed in Chapter 2). From the mathematical point of view, there is a function that supports the property (3.8) - it is exponent:

$$e^{\tau_1+\tau_2} = e^{\tau_1}e^{\tau_2}. \tag{3.9}$$

Based on this principle of memorylessness (and based on the Exponential distribution, which we will discuss in Section 4.5), Poisson distribution is derived. It is a discrete distribution, which is used for modelling number of patients arriving at specific time intervals, based on the average number of arrivals λ. Its PMF has the shape shown in Figure 3.8.

From the graphs in Figure 3.8, we can make several observations:

1. Zeroes are natural in Poisson distribution. This means that there is a chance that nobody will come in the next hour.

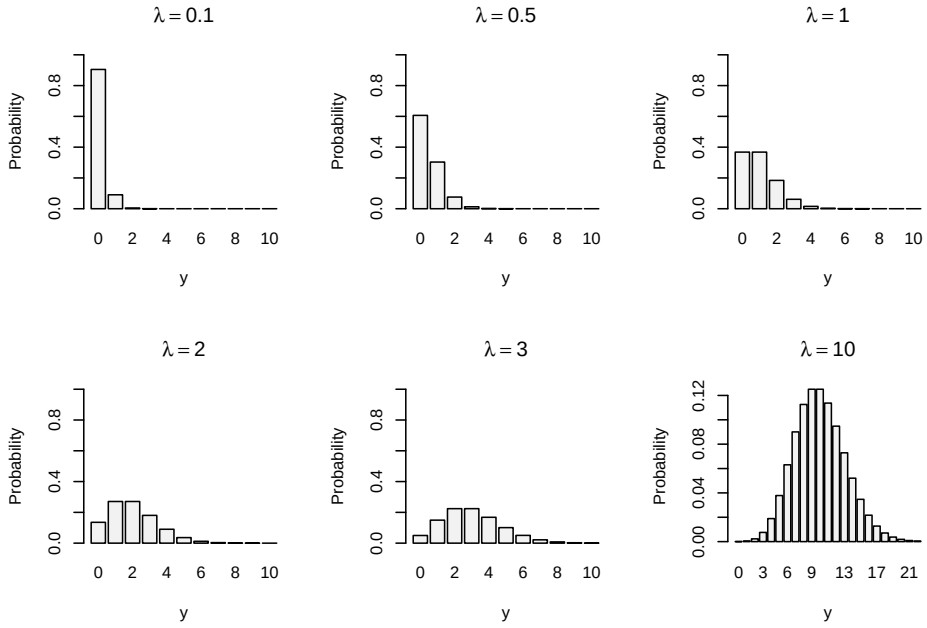


Figure 3.8: Probability Mass Function of Poisson distribution with different values of λ.

2. With the increase of the average number of arrivals λ, the chance to have more patients arriving increases. For example, with $\lambda = 0.1$, the probability of having no patients is approximately 0.9, while in case of $\lambda = 1$, it is approximately 0.4.
3. With the increase of λ, the shape of distribution becomes closer to the Normal one (discuss in Section 4.3).

Mathematically, the PMF of Poisson distribution is represented as:

$$f(y, \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \tag{3.10}$$

where e is the Euler's constant. The part $e^{-\lambda}$ represents the memoryless arrivals of patients. The Poisson distribution is characterised as $\mathcal{P}\wr\rangle f(\lambda)$ and has an additional property that its expectation is equal to variance:

$$\mathrm{E}(y) = \mathrm{V}(y) = \lambda, \tag{3.11}$$

which makes it convenient to work with and easy to estimate. If we have the parameters of the Poisson distribution, we can calculate probabilities for a variety of situations or generate quantiles - depending on what we need specifically. For example, for scheduling purposes we might need to understand what is the probability of having more than zero but up to four patients in any specific

hour. Assume that based on the available data we estimated that $\hat{\lambda} = 2$. In this situation we need to either use PMF or the CDF to calculate the following probability:

$$\begin{aligned} \mathrm{P}(0 < y \leq 4) =& \mathrm{P}(y = 1) + \mathrm{P}(y = 2) + \mathrm{P}(y = 3) + \mathrm{P}(y = 4) = \\ & \mathrm{P}(y \leq 4) - \mathrm{P}(y = 0) \end{aligned}. \tag{3.12}$$

The first line in (3.12) shows how the probability can be calculated via the PMF, while the second one is for the CDF. They both can be calculated in R via:

```
sum(dpois(1:4,2))
```

```
## [1] 0.8120117
```

```
ppois(4,2) - ppois(0,2)
```

```
## [1] 0.8120117
```

where `dpois()` is the PMF and `ppois()` is the CDF of the Poisson distribution. Mathematically, the CDF of the Poisson distributions is written as:

$$F(y, \lambda) = e^{-\lambda} \sum_{j=0}^{y} \frac{\lambda^j}{j!}, \tag{3.13}$$

for integer values of y. Finally, we could get quantiles of the distribution for the specified probability. In R, this is done via the `qpois()` function. For example, here is the 0.95 quantile for the Poisson distribution with $\hat{\lambda} = 2$:

```
qpois(0.95, 2)
```

```
## [1] 5
```

Remark. When identifying a suitable distribution, Poisson has several distinct characteristics:

1. There are many trials (in our example with the hospital, many people live in the area).
2. There is a small chance that each specific patient will arrive in a specific hour (number of patients arriving at the hospital each hour is small in comparison with the population).
3. Each specific event is independent of the others (memorylessness property: a patient arrives independent of another one).

These three criteria can be used to identify Poisson distribution in practice.

3.5 Discrete Uniform distribution (Rolling a dice)

In the second world war, the UK faced a problem of understanding how many tanks Nazi Germany had. Some tank components had serial numbers, and when

some of tanks were captured or destroyed, it was possible to track these numbers. But was it possible to say how many enemy tanks there were overall? As it appears, it was. From the stand point of the UK, it was equally possible to have a serial number 0001, 0042, 0500 or, say, 1984. The distribution of these serial numbers could be assumed uniform, which was then used to get an estimate of the maximum number of tanks that the enemy had. We will come back to the solution of this problem at the end of this section.

This distribution is one of the basic ones in the probability theory. For now we focus on the discrete version of it, keeping in mind that there also exists the continuous one (see Section 4.2).

The classical example of application of this distribution is dice rolling. The conventional dice has 6 sides and when rolled can give a value of 1 to 6. If the dice is fair then the probability of getting a score on it is the same for all the sides. This means that the PMF of the distribution can be written as:

$$f(y, k) = \frac{1}{k}, \tag{3.14}$$

where k is the number of outcomes (sides of the dice). The more outcomes there are, the lower the probability of having a specific outcome is. For example, on a dice with 10 sides, the probability of getting the score 5 is $\frac{1}{10}$, while on the 6-sided version it is $\frac{1}{6}$.

The PMF of the Uniform distribution is shown visually in Figure 3.9 on example of 1d6.

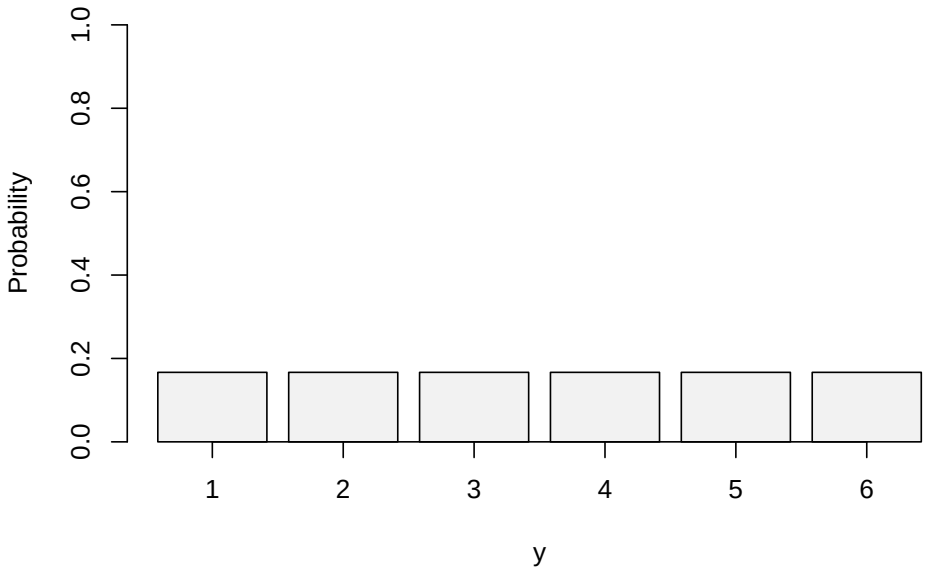


Figure 3.9: Probability Mass Function of Uniform distribution for 1d6.

The mean of this distribution is calculated as $\frac{a+b}{2}$, where a is the lowest and b is the highest possible values. So, for the 1d6, the mean is $\frac{1+6}{2} = 3.5$. This means that if we roll the dice many times the average score will be 3.5.

The variance of the uniform distribution depends on the number of outcomes and is calculated as:

$$\sigma^2(y, k) = \frac{k^2 - 1}{12}. \tag{3.15}$$

As can be seen from the formula, the variance of Uniform distribution is proportional to the number of outcomes.

Coming to the CDF of the Uniform distribution, it is calculated as:

$$f(y, k) = \frac{y - a + 1}{k}, \tag{3.16}$$

where a is the lowest possible value and k is the number of outcomes. This CDF can be visualised as shown in Figure 3.10.

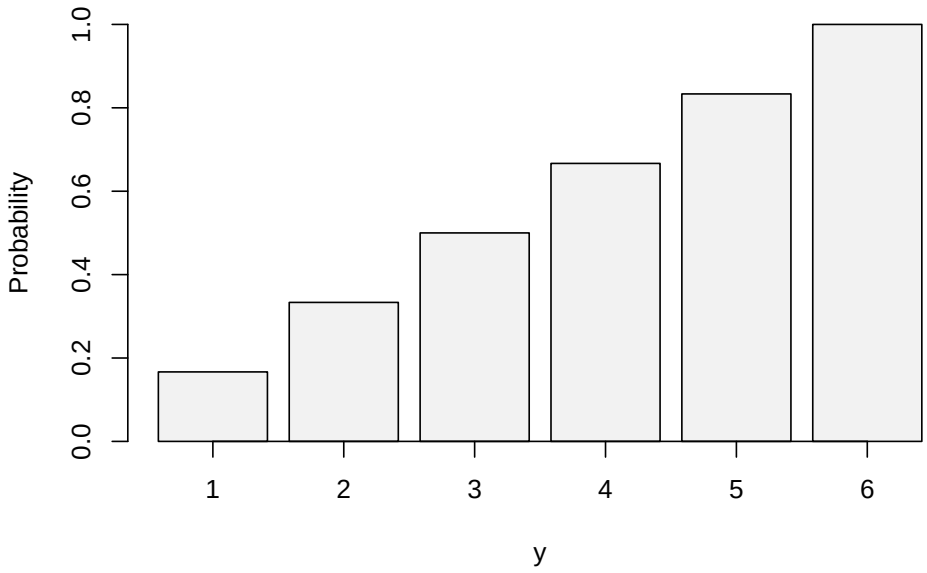


Figure 3.10: Cumulative Distribution Function of Uniform distribution for 1d6.

Given that the probability of each separate outcome in the Uniform distribution is always $\frac{1}{k}$, the CDF demonstrates a linear growth, reaching 1 at the highest point, which can be interpreted as rolling 1d6, we will always get a value up to 6 (less than or equal to 6). The CDF can be used to get probabilities of several events at the same time. For example, we can say that when rolling 1d6 the probability of getting 1 or 2 is $\frac{2-1+1}{6} = \frac{1}{3}$.

Bernoulli distribution (Section 3.2) with $p = 0.5$ can be considered as a special case of the Uniform distribution (with only two outcomes).

A company produces headphones, putting serial numbers on them. So far, it has produced 9,990 of them. If a customer buys headphones, what is the probability that they will get a serial number with three digits?

Solution. This is the task on Uniform distribution, because serial numbers do not repeat and we can assume that the probability of getting any of them is the same. In terms of parameters, $a = 1$ and $b = 9990$. To get a serial number with three digits, a customer needs to have anything between 100 and 999. This can be formulated as:

$$\mathrm{P}(100 \leq y \leq 999) = \mathrm{P}(y \leq 999) - \mathrm{P}(y \leq 99).$$

Inserting the values in the CDF of the Uniform distribution (3.16) we get:

$$\mathrm{P}(100 \leq y \leq 999) = \frac{999}{9990} - \frac{99}{9990} \approx 0.1 - 0.01 = 0.09.$$

Remark. Similarly how Binomial distribution is a generalisation of the Bernoulli, there is distribution describing the multiple dice rolls. It is called the Multinomial distribution. While we do not discuss it here, we note that this is a distribution, which is, for example, used to model respondents choices in survey, when the variable of interest is in a categorical scale and the probabilities for different options are not equal.

Coming back to the example with the German tanks, this problem was solved by estimating the maximum of the uniform distribution, i.e. getting the estimate of b. Goodman (1952) provides a solution. He showed how to find the maximum for a set of serial numbers. An unbiased and efficient estimate of the maximum value can be calculated using the following formula:

$$\hat{b} = m + \frac{m}{k} - 1,$$

where m is the maximum number observed in the sample, and k is the number of the observed values. For the sequence of serial numbers (from Goodman, 1954) 83, 135, 274, 380, 668, 895, 955, 964, 1113, 1174, 1210, 1344, 1387, 1414, 1610, 1668, 1689, 1756, 1865, 1874, 1880, 1936, 2005, 2006, 2065, 2157, 2220, 2224, 2396, 2543 and 2787, we have:

$$\hat{b} = 2787 + \frac{2787}{31} - 1 \approx 2876.$$

The real maximum number in that example was 2885, making the estimate above quite accurate. There are several solutions to this problem, and in the Second World War, statistical estimates were shown to be much more accurate than the ones obtained via intelligence.

3.6 Negative Binomial distribution

One of the tasks in manufacturing is to understand how many days it might take for a machine to break. In this case, everyday when it works we could encode as a

"failure to break", and when it breaks, we would encode it as "success". Assuming that there are many small unpredictable factors impacting the possibility to break, we can safely assume that there is a fixed probability that this will happen. Can we somehow model this process to make adequate decisions about the work of the machine?

Yes. The Negative Binomial distribution is one of the appropriate tools here. It models the number of failures in an experiment (no breaks in our example) before some defined number of successes occurs. In contrast with the Binomial distribution, in the NegBin, the number of the total trials is unknown. And in the example above, we could say that a success is when the machine stops working.

Visually, the PMF of the NegBin can be represented in the following way (Figure 3.11) with p being the probability of success, k being the number of successes before the experiment stops, and m being the number of failures.

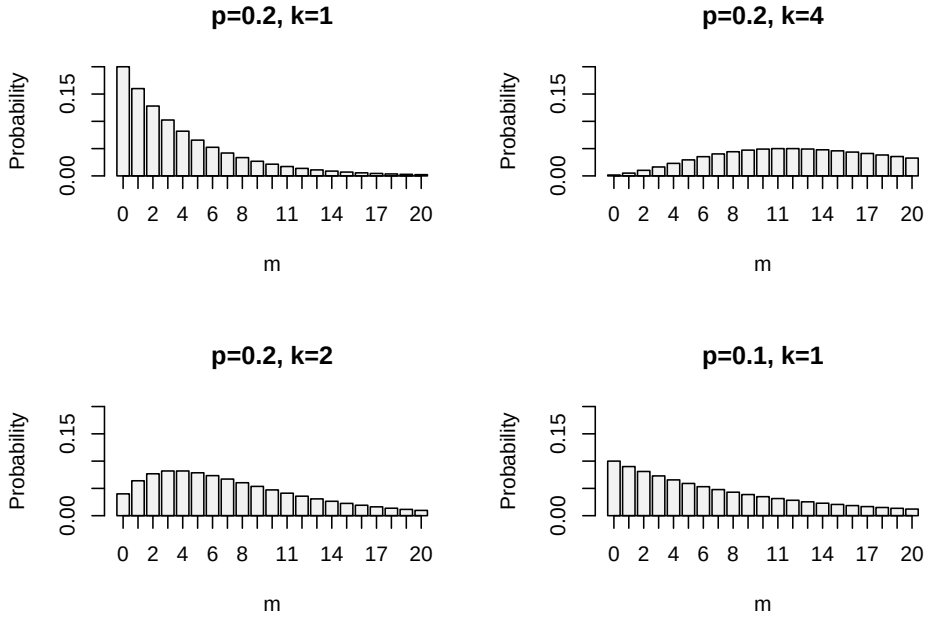


Figure 3.11: PMF of the Negative Binomial distribution with different probabilities and number of successes k.

Figure 3.11 shows that, for example, if p=0.2, and we want to have one success (k=1), we have the probability of 0.2 that we will succeed without having any failures. To make this easier to understand, if we have a 1d5 (five-sided dice), and we define success as rolling 5 on it, the probability that we will have it in the first roll is 0.2 (this is one out of five possible outcomes), which corresponds to the first bar in the first plot in Figure 3.11. The probability that we will roll 5 in the second trial is $0.2 \times (1 - 0.2) = 0.16$, which corresponds to the

second bar in the first plot in Figure 3.11, and so on. With $k = 2$, a situation is modelled when we need to get five on a dice two times in an experiment (not necessarily sequentially). The calculation of the specific probability becomes more complicated with the incrase of k. But in general, the PMF of the Negative Binomial distribution is written as:

$$f(m, k, p) = \binom{k + m - 1}{m} p^k (1 - p)^m, \tag{3.17}$$

where k is the number of successes and m is the number of failures. This PMF is very similar to the one of the Binomial distribution, but it is parametrised differently, which makes it useful in some contexts (like the one mentioned in the beginning of this section).

The characteristics of the Negative Binomial distribution are similar to the Binomial one, discussed in Section 3.3, except for the last (fifth) element: the number of overall trials in the NegBin is unknown.

The mean and standard deviations of the distribution are defined as:

$$\mathrm{E}(y) = k\frac{1 - p}{p} \tag{3.18}$$

and

$$\mathrm{V}(y) = k\frac{1 - p}{p^2} \tag{3.19}$$

respectively.

We do not provide the formula for the CDF here, because it is quite complicated.

A machine in a factory works only when at least one of the three independent components works. The probability of a break of each component is estimated to be 0.1. How many days would it take on average for the machine to stop working? What is the probability that the machine will work exactly for a month (30 days)?

Solution. In this task, we use the Negative Binomial distribution because:

1. There are many potential trials;
2. In each trial, there are only two outcomes: components work or fail;
3. The probability of success (a component breaks) is fixed between the trials;
4. The trials are independent: if one component fails, this should not impact the work of the other one;
5. The number of trials n is unknown.

The probability of failure (success in the terms of NegBin) is $p = 0.1$. The number of components, $k = 3$. This is all we need to know to use the distribution and answer the questions.

The average number of days until the machine stops working is calculated using (3.20):

$$\mathrm{E}(y) = 3\frac{1 - 0.1}{0.1} = 27 \tag{3.20}$$

As for the answer to the second question, working for exactly 30 days means that the machine will break on the 31st day. We can use the PMF, specifying $m = 30$:

$$f(30, 3, 0.1) = \binom{30 + 3 - 1}{3} 0.1^3 (1 - 0.1)^3 0 \approx 0.021.$$

In R, the Negative Binomial distribution is implemented in the functions `rnbinom()`, `dnbinom()`, `pnbinom()`, and `qnbinom()`, implementing the random function, the PMF, the CDF and the QF respectively. Here how the PMF can be used to get the answer to the question in the previous task:

```
dnbinom(x=30,size=3,prob=0.1)
```

```
## [1] 0.02102601
```

Chapter 4

Continuous distributions

After discussing the discrete probability distributions, we can now move to the continuous ones. The idea behind them is similar to the one discussed in Chapter 3. In this chapter, we will discuss the main properties of continuous distributions, focusing on several of them, including: Uniform, Normal, Exponential, Laplace, S, Generalised Normal, Asymmetric Laplace, Log Normal, Inverse Gaussian and Gamma.

4.1 What is continuous distribution?

The main difference arises from the idea discussed in Section 2.2: the probability that a continuous random variable will take a specific value is zero. Because of that we should be discussing the probability of a random variable taking a value in an interval. Figure 4.1 demonstrates an empirical distribution of continuous random variable.

Based on Figure 4.1, we can say that the probability of obtaining the value between 100 and 105 is higher than for the variable to get any other interval. It also looks like the variable is continuous on the interval between 75 and 125, and we do not observe any values outside of this interval.

Almost any continuous distribution can be characterised by several functions:

1. Probability density function (PDF);
2. Cumulative distribution function (CDF);
3. Quantile function (QF);
4. Moment Generation Function (MMF);
5. Characteristic function.

PDF has similarities with PMF, but does not return the probabilities, but rather the score for the variable in each specific point because (once again) the probability of continuous variable taking a specific value is equal to zero. PDF

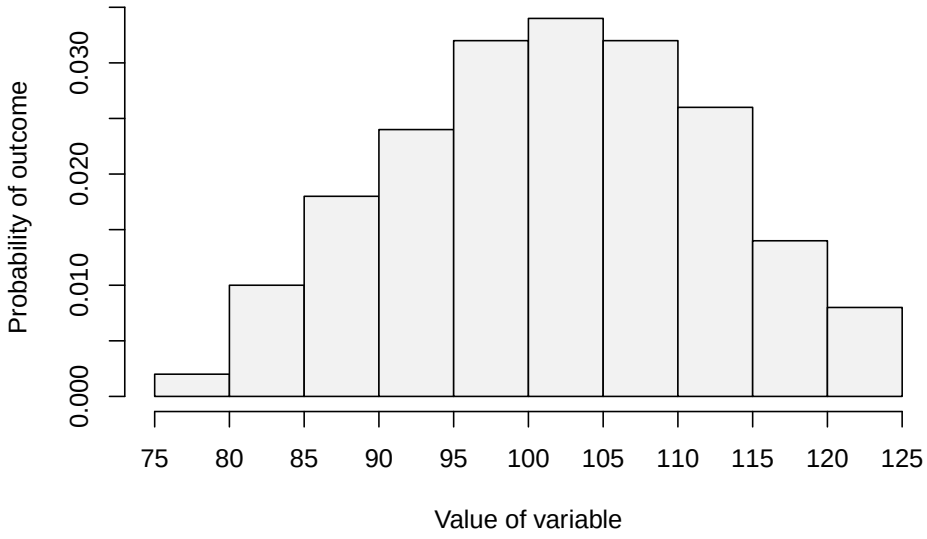


Figure 4.1: Distribution of a continuous random variable.

however is useful because it represents the shape of the distribution, showing where the values are concentrated. Figure 4.2 demonstrates an example of PDF of Normal distribution (discussed in more detail in Section 4.3).

It can be seen from the Figure 4.2 that the density of the distribution is higher at its centre, around zero, which means that it is more likely to get values around the centre rather than near the tails of the distribution.

If we want to work with probabilities in case of the continuous distribution, we need to use CDF, which is similar to the one discussed in the Chapter 3. Figure 4.3 shows example of CDF of Normal distribution.

The CDF of continuous variable has the same properties as CDF of the discrete one with the minor differences: in a general case it converges to one with the increase of the value y and converges to zero with the decrease of it. There are some continuous distributions that are restricted with an interval. For those distributions, the CDF reaches boundary values.

CDF is obtained by calculating the surface of PDF for each specific value of y. Figure 4.4 shows the connection between them.

The dark area in the first plot in Figure 4.4 is equal to the probability (the value on y-axis) in the second plot, which is approximately equal to 0.16.

Another important function is the Quantile Function. It returns the value of y for the given probability. By the definition, QF is the inverse of the CDF. It does not always have a closed form (thus cannot be represented mathematically), and for some distributions numerical optimisation is required to obtain the quantiles.

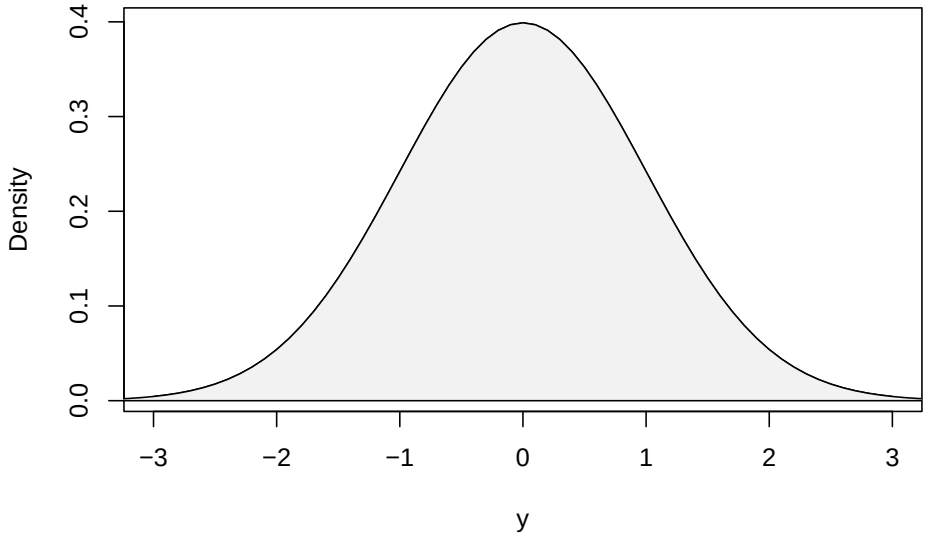


Figure 4.2: Probability Density Function of Normal distribution

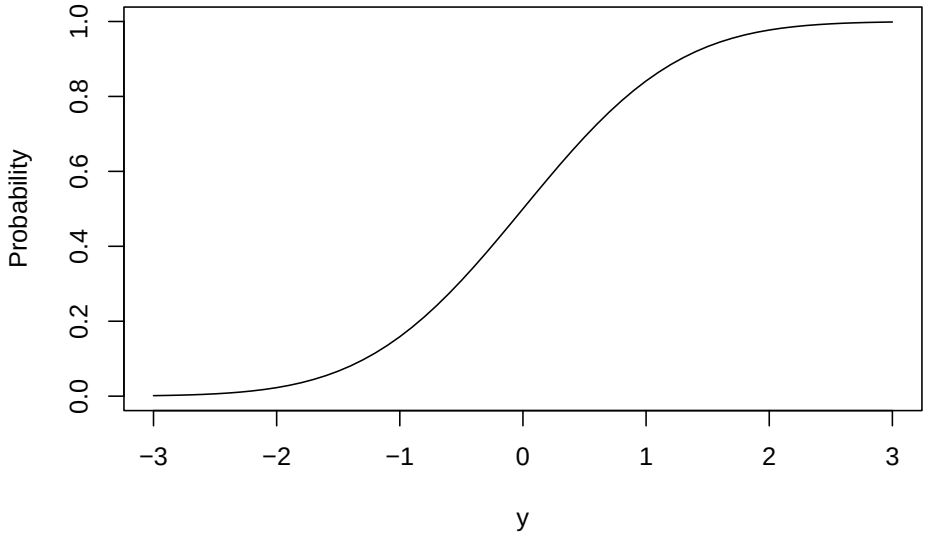


Figure 4.3: Cumulative Distribution Function of Normal distribution

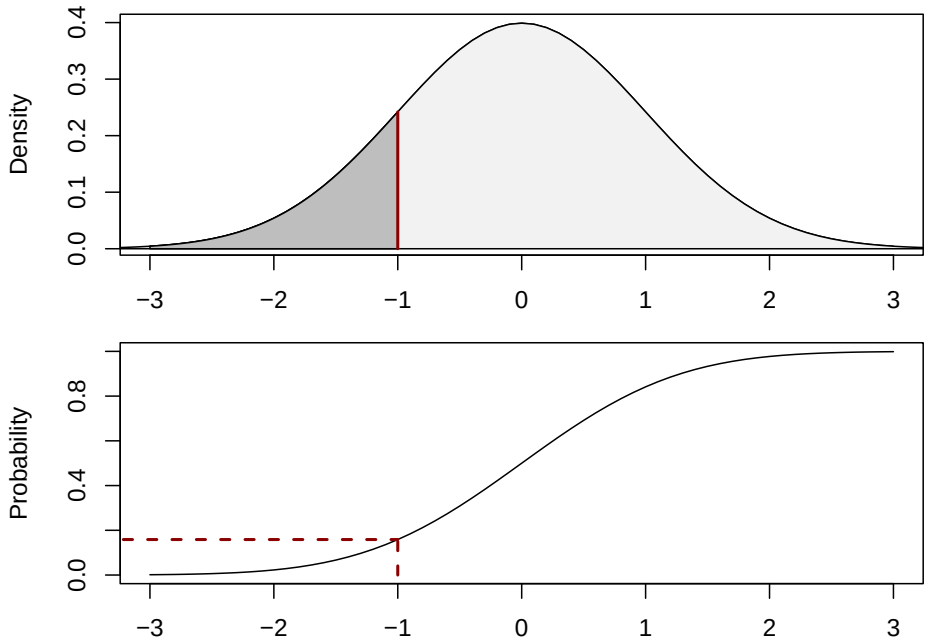


Figure 4.4: Cumulative Distribution Function of Normal distribution

Figure 4.5 demonstrates the quantile function of Normal distribution.

The dashed lines in Figure 4.5 show the value of y for the probability 0.25 according to the quantile function. In this example $y \approx -0.68$, meaning that in 25% of the cases the random variable y will lie below this value. The quantiles of distributions are discussed in more detail in Section 5.1.

Finally, we have already mentioned MGF and CF in the context of discrete distributions. They play similar roles to the ones already discussed and can be used to obtain mean, variance, skewness and other statistics.

4.2 Continuous Uniform distribution

Consider an example with elevator. You press the button to call it, and it arrives after some time. This time of arrival can be 0 seconds if the elevator is at your floor, or 1 minute if it is already moving and delivering people to a different floor. From our perspective, we do not know what the state of the elevator is, and we consider each of the possible scenarios equally probable. So, the elevator arriving in 0 seconds has the same probability as it arriving in 5, 15, 30, 55 or 60 seconds. In this case, we could use the continuous uniform distribution to model the lift arrival and to understand, for example, how much time we would need to wait on average.

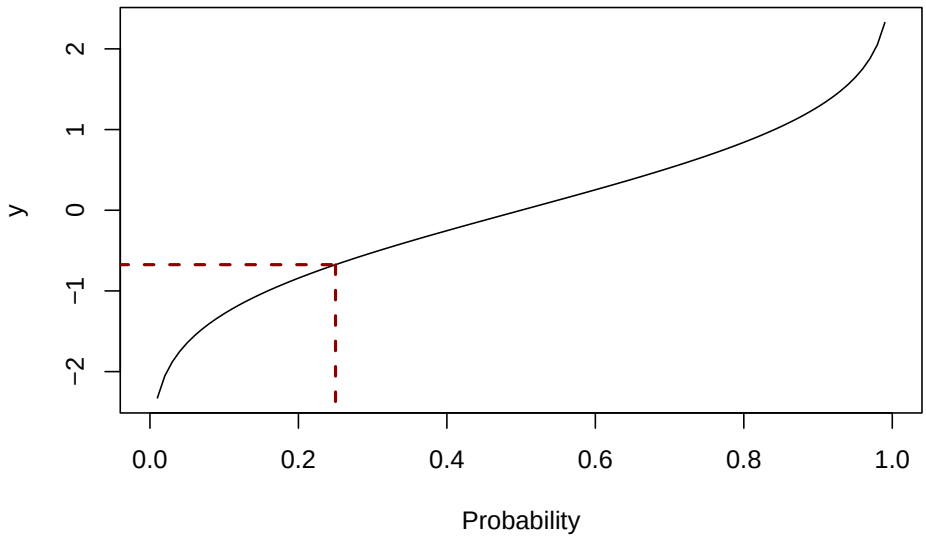


Figure 4.5: Quantile Function of Normal distribution

The contiunuous uniform distribution has similarities to the discrete one, which was discussed in Section 3.5, but due to the different nature of the random variable is parameterised differently. First, because we are discussing continuous variable, it is always defined on a segment of values, from a to b. For example, we can have a random variable which can take any value from 0 to 10 with equal likelihood. Mathematically, the PDF of continuous distribution can be written as:

$$f(y, b, a) = \begin{cases} \frac{1}{b-a} & \text{for } y \in [a, b] \\ 0 & \text{otherwise} \end{cases}. \tag{4.1}$$

It can be represented visually as shown in Figure 4.6.

According to this distribution, it is equally likely to have 1, 1.1, 1.0001, 9 etc values. The mean of this distribution coincides with the middle of the segment and can be calculated as:

$$\mathrm{E}(y) = \frac{1}{2}(a + b), \tag{4.2}$$

while the variance is calculated as:

$$\mathrm{V}(y) = \frac{1}{12}(b - a)^2. \tag{4.3}$$

The CDF of the Uniform distribution corresponds to the straight line going from the point (a,0) to the point (b,1), as shown in Figure 4.7.

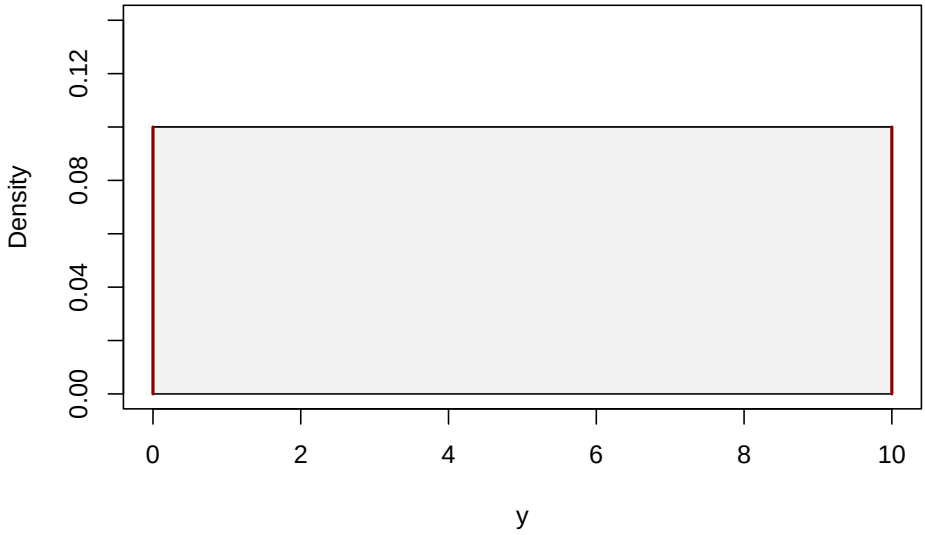


Figure 4.6: Probability Density Function of Continuous Uniform distribution.

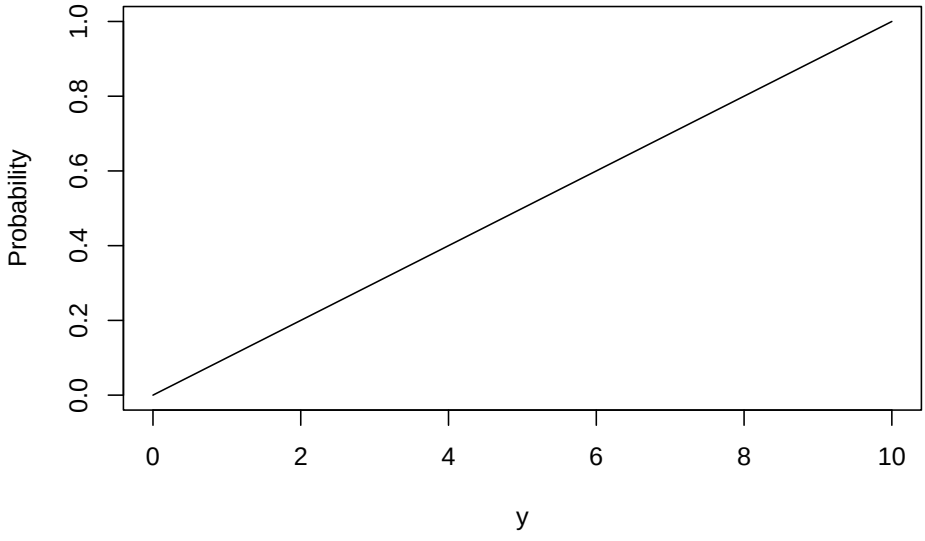


Figure 4.7: Cumulative Density Function of Continuous Uniform distribution.

Mathematically, the CDF can be represented as:

$$F(y,b,a) = \begin{cases} 0 & \text{if } y < a \\ \dfrac{y-a}{b-a} & \text{for } y \in [a,b] \\ 1 & \text{if } y > b \end{cases} . \tag{4.4}$$

Continuous uniform distribution is sometimes used in statistics as a prior, when a researcher does not have grounds to assume any other, more complicated distribution.

In R, this distribution is implemented in `stats` package with functions `dunif()`, `punif()`, `qunif()` and `runif()` for PDF, CDF, QF and random generator respectively.

4.3 Normal distribution

Every statistical textbook has Normal distribution. It is that one famous bell-curved distribution that every statistician likes because it is easy to work with and because it is an asymptotic distribution for many other well-behaved distributions under some conditions (see discussion of "Central Limit Theorem" in Section 6.2). For example, consider the the coin tossing example and Binomial distribution discussed in Section 3.3. If we toss the coin one time only, we get the Bernoulli distribution with two outcomes. If we do that ten times, the shape of distribution will change and there will be a score with higher probability than the others (that is $\mathcal{B}\rangle(10, 0.5)$). If we continue tossing the coin and do that for a hundred times, the shape of distribution will start converging to the bell-curve, reminding the Normal distribution. These three cases are shown in Figure 4.8.

This is one of the classical examples of a distribution converging to the Normal one with the increase of the number of trials n under some circumstances. This also tells us that in some circumstances we can use Normal distribution as an approximation of the real distribution: in Figure 4.8, the third graph corresponds to the $\mathcal{B}\rangle(100, 0.5)$, but it can be approximated by the normal distribution $\mathcal{N}(\mu_y, \sigma)$, where $\mu_y = 100 \times 0.5 = 50$ is the mean of the distribution and $\sigma^2 = 100 \times 0.5 \times 0.5 = 25$ is the variance (as discussed in Section 3.3), i.e. $\mathcal{N}(50, 25)$.

The probability density function (PDF) of the Normal distribution with some mean μ_y and variance σ^2 is:

$$f(y, \mu_y, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma}\right)^2\right), \tag{4.5}$$

This distribution can be represented in Figure 4.9.

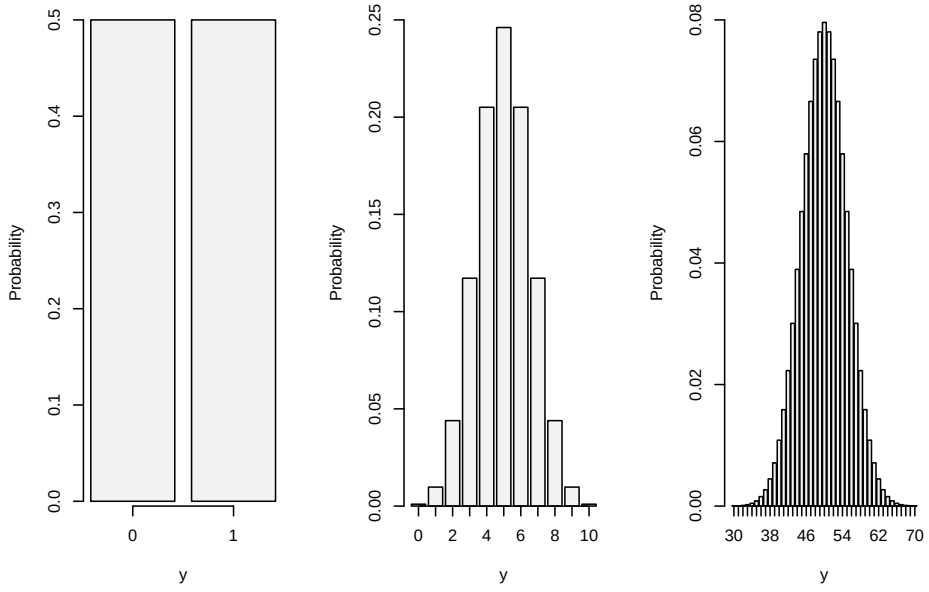


Figure 4.8: Probability Mass Functions for Binomial distribution with p=0.5 and n={1, 10, 100}.

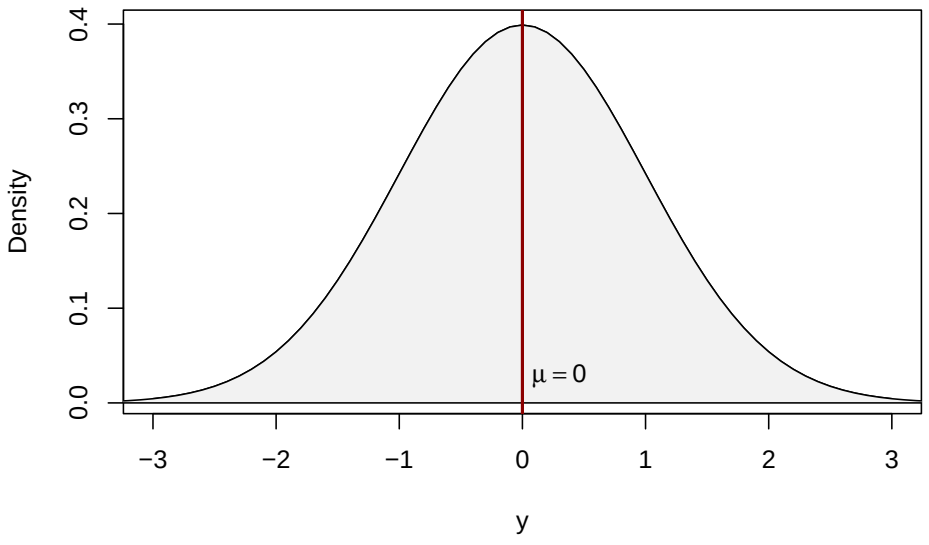


Figure 4.9: Probability Density Function of Standard Normal distribution

Figure 4.9 demonstrates a standard normal distribution, meaning that $\mu_y = 0$ and $\sigma^2 = 1$. A more general distribution with non-zero μ_y and a non-unity σ^2 will have the same shape, but will have different values on the axes. The shape itself demonstrates that there is a central tendency (in our case - the mean μ_y), around which the density of values is the highest and there are other potential cases, further away from the centre of distribution, but their probability of appearance reduces proportionally to the distance from the centre. As we can see from Figure 4.9, the Normal distribution is symmetric. It has skewness of zero and kurtosis of 3 (see discussion in Section 5.1).

When it comes to the cumulative distribution function, it has a form of an S-curve as shown in Figure 4.10.

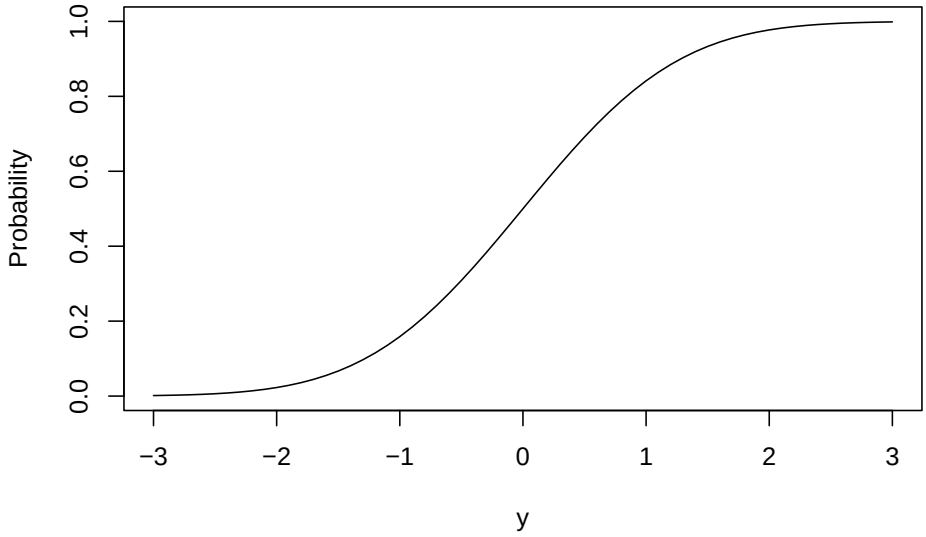


Figure 4.10: Cumulative Distribution Function of Standard Normal distribution

The CDF in Figure 4.10 has the properties of any other CDF: it converges to one with the increase of the value of y and reaches zero asymptotically with the decrease of the value of y. It can be used to solve problems of the style "what is the probability that y will lie between 20 and 30 for the $\mathcal{N}(20, 10)$". In R, this can be done by entering the values of y, μ and σ^2 in the following function (note that in R, the scale is the standard deviation σ, not the variance σ^2):

```
pnorm(q=30, mean=20, sd=10) - pnorm(q=20, mean=20, sd=10)
```

```
## [1] 0.3413447
```

which mathematically is typically represented as:

$$\Phi(y_2, \mu_y, \sigma^2) - \Phi(y_1, \mu_y, \sigma^2) = \Phi(30, 20, 100) - \Phi(20, 20, 100) = 0.841 - 0.5 \approx 0.341. \quad (4.6)$$

The CDF itself is difficult to summarise and involves a complicated integral:

$$\Phi(y, \mu_y, \sigma^2) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{y - \mu_y}{\sqrt{2\sigma^2}}\right)\right), \tag{4.7}$$

where $\text{erf(y)} = \frac{2}{\sqrt{\pi}} \int_0^{\text{y}} \text{e}^{-\text{x}^2} \text{dx}$ is the so called "error function''. This function does not have a closed form and is typically evaluated numerically or using some approximations. The probability (4.6) corresponds to the difference between the points shown in Figure 4.11.

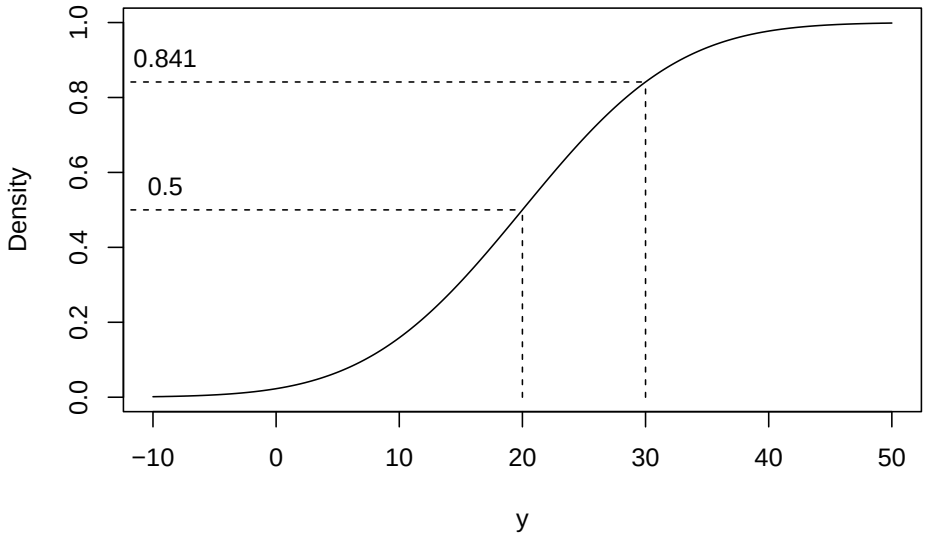


Figure 4.11: Values of CDF for the example with Normal distribution

The same number also corresponds to the dark area in PDF of the distribution as shown in Figure 4.12.

The relation between the area in Figure 4.12 and the difference between the two points in Figure 4.11 comes directly from the definition of CDF, the latter being the function of cumulative values of the density function.

Many distributions converge to the Normal one or can be approximated by it under some circumstances. For example, as shown earlier, Binomial distribution can be approximated by the normal in some cases. More specifically, the approximation works when $n > 20$, $np \geq 5$ and $n(1 - p) \geq 5$. Figure 4.13 demonstrates how the normal curve (the solid red line) approximates the barplots of the Binomial distribution.

As we can see from Figure 4.13, the normal curve fits the bars of the Binomial distribution very well, which means that we can use it, for example, to compute the probability that the variable y will be equal to 41, via the formula:

$$\text{P}(y = 41) \approx \Phi(41.5, 50, 25) - \Phi(40.5, 50, 25) \tag{4.8}$$

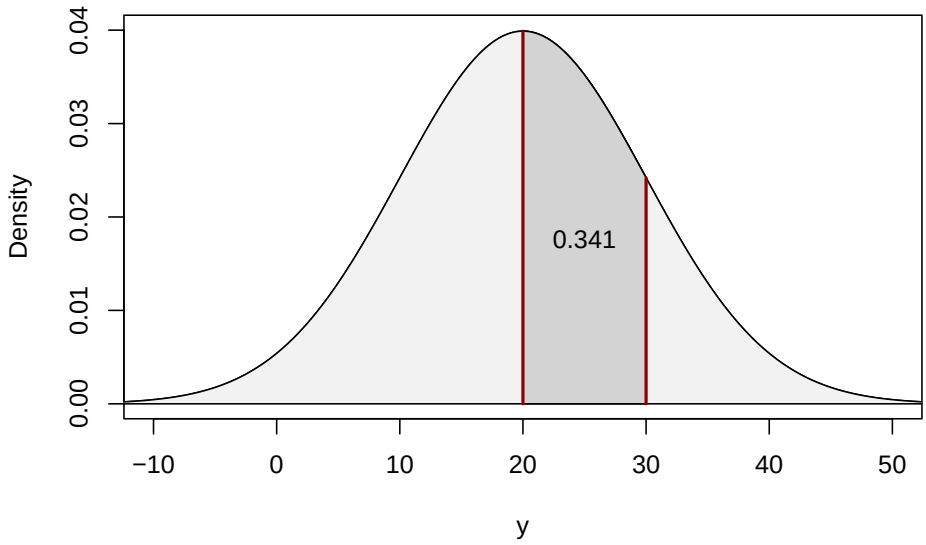


Figure 4.12: Values of PDF for the example with Normal distribution

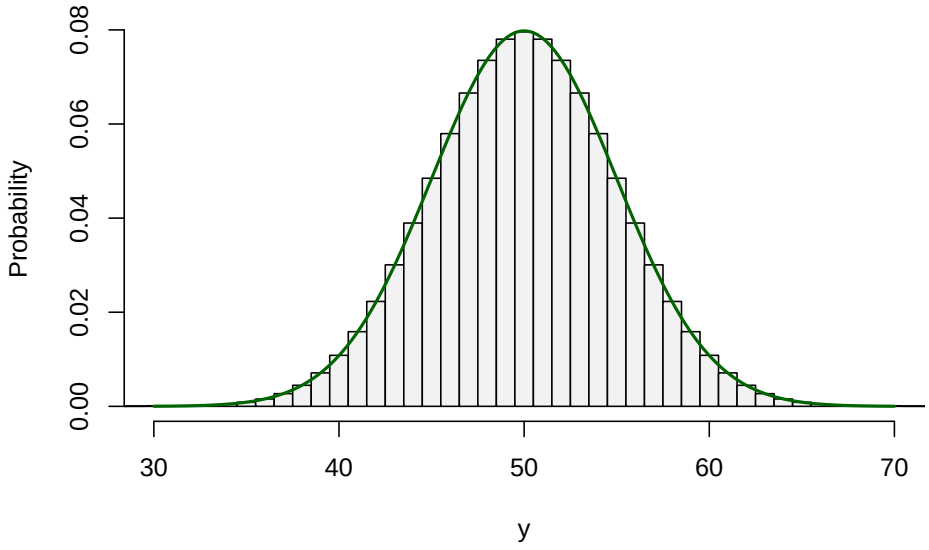


Figure 4.13: Binomial distribution and its approximation via the Normal one.

In the equation (4.8) by adding and subtraction 0.5, we calculate the surface of the area under the normal curve, corresponding roughly to the area of the respective bin in the barplot. We can see that this method of calculation gives a result very close to the one from the Binomial distribution:

```
c(dbinom(41,100,0.5),
  pnorm(41.5,50,5)-pnorm(40.5,50,5)) |>
    setNames(c("Binomial", "Normal"))
```

```
##   Binomial     Normal
## 0.01586907 0.01584890
```

This calculation based on the approximation is shown visually in Figure 4.13, where the second figure is the zoomed-in area in the first one. As we can see, the area under the curve of the Normal distribution is roughly equal to the area of the bar of the Binomial distribution.

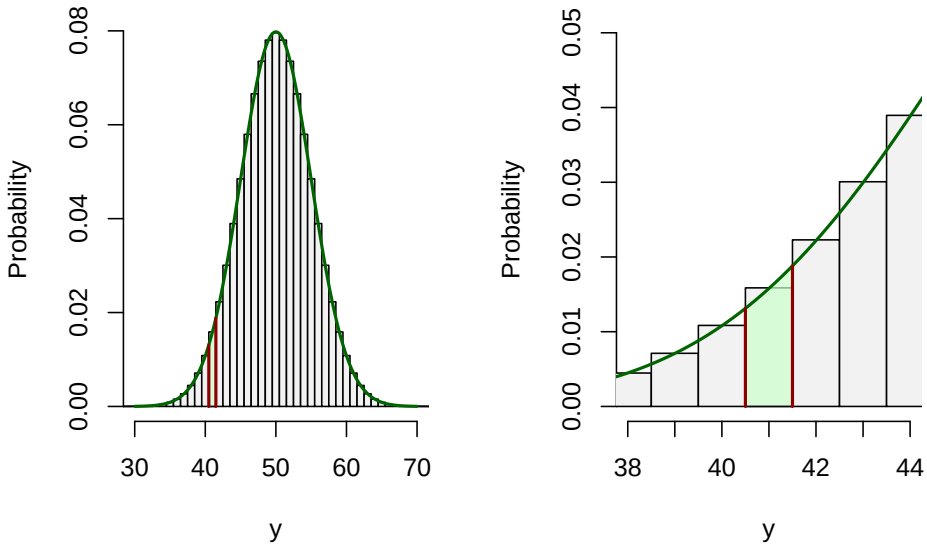


Figure 4.14: Calculating the probability based on Normal approximation.

Normal distribution is implemented in `dnorm()`, `qnorm()`, `pnorm()` and `rnorm()` functions from `stats` package in R.

Finally, as mentioned earlier, Normal distribution is popular among statisticians to use for the error term in a model of a type:

$$y_j = \mu_j + \epsilon_j, \tag{4.9}$$

where μ_j is some structure and $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$. The main reason for this is the ease of use of the distribution, and because it is described using its mean and variance. For example, based on (4.9), we can say that $y_j \sim \mathcal{N}(\mu_j, \sigma^2)$, because

as we remember from Chapter 2, $\mathrm{E}(y_j) = \mathrm{E}(\mu_j) + \mathrm{E}(\epsilon_j) = \mathrm{E}(\mu_j)$. In reality, the error term of a model might not follow normal distribution, it can be more complicated and sometimes might not follow any theoretical distribution.

4.4 Log-Normal distribution

Log-Normal distribution is closely related to the Normal one and is supported for the positive values of y. It is defined as a distribution arising after a transformation of a variable $x = \log(y)$ or equivalently $y = e^x$. It is said that $y = e^x \sim \log\mathcal{N}(\mu_y, \sigma^2)$ if $\log y = x \sim \mathcal{N}(\mu_y, \sigma^2)$. Figure 4.15 shows the connection between Normal and Log-Normal distributions. In that plot, we can see how the density changes because of the $y = e^x$ transformation.

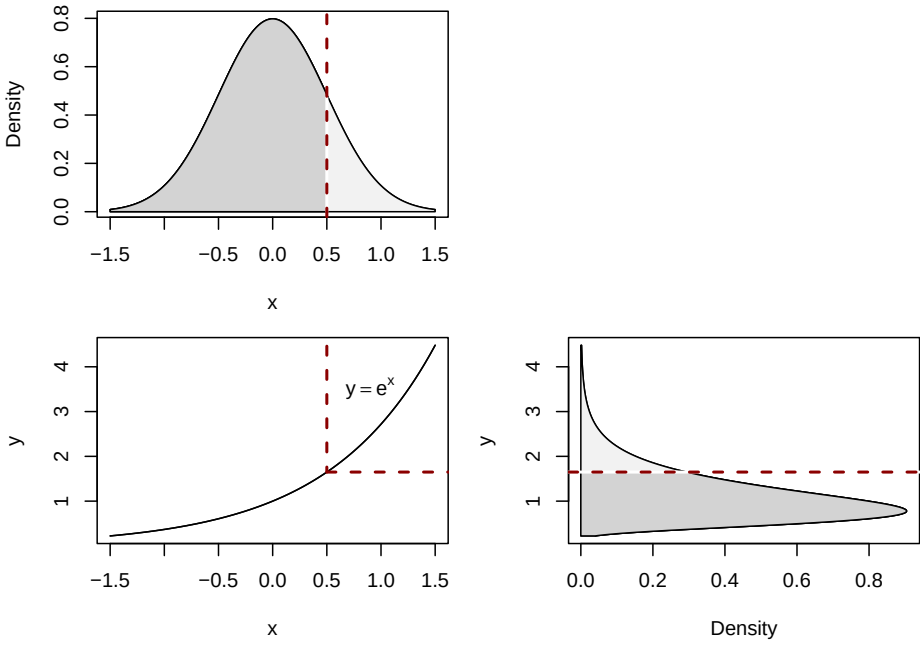


Figure 4.15: Connection between Normal and Log-Normal distributions

The dark areas on the plots in Figure 4.15 show equal probabilities for the Normal and the Log-Normal distributions obtained via specific quantiles. This demonstrates that in order to obtain a quantile of the Log-Normal distribution for y, we need to produce a quantile from the Normal one for x and then exponentiate the value.

Because of its shape and support of positive values only, the Log-Normal distribution is often used in multiplicative models of the style:

$$y_j = \mu_j \epsilon_j, \tag{4.10}$$

which is equivalent to:

$$\log y_j = \log \mu_j + \log \epsilon_j, \tag{4.11}$$

where $\epsilon_j \sim \log\mathcal{N}(0, \sigma^2)$ and $\log \epsilon_j \sim \mathcal{N}(0, \sigma^2)$. Log-Normal distribution is also used to model, for example, prices or income of households. When talking about the latter, conceptually, we expect it to have asymmetric distribution, because there will be a lot of households with low income and few with very high ones. Log-Normal distribution can be considered as a reasonable model in this case.

The PDF of the Log-Normal distribution is written mathematically as:

$$f(y, \mu_y, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{\log y - \mu_y}{\sigma}\right)^2\right). \tag{4.12}$$

Several PDFs of Log-Normal distribution are shown in Figure 4.16.

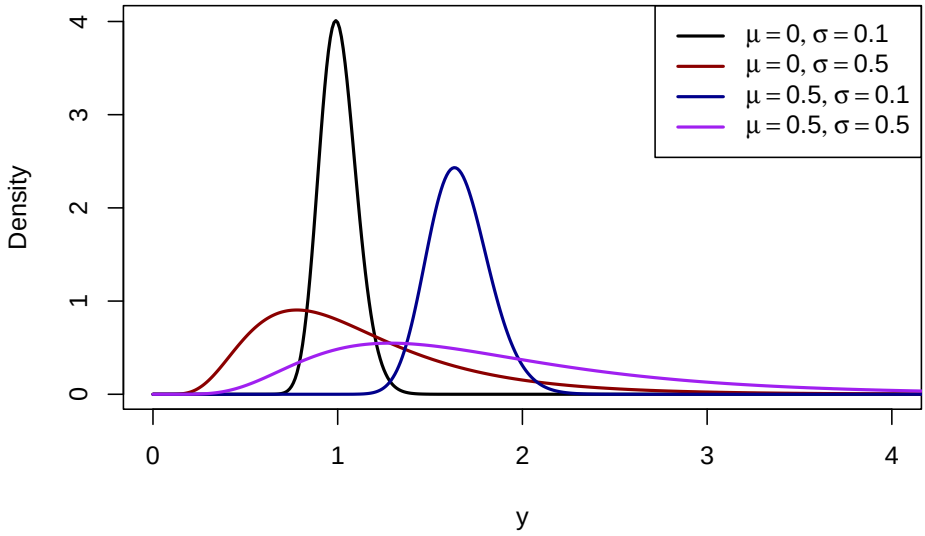


Figure 4.16: Probability Density Function of Log-Normal distribution with a variety of parameters.

The Figure 4.16 shows that with the increase of the location parameter μ, the distribution shifts to the right, while with the increase of the scale parameter σ it becomes more asymmetric with a longer right tail and its mode moves closer to zero. In fact, the skewness of the Log-Normal distribution depends solely on the value of σ^2 - the higher it is, the more skewed the distribution is. It can be calculated as:

$$\text{Sk}(y) = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1}. \tag{4.13}$$

Log-Normal distribution is supported by `dlnorm()`, `plnorm()`, `qlnorm()` and `rlnorm()` functions from `stats` package in R.

4.5 Exponential distribution

We have already touched upon the Exponential distribution, when we discussed the arrival times in Poisson distribution (Section 3.4). Exponential distribution is used in modelling time between arrivals, because it is memoryless. We mentioned earlier that if a process is memoryless, then the following holds:

$$\mathrm{P}(t > \tau_1 + \tau_2) = \mathrm{P}(t > \tau_1)\mathrm{P}(t > \tau_2).$$

Exponential distribution relies on this property. It has only one parameter, the rate λ and can be written as $\mathcal{E}\S\sqrt{}(\lambda)$. Here is its PDF:

$$f(t, \lambda) = \lambda e^{-\lambda t}, \tag{4.14}$$

where t is a positive number and λ is the rate parameter. This PDF is shown in Figure 4.17.

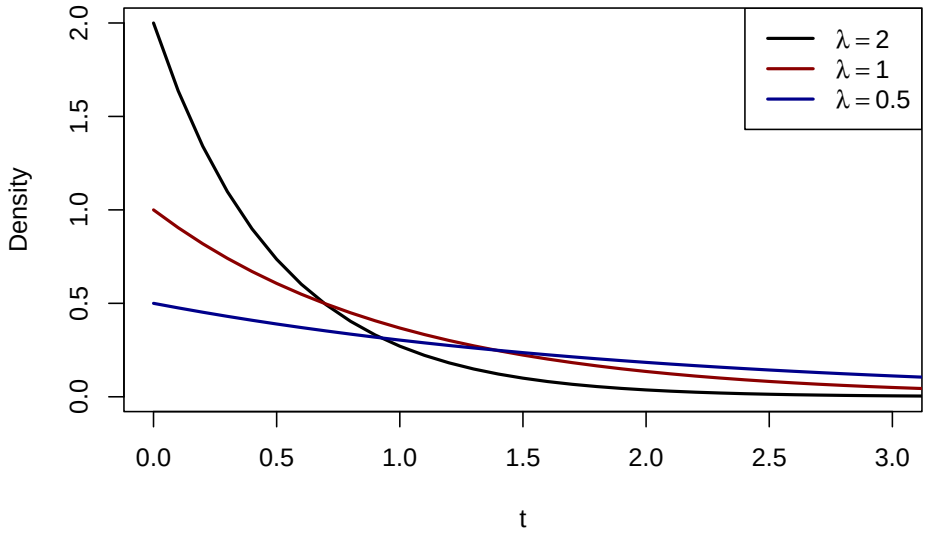


Figure 4.17: Probability Density Function of Exponential distribution with several values of rate parameter λ.

The plot in Figure 4.17 shows that there is more likely for the variable t to get lower values (closer to zero) than the higher ones. The parameter λ controls the steepness of decline of the density curve with increase of t: the higher the rate is, the more likely it is that the event will occur earlier and less likely that it will occur later.

The CDF of the distribution is shown in Figure 4.18.

The CDFs show how fast the probability of one is achieved with different rates. Mathematically, it is written as:

$$\mathrm{F}(t, \lambda) = 1 - e^{-\lambda t}. \tag{4.15}$$

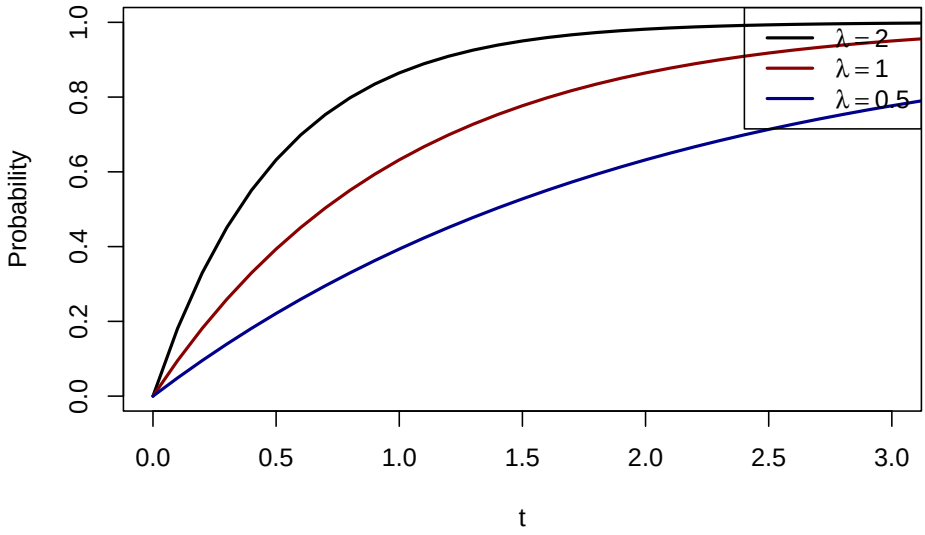


Figure 4.18: Cumulative Distribution Function of Exponential distribution with a variety of rate parameters.

Based on it, we can say, for example what is the probability that an event will occur in 1.5 seconds if the rate is $\lambda = 2$ per second. It is:

$$\mathrm{F}(t < 1.5, 2) = 1 - e^{-2 \times 1.5} \approx 0.95.$$

This can also be calculated in R:

```
pexp(1.5, rate=2)
```

```
## [1] 0.9502129
```

Coming back to the memorylessness property of the distribution, in order to show it visually, consider an example based on the property:

$$\mathrm{P}(t > 1.5) = \mathrm{P}(t > 0.5)\mathrm{P}(t > 1)$$

and $\lambda = 2$. Note that $P(t > a) = 1 - F(a)$ by definition of CDF. In terms of probabilities of Exponential distribution, this means that:

$$\mathrm{P}(t > 1.5) = (1 - \mathrm{F}(1, 2))(1 - \mathrm{F}(0.5, 2)).$$

Now, in order to show this property visually, we take logarithms of the left and right hand sides of the previous equation to get:

$$\log\left(\mathrm{P}(t > 1.5)\right) = \log\left(1 - \mathrm{F}(1, 2)\right) + \log\left(1 - \mathrm{F}(0.5, 2)\right). \tag{4.16}$$

We take logarithms to linearise relation, because it is easier to work with. If we now insert (4.15) in (4.16), we will get:

$$\log\left(\mathrm{P}(t > 1.5)\right) = \log\left(1 - 1 + e^{-2\times 1}\right) + \log\left(1 - 1 + e^{-2\times 0.5}\right) = -2 - 1 = -3, \tag{4.17}$$

which is obtained independently for $\log(1 - F(1.5, 2)) = \log(e^{-2\times 1.5}) = -3$. We can also plot the function $\log(1 - F(t, 2))$ in Figure 4.19 to show that (4.17) holds, when the values on y-axis are added.

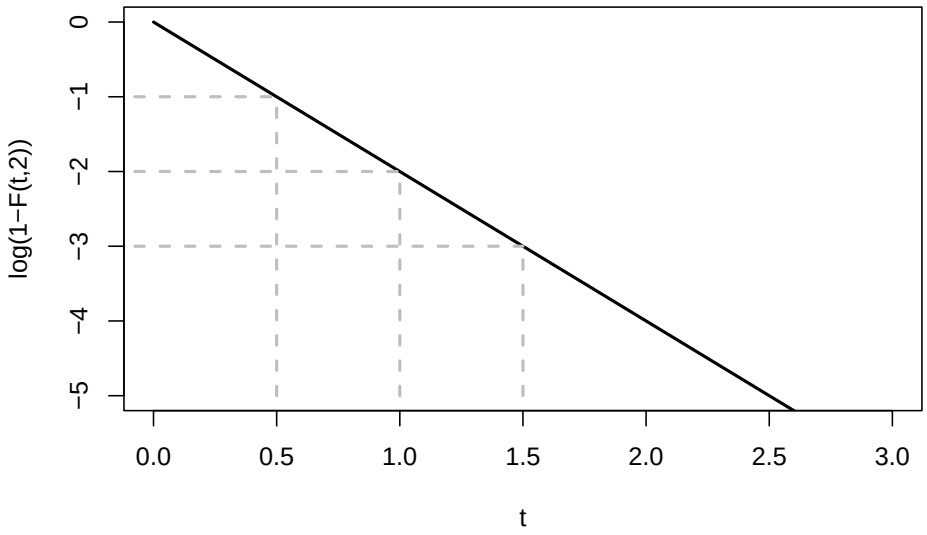


Figure 4.19: Memoryless property of Exponential distibution with $\lambda = 2$.

Note that because of the logarithm of the CDF, the function has now become linear, and the property becomes more apparent.

When it comes to the mean of the distribution, it is equal to:

$$\mathrm{E}(t) = \frac{1}{\lambda}, \tag{4.18}$$

which makes it easy to estimate on a sample of observations – just take the mean, and you get an estimate of the rate parameter λ. Furthermore, the rate parameter in Exponential distribution is directly related to the one in the Poisson (Section 3.4): the parameter in the latter equals the ratio of time interval t and λ:

$$\lambda_P = \frac{t}{\lambda}, \tag{4.19}$$

where λ_P is the parameter of the Poisson distribution. The main difference between the distributions is that Poisson explains the number of arrivals over a fixed period of time, while the Exponential represents the time between the arrivals.

