Analyzing data using linear models

Stéphanie van den Berg

Versie 0.1
(June 9, 2018)

Abstract

This book is intended to be of use to bachelor students in social sciences that want to learn how to analyze their data, with the specific aim to answer research questions. The book has a practical take on data analysis: how to do it, how to interpret the results, and how to report the results. All techniques are presented within the framework of linear models: this includes simple regression models, to linear mixed models, and generalized linear models. All methods can be carried out within one supermodel: the generalized linear mixed model. This approach is illustrated using SPSS.

Contents

Chapter 1

Exploring your data

1.1 Types of variables

Data analysis is about variables. In linear models there are different kinds of variables. One important distinction is between dependent variables and independent variables. The other important distinction is about the measurement level of the variable: continuous, ordinal or categorical.

1.1.1 Continuous, ordinal, and categorical variables

A typical example of a contiuous variable is age: in theory, you could calcualate your age in the number of minutes that have passed since your time of birth. It is continuous in the sense that it has an (almost) infinite number of possible values. For example, for two children born one minute a part, there could be a third child that was born just in between the other two. In practice of course, we measure age in days, and sometimes only in months in years, but given there are many values, we usually treat such an age variable in years as continuous. Other examples of continuous variables include height in inches, temperature in degrees Celcius, years of education, or systolic bloodpressure in millimeters of mercury. Note that in all these examples, quantities (age, height, temperature) are expressed as the number of a particlar unit (years, inches, degrees). Therefore continuous variables are often called quantitative variable, or quantitative measures. There is a further distinction into interval and ratio variables; this distinction is treated in the research methods course in Module 1.

With ordinal measures, there are no units. An example would be a variable that would quantify size, by stating whether a t-shirt is small, medium or large. Yes, there is a quantity here, size, but there is no unit to state EXACTLY how much of that quantity is available. Similar for age, we could code a number of people as young, middle-aged or old, but on the basis of such a variable we could not state by *how much* two individuals differ in age. Ordinal data are usually *discrete*: there are no infinite number of levels of the variable. It goes

up in discrete steps, for example, having values of 1, 2 and 3, and nothing in between.

Lastly, categorical variables are not about quantity at all. Categorical variables are about quality. A typical example of a categorical variable would be the colour of pencils: they can be either green, blue, black, white, red, yellow, etcetera. Nothing quantitative could be stated about a bunch of pencils that are only assessed regarding their colour, other than saying that a green pens are greener than other pens, and red pens are redder than other pens. There is usually no logical order in the values of such variables. Other examples include nationality (French, Turkish, Indian, other) or sex (male, female, other). Categorical variables are often called nominal variables, or qualitative variables.

Exercises

In the following, identify the type of variable in termes of continuous, ordinal discrete, or categorical:
Age: ... years
Weight: ... kilograms
Size: ... meters
Size: small, medium, large
Exercise intensity: low, moderate, high
Agreement: not agree, somewhat agree, agree
Agreement: totally not agree, somewhat not agree, neither disagree nor agree, somewhat agree, totally agree
Pain: 1, 2.. ..... , 99, 100
Quality of life: 1=extremely low, ..., ..., 7=extremely high
Colour: blue, green, yellow, other
Nationality: Chinese, Korean, Australian, Dutch, other
Gender: Female, Male, other
Gender: Female, Male Number of shoes:

1.1.2 Qualitative and quantitative treatment of variables in data analysis

There is a fundamental difference between continuous and ordinal variables, but it is possible to treat them the same way in data analysis. For data analysis with linear models, you have to decide for each variable whether you want to treat it as qualitative or quantitative. Continuous variables are always treated as quantitative. Categorical data are always treated as qualitative. The problem is with ordinal variables: you can either treat them as quantitative variables or as qualitative variables. The choice is usually based on common sense and whether the results are meaningful. For instance, if you have an ordinal variable with 8 levels, like a Likert scale, it usually does not make sense to treat it as qualitative. If the variable has only 3 levels, it is often meaningful to treat it as qualitative: assuming that the three levels can show qualitative differences.

In the coming chapters, we will come back to this distinction. Remember, in the coming chapters we will only speak of quantitative and qualitative treatment of variables, and remember that continuous variables are always treated as quantitative and categorical data are always treated as qualitative.

1.1.3 Dependent and independent variables

So now that we have discussed the distinction between continuous, ordinal and categorical variables, let's turn to dependent and independent variables. Determining whether a variable is treated as independent or not, is often either a case of logic or a case of theory. When studying the relationship between the height of a father and that if his child, the more logical it would be to see the height of the child **as a function** of the height of the father. This because we assume that the genes are transferred from the father to the child. The father comes first, and the height of the child is partly the *result* of the genes that were transmitted during fertilisation. Similarly, when predicting precipitation on the basis of the hours of sun light on the previous day, it seems natural to study the effect of hours of sunlight on the previous day on precipitation on the next day. That which is the result is usually taken as the dependent variable. The theoretical cause or antecedent is usually taken as the independent variable.
The dependent variable is often called the *response variable*. An independent variable is often called a *predictor variable* or simply *predictor*.

Examples: the effect of income on health
size is caused by inflation
size is influenced by weight
shoe size is predicted by sex

Exercises

From each of the following statements, identify the dependent variable and the independent variable:

The less you drink the more thirsty you become
The more calories you eat, the more you weigh
Weight is affected by food intake
Weight is affected by exercise
Food intake is predicted by time of year
There is an effect of exercise on heart rate
Inflation leads to higher wages
Unprotected sex leads to pregnancy
HIV-infection is caused by unprotected sex
The effect of alcohol intake on driving performance
Sunshine causes growth

1.2 Distributions

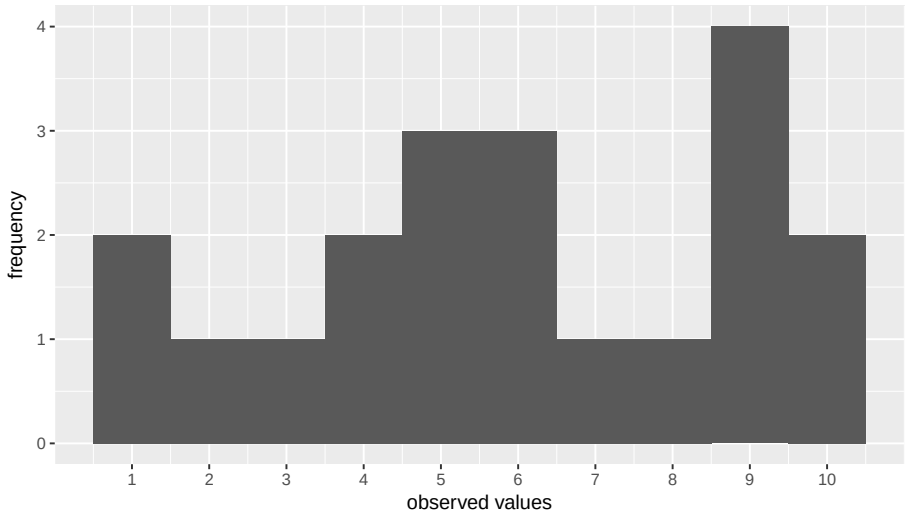


Figure 1.1: A frequency distribution

Variable have distributions. That means that if you put all the values you observed in order from low to high, you see a certain shape. For example, take the set of following numbers: 4, 8, 5, 9, 9, 1, 6, 9, 6, 5, 10, 5, 7, 6, 2, 9, 3, 1, 4, 10. If you plot these values on the horizontal axis, and how often they are observed (the frequency or count) on the y-axis you get the frequency plot in Figure 1.1.

Often a histogram is plotted. A histogram is very much like a frequency plot, except that its surface area adds up to one. For example, in Figure 1.1, the total area is equal to the total observed numbers, which is 20. If we divide all observed frequencies by 20, we get the plot in Figure 1.2.

This is called a histogram. We immediately see that 20% of the observations is a value of 9, and values of 5 make up 15% of the observations.

Figure 1.2 shows a histogram with 11 bins. Figure 1.3 we use the same data, but use only 5 bins: for the first bin, we take values of 1 and 2, for the second bin we take values 3 and 4 together, etcetera, until we take vales 9 and 10 for the fifth bin. For each bin, we compute how often we observe them and divide them by the total number of observations. Next we divide by the bin width. For example, we observe 4 nines and 2 tens, so 6 times a value of either 9 or 10. Dividing by the number of observations we get $6/20 = 0.3$. This proportion should be divided by 2, to get $0.33/0.15 =$. The binwidth is here 2: all values between 8.5 and 10.5 are taken to lie in the 5th bin. The distiance between these values is $10.5 - 8.5 = 2$. We have to divide the proportion by the binwidth because we want the total aree to sum to 1. For each bin, we take the binwidth and multiply it with its height (density), and sum these together.

When every observed value is unique, there is only one of it, then it's better

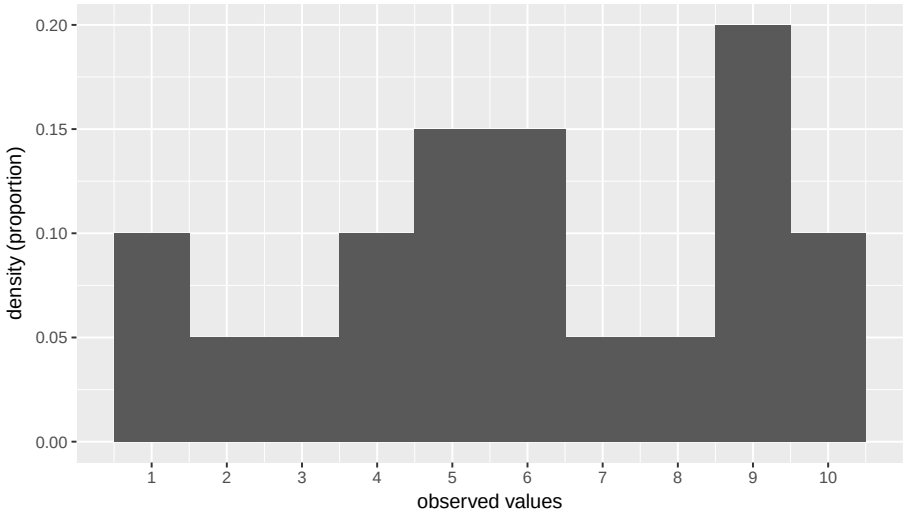


Figure 1.2: A histogram

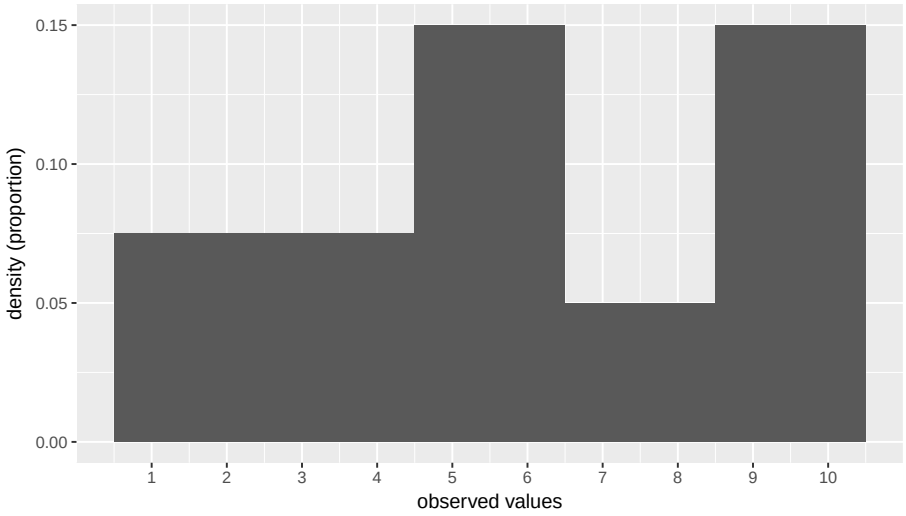


Figure 1.3: A histogram

to present a density: a line that shows how often values of more or less that value are observed, relative to other values.

Frequencies versus density.

uniform normal, z-scores briefly mention as examples: Student's t, chi-square, poisson

1.3 Mean, median and mode

1.3.1 The mean

The mean of set of values is the same as the average. Suppose we have the values 1, 2 and 3, then we compute the mean (or average) by first adding these numbers and then divide them by the number of values we have. In this case we have three values, so the mean is equal to $(1 + 2 + 3)/3 = 2$. In statistical formulas, the mean is indicated by a bar above the variable. So if our values of variable y are 1, 2 and 3, then we denote the mean by $\bar{y}$ (pronounced as y-bar). For taking the sum of a set of values, statistical formulas show a Σ (pronounced as sigma). So we often see the following formula for the mean of a set of n values for variable y:

$$\bar{y} = \frac{\Sigma_i^n y_i}{n} \tag{1.1}$$

In words, we take every value for y from 1 to n and sum them, and the result is divided by n.

1.3.2 The median

The mean is a measure of central tendency: if the mean is 100, it means the values tend to cluster around this value. A different measure of central tendency is the median. The median is nothing but the middle value. Suppose we have the values 45, 567, and 23. Then what value lies in the middle? Let's first order them from small to large to get a better look, then we get 23, 45 and 567. Then the value in the middle is of course 45.

Suppose we have the values 45, 45, 45, 65, and 23. What is the middle value? We first order them again and see what is in the middle: 23, 45, 45, 45 and 65. Obviously now 45 is the median.

What if we have two values in the middle? Suppose we have the values 46, 56, 45 and 34. If we order them we get 34, 45, 46 and 56. Now there are two values in the middle: 45 and 46. In that case, we take the average of these two middle values, so the median is 45.5.

1.3.3 The mode

A third measure of central tendency is the *mode*. The mode is defined as the value that we see most frequently in a series of values. For example, if we have

the series 4, 7, 5, 5, 6, 6, 6, 4, then the value observed most often is 6 (three times).

1.4 Variance

Suppose we measure the height of 3 children and their heights (in cms) are 120, 120, 120. There is no variation in height: all heights are the same. There are no differences. Then the average height is 120, the median height is 120, and the mode is 120.

Now suppose their heigths are 119, 120, 120. Now there are differences: one child is smaller than the other two, who have the same height. There is some variation now. We know how to quantify the mean, which is 119.6666667, we know how to quantify the median, which is 120, and we know how to quantify the mode, which is also 120. But how do we quantify the variation? Is there a lot of variation, or just a little, and how do we measure it?

One way you could think of is measuring the distance between the lowest value and the highest value. This we call the *range*. The lowest value is 119, and the highest value is 120, so the range of the data is equal to $120 - 119 = 1$. As another example, suppose we have the values 20, 20, 21, 20, 19, 20 and 454. Then the range is equal to $454 - 19 = 435$. That's a large range, for a series of values that for the most part hardly differ from another. Another measure for spread is *variance*, and variance is based on the *sum of squares*.

1.4.1 Sum of squares

What we call a sum of square is actually a sum of squared deviations. But deviations from what? First we have to know whether we are interested in the spread around what value. For instance we could be interested in how far the values 119.6666667 deviate from 0. The first differs 119, and the second and third differ 120. All values differ in a positive sense from 0: all values are positive. The deviations from zero are then 119, 120 and 120. Squaring these, we get the squared deviations, 119^2, 120^2 and 120^2 so 14161, 14400 and 14400. Adding these squared deviations, we obtain 42961 as the sum of squares.

We could also be interested in how much the values 119.6666667 vary around the *mean* of these values. The first value differs $119-119.6666667 = -0.6666667$, the second value differs $120-119.6666667 = 0.3333333$, and the third value also differs $120 - 119.6666667 = 0.3333333$.

Always when we look at deviations from the mean, some deviations are positive and some deviations will be negative (except when there is no variation). If we want to measure variation, it should not matter whether deviations are positive or negative: any deviation should add to the total variation in a postive way. So that is why we should better make all deviations positive, and this is done by taking the square of the deviations. So for our three values 119, 120 and 120, we get the deviations -0.67, +0.33 and +0.33, and if we square these

deviations, we get -0.67^2, $+0.33^2$ and $+0.33^2$, so -0.4489, 0.1089 and 0.1089. If we add these three squares, we obtain the sum $-0.67^2+0.33^2+0.33^2 = -0.2311$.

This is called the sum of squares, or SS. In most cases, the sum of squares refers to the sum of squared deviations from the mean. In brief, suppose you have n values of a variable y, you first take the mean of those values (this is $\bar{y}$), you subtract this mean from each of these n values $(y - \bar{y})$, then you take the squares of these deviations $((y - \bar{y})^2)$, and then add them toghether (take the sum of these squared deviations, $\Sigma(y-\bar{y})^2$). In formula form, this process looks like:

$$SS = \Sigma_i^n (y_i - \bar{y}) \tag{1.2}$$

As an example, suppose you have the values 10, 11 and 12, then the average value is 11. Then the deviations from the mean are -1, 0 and 1. If you square them you get 1, 0 and 1, and if you add these three values, you get $SS = 2$.

As another example, suppose you have the values 8, 10 and 12, then the average value is 10. Then the deviations from 10 are -2. 0 and +2. Taking the squares, you get 4, 0 and 4 and if you add them you get $SS = 8$.

Oftentimes, you are not interested in the total variation, but you're interesed in the average variation: how much does the avarage value differ from the mean? Suppose we have the values 10, 11 and 24. The mean is then $45/3 = 15$. Then we have two values that are smaller than the average and one value that is larger than the average, so two negative deviations and one positive deviation. Squaring them makes them all positive. The squared deviations are 25, 16, and 81. So the third value has a huge squared deviation (81) compared to the other two values. If we take the *average* squared deviation, we get $(25 + 16 + 81)/3 = 40.6666667$. So the average squared deviation is equal to 40.6666667. This we call the *variance*. So the variance of a bunch of values is nothing but the SS divided by the number of values, n. The variance is *the average squared deviation from the mean*. The symbol used for the variance is usually σ^2 (pronounced as "sigma squared").

$$\sigma^2 = \frac{SS}{n} = \frac{\Sigma_i^n (y_i - \bar{y})}{n} \tag{1.3}$$

As an example, suppose you have the values 10, 11 and 12, then the average value is 11. Then the deviations are -1, 0 and 1. If you square them you get 1, 0 and 1, and if you add these three values, you get $SS = 2$. If you divide this by 3, you get the variance: 0.67. Put differently, if the squared deviations are 1, 0 and 1, then the average squared deviation (i.e., the variance) is $\frac{1+0+1}{3} = 0.67$.

As another example, suppose you have the values 8, 10, 10 and 12, then the average value is 10. Then the deviations from 10 are -2, 0, 0 and +2. Taking the squares, you get 4, 0, 0 and 4 and if you add them you get $SS = 8$. To get the variance, you divide this by 4: $8/4 = 2$. Put differently, if the squared deviations are 4, 0, 0 and 4, then the average squared deviation (i.e., the variance) is $\frac{4+0+0+4}{4} = 2$.

Often we also see another measure of variation: the *standard deviation*. The standard deviation is nothing but the root of the variance and is therefore denoted as σ:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma_i^n (y_i - \bar{y})}{n}} \tag{1.4}$$

Chapter 2

Linear modelling: introduction

2.1 Linear equations

From secondary education you might remember linear equations. Suppose you have two quantities, x and y, and there is a straight line that describes best their relationship. An example is given in Figure 2.1

We see that for every value of x, there is only one value of y. Moreover, the larger the value of x, the larger the value of y. If we look more closely we see that for each increase of 1 in x, there is an increase of 2 in y. For instance, if x=1, we see a y value of 2, and if x=2 we see a y-value of 4. So if we move from x=1 to x=2 (a step of one on the x-axis), we move from 2 to 4 on the y-axis, which is an increase of 2.

This increase of 2 for every step of 1 in x is the same for all values of x and y. For instance, if we move from 9 to 10 on the x-axis, we go from 18 to 20 on the y-axis: an increase of 2. This constant increase is typical of linear relationships. The increase in y for every unit increase in x is called the *slope* of a straight line. In this figure, the slope is equal to 2.

The slope is one important characteristic of a straight line. The second important property of a straight line is the *intercept*. The intercept is the value of y, if $x = 0$. In Figure 2.1 we see that if $x = 0$, y is 0, too. Therefore the intercept of this straight line is 0.

With the intercept and the slope, we completely describe this straight line: no other information is necessary. Such a straight line describes a linear relationship between x and y. The linear relationship can be formalized using a linear equation. The general form of a linear equation for two variables x and y is the following:

$$y = intercept + slope \times x \tag{2.1}$$

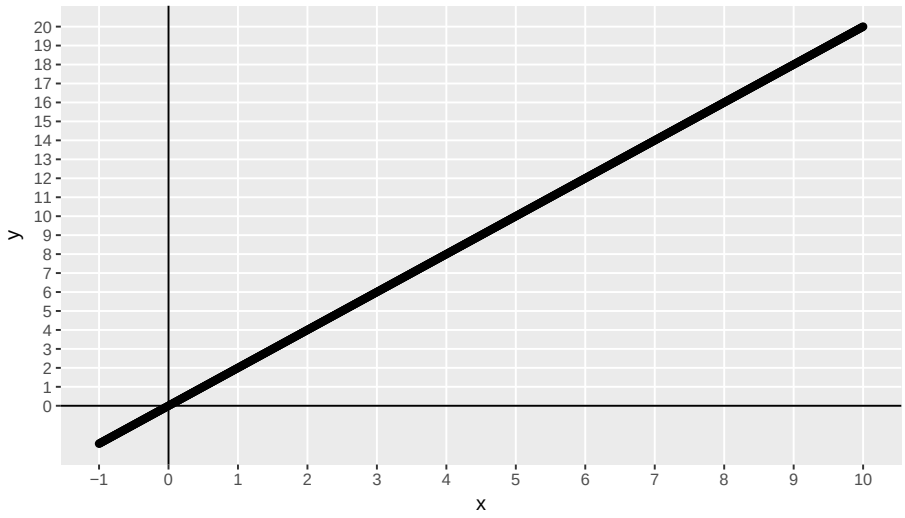


Figure 2.1: Straight line with intercept 0 and slope 2.

For the linear relationship between x and y in Figure 2.1 the linear equation is therefore

$$y = 0 + 2x \tag{2.2}$$

which can be simplified to

$$y = 2x \tag{2.3}$$

With this equation, we can find the y-value for all x-values. For instance, if we want to know the y-value for $x = 3.14$, then using the linear equation we know that $y = 2 \times 3.14 = 6.28$. If we want to know the y value for $x = 49876.6$, we use the equation to obtain $y = 2 \times 49876.6 = 99753.2$. In short, the linear equation is very helpful to quickly say what y-value is on the basis of the x-value, even if we don't have a graph of the relationship or if the graph does not extent to certain x-values.

Figure 2.2 shows a different linear relationship between x and y. First we look at the slope: we see that for every unit increase in x (from 1 to 2, or from 4 to 5) we see an increase of 0.5 in y. Therefore the slope is equal to 0.5. Second we look at the intercept: we see that if $x = 0$, y has the value -2. So the intercept is -2. Again, we can describe the linear relationship by a linear equation, which is now:

$$y = -2 + 0.5x \tag{2.4}$$

Linear relationships can also be negative, see Figure 2.3. There, we see that if we move from 0 to 1, we see a *decrease* of 2 in y (we move from y=-2 to y=-4),

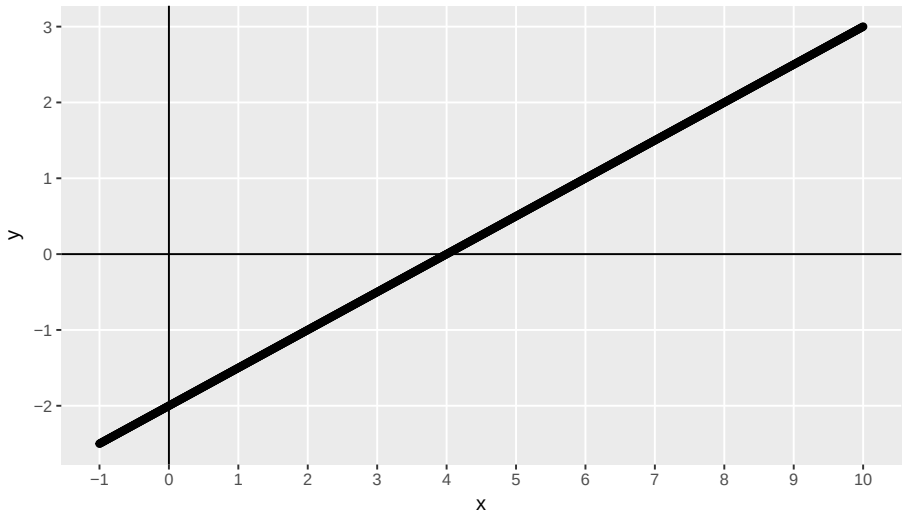


Figure 2.2: Straight line with intercept -2 and slope 0.5.

so that is our slope value. Further, if x=0, we see a y-value of -2, and that is our intercept. The linear equation is therefore:

$$y = -2 - 2x \tag{2.5}$$

2.1.1 Exercises

1. For Figures 2.4, 2.5 and 2.6, give the linear equations for the relationship between x and y.

2. Try to sketch the straight line for the equation $y = 1 - 2x$

Answers:
The equations are

$$y = 3 - 1x \tag{2.6}$$

$$y = 1.5 - 0.5x \tag{2.7}$$

$$y = -2 + 0.33x \tag{2.8}$$

The straight line for $y = 1 - 2x$ is presented in Figure 2.7.

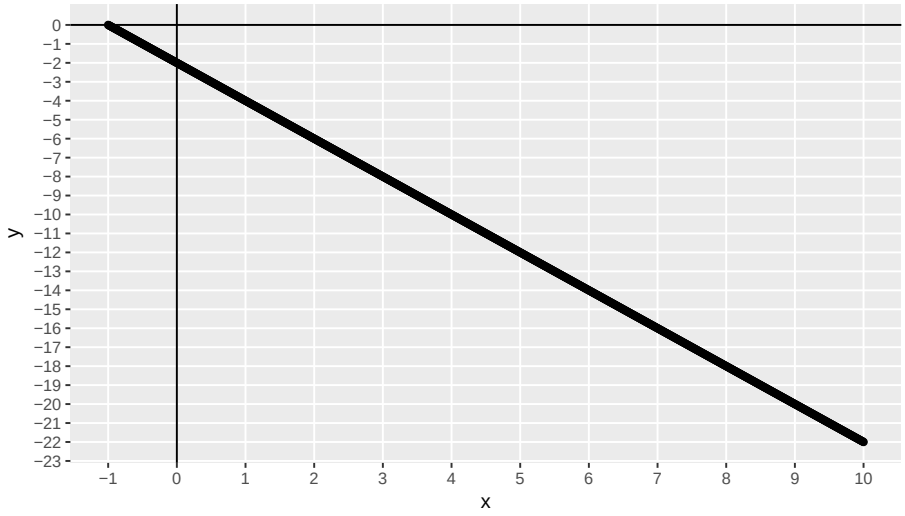


Figure 2.3: Straight line with intercept -2 and slope 0.5.

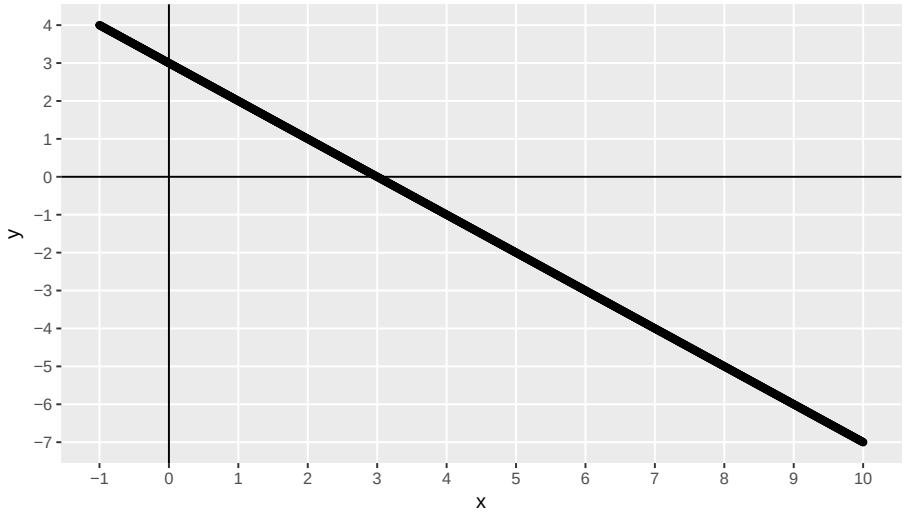


Figure 2.4: Straight line example.

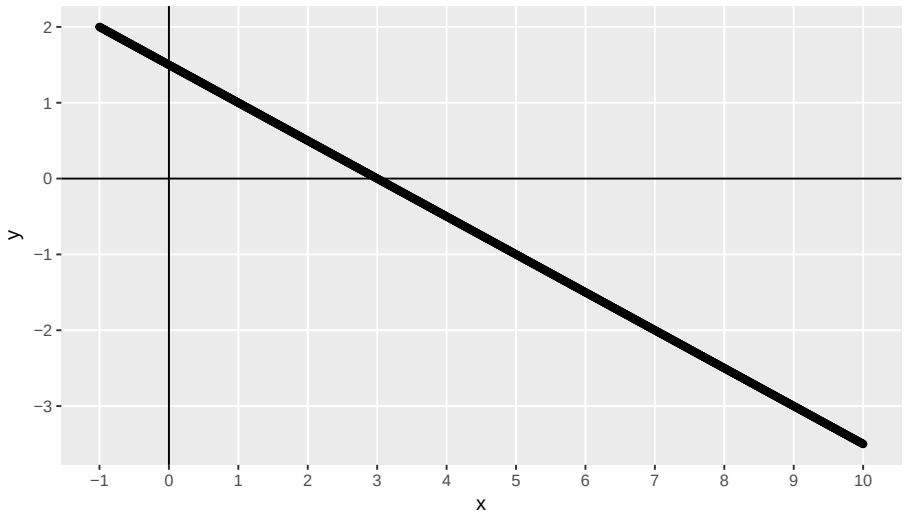


Figure 2.5: Straight line example.

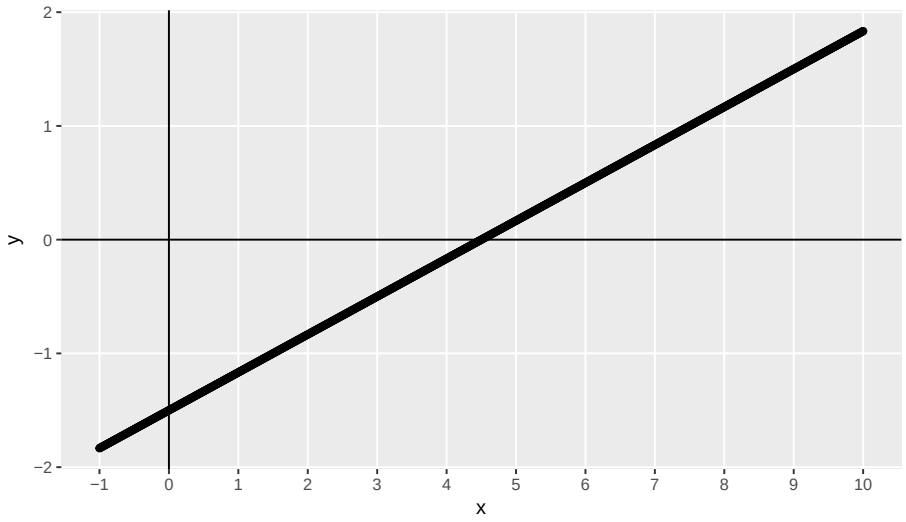


Figure 2.6: Straight line example.

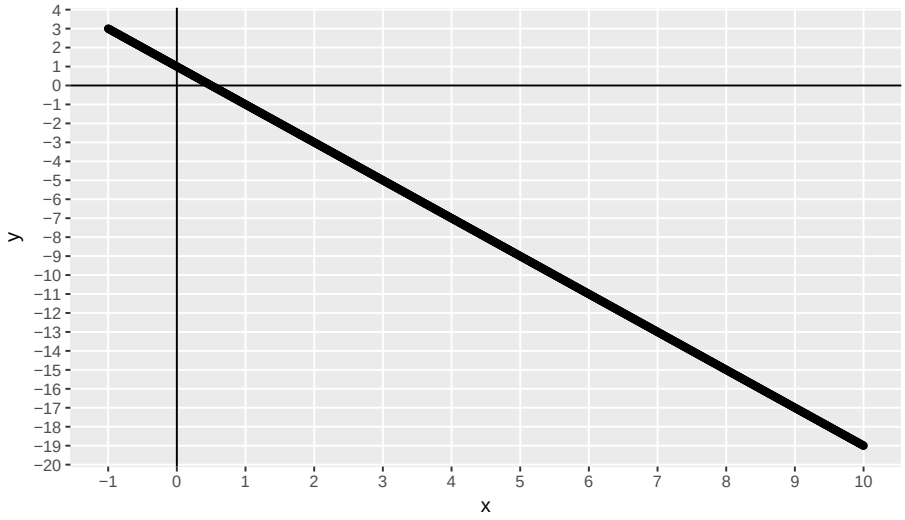


Figure 2.7: Straight line with based on y=1-2x.

2.2 Linear regression

In the previous section, we saw perfect linear relationships between quantities x and y. For each x-value there was only 1 y-value, and the values are all desribed by a straight line. Such relationships we often see in physics, or in mathematics.

In social sciences we hardly ever see such perfectly linear relationships between quantities (variables). For instance, let us plot the relationship between yearly income and the amount of Euros spent on holidays. Yearly income is measured in thousands of Euros (kEuros), and money yearly spent on holidays is measured in Euros. We plot yearly income on the x-axis (horizontal axis) and holiday spendings on the y-axis (vertical axis). Suppose we find the following data from 100 women between 30 and 40 years of age, plotted in Figure 2.8.

In the scatterplot, we see that one women has a yearly income of 100,000 Euros, and that she spends almost 1100 Euros per year on holidays. We also see a couple of women who earn less between 10,000 and 20,000 Euros a year and they spend between 200 and 300 Euros per year on holiday.

The data obviously do not form a straight line. However, we tend to think that the relationship between yearly income and holiday spending is more or less linear: for every increase of 10,000 Euros in yearly income, we an increase of about 100 Euros.

Lets plot such a straight line with a slope of 100 straight through the data points. The result is seen in Figure 2.9. We see that the line with a slope of 100 is a nice approximation of the relationship between yearly income and holiday spendings. We also see that the intercept of the line is 100.

Given the intercept and slope, the linear equation for the straight line approximating the relationship is

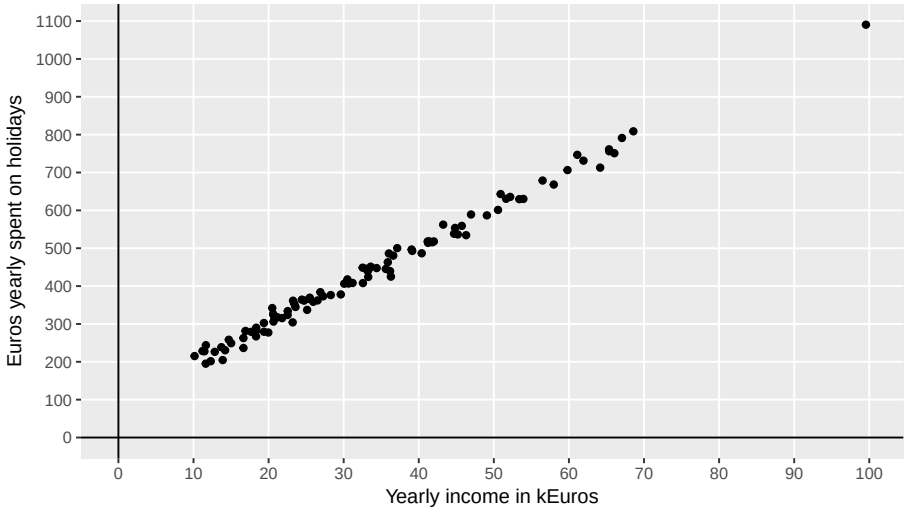


Figure 2.8: Data on holiday spending.

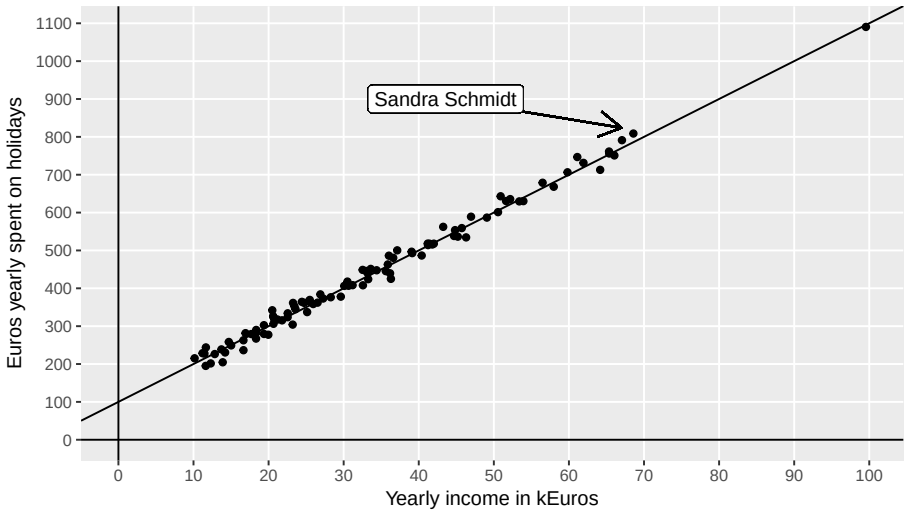


Figure 2.9: Data on holiday spending with an added straight line.

$$YearlyIncome = 100 + 100 \times HolidaySpendings \tag{2.9}$$

In summary, data on two variables may not show a perfect linear relationship, but in many cases, a perfect straight line can be a very reasonable approximation of the data. Finding such a straight line to approximate the data points is called *linear regression*. In linear regression we describe the behaviour of the dependent variable (the y-variable on the vertical axis) on the basis of the independent variable (the x-value on the horizontal axis) using a linear equation. We say that we regress variable y on variable x.

2.3 Residuals

Even though a straight line can be a good approximation of a data set, it is hardly ever perfect: there are always discrependencies between what the straight line describes and what the data actually tell us.

For instance, in Figure 2.9, we see a women, Sandra Schmidt, who makes 6.8584951×10^4 Euros a year and who spends 808.78 Euros on holidays. According to the linear equation that describes the straight line, a woman that earns 6.8584951×10^4 Euros a year would spend $100 + 100 * 6.8584951 \times 10^4 = 785.85$ Euros on holidays. The discrependy between the actual amount spent and the one described by the linear equation equals $808.78 - 785.85 = 22.93$ Euros. This difference is rather small and the same holds for all the other women in this data set.

Such descrepencies between the actual amount spent and the amount as described or predicted by the straight line are called *residuals* or *errors*. Using the linear equation, we could predict holiday spending even for yearly incomes that are not in the data set. For instance, in this data set there is no woman with an income of 80,000, but still we can use the linear equation with a prediction that such a woman would probably spend around $100 + 100 \times 80 = 8100$ Euros.

The residual (or error) is the difference between a certain data point and what the linear equation predicts.

Let us look at another fictitious data set where the residuals (errors) are a bit larger. Figure 2.10 shows the relationship between variables x and y. The dots are the actual data points and the blue straight line is an approximation of the actual relationship. The residuals are also visualized: sometimes the observed y-value is greater than the predicted y-value (dots above the line) and sometimes the oberved y-value is smaller than the predicted y-value. Let's denote the predicted y-value (the value of y predicted by the blue line) as $\hat{y}$ (pronounced as y-hat), then we can define a residual or error as the discrepency between the observed y and $\hat{y}$:

$$e = y - \hat{y} \tag{2.10}$$

where e stands for the error (residual).

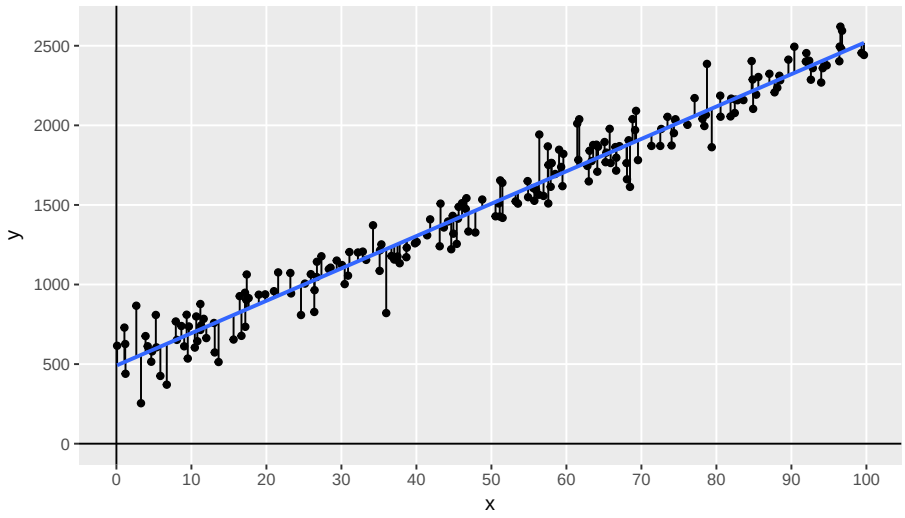


Figure 2.10: Data on variables x and y with an added straight line.

If we compute residual e for every y-value in the data set, we can plot them using a histogram, as displayed in Figure 2.11. We see that the residuals are on average 0, and that the histogram has the shape of the normal distribution, more or less. Such normally-shaped distributions of residuals we see often in research. Here, the residuals show a normal distribution with mean 0 and variance of 1.3336×10^4 (a standard deviation of 115).

2.4 Least squares regression lines

You may ask yourself how to draw a straight line through the data points: How do you decide the exact slope and the exact intercept? And what if you don't want to draw the data points and the straight line by hand? That can be quite cumbersome if you have more than 2000 data points to plot!

First, because we are lazy, we always use a computer to draw the data points and the regression line. Second, since we could draw many different straight lines through a scatter of points, we need a criterion to determine a nice combination of intercept and slope. With such a criterion we can then let the computer determine the straight line with its equation for us.

The criterion that we use in this chapter is called Least Squares, or Ordinary Least Squares (OLS). To explain the Least Squares principle, look again at Figure 2.10 where we see both small and large residuals. About half of them are positive (above the blue line) and half of them are negative (below the blue line).

The most reasonable idea is to draw a straight line that is more or less in the middle of the y-values, in other words, with about half of residuals positive

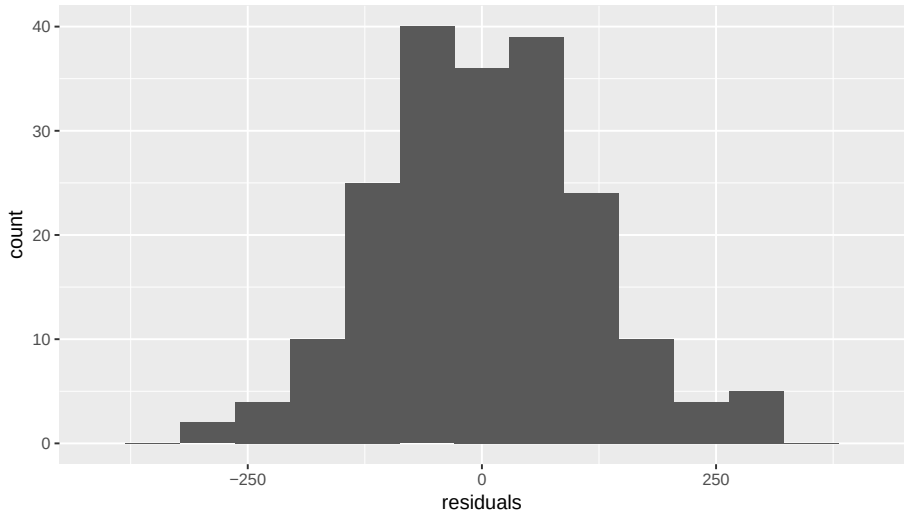


Figure 2.11: Histogram of the residuals (errors).

and about half of them negative. Or perhaps we could say that on average, the residuals should be 0. A third way of saying the same thing is that the sum of the residuals should be equal to 0.

However, the criterion that all residuals should sum to 0 is not sufficient. In Figure 2.12 we see a straight line with a slope of 0 where all residuals sum to 0. However, this regression line does not make intuitive sense: it does not describe the structure in the data very well. Moreover, we see that the residuals are much larger than in Figure 2.10.

We therefore need a second criterion to find a nice straight line. We want the residuals to sum to 0, but also want the residuals to be as small as possible: the descrepencies between what the linear equation predicts (the $\hat{y}$-values) and the actual y-values should be as small as possible.

So now we have two criteria: we want the sum of the residuals to be 0 (half of them negative, half of them positive), and we want the residuals to be as small as possible. We can achieve both of these when we use as our criterion the idea that the sum of the *squared* residuals be as small as possible. If the sum of the squared residuals is as small as possible, we know that the residuals are as small as possible. But we also know that the residuals then sum to zero[1]. Thus, as our criterion we can use the regression line for which the squared differences between predicted and observed y-values are as small as possible.

Figure 2.13 shows three different regression lines for the same data set. Figure 2.14 shows the respective distributions of the residuals. For the first line, we see that the residuals sum to 0, for the residuals are on average 0 (the red vertical line). However, we see quite large residuals. The residuals for the second

[1]For the proof of this, see ....

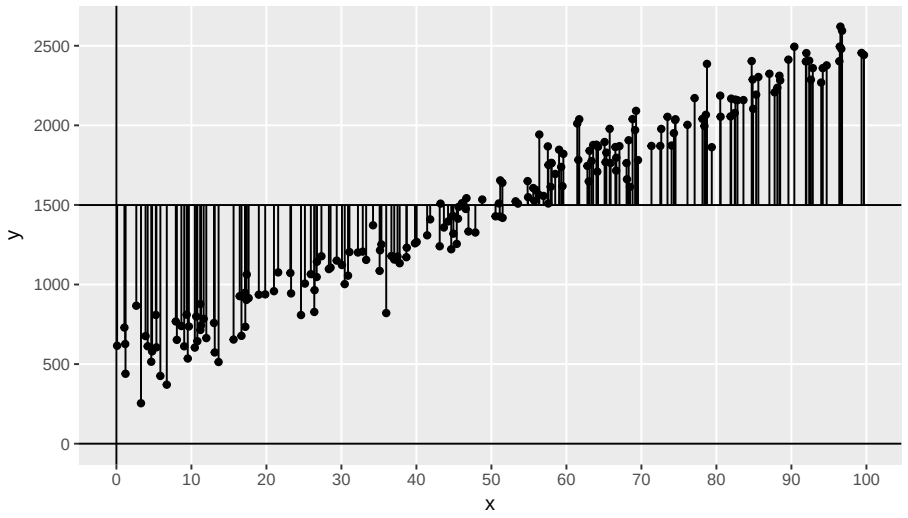


Figure 2.12: Data on variables x and y with an added straight line. The sum of the residuals equals 0.

line are smaller: we see very small positive residuals, but the negative residuals are still quite large. We also see that the residuals do not sum to 0. For the third line, we see both criteria optimized: the sum of the residuals is zero and the residuals are all very small. We see that for regression line 3, the sum of squared residuals is at its minimum value.

In summary, when we want to have a straight line that describes our data best, we'd like a line such that the residuals are on average 0 (i.e, sum to 0), and where we see the smallest residuals possible. We reach these criteria when we use the line in such a way that we have the lowest value for the sum of the squared residuals possible. This line is therefore called the least squares or OLS regression line. It turns out that this optimal regression slope can be found by a relatively simple computation using matrix algebra. In daily life, we do not do this by hand but let computers compute it for us, with software like for instance SPSS.

2.4.1 Exercises

1. In Table 2.1 you find a small data set on the price of homes with dependent variable price in kEuros and independent variable area in square meters. The least squares regression equation turns out to be $price = out.price\text{coef}[1] + \text{out.price}coef[2] \times area$. Add a third column with the expected prices based on the regression equation ($\hat{y}$. Put the difference between the observed price and the expected price in the fourth column (e). Then compute the squared residuals and put those in the fifth column (e^2). Take the sum of the squared residuals: How large is sum of the

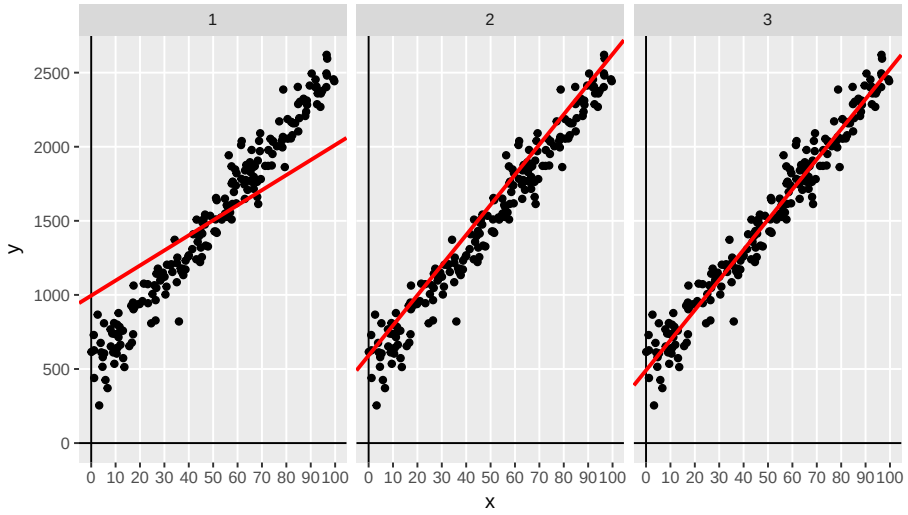


Figure 2.13: Three times the same data set, but with different regression lines.

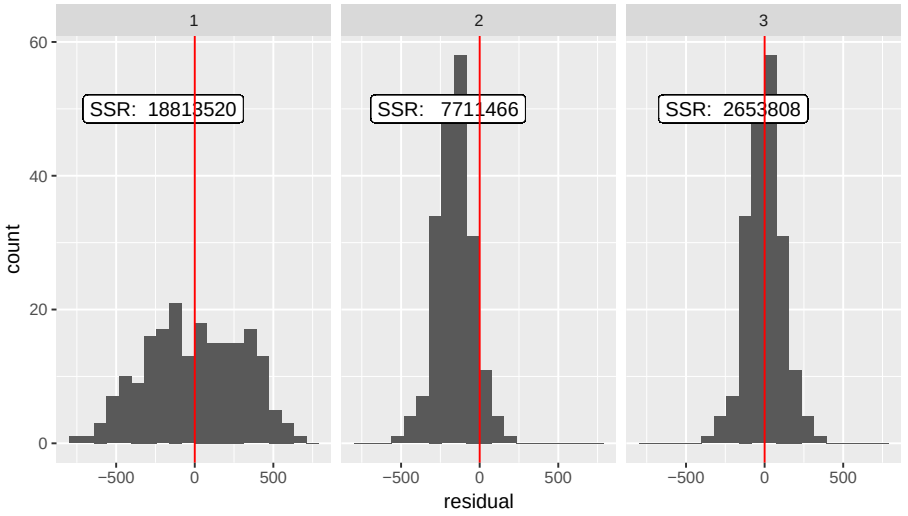


Figure 2.14: Histogram of the residuals (errors) for three different regression lines, and the respective sums of squared residuals (SSR).

Table 2.1: Home prices.

Area	Price	PredictedPrice	Residual	SquaredResidual
56.00	165.00			
101.00	180.00			
67.00	115.00			
109.00	164.00			
115.00	175.00			
34.00	135.00			

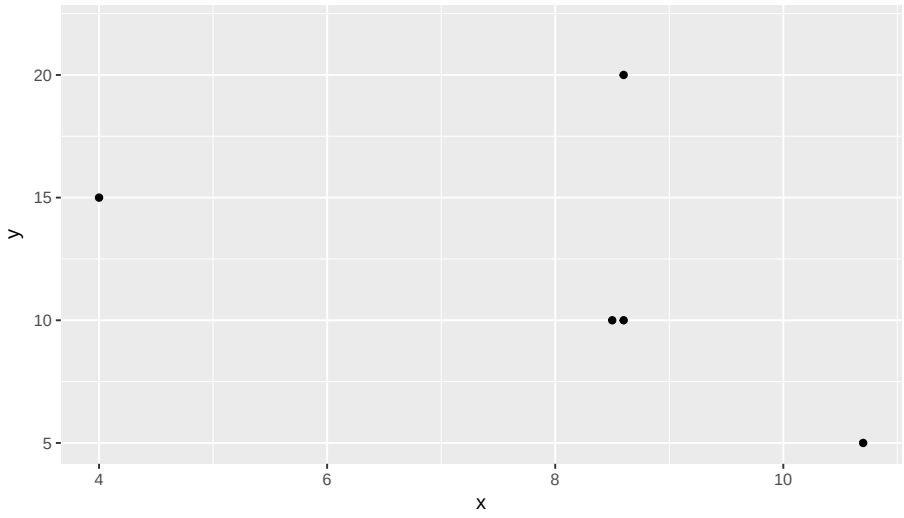


Figure 2.15: Plot of housing data.

squared residuals?

2. See website ....., try to find the Least Squares regression line for the given data set by changing both intercept and slope. How large is the sum of the squared residuals for that optimal regression line?

3. Do this exercice with one or more of your fellow students. Look at the data set plotted in Figure 2.15. Try to find the regression line with the lowest sum of squared residuals possible.

Answers:

1. The predicted prices, the residuals and the squared residuals are displayed in Table 2.2. The sum of the squared residuals equals 2397.4189795.

2.

3. The lowest sum of squared residuals is 97. This is the sum that you get with intercept 21.4 and slope -1.2.

Table 2.2: Home prices.

Area	Price	PredictedPrice	Residual	SquaredResidual
56.00	165.00	140.74	24.26	588.34
101.00	180.00	167.91	12.09	146.18
67.00	115.00	147.38	-32.38	1048.76
109.00	164.00	172.74	-8.74	76.37
115.00	175.00	176.36	-1.36	1.85
34.00	135.00	127.46	7.54	56.80

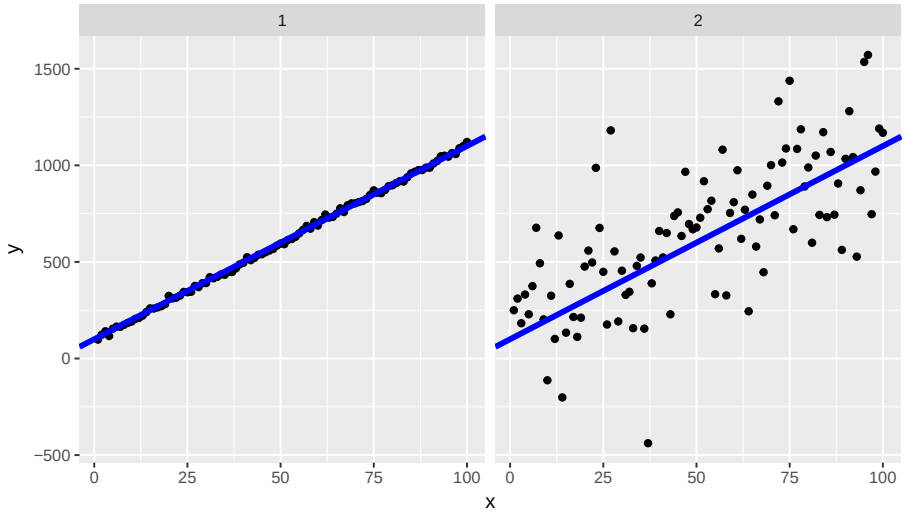


Figure 2.16: Plot of housing data.

2.5 Pearson correlation

For any set of two quantitative variables, we can determine the least squares regression line. However, it depends on the data set how well that regression line describes the data. Figure 2.16 shows two different data sets on variables x and y. Both plots also show the least squares regression line, and they both turn out to be exactly the same: $y = 100 + 10x$.

We see that the regression line describes data set very well (left panel): the observed dots are very close to the line, which means that the residuals are very small. The regression line does a worse job for the second data set (right panel) since there are quite large discrepencies between the observed y-values and the predicted y-values. Put differently, the regressin equation can be used to predict y-values in data set 1 very well, almost without error, whereas the regression line cannot be used to predict y-values very precisely. The regression line is also the least squares regression line for data set 2, so any improvement by choosing another slope or intercept is not possible.

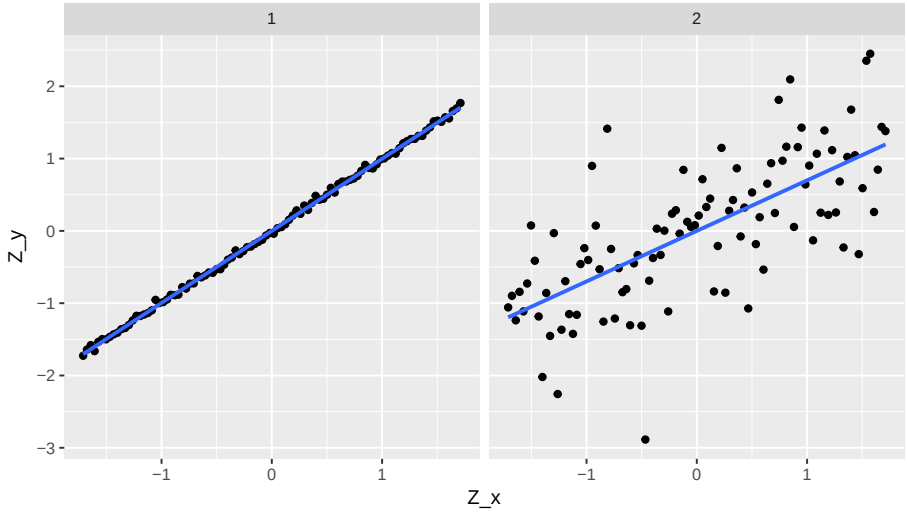


Figure 2.17: Plot of housing data.

Francis Galton was the first to think about how to quantify this difference in the ability of a regression line to predict the dependent variable. Karl Pearson later worked on this measure so that it became to be called Pearson's correlation coefficient. It is a standardized measure, so that it can be used to compare different data sets.

In order to get to Pearson's correlation coefficient, you first need to standardize both the independent variable, x, and the dependent variable, y. You standardize scores by taking their values, subtract the mean from them, and divide by the standard deviation. So, in order to obtain a standardized x-value we compute Z_x,

$$Z_x = \frac{x - \bar{x}}{\sigma_x} \tag{2.11}$$

and in order to obtain a standardized y-value we compute Z_y,

$$Z_y = \frac{y - \bar{y}}{\sigma_y}. \tag{2.12}$$

Let's do this both for data set 1 and data set 2, and plot the standardized scores. see Figure 2.17. If we then plot the least squares regression lines for the standardized values, we obtain different equations. For both data sets, the intercept is 0 because by standardizing the scores, the means become 0. But the slopes are different: in data set 1, the slope is 0.997 and in data set 2, the slope is 0.699.

$$Z_y = 0 + 0.997 Z_x = 0.997 Z_x \tag{2.13}$$

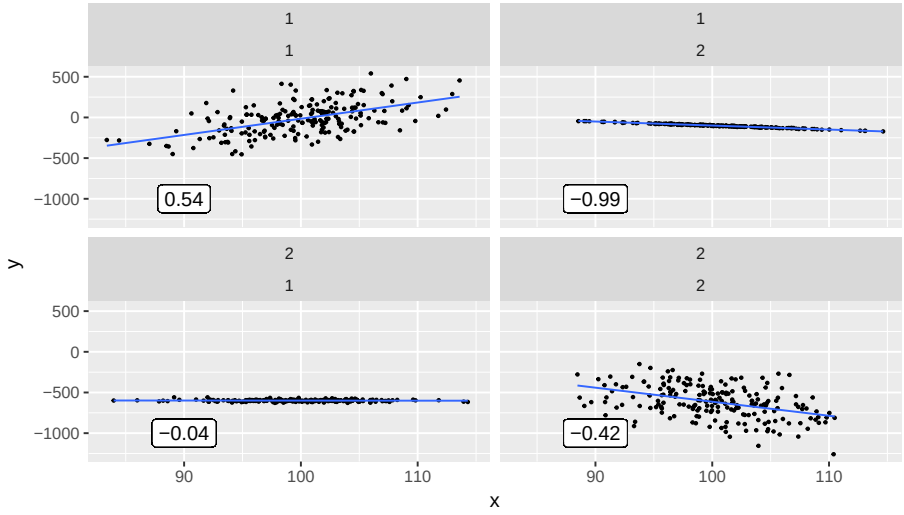


Figure 2.18: Plot of housing data.

$$Z_y = 0 + 0.699 Z_x = 0.699 Z_x \tag{2.14}$$

These two slopes, the slope for the regression of standardized y-values on standardized x-values, are the correlation coefficients for data sets 1 and 2, respectively. For obvious reasons, the correlation is sometimes also referred to as the *standardized slope coefficient*.

Correlation stands for the *co-relation* between two variables. It tells you how strongly one variable can be predicted from the other. The correlation is bi-directional: the correlation between y and x is the same as the correlation between x and y. For instance in Figure 2.17, if we would have put the Z_x-variable on the Z_y-axis, and the Z_y-variable on the Z_x-axis, the slopes would be exactly the same. This is true because the variances of the y and x-variables are equal after standardization (both variances equal to 1).

Since a slope can be negative, a correlation can be negative too. Furthermore, a correlation is always between -1 and 1. Look at Figure 2.17: the correlation between x and y is 0.997. The dots are almost on a straight line. If the dots would all be exactly on the straight line, the correlation would be 1.

Figure 2.18 shows a number of scatterplots of x and y with different correlations. Note that if dots are very close to the regression line, the correlation can still be close to 0. If the slope is 0 (bottom-left panel), then one variable cannot be predicted from the other variable, hence the correlation is 0, too.

In summary, the correlation coefficient indicates how well one variable can be predicted from the other variable. It is the slope of the regression line if both variables are standardized. If prediction is not possible (when the regression slope is 0), the correlation is 0, too. If the prediction is perfect, without errors (no residuals) and with a slope unequal to 0, then the correlation is either -1 or

+1, depending on the sign of the slope.

2.6 Covariance

The correlation is a standardized measure for how much two variables co-relate. There exists also an unstandardized measure for how much two variables co-relate: the *covariance*. The correlation is the slope when two variables have each variance 1. When you multiply the correlation by a number indicating the variances of the two variables, you get the covariance. This number is the product of the two respective standard deviations.

The covariance between variables x and y, Cov(x,y) can be computed as:

$$Cov(x, y) = Cor(x, y) \times \sigma_x \sigma_y \tag{2.15}$$

For example, if the variance of x equals 49 and the variance of y equals 25, then the respective standard deviations are 7 and 5. If the correlation between x and y equals 0.5, then the covariance between x and y is equal to $0.5 \times 7 \times 5 = 17.5$.

Similar to correlation, the covariance of two variables indicates by how much they co-vary. For instance, if the variance of x is 3 and the variance of y is 5, then a covariance of 2 indicates that x and y co-vary: if x increases by a certain amount, y also increases. If you want to know how many standard deviations y increases if x increases with one standard deviation, you can turn the covariance into a correlation my dividing the covariance by the respective standard deviations.

$$Cor(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{2}{\sqrt{3}\sqrt{5}} = 0.52 \tag{2.16}$$

Similar to correlations and slopes, covariances can also be negative.

2.6.1 Exercises

1. The correlation between brain size and intelligence in 9-year old children equals 0.30. Suppose the variance in brain size equals 45 and the variance in intelligence 225. Compute the covariance.

2. The covariance between intelligence and extraversion equals 1. The variance of intelligence is 225 and the variance of extraversion is 9. What is the correlation?

3. Suppose the correlation between intelligence and extraversion is 0.10. What does this mean?

4. Suppose the correlation between intelligence and extraversion is -0.05. What does this mean?

5. Suppose the correlation between intelligence and extraversion is 0.30. What is the regression slope if the variance of intelligence is 225 and the variance of extraversion is 9?

Answers:

1.
$$Cov(x, y) = Cor(x, y) \times \sigma_x \sigma_y = 0.30 \times \sqrt{45} \times \sqrt{225} = 30 \tag{2.17}$$

2.
$$Cor(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{1}{\sqrt{225}\sqrt{9}} = 0.02 \tag{2.18}$$

3. If you increase intelligence by 1 standard deviation, then extraversion increases with a tenth of a standard deviation.

4. If you increase intelligence by 1 standard deviation, then extraversion increases with 0.05 standard deviations.

5. The correlation is 0.30, so if you increase intelligence by one standard deviation (which is $\sqrt{225} = 15$), extraversion increases by 0.30 standard deviations (which equals $0.30 \times \sqrt{9} = 0.90$). Therefore, if you increase intelligence by 15 points, you increase extraversion by 0.90 points. Thus if you increase intelligence by 1 point, you increase extraversion by $0.90/15 = 0.06$ points. The slope for the regression of extraversion on intelligence is therefore 0.06.

2.7 Regression using SPSS

2.8 Linear models

By performing a regression analysis of y on x, we try to predict the y-value from a given x-value on the basis of a linear equation. We try to find an intercept and a slope for that linear equation such that our prediction is best. We define best as the linear equation for wich we see the lowest possible value for the sum of the squared residuals (least squares principle).

Thus, the predicted value of y ($\hat{y}$) can be computed by the linear equation

$$\hat{y} = b_0 + b_1 x \tag{2.19}$$

In reality, the predicted values of y always deviate from the observed values of y. So, there is always an error e that is the difference between $\hat{y}$ and y. Thus we have for the observed values of y

$$y = \hat{y} + e = b_0 + b_1 x + e \tag{2.20}$$

Typically, we assume that the residuals e are on average 0 and have a normal distribution with a certain variance σ_e^2. Taking the linear equation and the

normally distributed residuals together, we have *a linear model* for the two variables x and y.

$$y = b_0 + b_1 x + e \tag{2.21}$$

$$e \sim N(0, \sigma_e^2) \tag{2.22}$$

In the remainder of this book, we will see a great variety of linear models between 2 or more variables. They can all be seen as extensions and variations of this basic linear model. They all aim to predict as best as possible one dependent variable from a number of independent variables.

Chapter 3

Inference I: random samples, standard errors and confidence intervals

In the previous chapter on simple regression we saw how a linear equation can describe a data set: the linear equation describes the behaviour of one variable, the dependent variable, on the basis of one other variable, the independent variable. Sometimes we are indeed interested in the relationship between two variables in one given data set. For instance, a teacher wants to know whether her exam gradings in her class of last year predict how well her students do in a second course a year later.

But very often, researchers are not interested in the relationships between variables in one data set, but interested in the relationship between variables in general, not limited to only the observed data. For example, a researcher would like to know what the relationship is between the temperature in a brewery and the volume of beer that goes into one bottle. In order to study the effect of temperature on volume, the researcher measures the volume of beer in 200 bottles at 20 degrees Celsius and determines from log files the temperature in the factory during production for each measured bottle. The linear equation might be $volume = 31.72 - 0.088 \times t + e$, see Figure 3.1. But the question is what the equation would be if the researcher had used information about *all* bottles produced in the same factory.

In other words, we may know about the linear relationship between temperature and volume in a *sample* of bottles, but we might really be interested to know what the relationship would look like *had we been able to measure the volume in all bottles.*

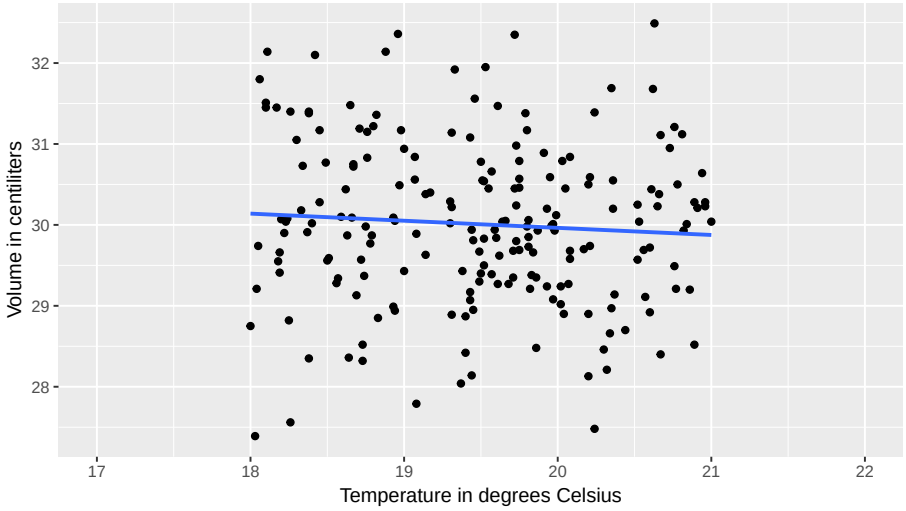


Figure 3.1: The relationship between temperature and volume in a sample of 200 bottles.

3.1 Population data and sample data

In the beer bottle example above, the volume of beer was measured in a total of 200 bottles. Let's do a thought experiment. Suppose we could have access to volume data about all bottles of beer on all days where the factory was operating, including information about the temperature for each day of production. Suppose that the total number of bottles produced is 80,000 bottles. When we plot the volume of each bottle against the temperature of the factory we get the scatter plot in Figure 3.2.

In our thought experiment, we could determine the regression equation using all bottles that were produced: all 80,000 of them. We then find the blue regression line displayed in Figure 3.2. Its equation is $Volume = 29.98 + 0.001 \times t$.

In the data example above, data was only collected on 200 bottles. These bottles were randomly selected[1]: there were many more bottles but we could measure only a limited number of them. This explains why the regression equation based on the sample differed from the regression equation based on all bottles: we only see part of the data.

Here we see a discrepency between the regression equation based on the sample, and the regresssion equation based on the population. Here, the *population* is the collection of all bottles produced in the factory. The *sample* is the collection of 200 randomly selected bottles. Here we have a slope of 0.001 in the population, and we see a slope of -0.088 in the sample. Also the intercepts differ. To distinguish between the coefficients of the population and coefficients

[1]Random selection means that each of the 80,000 bottles had an equal probability to end up in this sample of 200 bottles.

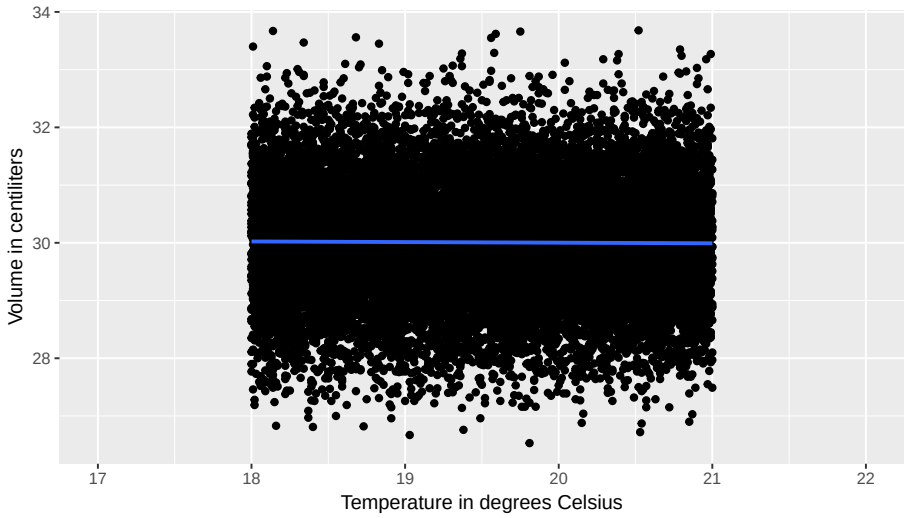


Figure 3.2: The relationship between temperature and volume in all 80,000 bottles.

of the sample, the population coefficient is often denoted by the Greek letter β and the sample coefficient by the Roman letter b.

$$\begin{aligned} Population : Volume &= 29.98 + 0.001 \times t \\ Sample : Volume &= 31.72 - 0.088 \times t \end{aligned}$$

The discrependency between the two equations is simply the result of chance: had we selected another sample of 200 bottles, we probably would have found a different sample equation with a different slope and a different intercept. The intercept and slope based on sample data, are the result of chance and therefore vary from sample to sample. The population intercept and slope (the true ones) are fixed, but unknown. If we want to know something about the population intercept and slope, we only have the sample equation to go on. Our best guess for the population equation is the sample equation, but how certain can we be about how close the sample intercept and slope are to the population intercept and slope?

3.2 Random sampling and the standard error

In order to know how close the intercept and slope in a sample are to their values in the population, we do another thought experiment. Let's see what happens if we take more than one random sample of 200 bottlees.

We put the 200 bottles that we selected earlier back into the population and we again blindly pick a new collection of 200 bottles. We then measure for each

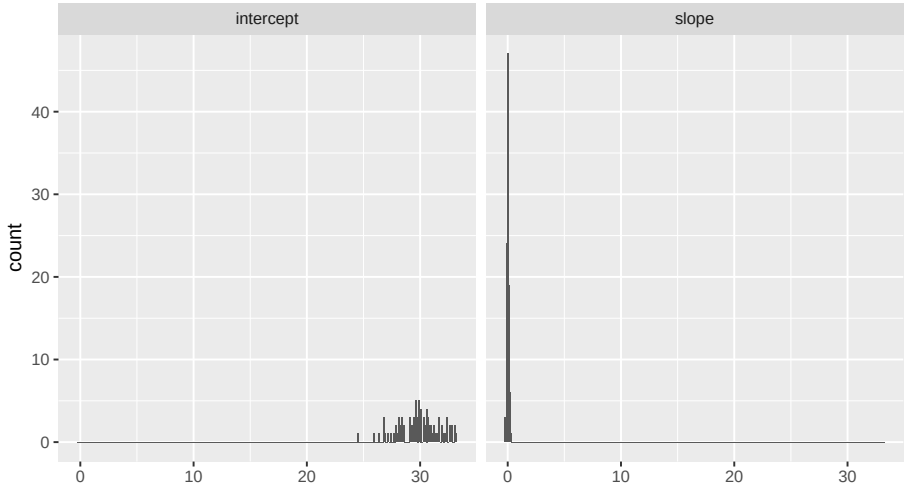


Figure 3.3: Distribution of the 100 sample intercepts and 100 sample slope.

bottle the volume of beer it contains and we determine the temperature in the factory on the day of its production. We then apply a regression analysis and determine the intercept and the slope. Next, we put these bottles back into the population and draw a next random sample of 200 bottles.

You can probably imagine that if we repeat this procedure of randomly picking 200 bottles from a large population of 80,000, each time we find a different intercept and a different slope. Let's carry out this procedure 100 times by a computer. If we then plot the histograms of the 100 sample intercepts and sample slopes we get Figure 3.3. We see a large variation in the intercepts, and only a small variation in the slopes (i.e., all values very close to 0).

For now, let's focus on the slope; this because we are mostly interested in the linear relationship between volume and temperature, but everything that follows also applies to the intercept. In Figure 3.4 we see the histogram of the slopes if we carry out the random sampling 1000 times. We see that on average the sample slope is around 0.001, which is the population slope (the slope if we analyze all bottles). But there is variation around that mean of 0: the standard deviation of all 1000 sample slopes turns out to be 0.084.

The standard deviation of the sample slopes is called the *standard error*. Had the population slope been 110 or -40, the sample slopes would cluster around 110 or -40, but the standard deviation of the sample slopes, the standard error, would be the same.

The standard error for a sample slope represents the uncertainty about the population slope. If the standard error is large, it means that if we would draw many different random samples from the same population data, we would get very different sample slopes. If the standard error is small, it means that if we

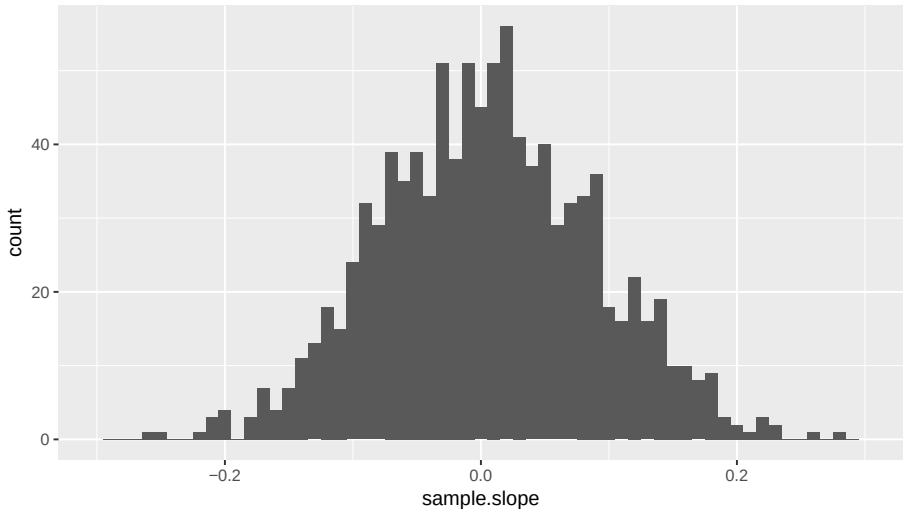


Figure 3.4: Distribution of 1000 sample slopes.

would draw many different random samples from the same population data, we would get sample slopes that are very close to one another, and very close to the population slope.[2]

3.2.1 Standard error and sample size

The standard error for a sample slope depends on many things, but the most important factor is the *sample size*: how many bottles there are in each random sample. The larger the sample size, the smaller the standard error, the more certain we are about the population slope. In the above example, the sample size is 200 bottles.

Imagine that you draw only 2 bottles from the population of 80,000 bottles. Then there is quite some probability that by sheer luck you find one bottle with a low temperature and a small volume, and another bottle with a high temperature and a large volume. This would yield a sample slope that is quite large and positive. But there is an equally high probability that you get one bottle with a low temperature with a large volume, and another bottle with a high temperature and a small volume. Then based on these two other bottles, the sample slope will be large and negative. In case of a sample size of only 2, you see that there will be quite a lot of variation in the sample slope if we draw various random samples. This large variation in sample slopes is then captured by the standard error, that will be large. With only 2 bottles per sample, the uncertainty about the population slope will then also be large. The left panel

[2]Because sample slopes cluster aournd the population slope, the sample slope is very close to the population slope when the standard error is small.

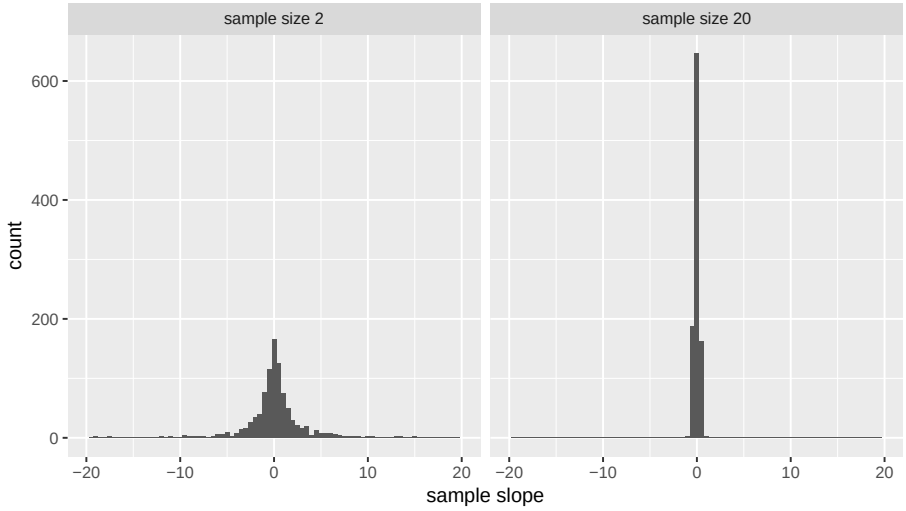


Figure 3.5: Distribution of the sample slope when sample size is 2 (left panel) and when sample size is 20 (right panel).

of Figure 3.5 shows the distribution of the sample slope where the sample size is 2. You see that for quite a number of samples, the lsope is larger than 10, even if the population slope is 0.001.

Now imagine that your sample size is 20. Then the probability that the 20 bottles will result in a large variation of slopes will be smaller: it would be very unlikely that *all* 20 bottles have either a high volume and a high temperature, or a low volume and a low temperature. If there happen to be a few of such bottles in the sample, the other bottles will average these effects out. Because of this averaging effect, the slope based on 20 bottles will then be closer to the population slope. The standard error therefore decreases with increasing sample size. With a sample size of 20, most slopes are between -0.6 and 0.6.

In Figure 3.5 we see the distributions of the sample slope where the sample size is either 2 (left panel) or 20 (right panel). We see quite a lot of variation in sample slopes with sample size equal to 2, and considerably less variation in sample slopes if sample size is 20. This shows that the larger the sample size, the smaller the standard error, the larger the certainty about the population slope.

3.2.2 From sample slope to population slope

In the previous section we saw that if we have a small standard error, we can be relatively certain that our sample slope is close to the population slope. We did a thought experiment where we knew everything about the population intercept and slope, and we drew many samples from this population. In reality, we don't know anything about the population: we only have one sample of data. So

suppose we draw a sample of 200 from an unknown population of bottles, and we find a slope of 1, we have to look at the standard error to know how close that sample slope is to the population slope.

For example, suppose we find a sample slope of 1 and the standard error is equal to 0.1. Then we know that the population slope is more likely to be in the neighbourhood of values like 0.9, 1.0, or 1.1 than in the neighbourhood of 10 or -10.

Now suppose we find a sample slope of 1 and the standard error is equal to 10. Then we know that the sample slope is more likely to be somewhere in the neighbourhood of values like -9, 1 or 11, than around values in the neighbourhood of -100 or +100. However, values like -9, 1 and 11 are quite far apart, so actually we have no idea what the population slope is; we don't even know whether the population slope is positive or negative! The standard error is simply too large.

As we have seen, the standard error depends very much on sample size. Apart from sample size, the standard error for a slope also depends on the variance of the independent variable, the variance of the dependent variable, and the correlations between the independent variable and other independent variables in the equation (in case of multiple regression and other linear models, see later chapters). We will not bore you with the complicated formula for the standard error for regression coefficients [3]. Instead, we look at the standard error that SPSS or other computer packages compute for us.

3.3 *t*-distributions

Above we saw that if there is a large collection of data points (population) with a particular slope that describes the relationship between two variables, and if you then take random samples out of this collection, each time you find a different value for the slope in the sample, the sample slope. We saw that the standard deviation of the distribution of all such slopes is called the standard error. The standard error gives us information about how certain we can be that the slope in the sample is close to the slope in the population. The smaller the standard error, the more certain we can be that the population slope has a value in the neighbourhood of the value for the sample slope.

When we look at the distribution of the sample slope, for instance in Figure 3.4, we notice that the distribution looks very much like a normal distribution. Well, actually it isn't quite a normal distribution. In reality it has the shape of a *t*-distribution. Figure 3.6 shows the difference between a *t*-distribution (in red) and a normal distribution (in blue). In this figure, the means are equal (0) and the areas under the curve are equal (1), but the shapes are clearly different. Compared to the *t*-distribution, the normal distribution has more observed values close to the mean (the distribution is more peaked). The *t*-distribution has relatively more observations in the tails of the distribution

[3] See https://www3.nd.edu/ rwilliam/stats1/x91.pdf for the formula. In this pdf, 'IV' means independent variable

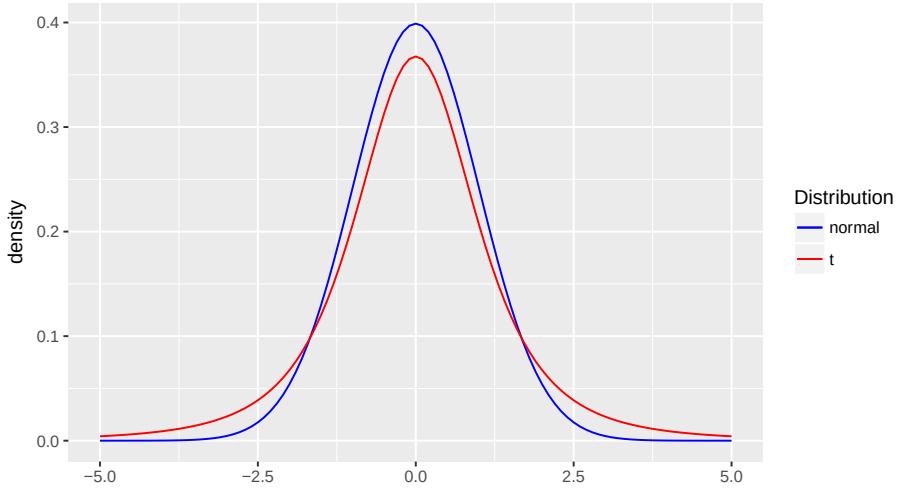


Figure 3.6: Difference in the shapes of a normal distribution and a t-distribution

(heavy tails).

Actually, the shape of the distribution of sample slopes depends on the size of the samples, the sample size. In Figure 3.7 we see what the distribution of sample slopes would look like if all samples would be of size 4 (the red line) and what the distribution would look like if sample size would be 200 (the blue line). If we compare the blue lines in Figures 3.6 and 3.7 we see that the shape of the t-distribution for a sample size of 200 looks extremely close to the normal distribution. Remember: we are talking here only about the *shape* of the distribution.[4]

In summary, when we draw many samples from a population, the standard deviation of the sample slopes (the standard error) will be smaller for a large sample size than for a small sample size. In addition, the *shape* of the distribution of sample slopes is that of a t-distribution. The shape of the t-distribution also depends on sample size. The larger the sample size, the more the shape of the t-distribution looks like a normal distribution. Thus, for large sample sizes, the distribution of sample slopes shows very little variance with a shape closely resembling a normal distribution.

3.4 T-statistics

Above we saw that sample slopes have a t-distribution, and that if sample size is large, say larger than 200, the t-distribution looks very much like a normal distribution. From the normal distribution, we know that if we standardize the

[4]The variance (i.e., the square of the standard error) will be smaller for larger samples sizes.

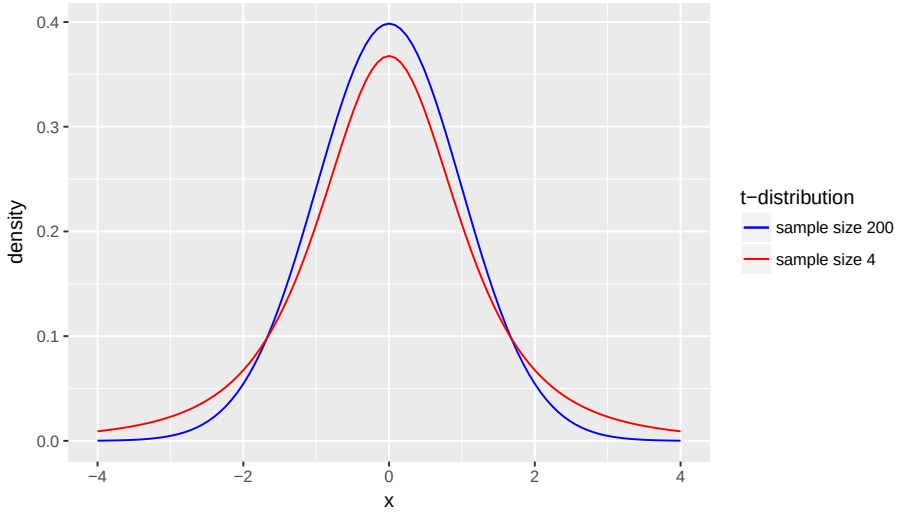


Figure 3.7: The shape of the distribution of sample slopes depends on sample size.

scores by computing Z-scores, that is, if we subtract the mean and then divide by the standard deviation, $Z = \frac{x-\bar{x}}{\sigma}$, then 2.5% of the Z-values is smaller than -1.96 and 2.5% of the z-values is larger than +1.96.

Therefore, if for large sample sizes the t-distribution is practically indistinguishable from the normal distribution, we know that if we standardize the sample slope values, we get a similar result. Instead of looking at the actual slope value, we can compute a standardized slope. Let's call that standardized result T. Then we get:

$$T = \frac{b - \bar{b}}{se} \tag{3.1}$$

In words: we take a particular sample slope b and we subtract the mean from all sample slopes. The result we divide by the standard deviation of the sample slopes, which is callled the standard error se.

But what is the mean sample slope? Since the sample slopes cluster around the population slope β, the average of all possible samples slopes is equal to β. Thus we have:

$$T = \frac{b - \beta}{se} \tag{3.2}$$

Let's go back to the example of the beer bottles. In our first random sample of 200 bottles, we found a sample slope of -0.088. We also happened to know the population slope, which was 0.001. From our computer experiment, we saw that the standard deviation of the sample slopes with sample size 200 was equal

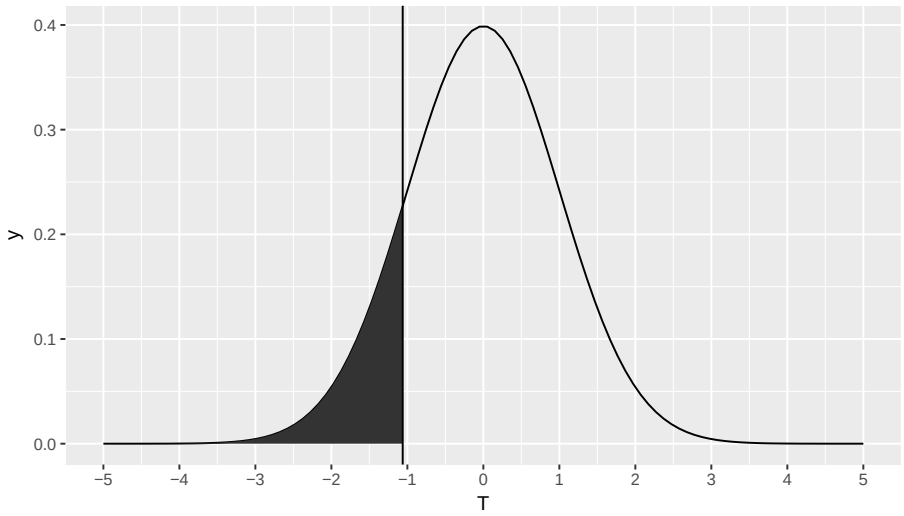


Figure 3.8: The standard normal distribution and the probability of a Z-score lower than -1.06

to 0.084. Thus, if we fill in the formula for the standardized slope T, we get for this particular sample

$$T = \frac{-0.088 - 0.001}{0.084} = -1.06 \tag{3.3}$$

Notice that we distinguish between a variable t that has a t-distribution, and a T-statistic that is based on a computation.

Now, what can we say about this T-value? Since with a sample size of 200 the distribution of sample slopes closely resembles a normal distribution, we can use normal tables published online or in computer packages to see how likely a value of $T = -1.06$ actually is. In normal tables we find that a Z-value of -1.06 is not that strange: in the standard normal distribution, 14.457% of the values is smaller than -1.06. The area is shown in Figure 3.8.

When would we say that a certain T-value would cause concern? Well, perhaps we could say that if the T-value were 3 standard deviations away from the population value, either 3 standard deviations above the population value or 3 standard deviations below the population value. From the normal tables, we know that that happens only 0.27% of the time.

Alternatively, we could say that we would perhaps also be worried if the sample slope were 2 standard deviations away from the population slope, corresponding to T-value of 2 or -2. We know that the probabilty that that happens is around 5%, small enough perhaps to raise concern about our knowledge about the population slope.

In this section, when discussing T-statistics, we assumed we knew the population slope β, that is, the slope of the linear equation based on all 80,000

bottles. In reality, we never know the population slope: the whole reason to look at the sample slope is to have an idea about the population slope. Let's look at some hypothetical population slopes.

3.5 Hypothetical population slopes

Since we don't know the actual value of the population slope β, we could ask the personnel in the beer factory what they think is a likely value for the slope. Suppose Mark says he believes that a slope of 2 could be true. Well, let's find out whether that is a reasonable guess. Now we *assume* that the population slope β is 2, and we compute the T-statistic for our sample slope:

$$T = \frac{-0.088 - 2}{0.084} = -24.8 \tag{3.4}$$

From the normal distribution, we know that such a T-value is very unlikely: the probability of finding a sample slope -24.8 standard deviations away form a population slope of 2 is less than 0.00000000000000000000000000000000001. Because we know that such a T-value of -24.8 is unlikely, we know that a sample slope of -0.0879535 is unlikely *if the population slope is equal to 2*. Therefore, we feel 2 is not a realistic value for the population slope.

Now let's ask Martha. She thinks a reasonable value for the population slope is 0, as she doesn't believe there is a linear relationship between temperature and volume. She feels that the fact that we found a sample slope that was not 0 was a pure coincidence. Based on that hypothesis, we compute T again and find:

$$T = \frac{-0.0879535 - 0}{0.0840496} = -1.0464475 \tag{3.5}$$

In other words, if we believe Martha, our sample slope is only about 1 standard deviation away from her hypothesized value. That's not a very bad idea, since from the normal distribution we know that a value more than 1.05 standard deviations away from the mean (above or below) is 29.54%. In other words, if the population is truly 0, then our sample slope of -0.0879535 is quite a reasonable finding. If we reverse this line of reasoning: if our sample slope is -0.0879535, with a standard error of 0.0840496, then a population slope of 0 is quite a reasonable guess! It is reasonable, since the difference between the sample slope and the hypothesised value is only -1.0464475 standard errors.

So when do we no longer feel that a value for the population slope is reasonable? Perhaps if the probability of finding a sample slope of at least a certain size given a hypothesised population slope is so small that we no longer believe that the hypothesised value is reasonable. We might for example choose a small probability like 1%. We know from the normal distribution that 1% of the values lie at least 2.58 standard deviations above and below the mean. So if our sample slope is more than 2.58 standard errors away from the hypothesised population slope, then that population slope is *not* a reasonable guess. In other words, if

the *distance* between the sample slope and the hypothesised population slope is more than 2.58 standard errors, then the hypothesised population slope is no longer reasonable.

This implies that *any* value closer than 2.58 standard errors from the sample slope is a collection of reasonable values for the population slope.

Thus, in our example of the 200 bottles with a sample slope of -0.088 and a standard error of 0.084, the interval from $-0.088 - 2.58 * 0.0840496$ to $-0.088 + 2.58 * 0.084$ contains reasonable values for the population mean. If we do the calculation, we get the interval from -0.3 to 0.13. If we would have to guess the value for the population slope, our guess would be that it would lie somewhere between between -0.3 and 0.13, *if we feel that 1% is a small enough probability.*

In data analysis, such an interval that contains reasonable values for the population value, if we only know the sample value, is called a *confidence interval*. Here we've chosen to use 2.58 standard deviations as our cut-off point, because we felt that 1% would be a small enough probability to dismiss the real population value as a reasonable candidate. Such a confidence interval based on this 1% cut-off point is called a 99% confidence interval.

One often also sees 95% confidence intervals, particularly in social and behavioural sciences. Because with the normal distribution, 5% of the observations lie more than 1.96 standard deviations away from the mean, the 95% confidence interval is constructed by subtracting/addding 1.96 standard errors from/to the sample value. Thus, in the case of our bottle sample, the 95% confidence interval for the population slope is from $-0.088 - 1.96 * 0.084$ to $-0.088 + 1.96 * 0.084$, so reasonable values for the population slope are those values between -0.25 and 0.08. Luckily, this corresponds to the truth, because we happen to know that the population slope is equal to 0.001. In real life, we don't know the population slope and of course it might happen that the true population value is not within the 95% confidence interval. If you want to make the probability of this being the case smaller, then you can use a 99%, a 99.9% or an even larger confidence interval.

3.6 Confidence intervals for smaller sample sizes

In the previous section we used the normal distribution to come up with 95% and 99% confidence intervals for the slope coefficient. These were constructed using 1.96 and 2.58 times the standard error, respectively. However, these numbers 1.96 and 2.58 can only be used when the sample size is large enough to say that the distribution of the sample slope is very close to a normal distribution. Earlier, we saw that the distribution of the sample slope is actually a t-distribution, that doesn't look normal at all for small sample sizes.

Therefore, for small sample sizes, we need to know the cut-off points that correspond to 5% and 1% probabilities for the t-distribution.

Figure 3.9 shows the case for the situation where the population slope is 0 and the sample size is 4. Suppose the standard error is equal to 1. Then this

figure shows that roughly 95% of the sample slopes lie between ± 3.18 standard errors below and above the mean (the red lines). In the same figure we also see that if sample size is 200, 95% of the sample means lie between ± 1.97 standard errors below and above the mean (the blue line). This is almost the same as for the normal distribution, where 95% of the observations lie between ± 1.96 standard deviations below and above the mean.

Because for every sample size, the middle region where 95% of the observations lie is different, there are tables available where these values can be found. However, these tables are built-in in every statistical package, so it is far easier to let SPSS construct the 95% confidence intervals for us.

Table 3.1: Quantiles for the standard normal and several t-distributions.

probs	norm	t198	t100	t50	t10	t2
0.0005	-3.29	-3.34	-3.39	-3.50	-4.59	-31.60
0.0010	-3.09	-3.13	-3.17	-3.26	-4.14	-22.33
0.0050	-2.58	-2.60	-2.63	-2.68	-3.17	-9.92
0.0100	-2.33	-2.35	-2.36	-2.40	-2.76	-6.96
0.0250	-1.96	-1.97	-1.98	-2.01	-2.23	-4.30
0.0500	-1.64	-1.65	-1.66	-1.68	-1.81	-2.92
0.1000	-1.28	-1.29	-1.29	-1.30	-1.37	-1.89
0.9000	1.28	1.29	1.29	1.30	1.37	1.89
0.9500	1.64	1.65	1.66	1.68	1.81	2.92
0.9750	1.96	1.97	1.98	2.01	2.23	4.30
0.9900	2.33	2.35	2.36	2.40	2.76	6.96
0.9950	2.58	2.60	2.63	2.68	3.17	9.92
0.9990	3.09	3.13	3.17	3.26	4.14	22.33
0.9995	3.29	3.34	3.39	3.50	4.59	31.60

But let us look at a few regularities. For several probabilities, the corresponding quantiles are presented in Table **??** for the standard normal distribution and several t-distributions.

The shape of the t-distribution is indicated by its *degrees of freedom*. The shape of the distribution of sample slopes when sample size is 200, is a t-distribution with 198 degrees of freedom. The shape of the distribution of sample slopes when sample size is 4, is a t-distribution with 2 degrees of freedom. In general, the shape of the distribution of sample slopes for sample size n, is a t-distribution with $n-2$ degrees of freedom. The higher the degrees of freedom, the more the corresponding t-distribution looks like a normal distribution. We will come back to degrees of freedom and the $n-2$ rule in the next section.

Table **??** shows for instance the cut-off points for 2.5% and 97.5% for the normal distribution (the 0.025 and 0.975 quantiles, respectively) and the t-distribution with 198 degrees of freedom: 1.96 and 1.97 standard deviations (standard errors) respectively. For the t-distribution with 100 degrees of freedom, the cutoff point is 1.98 standard errors. This would be the appropriate

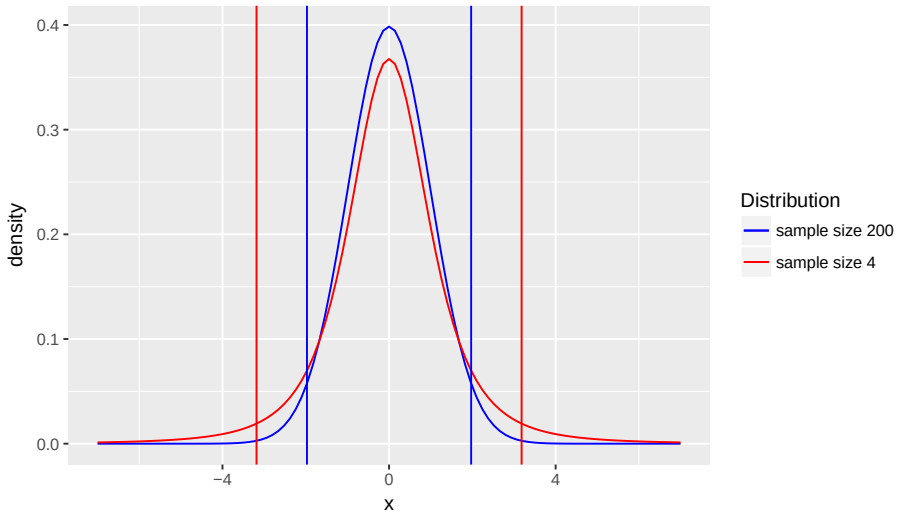


Figure 3.9: Two t-distributions when sample size is 4 or 200, with corresponding 95 percent intervals.

t-distribution for a sample size of 102. But for smaller sample sizes, the increase in number of standard errors goes up quickly: with 50 degrees of freedom (sample size 52), the cutoff is 2.01, for 10 degrees of freedom it is 2.23 and for 2 degrees of freedom it becomes even 4.30 standard errors. Thus, if we have a sample size of 4, we construct a 95% confidence interval of 4.30 standard errors below the sample slope and 4.30 standard errors above the sample slope.

If you want to have the 99% confidence interval, you look at the cut-off points for 0.005 and 0.995 which are -2.58 and +2.58, respectively, for the normal distribution, but -9.92 and +9.92 for a t-distribution with 2 degrees of freedom. Suppose we sample 4 bottles and find a sample slope of 5 with a standard error of 4, then the 99% confidence for the slope is from $5 - 9.92 \times 4$ to $5 + 9.92 \times 4$, so from -34.68 to 44.68, which is of course a huge interval. On the other hand, a sample of only 4 bottles is of course very small. It makes intuitive sense that if you have only 4 bottles to go on, you are very uncertain about the population slope: it could be anything!

In short, we can look up the cut-off points for 95%, 99% and other intervals from tables online, in books, or in statistical packages. Generally, the smaller the sample size, the lower the degrees of freedom, the larger the number of standard errors you need to construct your confidence intervals.

3.6.1 Exercises

1. Suppose we randomly pick 102 students from the University of Twente and determine the linear equation between age in years (independent variable) and height in cms (dependent variable). Suppose we find a slope coefficient

of 0.010, with a standard error of 0.009. Construct the 95% confidence interval for the slope in the entire population in UT students using table 12.1.

2. What can we say about values within this constructed confidence interval?

3. Suppose a professor believes the true slope is equal to 0: is that a reasonable belief given the finding of a sample slope of 0.010? Motivate your answer using the 95% confidence interval.

4. Suppose we randomly pick 52 adult inhabitants of Tuvalu and determine the linear equation between age in years (independent variable) and height in cms (dependent variable). Suppose we find an intercept of 168, with a standard error of 0.07. Construct the 99% confidence interval for the intercept in the entire population of adult inhabitants of Tuvalu using Table 12.1.

5. What can we say about values within this constructed confidence interval?

6. Suppose a Swedish diplomat stationed in Tuvalu believes the population intercept is equal to 169 cm: is that a reasonable belief given the finding of a sample intercept of 168? Motivate your answer using the 99% confidence interval.

Answers:

1. Sample size is 102, so degrees of freedom for the sample slope is 100. For a 95% interval, 2.5% of the observations should be on the left, and 2.5% of the observations should be on the right. The cut-off quantiles should therefore be 0.025 and 0.975. These cut-off values for the t-distribution with 100 degrees of freedom are -1.98 and 1.98. Therefore the 95% interval ranges from $0.010 - 1.98 \times 0.009$ to $0.010 + 1.98 \times 0.009$, so from -0.008 to 0.028.

2. These values are all reasonable values for the slope in the population of University of Twente students.

3. Yes, the value of 0 lies within the range from -0.008 to 0.028, so 0 is a reasonable value for the population slope.

4. Sample size is 52, so degrees of freedom for the sample slope is 50. For a 99% interval, 0.5% of the observations should be on the left, and 0.5% of the observations should be on the right. The cut-off quantiles should therefore be 0.005 and 0.995. The 99% cut-off values for the t-distribution with 100 degrees of freedom are therefore -2.68 and 2.68. Thus, the 99% interval ranges from $168 - 2.68 \times 0.07$ to $168 + 2.68 \times 0.07$, so from 167.8124 to 168.1876.

5. These values are all reasonable values for the slope in the population of all adult inhabitants of Tuvalu.

6. No, the value of 169 does not lie within the range from 167.8124 to 168.1876, so 169 is not a reasonable value for the population intercept.

3.7 Degrees of freedom

What does the term, "degrees of freedom" mean? It refers to the number of independent pieces of information in a sample of data.

Suppose that we have a sample with four values: 4, 2, 6, 8. There are four separate pieces of information here. There is no particular connection between these values. They are free to take any values, in principle. We could say that there are four degrees of freedom associated with this sample of data.

Now, suppose that I tell you that three of the values in the sample are 4, 2, and 6; and I also tell you that the sample average is 5. You can immediately deduce that the fourth value has to be 8. For any other value, the average would not be 5.

Once I tell you that the sample average is 5, I am effectively introducing a *constraint*. The value of the unknown fourth sample value is implicitly being determined from the other three values plus the constraint. That is, once the constraint is introduced, there are only three logically independent pieces of information in the sample. That is to say, there are only three "degrees of freedom", once the sample average is revealed.

Let's carry this example to regression analysis. Suppose I have four observations of variables x and y, where the values for x are 1, 2, 3 and 4. Each value of y is one piece of information. These y-values could be anything, so we say that we have 4 degrees of freedom. Now suppose I use a linear equation for these data points, and suppose I only use an intercept. Let the intercept be 5 so that we have $y = 5 + e$. Now the first bit of information for $x = 1$, y could be anything, say 2. The second and third bits of information for $x = 2$ and $x = 4$ could also be anything, say 6 and 2. Figure 3.10 shows these bits of information as dots in a scatterplot. Since we know that the intercept is equal to 5, with no slope (slope=0), we can also draw the regression line.

If we compute the residuals, we have residuals -3, 1 and -3 for these data points. When we sum them we get -3. Since we know that all residuals should sum to 0 in a regression analysis (see previous chapter), we can derive the fourth residual to be +5, since only then the residuals sum to 0. Therefore, the y-value for the fourth data point (for $x = 3$) has to be 10, since then the residual is equal to $10 - 5 = 5$.

In short, when we do a regression analysis with only an intercept, the degrees of freedom is equal to the number of data points (combinations of x and y) minus 1, or in short notation: $n - 1$, where n stans for sample size.

Now let's look at the situation where we do a regression analysis with both an intercept and a slope: suppose the intercept is equal to 3 and the slope is equal to 1: $y = 3 + 1x + e$. Then suppose we have the same x-values as the example above: 1, 2 and 4. When we give these x-values corresponding y-values, 2, 6, and 3, we get the plot in Figure 3.11.

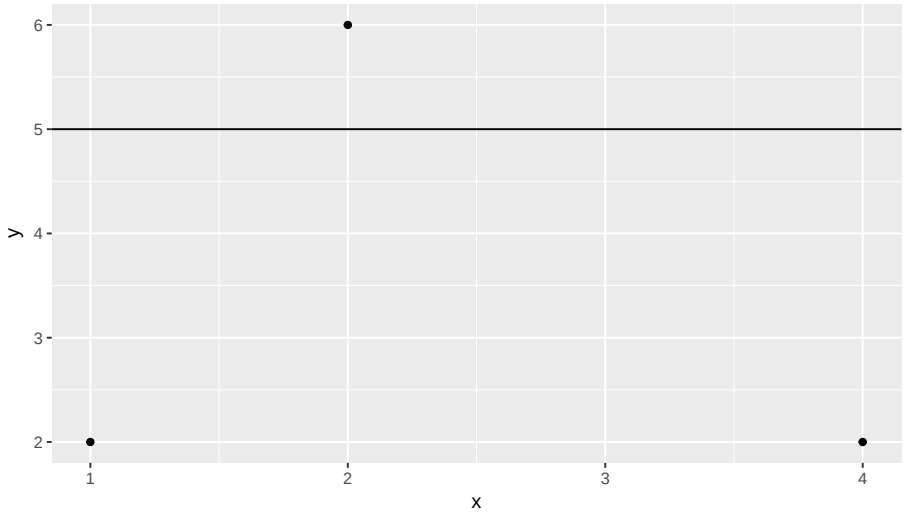


Figure 3.10: Distribution of the sample mean when sample size is 4 or 200

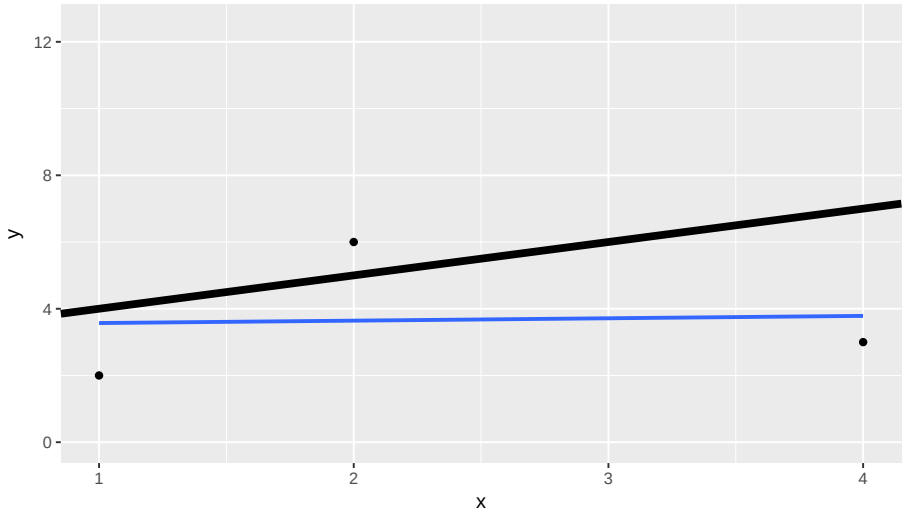


Figure 3.11: Distribution of the sample mean when sample size is 4 or 200

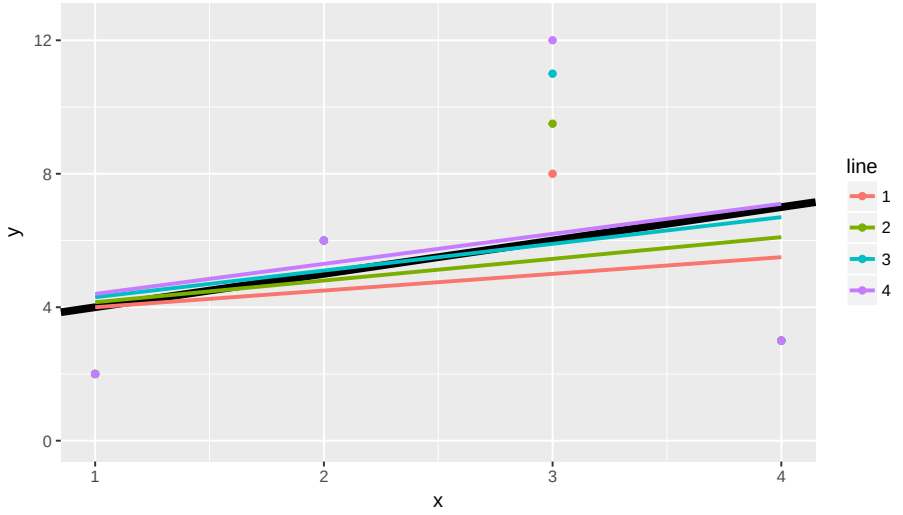


Figure 3.12: Different regression lines for different values of y if x=3.