Finally, the variance in Exponential distribution is calculated as:

$$\mathrm{V}(t) = \frac{1}{\lambda^2}, \tag{4.20}$$

In R, the Exponential distribution is implemented in `dexp()`, `pexp()`, `qexp()` and `rexp()` for PDF, CDF, QF and Random variables respectively.

Chapter 5

Preliminary data analysis

One of the basic thing that is worth doing before starting any modelling is the preliminary data analysis. This can be done either using numerical or graphical analysis. The former is useful when you want to have a summary information about the data without trying to find detailed information about it. The latter is useful when you can spend more time, investigating relations and issues in the data. In many cases, they compliment each other.

5.1 Numerical analysis

In this section we will use the classical `mtcars` dataset from `datasets` package for R. It contains 32 observations with 11 variables. While all the variables are numerical, some of them are in fact categorical variables encoded as binary ones. We can check the description of the dataset in R:

```
?mtcars
palette("default")
```

Judging by the explanation in the R documentation, the following variables are categorical:

1. vs - Engine (0 = V-shaped, 1 = straight),
2. am - Transmission (0 = automatic, 1 = manual).

In addition, the following variables are integer numeric ones:

1. cyl - Number of cylinders,
2. hp - Gross horsepower,
3. gear - Number of forward gears,
4. carb - Number of carburetors.

All the other variables are continuous numeric.

Takign this into account, we will create a data frame, encoding the categorical variables as factors for further analysis:

```
mtcarsData <- data.frame(mtcars)
mtcarsData$vs <- factor(mtcarsData$vs,levels=c(0,1),labels=c("V-shaped","Straight"))
mtcarsData$am <- factor(mtcarsData$am,levels=c(0,1),labels=c("automatic","manual"))
```

Given that we only have two options in those variables, it is not compulsory to do this encoding, but it will help us in the further analysis.

We can start with the basic summary statistics. We remember from the scales of information (Section 1.2) that the nominal variables can be analysed only via frequencies, so this is what we can produce for them:

```
table(mtcarsData$vs)
```

```
##
## V-shaped Straight
##       18       14
```

```
table(mtcarsData$am)
```

```
##
## automatic    manual
##        19        13
```

These tables are called **contingency tables**, they show the frequency of appearance of values of variables. Based on this, we can conclude that the cars with V-shaped engine are met more often in the dataset than the cars with the Straight one. In addition, the automatic transmission prevails in the data. The related statistics which is useful for analysis of categorical variables is called **mode**. It shows which of the values happens most often in the data. Judging by the frequencies above, we can conclude that the mode for the first variable is the value "V-shaped".

All of this is purely descriptive information, which does not provide us much. We could probably get more information if we analysed the contingency table based on these two variables:

```
table(mtcarsData$vs,mtcarsData$am)
```

```
##
##            automatic manual
##   V-shaped        12      6
##   Straight         7      7
```

For now, we can only conclude that the cars with V-shaped engine and automatic transmission are met more often than the other cars in the dataset.

Next, we can look at the numerical variables. As we recall from Section 1.2, this scale supports all operations, so we can use quantiles, mean, standard deviation

etc. Here how we can analyse, for example, the variable mpg:

```
setNames(mean(mtcarsData$mpg),"mean")
```

```
##     mean
## 20.09062
```

```
quantile(mtcarsData$mpg)
```

```
##     0%    25%    50%    75%   100%
## 10.400 15.425 19.200 22.800 33.900
```

```
setNames(median(mtcarsData$mpg),"median")
```

```
## median
##   19.2
```

The output above produces:

1. **Mean** - the average value of mpg in the dataset, $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$.
2. **Quantiles** - the values that show, below which values the respective proportions of the dataset lie. For example, 25% of observations have mpg less than 15.425. The 25%, 50% and 75% quantiles are also called 1st, 2nd and 3rd **quartiles** respectively.
3. **Median**, which splits the sample in two halves. It corresponds to the 50% quantile.

If median is greater than mean, then this typically means that the distribution of the variable is skewed (it has some rare observations that have large values).

Making a step back, we also need to mention the **variance**, which shows the overall variability of the variable around its mean:

$$\mathrm{V}(y) = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2. \tag{5.1}$$

Note that the division in (5.1) is done by $n-1$, and not by n. This is done in order to correct the value for the in-sample bias (we will discuss this in Subsection 6.3.1). The number itself does not tell us much about the variability, but having it allows calculating other more advanced measures. Typically the square root of variance is used in inferences, because it is measured in the same scale as the original data. It is called **standard deviation**:

$$\hat{\sigma} = \sqrt{\mathrm{V}(y)}, \tag{5.2}$$

it has the same scale as the variable y_j. In our example, both can be obtained via:

```
var(mtcarsData$mpg)
```

```
## [1] 36.3241
```

```
sd(mtcarsData$mpg)
```

```
## [1] 6.026948
```

Visually, standard deviation can be represented as a straight line, depicting the overall variability of the data (Figure 5.1).

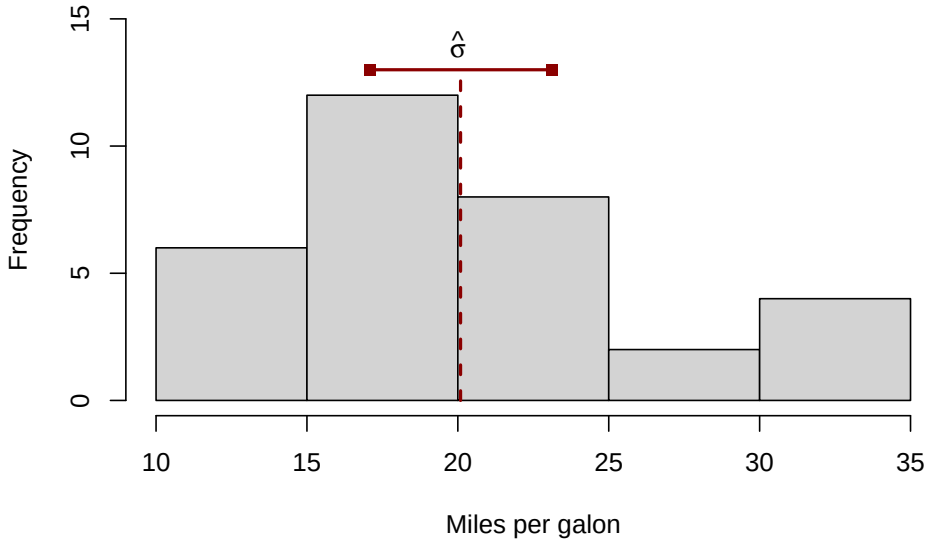


Figure 5.1: Visual presentation of standard deviation. The value of standard deviation corresponds to the segment between squares at the top of the histogram.

In Figure 5.1, we depict the distribution of the variable `mpg` using histogram (see Section 5.2) and show how the standard deviation relates to it – it is equal to the length of the segment above the histogram.

Coming back to the point about the asymmetry of the distribution of a variable in our example, we can investigate it further using skewness and kurtosis from `timeDate` package:

```
timeDate::skewness(mtcarsData$mpg)
```

```
## [1] 0.610655
## attr(,"method")
## [1] "moment"
```

```
timeDate::kurtosis(mtcarsData$mpg)
```

```
## [1] -0.372766
## attr(,"method")
## [1] "excess"
```

Skewness shows the asymmetry of distribution. If it is greater than zero, then the distribution has the long right tail. If it is equal to zero, then it is symmetric. It is calculated as:

$$\text{skew}(y) = \frac{1}{n}\sum_{j=1}^{n}\frac{(y_j - \bar{y})^3}{\hat{\sigma}^3}. \tag{5.3}$$

Kurtosis shows the excess of distribution (fatness of tails) in comparison with the normal distribution. If it is equal to zero, then it is the same as for the normal distribution. Here how it is calculated:

$$\text{kurt}(y) = \frac{1}{n}\sum_{j=1}^{n}\frac{(y_j - \bar{y})^4}{\hat{\sigma}^4} - 3. \tag{5.4}$$

Note that there is 3 in the formula. This is because the excess (which is the value without 3) of Normal distribution is equal to 3. Instead of dividing the value by 3 (which would make kurtosis easier to interpret), Karl Pearson has decided to use subtraction. The same formula (5.4) can be rewritten as:

$$\text{kurt}(y) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{y_j - \bar{y}}{\hat{\sigma}}\right)^4 - 3. \tag{5.5}$$

Analysing the formula (5.5), we see that the impact of the deviations lying inside one $\hat{\sigma}$ bounds are reduced, while those that lie outside, are increased. e.g. if the value $y_j - \bar{y} > \hat{\sigma}$, then the value in the ratio will be greater than one (e.g. 2), thus increasing the final value (e.g. $2^4 = 16$), while in the opposite case of $y_j - \bar{y} < \hat{\sigma}$, the ratio will be diminished (e.g. $0.5^4 = 0.0625$). After summing up all n values in (5.5), the values outside one $\hat{\sigma}$ will have a bigger impact on the resulting kurtosis than those lying inside. So, kurtosis will be higher for the distributions with longer tails and in the cases when there are outliers in the data.

Based on all of this, we can conclude that the distribution of `mpg` is skewed and has the longer right tail. This is expected for such variable, because the cars that have higher mileage are not common in this dataset.

All the conventional statistics discussed above can be produced using the following summary for all variables in the dataset:

```
summary(mtcarsData)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs              am
```

```
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   V-shaped:18   automatic:19
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   Straight:14   manual   :13
##  Median :3.695   Median :3.325   Median :17.71
##  Mean   :3.597   Mean   :3.217   Mean   :17.85
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90
##  Max.   :4.930   Max.   :5.424   Max.   :22.90
##       gear            carb
##  Min.   :3.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000
##  Median :4.000   Median :2.000
##  Mean   :3.688   Mean   :2.812
##  3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :8.000
```

Finally, one other moment that is often used in statistics is called "covariance". It measures the joint variability of two variables and is calculated as:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}). \tag{5.6}$$

Covariance is one of the more complicated moments to explain. It is equal to zero if one of the variables does not have any variability and otherwise can take any real value. We will discuss its meaning later in Section 10.2.

5.2 Graphical analysis

5.2.1 One categorical/discrete variable

Continuing our example with `mtcars` dataset, we now investigate what plots can be used for different types of data. As discussed earlier, we have two categorical variables: vs and am - and they need to be treated differently than the numerical ones. We can start by producing their barplots:

```
barplotVS <- barplot(table(mtcarsData$vs), xlab="Type of engine")
text(barplotVS,table(mtcarsData$vs)/2,table(mtcarsData$vs),cex=1.25)
```

This is just a graphical presentation of the contingency table we have already discussed earlier.

Remark. Histograms do not make sense in case of categorical variables, because they assume that variables are numerical and continuous (see Section 2.2) - they will split the values of a variable in the bins, based on the idea that the variable can take any of the values in each bin.

Barplots are useful when you deal with either categorical variables or integer numerical ones. Here is what we can produce in case of the integer variable `cyl`:

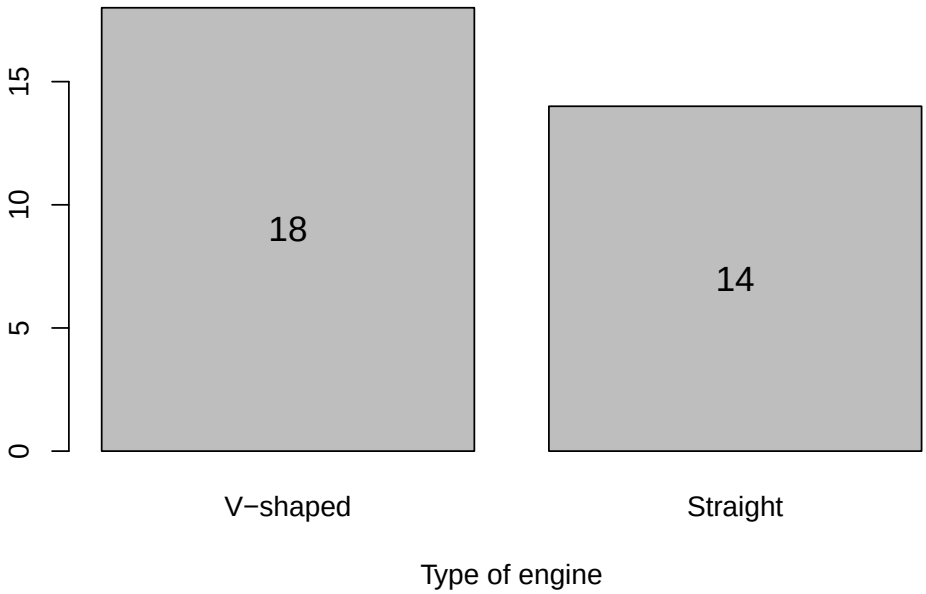


Figure 5.2: Barplot for the engine type.

```
barplotCYL <- barplot(table(mtcarsData$cyl), xlab="Number of cylinders")
text(barplotCYL,table(mtcarsData$cyl)/2,table(mtcarsData$cyl),cex=1.25)
```

Figure 5.3 shows that there are three types of cars in the data: with 4, 6 and 8 cylinders. The most frequently met is the car with 8 cylinders. Judging by the plot, half of cars have not more than 6 cylinders (median is equal to 6). All of this can be deducted from the barplot. And here how the histogram would look like for cylinders:

```
hist(mtcarsData$cyl)
```

Figure 5.4 is difficult to read, because on histogram, the bars show frequency at which continuous variable appears in pre-specified bins. In our case we would conclude that the most frequently cars in the dataset are those that have 7.5 - 8 cylinders, which is wrong and misleading. In addition, this basic plot does not have a readable label for x-axis and a meaningful title (in fact, we do not need one, given that we have caption). So, always label your axis and make sure that the text on plots is easy to understand for those people who do not work with the data.

5.2.2 Two categorical/discrete variables

Coming back to categorical variables, we can construct two-dimensional plots to investigate potential relations between variables. We will first try the same

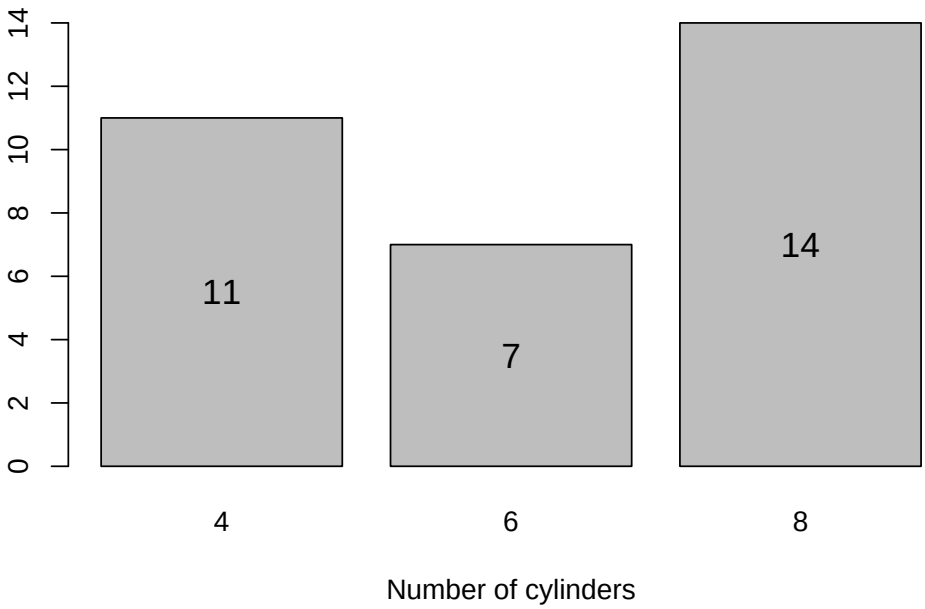


Figure 5.3: Barplot for the number of cylinders.

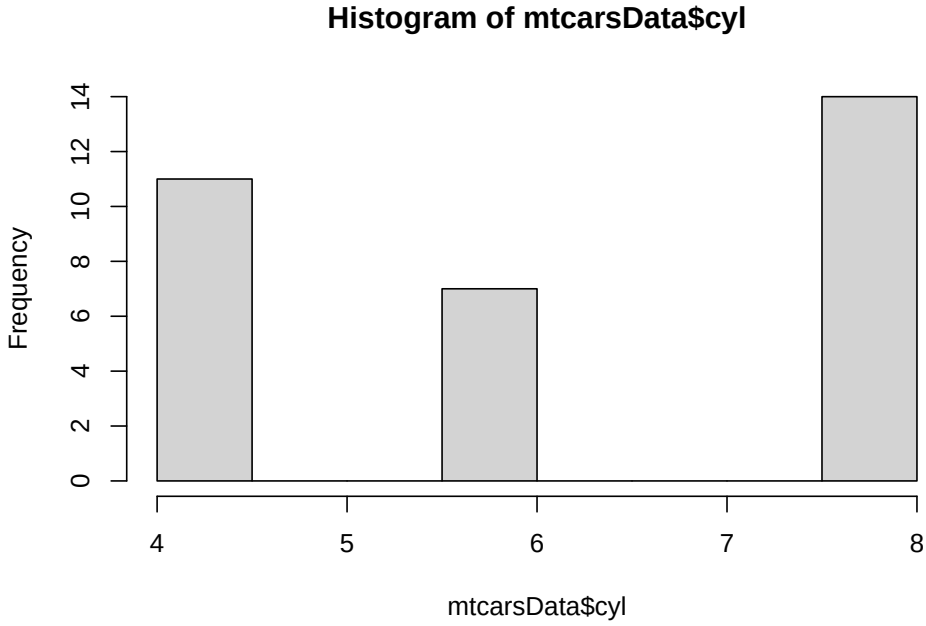


Figure 5.4: Histogram for the number of cylinders. Do not do this!

barplot as above, but with vs and am variables:

```
barplot(table(mtcarsData$vs,mtcarsData$am),
        xlab="Type of transmission", legend.text=levels(mtcarsData$vs))
```

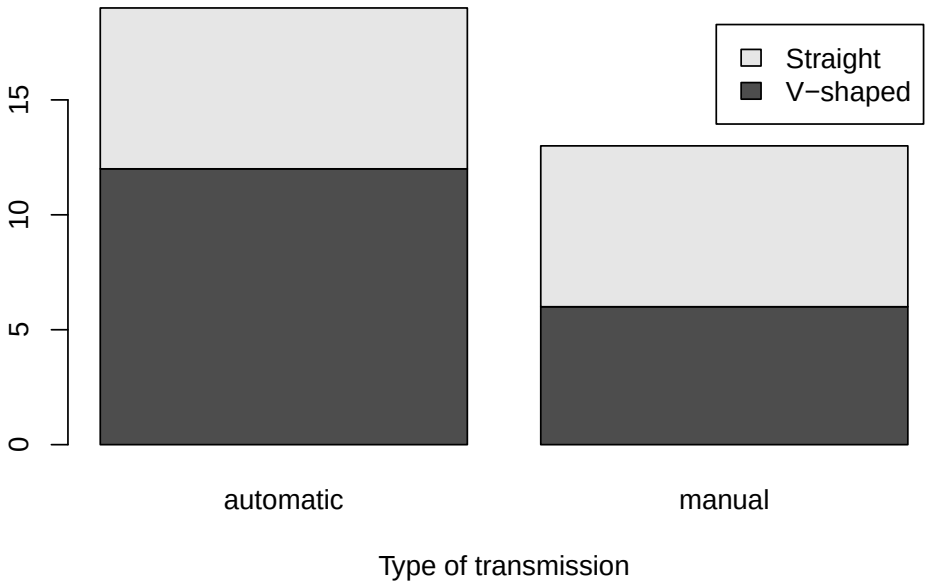


Figure 5.5: Barplot for the type of engine and transmission.

Figure 5.5 provides some information about the distribution of type of engine and transmission. For example, we can say that the most often met car in the dataset is the one with automatic transmission and V-shaped engine. However, it is not possible to say much about the relation between the two variables based on this plot. So, there is an alternative presentation, which uses the heat map (`tableplot()` from `greybox`):

```
tableplot(mtcarsData$vs,mtcarsData$am,
          xlab="Type of engine", ylab="Type of transmission")
```

The idea of this plot is that the darkness of areas shows the frequency of occurrence of each specific value. This message is duplicated by the number of dots in the plot (the more dots there are, the more observations there are in that specific area). The numbers inside the box show the proportions for each answer. So, we can conclude (again), that automatic transmission with V-shaped engine is met in 37.5% of cases. On the other hand, the least frequent type of car is the one with V-shaped engine and manual transmission. There might be some tendency in the dataset: the engine and transmission might be related (v-shaped with automatic vs Straigh with manual) - but it is not very well pronounced. The same table plot can be used for the analysis of relations between integer variables (and categorical). Here, for example, the plot between the number of

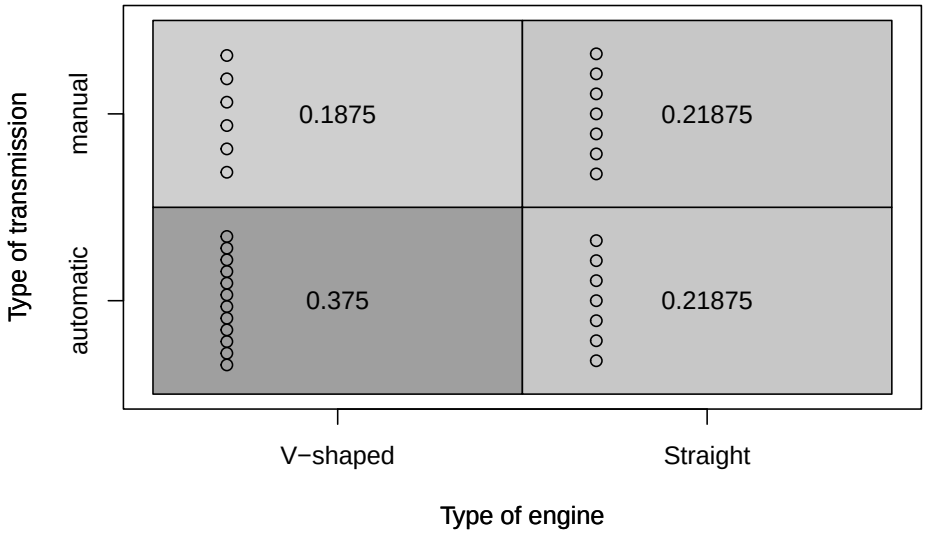


Figure 5.6: Heat map for the type of engine and transmission.

cylinders and the type of engine:

```
tableplot(mtcarsData$cyl,mtcarsData$vs,
          xlab="Number of cylinders", ylab="Type of engine")
```

Figure 5.7 allows making more solid conclusions about the relation between the two variables: we see that with the increase of the number of cylinders, the cars tend to switch from Straight to the V-shaped engines. This has an explanation: the engines with more cylinders need to have a different geometry to fit them all, and the V shape is more suitable for them. The table plot shows clearly this relation between the two variables.

5.2.3 One numerical continuous variable

Next, we can analyse the numerical continuous variables. We can start with the basic histogram:

```
hist(mtcarsData$wt, xlab="Weight", main="", probability=TRUE)
lines(density(mtcarsData$wt), col="red3")
```

The histogram 5.8 shows that there is a slight skewness in the data: the cars with weight from 3 to 4 thousands pounds are met more often than the cars with more than 5. The left tail of this distribution is slightly longer than the right one. Note that I have produced the probabilities on the y-axis of the plot in order to add the density curve, which smooths out the frequencies and shows how the distribution looks like.

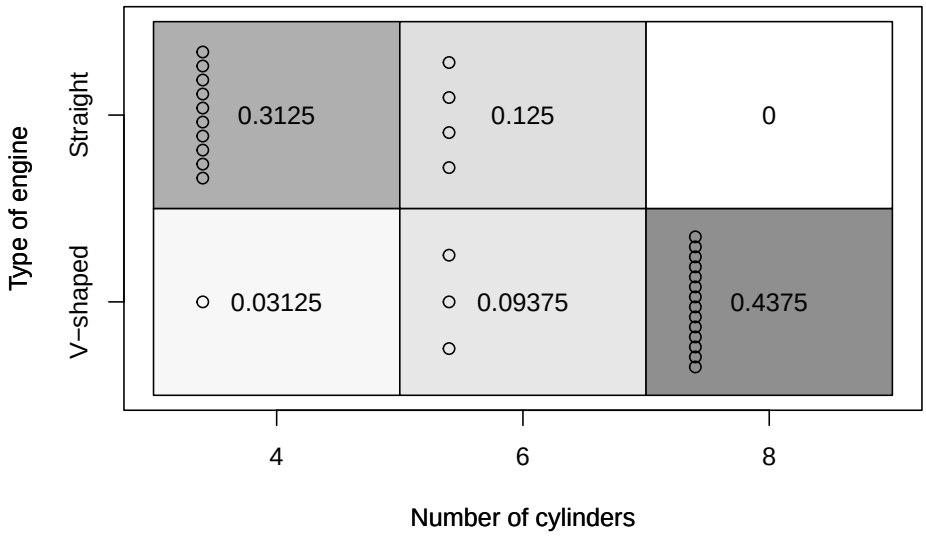


Figure 5.7: Heat map for the number of cylinders and the type of engine.

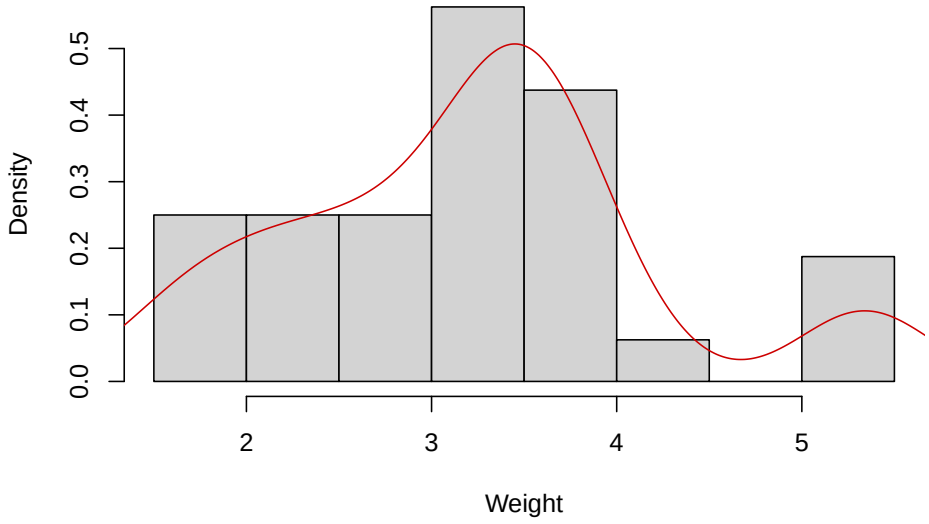


Figure 5.8: Distribution of the weights of cars.

An alternative presentation of the histogram is the boxplot, which graphically presents quantiles of distribution:

```
boxplot(mtcarsData$wt, ylab="Weight")
points(mean(mtcarsData$wt),col="red3", pch=16)
```

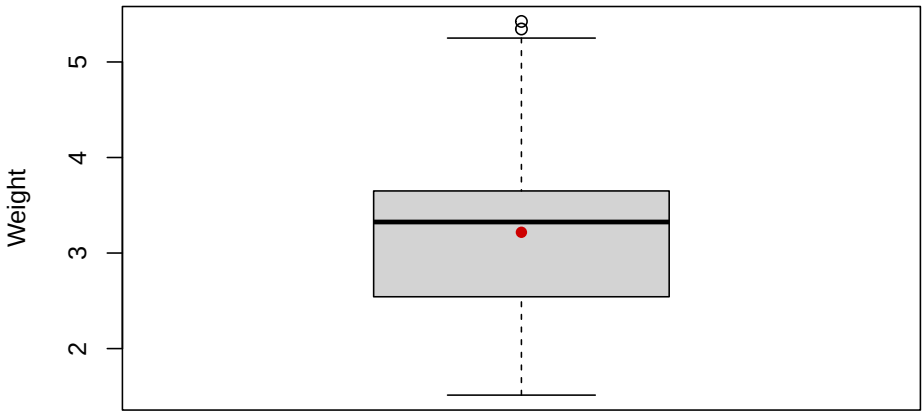


Figure 5.9: Boxplot of the variable weight.

This plot has the box in the middle, the whiskers on the sides, points at the top and the red point at the centre. The box shows 1st, 2nd and 3rd quartiles of distribution, thus the black line in the middle is the median. The distance between the 1st and the 3rd quartiles is called "Interquartile range" (IQR) and is used for the calculation of the interval (1st / 3rd quartile $\pm 1.5 \times$IQR), which corresponds roughly to the 99.3% interval (read more about this in Section **??**) from Normal distribution and is graphically drawn as the furthest observation in the interval. So, the lower whisker on our plot corresponds to the minimum value in the data, which is still in the interval, while the upper whisker corresponds to the bound of the interval. All the observations that lie beyond the interval are marked as potential outliers. Note that *this does not mean that the values are indeed outliers*, they just lie outside the 99.3% interval of Normal distribution. Finally, the red dot was added by me to show where the mean is. It is lower than median, this implies that there is a slight skewness in the distribution of weight.

There is also a way for producing the plots that would combine elements of histogram, density curve and boxplot. There is a plot called "violin plot". We will use `vioplot()` function from `vioplot` package in order to produce them:

```
vioplot(mtcarsData$wt, ylab="Weight")
points(mean(mtcarsData$wt),col="red3", pch=16)
```

Figure 5.10 unites the boxplot and the density curve from the plots above, providing not only information about the quantiles, but also about the shape of the distribution.

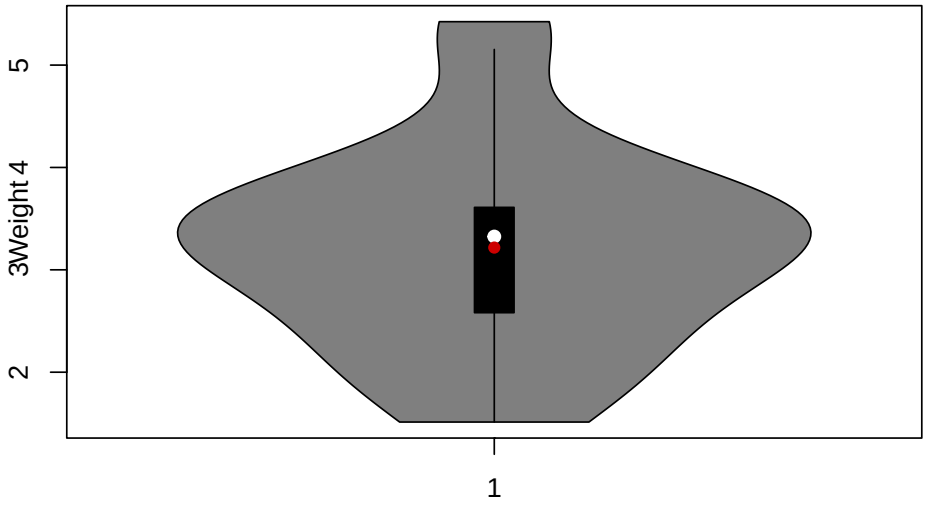


Figure 5.10: Violin plot together with boxplot of the variable weight.

Finally, if we want to compare the distribution of a variable with a known theoretical distribution, we can produce the QQ-plot. Here how it looks for Normal distribution:

```
qqnorm(mtcarsData$wt)
qqline(mtcarsData$wt)
```

The idea of the plot on Figure 5.11 is to compare theoretical quantiles with the empirical ones. If the variable would follow the specific distribution, then all the points would lie on the solid line. In our case, they do not: there are points in the right tail that are very far from the line - so we would conclude that the distribution of weight does not look Normal.

5.2.4 Two continuous numerical variables

So far, we have discussed the univariate analysis of numerical variables, but we can also produce plots showing potential relations between them. We start with the classical scatterplot:

```
plot(mtcarsData$wt, mtcarsData$mpg, xlab="Weight", ylab="Mileage")
lines(lowess(mtcarsData$wt, mtcarsData$mpg), col="red3")
```

The plot on Figure 5.12 shows the observations that have specific weight and mileage. Based on this, we can see if there is a relation between variables or not and what sort of relation this is. In order to simplify analysis, I have added the lowess line to the plot. It smooths the relation between variables, drawing the smooth line through the points and helps in understanding the existing relations in the data. Judging by Figure 5.12, there is a negative, slightly non-linear

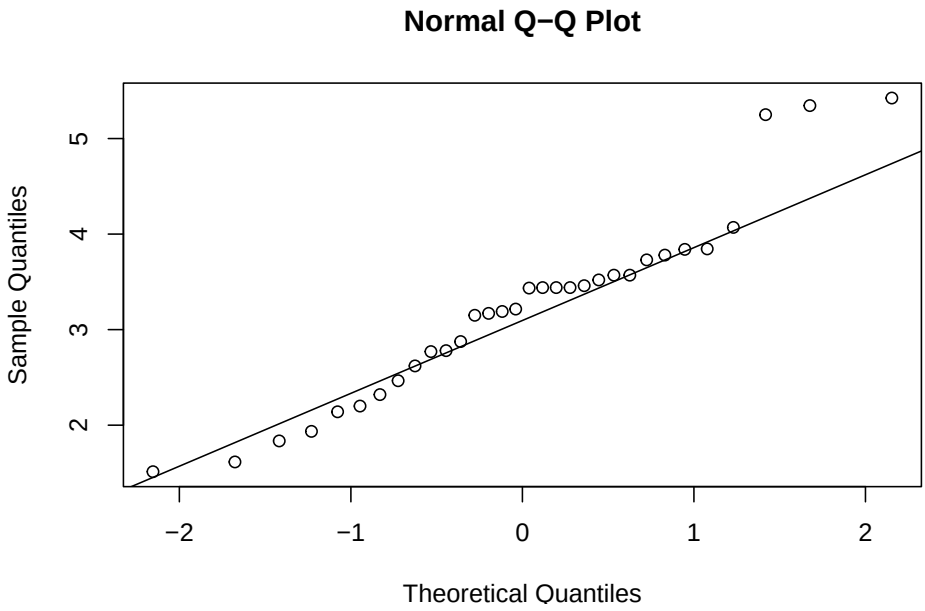


Figure 5.11: QQ plot of Normal distribution for variable weight.

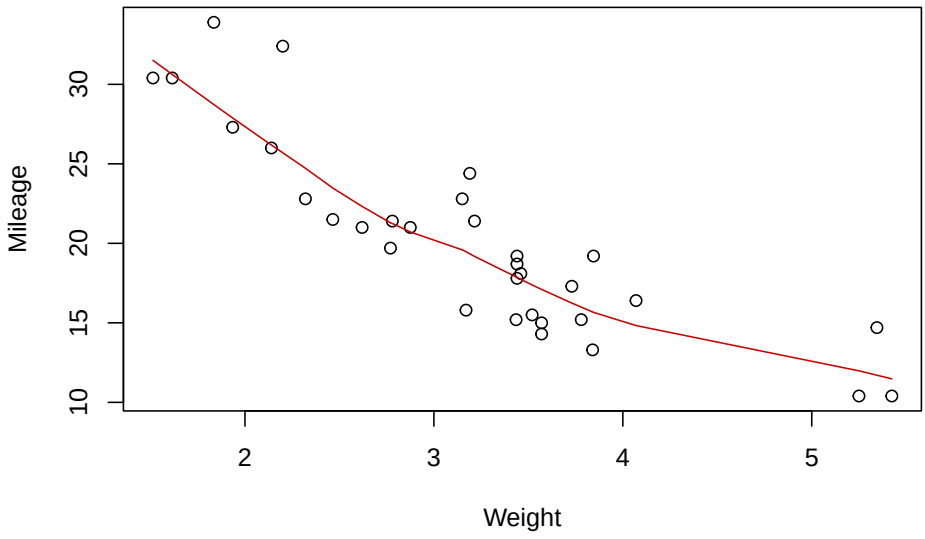


Figure 5.12: Scatterplot diagram between weight and mileage.

relation between the variables: the mileage decreases with reduced speed, when weight of a car increases. This relation makes sense, because heavier cars will consume more fuel and thus drive less on a gallon of petrol.

5.2.5 A mixture of variables

We could construct similar plots for all the other numerical variables, but not all plots would be helpful. For example, a plot of mileage versus number of forward gears would be very difficult to read (see Figure 5.13).

```
plot(mtcarsData$gear, mtcarsData$mpg, xlab="Number of gears", ylab="Mileage")
```

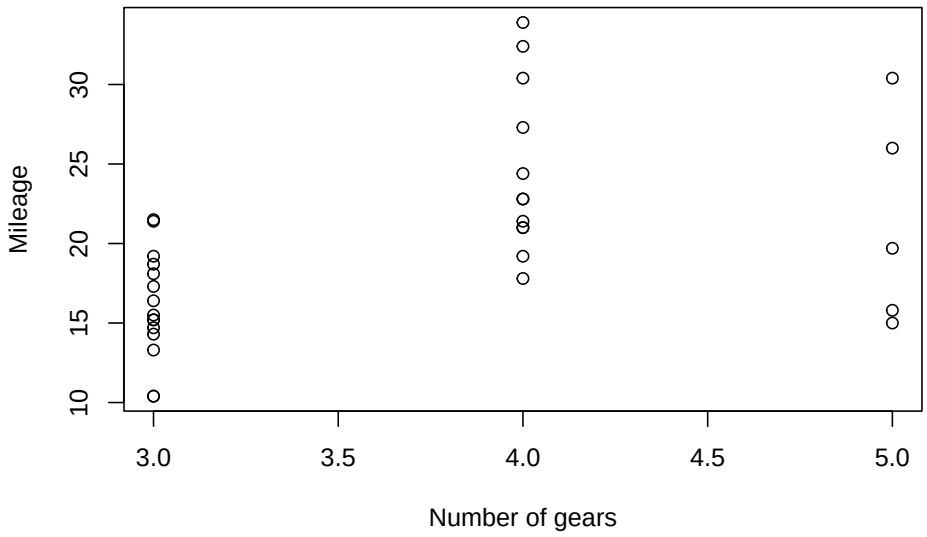


Figure 5.13: Scatterplot diagram between weight and mileage.

This is because one of the variables is integer and takes only a handful of values. In this case, a boxplot or a violin plot would be more useful:

```
boxplot(mpg~gear, mtcarsData, xlab="Number of gears", ylab="Mileage")
points(tapply(mtcarsData$mpg, mtcarsData$gear, mean), col="red3", pch=16)
```

The plot on Figure 5.14 is more informative than the one on Figure 5.13: it shows how the distribution of mileage changes with the increase of the numeric variable number of gears. We can also see that the mean value first increases and then goes down slightly. I do not have any good explanation of this phenomenon, but it might be related with how efficient the cars become with the increase fo the number of gears, or this could happen due to some latent, unobserved factors. So, the data tells us that there is a non-linear relation between number of gears and mileage.

Similarly, we can produce violin plots for the same data using the following code:

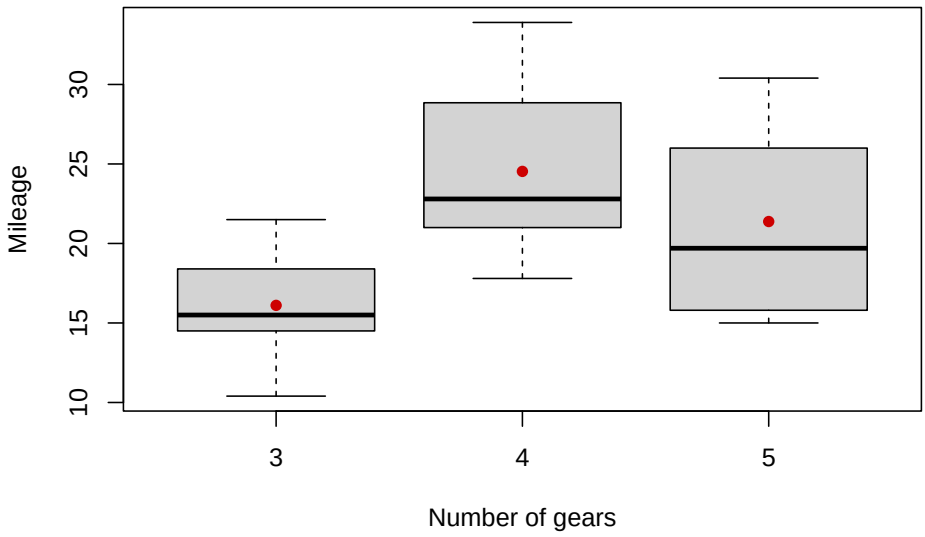


Figure 5.14: Boxplot of mileage vs number of gears.

```
vioplot(mpg~gear, mtcarsData, xlab="Number of gears", ylab="Mileage")
points(tapply(mtcarsData$mpg, mtcarsData$gear, mean), col="red3", pch=16)
```

Finally, using exactly the same idea with boxplots / violin plots, we can analyse relations between categorical and numerical variables. Figure 5.15 shows the relation between transmission type and mileage. We can conclude that the cars with manual transmission tend to have a higher mileage than the ones with the automatic one in our dataset.

```
vioplot(mpg~am, mtcarsData, xlab="Transmission type", ylab="Mileage")
points(tapply(mtcarsData$mpg, mtcarsData$am, mean), col="red3", pch=16)
```

5.2.6 Plot for several variables

Finally, producing plots one by one might be a tedious and challenging task, so it is good to have some instruments for producing several of them together. The `plot()` method will produce scatterplot matrix for numerical variables, but does not deal well with integer and categorical variables:

```
plot(mtcars)
```

Figure 5.16 is informative for the variables `mpg`, `cyl`, `disp`, `hp`, `drat`, `qsec` and `carb`, but is difficult to read for the others. In order to address this issue, we can use the `spread()` function from `greybox`, which will detect types of variables and produce the necessary plots automatically:

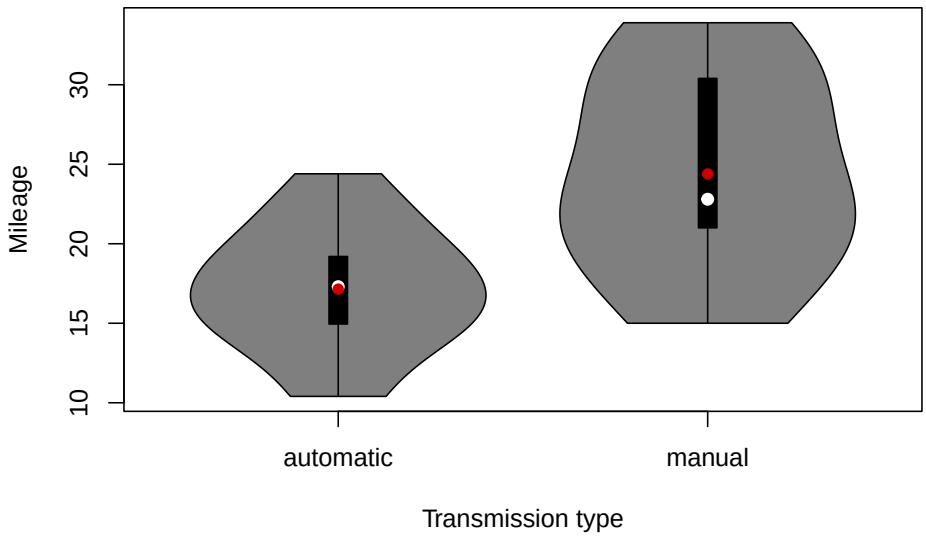


Figure 5.15: Violin plot of mileage vs transmission type.

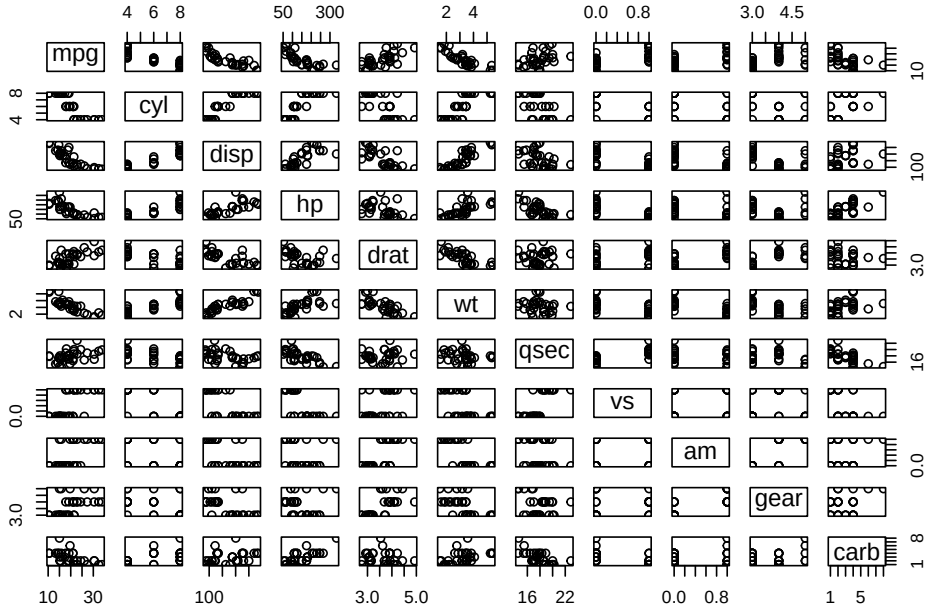


Figure 5.16: Scatterplot matrix for the mtcars dataset.

```
spread(mtcarsData, lowess=TRUE)
```

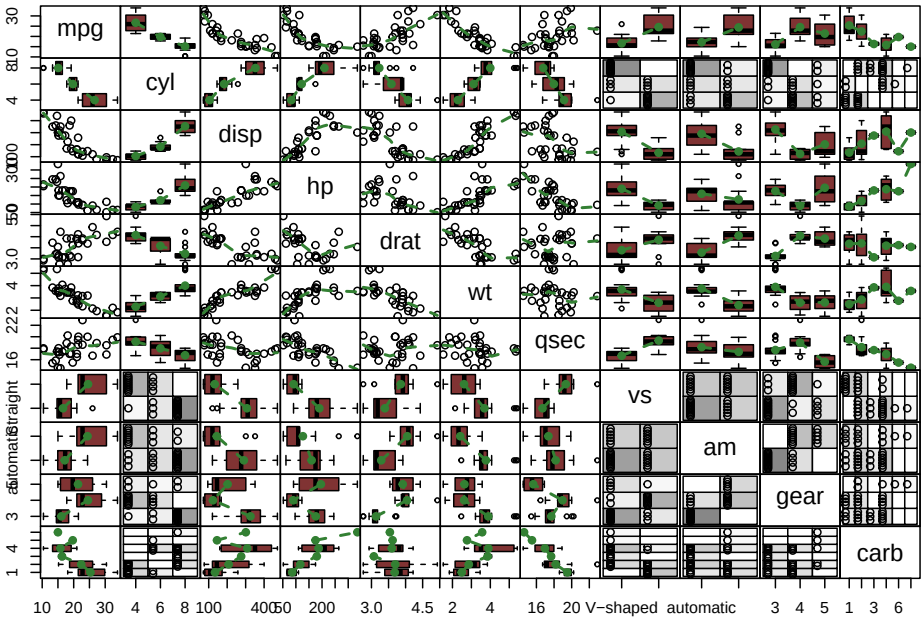


Figure 5.17: Spread plot for the mtcars dataset.

The plot on Figure 5.17 is the collection of the plots discussed above, so I will not stop on explaining what it shows.

As a final word for this section, when analysing data, it is critically important not to just describe what we see, but also explain why a result or a relationship is meaningful, otherwise this becomes an exercise of stating the obvious which does not have any value. So, for example, concluding based on Figure 5.17 that the mileage has a negative relation with displacement is not enough. If you want to analyse the data properly, you need to explain that this relation is meaningful, because with the increase of the size of engine, the fuel consumption will increase as well, and as a result the mileage will go down. Furthermore, the relation is non-linear because the change in decrease will slow down with cars with bigger engines. Inevitably, the car with a gigantic engine will be able to travel a short distance on a gallon of fuel - the mileage will not become negative, so the non-linearity is not an artefact of the data, but an existing phenomenon.

Chapter 6

Population and sampling

Consider a case, when you want to understand what is the average height of teenagers living in your town. It is very expensive and time consuming to go from one house to another and ask every single teenager (if you find one), what their hight is. If we could do that, we would get the true mean, true average height of teenagers living in the town. But in reality, it is more practical to ask a sample of teenagers and make conclusions about the "population" (all teenagers in the town) based on this sample. Indeed, you will spend much less time collecting the information about the height of 100 people rather than 100,000. However, when we take a sample of something, the statistics we work with will always differ from the truth: sample mean will never be equal to the true mean, but it can be shown mathematically that it will converge to the truth, when some specific conditions are met and when the sample size increases. If we set up the experiment correctly, then we can expect our statistics to follow some laws. In this chapter, we discuss these laws, how they work and what they imply.

6.1 Law of Large Numbers

The first law is called the **Law of Large Numbers** (LLN). It is the theorem saying that (under wide conditions) the average of a variable obtained over the large number of trials will be close to its expected value and will get closer to it with the increase of the sample size. This can be demonstrated with the following example:

```
obs <- 10000
# Generate data from normal distribution
y <- rnorm(obs,100,100)
# Create sub-samples of 50 and 100 observations
y30 <- sample(y, 30)
y1000 <- sample(y, 1000)
```

```
par(mfcol=c(1,2))
hist(y30, xlab="y")
abline(v=mean(y30), col="red")
hist(y1000, xlab="y")
abline(v=mean(y1000), col="red")
```

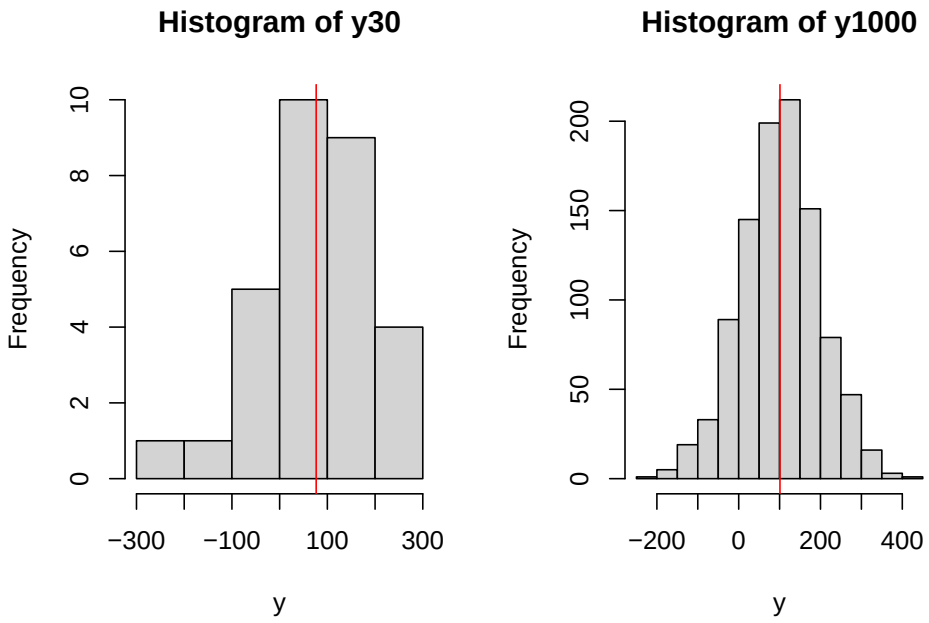


Figure 6.1: Histograms of samples of data from variable y.

What we will typically see on the plots above is that the mean (red line) on the left plot will be further away from the true mean of 100 than in the case of the right plot. Given that this is randomly generated, the situation might differ, but the idea would be that with the increase of the sample size the estimated sample mean will converge to the true one. We can even produce a plot showing how this happens:

```
yMean <- vector("numeric",obs)
for(i in 1:obs){
    yMean[i] <- mean(sample(y,i))
}
plot(yMean, type="l", xlab="Sample size", ylab="Sample mean")
```

We can see from the plot above that with the increase of the sample size the sample mean reaches the true value of 100. This is a graphical demonstration of the Law of Large Numbers: it only tells us about what will happen when the sample size increases. But it is still useful, because it used for many statistical inferences and if it does not work, then the estimate of mean would be incorrect,

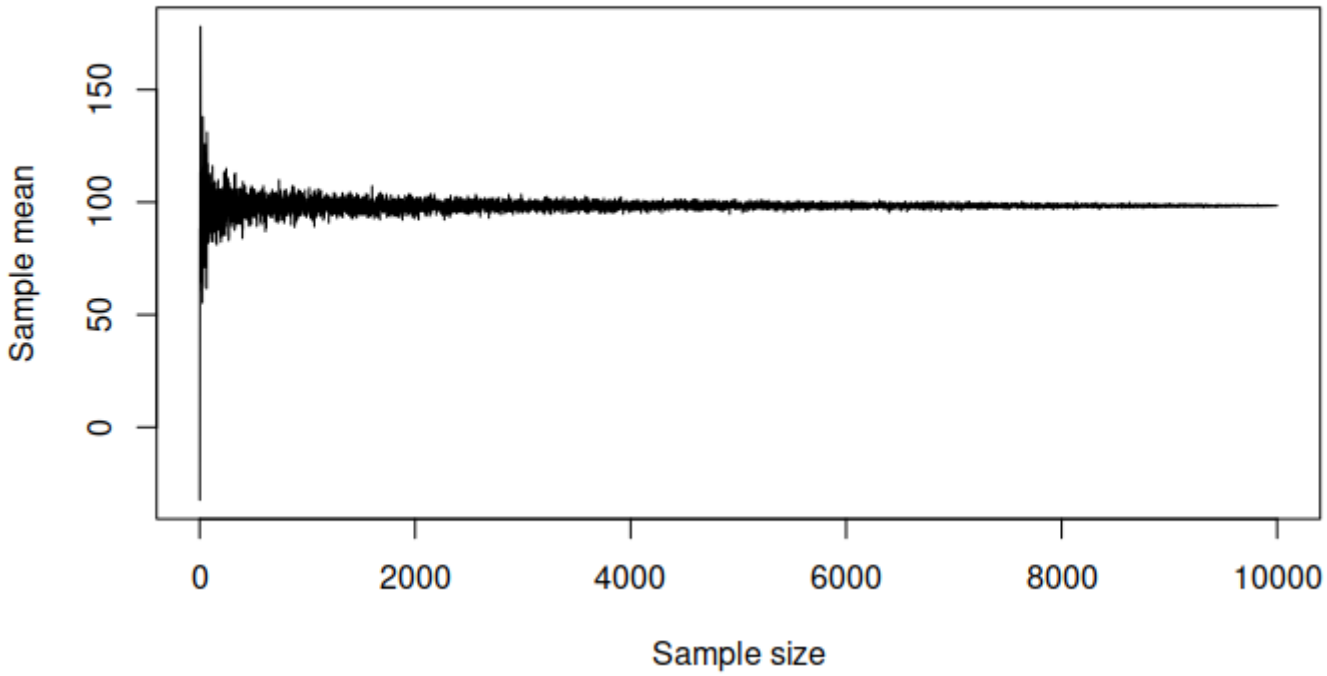


Figure 6.2: Demonstration of Law of Large Numbers.

meaning that we cannot make conclusions about the behaviour in population.

In order for LLN to work, the distribution of variable needs to have finite mean and variance. This is discussed in some detail in the next subsection.

In summary, what LLN tells us is that if we average things out over a large number of observations, then that average starts looking very similar to the population value. However, this does not say anything about the performance of estimators on small samples.

6.2 Central Limit Theorem

As we have already seen on Figure 6.2, the sample mean is not exactly equal to the population mean even when the sample size is very large (thousands of observations). There is always some sort of variability around the population mean. In order to understand how this variability looks like, we could conduct a simple experiment. We could take a random sample of, for instance, 1000 observations several times and record each of the obtained means. We then can see how the variable will be distributed to see if there are any patterns in the behaviour of the estimator:

```
nIterations <- 1000
yMean <- vector("numeric",nIterations)
for(i in 1:nIterations){
    yMean[i] <- mean(sample(y,1000))
}
hist(yMean, xlab="Sample mean", main="")
```

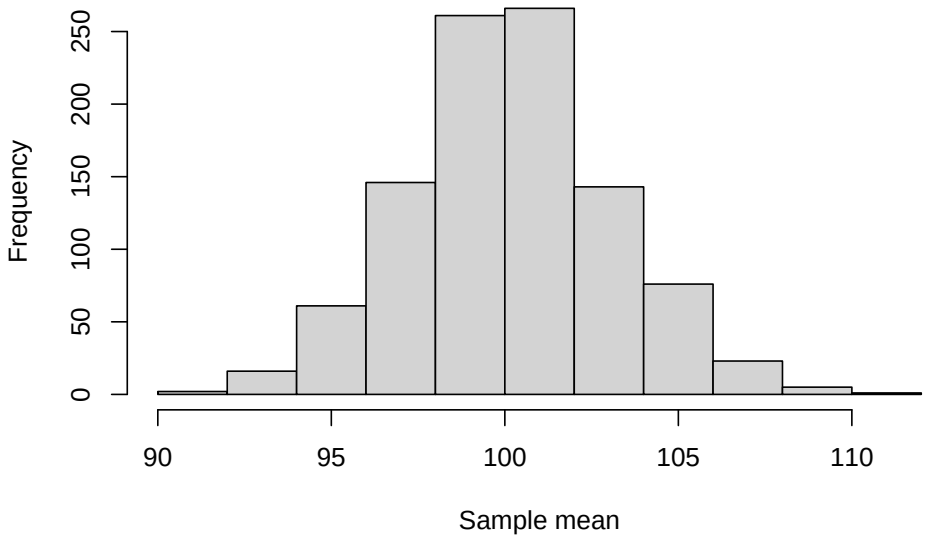


Figure 6.3: Histogram of the mean of the variable y.

There is a theorem that says that the distribution of mean in the experiment above will follow normal distribution under several conditions (discussed later in this section). It is called **Central Limit Theorem** (CLT) and very roughly it says that when independent random variables are added, their normalised sum will asymptotically follow normal distribution, even if the original variables do not follow it. Note that this is the theorem about *what happens with the estimate* (sum in this case), *not with individual observations*. This means that the error term might follow, for example, Inverse Gaussian distribution, but the estimate of its mean (under some conditions) will follow normal distribution. There are different versions of this theorem, built with different assumptions with respect to the random variable and the estimation procedure, but we do not dicuss these details in this textbook.

In order for CLT to hold, the following important assumptions need to be satisfied:

1. **The true value of parameter is not near the bound**. e.g. if the variable follows uniform distribution on (0, a) and we want to estimate a, then its distribution will not be Normal (because in this case the true value is always approached from below). This assumption is important in our context, because ETS and ARIMA typically have restrictions on their parameters.
2. **The random variables are identically independent distributed** (i.i.d.). If they are not, then their average might not follow normal distribution (in some conditions it still might).
3. **The mean and variance of the distribution are finite**. This might seem as a weird issue, but some distributions do not have finite moments, so the CLT will not hold if a variable follows them, just because the sample mean will be all over the plane due to randomness and will not converge to the "true" value. Cauchy distribution is one of such examples.

If these assumptions hold, then CLT will work for the estimate of a parameter, no matter what the distribution of the random variable is. This becomes especially useful, when we want to test a hypothesis or construct a confidence interval for an estimate of a parameter.

6.3 Properties of estimators

Before we move further, we need to agree what the term "estimator" means, which will be used several times further in this textbook:

- **Estimate** of a parameter is an in sample result of application of a statistical procedure to the data for obtaining some coefficients of a model. The value calculated using the arithmetic mean would be an estimate of the population mean;
- **Estimator** is the rule for calculating estimates of parameters based on a sample of data. For example, arithmetic mean is an estimator of the

population mean. Another example would be method of Ordinary Least Squares, which is a rule for producing estimates of parameters of a regression model and thus an estimator.

In this section, we discuss such terms as **bias**, **efficiency** and **consistency** of estimates of parameters, which are directly related to LLN and CLT. Although there are strict statistical definitions of the aforementioned terms (you can easily find them in Wikipedia or anywhere else), I do not want to copy-paste them here, because there are only a couple of important points worth mentioning in our context.

Note that all the discussions in this chapter relate to **the estimates of parameters**, not to the distribution of a random variable itself. A common mistake that students make when studying statistics, is that they think that the properties apply to the variable y_j instead of the estimate of its parameters (e.g. mean of y_j).

6.3.1 Bias

Bias refers to the expected difference between the estimated value of parameter (on a specific sample) and the "true" one (in the true model). Having unbiased estimates of parameters is important because they should lead to more accurate forecasts (at least in theory). For example, if the estimated parameter is equal to zero, while in fact it should be 0.5, then the model would not take the provided information into account correctly and as a result will produce less accurate point forecasts and incorrect prediction intervals. In inventory context this may mean that we constantly order 100 units less than needed only because the parameter is lower than it should be.

The classical example of bias in statistics is the estimation of variance in sample. The following formula gives biased estimate of variance in sample:

$$\mathrm{V}(y) = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2, \tag{6.1}$$

where n is the sample size and $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ is the mean of the data. There is a lot of proofs in the literature of this issue (even Wikipedia (2020a) has one), we will not spend time on that. Instead, we will see this effect in the following simple simulation experiment:

```
mu <- 100
sigma <- 10
nIterations <- 1000
# Generate data from normal distribution, 10,000 observations
y <- rnorm(10000,mu,sigma)
# This is the function, which will calculate the two variances
varFunction <- function(y){
   return(c(var(y), mean((y-mean(y))^2)))
```

```
}
# Calculate biased and unbiased variances for the sample of 30 observations,
# repeat nIterations times
varValues <- replicate(nIterations, varFunction(sample(y,30)))
```

This way we have generated 1000 samples with 30 observations and calculated variances using the formulae (6.1) and the corrected one for each step. Now we can plot it in order to see how it worked out:

```
par(mfcol=c(1,2))
# Histogram of the biased estimate
hist(varValues[2,], xlab="V(y)", ylab="y", main="Biased estimate of V(y)")
abline(v=mean(varValues[2,]), col="red")
legend("topright",legend=TeX(paste0("E$\\left(V(y)\\right)$=",round(mean(varValues[2,]),2))),lwd=

# Histogram of unbiased estimate
hist(varValues[1,], xlab="V(y)", ylab="y", main="Unbiased estimate of V(y)")
abline(v=mean(varValues[1,]), col="red")
legend("topright",legend=TeX(paste0("E$\\left(V(y)\\right)$=",round(mean(varValues[1,]),2))),lwd=
```

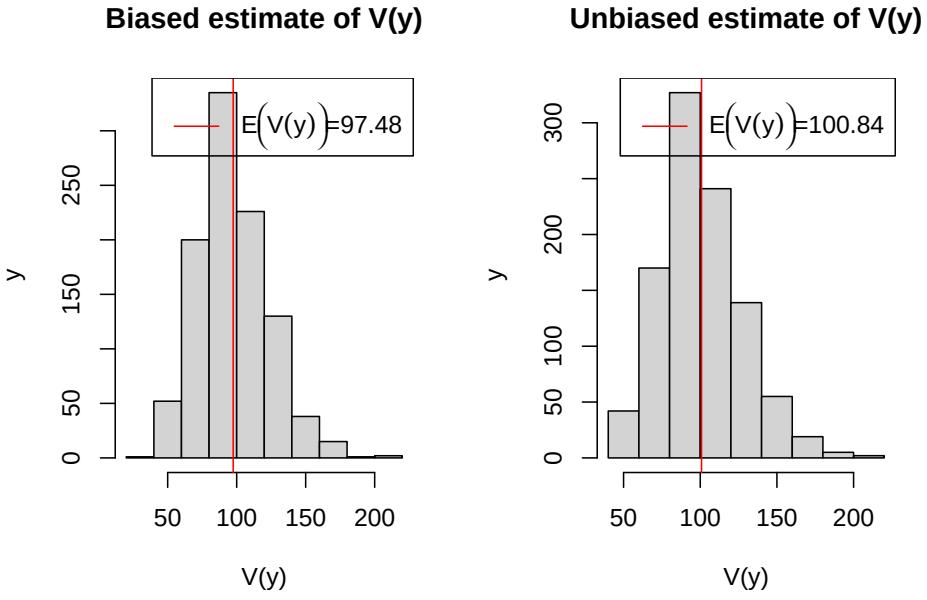


Figure 6.4: Histograms for biased and unbiased estimates of variance.

Every run of this experiment will produce different plots, but typically what we will see is that, the biased estimate of variance (the histogram on the right hand side of the plot) will have lower mean than the unbiased one. This is the graphical example of the effect of not taking the number of estimated parameters

into account. The correct formula for the unbiased estimate of variance is:

$$s^2 = \frac{1}{n-k}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2, \tag{6.2}$$

where k is the number of all independent estimated parameters. In this simple example $k = 1$, because we only estimate mean (the variance is based on it). Analysing the formulae (6.1) and (6.2), we can say that with the increase of the sample size, the bias will disappear and the two formulae will give almost the same results: when the sample size n becomes big enough, the difference between the two becomes negligible. This is the graphical presentation of the bias in the estimator.

6.3.2 Efficiency

Efficiency means, if the sample size increases, then the estimated parameters will not change substantially, they will vary in a narrow range (variance of estimates will be small). In the case with inefficient estimates the increase of sample size from 50 to 51 observations may lead to the change of a parameter from 0.1 to, let's say, 10. This is bad because the values of parameters usually influence both point forecasts and prediction intervals. As a result the inventory decision may differ radically from day to day. For example, we may decide that we urgently need 1000 units of product on Monday, and order it just to realise on Tuesday that we only need 100. Obviously this is an exaggeration, but no one wants to deal with such an erratically behaving model, so we need to have efficient estimates of parameters.

Another classical example of not efficient estimator is the median, when used on the data that follows Normal distribution. Here is a simple experiment demonstrating the idea:

```
mu <- 100
sigma <- 10
nIterations <- 500
obs <- 100
varMeanValues <- vector("numeric",obs)
varMedianValues <- vector("numeric",obs)
y <- rnorm(100000,mu,sigma)
for(i in 1:obs){
    ySample <- replicate(nIterations,sample(y,i*100))
    varMeanValues[i] <- var(apply(ySample,2,mean))
    varMedianValues[i] <- var(apply(ySample,2,median))
}
```

In order to establish the efficiency of the estimators, we will take their variances and look at the ratio of mean over median. If both are equally efficient, then this ratio will be equal to one. If the mean is more efficient than the median, then the ratio will be less than one:

```
options(scipen=6)
plot(1:100*100,varMeanValues/varMedianValues, type="l", xlab="Sample size",ylab="Relative efficie
abline(h=1, col="red")
```

What we should typically see on this graph, is that the black line should be below the red one, indicating that the variance of mean is lower than the variance of the median. This means that mean is more efficient estimator of the true location of the distribution μ than the median. In fact, it is easy to proove that asymptotically the mean will be 1.57 times more efficient than median (Wikipedia, 2020b) (so, the line should converge approximately to the value of 0.64).

6.3.3 Consistency

Consistency means that our estimates of parameters will get closer to the stable values (true value in the population) with the increase of the sample size. This follows directly from LLN and is important because in the opposite case estimates of parameters will diverge and become less and less realistic. This once again influences both point forecasts and prediction intervals, which will be less meaningful than they should have been. In a way consistency means that with the increase of the sample size the parameters will become more efficient and less biased. This in turn means that the more observations we have, the better.