The black line is the regression line that should be appropriate for these data set of four points. The blue line is the regression line based on only the three visible data points. Now the question is, is it possible for a fourth data point with $x = 3$, to think of a y-value such that the regression line based on these four data points is equal to $y = 3 + 1x$? In other words, can we choose a y-value such that the blue line exactly overlaps with the black line?

Figure 3.12 shows a number of possibilities for the value of y if $x = 3$. It can be seen, that it is impossible to pick a value for y such that we get a regression equation $y = 3 + 1x$. The blue line for instance comes closest to the black line. This is the regression line when $y = 11$. However, it does not exactly overlap the black line. If you lower values for y such as 9.5 (green line) or 8 (red line), the regression lines still not overlap, nor for a higher value of y such as 12 (purple line).

So, with 4 data points, we can never freely choose 3 residuals in order to satisfy the constraint that a particular regression equation holds. We have less then 3 degrees of freedom because it is impossible to think of a fitting fourth value. It turns out, that in this case we can only choose 2 residuals freely, and the remaining residuals are then already determined. To prove this requires matrix algebra, but the gist of it is that if you have a regression equation with both an intercept and a slope, the degrees of freedom is equal to the number of data points (sample size) minus 2: $n - 2$.

Generally, these degrees of freedom based on the number of residuals that could be freely chosen, given the constraints of the model, are termed *residuals degrees of freedom*. When using regression models, one usually only reports these residual degrees of freedom. Later on in this book, we will see instances

where one also should use *model degrees of freedom*. For now, it suffices to know what is meant by residuals degrees of freedom.

Chapter 4

Inference II: hypothesis testing, p-values and beyond

4.1 The null-hypothesis

Often, data analysis is about finding an answer to the question whether there is a relationship between two variables. In most cases, the question pertains to the population: is there a relationship between variable y and variable x in the population? In many cases, one looks for a linear relationship between two variables.

One common method to answer this question is to analyse a sample of data, apply a linear model, and look at the slope. However, one then knows the slope in the sample, but not the slope in the population. We have seen that the slope in the sample can be very different from the slope in the population. Suppose we find a slope of 1: does that mean there is a slope in the population or that there is no slope in the population?

In inferential data analysis, one often works with two hypotheses: the *null-hypothesis* and the *alternative hypothesis*. The null-hypothesis states that the population slope is equal to 0 and the alternative hypothesis states that there is a slope that is different from 0. Remember that if the population slope is equal to 0, that is saying that there is no linear relationship between x and y (that is, you cannot predict one variable on the basis of the other variable). Therefore, the null-hypothesis states there is no linear relationship between x and y in the population. If there is a slope, whether positive or negative, is the same as saying there is a linear relationship, so the alternative hypothesis states that there is a linear relationship between x and y in the population.

The null-hypothesis is often denoted as H_0 and the alternative hypothesis is often denoted as H_1. In formula form, we have

$$H_0 : \beta_{slope} = 0 \tag{4.1}$$

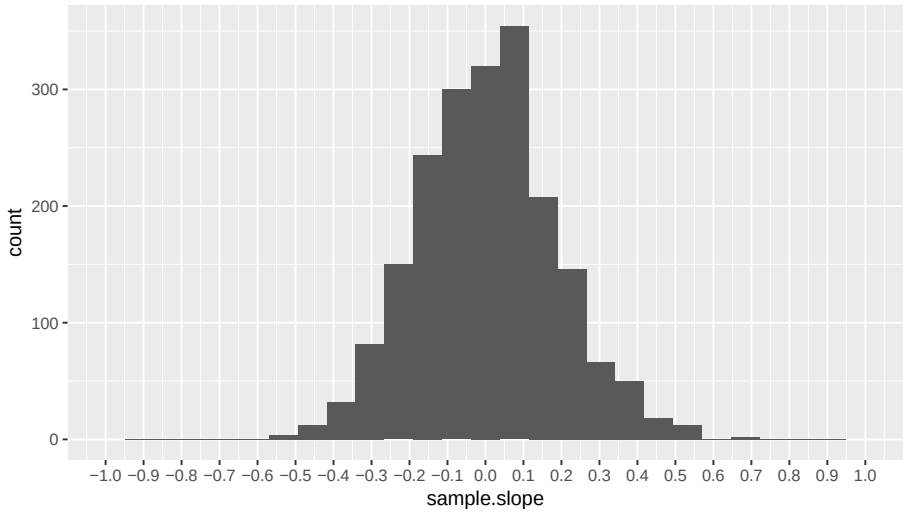


Figure 4.1: Distribution of the sample slope.

$$H_1 : \beta_{slope} \neq 0 \tag{4.2}$$

The population slope, β_{slope}, is either 0 or it is not. Our data analysis is then aimed at determining which of these two hypotheses is true. Key is that we do a thought experiment on the null-hypothesis: we wonder what would happen if the population slope would be really 0. In our imagination we draw many samples of a certain size, say 40 data points, and then determine the slope for each sample. Earlier we learned that the many sample slopes would form a histogram in the shape of a t-distribution with $n - 2 = 38$ degrees of freedom. For example, suppose we would draw 1000 samples of size 40, then the histogram of the 1000 slopes would be like depicted in Figure 4.1

From this histogram we see that all observed sample slopes are well between -0.8 and 0.8. This gives us the information we need. Of course, we have only one sample of data, and we don't know anything about the population data. But we *do* know that *if the population slope is equal to 0*, then it is very unlikely to find a sample slope of say 1 or -1. Thus, if we happen to find a sample slope of say -1, we know that this finding is very unlikely *if we hold the null-hypothesis to be true*. In other words, if the population slope is equal to 0, it would be quite improbable to find a sample slope of -1 or smaller. Therefore, we regard the null-hypothesis to be false, since it does not provide a good explanation of why we found a sample slope of -1. In that case, we say that *we reject the null-hypothesis*.

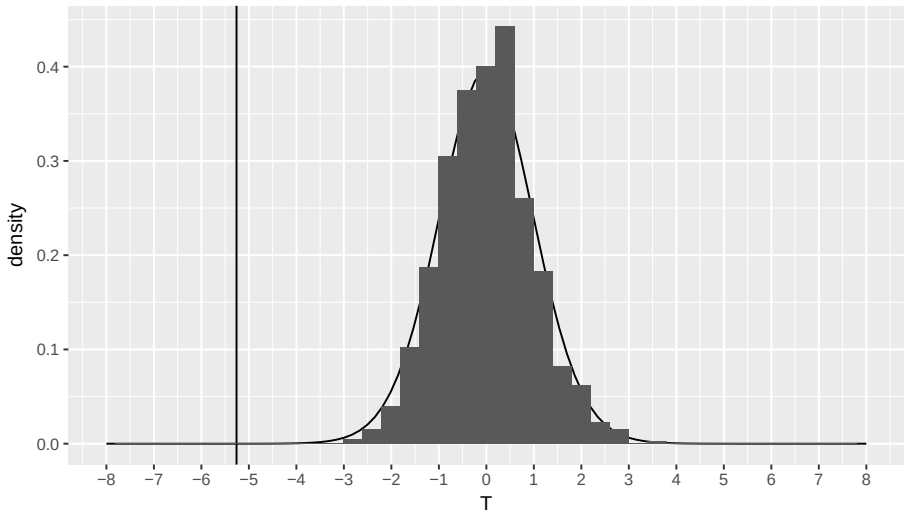


Figure 4.2: The histogram of 1000 sample slopes and its corresponding theoretical t-distribution with 38 degrees of freedom. The vertical line represents the T-value of -5.56.

4.2 The p-value

The p-value is a probability. It represents the probability of observing certain events, given that the null-hypothesis is true.

In the previous section we saw that if the population slope is 0, and we drew 1000 samples of size 40, we did not observe a sample slope of -1 or smaller. In other words, the frequency of observing a slope of -1 or smaller was 0. If we would draw more samples, we theoretically could observe a sample slope of -1, but the probability that that happens for any new sample we can estimate at less than 1 in a 1000, so less than 0.001.

This estimate of the p-value was based on 1000 randomly drawn samples of size 40 and then looking at the frequency of certain values in that data set. But there is a short-cut, for we know that the distribution of sample slopes has a t-distribution if we standardize the sample slopes. Therefore we do not have to take 1000 samples and estimate probabilities, but we can look at the t-distribution directly, using tables online or in statistical packages.

Figure 4.2 shows the t-distribution that is the theoretical distribution corresponding to the histogram in Figure 4.1. If the standard error is equal to 0.19, and the hypothetical population slope is 0, then the T-statistic associated with a slope of -1 is equal to $\frac{-1-0}{0.19} = -5.26$. With this value, we can look up in the tables, how often such a value of -5.26 or smaller occurs in a t-distribution with 38 degrees of freedom. In the tables we find that the probability that this occurs is 0.00000294. So, the fact that the T-statistic has a t-distribution gives us the opportunity to exactly determine certain probabilities, including the p-value.

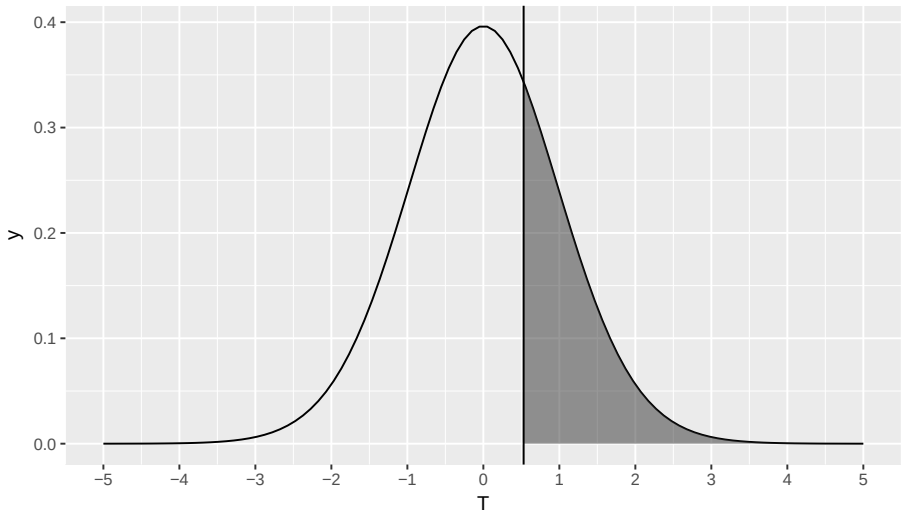


Figure 4.3: Probability of a T-value larger than 0.53.

Now let's suppose we have only one sample of 40 bottles, and we find a slope of 0.1 with a standard error of 0.19. Then this value of 0.1 is $(0.1-0)/0.19 = 0.53$ standard errors away from 0. Thus, the T-statistic is 0.53. We then look at the t-distribution with 38 degrees of freedom, and see that such a T-value of 0.53 is not very strange: it lies well within the middle 95% of the t-distribution (see Figure **??**.

Let's determine the p-value again for this slope of 0.1: we determine the probability that we obtain such a T-value of 0.53 or larger. Figure 4.3 shows the area under the curve for values of T that are larger than 0.53. This area under the curve can be seen as a probability. The total area under the curve of the t-distribution amounts to 1. If we know the area of the shaded part of the total area, we can compute the probability of finding T-values larger than 0.53.

In tables online, in books, or available in statistical packages, we can look up how large this area is. It turns out to be 0.3. So, if the population slope is equal to 0 and we draw an infinite number of samples of size 40 and compute the sample slopes, then 30% of them will be larger than our sample slope of 0.1. The proportion of the shaded area is what we call a *one-sided* p-value. We call it one-sided, because we only look at one side of the t-distribution: we only look at values that are larger than our T-value of 0.53.

We conclude that a slope value of 0.1 is not that strange to find if the population slope is 0. By the same token, it would also have been probable to find a slope of -0.1, corresponding to a T-value of -0.53. Since the t-distribution is symmetrical, the probability of finding a T-value of less than -0.53 is depicted in Figure 4.4, and of course this probability is also 0.3.

Remember that the null-hypothesis is that the population slope is 0, and the

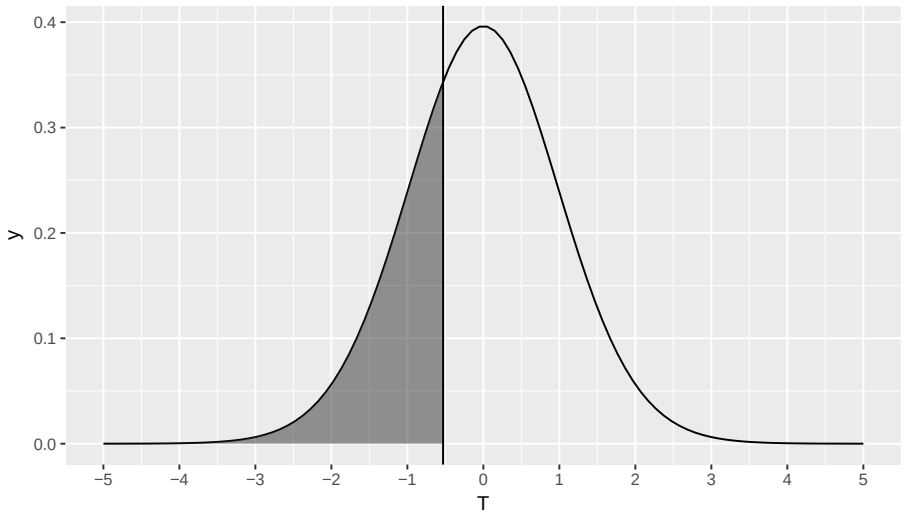


Figure 4.4: Probability of finding a T-value smaller than -0.53.

alternative hypothesis is that the population slope is *not* 0. We should therefore conclude that if we find a very large positive *or* negative slope, large in the sense of the number of standard errors away from 0, that the null-hypothesis is unlikely to be true. Therefore, if we find a slope of 0.1 or -0.1, then we should determine the probability of finding a T-value that is larger than 0.53 or smaller than -0.53. This probability is depicted in Figure 4.5 and is equal to twice the one-side p-value, $2 \times 0.2995977 = 0.5991953$.

This probability is called the *two-sided* p-value. This is the one that should always be used, since the alternative hypothesis is also two-sided: the population slope can be positive or negative. The question now is: is a sample slope of 0.1 enough evidence to reject the null-hypothesis? To determine that, we determine how many standard errors away from 0 the sample slope is and we look up in tables how often that happens. Thus in our case, we found a slope that is 0.53 standard errors away from 0 and the tables told us that the probability of finding a slope that is at least 0.53 standard deviations away from 0 (positive or negative) is equal to 0.5991953. We find this probability rather large, so we decide that we *do not reject the null-hypothesis*.

4.3 Hypothesis testing

In the previous section, we found a one-sided p-value of 0.00000294 for a sample slope of -1 and more or less concluded that this probability was rather small. The two-sided p-value would be twice this value, so 0.00000588, which is still very small. Next we determined the p-value associated with a slope of 0.1 and found a p-value of 0.60. This probability we found was rather large, and we

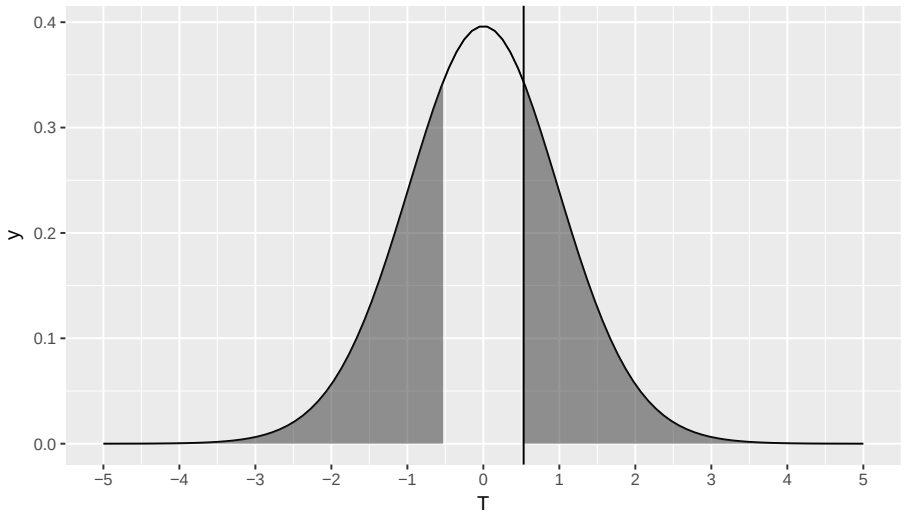


Figure 4.5: The vertical line represents a T-value of 0.53. The shaded area represents the two-sided p-value: the probability of obtaining a T-value smaller than -0.53 or larger than 0.53.

decided not to reject the null-hypothesis. In other words, the probability was so large that we thought that the hypothesis that the population slope is 0 should not be rejected based on our findings.

When should we think the p-value is small enough to conclude that the null-hypothesis can be rejected? When can we conclude that the hypothesis that the population slope is 0 is not supported by our sample data? This was a question posed to the founding father of statistical hypothesis testing, Sir Ronald Fischer. In his book *Statistical Methods for Research Workers* (1925), Fisher proposed a probability of 5%. He advocated 5% as a standard level for concluding that there is evidence against the null-hypothesis. However, he did not see it as an absolute rule: "If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05...". So Fisher saw the p-value as an informal index to be used as a measure of discrepancy between the data and the null-hypothesis: The null-hypothesis is never proved, but is possibly disproved.

Later, Jerzy Neyman and Egon Pearson saw the p-value as an instrument in decision making: is the null-hypothesis true, or is the alternative hypothesis true? You either reject the null-hypothesis or you don't, there is nothing in between. This view to data-analysis is still rather popular in the social and behavioural sciences, but also in particle physics. In order to make such a black-and-white decision, you decide before-hand, that is, before collecting data,

what *level of significance* you choose for your p-value to decide whether to reject the null-hypothesis. For example, as your significance level, you might want to choose 1%. Let's call this chosen significance level α. Then you collect your data, you apply your linear model to the data, and find that the p-value associated with the slope equals p. If this p is smaller than or equal to α you *reject the null-hypothesis*, and if p is larger than α then you *do not reject the null-hypothesis*. A slope with a $p \leq \alpha$ is said to be *significant*, and a slope with a $p > \alpha$ is said to be *non-significant*. If the slope is significant, then one should reject the null-hypothesis and say there is a slope in the population different from zero. If the slope is not significant, then one should not reject the null-hypothesis and say there is no slope in the population (i.e., the slope is 0).

4.3.1 Exercises

1. Suppose you test the null-hypothesis that in a linear equation describing the relationship between the mass of a planet and its volume, the slope equals 0:

$$mass = \beta_0 + \beta_1 volume + \epsilon \tag{4.3}$$

   State the null-hypothesis.

2. You set your significance level to 1%, so $\alpha = 0.01$. Next, you measure 52 planets and you find a sample slope of $b_1 = 6$, with a standard error of 2.24. Determine the T-statistic with which you test the null-hypothesis.

3. Determine the two-sided p-value on the basis of Table 12.1.

4. Do you reject or do you not reject the null-hypothesis? What does this mean?

5. A car manufacturer wants to build safe cars. One of the engineers conducts collision experiments: cars with a certain velocity are directed towards a parked car. Both the velocity and the deepness of the dent in the parked car is measured. She expects to see that high velocity creates deeper dents and she applies a regression model.

$$deepness = \beta_0 + \beta_1 velocity + \epsilon \tag{4.4}$$

   State the null-hypothesis.

6. The engineer sets her significance level to 5%, so $\alpha = 0.05$. Next, she measures 4 cars with speeds between 90 mph and 92 mph and she finds a sample slope of $b_1 = 2$, with a standard error of 1.5. Determine the T-statistic with which you test the null-hypothesis.

7. For her T-statistic with 2 degrees of freedom, she finds a two-sided p-value of 0.3140057. Is the effect of velocity on deepness of the dent significant?

8. Should the engineer reject or not reject the null-hypothesis? What does this mean?

9. Could you think of possible reasons why the engineer does not find an effect of velocity on the deepness of the dent?

Answers:

1. $$H_0 : \beta_1 = 0 \tag{4.5}$$

2. $T = \frac{6-0}{2.24} = 2.68$

3. Degrees of freedom is $52 - 2 = 50$. From the table for a t-distribution with 50 degrees of freedom, we see that a T-value of 2.68 is the 0.995 quantile. Thus, half a percent of the T-values are larger than 2.68. Because of symmetry, half a percent of the T-values is smaller than -2.68. So in total, 1% of the T-values are at least 2.68 away from the mean (both directions). Therefore, the two-sided p-value is 0.01.

4. Our p-value of 0.01 is equal to our α and we therefore reject the null-hypothesis. This means that we conclude that the slope coefficient in all planets in the universe is not 0. There is a relationship between the volume of a planet and its mass.

5. $$H_0 : \beta_1 = 0 \tag{4.6}$$

6. $T = (2 - 0)/1.5 = 2/1.5 = 1.33$

7. The p-value 0.3140057 is larger than her α, so the effect of velocity is not significant.

8. The effect is not significant so she should not reject the null-hypothesis. This means that the conclusion is that there is no relationship between the velocity of the incoming car and the deepness of the dent in the receiving car.

9. First of all, there were only 4 cars tested. A small sample size results in a relatively large standard error, so a relatively small T-statistic. The higher the T-value the lower the p-value. Second, there was hardly any variation in the speed of the incoming car: if you want to find an effect, there should be cars with both high and low velocities, otherwise you won't see any differences in the dents.

4.4 Type I and Type II errors in decision making

Since data-analysis is about probabilities, there is always a chance that you make the wrong decision: you can wrongfully reject the null-hypothesis, or you can wrongfully accept the null-hypothesis. Pearson and Neyman distinguished between two kinds of error: one could reject the null-hypothesis while it is actually true (error of the first kind, or type I error) and one could accept the null-hypothesis while it is not true (error of the second kind, or type II error).

To illustrate the difference between type I and type II errors, let's recall the famous fable by Aesop about the boy who cried wolf. The tale concerns a shepherd boy who repeatedly tricks other people into thinking a wolf is attacking his flock of sheep. The first time he cries "There is a wolf!", the men working in an adjoining field come to help him. But when they repeatedly find there is no wolf to be seen, they realise they are being fooled by the boy. One day, when a wolf *does* appear and the boy again calls for help, the men believe that it is another false alarm and the sheep are eaten by the wolf.

In this fable, we can think of the null-hypothesis as the hypothesis that there is no wolf. The alternative hypothesis is that there is a wolf. Now, when the boy cries wolf the first time, there is in fact no wolf. The men from the adjoining field make a type I error: they think there is a wolf while there isn't. Later, when they are fed up with the annoying shepherd boy, they don't react when the boy cries "There is a wolf!". Now they make a type II error: they think there is no wolf, while there actually is a wolf.

Let's return to regression analysis. Suppose you want to determine the slope for the effect of age on height in children. Let the slope now stand for the slope: either there is no slope (no wolf, H_0) or there is a slope (wolf, H_1). The null-hypothesis is that the slope is 0 in the population of all children (a slope of 0 means there is no slope) and the alternative hypothesis that the slope is not 0, so there is a slope. You might study a sample of children and you might find a certain slope. You might decide that if the p-value is lower than a critical value you conclude that the null-hypothesis is not true. Suppose you think a probability of 10% is small enough to reject the null-hypothesis as true. In other words, if $p \leq 0.10$ then we no longer think 0 is a reasonable value for the population slope. In this case, we have fixed our α or type I error rate to be $\alpha = 0.10$. This means that if we study a random sample of children, we look at the slope and find a p-value of 0.11, then we do not reject the null-hypothesis. If we find a p-value of 0.10, then we reject the null-hypothesis.

Note that the probability of a type I error is the same as our α for the significance level. Suppose we set our $\alpha = 0.05$. Then for any p-value equal or smaller than 0.05, we reject the null-hypothesis. Suppose the null-hypothesis is true, how often do we then find a p-value smaller than 0.05? We find a p-value smaller than 0.05 if we find a T-value that is above a certain threshold. For instance, for the t-distribution with 198 degrees of freedom, the critical value is ± 1.97, *because only in 5% of the cases we find a* T*-value of* ± 1.97 *or more if the*

null-hypothesis is true! Thus, if the null-hypothesis is true, we see a T-value of at least ± 1.97 in 5% of the cases. Therefore, we see a signficicant p-value in 5% of the cases if the null-hypothesis is true. This is exactly the definition of a Type I error: the probability that we reject the null-hypothesis (finding a significant p-value), given that the null-hypothesis is true. So we call our α-value the type I error rate.

Suppose 100 researchers are studying a particular slope. Unbeknownst to them, the population slope is exactly 0. They each draw a random sample from the population and test whether their sample slope is significantly different from 0. Suppose they all use different sample sizes, but they all use the same α of 0.05. Then we can expect that about 5 researchers will reject the null-hypothesis (finding a p-value less than or smaller than 0.05) and about 95 will not reject the null-hypothesis (finding a p-value of more than 0.05).

Fixing the type I error rate should always be done *before* data collection. How willing are you to take a risk of a type I error? You are free to make a choice about α, as long as you report it.

If α represents the probability of making a type I error, then we can use β to represent the probability of not rejecting the null-hypothesis while it is not true (type II error, thinking there is no wolf while there is). However, setting the β value prior to data collection is a bit trickier than choosing your α. It is not possible to compute the probability that we find a non-significant effect ($p > \alpha$), given that the alternative hypothesis is true, because the alternative hypothesis is only saying that the slope is not equal to 0. In order to compute β, we need to think first of a reasonable size of the slope that we expect. For example, suppose we believe that a slope of 1 is quite reasonable, given what we know about growth in children. Let that be our alternative hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 = 1$$

Next, we determine the distribution of sample slopes under the assumption that the population slope is 1. We know that this distribution has a mean of 1 and a standard deviation equal to the standard error. We also know it has the shape of a t-distribution. Let sample size be equal to 102 and the standard error 2. If we standardize the slopes by dividing by the standard error, we get the two t-distributions in Figure 4.6: one distribution of T-values if the population slope is 0 (centered around T=0), and one distribution of T-values if the population slope is 1 (centered around $T = 1/2 = 0.5$.

Let's fix α to 10%. The shaded areas represent the area where $p \leq \alpha$: for all values of T smaller than -1.6859545 and larger than 1.6859545, we reject the null-hypothesis. The probability that this happens, *if the null-hypothesis is true*, is equal to α which is 0.10 in this example. The probability that this happens *if the alternative hypothesis is true* (i.e., population slope is 1), is depicted in Figure 4.7.

The shaded area in Figure 4.7 turns out to be 0.1415543. This represents

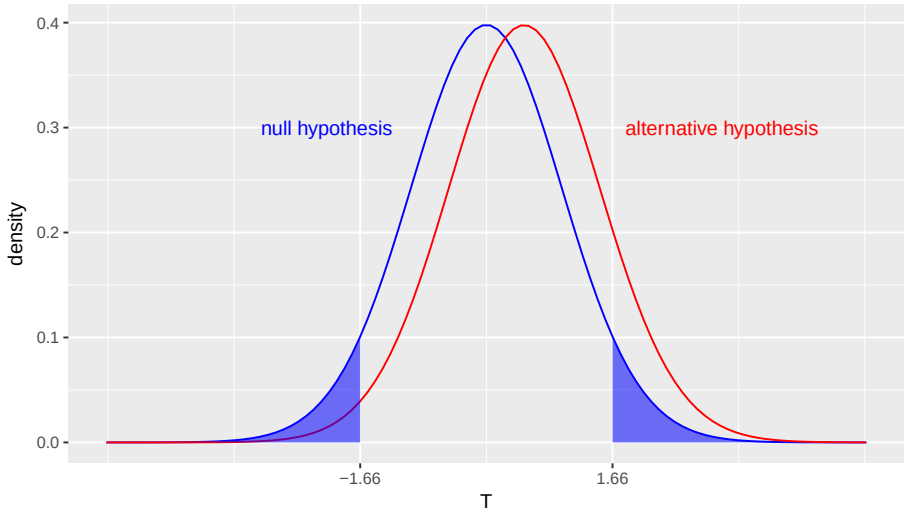


Figure 4.6: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Blue area depicts the probability that we find a p-value value smaller than 0.10 if the population slope is 0 (alpha).

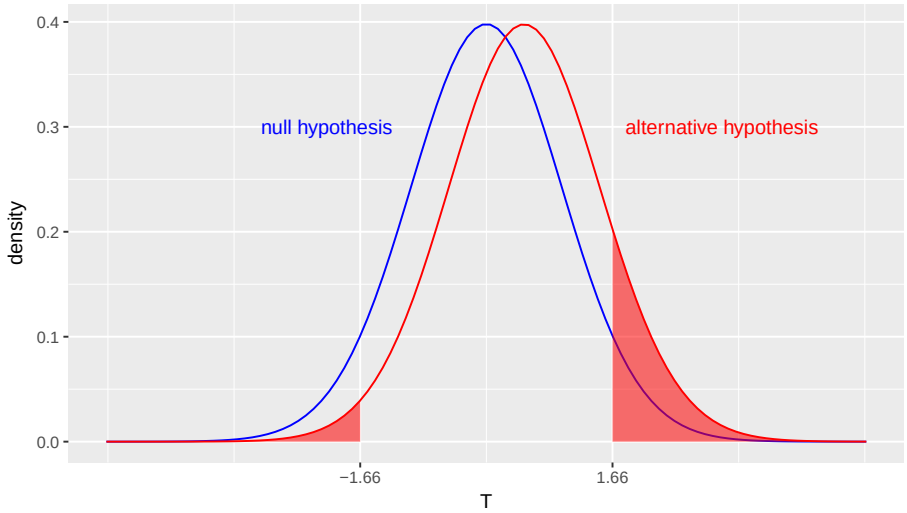


Figure 4.7: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Shaded area depicts the probability that we find a p-value value smaller than 0.10 if the population slope is 1 (1-beta).

the probability that we find a significant effect, *if the population slope is 1*. This is actually the *complement*[1] of the probability to find a *non*-significant effect, *if the population slope is 1*, which is defined as β. Therefore, the shaded area in Figure 4.7 represents $1-\beta$: the probability of finding a significant p-value, if the population slope is 1. In this example, $1-\beta$ is equal to 0.1415543, so β is equal to its complement, $1-0.1415543 = 0.8584457$.

In sum, in this example with an α of 0.10 and assuming a population slope of 1, we find that the probability of a type II error is 0.86: if there is a slope of 1, then we have a 86% chance of wrongly concluding that the slope is 0.

Type I and II error rates α and β are closely related. If we feel that a significance level of $\alpha = 0.10$ is too high, we could choose a level of 0.01. This ensures that we are less likely to reject the null-hypothesis when it is true. The critical value for our T-statistic is then equal to ± 2.6258905, see Figure 4.8. In Figure 4.9 we see that if we change α, we also get a different value for $1-\beta$, in this case 0.0196567.

Thus, if we use smaller values for α, we get smaller values for $1-\beta$, so we get larger values for β. This means that if we lower the probability of rejecting the null-hypothesis given that it is true (type I error) by choosing a lower value for α, we inadvertently increase the probability of failing to reject the null-hypothesis given that it is not true (type II error). One should therefore always strike a balance between the two types of errors. One should consider how bad it is to think that the slope is not 0 while it is, and how bad it is to think that the slope is 0, while it is not. If you feel that the first mistake is worse than the second one, then make sure α is really small, and if you feel that the second mistake is worse, then make α not too small. Another option, and a better one, to avoid type II errors, is to increase sample size, as we will see in the next section.

4.4.1 Exercises

1. When we talk about decision making in data analysis, what do we mean by β?
2. What do we mean by $1-\beta$?
3. What doe we mean by α?
4. What do we mean by making a type I error?
5. What do we mean by making a type II error?
6. What do we mean by $1-\alpha$?

Answers:

1. The type II error rate, or the probability of not rejecting the null-hypothesis while the null-hypothesis is not true.

[1] Explaination complement.....

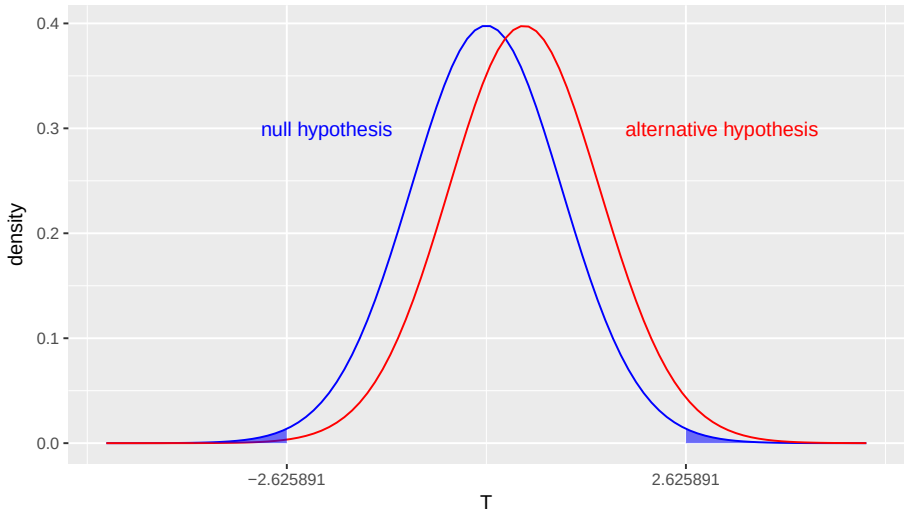


Figure 4.8: Different t-distributions of the sample slope if the population slope equals 0 (left curve), and if the population slope equals 1 (right curve). Grey area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 0.

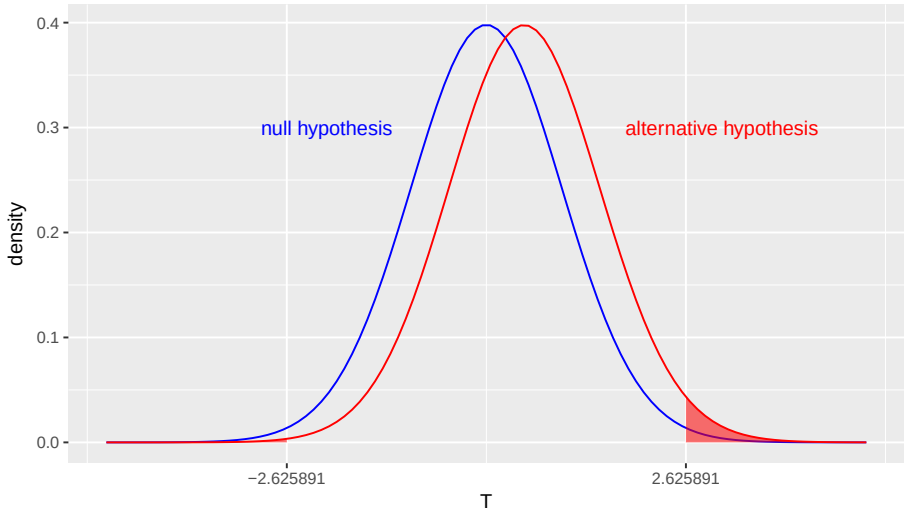


Figure 4.9: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Red area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 1: 1-beta=.

2. The probability of finding a significant effect if the alternative hypothesis is true.

3. The type I error rate, or the probability of rejecting while the null-hypothesis is true

4. Wrongly concluding that the null-hypothesis is not true.

5. Wrongly concluding that the null-hypothesis is true.

6. The probability of not rejecting the null-hypothesis while the null-hypothesis is true.

4.5 Statistical power

Null-hypothesis testing only involves the null-hypothesis: we look at the sample slope, compute the T-statistic and then see how often such a T-value and larger values occur given that the population slope is 0. Then we look at the p-value and if that p-value is smaller than or equal to α, we reject the null-hypothesis. Therefore, null-hypothesis testing does not involve testing the alternative hypothesis. We can decide what value we choose for our α, but not our β. The β is dependent on what the actual population slope is, and we simply don't know that.

As stated in the previous section, we can compute β only if we have a more specific idea of an alternative value for the population slope. We saw that we needed to think of a reasonable value for the population slope that we might be interested in. Suppose we have the intuition that a slope of 1 could well be the case. Then, we would like to find a p-value of less than α if indeed the slope were 1. We hope that the probability that this happens is very high: the conditional probability that we find a T-value large enough to reject the null-hypothesis, given that the population slope is 1. This probability is actually the *complement* of β, $1 - \beta$: the probability that we reject the null-hypothesis, given that the alternative hypothesis is true. This $1 - \beta$ is often called the *statistical power* of a null-hypothesis test. When we think again about the boy cried wolf: the power is the probability that the villagers come to the rescue if there is indeed a wolf attacking the sheep. The power of a test should always be high: if there is a population slope that is not 0, then of course you would like to detect it by finding a significant T-value!

In order to get a large value for $1 - \beta$, we should have large T-values in our data-analysis. There are two ways in which we can increase the value of the T-statistic. Since with null-hypothesis testing $T = (b - 0)/se = b/se$, we can get large values for T if we have a small standard error, se. We get a small standard error if we have a large sample size, see Section **??**.

If we go back to the example of the previous section where we had a sample size of 102 children and our alternative hypothesis was that the population slope was 1, we found that the t-distribution for the alternative hypothesis was centered around 0.5, because the standard error was 2. Suppose that we would

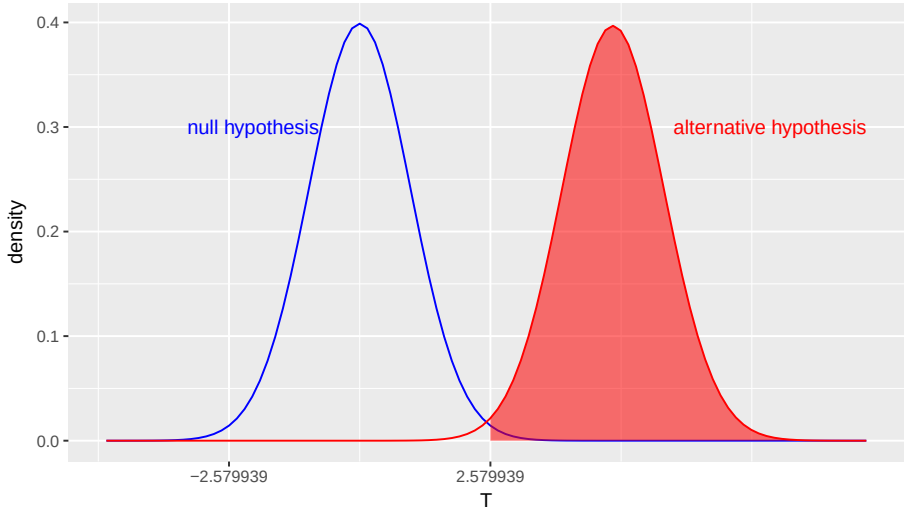


Figure 4.10: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Now for a larger sample size. Shaded area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 1.

increase sample size to 1200 children, then our standard error might be 0.2. Then our t-distribution for the alternative hypothesis is centerd at 5. This is shown in Figure 4.10.

We see from the shaded area that if the population slope is really 1, there is a very high chance that the T-value for the sample slope will be larger than 2.58, the cutoff point for an α of 0.01 and 1198 degrees of freedom. The probability of rejecting the null-hypothesis while it is not true, is therefore very large. This is our $1-\beta$ and we call this the power of the null-hypothesis test. We see that with increasing sample size, the power to find a significant T-value increases too.

By the same token, the power also increases with increasing population slope. If the population slope were 10, and we only had a sample of 102 children (resulting in a standard error of 2), the t-distribution for the alternative hypothesis that the population slope is centered around $B/se = 10/2 = 5$, resulting in the same plot as in Figure 4.10, with a large value for $1-\beta$.

In sum: the statistical power of a test is the probability that the null-hypothesis is rejected, given that it is not true. This probability is equal to $1-\beta$. The statistical power of a test increases with sample size, and depends on the actual population slope. The further away the population slope is from 0 (positive or negative), the larger the statistical power. Earlier we also saw that $1-\beta$ decreases with increasing α: the smaller α, the lower the power.

4.5.1 Exercises

1. Prior to an experiment with 100 participants, a researcher fixes α to 0.05. She expects to find a slope of at least 2. She then computes the power of the test and finds 0.50. What does this mean?

2. She would like to increase the power of her test, but is unable to increase sample size due to financial constraints. What can she do?

Answers:

1. A power of 0.50 means that if the alternative hypothesis is true (i.e. the slope is 2), then the probability of finding a significant p-value ($p \leq 0.05$) is 0.50.

2. She can't change sample size, nor can she change the population slope. She can only change her α. The lower the α, the lower the power. She could therefore use a higher α, for instance 0.10. However, this of course raises the probability of a type I error.

4.6 Power analysis

Because of these relationships between statistical power, α, sample size and the actual population slope, we can compute the statistical power for any combination of α, sample size and hypothetical population slope.

If you really care about the quality of your research, you carry out a *power analysis* prior to collecting data. With such an analysis you can ind out how large your sample size should be. You can find many tools online that can help you with that.

Suppose you want to minimize the probability of a type I error, so you choose an $\alpha = 0.01$. Next, you think of what kind of population slope you would like to find, if it indeed has that value. Suppose that you feel that if the population slope is 0.15, you would really like to find a significant T-value so that you can reject the null-hypothesis. Next, you have to specify how badly you want to reject the null-hypothesis if indeed the population slope is 0.15. If the population slope is really 0.15, then you would like to have a high probability to find a T-value large enough to reject the null-hypothesis. This is of course the power of the test, $1 - \beta$. Let's say you want to have a power of 0.90. Now you have enough information to calculate how large your sample size should be.

Let's look at G*power[2], an application that can be downloaded from the web. If we start the app, we can ask for the sample size required for a slope of 0.15, an α of 0.01, a power $(1 - \beta)$ of 0.90. Let the standard deviation of our dependent variable (y=height) be 3 and the standard deviation of our independent variable (x=age) be 2. Then we get the input as displayed in Figure 4.11. Note that you should use two-sided p-values, so tails=two. From the output we see that the required sample size is 1477 children.

[2] http://www.gpower.hhu.de/

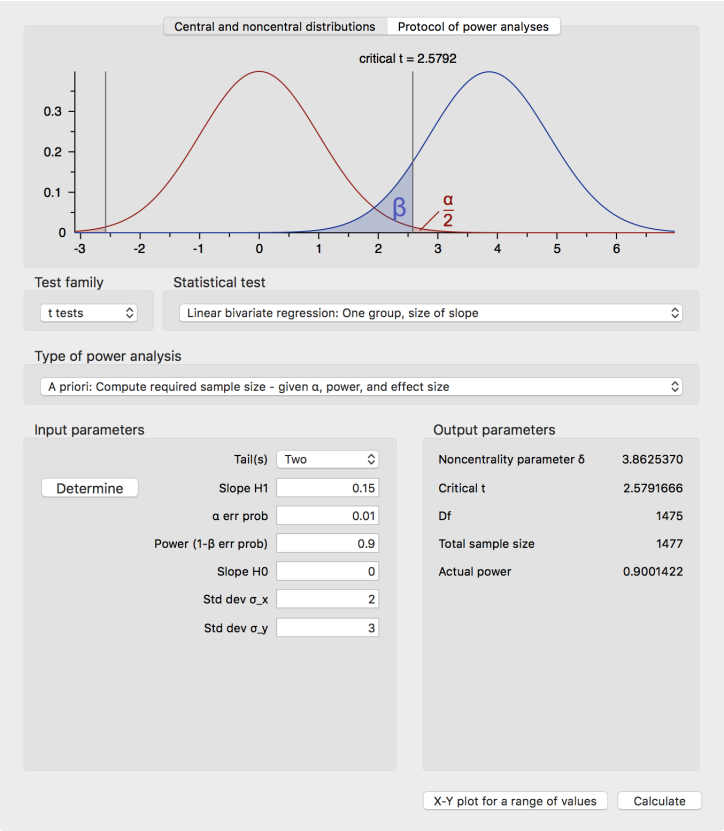


Figure 4.11: G*power output for a simple regression analysis.

4.6.1 Exercises

1. You want to predict height by age in children. Use G*power to find out how large sample size should be if you want to find a slope of 0.15 with a type I error rate of 0.01, and a power of 80%. Suppose the standard deviation of height is about 3 and the standard deviation of age is about 2.

2. A teacher friend says you can use her children, all having the age of 9. The standard deviation for age in a classroom of 9-year-olds is about 0.5, and their heights have a standard deviation of about 1. How many children of age 9 would you need in order to get your power of 80%?

3. A hockey friend says you can use children from his hockey club. They have ages between 5 and 15, and the standard deviation is about 2. The standard deviation of their heights is about 0.5. How many hockey club

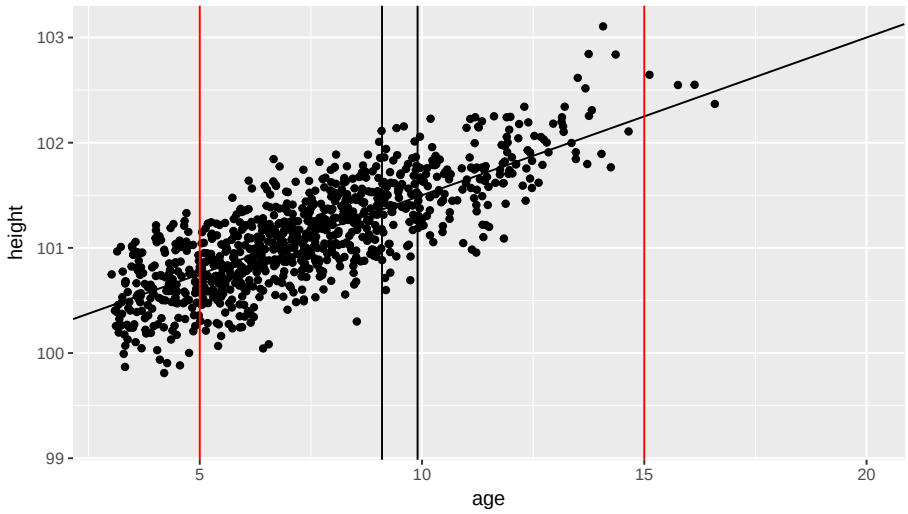


Figure 4.12: Illustration of exercise: the relation between age and height in children.

children would you need for your power of 80%?

4. Explain why you need so many children from age 9, and fewer children with ages between 5 and 15. Sketch a scatterplot, if that helps you.

Answers:

1. 1160 children.
2. 2068 children.
3. 25 children.
4. See Figure 4.12. In order to see a relationship between variation in variable x and variable y, there should at least be variation in one of them. So if you want to see an effect, make sure you see a lot of variation in one the two variables, for instance use a sample with a large spread in age.

4.7 Criticism on null-hypothesis testing and p-values

The practice of null-hypothesis significance testing (NHST) is widespread. However, from the beginning it has received much criticism. One of the first to critize the approach was the inventor of the p-value, Sir Ronald Fisher himself. Fisher explicitly contrasted the use of the p-value for statistical inference in science with the NeymanPearson approach, which he termed "Acceptance

Procedures". Whereas in the Pearson-Neyman approach the only relevance of the p-value is whether it is smaller or larger than the fixed significance level α, Fisher emphasized that the exact p-value should be reported to indicate the strength of evidence against the null-hypothesis. He emphasized that no single p-value can refute a hypothesis, since chance always allows for type I and type II errors. Conclusions can and will be revised with further experimentation; science requires more than one study to reach solid conclusions. Decision procedures with clear-cut decisions based on one study only hamper science and lead to tunnel-vision.

Apart from these science-theoretical considerations of the NHST, there are also practical reasons why pure NHST should be avoided. In at least a number of research fields, the p-value has become more than just the criterion for finding an effect or not: it has become the criterion of whether the research is publishable or not. Editors and reviewers of scientific journals have increasingly interpreted a study with a significant effect to be more interesting than a study with a non-significant effect. For that reason, in scientific journals you will find mostly studies reported with a significant effect. This has led to *the file-drawer problem*: the literature reports significant effects for a particular phenomenon, but there can be many unpublished studies with non-significant effects for the same phenomenon. Theses unpublished studies remain unseen in file-drawers (or these days on hard-drives). So based on the literature there might seem to exist a particular phenomenon, but if you would put all the results together, including the unpublished studies, the effect might disappear completely.

Remember that if the null-hypothesis is true and everyone uses an α of 0.05, then out of 100 studies of the same phenomenon, only 5 studies will be significant and are likely to be published. The remaining 95 studies with insignificant effects are more likely to remain invisible.

As a result of this bias in publication, scientists who want to publish their results are tempted to fiddle around a bit more with their data in order to get a significant result. Or if they obtain a p-value of 0.07, they decide to increase their sample size, and perhaps stop as soon as the p-value is 0.05 or less. This horrible malpractice is called *p-hacking* and is extremely harmful to science. As we saw earlier, if you want to find an effect and not miss it, you should carry out a power analysis *before* you collect the data and make sure that your sample size is large enough to obtain the power you want to have. Increasing sample size *after* you have found a non-significant increases your type I error rate dramatically: if you stop collecting data *until* you find a significant p-value, the type I error rate is equal to 1!

There have been wide discussions the last few years about the use and interpretation of p-values. In a formal statement, the American Statistical Association published six principles that should be well understood by anyone, including you, who uses them.

The six principles are:

1. P-values can indicate how incompatible the data are with a specified statistical model (usually the null-hypothesis).

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. Instead, they measure how likely it is to find a sample slope of at least the size that you found, given that the population slope is 0.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold. For instance, also look at the size of the effect: is the slope large enough to make policy changes worth the effort? Have other studies found effects of similar sizes?

4. Proper inference requires full reporting and transparency. Always report your sample slope, the standard error, the *T*-statistic, the degrees of freedom, and the *p*-value. Only report about null-hypotheses that your study was designed to test.

5. A *p*-value or statistical significance *does not* measure the size of an effect or the importance of a result. (See principle 1)

6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis. At least as important is the design of the study.

These six principles are further explained in the statement online[3]. The bottom line is, *p*-values have worth but only when used and interpreted in a proper way, although some disagree. The philosopher of science William Rozeboom once called NHST surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students. The scientific journal *Basic and Applied Social Psychology* even banned NHST altogether: *T*-values and *p*-values are not allowed if you want to publish your research in that journal.

Most researchers now realize that reporting confidence intervals is often a lot more meaningful than reporting whether a *p*-value is significant or not. A *p*-value only says something about evidence against the hypothesis that the slope is 0. In contrast, a confidence interval gives a whole range of reasonable values for the population slope. If 0 lies within the confidence interval, then 0 is a reasonable value; if it is not, then 0 is not a reasonable value so that we can reject the null-hypothesis.

Using confidence intervals also counters one fundamental problem of null-hypotheses: nobody believes in them! Remember that the null-hypothesis states that a particular effect (a slope) is exactly 0: not 0.0000001, not -0.000201, but exactly 0.00000000000000000000.

Sometimes a null-hypothesis doesn't make sense at all. Suppose we are interested to know what the relationship is between age and height in children. Nobody believes that the effect of age on height is 0. Why then test this hypothesis? More interesting would be to know *how large* the population slope is. A confidence interval would then be much more informative than a simple rejection of the null-hypothesis.

[3]https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

In some cases, a null-hypothesis can be slightly more meaningful: suppose you are interested in the effect of cognitive behavioural therapy on depression. You hope that the number of therapy sessions has a negative effect on the severity of the depression, but of course it is entirely possible that the effect is very close to nonexisting. Of course you can only look at a sample of patients and determine the sample slope. But think now about the population slope: think about all patients in the world with depression that theoretically could partake in the research. Some of them have 0 sessions, some have 1 session, and so on. Now imagine that there are 1 million of such people. How likely is it that in the population, the slope for the regression is exactly 0? Not 0.00000001, not -0.0000000002, but exactly 0.0000000000. Of course, this is extremely unlikely. The really interesting question in such research is whether there is a *meaningful* effect of therapy. For instance, an effect of at least half a point decrease on the Hamilton depression scale for 5 sessions. Also in this case, a confidence interval for the effect of therapy on depression would be more helpful than a simple p-value. A confidence interval of -2.30 to -0.01 says that a small population effect of -0.01 might be there, but that an effect of -0.0001 or 0.0000 is rather unlikely. The p-value less than α only tells you only that a value of exactly 0.0000 is not realistic.

So, instead of asking research questions like "Is there a linear relationship between x and y?" you might ask: "How large is the linear effect of x on y?" Instead of a question like "Is there an effect of the intervention?" it might be more interesting to ask: "How large is the effect of the intervention?"

Summarizing, remember the following principles when doing your own research or evaluating the research done by others:

- Inference about a population slope or intercept can be made on the basis of sample data, but only in probabilistic terms. This means that a simple statement like "the value of the population slope is definitely not zero" cannot be made. Only statements like "A population slope of 0 is not very likely given the sample data" can be made.

- Science is cumulative. No study is definitive. Effects should be replicated by independent researchers.