An example of inconsistent estimator is Chebyshev (or max norm) metric. It is formulated the following way:

$$\text{LMax} = \max\left(|y_1 - \hat{y}|, |y_2 - \hat{y}|, \dots, |y_n - \hat{y}|\right). \tag{6.3}$$

Minimising this norm, we can get an estimate $\hat{y}$ of the location parameter μ. The simulation experiment becomes a bit more tricky in this situation, but here is the code to generate the estimates of the location parameter:

```
LMax <- function(y){
    estimator <- function(par){
        return(max(abs(y-par)));
    }

    return(optim(mean(y), fn=estimator, method="Brent", lower=min(y), upper=max(y)));
}

mu <- 100
sigma <- 10
nIterations <- 1000
y <- rnorm(10000, mu, sigma)
LMaxEstimates <- vector("numeric", nIterations)
for(i in 1:nIterations){
```

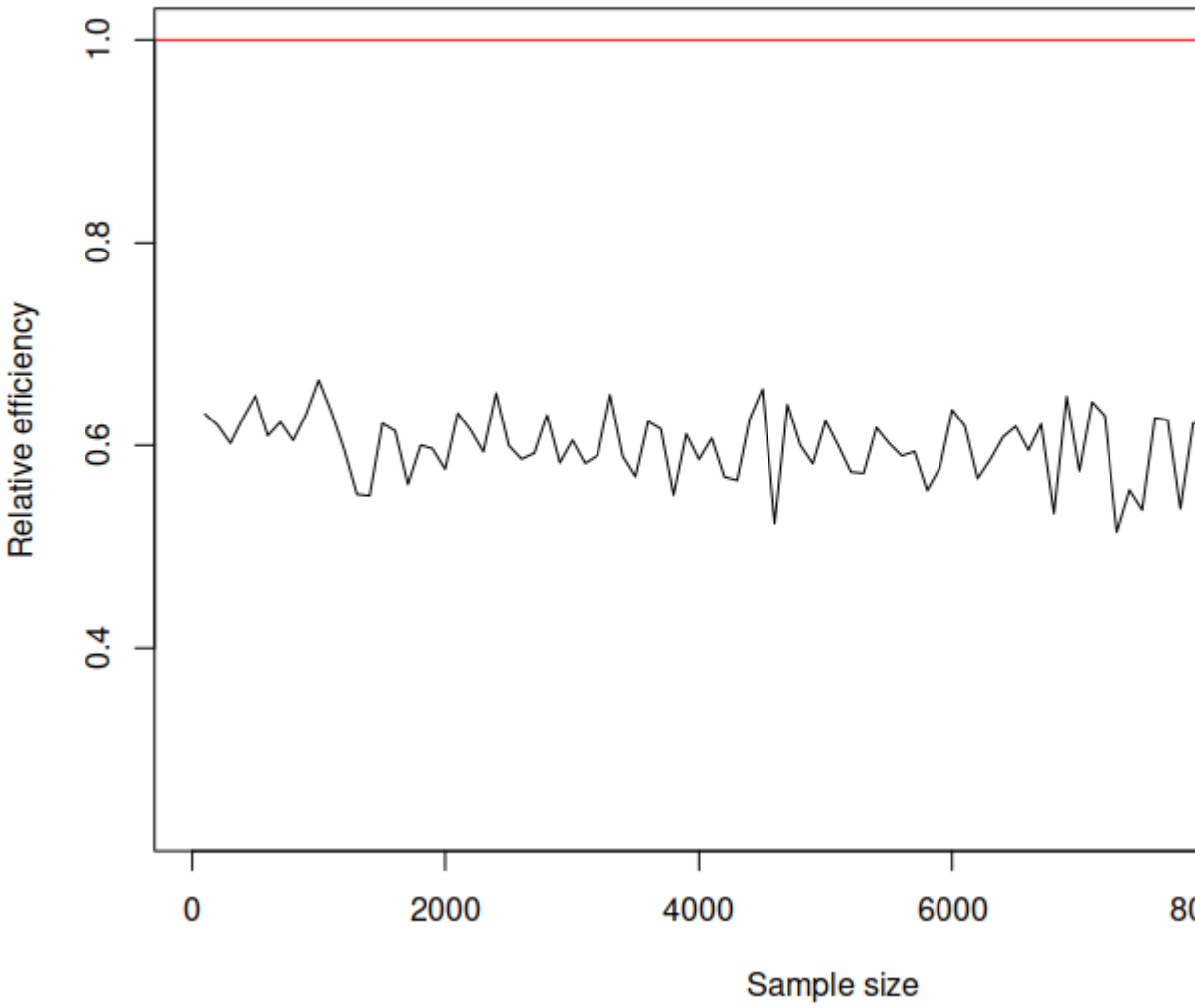


Figure 6.5: An example of a relatively inefficient estimator.

```
    LMaxEstimates[i] <- LMax(y[1:(i*10)])$par;
}
```

And here how the estimate looks with the increase of sample size:

```
plot(1:nIterations*10, LMaxEstimates, type="l", xlab="Sample size",ylab=TeX("Estimate of $\\mu$")
abline(h=mu, col="red")
```

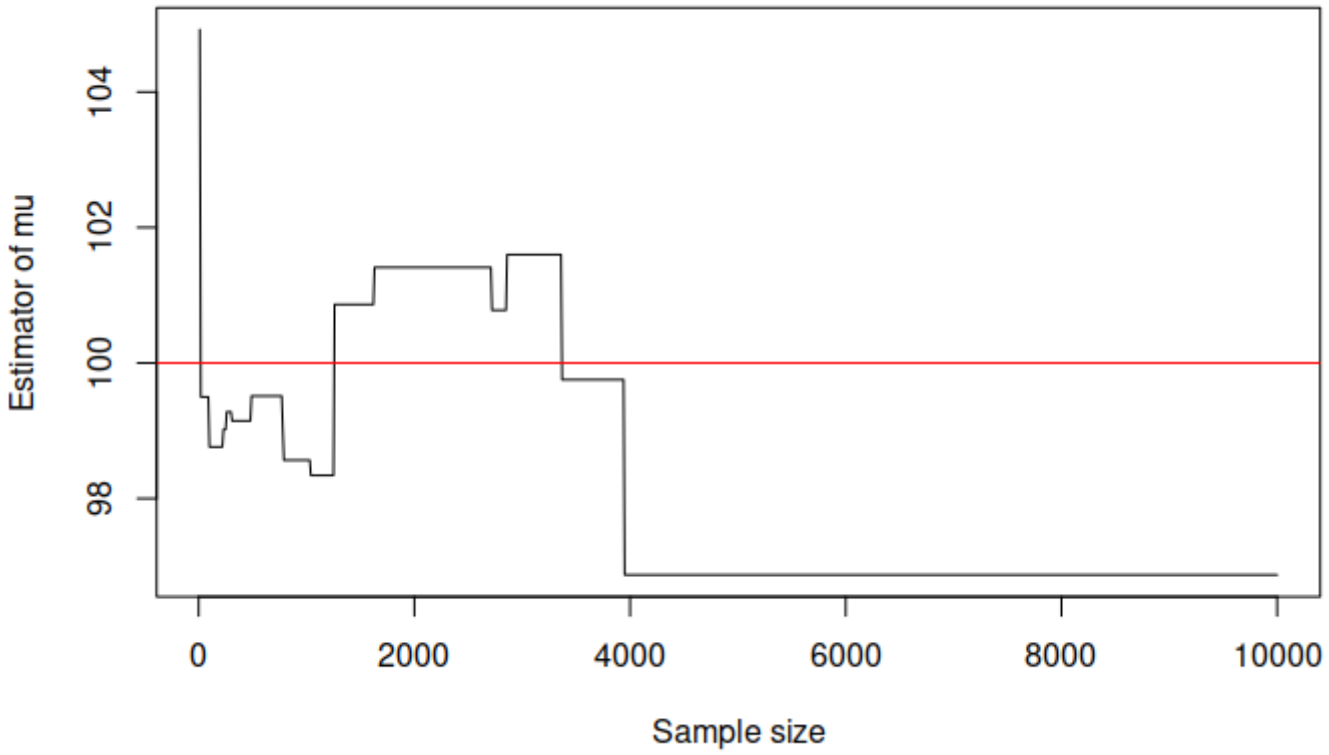


Figure 6.6: An example of inconsistent estimator.

While in the example with bias we could see that the lines converge to the red line (the true value) with the increase of the sample size, the Chebyshev metric example shows that the line does not approach the true one, even when the sample size is 10000 observations. The conclusion is that when Chebyshev metric is used, it produces inconsistent estimates of parameters.

Remark. There is a prejudice in the world of practitioners that the situation in the market changes so fast that the old observations become useless very

fast. As a result many companies just throw away the old data. Although, in general the statement about the market changes is true, the forecasters tend to work with the models that take this into account (e.g. Exponential smoothing, ARIMA, discussed in this book). These models adapt to the potential changes. So, we may benefit from the old data because it allows us getting more consistent estimates of parameters. Just keep in mind, that you can always remove the annoying bits of data but you can never un-throw away the data.

6.3.4 Asymptotic normality

Finally, **asymptotic normality** is not critical, but in many cases is a desired, useful property of estimates. What it tells us is that the distribution of the estimate of parameter will be well behaved with a specific mean (typically equal to μ) and a fixed variance. This follows directly from CLT. Some of the statistical tests and mathematical derivations rely on this assumption. For example, when one conducts a significance test for parameters of model, this assumption is implied in the process. If the distribution is not Normal, then the confidence intervals constructed for the parameters will be wrong together with the respective t- and p- values.

Another important aspect to cover is what the term **asymptotic**, which we have already used, means in our context. Here and after in this book, when this word is used, we refer to an unrealistic hypothetical situation of having all the data in the multiverse, where the time index $t \to \infty$. While this is impossible in practice, the idea is useful, because asymptotic behaviour of estimators and models is helpful on large samples of data. Besides, even if we deal with small samples, it is good to know what to expect to happen if the sample size increases.

6.3.5 Why having biased estimate can be better than having the inefficient one?

It might not be clear to everyone why the model with some bias in it might be better than the model with high variance. In order to answer this question, consider the situation, where we want to estimate the value of parameter μ, and we have two methods to do that. Given that we work on a sample of data, the estimates will have some sorts of distributions, shown in Figure 6.7.

Which of the two estimators would you prefer: the first one or the second one? The conventional statistician might choose Estimator 1, because it produces the unbiased estimates of parameter, meaning that on average we will have the correct value of the true parameter. However, if we rephrase the question slightly, making it more realistic, the answer would probably change: "Which of the two estimators would you prefer **on small sample**?". In this situation, we understand that we have limited data and need to make a decision based on what we have on hands, we might not be able to rely on asymptotic properties, on LLN and CLT (Chapter 6). If we choose Estimator 1, then on our specific sample, we might end up easily with a value for m of -2, 0 or 6, just due to

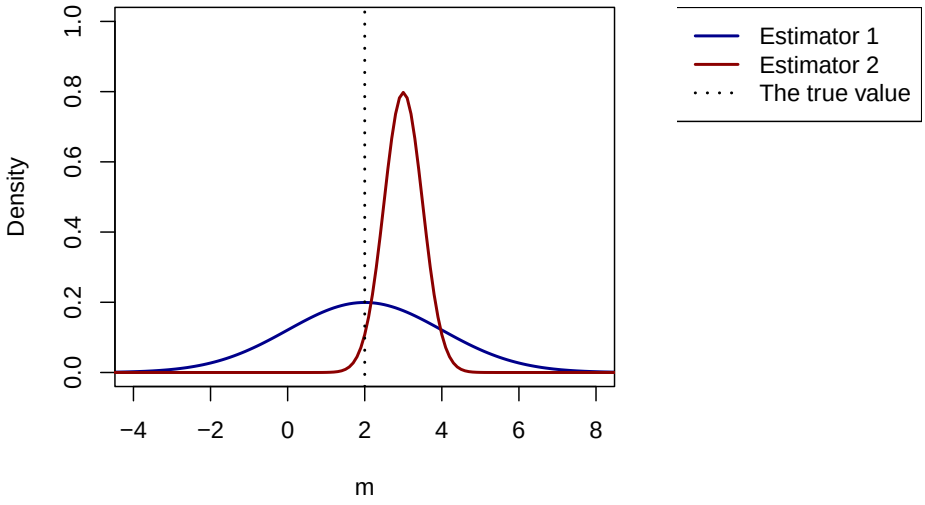


Figure 6.7: Example of two estimators of a parameter.

the pure chance - this is how wide the distribution is. On the other hand, if we choose the Estimator 2, we will end up with the value, which will be close to the true one: 2.5, 3 or 4. Yes, this value will be typically higher than needed, but at least it will not lead us to confusing conclusions on the data we have. Having said that, if the bias was too high (e.g. if the distribution of the Estimator 2 was placed around -4), the estimator might become unreliable, so there should be some balance in how much bias one should impose.

Example 6.1. In a computer game Diablo II (by Blizzard North), there are two spells, which might be considered as similar in terms of damage to monsters: Lightning and Glacial Spike. On the first level, the Lightning does random damage from 1 to 43, while the Glacial Spike does randomly 17 to 26. Assuming that the distributions of damage are uniform in both cases, we would conclude that on average the Lightning does slightly more damage than the Glacial Spike: $\frac{1}{2}(43+1) = 22$ vs $\frac{1}{2}(17+26) = 21.5$. However, the Lightning has much higher variability, and is less efficient in killing monsters than the Glacial Spike: it has variance of $\frac{1}{12}(43-1)^2 = 147$ versus $\frac{1}{12}(26-17)^2 = 6.75$ of the Glacial Spike. This means that each time a player shoots the Lightning, there is a chance that it will do less damage than the Glacial Spike (for example, in $\frac{(17-1)}{(43-1)} \approx 38\%$ of the cases Lightning will do less damage than the lowest possible damage of the Glacial Spike). This means that if one needs to choose, which of the spells to use in a battle, the Glacial Spike would be a safer option, as each specific shot will not be as weak as it could be in the case of the Lightning. But if a player casts both spells many times, then asymptotically the Lightning will be better than Glacial Spike, as it would do more damage on average.

6.4 Confidence interval

As mentioned in Section 1.4, we always work with samples and inevitably we deal with randomness just because of that even, when there are no other sources of uncertainty in the data. For example, if we want to estimate the mean of a variable based on the observed data, the value we get will differ from one sample to another. This should have become apparent from the examples we discussed earlier. And, if the LLN and CLT hold, then we know that the estimate of our parameter will have its own distribution and will converge to the population value with the increase of the sample size. This is the basis for the confidence and prediction interval construction, discussed in this section. Depending on our needs, we can focus on the uncertainty of either the estimate of a parameter, or the random variable y itself. When dealing with the former, we typically work with the **confidence interval** - the interval constructed for the estimate of a parameter, while in the latter case we are interested in the **prediction interval** - the interval constructed for the random variable y.

In order to simplify further discussion in this section, we will take the population mean and its in-sample estimate as an example. In this case we have:

1. A random variable y, which is assumed to follow some distribution with finite mean μ and variance σ^2;
2. A sample of size n from the population of y;
3. Estimates of mean $\hat{\mu} = \bar{y}$ and variance $\hat{\sigma}^2 = s^2$, obtained based on the sample of size n.

What we want to get by doing this is an idea about the population mean μ. The value $\bar{y}$ does not tell us much on its own due to randomness and if we do not capture its uncertainty, we will not know, where the true value μ can be. But using LLN and CLT, we know that the sample mean should converge to the true one and should follow normal distribution. So, the distribution of the sample mean would look like this (Figure 6.8).

On its own, this distribution just tells us that the variable is random around the true mean μ and that its density function has a bell-like shape. In order to make this more useful, we can construct the **confidence interval** for it, which would tell us where the true parameter is most likely to lie. We can cut the tails of this distribution to determine the width of the interval, expecting it to cover $(1-\alpha)\times 100\%$ of cases. In the ideal world, asymptotically, the confidence interval will be constructed based on the true value, like this:

Figure 6.9 shows the classical normal distribution curve around the population mean μ, confidence interval of the level $1-\alpha$ and the cut off tails, the overall surface of which corresponds to α. The value $1-\alpha$ is called **confidence level**, while α is the **significance level**. By constructing the interval this way, we expect that in the $(1-\alpha)\times 100\%$ of cases the value will be inside the bounds, and in $\alpha \times 100\%$ it will not.

In reality we do not know the true mean μ, so we do a slightly different thing: we

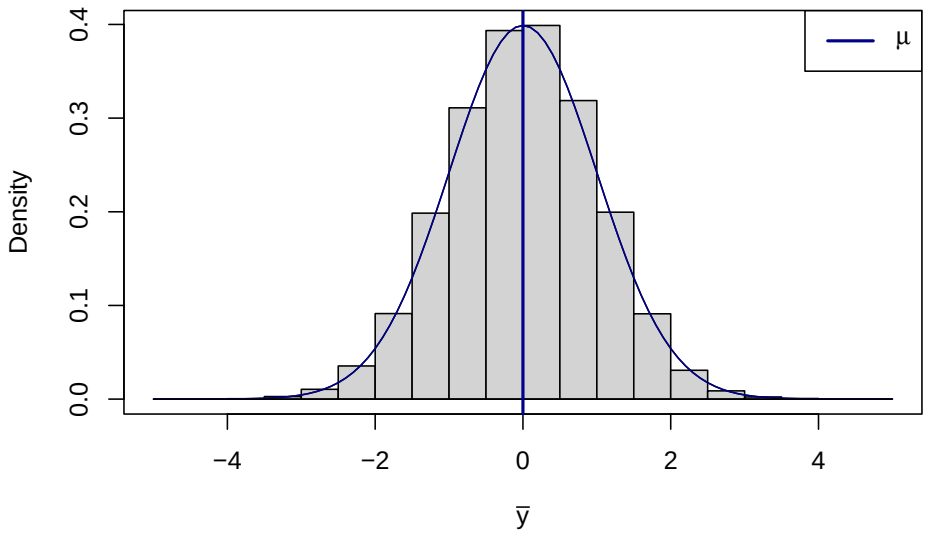


Figure 6.8: Distribution of the sample mean.

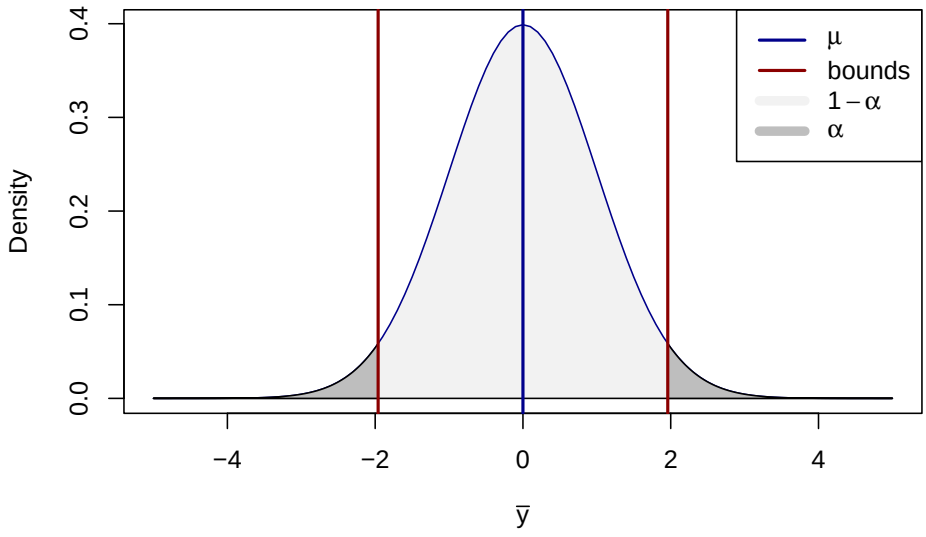


Figure 6.9: Distribution of the sample mean and the confidence interval based on the population data.

construct a confidence interval based on the sample mean $\bar{y}$ and sample variance s^2, hoping that due to LLN they will converge to the true values. We use Normal distribution, because we expect CLT to work. This process looks something like in Figure 6.10, with the bell curve in the background representing the true distribution for the sample mean and the curve on the foreground representing the assumed distribution based on our sample:

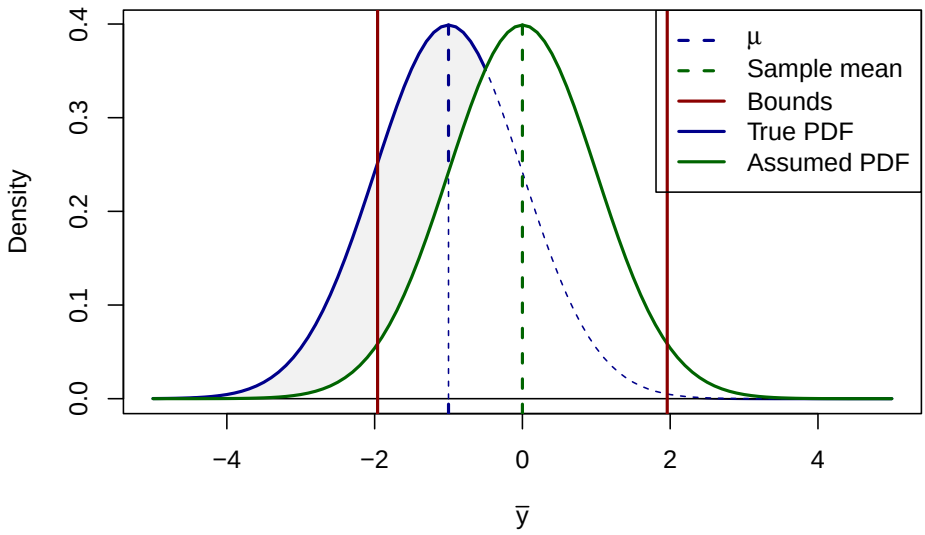


Figure 6.10: Distribution of the sample mean and the confidence interval based on a sample.

So, what the confidence interval does in reality is tries to cover the unknown population mean, based on the sample values of $\bar{y}$ and s^2. If we construct the confidence interval of the width $1-\alpha$ (e.g. 0.95) for thousands of random samples (thousands of trials), then in $(1-\alpha)\times 100\%$ of cases (e.g. 95%) the true mean will be covered by the interval, while in $\alpha \times 100\%$ cases it will not be. The interval itself is random, and we rely on LLN and CLT, when constructing it, expecting for it to work asymptotically, with the increase of the number of trials.

Mathematically the red bounds in Figure 6.10 are represented using the following well-known formula for the confidence interval:

$$\mu \in (\bar{y} + t_{\alpha/2}(df)s_{\bar{y}}, \bar{y} + t_{1-\alpha/2}(df)s_{\bar{y}}), \tag{6.4}$$

where $t_{\alpha/2}(df)$ is Student's t-statistics for $df = n - k$ degrees of freedom (n is the sample size and k is the number of estimated parameters, e.g. $k = 1$ in our case) and level $\alpha/2$, and $s_{\bar{y}} = \frac{1}{\sqrt{n}}s$ is the estimate of the standard deviation of the sample mean (see proof below). If we knew for some reason the true variance σ^2, then we could use z-statistics instead of t, but we typically do not, so we need to take the uncertainty about the variance into account as well, thus the use of t-statistics (see discussion of sample mean tests in Section 8.1).

Proof. We are interested in calculating the variance of $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$. If s is the standard deviation of the random variable y, then $\mathrm{V}(y_j) = s^2$. The following derivations assume that y_j is i.i.d. (specifically, not correlated with each other).

$$\begin{aligned}\mathrm{V}(\bar{y}) =& \mathrm{V}\left(\frac{1}{n}\sum_{j=1}^{n} y_j\right) = \frac{1}{n^2}\mathrm{V}\left(\sum_{j=1}^{n} y_j\right) = \\ & \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{V}\left(y_j\right) = \frac{1}{n^2}\sum_{j=1}^{n} s^2 = \\ & \frac{1}{n}s^2\end{aligned}$$

Based on this, we can conclude that $s_{\bar{y}} = \sqrt{\mathrm{V}(\bar{y})} = \frac{1}{\sqrt{n}}s$. □

Note, that in order to construct *confidence interval*, we do not care what distribution y follows, as long as LLN and CLT hold.

6.5 Prediction interval

If we are interested in capturing the uncertainty about the random variable y, then we should refer to prediction interval. In this case, we typically rely on LLN and the assumed distribution for the random variable y. For example, if we know that $y \sim \mathcal{N}(\mu, \sigma^2)$, then based on our sample we can construct a prediction interval of the width $1 - \alpha$:

$$y \in (\bar{y} + z_{\alpha/2}s, \bar{y} + z_{1-\alpha/2}s), \tag{6.5}$$

where $z_{\alpha/2}$ is the z-statistics (quantile of standard normal distribution) for the level $\alpha/2$ and $\bar{y}$ is the sample estimate of μ and s is the sample estimate of σ. The graphical presentation of such interval can be shown as in Figure 6.11.

Figure 6.11 shows the 95% prediction interval on two plots: the linear plot of values vs observations id and on the histogram. In both cases the prediction intervals are the dashed orange lines, lying further away from the sample mean (the solid blue line). The two solid red lines around the mean represent the 95% confidence intervals for the mean (discussed in Section 6.4). As can be seen, the prediction intervals show, where the 95% of observations are expected to lie. As a result, several observations lie outside the bounds (given the sample of 100 observations, we would expect 5 of them to lie outside, but this will vary from one sample to another). In contrast, confidence interval shows, where the expectation of the population will lie in 95% of the cases, if the interval is constructed many times for random samples.

The formula (6.5) relies on the assumption of normality. If it does not hold, the formula would change. In a way, the prediction interval just comes to getting the quantiles of the assumed distribution based on estimated parameters. In

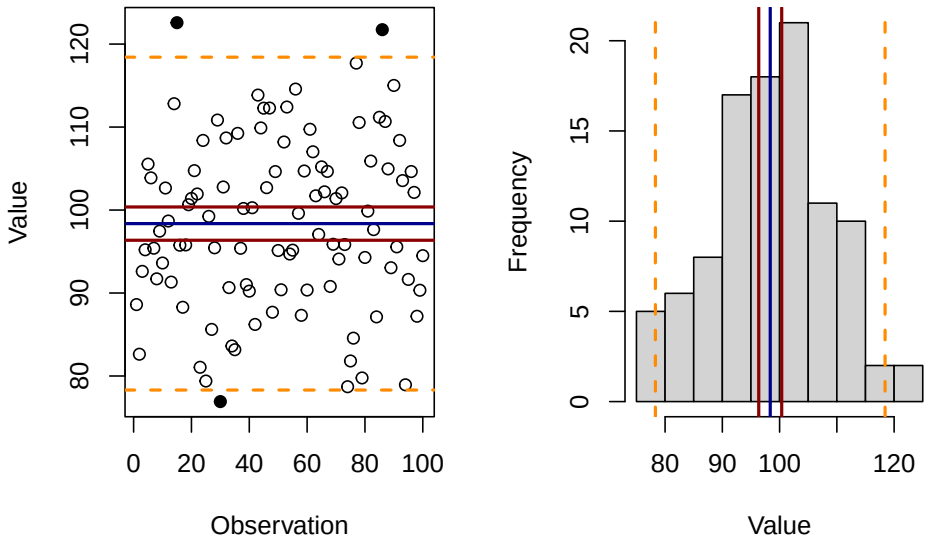


Figure 6.11: Artificial data, mean, confidence and prediction intervals.

some cases, when some of the assumptions do not hold, we might switch to more advanced methods for prediction interval construction.

Chapter 7

Hypothesis testing

Hypothesis testing arises naturally from the idea of confidence intervals discussed in Section 6.4: instead of constructing the interval and getting the idea about the uncertainty of the parameter, we could check, whether the sample agrees with our expectations or not. For example, we could test, whether the population mean is equal to zero based on our sample. We could either construct a confidence interval for the sample mean and see if zero is included in it (in which case it might indicate that zero is one of the possible values of the population mean), or we could reformulate the problem and compare some calculated value with the theoretical threshold. The latter approach is in the nutshell what hypothesis testing does.

In this Chapter we will discuss the main mechanism of hypothesis testing, then move to the discussion of type 0, I and II errors that might arise in the process. We then will discuss the idea of a Power of a Test and investigate what it is influenced by. After that we will discuss several basic popular parametric and non-parametric tests and how to select the most appropriate one between them.

7.1 Basic idea

Fundamentally, the hypothesis testing process relies on the ideas of induction and dichotomy: we have a null (H_0) and alternative (H_1) hypotheses about the process or a property in the population, and we want to find some evidence to reject the H_0. Rejecting a hypothesis is actually more useful than not rejecting it, because in the former case we know what not to expect from the data, while in the latter we just might not have enough evidence to make any solid conclusion. For example, we could formulate H_0 that all cats are white. Failing to reject this hypothesis based on the data that we have (e.g. a dataset of white cats) does not mean that they are all (in the universe) indeed white, it just means that we have not observed the non-white ones. If we collect enough evidence

to reject H_0 (i.e. encountered a black cat), then we can conclude that not all cats are white. This is a more solid conclusion than the one in the previous case. So, if you are interested in a specific outcome, then it makes sense to put this in the alternative hypothesis and see if the data allows to reject the null. For example, if we want to see if the average salary of professors in the UK is higher than £100k per year we would formulate the hypotheses in the following way:

$$H_0 : \mu \leq 100, H_1 : \mu > 100.$$

Having formulated hypotheses, we can check them, but in order to do that, we need to follow a proper procedure, which can be summarised in the six steps:

1. Formulate null and alternative hypotheses (H_0 and H_1) based on your understanding of the problem;
2. Select the significance level α on which the hypothesis will be tested;
3. Select the test appropriate for the formulated hypotheses (1);
4. Conduct the test (3) and get the calculated value;
5. Compare the value in (4) with the threshold one;
6. Make a conclusion based on (5) on the selected level (2).

Note that the order of some elements might change depending on the circumstances, but (2) should always happen before (4), otherwise we might be dealing with so called "p-hacking", trying to make results look nicer than they really are.

Consider an example, where we want to check, whether the population mean μ is equal to zero, based on a sample of 36 observations, where $\bar{y} = -0.5$ and $s^2 = 1$. In this case, we formulate the null and alternative hypotheses:

$$H_0 : \mu = 0, H_1 : \mu \neq 0.$$

We then select the significance level $\alpha = 0.05$ (just as an example) and select the test. Based on the description of the task, this can be either a t-test, or a z-test, depending on whether the variance of the variable is known or not. Usually it is not, so we tend to use t-test. We then conduct the test using the formula:

$$t = \frac{\bar{y} - \mu}{s_{\bar{y}}} = \frac{-0.5 - 0}{\frac{1}{\sqrt{36}}} = -3. \tag{7.1}$$

After that we get the critical value of t with $df = 36 - 1 = 35$ degrees of freedom and significance level $\alpha/2 = 0.025$, which is approximately equal to -2.03. We compare this value with the (7.1) by absolute and reject H_0 if the calculated value is higher than the critical one. In our case it is, so it appears that we have enough evidence to say that the population mean is not equal to 0, on the 5% significance level.

Visually, the whole process of hypothesis testing explained above can be represented in the following way:

If the blue line on Figure 7.1 would lie inside the red bounds (i.e. the calculated value is less than the critical value by absolute), then we would fail to reject

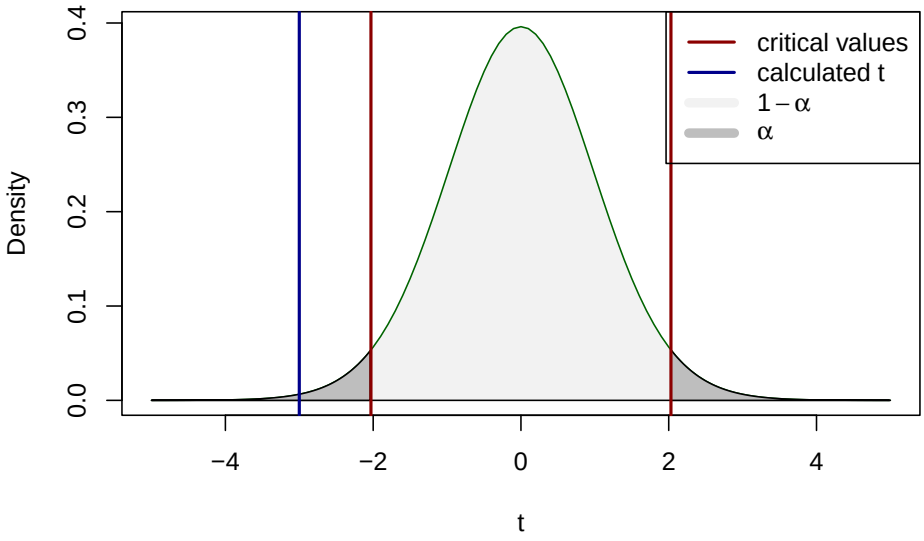


Figure 7.1: The process of hypothesis testing with t value.

H_0. But in our example it is outside the bounds, so we have enough evidence to conclude that the population mean is not equal to zero on 5% significance level. Notice, how similar the mechanisms of confidence interval construction and hypothesis testing are. This is because they are one and the same thing, presented differently. In fact, we could test the same hypothesis by constructing the 95% confidence interval using (6.4) and checking, whether the interval covers the $\mu = 0$:

$$\mu \in \left(-0.50 - 2.03\frac{1}{\sqrt{36}}, -0.50 + 2.03\frac{1}{\sqrt{36}}\right),$$
$$\mu \in (-0.84, -0.16).$$

In our case it does not, so we conclude that we reject H_0 on 5% significance level. This can be roughly represented by the graph on Figure 7.2:

Note that the positioning of the blue line has changed in the case of confidence interval, which happens because of the transition from (7.1) to (6.4). The idea and the message, however, stay the same: if the value is not inside the light grey area, then we reject H_0 on the selected significance level.

Also **note** that we never say that we accept H_0, because this is not what we do in hypothesis testing: if the value would lie inside the interval, then this would only mean that our sample shows that the tested value is covered by the region - the true value can be any of the numbers between the bounds.

Finally, there is a third way to test the hypothesis. We could calculate how much surface is left in the tails with the cut off of the assumed distribution by the blue line on Figure 7.1 (calculated value). In R this can be done using the

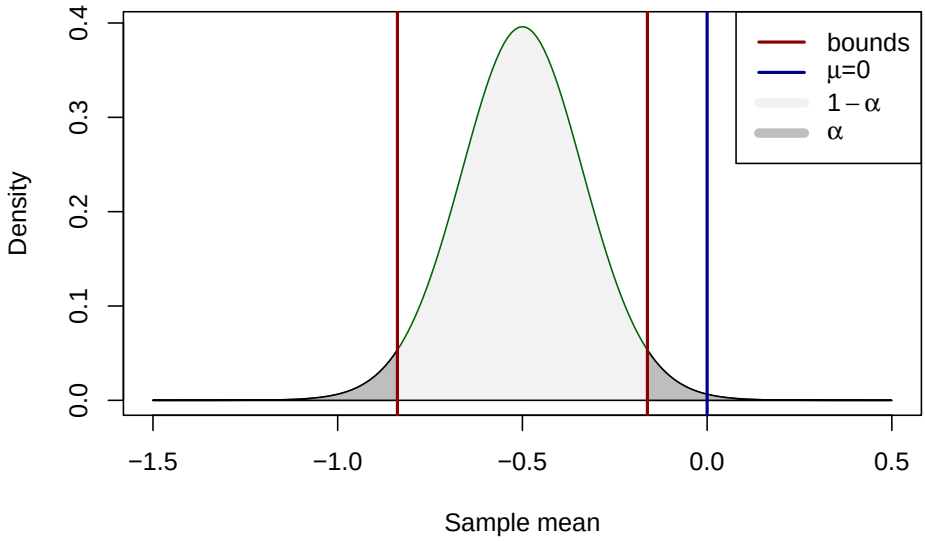


Figure 7.2: Confidence interval for for the population mean example.

`pt()` function:

```
pt(-3, 36-1)
```

```
## [1] 0.002474416
```

Given that we had the inequality in the alternative hypothesis, we need to consider both tails, multiplying the value by 2 to get approximately 0.0049. This is the significance level, for which the switch from "reject" to "do not reject" happens. We could compare this value with the pre-selected significance level directly, rejecting H_0 if it is lower than α. This value is called "p-value" and simplifies the hypothesis testing, because we do not need to look at critical values or construct the confidence interval. There are different definitions of what it is, I personally find the following easier to comprehend: **p-value** is the smallest significance level at which a null hypothesis can be rejected.

Despite this simplification, we still need to follow the procedure and select α *before conducting the test*! We should not change the significance level after observing the p-values, otherwise we might end up bending reality for our needs. Also *note* that if in one case p-value is 0.2, while in the other it is 0.3, it does not mean that the the first case is more significant than the second! P-values are not comparable with each other and they do not tell you about the size of significance. *This is still a binary process*: we either reject, or fail to reject H_0, depending on whether p-value is smaller or greater than the selected significance level.

While p-value is a comfortable instrument, I personally prefer using confidence

intervals, because they show the uncertainty clearer and are less confusing. Consider the following cases to see what I mean:

1. We reject H_0 because t-value is -3, which is smaller than the critical value of -2.03 (or equivalently the absolute of t-value is 3, while the critical is 2.03);
2. We reject H_0 because p-value is 0.0049, which is smaller than the significance level $\alpha = 0.05$;
3. The confidence interval for the mean is $\mu \in (-0.84, -0.16)$. It does not include zero, so we reject H_0.

In case of (3), we not only get the same message as in (1) and (2), but we also see how far the bound is from the tested value. In addition, in the situation, when we fail to reject H_0, the approach (3) gives more appropriate information. Consider the case, when we test, whether $\mu = -0.6$ in the example above. We then have the following three approaches to the problem:

1. We fail to reject H_0 because t-value is 0.245, which is smaller than the critical value of 2.03;
2. We fail to reject H_0 because p-value is 0.808, which is greater than the significance level $\alpha = 0.05$;
3. The confidence interval for the mean is $\mu \in (-0.84, -0.16)$. It includes -0.6, so we fail to reject H_0. *This does not mean that the true mean is indeed equal to -0.6*, but it means that the region will cover this number in 95% of cases if we do resampling many times.

In my opinion, the third approach is more informative and saves from making wrong conclusions about the tested hypothesis, making you work a bit more (you cannot change the confidence level on the fly, you would need to reconstruct the interval). Having said that, either of the three is fine, as long as you understand what they really imply.

Furthermore, if you do hypothesis testing and use p-values, it is worth mentioning the statement of American Statistical Association about p-values (Wasserstein and Lazar, 2016). Among the different aspects discussed in this statement, there is a list of principles related to p-values, which I cite below:

1. P-values can indicate how incompatible the data are with a specified statistical model;
2. P-values do not measure:

- the probability that the studied hypothesis is true,
- or the probability that the data were produced by random chance alone;

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold;
4. Proper inference requires full reporting and transparency;
5. A p-value, or statistical significance, does not measure:

- the size of an effect

- or the importance of a result;

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The statement provides more details about that, but summarising, whatever hypothesis you test and however you test it, you should have apriori understanding of the problem. Diving in the data and trying to see what floats (i.e. which of the p-values is higher than α) is not a good idea (Wasserstein and Lazar, 2016). Follow the proper procedure if you want to test the hypothesis.

Furthermore, the hypothesis testing mechanism has been criticised by many scientists over the years. For example, Cohen (1994) discussed issues with the procedure, making several important points, some of which are outlined above. He also points out at the fundamental problem with hypothesis testing, which is typically neglected by proponents of the procedure: in practice, null hypothesis is always wrong. In reality, it is not possible for a value to be equal, for example, to zero. Even an unimportant effect of one variable on another would be close to zero, but not equal to it. This means that with the increase of the sample size, H_0 will inevitably be rejected.

Remark. When formulating the null hypothesis as equality, it is in fact almost always wrong, because it is close to impossible for a parameter to be equal to a specific value. In the light of this, we should understand that the null hypothesis really means that the true value of parameter is in a proximity of the tested value. So, the hypotheses in this case should be understood as:

$$H_0 : \mu \approx a, H_1 : \mu \not\approx a,$$

where a is the value we are comparing our parameter with. Note that in case of one-tailed tests, this is no longer an issue, because the null hypothesis in that case compares the value with a set of values (e.g. $H_0 : \mu \leq a$).

Finally, our mind operates with binary constructs: true / not true - while the hypothesis testing works in the dichotomy "**I know / I don't know**", with the latter appearing when there is not enough evidence to reject H_0. As a result of this, people tend to make wrong conclusions, because it is difficult to distinguish "not true" from "I don't know", especially for those who do not know statistics well.

Summarising the discussion above, in my opinion, it makes sense to move away from hypothesis testing if possible and switch to other instruments for uncertainty measurement, such as confidence intervals.

7.1.1 Common mistakes related to hypothesis testing

Over the years of teaching statistics, I have seen many different mistakes, related to hypothesis testing. No wonder, this is a difficult topic to grasp. Here, I have decided to summarise several typical erroneous statements, providing

explanations why they are wrong. They partially duplicate the 6 principles from ASA discussed above, but they are formulated slightly differently.

1. "Calculated value is lower than the critical one, so the null hypothesis is true".

- This is wrong on so many level, that I do not even know where to start. We can never know if the hypothesis is true or wrong. All the evidence might point towards the H_0 being correct, but it still can be wrong and at some point in future one observation might reject it. The classical example is the "Black swan in Australia". Up until the discover of Australia, the Europeans thought that there only exist white swans. This was supported by all the observations they had. Wise people would say that "We fail to reject H_0 that all swans are white". Uneducated people would be tempted to say that " H_0: All swans are white" is true. After discovering Australia in 1606, Europeans have collected evidence of existence of black swans, thus rejecting H_0 and showing that "not all swans are white", which implies that actually the alternative hypothesis is true. This is the essence of scientific method: we always try rejecting H_0, collecting some evidence. If we fail to reject it, it might just mean that we have not collected enough evidence or have not modelled it correctly.

2. "Calculated value is lower than the critical one, so we accept the null hypothesis".

- We **never** accept null hypothesis. Even if your house is on fire or there is a tsunami coming, you should not "accept H_0". This is a fundamental statistical principle. We collect evidence to reject the null hypothesis. If we do not have enough evidence, then we just fail to reject it, but we can never accept it, because failing to reject just means that we need to collect more data. As mentioned earlier, we focus on rejecting hypothesis, because this at least tells us, what the phenomenon is not (e.g. that not all swans are white).

3. "The parameter in the model is significant, so we can conclude that..."

- You cannot conclude if something is significant or not without specifying the significance level. Things are only significant if they pass specific test on a specified level α. The correct sentence would be "The parameter in the model is significant on 3%, so we can conclude that...", where 3% is the selected significance level α.

4. "The parameter in the model is **highly** significant, so we can conclude that..."

- This one is similar to (3), with the only difference being the word "highly", which is supposed to show that we obtained a very low p-value and thus the hypothesis can be rejected on a very low significance level. However, this is wrong because the outcome of the hypothesis testing is always binary, so the conclusion should be either "reject" (significant) or "fail to reject" (not

significant) on the selected level α. The significance level should always be selected prior the hypothesis testing.

5. "The parameter in the model is significant because p-value<0.0000"

- Indeed, some statistical software will tell you that `p-value<0.0000`, but this just says that the value is very small and cannot be printed. Even if it is that small, you need to state your significance level and compare it with the p-value. You might wonder, "why bother if it is that low?". Well, if you change the sample size or change model specification, your p-value will change as well, and in some cases it might all of a sudden become higher than your significance level. So, you always need to keep it in mind and make conclusions based on the significance level, not just based on what software tells you.

6. "The parameter is not significant, so we remove the variable from the model".

- This is one of the worst motivations for removing variables that there is (statistical blasphemy!). There are thousands of reasons, why you might get p-value greater than your significance level (assumptions do not hold, sample is too small, the test is too weak, the true value is small etc) and only one of them is that the explanatory variable does not impact the response variable and thus you fail to reject H_0. Are you sure that you face exactly this one special case? If yes, then you already have some other (better) reasons to remove the variable. This means that you should not make decisions just based on the results of a statistical test. *You always need to have some fundamental reason to include or remove variables in the model*. Hypothesis testing just gives you additional information that can be helpful for decision making.

7. "The parameter in the new model is more significant than in the old one".

- There is no such thing as "more significant" or "less significant". **Significance is binary** and depends on the selected level. The only thing you can conclude is whether the parameter is significant on the chosen level α or not.

8. "The p-value of one variables is higher than the p-value of another, so...".

- p-values are not comparable between variables. They only show on what level the hypothesis is rejected and only work together with the chosen significance level. (6) is similar to this mistake.

Remember that the p-value itself is random and will change if you change the sample size or the model specification. Always keep this in mind, when conducting statistical tests. All these mistakes typically arise because of the misuse of p-values and hypothesis testing mechanism. This is one of the reasons, why I prefer confidence intervals (see Section 6.4), when possible (as discussed above).

Table 7.1: Four outcomes in hypothesis testing.

		Reality	
		H_0 is true	$\mathrm{H
The data tells us	Fail to reject H_0	Correct decision,	Type I
		Probability is $1-\alpha$	Probability
	Reject H_0	Type I error,	Correct
		Probability is α	Probability

7.2 Errors of types 0, I and II

When conducting a conventional statistical test, we can have one of the four situations, depending on what happens in real life and what results we obtain. They are summarised in Table 7.1.

The Table 7.1 shows two hypothetical outcomes in reality (we never know, which one we have) and two outcomes of hypothesis testing. This gives us the 2×2 matrix, where α is the significance level and $1 - \beta$ is so called "Power of the Test" (discussed in detail in Subsection 7.3).

Type I error (aka "false positive", i.e. we find a positive effect, when we should not have found it) happens when the null hypothesis is actually true, but we reject it. The probability of this event is equal to α. This is one of the definitions of the significance level α (in how many cases we are ready to make mistakes, when the null hypothesis is true).

Type II error (aka "false negative", i.e. we do not find effect, while we should have found it) happens when we fail to reject the wrong hypothesis (H_0 is not true). The probability of this event equals to β, which can be calculated based on the assumed distribution, the critical and calculated values for the hypothesis.

In order to remember what Type I and Type II errors stand for, there is a good mnemonic with a story of a boy who cried "wolf".

Example 7.1. Just as a reminder, in a village, there lived a boy who one day decided to make a practical joke of his fellow villagers. He ran around the main street crying "Wolf!". We should acknowledge that there was no wolf at that stage, so in our terms we would say that the H_0: $\mu = 0$ was true. But the villagers who have heard the boy went on the streets to help. They rejected the correct null hypothesis in order to help the boy, and they were surprised to find that there were no wolves on the streets. Thus the villagers made a **Type I error**.

Next week, the boy encountered a wolf on the main street and started crying "Wolf!", calling for help. Alas, this time nobody believed the boy and nobody came out to help, and thus the villagers rejected the correct null hypothesis in

favour of the wrong one, H_1: $\mu \neq 0$. By doing so they have made the **Type II error**. If the villagers knew statistics, they would understand that failing to reject H_0 once does not mean that it is true.

While we can regulate the probability of Type I error by changing α, the probability of Type II error cannot be controlled directly. Ideally, we want it to be as low as possible. In general, the more information about the "true" parameters and model you can provide to the test, the lower the Type II error will be. For example, if we want to conduct a test to compare mean with a value and the CLT holds (see Section 6.2), then you might want to choose between t-test and z-test. The latter assumes that the population standard deviation is known (and you can provide it), and as a result has a lower probability of Type II error than the former. We will discuss specific tests in the Section 8.

All the four situations in Table 7.1 rely on the idea that the reality is somehow known. But in real life, we never know whether the null hypothesis is true or not. However, the Table is still useful because it gives an understanding of what to expect from a statistical test and what test to select in each specific situation.

Finally, sometimes analysts refer to the **Type 0** error (it is sometimes called "type III" error, but it is more fundamental than Type I or Type II, so I prefer "Type 0"). This is the error that arises when an answer is obtained to the wrong question. This does not have any mathematics behind it but is important in general: we need to understand what questions to ask and how to formulate them correctly before doing the test.

7.3 Power of a test

Consider a situation when we want to know average income of people living in a country across two regions, and we want to find out whether those averages are similar or not. We could ask "what's your yearly income" for ten people in each of the regions, calculate means and compare them using some statistical test (these will be discussed in Chapter 8). If the average incomes are in reality indeed different, we would expect the test to tell us that. However, having such a small sample, we might not be able to detect the difference correctly. For example, although the averages of 25000 and 27000 could look very different, the specific values could have happened at random, and if we asked another ten people in the regions, the samples could well be 28000 and 24000 respectively. This happens because of the small sample inefficiency of the averages (discussed in Subsection 6.3.2). In this situation of testing a hypothesis on a small sample, we would say that the *power of the test is low*.

If we were to collect larger samples, say 100 in each region, the estimates of the average income will be more reliable (efficient, less biased), and thus the same statistical test will find it easier to detect the differences between them. In that situation we would say that its power has increased in comparison with the small sample case.

This idea of the power of the test is important to understand what test to use in different circumstances. Its definition is as follows: **power of the test is the probability of correctly rejecting the wrong hypothesis**. By definition, it is equal to $1 - \beta$ in Table 7.1 and thus lies between 0 and 1, where the higher number reflects the higher power. While it is possible to calculate the power of a test, when the true value is known, in practice this does not make sense, because in reality we never know it. But the idea itself is useful because it allows comparing the theoretical properties of different statistical tests.

Given that the power of the test is a probability, it can be calculated for a specific test with specific parameters, assuming that a specific hypothesis is true. In this section, we will explain how to do that.

7.3.1 Visual explanation

We start the explanation of the idea of the Power of a Test with a visual example. Consider an example of a z-test, which can be used for comparison of means if the population standard deviation is known (see Section 8 for the explanation of the test). We will use an example of a one sided test for the following hypothesis (where number 3 is selected arbitrarily, just for the example):

$$\mathrm{H}_0 : \mu \geq 3, \mathrm{H}_1 : \mu < 3. \tag{7.2}$$

In this example we will test the hypothesis on 5% significance level. The general process of hypothesis testing can be visualised as shown in Figure 7.3.

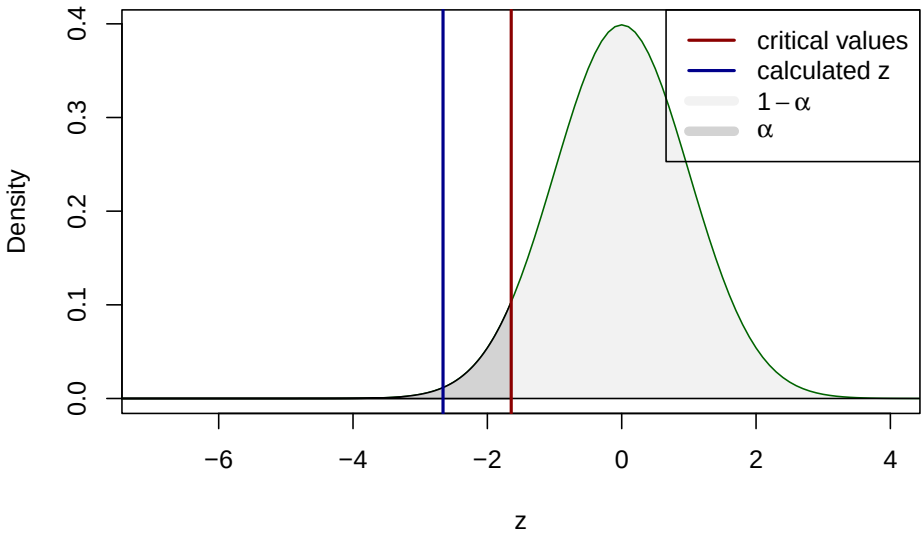


Figure 7.3: The process of hypothesis testing of one sided hypothesis (7.2).

The plot in Figure 7.3 shows the theoretical distribution of z-values, the 5% critical value and the calculated one. On the plot, we see that the calculated

value lies in the tail of the distribution, which means that we can reject the null hypothesis on 5% significance level. The situation, when we correctly reject H_0 in our example corresponds to the case, when the true distribution lies to the left of the assumed one as shown in Figure 7.4.

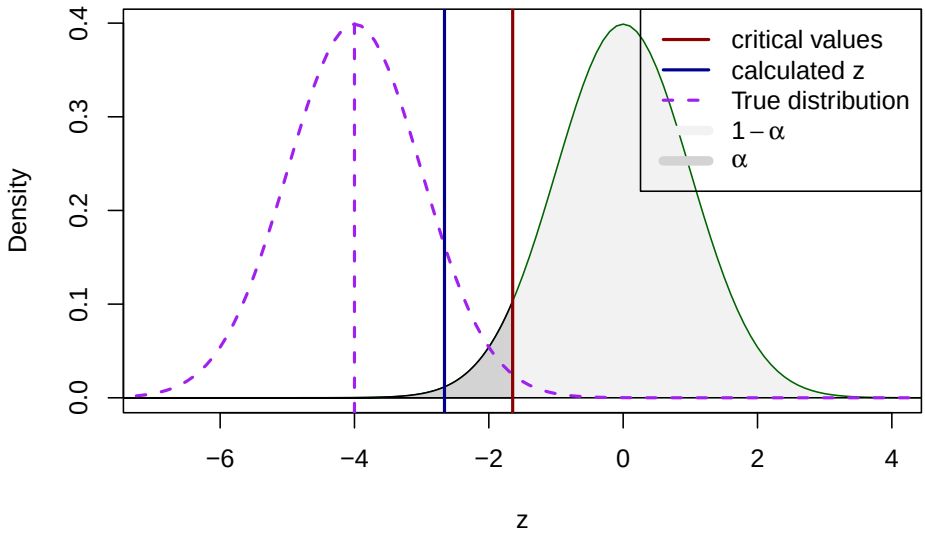


Figure 7.4: Hypothetical "true" and the assumed distributions.

In the example of Figure 7.4 we consider a hypothetical situation, where the true mean is such that the standard normal distribution is centered around the value -4. In this example we correctly reject the wrong null hypothesis, which corresponds to the correct decision in Table 7.1 and the probability of this case is the "Power of a Test". Visually, it corresponds to the surface of the "true" distribution to the left of the critical value that we have chosen in the beginning - any calculated value below this will lead to the rejection of H_0 and thus to the correct decision. This is shown visually in Figure 7.5.

The surface to the left of the critical value in Figure 7.5 is the Power of the Test for the example of a specific value of assumed μ. Given that we never know the true value, we could try other values, which would shift the distribution to the left or to the right, meaning either the increase or the decrease in power of the test. For example, the situation shown in Figure 7.6 corresponds to the smaller Power of the Test (because the surface of the distribution to the left is smaller than the surface of the distribution in Figure 7.5).

Analysing Figures 7.5 and 7.6, we can already outline two factors impacting the power of a test:

1. The location of the true mean. The further it is away to the tested one, the easier it is to detect the difference and reject the H_0, which implies a higher power of a test;

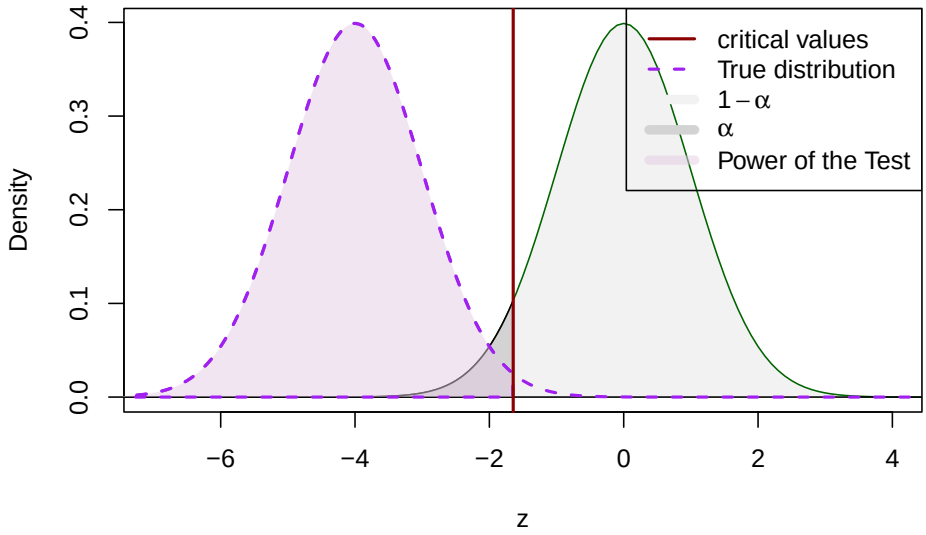


Figure 7.5: Power of the Test based on the hypothetical true value of μ.

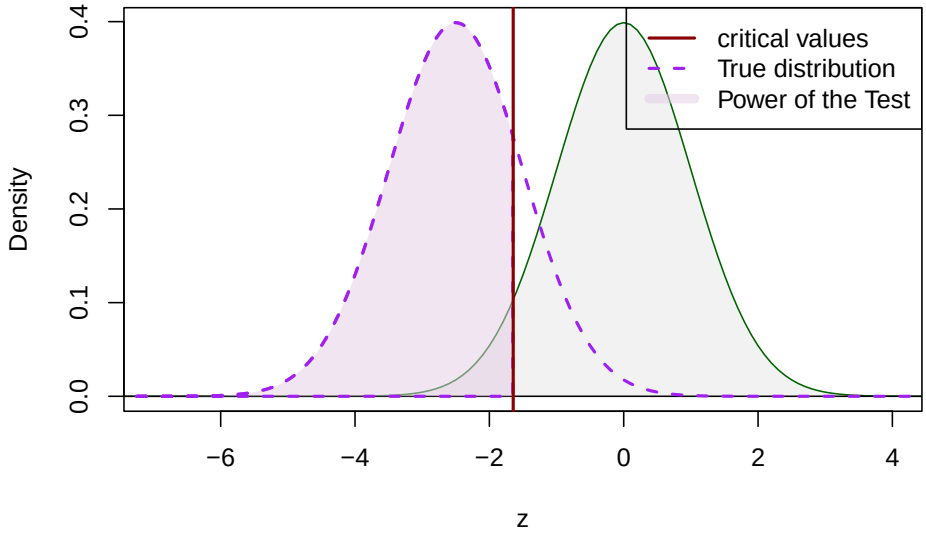


Figure 7.6: Power of the Test based on another hypothetical true value of μ.

2. Significance level α. With the lower significance level, the critical value (vertical line in Figures 7.5 and 7.6) will be further to the left and thus the power of the test will be lower.

There are other factors, which are not as obvious as the two above. For example, if we did not know the true value of σ, we would need to estimate it, and as a result we would need to use a less powerful t-test instead of the z-test. On smaller samples the critical value of the t-test is typically higher than the one from the z-test by absolute value. For example, in case of 36 observations, on 5% significance level it is equal to -1.6896, which is lower than the similar value of -1.6449 from the standard normal distribution. This means that if we used the t-test instead of the z-test in the example above, the vertical line on the plots in Figures 7.5 and 7.6 would be firther to the left of the one that we had. As a result, we would conclude that the power of the t-test is lower than the power of the z-test.

Furthermore, with the increase of the sample size the distribution of means tends to become narrower due to the Law of Large Numbers (see Section 6.1) and thus the power of tests grow, because the critical value moves closer to the centre of distribution.

7.3.2 Mathematical explantion

Moving from the visual explanation to the mathematical one, we can present the Power of a Test as the following probability:

$$\mathrm{P}(\text{reject } \mathrm{H}_0 | \mathrm{H}_0 \text{ is wrong}) = 1 - \mathrm{P}(\text{Type II error}) = 1 - \beta. \tag{7.3}$$

As shown in the previous Subsection, the calculation of the Power of the Test is done based on the parameters of the specific test. In this Subsection, we continue with the same example as before and the same hypothesis:

$$\mathrm{H}_0 : \mu \geq 3, \mathrm{H}_1 : \mu < 3.$$

For the example, we assume that the population standard deviation is known and is equal to $\sigma = 0.18$, that we work with a sample of 36 observations and that we use a 5% significance level to test the hypothesis. In this case, the test statistics is:

$$z = \frac{\bar{y} - 3}{\sigma/\sqrt{n}} = \frac{\bar{y} - 3}{0.03}, \tag{7.4}$$

where $\bar{y}$ is the sample mean. The rule for rejecting the null hypothesis in this situation is that if the calculated value z is lower than or equal to the critical one, which in our case (for the chosen 5% significance level) is -1.645:

$$z = \frac{\bar{y} - 3}{\sigma/\sqrt{n}} = \frac{\bar{y} - 3}{0.03} \leq -1.645$$

Solving this inequality for $\bar{y}$, we will reject H_0 if the sample mean is

$$\bar{y} \leq 3 - 1.645 \times 0.03 = 2.95. \tag{7.5}$$

This can be interpreted as "we will fail to reject the null hypothesis in the cases, when the sample mean is greater than 2.95". Now that we have this value, we can calculate the theoretical power of the test for a variety of cases. For example, we can see how powerful the test is in rejecting the wrong null hypothesis if the true mean is in fact equal to 2.87:

$$\begin{aligned} 1-\beta =& \mathrm{P}(\bar{y} \leq 2.95 | \mu = 2.87) = \\ & \mathrm{P}\left(z \leq \frac{2.95-2.87}{0.03}\right) = \\ & \mathrm{P}\left(z \leq 2.67\right) = \\ & 0.9962 \end{aligned} \tag{7.6}$$

We could do similar calculations for other cases of the true mean and see how powerful the test is in those situations. In fact, we could create a **power curve**, showing how the power of the test changes in a variety of cases of hypothetical true mean. In R, this can be construct in the following way:

```
# Set all the parameters
yMean <- 3
yMeanSD <- 0.18 / sqrt(36)
levelSignigicance <- 0.05
zValue <- 3 + qnorm(levelSignigicance) * yMeanSD
# Vector of hypothetical population means
muValues <- seq(2.8,3.1,length.out=100)
# Vector of values for power curve
powerValues <- vector("numeric",100)

# Calculate the power values
powerValues <- pnorm((zValue - muValues)/yMeanSD)

# Plot the power curve
plot(muValues, powerValues, type="l",
     xlab=latex2exp::TeX("$\\mu$"),
     ylab="Power of the test")
# Add lines for the case of 1-beta=0.05
lines(rep(3,2),c(0,0.05), lty=3)
lines(c(0,3),rep(0.05,2), lty=3)
# And provide a description
text(3, 0.05+0.05,
     latex2exp::TeX("$\\alpha =1- \\beta =0.05$"),
     pos=4)
```

The plot in Figure 7.7 shows how powerful the z-test is for each specific value of population mean. We can see that the test becomes more powerful the further the true mean is away from the tested value (we chose 3 as the tested value). This means that it is easier for the test to detect the distance of the sample

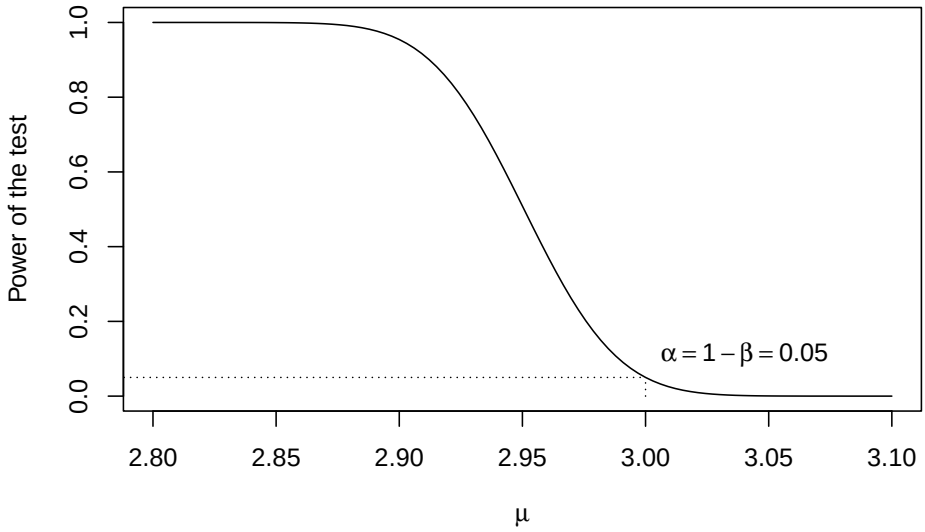


Figure 7.7: Power curve for the z-test in the example.

mean from the true value, when the true value is, for instance, 2.8 than in the case, when it is 2.95.

There is one specific point, where the power of the test coincides with the significance level. It is the case, when the population mean is indeed equal to 3. In this situation rejecting the null hypothesis would result in Type I error, which is equivalent to the significance level α, which we chose to be equal to 0.05.

In general, there are several things that influence the power of any statistical test:

1. The value of the true parameter;
2. The significance level;
3. The sample size;
4. The amount of information available for the test.

The element (1) is depicted on the plot in Figure 7.7. If we conduct the test about the wrong value of the true mean, then the distance from it will impact the power: the further it is away, the more powerful the test will be, being able to tell the difference between the sample mean and the true mean.

The element (2) will define the critical value of a statistical test, and in general the smaller the significance level is, the less powerful the test will be, as we will not be able to spot small discrepancies from the true mean.

The larger the sample size (element (3)), the more powerful the test becomes in general. In our example, we can see that from the equation (7.4), where the sample size n is in the denominator of the denominator. The higher values of n

will lead to the higher values of z and as a result, the higher chance of rejecting the H_0 if it is wrong. Figure 7.8 demonstrates how the power curve changes with the increase of the sample size. we see that the power of the test increases much faster with the decrease of the hypothetical value of μ, when the sample size is large (for example, $n = 1000$) than in the case of small samples (e.g., $n = 25$).

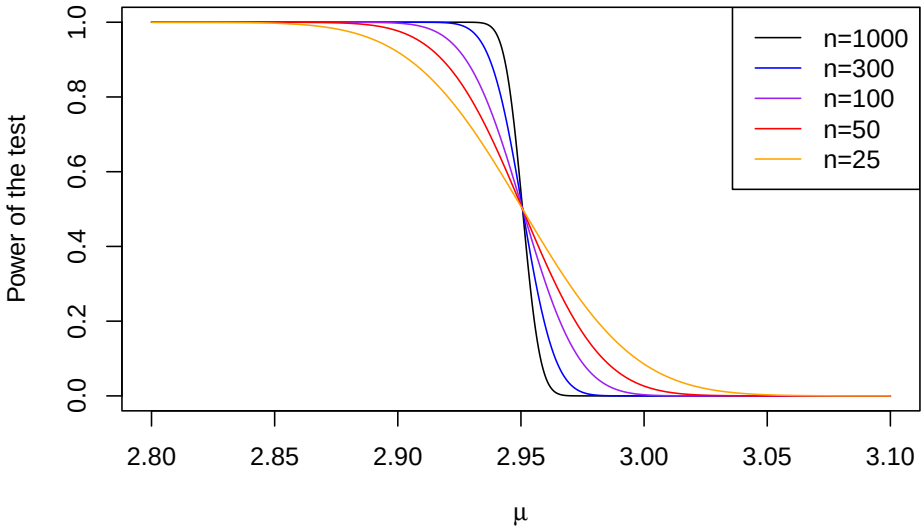


Figure 7.8: Power curves with different sample sizes.

Finally, the more general point about the "amount of information" applies to the selection between the tests. In the Section 8 we will discuss a variety of tests and we will discuss the conditions under which some of them are more powerful than the others. But in general the rule is: the more a priori information you can provide to the test, the easier it becomes to detect deviations from the tested value, because the uncertainty caused by estimation of additional parameters is decreased in this case.

7.3.3 Expected power of a test

As discussed in the previous subsections, the power of a test is calculated for each specific value of μ, measuring the probability of rejecting the wrong hypothesis for the specific parameters. This approach has a limitation, because typically we do not know the true value of μ, and sometimes selecting the correct one is challenging. One of the solutions in this case is calculating the **expected power of a test** or **Assurance** (O'Hagan et al., 2005), which is the expectation of the probability (7.3) for all possible values of μ for which the hypothesis would be correctly rejected. In a special case, when μ can only take discrete values, this can be roughly represented as a mean of powers for all values of μ that are lower than the value μ^* that corresponds to the critical one (in our example in

Subsection 7.3.2 μ^* was equal to 2.95):

$$\mathrm{E}(\mathrm{P}(\text{reject wrong } \mathrm{H}_0)) = \sum_{\mu_j < \mu^*} \mathrm{P}(\text{reject } \mathrm{H}_0 | \mu = \mu_j) \times \mathrm{P}(\mu = \mu_j). \tag{7.7}$$

Note that we sum up all the values up until μ^*, because the value higher than that would imply that we fail to reject the correct hypothesis. Note that in reality μ or any other parameter of distribution is typically continuous, which means that a probability density function should be used instead of probabilities and an integral should be used instead of the sum in (7.7), i.e.:

$$\mathrm{E}(\mathrm{P}(\text{reject wrong } \mathrm{H}_0)) = \int_{-\infty}^{\mu^*} \mathrm{P}(\text{reject } \mathrm{H}_0 | \mu = x) \times f(x) dx. \tag{7.8}$$

Remark. Note that the integration is done until μ^*, covering the values, where the hypothesis would be correctly rejected. In the example discussed in Subsection 7.3.2, we calculated in equation (7.5) that $\mu^* = 2.95$. If we had a different null hypothesis, we would do integration differently. For example, in case of $\mathrm{H}_0 : \mu \leq 3$, the integration would be done from 2.95 to ∞.

The calculation of the assurance in some cases can be done analytically. For example, O'Hagan et al. (2005) derive formulae for normal distribution for several tests. However, the analytical solution in some cases might be either too complicated, or unavailable. In these situations, the integration can be done numerically, for example, using Monte-Carlo simulations.

Finally, in practice, the assurance is used to determine the sample size for trials. The canonical example of application is deciding how many participants are needed to establish the effectiveness of a medical treatment, by setting the desired assurance to the pre-specified value.

7.4 Statistical and practical significance

We have already discussed what the significance level means and how it connects with Type I and Type II errors in the previous sections. We have noticed that the significance level α is non-linearly related with the Type II probability β and have spotted that there is a relation between them and subsequently relation between significance level and the power of a test. In fact, coming back to Table 7.1, we can say that there is a trade-off between them. If we use a very low significance level then we will make fewer mistakes when testing the correct hypothesis (failing to reject the correct H_0), but we will make more mistakes when the null hypothesis is wrong, because the power of the test will be lower. Sometimes, this can be compensated by choosing more powerful statistical tests (see Section 8) or increasing the sample size, but this is not always possible to do. On the other hand, choosing a high significance level means that we will make more Type I errors, rejecting the null hypothesis when it is correct, but at

the cost of making fewer Type II errors, rejecting the H_0, when it is wrong even if the difference between the true unknown mean and the sample one is small. This trade-off can be taken into account, when an analyst needs to decide what significance level to use.

If you are unsure what significance level to choose, Dave Worthington, a colleague of mine and a Statistics mentor at Lancaster University, has proposed an interesting motivation for that. If you do not have a level, driven by the problem (e.g. we need to satisfy 99% of demand, thus the significance level is 1%), then select the one for your life time. In how many times in your life would you be ready to make a mistake? Would it be 5%? 3%? 1%? Select something and stick with it. Then over the years you will know that you have made the selected proportion of mistakes, when conducting different statistical tests in various circumstances.

However, there is an important aspect that should also be considered when making decisions based on results of statistical tests - "practical significance". While it is not universally measurable, it is an aspect that is worth considering when making decisions. To better understand it, consider the following artificial example. A company creating helmets for cyclists has collected data about cyclists injuries for two cases: when they wear helmets and when they do not. They found that the cyclists that do not wear helmets get in car accidents less often than the cyclists that do. The probability of the event for them is just 1%, while it is 5% for the latter group (and the difference was statistically significant on 5%). Based on this a cyclist that have only started learning a statistics course can make a conclusion that they should not wear a helmet because it will decrease the chance of accident. However, the company has also analysed the types of injuries that cyclists get in case of accidents, and found that in 2% of the cases, those cyclists that do not wear helmets die in accidents, while this number for those that wear helmets is 1%. The company pointed out that the difference between the two situations was not statistically significant on 5% level. So, a person learning statistics would be inclined to conclude based on that they should not wear a helmet, because the two situations are not statistically different. This would be a wrong decision because it does not consider the practical significance.

Indeed, in the example above, based on the data collected by a company, the mortality in two cases is similar in terms of statistical significance. However, having the two times higher probability of death when not wearing a helmet than in the case with a helmet has serious practical implications. Getting in an accident is unpleasant and is associated with some costs (financial, moral etc), but dying has a much higher cost, incomparable with that. And even though the probability of dying in an accident without helmet is low (only 2%), it does not mean that a person not wearing a helmet will be lucky enough to appear in that 98% of cases, when an accident happens. Even 2% is enough for an event with such a critical outcome - this can happen any time with anyone. And although the probability of death in case of "wearing a helmet" is just 1% lower, given

Table 7.2: Making decisions in the case of practical vs statistical significance.

	Statistically significant	Statist
Practically significant	Make a decision	Think about
Practically insignificant	Think about it and do not make a decision	Do no

the asymmetry of costs, it is better to wear a helmet and increase the odds of survival by that one percent than to continue gambling. After all, your head is one of the most important parts of your body.

The situation above is artificial, I could not find appropriate data for this example. In reality, the numbers might be different, but the message is the same: you should consider practical implications of statistical analysis, when making decisions. Taking both statistical and practical significance into account, we can crate a table demonstrating the four possible cases for decision making (see Table 7.2)

In Table 7.2, there are two situations, when there is nothing to argue about: when practical and statistical significances agree with each other (either they are both significant or not). However, I argue that the practical significance is in general more important than the statistical one. If you find that a new decision will reduce costs but the reduction will not be statistically significant, then it makes sense to make that decision anyway. On the other hand, if the decision is statistically significant (for instance, it improves the process by 1%, being significant on the selected level), but it is not practically significant (the costs of implementing it are higher than the savings from it) then it should not be made. This is because the statistical outcomes are always associated with potential Type I and Type II errors discussed in Section 7.2 and thus not finding difference could be due to Type II error, while finding one could be due to Type I error. When it comes to making decisions, the results of statistical testing should only help in supporting them, rather than guiding them.

Chapter 8

Statistical tests

Having discussed the idea of hypothesis testing, we can now move to the discussion of specific tests. In this chapter, we will start the discussion with the tests for the means of random variables, then move towards the variance and to the comparison of several variables. The tests in this chapter are introduced based on the needs of an analyst. We finish the chapter with a discussion of how to select the appropriate statistical test for your problem.

Before we proceed, we need to define two terms, which will be used in this chapter. **Parametric statistical test** is the test that fully relies on distributional assumptions about the random variable. For example, we can assume that the variable follows Normal distribution and thus we can use a parametric test. **Non-parametric statistical test** is the test that does not rely on distributional assumptions. These types of test are typically conducted on ranked data rather than on the original one, reducing the scale of information to the ordinal one (see Section 1.2). The parametric ones are typically more powerful than their non-parametric counterparts if the assumptions hold. If they do not, the non-parametric ones become more powerful.

8.1 One-sample tests about mean

Consider an example, where we collected the data of height of 100 males across England for the 2021. Based on that sample we want to see if the average height is 175.3 cm, as it was claimed by NHS back in 2012. Based on our sample, we found that the mean height was 176.1 cm. Can we say that the height has increased or did we get this value just because of the pure randomness? How do we formulate and test such hypothesis? We need to follow the procedure described in Section 7.1. We start by formulating null and alternative hypotheses:

$$\mathrm{H}_0 : \mu = 175.3, \mathrm{H}_1 : \mu \neq 175.3.$$

Next, we select the significance level. If the level is not given to us, then we need to choose one based on our preferences. I like 1%, because I am risk averse.

After that, we select the appropriate test. The task described above focuses on investigating the hypothesis about the mean. If we can assume that the CLT holds (see Section 6.2), then we can use a parametric statistical test, because we know that the sample mean will follow Normal distribution in that case. In our example, we can indeed assume that it holds, because we deal with a sample of 100 observations, and we can also assume that the distribution of height across England is symmetric (we do not expect people to have extreme heights of, let's say, 5 meters or 0.5 meters). The next thing to consider is whether the population standard deviation of the mean height is known or not. Based on that, we would choose either z-test or t-test.

8.1.1 z-test

Consider the situation, where we know from NHS that the population standard deviation of height is $\sigma = 5$ (typically, we do not know it, except for some very rare cases, when this is given by design, e.g. due to how some machine is calibrated).

Remark. As shown in Section 6.4, if the standard deviation of y is σ, then the standard deviation of $\bar{y}$ is $\frac{1}{\sqrt{n}}\sigma$. This means that in our case $\sigma_{\bar{y}} = \frac{1}{\sqrt{100}} \times 5 = 0.5$.

If we assume that the mean follows Normal distribution and know the population standard deviation, i.e. $\bar{y} \sim \mathcal{N}\left(\mu, \sigma^2_{\bar{y}}\right)$, then the standardised value z:

$$z = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}}. \tag{8.1}$$

will follow standard normal distribution: $z \sim \mathcal{N}(0, 1)$. Knowing these properties, we can conduct the test using one of the three approaches discussed in Section 7.1. First, we could calculate the z value based on formula (8.1):

$$z = \frac{176.1 - 175.3}{0.5} = 1.6$$

and compare it with the critical one. Given that we test the two-sided hypothesis (because the alternative is "inequality"), he critical value should be split into two parts, to have $\frac{\alpha}{2}$ in one tail and $\frac{\alpha}{2}$ in the other one. The critical values can be calculated, for example, in R using the code below:

```
# I have previously selected significance level of 1%
alpha <- 0.01
qnorm(c(alpha/2, 1-alpha/2), 0, 1)
```

```
## [1] -2.575829  2.575829
```

We see that the calculated value lies inside the interval, so we fail to reject the null hypothesis on 1% significance level. The simplified version for this procedure

is to compare the absolute of the calculated value with the absolute of the critical one. If the calculated is greater than the critical, then we reject H_0. In our case $1.6 < 2.58$, so we fail to reject H_0.

Another way of testing the hypothesis is by calculating the p-value and comparing it with the significance level. In R, this can be done using the command:

```
(1-pnorm(abs(1.6)))*2
```

```
## [1] 0.1095986
```

In the R code above we are calculating the surface in the tails of distribution. Thus the appearance of the number 2, to add the surfaces in two tails. This procedure is shown in Figure 8.1.

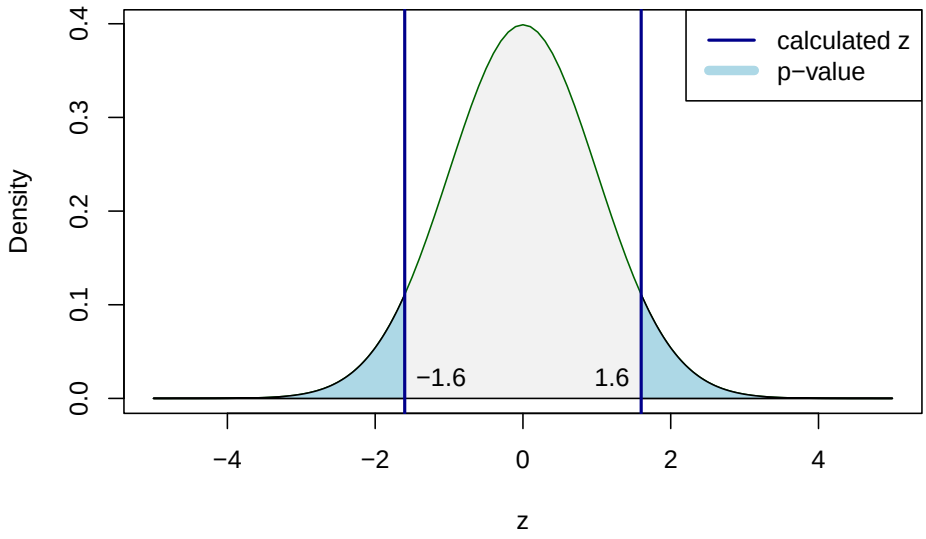


Figure 8.1: The process of p-value calculation for the z-test.

Comparing the p-value of 0.1096 with the significance level of 0.01, we can conclude that we fail to reject the null hypothesis. This means that based on the collected sample, we cannot tell the difference between the population mean height of 175.3 and the sample height of 176.1.

8.1.2 t-test

The population standard deviation is rarely known. The more practical test is the t-test, which relies on the relation between the normal distribution and the Student's distribution. If $y \sim \mathcal{N}(\mu, \sigma^2)$ and $s = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$ is the estimate of the sample standard deviation, then $t \sim \sqcup(n-1)$ where

$$t = \frac{\bar{y} - \mu}{s_{\bar{y}}} \tag{8.2}$$

and $n-1$ is the number of degrees of freedom. The reason why we need to use Student's distribution in this case rather than the Normal one is because of the uncertainty arising from the estimation of the standard deviation s. The hypothesis testing procedure in this case is exactly the same as in the case of z-test. We insert the values in (8.2) to get the calculated t, then compare it with the critical and make a conclusion. Consider an example, where the estimated standard deviation $s = 4$. We then get:

$$t = \frac{176.1 - 175.3}{0.4} = 2,$$

while the critical value for the chosen 1% significance level is:

```
qt(c(alpha/2, 1-alpha/2), 100-1)
```

```
## [1] -2.626405  2.626405
```

Given that the calculated value of 2 is lower than 2.626, we fail to reject the null hypothesis on 1% significance level. So, we again conclude that we cannot tell the difference between the mean in the data and the assumed population mean.

In R, the same procedures can be done using the `t.test()` function from `stats` package. Here is an example, demonstrating how the test can be done for the generated data:

```
y <- rnorm(100, 175.3, 5)
# Note that our significance level is 1%,
# so we ask to produce 99% confidence interval
t.test(y, mu=175.3, alternative="two.sided", conf.level=0.99)
```

```
##
##  One Sample t-test
##
## data:  y
## t = 1.8509, df = 99, p-value = 0.06717
## alternative hypothesis: true mean is not equal to 175.3
## 99 percent confidence interval:
##  174.9013 177.6018
## sample estimates:
## mean of x
##  176.2516
```

Remark. If we were to test a one-sided hypothesis (e.g. $H_0 : \mu < 175.3, H_1 : \mu \geq 175.3$), then we would need to change the `alternative` parameters in the `t.test()` function to correspond to the formulated H_1.

The output above shows the calculated t (1.8509), the number of degrees of freedom and the p-value. It also constructs the 99% confidence interval (174.9013, 177.6018). The conclusions can be made using one of the three approaches, discussed above and in Section 7.1:

1. The calculated value is 1.8509, which is lower than the critical one of 2.6264 as discussed earlier, so we fail to reject H_0;
2. The p-value is 0.0672, which is greater than the selected significance level of 1%, so we fail to reject H_0;
3. The 99% confidence interval includes the tested value of 175.3, so we fail to reject the H_0 on the 99% confidence level.

As mentioned in Section 7.1, I personally prefer the last approach of the three because it gives more information about the uncertainty around the estimate of the sample mean.

8.1.3 Non-parametric, one-sample Wilcoxon test

In some situations, the CLT might not hold due to violation of some of assumptions. For example, the distribution of the random variable is expected to be asymmetric with a long tail. In this case, the mean might not be finite and thus the CLT would not hold. Alternatively, the sample size might be too small to assume that CLT has started working. In these cases, the parametric tests for the mean will not be powerful enough (see discussion in Section 7.3) and would lead to wrong conclusions about the tested hypothesis.

One of the solutions in this situation is a non-parametric test that does not have distributional assumptions. In the case of the hypothesis about the mean of a random variable, we could use Wilcoxon test. The null hypothesis in this test can be formulated in a variety of ways, one of which is the following:

$$\begin{aligned} H_0 &: \text{ distribution is symmetric around } \mu = 175.3, \\ H_1 &: \text{ distribution is not symmetric around } \mu = 175.3. \end{aligned}$$

If the H_0 is true in this case, then it means that the mean will coincide with the centre of distribution, which should be around the tested value. If it is not symmetric, then possibly the centre of distribution is not around the tested value. The test is done on the ranked data, sorting the values of y from the lowest to the highest and assigning the numerical values to them. After that the test values is calculated.

Given that the test does not rely on distributional assumptions, it is less powerful than the parametric tests on large samples, but it is also more powerful on the small ones.

In R, the test can be conducted using `wilcox.test()` function from `stats` package:

```
wilcox.test(y, mu=175.3, alternative="two.sided")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  y
```

```
## V = 3039, p-value = 0.07747
## alternative hypothesis: true location is not equal to 175.3
```

Similar to how we have done that with t-test, we can compare the p-value with the significance level (reminder: we have chosen 1%) and make a conclusion. Based on the output above, we fail to reject H_0 because 0.0775>0.01.This means that once again, we cannot tell the difference between the sample mean and the population mean of 175.3.

8.2 One-sample test about variance

Another example of a situation that could be potentially interesting in practice is when we are not sure about the estimated variance of a distribution. We might want to understand, whether the variance in population is similar to the value we obtained based on our sample. For illustrative purposes, consider the continued example with the height of humans. After collecting the sample of 30 observations, it was found that the variance of height is 100. However, based on previous survey, we know that the population variance of the height is 121. We want to know whether the in-sample estimate of variance is significantly lower from the population one on 1% significance level. This hypothesis can be formulated as:

$$H_0 : \sigma^2 \geq 121, H_1 : \sigma^2 < 121.$$

where σ^2 is the population variance. The conventional test for this hypothesis is the Chi-squared test (χ^2). This is a parametric test, which means that it relies on distributional assumptions. However, unlike the test about the mean from the Section 8.1 (which assumed that CLT holds, see Section 6.2), the Chi-squared test relies on the assumption about the random variable itself:

$$y_j \sim \mathcal{N}(\mu, \sigma^2),$$

which means that (as discussed in Section 8.1.1):

$$z_j = \frac{y_j - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Coming back to the variance, we would typically use the following formula to estimate it in sample:

$$\mathrm{V}(y) = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2,$$

where $\bar{y}$ is an unbiased, efficient and consistent estimate of μ (Section 6.3). If we divide the variance by the true value of the population one, we will get the following:

$$\frac{\mathrm{V}(y)}{\sigma^2} = \frac{1}{n-1} \sum_{j=1}^{n} \frac{(y_j - \bar{y})^2}{\sigma^2},$$

or after multiplying it by $n-1$:

$$\chi^2 = (n-1)\frac{\mathrm{V}(y)}{\sigma^2} = \sum_{j=1}^{n} z_j^2. \tag{8.3}$$

Given the assumption of normality of the random variable z_j, the variable (8.3) will follow the Chi-squared distribution with $n-1$ degrees of freedom (this is the definition of the Chi-squared distribution). This property can be used to test a statistical hypothesis about the variance.

The hypothesis testing itself is done using the same procedure as before (Section 7.1). After formulating the hypothesis and selecting the significance level, we use the formula (8.3) to get the calculated value:

$$\chi^2 = (30-1) \times \frac{100}{121} \approx 23.97$$

Given that the alternative hypothesis in our example is "lower than", we should compare the obtained value with the critical one from the left tail of the distribution, which on 1% significance level is:

```
qchisq(0.01,29)
```

```
## [1] 14.25645
```

Comparing the critical and calculated values ($23.97 > 14.26$), we conclude that we fail to reject the null hypothesis. This means that the sample variance is not statistically different from the population one. The test of this hypothesis is shown visually in Figure 8.2, which shows that the calculated value is not in the tail of the distribution (thus "fail to reject").

A thing to keep in mind about the Chi-square distribution is that it is in general asymmetric (it converges to normal with the increase of the sample size, see discussion in Chapter 4). This means that if the alternative hypothesis is formulated as "not equal", the absolutes of critical values for the left and right tails will differ. This situation is shown in Figure 8.3 with the example of 5% significance level, which is split into two equal parts of 2.5% with critical values of 16.05 and 45.72.

The surfaces in the tails in Figure 8.3 are equal, but because of the asymmetry of distribution they look different.

Remark. If an analyst needs to conduct the test about the variance of the mean, then only CLT is required to hold, no additional assumptions about the distribution of the random variable y_j are required.

Chi-squared test is also used in many other situations, for example to test the relation between categorical variables (see Section 9.1) or to test the "goodness of fit". We do not discuss them in this Chapter.

Finally, there are non-parametric analogues of the Chi-squared test, but their discussion lies outside of the scope of this textbook.

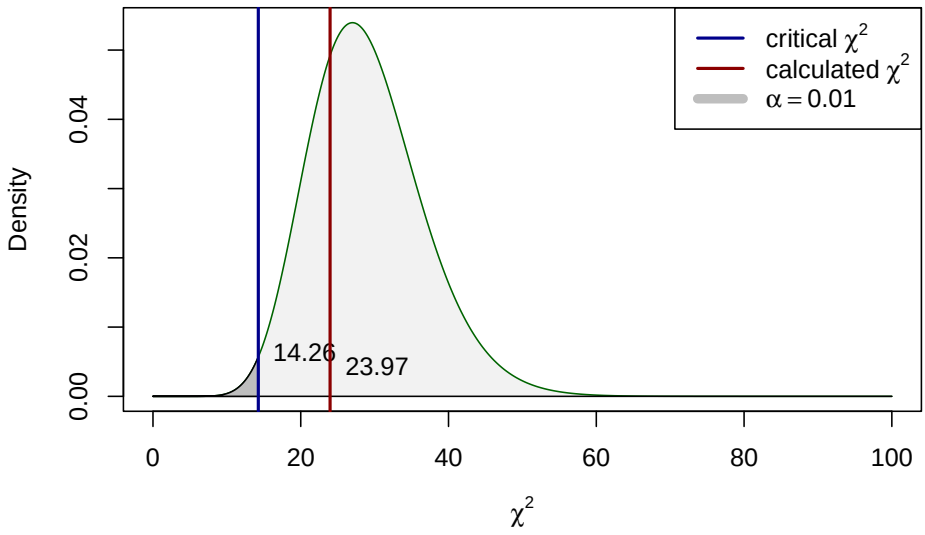


Figure 8.2: The process of hypothesis testing in Chi-squared distribution (one-tailed case).

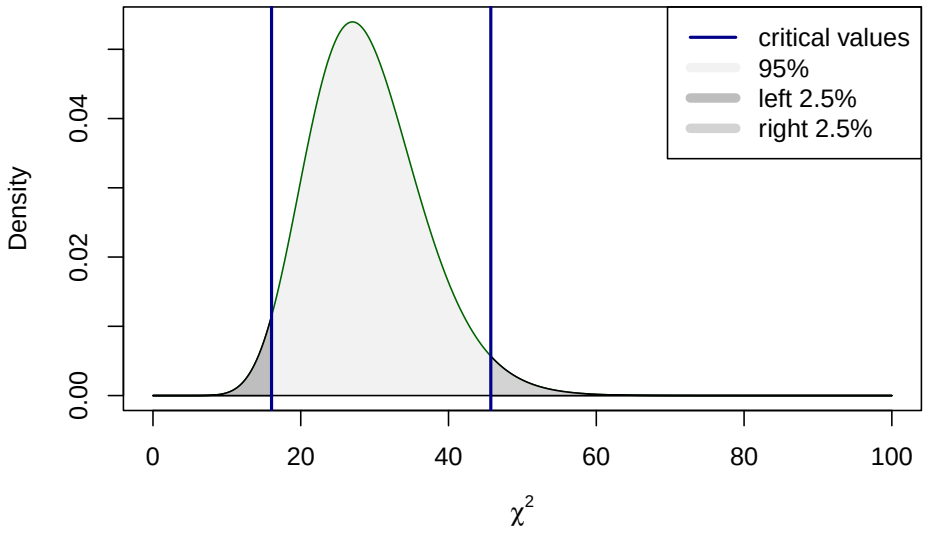


Figure 8.3: The process of hypothesis testing in Chi-squared distribution (two-tailed case).

Chapter 9

Measuring relations between variables

Now that we have discussed confidence intervals and hypothesis testing, we can move towards the analysis of relations between variables, in a way continuing the preliminary data analysis that we finished in Section 5. We continue using the same dataset `mtcarsData` with the two categorical variables, `am` and `vs`.

9.1 Nominal scale

As discussed in Section 1.2, not all scales support the more advanced operations (such as taking mean in ordinal scale). This means that if we want to analyse relations between variables, we need to use appropriate instrument. The coefficients that show relations between variables are called "**measures of association**". We start their discussions with the simplest scale - nominal.

There are several measures of association for the variables in nominal scale. They are all based on calculating the number of specific values of variables, but use different formulae. The first one is called **contingency coefficient** ϕ and can only be calculated between variables that have only two values. As the name says, this measure is based on the contingency table. Here is an example:

```
table(mtcarsData$vs,mtcarsData$am)
```

```
##
##            automatic manual
##   V-shaped        12      6
##   Straight         7      7
```

The ϕ coefficient is calculated as:

$$\phi = \frac{n_{1,1}n_{2,2} - n_{1,2}n_{2,1}}{\sqrt{n_{1,\cdot} \times n_{2,\cdot} \times n_{\cdot,1} \times n_{\cdot,2}}}, \tag{9.1}$$

where $n_{i,j}$ is the element of the table on row i and column j, $n_{i,\cdot} = \sum_j n_{i,j}$ - is the sum in row i and $n_{\cdot,j} = \sum_i n_{i,j}$ - is the sum in column j. This coefficient lies between -1 and 1 and has a simple interpretation: if will be close to 1, when the elements on diagonal are greater than the off-diagonal ones, implying that there is a relation between variables. The value of -1 can only be obtained, when off-diagonal elements are non-zero, while the diagonal ones are zero. Finally, if the values in the contingency table are distributed evenly, the coefficient will be equal to zero. In our case the value of ϕ is:

```
(12*7 - 6*7)/sqrt(19*13*14*18)
```

```
## [1] 0.1683451
```

This is a very low value, so even if the two variables are related, the relation is not well pronounced. In order to see, whether this value is statistically significantly different from zero, we could test a statistical hypothesis (hypothesis testing was discussed in Section 7):

H_0: there is no relation between variables

H_1: there is some relation between variables

This can be done using χ^2 **test** (we discussed it in a different context in Section 8.2), the statistics for which is calculated via:

$$\chi^2 = \sum_{i,j} \frac{n \times n_{i,j} - n_{i,\cdot} \times n_{\cdot,j}}{n \times n_{i,\cdot} \times n_{\cdot,j}}, \tag{9.2}$$

where n is the sum of elements in the contingency table. The value calculated based on (9.2) will follow χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in contingency table. This is a proper statistical test, so it should be treated as one. We select my favourite significance level, 1% and can now conduct the test:

```
chisq.test(table(mtcarsData$vs,mtcarsData$am))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(mtcarsData$vs, mtcarsData$am)
## X-squared = 0.34754, df = 1, p-value = 0.5555
```

Given that p-value is greater than 1%, we fail to reject the null hypothesis and can conclude that the relation does not seem to be different from zero - we do not find a relation between the variables in our data.

The main limitation of the coefficient ϕ is that it only works for the 2×2 tables. In reality we can have variables in nominal scale that take several values and it might be useful to know relations between them. For example, we can have a variable `colour`, which takes values `red`, `green` and `blue` and we would want to know if it is related to the transmission type. We do not have this variable in the data, so just for this example, we will create one (using multinomial distribution):

```
colour <- c(1:3) %*% rmultinom(nrow(mtcars), 1,
                               c(0.4,0.5,0.6))
colour <- factor(colour, levels=c(1:3),
                 labels=c("red","green","blue"))
barplot(table(colour), xlab="Colour")
```

In order to measure relation between the new variable and the `am`, we can use Cramer's V coefficient, which relies on the formula of χ^2 test (9.2):

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}. \tag{9.3}$$

Cramer's V always lies between 0 and 1, becoming close to one only if there is some relation between the two categorical variables. `greybox` package implements this coefficient in `cramer()` function:

```
cramer(mtcarsData$am,colour)
```

```
## Cramer's V: 0
## Chi^2 statistics = 0.3222, df: 2, p-value: 0.8512
```

The output above shows that the value of the coefficient is approximately 0.1, which is low, implying that the relation between the two variables is very weak. In addition, the p-value tells us that we fail to reject the null hypothesis on 1% level in the χ^2 test (9.2), and the relation does not look statistically significant. So we can conclude that according to our data, the two variables are not related (no wonder, we have generated one of them).

The main limitation of Cramer's V is that it is difficult to interpret beyond "there is a relation". Imagine a situation, where the colour would be related to the variable "class" of a car, that can take 5 values. What could we say more than to state the fact that the two are related? After all, in that case you end up with a contingency table of 3×5, and it might not be possible to say how specifically one variable changes with the change of another one. Still, Cramer's V at least provides some information about the relation of two categorical variables.

9.2 Ordinal scale

As discussed in Section 1.2, ordinal scale has more flexibility than the nominal one - its values have natural ordering, which can be used, when we want to

measure relations between several variables in ordinal scale. Yes, we can use Cramer's V and χ^2 test, but this way we would not be using all advantages of the scale. So, what can we use in this case? There are three popular measures of association for variables in ordinal scale:

1. Goodman-Kruskal's γ,
2. Yule's Q,
3. Kendall's τ.

Given that the ordinal scale does not have distances, the only thing we can do is to compare values of variables between different observations and say, whether one is greater than, less than or equal to another. What can be done with two variables in ordinal scale is the comparison of the values of those variables for a couple respondents. Based on that the pairs of the observations can be called:

1. **Concordant** if both $x_1 < x_2$ and $y_1 < y_2$ or $x_1 > x_2$ and $y_1 > y_2$ - implying that there is an agreement in order between the two variables (e.g. with a switch from a younger age group to the older one, the size of the T-shirt will switch from S to M);
2. **Discordant** if for $x_1 < x_2$ and $y_1 > y_2$ or for $x_1 > x_2$ and $y_1 < y_2$ - implying that there is a disagreement in the order of the two variables (e.g. with a switch from a younger age group to the older one, the satisfaction from drinking Coca-Cola will switch to the lower level);
3. **Ties** if both $x_1 = x_2$ and $y_1 = y_2$;
4. Neither otherwise (e.g. when $x_1 = x_2$ and $y_1 < y_2$).

All the measures of association for the variables in ordinal scale rely on the number of concordant, discordant variables and number of ties. All of these measures lie in the region of [-1, 1].

Goodman-Kruskal's γ is calculated using the following formula:

$$\gamma = \frac{n_c - n_d}{n_c + n_d}, \tag{9.4}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs. This is a very simple measure of association, but it only works with scales of the same size (e.g. 5 options in one variable and 5 options in the other one) and ignores the ties.

In order to demonstrate this measure in action, we will create two artificial variables in ordinal scale:

1. Age of a person: young, adult and elderly;
2. Size of t-shirt they wear: S, M or L.

Here how we can do that in R:

```
age <- c(1:3) %*% rmultinom(nrow(mtcars), 1,
                            c(0.4,0.5,0.6))
age <- factor(age, levels=c(1:3),
```

```
                  labels=c("young","adult","elderly"))
size <- c(1:3) %*% rmultinom(nrow(mtcars), 1,
                             c(0.3,0.5,0.7))
size <- factor(size, levels=c(1:3),
                  labels=c("S","M","L"))
```

And here is how the relation between these two artificial variables looks:

```
tableplot(age,size,xlab="Age",ylab="T-shirt size")
```

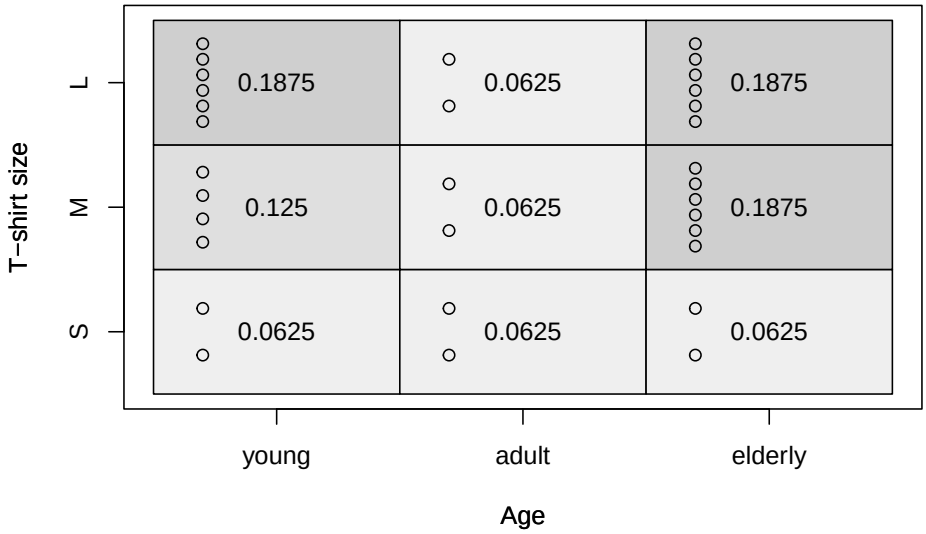


Figure 9.1: Heat map for age of a respondent and the size of their t-shirt.

The graphical analysis based on Figure 9.1 does not provide a clear information about the relation between the two variables. But this is where the Goodman-Kruskal's γ becomes useful. We will use `GoodmanKruskalGamma()` function from `DescTools` package for R for this:

```
DescTools::GoodmanKruskalGamma(age,size,conf.level=0.95)
```

```
##       gamma      lwr.ci      upr.ci
## -0.03846154 -0.51302449  0.43610141
```

This function returns three values: the γ, which is close to zero in our case, implying that there is no relation between the variables, lower and upper bounds of the 95% confidence interval. Note that the interval shows us how big the uncertainty about the parameter is: the true value in the population can be anywhere between -0.51 and 0.44. But based on all these values we can conclude that we do not see any relation between the variables in our sample.

The next measure is called **Yule's Q** and is considered as a special case of

Goodman-Kruskal's γ for the variables that only have 2 options. It is calculated based on the resulting contingency 2×2 table and has some similarities with the contingency coefficient ϕ:

$$\mathrm{Q} = \frac{n_{1,1}n_{2,2} - n_{1,2}n_{2,1}}{n_{1,1}n_{2,2} + n_{1,2}n_{2,1}}. \tag{9.5}$$

The main difference from the contingency coefficient is that it assumes that the data has ordering, it implicitly relies on the number of concordant (on the diagonal) and discordant (on the off diagonal) pairs. In our case we could calculate it if we had two simplified variables based on age and size (in real life we would need to recode them to "young", "older" and "S", "Bigger than S" respectively):

```
table(age,size)[1:2,1:2]
```

```
##        size
## age     S M
##   young 2 4
##   adult 2 2
```

```
(2*2-4*2)/(2*2+4*2)
```

```
## [1] -0.3333333
```

In our toy example, the measure shows that there is a weak negative relation between the trimmed age and size variables. We do not make any conclusions based on this, because this is not meaningful and is shown here just for purposes of demonstration.

Finally, there is **Kendall's** τ. In fact, there are three different coefficients, which have the same name, so in the literature they are known as τ_a, τ_b and τ_c.

τ_a coefficient is calculated using the formula:

$$\tau_a = \frac{n_c - n_d}{\frac{T(T-1)}{2}}, \tag{9.6}$$

where T is the number of observations, and thus in the denominator, we have the number of all the pairs in the data. In theory this coefficient should lie between -1 and 1, but it does not solve the problem with ties, so typically it will not reach the boundary values and will say that the relation is weaker than it really is. Similar to Goodman-Kruskal's γ, it can only be applied to the variables that have the same number of levels (same sizes of scales). In order to resolve some of these issues, τ_b was developed:

$$\tau_b = \frac{n_c - n_d}{\sqrt{\left(\frac{T(T-1)}{2} - n_x\right)\left(\frac{T(T-1)}{2} - n_y\right)}}, \tag{9.7}$$

where n_x and n_y are the number of ties calculated for both variables. This coefficient resolves the problem with ties and can now reach the boundary values

in practice. However, this coefficient does not resolve the issue with different scale sizes. And in order to address this problem, we have τ_c (**Stuart-Kendall's** τ_c):

$$\tau_c = \frac{n_c - n_d}{\frac{n^2}{2}\frac{\min(r,c)-1}{\min(r,c)}}, \tag{9.8}$$

where r is the number of rows and c is the number of columns. This coefficient works for variables with different lengths of scales (e.g. age with 5 options and t-shirt size with 7 options). But now we are back to the problem with the ties...

In R, the `cor()` function implements Kendall's τ_a and τ_b (the function will select automatically based on the presence of ties). There are also functions `KendallTauA()`, `KendallTauB()` and `StuartTauC()` in `DescTools` package that implement the three respective measures of association. The main limitation of `cor()` function is that it only works with numerical variables, so we would need to transform variables before applying the function. The functions from `DescTools` package, on the other hand, work with factors. Here are the values of the three coefficients for our case:

```
DescTools::KendallTauA(age,size,conf.level=0.95)
```

```
##       tau_a      lwr.ci      upr.ci
## -0.01612903 -0.15347726  0.12121920
```

```
DescTools::KendallTauB(age,size,conf.level=0.95)
```

```
##       tau_b      lwr.ci      upr.ci
## -0.02469136 -0.32938991  0.28000720
```

```
DescTools::StuartTauC(age,size,conf.level=0.95)
```

```
##       tauc     lwr.ci     upr.ci
## -0.0234375 -0.3126014  0.2657264
```

Given that both variables have the same scale sizes, we should use either τ_a or τ_b for the analysis. However, we do not know if there are any ties in the data, so the safer option would be to use τ_b coefficient. The value of the coefficient and its confidence interval tell us that there is no obvious association between the two variables in our sample. This is expected, because the two variables were generated independently of each other.

9.3 Numerical scale

Finally we come to the discussion of relations between variables measured in numerical scales. The most famous measure in this category is the **Pearson's correlation coefficient**, which population value is:

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}, \tag{9.9}$$

where $\sigma_{x,y}$ is the covariance between variables x and y (see discussions in Sections 5.1 and 10.2), while σ_x and σ_y are standard deviations of these variables. Typically, we do not know the population values, so this coefficient can be estimated in sample via:

$$r_{x,y} = \frac{\text{cov}(x,y)}{\sqrt{V(x)V(y)}}, \tag{9.10}$$

where all the values from (9.9) are substituted by their in-sample estimates. This coefficient measures the **strength of linear relation** between variables and lies between -1 and 1, where the boundary values correspond to perfect linear relation and 0 implies that there is no **linear** relation between the variables. In some textbooks the authors claim that this coefficient relies on Normal distribution of variables, but nothing in the formula assumes that. It was originally derived based on the simple linear regression (see Section 10) and its rough idea is to get information about the angle of the straight line drawn on the scatterplot. It might be easier to explain this on an example:

```
plot(mtcarsData$disp,mtcarsData$mpg,
     xlab="Displacement",ylab="Mileage")
abline(lm(mpg~disp,mtcarsData),col="red")
```

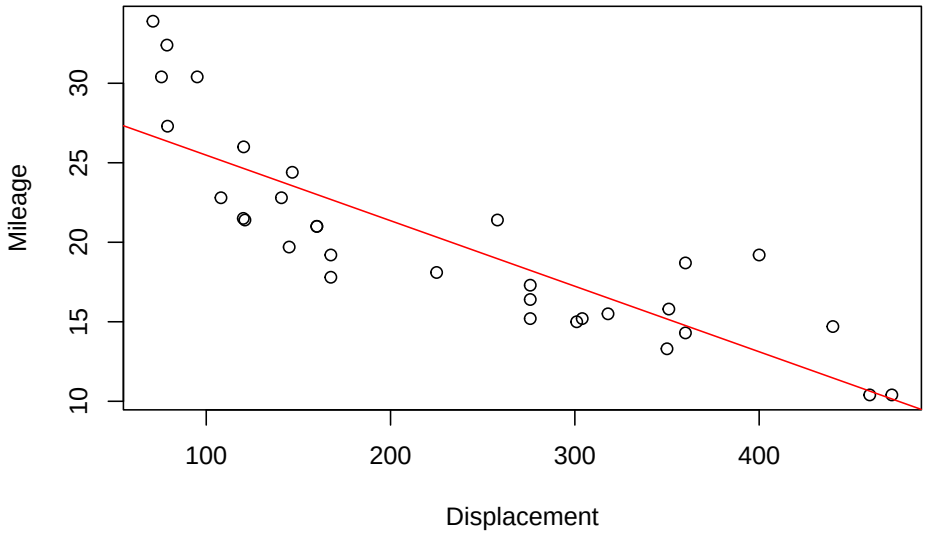


Figure 9.2: Scatterplot for dispalcement vs mileage variables in mtcars dataset

Figure 9.2 shows the scatterplot between the two variables and also has the straight line, going through the cloud of points. The closer the points are to the line, the stronger the linear relation between the two variables is. The line corresponds to the formula $\hat{y} = a_0 + a_1 x$, where x is the displacement and $\hat{y}$ is the line value for the Mileage. The same relation can be presented if we swap the axes and draw the line $\hat{x} = b_0 + b_1 y$:

```
plot(mtcarsData$mpg,mtcarsData$disp,
     xlab="Mileage",ylab="Displacement")
abline(lm(disp~mpg,mtcarsData),col="red")
```

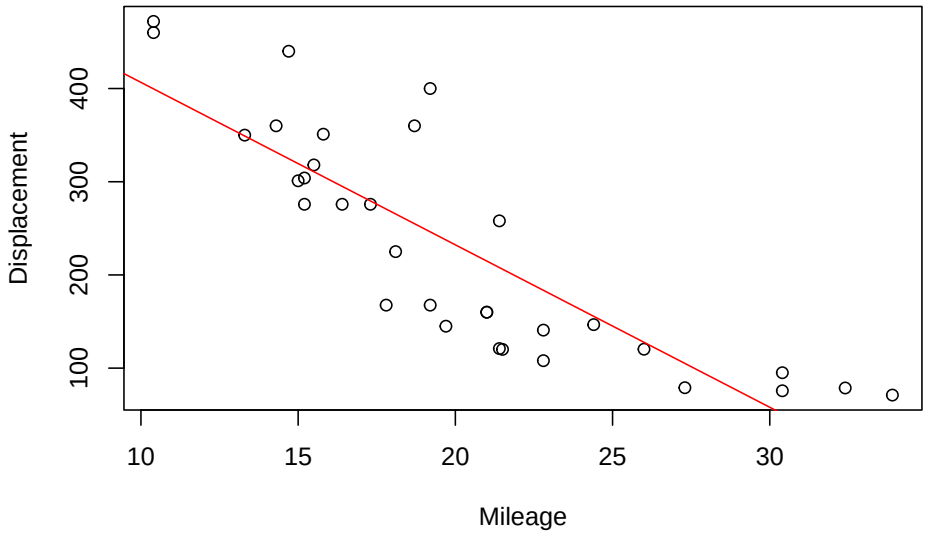


Figure 9.3: Scatterplot for mileage vs dispalcement

The slopes for the two lines will in general differ, and will only coincide if the two variables have functional relations (all the point lie on the line). Based on this property, the correlation coefficient was originally constructed, as a geometric mean of the two parameters of slopes: $r_{x,y} = \sqrt{a_1 b_1}$. We will come back to this specific formula later in Section 10. But this idea provides an explanation why the correlation coefficient measures the strength of linear relation. For the two variables of interest it will be:

```
cor(mtcarsData$mpg,mtcarsData$disp)
```

```
## [1] -0.8475514
```

Which shows strong negative linear relation between the displacement and mileage. This makes sense, because in general the cars with bigger engines will have bigger consumption and thus will make less miles per gallon of fuel. The more detailed information about the correlation is provided by the `cor.test()` function:

```
cor.test(mtcarsData$mpg,mtcarsData$disp)
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  mtcarsData$mpg and mtcarsData$disp
## t = -8.7472, df = 30, p-value = 9.38e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9233594 -0.7081376
## sample estimates:
##        cor
## -0.8475514
```

In addition to the value, we now have results of the hypothesis testing (where null hypothesis is $\rho_{x,y} = 0$) and the confidence interval for the parameter. Given that the value of the parameter is close to its bound, we could conclude that the linear relation between the two variables is strong and statistically significant on 1% level.

Note that the value of correlation coefficient only depends on the distance of points from the straight line, it does not depend on the slope (excluding case, when slope is equal to zero and thus the coefficient is equal to zero as well). So the following two cases will have exactly the same correlation coefficients:

```
error <- rnorm(100,0,10)
x <- c(1:100)
y1 <- 10+0.5*x+0.5*error
y2 <- 2+1.5*x+1.5*error
# Produce the plots
par(mfcol=c(1,2))
plot(x,y1,ylim=c(0,200))
abline(lm(y1~x),col="red")
text(30,150,paste0("r=",round(cor(x,y1),5)))
plot(x,y2,ylim=c(0,200))
abline(lm(y2~x),col="red")
text(30,150,paste0("r=",round(cor(x,y2),5)))
```

There are other examples of cases, when correlation coefficient would be misleading or not provide the necessary information. One of the canonical examples is the Anscombe's quartet (Wikipedia, 2021), which shows very different types of relations, for which the Pearson's correlation coefficient would be exactly the same. An important lesson from this is to always do graphical analysis (see Section 5.2) of your data, when possible - this way misleading situations can be avoided.

Coming back to the scatterplot in Figure 9.2, it demonstrates some non-linearity in the relation between the two variables. So, it would make sense to have a different measure that could take it into account. This is where **Spearman's correlation coefficient** becomes useful. It is calculated using exactly the same formula (9.10), but applied to the data in ranks. By using ranks, we loose information about the natural zero and distances between values of the variable, but at the same time we linearise possible non-linear relations. So, Spearman's

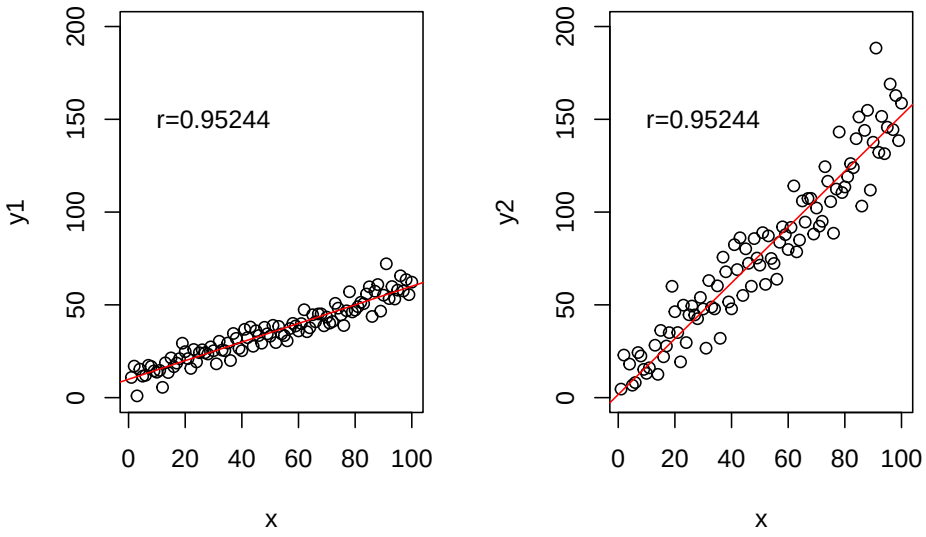


Figure 9.4: Example of relations with exactly the same correlations, but different slopes.

coefficient shows the strength of monotonic relation between the two variables:

```
cor.test(mtcarsData$mpg,mtcarsData$disp,
         method="spearman")
```

```
## Warning in cor.test.default(mtcarsData$mpg, mtcarsData$disp, method =
## "spearman"): Cannot compute exact p-value with ties
```

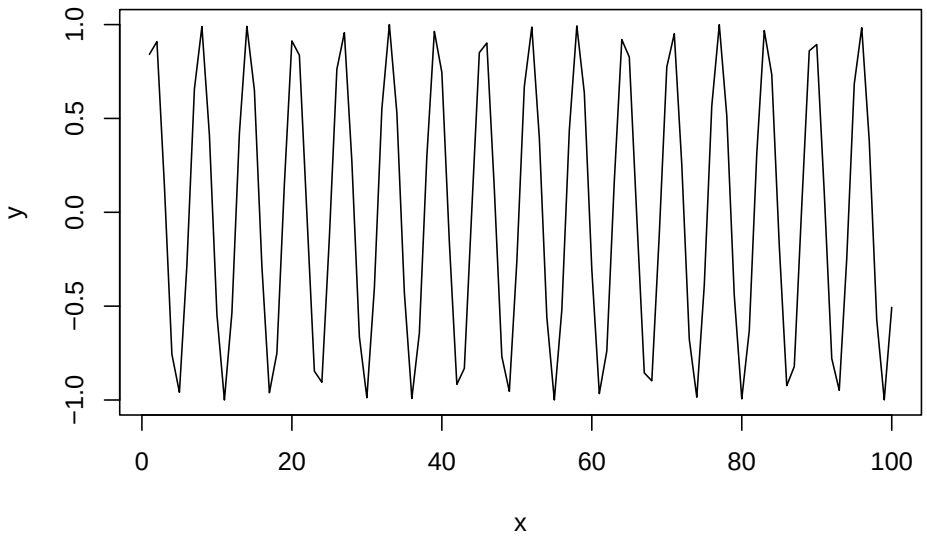
```
##
##  Spearman's rank correlation rho
##
## data:  mtcarsData$mpg and mtcarsData$disp
## S = 10415, p-value = 6.37e-13
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.9088824
```

We can notice that the value of the Spearman's coefficient in our case is higher than the value of the Pearson's correlation, which implies that there is indeed non-linear relation between variables. The two variables have a strong monotonic relation, which makes sense for the reasons discussed earlier. The non-linearity makes sense as well because the car with super powerful engines would still be able to do several miles on a gallon of fuel, no matter what. The relation will never be zero or even negative.

Note that while Spearman's correlation will tell you something about monotonic

relations, it will fail to capture all other non-linear relations between variables. For example, in the following case the true relation is trigonometric:

```
x <- c(1:100)
y <- sin(x)
plot(x,y,type="l")
```



But neither Pearson's nor Spearman's coefficients will be able to capture it:

```
cor(x,y)
```

```
## [1] -0.04806497
```

```
cor(x,y,method="spearman")
```

```
## [1] -0.04649265
```

In order to correctly diagnose such non-linear relation, either one or both variables need to be transformed to linearise the relation. In our case this implies measuring the relation between y and $\sin(x)$ instead of y and x:

```
cor(sin(x),y)
```

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

9.4 Mixed scales

Finally, when we have two variables measured in different scales, the general recommendation is to use the measure of association for the lower scale. For example, if we have the nominal variable colour and the ordinal variable size (both related to T-shirts people prefer), we should use Cramer's V in order to

measure the relation between them:

```
cramer(colour,size)
```

```
## Cramer's V: 0.1632
## Chi^2 statistics = 5.7241, df: 4, p-value: 0.2207
```

Similarly, if we have a numerical and ordinal variables, we should use one of the measures for ordinal scales.

However, in some cases we might be able to use a different measure of association. One of those is called multiple correlation coefficient and can be calculated for variables in numerical vs categorical scales. This coefficient can be calculated using different principles, the simplest of which is constructing a regression model (discussed later in Section 11) of numerical variable from the dummy variables (see Section 13) created from the categorical one and then extracting the square root of coefficient of determination (discussed in Section 10.4). The resulting coefficient lies between 0 and 1, where 1 implies perfect linear relation between the two variables and 0 implies no linear relation between them. `mcor()` function from `greybox` implements this:

```
mcor(mtcars$am, mtcars$mpg)
```

```
## Multiple correlations value: 0.5998
## F-statistics = 16.8603, df: 1, df resid: 30, p-value: 3e-04
```

Based on the value above, we can conclude that the type of transmission has a linear relation with the mileage. This aligns with what we have already discovered earlier, in preliminary analysis section (Section 5.2) in Figure 5.15.

Finally, there is a function `assoc()` (aka `association()`) in `greybox` package, which will automatically select the necessary measure of association based on the type of a variable and produce three matrices: 1. measures of association, 2. p-values for testing H_0 that there measure is equal to zero, 3. names of functions used for each pair. Here how it works for the `mtcarsData` example:

```
assoc(mtcarsData)
```

```
## Associations:
## values:
##          mpg    cyl    disp      hp    drat      wt    qsec     vs     am
## mpg   1.0000 0.8558 -0.8476 -0.7762  0.6812 -0.8677  0.4187 0.6640 0.5998
## cyl   0.8558 1.0000  0.9152  0.8449  0.7018  0.7826  0.5913 0.7889 0.4643
## disp -0.8476 0.9152  1.0000  0.7909 -0.7102  0.8880 -0.4337 0.7104 0.5912
## hp   -0.7762 0.8449  0.7909  1.0000 -0.4488  0.6587 -0.7082 0.7231 0.2432
## drat  0.6812 0.7018 -0.7102 -0.4488  1.0000 -0.7124  0.0912 0.4403 0.7127
## wt   -0.8677 0.7826  0.8880  0.6587 -0.7124  1.0000 -0.1747 0.5549 0.6925
## qsec  0.4187 0.5913 -0.4337 -0.7082  0.0912 -0.1747  1.0000 0.7445 0.2299
## vs    0.6640 0.7889  0.7104  0.7231  0.4403  0.5549  0.7445 1.0000 0.0000
## am    0.5998 0.4643  0.5912  0.2432  0.7127  0.6925  0.2299 0.0000 1.0000
```

```
## gear  0.6551 0.4820  0.7671  0.6638  0.8319  0.6587  0.6334 0.5728 0.7808
## carb  0.6667 0.4847  0.5605  0.7873  0.3344  0.6129  0.6695 0.5733 0.1864
##        gear   carb
## mpg  0.6551 0.6667
## cyl  0.4820 0.4847
## disp 0.7671 0.5605
## hp   0.6638 0.7873
## drat 0.8319 0.3344
## wt   0.6587 0.6129
## qsec 0.6334 0.6695
## vs   0.5728 0.5733
## am   0.7808 0.1864
## gear 1.0000 0.3217
## carb 0.3217 1.0000
##
## p-values:
##         mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear
## mpg  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0171 0.0000 0.0003 0.0003
## cyl  0.0000 0.0000 0.0000 0.0000 0.0001 0.0000 0.0020 0.0000 0.0126 0.0012
## disp 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0131 0.0000 0.0004 0.0000
## hp   0.0000 0.0000 0.0000 0.0000 0.0100 0.0000 0.0000 0.0000 0.1798 0.0002
## drat 0.0000 0.0001 0.0000 0.0100 0.0000 0.0000 0.6196 0.0117 0.0000 0.0000
## wt   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.3389 0.0010 0.0000 0.0003
## qsec 0.0171 0.0020 0.0131 0.0000 0.6196 0.3389 0.0000 0.0000 0.2057 0.0006
## vs   0.0000 0.0000 0.0000 0.0000 0.0117 0.0010 0.0000 0.0000 0.5555 0.0022
## am   0.0003 0.0126 0.0004 0.1798 0.0000 0.0000 0.2057 0.5555 0.0000 0.0000
## gear 0.0003 0.0012 0.0000 0.0002 0.0000 0.0003 0.0006 0.0022 0.0000 0.0000
## carb 0.0065 0.0066 0.0662 0.0001 0.6607 0.0242 0.0061 0.0090 0.2838 0.0857
##        carb
## mpg  0.0065
## cyl  0.0066
## disp 0.0662
## hp   0.0001
## drat 0.6607
## wt   0.0242
## qsec 0.0061
## vs   0.0090
## am   0.2838
## gear 0.0857
## carb 0.0000
##
## types:
##      mpg       cyl      disp      hp        drat      wt        qsec
## mpg  "none"    "mcor"   "pearson" "pearson" "pearson" "pearson" "pearson"
## cyl  "mcor"    "none"   "mcor"    "mcor"    "mcor"    "mcor"    "mcor"
## disp "pearson" "mcor"   "none"    "pearson" "pearson" "pearson" "pearson"
```

```
## hp   "pearson" "mcor"   "pearson" "none"    "pearson" "pearson" "pearson"
## drat "pearson" "mcor"   "pearson" "pearson" "none"    "pearson" "pearson"
## wt   "pearson" "mcor"   "pearson" "pearson" "pearson" "none"    "pearson"
## qsec "pearson" "mcor"   "pearson" "pearson" "pearson" "pearson" "none"
## vs   "mcor"    "cramer" "mcor"    "mcor"    "mcor"    "mcor"    "mcor"
## am   "mcor"    "cramer" "mcor"    "mcor"    "mcor"    "mcor"    "mcor"
## gear "mcor"    "cramer" "mcor"    "mcor"    "mcor"    "mcor"    "mcor"
## carb "mcor"    "cramer" "mcor"    "mcor"    "mcor"    "mcor"    "mcor"
##      vs        am       gear      carb
## mpg  "mcor"    "mcor"   "mcor"    "mcor"
## cyl  "cramer"  "cramer" "cramer"  "cramer"
## disp "mcor"    "mcor"   "mcor"    "mcor"
## hp   "mcor"    "mcor"   "mcor"    "mcor"
## drat "mcor"    "mcor"   "mcor"    "mcor"
## wt   "mcor"    "mcor"   "mcor"    "mcor"
## qsec "mcor"    "mcor"   "mcor"    "mcor"
## vs   "none"    "cramer" "cramer"  "cramer"
## am   "cramer"  "none"   "cramer"  "cramer"
## gear "cramer"  "cramer" "none"    "cramer"
## carb "cramer"  "cramer" "cramer"  "none"
```

Chapter 10

Simple Linear Regression

Example 10.1. A timber harvesting company "Timber Lend" needs to measure the volume of trees they cut. While they could measure the volume using physics principles, this is time consuming and they want to speed up the process. They have collected data of 31 trees, which includes:

1. `volume` measured manually by a special group of tree surgeon,
2. `height` of the tree, measured from the bottom to the top of the cut trunk,
3. `diameter` of the trunk, measured at the bottom.

They want to improve their timber harvesting process by speeding up the volume measurement. How can they do that based on the available data?

The data in this example is available online from here and can be loaded in R the following way:

```
load(url("https://github.com/config-i1/sba/raw/refs/heads/master/data/SBA_Chapter_10_Trees.Rdata"
```

To answer the question of the company "Timber Lend", we need to understand how we can capture relations between different variables numerically, so that we would be able to say what volume the company can expect based on the height and/or diameter of each trunk. Yes, we already know how to do graphical and correlations analysis. But this will not provide us sufficient information to answer the question in the beginning of this chapter. Still, the first thing to do is to plot the relation between the variables. We start with analysing the relation between height and volume, which is plotted in Figure 10.1.

We can see that there is a relation between the height and volume (Figure 10.1), which is mildly linear: with the increase of the height, volume of trees tends to increase on average. In addition to that, we can spot that there is a higher variability in the volume of trees with larger height in comparison with the ones with the lower one. This effect is called "heteroscedasticity" and we will come back to it in Section 15. But the important question for us now is whether

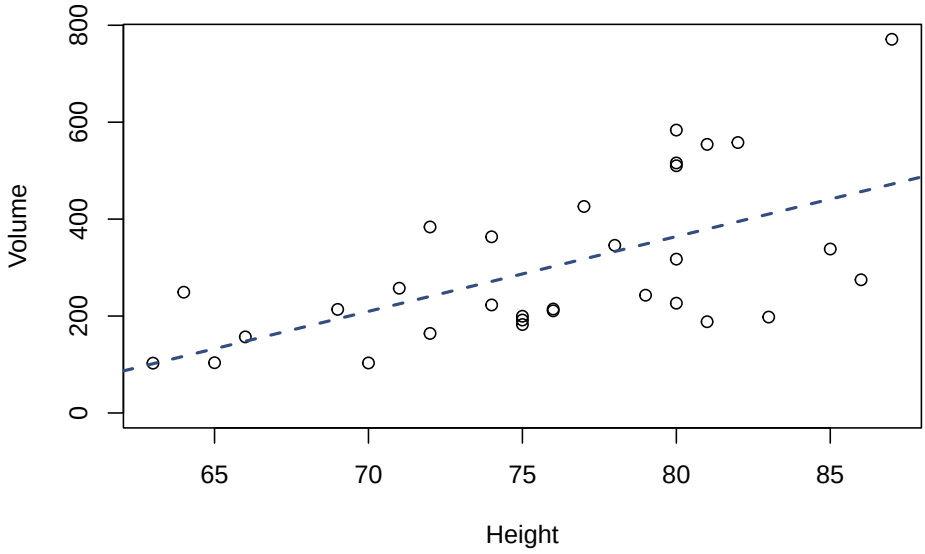


Figure 10.1: Scatterplot matrix of the trees volume and height.

we can quantify this relation between the variables, so that we could say, for example, that a tree that has a height of 75 is expected to have some specific volume?

To answer this question, we need to mathematically describe this relation. This can be done by finding coefficients of the line going through the cloud of points in Figure 10.1. The general mathematical form of this line (called "regression line") is:

$$\hat{y}_j = \beta_0 + \beta_1 x_j, \tag{10.1}$$

where $\hat{y}_j$ is the expected value of the response variable (expected volume in the example above), β_0 is the intercept (constant term), showing where the line intersects the y-ays, and β_1 is the coefficient for the slope parameter, which regulates how fast the expected volume increases with the increase of height. If β_1 is negative, the line would go down, showing that with the increase of one variable, the other tends to decrease. β_1 is also the tangent of the angle ϕ between the line drawn through the cloud of points and the x-ays. The two parameters are visualised in Figure 10.2 for an example of some artificial data.

Based on this regression line, we could explain every observation in sample as:

$$y_j = \hat{y}_j + \epsilon_j = \beta_0 + \beta_1 x_j + \epsilon_j, \tag{10.2}$$

where ϵ_j is the deviation of each specific point from the line. This variable is also called the "error term" and can be shown visually as in Figure 10.3, where each error corresponds to the size of each vertical line. In that figure, we only showed three errors for observations 17, 18 and 31, which all have large heights

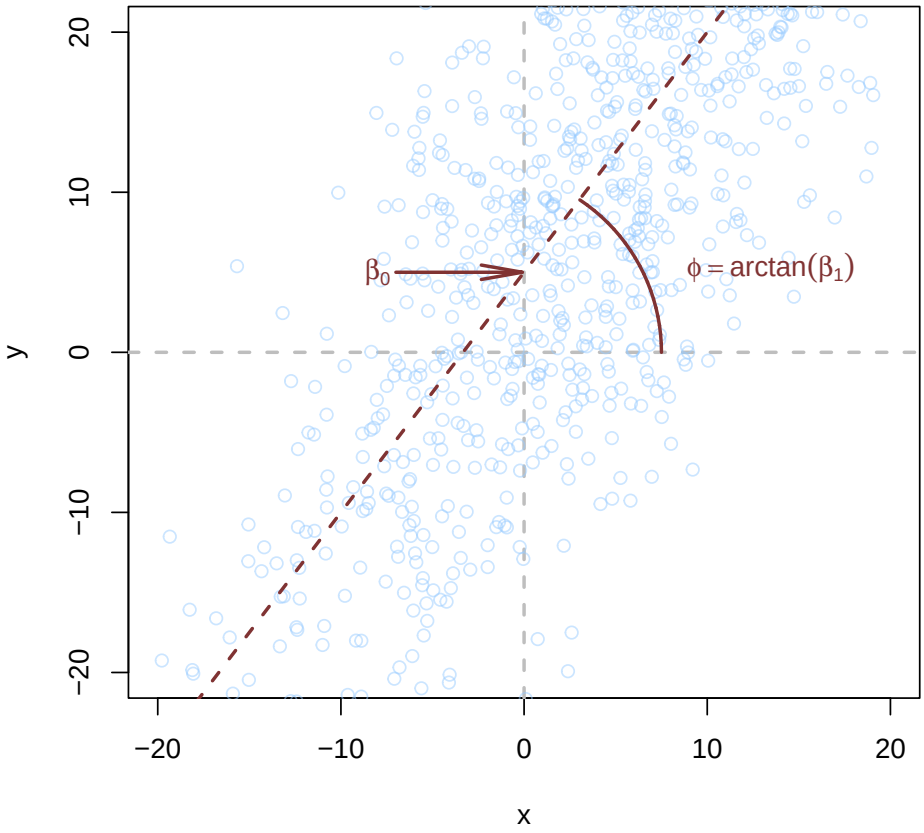


Figure 10.2: Visualisation of regression line drawn for some artificial data.

of 85, 86 and 87 respectively. But we could calculate such errors for all the other points in Figure 10.3.

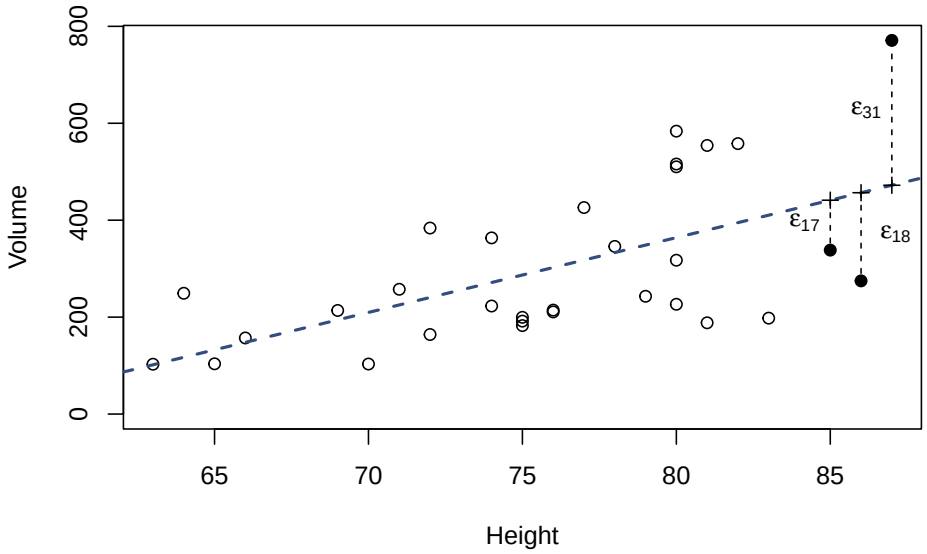


Figure 10.3: Scatterplot diagram between height and volume, together with an error term

The mathematical formula (10.2) is called "simple regression model", and is one of the basic *statistical model* (discussed in Subsection 1.1.1) that captures the relation between an explanatory variable x_j and the response variable y_j and explains what composes the response variable. In our example, the volume is impacted by the height and some individual errors that happen due to randomness.

Remark. The line in Figure 10.3 captures the averaged-our relation between the height and the volume. We might find some specific points, where the increase of height would not increase volume (e.g. switch from the observation 17 to 18 at the right-hand side of the image), but this can be considered as a random fluctuation. But overall, the average tendency is described by the increasing line.

Now the question is how to capture and quantify this relation correctly, so that we could help the "Timber Lend" company with its problem. One of the simplest techniques for this is called "Ordinary Least Squares".

10.1 Ordinary Least Squares (OLS)

For obvious reasons, we do not have the values of parameters from the population: it would be simply impossible to measure heights, diameters and volumes of all eysting trees in the world. This means that we will never know what the

true intercept and slope are. But we can get some estimates of these parameters based on the sample of data we have. There are different ways of doing that, and the most popular one is called "Ordinary Least Squares" method. This is the method that was used in the estimation of the model in Figure 10.3. So, how does it work?

Having the sample of data, we can draw a line through the cloud of points and then change the parameters for the intercept and slope until we are satisfied with how the line looks like. This would not be a reliable approach, but what we would be doing in this case is probably just making sure that the line goes somehow in the middle of data. To make this more rigorous, we could use the following simple method:

Remark. This is not how OLS works, but this gives an idea what it implies.

1. Sort all values in ascending order;
2. Split the sample in two halves based on the middle of the explanatory variable (in our case that would be `height=76`);
3. Calculate mean height and volume in the first half of the data;
4. Calculate mean height and volume of the second half;
5. Draw the line through the points on the plane.

The resulting line is shown in Figure 10.4

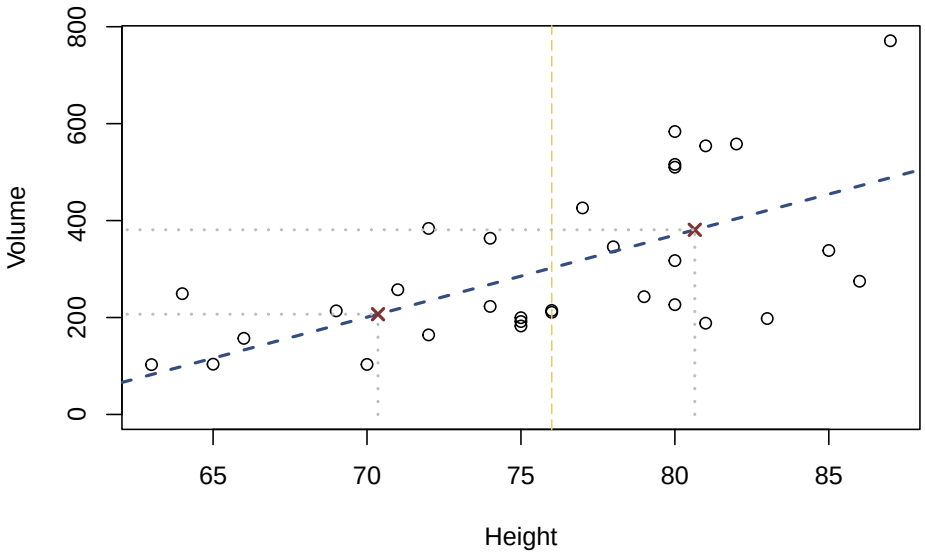


Figure 10.4: Scatterplot diagram between height and volume and the line drawn through two middle points of the data.

The line in the figure represents the average change of volume with the increase of height of trunks. The red crosses show the middle points, the vertical line in the middle shows where the sample is split into two halves. We could improve

this method by splitting each half into two halves again, and calculating points for them, or even further splitting each resulting half in halves etc. This method might not be reasonable for the specific sample, but if we had the population data, eventually we would be able to get a collection of points, each one of them representing the mean volume of trees given specific height.

But there is an easier way to do something similar but more practical. We could draw an arbitrary line, picking some estimates of parameters b_0 and b_1.

Remark. We never know the true values of β_0 and β_1, which is why when we estimate a model, we should substitute them with b_0 and b_1. This way we show that we deal with just some estimates of the true parameters.

After that we can calculate errors for each of observations, as we did before, but this time, because we do not know the true line, and we are only trying to get the best possible estimates of parameters, we should denote each error as e_j instead of ϵ_j, which in general they can be calculated as $e_j = y_j - \hat{y}_j$, where $\hat{y}_j$ is the value of the regression line (aka "fitted" value) for each specific value of explanatory variable.

For example, for the height of tree of 64 meters, the actual volume is 249.2, while the fitted value would be 117.052. The resulting error (or residual of model) is 249.2 - 117.052 = 132.148. We could collect all these errors of the model for all available trees based on their heights and this would result in a vector of positive and negative values like this:

```
##           1           2           3           4           5           6
## -106.477610  -28.789597    1.185608  -76.452815 -191.191239 -212.466444
##           7           8           9          10          11          12
##    9.072800 -104.365623 -137.553636  -87.265623 -105.616034  -91.603226
##          13          14          15          16          17          18
##  -87.803226   19.459993  -95.165623  -48.528021 -102.941649 -181.879252
##          19          20          21          22          23          24
##   32.284787  132.148006   12.721569  -46.653636   92.071979  143.247185
##          25          26          27          28          29          30
##  108.359172  174.708761  163.071158  219.646364  151.846364  146.046364
##          31
##  298.883145
```

These residuals are obtained from the following mathematical formula, given some values of b_0 and b_1:

$$e_j = y_j - b_0 - b_1 x_j. \tag{10.3}$$

If we needed to estimate parameters b_0 and b_1 of the model, we would want to minimise those distances by changing the parameters of the model. This would correspond to drawing a line going through the middle of the series, in a way connecting all the possible mean points in the data. Visually this is shown in Figure 10.5, where the line somehow goes through the data, and we calculate errors from it.

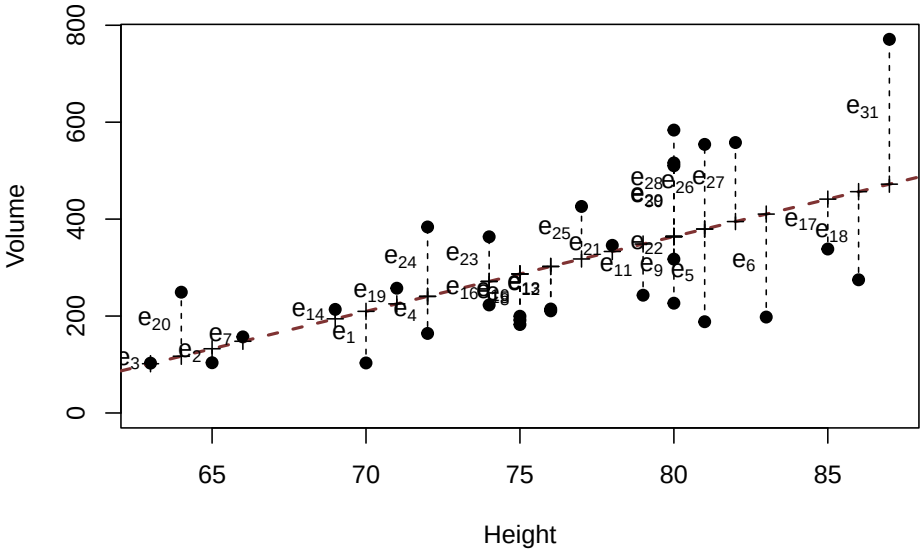


Figure 10.5: Scatterplot diagram between height and volume and the OLS line.