- Always report your regression slope or intercept, with the standard error and the sample size. Based on these, the T-statistics can be computed with the degrees of freedom. Then if several other researchers have done the same type of research, the results can be combined in a so-called meta-analysis, so that a stronger statement about the population can be made, based on a larger total sample size. The standard error and sample size moreover allow for the construction of confidence intervals. But better is to report confidence intervals yourself.

4.7.1 Exercise

Why is it that the type I error rate becomes 1 if you keep increasing your sample size until the p-value is smaller than α?

4.8 Relationship between p-values and confidence intervals

In previous sections we stated that if the value 0 lies within a confidence interval, it is a reasonable value for the population slope. If 0 is not within the interval, 0 is not a reasonable value for the population slope, so we have to reject the null-hypothesis. Here we will elaborate a little on this theme.

Both the confidence interval and the p-value are based on the same t-distribution. Suppose we set our α to 0.05, and our sample size is 102. This means that if we find a p-value $p \leq 0.05$ we reject the null-hypothesis that the slope is 0. The p-value depends on how many standard deviations our sample slope deviates from 0. We calculate this by computing a standardized slope. For example, for a sample slope of 1 and a standard error of 0.5, our standardized slope is $T = (1-0)/0.5 = 2$. In other words, our sample slope of 1 is 2 standard errors away from 0. From t-tables, we know that with 100 degrees of freedom, the 2.5th and 97.5th percentiles are -1.98 and 1.98, respectively (see Table **??**). Therefore, the p-value depends on the size of the T-statistic. If it is equal to -1.98 or 1.98, the p-value is exactly 0.05. If the T-statistic is smaller than -1.98 or larger than 1.98, the p-value is smaller than 0.05.

The values -1.98 and 1.98 are also used for the construction of the 95% confidence interval. The lower bound lies at 1.98 times the standard error below the sample slope, and the upper bound lies at 1.98 times above the sample slope. Therefore, if 0 lies more than 1.98 standard erros away from the mean, it lies outside the confidence interval. But if 0 lies more than 1.98 standard erros away from the mean, this implies that the sample slope lies more than -1.98 standard erros away from 0, which corresponds to a T-statistic of more than ± 1.98. Thus, if 0 is not within the 95% confidence interval, we know that the p-value is smaller than 0.05.

Using the same reasoning as above, we also know that if 0 is not within the 99% confidence interval, we know that the p-value is smaller than 0.01, and if 0 is not within the 99.9% confidence interval, we know that the p-value is smaller than 0.001, etcetera.

A 95% confidence interval can therefore also be seen as the range of possible values for the null-hypothesis that cannot be rejected with an α of 5%. By the same token, a 99% confidence interval can be seen as the range of possible values for the null-hypothesis that cannot be rejected with an α of 1%, etcetera.

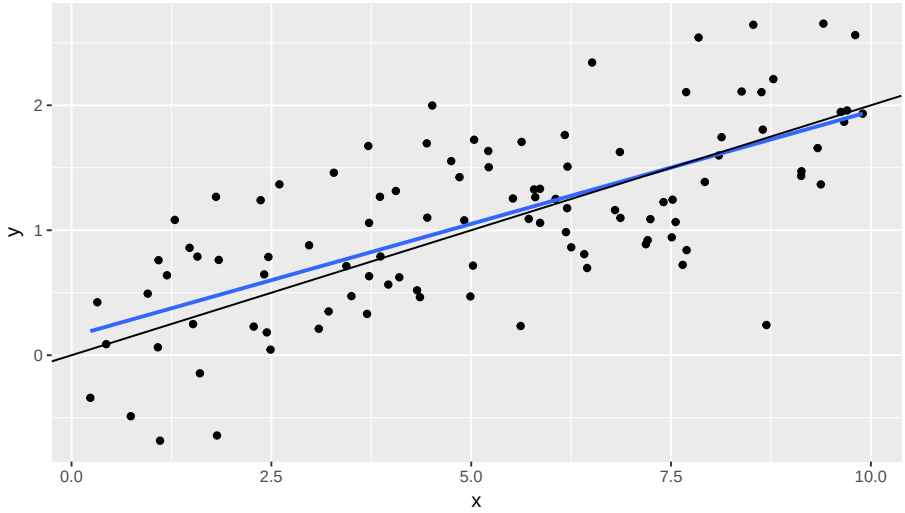


Figure 4.13: Regression of y on x, with the population line in black and the sample line in blue. The grey area is the 95 percent confidence interval for the population line based on the sample line.

4.9 Inference using SPSS

Figure 4.13 shows an example of a regression analysis on 102 datapoints. The dependent variable is y and the independent variable is x. The black line represents the linear equation in the population, whereas the blue line represents the sample equation:

$$Population: \quad y = 0 + 0.2 \times x + \epsilon \tag{4.7}$$

$$Sample: \quad y = 0.1500534 + 0.180198 + e \tag{4.8}$$

The syntax that we can use for these sampe data is

```
UNIANOVA y WITH x
  /DESIGN=x
  /PRINT=PARAMETER
  /CRITERIA=ALPHA(.01).
```

Note that we have set the significance level α to 0.01 with the statement `CRITERIA=ALPHA(0.01)`. Figure 4.14 shows the SPSS output. Look at the Parameter Estimates table. It shows the intercept, with a standard error of 0.1107062. The t-value in the output is the T-statistic for the null-hypothesis, and is equal to $(B - 0)/SE = 0.15/0.111 = 1.355$. We had 102 data points, so

the degrees of freedom is equal to $102 - 2 = 100$. [4] From online tables it is known that with 100 degrees of freedom, 0.089 of t-values are larger than 1.355 and 0.089 of t-values are smaller than -1.355. SPSS knows this and automatically calculates the two-sided p-value. Therefore, the two-sided p-value for a T-statistic of 1.355 with 100 degrees of freedom is equal to $2 \times 0.089 = 0.178$. Since $p > 0.01$, we cannot reject the null-hypothesis that the intercept in the population data is equal to 0. SPSS also shows the 99% confidence interval that runs from -0.141 to 0.441. All these values in this interval are reasonable values for the population intercept.

Let's now turn to the output for the effect of x. The table shows a slope of 0.18 with a standard error of 0.019. Therefore, the T-value for the null-hypothesis equals $(0.18 - 0)/0.02 = 9.515$. With 100 degrees of freedom, a proportion of 0 of the t-values is larger than 9.515 and 0 of the t-values is smaller than -9.515. Therefore, the associated p-value is 0. Since $p \leq 0.01$, we reject the null-hypothesis that the population slope is 0. The 99% confidence interval for the population slope is from 0.13 to 0.23.

Tests of Between-Subjects Effects

Dependent Variable: y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	23.877[a]	1	23.877	90.526	.000
Intercept	.485	1	.485	1.837	.178
x	23.877	1	23.877	90.526	.000
Error	26.376	100	.264		
Total	170.479	102			
Corrected Total	50.252	101			

a. R Squared = .475 (Adjusted R Squared = .470)

Parameter Estimates

Dependent Variable: y

Parameter	B	Std. Error	t	Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	.150	.111	1.355	.178	-.141	.441
x	.180	.019	9.515	.000	.130	.230

Figure 4.14: Output for a simple regression analysis.

[4] Note that this is not shown in the Parameter Estimates table, but in the Tests of Between-Subjects Effects table in the row for Error (error is another word for residual). In that row we see the error degrees of freedom (df) of 100.

4.9.1 Exercises

1. Suppose you want to predict the personality trait aggressiveness on the basis of yearly income in Euros. The variable that measures aggressiveness is **aggr** and the variable that measures income is **yearincome**. Give the linear equation for the relationship between these two variables in the population.

2. You want to know whether there is a relationship between income and aggressiveness. State the null-hypothesis in terms of the linear equation you gave.

3. Suppose you want to test this null-hypothesis using a type I error rate of 0.05. Provide the SPSS syntax that is needed to perform this test.

4. Suppose someone else has done the analysis for you using a different software package and gives you the output in Figure 4.14. See if you can find the 95% confidence interval for the effect of yearly income on aggressiveness.

```
##
## Call:
## lm(formula = aggr ~ yearincome)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8470 -1.2149 -0.1420  0.8981  4.0501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.13817    0.32236  -0.429    0.669
## yearincome  -0.15573    0.06534  -2.383    0.019 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 on 100 degrees of freedom
## Multiple R-squared:  0.05375,Adjusted R-squared:  0.04429
## F-statistic:  5.68 on 1 and 100 DF,  p-value: 0.01905
##              2.5 % 97.5 %
## (Intercept) -0.778  0.501
##             2.5 % 97.5 %
## yearincome -0.285 -0.026
```

5. On the basis of this confidence interval, is there a significant effect of income on aggressiveness? Explain your answer.

6. On the basis of the T-statistic and its p-value, is there a significant effect of income on aggressiveness? Explain your answer.

Answers:

1.
$$aggr = \beta_0 + \beta_1 \times yearincome + \epsilon \tag{4.9}$$

2.
$$H_0 : \beta_1 = 0 \tag{4.10}$$

3.
```
UNIANOVA aggr WITH yearincome
  /DESIGN=yearincome
  /PRINT=PARAMETER
  /CRITERIA=ALPHA(.05).
```

4. -0.285, -0.026

5. The 95% confidence interval for the slope does not contain 0, so we can reject the null-hypothesis. We therefore call the effect of income on aggressiveness significant.

6. The p-value associated with the t-value for the slope is smaller than 0.05. Therefore, we can reject the null-hypothesis. We call the effect of income on aggressiveness significant.

Chapter 5

Multivariate regression

5.1 Explained and unexplained variance

In the previous chapter we have seen relationships between two variables: one dependent variable and one independent variable. The dependent variable we usually denote as y, and the indepedent variable we denote by x. The relationship was modelled by a linear equation: an equation with an intercept b_0 and a slope parameter b_1:

$$y = b_0 + b_1 x \tag{5.1}$$

Further, we argued that in most cases, the relationship between x and y cannot be completely described by a straight line. Not all of the variation in y can be explained by the variation in x. Therefore, we have *residuals* e: the difference between the y-values that are predicted by the straight line, (denoted by $\hat{y}$), and the observed y-value:

$$e = \hat{y} - y \tag{5.2}$$

Therefore, the relationship between x and y is denoted by a regression equation, where the relationship is approached by a linear equation, plus a residual part e:

$$y = b_0 + b_1 x + e \tag{5.3}$$

The linear equation only gives us only the expected y-value, $\hat{y}$:

$$\hat{y} = b_0 + b_1 x \tag{5.4}$$

We've also seen that the residual e is assumed to have a normal distribution, with mean 0 and variance σ^2:

$$e \sim N(0, \sigma^2) \tag{5.5}$$

Remember that linear models are used to explain (or predict) the variation in y: why are there both high values of y and some low values? Where does the variance in y come from? Well, the linear model tells us that the variation is in part explained by the variation in x. If b_1 is positive, we predict a relatively high value for y for a high value of x, and we predict a relatively low value for y if we have a low value for x. If b_1 is negative, it is of course in the opposite direction. Thus, the variance in y is in part explained by the variance in x, and the rest of the variance can only be explained by the residuals e.

$$Var(y) = Var(\hat{y}) + Var(e) = Var(b_0 + b_1 x) + \sigma^2 \tag{5.6}$$

Because the residuals do not explain anything (we don't know where these residuals come from), we say that the *explained* variance of y is only that part of the variance that is explained by independent variable x: $Var(b_0 + b_1 x)$. The *unexplained* variance of y is the variance of the residuals, σ^2. The explained variance is often denoted by a ratio: the explained variance divided by the total variance of y:

$$Var_{explained} = \frac{Var(b_0 + b_1 x)}{Var(y)} = \frac{Var(b_0 + b_1 x)}{Var(b_0 + b_1 x) + \sigma^2} \tag{5.7}$$

From this equation we see that if the variance of the residuals is large, then the explained variance is small. If the variance of the residuals is small, the variance explained is large.

5.2 More than one predictor

In regression analysis, and in linear models in general, we try to make the explained variance as large as possible. In other words, we try to minimize the residual variance, σ^2.

One way to do that is to use a second independent variable. If not all of the variance in y is explained by x, then why not try an extra independent variable?

Let's use an example with data on the weight of books, the size of books (area), and the volume of books. Let's try first to predict the weight of a book, *weight*, on the basis of the volume of the book, *volume*. Suppose we find the following regression equation and a value for σ^2:

$$weight = 107.7 + 0.71 \times volume + e \tag{5.8}$$

$$e \sim N(0, 15362) \tag{5.9}$$

In the data set, we see that the variance of the weight, $Var(weight)$ is equal to 72274. Since we also know the variance of the residuals, we can solve for the variance explained by **volume**:

$$Var(weight) = 72274 = Var(107.7 + 0.7 \times volume) + 15362$$
$$Var(107.7 + 0.7 \times volume) = 72274 - 15362 = 56912$$

So the proportion of explained variance is equal to $\frac{56912}{72274} = 0.7874478$. This is quite a high proportion: nearly all of the variation in the number of houses per city is explained by how many inhabitants a city has.

But let's see if we can explain even more variance if we add an extra independent variable. Suppose we know the area of each book. We expect that books with a large area weigh more. Our linear equation might look like this:

$$weight = 22.4 + 0.71 \times volume + 0.5 \times area + e \tag{5.10}$$

$$e \sim N(0, 6031) \tag{5.11}$$

How much of the variance in weight does this equation explain? The proportion of explained variance is equal to $\frac{66243}{72274} = 0.9165537$. So the proportion of explained variance has increased!

Note that the variance of the residuals has decreased; this is the main reason why the proportion of explained variance has increased. By adding the extra independent variable, we can explain some of the variance that without this variable could not be explained! In summary, by adding independent variables to a regression equation, we can explain more of the variance of the dependent variable. A regression analysis with more than one independent variable we call *multiple regression*. Regression with only one indendent variable is often called *simple regression*.

5.3 R-squared

With regression analysis, we try to explain variance of the dependent variable. With multiple regression, we use more than one independent variable to try to explain this variance. In regression analysis, we use the term R-squared to refer to the proportion of explained variance, usually with the symbol R^2. The unexplained variance is of course the variance of the residuals, $Var(e)$, usually denoted as σ_e^2. So suppose the variance of dependent variable y equals 100, and the residual variance in a regression equation equals say 80, then R^2 or the proportion of explained variance is $(100 - 80)/100 = 0.20$.

$$R^2 = \sigma^2_{explained}/\sigma^2_y = (1 - \sigma^2_{unexplained})/\sigma^2_y = (1 - \sigma^2_e)/\sigma^2_y \tag{5.12}$$

This is the defintion of R-squared at the population level, where we know the exact values of the variances. However, regression analysis is most often based on a random sample of the population, and we don't know the values exactly, we can only try to estimate them.

For σ^2_y we take as an adjusted estimate the variance of y in our sample data, Var(y), which is calculated by

$$\widehat{\sigma^2_y} = \frac{\Sigma(y - \bar{y})^2}{n - 1} \tag{5.13}$$

where n is sample size. We divide by $n-1$ and not by n, because we want to estimate the variance of y in the population data.

For σ_e^2 we take as an adjusted estimate the variance of the residuals e in our sample data, Var(e), which is calculated by

$$\widehat{\sigma_e^2} = \frac{\Sigma e^2}{n-1} \tag{5.14}$$

Here we do not have to subtract the mean of the residuals, because this is 0 by definition.

So our estimate for R^2 in the population is then

$$\begin{aligned} \widehat{R^2} &= \frac{\frac{\Sigma(y-\bar{y})^2}{n-1} - \frac{\Sigma e^2}{n-1}}{\frac{\Sigma(y-\bar{y})^2}{n-1}} \\ &= \frac{\Sigma(y-\bar{y})^2 - \Sigma e^2}{\Sigma(y-\bar{y})^2} = 1 - \frac{SSE}{SST} \end{aligned} \tag{5.15}$$

where SST refers to the total sum of squares.

As we saw previously, in a regression analysis, the intercept and slope parameters are found by minimizing the sum of squares of the residuals, SSE. Since the variance of the residuals is based on this sum of squares, in any regression analysis, the variance of the residuals is always as small as possible. The values of the parameters for which the SSE (and by consequence the variance) is smallest, are the least squares regression parameters. And if the variance of the residuals is always minimized in a regression analysis, the explained variance is always maximized!

Because in any least squares regression analysis based on a sample of data, the explained variance is always maximized, we may overestimate the variance explained in the population data. Therefore very often in regression analysis we use an *adjusted R-squared* that takes this possible overestimation (*inflation*) into account. The adjustment is based on the number of independent variables and sample size.

The formula is

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n is sample size and p is the number of independent variables. For example, if R^2 equals 0.10 and we have a sample size of 100 and 2 independent variables, the adjusted R^2 is equal to $1 - (1 - 0.10)\frac{100-1}{100-2-1} = 1 - (0.90)\frac{99}{97} = 0.08$. Thus the estimated proportion of variance explained at population level equals 0.08. Remember that the adjusted R-squared is *never larger* than the unadjusted R-squared.

5.4 Multicollinearity

In general, if you add independent variables to a regression equation, the proportion explained variance, R^2, increases. Suppose you have the following three regression equations:

$$weight = b_0 + b_1 \times volume + e \tag{5.16}$$

$$weight = b_0 + b_1 \times area + e \tag{5.17}$$

$$weight = b_0 + b_1 \times volume + b_1 \times area + e \tag{5.18}$$

If we carry out these three analyses, we obtain an R^2 of 0.8026346 if we only use **volume** as predictor, and an R^2 of 0.1268163 if we only use **area** as predictor. So perhaps you'd think that if we take both **volume** and **area** as predictors in the model, we would get an R^2 of $0.8026346 + 0.1268163 = 0.9294509$. However, if we carry out the multiple regression with **volume** and **area**, we obtain an R^2 of 0.9284738, which is slightly less! This is not a rounding error, but the result of the fact that there is a correlation between the volume of a book and the area of a book. Here it is a tiny correlation of $round(cor(allbacks\text{area}, \text{allbacks}volume), 3)$, but nevertheless it affects the proportion of variance explained when you use both these variables.

Let's look at what happens when indendent variables are strongly correlated. Table 5.1 shows measurements on a breed of seals (only measurements on the first 6 seals are shown). Often, the age of an animal is gaged from its weight: we assume that heavier seals are older than lighter seals. If we carry out a simple regression analysis, we get the following equation:

Table 5.1: Part of Cape Fur Seal Data.

age	weight	heart
33.00	27.50	127.70
10.00	24.30	93.20
10.00	22.00	84.50
10.00	18.50	85.40
12.00	28.00	182.00
18.00	23.80	130.00

Table 5.2: Regression table for predicting age from height.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.4419	4.6974	2.44	0.0215
weight	0.8169	0.0716	11.41	0.0000

$$age = 11.4 + 0.82 \times weight + e \tag{5.19}$$

$$e \sim N(0, 200) \tag{5.20}$$

USE regression table instead of formula

From the data we calculate the variance of age, and we find that it is 1090.8551724. The variance of the residuals is 200, so that the proportion of explained variance is $(1090.8551724 - 200)/1090.8551724 = 0.8166576$.

Since we also have data on the weight of the heart alone, we could try to predict the age from the weight of the heart. Then we get:

Table 5.3: Regression table for predicting age from heart

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.4419	4.6974	2.44	0.0215
weight	0.8169	0.0716	11.41	0.0000

$$age = 20.6 + 0.11 \times heart + e \tag{5.21}$$

$$e \sim N(0, 307) \tag{5.22}$$

USE regression table instead of formula

Here the variance of the residuals is 307, so the proportion of explained variance is $(1090.8551724 - 370)/1090.8551724 = 0.6608166$.

Now let's see what happens if we include both total weight and weight of the heart into the linear model. This results in the following model equation:

$$age = 10.3 + 0.99 \times weight - 0.03 \times heart + e \tag{5.23}$$

$$e \sim N(0, 204) \tag{5.24}$$

Table 5.4: Regression table for predicting age from heart and weight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.4419	4.6974	2.44	0.0215
weight	0.8169	0.0716	11.41	0.0000

USE regression table instead of formula

Here we see that the regression parameter for total weight has increased from 0.82 to 0.99. At the same time, the regression parameter for the weight of the heart has decreased, has even become negative, from 0.11 to -0.03. From this equation we see that there is a strong relationship between the total weight and the age of a seal, but on top of that, for every unit increase in the weight of the heart, there is a very small decrease in the expected age. In fact, we find that the effect of **heart** is no longer significant, so we could say that on top of the effect of total weight, there is no remaining relationship between the weight of the heart and age. In other words, once we can use the total weight of a seal, there is no more information coming from the weight of the heart.

This is because the total weight of a seal and the weight of its heart are strongly correlated: heavy seals have generally heavy hearts. Here the correlation turns out to be 0.9587873, almost perfect! If you know the weight of seal, you practically know the weight of the heart. This is logical of course, since the total weight is a composite of all the weights of all the parts of the animal: the total weight variable *includes* the weight of the heart.

Here we have seen, that if we use multiple regression, we should be aware of how strongly the independent variables are correlated. Heavily correlated predictor variables do not add extra predictive power. Worse: they can cause problems in estimating regression parameters because it becomes hard to tell which variable is more important: if they are strongly correlated (positive or negative), than they measure almost the same thing!

When two predictor variabels are perfectly correlated, either 1 or -1, estimation is no longer possible, the software stops and you get a warning. We call such a situation *multiple collinearity*. But also if the correlation is close to 1 or -1, you should be very careful interpeting the regression parameters. You will then see there are very wide confidence intervals (very large standard errors). If this happens, try to find out what variables are highly correlated, and select the variable that makes most sense.

In our seal data, there is a very high correlation between the variables **heart** and **weight** that results in estimation problems and very large standard errors (wide confidence intervals), so a lot of uncertaintly. The standard errors were about 3 times as large with the multiple regression than with simple regressions. It makes therefore more sense to use only the total weight variable, since when seals get older, *all* their organs and limbs get larger, not just their heart.

5.5 Multiple regression and inference

In an earlier chapter on inference, we saw that if we want to say something about the population slope on the basis of the sample slope, we can use t-distribtutions. The sahpe of the t-distribtution depends on the degrees freedom and we saw that these depend on sample size. For simple regression (one intercept and one slope), we saw that the number of degrees of freedom, the residual degrees of freedom, was equal to sample size minus 2 ($n - 2$).

In the more general case of multiple regression, with the number of independent variables equal to K and including an intercept, the degrees of freedom for the t-distribution of sample slopes is equal to $n - K - 1$. One could also say, the degrees of freedom is equal to sample size minus the number of parameters (coefficients) in your model.

For example, suppose you have 200 data points and 4 independent variables. Then you have 4 slope parameters and 1 intercept parameter in your model, so 5 parameters in total. The (residual) degrees of freedom is in that case $n - 5 = 195$.

5.6 Multiple regression in SPSS

Let's use the book data and run the multiple regression in SPSS. The syntax looks very similar to simple regression, except that we now specify two independent variables, volume and area, instead of one.

```
UNIANOVA weight WITH volume area
  /DESIGN = volume area
  /PRINT = PARAMETER R-Squared.
```

Tests of Between-Subjects Effects

Dependent Variable: weight

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	939460.71 [a]	2	469730.354	77.885	.000
Intercept	888.274	1	888.274	.147	.708
volume	811143.719	1	811143.719	134.495	.000
area	127328.290	1	127328.290	21.112	.001
Error	72372.626	12	6031.052		
Total	8502500.00	15			
Corrected Total	1011833.33	14			

a. R Squared = .928 (Adjusted R Squared = .917)

Parameter Estimates

Dependent Variable: weight

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	22.413	58.402	.384	.708	-104.835	149.661
volume	.708	.061	11.597	.000	.575	.841
area	.468	.102	4.595	.001	.246	.691

Figure 5.1: SPSS output of a linear model (multiple regression) for predicting the weight of books.

Figure 5.1 shows the output. There we see an intercept, a slope parameter for volume and a slope parameter for area. These numbers tell us that the expected or predicted weight of a book that has a volume of 0 and an area of 0 is 22.413. For every unit increase in volume, the predicted weight increases by 0.708, and for every unit increase in area, the predicted weight increases by 0.468.

So the linear model looks like:

$$weight = 22.413 + 0.708 \times volume + 0.468 \times area + e \tag{5.25}$$

Thus, the predicted weight of a book that has a volume of 10 and an area of 5, the expected weight is equal to $22.413 + 0.708 \times 10 + 0.468 \times 5 = 31.833$.

In the output, there is also another table, and there we see the R-squared and the Adjusted R-squared. In Figure 5.1 we see that the R squared is equal to 0.928. As seen earlier, this value can be computed from the sums of squares: $(SST - SSE)/SST$. From the table we see that the SST is 8502500 (corrected total sum of squares)[1], and the SSE is 72372.626. If we do the math, we see that we get $(1011833 - 72372.626)/1011833 = 0.928$.

5.7 Simpson's paradox

With muliple regression, you may uncover very surprising relationships between two variables, that can never be found using simple regression. Here's an example from Paul van der Laken[2], who simulated a data set on the topic of Human Resources (HR).

Assume you run a company of 1000 employees and you have asked all of them to fill out a Big Five personality survey. Per individual, you therefore have a score depicting his/her personality characteristic Neuroticism, which can run from 0 (not at all neurotic) to 7 (very neurotic). Now you are interested in the extent to which this **Neuroticism** of employees relates to their **salary** (measured in Euros per year).

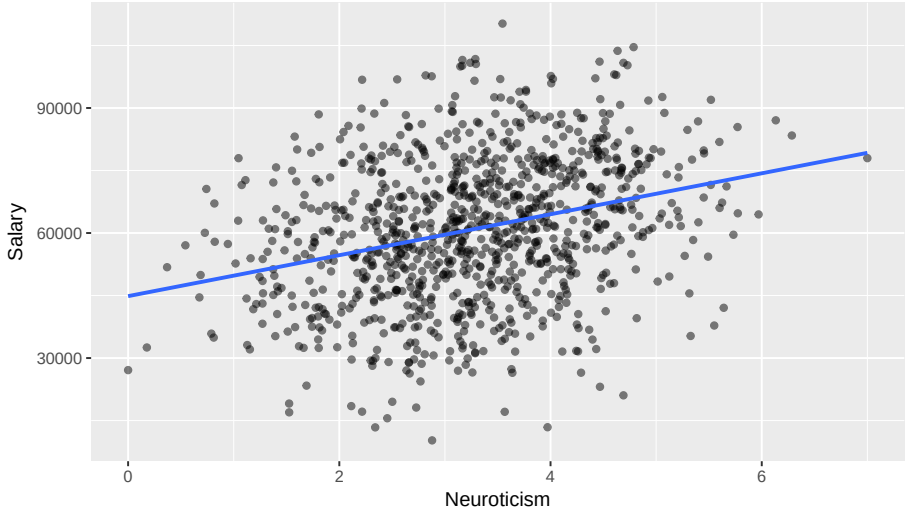
We carry out a simple regression, with salary as our dependent variable and Neuroticism as our independent variable. We then find the following regression equation:

$$salary = 44857 + 4912 \times Neuroticism + e \tag{5.26}$$

Figure **??** shows the data and the regression line. From this visualizations it would look like Neuroticism relates significantly and *positively* to their yearly salary: the more neurotic people earn more salary than less neurotic people.

[1] In SPSS, the total sum of squares reports the sum of the squared deviations from 0, whereas the *corrected* total sum of squares reports the squared deviations from the mean of the dependent variable, $\bar{y}$

[2] https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/



Now we run a multiple regression analysis. We assume that one very important cause of how much people earn is their educational background. If we include both Education and Neuroticism as independent variables and run the analysis, we obtain the following regression equation:

$$salary = 50249 - 3176 \times Neuroticism + 20979 \times Education + e \qquad (5.27)$$

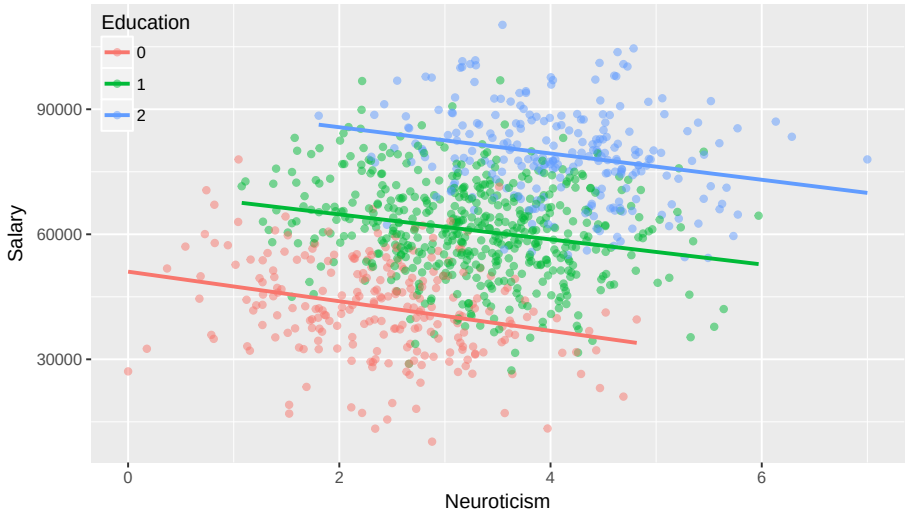
Note that we now find a *negative* slope parameter for the effect of Neuroticism! This implies there is a relationship in the data where neurotic employees earn *less* than their less neurotic colleagues! How can we reconcile this seeming paradox: which result should we trust: the one from the simple regression, or the one from the multiple regression?

The answer is: neither. Or perhaps: both! Both analyses give us different information.

Let's look at the last equation more closely. Suppose we make a prediction for a person with a low educational background (Education=0). Then the equation tells us that the expected salary of a person with neuroticism score of 0 is around 50249, and of a person with a neuroticism score of 7 is around 28019. So for employees with low education, the more neurotic employees earn less! If we do the same exercise for average ecudation and high education employees, we find exactly the same pattern: for each unit increase in neuroticism, the yearly salary drops by 3176 Euros.

It is true that in this company, the more neurotic persons generally earn a higher salary. But if we take into account educational background, the relationship flips around. This can be seen from Figure **??**: looking only at the people with a low educational background (Education=0), then the more neurotic people earn less than they less neurotic colleagues with a similar educational background. And the same is true for people with an average education (Education=1) and a high education (Education=3). Only when you put all employees

together in one group, you see a positive relationship between Neuroticism and salary.



Simpson's paradox tells us that we should always be careful when interpreting positive and negative correlations between two variables: what might be true at the total group level, might not be true at the level of smaller subgroups. Multiple linear regression helps us investigate correlations more deeply and uncover exciting relationships between multiple variables.

5.8 Exercises

Two neighbours, Elsa and John, are chopping trees in the forest for their respective fireplaces. They pick their trees to chop down, based on the expected volume of wood they can get from that tree. However, Elsa and John disagree on what is the most important aspect of trees for selection. Elsa believes that the tallest tree will give the biggest volume of wood for the fireplace, but John believes that the tree with the largest girth gives the most volume of wood. Luckily there is a data set with three variables: Volume, Girth and Height.

1. What would the SPSS syntax look like to run a multiple regression, if you want to find out which predictor is most important for the volume of wood that comes from a tree?

```
UNIANOVA ..... WITH ........
  /DESIGN = ........
  /PRINT = PARAMETER R-Squared.
```

2. Suppose you find the output in Table 5.5: what would your linear equation look like?

$$\ldots\ldots = \ldots\ldots\ldots\ldots\ldots + e \tag{5.28}$$

Table 5.5: Regression table for predicting volume from height and girth.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.9877	8.6382	-6.71	0.0000
Girth	4.7082	0.2643	17.82	0.0000
Height	0.3393	0.1302	2.61	0.0145

3. On the basis of the output, what would be the predicted volume for a tree with a height of 10 and a girth of 5?

4. On the basis of the output, what would be the predicted volume for a tree with a height of 5 and a girth of 10?

5. For each unit increase of height, how much does the volume increase? Give the approximate 95% confidence interval for this increase.

6. For each unit increase of girth, how much does the volume increase? Give the approximate 95% confidence interval for this increase.

7. On the basis of the SPSS output, do you think Lisa is right in saying that height is an important predictor of volume? Explain your answer.

8. On the basis of the SPSS output, do you think John is right in saying that girth is an important predictor of volume? Explain your answer.

9. On the basis of the plots in Figures 5.2 and 5.3, which do you think is the most reliable predictor for Volume: Height or Girth? Explain your answer.

10. How large is the proportion of variance explained in volume, by girth and height?

11. How would you summarize this multiple regression analysis in a research report?

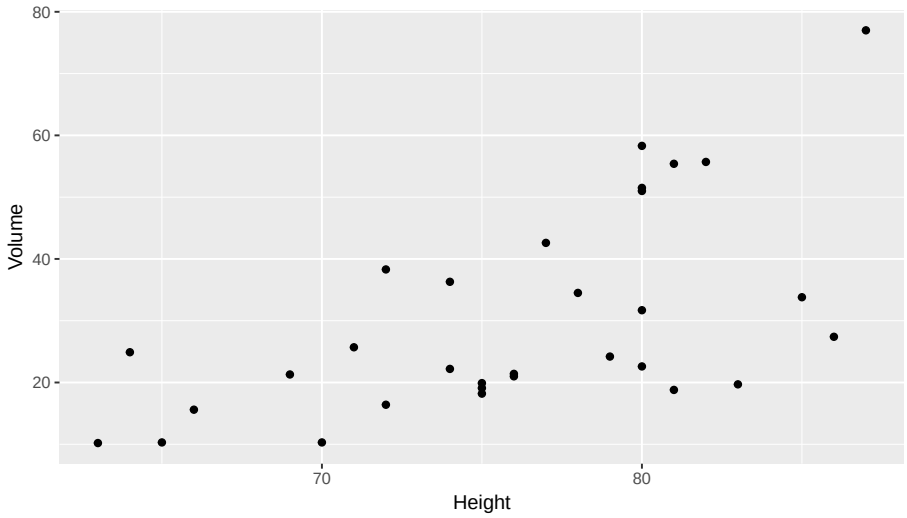


Figure 5.2: A scatterplot for the relationship between height and volume of a tree.

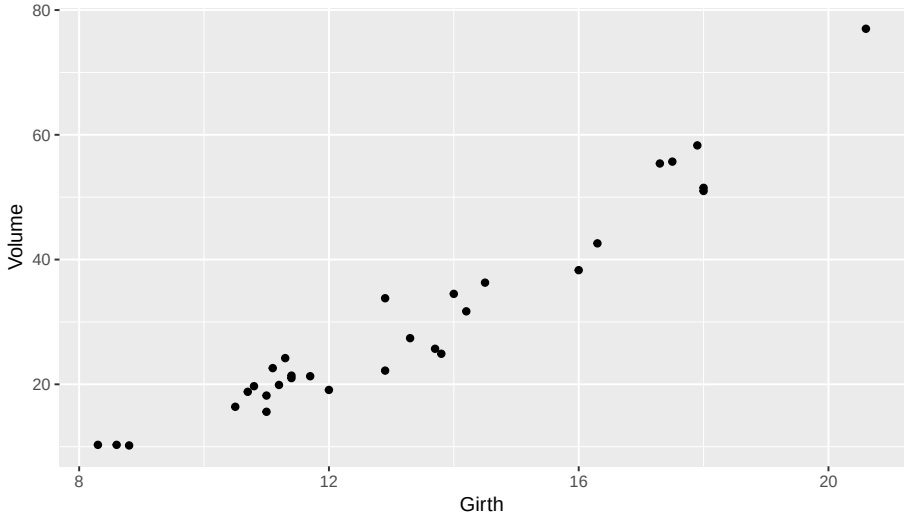


Figure 5.3: A scatterplot for the relationship between girth and volume of a tree.

Chapter 6

Categorical predictor variables

6.1 Dummy coding

As we have seen in Chapter 1, there are largely two different types of variables: quantitative variables and qualitative variables. Quantitative variables say something about how much of an attribute is in an object: for instance height (measured by inches) or heat (measured in degrees Kelvin). Qualitative variables say something about the quality of an attribute: for instance colour (red, green, yellow) or placement (aisle seat, window seat).

In the chapters on simple and multiple regression we have seen that both the dependent and the independent variables were all quantitative. The linear model used in regression analysis always involves a quantitiative dependent variable. However, in such analyses it is possible to use qualitative independent variables. In this chapter we explain how to do that and how to interpret the results.

The basic trick that we need is *dummy coding*. Dummy coding involves making one or more new variables, that reflect the different categories of a qualitative variable. First we focus on qualitative variables with only two variables. Later in this chapter, we will explain what to do with qualitative variables with more than two categories.

Suppose we have two different placements in a bus: aisle seats and window seats. Suppose we ask 5 people who have travelled from Amsterdam to Paris during the last 12 months, whether they had an aisle seat or a window seat, and how much they payed for the trip. Suppose we have the variables, person, seat and price. Table 6.1 shows part of the anonymized data.

With dummy coding, we make a new variable that only has values 0 and 1, and that conveys the same information as the **seat** variable. The resulting variable is called a dummy variable. Let's call this dummy variable **window** and give it the value 1 for all persons that travelled in a window seat. We give the value 0 for all persons that travelled in an aisle seat. We can also call

Table 6.1: Bus trips to Paris.

person	seat	price
001	aisle	57.00
002	aisle	59.00
003	window	68.00
004	window	60.00
005	aisle	61.00

the new variable **window** a *boolean variable* with TRUE and FALSE, since in computer science, TRUE is coded by a 1 and FALSE by a 0. Another name that is sometimes used is an *indicator variable*. Whatever you want to call it, the data matrix including the new variable is displayed in Table 6.2.

Table 6.2: Bus trips to Paris.

person	seat	window	price
001	aisle	0.00	57.00
002	aisle	0.00	59.00
003	window	1.00	68.00
004	window	1.00	60.00
005	aisle	0.00	61.00

What we have done now is coding the old qualitative variable **seat** into a quantitative variable **window** with values 0 and 1. Let's see what happens if we use a linear model for the variables price (dependent variable) and aisle (independent variable). The linear model is:

$$\begin{aligned} price &= b_0 + b_1 window + e && (6.1) \\ e &\sim N(0, \sigma_e^2) && (6.2) \end{aligned}$$

Let's use the bus trip data and determine the least squares regression line. We then find the following linear equation:

$$\widehat{price} = 59 + 5window \tag{6.3}$$

If the variable **window** is 1, then the expected or predicted price of the bus ticket is, according to this equation, $59 + 5 \times 1 = 64$. What does this mean? Well, all tickets from persons with a window seat were coded as a 1. Therefore the expected price of a window seat equals 64. By the same token, the expected price of an aisle seat (window = 0) is $59 + 5 \times 0 = 59$.

You see that by coding a qualitative variable into a quantitative dummy variable, we can describe the linear relationship between the type of seat and the price of the ticket. Figure 6.1 shows the relationship between the quantitative variables window and price.

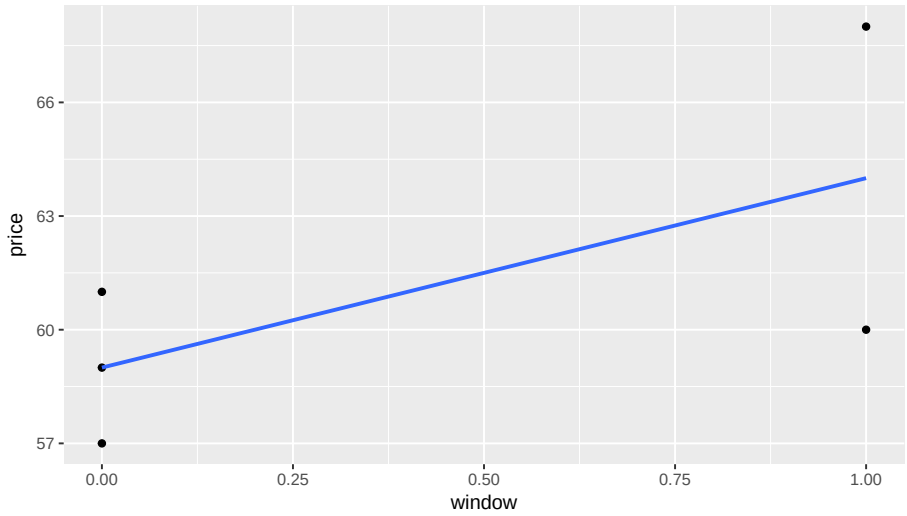


Figure 6.1: Relation between dummy variable window and price.

Note that the blue regression line goes straight through the mean of the prices for window seats (window=1) and the mean of the prices for aisle seats (window=0). In other words, using a dummy variable actually models the group means of window and aisle seats.

Figure 6.2 shows the same regression line but now for the original variable **seat**.

6.2 Using regression to describe group means

In the previous section we saw that if we replace a qualitative variable with a quantitative dummy variable with values 0 and 1, we can use a linear model to describe the relationship between a qualitative independent variable and a quantitative dependent variable. We also saw that if we take the least squares regression line, this line goes straight through the averages, the group means. The line goes straight through the group means because then the sum of the squared residuals is then at its smallest value. Let's look at the bus trip data again and compute the residuals and the squared residuals, see Table 6.3.

If we take the sum of the squared residuals we obtain 40. Now if we use a slightly different slope, so that we no longer go straight through the average prices for aisle and window seats (see Figure 6.3) and we compute the predicted values, the residuals and the squared residuals (see Table 6.4), we obtain a higher sum: 40.05.

Only the least squares regression line goes through the average seat prices of aisle and window seats. Thus, we can use the least squares regression equation to describe group means for dummy-coded qualitative variables. Now you also

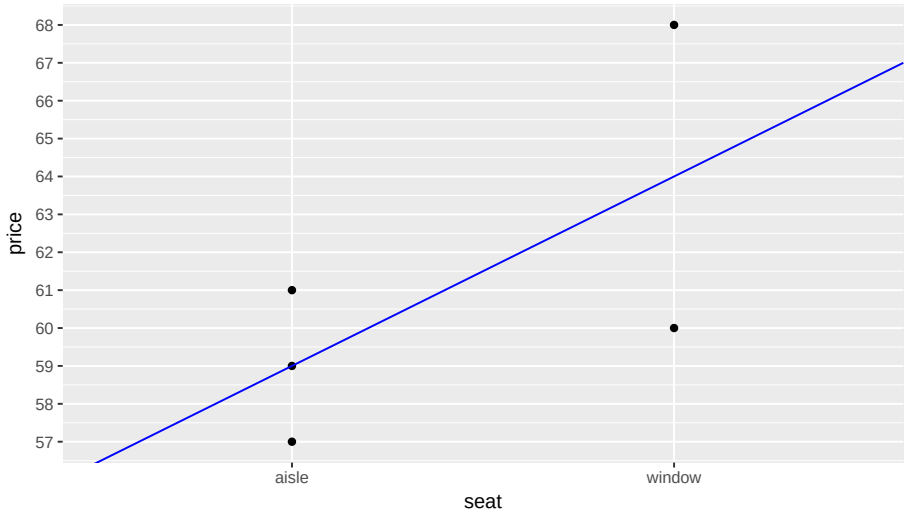


Figure 6.2: Relation between type of seat and price.

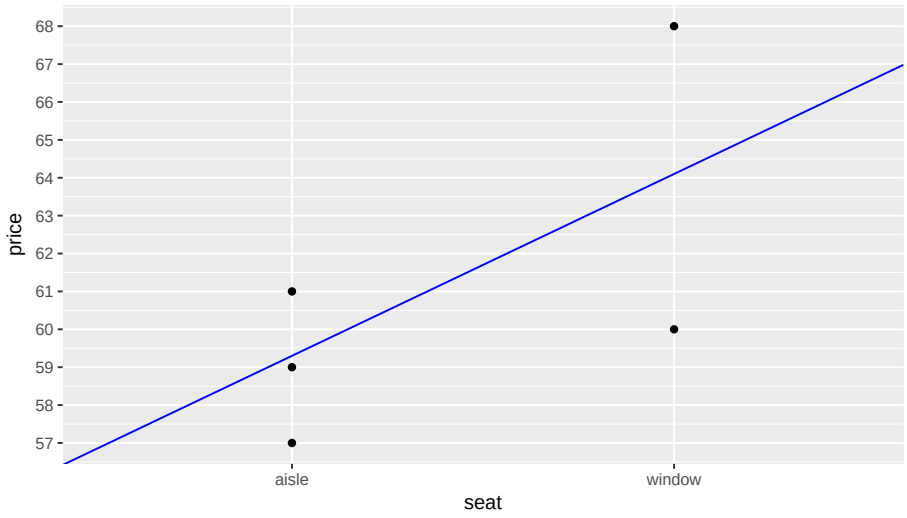


Figure 6.3: Relation between type of seat and price, with the regression line being not quite the least squares .

Table 6.3: Bus trips to Paris.

person	seat	window	price	e	e2
001	aisle	0.00	57.00	-2.00	4.00
002	aisle	0.00	59.00	0.00	0.00
003	window	1.00	68.00	4.00	16.00
004	window	1.00	60.00	-4.00	16.00
005	aisle	0.00	61.00	2.00	4.00

Table 6.4: Bus trips to Paris.

person	seat	window	price	wrongpredict	e	e2
001	aisle	0.00	57.00	59.10	-2.10	4.41
002	aisle	0.00	59.00	59.10	-0.10	0.01
003	window	1.00	68.00	63.90	4.10	16.81
004	window	1.00	60.00	63.90	-3.90	15.21
005	aisle	0.00	61.00	59.10	1.90	3.61

know that when you know the group means, it is very easy to draw the regression line.

Let's look at another data example. Results obtained from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment condition. Let's plot the data first, where we only compare the two experimental conditions.

With treatment 1, the average yield turns out to be 4.661, and with treatment 2, the average yield is 5.526. Suppose we make a new dummy variable treatment that is 0 for treatment 1 and 1 for treatment 2. Then we have the linear equation:

$$\widehat{weight} = b_0 + b_1 \times treatment \tag{6.4}$$

If we fill in the dummy variables and the expected weights (the means!), then we have the linear equations:

$$4.661 = b_0 + b_1 \times 0 = b_0 \tag{6.5}$$

$$5.526 = b_0 + b_1 \times 1 = b_0 + b_1 \tag{6.6}$$

So from this, we know that $b_0 = 4.661$, and if we fill that in for the second equation above, we get $b_1 = 5.526 - b_0 = 5.526 - 4.661 = 0.865$.

Thus, we get the linear equation

$$\widehat{weight} = 4.661 + 0.865 \times treatment \tag{6.7}$$

Since this regression line goes straight through the average yield for each treatments, we know that this is the least square regression equation. Therefore

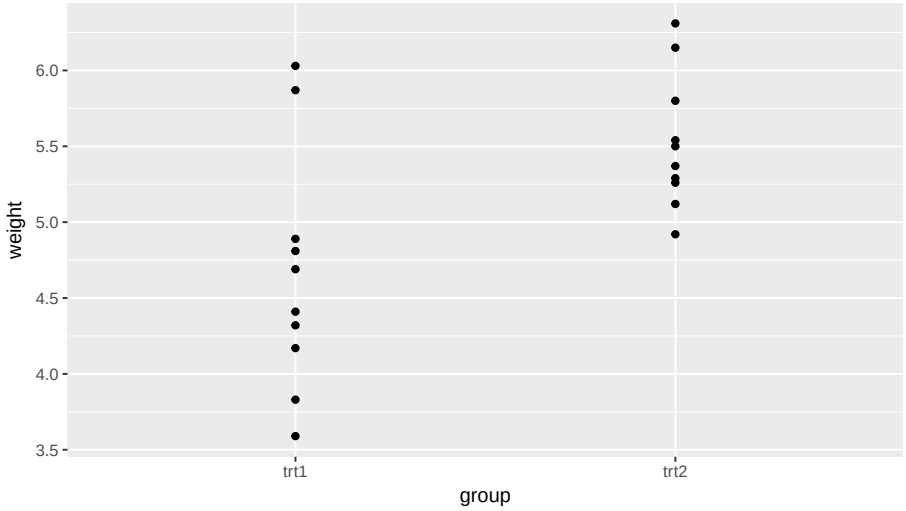


Figure 6.4: Relation between type of seat and price, with the regression line being not quite the least squares .

we could have obtained the exact same result with a regression analysis using SPSS.

The interesting thing about a dummy variable is that the slope of the regression line is exactly equal to the differences between the two averages. If we look at Equation 6.7, we see that the slope coefficient is 0.865 and this is exactly equal to the difference in mean weight for treatment 1 and treatment 2. Thus, the slope coefficient for a dummy variable indicates how much the average of the treatment that is coded as 1 differs from the treatment that is coded as 0. Here the slope is positive so that we know that the treatment coded as 1 (trt2), leads to a higher average yield than the treatment coded as 0 (trt1). This makes it possible to test null-hypotheses about group means.

6.3 Testing hypotheses about group means

In the previous section we saw that the slope in a dummy regression is equal to the difference in group means.

Suppose researchers are interested in the effects of different treatments on yield. They'd like to know what the difference is in yield between treatments 1 and 2, using a sample of 30 data points. Based on this sample, they'd like to generalize to the population of all yields based on treatments 1 and 2. They adopt a type I error rate of $\alpha = 0.05$.

The researchers analyze the data and they find the results as displayed in Table 6.5

The 95% confidence interval for the slope is from 0.2612664 to 1.4687336.

Table 6.5: Yield by treatment.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6610	0.2032	22.94	0.0000
grouptrt2	0.8650	0.2874	3.01	0.0075

This means that reasonable values for the difference between the two treatments on yield lie within this interval. All these values are positive, so we reasonably believe that treatment 2 leads to higher yield. We know that it is treatment 2 that leads to higher yields, because the slope in the regression equation has been coded as 'grouptr2'. Thus, a dummy variable has been computed, 'grouptrt2', where trt2 has been coded as 1 (and trt1 consequently coded as 0).

We could also have coded the dummy variable by hand first, and then use this variable in the linear regression. In the next section, we will see how to do that in SPSS.

The 95% confidence interval for the slope does not contain 0, so we can therefore reject the null-hypothesis that there is no difference in group means at an α of 5%. The exact p-value can be read from Table 6.5 and is equal to 0.0075.

Thus, based on this regression analysis the researchers can write in a report that there is a significant difference between the yield after treatment 1 and treatment 2, $p = 0.01$. Treatment 2 leads to a higher yield of about 0.87 (SE=0.29) more than treatment 1 (95% CI: 0.26, 1.47).

6.4 Regression analysis using a dummy variable in SPSS

In SPSS there are two ways in which you can use a linear model with a qualitative independent variable. The first and easiest way is to tell SPSS that your variable is to be treated qualitatively, and you do that by the keyword BY. For instance, for the data set on treatment 1 and 2 and yield, you get the following syntax:

```
UNIANOVA growth BY group
/DESIGN group
/PRINT parameter.
```

All variables after the BY keyword are automatically turned into dummy variables. SPSS chooses automatically what value of group is coded as 1 and what value of group is coded as 0. SPSS uses alphabetical and numerical order. Thus, if we have trt1 and trt2 as values, SPSS automatically picks trt1 as the one that is coded as 1, because that is first in the alphabetical-numerical order.

The output then looks like as displayed in Figure **??**. Here, we see an intercept of ... , a slope of .... for dummy variable [group=trt1] and a slope of 0 for [group=trt2]. Of course we don't need this slope, therefore SPSS tells us it is

redundant. So we have to read the SPSS output table like this: If group=trt1, then the expected weight differs from trt2 (the reference category) by .... .

We call trt2 as the *reference category* since it is the category that we use for comparison. The reference category is the category that is coded 0.

Parameter Estimates

Dependent Variable: weight

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.526	.203	27.195	.000	5.099	5.953
[group=trt1]	-.865	.287	-3.010	.008	-1.469	-.261
[group=trt2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 6.5: SPSS output of a regression analysis of weight on treatment.

Sometimes, the automatic choice by software is something you don't want. Suppose that you'd like to compare two treatments: the old treatment, and a new treatment, one that avoids all the insectides that are so bad for bees and bumblebees. Here, the most interesting question is how the new treatment differs from the old one. So you'd like to use the old treatment as the reference category (coded as 0) to which you want to compare the yield of the new treatment. In order to choose your own way of dummy coding, and thereby choosing your reference category you can use the syntax to create a new variable treatment2

```
RECODE group ('trt2'=1) (ELSE=0) INTO treatment2.
EXECUTE.
```

Then you use a slightly altered syntax. First, you now use the new variable treatment2. Second, you use the keyword WITH instead of BY.

```
UNIANOVA weight WITH treatment2
/DESIGN treatment2
/PRINT parameter.
```

With the keyword BY, SPSS chooses its own reference category. With the keyword WITH, you treat the new variable treatment2 as a quantitative variable. The result is the output in Figure 6.6.

Because you now use your own dummy variable and use it as any quantitative variable, the output looks a bit simpler. You see an intercept of 4.661 and a slope of 0.865 for the treatment2 variable. Thus, if this treatment2 variable has value 0, the expected weight equals the intercept 4.661, and if this treatment2 variable has value 1, the expected weight equals the sum of the intercept and the slope, $4.661 + 0.865 = 5.526$. Notice that by using a different reference category

Parameter Estimates

Dependent Variable: weight

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4.661	.203	22.938	.000	4.234	5.088
treatment2	.865	.287	3.010	.008	.261	1.469

Figure 6.6: SPSS output of a regression analysis of weight on treatment.

(now trt1), the sign of the slope has changed. The intercept has also changed, as the intercept is the expected weight for the reference category.

In general, we advise to use the BY keyword to indicate you'd like to have an automatically coded dummy variable. If however the output is very hard to interpret, think of the best way to code your own dummy variable.

For experimental designs, it makes sense to code control conditions as 0, and experimental conditions as 1. For surveys, if you want to compare how a social minority scores relative to a social majority, it makes sense to code the minority group as 1 and the social majority as 0. In eduational studies, it makes sense to code an old teaching method as 0 and a new method as 1.

6.4.1 Exercise

1. Look at the output in Figure 6.7. It describes the comparison between average height in males and females. Based on this output: what is the average height in females? And what is the average height in males?

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	182.400	1.233	147.946	.000	179.910	184.890
[sex=female]	-12.067	1.906	-6.332	.000	-15.915	-8.218
[sex=male]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 6.7: SPSS output of a regression analysis of height on sex.

2. Look at the output in Figure 6.8. It describes the comparison between average height in Ethiopeans and Japanese people. A variable ethnicity was used that equals 1 for Ethiopeans and 0 for Japanese people. The analysis was run using the BY keyword, treating the ethnicity variable as qualitative. Based on this output: what is the average height in the Japanese in this data set? And what is the average height in the Ethiopeans?

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	167.167	1.517	110.191	.000	164.103	170.230
[ethnicity=.00]	15.233	1.990	7.656	.000	11.215	19.251
[ethnicity=1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 6.8: SPSS output of a regression analysis of height on ethnicity.

3. A study looks into the effects of drinking milk during childhood on adult height. A number of adults are categorized into those that that have been drinking less than 1 liter of milk per month during childhood (coded as milk=0) and into those that have been drinking at least 1 liter of milk per month during childhood (coded as milk=1). A regression analysis treating the milk variable as quantitative (using the WITH keyword) yields the output in Figure 6.9.