The problem is that some errors are positive, while the others are negative (see the middle image in Figure 10.6). If we just sum them up, they will cancel each other out, and we would loose the information about the distance. The simplest way to get rid of sign and keep the distance is by taking squares of each error, as shown in the bottom image in Figure 10.6.

If we then sum up all the squared residuals, we will end up with something called "Sum of Squared Errors":

$$\text{SSE} = \sum_{j=1}^{n} e_j^2. \tag{10.4}$$

If we now minimise SSE by changing values of parameters b_0 and b_1, we will find the parameters that would guarantee that the line goes through the cloud of points. Luckily, we do not need to use any fancy optimisers for this, as there is an analytical solution to this:

$$\begin{aligned} b_1 =& \frac{\text{cov}(x, y)}{\text{V}(x)}, \\ b_0 =& \bar{y} - b_1 \bar{x} \end{aligned} \tag{10.5}$$

where $\bar{x}$ is the mean of the explanatory variable x_j (height in our example) and $\bar{y}$ is the mean of the response variables y_j (volume).

Proof. In order to get (10.5), we should first insert (10.3) in (10.4) to get:

$$\text{SSE} = \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2.$$

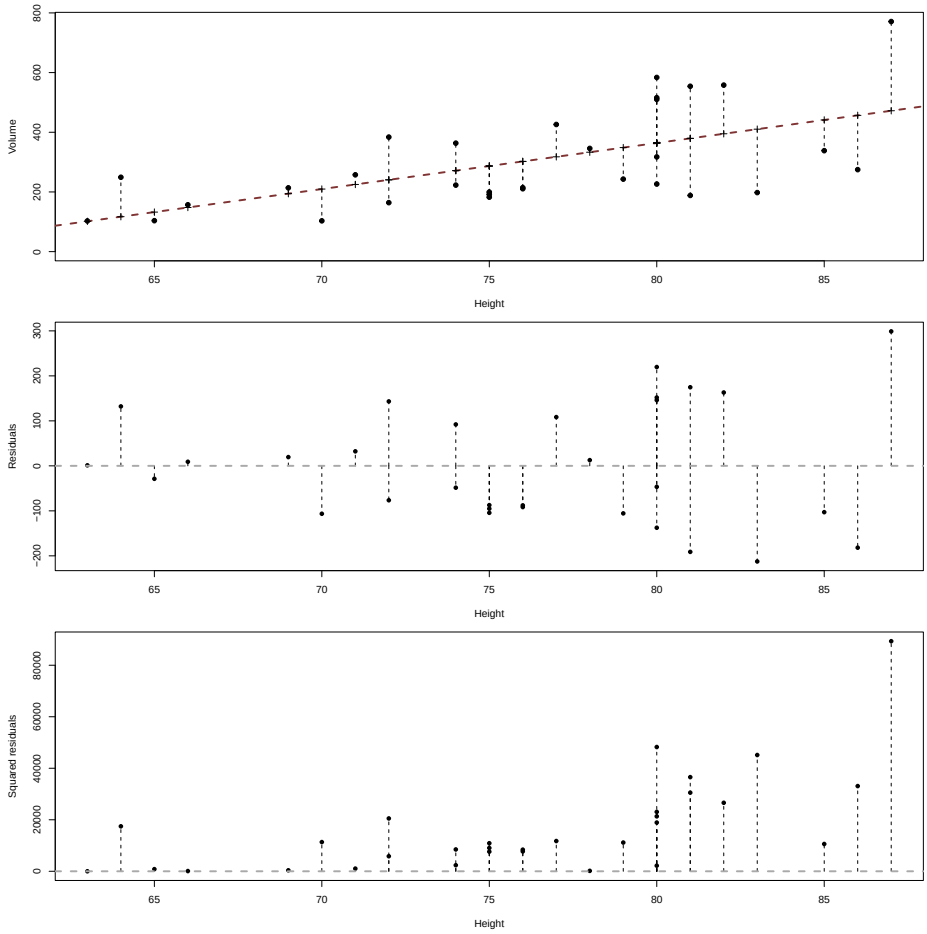


Figure 10.6: Volume, residuals and their squared values plotted againts the height of trees.

This can be expanded to:

$$\text{SSE} = \sum_{j=1}^{n} y_j^2 - 2b_0 \sum_{j=1}^{n} y_j - 2b_1 \sum_{j=1}^{n} y_j x_j + nb_0^2 + 2b_0 b_1 \sum_{j=1}^{n} x_j + b_1^2 \sum_{j=1}^{n} x_j^2$$

Given that we need to find the values of parameters b_0 and b_1 minimising SSE, we can take a derivative of SSE with respect to b_0 and b_1, equating them to zero to get the following **System of Normal Equations**:

$$\begin{aligned} \frac{d\text{SSE}}{db_0} &= -2\sum_{j=1}^{n} y_j + 2nb_0 + 2b_1 \sum_{j=1}^{n} x_j = 0 \\ \frac{d\text{SSE}}{db_1} &= -2\sum_{j=1}^{n} y_j x_j + 2b_0 \sum_{j=1}^{n} x_j + 2b_1 \sum_{j=1}^{n} x_j^2 = 0 \end{aligned}$$

Solving this system of equations for b_0 and b_1 we get:

$$\begin{aligned} b_0 &= \frac{1}{n}\sum_{j=1}^{n} y_j - b_1 \frac{1}{n}\sum_{j=1}^{n} x_j \\ b_1 &= \frac{n\sum_{j=1}^{n} y_j x_j - \sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j}{n\sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2} \end{aligned} \tag{10.6}$$

In the system of equations (10.6), we have the following elements:

1. $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$,
2. $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$,
3. $\text{cov}(x, y) = \frac{1}{n}\sum_{j=1}^{n} y_j x_j - \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j$,
4. $\text{V}(x) = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2$,

which after inserting in (10.6) lead to (10.5). □

Remark. If for some reason $b_1 = 0$ in (10.5) (for example, because the covariance between x and y is zero, implying that they are not correlated), then the intercept $b_0 = \bar{y}$, meaning that the global average of the data would be the best predictor of the variable y_j.

This method of estimation of parameters based on the minimisation of SSE, is called "Ordinary Least Squares", because by using this method we get the least possible squares of errors for the data. The word "ordinary" means that this is one of the basic estimation techniques. There are other least squares techniques, which we are not yet discussing in this book. The method is simple and does not require any specific assumptions: we just minimise the overall distance between the line and the points by changing the values of parameters.

Example 10.2. For the problem with trees, we can use the `lm()` function from the `stats` package in R to get the OLS estimates of parameters. This is done in the following way:

```
slmTrees <- lm(volume~height, SBA_Chapter_10_Trees)
slmTrees
```

```
##
## Call:
## lm(formula = volume ~ height, data = SBA_Chapter_10_Trees)
##
## Coefficients:
## (Intercept)       height
##     -870.95        15.44
```

The syntax of the function implies that we use volume as the response variable and height as the explanatory one. The resulting model in our notations has b_0 = -870.95 and b_1 = 15.44, its equation can be written as: $volume_j$ = -870.95 + 15.44 $height_j + e_j$.

While we can make some conclusions based on the simple linear regression, we know that in real life we rarely see bivariate relations - typically a variable is influenced by a set of variables, not just by one. This implies that the correct model would typically include many explanatory variables. This is why we only discuss the simple linear regression for educational purpose and generally, do not recommend to use it in real life situations.

10.2 Covariance, correlation and SLR

Now that we have introduced a simple linear regression, we can take a step back to better understand some statistics related to it, which we discussed in previous chapters.

10.2.1 Covariance

Covariance is one of the most complicated things to explain to general audience. We will need to use a bit of mathematics that we introduced in Section 10.1, specifically the formula (10.5), where b_1 is calculated as:

$$b_1 = \frac{\text{cov}(x,y)}{\hat{\sigma}_x^2},$$

where $\hat{\sigma}_x$ is the in-sample estimate of standard deviation. Using a simple manipulation, we can express $\text{cov}(x,y)$ as:

$$\text{cov}(x,y) = b_1 \hat{\sigma}_x \hat{\sigma}_x.$$

Visually, this can be represented as the areas of a rectangular shown on right pane in Figure 10.7.

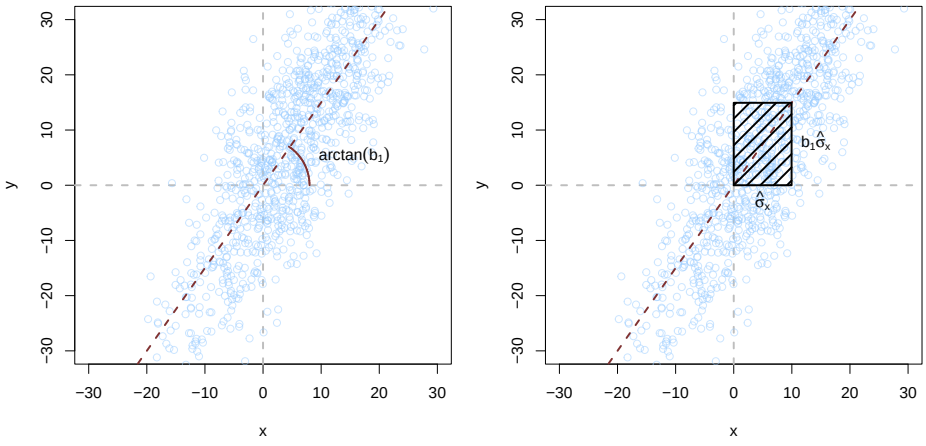


Figure 10.7: Visualisation of covariance between two random variables, x and y.

In Figure 10.7, the data is centred around the means of x and y. We draw a regression line with the angle $\arctan(b_1)$ through the cloud of points (the left-hand side image). After that we draw a segment of the length $\hat{\sigma}_x$ parallel to the x-ays (righ-hand side image). The multiplication of b_1 by $\hat{\sigma}_x$ gives the side denoted as $b_1\hat{\sigma}_x$ (because b_1 equals to tangent of the angle of the line to the x-ays). And finally, the $b_1\hat{\sigma}_x \times \hat{\sigma}_x = \text{cov}(x, y)$ is the area of the rectangular in the right-hand side image of Figure 10.7. The higher the standard deviation of x is, the bigger the area will be, implying that the covariance becomes larger. At the same time, for the same values of $\hat{\sigma}_x$, the higher b_1 is, the larger the area becomes, increasing the covariance as well. In this interpretation the covariance becomes equal to zero in one of the two cases:

$$b_1 = 0$$
$$\hat{\sigma}_x = 0,$$

which implies that either the angle of the regression line is zero, i.e. there is no linear relation between x and y, or there is no variability in the variable x.

Similar visualisations can be done if the axes are swapped and the regression $x = a_0 + a_1 y + u$ is constructed. The logic would be similar, only changing the value of the slope parameter b_1 by a_1 and substituting $\hat{\sigma}_x$ with $\hat{\sigma}_y$. Finally, in case of negative relation between the variables, the rectangular area will be drawn below the zero line and thus could be considered as being negative (although the surface of the area itself cannot be negative). Still, the logic explained for the positive case above could be transferred on the negative case as well.

10.2.2 Correlation

Another thing to discuss is the connection between the parameter b_1 and the correlation coefficient. We have already briefly mentioned that in Section 9.3,

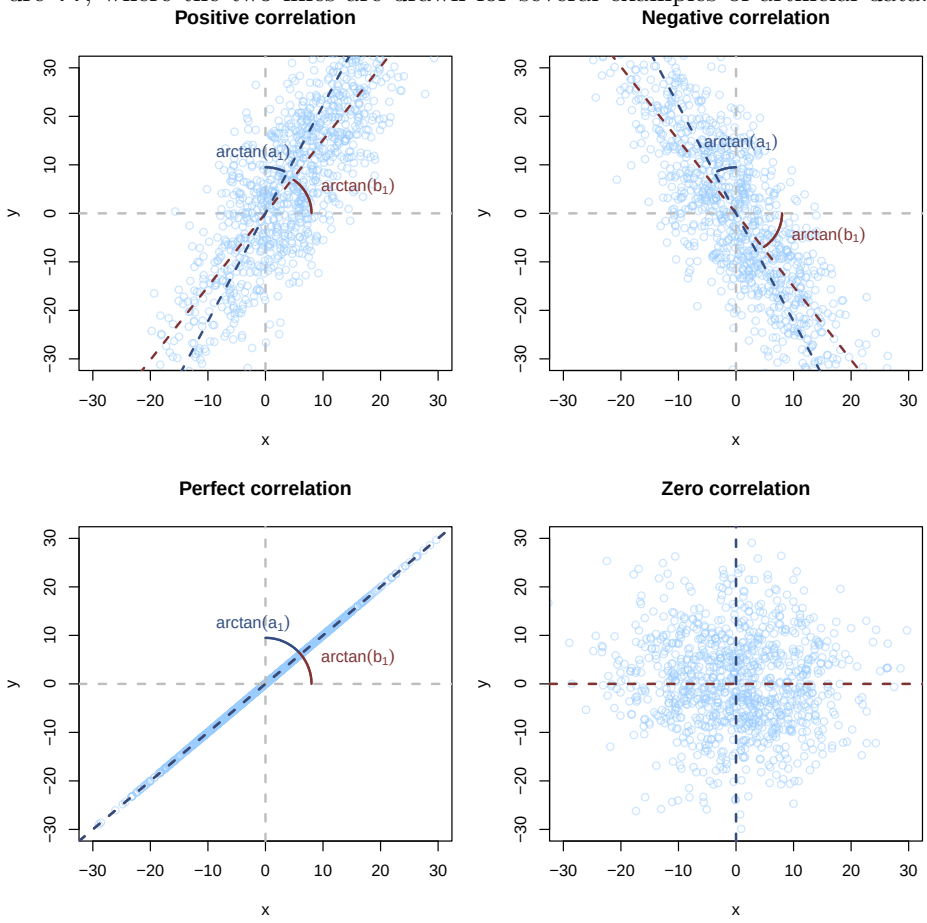
but here we can now spend more time on it. First, we could estimate two models given the pair of variable x and y:

1. Model (**??**) $y_j = b_0 + b_1 x_j + e_j$;
2. The inverse model $x_j = a_0 + a_1 y_j + u_j$.

We could then extract the slope parameters of the two models via (10.5) and get the value of correlation coefficient as a geometric mean of the two:

$$r_{x,y} = \text{sign}(b_1)\sqrt{b_1 a_1} = \text{sign}(\text{cov}(x,y))\sqrt{\frac{\text{cov}(x,y)}{\text{V}(x)}\frac{\text{cov}(x,y)}{\text{V}(y)}} = \frac{\text{cov}(x,y)}{\sqrt{V(x)V(y)}}, \tag{10.7}$$

which is the formula (9.10). This is how the correlation coefficient was originally derived by Karl Pearson (**?**). Visually this is shown in Figure **??**, where the two lines are drawn for several examples of artificial data.



If the lines have positive slopes (as shown in the left top pane in Figure **??**) then the resulting coefficient of correlation will be positive. If they are both negative,

the correlation will be negative as well (right top pane in Figure **??**). If the lines coincide then the product of tangents of their angles will be equal to 1 (thus we would have a perfect correlation of 1, left bottom pane in Figure **??**). Finally, in the case, when there is no linear relation between variables, the lines will coincide with the x- and y- axes respectively, producing $a_1 = b_1 = 0$ and thus leading to the zero correlation coefficient (right bottom pane in Figure **??**).

Finally, another way to look at the correlation is to consider the visualisation of covariance from Figure 10.7 and to expand it to the correlation coefficient, for which the part $\hat{\sigma}_x \times \hat{\sigma}_y$, corresponds to the denominator in the formula (10.7), and is shown in Figure 10.8 as a red area.

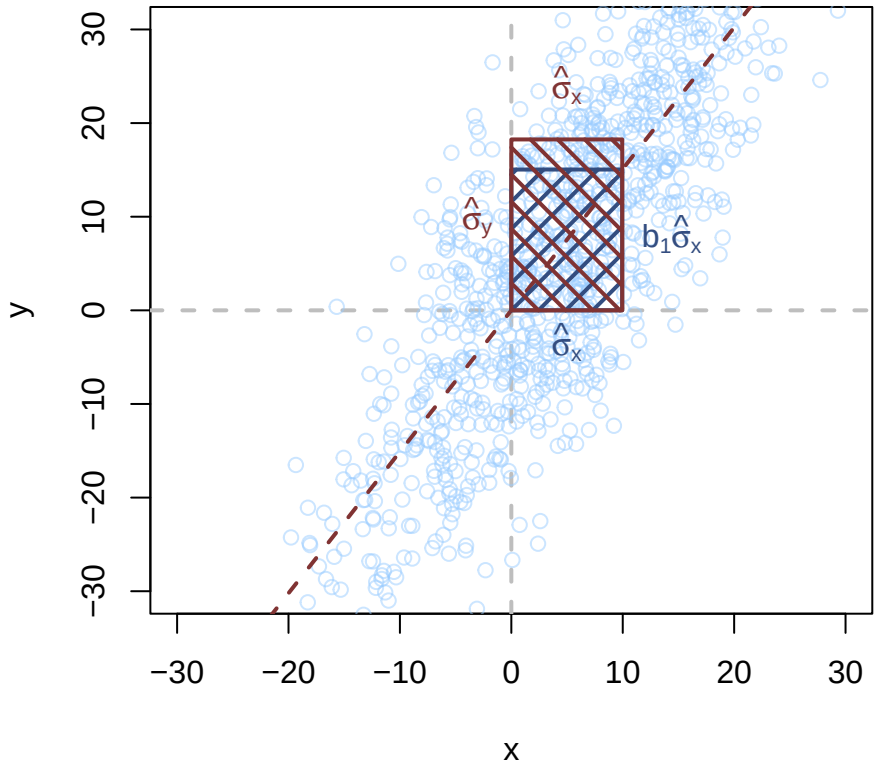


Figure 10.8: Visualisation of correlation between two random variables, x and y.

The correlation in Figure 10.8, corresponds to the ratio of the two areas: blue one (covariance) to the red one (the product of standard deviations). If the areas coincide, the correlation is equal to one. This would only happen if all the observations lie on the straight line, the case for which $\hat{\sigma}_y = b_1\hat{\sigma}_x$. Mathematically, this can be seen if we take the variance of the response variable conditional on the slope parameter:

$$\mathrm{V}(y|b_1) = b_1^2\mathrm{V}(x) + V(e), \tag{10.8}$$

which leads to the following equality for the conditional standard deviation of y:

$$\hat{\sigma}_y = \sqrt{b_1^2 \hat{\sigma}_x^2 + \hat{\sigma}_e^2}, \tag{10.9}$$

where $\hat{\sigma}_e$ is the standard deviation of the residuals. In this case, it becomes clear that the correlation is impacted by the variance of the error term $\hat{\sigma}_e^2$. If it is equal to zero, we get the equality: $\hat{\sigma}_y = b_1 \hat{\sigma}_x$, for which the areas in Figure 10.8 will coincide and correlation becomes equal to one. The bigger the variance of residuals is, the lower the correlation coefficient becomes. Note that the value of b_1 does not impact the strength of correlation, it only regulates, whether the correlation is positive, negative or zero. Several correlation coefficients and respective rectangular areas are shown in Figure 10.9.

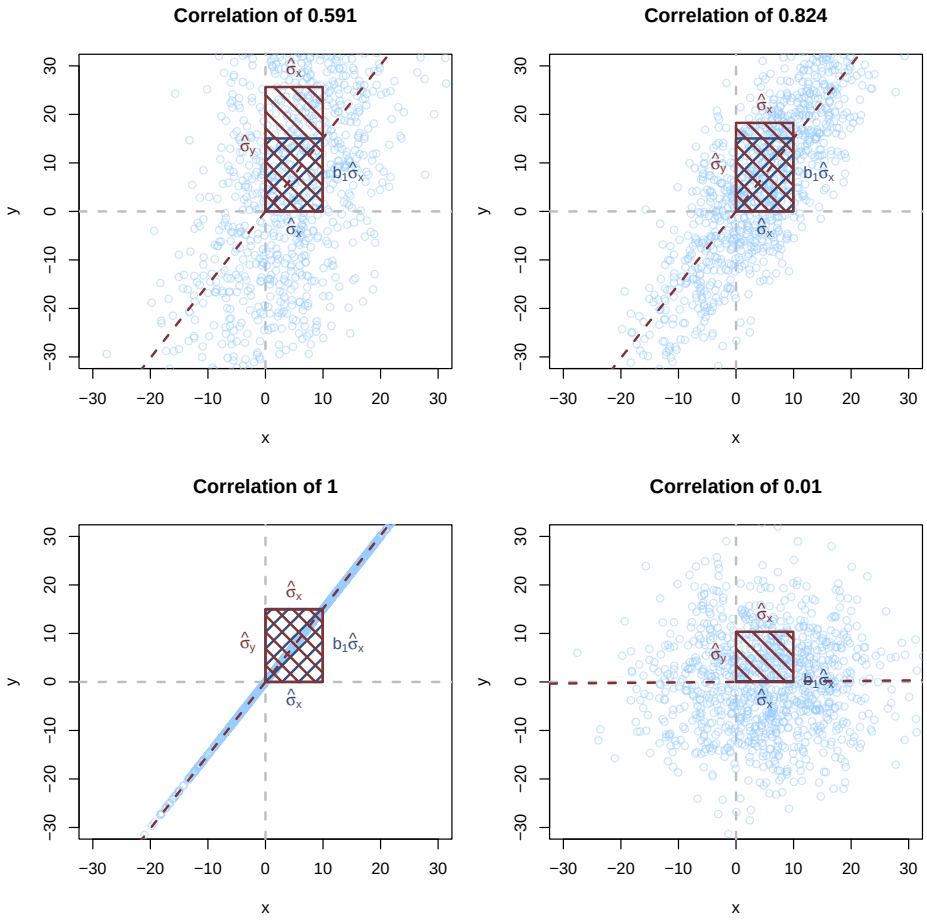


Figure 10.9: Visualisation of several correlation coefficients.

10.3 Residuals of model estimated via OLS

OLS applied to any model guarantees two important properties about its residuals:

1. $\mathrm{E}(e_j) = \frac{1}{n}\sum_{j=1}^{n} e_j = 0$,
2. $\mathrm{E}(e_j x_{i,j}) = \frac{1}{n}\sum_{j=1}^{n} e_j x_{i,j} = 0$ for any i.

The first property means that the in-sample mean of residuals is always equal to zero, while the second implies that the estimation is done in a way that the in-sample correlation between the residuals and any explanatory variable in the model is equal to zero. These two conditions happen automatically, and there is no point in testing them or trying to see whether they have been violated or not. On the other hand, if a model was estimated using some other method, these properties might not hold anymore, and it might be the case that the mean of the in-sample residuals and/or the correlation between the error and explanatory variables would not be equal to zero. We will come back to these properties in Chapter 15, when we discuss the standard assumptions of statistical models.

It is possible to prove mathematically that these two conditions hold. Here is a proof of the first one:

Proof. Consider the sum of residuals of a simple linear regression model estimated using OLS:

$$\sum_{j=1}^{n} e_j = \sum_{j=1}^{n}(y_j - b_0 - b_1 x_j) = \sum_{j=1}^{n} y_j - n b_0 - b_1 \sum_{j=1}^{n} x_j \tag{10.10}$$

Inserting the formula for b_0 from (10.6) in (10.10) we get:

$$\sum_{j=1}^{n} e_j = \sum_{j=1}^{n} y_j - n\frac{1}{n}\sum_{j=1}^{n} y_j + n b_1 \frac{1}{n}\sum_{j=1}^{n} x_j - b_1 \sum_{j=1}^{n} x_j \tag{10.11}$$

which after some cancelations leads to:

$$\sum_{j=1}^{n} e_j = \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} y_j + b_1 \sum_{j=1}^{n} x_j - b_1 \sum_{j=1}^{n} x_j = 0 \tag{10.12}$$

Given that the sum of errors is equal to zero, their mean will be equal to zero as well. □

The second property is less straightforward, but it can be proven as well, using similar logic:

Proof. For the same simple linear regression, estimated using OLS, consider:

$$\sum_{j=1}^{n} e_j x_j = \sum_{j=1}^{n}(y_j x_j - b_0 x_j - b_1 x_j^2) = \sum_{j=1}^{n} y_j x_j - b_0 \sum_{j=1}^{n} x_j - b_1 \sum_{j=1}^{n} x_j^2. \tag{10.13}$$

Inserting the formula for b_0 from (10.6) in (10.13) leads to:

$$\begin{aligned}\sum_{j=1}^{n} e_j x_j = &\sum_{j=1}^{n} y_j x_j - \frac{1}{n}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j + b_1 \frac{1}{n}\sum_{j=1}^{n} x_j \sum_{j=1}^{n} x_j - b_1 \sum_{j=1}^{n} x_j^2 = \\ &\sum_{j=1}^{n} y_j x_j - \frac{1}{n}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j + b_1 \left(\frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2 - \sum_{j=1}^{n} x_j^2 \right).\end{aligned} \tag{10.14}$$

Now we insert the formula for b_1 from (10.6) in (10.14) to get:

$$\begin{aligned}\sum_{j=1}^{n} e_j x_j = &\sum_{j=1}^{n} y_j x_j - \frac{1}{n}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j + \\ &\frac{n\sum_{j=1}^{n} y_j x_j - \sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j}{n\sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2} \left(\frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2 - \sum_{j=1}^{n} x_j^2 \right).\end{aligned} \tag{10.15}$$

The ratio in the right-hand side of (10.15) can be regrouped and rewritten as:

$$-\frac{n\sum_{j=1}^{n} y_j x_j - \sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j}{n\left(\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right)} \left(\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2 \right) = -\sum_{j=1}^{n} y_j x_j - \sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j, \tag{10.16}$$

which after inserting it back in (10.15) leads to:

$$\sum_{j=1}^{n} e_j x_j = \sum_{j=1}^{n} y_j x_j - \frac{1}{n}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} y_j x_j + \frac{1}{n}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j = 0 \tag{10.17}$$

Given that the sum (10.17) is equal to zero, the mean of $e_j x_j$ will be equal to zero as well. □

In order to see that the second property implies that the correlation between the residuals and regressors is equal to zero, we need to take a step back and consider the covariance between e_j and x_j (because it is used in correlation coefficient as discussed in Section 9.3):

$$\text{cov}(e_j, x_j) = \sum_{j=1}^{n} (e_j - \bar{e})(x_j - \bar{x}) \tag{10.18}$$

The first thing to notice in (10.18) is that $\bar{e} = 0$ because of the property (1) discussed in the beginning of this section. This simplifies the formula and leads to:

$$\text{cov}(e_j, x_j) = \sum_{j=1}^{n} e_j (x_j - \bar{x}) = \sum_{j=1}^{n} e_j x_j - \bar{x}\sum_{j=1}^{n} e_j = \sum_{j=1}^{n} e_j x_j, \tag{10.19}$$

because the second sum in (10.19) is equal to zero due to the same property (1).

These two basic properties on one hand are useful for further derivations and on the other one show what to expect from the residuals of a regression model estimated via the OLS. The latter means, for example, that there is no point in testing whether the two properties hold, they will be satisfied automatically in case of OLS.

10.4 Quality of a fit

The term "Quality of a fit" is used often in statistics to outline approaches that provide some information about how the applied models fit the data. We find it misleading, because the word "quality" is not appropriate here. The measures discussed in this section only show how well the actual values are approximated by the model, but their values do not tell us whether a model is good or bad. Still, to get a general impression about the performance of the estimated model, we can calculate several in-sample measures, which could provide us insights about the approximating properties of the model.

10.4.1 Sums of squares

The fundamental measure that lies in the basis of many other ones is SSE, which is the value of the OLS criterion (10.4). It cannot be interpreted on its own and cannot be used for model comparison, but it shows the overall variability of the data around the regression line. In a more general case, it is written as:

$$\text{SSE} = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2. \tag{10.20}$$

This sum of squares is related to another two, the first being the Sum of Squares Total:

$$\text{SST} = \sum_{j=1}^{n} (y_j - \bar{y})^2, \tag{10.21}$$

where $\bar{y}$ is the in-sample mean. If we divide the value (10.21) by $n-1$, we get the in-sample variance (introduced in Section 5.1):

$$\text{V}(y) = \frac{\text{SST}}{n-1} = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2.$$

The last sum of squares is Sum of Squares of Regression:

$$\text{SSR} = \sum_{j=1}^{n} (\bar{y} - \hat{y}_j)^2, \tag{10.22}$$

which shows the variability of the regression line. It is possible to show that in *the linear regression* (**this is important**, this property might be violated

in other models), the three sums are related to each other via the following equation:

$$\text{SST} = \text{SSE} + \text{SSR}. \tag{10.23}$$

Proof. This involves manipulations, some of which are not straightforward. First, we assume that SST equals to SSE + SSR, and see whether we reach the original formula of SST:

$$\begin{aligned}
\text{SST} = \text{SSR} + \text{SSE} &= \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \\
&= \sum_{j=1}^{n}\left(\hat{y}_j^2 - 2\hat{y}_j\bar{y} + \bar{y}^2\right) + \sum_{j=1}^{n}\left(y_j^2 - 2y_j\hat{y}_j + \hat{y}_j^2\right) \\
&= \sum_{j=1}^{n}\left(\hat{y}_j^2 - 2\hat{y}_j\bar{y} + \bar{y}^2 + y_j^2 - 2y_j\hat{y}_j + \hat{y}_j^2\right) \\
&= \sum_{j=1}^{n}\left(\bar{y}^2 - 2\bar{y}y_j + y_j^2 + 2\bar{y}y_j + \hat{y}_j^2 - 2\hat{y}_j\bar{y} - 2y_j\hat{y}_j + \hat{y}_j^2\right) \\
&= \sum_{j=1}^{n}\left((\bar{y} - y_j)^2 + 2\bar{y}y_j + 2\hat{y}_j^2 - 2\hat{y}_j\bar{y} - 2y_j\hat{y}_j\right)
\end{aligned} \tag{10.24}$$

We can then substitute $y_j = \hat{y}_j + e_j$ in the right hand side of (10.24) to get:

$$\begin{aligned}
\text{SST} &= \sum_{j=1}^{n}\left((\bar{y} - y_j)^2 + 2\bar{y}(\hat{y}_j + e_j) + 2\hat{y}_j^2 - 2\hat{y}_j\bar{y} - 2(\hat{y}_j + e_j)\hat{y}_j\right) \\
&= \sum_{j=1}^{n}\left((\bar{y} - y_j)^2 + 2\bar{y}\hat{y}_j + 2\bar{y}e_j + 2\hat{y}_j^2 - 2\hat{y}_j\bar{y} - 2\hat{y}_j\hat{y}_j - 2e_j\hat{y}_j\right) \\
&= \sum_{j=1}^{n}\left((\bar{y} - y_j)^2 + 2\bar{y}e_j + 2\hat{y}_j^2 - 2\hat{y}_j^2 - 2e_j\hat{y}_j\right) \\
&= \sum_{j=1}^{n}\left((\bar{y} - y_j)^2 + 2\bar{y}e_j - 2e_j\hat{y}_j\right)
\end{aligned} \tag{10.25}$$

Now if we split the sum into three elements, we will get:

$$\begin{aligned}
\text{SST} &= \sum_{j=1}^{n}(\bar{y} - y_j)^2 + 2\sum_{j=1}^{n}(\bar{y}e_j) - 2\sum_{j=1}^{n}(e_j\hat{y}_j) \\
&= \sum_{j=1}^{n}(\bar{y} - y_j)^2 + 2\bar{y}\sum_{j=1}^{n}e_j - 2\sum_{j=1}^{n}(e_j\hat{y}_j)
\end{aligned} \tag{10.26}$$

The second sum in (10.26) is equal to zero, because OLS guarantees that the in-sample mean of error term is equal to zero (see proof in Subsection 10.3). The

third one can be expanded to:

$$\sum_{j=1}^{n}(e_j\hat{y}_j) = \sum_{j=1}^{n}(e_j b_0 + b_1 e_j x_j). \tag{10.27}$$

We see the sum of errors in the first sum of (10.27), so the first elements is equal to zero again. The second term is equal to zero as well due to OLS estimation (this was also proven in Subsection 10.3). This means that:

$$\text{SST} = \sum_{j=1}^{n}(\bar{y} - y_j)^2, \tag{10.28}$$

which is the formula of SST (10.21). □

The relation between SSE, SSR and SST can be visualised and is shown in Figure 10.10. If we take any observation in that figure, we will see how the deviations from the regression line and from the mean are related.

10.4.2 Coefficient of determination, R^2

While the sums of squares do not have a nice interpretation and are hard to use for diagnostics, they can be used in calculating the measure called "Coefficient of Determination". It is calculated in the following way:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}. \tag{10.29}$$

Given the fundamental property (10.23), we can see that R^2 will always lie between zero and one. To better understand its meaning, imagine the following two extreme situations:

1. The model fits the data in the same way as the mean line (black line coincides with the grey line in Figure 10.10). In this case SSE would be equal to SST and SSR would be equal to zero (because $\hat{y}_j = \bar{y}$) and as a result the R^2 would be equal to zero.
2. The model fits the data perfectly, without any errors (all points lie on the black line in Figure 10.10). In this situation SSE would be equal to zero and SSR would be equal to SST, because the regression would go through all points (i.e. $\hat{y}_j = y_j$). This would make R^2 equal to one.

So, the zero value of the coefficient of determination means that the model does not explain the data at all and one means that it overfits the data. The value itself is usually interpreted as a percentage of variability in data explained by the model.

Remark. The properties above provide us an important point about the coefficient of determination: *it should not be equal to one, and it is alarming if it is very close to one.* This is because in this situation we are implying that there is no

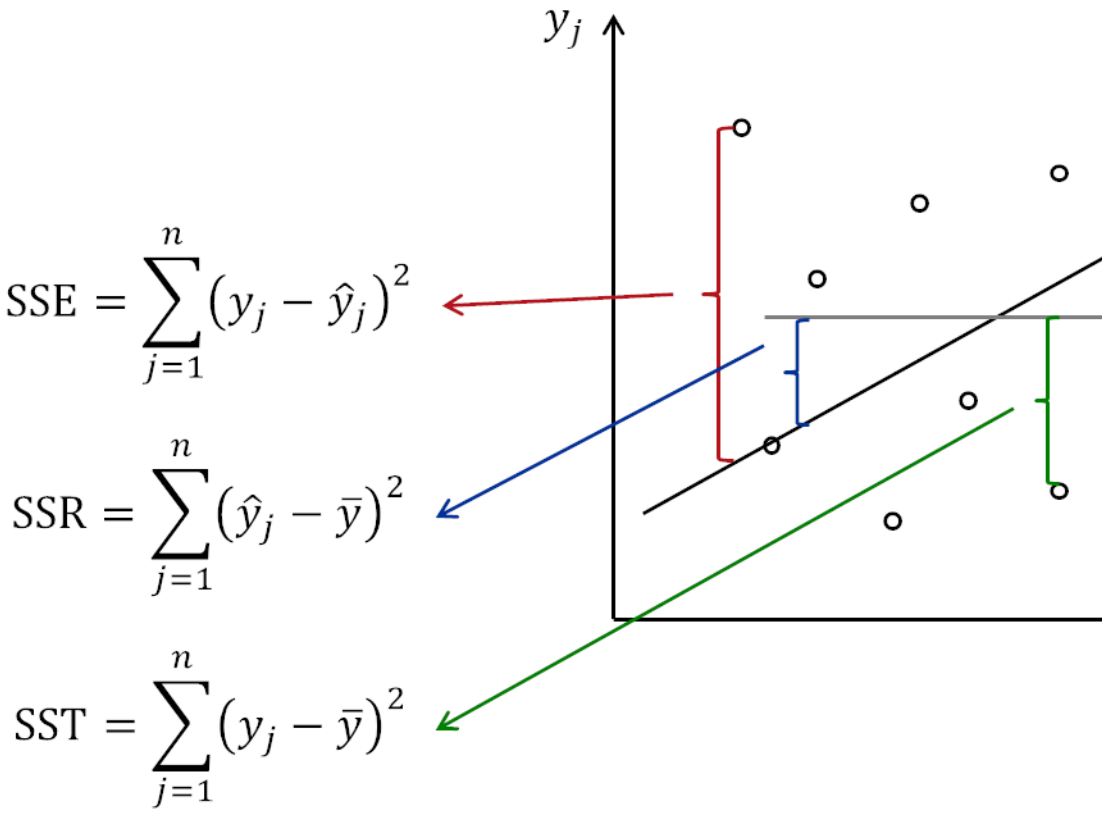


Figure 10.10: Relation between different sums of squares.

randomness in the data, which contradicts our definition of the statistical model (see Section 1.1.1). The adequate statistical model should always have some randomness in it. The situation of $R^2 = 1$ implies mathematically:

$$y_j = b_0 + b_1 x_j,$$

which means that all $e_j = 0$, being unrealistic and only possible if there is a functional relation between y and x (no need for statistical inference then). So, in practice we should not maximise R^2 and should be careful with models that have very high values of it. At the same time, too low values of R^2 are also alarming, as they tell us that the model becomes:

$$y_j = b_0 + e_j,$$

meaning that it is not different from the simple mean of the data, because in that case $b_0 = \bar{y}$. So, coefficient of determination in general is not a very good measure for assessing performance of a model. It can be used for further inferences, and for a basic indication of whether the model overfits (R^2 close to 1) or underfits (R^2 close to 0) the data. But no serious conclusions should be solely based on it.

Here how this measure can be calculated in R based on the model that we estimated in Section 10.1:

```
n <- nobs(slmTrees)
R2 <- 1 - sum(resid(slmTrees)^2) / (var(actuals(slmTrees))*(n-1))
R2
```

```
## [1] 0.357975
```

Note that in this formula we used the relation between SST and $V(y)$, multiplying the value by $n - 1$ to get rid of the denominator. The resulting value tells us that the model has explained 35.8% deviations in the data.

Finally, based on coefficient of determination, we can also calculate the coefficient of multiple correlation, which we have already discussed in Section 9.4:

$$R = \sqrt{R^2} = \sqrt{\frac{\text{SSR}}{\text{SST}}}. \tag{10.30}$$

It shows the closeness of relation between the response variable y_j and the explanatory variables to the linear one. The coefficient has a positive sign, no matter what the relation between the variables is. In case of the simple linear regression, it is equal to the correlation coefficient (from Section 9.3) with the sign equal to the sign of the coefficient of the slop b_1 (this was discussed in Subsection 10.2.2):

$$r_{x,y} = \text{sign}(b_1) R. \tag{10.31}$$

Here is a demonstration of the formula above in R:

```
sign(coef(slmTrees)[2]) * sqrt(R2)
```

```
##    height
## 0.5983101
```

```
cor(SBA_Chapter_10_Trees$height,SBA_Chapter_10_Trees$volume)
```

```
## [1] 0.5983101
```

10.5 What about the "Timber Lend" company?

Coming back to the example that motivated this chapter, there is a way we can improve the model for the company and help them in making better decisions. After all, the determination coefficient of the model of volume from height was just 0.358.

First, we check how the relation between the diameter and the volume looks (Figure 10.11)

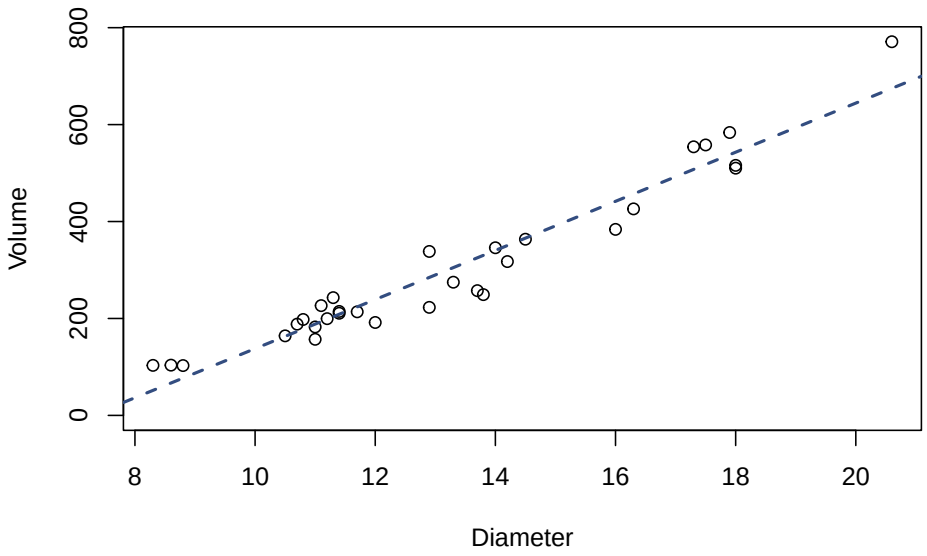


Figure 10.11: Scatterplot matrix of the trees volume and dimeter.

It might not be apparent for an inexperienced analyst, but the relation between the diameter and volume is non-linear, because the points for the lowest and highest diameters lie consistently above the straight line. Even if this relation is not apparent visually, there is a fundamental reason for its existence: trunks of trees have a shape close to cylinder. Some of you might remember from geometry that the volume of a cylinder is calculated as:

$$V = h\pi r^2, \tag{10.32}$$

where V is the volume, h is the height, r is the radius and π is the constant number. The diameters that we have in the data equal to $d = 2 \times r$. Having this fundamental formula, implies that the relation between the diameter and volume should be indeed non-linear. Based on that, we can create a new variable, which could be called `cylinder`:

```
SBA_Chapter_10_Trees$cylinder <- (SBA_Chapter_10_Trees$height *
                                    SBA_Chapter_10_Trees$diameter^2)
```

Furthermore, because trunks of trees are not exactly cylinders, our model can be represented as:

$$volume = \beta_0 + \beta_1 cylinder + \epsilon_j.$$

This model in R gives is much more reasonable than either the model of volume from height or volume from diameter. We can fit it and see how much variability it explains:

```
# Fit the model
slmTreesCyl <- lm(volume~cylinder, SBA_Chapter_10_Trees)
# Get the number of observations
n <- nobs(slmTreesCyl)
# Calculate R^2
1 - sum(resid(slmTreesCyl)^2) / (var(actuals(slmTreesCyl))*(n-1))
```

```
## [1] 0.9777898
```

Remark. While in general, we should not compare models based on R^2, in this specific case we can because the number of explanatory variables in the two models is exactly the same.

As we see the model of volume from cylinders explains the data much better than the previous one.

In the formula of the cylinder (10.32), we do not have any error term. Why do we expect it to be in the model that we fit? Shouldn't R^2 be equal to one in our example?

We do not expect to have zero error in this case, because the trunks of trees are not perfect cylinders: the diameter at the top of the tree is smaller than the diameter at the bottom, and trees have some curvature, deviating in shape from the perfect cylinder. Due to these factors, the formula (10.32) does not perfectly describe the volume of the tree, but instead is a good approximation of it.

As for the coefficients of the model, based on our sample, they were:

```
coef(slmTreesCyl)
```

```
## (Intercept)    cylinder
## -2.43730518  0.02124751
```

Using some specific measurements of a tree, we can get its expected volume. For example, if a tree has the height of 80 and diameter of 10.7, our new variable cylinder would be $80 \times 11.1^2 = 9856.8$. Inserting this value in the equation, we get the expected volume:

$$volume = -2.44 + 0.02 \times 9856.8 \approx 207.00,$$

which is not too far from the real volume of 226.5 that we have in the data (observation number 9). While this is not a perfect estimate of volume, it allows improving the operational process for the "Timber Lend" company, hopefully reducing some costs.

Chapter 11

Multiple Linear Regression

Example 11.1. One of the problems that construction companies face is getting a good estimate of the budget needed to build something. Many companies tend to underestimate the costs and time the project will take. To address this, a mid-size company called "Eden city", which specialises on construction of residential buildings, decided to take a more analytical approach to the problem. They have collected the data of their previous projects and needs help in building a model that would explain what forms the costs for different types of buildings. Their idea is to use this model during the business plan write-up phase to get an estimate of the future project, which they hope will be better than the ones they used before based on their pure judgment.

In this example, we are interested in the overall costs of construction (in thousands of pounds), which can be impacted by:

- The size of a building in squared meters,
- The cost of materials (in thousands of pounds),
- Type of building (detached, semi-detached, bungalow etc),
- How many projects the specific crew did before,
- Year when the project was started.

What else do you think can impact such costs in theory?

This data is available online:

```
load(url("https://github.com/config-i1/sba/raw/refs/heads/master/data/SBA_Chapter_11_Costs.Rdata"
```

Based on what we have discussed before, we can do analysis of measures of association and even build simple regression model (or several of them), but we acknowledge that in many real life situations, there are many factors that impact the variable of interest. In the example above, we have listed five explanatory variables that can be connected to the overall costs. This means that a basic bivariate analysis (one variable vs the other) might not be sufficient. Furthermore,

the relations between variables are typically complicated. So analysing, for example, only the relation between the cost of project and the size of building without considering the cost of materials might be misleading.

Here is how the bi-variate relations between the variables in our dataset look like (Figure 11.1):

```
spread(SBA_Chapter_11_Costs)
```

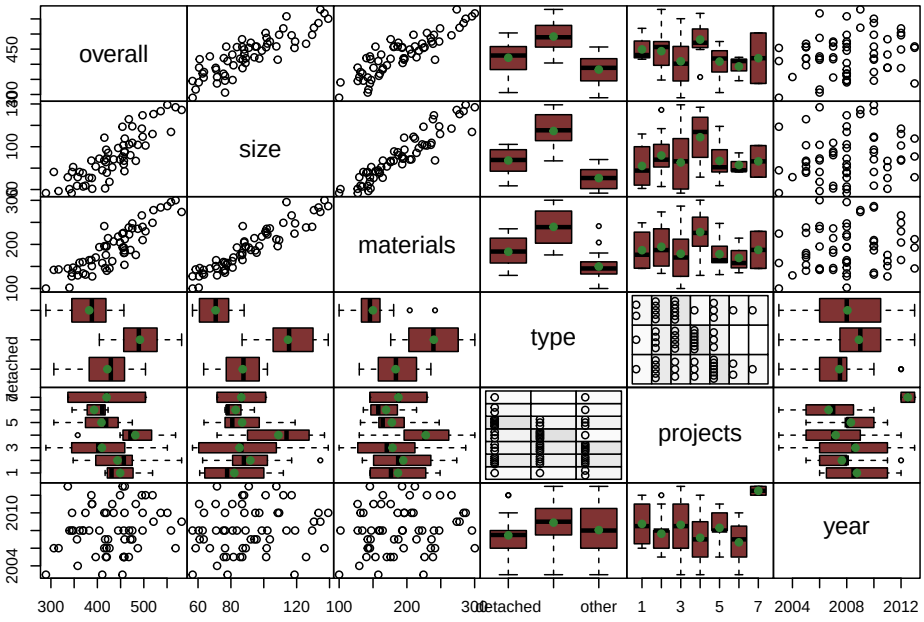


Figure 11.1: Spread plot between variables in the building costs dataset.

While the plot in Figure 11.1 gives a good idea about the relations, for example, between the size of property and the overall costs (it seems to be linear positive) or between material and the overall costs (linear positive again), it does not tell us much about the complex relation between one variable (overall costs) and several others.

All of this gives a motivation to having a so called "Multiple Linear Regression", the model that expresses the relation between one variable and several of others. Mathematically, this is a straight forward extension of the simple linear regression model from Chapter 10, where we just add variables to the right-hand side of the equation. For example, if we had two variables impacting one (e.g. *size* of project and cost of *materials* vs *overall* cost of the project), we could write:

$$overall_j = \beta_0 + \beta_1 size_j + \beta_2 material_j + \epsilon_j, \tag{11.1}$$

where $beta_0$ is the intercept, and $beta_1$, $beta_2$ are the coefficients for the respective variables. The predicted overall costs can be calculated based on this model by

dropping the error term ϵ_j:

$$\widehat{overall}_j = \beta_0 + \beta_1 size_j + \beta_2 material_j.$$

While in the example with the Simple Linear Regression the predicted (or fitted) values implied drawing a line through the cloud of dots on the plane of the two variables, now we are talking about drawing a plane through the point in the three-dimensional space. It can be visualised in the following way (Figure 11.2):

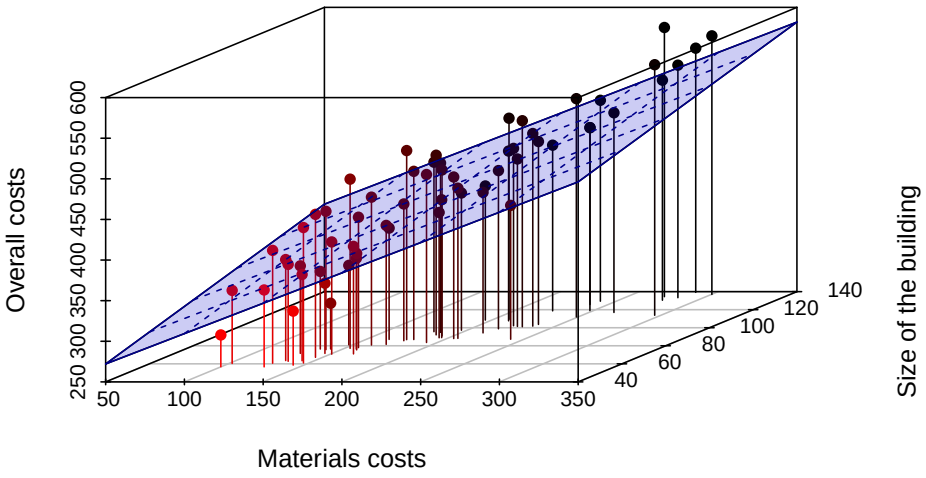


Figure 11.2: 3D scatterplot of Overall costs vs size of project and costs of materials.

The 3d image in Figure 11.2 is already hard to analyse, but at least it gives an idea of how the overall costs change with the change of materials costs and size of buildings. However, it would be impossible to produce a meaningful plot of overall costs from more than two variables. What the figure above gives us is the connection between the simple linear regression (which is just a straight line in the two-dimensional plane) and the multiple one (which is a plane in a multi-dimensional space).

In a more general way, the multiple linear regression can be written as:

$$y_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \cdots + \beta_{k-1} x_{k-1,j} + \epsilon_j, \tag{11.2}$$

where β_i is a i-th parameter for the respective i-th explanatory variable and there is $k-1$ of them in the model, meaning that when we want to estimate this model, we will have k unknown parameters. The regression line of this model in population (aka expectation conditional on the values of explanatory variables) is:

$$\mu_{y,j} = \mathrm{E}(y_j|\mathbf{x}_j) = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \cdots + \beta_{k-1} x_{k-1,j}. \tag{11.3}$$

Furthermore, similar to how we discussed it in Chapter 10, when we want to estimate model @ref{eq:MLRFormula}, we should substitute all parameters β_j

with their estimates b_j:

$$\hat{y}_j = b_0 + b_1 x_{1,j} + b_2 x_{2,j} + \cdots + b_{k-1} x_{k-1,j}. \tag{11.4}$$

Similar to the Simple Linear Regression, each parameter in equation (11.4) represents the slope for the respective variable, showing how on average the value of the response variable (overall costs in our example) changes with the change of each variable.

11.1 OLS estimation

We have already discussed the idea of the OLS in Section 10.1 on the example of the Simple Linear Regression. The logic with the multiple one is exactly the same: we want to minimise the sum of squared errors by changing the values of parameters. The main difference now is that we can have more than two parameters and as a result, the formulae become more complicated.

11.1.1 A bit of maths

There are several ways how we can get the formula for the OLS of the Multiple Linear Regression. We could start with the same SSE value, expanding it based on the equation of the regression (11.4):

$$\text{SSE} = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 = \sum_{j=1}^{n} (y_j - b_0 - b_1 x_{1,j} - b_2 x_{2,j} - \cdots - b_{k-1} x_{k-1,j})^2.$$

Taking derivatives of SSE with respect to b_0, b_1, b_2, …, b_{k-1} and then equating them to zero, we would get a so-called **System of Normal Equations** (we have discussed it when providing a proof in Section 10.1), which in general case has form:

$$\begin{aligned}
&\sum_{j=1}^{n} y_j - n b_0 - b_1 \sum_{j=1}^{n} x_{1,j} - b_2 \sum_{j=1}^{n} x_{2,j} - \cdots - b_{k-1} \sum_{j=1}^{n} x_{k-1,j} = 0 \\
&\sum_{j=1}^{n} y_j x_{1,j} - b_0 \sum_{j=1}^{n} x_{1,j} - b_1 \sum_{j=1}^{n} x_{1,j}^2 - b_2 \sum_{j=1}^{n} x_{2,j} x_{1,j} - \cdots - b_{k-1} \sum_{j=1}^{n} x_{k-1,j} x_{1,j} = 0 \\
&\sum_{j=1}^{n} y_j x_{2,j} - b_0 \sum_{j=1}^{n} x_{2,j} - b_1 \sum_{j=1}^{n} x_{1,j} x_{2,j} - b_2 \sum_{j=1}^{n} x_{2,j}^2 - \cdots - b_{k-1} \sum_{j=1}^{n} x_{k-1,j} x_{2,j} = 0 \\
&\vdots \\
&\sum_{j=1}^{n} y_j x_{k-1,j} - b_0 \sum_{j=1}^{n} x_{k-1,j} - b_1 \sum_{j=1}^{n} x_{1,j} x_{k-1,j} - b_2 \sum_{j=1}^{n} x_{2,j} x_{k-1,j} - \cdots - b_{k-1} \sum_{j=1}^{n} x_{k-1,j}^2 = 0
\end{aligned}$$

Solving this system of equations gives us formulae for parameters b_0, b_1, b_2, …, b_{k-1}, that guarantee that the SSE is minimal.

However, there is a more compact and easier in logic way of getting the formulae, but it requires some basic knowledge of linear algebra. To explain it, we need to

present the multiple linear regression in a more compact form. In order to do that we will introduce the following vectors:

$$\mathbf{x}_j' = \begin{pmatrix} 1 & x_{1,j} & \dots & x_{k-1,j} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{pmatrix}, \tag{11.5}$$

where $'$ symbol is the transposition. This can then be substituted in (11.2) to get:

$$y_j = \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_j. \tag{11.6}$$

This form is just convenient, but it denotes exactly the same model as in equation (11.2). All we need to remember in case of the equation (11.6) is that it represents the sum of products of variables by their coefficients, just in a compact way.

But this is not over yet, we can make it even more compact, if we pack all those values with index j in vectors and matrices:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{k-1,1} \\ 1 & x_{1,2} & \dots & x_{k-1,2} \\ \vdots & & & \\ 1 & x_{1,n} & \dots & x_{k-1,n} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{11.7}$$

where n is the sample size. This leads to the following even more compact form of the multiple linear regression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{11.8}$$

If you compare (11.8) with the original one (11.2):

$$y_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \cdots + \beta_{k-1} x_{k-1,j} + \epsilon_j,$$

you will probably see the connection. But the form (11.8) is just more abstract. This abstraction, however, allows us getting an analytical formula for the calculation of the estimates of parameters of the model (remember, we substitute the true values β_j by their sample estimates b_j):

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{11.9}$$

Proof. Sum of squared errors based on the model (11.8) applied to the data can be expressed as:

$$\text{SSE} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}),$$

where $\mathbf{e}$ is the estimate of $\boldsymbol{\epsilon}$ and $\mathbf{b}$ is the estimate of $\boldsymbol{\beta}$. This can be expanded by opening brackets to:

$$\begin{aligned} \text{SSE} = & \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \\ & \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \end{aligned}, \tag{11.10}$$

which can be done because $\mathbf{y}'\mathbf{X}\mathbf{b}$ is a scalar and $\mathbf{y}'\mathbf{X}\mathbf{b} = (\mathbf{y}'\mathbf{X}\mathbf{b})' = \mathbf{b}'\mathbf{X}'\mathbf{y}$. Now we need to minimise (11.10) with respect to parameters to find their estimates. This can be done by taking derivative of (11.10) with respect to $\mathbf{b}$ and equating it to zero:

$$\frac{\partial \text{SSE}}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0. \tag{11.11}$$

After that, we can regroup the elements in (11.11) to get:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

And then, we can multiply both parts of the equation by the inverse of $\mathbf{X}'\mathbf{X}$ to get rid of that part in the left-hand side of the equation:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which then leads to the final formula:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

□

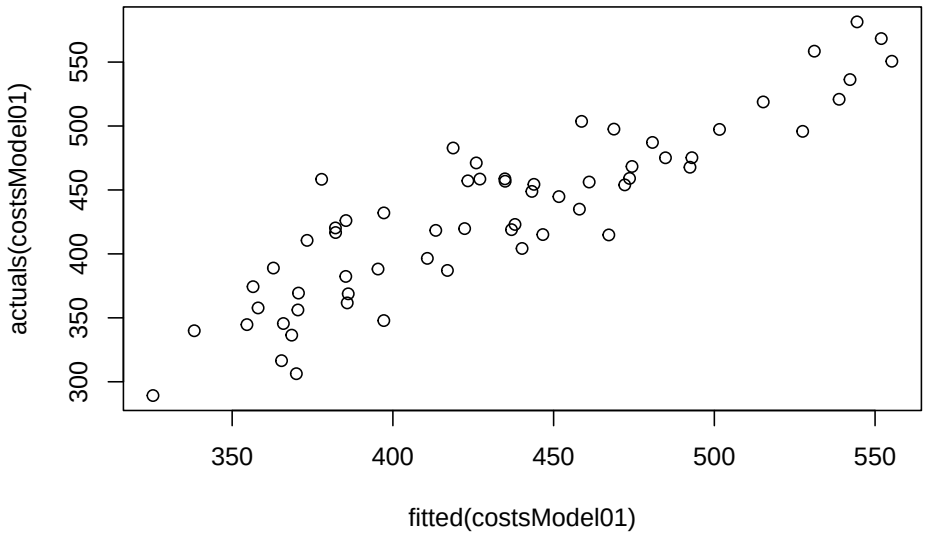
11.1.2 Application

Luckily, you do not need to remember the formula (11.9) and should not need to use it in real life, because it is used in all statistical software, including `lm()` function from `stats` package for R. Here is an example with the same dataset:

```
costsModel01 <- lm(overall~size+materials+projects+year, SBA_Chapter_11_Costs)
```

To better understand what fitting such model to the data implies, we can produce a plot of fitted vs actuals values, with $\hat{y}_j$ on x-axis and y_j on the y-axis:

```
plot(fitted(costsModel01),actuals(costsModel01))
```



The same plot is produced via `plot()` method if we use `alm()` function from `greybox` instead:

```
costsModel02 <- alm(overall~size+materials+projects+year, SBA_Chapter_11_Costs, loss="MSE")
plot(costsModel02,1)
```

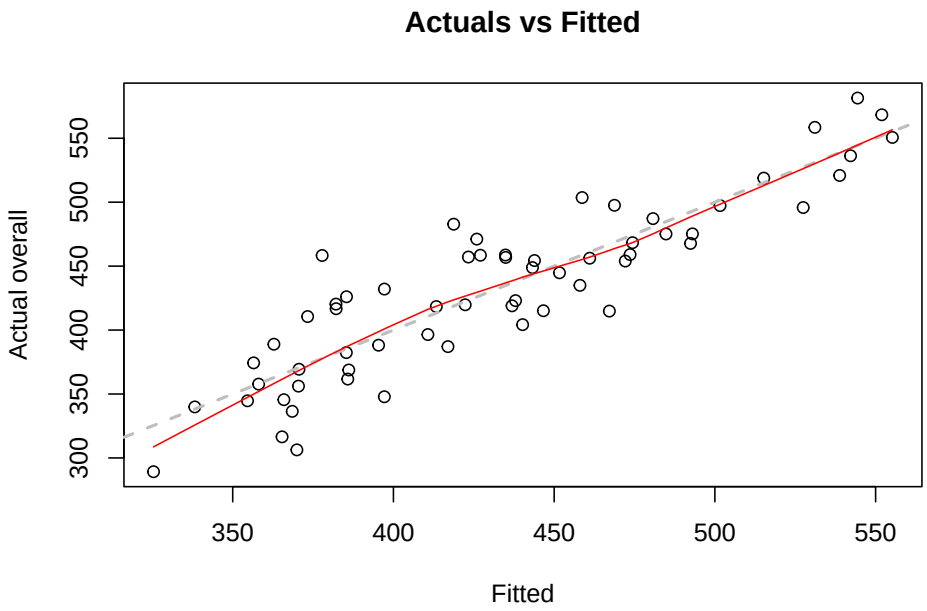


Figure 11.3: Actuals vs fitted values for multiple linear regression model on mtcars data.

We use `loss="MSE"` in this case, to make sure that the model is estimated via OLS. We will discuss the default estimation method in `alm()`, likelihood, in Section 16.

The plot on Figure 11.3 can be used for diagnostic purposes and in ideal situation the red line (LOWESS line) should coincide with the grey one, which would mean that we have correctly capture the tendencies in the data, so that all the regression assumptions are satisfied (see Chapter 15).

11.1.3 Properties of the OLS Estimators

There are several important properties of the OLS estimated regression that are worth keeping in mind:

1. The mean of residuals of the model is always equal to zero as long as it contains intercept.
2. The explanatory variables in the model are not correlated with the residuals of the model.
3. The mean of the fitted values coincides with the mean of the actual values.

They all follow directly from the derivation of the OLS formula.

Proof. Consider the system of normal equations

$$(\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{11.12}$$

Given that we estimated the model in sample, we can rewrite the multiple linear regression as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{e}$$

and substitute it in (11.12) to get:

$$(\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\mathbf{X}\boldsymbol{b} + \mathbf{X}'\boldsymbol{e}. \tag{11.13}$$

Equation (11.13) implies that $\mathbf{X}'\boldsymbol{e} = 0$, which given how the matrix $\mathbf{X}$ is formed prooves the first two properties:

1. The first column of the matrix (which contains ones) corresponds to the intercept and the multiplication of it by the error implies that $\sum_{j=1}^{n} e_j = 0$, which also means that the mean of the residuals is zero as well: $\bar{e} = 0$;
2. All the other columns contain the explanatory variables and for each one of them this comes to the equation: $\sum_{j=1}^{n} x_{i,j} e_j$ for all $i = \{1, \dots, k-1\}$. Given that the first property holds, the same equation can be rewritten to:

   $$\sum_{j=1}^{n} x_{i,j}(e_j - \bar{e}) = \sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)(e_j - \bar{e}) = \sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)(e_j - \bar{e}) + \bar{x}_i \sum_{j=1}^{n}(e_j - \bar{e})$$

   The right hand side of this equation equals to zero due to the first property, which implies that:

   $$\sum_{j=1}^{n} x_{i,j} e_j = \sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)(e_j - \bar{e}), \tag{11.14}$$

   which is the covariance between the residuals and the explanatory variable x_i. And because the equation (11.13) implies that $\sum_{j=1}^{n} x_{i,j} e_j = 0$, the covariance (11.14) should be equal to zero as well.

Finally, it is easy to show the third property. All we need to do is to take the mean of the response variable:

$$\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{X}\boldsymbol{b} + \boldsymbol{e}) = \mathrm{E}(\mathbf{X}\boldsymbol{b}) + \mathrm{E}(\boldsymbol{e}),$$

where the expectation of the error term equals to zero, implying that:

$$\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{X}\boldsymbol{b}) = \mathrm{E}(\hat{\mathbf{y}}).$$

□

These three properties are useful because, first, they show that it does not make sense to test whether they hold or not, with OLS they will be satisfied automatically, and second they allow measuring quality of the fit via the squares of regression, which will be discussed in Section 11.2.

11.1.4 Gauss-Markov theorem

OLS is a very popular estimation method for linear regression for a variety of reasons. First, it is relatively simple (much simpler than other approaches) and conceptually easy to understand. Second, the estimates of OLS parameters can be found analytically (using formula (10.5)). Furthermore, there is a mathematical proof that the OLS estimates of parameters are efficient (Subsection 6.3.2), consistent (Subsection 6.3.3) and unbiased (Subsection 6.3.1). The theorem that states that is called "Gauss-Markov theorem", here is one of versions of it:

Theorem 11.1. *If regression model is correctly specified then OLS will produce Best Linear Unbiased Estimates (BLUE) of parameters.*

The term "correctly specified" implies that all main statistical assumptions about the model are satisfied (such as no omitted important variables, no autocorrelation and heteroscedasticity in the residuals, see details in Chapter 15). The "BLUE" part means that OLS guarantees the most efficient and the least biased estimates of parameters amongst all possible estimators of a linear model. For example, if we used a criterion of minimisation of Mean Absolute Error (MAE), then the estimates of parameters would be less efficient than in case of OLS. This is because OLS gives "mean" estimates, while the minimum of MAE corresponds to the median (see Subsection 6.3.2).

Practically speaking, the theorem implies that when you use OLS, the estimates of parameters will have good statistical properties (given that the model is correctly specified), in some cases better than the estimates obtained using other methods.

11.2 Quality of a fit

Building upon the discussion of the quality of the fit in Section 10.4, we can introduce a measure, based on the OLS criterion, (10.4), which is called either "Root Mean Squared Error" (RMSE) or a "standard error" or a "standard deviation of error" of the regression:

$$\hat{\sigma}^2 = \sqrt{\frac{1}{n-k}\sum_{j=1}^{n} e_j^2}. \tag{11.15}$$

The denominator of (11.15) contains the number of degrees of freedom in the model, $n-k$, not the number of observations n, so technically speaking this is not a "mean" any more. This is done to correct the in-sample bias (Section 6.3.1) of the measure. Standard error does not tell us much about the in-sample performance but can be used to compare several models with the same response variable between each other: the lower it is, the better the model fits the data, given the number of estimated parameters. However, this measure is not aware

of the randomness in the true model (Section 1.1.1) and thus will be equal to zero in a model that fits the data perfectly (thus ignoring the existence of error term). This is a potential issue, as we might end up with a poor model that would seem like the best one.

Here is how this can be calculated for our model, estimated using `alm()` function:

```
sigma(costsModel02)
```

```
## [1] 30.56428
```

The value of RMSE does not provide any important insights on its own, but it can be compared to the RMSE of another model to decide, which one of the two fits the data better.

Similarly to the simple linear regression, we can calculate the R^2 (see Section 10.4). The problem is that the value of coefficient of determination would always increase with the increase of number of variables included in the model. This is because every variable will explain some proportion of the data due to randomness. So, if we add redundant variables, the fit will improve, but the quality of model will deteriorate. Here is an example:

```
# Record number of observations
n <- nobs(costsModel02)
# Generate white noise
SBA_Chapter_11_Costs$noise <- rnorm(n,0,10)
# Add it to the model
costsModel02WithNoise <- alm(overall~size+materials+projects+year+noise,
                             SBA_Chapter_11_Costs, loss="MSE")
```

The code above introduces a new variable, `noise`, which has nothing to do with the `overall` costs. We would expect that this variable would not bring value to the model. And here is the value of determination coefficient of the new model:

```
1 - sum(resid(costsModel02WithNoise)^2) /
    (var(actuals(costsModel02WithNoise))*(n-1))
```

```
## [1] 0.8029458
```

Compare it with the previous one:

```
1 - sum(resid(costsModel02)^2) /
    (var(actuals(costsModel02))*(n-1))
```

```
## [1] 0.8016625
```

The value in the new model will always be higher than in the previous one (or equal to it in some very special cases), no matter how we generate the random fluctuations. This means that some sort of penalisation of the number of estimated parameters is required to make the measure more reasonable. This

is what adjusted coefficient of determination does:

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\text{V}(y)} = 1 - \frac{(n-1)\text{SSE}}{(n-k)\text{SST}}. \tag{11.16}$$

So, instead of dividing sums of squares, in the adjusted R^2 we divide the entities that are based on degrees of freedom. Given the presence of k in the formula (11.16), the coefficient will not necessarily increase with the addition of variables – when the variable does not contribute in the reduction of SSE of model substantially, R^2 will not go up. Furthermore, if one model has higher $\hat{\sigma}^2$ than the other one, then the R^2_{adj} of that model will be lower, which becomes apparent, given that we have $-\hat{\sigma}^2$ in the formula (11.16).

Here how the adjusted R^2 can be calculated for a model in R:

```
setNames(c(1 - sigma(costsModel02)^2 / var(actuals(costsModel02)),
           1 - sigma(costsModel02WithNoise)^2 / var(actuals(costsModel02WithNoise))),
         c("R^2-adj","R^2-adj, Noise"))
```

```
##         R^2-adj R^2-adj, Noise
##       0.7874955      0.7850317
```

What we will typically see in the output above is that the model with the noise will have a lower value of adjusted R^2 than the model without it. However, given that we deal with randomness, if you reproduce this example many times, you will see different situation, including those, where introducing noise still increases the value of the parameter just due to pure chance. So, you should not fully trust R^2_{adj} either. When constructing a model or deciding what to include in it, you should always use your judgement - make sure that the variables included in the model are meaningful. Otherwise you can easily overfit the data, which would lead to inefficient estimates of parameters (see Section 15 for details) and inaccurate forecasts.

11.2.1 Common mistakes related to quality of a fit

There are several common mistakes that arise when people measure quality of a fit of regression. We have seen these mistakes done by students, but they also appear on social media and sometimes even in scientific papers. Here they are:

1. "Model is good because R^2/Adjusted R^2 is 0.9876/greater than some arbitrary threshold"

- Neither R^2, nor Adjusted R^2 tells anything about the quality of the model. They only tell us how well it fits the data. In case of the former, it shows the "percentage of the explained variance in the response variable", but they both do not know that any model has an irreducible error, and thus a high value of R^2 does not mean that the model is good. R^2 should not be used for model selection and provides little to no useful diagnostic

information. R^2 adjusted can be used for model comparison, but on its own does not provide useful information either.

2. "Although R^2/Adjusted R^2 is very low (e.g. 0.05), the model is statistically significant and thus makes sense"

- This argument relates to what we will discuss in Section 12.2, but we can say now that the low value of the coefficient of determination indicates that the model does not differ much from the straight line (global mean). To that extent, its low value could be alarming.

In general, we think that both R^2 close to one and close to zero are alarming, the former implies that the model might overfit the data, while the latter means that it underfits it. But there is no proper threshold value, with which you should compare your coefficient of determination.

3. "Model A is better than model B because its R^2 value is higher"

- As discussed in this section, R^2 will increase with the increase of the number of parameters, so this statement is in general wrong. It only works if you have two models with exactly the same number of parameters. But usually, we see such statements in the context of variable selection, when models have different number of parameters. If you want to use basic statistics for model selection then at least use RMSE or the adjusted R^2. But there is a better way, which we will discuss in Section 16.4.

11.3 Interpretation of parameters

Finally, we come to the discussion of parameters of a model. As mentioned earlier, each one of them represents the slope of the model. But there is more to the meaning of parameters of the model. Consider the coefficients of the previously estimated model:

```
coef(costsModel02)
```

```
##   (Intercept)          size     materials      projects          year
## -2964.6192397     0.8970045     0.7743759    -5.3095572     1.5864555
```

Each of the parameters of this model shows an **average** effect of each variable on the overall costs. They have a simple interpretation and show how the response variable will change **on average** with the increase of a variable by 1 unit, keeping all the other variables constant.

For example, the parameter for `size` shows that with the increase of size of the building by on squared meter, the overall cost tends to increase **on average** by 0.897 thousand pounds, if all the other variables do not change.

I have made the word "average" boldface three times in this section for a reason. This is a very important point to keep in mind - the parameters will not tell you how variable will change for any specific observation. Any regression model

captures mean tendencies and thus the word "average" is very important in the interpretation. In each specific case, the increase of size by 1 squared meter will lead to different increases (and even decreases in some cases) of the overall costs. But if we take the arithmetic mean of those individual effects, it should be close to the value of the parameter in the model. This however is only possible if all the assumptions of regression hold (see Section 15).

Finally, it is worth discussing what the interpretation of the intercept in the model is. If we set all the explanatory variables to zero, the overall costs will be equal to the value of the intercept. In our example, where we fitted the basic linear model, the interpretation is meaningless: there is no such thing as a house with no costs, size of zero, and it definitely cannot have negative overall costs. In this case, intercept plays purely technical role, showing where the regression line intersects the y-axis. However, if we were to build a different model, the value might have a meaning. Still, we personally prefer avoiding interpreting the intercept.

Chapter 12

Uncertainty in regression

Coming back to the example of mileage vs weight of cars, the estimated simple linear regression on the data was mpg=37.29-5.34wt+et. But what would happen if we estimate the same model on a different sample of data (e.g. 15 first observations instead of 32)?

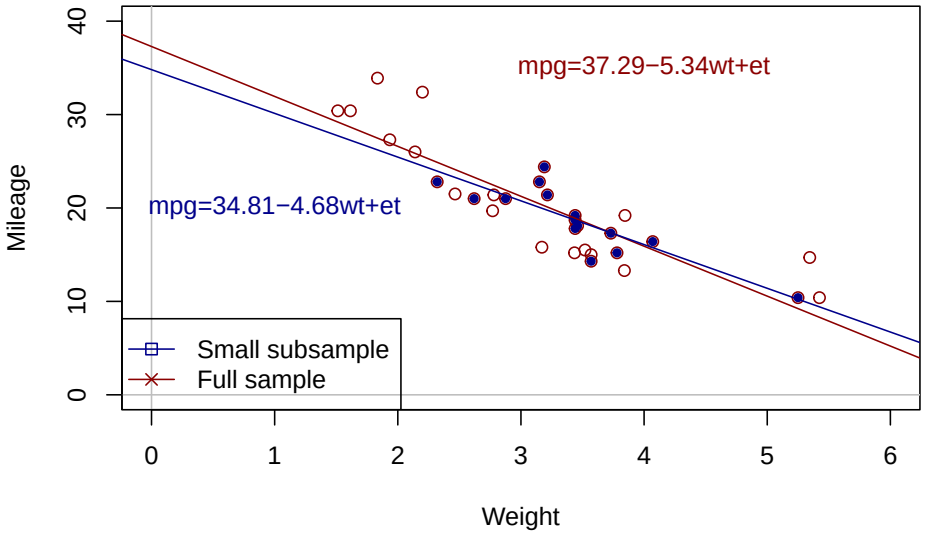


Figure 12.1: Weight vs mileage and two regression lines.

Figure 12.1 shows the two lines: the red one corresponds to the larger sample, while the blue one corresponds to the small one. We can see that these lines have different intercepts and slope parameters. So, which one of them is correct? An amateur analyst would say that the one that has more observations is the correct model. But a more experienced statistician would tell you that none of

the two is correct. They are both estimated on a sample of data and they both inevitably inherit the uncertainty of the data, making them both incorrect if we compare them to the hypothetical true model. This means that whatever regression model we estimate on a sample of data, it will be incorrect as well.

This uncertainty about the regression line actually comes to the uncertainty of estimates of parameters of the model. In order to see it more clearly, consider the example with Speed and Stopping Distances of Cars dataset from `datasets` package (`?cars`):

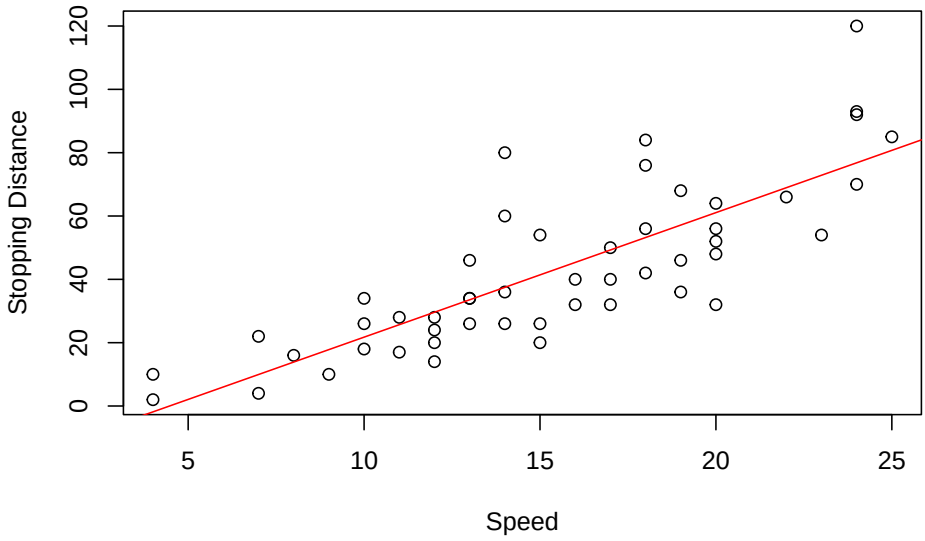


Figure 12.2: Speed vs stopping distance of cars

While the linear relation between these variables might be not the the most appropriate, it suffices for demonstration purposes. What we will do for this example is fit the model and then use a simple bootstrap technique to get estimates of parameters of the model. We will do that using `coefbootstrap()` method from `greybox` package. The bootstrap technique implemented in the function applies the same model to subsamples of the original data and returns a matrix with parameters. This way we get an idea about the empirical uncertainty of parameters:

```
slmSpeedDistanceBootstrap <- coefbootstrap(slmSpeedDistance)
```

Based on that we can plot the histograms of the estimates of parameters.

```
par(mfcol=c(1,2))
hist(slmSpeedDistanceBootstrap$coefficients[,1],
     xlab="Intercept", main="")
hist(slmSpeedDistanceBootstrap$coefficients[,2],
     xlab="Slope", main="")
```

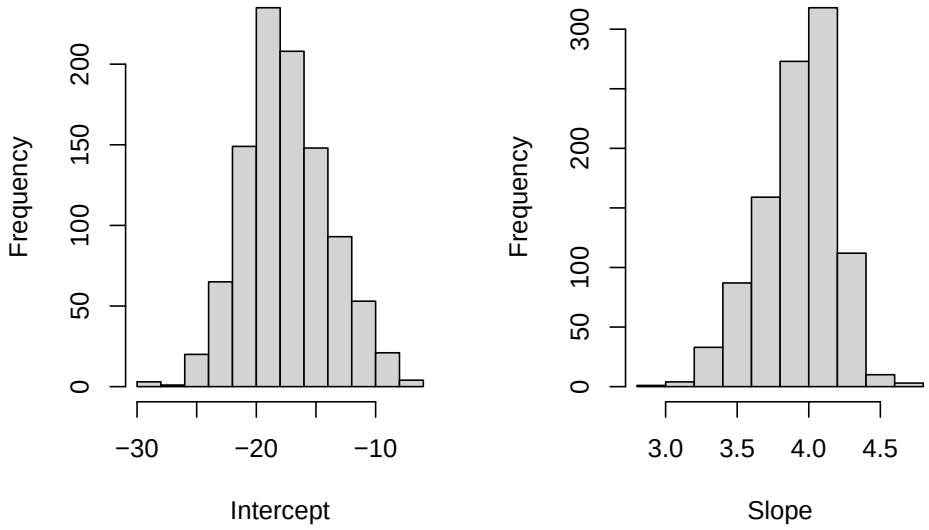


Figure 12.3: Distribution of bootstrapped parameters of a regression model

Figure 12.3 shows the uncertainty around the estimates of parameters. These distributions look similar to the normal distribution. In fact, if we repeated this example thousands of times, the distribution of estimates of parameters would indeed follow the normal one due to CLT (if the assumptions hold, see Sections 6.2 and 15). As a result, when we work with regression we should take this uncertainty about the parameters into account. This applies to both parameters analysis and forecasting.

12.1 Confidence intervals

In order to take this uncertainty into account, we could construct confidence intervals for the estimates of parameters, using the principles discussed in Section 6.4. This way we would hopefully have some idea about the uncertainty of the parameters, and not just rely on average values. If we assume that CLT holds, we could use the t statistics for the calculation of the quantiles of distribution (we need to use t because we do not know the variance of estimates of parameters). But in order to do that, we need to have variances of estimates of parameters. One of possible ways of getting them would be the bootstrap used in the example above. However, this is a computationally expensive operation, and there is a more efficient procedure, which however only works with linear regression models either estimated using OLS or via Maximum Likelihood Estimation assuming Normal distribution (see Section 16). In these conditions the covariance matrix

of parameters can be calculated using the following formula:

$$\mathrm{V}(\hat{\boldsymbol{\beta}}) = \frac{1}{n-k}\sum_{j=1}^{n} e_j^2 \times \left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{12.1}$$

This matrix will contain variances of parameters on the diagonal and covariances between the parameters on off-diagonals. In this specific case, we only need the diagonal elements. We can take square root of them to obtain standard errors of parameters, which can then be used to construct confidence intervals for each parameter i via:

$$\beta_i \in (b_i + t_{\alpha/2}(n-k)s_{b_i}, b_i + t_{1-\alpha/2}(n-k)s_{b_i}), \tag{12.2}$$

where s_{b_i} is the standard error of the parameter b_i. All modern software does all these calculations automatically, so we do not need to do them manually. Here is an example:

```
vcov(slmSpeedDistance)
```

```
##             (Intercept)      speed
## (Intercept)   45.676514 -2.6588234
## speed         -2.658823  0.1726509
```

This is the covariance matrix of parameters, the diagonal elements of which are then used in the `confint()` method:

```
confint(slmSpeedDistance)
```

```
##                   S.E.       2.5%     97.5%
## (Intercept) 6.7584402 -31.167850 -3.990340
## speed       0.4155128   3.096964  4.767853
```

The confidence interval for speed above shows, for example, that if we repeat the construction of interval many times, the true value of parameter speed will lie in 95% of cases between 3.08 and 4.78. This gives an idea about the real effect in the population. We can also present all of this in the following summary (this is based on the `alm()` model, the other functions will produce different summaries):

```
summary(slmSpeedDistance)
```

```
## Response variable: dist
## Distribution used in the estimation: Normal
## Loss function used in estimation: MSE
## Coefficients:
##             Estimate Std. Error Lower 2.5% Upper 97.5%
## (Intercept) -17.5791     6.7584   -31.1678     -3.9903 *
## speed         3.9324     0.4155     3.0970      4.7679 *
##
```

```
## Error standard deviation: 15.3796
## Sample size: 50
## Number of estimated parameters: 2
## Number of degrees of freedom: 48
## Information criteria:
##      AIC     AICc      BIC     BICc
## 419.1569 417.4122 424.8929 421.4803
```

This summary provide all the necessary information about the estimates of parameters: their mean values in the column "Estimate", their standard errors in "Std. Error", the bounds of confidence interval and finally a star if the interval does not contain zero. This typically indicates that we are certain on the selected confidence level (95% in our example) about the sign of the parameter and that the effect really exists.

12.2 Hypothesis testing

Another way to look at the uncertainty of parameters is to test a statistical hypothesis. As it was discussed in Section 7, I personally think that hypothesis testing is a less useful instrument for these purposes than the confidence interval and that it might be misleading in some circumstances. Nonetheless, it has its merits and can be helpful if an analyst knows what they are doing. In order to test the hypothesis, we need to follow the procedure, described in Section 7.

12.2.1 Regression parameters

The classical hypotheses for the parameters are formulated in the following way:

$$\begin{aligned} &\mathrm{H}_0 : \beta_i = 0 \\ &\mathrm{H}_1 : \beta_i \neq 0 \end{aligned}. \tag{12.3}$$

This formulation of hypotheses comes from the idea that we want to check if the effect estimated by the regression is indeed there (i.e. statistically significantly different from zero). Note however, that as in any other hypothesis testing, if you fail to reject the null hypothesis, this only means that you do not know, we do not have enough evidence to conclude anything. This **does not mean** that there is no effect and that the respective variable can be removed from the model. In case of simple linear regression, the null and alternative hypothesis can be represented graphically as shown in Figure 12.4.

The graph on the left in Figure 12.4 demonstrates how the true model could look if the null hypothesis was true - it would be just a straight line, parallel to x-axis. The graph on the right demonstrates the alternative situation, when the parameter is not equal to zero. We do not know the true model, and hypothesis testing does not tell us, whether the hypothesis is true or false, but if we have enough evidence to reject H_0, then we might conclude that we see an effect of one variable on another in the data. Note, as discussed in Section 7, the null

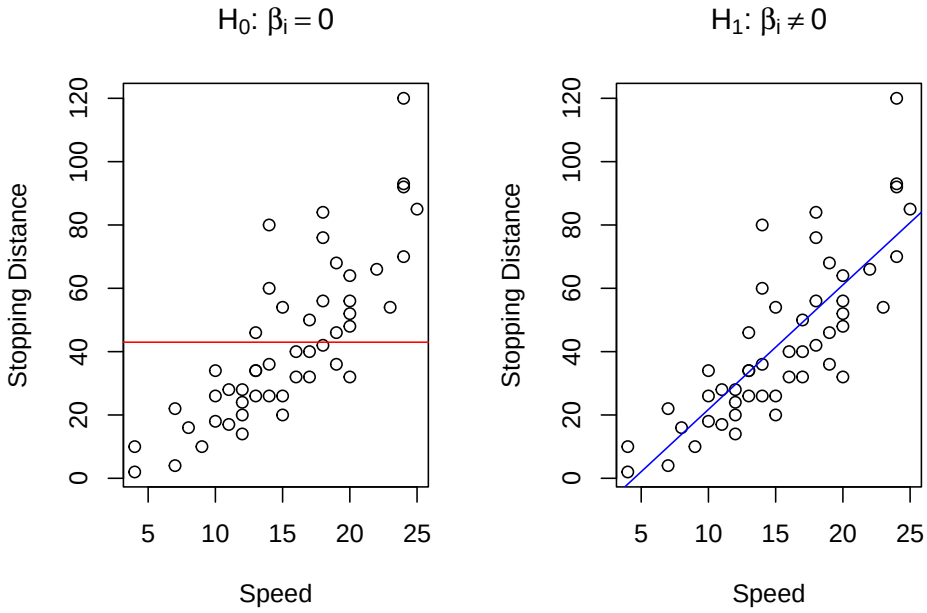


Figure 12.4: Graphical presentation of null and alternative hypothesis in regression context

hypothesis is always wrong, and it will inevitably be rejected with the increase of sample size.

Given the discussion in the previous subsection, we know that the parameters of regression model will follow normal distribution, as long as all assumptions are satisfied (including those for CLT). We also know that because the standard errors of parameters are estimated, we need to use Student's distribution, which takes the uncertainty about the variance into account. Based on this, we can say that the following statistics will follow t with $n - k$ degrees of freedom:

$$\frac{b_i - 0}{s_{b_i}} \sim t(n-k). \tag{12.4}$$

After calculating the value and comparing it with the critical t-value on the selected significance level or directly comparing p-value based on (12.4) with the significance level, we can make conclusions about the hypothesis.

The context of regression provides a great example, why we never accept hypothesis and why in the case of "Fail to reject H_0", we should not remove a variable (unless we have more fundamental reasons for doing that). Consider an example, where the estimated parameter $b_1 = 0.5$, and its standard error is $s_{b_1} = 1$, we estimated a simple linear regression on a sample of 30 observations, and we want to test, whether the parameter in the population is zero (i.e. hypothesis (12.3))

on 1% significance level. Inserting the values in formula (12.4), we get:

$$\frac{|0.5-0|}{1}=0.5,$$

with the critical value for two-tailed test of $t_{0.01}(30-2)\approx 2.76$. Comparing t-value with the critical one, we would conclude that we fail to reject H_0 and thus the parameter is not statistically different from zero. But what would happen if we check another hypothesis:

$$\begin{aligned}H_0 &: \beta_1 = 1\\ H_1 &: \beta_1 \neq 1\end{aligned}.$$

The procedure is the same, the calculated t-value is:

$$\frac{|0.5-1|}{1}=0.5,$$

which leads to exactly the same conclusion as before: on 1% significance level, we fail to reject the new H_0, so the value is not distinguishable from 1. So, which of the two is correct? The correct answer is "we do not know". The non-rejection region just tells us that uncertainty about the parameter is so high that it also include the value of interest (0 in case of the classical regression analysis). If we constructed the confidence interval for this problem, we would not have such confusion, as we would conclude that on 1% significance level the true parameter lies in the region $(-2.26, 3.26)$ and can be any of these numbers.

In R, if you want to test the hypothesis for parameters, I would recommend using `lm()` function for regression:

```
lmSpeedDistance <- lm(dist~speed,cars)
summary(lmSpeedDistance)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
```

```
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

This output tells us that when we consider the parameter for the variable speed, we reject the standard H_0 on the pre-selected 1% significance level (comparing the level with p-value in the last column of the output). Note that we should first select the significance level and only then conduct the test, otherwise we would be bending reality for our needs.

12.2.2 Regression line

Finally, in regression context, we can test another hypothesis, which becomes useful, when a lot of parameters of the model are very close to zero and seem to be insignificant on the selected level:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0 \\ H_1 &: \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \cdots \vee \beta_{k-1} \neq 0' \end{aligned} \tag{12.5}$$

which translates into normal language as "H_0: all parameters (except for intercept) are equal to zero; H_1: at least one parameter is not equal to zero". This hypothesis is only needed, when you have a model with many statistically insignificant variables and want to see if the model explains anything. This is done using F-test, which can be calculated based on sums of squares:

$$F = \frac{SSR/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k),$$

where the sums of squares are divided by their degrees of freedom. The test is conducted in the similar manner as any other test (see Section 7): after choosing the significance level, we can either calculate the critical value of F for the specified degrees of freedom, or compare it with the p-value from the test to make a conclusion about the null hypothesis.

This hypothesis is not very useful, when the parameter are significant and coefficient of determination is high. It only becomes useful in difficult situations of poor fit. The test on its own does not tell if the model is adequate or not. And the F value and related p-value is not comparable with respective values of other models. Graphically, this test checks, whether in the true model the slope of the straight line on the plot of actuals vs fitted is different from zero. An example with the same stopping distance model is provided in Figure 12.5.

What the test is tries to get insight about, is whether in the true model the blue line coincides with the red line (i.e. the slope is equal to zero, which is only possible, when all parameters are zero). If we have enough evidence to reject the null hypothesis, then this means that the slopes are different on the selected significance level.

Here is an example with the speed model discussed above with the significance level of 1%:

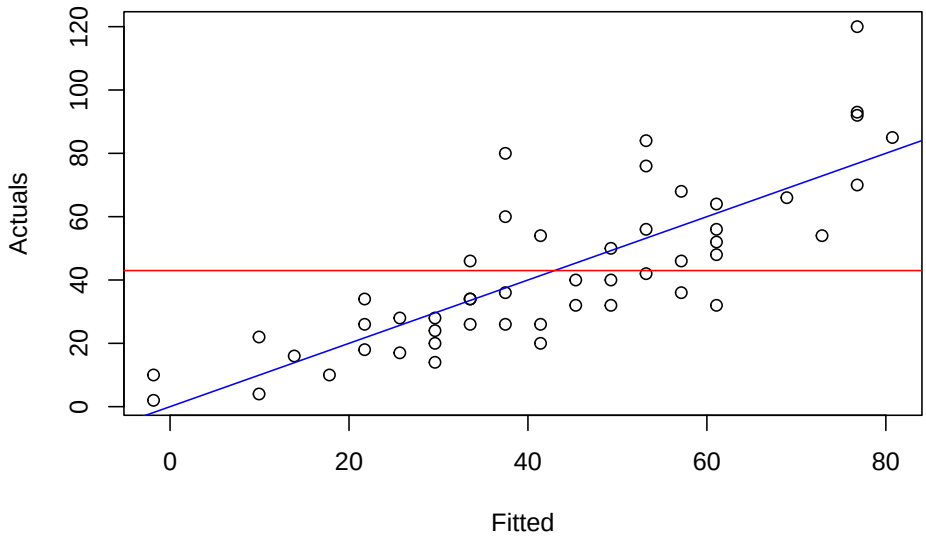


Figure 12.5: Graphical presentation of F test for regression model.

```
lmSpeedDistanceF <- summary(lmSpeedDistance)$fstatistic
# F value
lmSpeedDistanceF[1]
```

```
##    value
## 89.56711
```

```
# F critical
qf(0.99,lmSpeedDistanceF[2],lmSpeedDistanceF[3])
```

```
## [1] 7.194218
```

```
# p-value from the test
1-pf(lmSpeedDistanceF[1],lmSpeedDistanceF[2],lmSpeedDistanceF[3])
```

```
##        value
## 1.489919e-12
```

In the output above, the critical value is lower than the calculated, so we can reject the H_0, which means that there is something in the model that explains the variability in the variable `dist`. Alternatively, we could focus on p-value. We see that the it is lower than the significance level of 1%, so we reject the H_0 and come to the same conclusion as above.

12.3 Regression line uncertainty

Given the uncertainty of estimates of parameters, the regression line itself and the points around it will be uncertain. This means that in some cases we should not just consider the predicted values of the regression $\hat{y}_j$, but also the uncertainty around them.

The uncertainty of the regression line builds upon the uncertainty of parameters and can be measured via the conditional variance in the following way:

$$\mathrm{V}(\hat{y}_j|\mathbf{x}_j) = \mathrm{V}(b_0 + b_1 x_{1,j} + b_2 x_{2,j} + \cdots + b_{k-1} x_{k-1,j}), \tag{12.6}$$

which after some simplifications leads to:

$$\mathrm{V}(\hat{y}_j|\mathbf{x}_j) = \sum_{l=0}^{k-1} \mathrm{V}(b_j) x_{l,j}^2 + 2 \sum_{l=1}^{k-1} \sum_{i=0}^{l-1} \mathrm{cov}(b_i, b_l) x_{i,j} x_{l,j}, \tag{12.7}$$

where $x_{0,j} = 1$. As we see, the variance of the regression line involves variances and covariances of parameters. This variance can then be used in the construction of the confidence interval for the regression line. Given that each estimate of parameter b_i will follow normal distribution with a fixed mean and variance due to CLT, the predicted value $\hat{y}_j$ will follow normal distribution as well. This can be used in the construction of the confidence interval, in a manner similar to the one discussed in Section 6.4:

$$\mu_j \in (\hat{y}_j + t_{\alpha/2}(n-k) s_{\hat{y}_j}, \hat{y}_j + t_{1-\alpha/2}(n-k) s_{\hat{y}_j}), \tag{12.8}$$

where $s_{\hat{y}_j} = \sqrt{\mathrm{V}(\hat{y}_j|\mathbf{x}_j)}$.

In R, this interval can be constructed via the function `predict()` with `interval="confidence"`. It is based on the covariance matrix of parameters, extracted via `vcov()` method in R (it was discussed in a previous subsection). Note that the interval can be produced not only for the in-sample value, but for the holdout as well. Here is an example with `alm()` function:

```
slmSpeedDistanceCI <- predict(slmSpeedDistance,interval="confidence")
plot(slmSpeedDistanceCI, main="",
     xlab="Observation", ylab="Distance")
```

The same fitted values and interval can be presented differently on the actuals vs fitted plot:

```
plot(fitted(slmSpeedDistance),actuals(slmSpeedDistance),
     xlab="Fitted",ylab="Actuals")
abline(a=0,b=1,col="darkblue",lwd=2)
lines(sort(fitted(slmSpeedDistance)),
      slmSpeedDistanceCI$lower[order(fitted(slmSpeedDistance))],
      col="darkred", lwd=2)
lines(sort(fitted(slmSpeedDistance)),
```

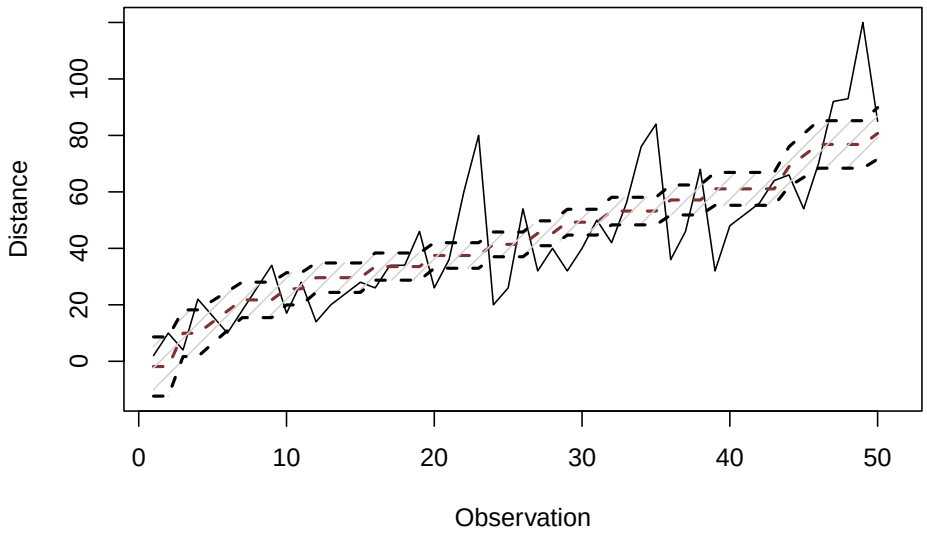


Figure 12.6: Fitted values and confidence interval for the stopping distance model.

```
      slmSpeedDistanceCI$upper[order(fitted(slmSpeedDistance))],
      col="darkred", lwd=2)
```

Figure 12.7 demonstrates the actuals vs fitted plot, together with the 95% confidence interval around the line, demonstrating where the line would be expected to be in 95% of the cases if we re-estimate the model many times. We also see that the uncertainty of the regression line is lower in the middle of the data, but expands in the tails. Conceptually, this happens because the regression line, estimated via OLS, always passes through the average point of the data $(\bar{x}, \bar{y})$ and the variability in this point is lower than the variability in the tails.

If we are not interested in the uncertainty of the regression line, but rather in the uncertainty of the observations, we can refer to prediction interval. The variance in this case is:

$$\mathrm{V}(y_j|\mathbf{x}_j) = \mathrm{V}(b_0 + b_1 x_{1,j} + b_2 x_{2,j} + \cdots + b_{k-1} x_{k-1,j} + e_j), \tag{12.9}$$

which can be simplified to (if assumptions of regression model hold, see Section 15):

$$\mathrm{V}(y_j|\mathbf{x}_j) = \mathrm{V}(\hat{y}_j|\mathbf{x}_j) + \hat{\sigma}^2, \tag{12.10}$$

where $\hat{\sigma}^2$ is the variance of the residuals e_j. As we see from the formula (12.10), the variance in this case is larger than (12.7), which will result in wider interval than the confidence one. We can use normal distribution for the construction of the interval in this case (using formula similar to (12.8)), as long as we can assume that $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$.

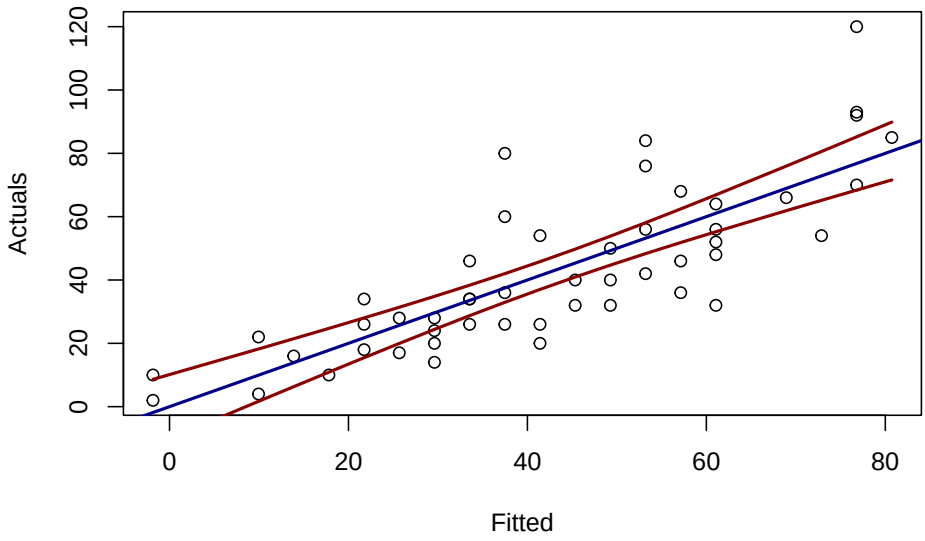


Figure 12.7: Actuals vs Fitted and confidence interval for the stopping distance model.

In R, this can be done via the very same `predict()` function with `interval="prediction"`:

```
slmSpeedDistancePI <- predict(slmSpeedDistance,interval="prediction")
```

Based on this, we can construct graphs similar to 12.6 and 12.7.

Figure 12.8 shows the prediction interval for values over observations and for actuals vs fitted. As we see, the interval is wider in this case, covering only 95% of observations (there are 2 observations outside it).

In forecasting, prediction interval has a bigger importance than the confidence interval. This is because we are typically interested in capturing the uncertainty about the observations, not about the estimate of a line. Typically, the prediction interval would be constructed for some holdout data, which we did not have at the model estimation phase. In the example with stopping distance, we could see what would happen if the speed of a car was, for example, 30mph:

```
slmSpeedDistanceForecast <- predict(slmSpeedDistance,newdata=data.frame(speed=30),
                                    interval="prediction")
plot(slmSpeedDistanceForecast)
```

Figure 12.9 shows the point forecast (the expected stopping distance if the speed of car was 30mph) and the 95% prediction interval (we expect that in 95% of the cases, the cars will have the stopping distance between 66.865 and 133.921 feet.

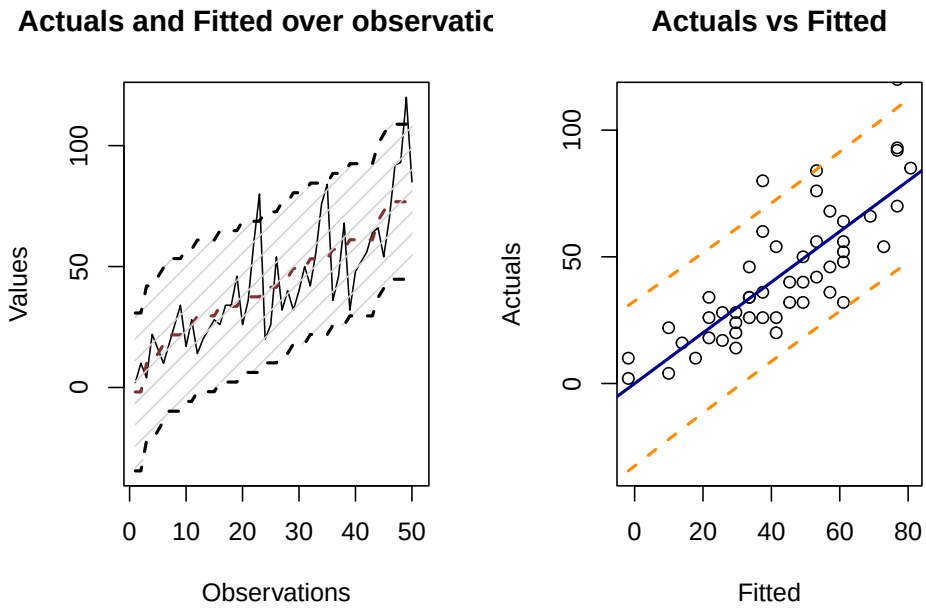


Figure 12.8: Fitted values and prediction interval for the stopping distance model.

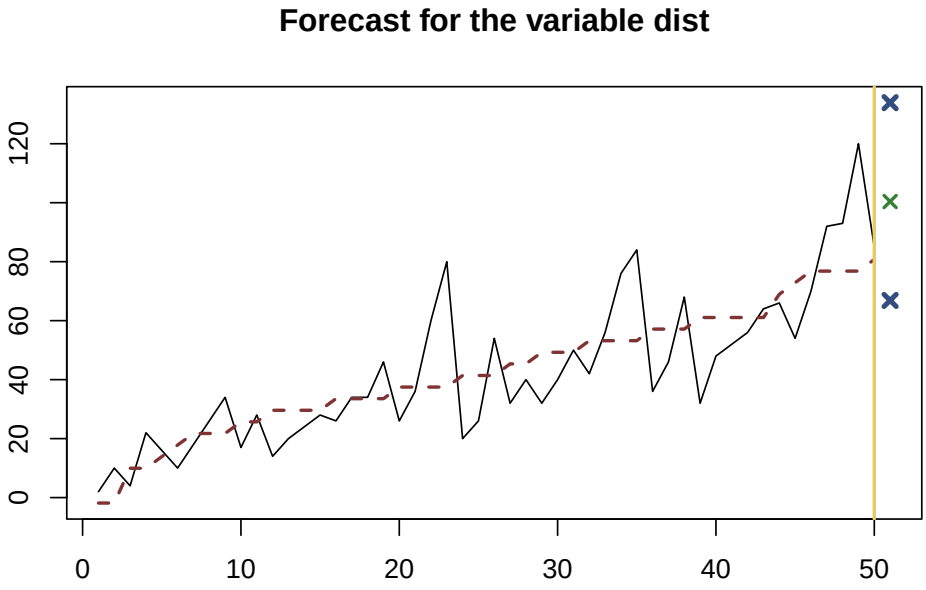


Figure 12.9: Forecast of the stopping distance for the speed of 30mph.

Chapter 13

Regression with categorical variables

So far we assumed that the explanatory variables in the model are numerical. But is it possible to include somehow in regression model variables in categorical scales, for example, colour and size of t-shirts? Yes, it is. This is done using so called "dummy variables".

13.1 Dummy variables for the intercept

As we remember from Section 1.2, the variables in categorical scale do not have distance or natural zero. This means that if we encode the values in numbers (e.g. "red" - "1", "green" - "2", "blue" - "3"), then these numbers will not have any proper mathematical meaning - they will only represent specific values (and order in case of ordinal scale), but we would be limited in operations with these values. In order to overcome this limitation, we could create a set of dummy variables, each of which would be equal to one if the value of the original variable is equal to a specific value and zero otherwise. Consider the example with colours, where we have three types of t-shirts to sell:

1. Red,
2. Green,
3. Blue.

Every t-shirt in our dataset would have one of these colours, and based on this we could create three dummy variables:

1. colourRed, which would be equal to one if the t-shirt is Red and zero otherwise,
2. colourGreen: 1 if the t-shirt is Green and 0 otherwise,
3. colourBlue: 1 if the t-shirt is Blue and 0 otherwise.

These dummy variables can then be added to a model instead of the original variable colour, resulting, for example, in the model:

$$sales_j = \beta_0 + \beta_1 price_j + \beta_2 colourRed_j + \beta_3 colourGreen_j + \epsilon_j. \quad (13.1)$$

Notice that I have only included two dummy variables out of the three. This is because we do not need to have all of them to be able to say what colour of t-shirt we have: if it is not Red and not Green, then it must be Blue. Furthermore, while some models and estimation methods could handle all the dummy variables in the model, the linear regression cannot be estimated via the conventional methods if they are all in. This is exactly because of this situation with "not Red, not Green". If we introduce all three, the model will have so called "dummy variables trap", implying perfect multicollinearity (see Subsection 15.3), because of the functional relation between variables:

$$colourBlue_j = 1 - colourRed_j - colourGreen_j \text{ for all } j = 1, \dots, n. \quad (13.2)$$

This is a general rule: if you have created a set of dummy variables from a categorical one, then one of them needs to be dropped, in order not to have the dummy variables trap.

So, what does the inclusion of dummy variables in the regression model means? We can see that on the following example of artificial data:

```
tShirts <- cbind(rnorm(150,20,2),0,0,0)
tShirts[1:50,2] <- 1
tShirts[1:50+50,3] <- 1
tShirts[1:50+50*2,4] <- 1
tShirts <- cbind(1000 + tShirts %*% c(-2.5, 30, -20, 50) + rnorm(150,0,5), tShirts)
colnames(tShirts) <- c("sales","price","colourRed","colourGreen","colourBlue")
```

We can produce spread plot to see how the data looks like:

```
spread(tShirts)
```

Figure @red(fig:tShirtsSpread) demonstrates that the sales differ depending on the type of colour (the boxplots). The scatterplot between sales and price is not very clear, but there are actually three theoretical lines on that plot. We can enlarge the plot and draw them:

```
plot(tShirts[,2:1])
abline(a=1000+30, b=-2.5, col="red")
abline(a=1000-20, b=-2.5, col="green")
abline(a=1000+50, b=-2.5, col="blue")
```

Now, if we want to construct the regression that would take these differences into account, we need to estimate the model (13.1):

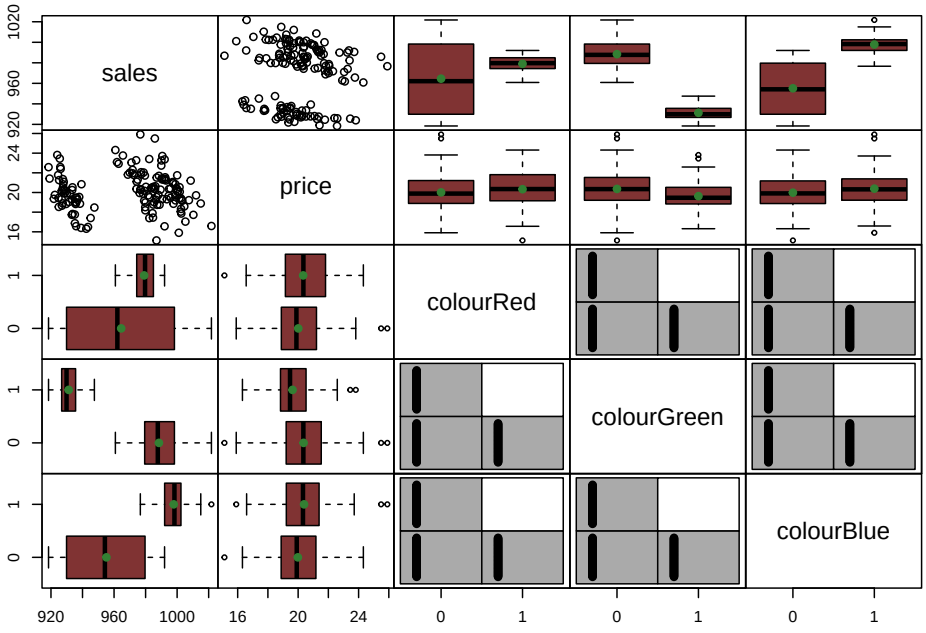


Figure 13.1: Spread plot of t-shirts data.

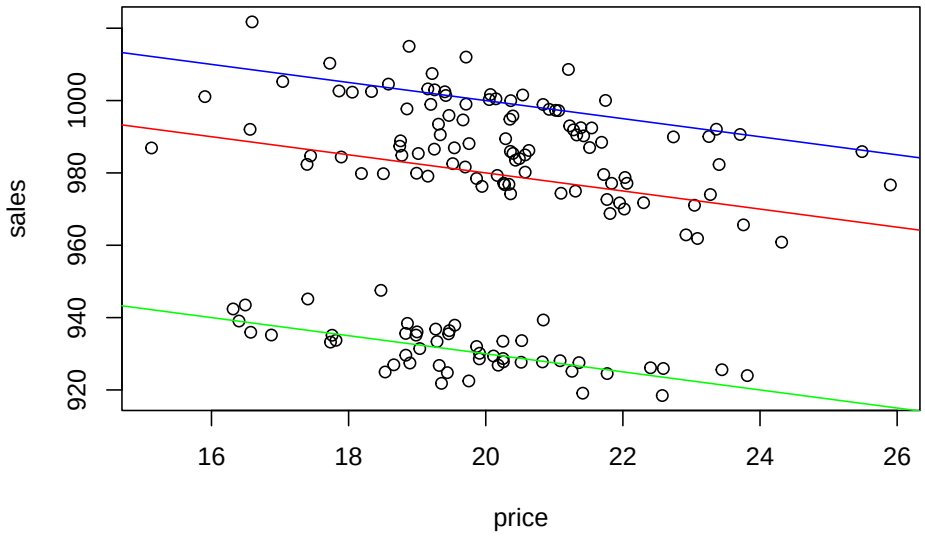


Figure 13.2: Scatterplot of Sales vs Price of t-shirts of different colour.

```
tShirtsALM <- alm(sales~price+colourRed+colourGreen, tShirts, loss="MSE")
summary(tShirtsALM)
```

```
## Response variable: sales
## Distribution used in the estimation: Normal
## Loss function used in estimation: MSE
## Coefficients:
##              Estimate Std. Error Lower 2.5% Upper 97.5%
## (Intercept) 1057.3799     4.5380  1048.4113   1066.3484 *
## price         -2.9201     0.2198    -3.3545     -2.4857 *
## colourRed    -18.9248     1.0048   -20.9107    -16.9389 *
## colourGreen  -68.8138     1.0186   -70.8270    -66.8007 *
##
## Error standard deviation: 5.0238
## Sample size: 150
## Number of estimated parameters: 4
## Number of degrees of freedom: 146
## Information criteria:
##      AIC     AICc      BIC     BICc
## 915.8849 914.1608 930.9381 926.6186
```

Notice that the intercept in this model is not 1000, as we used in the generation of the data, but is 1057. This is because it now also contains the effect of blue colour on sales in it. So, the sales of blue coloured t-shirt is now the baseline category, and each dummy variable now represents the shifts of sales, when we switch from one colour to another. For example, we can say that *the sales of red colour t-shirt are on average lower than the sales of the blue one by approximately 19 units*. What dummy variables do in the model is just shift the line from one level to another. This becomes clear if we consider special cases of models for the three t-shirts:

1. For the blue t-shirt, our model is: sales=1057.38-2.92price+et. This is because both `colourRed` and `colourGreen` are zero in this case;
2. For the red t-shirt the model is: sales=1057.38+-18.92-2.92price+et or sales=1038.46-2.92price+et;
3. Finally, for the green one, the model is: sales=1057.38+-68.81-2.92price+et or sales=988.57-2.92price+et.

In a way, we could have constructed three different regression models for the sub-samples of data, and in the ideal situation (all the data in the world) we would get the same set of estimates of parameters. However, this would be a costly procedure from the statistical perspective, because three separate models will have lower number of degrees of freedom, then the model with dummy variables. Thus, the estimates of parameters will be more uncertain in those three models than in one model `tShirtsALM`.

One thing that we can remark is that the estimated parameters differ from

the ones we used in the data generation. This is because the intercepts of the three models above intersect the y-axis in the points 1057.38, 1038.46 and 988.57 respectively. Furthermore, in general it is not possible to extract the specific effect of blue colour on sales based on the estimates of parameters, unless we impose some restrictions on parameters. The closest we can get to the true parameters is if we normalise them (assuming that there is some baseline and that the colours build upon it and add up to zero):

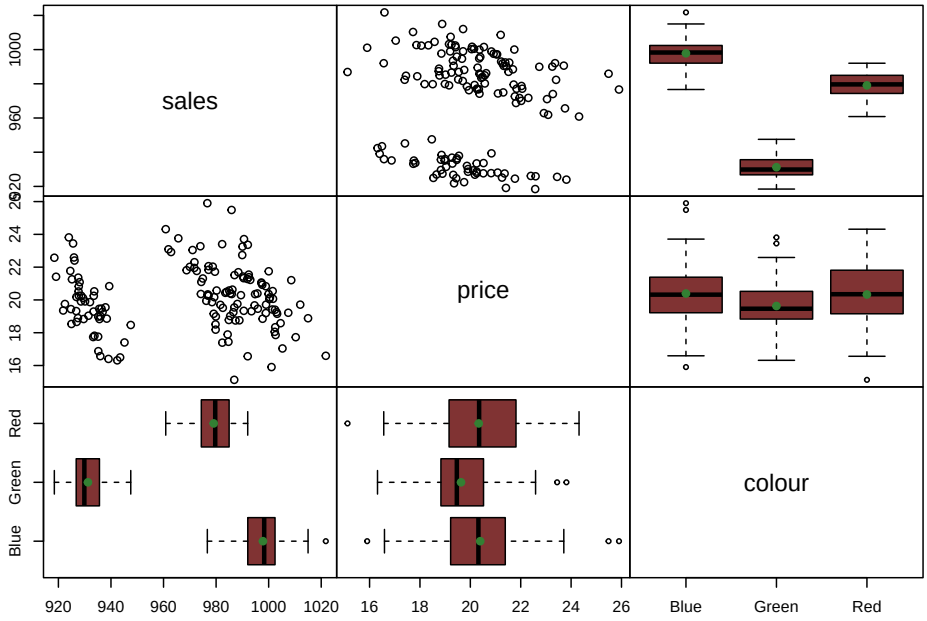
```
colourParameters <- c(coef(tShirtsALM)[3:4]+coef(tShirtsALM)[1],coef(tShirtsALM)[1])
names(colourParameters)[3] <- "colourBlue";
colourParameters - mean(colourParameters)
```

```
##   colourRed colourGreen  colourBlue
##    10.32141   -39.56763    29.24622
```

The meaning of these effects is that on average they change the baseline sales of colourless t-shirts according to these values. For example, the specific increase of sales due to the red colour of t-shirt is 10 units. In general, it is not worth bothering with these specific effects, and we can just stick with parameters of model, keeping in mind that we only have effects comparative to the selected baseline category.

In R, we can also work with factor variables, without a need to expand variables in a set of dummies - the program will do the expansion automatically and drop the first level of the variable. In order to see how it works, we create a data frame with the factor variable `colour`:

```
tShirtsDataFrame <- as.data.frame(tShirts[,1:2])
tShirtsDataFrame$colour <- factor(c("Red","Green","Blue")[tShirts[,3:5] %*% c(1:3)])
spread(tShirtsDataFrame)
```



Notice that the "Blue" was automatically set as the first level, because `factor()` function would sort labels alphabetically unless the levels are provided explicitly. The estimated model in this case will be exactly the same as the `tShirts` model above:

```
tShirtsDataFrameALM <- alm(sales~price+colour, tShirtsDataFrame, loss="MSE")
summary(tShirtsDataFrameALM)
```

```
## Response variable: sales
## Distribution used in the estimation: Normal
## Loss function used in estimation: MSE
## Coefficients:
##              Estimate Std. Error Lower 2.5% Upper 97.5%
## (Intercept) 1057.3799     4.5380  1048.4113   1066.3484 *
## price         -2.9201     0.2198    -3.3545     -2.4857 *
## colourGreen  -68.8138     1.0186   -70.8270    -66.8007 *
## colourRed    -18.9248     1.0048   -20.9107    -16.9389 *
##
## Error standard deviation: 5.0238
## Sample size: 150
## Number of estimated parameters: 4
## Number of degrees of freedom: 146
## Information criteria:
##      AIC     AICc      BIC     BICc
## 915.8849 914.1608 930.9381 926.6186
```

Finally, it is recommended in general not to drop dummy variables one by one,

if for some reason you decide that some of them are not helping. If, for example, we decide not to include `colourRed` and only have the model with `colourGreen`, then the meaning of the dummy variables will change - we will not be able to distinguish the Blue from Red. Furthermore, while some dummy variables might not seem important (or significant) in regression, their combination might be improving the model, and dropping some of them might be damaging for the model in terms of its predictive power. So, it is more common either to include all levels (but one) of categorical variable or not to include any of them.

13.2 Categorical variables for the slope

In reality, we can have more complicated situations, when the change of price would lead to different changes in sales for different types of t-shirts. In this case, we are talking about an **interaction effect** between price and colour. The following artificial example demonstrates the situation:

```
tShirtsInteraction <- cbind(rnorm(150,20,2),0,0,0)
tShirtsInteraction[1:50,2] <- tShirtsInteraction[1:50,1]
tShirtsInteraction[1:50+50,3] <- tShirtsInteraction[1:50+50,1]
tShirtsInteraction[1:50+50*2,4] <- tShirtsInteraction[1:50+50*2,1]
tShirtsInteraction <- cbind(1000 + tShirtsInteraction %*% c(-2.5, -1.5, -0.5, -4) +
                            rnorm(150,0,5), tShirtsInteraction)
colnames(tShirtsInteraction) <- c("sales","price","price:colourRed",
                                  "price:colourGreen","price:colourBlue")
```

This artificial data can be plotted in the following way to show the effect:

```
plot(tShirtsInteraction[,2:1])
abline(a=1000, b=-2.5-1.5, col="red")
abline(a=1000, b=-2.5-0.5, col="green")
abline(a=1000, b=-2.5-4, col="blue")
```

The plot on Figure 13.3 shows that there are three categories of data and that for each of it, the price effect will be different: the increase in price by one unit leads to the faster reduction of sales for the blue t-shirts than for the others. Compare this with Figure 13.2, where we had the difference only in intercepts. This implies a different model:

$$sales_j = \beta_0 + \beta_1 price_j + \beta_2 price_j \times colourRed_j + \beta_3 price_j \times colourGreen_j + \epsilon_j. \tag{13.3}$$

Notice that we still include only two dummy variables out of three in order to avoid the dummy variables trap. What is new in this case is the multiplication of price by the dummy variables. This trick allows changing the slope of price, depending on the colour of t-shirt. For example, here what the model (13.3) would look like for the three types of colours:

1. Red colour: $sales_j = \beta_0 + \beta_1 price_j + \beta_2 price_j + \epsilon_j$ or $sales_j = \beta_0 + (\beta_1 + \beta_2) price_j + \epsilon_j$;

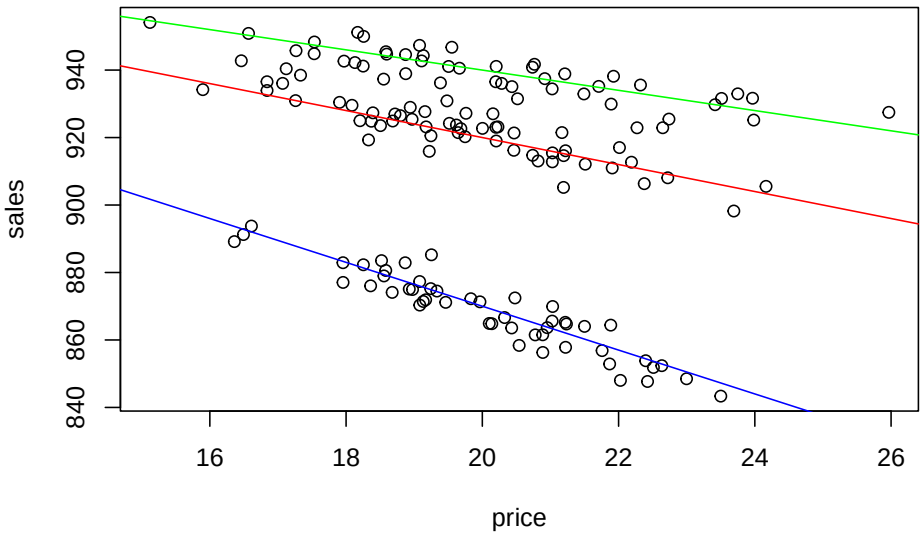


Figure 13.3: Scatterplot of Sales vs Price of t-shirts of different colour, interaction effect.

2. Green colour: $sales_j = \beta_0 + \beta_1 price_j + \beta_3 price_j + \epsilon_j$ or $sales_j = \beta_0 + (\beta_1 + \beta_3) price_j + \epsilon_j$;
3. Blue colour: $sales_j = \beta_0 + \beta_1 price_j + \epsilon_j$.

In R, the interaction effect can be introduced explicitly in the formula via : symbol if you have a proper factor variable:

```
tShirtsInteractionDataFrame <- as.data.frame(tShirtsInteraction[,1:2])
tShirtsInteractionDataFrame$colour <- tShirtsDataFrame$colour
# Fit the model
tShirtsInteractionDataFrameALM <- alm(sales~price+price:colour,
                                      tShirtsInteractionDataFrame, loss="MSE")
summary(tShirtsInteractionDataFrameALM)
```

```
## Response variable: sales
## Distribution used in the estimation: Normal
## Loss function used in estimation: MSE
## Coefficients:
##                   Estimate Std. Error Lower 2.5% Upper 97.5%
## (Intercept)       997.8761     3.5949   990.7713   1004.9809 *
## price              -6.4493     0.1803    -6.8057     -6.0929 *
## price:colourGreen   3.4919     0.0418     3.4094      3.5745 *
## price:colourRed     2.5696     0.0421     2.4865      2.6528 *
##
## Error standard deviation: 4.207
## Sample size: 150
```

```
## Number of estimated parameters: 4
## Number of degrees of freedom: 146
## Information criteria:
##      AIC     AICc      BIC     BICc
## 862.6541 860.9300 877.7073 873.3878
```

Note that the interpretation of parameters in such model will be different, because now the `price` shows the baseline effect for the blue t-shirts, while the interaction effects show how this effect will change for other colours. So, for example, in order to see what would be the effect of price change on sales of red t-shirts, we need to sum up the parameter for `price` and `price:colourRed`. We then can say that if price of red t-shirt increases by £1, the sales will decrease on average by 3.88 units.

Chapter 14

Variables transformations

So far we have discussed linear regression models, where the response variable linearly depends on a set of explanatory variables. These models work well in many contexts, especially when the response variable is measured in high volumes (e.g. sales in thousands of units). However, in reality the relations between variables can be non-linear. In this chapter we consider an example of application to see how transformations can be motivated by a real life example and then discuss different types of transformations and what they imply for analytics and forecasting

14.1 Example of application

Consider, for example, the stopping distance vs speed of the car, the case we have discussed in the previous sections. This sort of relation in reality is non-linear. We know from physics that the distance travelled by car is proportional to the mass of car, the squared speed and inversely proportional to the breaking force:

$$distance \propto \frac{mass}{2breaking} \times speed^2. \tag{14.1}$$

If we use the linear function instead, then we might fail in capturing the relation correctly. Here is how the linear regression looks like, when applied to the data (Figure 14.1).

The model on the plot in Figure 14.1 is misleading, because it predicts that the stopping distance of a car, travelling with speed less than 4mph will be negative. Furthermore, the modelunderestimates the real stopping distance for cars with higher speed. If a decision is made based on this model, then it will be inevitably wrong and might potentially lead to serious repercussions in terms of road safety. Given the relation (14.1), we should consider a non-linear model. In this specific

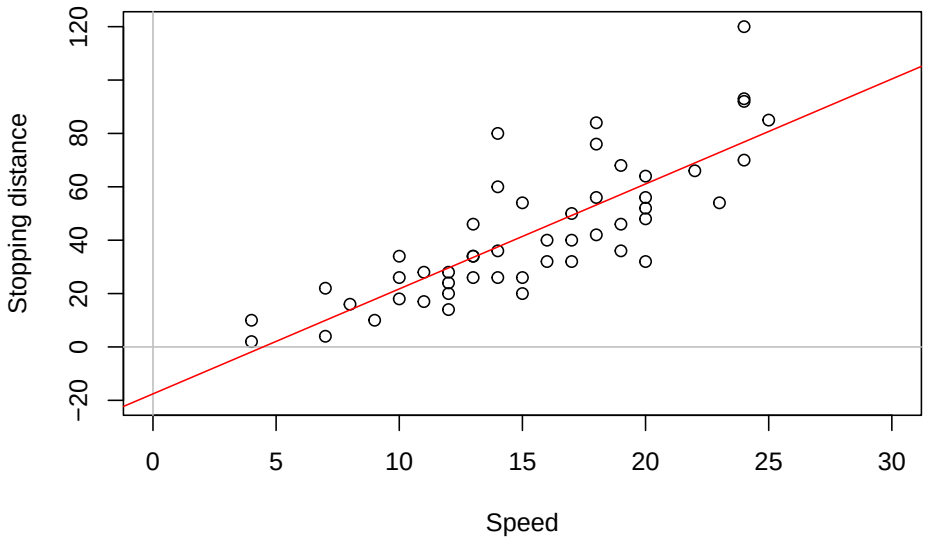


Figure 14.1: Speed vs stopping distance and a linear model

case, we should consider the model of the type:

$$distance = \beta_0 speed^{\beta_1} \times (1 + \epsilon). \tag{14.2}$$

The multiplication of speed by the error term is necessary, because the effect of randomness will have an increasing variability with the increase of speed: if the speed is low, then the random factors (such as road conditions, breaks condition etc) will not have a strong effect on distance, while in case of the high speed these random factors might lead either to the serious decrease or increase of distance (a car on a slippery road, stopping from 50mph will have much longer distance than the same car on a dry road). Note that I have left the parameter β_1 in **??**eq:speedDistanceModel) and did not set it equal to two. This is done for the case we want to estimate the parameter based on the data. The problem with the model (14.2) is that it is difficult to estimate due to the non-linearity. In order to resolve this problem, we can linearise it by taking logarithms of both sides, which will lead to:

$$\log(distance) = \log \beta_0 + \beta_1 \log(speed) + \log(1 + \epsilon). \tag{14.3}$$

If we substituted every element with log in (14.3) by other names (e.g. $\log(\beta_0) = \beta_0'$ and $\log(speed) = x$), it would be easier to see that this is a linear model, which can be estimated via OLS. This type of model is called "log-log", reflecting that it has logarithms on both sides. Even the data will be much better behaved if we use logarithms in this situation (see Figure 14.2).

What we want to see on Figure 14.2 is the linear relation between the variables with points having fixed variance. However, in our case we can notice that the

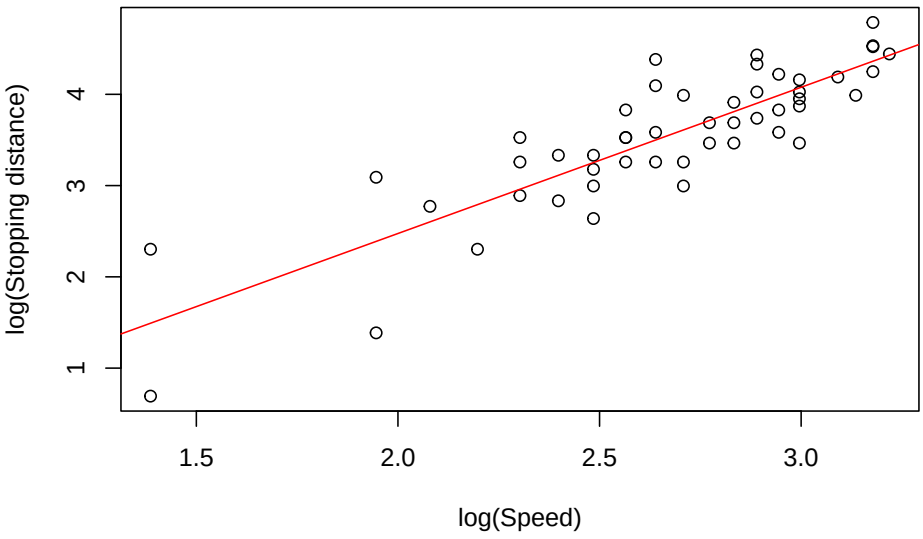


Figure 14.2: Speed vs stopping distance in logarithms

variance of the stopping distances does not seem to be stable: the variability around 2.0 is higher than the variability around 3.0. This might cause issues in the model due to violation of assumptions (see Section 15). For now, we acknowledge the issue but do not aim to fix it. And here how the model (14.3) can be estimated using R:

```
slmSpeedDistanceModel01 <- alm(log(dist)~log(speed), cars, loss="MSE")
```

The values of parameters of this model will have a different meaning than the parameters of the linear model. Consider the example with the model above:

```
summary(slmSpeedDistanceModel01)
```

```
## Response variable: logdist
## Distribution used in the estimation: Normal
## Loss function used in estimation: MSE
## Coefficients:
##             Estimate Std. Error Lower 2.5% Upper 97.5%
## (Intercept)  -0.7297     0.3758    -1.4854       0.026
## log(speed)    1.6024     0.1395     1.3218       1.883 *
##
## Error standard deviation: 0.4053
## Sample size: 50
## Number of estimated parameters: 2
## Number of degrees of freedom: 48
## Information criteria:
##     AIC    AICc     BIC    BICc
```

```
## 55.5318 53.7872 61.2679 57.8553
```

The value of parameter for the variable `log(speed)` now does not represent the marginal effect of speed on distance, but rather shows the elasticity, i.e. if the speed of a car increases by 1%, the travel distance will increase on average by 1.6%.

In order to analyse the fit of the model on the original data, we would need to produce fitted values and exponentiate them. Note that in this case they would correspond to geometric rather than arithmetic means:

```
plot(cars, xlab="Speed", ylab="Stopping distance")
lines(cars$speed,exp(fitted(slmSpeedDistanceModel01)),col="red")
```

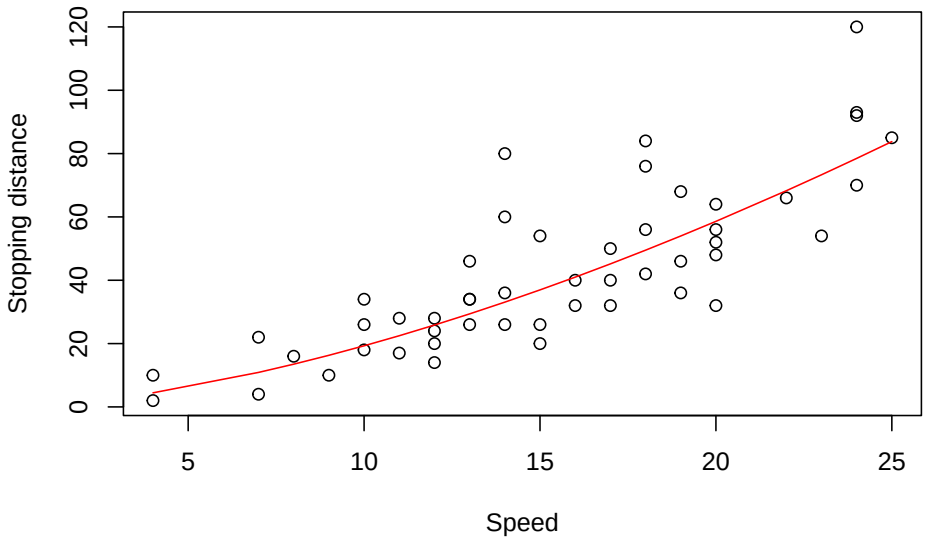


Figure 14.3: Speed vs stopping distance and the log-log model fit.

The resulting model in Figure 14.3 is the power function, which exhibits the increase in speed of change of one parameter with a linear change of another one. Note that technically speaking, the log-log model only makes sense, when the data is strictly positive. If it also contains zeroes (the speed is zero, thus the stopping distance is zero), then some other transformations might be in order. For example, we could square the speed in the model and try constructing the linear model, aligning it better with the physical model (14.1):

$$distance = \beta_0 + \beta_1 speed^2 + \epsilon. \tag{14.4}$$

The issue of this model would be that the error term is additive and thus the model would assume that the variability of the error does not change with the speed, which is not realistic.

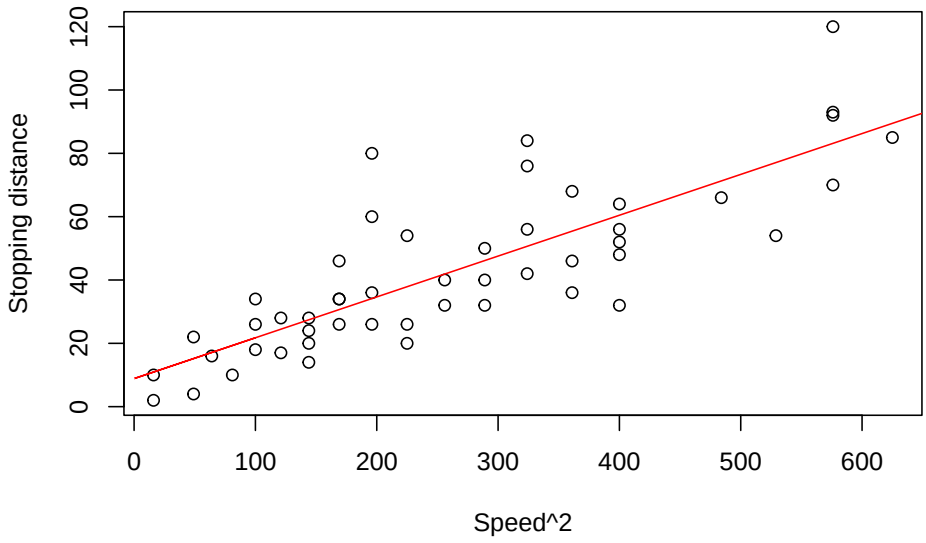


Figure 14.4: Speed squared vs stopping distance.

Figure 14.4 demonstrates the scatterplot for squared speed vs stopping distances. While we see that the relation between variables is closer to linear, the problem with variance is not resolved. If we want to estimate this model, we can use the following command in R:

```
slmSpeedDistanceModel02 <- alm(dist~I(speed^2), cars, loss="MSE")
```

Note that we use `I()` in the formula to tell R to square the variable - it will not do the necessary transformation otherwise. Also note that in our specific case we did not include the non-transformed speed variable, because we know that the lowest distance should be, when speed is zero. But this might not be the case in other cases, so in general instead of the formula used above we should use: `y~x+I(x^2)`. Furthermore, if we know for sure that the intercept is not needed (i.e. we know that the distance will be zero, when speed is zero), then we can remove it and estimate the model:

```
slmSpeedDistanceModel03 <- alm(dist~I(speed^2)-1, cars, loss="MSE")
```

```
## Warning: You have asked not to include intercept in the model. We will try to
## fit the model, but this is a very naughty thing to do, and we cannot guarantee
## that it will work...
```

`alm()` function will complain about the exclusion of the intercept, but it should estimate the model nonetheless. The fit of the model to the data would be similar in its shape to the one from the log-log model (see Figure 14.5).

The plot in Figure 14.5 demonstrates how the two models fit the data. The Model 2, as we see goes through the origin, which makes sense from the physical

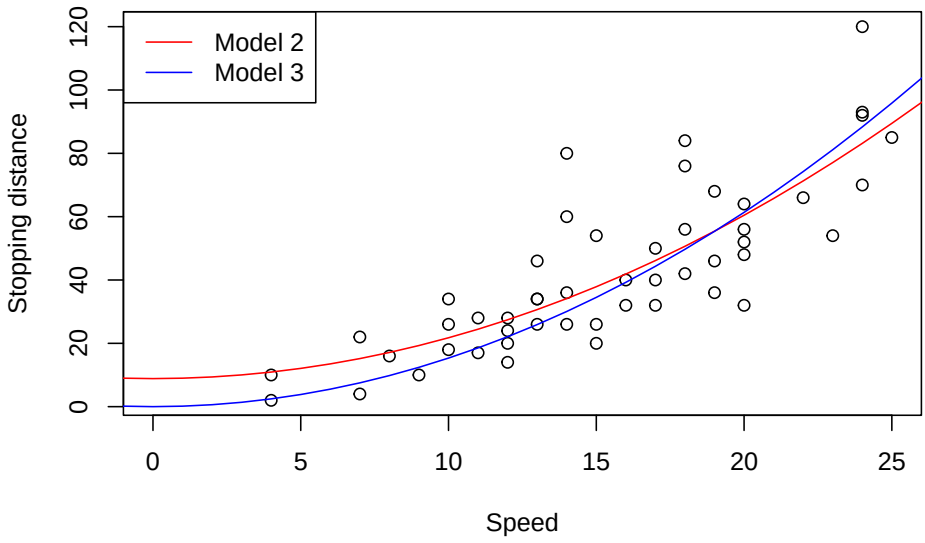


Figure 14.5: Speed squared vs stopping distance with models with speed^2.

point of view. However, because of that it might fit the data worse than the Model 1 does. Still, it it better to have a more meaningful model than the one that potentially overfits the data.

Another way to introduce the squares in the model is to take square root of distance. This would potentially align better with the physical model of stopping distance (14.1):

$$\sqrt{distance} = \beta_0 + \beta_1 speed + \epsilon, \tag{14.5}$$

which will be equivalent to:

$$distance = (\beta_0 + \beta_1 speed + \epsilon)^2. \tag{14.6}$$

The good news is, the error term in this model will change with the change of speed due to the interaction effect, cause by the square of the sum in (14.6). And, similar to the previous models, the parameter β_0 might not be needed. Graphically, this transformation is present on Figure 14.6.

As the plot in Figure 14.6 demonstrates, the relation has become linear and the variance seems to be constant, no matter what the speed is. This means that the proposed model might be more appropriate to the data than the previous ones. This is how we can estimate this model:

```
slmSpeedDistanceModel04 <- alm(sqrt(dist)~speed, cars, loss="MSE")
```

Similar to the Model 2 with squares, we will also consider the model without intercept on the grounds that if we capture the relation correctly, the zero speed should result in zero distance.

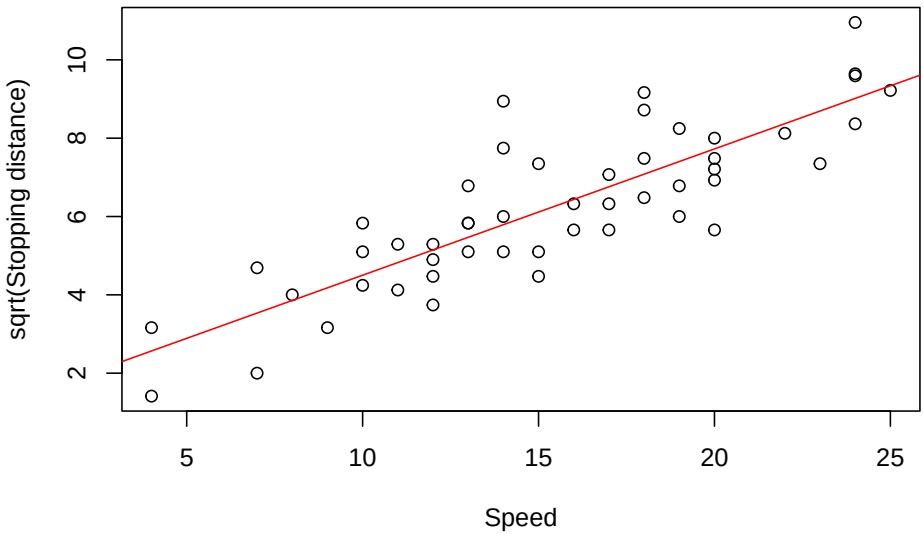
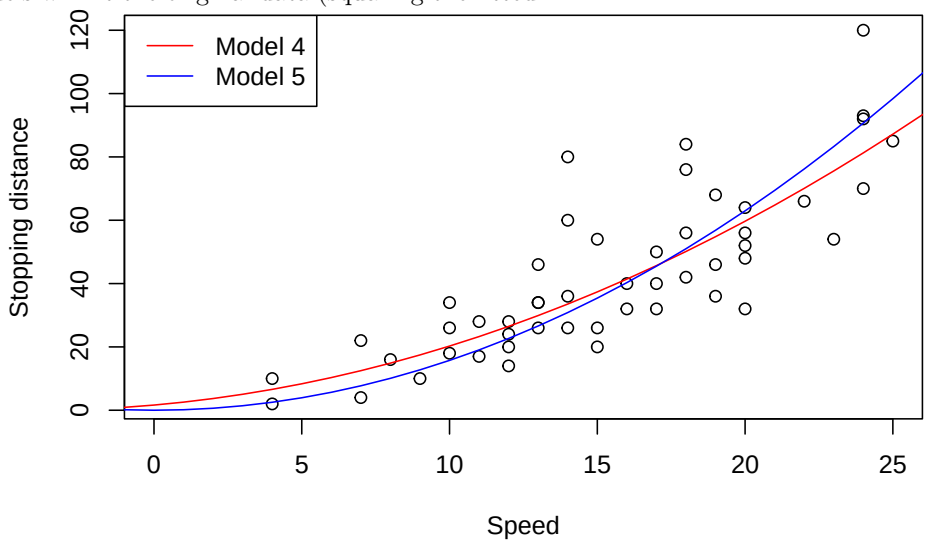


Figure 14.6: Speed vs square root of stopping distance.