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	169.278	1.806	93.710	.000	165.630	172.926
milk	10.482	2.369	4.425	.000	5.698	15.267

Figure 6.9: SPSS output of a regression analysis of height on milk.

What is the average height in people who drank less than 1 liter of milk per month during childhood? And what is the average height in people who drank at 1 liter of milk or more per month?

4. A study looks into the effect of vitamin B2 (riboflavin) on the frequency of migraine attacks. It compares 100 patients who take a pill containing 50 mg of riboflavin per day for a month and 100 patients who take a pill containing 0 mg of riboflavin per day for a month. Suppose you want to code a dummy variable called **riboflavin** in order to perform a regression analysis. Which patients would you code as 1, and which patients as 0? Motivate your answer.

Answers:

1. The intercept equals 182.4. For sex=female, there is an extra effect of -12.067. For sex=male, the extra effect is fixed to 0. Therefore, the average

height in females in this data set is 182.4-12.067=170.333 and the average height in males equals 182.4.

2. The intercept is 167.167. If ethnicity=0, then there is an extra height of 15.233. Since the Japanese are coded as 0, the average height in the Japanese people in the data set equals $167.167 + 15.233 = 182.4$. The average height in the Ethiopeans in this data set equals 167.167.

3. The intercept equals 169.278 and the slope of milk is 10.482. That means that people who score 0 on milk, have an average height of $169.278 + 10.482 \times 0 = 169.278$. The ones that score 0 on the milk variable drank less than 1 liter of milk per month. The ones that score 1 on the milk variable drank at least 1 liter of milk per month, and their average is $169.278 + 10.482 \times 1 = 179.76$.

4. If you're interested in the effect of riboflavin, you'd like to compare the people who took 50 mg to those who took 0 mg. How much more or less frequent are the migraine attacks in people who took riboflavin relative to those that did not take extra riboflavin? Then the natural reference cateogory is the group with 0 mg riboflavin. If you then analyze the output, the effect of **riboflavin** is then the increase or decrease in migraine frequency in people who took riboflavin.

6.5 Dummy coding for more than two groups

In the previous sections we saw how to code a qualitative variable with 2 categories into 1 dummy variable. In this section, we see how to code a qualitative variable with 3 categories into 2 dummy variables, and to code a qualitative variable with 4 categories into 3 dummy variables, etcetera.

Take for instance the variable Country, where in your data set, there are three different values for this variable, for instance, Norway, Sweden and Finland, or perhaps Zimbabwe, Congo and South-Africa. Let's call these countries A, B and C. Here's a data example:

ID	Country	height
001	A	120
002	A	160
003	B	121
004	B	125
005	C	140
...	...	...

We can code this Country variable with three categories into two dummy variables in the following way. First we create a variable countryA. This is a dummy variable, or indicator variable, that indicates whether a person comes from country A or not. Those that do are coded 1, and those that don't are coded as 0. Next, we create a dummy variable countryB that indicates whether or not

people come from country B. Again, those that do are coded 1 and those that don't are coded 0. The resulting variables are displayed in Table 6.5

ID	Country	height	countryA	countryB
001	A	120	1	0
002	A	160	1	0
003	B	121	0	1
004	B	125	0	1
005	C	140	0	0
...	...	...	...	...

Note that we have now for every value of Country (A, B, or C) a unique combination of the variables countryA and countryB. All those from country A have a 1 for countryA and a 0 for countryB; all those from country B have a 0 for countryA and a 1 for countryB, and all those from country C have a 0 for countryA and a 0 for countryB. Therefore a third dummy variable countryC is not necessary.

Remember that with two categories, you only need one dummy variable, where one category gets 1s and another category gets 0s. In this way both categories are uniquely identified. Here with three categories we also have unique codes for every category. Similarly, if you have 4 categories, you can code this with 3 dummy variables. In general, when you have a variable with K categories, you can code them with $K - 1$ dummy variables.

6.5.1 Exercise

Below you see a table with data on the favorite colour of 10 children. The only colours mentioned are blue, pink, purple, red and green.

ID	Colour
001	purple
002	green
003	red
004	blue
005	red
006	pink
007	pink
008	green
009	blue
010	red

How many dummy variables do you need in order to uniquely identify each colour? Construct the dummy variables by hand and add them to the table. Don't forget the variable names. You can start with any colour you'd like.

Answer: You have 5 different colours, so you need 4 dummy variables.

ID	Colour	purple	green	red	blue
001	purple	1	0	0	0
002	green	0	1	0	0
003	red	0	0	1	0
004	blue	0	0	0	1
005	red	0	0	1	0
006	pink	0	0	0	0
007	pink	0	0	0	0
008	green	0	1	0	0
009	blue	0	0	0	1
010	red	0	0	1	0

6.6 Analyzing categorical predictor variables in SPSS

Suppose we have data on height based on a sample of thirty people ($N = 30$) that come from three different countries. We want to know whether the average height is different for each country, or whether the average height is the same (null-hypothesis). Since we know that applying a linear model to a categorical independent variable is the same as modelling group means, we can test the null-hypothesis that all group means are equal in the population. Let μ_A be the mean height in the population of country A, μ_B be the mean height in the population of country B, and μ_C be the mean height in the population of country C. Then we can specify the null-hypothsis using symbols in the following way:

$$H_0 : \mu_A = \mu_B = \mu_C \tag{6.8}$$

In all group means are equal in the population, then all population intercepts would be 0. We want to test this null-hypothesis with a linear model in SPSS. Now there are two ways of doing this. First option is that you can use dummy coding first, and then treat these dummy variables in a quantitative way. The second option is that you let SPSS do the dummy coding for you, by indicating that you want to treat the original variable as qualitative. Let's start with the first option and then discuss the second option. Afterwards we will compare these two options.

6.6.1 Treating dummy variables quantitatively

First we create two new dummy variables, and then perform a linear model analysis using these. Note that we actually perform a multiple regression with two dummy variables. We use the keyword WITH to indicate that we want treat the dummy variables quantitatively.

```
RECODE Country ('A'=1) ('B'=0) ('C'=0) INTO CountryA.
```

```
RECODE Country ('A'=0) ('B'=1) ('C'=0) INTO CountryB.
EXECUTE.
UNIANOVA height WITH CountryA CountryB
/ design = CountryA CountryB
/ print = parameter.
```

Univariate Analysis of Variance

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	880.067[a]	2	440.033	9.763	.001
Intercept	297217.600	1	297217.600	6594.523	.000
CountryA	28.800	1	28.800	.639	.431
CountryB	510.050	1	510.050	11.317	.002
Error	1216.900	27	45.070		
Total	920497.000	30			
Corrected Total	2096.967	29			

a. R Squared = .420 (Adjusted R Squared = .377)

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	172.400	2.123	81.207	.000	168.044	176.756
CountryA	-2.400	3.002	-.799	.431	-8.560	3.760
CountryB	10.100	3.002	3.364	.002	3.940	16.260

Figure 6.10: Output of a regression analysis on two dummy variables, using the keyword WITH.

In the Parameter Estimates table in Table 6.10, we see the effects of the two dummy variables. All observations with a 1 for variable CountryA get an extra predicted height of -2.4, and all observations with a 1 for variable CountryB get an extra predicted height of 10.1. So the expected height in country A equals $172.4 - 2.4 = 174.8$, and the expected height in country B equals $172.4 + 10.1 = 182.5$. Observations in country C have a 0 for both variables CountryA and CountryB, so the expected height in country C equals the intercept 172.4.

In the Tests of between-subjects Effects table, we see other stuff going on. This is not regression output, but output based on a so-called Analysis Of VAriance, or ANOVA for short. First note that the significance levels (the p-values) for the two effects are exactly the same as those from the regression table. Second, note that the reported values of F are the square of the t values in the regression table: $-.799^2 = .619$ and $3.364^2 = 11.317$.

ANOVA is a particular case of a linear model. The F-statistic is constructed on the basis of Sums of Squares. For instance, take a look at the row for the effect of CountryA. The sum of squares is equal to 28.80. If you divide this by the degrees of freedom for this effect, your get the Mean Square: $28.80/1 = 28.80$. Now look at the row for Error. The sum of squares equals 1216.90. Divided by the corresponding degrees of freedom you get the Mean Square: $12.16.90/27 = 45.07$. You obtain the F-statistic by dividing the CountryA Mean Square by the Error Mean Square: $F = 28.80/45.07 = 0.639$. It is not a coincidence that this F-value is exactly equal to the square of the corresponding t-value: $F = t^2$. Remember that the t-value is equal to the B parameter divided by the standard error: $t = -2.400/3.002 = -.799 = \sqrt{0.639}$. To obtain the regression coefficient we minimize the sums of squares of the residuals. So both the F-statistic and the t-statistic come from computing sums of squares and are thus based on the same general logic of the linear model.

Since ANOVA is a special case of the linear model, we believe that it is not necessary to understand ANOVA fully: if you understand the linear model, that is good enough. Just remember that sometimes you see ANOVAs reported in the literature. Be aware that what they are actually doing is running a linear model.

Returning back to our null-hypothesis that all group means are equal in the population: if the group means are equal in the population, then the slope parameters for countryA and countryB should consequently be 0 in the population. Looking at the 95% confidence intervals, we see that 0 is a reasonable value for the difference between country C (the reference category) and country A, but 0 is *not* a reasonable value for the difference between country C and country B. But how can we rigourously test the null-hypothesis that all three group means are the same? Now we have two p-values, one for countryA and countryB, but we need one p-value for the null-hypothesis of three equal means. Let's see if we can get one p-value if we try the second way to perform this analysis.

6.6.2 Treating the original variable qualitatively

In the second approach, we let SPSS do the dummy variable coding automatically. In that case we use the original variable Country with its three categories directly, and change the WITH keyword into BY in the following way:

```
UNIANOVA height BY Country
/ design = Country
/ print = parameter.
```

All variables named after BY are treated as categorical variables and automatically coded into dummy variables. The output then looks like the following:

The Parameter Estimates table in Figure 6.12 now looks slightly different: The intercept is the same, the dummy effects are presented in a slightly different way, and there is an extra row for country C where a regression coefficient B

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	880.067[a]	2	440.033	9.763	.001
Intercept	918400.033	1	918400.033	20377.024	.000
country	880.067	2	440.033	9.763	.001
Error	1216.900	27	45.070		
Total	920497.000	30			
Corrected Total	2096.967	29			

a. R Squared = .420 (Adjusted R Squared = .377)

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	172.400	2.123	81.207	.000	168.044	176.756
[country=A]	-2.400	3.002	-.799	.431	-8.560	3.760
[country=B]	10.100	3.002	3.364	.002	3.940	16.260
[country=C]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 6.11: Output of a regression analysis on the original variable, using the keyword BY.

of 0 is reported, with no other information. The values for the other effects are exactly the same as with the previous analysis. This means we can interpret these country=A and country=B effects as the effects of dummy variables: all observations start from an intercept of 172.40 and depending on whether the observation from country A or country B, you get an extra predicted height of -2.4 or 10.1, respectively. Observations from country C get an extra height of 0, so in effect nothing extra. (SPSS creates an extra dummmy variable for country C, but because this is not necessary, the effect is fixed to 0).
Also the Tests of Between-Subjects Effects table looks slightly different: instead of two separate effects for two dummy variables, we now see one row for the original variable Country. And in the column df (degrees of freedom), instead of 1 degree of freedom for a country effect, we see 2 degrees of freedom. So this suggests that *the effects of the two dummy variables are now combined into one effect*, with one particular F-value, and one p-value that is also different from those of the two separate dummy variable. This is actually the p-value test for the null-hypothesis that all 3 means are equal.

$$H_0 : \mu_A = \mu_B = \mu_C \tag{6.9}$$

This is very different from the t-tests in the Parameter Estimates table. The t-test for the country=A effect specificically tests whether the average height in country A is different from the average height in country C (the reference country). The t-test for the country=B effect specifically tests whether the

average height in country B is different from the average height in country C (the reference country). Since these do not refer to our research question regarding overall differences across all three countries, we do not report these t-tests, but report the overal F-test from the Tests of Between-Subjects Effects table.

6.6.3 Reporting ANOVA

In all cases where you have a categorical predictor variable with more than two categories, and where the null-hypothesis is about the equality of all means, you always report the F-statistic from the Tests of Between-Subjects Effects. You do that in the following way for this particular example:

> "The null-hypothesis that all 3 population means were equal was tested with a linear model (analysis of variance). The results showed that the null-hypothesis can be reject: the means in the population are not equal, $F(2, 27) = 9.76, MSE = 45.07, p = 0.001$."

Always check the degrees of freedom for your F-statistic. The first number refers to the number of dummy variables that are tested at once: this is the number of categories minus 1. The second number refers to the error degrees of freedom: this is the number of observations minus the number of effects in your model. In this model you have 30 data points and you have three effects (parameters): one intercept, one effect for Country=A, and one effect for Country=B. So your error degrees of freedom is $30 - 3 = 27$. Note that this error degrees of freedom is equal to that of the t-statistic for multiple regression.

6.7 F-test for multiple group comparisons

Here we slightly elaborate on the F-test for testing null-hypotheses about group means. Remember that the T-statistic was based on the slope divided by its standard error. Above we saw that the F-statistic is based on the ratio of mean squared errors, that are in turn based on sums of squares.

If we go back to Chapter 2 on the inference about population slopes, we remember that given that the population slope is 0, and if one draws many random samples, the distribution of T-statistics shows a t-distribution with a certain degrees of freedom that depends on sample size. Similarly for inference about population group means, given a null-hypothesis that three group means are equal in the population, and if one draws many random samples from this population, the F-statistic shows an F-distribution with a certain model degrees of freedom of 2 and an error degrees of freedom that depends on the sample size.

Figure 6.12 shows the F-distribution with 2 model degrees of freedom and 156 residual degrees of freedom. F-values are always positive because they are based on sums of squares. The larger the F-value, the less likely it is to be the result of sampling error. Thus, if the F-value is very large, it is not likely that the population means are equal. When is an F-value large enough to think

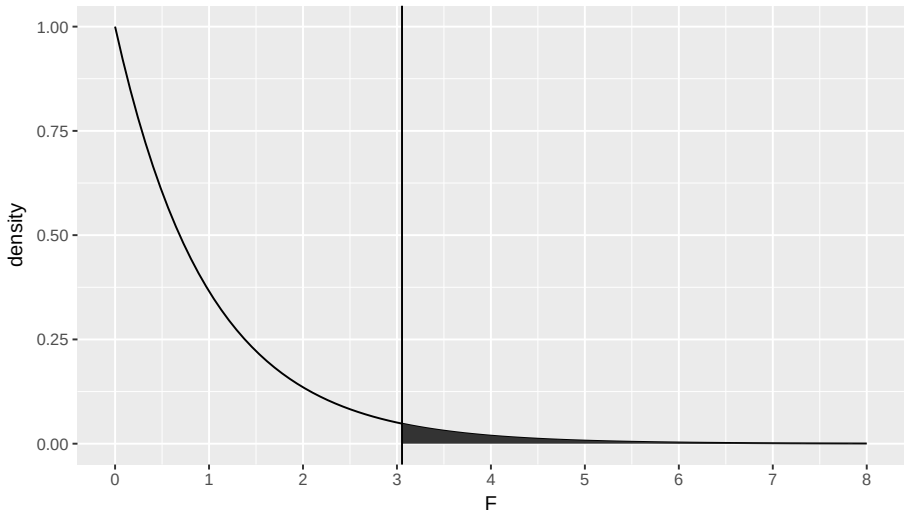


Figure 6.12: The F-distribution with 2 model degrees of freedom and 156 error degrees of freedom. The shaded area is the upper 5 percent of the distribution. The critical F-value for alpha=0.05 is depicted by the vertical line.

that the null-hypothesis is not true? Similar to T-statistics, we can choose an arbitrary level of significance, say $\alpha = 0.05$, and reject the null-hypothesis when the F-value is beyond the critical value associated with the α. For this particular F-distribution, the critical F-value is 3.05. This number can be looked up in tables or is available in software packages like SPSS. Thus, if we find a F-value equal to or larger than 3.05, we reject the null-hypothesis. If the F-value is less than 3.05, we do not reject the null-hypothesis.

We've also stated that the t-distribution and the F-distribution have much in common. Here we will illustrate this. Suppose that we test the null-hypothesis that a certain population slope is 0. We perform a regression analysis and obtain a T-statistic of -2.40. Suppose our sample size was 42, so that our residual degrees of freedom equals $42 - 2 = 40$. Figure 6.13 shows the theoretical t-distribution with 40 degrees of freedom. It also shows our value of -2.40. The shaded area represent the values for T that would be significant at an $\alpha = 0.05$.

Now if we take every possible T-value and square it, we can plot these and look at the distribution (see Figure 6.14).

It turns out that with an F-distribution with 1 model degrees of freedom and 40 residuals degrees of freedom, the proportion of F-values larger than 5.76 equals 0.0211408. The proportion of T-values larger than 2.40 or smaller than-2.40 equals also 0.0211408. Thus, the two-sided p-value associated with a certain T-value, is equal to the p-value associated with an F-value that is the square of the T-value.

This means that if you see a T-statistic of say 3 reported with a residu-

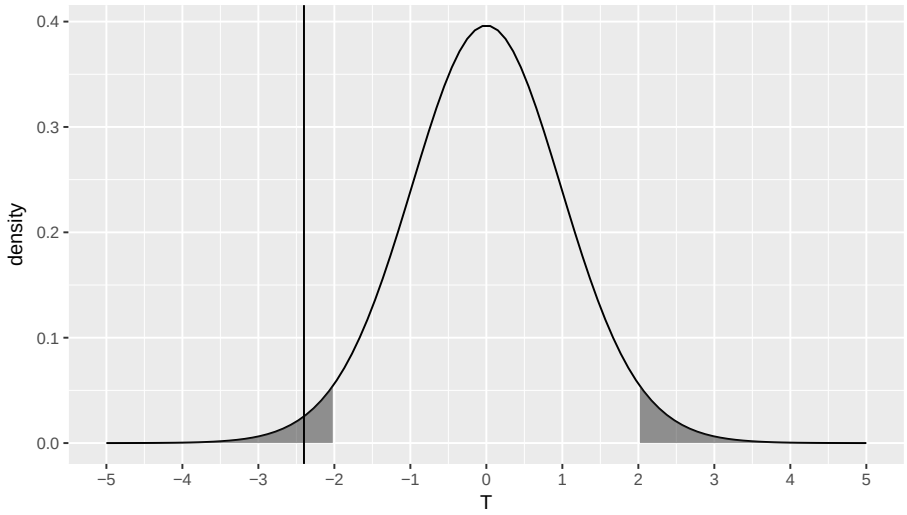


Figure 6.13: The vertical line represents a T-value of -2.40. The shaded area represents the extreme 5 percent of the possible T-values

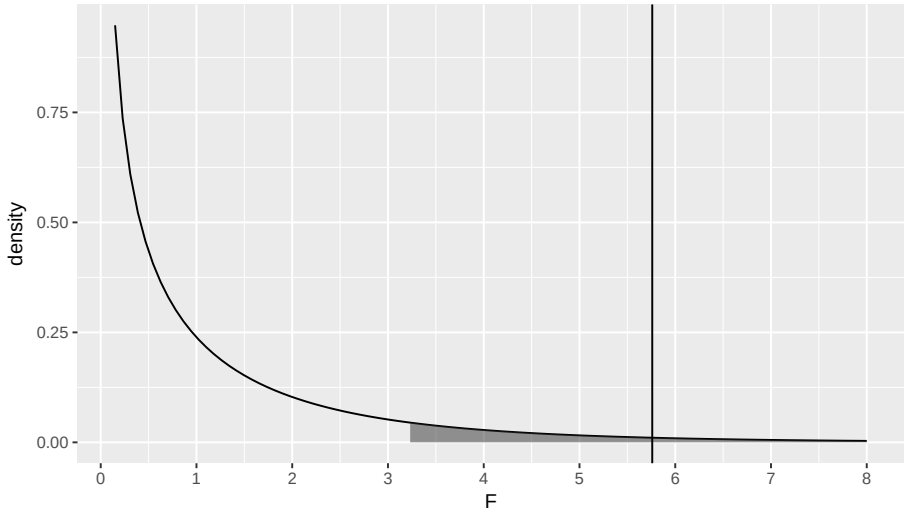


Figure 6.14: The F-distribution with 1 model degrees of freedom and 40 error degrees of freedom. The shaded area is the upper 5 percent of the distribution. The vertical line represents the the square of -2.40: 5.76

als degrees of freedom of 50, you can equally report this as an $F(1, 50) = 9$. Conversely, if you see a reported F-value of $F(1, 67) = 49$, you could without problems turn this into a $t(67) = 7$. Note however that this only the case if the model degrees of freedom of the F-statistic is equal to 1. Next time you look at UNIANOVA output, watch the t-statistics and F-statistics carefully and check whether the F-statistic is the square of the T-statistic. Check for instance Figure 6.10 again.

Chapter 7

Moderation: testing interaction effects

7.1 Categorical by linear interaction

Suppose there is a linear relationship between age and vocabulary: the older you get, the more words you know. Suppose we have the following linear regression equation for this relationship:

$$vocab = 200 + 500 \times age + e \tag{7.1}$$

So according to this model, the expected number of words for a newborn baby (age=0) equals 200. This may sound silly, but suppose this model is a very good model for vocabulary size in children between 2 and 5 years of age. Then this equation tells us that the expected increase in vocabulary size is 500 words per year.

This model is meant for everybody in the Netherlands. But suppose that one researcher expects that the increase in words is much faster in children from high SES families than in children from low SES families. First he believes that vocabulary will be larger in higher SES children than in low SES children. In other words, he expects an effect of SES, over and above the effect of age:

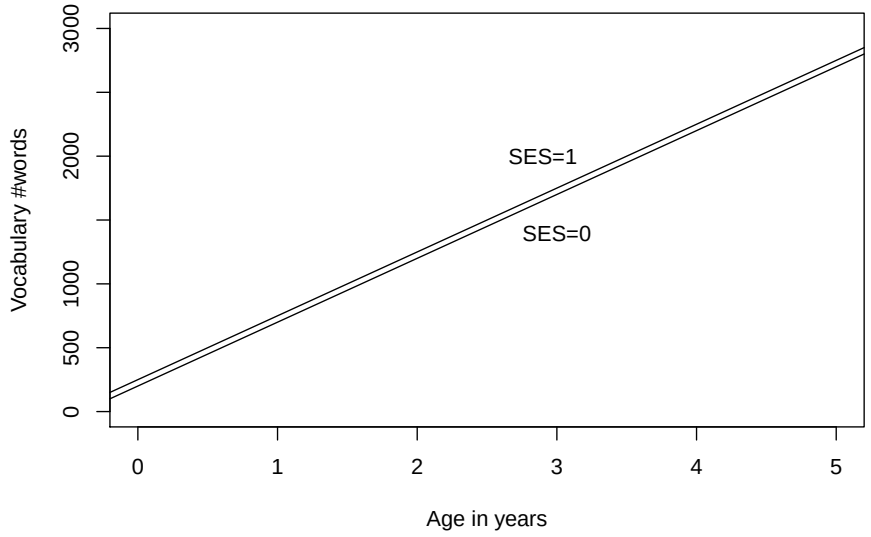
$$vocab = 200 + 500 \times age + \beta_2 \times SES + e \tag{7.2}$$

This main effect of SES is yet unknown and denoted by β_2.This linear equation is an example of a multiple regression.

Let's use some numerical example. Suppose age is coded in years, and SES is dummy coded, with a 1 for high SES and a 0 for low SES. Let β_2, the effect of SES, be 10. Then we can write out the linear equation for low SES and high SES separately.

$$lowSES : vocab = 200 + 500 \times age + 10 \times 0 + e = 200 + 500 \times age + e \quad (7.3)$$
$$lowSES : vocab = 200 + 500 \times age + 10 \times 1 + e = (200 + 10) + 500 \times age + e \quad (7.4)$$

Figure **??** depicts the two regression lines for the high and low SES children separately. So we see that the effect of SES involves a change in the intercept: the intercept equals 200 for low SES children and the intercept for high SES children equals 210. The difference in intercept is indicated by the coefficient for SES. Note that the two regression lines are parallel: for every age, the difference between the two lines is equal tot 10. For every age therefore, the predicted number of words is 10 words more for high SES children than for low SES children.



However, our researcher also expects that the *yearly increase* in vocabulary is a bit lower than 500 words in low SES families, and a little bit higher than 500 words in high SES families. In other words, he believes that SES might *moderate* (affects, changes) the slope coefficient. Let's call the slope coefficent in this case β_1. In the above equation this slope parameter is equal to 500, but let's now let it be a linear function of SES:

$$\beta_1 = \alpha^* + \beta_3 \times SES \quad (7.5)$$

In words: the slope coefficient for the regression of vocabulary on age, is itself linearly related to SES. We predict the slope on the basis of SES. So now we have two linear equations for the relationship between vocabulary, age and SES:

$$
\begin{aligned}
vocab &= 200 + \beta_1 \times age + \beta_2 \times SES + e && (7.6)\\
\beta_1 &= \alpha^* + \beta_3 \times SES && (7.7)
\end{aligned}
$$

We can rewrite this by plugging the second equation into the first one:

$$vocab = 200 + (\alpha^* + \beta_3 \times SES) \times age + \beta_2 \times SES + e \tag{7.8}$$

Multiplying this out gets us:

$$vocab = 200 + \alpha^* \times age + \beta_3 \times SES \times age + \beta_2 \times SES + e \tag{7.9}$$

If we rearrange the terms a bit, we get:

$$vocab = 200 + \alpha^* \times age + \beta_2 \times SES + \beta_3 \times SES \times age + e \tag{7.10}$$

Now this very much looks like a regression equation with one intercept and *three* slope coefficients: one for age (α^*), one for SES (β_2) and one for SES$\times$ age (β_3).

we might want to change the label α^* into β_1 to get a more familiar looking form:

$$vocab = 200 + \beta_1 \times age + \beta_2 \times SES + \beta_3 \times SES \times age + e \tag{7.11}$$

So the first slope coefficient is the increase in vocabulary for every year that age increases (β_1), the second slope coefficient is the increase in vocabulary for an increase of 1 on the SES variable (β_2), and the third slope coefficient is the increase in vocabulary for every increase of 1 on the *product* of age and SES (β_3).
So what does this mean exactly?

If we look at this equation:

$$\beta_1 = \alpha^* + \beta_3 \times SES \tag{7.12}$$

we see that a high positive value of β_3 increases the size of β_1, which is the effect of age on vocabulary.

Suppose we find the following solution for the regression equation:

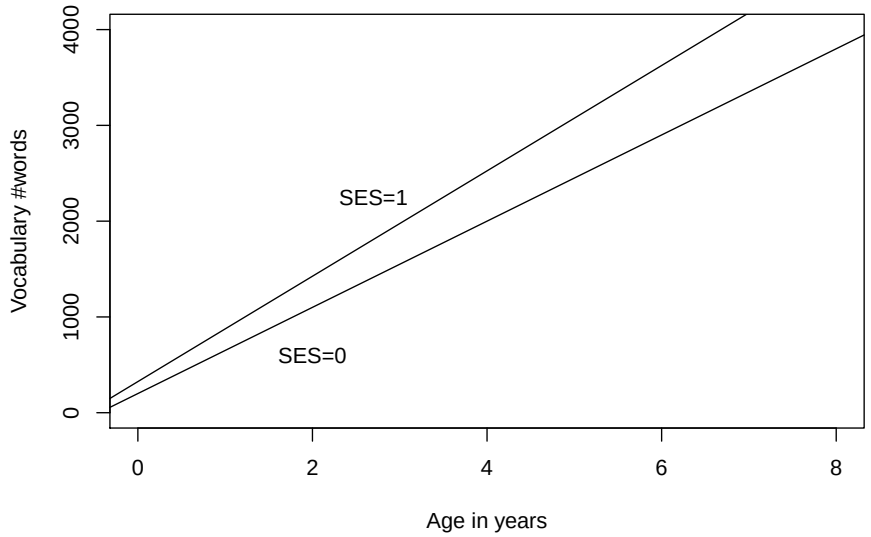
$$
\begin{aligned}
vocab &= 200 + \beta_1 \times age + \beta_2 \times SES + \beta_3 \times SES \times age + e && (7.13)\\
vocab &= 200 + 450 \times age + 125 \times SES + 100 \times SES \times age + e && (7.14)
\end{aligned}
$$

If we code low SES children as SES=0, and high SES children as SES=1, we can write the above equation into two regression equations, one for low SES children (SES=0) and one for high SES chilrden (SES=1):

$$lowSES : vocab = 200 + 450 \times age + e \quad (7.15)$$
$$highSES : vocab = 200 + 450 \times age + 125 + 100 \times age + e \quad (7.16)$$

So for low SES children, the intercept is 200 and the regression slope for age is 450, so they learn 450 words per year. For high SES children, we see the same intercept of 200, with an extra 125 (this is the main effect of SES). So effectively their intercept is now 325. For the regression slope, we now have $450 \times age + 100 \times age$ which is of course equal to $550 \times age$. So we see that the high SES group has both a different intercept, and a different slope: the increase in vocabulary is 550 per year: somewhat steeper than in low SES children. So yes, the researcher was right: vocabulary increase per year is faster in high SES children than in low SES children.

These two different regression lines are depicted below. It can be clearly seen that the lines have two different intercepts and two different slopes. That they have two different slopes can be seen from the fact that the lines are not parallel. One has a slope of 450 words per year and the other has a slope of 550 words per year. This difference in slope of 100 is exactly the size of the slope coefficient pertaining to the product $SES \times age$, β_3.



The observation that the slope coefficient is different for different groups is called an *interaction effect*, or *interaction* for short. Other words for this

phenomenon are *modification* and *moderation*. In this case, SES is called the *modifier variable*: it modifies the relationship between age on vocabulary. (Note however that you could also interpret age as the modifier variable: the effect of SES is larger for older children than for younger children. In the plot you see that the difference between vocab for high and low SES children of age 6 is larger than it is for children of age 2.)

So, what do you have to do if you want to know if there is an interaction effect between age and SES on vocabulary size?

First you dummy code the nominal variable SES:

```
RECODE SES ('low'=0) ('high'=1) INTO SES_dummy.
EXECUTE.
```

Next we compute a new variable, that is, the product $SES \times age$ (but use the dummy variable):

```
COMPUTE SESage = SES_dummy * age .
EXECUTE.
```

So now you have three variables for a multiple regression:

```
UNIANOVA vocab WITH age SES_dummy SESage
/ design=age SES_dummy SESage.
```

Note there is also a faster way of analyzing interaction effects in SPSS. The following syntax is exactly equivalent, but does not require the computation of the interaction variable $SESage$:

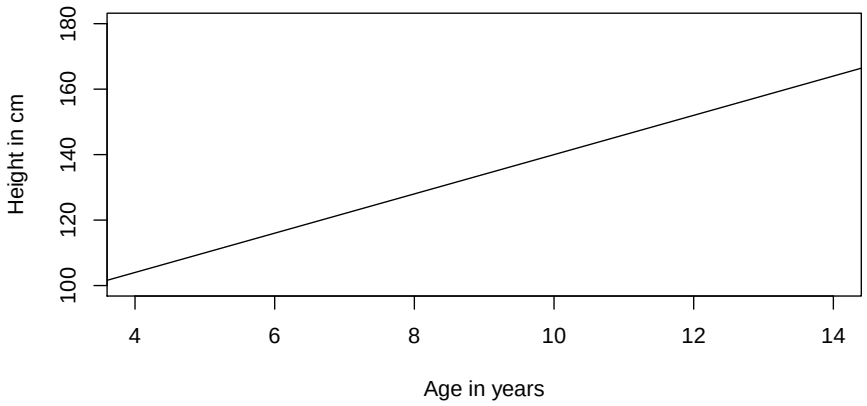
```
UNIANOVA vocab WITH age SES_dummy
/ design = age SES_dummy age*SESdummy
/ print = parameter.
```

With this design specification, SPSS computes the interaction variable automatically for you.

Let's look at some example output for another data set. A researcher is interested in childrens' height. She has data on children between the ages of 4 and 8, with measures on their height. She wants to know whether children growing up in the city grow just as fast as in the countryside. So the data might look something like this.

child	location	age	height
001	city	5	120
002	country	14	160
003	city	4	121
004	city	6	125
005	country	9	140
...	...	...	...

The general regression of height on age might something like this:



This regression line for the entire sample of children has a slope of around 6 cm per year. Now the researcher wants to know whether this slope is the same for children in the cities and in the countryside, in other words, do children grow as fast in the city as in the countryside? We might expect that location (city vs countryside) *moderates* the effect of age on height. We use the following SPSS syntax to study this *location* × *age* effect, first creating a dummy variable for location, arbitrarily coding *country* as 1:

```
RECODE location ('city'=0) ('country'=1) INTO location_dummy.
EXECUTE.
UNIANOVA height WITH age location_dummy
/ design = age location_dummy  age*location_dummy
/ print = parameter.
```

Below we find the corresponding SPSS output. So the null-hypothesis is that the two slopes are equal, in other words, that the interaction effect equals zero. In the output, this is the age * location_dummy effect.

In the table with the parameter estimates, we find the regression coefficients.

So we can fill in the regression equation:

$$height = 96 + 4.6 \times age + 3.8 \times locationdummy - 0.368 \times age \times locationdummy + e$$

If we fill in 0s for the location dummy, we get the equation for city children:

$$height = 96 + 4.6 \times age + e$$

Univariate Analysis of Variance

[DataSet1]

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1487.283[a]	3	495.761	17.616	.004
Intercept	7139.880	1	7139.880	253.696	.000
age	1214.207	1	1214.207	43.144	.001
location_dummy	3.079	1	3.079	.109	.754
age * location_dummy	1.450	1	1.450	.052	.829
Error	140.717	5	28.143		
Total	154509.000	9			
Corrected Total	1628.000	8			

a. R Squared = .914 (Adjusted R Squared = .862)

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	96.652	6.068	15.928	.000	81.053	112.251
age	4.626	.704	6.568	.001	2.815	6.436
location_dummy	3.848	11.633	.331	.754	-26.055	33.751
age * location_dummy	-.368	1.621	-.227	.829	-4.535	3.799

So the intercept equals 96 and the slope equals 4.6.

If we fill in 1s for the location dummy variable, we get the equation for country side children:

$$\begin{aligned} height &= 96 + 4.6 \times age + 3.8 - 0.368 \times age + e \\ &= (96 + 3.8) + (4.6 - 0.368) \times age + e \end{aligned} \tag{7.17}$$

We see that that the intercept is now equal to the intercept is $96 + 3.8$, and the slope equals $4.6 - 0.368$.

So, we know that the slope for countryside children is 0.368 less steep than for city children. In this sample, it seems that children in the city grow 4.626 centimeters per year (on average), but that children in the countryside grow $4.626 - 0.368 = 4.258$ centimeters per year (on average). Is this value of 0.368 possible if the value in the entire population of children equals 0? In other words, is the value of 0.368 significantly different from 0? No, the effect of 0.368 is not significant, $t(5) = -0.23, p > 0.05$. We therefore do not reject the null-hypothesis and conclude that there is *no* evidence that children in the city grow at a different pace than children in the countryside.

In this section we discussed the situation that regression slopes might be different in two groups: the regression slope might be steeper in one group than in another group. So suppose that we had a continuous predictor x for a continuous dependent variable variable y, we said that a particularly dummy variable z moderated the effect of x on y. This moderation was quantified by an interaction effect, $x \times z$.

So suppose we have the following regression equation:

$$y = -b_0 + b_1 \times x + b_2 \times dummy + b_3 \times x \times dummy + e$$

Here, we call b_0 the intercept, b_1 the main effect of x, b_2 the main effect of the dummy variable, and b_3 the interaction effect of x and the dummy.

7.1.1 Exercises

We have the following regression equation, with y as dependent variable, x as a continuous predictor variable, and a dummy variable *dummy*.

$$y = 5.3 + 3.6 \times x + 3.8 \times dummy + 8.2 \times x \times dummy + e$$

Write down the regression equation in the case the dummy variable equals 0.
Write down the regression equation in the case the dummy variable equals 1.
What is the intercept if the dummy variable equals 0?
What is the intercept if the dummy variable equals 1?
What is the slope if the dummy variable equals 0?
What is the slope if the dummy variable equals 1?
How large is the difference in intercepts between the two groups?
Where can we find this value in the equation?
How large is the difference in slopes between the two groups?
Where can we find this value in the equation?

We have the following regression equation, with y as dependent variable, x as a continuous predictor variable, and a dummy variable *dummy*.

$$y = -4.1 + 1.2 \times x - 6.5 \times dummy - 1.3 \times x \times dummy + e$$

Write down the regression equation in the case the dummy variable equals 0.
Write down the regression equation in the case the dummy variable equals 1.
What is the intercept if the dummy variable equals 0?
What is the intercept if the dummy variable equals 1?

What is the slope if the dummy variable equals 0?
What is the slope if the dummy variable equals 1?
How large is the difference in intercepts between the two groups?
Where can we find this value in the equation?
How large is the difference in slopes between the two groups?
Where can we find this value in the equation?

Suppose we find the following linear equation:

$$mathscore = 16.3 + 5.5 \times age - 0.8 \times sex - 1.2 \times age \times sex + e$$

What is the main effect of *age* on mathscore?
What is the main effect of the *sex* on mathscore?
How large is the interaction effect of *age* and *sex* on mathscore?
What is the predicted mathscore for a girl of age 12, if sex is coded 1 for boys?
What is the predicted mathscore for a boy of age 22, if sex is coded 1 for boys?

7.2 Interaction with two dummy variables

In the previous section we discussed the situation that regression slopes might be different in two groups. Now we discuss the situation that we have two dummy variables, and that we're interested whether there is an interaction effect. In other words, does one dummy variable moderate the effect of the other dummy variable?

Suppose in country A, men are on average taller than women. In order to study this effect, we analyze data from a random sample of inhabitants, and we come up with the following regression equation:

$$height = 165 + 10 \times sex + e$$

In this equation, sex is coded 0 for females, and 1 for males. So, the predicted height for a female from country A equals 165 and the predicted height for a male equals $165 + 10 \times 1 = 175$.

Suppose we also study height in country B. Again with a random sample of inhabitants, we find the following regression equation:

$$height = 175 + 15 \times sex + e$$

In this equation, the predicted height for a female from country B equals 175

and the predicted height for a male equals $175 + 15 \times 1 = 190$.

So it seems that in general, the people in the random sample from country B are taller than the people in the random sample from country A: both men and women show taller averages in country B. But we also see another difference between the two countries: the average difference between men and women is 10 cm in country A, but 15 cm in country B. So we can say that in these samples, the effect of sex on height is a little bit different in both countries. Now of course this difference could be a coincidence, a random result from sampling, or it could be a real thing in the populations. Suppose we'd like to know whether the effect of sex on height is different in the two countries at population level. We'd like to know whether country is a moderator of the effect of age on height. So we use the following regression equation:

$$height = b_0 + b_1 \times sex + b_2 \times country + b_3 \times sex \times country + e$$

and perform a regression equation. We *could* use the same SPSS syntax as in the previous section, making dummy variables ourselves and analysing them quantitatively using the WITH syntax:

```
RECODE country ('A'=0) ('B'=1) INTO country.
RECODE sex ('female'=0) ('male'=1) INTO sex.
EXECUTE.
UNIANOVA height WITH sex country
/ design = sex country sex*country
/ print = parameter.
```

However, the easier option, as we have seen in the previous section, is to let SPSS do the dummy coding. Simply omit the RECODE lines and use the BY syntax to indicate that you want to use country and sex in a qualitative way using dummy coding:

```
UNIANOVA height BY sex country
/ design = sex country sex*country
/ print = parameter.
```

From now on, we recommond using the BY syntax for variables that you wish to analyze qualitatively (all categorical variables, and sometimes ordinal variables).

In the output we find the following values:

$$height = 165 + 10 \times sex + 10 \times country + 5 \times sex \times country + e$$

So the predicted value for specific subgroups are the following:

Sex	Country	equation	predicted height
Female	A	$165 + 10 \times 0 + 10 \times 0 + 5 \times 0 \times 0$	165
Male	A	$165 + 10 \times 1 + 10 \times 0 + 5 \times 1 \times 0$	175
Female	B	$165 + 10 \times 0 + 10 \times 1 + 5 \times 0 \times 1$	175
Male	B	$165 + 10 \times 1 + 10 \times 1 + 5 \times 1 \times 1$	190

Note that we see exactly the same predicted values for the subgroups as we saw in the separate analyses for countries A and B. The interaction effect in this example is equal to 5: it means that the effect of sex (being a male) on height is 5 cm larger in country A than in country B. See that the difference in height between males and females is 10 cm in country A and 15 cm in country B. So the difference in the differences equals 5 cm. But note that you can also look at it from another angle: the difference between country A and B equals 10 cm for females, and 15 cm for males. So you can equally say that Sex moderates the effect of country: the effect of country is larger for males than for females, and this difference is again 5 cm.
Whether the interaction effect also exists at the population level, we can see from SPSS output. If the effect is significant, we conclude that the difference between males and females in height is different in two countries. Or, equivalently, we conclude that the difference in height between the two countries is different for males and famales. If the effect is not significant, we conclude that that the difference in height between females and males is the same in country A and B. Or, equivalently, we conclude that the difference in height between the two countries is the same for males and females.

7.3 More than two groups

Now what happens is we have categorical variables with more than two levels? Suppose we want to do the same study on height but now in countries A, B and C. As we saw earlier, in SPSS we can treat variables in a regression analysis either as quantitative or qualitative. If we want to treat variable as quantitative, we use the word WITH, and if we want to treat the variable as qualitative, we use the word BY in the SPSS syntax. For dummy variables, both options are possible, but we generally recommend using the WITH word. When you have a variable with more than two levels, say country with three levels, we generally recommend using the BY word. This makes SPSS turn the categorical variable into two dummy variables automatically. In general, if you have K levels in a categorical variable, SPSS computes $K - 1$ dummy variables.

Suppose you have the categorical variable country with levels A, B and C, and you have the sex variable dummy coded as 1 for males and 0 for females. You want to treat the dummy variable quantitatively, and the country variable

qualitatively. Then with the next syntax you can run a regression analysis with a main effect of sex, a main effect of country and an interaction effect of sex by country in the following way.

```
UNIANOVA height BY country WITH sex
/ design = sex country sex*country
/ print = parameter.
```

The SPSS output might look something like this:

Between-Subjects Factors

		N
country	A	10
	B	10
	C	10

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1712.167[a]	5	342.433	21.358	.000
Intercept	439984.067	1	439984.067	27441.834	.000
sex	410.700	1	410.700	25.615	.000
country	298.133	2	149.067	9.297	.001
country * sex	421.400	2	210.700	13.141	.000
Error	384.800	24	16.033		
Total	920497.000	30			
Corrected Total	2096.967	29			

a. R Squared = .816 (Adjusted R Squared = .778)

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	173.800	1.791	97.056	.000	170.104	177.496
sex	-2.800	2.532	-1.106	.280	-8.027	2.427
[country=A]	-8.800	2.532	-3.475	.002	-14.027	-3.573
[country=B]	1.200	2.532	.474	.640	-4.027	6.427
[country=C]	0[a]	.	.	.	.	.
[country=A] * sex	12.800	3.581	3.574	.002	5.408	20.192
[country=B] * sex	17.800	3.581	4.970	.000	10.408	25.192
[country=C] * sex	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Here we see that 2 dummy variables have been computed, automatically by SPSS. One for being in country A, and one for being in country B. Country C is here used as the so-called reference category. This SPSS output is therefore equivalent to the equation:

$$height = 173.8 - 2.8 \times sex - 8.8 \times CountryA + 1.2 \times CountryB$$

$$+ \quad 12.8 \times CountryA \times sex + 17.8 \times CountryB \times sex + e$$

All observations done in country C for variables CountryA and CountryB are coded as 0. So let's do the math to get the predicted heights for each subgroup. Females are coded as 0 and males as 1, so a Female from country C gets the predicted value 173.8. Let's do the computations for all subgroups:

Sex	Country	equation	height
Female	A	$173.8 - 2.8 \times 0 - 8.8 \times 1 + 1.2 \times 0 + 12.8 \times 1 \times 0 + 17.8 \times 0 \times 0$	165
Male	A	$173.8 - 2.8 \times 1 - 8.8 \times 1 + 1.2 \times 0 + 12.8 \times 1 \times 1 + 17.8 \times 0 \times 1$	175
Female	B	$173.8 - 2.8 \times 0 - 8.8 \times 0 + 1.2 \times 1 + 12.8 \times 0 \times 0 + 17.8 \times 1 \times 0$	175
Male	B	$173.8 - 2.8 \times 1 - 8.8 \times 0 + 1.2 \times 1 + 12.8 \times 0 \times 1 + 17.8 \times 1 \times 1$	190
Female	C	$173.8 - 2.8 \times 0 - 8.8 \times 0 + 1.2 \times 0 + 12.8 \times 0 \times 0 + 17.8 \times 0 \times 0$	173.8
Male	C	$173.8 - 2.8 \times 1 - 8.8 \times 0 + 1.2 \times 0 + 12.8 \times 0 \times 1 + 17.8 \times 0 \times 1$	171

Note that we now have very different values for the regression parameters than in the analysis with only countries A and B (see Table ...), but nevertheless we end up with the same expected heights in Countries A and B. The difference in the parameter values stems from the fact that we have now treated country C as the reference category (dummy variable equal to 0), whereas in the previous two country analysis, we treated country A as the reference category (dummy equal to 0). In the output we see that the CountryA by sex interaction effect is significant: there is an extra height of 12.8 cms seen in males from country A, over and above the main effects of being male in general and being from country A. In other words, the effect of being male is larger in country A than it is in Country C (the reference country). We also see this in the predicted means: male-female difference in country C is -2.8 (males shorter), but in country A it is +10 (males larger). In the output we also see that the CountryB by sex interaction effect is significant: the effect of being male is 17.8 cm larger in country B than in Country C (the reference category). From the means we see that the male-female difference is 15 in country B, which is 17.8 cm more than the -2.8 in country C. Both these effects are significant. Moreover, from the ANOVA table (Tests of Between-Subjects Effects) we see that these two interaction effects overall are significantly different from 0. So we conclude that in the populations of countries A, B and C, the difference in height between males and females are significantly different, $F(2, 24) = 13.141, MSE = 210.70, p < 0.05$.
Alternatively, but equivalently, we may conclude that the differences in height across the three countries, are significantly different for males than for females, $F(2, 24) = 13.141, MSE = 210.70, p < 0.05$.

Exercises

From a sample of data on height, country, and weight, we get the following linear equation:

$$\widehat{weight} = 40 + 30 \times CountryA + 0.4 \times height + 0.1 \times CountryA \times height$$

1. What is the expected weight for an individual from country A with a height of 1.5?

2. What is the expected weight for an individual from country B with a height of 1.0?

3. How large is the slope coefficient of height in country A?

4. How large is slope coefficient of height in country B?

Answers

1. $$\widehat{weight} = 40 + 30 \times 1 + 0.4 \times 1.5 + 0.1 \times 1 \times 1.5 = 70.75$$

2. $$\widehat{weight} = 40 + 30 \times 0 + 0.4 \times 1.0 + 0.1 \times 0 \times 1.0 = 40.4$$

3. $0.4 + 0.1 = 0.5$

4. 0.4

7.4 Linear by linear interaction

Chapter 8

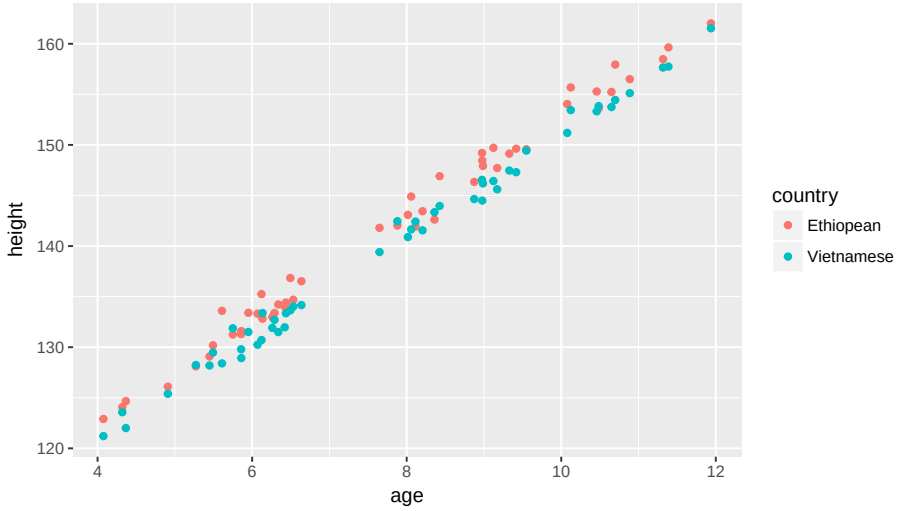
Assumptions of linear models RUYA

8.0.1 Independence

The assumption of independence is about the way in which observatins are similar and dissimilar. Take for instance the following regression equation for children's height predicted by their age:

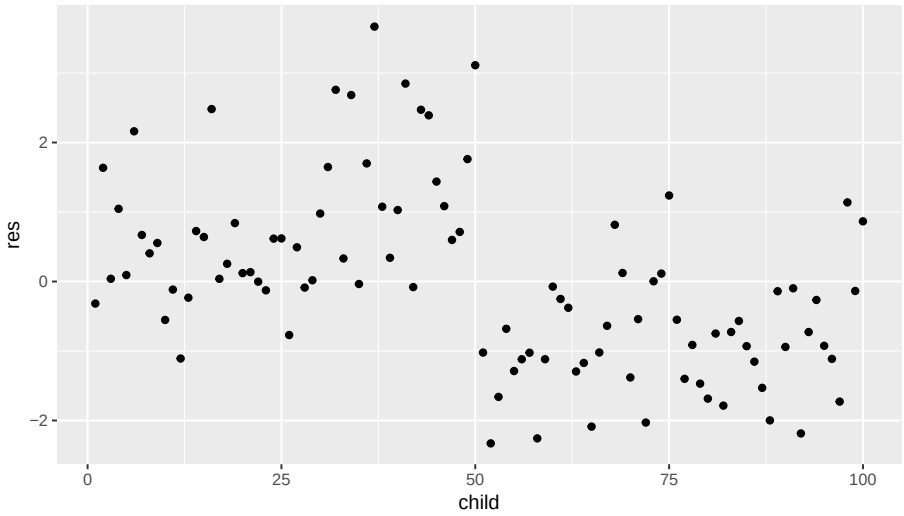
$$height = 100 + 5 \times age + e \tag{8.1}$$

This regression equation predicts that a child of age 5 has a height of 125 and a child of age 10 has a height of 150. In fact, all children of age 5 have the same predicted height of 125 and all children of age 10 have the same predicted height of 150. Of course, in reality, children of the same age will have very different heights: they differ. According to the above regression equation, children are similar in height because they have the same height, but they differ because of the random term e that has a normal distribution: predictor age makes them similar, residual e makes them dissimilar. Now, if this is all there is, then this is a good model. But let's suppose that we're studying height in an international group of 50 Ethiopian children and 50 Vietnamese children. Let's plot their heights:




```
##
## Call:
## lm(formula = height ~ age + country + age * country, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9692 -0.7366 -0.1769  0.6061  2.8313
##
## Coefficients:
##                         Estimate Std. Error t value            Pr(>|t|)
## (Intercept)           102.57196    0.53708 190.981 <0.0000000000000002 ***
## age                     5.02598    0.06686  75.168 <0.0000000000000002 ***
## countryVietnamese      -1.57419    0.75954  -2.073              0.0409 *
## age:countryVietnamese  -0.01771    0.09456  -0.187              0.8518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.988 on 96 degrees of freedom
## Multiple R-squared:  0.9916,Adjusted R-squared:  0.9913
## F-statistic:  3779 on 3 and 96 DF,  p-value: < 0.00000000000000022
```

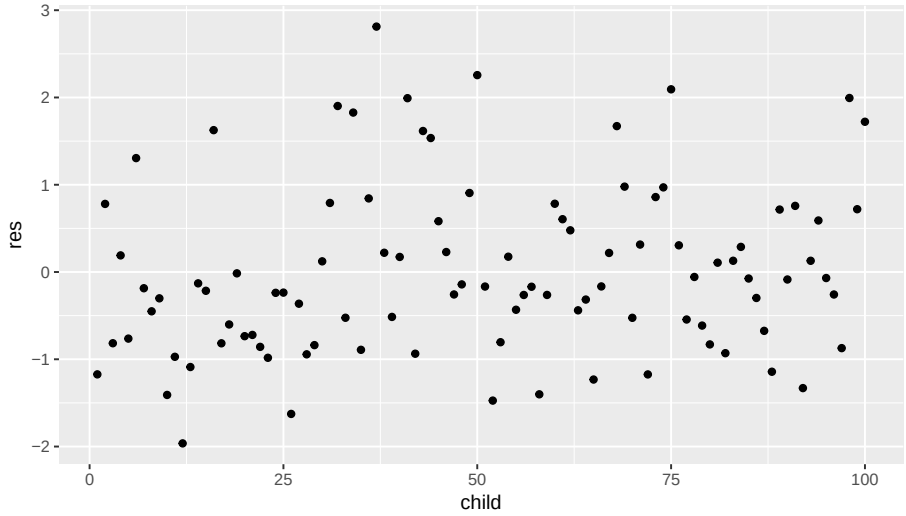
From this graph, we see that heights are similar because of age: older children are taller than younger children. But we also see that children are similar because of their national background: Ehtiopian children are systematically taller than Vietnamese children, irrespective of age. So here we see that a simple regression of height on age is not a good model. We see that when we estimate the simple regression on age and look at the residuals:



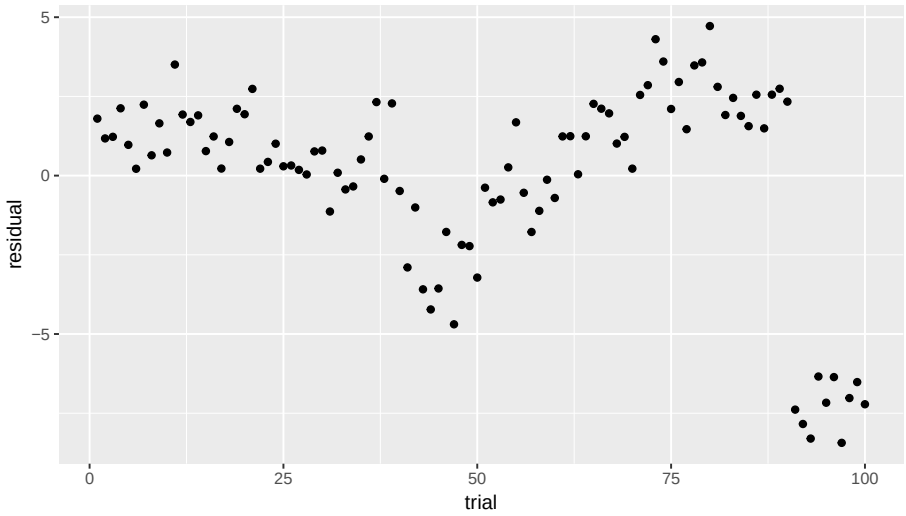
As our model predicts random residuals, we expect a random scatter of residuals. However, what we see here is a systematic order in the residuals: they tend to be positive for the first 50 children and negative for the last 50 children. These turn out to be the Ethiopean and the Vietnamese children, respectively. This systematic order in the residuals is a violations of independence: the residuals should be random, and they are not. The residuals are dependent on country: positive for Ethiopeans, negative for Vietnamese children. Thus, there is more than just age that makes children similar. If we use multiple regression, including both age and country, we get the following regression equation:

$$height = 102.641 + 5.017 \times age - 1.712 \times country + e \tag{8.2}$$

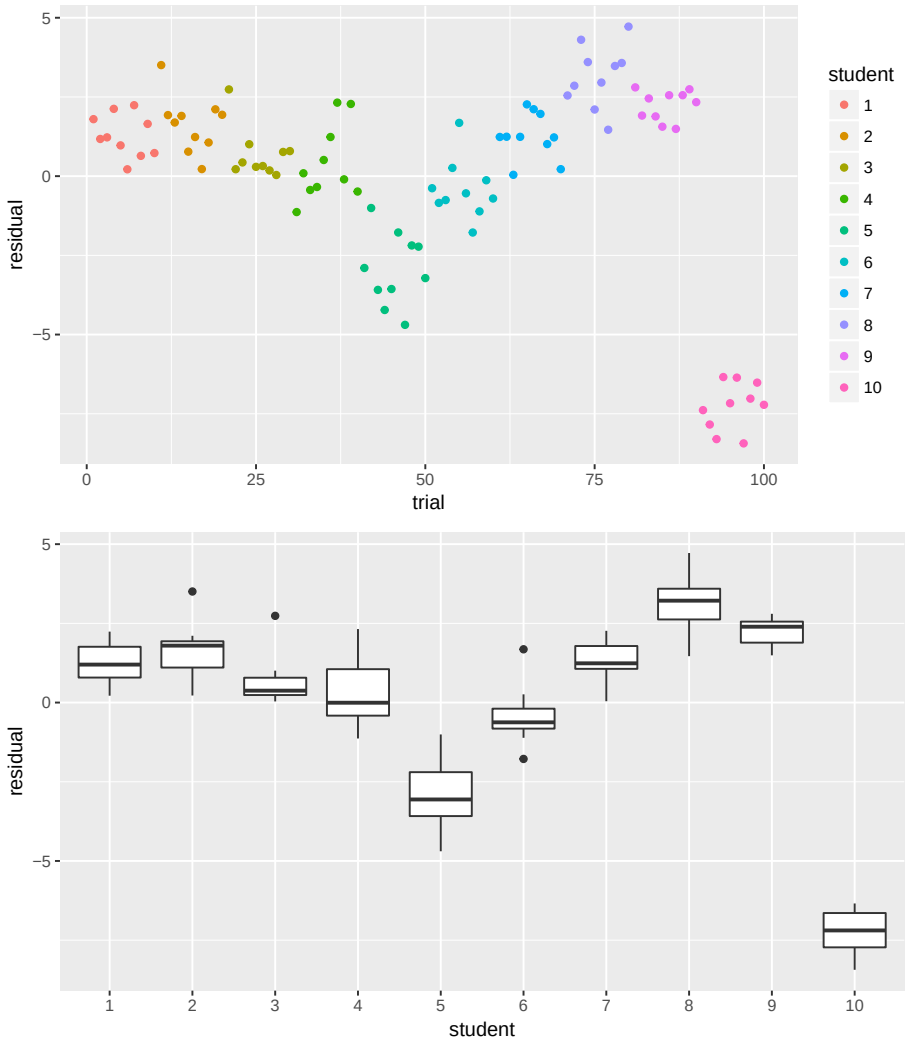
When we now plot the residuals we get a nice random scatter:



Another typical example of non random scatter of residuals is the following:



They come from an analysis of reaction times, done on 10 students where we also measured their IQ. Each student was measured on 10 trials. We predicted reaction time on the basis of student's IQ using a simple regression analysis. The residuals are clearly not random, and if we look more closely, we see some clustering if we give different colours for the data from the different students:



We see that residuals that are close toghether come from the same student. So, reaction time are not only similar because of IQ, but also because they come from the same student: clearly something else other than IQ explains why reaction times are dissimilar across individuals. The residuals in this analysis are not independent given IQ, they are dependent on the student. Thus, the assumption of independently distributed residuals is violated.

8.0.2 Linearity (additivity)

8.0.3 Homogeneity of variance

8.0.4 Residuals normally distributed

8.1 Testing assumptions

8.1.1 Independence

Check the residuals, do you see any regularities? More specifically: Do you see any clustering?

8.1.2 Linearity (additivity)

8.1.3 Homogeneity of variance

8.1.4 Residuals normally distributed

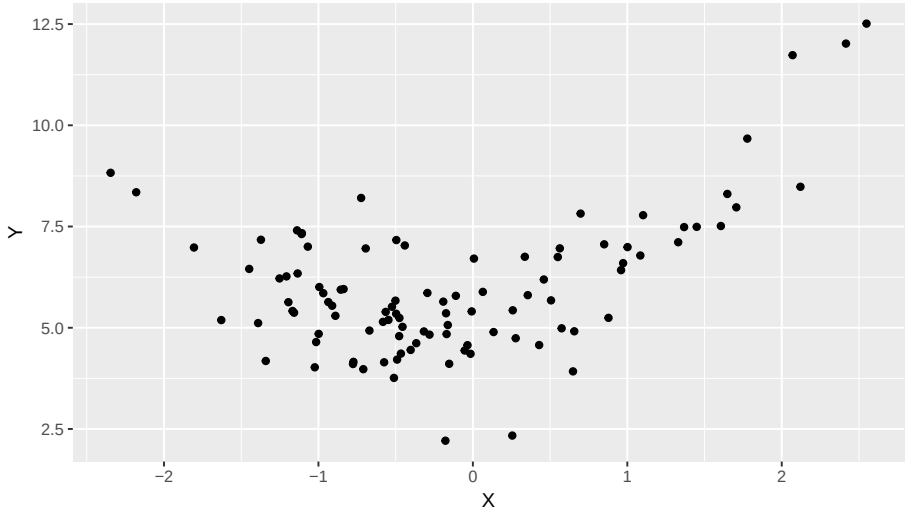
Make a histogram of the residuals: is the shape symmetric or skew? Is it more or less bell-shaped? Note that for small data sets, the shape is never perfectly normal. It should however have one peak and be symmetric.

8.1.5 What to do when assumptions are violated?

When the assumption of independence is violated, try including more predictors in your regression model. Like in the example of height in children, adding country into the equation solved the problem. In the case of reaction times, it was not clear what explains the clustering of residuals: we only know that reaction times from the same person were very similar. In such situations, consider linear *mixed* models, to be discussed in a later chapter, and include a so-called *random factor*, in this case for students.

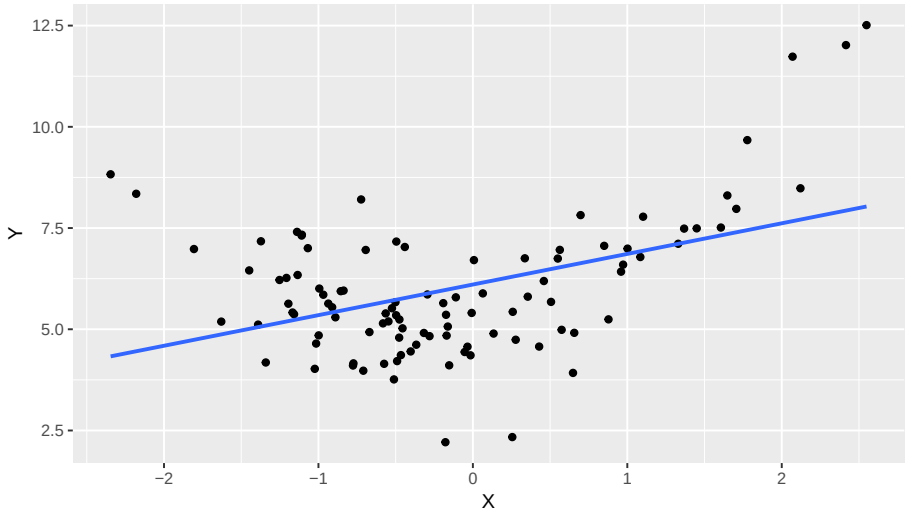
8.1.6 nonlinearity

If we have data and we analyze these with a linear (mixed) model, we can find nonlinearity. In that case we might introduce a quadratic term to make it more linear. For example, suppose we have the following data set, summarized in a scatter plot:



We might then fit the following linear model, and find that it does not fit the data well:

$$y = b_0 + b_1 X + e \tag{8.3}$$



A better model might be gained by introducing a new variable $X2$ that is computed by multiplying variable X by itself: $X2 = X^2$, and use this as an extra predictor:

$$y = b_0 + b_1 X + b_2 X2 + e \tag{8.4}$$

Chapter 9

Linear modelling: more advanced topics

9.0.1 Planned comparisons

Suppose you have height data from three countries: Greece, Italy, and Norway. You might wish to know whether in these populations there is a difference in average height. If that is all you want to know, you can perform the SPSS UNIANOVA analysis described above. In that case, the null-hypothesis is

$H_0 : \mu_{Greece} = \mu_{Italy} = \mu_{Norway}$

However, suppose your most important hypothesis is really much more specific: you only want to know whether the average height in Italy is different from the average height in Greece. The corresponding hypothesis would then be:

$H_0 : \mu_{Greece} = \mu_{Italy}$

In such cases, where the null-hypothesis is more specific than simply stating "there are differences among the groups", then you should perform *planned comparisons*. Here you would like to make a comparison between the average heights of Greece and Italy You could also say you'd like to *contrast* the average height of Greece with that of Italy.

We then have to define this contrast in such a way that SPSS knows what we want. So we could define our contrast in a similar vein as the null-hypothesis. Let's call the contrast γ_1.

$\gamma_1 : \mu_{Greece} = \mu_{Italy}$

This contrast could also be written such that there is zero on the right-hand

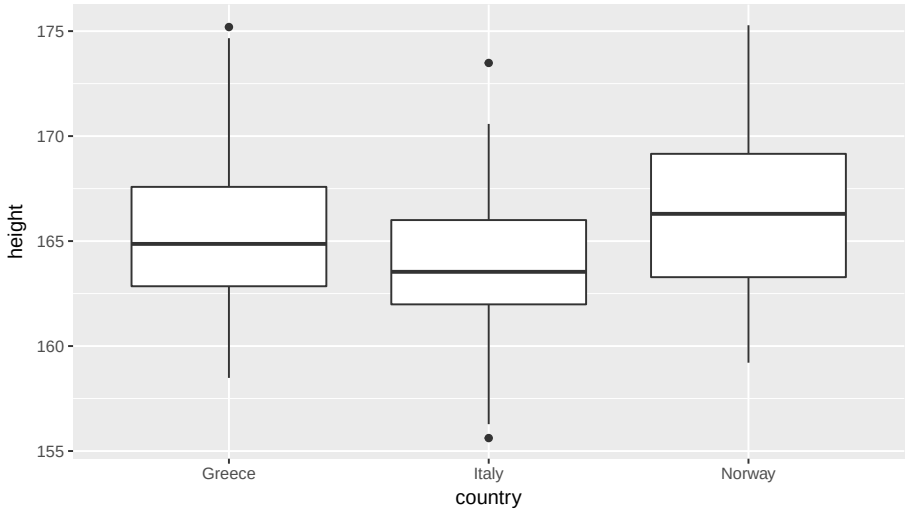
side of the equation, like this:

$\gamma_1 : \mu_{Greece} - \mu_{Italy} = 0$

This is the preferred way of specifying contrasts: having a zero on the right-hand side. And how about Norway? How do we add Norway into this contrast? Well, notice that you could also write the contrast like this:

$\gamma_1 : (1) \times \mu_{Greece} + (-1) \times \mu_{Italy} + (0) \times \mu_{Norway} = 0$

So we could code this specific contrast with the numbers preceding the group means: A 1 for Italy, a -1 for Greece and a 0 for Norway. For SPSS therefore, to specify a contrast, we only need this coding. This can be done in the following way. Suppose you have the following data on height, summarized in a boxplot:



Note that in this SPSS data set, the country variable is coded 1 for Greece, 2 for Italy and 3 for Norway. Then we could use the following syntax to ask for the specific comparison (or contrast) of the first and the second group, that is, Greece and Italy, respectively. So group 1 (=Greece) gets a 1, group 2 (=Italy) gets a -1, and group 3 (=Norway) gets a 0, so our coded contrast looks like $(1, -1, 0)$.

```
UNIANOVA height BY country
/DESIGN=country
/CONTRAST(country)=SPECIAL(1  -1  0).
```

Below we see the output:

From the output, we see first see that the general null-hypothesis that all three countries have the same average height is rejected, $F(2, 147) = 6.14, MSE = 14.77, p = 0.003$. This test has 2 degrees of freedom, one for each of the two

country	1.00	Greece	50
	2.00	Italy	50
	3.00	Norway	50

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	181.382[a]	2	90.691	6.142	.003
Intercept	4097387.37	1	4097387.37	277508.194	.000
country	181.382	2	90.691	6.142	.003
Error	2170.444	147	14.765		
Total	4099739.20	150			
Corrected Total	2351.826	149			

a. R Squared = .077 (Adjusted R Squared = .065)

Contrast Results (K Matrix)

country Special Contrast			Dependent Variable: height
L1	Contrast Estimate		1.008
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		1.008
	Std. Error		.769
	Sig.		.192
	95% Confidence Interval for Difference	Lower Bound	-.511
		Upper Bound	2.527

Test Results

Dependent Variable: height

Source	Sum of Squares	df	Mean Square	F	Sig.
Contrast	25.406	1	25.406	1.721	.192
Error	2170.444	147	14.765		

dummy variables that are needed to test this model. The error degrees of freedom equals 147: that is the number of data points (150) minus the number of parameters in the model: one for the intercept and two for the dummy variables, so $150 - 3 = 147$.

Next, we see the results for our *specific* null-hypothesis (the contrast): that the means for Greece and Italy are equal, irrespective of Norway. We see a Contrast Estimate of 1.008 and a significance level of $p = 0.19$. So the contrast is positive: what does this mean? Well, let's put the 1.008 into our contrast above:

$$\gamma_1 : (1) \times \mu_{Greece} + (-1) \times \mu_{Italy} + (0) \times \mu_{Norway} = 1.008$$

which can be simplified to

$$\gamma_1 : \mu_{Greece} - \mu_{Italy} = 1.008$$

So the 1.008 indicates that the average height in Italy is 1.008 cm shorter than

in Greece, at least in our sample. The relatively high p-value indicates that the population means are however not different: we do not reject the null-hypothesis that the average heights are the same. We also see an F-test for this contrast, $F(1, 147) = 1.721, MSE = 14.76, p = 0.19$. Since the first degrees of freedom number is a 1, we know that F is the same as a squared t, t^2, So, an equivalent presentation of the contrast effect would be $t(147) = 2.96, p = 0.19$

Now perhaps you realize, why didn't we do a t-test in the first place? We might have taken the Greece and Italy data separately, used a dummy variable and then run an ordinary linear model. Well, let's see what would happen then. First we let SPSS select only the Greece and Italy data with the SELECT IF syntax, and next run an UNIANOVA, using the BY keyword for country to indicate that we treat it as a qualitative variable. We also let the selection be preceded by a TEMPORARY commmand, to indicate that the selection only applies to the analysis that follows.

```
TEMPORARY.
SELECT IF (country<3).
UNIANOVA height BY country
/DESIGN=country
/PRINT = PARAMETER.
```

Below we see the output:

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
country	1.00	Greece	50
	2.00	Italy	50

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	25.406[a]	1	25.406	1.880	.173
Intercept	2707809.07	1	2707809.07	200384.457	.000
country	25.406	1	25.406	1.880	.173
Error	1324.281	98	13.513		
Total	2709158.76	100			
Corrected Total	1349.687	99			

a. R Squared = .019 (Adjusted R Squared = .009)

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	164.050	.520	315.562	.000	163.019	165.082
[country=1.00]	1.008	.735	1.371	.173	-.451	2.467
[country=2.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

We see that now only the data from Greece and Italy are compared. From the Parameter Estimates table, we see that country 2 is the reference category, which is Italy. We also see that the average height in Greece is 1.008 cm taller

than Italy. This effect is not significant, $t(98) = 1.371, p = 0.17$. We use 98 degrees of freedom because we have 100 people and loose 2 degrees of freedom, one for the intercept and once for the regression coefficient for country=1. Note that this number is also indicated in the Tests of Between-Subjects Effects table, as the df of the Error.

So what is different from the contrast effect? First of all, we see that the p-value is different: the p-value for the contrast effect analysis is somewhat larger than the one for the 'ordinary' t-test. Second, we see that the t-statistic is different: In the contrast analysis the t-value was 2.96 and in the ordinary analysis it was 1.371. So part of the reason that the p-value is different for the contrast is that the t-value is higher. Third, we see that the degrees of freedom has changed from 147 to 98. Well, we know that the significance level of a t-value depends on its size, the higher the t-value the lower the p-value, and we know that significance also depends on the degrees of freedom: the more degrees of freedom, the lower the p-value.

So why has the t-value changed? Well, we know that the t-value is nothing but the regression coefficient divided by its standard error, $t = B/SE$. If we compare the two outputs, we see that the effect of B is the same, the regression coefficient is equal to 1.008, and the contrast estimate in the Contrast Results (K-Matrix) is 1.008. So the sizes are the same.

So the only difference between the two analyses can be the standard error. It turns out the standard error is computed in different ways: in the first contrast analysis, this SE is computed on the basis of *all* the data, including the Norway data (150 people). We can see that from the degrees of freedom, the number of people minus 3. Why three? Because we also compute the variance of the Norway data, for which we have to estimate the mean (variance is the average squared difference from the mean). In the second analysis, we only used the data from 100 Greeks and Italians. The Norway data was not used at all, that's why the degrees of freedom is different and also the standard error for our t-statistic.

Thus in summary: if we have a very specific hypothesis about the difference in means among two groups, it's better to use a contrast analysis, rather than a simple analysis regarding only those two groups. The reason for this is that *more information is used*, even from groups for which we have no specific hypothesis.
Note that here we saw a higher p-value for the contrast analysis, but still, in general it is wiser to use as much data as possible, so we prefer a contrast analysis over a simple analysis excluding other groups.

The above example of comparing only the means of two groups, at the same time making use of the data in other groups, is called a *simple contrast*. Now let's look at a *complex contrast*. Suppose your hypothesis is that average height is different in Northern-European countries than in Southern-European countries, then you would like to know whether Italy AND Greece taken together differ regarding the average height from Norway. So we'd like to compare the

average height in Norway to the average height in Greece and Italy together. We could write that null-hypothesis as follows:

$$H_0 : \frac{\mu_{Greece} + \mu_{Italy}}{2} = \mu_{Norway} \tag{9.1}$$

In other words, the mean height in Norway is equal to the *average* mean height in the two other countries. If we want to test this hypothesis, we have to define a contrast. So we get the zero on the right-hand side in the following way:

$$\gamma_2 : \frac{\mu_{Greece} + \mu_{Italy}}{2} - \mu_{Norway} = 0 \tag{9.2}$$

This in turn we can write as

$$\gamma_2 : 0.5\mu_{Greece} + 0.5\mu_{Italy} - 1\mu_{Norway} = 0 \tag{9.3}$$

In SPSS, we can code this in the following way $(\frac{1}{2}\frac{1}{2} - 1)$[1]. We use the following syntax:

```
UNIANOVA height BY country
/DESIGN=country
/CONTRAST(country)=SPECIAL(0.5  0.5  -1).
```

Now let's compare this analysis where we use a dummy variable NrthrnEr, where we code 1s for Norwegians and 0s for Greeks and Italians. We then run an ordinary linear model on the data using this dummy that is a quantitative predictor of height:

```
UNIANOVA height WITH NrthrnEr
/PRINT=PARAMETER
/DESIGN=NrthrnEr.
```

From the contrast analysis we obtain $F(1, 147) = 10.56, MSE = 14.77, p = 0.001$, which is equivalent to $t(147) = 3.25, p = 0.001$. From the dummy variable analysis, we obtain $t(148) = 3.24, p = 0.001$. So in this case, the results are very close: the degrees of freedom is larger for the dummy variable analysis, but the t-value is lower. All in all, we gain nothing much, and that is because in the dummy analysis we also use all of the data: we put the Greeks and Italians in one group (N=100), and the Norwegians in another group (N=50). The degrees of freedom therefore differs cause we only need to estimate 2 means in the dummy analysis, instead of 3 in the contrast analysis.

[1] Note that for the SPSS variable country variable, the first group is Greece, the second is Italy and the third is Norway. The contrast code should reflect the same order.

In general: if you run a model where you compare various groups, AND you have very specific hypotheses that you'd like to test, it is generally advised to run a contrast analysis, including a small number of them, and then report only the tests and p-values of those contrast. Do not then also report the p-values of the parameters of your model. Actually, your contrasts are a respecification of your model: either report the contrasts or the parameters, but not both, since they contain the same information. If you report too many p-values, the probability that you make a Type I error (concluding that you have a significant difference while there is really no difference) becomes too large. That is also the reason why you should report not more contrasts than the number of your parameters for your variable. For example, if you compare 5 groups, you will have 4 parameters for these groups. In that case specify no more than 4 contrasts. SPSS has a number of prespecified sets of contrasts, like Helmert, Deviation, Difference, etcetera. Check out the SPSS manual for more details. If you want something more specific, use the SPECIAL option as indicated above.

Also very important: only go for contrasts when the overall ANOVA is significant. If the group means are not significantly different from eachother in a general sense, it is generally not advised to test specific contrasts.

9.0.2 Testing more than one contrast

In some cases you have a number of research hypotheses about group differences. For instance, you might have the a priori hypothesis that people in Northern countries are taller than in Southern countries, and another a priori hypothesis that people in Western countries are taller than in Eastern countries. So, a priori you have the hypothesis that the mean height in Norway is different from the mean heights in Greece and Italy combined. Second, you expect that the mean height in Italy is higher than in Greece.

You could test these two hypotheses at once in SPSS by specifying a K matrix, like so:

```
UNIANOVA height BY country
/DESIGN=country
/CONTRAST(country)=SPECIAL(-0.5  -0.5  1
                           1    -1   0).
```

Remember that the countries were coded like 1=Greece, 2=Italy, and 3= Norway. So the first null-hypothesis that is tested is that Norway has the same mean as the average of Greece and Italy. The second null-hypothesis is that Greece and Italy have the same mean.

This set of contrasts is said to be *orthogonal*: whether or not we find a significant result for the first contrast has nothing to do with whether we find a significant result for the second contrast. Why this is the case can be seen from the K matrix: if we take the first elements of the first and second row and multiply them we get $-0.5 \times 1 = -0.5$. If we take the second elements of the first and second row and multiply them we get $-0.5 \times -1 = 0.5$. If we

take the third elements of the first and second row and multiply them we get $1 \times 0 = 0$. If we add these numbers we get $-0.5 + 0.5 + 0 = 0$. Here we get a total of 0, which indicates that the contrasts are orthogonal, implying that the statistical results for contrast 1 and 2 are independent of eachother. If the sum is unequal to 0, the contrasts are said to be *dependent*. Here's an example of a non-orthogonal set of contrasts:

```
UNIANOVA height BY country
/DESIGN=country
/CONTRAST(country)=SPECIAL( 1    0   -1
                            1   -1   0).
```

Here the sum of the products equals $1 \times 1 + 0 \times -1 + -1 \times 0 = 1$. This means that the set of statistical results is not independent of each other, so if the null-hypothesis if the first contrast is significant, this yields some information about the probablity of obtaining a significant result for the second contrast. This you do not want, of course. So generally you would want to use independent sets of contrast. However, research questions are always more important: if you have good theoretical reason to specify a set of non-orthogonal contrast, just go for it (Stevens,).

As stated earlier, SPSS has a number of pre-specified sets of contrasts. One of them is the Helmert set of contrasts. In Helmert contrasts, the first group is contrasted with the average of all later groups, the second group is compared to the average of the later groups (ignoring group 1), the third group is compared with the average of the later groups (ignoring groups 1 and 2), etcetera. For a five country analysis, the syntax would be like

```
UNIANOVA height BY country
/DESIGN=country
/CONTRAST(country)=SPECIAL(1     -0.25 -0.25 -0.25 -0.25
                           0 1 -0.33 -0.33 -0.33
                           0 0 1 -0.5 -0.5
                           0 0 0 1 -1).
```

and this is equivalent to the syntax

```
UNIANOVA height BY country
/DESIGN=country
/CONTRAST(country)=HELMERT.
```

This set of 4 contrasts is also completely orthogonal: all pairs of contrasts are orthogonal.

9.0.3 Post-hoc comparisons

In some cases, you compare 3 or more groups, and you find some interesting differences. For instance, in the above example, when you look at the boxplot

of the differences between Greece, Italy and Norway, you might wonder whether there is a real difference in Italy and Greece. Or perhaps there is a difference between Norway and Italy, or even between Norway and Greece. There might be all kinds of interesting things to find out from these data. In this case note, we now formulate these hypotheses *after* looking at the difference in our data. In this case, suppose that there was no specific hypothesis before collecting our data, and we merely wanted to find out whether there are differences between mean heights across these countries. So, our null-hypothesis before looking at the data was that there were no differences in mean height across Greek, Italian and Norwegian populations. To test this we perform a regular linear model analysis, with height as the dependent variable and a categorical variable country as independent variable. We want SPSS to make dummy variables automatically, so we use the following syntax using BY:

```
UNIANOVA height BY country
/PRINT=PARAMETER
/DESIGN=country.
```

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	181.382 [a]	2	90.691	6.142	.003
Intercept	4097387.37	1	4097387.37	277508.194	.000
country	181.382	2	90.691	6.142	.003
Error	2170.444	147	14.765		
Total	4099739.20	150			
Corrected Total	2351.826	149			

a. R Squared = .077 (Adjusted R Squared = .065)

From the output we use the F-test to test the overall hypothesis about equality of means, and report a significant difference between the three means, $F(2, 147) = 6.14, MSE = 14.77, p < 0.05$. Now *given* that we have this rejection of the null-hypothesis, we might be very interested where this significance comes from: is it that Norway is very different from the other two countries? Perhaps there are no differences between Greece and Italy? Perhaps there is only a real difference between Norway and Italy, but no real difference between Greece and Italy. And so on, and so forth. Note that here you could make 3 pair-wise simple comparisons: Greece vs Italy, Greece vs Norway and Italy vs Norway. You or anyone else intested in your research would like to know if these pair-wise differences are significant. In that case you can report so-called *post hoc* pairwise comparisons.

Note that we do not perform planned comparisons using contrasts here. Planned comparisons are very powerful tools that are only allowed for hypotheses that are specificied *a priori*, that is, *before* doing any analysis, and preferably before any data collection. *Post hoc* comparisons are done *after* the fact: after testing the research hypothesis and after having looked at the data (seeing means or boxplots!), you can test extra hypotheses that are of secondary interest. First we will show you how to do it, and second we will explain why we do it like that.

```
UNIANOVA height BY country
/PRINT=PARAMETER
/DESIGN=country
/POSTHOC=country(BONFERRONI).
```

So we state that we wish to see posthoc comparisons for the variable *country*. In parentheses we indicate *Bonferroni*, which we will explain later.

Bonferroni

(I) country	(J) country	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Greece	Italy	1.0081	.76850	.575	-.8530	2.8692
	Norway	-1.6591	.76850	.097	-3.5202	.2020
Italy	Greece	-1.0081	.76850	.575	-2.8692	.8530
	Norway	-2.6672*	.76850	.002	-4.5283	-.8061
Norway	Greece	1.6591	.76850	.097	-.2020	3.5202
	Italy	2.6672*	.76850	.002	.8061	4.5283

Based on observed means.
The error term is Mean Square(Error) = 14.765.
*. The mean difference is significant at the 0

From the figure we see that there are 6 comparisons, but by closer inspection we see that all three possible comparisons are reported twice. We see a significant difference between Italy and Norway, $p < 0.002$, and that the other two comparisons are not significant. So average height in Greece is not different from the average height in Italy, $p > 0.575$, nor from that in Norway, $p > 0.097$. Now note the difference from the Planned comparisons analysis. There we found a p-value of 0.19 for the hypothesis that Greece has the same average height as Italy. This contrast had a somewhat higher p-value than a simple dummy analysis for Greece and Italy, ignoring the Norwegian data, $p = 0.17$. And now we see a much higher p-value of 0.58. The reason is that a correction has been applied to the p-values. This correction is needed because otherwise we too easily conclude that there are true differences between Greece and Italy.

Remember that the probability of a Type I error is very often chosen to be 5%. If we have one null-hypothesis that we want to test, the probability that we incorrectly conclude that there *is* a difference in means (but there is really *no* difference at population level!) is 5%. But suppose we have 10 hypotheses that we wish to test. Then what is the probability of finding at least 1 signficant result while there is no difference? Well, it could be that our first hypothesis is falsely rejected, or our second, or our third, or perhaps even both our second and fourth hypothesis, and so on and so forth. With 10 hypotheses to test, there will be a high probablity that *at least* 1 will be falsely rejected. If *each* hypothesis has a probability of a Type I error of 5%, the probability that none of the hypotheses is falsely rejected equals $0.95^10 = 0.60$ (if we assume that all probablities are independent). So the probability that at least one is falsely rejected is the complement of that, so 0.40. If we carry out such research with 10 hypotheses, each using a significance level of 5%, we actually have a probability of 40% of making at least one Type I error! That is awful, we don't want that. That's why in research, with a lot of hypotheses to be tested, we generally adjust the p-value in order to be more careful rejecting null-hypotheses. Theoretically,

the p-value of our posthoc comparison of Greece and Italy should be equal to 0.19, corresponding to the simple dummy variable analysis ignoring Norway, but we report 0.575, because we also test two other hypotheses here. Actually, the p-value of the simple analysis of 0.19 is multiplied by the total number of tests, which is 3.

9.0.4 Posthoc tests for complex contrasts

9.0.5 Fishing expeditions

The practice of testing a lot of hypotheses is often described as a fishing expedition. Just set out with large nets, throw them out, and catch whatever you can. In some extreme cases, in genetics for example, researchers test thousands or even millions of hypotheses on the basis of only one data set. Imagine that you collect height data on 70 countries and you want to know what countries differ from what other countries. The total number of pairs of countries equals 70 over 2, which is equal to 2415. So with 2415 p-values, what is the a priori probablity of a significant result? If in reality there ARE no differences in means, and a fixed significance level of 0.05, 5% of the p-values will be significant! So with such a data set, there will be at least $0.05 \times 2415 = 121$ significant p-values. At least, because there might be some true ones too. So in that scenario, it is impossible to know which p-values are too be trusted: many of them will involve false rejections.

For this reason, always be very specific about the null-hypothesis that you want to test with your data. If you have a very specific hypothesis about the differences in means, following a specific pattern, then always use a planned contrasts analysis. If after your analysis, there are some secondary hypotheses that you'd like to check (but for which you had no specific expectation) then report posthoc tests. The Bonferroni post hoc test is a good choice, as it is very conservative: it is very unlikely that you will falsely reject a hypothesis. Alternatively, there are some other post hoc tests, for further reading see the SPSS manual.

In general do a contrast analysis (planned comparisons) if:

- the overall test for the equality of all means is significant
- the comparisons are chosen *before* looking at the results (means, plots, statistical tests): they should be planned ahead!
- the number of planned contrasts should not exceed the degrees of freedom, that is, the number of groups minus 1.

Otherwise, do posthoc analyses, or better still, perform as few tests as possible! Only do posthoc tests if you are in an exploratory mood (you're not having a specific hypothesis but you would like to get some new ideas for future research) or when your supervisor asks for them.

Exercises

You compare 4 groups. You'd like to know whether the averages observed in groups 1 and 2 differ from the averages observed in groups 3 and 4.

- State the null hypothesis
- Define the contrast
- Provide the SPSS syntax for this contrast

answers:

- $H_0 : \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2} = 0$
- (0.5 0.5 -0.5 -0.5)
- ```
  UNIANOVA height BY group
  /DESIGN=country
  /CONTRAST(country)=SPECIAL(0.5 0.5 -0.5 -0.5).
  ```

You compare 5 groups. You'd like to know whether the average observed in group 1 differs from the averages observed in groups 3, 4 and 5.

- State the null hypothesis
- Define the contrast
- Provide the SPSS syntax for this contrast

answers:

- $H_0 : \frac{\mu_1}{1} - \frac{\mu_3+\mu_4+\mu_5}{3} = 0$
- (1 0 -0.33 -0.33 -0.33)
- ```
  UNIANOVA height BY group
  /DESIGN=group
  /CONTRAST(group)=SPECIAL((1 0 -0.33 -0.33 -0.33)).
  ```

A student has run the following SPSS syntax:

```
UNIANOVA score BY school
/DESIGN=school
/CONTRAST(school)=SPECIAL((0 0 1 -0.5 -0.5)).
```

What null-hypothesis is tested using this syntax?
answer: The hypothesis that the average score in school 3 is the same as the mean average score in schools 4 and 5 or $H_0 : \mu_3 = \frac{\mu_4+\mu_5}{2}$

A student has tested the research hypothesis that height is different in the

Benelux countries: The Netherlands, Belgium and Luxemburg, and finds a significant result. His supervisor asks then where the differences come from: is it that height is different in the Netherlands, or is it perhaps Luxemburg that deviates from the other two countries? She would like to have more specific information where the differences are between these three countries. What would you advise this student to do?

answer: the supervisor does not seem to have any clearcut hypthesis about height differences in the Benelux countries. You therefore advise to carry a number of posthoc tests, that take into account the increase in the probability of a Type I error by adjusting p-values.

A student has tested the research hypothesis that height is different in the Benelux countries: The Netherlands, Belgium and Luxemburg, and finds a significant result. His supervisor then says that the student is not finished yet. She would like to know whether the theory is correct that the larger the country, the taller the people. She would therefore like to know whether the average height in small country Luxemburg is different from the height averages in Belgium and The Netherlands.

answer: the supervisor has a clearcut hypothesis about height differences in the Benelux countries. You therefore advise to carry out a planned comparison (a contrast analysis), that specifically tests the null hypothesis that the average in Luxemburg is the same as the mean average of Belgium and the Netherlands together.

Chapter 10

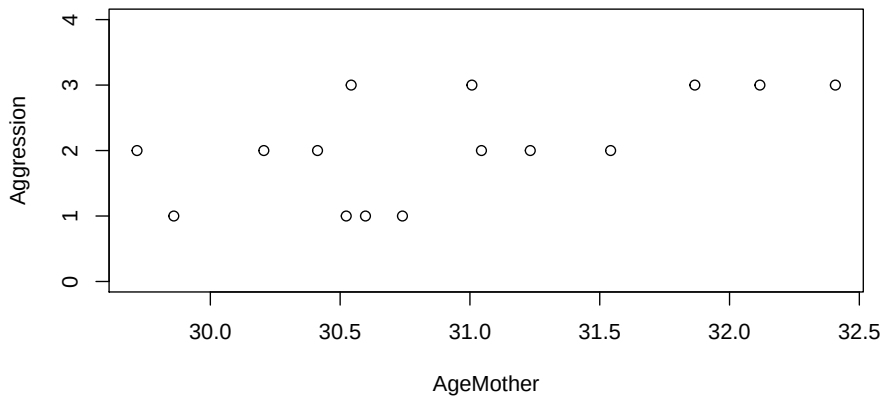
When assumptions are not met: non-parametric alternatives

Linear models do not apply to every data set. As discussed before, sometimes the assumptions of linear modelling are not met. One of the assumptions is linearity or additivity. Additivity requires that one unit change in variable X leads to the same amount of change in Y, no matter what value X has. For bivariate relationships this leads to a linear shape. But sometimes you can only expect that Y will change in the same direction, but you don't believe that this amount is the same for all values of X. This is the case for example with ordinal dependent variable. Suppose we wish to model the relationship between age of the mother and an aggression score of her 7-year-old child. Suppose aggression is measured on a three-point ordinal scale: not aggressive, sometimes aggressive, often aggressive. Since we do not know the quantitative differences between these three levels there are many graphs we could draw for a given data set.

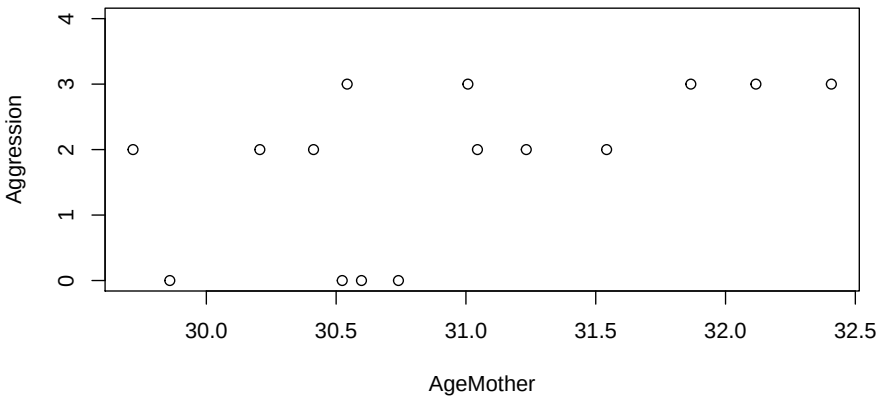
Suppose we have the following data set:

AgeMother	Aggression
32	Sometimes aggressive
31	Often aggressive
32	Often aggressive
30	Not aggressive
31	Sometimes aggressive
30	Sometimes aggressive
31	Not aggressive
31	Often aggressive
31	Not aggressive
30	Sometimes aggressive
32	Often aggressive
32	Often aggressive
31	Sometimes aggressive
30	Sometimes aggressive
31	Not aggressive

If we want to make a scatter plot, we could choose the values 1, 2, and 3 for the three categories respectively. We would then get the following plot:



But since the Aggression data are ordinal, we could also choose values 0, 2, and 3:



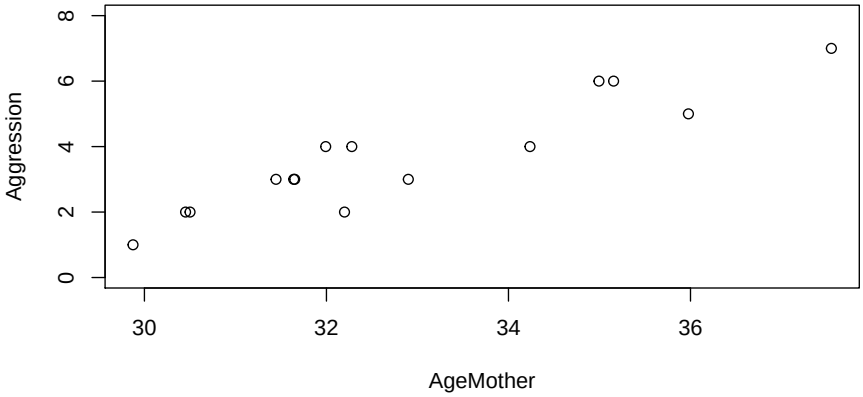
As we change the scale for the ordinal variable, we also see the best fitting regression line changing.

So with ordinal data, the assumption of additivity is often not met, since the values for a quantitative analysis are arbitrarily chosen.

In some case though, ordinal data could be modelled linearly. Look at the following example where we measured aggression with a 7-point Likert scale:

AgeMother	Aggression
35	6
32	4
35	6
36	5
33	3
30	1
32	4
32	2
34	4
30	2
32	3
31	2
32	3
31	3
38	7

When we plot these data, using the values 1 through 7, we see a nice linear relationship. So even when the values are arbitrarily chosen, a linear model can be a good model for a given data set with ordinal variables.

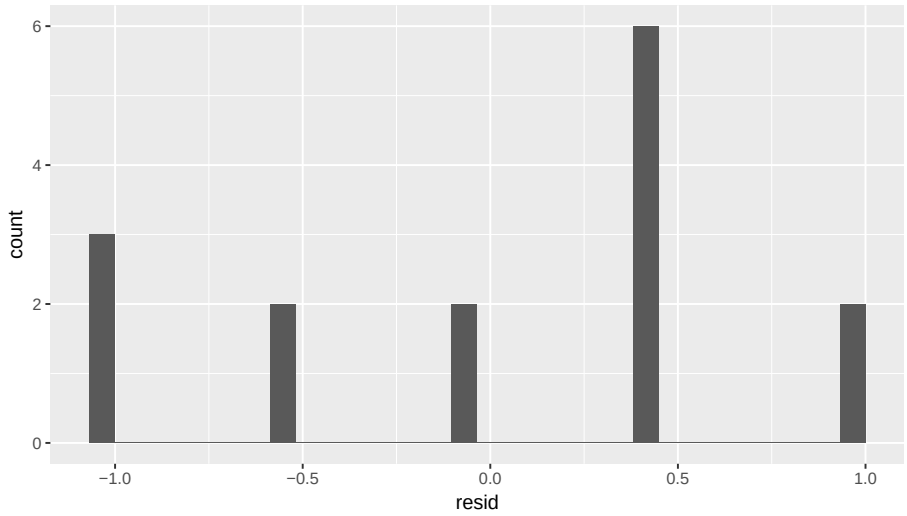


So with ordinal data, always check that your data indeed conform to a linear model, but realize at the same time that you're assuming a quantitative relationship between the variables that may or may not make sense.

If you believe that a quantitative analysis is meaningless then consider nonparametric analysis.

Another case where we favour a nonparameteric analysis, is when the assumption of normally distributed residuals is not tenable. For instance look again at Figure ????. When we perform a regression analysis and plot a histogram of the residuals, we see the following:

```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



Everytime we see a distribution of residuals that is either very skew, or has

very few different values, we should consider a nonparametric analysis. Note that the shape of the distribution of the residuals is directly related to what scale values we choose for the ordinal categories. By changing the values we change the regression line, and that directly affects the relative sizes of the residuals.

First, we will discuss a nonparametric alternative for two quantitative variables. We will start with Spearman's rho, or Spearmams rank-order correlation coefficient r_s. Next we will discuss an alternative to r_s, Kendall's T. After that we will discuss categorical variables, when comparing group differences.

10.1 Spearman's rho

Suppose we have 10 students and we ask their teachers to rate them. One teachers rates them on geography and the other teachers on history. We only ask to give rankings: indicate the brightest student with a 1 and the dullest student with a 10. Then we might have the following data set:

rank.geography	rank.history
5	4
4	5
6	7
7	8
8	6
9	9
10	10
2	3
1	1
3	2

Now we acknowledge the ordinal nature of the data: we only have rankings: a person with rank 1 is brighter than a person with rank 2, but we do not how large the difference in brightness really is. Now we want to establish whether there is a relationship between rankings on geography and the rankings on history: is it true that the higher the ranking on geography, the higher the ranking on history?

By eyeballing the data, we see that the brightest student in geography is also the brightest student in history (rank 1). We also see that the dullest student in history is the dullest student in geography (rank 10). Furthermore, we see relatively small differences between the rankings on the two subjects: high rankings on geography seem to go together with high rankings on history. Let;s look at these differences between rankings more closely by computing them:

rank.geography	rank.history	difference
5	4	-1
4	5	1
6	7	1
7	8	1
8	6	-2
9	9	0
10	10	0
2	3	1
1	1	0
3	2	-1

So theoretically the difference could be as large as 9, but here we see a biggest difference of -2. So the average difference is the sum of these differences, divided by 10, so we get 0. This is because we plus and minus values. If we would take the square of the differences, we would get positive values:

rank.geography	rank.history	difference	squared.difference
5	4	-1	1
4	5	1	1
6	7	1	1
7	8	1	1
8	6	-2	4
9	9	0	0
10	10	0	0
2	3	1	1
1	1	0	0
3	2	-1	1

Now we can compute the average squared difference, which is equal to 10/10 = 1. Generally, the smaller this value, the closer the rankings are together, and the more correlation there is between the two subjects.

A clever mathematician like Spearman has shown that is even better to use a somewhat different measure for a correlation between ranks. He showed that it is wiser to compute the following statistic:

$$r_s = 1 - \frac{6 \sum d^2}{N^3 - N} \tag{10.1}$$

because then you get a value between -1 and 1, just like a Pearson correlation. So in this case the sum of the squared difference is equal to 10, N is the number of students, so we get:

$$r_s = 1 - \frac{6 \times 10}{10^3 - 10} = 1 - 60/990 = 0.94 \tag{10.2}$$

This is called the Spearman rank-order correlation coefficient r_s. It can be used for any two variables of which one is at most ordinal. The trick is to convert

the scale values into ranks, and then apply the formula above. For instance, if we have the variable Grade with the following values (C, B, D, A, F), we convert them into rankings by saying the A is the highest value (1), B is the second highest value (2), C is the third highest value (3), D is the fourth highest value (4) and F is the fifth highest value (5). So tranformed into rankings we get (3, 2, 4, 1, 5). We can also let SPSS do the computations for us. Suppose we have two variables shoe size and aggression, we use the syntax:

```
NONPAR CORR
  /VARIABLES=shoesize aggression
 /PRINT=SPEARMAN .
```

In the output you will see a correlation matrix very similar the one for a Pearson correlation. Spearman's rho is equal to the r_2 mentioned above. You will also see whether the correlation is significantly different from 0, indicated by a p-value. If the p-value is very small, you may concluded that on the basis of these data, the correlation in the population is not equal to 0, ergo, in the population there is a relationship between shoe size and aggression.

Below we discuss an alternative measure for a correlation for ordinal data, the Kendall rank-order correlation coefficient T.

10.2 Kendall rank-order correlation coefficient T

If you want to know if there is a relationship between two variables, of which one is at most ordinal, you can either use Spearman's rho or Kendall's T. However, if you have three variables, and you want to know whether there is a relationship between A and B, over and above the effect of C, you can use a extension of Kendall's T. Note that this is very similar to multiple regression: a coefficient for variable X_1 in multiple regression with two predictors is the effect of X_1 on Y over and above the effect of X_2 on Y. The logic of Kendall's T is also based on rank orderings, but it involves a different computation. Let's look at the student data again with the teacher's rankings of the students on two subjects:

student	rank.geography	rank.history
1	5	4
2	4	5
3	6	7
4	7	8
5	8	6
6	9	9
7	10	10
8	2	3
9	1	1
10	3	2

First we put the students in a new order, such that the brightest student in geography comes first, and the dullest last. This also changes the order in the variable history:

student	rank.geography	rank.history
9	1	1
8	2	3
10	3	2
2	4	5
1	5	4
3	6	7
4	7	8
5	8	6
6	9	9
7	10	10

From this we see that the history teacher agrees with the geography teacher that student 9 is brighter than student 8. They also agree that student 9 is brighter than sutduent 10. If we do this for all possible pairs of students, we can count the number of times that they agree and we can count the number of times that they disagree. The total number of possible pairs is equal to $\binom{10}{2}$ $= 90/2 = 45$. So for example we might find for a data set on 10 students, that of all 45 pairs, the teachers agree 30 times, and disagree 15 times. Then we can compute Kendall's T as follows:

$$T = \frac{agreements - disagreements}{totalnumberofpairs} = \frac{30 - 15}{45} = 0.33 \tag{10.3}$$

This T statistic varies between -1 and 1 and can therefore be seen as a nonparametric analog of a Pearson correlation. Here, the teachers more often agree than disagree, so the correlation is positive. A negative correlation means that the teachers more often disagree than agree on the relative brightness of their students.

The computations are quite involving so we're very lucky that SPSS can do the tedious job for us, with the following syntax:

```
NONPAR CORR
  /VARIABLES=shoesize aggression
 /PRINT=KENDALL .
```

As said, the advantage of Kendall's T over Spearman's r is that Kendall's T can be extended to cover the case that you wish to establish the strength of the relationships of two variables A and B, over and above the relationship with C.

Now that we have discuss relationships between quantitative variables, let's have look at the case where we have categorical variables.

10.3 Kruskall-Wallis test for group comparisons

Suppose we have three groups of students that go on a field trip together: mathematicians, psychologists and engineers. Each can pick a rain coat, with five possible sizes: extra small, small, medium, large or extra large. We want to know if preferred size is different in the three populations, so that we can be better prepared in the future. Now we have information about size, but this knowledge is not quantitative: we do not know the difference in size between medium and large, only that large is larger than medium. We have ordinal data, so computing a mean is impossible here. Even we would assign values like 1= extra small, 2=small, 3= medium, etcetera, the mean would be rather meaningless as these values are arbitrary. So instead of focussing on means, we can focus on medians: the middle value. For instance, the median value for our sample of mathematicians could be medium, for our sample of psychologists small, and for our sample of engineers large. Our question might then be whether the median values in the populations are really different.

This can be assessed using the Kruskall-Wallis test. Similar to Spearman's r and Kendall's T, the data are transformed into ranks. This is done for all data at once, so for all students toghether.

For example, if we had the following data:

student	group	size
001	math	xs
002	math	xl
003	psych	m
004	psych	s
005	engineer	l
...	...	...

We would transform it into ranks, from smallest to largest, like this:

student	group	size	rank
001	math	xs	1
002	math	xl	5
003	psych	m	3
004	psych	s	2
005	engineer	l	4
...	...	...	...

Next, we could compute the average rank per group. The group with the smallest sizes would have the lowest average rank, etcetera. Under the null-hypothesis, if the distribution of size was the same in all three groups, the average ranks would be about the same. If they are very different, this is an indication that size is not distributed equally among the three groups. In order to have a proper statistical test, a rather complex formula is used to compute the so-called KW statistics. We know the distribution of this KW statistic under the null-hypothesis, so we know what extreme values are, and consequently

can compute p-values. This tedious computation is also done by SPSS using the following syntax.