```
slmSpeedDistanceModel05 <- alm(sqrt(dist)~speed-1, cars, loss="MSE")
```

```
## Warning: You have asked not to include intercept in the model. We will try to
## fit the model, but this is a very naughty thing to do, and we cannot guarantee
## that it will work...
```

Finally, we can see how both models will fit the original data (squaring the fitted values to get to the original scale):



Subjectively, I would say that Model 5 is the most appropriate from all the

models under consideration: it corresponds to the physical model on one hand, and has constant variance on the other one. Here is its summary:

```
summary(slmSpeedDistanceModel05)
```

```
## Response variable: sqrtdist
## Distribution used in the estimation: Normal
## Loss function used in estimation: MSE
## Coefficients:
##       Estimate Std. Error Lower 2.5% Upper 97.5%
## speed   0.3967     0.0102     0.3764      0.4171 *
##
## Error standard deviation: 1.1674
## Sample size: 50
## Number of estimated parameters: 1
## Number of degrees of freedom: 49
## Information criteria:
##      AIC     AICc      BIC     BICc
## 160.3623 158.4456 164.1863 160.4373
```

Its parameter contains some average information about the mass of cars and their breaking forces (this is based on the formula (14.1)). The interpretation of the parameter in this model, however, is challenging. In order to get to some crude interpretation, we need to revert to maths. Model 5 can be written as:

$$distance = (\beta_1 speed + \epsilon)^2. \tag{14.7}$$

If we take the first derivative of distance with respect to speed, we will get:

$$\frac{\mathrm{d}distance}{\mathrm{d}speed} = 2(\beta_1 speed + \epsilon), \tag{14.8}$$

which is now closer to what we need. We can say that if speed increases by 1mph, the distance will change on average by $2\beta_1 speed$. But this does not explain what the meaning of β_1 in the model is. So we take the second derivative with respect to speed:

$$\frac{\mathrm{d}^2 distance}{\mathrm{d}^2 speed} = 2\beta_1. \tag{14.9}$$

The meaning of the second derivative is that it shows the change of change of distance with a change of change of speed by 1. This implies a tricky interpretation of the parameter. Based on the summary above, the only thing we can conclude is that when the change of speed increases by 1mph, the change of distance will increase by 0.7934 feet. An alternative interpretation would be based on the model (14.5): with the increase of speed of car by 1mph, the square roo tof stopping distance would increase by 0.3967 square root feet. Neither of these two interpretations are very helpful, but this is the best we have for the parameter β_1 in the Model 5.

14.2 Types of variables transformations

Having considered this case study, we can summarise the possible types of transformations of variables in regression models and what they would mean. Here, we only discuss monotonic transformations, i.e. those that guarantee that if x was increasing before transformations, it would be increasing after transformations as well.

14.2.1 Linear model

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{14.10}$$

As discussed earlier, in this model, β_1 can be interpreted as a marginal effect of x on y. The typical interpretation is that with the increase of x by 1 unit, y will change on average by β_1 units. In case of dummy variables, their interpretation is that the specific category of product will have a different (higher or lower) impact on y by β_1 units. e.g. "sales of red mobile phones are on average higher than the sales of the blue ones by 100 units".

14.2.2 Log-Log model

Or power model or a multiplicative model:

$$\log y = \beta_0 + \beta_1 \log x + \log(1 + \epsilon). \tag{14.11}$$

It is equivalent to

$$y = \beta_0 x^{\beta_1}(1 + \epsilon). \tag{14.12}$$

The parameter β_1 is interpreted as elasticity: If x increases by 1%, the response variable y changes on average by β_1%. Depending on the value of β_1, this model can capture non-linear relations with slowing down or accelerating changes. Figure 14.7 demonstrates several examples of artificial data with different values of β_1.

As discussed earlier, this model can only be applied to positive data. If there are zeroes in the data, then logarithm will be equal to $-\infty$ and it would not be possible to estimate the model correctly.

14.2.3 Log-linear model

Or exponential model:

$$\log y = \beta_0 + \beta_1 x + \log(1 + \epsilon). \tag{14.13}$$

is equivalent to

$$y = \beta_0 \exp(\beta_1 x)(1 + \epsilon). \tag{14.14}$$

The parameter β_1 will control the change of speed of growth / decline in the model. If variable x increases by 1 unit, then the variable y will change on average

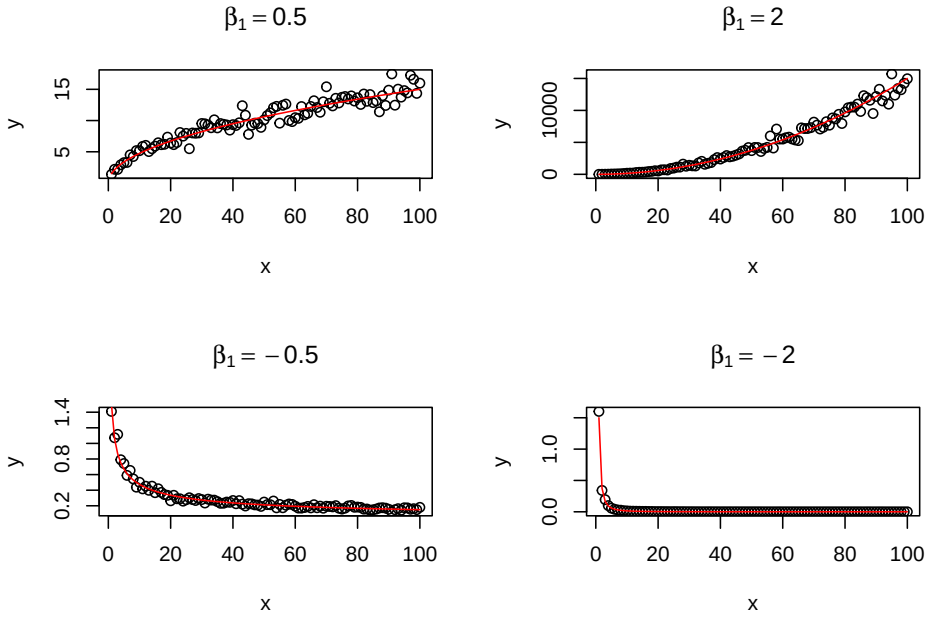
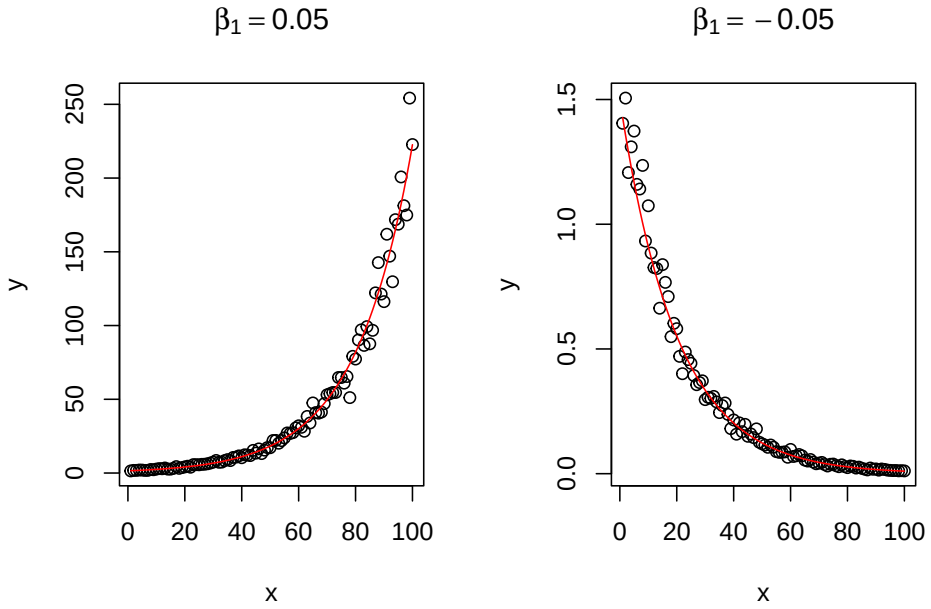


Figure 14.7: Examples of log-log relations with different values of elasticity parameter.

by $(\exp(\beta_1) - 1) \times 100\%$. If the value of β_1 is small (roughly $\beta_1 \in (-0.2, 0.2)$), then due to one of the limits the interpretation can be simplified to: when x increases by 1 unit, the variable y will change on average by $\beta_1 \times 100\%$. The exponent is in general a dangerous function as it exhibits either explosive (when $\beta_1 > 0$) or implosive (when $\beta_1 < 0$) behaviour. This is shown in Figure **??**, where the values of β_1 are -0.05 and 0.05, and we can see how fast the value of y changes with the increase of x.



If x is a dummy variable (see Section 13), then its interpretation is slightly different: the presence of the effect x leads on average to the change of variable y by $\beta_1 \times 100\%$. e.g. "sales of red laptops are on average 15% higher than sales of blue laptops".

14.2.4 Linear-Log model

Or logarithmic model.

$$y = \beta_0 + \beta_1 \log x + \epsilon. \tag{14.15}$$

This is just a logarithmic transform of explanatory variable. The parameter β_1 in this case regulates the direction and speed of change. If x increases by 1%, then y will change on average by $\frac{\beta_1}{100}$ units. Figure 14.8 shows two cases of relations with positive and negative slope parameters.

The logarithmic model assumes that the increase in x always leads on average to the slow down of the value of y.

14.2.5 Square root model

$$y = \beta_0 + \beta_1 \sqrt{x} + \epsilon. \tag{14.16}$$

The relation between y and x in this model looks similar to the on in linear-log model, but the with a lower speed of change: the square root represents the slow down in the change and might be suitable for cases of diminishing returns of scale in various real life problems. There is no specific interpretation for the parameter β_1 in this model - it will show how the response variable y will change on average wih increase of square root of x by one. Figure 14.9 demonstrates square root relations for two cases, with parameters $\beta_1 = 1.5$ and $\beta_1 = -1.5$.

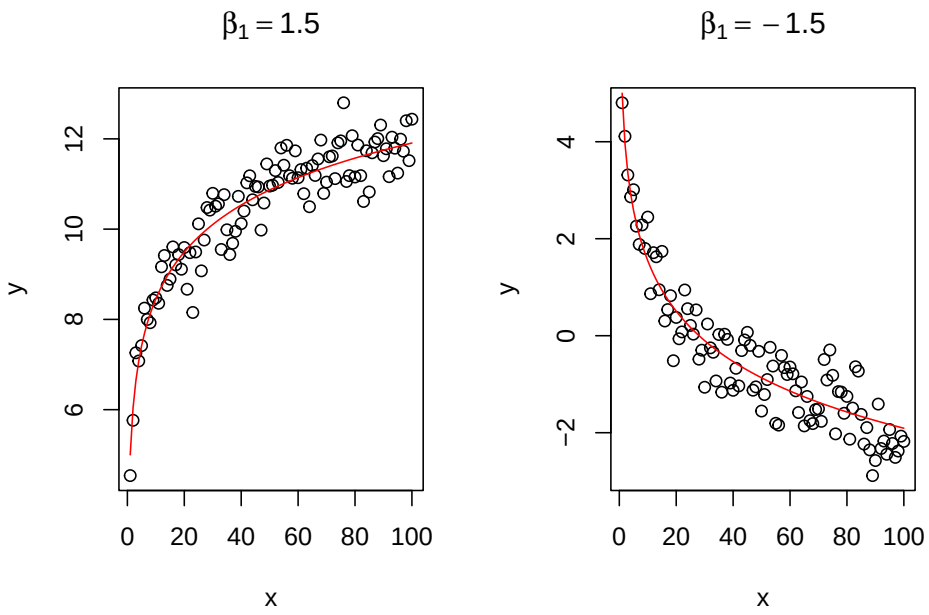


Figure 14.8: Examples of linear-log relations with two values of slope parameter.

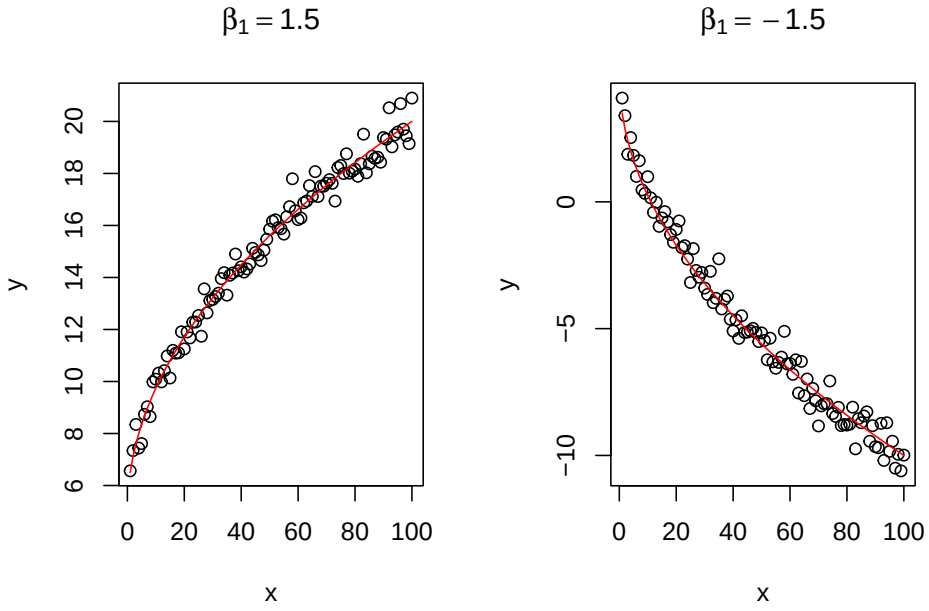


Figure 14.9: Examples of linear - square root relations with two values of slope parameter.

14.2.6 Quadratic model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon. \tag{14.17}$$

This relation demonstrates increase or decrease with an acceleration due to the present of squared x. This model has an extremum (either a minimum or a maximum), when $x = \frac{-\beta_1}{2\beta_2}$. This means that the growth in the data will be changed by decline or vice versa with the increase of x. This makes the model potentially prone to overfitting, so it needs to be used with care. Note that in general the quadratic equation should include both x and x^2, unless we know that the extremum should be at the point $x = 0$ (see the example with Model 5 in the previous section). Furthrmore, this model is close to the one with square root of y: $\sqrt{y} = \beta_0 + \beta_1 x + \epsilon$, with the main difference being that the latter formulation assumes that the variability of the error term will change together with the change of x (so called "heteroscedasticity" effect, see Section 15.2). This model was used in the examples with stopping distance above. Figure 14.10 shows to classical examples: with branches of the function going down and going up.

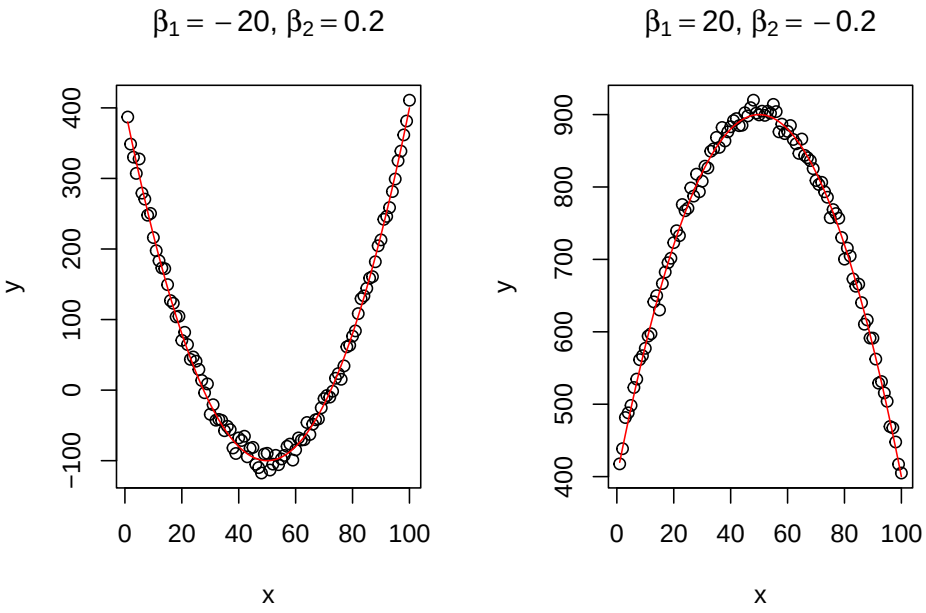


Figure 14.10: Examples of linear-log relations with two values of slope parameter.

14.2.7 Polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots \ \beta_k x^k + \epsilon. \tag{14.18}$$

This is a more general model than the quadratic one, introducing k polynomials. This is not used very often in analytics, because any data can be approximated by

a high order polynomial, and because the branches of polynomial will inevitably lead to infinite increase / decrease, which is not a common tendency in practice.

14.2.8 Box-Cox transform

Or power transform:

$$\frac{y^\lambda - 1}{\lambda} = \beta_0 + \beta_1 x + \epsilon. \tag{14.19}$$

This type of transform can be applied to either response variable or any of explanatory variables and can be considered as something more general than linear, log-linear, quadratic and square root models. This is because with different values of λ, the transformation would revert to one of the above. For example, with $\lambda = 1$, we end up with a linear model, just with a different intercept. If $\lambda = 0.5$, then we end up with square root, and when $\lambda \to 0$, then the relation becomes equivalent to logarithmic. The choice of λ might be a challenging task on its own, however it can be estimated via likelihood. If estimated and close to either 0, 0.5, 1 or 2, then typically a respective transformation should be applied instead of Box-Cox. For example, if $\lambda = 0.49$, then taking square root might be a preferred option.

14.2.9 Logistic transform

In some cases, the variable of interest might lie in a specific region, for example between 0 and 100. In that case a non-linear transform is required to change the range to the conventional one $(-\infty, \infty)$ used in the classical regression. The logistic transform is supposed to do that. Assuming that $y \in (0, 1)$ the model based on it can be written as:

$$y = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x + \epsilon)\right)}. \tag{14.20}$$

Remark. If y lies in a different fixed range, then a scaling can be applied to it to make it lie between zero and one. For example, if it lies between 0 and 100, division by 100 will fix the scale.

The inverse logistic transform might also be useful and allows estimating the model using the conventional methods after transforming the response variable:

$$\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 x + \epsilon. \tag{14.21}$$

The logistic function is used in models with binary response variable and is also one of the conventional functions used in more advanced machine learning techniques (e.g. Artificial Neural Networks). Figure **??** demonstrates how the response variable might look in the case of the model (14.20).

Sometimes the value of y in case of logistic model is interpreted as a probability of outcome. We will discuss models based on logistic function later in this textbook.

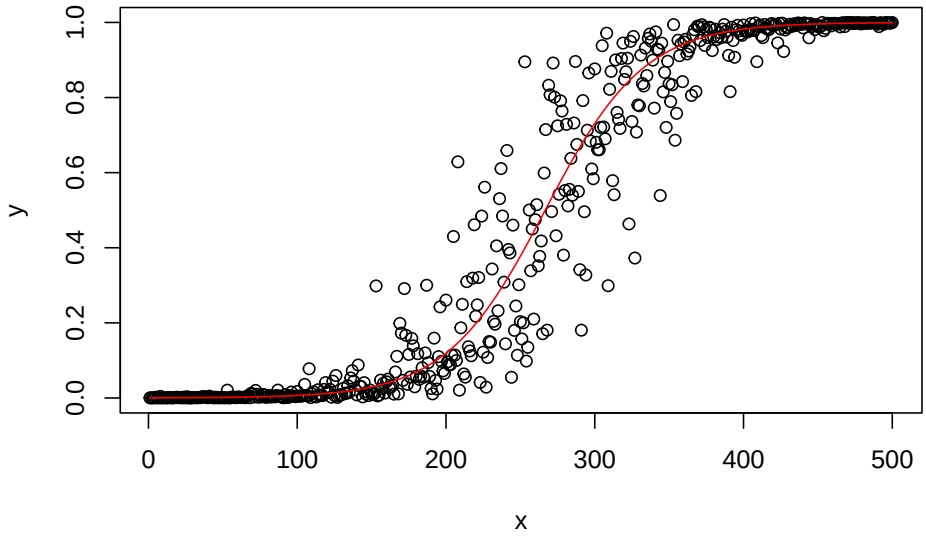


Figure 14.11: Examples of linear-log relations with two values of slope parameter.

14.2.10 Summary

In this subsection we discussed the basic types of variables transformations on examples with simple linear regression. The more complicated models with multiple explanatory variables and complex transformations can be considered as well. However, whatever transformation is considered, it needs to be meaningful and come from the theory, not from the data. Otherwise we may overfit the data, which will lead to a variety of issues, some of which are discussed in Section 15.1.

Chapter 15

Statistical models assumptions

In order for a statistical model to work adequately and not to fail, when applied to a data, several assumptions about it should hold. If they do not, then the model might lead to biased or inefficient estimates of parameters and inaccurate forecasts. In this section we discuss the main assumptions, united in three big groups:

1. Model is correctly specified;
2. Residuals are independent and identically distributed (i.i.d.);
3. The explanatory variables are not correlated with anything but the response variable.

We do not aim to explain why the violation of assumptions would lead to the discussed problem, and refer a curious reader to econometrics textbooks (for example Hanck et al., 2022). In many cases, in our discussions in this textbook, we assume that all of these assumptions hold. In some of the cases, we will say explicitly, which are violated and what needs to be done in those situations.

15.1 Model is correctly specified

This is one of the fundamental group of assumptions, which can be summarised as "we have included everything necessary in the model in the correct form". It implies that:

1. We have not omitted important variables in the model (underfitting the data);
2. We do not have redundant variables in the model (overfitting the data);
3. The necessary transformations of the variables are applied;
4. We do not have outliers in the residuals of the model.

15.1.1 Omitted variables

If there are some important variables that we did not include in the model, then the estimates of the parameters might be *biased* and in some cases quite seriously (e.g. positive sign instead of the negative one). A classical example of model with omitted important variables is simple linear regression, which by definition includes only one explanatory variable. Making decisions based on such model might not be wise, as it might mislead about the significance and sign of effects. Yes, we use simple linear regression for educational purposes, to understand how the model works and what it implies, but it is not sufficient on its own. Finally, when it comes to forecasting, omitting important variables is equivalent to underfitting the data, ignoring significant aspects of the model. This means that the point forecasts from the model might be *biased* (systematic under or over forecasting), the variance of the error term will be higher than needed, which will result in wider than necessary prediction interval.

In some cases, it is possible to diagnose the violation of this assumption. In order to do that an analyst needs to analyse a variety of plots of residuals vs fitted, vs time (if we deal with time series), and vs omitted variables. Consider an example with `mtcars` data and a simple linear regression:

```
mtcarsSLR <- alm(mpg~wt, mtcars, loss="MSE")
```

Based on the preliminary analysis that we have conducted in Sections 5 and 9, this model omits important variables. And there are several basic plots that might allow us diagnosing the violation of this assumption.

```
par(mfcol=c(1,2))
plot(mtcarsSLR,c(1,2))
```

Figure 15.1 demonstrates actuals vs fitted and fitted vs standardised residuals. The standardised residuals are the residuals from the model that are divided by their standard deviation, thus removing the scale. What we want to see on the first plot in Figure 15.1, is for all the point lie around the grey line and for the LOWESS line to coincide with the grey line. That would mean that the relations are captured correctly and all the observations are explained by the model. As for the second plot, we want to see the same, but it just presents that information in a different format, which is sometimes easier to analyse. In both plot of Figure 15.1, we can see that there are still some patterns left: the LOWESS line has a u-shaped form, which in general means that something is wrong with model specification. In order to investigate if there are any omitted variables, we construct a spread plot of residuals vs all the variables not included in the model (Figure 15.2).

```
spread(data.frame(residuals=resid(mtcarsSLR), mtcars[,-c(1,6)]))
```

What we want to see in Figure 15.2 is the absence of any patterns in plots of residuals vs variables. However, we can see that there are still many relations. For example, with the increase of the number of cylinders, the mean of residuals

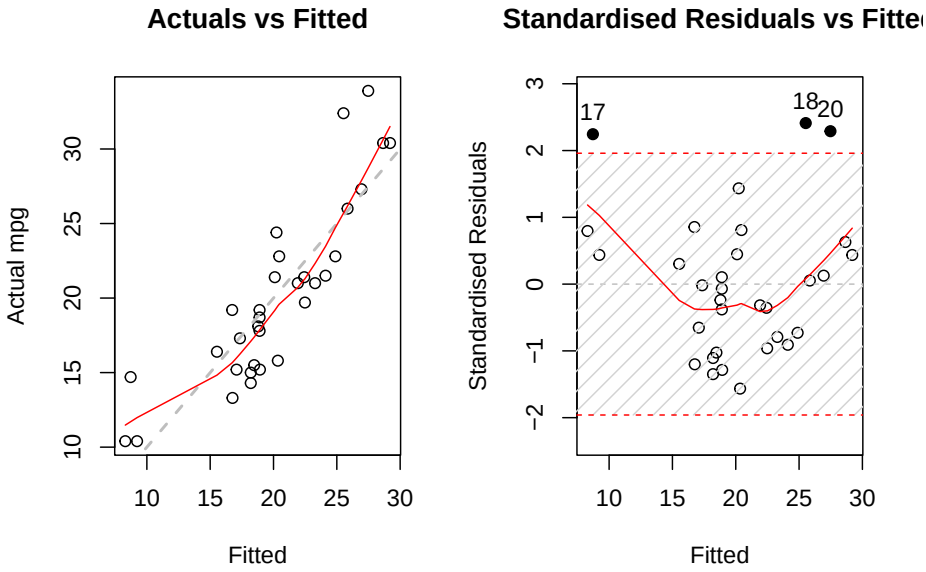


Figure 15.1: Diagnostics of omitted variables.

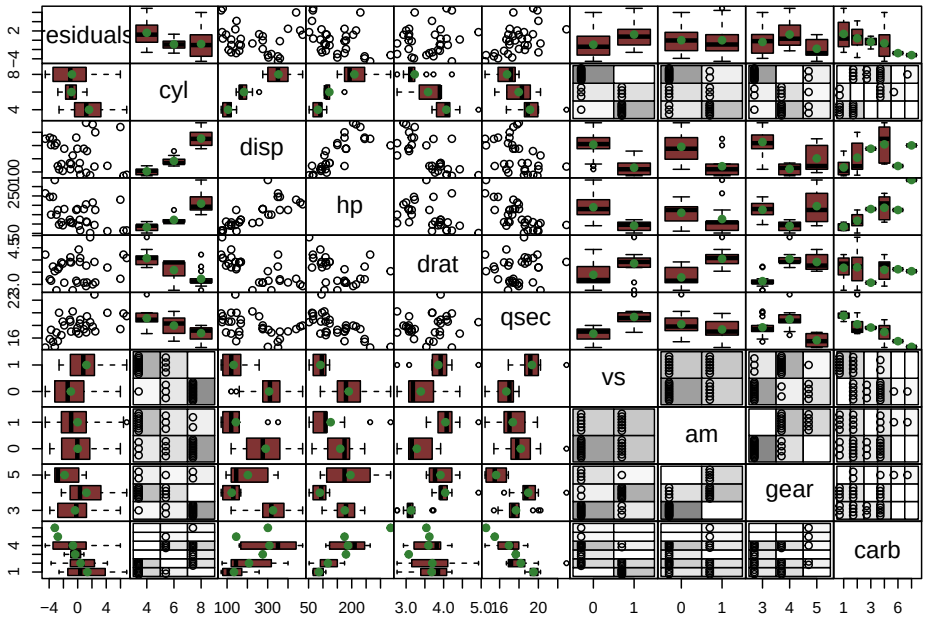


Figure 15.2: Diagnostics of omitted variables.

decreases. This might indicate that the variable is needed in the model. And indeed, we can imagine a situation, where mileage of a car (the response variable in our model) would depend on the number of cylinders because the bigger engines will have more cylinders and consume more fuel, so it makes sense to include this variable in the model as well.

Note that we do not suggest to start modelling from simple linear relation! You should construct a model that you think is suitable for the problem, and the example above is provided only for illustrative purposes.

15.1.2 Redundant variables

If there are redundant variables that are not needed in the model, then the estimates of parameters and point forecasts might be *unbiased*, but *inefficient*. This implies that the variance of parameters can be lower than needed and thus the prediction intervals will be narrower than needed. There are no good instruments for diagnosing this issue, so judgment is needed, when deciding what to include in the model.

15.1.3 Transformations

This assumption implies that we have taken all possible non-linearities into account. If, for example, instead of using a multiplicative model, we apply an additive one, the estimates of parameters and the point forecasts might be *biased*. This is because the model will produce linear trajectory of the forecast, when a non-linear one is needed. This was discussed in detail in Section 14. The diagnostics of this assumption is similar to the diagnostics shown above for the omitted variables: construct actuals vs fitted and residuals vs fitted in order to see if there are any patterns in the plots. Take the multiple regression model for mtcars, which includes several variables, but is additive in its form:

```
mtcarsALM01 <- alm(mpg~wt+qsec+am, mtcars, loss="MSE")
```

Arguably, the model includes important variables (although there might be some others that could improve it), but the residuals will show some patterns, because the model should be multiplicative (see Figure 15.3), because mileage should not reduce linearly with increase of those variables. In order to understand that, ask yourself, whether the mileage can be negative and whether weight and other variables can be non-positive (a car with $wt = 0$ just does not exist).

```
par(mfcol=c(1,2))
plot(mtcarsALM01,c(1,2))
```

Figure 15.3 demonstrates the u-shaped pattern in the residuals, which is one of the indicators of a wrong model specification, calling for a non-linear transformation. We can try a model in logarithms:

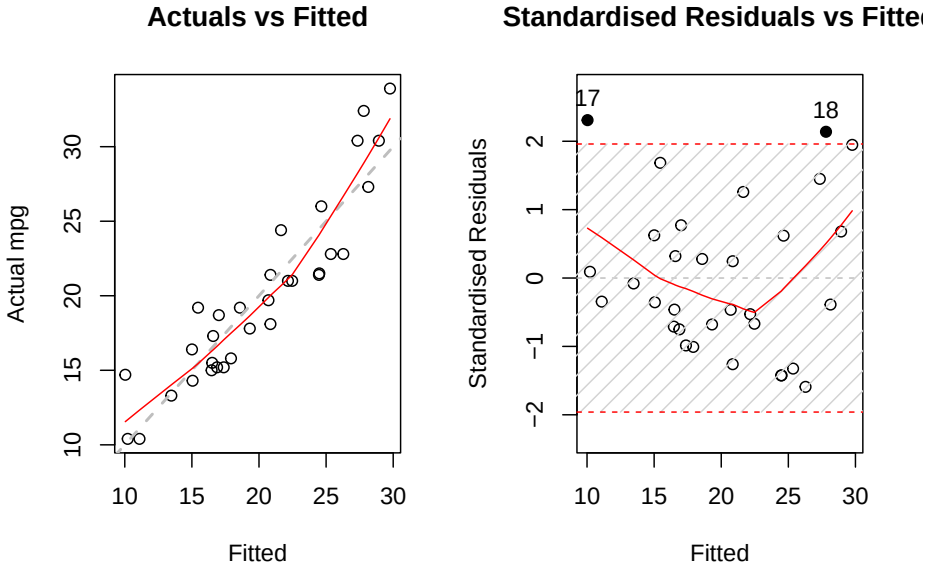


Figure 15.3: Diagnostics of necessary transformations in linear model.

```
mtcarsALM02 <- alm(log(mpg)~log(wt)+log(qsec)+am, mtcars, loss="MSE")
```

And see what would happen with the diagnostics of the model in logarithms:

```
par(mfcol=c(1,2))
plot(mtcarsALM02,c(1,2))
```

Figure 15.4 demonstrates that while the LOWESS lines do not coincide with the grey lines, the residuals do not have obvious patterns. The fact that the LOWESS line starts from below, when fitted values are low in our case only shows that we do not have enough observations with low actual values. As a result, LOWESS is impacted by 2 observations that lie below the grey line. This demonstrates that LOWESS lines should be taken with a pinch of salt and we should abstain from finding patterns in randomness, when possible. Overall, the log-log model is more appropriate to this data than the linear one.

15.1.4 Outliers

In a way, this assumption is similar to the first one with omitted variables. The presence of outliers might mean that we have missed some important information, implying that the estimates of parameters and forecasts would be *biased*. There can be other reasons for outliers as well. For example, we might be using a wrong distributional assumption. If so, this would imply that the prediction interval from the model is narrower than necessary. The diagnostics of outliers comes to producing standardised residuals vs fitted, to studentised vs fitted and

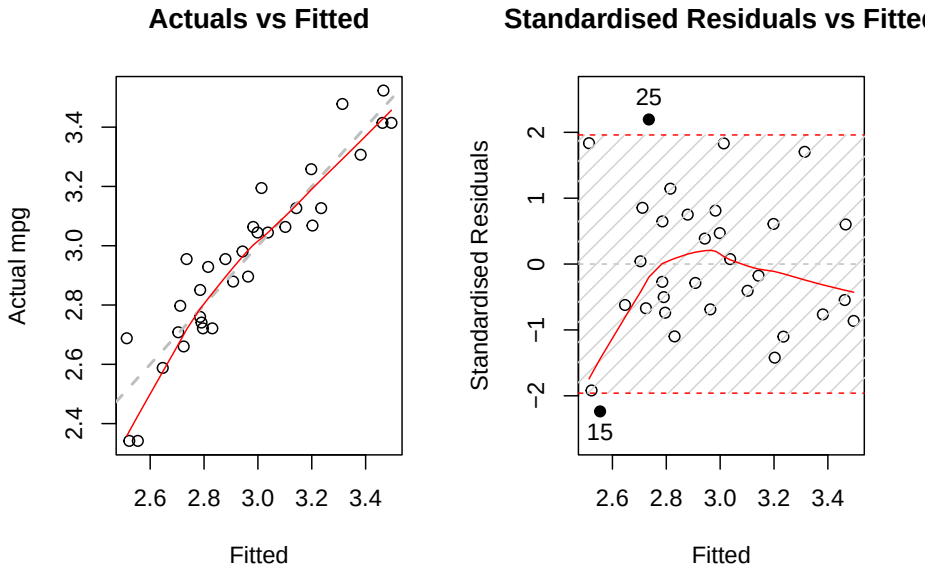


Figure 15.4: Diagnostics of necessary transformations in log-log model.

to Cook's distance plot. While we are already familiar with the first one, the other two need to be explained in more detail.

Studentised residuals are the residuals that are calculated in the same way as the standardised ones, but removing the value of each residual. For example, the studentised residual on observation 25 would be calculated as the raw residual divided by standard deviation of residuals, calculated without this 25th observation. This way we diminish the impact of potential serious outliers on the standard deviation, making it easier to spot the outliers.

As for the Cook's distance, its idea is to calculate measures for each observation showing how influential they are in terms of impact on the estimates of parameters of the model. If there is an influential outlier, then it would distort the values of parameters, causing bias.

```
par(mfcol=c(1,2))
plot(mtcarsALM02,c(2,3))
```

Figure 15.5 demonstrates standardised and studentised residuals vs fitted values for the log-log model on mtcars data. We can see that the plots are very similar, which already indicates that there are no strong outliers in the residuals. The bounds produced on the plots correspond to the 95% prediction interval, so by definition it should contain $0.95 \times 32 \approx 30$ observations. Indeed, there are only two observations: 15 and 25 - that lie outside the bounds. Technically, we would suspect that they are outliers, but they do not lie far away from the bounds and their number meets our expectations, so we can conclude that there are no

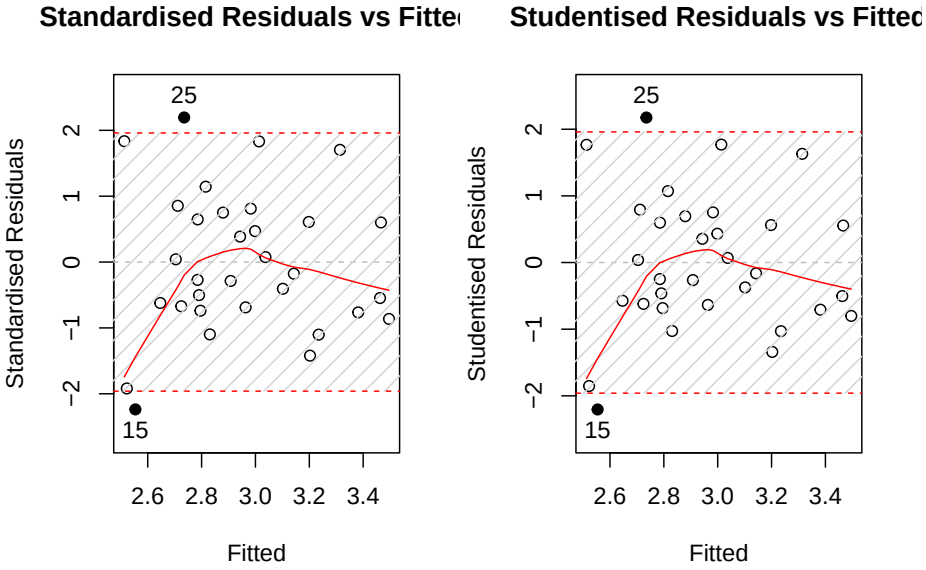


Figure 15.5: Diagnostics of outliers.

outliers in the data.

```
plot(mtcarsALM02,12)
```

Finally, we produce Cook's distance over observations in Figure 15.6. The x-axis says "Time", because `alm()` function is tailored for time series data, but this can be renamed into "observations". The plot shows how influential the outliers are. If there were some significantly influential outliers in the data, then the plot would draw red lines, corresponding to 0.5, 0.75 and 0.95 quantiles of Fisher's distribution, and the line of those outliers would be above the red lines. Consider the following example for demonstration purposes:

```
mtcarsData[28,6] <- 4
mtcarsALM03 <- alm(log(mpg)~log(wt)+log(qsec)+am, mtcarsData, loss="MSE")
```

This way, we intentionally create an influential outlier (the car should have the minimum weight in the dataset, and now it has a very high one).

```
plot(mtcarsALM03, 12, ylim=c(0,1.5), xlab="Observations", main="")
```

Figure 15.7 shows how Cook's distance will look in this case - it detects that there is an influential outlier, which is above the norm. We can compare the parameters of the new and the old models to see how the introduction of one outlier leads to bias in the estimates of parameters:

```
rbind(coef(mtcarsALM02),
      coef(mtcarsALM03))
```

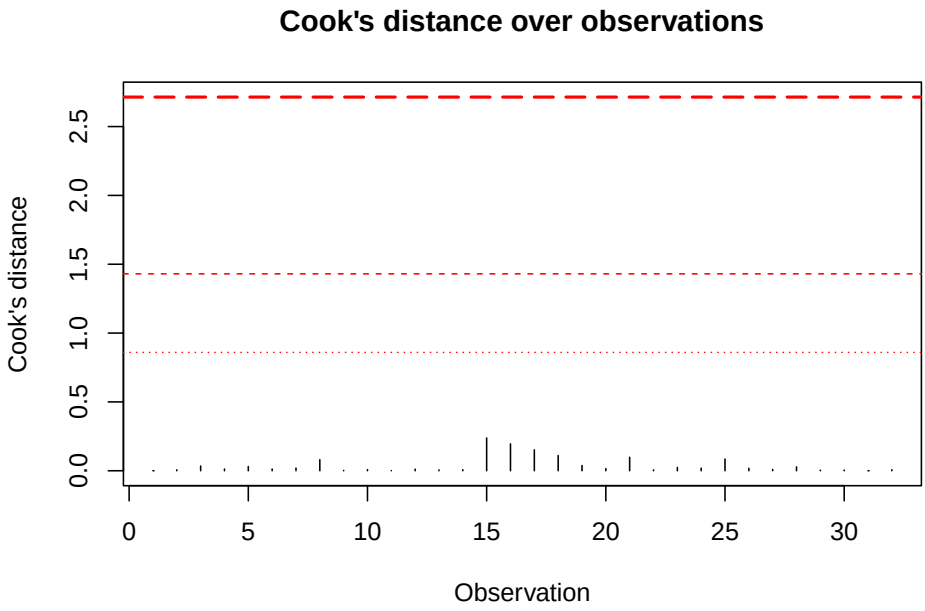


Figure 15.6: Cook's distance plot.

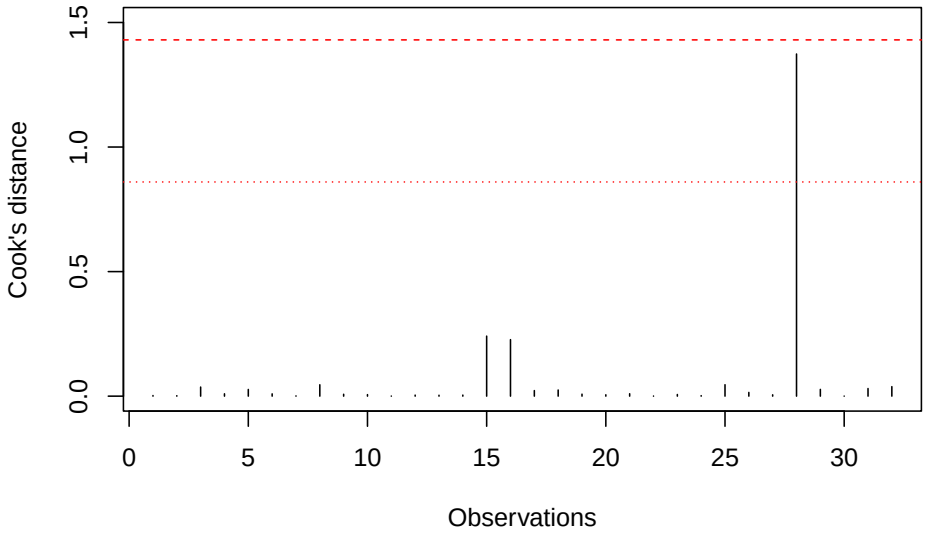


Figure 15.7: Cook's distance plot for the data with influential outlier.

```
##      (Intercept)    log(wt) log(qsec)         am
## [1,]   1.2095788 -0.7325269 0.8857779 0.05205307
## [2,]   0.1382442 -0.4852647 1.1439862 0.21406331
```

15.2 Residuals are i.i.d.

There are five assumptions in this group:

1. There is no autocorrelation in the residuals;
2. The residuals are homoscedastic;
3. The expectation of residuals is zero, no matter what;
4. The variable follows the assumed distribution;
5. More generally speaking, distribution of residuals does not change over time.

15.2.1 No autocorrelations

This assumption **only applies to time series data**, and in a way comes to capturing correctly the dynamic relations between variables. The term "autocorrelation" refers to the situation, when variable is correlated with itself from the past. If the residuals are autocorrelated, then something is neglected by the applied model. Typically, this leads to *inefficient* estimates of parameters, which in some cases might also become *biased*. The model with autocorrelated residuals might produce inaccurate point forecasts and prediction intervals of a wrong width (wider or narrower than needed).

There are several ways of diagnosing the problem, including visual analysis and statistical tests. In order to show some of them, we consider the `Seatbelts` data from `datasets` package for R. We fit a basic model, predicting monthly totals of car drivers in the Great Britain killed or seriously injured in car accidents:

```
SeatbeltsALM01 <- alm(drivers~PetrolPrice+kms+front+rear+law, Seatbelts)
```

In order to do graphical diagnose, we can produce plots of standardised / studentised residuals over time:

```
plot(SeatbeltsALM01,8,main="")
```

If the assumption is not violated, then the plot in Figure 15.8 would not contain any patterns. However, we can see that, first, there is a seasonality in the residuals and second, the expectation (captured by the red LOWESS line) changes over time. This indicates that there might be some autocorrelation in residuals caused by omitted components. We do not aim to resolve the issue now, it is discussed in more detail in Section 14.5 of Svetunkov (2021).

The other instrument for diagnostics is ACF / PACF plots, which are produced in `alm()` via the following command:

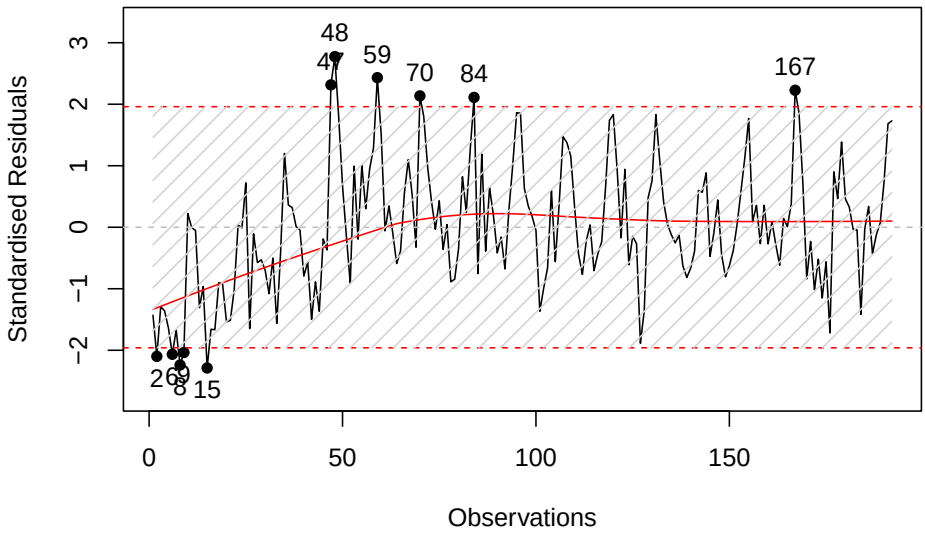


Figure 15.8: Standardised residuals over time.

```
par(mfcol=c(1,2))
plot(SeatbeltsALM01,c(10,11),main="")
```

These are discussed in more detail in Sections **??** and **??**.

15.2.2 Homoscedastic residuals

In general, we assume that the variance of residuals is constant. If this is violated, then we say that there is a **heteroscedasticity** in the model. This means that with a change of a variable, the variance of the residuals will change as well. If the model neglects this, then typically the estimates of parameters become *inefficient* and prediction intervals are wrong: they are wider than needed in some cases (e.g.m when the volume of data is low) and narrower than needed in the other ones (e.g. on high volume data).

Typically, this assumption will be violated if the model is not specified correctly. The classic example is the income versus expenditure on meals for different families. If the income is low, then there are not many options for buying, and the variability of expenses would be low. However, with the increase of income, the mean expenditures and their variability would increase because there are more options of what to buy, including both cheap and expensive products. If we constructed a basic linear model on such data, then it would violate the assumption of homoscedasticity and, as a result, will have issues discussed in section 15.2. But arguably, this would typically appear because of the misspecification of the model. For example, taking logarithms might resolve the issue in many cases, implying that the effect of one variable on the other

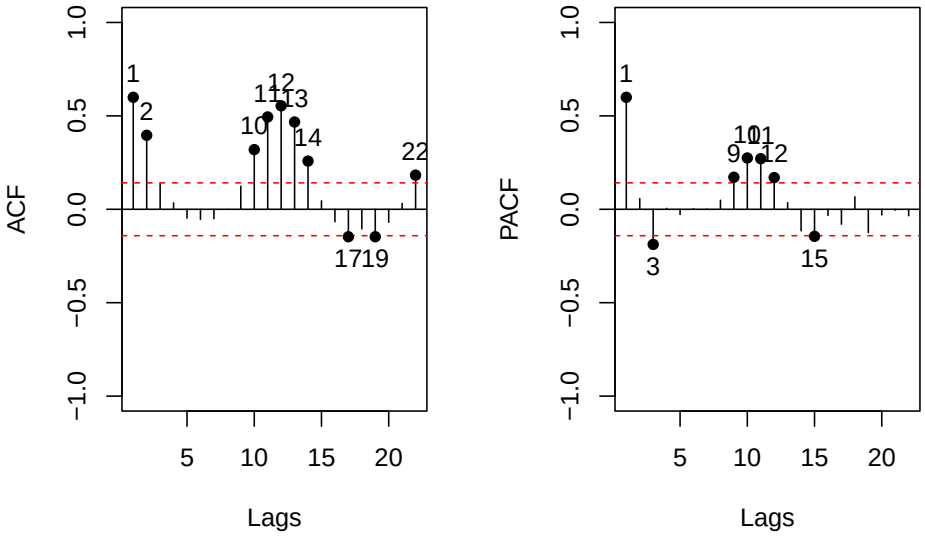


Figure 15.9: ACF and PACF of the residuals of a model.

should be multiplicative rather than additive. Alternatively, dividing variables by some other variable might (e.g. working with expenses per family member, not per family) resolve the problem as well. Unfortunately, the transformations are not the panacea, so in some cases, the analyst would need to construct a model, taking the changing variance into account (e.g. GARCH or GAMLSS models). This is discussed in Section **??**.

While forecasting, we are more interested in the holdout performance of models, in econometrics, the parameters of models are typical of the main interest. And, as we discussed earlier, in the case of a correctly specified model with heteroscedastic residuals, the estimates of parameters will be unbiased but inefficient. So, econometricians would use different approaches to diminish the heteroscedasticity effect on parameters: either a different estimator for a model (such as Weighted Least Squares) or a different method for calculating standard errors of parameters (e.g. Heteroskedasticity-Consistent Standard Errors). This does not resolve the problem but instead corrects the model's parameters (i.e. does not heal the illness but treats the symptoms). Although these approaches typically suffice for analytical purposes, they do not fix the issues in forecasting.

The diagnostics of heteroscedasticity can be done via plotting absolute and / or squared residuals against the fitted values.

```
par(mfcol=c(1,2))
plot(mtcarsALM01,4:5)
```

If your model assumes that residuals follow a distribution related to the Normal one, then you should focus on the plot of squared residuals vs fitted, as this

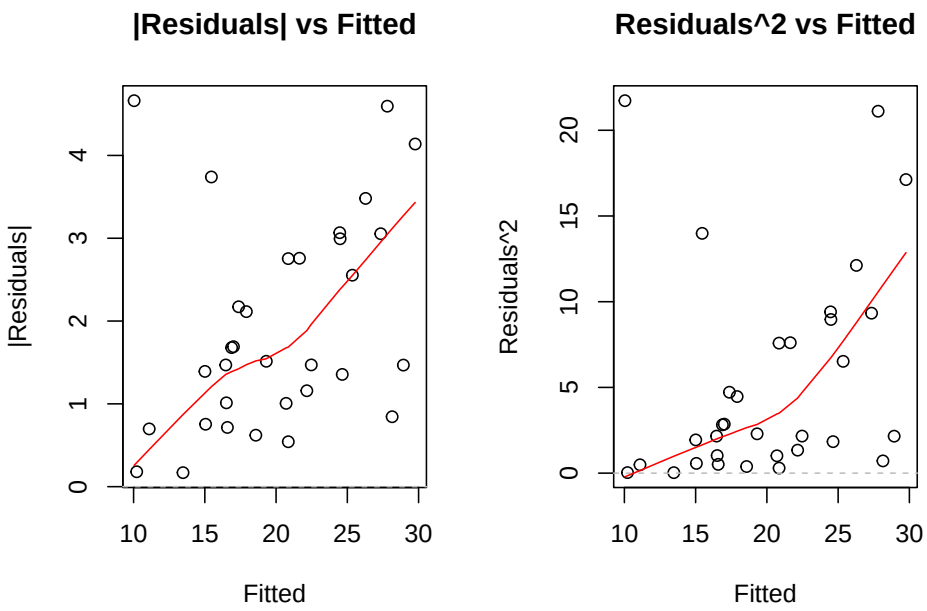


Figure 15.10: Detecting heteroscedasticity. Model 1.

would be closer related to the variance of the distribution. In the example of mtcars model in Figure 15.10 we see that the variance of residuals increases with the increase of Fitted values (the LOWESS line increases and the overall variability around 1200 is lower than the one around 2000). This indicates that the residuals are heteroscedastic. One of the possible solutions of the problem is taking the logarithms, as we have done in the model `mtcarsALM02`:

```
par(mfcol=c(1,2))
plot(mtcarsALM02,4:5)
```

While the LOWESS lines on plots in Figure 15.11 demonstrate some dynamics, the variability of residuals does not change significantly with the increase of fitted value, so non-linear transformation seems to fix the issue in our example. If it would not, then we would need to consider either some other transformations or finding out, which of the variables causes heteroscedasticity and then modelling it explicitly via the scale model (Section **??**).

15.2.3 Mean of residuals

While in sample, this holds automatically in many cases (e.g. when using Least Squares method for regression model estimation), this assumption might be violated in the holdout sample. In this case the point forecasts would be *biased*, because they typically do not take the non-zero mean of forecast error into account, and the prediction interval might be off as well, because of the wrong

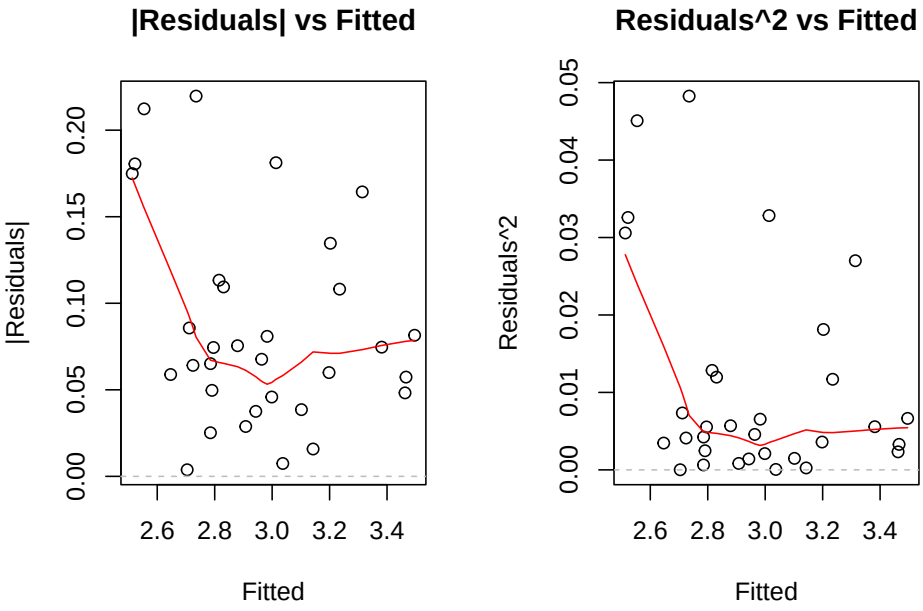


Figure 15.11: Detecting heteroscedasticity. Model 2.

estimation of the scale of distribution (e.g. variance is higher than needed). This assumption also implies that the expectation of residuals is zero even conditional on the explanatory variables in the model. If it is not, then this might mean that there is still some important information omitted in the applied model. This implies that the following holds for all x_i:

$$\text{cov}(x_i, e) = 0,$$

which in the case of $\text{E}(e) = 0$ is equivalent to:

$$\text{E}(x_i e) = 0.$$

If OLS is used in linear model estimation, then this condition is satisfied in sample automatically and does not require checking.

Note that some models assume that the expectation of residuals is equal to one instead of zero (e.g. multiplicative error models). The idea of the assumption stays the same, it is only the value that changes.

The diagnostics of the problem would be similar to the case of non-linear transformations or autocorrelations: plotting residuals vs fitted or residuals vs time and trying to find patterns. If the mean of residuals changes either with the change of fitted values of with time, then the conditional expectation of residuals is not zero, and something is missing in the model.

15.2.4 Distributional assumptions

In some cases we are interested in using methods that imply specific distributional assumptions about the model and its residuals. For example, it is assumed in the classical linear model that the error term follows Normal distribution. Estimating this model using MLE with the probability density function of Normal distribution or via minimisation of Mean Squared Error (MSE) would give *efficient* and *consistent* estimates of parameters. If the assumption of normality does not hold, then the estimates might be *inefficient* and in some cases *inconsistent*. When it comes to forecasting, the main issue in the wrong distributional assumption appears, when prediction intervals are needed: they might rely on a wrong distribution and be narrower or wider than needed. Finally, if we deal with the wrong distribution, then the model selection mechanism might be flawed and would lead to the selection of an inappropriate model.

The most efficient way of diagnosing this, is constructing QQ-plot of residuals (discussed in Section 5.2).

```
plot(mtcarsALM02,6)
```

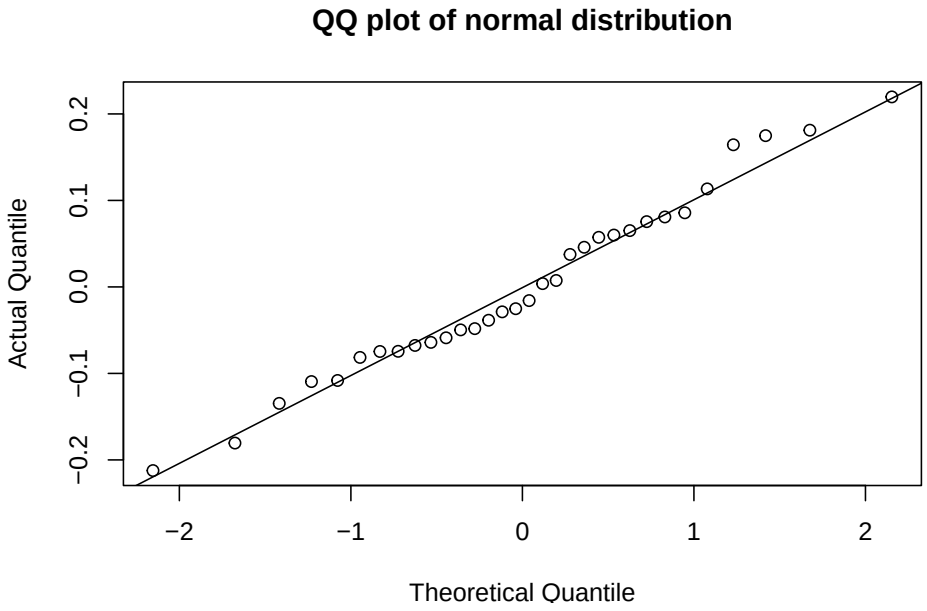


Figure 15.12: QQ-plot of residuals of model 2 for mtcars dataset.

Figure 15.12 shows that all the points lie close to the line (with minor fluctuations around it), so we can conclude that the residuals follow the normal distribution. In comparison, Figure 15.12 demonstrates how residuals would look in case of a wrong distribution. Although the values lie not too far from ths straight line, there are several observations in the tails that are further away than needed.

Comparing the two plots, we would select the on in Figure 15.12, as the residuals are better behaved.

```
plot(mtcarsALM01,6)
```

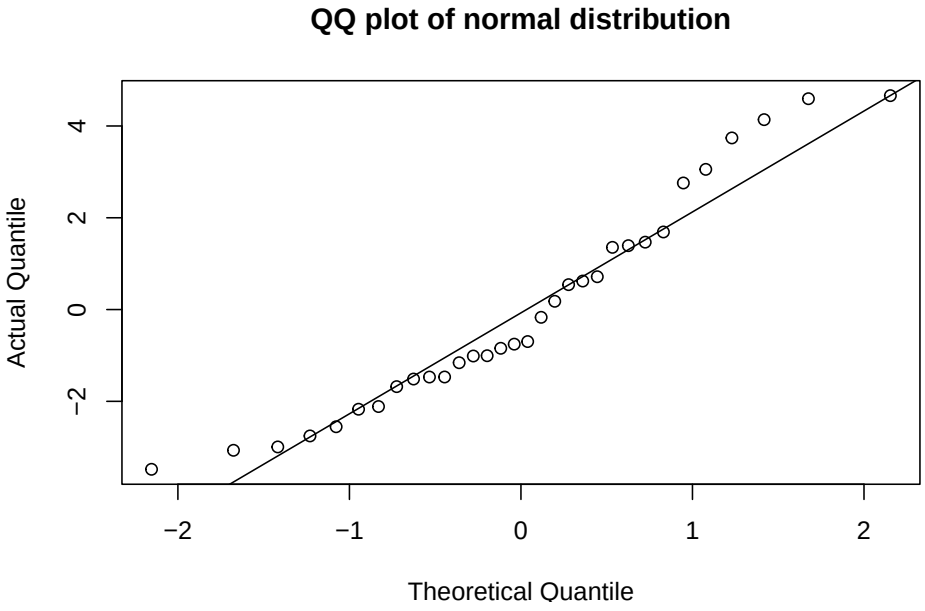


Figure 15.13: QQ-plot of residuals of model 1 for mtcars dataset.

15.2.5 Distribution does not change

This assumption aligns with the Subsection 15.2.4, but in this specific context implies that all the parameters of distribution stay the same and the shape of distribution does not change. If the former is violated then we might have one of the issues discussed above. If the latter is violated then we might produce *biased* forecasts and underestimate / overestimate the uncertainty about the future. The diagnosis of this comes to analysing QQ-plots, similar to Subsection 15.2.4.

15.3 The explanatory variables are not correlated with anything but the response variable

There are two assumptions in this group:

1. No multicollinearity;
2. No endogeneity;

Technically speaking, both of them are not assumptions, but rather potential issues of a model. This is because they have nothing to do with properties of the "true model". Indeed, it is unreasonable to assume that the explanatory variables do not have any relation between them or that they are not impacted by the response variable – they are what they are. However, the two issues cause difficulties in estimating parameters of models and lead to issues with estimates of parameters. So, they are worth discussing.

15.3.1 Multicollinearity

Multicollinearity appears, when either some of explanatory variables are correlated with each other (see Section 9.3), or their linear combination explains another explanatory variable included in the model. Depending on the strength of this relation and the estimation method used for model construction, the multicollinearity might cause issues of varying severity. For example, in the case, when two variables are perfectly correlated (correlation coefficient is equal to 1 or -1), the model will have perfect multicollinearity and it would not be possible to estimate its parameters. Another example is a case, when an explanatory variable can be perfectly explained by a set of other explanatory variables (resulting in R^2 being close to one), which will cause exactly the same issue. The classical example of this situation is the dummy variables trap (see Section 13), when all values of categorical variable are included in regression together with the constant resulting in the linear relation $\sum_{j=1}^{k} d_j = 1$. Given that the square root of R^2 of linear regression is equal to multiple correlation coefficient, these two situations are equivalent and just come to "absolute value of correlation coefficient is equal to 1". Finally, if correlation coefficient is high, but not equal to one, the effect of multicollinearity will lead to less efficient estimates of parameters. The loss of efficiency is in this case proportional to the absolute value of correlation coefficient. In case of forecasting, the effect is not as straight forward, and in some cases might not damage the point forecasts, but can lead to prediction intervals of an incorrect width. The main issue of multicollinearity comes to the difficulties in the model estimation in a sample. If we had all the data in the world, then the issue would not exist. All of this tells us how this problem can be diagnosed and that this diagnosis should be carried out before constructing regression model.

First, we can calculate correlation matrix for the available variables. If they are all numeric, then `cor()` function from `stats` should do the trick (we remove the response variable from consideration):

```
cor(mtcars[,-1])
```

```
##             cyl       disp         hp        drat         wt        qsec
## cyl   1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958 -0.59124207
## disp  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
## hp    0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
## drat -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
```

```
## wt    0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
## qsec -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
## vs   -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157  0.74453544
## am   -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953 -0.22986086
## gear -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870 -0.21268223
## carb  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059 -0.65624923
##              vs          am       gear        carb
## cyl  -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    1.0000000  0.16834512  0.2060233 -0.56960714
## am    0.1683451  1.00000000  0.7940588  0.05753435
## gear  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.5696071  0.05753435  0.2740728  1.00000000
```

This matrix tells us that there are some variables that are highly correlated and might reduce efficiency of estimates of parameters of regression model if included in the model together. This mainly applies to `cyl` and `disp`, which both characterise the size of engine. If we have a mix of numerical and categorical variables, then `assoc()` (aka `association()`) function from `greybox` will be more appropriate (see Section 9).