```
NPTESTS
  /INDEPENDENT TEST (size) GROUP (group) KRUSKAL_WALLIS(COMPARE=NONE).
```

The output gives you a significance level (p-value) of the test that size is distributed equally among psychology students, engineering students and mathematics students. If it is a very low number, you may conclude that in the population, students in psychology, mathematics in engineering have different preferences regarding the size of their rain coat on field trips.

Chapter 11

Linear mixed modelling: introduction

11.1 Fixed effects and random effects

In the simplest form of linear modelling, we have one dependent continuous variable, one intercept and one or more independent variables. Let's look at a simple regression equation where dependent variable y is predicted by an intercept b_0 and a linear effect of independent variable x_1 with regression slope parameter b_1, and an error term e, where we assume that the error term e comes from a normal distribution.

$$y = b_0 + b_1 x_1 + e \tag{11.1}$$

$$e \sim N(0, \sigma^2) \tag{11.2}$$

Using this model, we know that for a person with a value of 5 for x_1, we expect y to be equal to $b_0 + b_1 \times 5$. As another example, if y is someone's IQ score, x_1 is someone's brain size in cubic milliliters, b_0 is equal to 70, and b_1 is equal to 0.1, we expect on the basis of this model that a person with a brain size of 1500 cubic millimeters has an IQ score of $70+0.01\times1500$, which equals 85.

Now, for any model the predicted values usually are not the same as the observed values. If the model predicts on the basis of my brain size that my IQ is 140, my true IQ might be in fact 130. This discrepancy is termed the residual: the observed y, minus the predicted y, or $\hat{y}$, so in this case the residual is $y - \hat{y} = 130 - 140 = -10$.

Here we have the model for the relationship between IQ and brain size.

$$IQ = 70 + 0.1 \times Brainsize + e \tag{11.3}$$

$$e \sim N(0, \sigma^2) \tag{11.4}$$

Note that in this model, the values of 70 and 0.1 are *fixed*, that is, we use the same intercept and the same slope for everyone. You use these values for any person, for Henry, Jake, Lizz, and Margaret. We therefore call these effects of intercept and slope *fixed effects*, as they are all the same for all research units. In contrast, we call the e term, the random error term or the residual in the regression, a *random effect*. This is because the error term is *different for every research unit*. We don't know the specific values of these random errors or residuals for every person, but nevertheless, we assume that they come from a distribution, in this case a normal distribution with mean 0 and an unknown variance. This unknown variance is given the symbol σ^2

Here are a few more examples.

1. Suppose we study a number of schools, and for every school we use a simple linear regression equation to predict the number of students (dependent variable) on the basis of the number of teachers (independent variable). For every research unit (in this case: school), the intercept and the regression slope are the same (fixed effects), but the residuals are different (random effect).

2. Suppose we study a number of students, and for every student we use a simple linear regression equation to predict the math test score on the basis of the number of hours of study the student puts in. Here, the research unit is student, and for every student, the intercept and the regression slope are the same (fixed effects), but the residuals are different (random effect).

3. Suppose we study reaction times, and for every measure of reaction time – a trial – we use a simple linear regression equation to predict reaction time in milliseconds on the basis of the characteristics of the stimulus. Here, the research unit is trial, and for every trial, the intercept and the regression slope are the same (fixed effects), but the residuals are different (random effect).

Now, what happens when we have a lot of data on students, but the students come from different schools? Suppose we want to predict average grade for every student, on the basis of the number of hours of study the student puts in. We again could use a simple linear regression equation.

$$y = b_0 + b_1 hourswork + e \tag{11.5}$$
$$e \sim N(0, \sigma^2) \tag{11.6}$$

That would be fine if all schools would be all very similar. But suppose that some schools have a lot of high scoring students, and some schools have a lot of low scoring students? Then school itself would also be a very important

predictor, apart from the number of hours of study. One thing we could therefore do is to include school as a categorical predictor. We would then have to code this school variable into a number of dummy variables. The first dummy variable called *school*1 would indicate whether students are in the first school (school1=1) or not (school1=0). The second dummy variable *school*2 would indicate whether students are in the second school (school2=1) or not (school2=0), etcetera. You can then add these dummy variables to the regression equation like this:

$$y = b_0 + b_1 hourswork + b_2 school1 + b_3 school2 + b_4 school3 + ... + e$$
$$e \sim N(0, \sigma^2)$$

In the output we would find a large number of effects, one for each dummy variable. For example, if the students came from 100 different schools, I would get 99 fixed effects for the 99 dummy variables. However, one could wonder whether this is very useful. As stated earlier, fixed effects are called fixed because they are the same for every unit of research, in this case every student. But working with 99 dummy variables, where students mostly score 0, this seems very much over the top. In fact, we're not even interested in these 99 effects. We're interested in the relationship between test score and hours of work, meanwhile taking into account that there are test score differences across schools. The dummy variables are only there to account for differences across schools; the prediction for one school is a little bit higher or lower than for another school, depending on how well students generally perform in each school.

We could therefore try an alternative model, where we treat the school effect as *random*: we assume that every school has a different average test score, and that these averages are normally distributed. We call these average test score deviations *school effects*:

$$y = b_0 + b_1 hourswork + schooleffect + e \tag{11.7}$$
$$schooleffect \sim N(0, \sigma_s^2) \tag{11.8}$$
$$e \sim N(0, \sigma_e^2) \tag{11.9}$$

So in this equation, the intercept is fixed, that is, the intercept is the same for all observed test scores. The regression coefficient b_1 for the effect of hours of work is also fixed. But the schooleffect is random, since it is different for every school. The residual e is also random, being different for every student. It could also be written like this:

$$y = (b_0 + schooleffect) + b_1 hourswork + e \tag{11.10}$$
$$schooleffect \sim N(0, \sigma_s^2) \tag{11.11}$$
$$e \sim N(0, \sigma_e^2) \tag{11.12}$$

This representation emphasizes that for every school, the intercept is a little bit different: for school A the intercept might be $b_0 + 2$, and for school R the intercept might be $b_0 - 3$.

So, this equation states that every observed test score is 1) partly influenced by an intercept that is random, with a certain average b_0 and variance σ_s^2, that is dependent on which school students are in, 2) partly influenced by the number of hours of work, an effect that is the same no matter what school a student is in (fixed), and 3) partly influenced by unknown factors, indicated by a random residual e coming from a normal distribution with variance σ_e^2.

To put it more formally: test score y_{ij}, that is, the test score from student j in school i, is the sum of an effect of the school $b_0 + schooleffect_i$ (the average test score in school i), plus an effect of hours of work, $b_1 \times hourswork$, and an unknown residual e_{ij} (a specific residual for the test score for studuent j in school i).

$$y_{ij} = b_0 + schooleffect_i + b_1 hourswork + e_{ij} \tag{11.13}$$

$$schooleffect_i \sim N(0, \sigma_s^2) \tag{11.14}$$

$$e_{ij} \sim N(0, \sigma_e^2) \tag{11.15}$$

So in addition to the assumption of residuals that have a normal distribution with mean 0 and variance σ_s^2, we also have an assumption that the school averages are normally distributed, in this case with mean b_0 and variance σ_s^2.

Let's go back to the example of reaction times. Suppose in an experiment we measure reaction time in a large number of trials. We want to know whether the size of the stimulus (large/small) has an effect on reaction time. Let's also suppose that we carry out this experiment with 20 participants, where every participant is measured during 100 trials: 50 large stimuli and 50 small stimuli in random order. Now probably, some participants show generally very fast responses, and some participants show generally very slow responses. In other words, the average reaction time for the 100 trials may vary from participant to participant. This means that we can use participant as an important predictor of reaction times. To take this into account we can use the following equation:

$$y_{ij} = b_0 + speed_i + b_1 size + e_{ij} \tag{11.16}$$

$$speed_i \sim N(0, \sigma_s^2) \tag{11.17}$$

$$e_{ij} \sim N(0, \sigma_e^2), \tag{11.18}$$

where y_{ij}, is the reaction time j from participant i, $(b_0 + speed_i)$ is a random effect representing the average speed for each participant i (where b_0 is the overall average across all participants), b_1 is the fixed effect of the size of the stimulus, and unknown residual e_{ij} is a specific residual for the reaction time for trial j of participant i.

The reason for introducing random effects is that when your observed data are clustered, for instance student scores clustered within schools, or trial response times are clustered within participants, you violate the assumption of independence: two reaction times from the same person are more similar than two reaction times from different persons. Two test scores from students from the same school may be more similar than two scores from students in different schools. When this is the case, when data are clustered, it is very important to take this into account. When the assumption of independence is violated, you are making wrong inference if you only use a simple linear model. With clustered data, it is therefore necessary to work with an extension of the linear model, the so-called linear mixed model. The above models for test scores across different schools and reaction times across different participants, are examples of *linear mixed models*. The term *mixed* comes from the fact that the models contain a mix of both fixed and random effects.

If you have clustered data, you should take this clustering into account, either by using the grouping variable as a qualitative predictor (using a number of dummy variables) or by using a linear mixed model. As a rule of thumb: if you have fewer than 10 groups, consider dummy variables; if you have 10 or more groups, consider a linear mixed model. Use a linear mixed model if the assumption of normally distributed group differences is tenable. Use dummy variables if you are actually interested in the size of group differences.

Below, we will start with a very simple example of linear mixed model, one that we use for a simple pre-post intervention design.

11.2 Pre-post intervention design

Imagine a study where we hope to show that aspirin helps reduce headache. For 100 patients we ask to rate the severity of their headache before they use aspirin (on a scale from 1 - 100), and to rate the severity again 3 hours after taking 500 mg of aspirin. These patients were randomly selected among people who read the NY Times and suffer from regular headaches. So here we have clustered data: we have 100 patients, and for each patient we have two scores, one before (pre) and one after (post) the intervention of taking aspirin. Of course headaches differ from person to person, so we might have to take into account that some patients have a higher average level of headache than other patients. Now, the data could be represented in different ways, but suppose we have the following data matrix (showing only the first 5 patients):

patient	pre	post
001	55	45
002	63	50
003	66	56
004	50	37
005	63	50
...	...	...

What we observe here is that the severity seems generally lower after the intervention than before the intervention. But you may also notice that the severity of the headache also varies across patients: some have generally high scores (for instance patient 003), and some have generally low scores (for example patient 001). Therefore, the headache scores seem to be clustered, violating the assumption of independence. We can quantify this clustering by computing a correlation between the pre-intervention scores and the post-intervention scores. Here it appears that there is a strong positive correlation, indicating that the higher the pain score before the intervention, the higher the pain score after the intervention.

There is an alternative way of representing the same data. Let's look at the same data in a new format:

patient	measure	headache
1	1	55
1	2	45
2	1	63
2	2	50
3	1	66
3	2	56
4	1	50
4	2	37
5	1	63
5	2	50

Here we acknowledge that there is really only one dependent measure: headache severity. The other two variables indicate that this variable varies across both patients and time point (pre intervention and post intervention). Here we might consider applying a simple linear regression model, using severity as the dependent variable and **measure** (1st or 2nd) as a qualitative predictor (using a dummy variable). However, since we know that there is a correlation between the pre and post severity measures, we know that measures systematically vary across patients: some score high on average and some score low on average. Therefore we have to run a linear *mixed* model, including not only the fixed effect of **measure** but also a random effect for each patient. Since we are really interested in the effect of the intervention, that is, we want to know how large the effect of aspirin is, we use a fixed effect for the time effect (the variable **measure**). For the patient effect, because there are so many patients (100) and we're really not interested in all of these individual differences, we use a random

effect. This means that we only assume there is a normal distribution for all of the patient differences. So we get the following equation:

$$y_{ij} = b_0 + patient_i + b_1 measure + e_{ij} \tag{11.19}$$

$$patient_i \sim N(0, \sigma_p^2) \tag{11.20}$$

$$e_{ij} \sim N(0, \sigma_e^2) \tag{11.21}$$

y_{ij} is the jth headache severity score (first or second) for patient i, ($b_0 + patient_i$) is the average headache for patient i, *measure* is a dummy variable for which measure (first or second), and b_1 is the effect of the intervention (by how much the severity changes from pre to post). We assume that the average pain level for each patient shows a normal distribution with average b_0 and variance σ_p^2. And of course we assume that the residuals show a normal distribution.

This analysis can be done with the following SPSS syntax, treating time as a qualitative variable (using BY) for which SPSS will create a dummy variable automatically.

```
MIXED headache  BY measure
  /FIXED=measure
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(patient) COVTYPE(VC).
```

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	99	9494.826	.000
measure	1	99.000	648.284	.000

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	49.320000	.595162	124.746	82.868	.000	48.142077	50.497923
[measure=1.00]	10.360000	.406890	99.000	25.461	.000	9.552642	11.167358
[measure=2.00]	0[b]	0	.	.	.	.	.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		8.277980	1.176581
Intercept [subject = patient]	Variance	27.143838	4.485097

The most interesting output is given here. We're mostly interested in the fixed effect of the intervention: does aspirin reduce headache? After an F-test,

we see the linear model coefficients, with an intercept of around 49 and a positive effect of the intervention dummy variable, around +10. We see that the dummy variable was coded 0 for the second measure and 1 for first meausure. So, for our dependent variable headache, we see that the expected headache severity for the observations with a 0 for the dummy variable (that is, measure 2, which is *after* taking aspirin), is equal to $49 + (10) \times 0 = 49$. Similarly, we see that the expected headache severity for the observations with a 1 for the dummy variable (that is, *before* taking aspirin), is equal to $49 + (10) \times 1 = 49 + 10 = 59$. So, average pain severity is 10 points higher before the intervention than after the intervention. Whether this difference is significant is indicated by a t-test. We see here that the average headache severity after taking an aspirin is significantly different from the average headache severity before taking an aspirin, $t(99) = 25.46, p < 0.01$. The degrees of freedom are taken from the Test of Fixed Effects table with the F-statistics. Alternatively we can write

> The average headache severity after taking an aspirin is significantly different from the average headache severity before taking an aspirin, $F(1, 99) = 648.28, p < 0.01$

We might therefore carefully conclude that aspirin reduces headache in the population of NY Times readers with headache problems, where the reduction is around 10 points on a 1...100 scale (95% CI: 9.55 – 11.17).

Now, this looks like reporting the output of a regular linear model, but of course it isn't. We also have some extra output, about the random effect of patient. We assumed that the individual differences in headache severity in the 100 patients came from a normal distribution. How large are these individual differences actually? This can be gleaned from the Covariance Parameters part of the SPSS output. We there see two random effects: the one for the residuals and one for the patients. The intercept seems to vary with a variance of 27, which is equivalent to a standard deviation of $\sqrt{27}$ which is around 5.2. What does that mean exactly? Well let's look at the equation again and fill in the numbers:

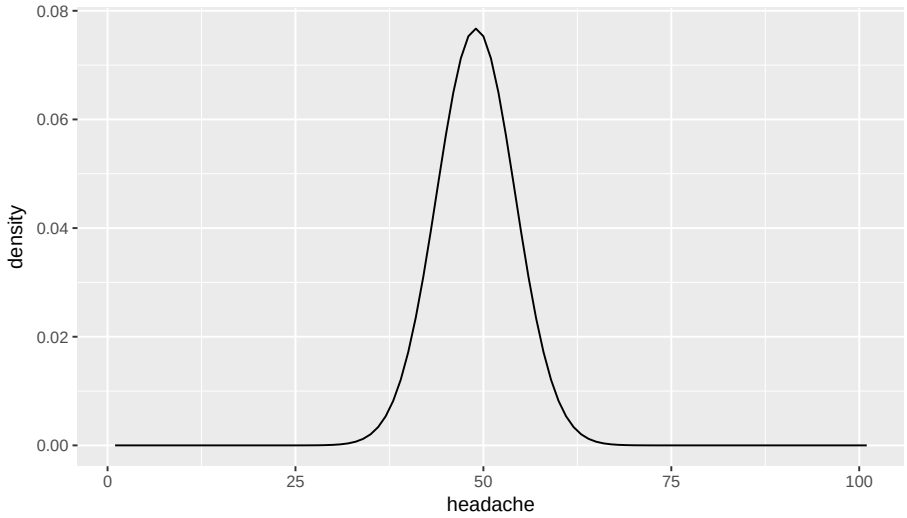
$$y_{ij} = b_0 + patient_i + b_1 measure + e_{ij} \tag{11.22}$$

$$y_{ij} = 49 + patient_i + 10 measure + e_{ij} \tag{11.23}$$

$$patient_i \sim N(0, 27) \tag{11.24}$$

$$e_{ij} \sim N(0, 8) \tag{11.25}$$

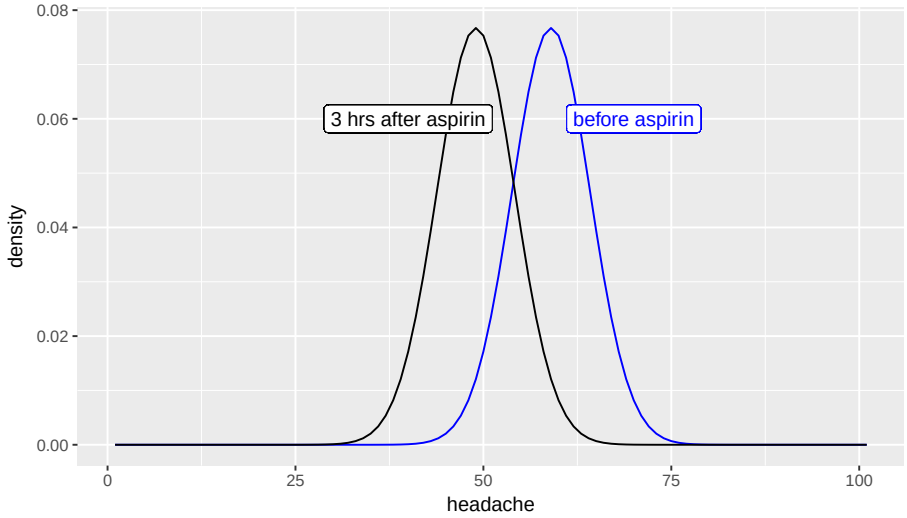
Since measure is coded 0 for the headache level after the intervention, we conclude that the average pain level after taking aspirin is 49. However, not everybody's pain level after taking aspirin is 49: people show variance. The pain level after aspirin varies with a variance of 27, which is equavalent to a standard deviation of arond 5.2. Figure below shows how much this variance actually is. It depicts a normal distribution with mean 49 and a standard deviation of 5.2.



So *after* taking aspirin, most patients show headache levels between 30 and 60. More specificially, if we would take the middle 95% by using plus or minus twice the standard deviation, we can estimate that 95% of the patients shows levels between $49 - 2 \times 5.2 = 38.6$ and $49 + 2 \times 5.2 = 59.4$

Now let's look at the levels *before* taking aspirin. The average headache leves is equal to $49 + 10 = 59$. So 95% of the patients shows headache levels between $59 - 2 \times 5.2 = 48.6$ and $59 + 2 \times 5.2 = 49.4$ before taking aspirin.

Together these results can be graphically explained in the plot below:



In this plot you see there is variability in headache levels before taking aspirin, and there is variation in headache levels after taking aspirin. We also

see that these distributions have the same spread (variance): in the model we assume that the variability in headache before aspirin is equal to the variability after aspirin. The distributions are equal, except for a horizontal shift: the distribution for heachache after aspirin is the same as the distribution before apsirin, except for a shift to the left of about 10 points. This is of course the effect of aspirin in the model, the b_1 parameter in our model above.

The fact that the two distributions before and after aspirin show the same spread (variance) was an inherent assumption in our model: we only have one random effect for patient in our output. If the assumption of equal variance (homoscedasticity) is not tenable, then one should consider other linear mixed models. But this is beyond the scope of this chapter. The assumption can be checked by plotting the residuals, using differnt colours for residuals from before taking aspirin and for residuals from after taking aspirin.

11.2.1 Exercises

Suppose an intervention study looked at the effect of therapy on depression levels. A random sample of patients were measured before and after the therapy. Given the following equation, based on output of the statistical software package R. The dummy variable *measure* was coded 0 for before therapy and 1 for after therapy.

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	129.673	257.700	.000
measure	1	99.000	31.263	.000

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	10.570000	.658442	129.673	16.053	.000	9.267319	11.872681
measure	-2.280000	.407773	99.000	-5.591	.000	-3.089110	-1.470890

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		8.313939	1.181692
Intercept [subject = patient]	Variance	1.784949	1.030708

Look at the output below. You see information about random effects and you see information about fixed effects.

- 1. What is the intercept of the model?

- 2. What is the slope coefficent for the measure variable?

- 3. What is the variance of the residuals? What is the standard deviation?

- 4. What is the variance of the individual differences among patients? What is the standard deviation?

- 5. Fill in the values in the linear mixed model below:

$$depression_{ij} = \ldots + patient_i + \ldots \times measure + e_{ij} \tag{11.26}$$
$$patient_i \sim N(0, \sigma_p^2 = \ldots) \tag{11.27}$$
$$e_{ij} \sim N(0, \sigma_e^2 = \ldots) \tag{11.28}$$

- 6. what can you say about the average depression level before therapy?

- 7. what can you say about the average depression level after therapy?

- 8. How much variance in depression level before therapy does this model predict? What is the standard deviation?

- 9. Between what values do depression levels before therapy in the middle 95% of patients show?

- 10. How much variance in depression level after therapy does this model predict? What is the standard deviation?

- 11. Between what values do depression levels after therapy in the middle 95% of patients show?

- 12. Does therapy help to alleviate depression in patients? You may use an approximation to construct a confidence interval.

- 13. A researcher has two groups of patients: one group receives medicine and one group receives therapy. The null-hypothesis is that depression levels after medicine are as high as depression levels after therapy. Do we analyse these data with an ordinary linear model, or with a linear mixed model? Explain your answer.

- 14. A researcher studies one group of students: they first get lectures from teacher A and then they get lectures from teacher B. The null-hypothesis is that the average teacher evaluation for teacher A is the same as the average teacher evaluation for teacher B. Do we analyse these data with an ordinary linear model, or with a linear mixed model? Explain your answer.

- 15. For a study to the effect of light on mood, we have data on 100 teachers They were asked to rate their mood on a cloudy day and asked to rate their mood on a sunny day. We have the variable **mood**, the dummy variable **sunny** and we want to include a random effect for **teacher** From the three syntaxes below, choose the one that is most suitable for your analysis and fill in the blanks.

```
MIXED ...  WITH ...
  /FIXED=...
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(...) COVTYPE(VC).

UNIANOVA ... WITH ...
/ design = ...
/ print = parameter.

UNIANOVA ... BY ...
/ design = ...
/ print = parameter.
```

- 16. A researcher wants to know whether students in green classrooms (colour = 1) perform better than students in yellow classrooms (colour = 2). The following data were collected (showing only a part):

student	colour	performance
1	1	6.79
2	2	8.28
3	1	9.08
4	2	5.65
5	1	8.43
6	2	8.51
7	1	7.43
8	2	7.45
9	1	7.44
10	2	7.11

Would you use an ordinary linear model or a linear mixed model to analyze these data? Explain your answer.

- 17. A researcher wants to know whether students in dark classrooms (brightness = 0) perform better than students in bright classrooms (brightness = 1). The following data were collected (showing only a part):

student	brightness	performance
1	0	8.66
1	1	5.95
2	0	6.50
2	1	9.47
3	0	9.46
3	1	8.14
4	0	8.21
4	1	4.96
5	0	7.51
5	1	6.91

  Would you use an ordinary linear model or a linear mixed model to analyze these data? Explain your answer.

- 18. A landscaper believes that people get more creative once the environment becomes greener. She measures creativity before and after the introduction of new trees around the office building in a random sample of employees. Because creativity can also be influenced by the weather she also uses a dummy variable **sunny** to correct for these effects. Whether creativity is meassured before or after the introduction of the trees is indicated by the variable **green** that is coded green=1 for after the introduction and green=0 for before the introduction. The model that she therefore runs in SPSS is the following:

```
MIXED creativity  WITH green sunny
  /FIXED= green sunny
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(employee) COVTYPE(VC).
```

  We get the following output:

  Write a short paragraph describing these results and the conclusions in APA format.

Answers:

- Ad1: 10.57

- Ad2: -2.28

- Ad3: 8.31,2.88

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	161.896	9073.526	.000
green	1	98.273	109.513	.000
sunny	1	126.083	5.377	.022

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	59.971909	.629593	161.896	95.255	.000	58.728636	61.215182
green	-4.501580	.430162	98.273	-10.465	.000	-5.355194	-3.647967
sunny	-1.280658	.552300	126.083	-2.319	.022	-2.373637	-.187679

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		9.238257	1.318161
Intercept [subject = employee]	Variance	23.662234	4.073012

- Ad4: 1.79,1.34
- Ad5:

$$depression_{ij} = 10.57 + patient_i + (-2.28) \times measure + e_{ij} \quad (11.29)$$
$$patient_i \sim N(0, \sigma_p^2 = 1.79) \quad (11.30)$$
$$e_{ij} \sim N(0, \sigma_e^2 = 8.31) \quad (11.31)$$

- Ad 6: 10.57
- Ad 7: 10.57 + -2.28 = 8.29
- Ad 8: 1.79, 1.34
- Ad 9: $10.57 \pm 2 \times 1.34 = 7.89, 13.25$
- Ad 10: 1.79, 1.34
- Ad 11: $(10.57 - 2.28) \pm 2 \times 1.34 = 5.61, 10.79$

- Ad 12: For the effect of therapy (the measure variable), we see a b-value of -2.28 with a standard error of 0.407773, so if we use the +/-2 rule to compute a 95% confidence interval, we get $[-2.28 - 2 \times 0.41, -2.28 + 2 \times 0.41] = [-3.1, -1.46]$. The 95% interval does NOT contain the value 0 so we can reject the null-hypothesis that the effect of therapy is zero. Therefore, we conclude that therapy has an influence on depression. In this case we saw a decrease in depression levels after therapy.

- Ad 13: Two groups of patients are studied, and for each patient we have only one measure. Because we only have one measure for each unit of observation we conduct an ordinary linear model.

- Ad 14: One group of students is studied, and for each student we have two evaluations: one for teacher A and one for teacher B. Because we have more than one measure for each unit of observation, we have to use a linear mixed model.

- Ad15:

```
MIXED mood  WITH sunny
  /FIXED=sunny
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(teacher) COVTYPE(VC).
```

- Ad16: it seems as if each student was only measured once, there is no clustering, so we can use an ordinary linear model.

- Ad17: it seems as if each student was measured twice, in both dark and bright conditions, so we use a linear mixed model to account for clustering.

- Ad18:

> A linear mixed model was run to test the effect of green surroundings on creativity. The analysis was corrected for the effects of weather (sunny or not sunny) and random effects for employees. The results showed a significant but negative effect of the introduction of trees on creativity: creativity was on average 4.5 points lower after the introduction, $t(98) = -10.47, p < 0.001$. This effect was present over and above the effect of the weather which by itself had also an effect, where creativity was 1.28 points lower on sunny days than on not cloudy days, $t(98) = -2.32, p = 0.02$. The variance not explained by weather and greenness was largely explained by individual differences in creativity among employees, with an intraclass correlation of $\frac{23.7}{23.7+9.2} = 0.72$. We conclude that the introduction of trees has a negative influence on creativity in the employees that worked in the building studied in this research.

11.3 Pre-mid-post intervention design

In many intervention studies, one has more than two measurement moments. For instance, you'd like to know if there is not only a short term effect of aspirin, but also a long-term effect. Suppose that the study mentioned in the previous section on headache among NY Times readers was extended by asking patients not only to rate their headache before aspirin and 3 hours after intake, but also 24 hours after intake. In this case our data could look like this:

patient	measure1	measure2	measure3
1	52	45	47
2	59	50	55
3	65	56	58
4	51	37	42
5	62	50	55
6	61	53	57
7	56	44	55
8	62	48	53
9	56	48	49
10	58	45	44

So for each patient we have three measures: pre, post1 and post2. To see if there is some significant clustering, it is no longer possible to study this by computing a single correlation. We could however compute 3 different correlations: pre-post1, pre-post2, and post1-post2, but this is rather tedious, and moreover does not give us a single measure of the extent of clustering of the data. But there is an alternative: one could compute not a Pearson correlation, but an *intraclass correlation* (ICC). To do this, we need to bring the data again into a regression type format, like this (we call this *long format*, as opposed to *wide format*):

patient	measure	headache
1	1	52
1	2	45
1	3	47
2	1	59
2	2	50
2	3	55
3	1	65
3	2	56
3	3	58
4	1	51

We can perform an analysis using a MIXED analysis in SPSS:

```
MIXED headache BY measure
  /FIXED=measure
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(patient) COVTYPE(VC).
```

The output is given below:

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	99.000	9162.260	.000
measure	2	198	309.580	.000

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	51.680000	.606686	136.283	85.184	.000	50.480263	52.879737
[measure=1.00]	7.490000	.413362	198	18.120	.000	6.674843	8.305157
[measure=2.00]	-2.360000	.413362	198	-5.709	.000	-3.175157	-1.544843
[measure=3.00]	0[b]	0	.	.	.	.	.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		8.543401	.858644
Intercept [subject = patient]	Variance	28.263434	4.431211

In the output we see the fixed effects of two automatically created dummy variables **measure=1** and **measure=2**, and the intercept. We also see the variances of the random effects: the variance of the residuals and the variance of the random effects for each patient.

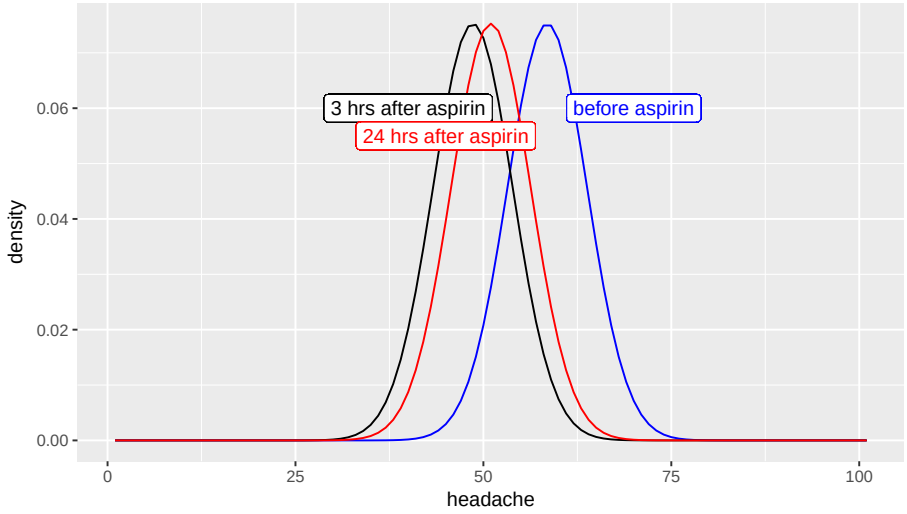
From this output, we can plug in the values into the equation:

$$headache_{ij} = 51 + patient_i + 7.5 \times measure1 - 2.4 \times measure2 + e_{ij}$$
$$patient_i \sim N(0, 28.3)$$
$$e_{ij} \sim N(0, 8.5)$$

Based on this equation, the expected headache severity score in the population 24 hours after aspirin intake is 51 (the third measure is the reference group). Dummy variable **measure=1** is coded 1 for the measurements before taking aspirin. Therefore, the expected headache score before aspirin intake is equal to $51 + 7.5 = 58.5$. Dummy variable **measure=2** was coded 1 for the measurements 3 hours after aspirin intake. Therefore, the expected headache score 3 hours after aspirin intake is equal to $51 - 2.4 = 48.6$. In sum, in this sample we see that the average headache level decreases directly after aspirin intake from 58.5 to 48.6, but then increases again to 51.

There was quite some variation in individual headache levels: the variance is equal to 28.3, so the standard deviation (its square root) is equal to about 5.3. Therefore, if we look at roughly 95% of the sample, we see that prior to taking

aspirin, the scores vary between $58.5 - 2 \times 5.3 = 47.9$ and $58.5 + 2 \times 5.3 = 69.1$. For the short-term effect of aspirin after 3 hours, we see that roughly 95% of the scores lie between $48.6 - 2 \times 5.3 = 38.0$ and $48.6 + 2 \times 5.3 = 59.2$. The normal distributions, predicted by this model, are depicted in the figure below:



So, are these distributions significantly different, in other words, do the means differ significantly before aspirin, 3hrs after aspirin and 24 hrs after aspirin? The answer is yes, because the F-test on the group means in the SPSS output is significant. Note the degrees of freedom: 2, because we compare 3 groups of data, so we need two dummy variables. Thus we report that aspirin has an effect on headache levels in NY Times readers, $F(2, 198) = 309.58, p < 0.001$.

If one has specific hypotheses regarding short-term and long-term effects, one could perform a planned contrast analysis, comparing the first measure with the second measure, and the first measure with the third measure. If one is just interested in whether aspirin has an effect on headache, then the F-test should suffice. If apart from this general effect one wishes to explore whether there are significant differences between the three groups of data, without any prior research hypothesis about this, then one could perform a post hoc analysis of the three means. See the relevant chapter on how to perform planned comparisons and post hoc tests.

Now recall that we mentioned an intraclass correlation, or ICC. An intraclass correlation indicates how much clustering there is within the groups, in this case, clustering of headache scores within NY Times readers. How much are the three scores alike that come from the same patient? This correlation can be computed on the basis of the SPSS output, using the following formula:

$$ICC = \frac{\sigma^2_{patient}}{\sigma^2_{patient} + \sigma^2_e} \tag{11.32}$$

Here, the variance of the **patient** random effects is equal to 28.3, and the variance of the residuals e is equal to 8.5, so the intraclass correlation for the headache severity scores is equal to

$$ICC = \frac{28.3}{28.3 + 8.5} = 0.80 \tag{11.33}$$

As this correlation is quite higher than 0, there seems to be quite a lot of clustering. Therefore it's a good thing that we used random effects for the individual differences in headache scores among NY Times readers. Had this correlation been 0 or very close to 0, however, then it would not have mattered to include these random effects. In that case, we might as well use an ordinary linear model, using the UNIANOVA syntax for example. Note from the formula that the correlation becomes 0 when the variance of the random effects for patients is 0.

11.3.1 Exercises

Suppose you let a sample of students do a math test in three different rooms: one with yellow walls, one with red walls and one with blue walls. All students do the math test three times, once in every room. The data are as follows:

student	colour	score
001	yellow	60
001	red	66
001	blue	60
002	yellow	24
002	red	15
002	blue	30
003	yellow	90
003	red	90
003	blue	89
004	yellow	10
004	red	20
004	blue	15
005	yellow	23
005	red	13
005	blue	18
...	...	...

- 1. If you want to test the hypothesis that the colour of the walls do not affect math test scores, and at the same time you want to take into account that some students are generally better at math than others, what

would the SPSS syntax be?

- 2. In the output that would result from that syntax from question 1, would you look at a t-test or or an F-test? Explain your answer.
- 3. How many degrees of freedom would you see for the denominator?

- 4. Suppose you see this in the output for this colour experiment. How important are the individual difference in math performance in the population of students? Can you quantify the amount of clustering?

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Residual		269.529502	50.050372
Intercept [subject = student]	Variance	228.263602	85.188559

a. Dependent Variable: mathscore.

Answers:

- 1.

```
MIXED score BY colour
  /FIXED=colour
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(student) COVTYPE(VC).
```

- 2. F-test. There will be two dummy variable and I want to know if the effects of both of these are significantly different from 0. The t-tests give me only information about the dummy variables separately.

- 3. 2, because there are 3 different colours, which can be represented by 2 dummy-variables.

- 4. In the table with the data you generally see that students who score high in one room also score high in another room (for instance, students 001 and 003). Students who score low in one room also score low in another room (for instance students 002, 004 and 005). This clustering can be quantified using an intraclass correlation, in this case equal to $\frac{228}{228+270} = 0.46$.

11.4 Pre-mid-post intervention design: linear effects

In the previous section, we've looked at *qualitative* variables: **measure** (pre intervention, 3 hours after, and 24 hours after), or **colour** (yellow, red, and blue rooms). We can use the same type of analysis for *quantitative* variables. In fact, we could have used a linear effect for time in the headache example: using time of measurement as a variable. Let's look at the headache data again. But now we've created a new variable **time** that is based on the measure **variable**: all first measurements are coded as **time**=0, all second measurements after 3 hours are coded as **time**=3, and all third measurements after 24 hours are coded as **time**=24.

patient	measure	headache	time
1	1	52	0
1	2	45	3
1	3	47	24
2	1	59	0
2	2	50	3
2	3	55	24
3	1	65	0
3	2	56	3
3	3	58	24
4	1	51	0

Instead of using a qualitative variable intervention, with three levels, we now use a quantitative variable, time, indicating the number of hours that have elapsed after aspirin intake. At point 0 hours, we measure headache severity, and patients take an aspirin. Next we measure headache after 3 hours and 24 hours. Above, we wanted to know if there were differences in average headache between before intake and 3hrs and 24 hrs after intake. Another question we might ask ourselves: is there a *linear* reduction in headache severity after taking aspirin?

For this we can do a linear regression type of analysis. We want to take into account individual differences in headache severity levels among patients, so we perform a MIXED analysis in SPSS, using the following syntax, replacing the key word BY with WITH, and the variable **measure** by **time**:

```
MIXED headache WITH time
  /FIXED=time
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(patient) COVTYPE(VC).
```

Below we see the corresponding output:

Based on the output, we see that the model for our data is equivalent to

$$headache_{ij} = 54 + patient_i - 0.16 \times time + e_{ij} \tag{11.34}$$

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	147.032	7811.150	.000
time	1	199.000	26.818	.000

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	54.791316	.619947	147.032	88.381	.000	53.566159	56.016473
time	-.155702	.030066	199.000	-5.179	.000	-.214991	-.096413

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		30.915686	3.099327
Intercept [subject = patient]	Variance	20.806006	4.541038

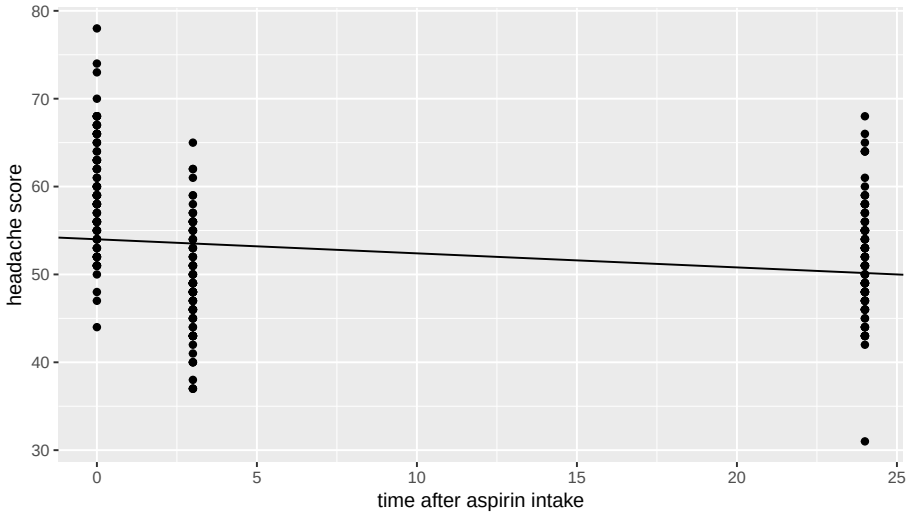
$$patient_i \sim N(0, 21) \tag{11.35}$$

$$e_{ij} \sim N(0, 31) \tag{11.36}$$

This model predicts that at time 0, the average headache severity score equals 54, and that for every hour after intake, the headache level drops by 0.16 points. So it predicts for example that after 10 hours, the headache has dropped 1.6 points to 52.4.

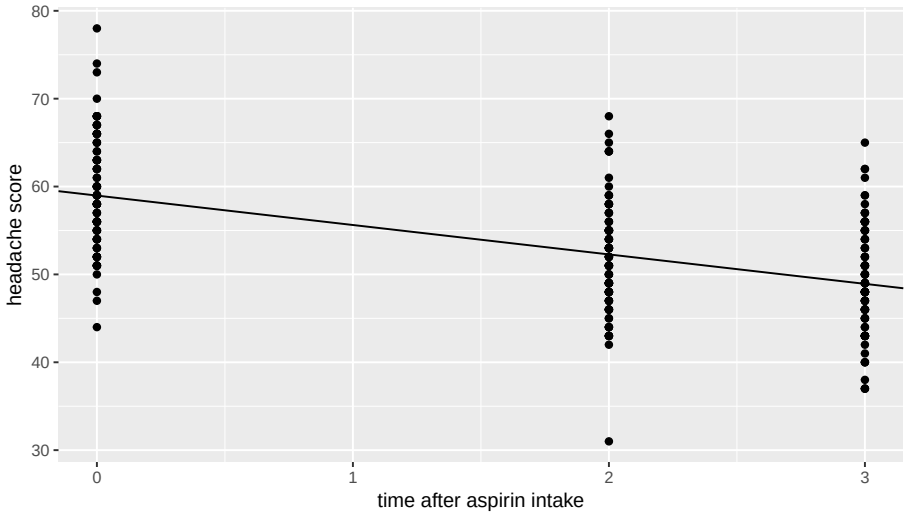
Is this a good model for the data? Probably not, look at the variance of the residuals: with its 31 it is now a lot bigger than in the previous analysis with the same data (see previous section). Larger variance of residuals means that the model explains the data worse: predictions are worse, so the residuals increase in size.

That the model is not appropriate for this data set is also obvious when we plot the data, focusing on the relationship between time and headache levels:



The line shown is the fitted line based on the SPSS output. It can be seen that the prediction for time=0 is too low, for time=2 too high, and for time=24 again too low. So for this particular data set on headache, it would be better to use a qualitative predictor for the effect of time on headache, like we did in the previous section.

As an example of a data set where a linear effect would have been appropriate, imagine that we measured headache 2 hours and 3 hours after aspirin intake (but not after 24 hours). Suppose these data would look like this:



Here we see a gradual increase of headache levels right after aspirin intake. Here, a quantitative treatment of the time variable would be more appropriate. The SPSS output is given below.

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	133.169	9571.257	.000
time	1	199.000	596.514	.000

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	58.972143	.602785	133.169	97.833	.000	57.779871	60.164414
time	-3.349286	.137133	199.000	-24.424	.000	-3.619706	-3.078865

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		8.775878	.879790
Intercept [subject = patient]	Variance	28.185942	4.431672

From the output we see that the intercept is 59 and that the slope is -3.3. So this model predicts an hourly decrease of 3.3 points in headache level. This regression line is also depicted in the above figure.

Because we are confident that this model is appropriate for our data, we can interpret the statistical output from SPSS.

> A linear mixed model was run, using a quantitive variable time and random effects for the variable patient. We saw a significant linear effect of time on headache level, $t(199) = -24.42, p < 0.001$. The estimated effect of time based on this analysis is negative, -3.3, so with every hour that elapses after aspirin intake, the predicted headache score decreases with 3.3 points.

11.4.1 Exercises

Suppose you have a number of CEOs with smart watches and you have these smart watches log skin conductance. Skin conductance is a good measure for stress. These measurements are done at random intervals, for at most 4 times during one day. The experiment starts at 7am and stops at 7pm. The **time** variable measures how many hours have passed since 7am.

CEO	time	conductance
001	2	80
001	3	65
001	10	60
001	11	60
002	4	34
002	6	25
002	9	30
002	12	30
003	3	23
003	4	15
003	5	20
003	8	20
004	0	90
004	3	70
004	4	65
004	11	65
...	...	...

Now you'd like to know if skin conductance in CEOs shows a general decrease during the day. Your null-hypothesis is therefore that there is no linear effect of time on skin conductance. Now, you have multiple measures for each CEO (repeated measures), and there might be individual differences in the average skin conductance that you would like to take into account. Therefore you perform a MIXED analysis in SPSS.

- 1. What would the SPSS syntax look like?

- 2. If you got the following output, what would your predicted skin conductance be for a CEO at 15.00 hrs?

- 3. Look at the data plotted: do you think a linear effect is reasonable for this data set?

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	57.524	242.222	.000
time	1	59.000	17.981	.000

a. Dependent Variable: conductance.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	61.738462	3.966881	57.524	15.563	.000	53.796487	69.680436
time	-4.130769	.974140	59.000	-4.240	.000	-6.080020	-2.181519

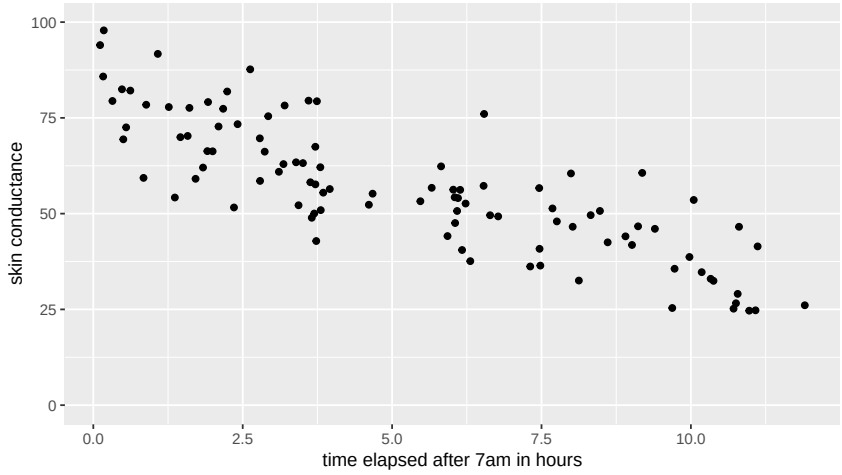
a. Dependent Variable: conductance.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Residual		246.726901	45.426105
Intercept [subject = CEO]	Variance	234.846972	84.637232

a. Dependent Variable: conductance.



- 4. How much clustering is there for skin conductance across CEOs?

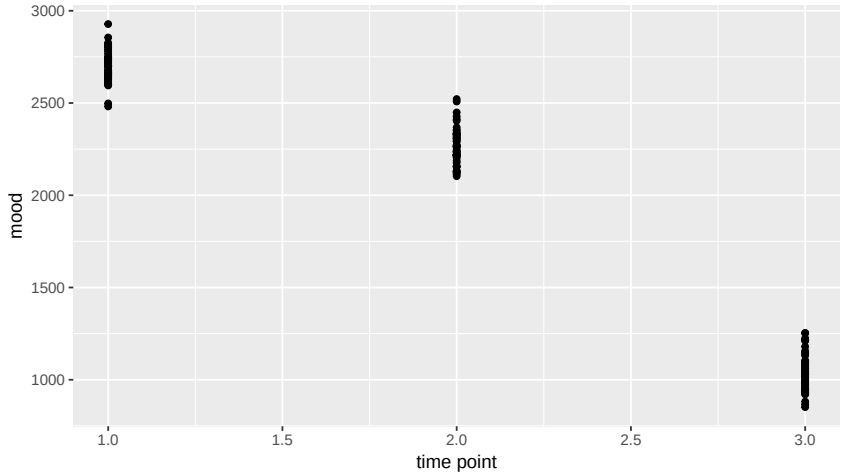
- 5. Would you say these individual differences are very important to take into account?

- 6. Is there a significant effect of time of day on skin conductance in CEOs?

- 7. What is the effect of time of day on skin conductance in CEOs? Also give the 95% confidence interval of this effect.

- 8. Write a short paragraph that describes the results in APA format.

- 9. Given a new data set where every student's mood was tested at three points in time: During Christmas holidays (time points 1), during Easter holidays (time point 2) and at the start of the academic year, September 1 (time point 3). Look at the data plotted: do you think a linear effect is reasonable for this data set? Explain your answer.



- 10. Provide the syntax you would use to analyse the problem of question 9.

Answers:

- 1.

```
MIXED conductance WITH time
  /FIXED=time
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(CEO) COVTYPE(VC).
```

- 2. 15 hrs is equal to 8 hours after 7am, so the expected skin conductance is equal to $62 - 4 \times 8 = 30$

- 3. Yes, a general linear downward trend is observed for the skin conductance.

- 4. The intraclass correlation coefficient is equal to $\frac{235}{235+247} = 0.49$,
- 5. The correlation is quite different from 0, so there is certainly some clustering in the data and it is important to take these individual differences into account.
- 6. Yes, there is a signficant linear effect of time on skin conductance in CEOs, $t(59) = -4.24, p < 0.01$.
- 7. The linear effect of time of day on skin conductance in CEOs is around -4.13 points per hour after 7am (95 % CI: -6.08 – -2.18).
- 8.

  > A linear mixed model was run with time as a quantitative predictor for skin conductance, including random effects for CEO. We found an effect of time of -4.13 points per hour which was significantly different from 0, $t(59) = -4.24, p < 0.001$. Therefore we conclude that time of day has an effect on skin conductance in the entire population of CEOs.

- 9. The relationship is not linear: you cannot draw a straight line through the means of the three measurements.
- 10. Because we have multiple measurements from the same students we should use a MIXED analysis. Furthermore, a qualitave analysis would be more suitable, given the nonlinear relationship between time and mood. So we use the syntax:

```
MIXED mood BY time
  /FIXED=time
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(student) COVTYPE(VC).
```

11.5 Linear mixed models and interaction effects

Suppose we carry out the aspirin and headache study not only with a random sample of NY Times readers that suffer from regular headaches, but also with a random sample of readers of the Wall Street Journal that suffer from regular headaches. We'd like to know whether aspirin works, but we are also interested to know whether the effect of aspirin is similar in the two groups of readers. Our null-hypothesis is that the effect of aspirin in affecting headache severity is the same in NY Times and Wall Street Journal readers that suffer from headache.

H_0: The effect of aspirin is the same for NY Times readers as for Wall Street Journal readers.

Suppose we have the following data set (we only show the first six patients), and we only look at the measurements before aspirin intake and 3 hours after aspirin intake (pre-post design):

patient	group	pre	post
1	NYTimes	55	45
2	WallStreetJ	63	50
3	NYTimes	66	56
4	WallStreetJ	50	37
5	NYTimes	63	50
6	WallStreetJ	65	53

In this part of the data set, patients 2, 4, and 6 read the Wall Street Journal, and patients 1, 3 and 5 read the NY Times. We assume that people only read one of these newspapers. We measure their headache before and after the intake of aspirin (a pre-post design). The data are now in what we call *wide format*: the dependent variable **headache** is spread over two columns, **pre** and **post**. In order to analyse the data with linear models, we need them in *long format*, like this:

patient	group	measure	headache
1	NYTimes	1	55
1	NYTimes	2	45
2	WallStreetJ	1	63
2	WallStreetJ	2	50
3	NYTimes	1	66
3	NYTimes	2	56

The new variable **measure** now indicates whether a given measurement of headache refers to a measurement before intake (first measurement) or after intake (second measurement). Again we could investigate whether there is an effect of aspirin with a linear mixed model, with **measure** as our qualitative predictor, but that is not really what we want to test: we only want to know whether the effect of aspirin (being small, large, negative or non-existent) *is the same for both groups*. Remember that this hypothesis states that there is no interaction effect of aspirin (**measure**) and group. The null-hypothesis is that group is *not* a moderator of the effect of aspirin on headache. There may be an effect of aspirin or there may not, and there may be an effect of newspaper (**group**) or there may not, but we're interested in the *interaction* of aspirin and group membership. Is the effect of aspirin different for NY times readers than for Wall Street Journal readers?

In our analysis we therefore need to specify an interaction effect. Since the data are clustered (2 measures per patient), we use a linear *mixed* model. First we show how to analyse these data using dummy variables, later we will show the results using a different approach.

We recode the data into two dummy variables, one for the aspirin intervention (measure), and one for group membership.

```
RECODE measure (1=0) (2=1) INTO post.
RECODE group ('WallStreetJ'=0) ('NYTimes'=1) INTO NYTimes.
EXECUTE.
```

Next we need to compute the product of these two dummies to code for the interaction effect. Since with the above dummy coding, all post measures get a 1, and all NYTimes readers get a 1, only the observations that are post aspirin and that are from NYTimes readers get a 1 for the product, the interactiondummy . That's why it is best to name this interaction effect PostNYTimes.

```
COMPUTE PostNYTimes=post*NYTimes.
EXECUTE.
```

With these three new dummy variables we can specify the linear mixed model.

```
MIXED headache WITH post NYTimes PostNYTimes
  /FIXED= post NYTimes PostNYTimes
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(patient) COVTYPE(VC).
```

In the output below, we recognize the three fixed effects for the three dummy variables. Since we're interested in the interaction effect, we look at the effect of PostNYTimes. The effect is in the order of +0.6. So what does this mean?

Remember that a reader from the Wall Street Journal gets a 0 for the group dummy **NYTimes**. All headache measures before aspirin intake are given a 0 for the intervention dummy **post**. Since the product of 0×0 equals 0, all these measures before aspirin in Wallstreet Journal readers get a 0 for the interaction dummy **PostNYTimes**. Therefore, the intercept of 59.5 refers to the expected headache severity of Wall Street Journal readers *before* they take their aspirin. This is significantly different from zero, meaning that in the population of Wall Street Journal readers, headache before aspirin intake is different from zero.

Furthermore, we see that the effect of the intervention is -10.7. So, relative to Wall Street Journal readers prior to aspirin intake, the level of post intake headache is 10.7 points *less*. So in the population of Wall Street Journal readers, the effect of aspirin is different from 0, since the effect of -10.7 is significant.

If we look further down in the table, we see the effect of NYTimes equals 0.32. So, relative to Wall Street Journal readers, before aspirin intake (the reference group), NY Times readers score on average 0.32 points higher on the headache scale before aspirin intake.

However, we're not interested in a general difference between those two groups of readers, we're interested in the effect of aspirin and whether it is different in the two groups of readers. In the last row we see the interaction

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	123.427	4966.920	.000
post	1	98.000	341.596	.000
NYTimes	1	123.427	.072	.789
PostNYTimes	1	98.000	.541	.464

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	59.520000	.844538	123.427	70.476	.000	57.848346	61.191654
post	-10.660000	.576768	98.000	-18.482	.000	-11.804577	-9.515423
NYTimes	.320000	1.194357	123.427	.268	.789	-2.044076	2.684076
PostNYTimes	.600000	.815673	98.000	.736	.464	-1.018677	2.218677

Estimates of Covariance Parameters

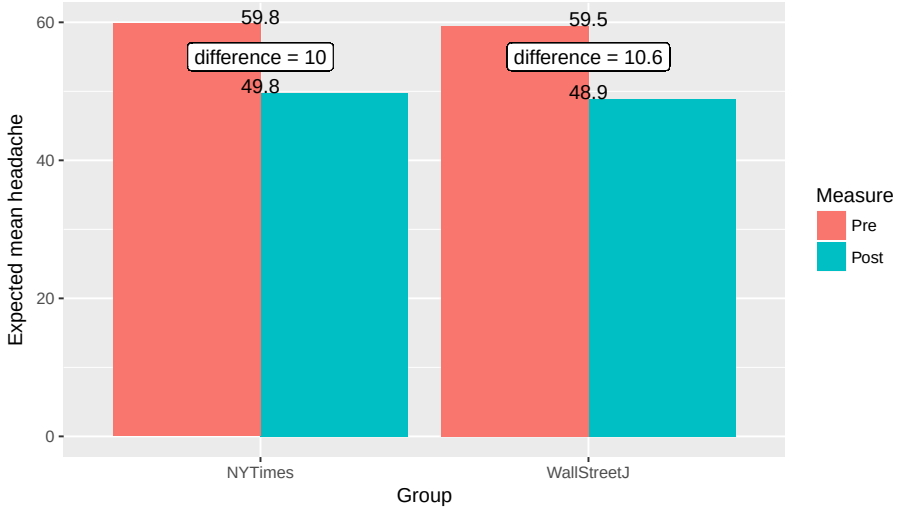
Parameter		Estimate	Std. Error
Residual		8.316531	1.188076
Intercept [subject = patient]	Variance	27.345714	4.539603

effect: being a reader of the NY Times AND at the same time being a measure after aspirin intake, the expected increase in mean headache equals 0.60. So the effect of aspirin is -10.7 in Wall Street Journal readers, as we saw above, but the effect is $-10.7 + 0.6 = -10.1$ in NY Times readers. So in this sample the effect of aspirin on headache is 0.6 *smaller* than in Wall Street Journal readers (note that even while the interaction effect is positive, it is positive on a scale where a high score means more headache).

Let's look at it in the different way, using a table with the dummy codes. For each group of data, pre or post aspirin and New York Times readers and Wall Street Journal readers, we note the dummy codes for the new dummy variables. In the last column we use the output estimates and multiply them with the respective dummy codes (1 and 0) to obtain the expected headache level (using rounded numbers):

measure	group	post	NYTimes	PostNYT	exp mean
pre	WallStreet	0	0	0	60
post	WallStreet	1	0	0	$60 + (-11) = 49$
pre	NYtimes	0	1	0	$60 + 0.3 = 60.3$
post	NYtimes	1	1	1	$60 + (-11) + 0.3 + 0.6 = 49.9$

The exact numbers are displayed in the graph below:



We see that the specific effect of aspirin in NYTimes readers is 0.6 smaller than the effect of aspirin in Wall Street Journal readers. This difference in the effect of aspirin between the groups was not significantly different from 0. The null-hypothesis that the effect is the same in the two populations of readers cannot be rejected. We therefore conclude that the effect that aspirin has on patients is the same for NY Times and Wall Street Journal readers.

Note that we could have done the analysis in another way, not treating the variables in a quantitative way and using dummy variables, but by treating them qualitatively using the key word BY. The SPSS syntax would then be:

```
MIXED headache BY measure group
  /FIXED=measure group measure*group
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(patient) COVTYPE(VC).
```

The output would then look like below:

Here SPSS has automatically created dummy variables, one for **measure=1**, one for **group=1**, and one for the interaction effect, **group=1 AND measure=1**. Because the dummy coding is different, the intercept and the main effects of group and measure have changed, but you see that the interaction effect is still 0.6, albeit now negative. We also see that the significance level of the interaction effect is still the same. You are always free to choose to either construct your own dummy variables and analyze them in a quantitative way (using WITH), or to let SPSS construct the dummy variables for you (using BY): the p-value for the interaction effect will always be the same (this is not true for the intercept and the main effects).

Because the two analyses are equivalent (they end up with exactly the same predictions, feel free to check!), we can safely report that we've found a non-significant group by measure interaction effect, $t(98) = 0.74, p = 0.46$. We

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	98.000	9428.174	.000
measure	1	98.000	645.279	.000
group	1	98.000	.305	.582
measure * group	1	98.000	.541	.464

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	48.860000	.844538	123.427	57.854	.000	47.188346	50.531654
[measure=1.00]	10.660000	.576768	98.000	18.482	.000	9.515423	11.804577
[measure=2.00]	0[b]	0	.	.	.	.	.
[group=1.00]	.920000	1.194357	123.427	.770	.443	-1.444076	3.284076
[group=2.00]	0[b]	0	.	.	.	.	.
[measure=1.00] * [group=1.00]	-.600000	.815673	98.000	-.736	.464	-2.218677	1.018677
[measure=1.00] * [group=2.00]	0[b]	0	.	.	.	.	.
[measure=2.00] * [group=1.00]	0[b]	0	.	.	.	.	.
[measure=2.00] * [group=2.00]	0[b]	0	.	.	.	.	.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		8.316531	1.188076
Intercept [subject = patient]	Variance	27.345714	4.539603

therefore conclude that in the populations of NY Times readers and Wall Street Journal readers, the short-term effect of aspirin on headache is the same.

11.5.1 Exercises

Below we see data from a study on the effects of the financial crisis on the number of employees in specific Dutch companies. The companies are distinguised into food and non-food related companies. The number of employees are recorded in January 2008 and January 2011.

company	food	2008	2011
1	nonfood	42	63
2	food	104	126
3	nonfood	76	58
4	food	65	131

1. These data are in wide format. Rewrite the datamatrix in such a way that

we have the same data in long format. Provide column (variable) names.

...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

2. Do we need to use a linear mixed model, or can we analyse these data with an ordinary linear model?

3. We want to test the null-hypothesis that the effects of the financial crisis in 2008 has the same effect on the number of employees in the food sector as in the non-food sector. Provide the syntax that helps you test this hypothesis.

4. Suppose the following output results from an analysis done by a colleague:

   She provides you with the information that food=1 means the food sector and food=2 is the nonfood sector.

   What does the model predict regarding the number of employees in 2008 in the non-food sector?

5. What does the model predict regarding the number of employees in 2011 in the non-food sector?

6. What does the model predict regarding the number of employees in 2008 in the food sector?

7. What does the model predict regarding the number of employees in 2011 in the food sector?

8. How large is the effect of the crisis in the food sector?

9. How large is the effect of the crisis in the non-food sector

10. How large is the intraclass correlation (ICC)? Give the computation.

11. Could we have done the analysis with an ordinary linear model? Explain your answer.

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	998.000	70192.133	.000
food	1	998.000	3389.819	.000
year	1	998	1122.117	.000
food * year	1	998	.437	.509

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	81.574000	.664366	1989.879	122.785	.000	80.271074	82.876926
[food=1.00]	39.312000	.939556	1989.879	41.841	.000	37.469384	41.154616
[food=2.00]	0[b]	0	.	.	.	.	.
[year=2008.00]	-22.056000	.913130	998	-24.154	.000	-23.847874	-20.264126
[year=2011.00]	0[b]	0	.	.	.	.	.
[food=1.00] * [year=2008.00]	.854000	1.291360	998	.661	.509	-1.680093	3.388093
[food=1.00] * [year=2011.00]	0[b]	0	.	.	.	.	.
[food=2.00] * [year=2008.00]	0[b]	0	.	.	.	.	.
[food=2.00] * [year=2011.00]	0[b]	0	.	.	.	.	.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		208.451418	9.331567
Intercept [subject = company]	Variance	12.239802	6.996594

12. Can we reject the null-hypothesis that the effects of the crisis were the same in the food and non-food sectors? Explain your answer.

Answers:

1. It could look like this:

company	sector	year	NEmployees	...
1	nonfood	2008	42	...
1	nonfood	2011	63	...
2	food	2008	104	...
2	food	2011	126	...
3	nonfood	2008	76	...
3	nonfood	2011	58	...
4	food	2008	65	...
4	food	2011	131	...

2. The data are clustered into companies: for each company we have two data points, so we should at least try a linear mixed model. Only if the variance of the company random effects is extremely small, we could use a linear model without random effects.

3. One option is to let SPSS construct the dummy variables:

```
MIXED employees BY year sector
  /FIXED=year sector year*sector
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(company) COVTYPE(VC).
```

   Or you do the dummy coding yourself, for example like this:

```
RECODE year (2008=0) (2011=1) INTO year2011.
RECODE sector ('Nonfood'=0) ('food'=1) INTO food.
EXECUTE.