```
assoc(mtcars)
```

In order to cover the second situation with linear combination of variables, we can use the `determ()` (aka `determination()`) function from `greybox`:

```
determ(mtcars[,-1])
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 0.9349544 0.9537470 0.8982917 0.7036703 0.9340582 0.8671619 0.7986256 0.7848763
##      gear      carb
## 0.8133441 0.8735577
```

This function will construct linear regression models for each variable from all the other variables and report the R^2 from these models. If there are coefficients of determination close to one, then this might indicate that the variables would cause multicollinearity in the model. In our case, we see that `disp` is linearly related to other variables, and we can expect it to cause the reduction of efficiency of estimate of parameters. If we remove it from the consideration (we do not want to include it in our model anyway), then the picture will change:

```
determ(mtcars[,-c(1,3)])
```

```
##       cyl        hp      drat        wt      qsec        vs        am      gear
## 0.9299952 0.8596168 0.6996363 0.8384243 0.8553748 0.7965848 0.7847198 0.8121855
```

```
##      carb
## 0.7680136
```

Now `cyl` has linear relation with some other variables, so it would not be wise to include it in the model with the other variables. We would need to decide, what to include based on our understanding of the problem.

Instead of calculating the coefficients of determination, econometricians prefer to calculate Variance Inflation Factor (VIF), which shows by how many times the estimates of parameters will loose efficiency. Its formula is based on the R^2 calculated above:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

for each model i. Which in our case can be calculated as:

```
1/(1-determ(mtcars[,-c(1,3)]))
```

```
##       cyl        hp      drat        wt      qsec        vs        am      gear
## 14.284737  7.123361  3.329298  6.189050  6.914423  4.916053  4.645108  5.324402
##      carb
##  4.310597
```

This is useful when you want to see the specific impact on the variance of parameters, but is difficult to work with, when it comes to model diagnostics, because the value of VIF lies between zero and infinity. So, I prefer using the determination coefficients instead, which is always bounded by $(0, 1)$ region and thus easier to interpret.

Finally, in some cases nothing can be done with multicollinearity, it just exists, and we need to include those correlated variables. This might not be a big problem, as long as we acknowledge the issues it will cause to the estimates of parameters.

15.3.2 Endogeneity

Endogeneity applies to the situation, when the dependent variable y_j influences the explanatory variable x_j in the model on the same observation. The relation in this case becomes bi-directional, meaning that the basic model is not appropriate in this situation any more. The parameters and forecasts will typically be *biased*, and a different estimation method would be needed (for example, instrumental variables) or maybe a different model would need to be constructed in order to fix this.

In econometrics, one of the definitions of the endogeneity is that the correlation between the error term and an explanatory variable is not zero, i.e. $\text{E}(\epsilon_j, x_{i,j}) \neq 0$ for at least some variable x_i. In my personal opinion, this is a very confusing definition. First, if this applies to the "true" model then this is absurd, because by definition the error term in the true model is not related with anything (because the true model is correctly specified). Second, if this applies to the

applied model, then this condition does not hold in sample if OLS is used for the estimation of parameters (this was discussed in Subsection 10.3). Third, even if we are talking about working with an incorrect model on the population data, the OLS will guarantee that $\mathrm{E}(e_j, x_{i,j}) = 0$. So, the only case when this makes sense is for the relation between the explanatory variables and the forecast errors from the model generated on the holdout sample of data. This is why I think that this definition is not useful.

To make things even more complicated, endogeneity cannot be properly diagnosed and comes to the judgment of analyst: do we expect the relation between variables to be one directional or bi-directional? From the true model perspective, the latter might imply that we need to consider a system of equations of the style:

$$\begin{matrix} y_j = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_{k-1} x_{k-1,j} + \epsilon_j \\ x_{1,j} = \gamma_0 + \gamma_1 y_j + \gamma_2 x_{2,j} + \cdots + \gamma_{k-1} x_{k-1,j} + \upsilon_j \end{matrix} . \tag{15.1}$$

In the equation (15.1), the response variable y_j depends on the value of $x_{1,j}$ (among other variables), but that variable depends on the value of y_j at the same time. In order to estimate such system of equations and break this loop, an analyst would need to find an "instrumental variable" – a variable that would be correlated with $x_{1,j}$ but would not be correlated with y_j and then use a different estimation procedure (e.g. two-stage least squares). We do not aim to cover possible solutions of this issue, because they lie outside of the scope of this textbook, but an interested reader is referred to Chapter 12 of Hanck et al. (2022).

Remark. Note that if we work with time series then endogeneity would only appear when the bi-directional relation happens at the same time t, not over time. In the latter case we would be dealing with recursive relation (y_t depends on x_t, but x_t depends on y_{t-1}) rather than the contemporaneous and thus the estimation of such a model would not lead to the issues discussed in this subsection.

Chapter 16

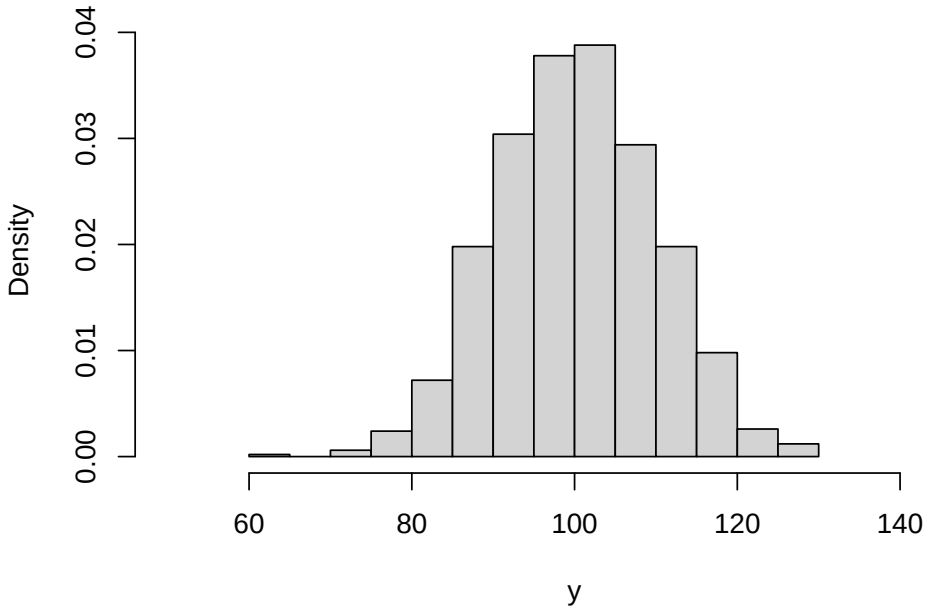
Likelihood Approach

We will use different estimation techniques throughout this book, one of the main of which is **Maximum Likelihood Estimate** (MLE). The very rough idea of the approach is to maximise the chance that each observation in the sample follows a pre-selected distribution with specific set of parameters. In a nutshell, what we try to do when using likelihood for estimation, is fit the distribution function to the data. In order to demonstrate this idea, we start in a non-conventional way, with an example in R. We will then move to the mathematical side of the problem.

16.1 An example in R

We consider a simple example, when we want to estimate the model $y_j = \mu_y + \epsilon_j$ (global average), assuming that the error term follows normal distribution: $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, which means that $y_j \sim \mathcal{N}(\mu_y, \sigma^2)$. In this case we want to estimate two parameters using likelihood: location $\hat{\mu}_y$ and scale $\hat{\sigma}^2$. First, we generate the random variable in R and plot its distribution:

```
y <- rnorm(1000, 100, 10)
hist(y, xlim=c(50,150), main="", probability=TRUE)
```



As expected, the distribution of this variable (1000 observations) has the bell shape of Normal distribution. In order to estimate the parameters, for the distribution, we will try them one by one and see how the likelihood and the shape of the fitted curve to this histogram change. We start with $\hat{\mu}_y = 80$ and $\hat{\sigma} = 10$ just to see how the probability density function of normal distribution fits the data:

```
hist(y, xlim=c(50,150), main="", probability=TRUE)
lines(c(50:150),dnorm(c(50:150),80,10),col="red",lwd=2)
abline(v=80,col="red",lwd=2)
```

and we get the following log-likelihood value (we will discuss how this formula can be obtained later):

```
sum(dnorm(y,80,10,log=T))
```

```
## [1] -5757.199
```

In order for the normal distribution on 16.1 to fit the data well, we need to shift the estimate of μ_y to the right, thus increasing the value to, let's say, $\hat{\mu}_y = 90$:

```
hist(y, xlim=c(50,150), main="", probability=TRUE)
lines(c(50:150),dnorm(c(50:150),90,10),col="orange",lwd=2)
abline(v=90,col="orange",lwd=2)
```

Now, in Figure 16.2, the normal curve is much closer to the data, but it is still a bit off. The log-likelihood value in this case is -4227.945, which is higher than the previous one, indicating that we are moving towards the maximum of the likelihood function. Moving it further, setting $\hat{\mu}_y = 100$, we get:

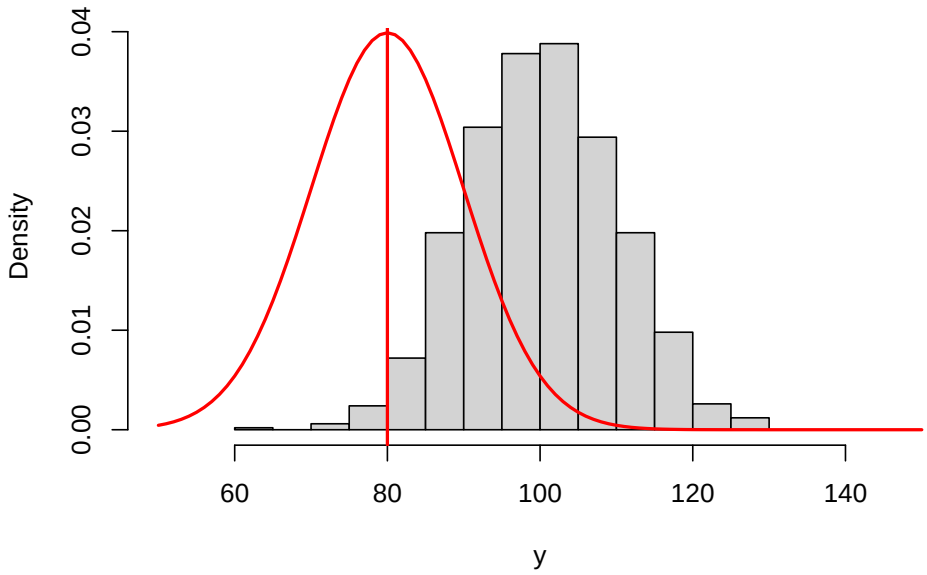


Figure 16.1: ML example with Normal curve and $\hat{\mu}_y = 80$ and $\hat{\sigma} = 10$

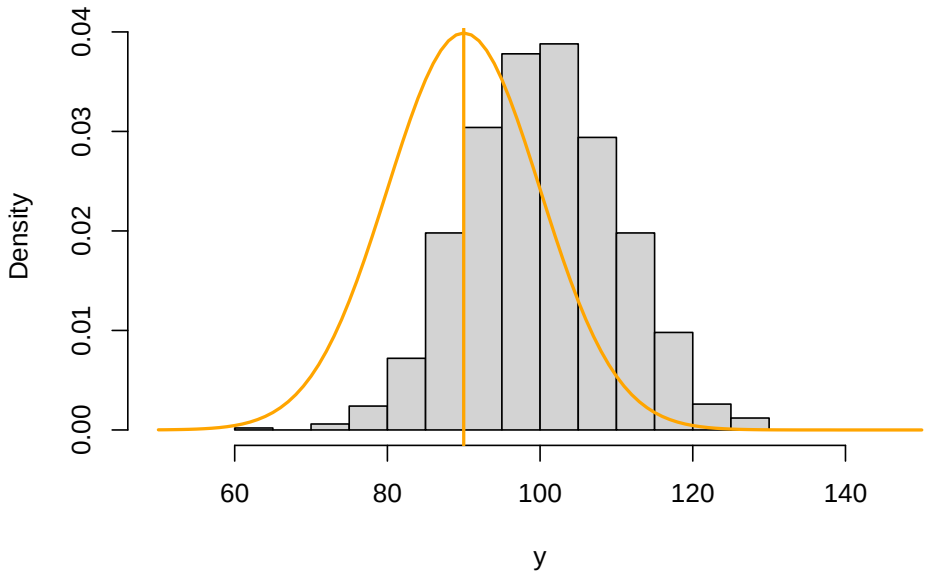


Figure 16.2: ML example with Normal curve and $\hat{\mu}_y = 90$ and $\hat{\sigma} = 10$

```
hist(y, xlim=c(50,150), main="", probability=TRUE)
lines(c(50:150),dnorm(c(50:150),100,10),col="green3",lwd=2)
abline(v=100,col="green3",lwd=2)
```

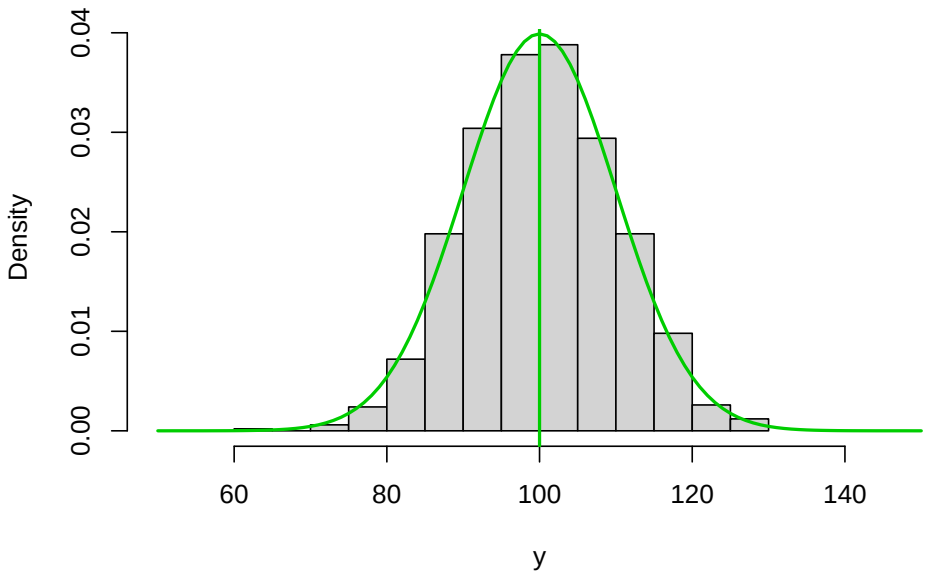


Figure 16.3: ML example with Normal curve and $\hat{\mu}_y = 100$ and $\hat{\sigma} = 10$

Figure 16.2 demonstrates a much better fit than in the previous cases with the log-likelihood of -3698.691, which is even higher than in the previous case. We are almost there. In fact, in order to maximise this likelihood, we just need to calculate the sample mean of the variable (this is the MLE of the location parameter in normal distribution) and insert it in the function to obtain:

```
hist(y, xlim=c(50,150), main="", probability=TRUE)
lines(c(50:150),dnorm(c(50:150),mean(y),10),col="darkgreen",lwd=2)
abline(v=mean(y),col="darkgreen",lwd=2)
```

So the value of $\hat{\mu}_y = \bar{y} = 100.293$ (where $\bar{y}$ is the sample mean) maximises the likelihood function, resulting in log-likelihood of -3698.263.

In a similar fashion we can get the MLE of the scale parameter σ^2 of the model. In this case, we will be changing the height of the distribution. Here is an example with $\hat{\mu}_y = 100.293$ and $\hat{\sigma} = 15$:

```
hist(y, xlim=c(50,150), main="", probability=TRUE)
lines(c(50:150),dnorm(c(50:150),mean(y),15),col="royalblue",lwd=2)
abline(v=mean(y),col="royalblue",lwd=2)
```

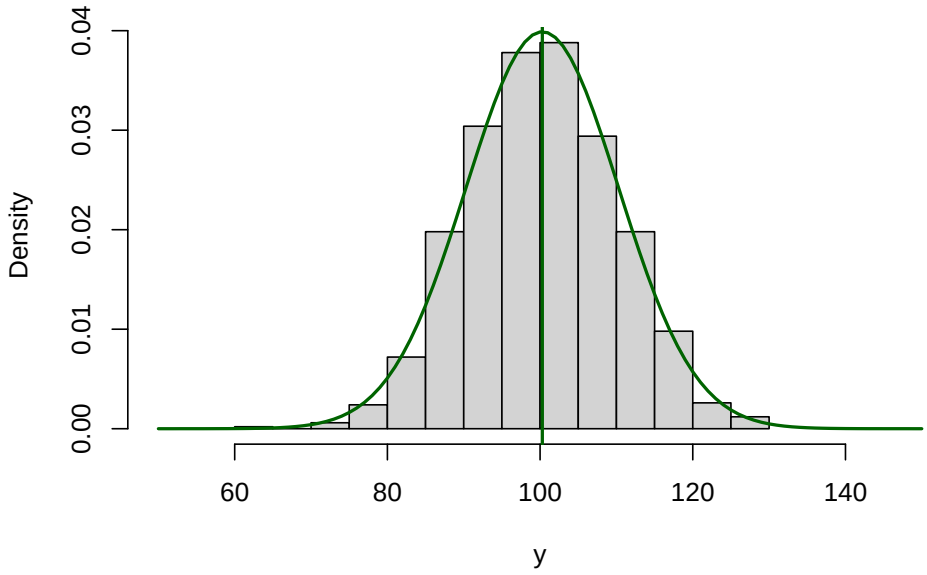


Figure 16.4: ML example with Normal curve and $\hat{\mu}_y = \bar{y}$ and $\hat{\sigma} = 10$

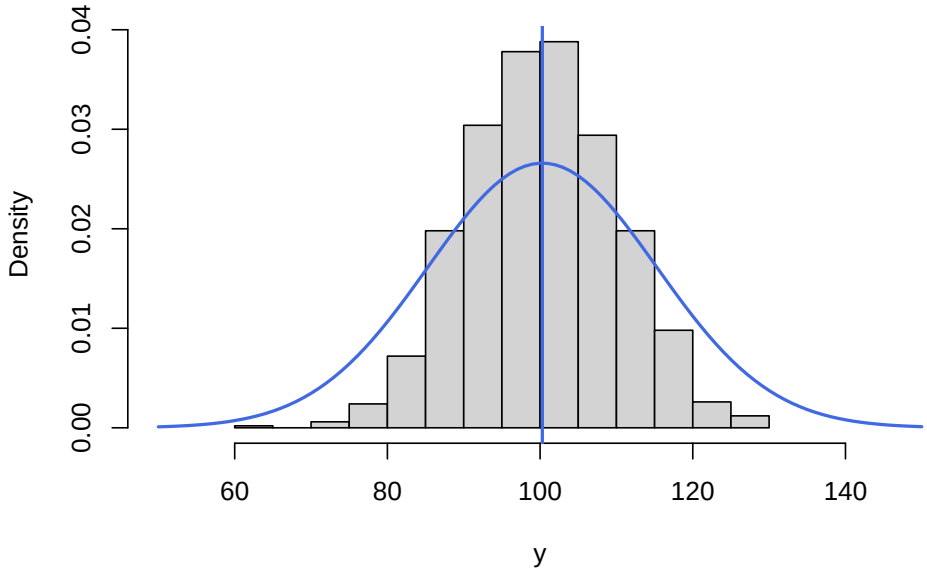


Figure **??** demonstrates that the curve is located lower than needed, which implies that the scale parameter $\hat{\sigma}$ is too high. The log-likelihood value in this case is -3838.873. In order to get a better fit of the curve to the data, we need to reduce the $\hat{\sigma}$. Here how the situation would look for the case of $\hat{\sigma} = 10$:

```
hist(y, xlim=c(50,150), main="", probability=TRUE)
lines(c(50:150),dnorm(c(50:150),mean(y),10),col="darkblue",lwd=2)
```

```
abline(v=mean(y),col="darkblue",lwd=2)
```

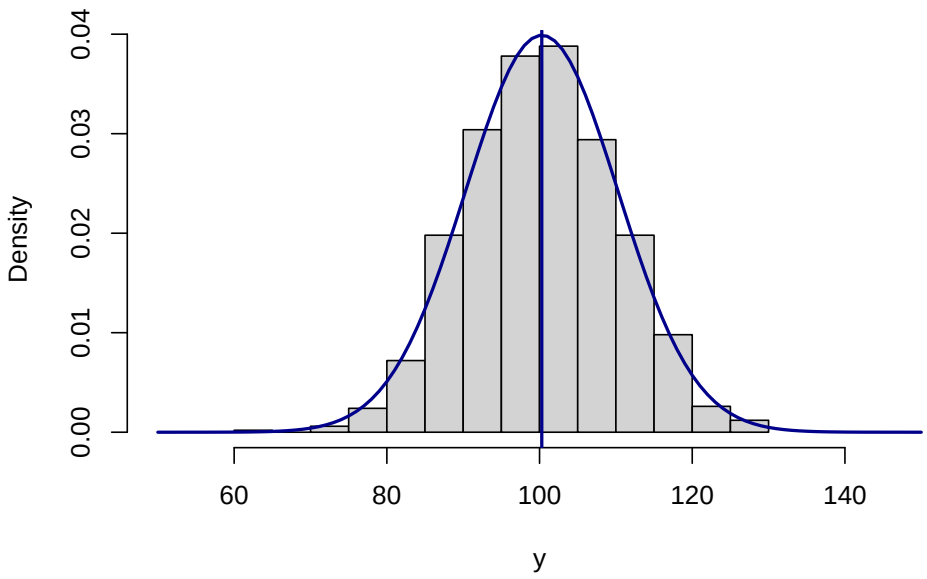


Figure 16.5: ML example with Normal curve and $\hat{\mu}_y = \bar{y}$ and $\hat{\sigma} = 10$

The fit on Figure 16.5 is better than on Figure **??**, which is also reflected in the log-likelihood value being equal to -3698.263 instead of -3838.873. The best fit and the maximum of the likelihood is obtained, when the scale parameter is estimated using the formula $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2$, resulting in log-likelihood of -3697.705. Note that if we use the unbiased estimate of the variance $\hat{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$, the log-likelihood will not reach the maximum and will be equal to -3697.705. In our special case the difference between the two is infinitesimal, because of the large sample (1000 observations), but it will be more substantial on small samples. Still, the two likelihood values are different, which can be checked in R via the following commands:

```
# The maximum log-likelihood with the biased variance
logLik01 <- sum(dnorm(y,mean(y),sqrt(mean((y-mean(y))^2)),log=TRUE))
# The log-likelihood value with the unbiased variance
logLik02 <- sum(dnorm(y,mean(y),sd(y),log=TRUE))
# The difference between the two
logLik01 - logLik02
```

```
## [1] 0.0002501668
```

All of this is great, but so far we have discussed a very special case, when the data follows normal distribution and we fit the respective model. But what if the model is wrong (no kidding!)? In that case the idea stays the same: we need to find the parameters of the normal distribution, that would guarantee the best

possible fit to the non-normal data. Here is an example with MLE of parameters of Normal distribution for the data following Log Normal one:

```
y <- rlnorm(1000, log(80), 0.4)
hist(y, main="", probability=T, xlim=c(0,300))
lines(c(0:300),dnorm(c(0:300),mean(y),sd(y)),col="blue",lwd=2)
```

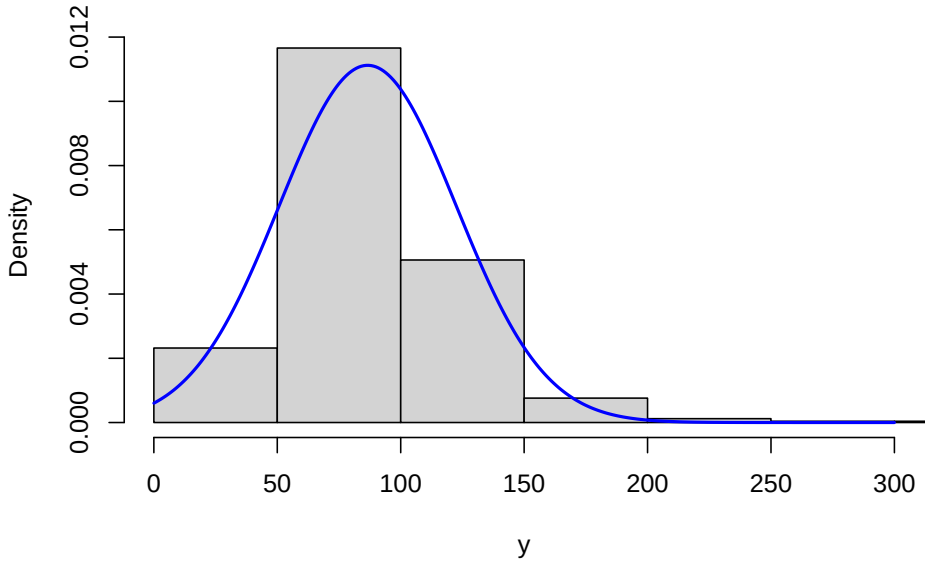


Figure 16.6: ML example with Normal curve on Log Normal data

Figure 16.6 shows that the Normal model does not fit the Log Normal data properly, but this is the best we can get, given our assumptions. The log-likelihood in this case is -4998.395. The much better model would be the Log Normal one:

```
hist(y, main="", probability=T, xlim=c(0,300))
lines(c(0:300),dlnorm(c(0:300),mean(log(y)),sd(log(y))),col="red",lwd=2)
```

The model in Figure 16.7 has the log likelihood of -4884.452. This indicates that the Log Normal model is more appropriate for the data and gives us an idea that it is possible to compare different distributions via the likelihood, finding the better fit to the data. This idea is explored further in the next section.

As a final word, when it comes to more complicated models with more parameters and dynamic structure, the specific curves and data become more complicated, but the logic of the likelihood approach stays the same.

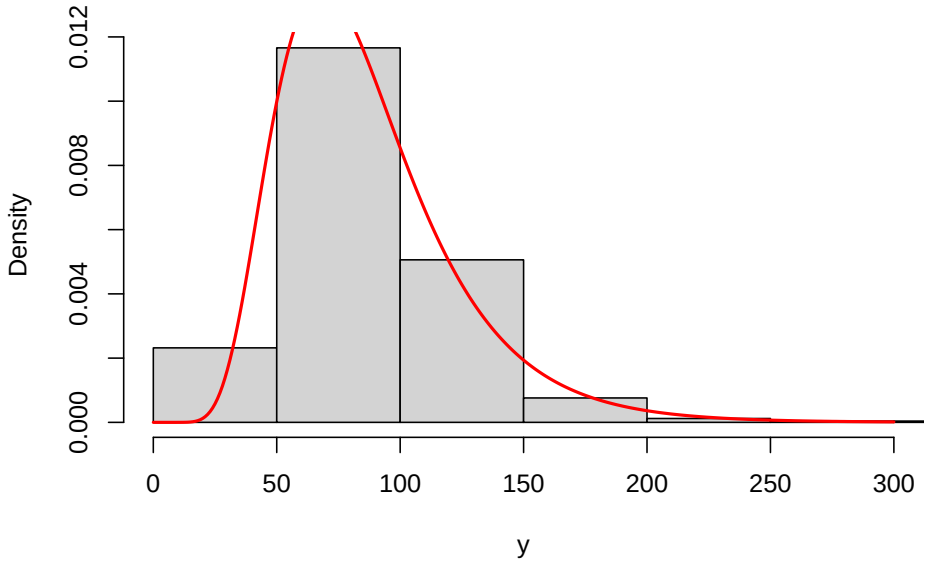


Figure 16.7: ML example with Log Normal curve on Log Normal data

16.2 Mathematical explanation

Now we can discuss the same idea from the mathematical point of view. We estimated the following simple model:

$$y_j = \mu_y + \epsilon_j, \tag{16.1}$$

assuming normal distribution of the residuals (see Section 4.3). In order to make things closer to the regression context, we will introduce changing location, which is defined by the regression line (thus, it is conditional on the set of $k-1$ explanatory variables):

$$y_j = \mu_{y,j} + \epsilon_j, \tag{16.2}$$

where $\mu_{y,j}$ is the population regression line, defined via:

$$\mu_{y,j} = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \cdots + \beta_{k-1} x_{k-1,j}. \tag{16.3}$$

The typical assumption in regression context is that $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$ (normal distribution with zero mean and fixed variance), which means that $y_j \sim \mathcal{N}(\mu_{y,j}, \sigma^2)$. We can use this assumption in order to calculate the point likelihood value for each observation based on the PDF of Normal distribution (Subsection 4.3):

$$\mathcal{L}(\mu_{y,j}, \sigma^2 | y_j) = f(y_j | \mu_{y,j}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - \mu_{y,j})^2}{2\sigma^2}\right). \tag{16.4}$$

Very roughly, what the value (16.4) shows is how likely it is that the specific observation comes from the assumed model with specified parameters (we know

that in real world data does not come from any model, but this interpretation is easier to work with). Note that the likelihood is not the same as probability, because for any continuous random variables the probability for it to be equal to any specific number is equal to zero (as discussed in Section 4.1). The point likelihood (16.4) is not very helpful on its own, but we can get n values like that, based on our sample of data. We can then summarise them in one number, that would characterise the whole sample, given the assumed distribution, applied model and selected values of parameters:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2|\mathbf{y}) = \prod_{j=1}^{n} \mathcal{L}(\mu_{y,j}, \sigma^2|\mathbf{y}) = \prod_{j=1}^{n} f(y_j|\mu_{y,j}, \sigma^2), \tag{16.5}$$

where $\boldsymbol{\theta}$ is the vector of all parameters in the model (in our example, it is $k+1$ of them: all the coefficients of the model and the scale σ^2). We take the product of likelihoods in (16.5) because we need to get the joint likelihood for all observations and because we can typically assume that the point likelihoods are independent of each other (for example, the value on observation j will not be influenced by the value on $j-1$). The value (16.5) shows roughly how likely on average it is that the data comes from the assumed model with specified parameters.

Remark. Technically speaking, the "on average" element will be achieved if we divide (16.5) by the number of observations n.

Having this value, we can change the values of parameters of the model, getting different value of (16.5) (as we did in the example in Section 16.1). Using an iterative procedure, we can get such estimates of parameters that would maximise the likelihood (16.5). These estimates of parameters are called "Maximum Likelihood Estimates'' (MLE). However, working with the products in formula (16.5) is challenging, so typically we linearise it using natural logarithm, obtaining log-likelihood. For the normal distribution, it can be written as:

$$\ell(\boldsymbol{\theta}, \sigma^2|\mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta}, \sigma^2|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{j=1}^{n} \frac{(y_j - \mu_{y,j})^2}{2\sigma^2}. \tag{16.6}$$

Based on that, we can find some of parameters of the model analytically. For example, we can derive the formula for the estimation of the scale based on the provided sample. Given that we are estimating the parameter, we should substitute σ^2 with $\hat{\sigma}^2$ in (16.6). We can then take derivative of (16.6) with respect to $\hat{\sigma}^2$ and equate it to zero in order to find the value that maximises the log-likelihood function in our sample:

$$\frac{d\ell(\boldsymbol{\theta}, \hat{\sigma}^2|\mathbf{y})}{d\hat{\sigma}^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{j=1}^{n}(y_j - \mu_{y,j})^2 = 0, \tag{16.7}$$

which after multiplication of both sides by $2\hat{\sigma}^4$ leads to:

$$n\hat{\sigma}^2 = \sum_{j=1}^{n}(y_j - \mu_{y,j})^2, \tag{16.8}$$

or

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \mu_{y,j}\right)^2 . \tag{16.9}$$

The value (16.9) is in fact a Mean Squared Error (MSE) of the model. If we calculate the value of $\hat{\sigma}^2$ using the formula (16.9), we will maximise the likelihood with respect to the scale parameter. In fact, we can insert (16.9) in (16.6) in order to obtain the so called "concentrated" (or profile) log-likelihood for the normal distribution:

$$\ell^*(\boldsymbol{\theta}|\mathbf{y}) = -\frac{n}{2}\left(\log(2\pi e) + \log\hat{\sigma}^2\right). \tag{16.10}$$

Remark. Sometimes, statisticians drop the $2\pi e$ part from the (16.10), because it does not affect any inferences, as long as one works only with Normal distribution. However, in general, it is not recommended to do (Burnham and Anderson, 2004), because this makes the comparison with other distributions impossible.

This function is useful because it simplifies some calculations and also demonstrates the condition, for which the likelihood is maximised: the first part on the right hand side of the formula does not depend on the parameters of the model, it is only the $\log\hat{\sigma}^2$ that does. So, the maximum of the concentrated log-likelihood (16.10) is obtained, when $\hat{\sigma}^2$ is minimised, implying the minimisation of MSE, which is the mechanism behind the "Ordinary Least Squares" (OLS from Section 10.1) estimation method. By doing this, we have just demonstrated that if we assume normality in the model, then the estimates of its parameters obtained via the maximisation of the likelihood coincide with the values obtained from OLS. So, why bother with MLE, when we have OLS?

First, the finding above holds for the Normal distribution only. If we assume a different distribution, we would get different estimates of parameters. In some cases, it might not be possible or reasonable to use OLS, but MLE would be a plausible option (for example, logistic, Poisson and any other non-standard model).

Second, the MLE of parameters have good statistical properties: they are consistent (Subsection 6.3.3) and efficient (Subsection 6.3.2). These properties hold almost universally for many likelihoods under very mild conditions. Note that the MLE of parameters are not necessarily unbiased (Subsection 6.3.1), but after estimating the model, one can de-bias some of them (for example, calculate the standard deviation of the error via division of the sum of squared errors by the number of degrees of freedom $n-k$ instead of n as discussed in Section 11.2).

Third, likelihood can be used for the model assessment, even when the standard statistics, such as R^2 or F-test are not available. We do not discuss these aspects in this textbook, but interested reader is directed to the topic of likelihood ratios.

Finally, likelihood permits the model selection (which will be discussed in Section **??**) via information criteria. In general, this is not possible to do unless you

assume a distribution and maximise the respective likelihood. In some statistical literature, you can notice that information criteria are calculated for the models estimated via OLS, but what the authors of such resources do not tell you is that there is still an assumption of normality behind this (see the link between OLS and MLE of Normal distribution above).

Note that the likelihood approach assumes that all parameters of the model are estimated, including location, scale, shape, shift of distribution etc. So typically it has more parameters to estimate than, for example, the OLS. This is discussed in some detail later in the Section 16.3.

16.3 Calculating number of parameters in models

When performing model selection and calculating different statistics, it is important to know how many parameters were estimated in the model. While this might seems trivial there are a number of edge cases and wrinkles that are seldom discussed in detail.

When it comes to inference based on regression models, the general idea is to calculate the number of **all the independent estimated parameters** k. This typically includes all initial components and all coefficients of the model together with the scale, shape and shift parameters of the assumed distribution (e.g. variance in the Normal distribution).

Example 16.1. In a simple regression model: $y_j = \beta_0 + \beta_1 x_j + \epsilon_j$ - assuming Normal distribution for ϵ_j, using the MLE will result in the estimation of $k = 3$: the two parameters of the model (β_0 and β_1) and the variance of the error term σ^2.

If likelihood is not used, then the number of parameters might be different. For example, if we estimate the model via the minimisation of MSE (similar to OLS), then the number of all estimated parameters does not include the variance anymore - it is obtained as a by product of the estimation. This is because the likelihood needs to have all the parameters of distribution in order to be maximised, but with MSE, we just minimise the mean of squared errors, and the variance of the distribution is obtained automatically. While the values of parameters might be the same, the logic is slightly different.

Example 16.2. This means that for the same simple linear regression, estimated using OLS, the number of parameters is equal to 2: estimates of β_0 and β_1.

Remark. For the calculation of information criteria, the number of parameters in the example above should be still considered 3 (parameters and scale). See explanation in Section 16.4.

In addition, all the restrictions on the parameters can reduce the number of estimated parameters, when they get to the boundary values.

Example 16.3. If we know that the parameter β_1 lies between 0 and 1, and in the estimation process it gets to the value of 1 (due to how the optimiser works), it can be considered as a restriction $\beta_1 = 1$. So, when estimated via the minimum of MSE with this restriction, this would imply that $k = 1$.

In general, if a parameter is provided in the model, then it does not count towards the number of all estimated parameters. So, setting $b_1 = 1$ acts in the same fashion.

Finally, if a parameter is just a function of another one, then it does not count towards the k as well.

Example 16.4. If we know that in the same simple linear regression $\beta_1 = \frac{\beta_0}{\sigma^2}$, then the number of all the estimated parameter via the maximum likelihood is 2: β_0 and σ^2.

We will come back to the number of parameters later in this textbook, when we discuss specific models.

A final note: typically, the standard maximum likelihood estimators for the scale, shape and shift parameters are biased in small samples and do not coincide with the OLS estimators. For example, in case of Normal distribution, OLS estimate of variance has $n - k$ in the denominator, while the likelihood one has just n. This needs to be taken into account, when the variance is used in forecasting.

16.4 Information criteria

There are different ways how to select the most appropriate model for the data. One can use judgment, statistical tests, cross-validation or meta learning. The state of the art one in the field of exponential smoothing relies on the calculation of information criteria and on selection of the model with the lowest value. This approach is discussed in detail in Burnham and Anderson (2004). Here we briefly explain how this approach works and what are its advantages and disadvantages.

16.4.1 The idea

Before we move to the mathematics and well-known formulae, it makes sense to understand what we are trying to do, when we use information criteria. The idea is that we have a pool of model under consideration, and that there is a true model somewhere out there (not necessarily in our pool). This can be presented graphically in the following way:

This plot 16.8 represents a space of models. There is a true one in the middle, and there are four models under consideration: Model 1, Model 2, Model 3 and Model

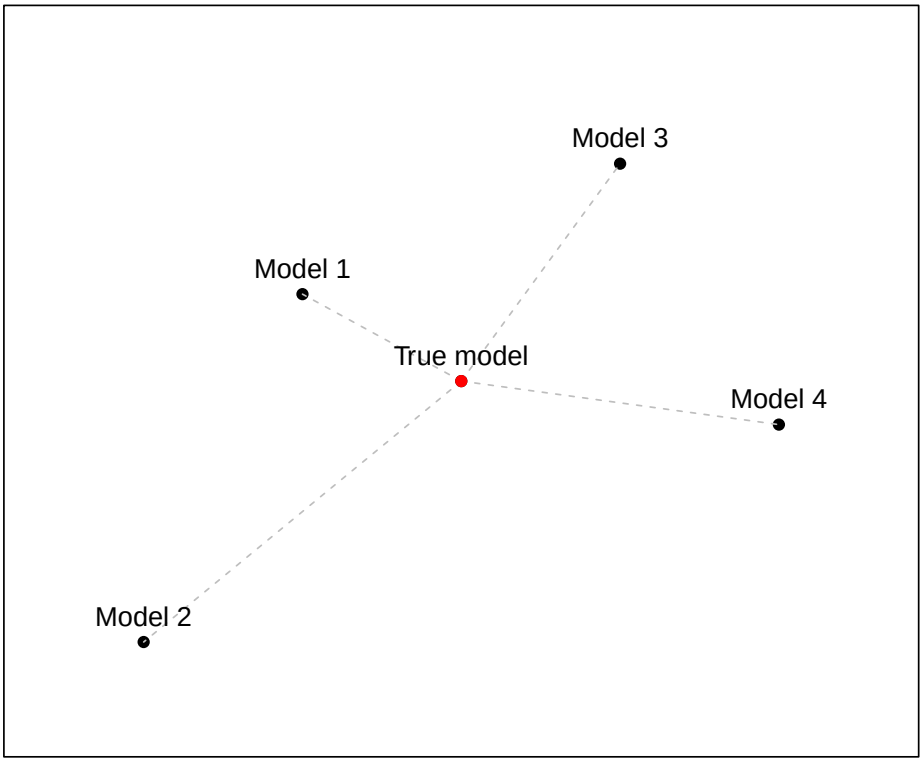


Figure 16.8: An example of a model space

4. They might differ in terms of functional form (additive vs. multiplicative), or in terms of included/omitted variables. All models are at some some distance (the grey dashed lines) from the true model in this hypothetic model space: Model 1 is closest while Model 2 is farthest. Models 3 and 4 have similar distances to the truth.

In the model selection exercise what we typically want to do is to select the model closest to the true one (Model 1 in our case). This is easy to do when you know the true model: just measure the distances and select the closest one. This can be written very roughly as:

$$\begin{aligned} d_1 &= \ell^* - \ell_1 \\ d_2 &= \ell^* - \ell_2 \\ d_3 &= \ell^* - \ell_3 \\ d_4 &= \ell^* - \ell_4 \end{aligned}, \tag{16.11}$$

where ℓ_j is the position of the j^{th} model and ℓ^* is the position of the true one. One of ways of getting the position of the model is by calculating the log-likelihood (logarithms of likelihood) values for each model, based on the assumed distributions. The likelihood of the true model will always be fixed, so if it is known it just comes to calculating the values for the models 1 - 4, inserting them in the equations in (16.11), and selecting the model that has the lowest distance d_j.

In reality, however, we *never* know the true model. We therefore need to find some other way of measuring the distances. The neat thing about the maximum likelihood approach is that the true model has the highest possible likelihood by definition! This means that it is not important to know ℓ^* – it will be the same for all the models. So, we can drop the ℓ^* in the formulae (16.11) and compare the models via their likelihoods ℓ_1, ℓ_2, ℓ_3 and ℓ_4 alone:

$$\begin{aligned} d_1 &= -\ell_1 \\ d_2 &= -\ell_2 \\ d_3 &= -\ell_3 \\ d_4 &= -\ell_4 \end{aligned}, \tag{16.12}$$

This is a very simple method that allows us to get to the model closest to the true one in the pool. However, we should not forget that we usually work with samples of data instead of the entire population and correspondingly will have only *estimates* of likelihoods and not the true ones. Inevitably, they will be biased and will need to be corrected. Akaike (1974) showed that the bias can be corrected if the number of parameters in each model is added to the distances (16.12) resulting in the bias corrected formula:

$$d_j = k_j - \ell_j, \tag{16.13}$$

where k_j is the number of estimated parameters in model j (this typically includes scale parameters when dealing with Maximum Likelihood Estimates).

Remark. If your model is estimated using OLS then you can use the idea that the maximum of the likelihood of the normal distribution is achieved by the minimum of the MSE (OLS criterion). In that case you can jump from one thing to another and calculate AIC by inserting the estimates of parameters to the likelihood formula. However, there are two things to consider:

1. The scale of distribution used in likelihood needs to be based on (16.9), i.e. without the bias correction, because it maximises the likelihood;
2. The number of parameters k_j should also include the scale of distribution. So, if you had a linear regression model, which estimated 3 parameters (intercept and two coefficients for explanatory variables) then $k_j = 3+1 = 4$.

Akaike (1974) suggests "An Information Criterion" which multiplies both parts of the right-hand side of (16.13) by 2 so that there is a correspondence between the criterion and the well-known likelihood ratio test (Wikipedia, 2020c):

$$\text{AIC}_j = 2k_j - 2\ell_j. \tag{16.14}$$

This criterion now more commonly goes by the "Akaike Information Criterion".

Various alternative criteria motivated by similar ideas have been proposed. The following are worth mentioning:

- AICc (Sugiura, 1978), which is a sample corrected version of the AIC for normal and related distributions, which takes the number of observations into account:
$$\text{AICc}_j = 2\frac{n}{n-k_j-1}k_j - 2\ell_j, \tag{16.15}$$
where n is the sample size.
- BIC (Schwarz, 1978) (aka "Schwarz criterion"), which is derived from Bayesian statistics:
$$\text{BIC}_j = \log(n)k_j - 2\ell_j. \tag{16.16}$$
- BICc (McQuarrie, 1999) - the sample-corrected version of BIC, relying on the assumption of normality:
$$\text{BICc}_j = \frac{n\log(n)}{n-k_j-1}k_j - 2\ell_j. \tag{16.17}$$

In general, the use of the sample-corrected versions of the criteria (AICc, BICc) is recommended unless sample size is very large (thousands of observations), in which case the effect of the number of observations on the criteria becomes negligible. The main issue is that corrected versions of information criteria for non-normal distributions need to be derived separately and will differ from (16.15) and (16.17). Still, Burnham and Anderson (2004) recommend using formulae (16.15) and (16.17) in small samples even if the distribution of variables is not normal and the correct formulae are not known. The motivation for this

is that the corrected versions still take sample size into account, correcting the sample bias in criteria to some extent.

A thing to note is that the approach relies on asymptotic properties of estimators and assumes that the estimation method used in the process guarantees that the likelihood functions of the models are maximised. In fact, it relies on asymptotic behaviour of parameters, so it is not very important whether the maximum of the likelihood in sample is reached or not or whether the final solution is near the maximum. If the sample size changes, the parameters guaranteeing the maximum will change as well so we cannot get the point correctly in sample anyway. However, it is much more important to use an estimation method that will guarantee consistent maximisation of the likelihood. This implies that we might select wrong models in some cases in sample, but that is okay, because if we use the adequate approach for estimation and selection, with the increase of the sample size, we will select the correct model more often than an incorrect one. While the "increase of sample size" might seem as an unrealistic idea in some real life cases, keep in mind that this might mean not just the increase of n, but also the increase of the number of series under consideration. So, for example, the approach should select the correct model on average, when you test it on a sample of 10,000 SKUs.

Summarising, the idea of model selection via information criteria is to:

1. form a pool of competing models,
2. construct and estimate them,
3. calculate their likelihoods,
4. calculate the information criteria,
5. and finally, select the model that has the lowest value under the information criterion.

This approach is relatively fast (in comparison with cross-validation, judgmental selection or meta learning) and has good theory behind it. It can also be shown that for normal distributions selecting time series models on the basis of AIC is asymptotically equivalent to the selection based on leave-one-out cross-validation with MSE. This becomes relatively straightforward, if we recall that typically time series models rely on one step ahead errors ($e_t = y_t - \mu_{t|t-1}$) and that the maximum of the likelihood of Normal distribution gives the same estimates as the minimum of MSE.

As for the disadvantages of the approach, as mentioned above, it relies on the in-sample value of the likelihood, based on one step ahead error, and does not guarantee that the selected model will perform well for the holdout for multiple steps ahead. Using the cross-validation or rolling origin for the full horizon could give better results if you suspect that information criteria do not work. Furthermore, any criterion is random on its own, and will change with the sample This means that there is model selection uncertainty and that which model is best might change with new observations. In order to address this issue, combinations of models can be used, which allows mitigating this uncertainty.

16.4.2 Common confusions related to information criteria

Similar to the discussion of hypothesis testing, I have decided to collect common mistakes and confusions related to information criteria. Here they are:

1. "AIC relies on Normal distribution".

- This is not correct. AIC relies on the value of maximised likelihood function. It will use whatever you provide it, so it all comes to the assumptions you make. Having said that, if you use the sample corrected versions of information criteria, such as AICc or BICc, then you should keep in mind that the formulae (16.15) and (16.17) are derived for Normal distribution. If you use a different one (not related to Normal, so not Log Normal, Box-Cox Normal, Logit Normal etc), then you would need to derive AICc and BICc for it. Still Burnham and Anderson (2004) argue that even if you do not have the correct formula for your distribution, using (16.15) and (16.17) is better than using the non-corrected versions, because there is at least some correction of the bias caused by sample size.

2. "We have removed outlier from the model, AIC has decreased".

- AIC will always decrease if you decrease the sample size and fit the model with the same specification. This is because likelihood function relies on the joint PDF of all observations in sample. If the sample decreases, the likelihood increases. This effect is observed not only in cases, when outliers are removed, but also in case of taking differences of the data. So, when comparing models, make sure that they are constructed on exactly the same data.

3. "We have estimated model with logarithm of response variable, and AIC has decreased" (in comparison with the linear one).

- AIC is comparable only between models with the same response variable. If you transform the response variable, you inevitably assume a different distribution. For example, taking logarithm and assuming that error term follows normal distribution is equivalent to assuming that the original data follows log-normal distribution. If you want to make information criteria comparable in this case, either estimate the original model with a different distribution or transform AIC for the multiplicative model.

4. "We have used quantile regression, assuming normality and AIC is..."

- Information criteria only work, when the likelihood with the assumed distribution is maximised, because only then it can be guaranteed that the estimates of parameters will be consistent and efficient. If you assume normality, then you either need to maximise the respective likelihood or minimise MSE - they will give the same solution. If you use quantile regression, then you should use likelihood of Asymmetric Laplace. If you estimate parameters via minimisation of MAE, then Laplace distribution of residuals is a suitable assumption for your model. In the cases when

distribution and loss are not connected, the selection mechanism might break and not work as intended.

Chapter 17

Uncertainty about the model form

In this Chapter, we discuss more advanced topics related to regression modelling. In a way, this part builds upon elements of Statistical Learning (see, for example, the textbook of Hastie et al., 2009) and focuses on how to select variables for regression model. We start with a fundamental idea of bias-variance tradeoff, which lies in the core of many selection methods. We then move to the discussion of information criteria, explaining what they imply, after that - to several existing variable selection approaches, explaining their advantages and limitations. Furthermore, we discuss combination approaches and what they mean in terms of parameters of models. We finish this chapter with an introductory discussion of regularisation techniques (such as LASSO and RIDGE).

17.1 Bias-variance tradeoff

17.1.1 Graphical explanation

In order to better understand, why we need to bother with model selection, combinations and advanced estimators, we need to understand the principle of bias-variance tradeoff. Consider a simple example of relation between fuel consumption of a car and the engine size based on the `mtcars` dataset in R (Figure 17.1):

The plot in Figure 17.1 demonstrates clear non-linearity. Indeed, we would expect the relation between these variables to be non-linear in real life: it is difficult to imagine the situation, where a car with no engine will be able to drive at all. On the other hand, a car with a huge engine will still be able to drive some distance, although probably very small. The linear model would assume

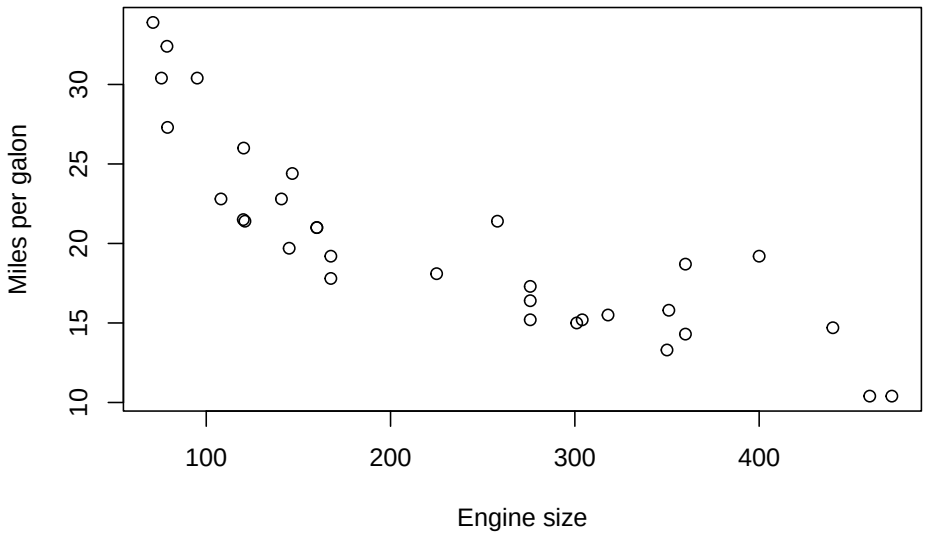


Figure 17.1: Fuel consumption vs engine size

that the "no engine" case would correspond to the value of approximately 30 miles per gallon (the intersection with y-axis), while the case of "huge engine" would probably result in negative mileage. So, the theoretically suitable model should be multiplicative, which for example can be formulated in logarithms:

$$\log mpg_j = \beta_0 + \beta_1 \log disp_j + \epsilon_j, \tag{17.1}$$

where mpg_j is the miles per galon of a car and $disp_j$ is the displacement (size) of engine. We will assume for now that this is the "true model", which would fit the data in the following way if we knew all the data in the universe (Figure 17.2):

While being wrong, we could still use the linear model to capture some relations in some parts of the data. It would not be a perfect fit (and would have some issues in the tails of our data), but it would be an acceptable approximation of the true model in some situations. If we vary the sample, we will see how the model would behave, which would help us in understanding of the uncertainty associated with it (Figure 17.3).

Figure 17.3 demonstrates the situation, where the linear model was fit to randomly peaked sub-samples of the original data. We can see that the linear model would exhibit some sort of bias in comparison with the true one: it is consistently above the true model in the region in the middle and is consistently below it in the tails of the sample.

Alternatively, we could fit a high order polynomial model to approximate the data and repeat the same procedure with sub-samples as before (Figure 17.4):

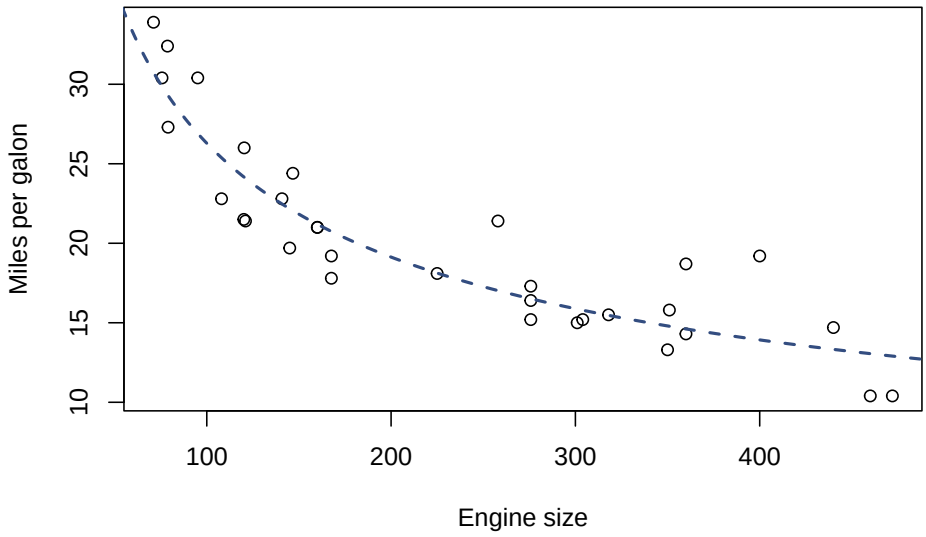


Figure 17.2: Fuel consumption vs engine size and the true model

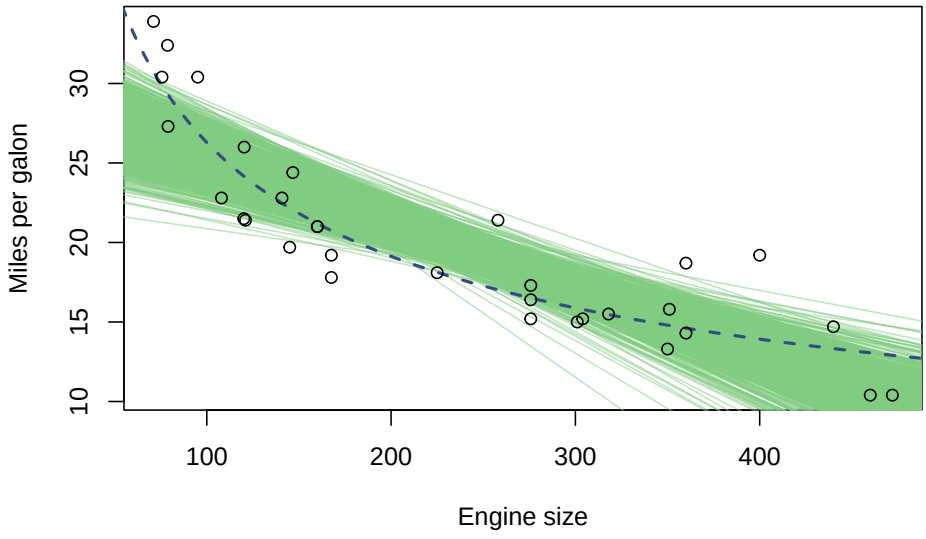


Figure 17.3: Fuel consumption vs engine size, the true and the linear models

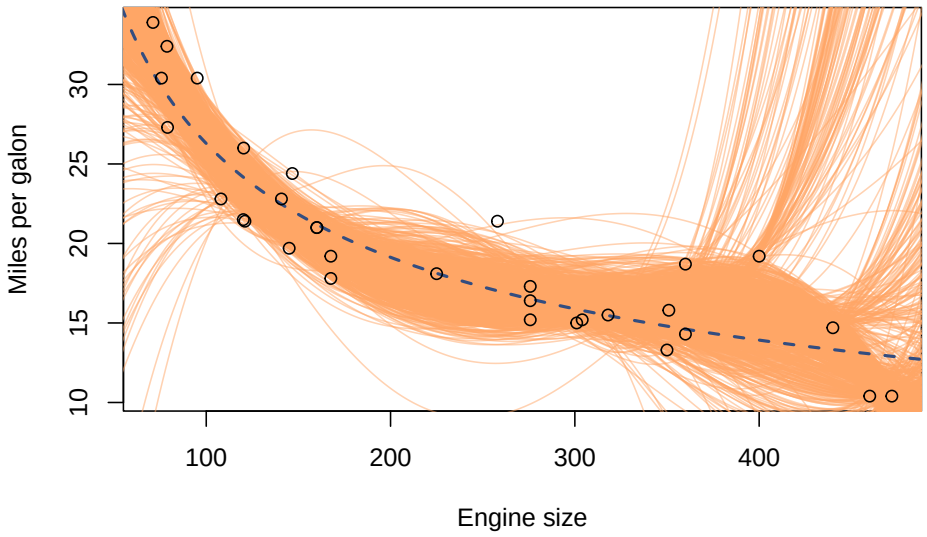


Figure 17.4: Fuel consumption vs engine size, the true, the linear and the polynomial models.

The new polynomial model on the plot in Figure 17.4 has a lower bias than the linear one, because on average it is closer to the true model in sample, getting closer to the green line (true model) even in the tails. However, it is also apparent that it has higher variability. This is because it is a more complex model than the linear one: it includes more variables (polynomial terms), making it more sensitive to specific observations in the sample. If we were to introduce even more polynomial terms, the model would have even more variance around the true model than before (Figure 17.5).

The model on plot in Figure 17.5 exhibits even higher variance in comparison with the true model, but it is still less biased than the linear model. The pattern that we observe in this demonstration is that the variance of the model in comparison with the true one increases with the increase of complexity, while the bias either decreases or does not change substantially. This is bias-variance tradeoff in action. It is the principle that states that with the increase of complexity of model, its variance (with respect to the true one) increases, while the bias decreases. This implies that typically you cannot minimise both variance and bias at the same time - depending on how you formulate and estimate a model, it will either have bigger variance or a bigger bias. This principle can be applied not only to models, but also to estimates of parameters or to forecasts from the models. It is one of the fundamental basic modelling principles.

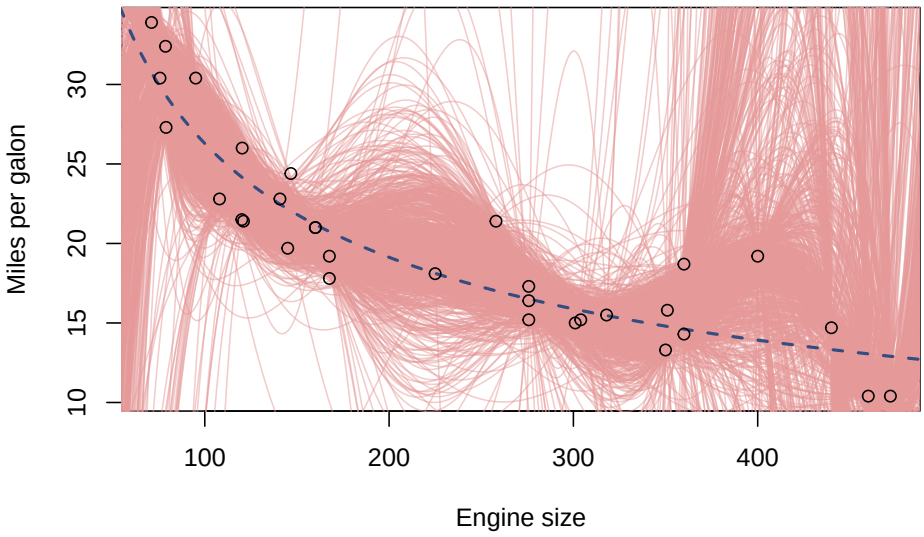


Figure 17.5: Fuel consumption vs engine size, the true and the polynomial (7th order) models

17.1.2 Mathematical explanation

Mathematically, it is represented for an estimate as parts of Mean Squared Error (MSE) of that estimate (we drop the index of observations j in $\hat{y}_j$ for convenience):

$$\text{MSE} = \text{Bias}(\hat{y})^2 + \text{V}(\hat{y}) + \sigma^2, \tag{17.2}$$

where $\hat{y}$ is the fitted value of our model, $\text{Bias}(\hat{y}) = \text{E}(\mu_y - \hat{y})$, $\text{V}(\hat{y}) = \text{E}\left((\mu_y - \hat{y})^2\right)$, μ_y is the fitted value of the true model, and σ^2 is the variance of the white noise of the true model.

Proof. The Mean Squared Error of a model by definition is the expectation of the squared difference between the actual and the fitted values:

$$\text{MSE} = \text{E}\left((y - \hat{y})^2\right) = \text{E}\left((\mu_y + \epsilon - \hat{y})^2\right) \tag{17.3}$$

The expectation of square of sum above can be expanded as:

$$\begin{aligned} \text{MSE} =& \text{E}\left((\mu_y + \epsilon - \hat{y} + \text{E}(\hat{y}) - \text{E}(\hat{y}))^2\right) = \\ & \text{E}\left((\mu_y - \text{E}(\hat{y}))^2\right) + \\ & \text{E}\left((\text{E}(\hat{y}) - \hat{y})^2\right) + \\ & \text{E}(\epsilon^2) + \\ & 2\text{E}\left((\mu_y - \text{E}(\hat{y}))\epsilon\right) + \\ & 2\text{E}\left((\text{E}(\hat{y}) - \hat{y})\epsilon\right) + \\ & 2\text{E}\left((\mu_y - \text{E}(\hat{y}))(\text{E}(\hat{y}) - \hat{y})\right) \end{aligned} \tag{17.4}$$

We now can consider each element of the sum of squares in (17.4) to understand what they are equal to. We start with the first one, which can be expanded to:

$$\mathrm{E}\left((\mu_y - \mathrm{E}(\hat{y}))^2\right) = \mathrm{E}(\mu_y^2) - 2\mathrm{E}(\mu_y \mathrm{E}(\hat{y})) + \mathrm{E}(\mathrm{E}(\hat{y})^2). \tag{17.5}$$

Given that μ_y is the value of the model in the population, which is fixed, its expectation will be equal to itself. In addition, the expectation of expectation is just an expectation: $\mathrm{E}(\mathrm{E}(\hat{y})^2) = \mathrm{E}(\hat{y})^2$. This leads to the following, which is just the bias of the estimated model:

$$\begin{aligned} \mathrm{E}\left((\mu_y - \mathrm{E}(\hat{y}))^2\right) =& \mu_y^2 - 2\mu_y \mathrm{E}(\hat{y}) + \mathrm{E}(\hat{y})^2 = \\ & (\mu_y - \mathrm{E}(\hat{y}))^2 = \mathrm{Bias}(\hat{y})^2 \end{aligned} \tag{17.6}$$

The second term in (17.4) is the variance of the model (by the definition of variance):

$$\mathrm{E}\left((\mathrm{E}(\hat{y}) - \hat{y})^2\right) = \mathrm{V}(\hat{y}) \tag{17.7}$$

The third term is equal to the variance of the error term as long as the expectation of the error is zero (which is one of the conventional assumptions, discussed in Subsection 15.2.3):

$$\mathrm{E}(\epsilon^2) = \sigma^2 \tag{17.8}$$

The more complicated thing is to show that the the other three elements are equal to zero. For the elements number four and five, we can use the assumption that the error term of the true model is independent of anything else (see discussion in Section 15.2), leading respectively to:

$$\mathrm{E}\left((\mu_y - \mathrm{E}(\hat{y}))\epsilon\right) = \mathrm{E}(\mu_y - \mathrm{E}(\hat{y})) \times \mathrm{E}(\epsilon) \tag{17.9}$$

and

$$\mathrm{E}\left((\mathrm{E}(\hat{y}) - \hat{y})\epsilon\right) = \mathrm{E}(\mathrm{E}(\hat{y}) - \hat{y}) \times \mathrm{E}(\epsilon). \tag{17.10}$$

Given that $\mathrm{E}(\epsilon) = 0$ due to one of the assumptions (Subsection 15.2.3), both terms will be equal to zero. Finally, the last term can be expanded to:

$$\begin{aligned} \mathrm{E}\left((\mu_y - \mathrm{E}(\hat{y}))(\mathrm{E}(\hat{y}) - \hat{y})\right) =& \mathrm{E}(\mu_y \mathrm{E}(\hat{y}) - \mu_y \hat{y} - \mathrm{E}(\hat{y})\mathrm{E}(\hat{y}) + \mathrm{E}(\hat{y})\hat{y}) = \\ & \mu_y \mathrm{E}(\hat{y}) - \mu_y \mathrm{E}(\hat{y}) - \mathrm{E}(\hat{y})^2 + \mathrm{E}(\hat{y})^2 = \\ & 0 \end{aligned} \tag{17.11}$$

So, this means that the last three terms in (17.4) are equal to zero and thus, inserting (17.5), (17.6) and (17.7) in (17.4), we get:

$$\mathrm{MSE} = \mathrm{Bias}(\hat{y})^2 + \mathrm{V}(\hat{y}) + \sigma^2$$

□

The similar mathematical formula holds for any other estimate, for example for an estimate of parameter. What this formula tells us is that there are two forces in the MSE of estimate that impact its value. Minimisation MSE does not imply

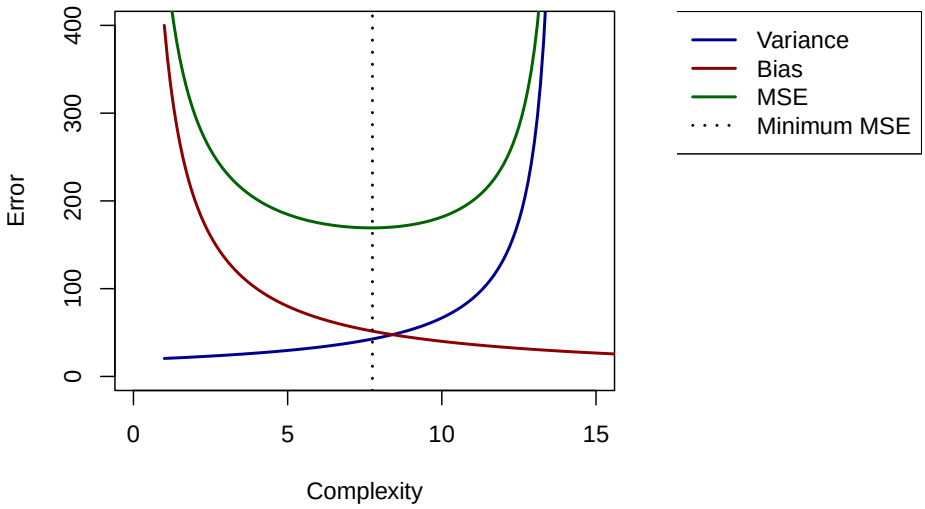


Figure 17.6: Bias, Variance and MSE as functions of model complexity.

that we reduce both of them, but most probably we are reducing one at the cost of another. The classical plot based on this looks as shown in Figure 17.6.

This plot shows the basic principles: with increase of complexity, the bias of estimate decreases, but its variance increases. There is typically the specific type of model (the point on Complexity axis) that minimises MSE (the vertical line on the plot), which will have some combination of variance and bias. But in practice, this point does not guarantee that we will have an accurate adequate model. In some situations we might prefer moving to the left on the plot in Figure 17.6, sacrificing unbiasedness of model to get the reduced variance. The model selection, model combinations and regularisation methods all aim to find the sweet spot on the plot for the specific sample available to the analyst.

The bias-variance trade-off is also related to the discussion we had in Subsection 6.3.5: having a more biased but more efficient estimate of parameter (thus going to the left in Figure 17.6) might be more desirable than having the unbiased but inefficient estimate. This is because on small samples the former estimate will be typically closer to the true value than the latter one. This explains why model selection, model combinations and regularisation are important topics and have become so popular over the last few decades: they all allow deciding, where to be in the bias-variance trade-off in order to produce more accurate estimates of parameters based on the available sample of data.

Bibliography

, ???? Data.
URL https://dictionary.cambridge.org/dictionary/english/data

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Burnham, K. P., Anderson, D. R., 2004. Model Selection and Multimodel Inference. Springer New York.

Chatfield, C., 1996. Model uncertainty and forecast accuracy. Journal of Forecasting 15 (7), 495–508.

Cohen, J., 1994. The earth is round ($p < .05$). American Psychologist 49 (12), 997–1003.

Goodman, L. A., Dec. 1952. Serial Number Analysis. Journal of the American Statistical Association 47 (260), 622–634.

Goodman, L. A., 1954. Some practical techniques in serial number analysis. Journal of the American Statistical Association 49 (265), 97–112.

Hanck, C., Arnold, M., Gerber, A., Schmelzer, M., 2022. Introduction to Econometrics with R. Bookdown, (version: 2022-04-17).
URL https://www.econometrics-with-r.org/index.html

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics. Springer New York, New York, NY.

Kotz, S., Balakrishnan, N., Read, C. B., Vidakovic, B., 2005. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc., Hoboken, New Jersey.

McQuarrie, A. D., 1999. A small-sample correction for the Schwarz SIC model selection criterion. Statistics {&} Probability Letters 44 (1), 79–86.

O'Hagan, A., Stevens, J. W., Campbell, M. J., 2005. Assurance in clinical trial design 4, 187–201.

Pidd, M., 2010. Why modelling and model use matter. Journal of the Operational Research Society 61 (1), 14–24.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6 (2), 461–464.

Sugiura, N., 1978. Further analysis of the data by akaike' s information criterion and the finite corrections. Communications in Statistics - Theory and Methods 7 (1), 13–26.

Svetunkov, I., 2021. Forecasting and analytics with adam. OpenForecast, (version: 2021-07-30).
URL https://openforecast.org/adam/

Svetunkov, I., 2025. greybox: Toolbox for Model Building and Forecasting. R package version 2.0.4.41004.
URL https://github.com/config-i1/greybox

Svetunkov, I., Boylan, J. E., 2019. Multiplicative state-space models for intermittent time series.

Wasserstein, R. L., Lazar, N. A., 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. American Statistician 70 (2), 129–133.

Wikipedia, 2020a. Bias of estimator: Sample variance. Wikipedia, (version: 2020-08-12).
URL https://en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance

Wikipedia, 2020b. Efficiency (statistics): Asymptotic efficiency. Wikipedia, (version: 2020-08-12).
URL https://en.wikipedia.org/wiki/Efficiency_(statistics)#Asymptotic_efficiency

Wikipedia, 2020c. Likelihood-ratio test. Wikipedia, (version: 2020-09-04).
URL https://en.wikipedia.org/wiki/Likelihood-ratio_test

Wikipedia, 2021. Anscombe's quartet. Wikipedia, (version: 2021-06-23).
URL https://en.wikipedia.org/wiki/Anscombe%27s_quartet