COMPUTE food2011=year2011*food.
EXECUTE.

MIXED employees WITH year2011 food food2011
  /FIXED= year2011 food food2011
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(company) COVTYPE(VC).
```

4. the nonfood sector is food=2, so the predicted number of employees in 2008 in the nonfood sector is equal to $81.57 + 0 - 22.056 + 0 = 59.514$

5. the nonfood sector is food=2, so the predicted number of employees in 2011 in the nonfood sector is equal to $81.57 + 0 + 0 + 0 = 81.57$

6. the food sector is food=1, so the predicted number of employees in 2008 in the food sector is equal to $81.57 + 39.31 - 22.056 + 0.85 = 99.674$

7. the food sector is food=1, so the predicted number of employees in 2011 in the food sector is equal to $81.57 + 39.31 + 0 + 0 = 120.88$

8. in the food sector the effect is a $120.88 - 99.674 = 21.206$ increase in number of employees

9. in the non-food sector the effect is a $81.57 - 59.514 = 22.056$ increase in number of employees

10. the ICC is $\frac{12}{12+208} = 0.05$

11. we have clustering, with multiple data point per company, so in general a linear mixed model is better than an ordinary linear model. However, since the intraclass correlation is rather low, the results would be very similar if we would use an ordinary linear model.

12. The null-hypothesis cannot be reject as the year by sector interaction effect is not signifcantly different from 0, $t(998) = 0.66, p = 0.51$. (alternatively, $F(1, 998) = 0.44, p = 0.51$). Note however that the statistical results are in terms of absolute number of employees. These data show that the average number of employees in 2008 is larger in the food sector than in the non-food sector. Perhaps it would be wiser to look at percentage increase in number of employees: A change from 100 to 102 reflects a larger impact than a change from 1000 to 1002.

11.6 Mixed designs

The design in the previous section where we had both a grouping variable and a pre-post or repeated measures design, is often called a *mixed design*. It is a mixed design in the sense that there are two kinds of variables: one is a *between-individuals* variable, and one variable is a *within-individual* variable. Here the between-individuals variable is **group**: two different populations of readers. It is called *between* because one individual can only be part of one group. When we study the effect of the group effect we are essentially comparing the scores of one group of individuals with the scores of another group of individuals, so the comparison is *between different individuals*. The two groups of data are said to be *independent*, as we knew that none of the readers in this data set reads both journals.

The within-variable in this design is the aspirin intervention, indicated by the variable **measure**. For each individual we have two observations: all individuals are present in both the pre condition data as well as in the post condition data. With this intervention variable, we are comparing the scores of a group of indiviudals with the scores *of that same group of individuals* at another time point. The comparison of scores is within a particular individual, at timepoint 1 and at timepoint 2. So the pre and post sets of data are not independent: the headache scores in both conditions are coming from the same individuals.

Mixed designs are often seen in psychological experiments. For instance, you want to know how large the effect of alcohol intake is on driving performance. You want to know whether the effect of alcohol on driving performance is the same in a Fiat 600 as in a Porsche. Suppose you have 100 participants for your study. There are many choices you can make regarding the design of your study. Here we discuss 4 alternative research designs:

1. One option is to have all participants participate in all four conditions: they all drive a Fiat with and without alcohol, and they all drive a Porsche,

with and without alcohol. In this case, both the car and the alcohol are within-participant variables.

2. The second option is to have 50 participants drive a Porsche, with and without alcohol, and to have the other 50 participants drive the Fiat, with and without alcohol. In this case, the car is the between-participants variable, and alcohol is the within-participant variable.

3. The third option is to have 50 participants without alcohol drive both the Porsche and the Fiat, and to have the other 50 participants drive the Porsche and the Fiat with alcohol. Now the car is the within-participant variable, and the alcohol is the between-participants variable.

4. The fourth option is to have 25 participants drive the Porsche with alcohol, 25 other participants drive the Porsche without alcohol, 25 participants drive the Fiat with alcohol, and the remaining 25 participants drive the Fiat without alcohol. Now both the car variable and the alcohol variable are between-participant variables: none of the participants is present in more than 1 condition.

Only the second and the third design described here are mixed designs, having at least one between-participants variable and at least one within-participant variable.

Remember that when there is at least one within variable in your design, you have to use a linear mixed model. If all variables are between variables, one can use an ordinary linear model. Note that the term *mixed* in linear mixed model refers to the effects in the model that can be both random and fixed. The term *mixed* in mixed designs refers to the mix of two kinds of variables: within variables and between variables.

Also note that the within and between distinction refers to the units of analysis. If the unit of analysis is school, then the location of the school building is a between-school variable. An example of a within-school variable could be time: before a major curriculum reform and after a major curriculum reform.

11.6.1 Exercises

1. A psychologist studies whether age affects math performance. In 2017, she measures math performance (one score) in a group of 80-year-olds and she measures math performance (one score) in a group of 90-year-olds.
   1. In this design, is the age variable a between-participants variable or a within-participant variable?
   2. Would you analyze these data with a linear model, or with a linear mixed model? Explain.

2. A psychologist studies whether age affects math performance. She measures math performance (one score) in a group of 7-year-olds and she

measures math performance again when the same children are 8 years old.
1. In this design, is the age variable a between-participants variable or a within-participant variable?
2. Would you analyze these data with a linear model, or with a linear mixed model? Explain.

3. Look at the data table below.

ID	Nationality	Sex	Mathscore
1	Dutch	Male	67
2	Dutch	Female	88
3	German	Male	50
4	German	Female	98
...	...	...	...

In this data set on Math performance, we see two variables, nationality and sex. What kind of variables are these: within-participant variables or between-participants variables? Explain.
1. Would you call this a mixed design? Explain.
2. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

4. Look at the data table below.

ID	Nationality	Age	Mathscore
1	Dutch	3	67
1	Dutch	5	88
2	German	4	50
2	German	6	98
...	...	...	...

In this data set on Math performance, we see two variables, nationality and age. What kind of variables are these: within-participant variables or between-participants variables? Explain.
1. Would you call this a mixed design? Explain.
2. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

5. Look at the data table below.

ID	Subject	Sex	Mood
1	Psychology	Male	67
1	Psychology	Female	88
2	Sociology	Female	50
2	Sociology	Male	98
...	...	...	...

In this data set on mood in transsexuals, we see two variables, the subject they have a Master's degree in, and sex. What kind of variables are these: within-participant variables or between-participants variables? Explain.
1. Would you call this a mixed design? Explain.
2. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

6. Look at the data table below.

SchoolID	Country	Year	Avarage Mathscore
1	The Netherlands	2010	67
1	The Netherlands	2011	88
1	The Netherlands	2012	50
1	The Netherlands	2013	98
2	Germany	2010	67
2	Germany	2011	88
2	Germany	2012	50
2	Germany	2013	98
...	...	...	...

In this data set on average Math performance in schools, we see two variables, country of the school and year of data collection. What kind of variables are these: within-school variables or between-schools variables? Explain.
1. Would you call this a mixed design? Explain.
2. Would you analyze this data set with a linear model or with a linear mixed model? Explain.

Answers:

1. 1. The age variable is a between-participants variable: some of the participants are 80 years old and some are 90 years old: none are both at the same time. Age discriminates between two sets of participants, so it is a between-participants variable. 2. Two groups of participants were studied. Because we only have one measure for each participant, there is no clustering, and we use an ordinary linear model.

2. 1. The age variable is a within-participants variable: children are studied twice and scores can therefore be compared within an individual. 2. One

group of participants was studied and for each participant we have two math scores. Because we have more than one measure for each participant, we have to use a linear mixed model to account for clustering.

3. Each participant is either Dutch or German. This is a between-participants variable. Each participant is either male or female, sex discriminates between separate groups of participants, so sex is a between-participants variable. 1. This is *not* a mixed design as it does not have both within-participant and between-participants independent variables. 2. Because we only have one measure for each participant, there is no clustering, and we use an ordinary linear model.

4. Each participant is either Dutch or German. This is a between-participants variable. On measurement 1 participants have a different age than on measurement 2. This is a within-participant variable. 1. This is a mixed design as it has both a within-participant and a between-participants independent variable. 2. For each participant we have two math scores, so we would have to use a linear mixed model to account for clustering.

5. Each participant has only one Masters degree. This is a between-participants variable. Between the two measurements, participants change their sex. This is a within-participant variable: we can compare people's mood when they are male and when they are female. 1. This is a mixed design as it has both a within-participant and a between-participants independent variable. 2. For each participant we have two mood scores, so we would have to use a linear mixed model to account for clustering.

6. Each school is based in only one country and has measurements across four years. Country is a between-schools variable and year is a within-school variable. 1. This is a mixed design as it has both a within-school and a between-schools independent variable. 1. For each school we have four average math scores, so we would have to use a linear mixed model to account for clustering.

11.7 Mixed design with a linear effect

In an earlier section we looked at a mixed design where the between variable was **newspaper** and the within variabe was **measure**: pre or post. It was a 2 by 2 design (2×2) design: 2 measures and 2 newspapers, where we were interested in the interaction effect. We wanted to know whether newspaper moderated the effect of aspirin on headache. We used the within variable **measure** in a qualitative way by dummy coding it.

In an earlier section in this chapter we saw that we can also model linear effects in linear mixed models, where we treated the time variable quantitatively: 0hrs, 3hrs after aspirin intake and 24 hrs after intake. Here we will give an example of a 3×20 mixed design: we have a qualitative group (between) variable with 3 levels and a quantitative time (within) variable with 20 levels. The

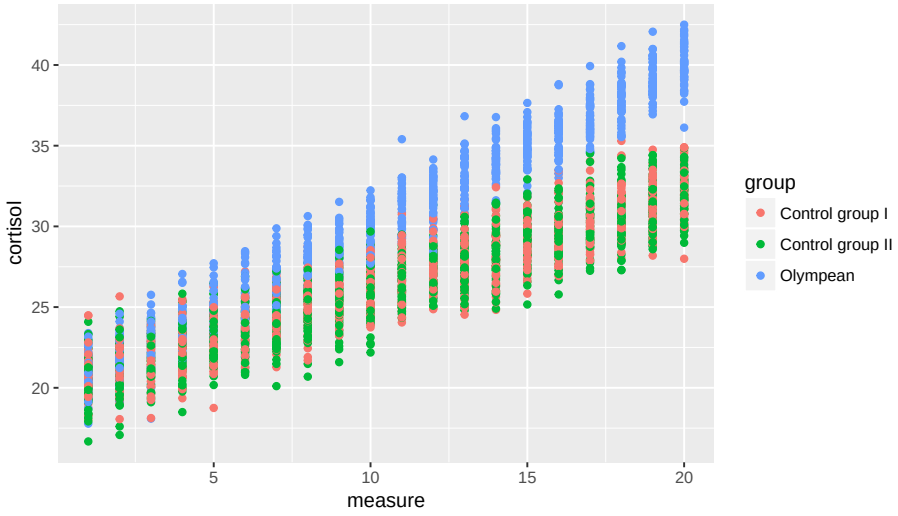
example is about stress in athletes that are going to partake in the 2018 Winter Olympics. Stress can be revealed in morning cortisol levels. In the 20 days preceding the start of the Olympics, each athlete was measured every morning after waking and before breakfast by letting them chew on cotton. The cortisol level in the saliva was then measured in the lab. Our research question is whether cortisol levels rise in athletes that prepare for the Olympics.

Three groups were studied. One group consisted of 50 athletes who were selected to partake in the Olympics, one group consisted of 50 athletes that were very good but were not selected to partake (Control group I) and one group consisted of 50 non-athlete spectators that were going to watch the games (Control group II). The null-hypothesis was that the linear change in cortisol levels during those 20 days was the same for the three groups: the Olympeans, Control group I and Control group II.

Below you see part of the data, the first 6 measurements on person 1 that belongs to the group of Olympeans.

person	group	measure	cortisol
1	Olympean	1	19.41570
1	Olympean	2	20.21298
1	Olympean	3	22.15560
1	Olympean	4	23.10641
1	Olympean	5	23.69396
1	Olympean	6	21.91889

When we plot the data, and use different colours for the three different groups, we already notice that the Olympeans show generally higher cortisol levels, but particulary at the end of the 20-day period.



So we want to know whether the linear effect of time is moderated by group. Since for every person we have 20 measurements, the data are clustered so we use a linear mixed model. We're looking for a linear effect of time, so we use the WITH keyword to indicate that we want to use the **measure** variable in a

quantitative way. We also use **group** as a predictor, but in a qualitative way, by using the keyword BY, so that SPSS will automatically make dummy variables. Because we're interested in an interaction effect, we include both main effects of **group** and **measure** and their interaction under the DESIGN subcommand. Lastly, we control for individual differences in cortisol levels by introducing a random effect for **person**.

```
MIXED cortisol WITH measure BY group
  /FIXED=measure group measure*group
  /PRINT=DESCRIPTIVES  SOLUTION
  /RANDOM=intercept | SUBJECT(person) COVTYPE(VC).
```

The SPSS output is presented below.

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	197.391	49368.492	.000
measure	1	2847	54256.135	.000
group	2	197.391	1.698	.186
group * measure	2	2847	1857.202	.000

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	19.691068	.155004	197.391	127.036	.000	19.385393	19.996744
measure	1.008885	.005476	2847	184.239	.000	.998148	1.019622
[group=1.00]	.402950	.219208	197.391	1.838	.068	-.029340	.835241
[group=2.00]	.176148	.219208	197.391	.804	.423	-.256142	.608439
[group=3.00]	0[b]	0	.	.	.	.	.
[group=1.00] * measure	-.411896	.007744	2847	-53.188	.000	-.427081	-.396711
[group=2.00] * measure	-.405513	.007744	2847	-52.364	.000	-.420698	-.390328
[group=3.00] * measure	0[b]	0	.	.	.	.	.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters

Parameter		Estimate	Std. Error
Residual		.997033	.026426
Intercept [subject = person]	Variance	.986156	.120850

In the output we see an intercept of 19.7, a slope of 1.0 for the effect of measure, two main effects for the group variable (group3 is the reference group, in this case the Olympeans, see the plot above), and two effects for the interaction

effect (one for control group I and one for control group II). Let's fill in the linear equation based on this output:

$$\begin{aligned} cortisol_{ij} = 19.7 + person_i + 1 \times measure + .4 ContrG1 + \\ 0.18 ContrG2 - .4 ContrG1 \times measure - .4 ContrG2 \times measure + e_{ij} \\ person_i \sim N(0, \sigma_p^2 = 0.99) \\ e_{ij} \sim N(0, \sigma_e^2 = 1.00) \end{aligned}$$

We see a clear intraclass correlation of around $\frac{0.986}{0.986+0.997} = 0.5$ so it's a good thing we've included a random effect for persons. The expected means at various time points and for various groups can be made with the use of the above equation.

It's easier to see what linear effects we have for the three different groups. Filling in the above equation for Control group 1, we get:

$$\begin{aligned} cortisol_{ij} &= 19.7 + person_i + 1 \times measure + .4 - .4 \times measure + e_{ij} \\ &= 20.1 + person_i + 0.6 \times measure + e_{ij} \end{aligned}$$

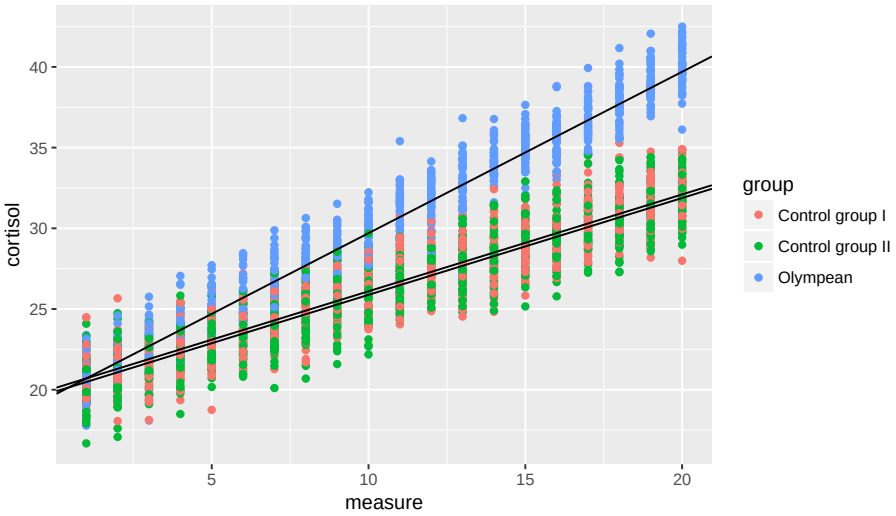
For Control group 2 we get:

$$\begin{aligned} cortisol_{ij} &= 19.7 + person_i + 1 \times measure + 0.18 - .4 \times measure + e_{ij} \\ &= 19.88 + person_i + 0.6 \times measure + e_{ij} \end{aligned}$$

And for the Olympeans we get:

$$cortisol_{ij} = 19.7 + person_i + 1 \times measure + e_{ij} \tag{11.37}$$

In these equations all intercepts are around 20. The slopes are 0.6 in both Control groups I and II, whereas the slope is 1.0 in the group of Olympean athletes. For illustration, these implied linear regression lines are depicted below:



So based on the linear equation, we see that in this sample the rise in cortisol levels is much steeper in Olympeans than in the two control groups. But is this true for all Olympeans and the rest of the populations of high performing athletes and spectators? Note that in the regression table we see two interaction effects: one for **group1*measure** and one for **group2*measure**. Here we're interested in the overall signficance of the interaction effects. That answer we find in the top table with the F-statistics: we see a significant group by measure interaction effect, $F(2, 28) = 18.57, p < 0.001$. The null-hypothesis of the same cortisol change in three different populations can be rejected, and we conclude that Olympean athletes, non-Olympean athletes and spectators show a different change in cortisol levels in the weeks preceding the games.

Chapter 12

Non-parametric alternatives for linear mixed models

12.1 Checking assumptions

In previous chapters we have discussed the assumptions of linear models and linear mixed models: linearity (in parameters), homoscedasticity (equal variance), normal distribution of residuals, normal distribution of random effects (relevant for linear mixed models only), and independence (no clustering unaccounted for).

The problem of nonlinearity can be solved by introducing quadratic terms, for instance by replacing a linear model $Y = b_0 + b_1X + e$ by another linear model $Y = b_0 + b_1X + b_2X^2 + e$.

If we have nonindependence, then you can introduce either an extra fixed effect or a random effect for this clustering. For example, if you see that cars owned by low income families have much more mileage than cars owned by high income families, you can account for this by adding a fixed effect of an income variable as predictor. If you see that average milage is rather similar within municipality but that average mileage can vary quite a lot across municipalities, you can introduce a random effect for municipality (if you have data say from 30 different municipalities).

Unequal variance of residuals and nonnormal distribution of residuals are harder to tackle. Unequal variance can be tackled sometimes by using linear models, but with more advanced options, or by making corrections to p-values that make inference more robust against model violations. Violations of normality are even a bigger problem. Nonnormality can sometimes be solved by using generalized linear models (see next chapter). A combination of nonnormality and unequal variance can sometimes be solved by using a transformation of the data, for instance not analyzing $Y = b_0 + b_1X + e$ but analyzing $log(Y) = b_0 + b_1X + e$ or $\sqrt{Y} = b_0 + b_1X + e$.

If these data transformations or advanced options don't work (or if you're

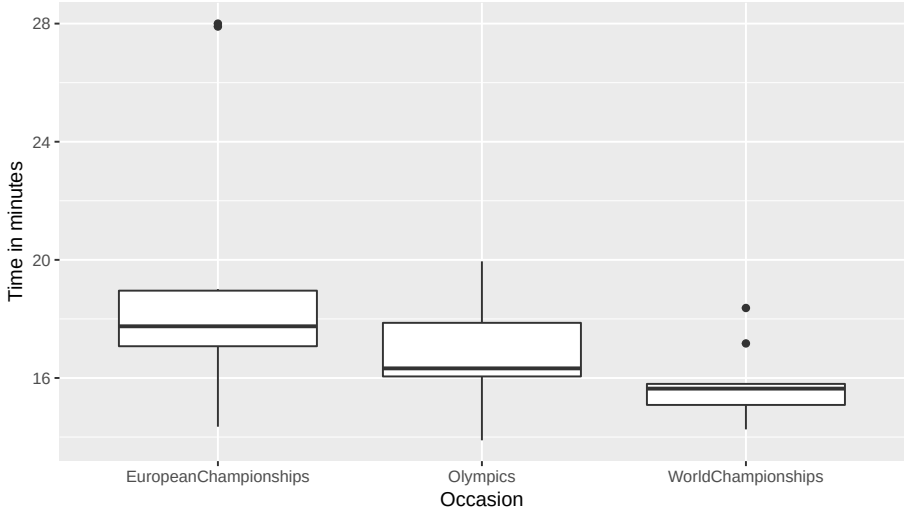


Figure 12.1: Boxplot of the imaginary speed skating data.

not acquainted with them), and your data show nonequal variance and/or non-normally distributed residuals, there are nonparametric alternatives. Here we discuss two: Friedman's test and Wilcoxon's signed rank test. We explain them using an imaginary data set on speedskating.

Suppose we have data on 12 speedskaters that participate on the 10 kilometers distance in three separate championships in 2017-2018: the European Championships, the Winter Olympics and the World Championships. Your friend expects that speedskaters will perform best at the Olympic games, so there she expects the fastest times. So you decide to test the null-hypothesis that average times are the same at the three occasions. In Figure 12.1 we see a boxplot of the data.

In order to test this null-hypothesis, we run a linear mixed model with dependent variable time, and independent variable occasion. We use random effects for the differences in speed across skaters. In Figure 12.2 we see the residuals:

From this plot we clearly see that the assumption of equal variance (homogeneity of variance) is violated: the variance of the residuals in the World-championships condition is clearly smaller than the variance of the European championships condition. From the histogram of the residuals in Figure 12.3 we also see that the distribution of the residuals is not bell-shaped: it is positively skewed (skewed to the right).

Since the assumptions of homogeneity of variance and of normally distributed residuals are violated[1], the results from the linear mixed model cannot be

[1]Remember that assumptions relate to the population not samples: oftentimes your data

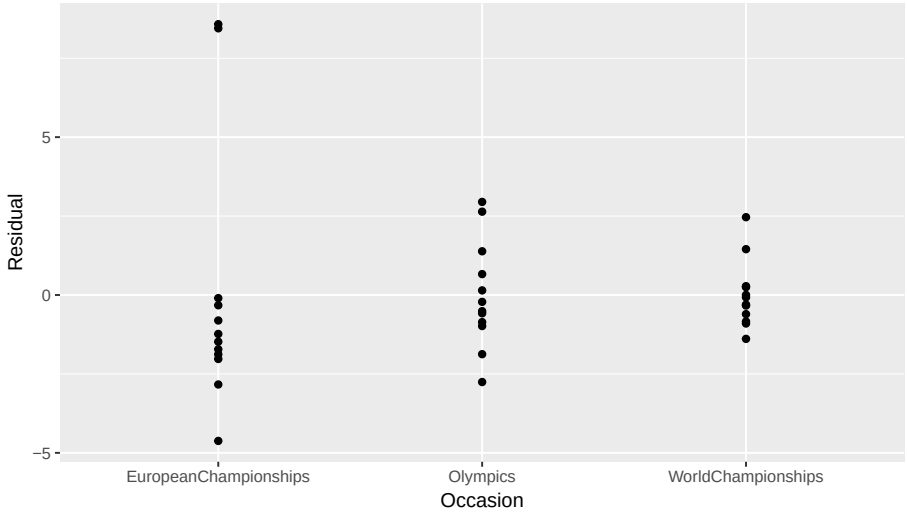


Figure 12.2: Residuals of the speedskating data with a linear mixed model.

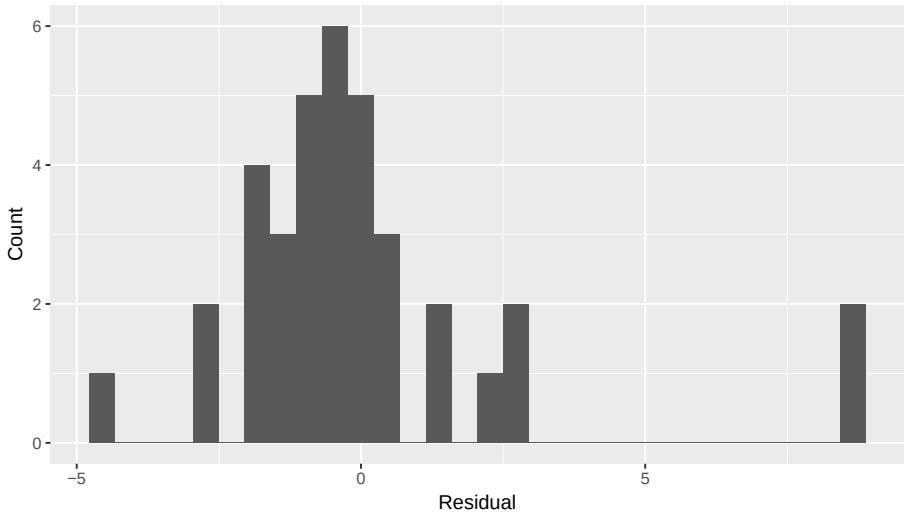


Figure 12.3: Histogram of the residuals of the speedskating data with a linear mixed model.

trusted. In order to answer our research question, we therefore have to resort to another kind of test. Here we discuss Friedman's test, a non-parametric test, for testing the null-hypothesis that the *medians* of the three groups of data are the same. This Friedman test can be used in all situations where you have at least 2 levels of the within variable. In other words, you can use this test when you have data from three occasions, but also when you have data from 10 occassions or only 2. In the following section the Wilcoxon signed ranks test is discussed. This test is often used in social and behavioural sciences. The downside of this test is that it can only handle data sets with 2 levels of the within variable. In other words, it can only be used when we have data from two occassions. Friedman's test is therefore more generally applicable than Wilcoxon's. We therefore advise to always go with the Friedman test, but for the sake of completeness, we will also explain the Wilcoxon test.

12.2 Friedman's test for k measures

Similar to many other nonparametric tests for testing the equality of medians, Friedman's test is based on ranks. Figure **??** shows the speedskating data in wide format.

Table 12.1: The speedskating data in wide format.

athlete	EuropeanChampionships	Olympics	WorldChampionships
1	14.35	16.42	15.79
2	17.36	18.13	14.26
3	19.01	19.95	18.37
4	27.90	17.78	15.12
5	17.67	16.96	17.17
6	17.83	16.15	15.30
7	16.30	19.44	15.63
8	28.00	16.23	15.69
9	18.27	15.76	15.65
10	17.00	16.18	14.99
11	17.10	13.89	15.83
12	18.94	14.83	14.77

We rank all of these time measures by determining the fastest time, then the next to fastest time, etcetera, until the slowest time. But because the data in each row belong together (we compare individuals with themselves), we do the ranking *row-wise*. For each athlete separately, we determine the fastest time (1), the next fastest time (2), and the slowest time (3) and put the ranks in a table.

set is too small to say anything about assumptions at the populationlevel. Residuals for a data set of 8 persons might show very nonnormal residuals, or very different variances for two subgroups of 4 persons each, but that might just be a coincidence, a random result because of the small sample size. If in doubt, it is best to use nonparametric methods.

Table 12.2: Row-wise ranks of the speedskating data.

athlete	EuropeanChampionships	Olympics	WorldChampionships
1	1.00	3.00	2.00
2	2.00	3.00	1.00
3	2.00	3.00	1.00
4	3.00	2.00	1.00
5	3.00	1.00	2.00
6	3.00	2.00	1.00
7	2.00	3.00	1.00
8	3.00	2.00	1.00
9	3.00	2.00	1.00
10	3.00	2.00	1.00
11	3.00	1.00	2.00
12	3.00	2.00	1.00

From this table we see for example that athlete 1 had the fastest time on the European Championships (14.35, rank 1) and the slowest at the Olympics (16.42, rank 3).

Next we compute the sum of the ranks column-wise: the sum of the ranks for the European Championships data is 31, for the Olympic data it's 26 and for the World Championships data it is 15.

From these sums we can gather that in general, these athletes showed their best times (many rank 1s) at the World Championships, as the sum of the ranks is lowest. We also see that in general these athletes showed their worst times (many rank 2s and 3s) at the European Championships, as the relevant column showed the highest sum of ranks.

In order to know whether these sums of ranks are significantly different from eachother, we may compute an F_r-value based on the following formula:

$$F_r = \left[\frac{12}{Nk(k+1)}\Sigma_{j=1}^{k} S_j^2\right] - 3N(k+1) \tag{12.1}$$

In this formula, N stands for the number of rows (12 athletes), k stands for the number of columns (3 occasions), and S_j^2 stands for the squared sum of column j (31^2, 26^2 and 15^2). If we fill in these numbers, we get:

$$\begin{aligned} F_r &= \left[\frac{12}{12 \times 3(3+1)} \times (31^2 + 26^2 + 15^2)\right] - 3 \times 12(3+1) \\ &= \left[\frac{12}{144} \times 1862\right] - 144 = 11.17 \end{aligned}$$

What can we tell from this F_r-statistic? In order to say something about significance, we have to know what values are to be expected under the null-hypothesis that there are no differences across the three groups of data. Suppose

we randomly mixed up the data by taking all the speedskating times and randomly assigning them to the three contests and the twelve athletes, until we have a newly filled datamatrix in Table 12.3:

Table 12.3: The raw skating data in random order.

athlete	EuropeanChampionships	Olympics	WorldChampionships
1	18.37	15.79	17.83
2	15.12	14.83	17.67
3	14.35	14.99	15.63
4	14.26	17.00	17.36
5	19.01	16.30	17.17
6	16.23	15.30	14.77
7	15.83	15.69	27.90
8	15.76	19.44	13.89
9	15.65	18.27	16.18
10	28.00	17.78	16.15
11	19.95	16.42	17.10
12	18.13	16.96	18.94

If we then compute F_r for these mixed up data, we get another value. If we do this say 1000 times, we get the following values for F_r, summarized in the histogram in Figure 12.4.

So if the data is just randomly distributed over the three columns in the data matrix, we expect no systematic differences and so the null-hypothesis is true. So now we know what the distribution of F_r looks like when the null-hypothesis is true. Remember that for the true data that we actually gathered, we found an F_r-value of 11.17. From the histogram, we see that only very few values of 11.17 or larger are observed when the null-hypothesis is true. If we look more closely, we find that only 0.4% of the values are larger than 11.17, so we have a p-value of 0.004. The 95th percentile of these 1000 F_r-values is 5.1666667, meaning that of the 1000 values for F_r, 5% are larger than 5.1666667. So if we use a signficance level of 5%, our observed value of 11.17 is larger than the critical value for F_r, and we conclude that the null-hypothesis can be rejected.

Now this p-value of 0.004 and the critical value of 5.1666667 are based on our own computations. Actually there are better ways. One is to look up critical values of F_r in tables, for instance in Kendall M.G. (1970) *Rank correlation methods.* (fourth edition). The p-value corresponding to this F_r-value depends on k, the number of groups of data (here 3 columns) and N, the number of rows (12 individuals). If we look up that table, we find that for $k = 3$ and $N = 12$ the critical value of F_r for a type I error rate of 0.05 equals 6.17. Our observed F_r-value of 11.17 is larger than that, therefore we can reject the null-hypothesis that the median skating times are the same at the three different championships. So we have to tell your friend that there are general differences in skating times at different contests, $F_r = 11.17, p < 0.05$, but it is not the case that the fastest times were observed at the Olympics.

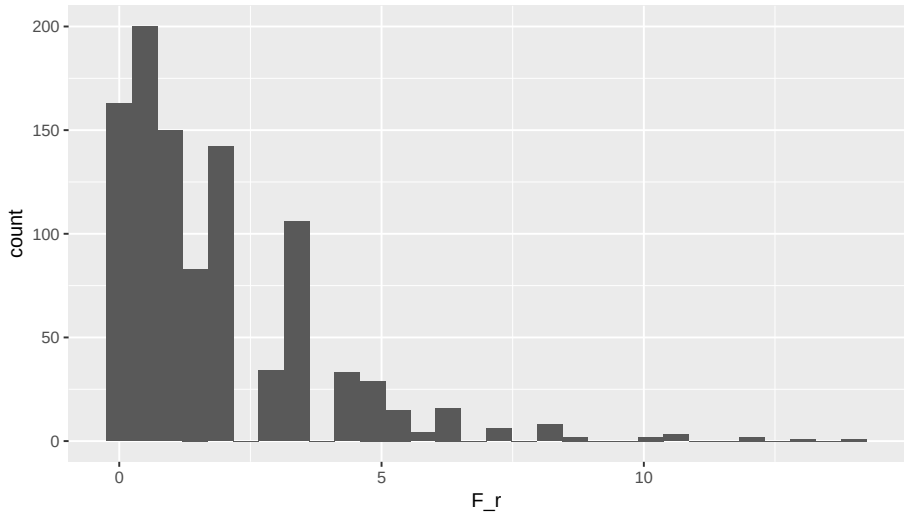


Figure 12.4: Histogram of 1000 possible values for Fr given that the null-hypothesis is true, for 12 speedskaters.

Another way is to make an approximation of the distribution of F_r. Note that the distribution in the histogram is very strangely shaped. The reason is that the data set is quite limited. Suppose we have not data on 12 speedskaters, but on 120. If we then randomly mix up data again and compute 1000 different values for F_r, we get the histogram in Figure 12.5.

The shape becomes more regular. It also starts to resemble another distribution, that of the χ^2 (chi-square). It can be shown that the distribution of the F_r for a large number of rows in the data matrix, and at least 6 columns, approaches the shape of the χ^2-distribution with $k-1$ degrees of freedom. This is shown in Figure 12.6.

The line of the χ^2-distribution with 2 degrees of freedom approaches the histogram quite well, but not perfectly. In general, for large N and $k > 5$, the approximation is good enough. In that way it gets easier to look up p-values for certain F_r-values, because the χ^2-distribution is well-known[2], so we don't have to look up critical values for F_r in old tables. For a significance level of 5%, the critical value of a χ^2 with 2 degrees of freedom is 5.991. This is close to the value in the table for F_r in old books: 6.17. The part of the χ^2-distribution with 2 degrees of freedom that is larger than the observed 11.17 is 0.0037537, so our approximate p-value for our null-hypothesis is 0.004.

[2]The χ^2-distribution is based on the normal distribution: the χ^2-distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

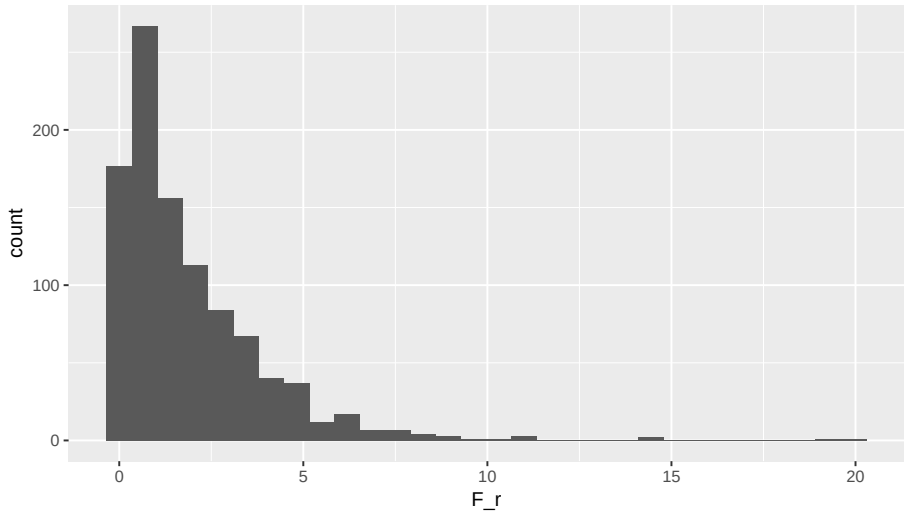


Figure 12.5: Histogram of 1000 possible values for Fr given that the null-hypothesis is true, for 120 speedskaters.

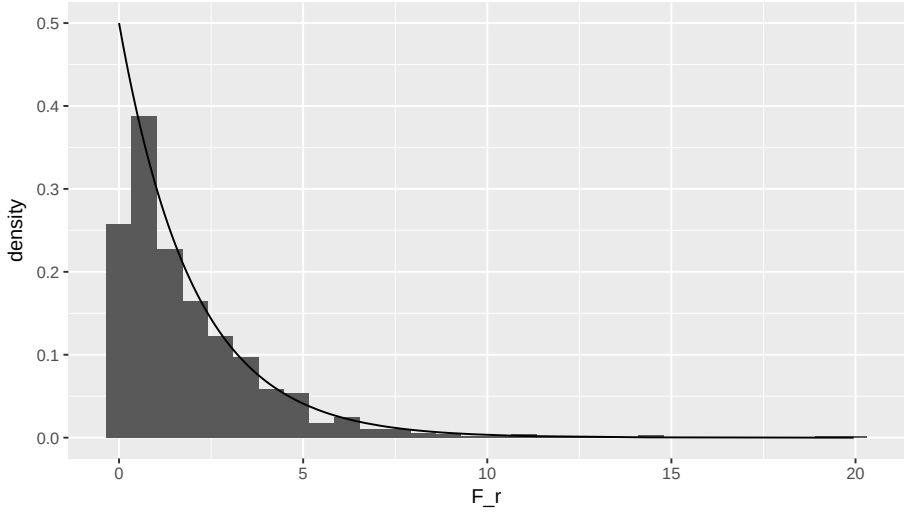


Figure 12.6: The distrbution of Fr under the null-hypothesis, overlain with a chi-square distribution with 2 degrees of freedom.

12.3 How to perform Friedman's test in SPSS

First of all, you need data in wide format. If your data happens to be in long format, use the CASETOVARS procedure to get the data in wide format. CASETOVARS requires your data to be ordered, so use the SORT CASE BY procedure before CASETOVARS. Suppose your data is in long format, as in Table 12.4.

Table 12.4: The raw skating data in long data format.

athlete	occasion	time
1	1.00	14.35
1	2.00	16.42
1	3.00	15.79
2	1.00	17.36
2	2.00	18.13
2	3.00	14.26

Then the following syntax turns the data into wide format:

```
SORT CASES BY athlete occasion.
CASESTOVARS
  /ID=athlete
  /INDEX=occasion
  /GROUPBY=VARIABLE
 /SEPARATOR = "_".
```

This creates the wide format data matrix in Table 12.5:

Table 12.5: The raw skating data in wide data format after CASETOVARS

athlete	time_1.00	time_2.00	time_3.00
1	14.35	16.42	15.79
2	17.36	18.13	14.26
3	19.01	19.95	18.37
4	27.90	17.78	15.12
5	17.67	16.96	17.17
6	17.83	16.15	15.30
7	16.30	19.44	15.63
8	28.00	16.23	15.69
9	18.27	15.76	15.65
10	17.00	16.18	14.99
11	17.10	13.89	15.83
12	18.94	14.83	14.77

Note the variable names: they start with the dependent variable time and are then indexed by the number of the occasion, 1.00, 2.00 and 3.00, that re-

late to European Championships, Olympic Games and World Championships, respectively.

We can then specify that we want Friedman's test by using the NPAR TESTS procedure with the FRIEDMAN subcommand and indicating which variables we want to use:

```
NPAR TESTS
/FRIEDMAN=time_1.00 time_2.00 time_3.00.
```

Ranks

	Mean Rank
time_1.00: time	2.58
time_2.00: time	1.25
time_3.00: time	2.17

Test Statistics

N	12
Chi-Square	11.167
df	2
Asymp. Sig.	.004

Figure 12.7: SPSS output of the Friedman test.

In the output in Figure 12.7 you first see the mean ranks. Note that if you multiply these by 12 (the number of rows), you get the sum of the ranks per column that we also computed above. Next you see a chi-square statistic, degrees of freedom, and an asymptotic p-value (Asymp. Sig.). Why don't we see an F_r-statistic?

The reason is, as discussed in the previous section, that for large number of measurements (columns) and a large number of individuals (rows), the F_r statistic tends to behave like a chi-square, χ^2, with $k-1$ degrees of freedom. So what we are looking at in this output is really an F_r-value of 11.17 (exactly the same value as we computed by hand in the previous section). In order to approximate the p-value, this value of 11.17 is interpreted as a chi-square (χ^2), which with 2 degrees of freedom has a p-value of 0.004.

This asymptotic (approximated) p-value is the correct p-value if you have a lot of rows (large N) and at least 6 variables ($k > 5$). If you do not have that, as we have here, this asymptotic p-value is only what it is: an approximation. If you want to have the exact p-value, then do

```
NPAR TESTS
```

```
/FRIEDMAN=time_1.00 time_2.00 time_3.00
/METHOD=EXACT.
```

and then use the p-value under *exactsign.*, in this case 0.002, see Figure 12.8.

Ranks

	Mean Rank
time_1.00: time	2.58
time_2.00: time	1.25
time_3.00: time	2.17

Test Statistics

N	12
Chi-Square	11.167
df	2
Asymp. Sig.	.004
Exact Sig.	.002
Point Probability	.001

Figure 12.8: SPSS output of the Friedman test with the exact p-value.

Thus, a Friedman's test of equal medians showed that speedskaters show significantly different median times on the 10 kilometer distance at the three types of contests, $F_r = 11.17, p = 0.002$.

12.4 Wilcoxon's signed ranks test for 2 measures

Friedman's test can be used for 2 measures, 3 measures or even 10 measures. As stated earlier, the well-known Wilcoxon's test can only be used for 2 measures. For completeness, we also discuss that test here.

For each athlete, we take the difference in skating times and call it d, see Table 12.6. Next we rank these d-values, irrespective of sign, and call these ranks $rank_d$. From the table **??** we see that athlete 12 shows the smallest difference in skating times (d= 0.06, rank = 1) and athlete 2 the largest difference.

Next we indicate for each rank whether it belongs to a positive or a negative difference d and call that variable **ranksign**.

Under the null-hypothesis, we expect that some of the larger d-values are positive and some of them negative, in a fairly equal amount. If we sum the ranks having plus-signs and sum the ranks having minus-signs, we would expect

Table 12.6: The raw skating data and the computations for Wilcoxon signed ranks test

athlete	Olympics	WorldChampionships	d	rank_d	ranksign
1	16.42	15.79	0.63	5.00	5.00
2	18.13	14.26	3.87	12.00	12.00
3	19.95	18.37	1.58	8.00	8.00
4	17.78	15.12	2.66	10.00	10.00
5	16.96	17.17	-0.21	3.00	-3.00
6	16.15	15.30	0.85	6.00	6.00
7	19.44	15.63	3.81	11.00	11.00
8	16.23	15.69	0.54	4.00	4.00
9	15.76	15.65	0.11	2.00	2.00
10	16.18	14.99	1.19	7.00	7.00
11	13.89	15.83	-1.94	9.00	-9.00
12	14.83	14.77	0.06	1.00	1.00

that these two sums are about equal, but only if the null-hypothesis is true. If the sums are very different, then we should reject this null-hypothesis. In order to see if the difference in sums is too large, we compute them as follows:

$$\begin{aligned} T^+ &= 5 + 12 + 8 + 10 + 6 + 11 + 4 + 2 + 7 + 1 = 66 \\ T^- &= 3 + 9 = 12 \end{aligned}$$

To know whether T^+ is significantly larger than T^-, the value of T^+ can be looked up in a table, for instance in Siegel & Castellan (1988). There we see that for T^+, with 12 rows, the probability of obtaining a T^+ of at least 66 is 0.0171. For a two-sided test (if we would have switched the columns of the two championships, we would have gotten a T^- of 66 and a T^+ of 12!), we have to double this probability. So we end up with a p-value of $2 \times 0.0171 = 0.0342$.

In the table we find no critical values for large sample size N, but fortunately, similar to the Friedman test, we use an approximation using the normal distribution. It can be shown that for large sample sizes, the statistic T^+ is approximately normally distributed with mean

$$\mu = \frac{N(N+1)}{4} \tag{12.2}$$

and variance:

$$\sigma^2 = \frac{N(N+1)(2N+1)}{24} \tag{12.3}$$

If we therefore standardize the T^+ by subtracting the μ and then dividing by the square root of the variance $\sqrt{(\sigma^2)} = \sigma$, we get a Z-value with mean 0 and standard deviation 1. To do that, we use the following formula:

$$Z = \frac{T^+ - \mu}{\sigma} = \frac{T^+ - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \tag{12.4}$$

Here T^+ is 66 and N equals 12, so if we fill in the formula we get $Z = 2.1180543$. From the standard normal distribution we know that 5% of the observations lie above 1.96 and below -1.96. So a value for Z larger than 1.96 or smaller than -1.96 is enough evidence to reject the null-hypothesis. Here our Z-statistic is larger than 1.96, therefore we reject the null-hypothesis that the median skating times are the same at the World Championships and the Olympics. The p-value associated with a Z-score of 2.1180543 is 0.034.

12.5 How to perform Wilcoxon's signed ranks test in SPSS

If you want to use the Wilcoxon test, then use the following syntax:

```
NPAR TESTS
/WILCOXON=time_2.00 time_3.00
/METHOD=EXACT.
```

In the output in Figure 12.9 we see a Z-statistic, an asymptotic p-value, and two exact p-values. The reason that we see a Z-statistic is that the Wilcoxon T^+ statistic approaches a normal distribution in case we have a large number of observations (many rows). If $N > 15$, the approximation is good enough so that the statistic can be interpreted as a z-score (standardized score with a normal distribution). That means that a z-score of 1.96 or larger or -1.96 or smaller can be regarded as significant at the 5% significance level. Since the standard normal distribution is only an approximation, and we have $N = 12$, we have to look at the exact significance level, which is in this case 0.034. We see that the exact p-value is in this case equal to the approximate p-value. Note that we use a two-sided test, to allow for the fact that random sampling could lead to a higher median for the Olympic Games or a higher median for the World Championships. We just want to know whether the null-hypothesis that the two medians differ can be rejected (in whatever direction) or not.

Let's compare the output with the Friedman test, but then only use the relevant variables in your syntax:

```
NPAR TESTS
/FRIEDMAN=  time_2.00 time_3.00
/METHOD=EXACT.
```

In the output in Figure 12.10 we see that the null-hypothesis of equal medians at the World Championships and the Olympic Games can be rejected, with a p-value of 0.039.

Ranks

		N	Mean Rank	Sum of Ranks
time_3.00: time - time_2.00: time	Negative Ranks	2[a]	6.00	12.00
	Positive Ranks	10[b]	6.60	66.00
	Ties	0[c]		
	Total	12		

a. time_3.00: time < time_2.00: time

b. time_3.00: time > time_2.00: time

c. time_3.00: time = time_2.00: time

Test Statistics

	time_3.00: time - time_2.00: time
Z	-2.118[b]
Asymp. Sig. (2-tailed)	.034
Exact Sig. (2-tailed)	.034
Exact Sig. (1-tailed)	.017
Point Probability	.004

b. Based on negative ranks.

Figure 12.9: SPSS output of the Wilcoxon test.

Note that both the Friedman and Wilcoxon tests come up with very similar p-values. Their rationales are very similar: Friedman's test is based on ranks and Wilcoxon's test is based on positive and negative differences between measures 1 and 2, so in fact ranks 1 and 2 for each row in the data matrix. Both can therefore be used in the case you have two measures. We recommend to use the Friedman test, since that test can be used in all situations where you have 2 or more measures per row. Wilcoxon's test can only be used if you have 2 measures per row.

In sum, we can report in two ways on our hypothesis regarding similar skating times at the World Championships and at the Olympics:

1. A Friedman test showed a significant difference between the 10km skating times at the World Championships and at the Olympics, $F_r = 5.33, p = 0.04$. Athletes more often show their fastest times at the World Championships than can be expected due to chance.

2. A Wilcoxon signed ranks test showed a significant difference between the 10km skating times at the World Championships

Ranks

	Mean Rank
time_2.00: time	1.17
time_3.00: time	1.83

Test Statistics

N	12
Chi-Square	5.333
df	1
Asymp. Sig.	.021
Exact Sig.	.039
Point Probability	.032

Figure 12.10: SPSS output of the Friedman test for two measures.

> and at the Olympics, $Z = -2.12, p = 0.03$. Athletes more often show their fastest times at the World Championships than can be expected due to chance.

How do we know that the fastest times were at the World Championships? If we look at raw data above, that does not seem that obvious. But this conlusion is based on the sum of ranks: we saw a sum of ranks of 26 for the Olympics and 15 for the World Championships. So the average rank is lower at the World Championships.

12.6 Ties

Many nonparametric tests are based on ranks. For example, if we have the data sequence 0.1, 0.4, 0.5, 0.2, we give these values the ranks 1, 3, 4, 2, respectively. But in may data cases, data sequences cannot be ranked unequivocally. Let's look at the sequence 0.1, 0.4, 0.4, 0.2. Here we have 2 values that are exactly the same. We say then that we have *ties*. If we have ties in our data like the 0.4 in this case, one very often used option is to arbitrarily choose one of the 0.4 values as smaller than the other, and then average the ranks. Thus, we rank the data into 1, 3, 4, 2 and then average the tied observations: 1, 3.5, 3.5, 2. As another example, suppose we have the sequence 23, 54, 54, 54, 19, we turn this into ranks 2, 3, 4, 5, 1 and take the average of the ranks of the tied observations of 54: 2, 4, 4, 4, 1. These ranks corrected for ties can then be used to compute the test statistic, for instance Friedman's F_r or Wilcoxon's Z. However, in many cases, because of these corrections, a slightly different formula is to be used. So the formulas become a little bit different. This is all

done in SPSS automatically. If you want to know more, see Siegel and Castellan (1988).

12.7 Exercises

A researcher is interested in the relationship between mood and day of the week: are people generally moodier on Monday than on Wednesday or Friday?

Below we see the data on 4 people that rated their mood from 1 (very moody) to 10 (not moody at all) on three separate days in a week in February: Day 1 is Monday, day 2 is Wednesday and day 3 is Friday:

ID	Day	Mood
1	1	3
1	2	5
1	3	8
2	1	4
2	2	7
2	3	6
3	1	2
3	2	4
3	3	1
4	1	9
4	2	5
4	3	3

1. Put the data into wide format, and think of appropriate variable names

...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

2. Rank these data row-wise: for each row determine the lowest mood (1), the second lowest mood (2) and the highest mood score (3)

...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

3. Determine the column sums: the sum of the ranks for Monday, Wednesday and Friday.

4. How many rows do you have (N) and how many columns of data do you have (k)?

5. Compute F_r.

6. Copy the data into SPSS and run a Friedman's test. Should you ask for an exact p-value? Provide the syntax.

7. Suppose you get the SPSS output in Figure . What would your conclusion be regarding the research question about the relationship between moodiness and the day of the week?

8. In this data set, for which day did we observe the personal best mood? How many of the individuals showed their best mood on that day?

9. A linear mixed model was run on this data set. When checking model assumptions, we saw the following graphs in Figures **??** and 12.14.

   Would you prefer to stick to the Friedman's test for this data set, or would you prefer to report a linear mixed model? Explain your answer.

10. Could you have performed a Wilcoxon test on these data? Why, or why not?

Answers:

1. The raw data in wide format:

ID	Mood_1	Mood_2	Mood_3
1	3	5	8
2	4	7	6
3	2	4	1
4	9	5	3

Ranks

	Mean Rank
Mood_1.00: Mood	1.75
Mood_2.00: Mood	2.50
Mood_3.00: Mood	1.75

Test Statistics

N	4
Chi-Square	1.500
df	2
Asymp. Sig.	.472
Exact Sig.	.653
Point Probability	.222

Figure 12.11: SPSS output of a Friedman test.

2. The row-wise ranked data:

ID	Mood_1	Mood_2	Mood_3
1	1	2	3
2	1	3	2
3	2	3	1
4	3	2	1

3. Day 1: 7, Day 2: 10 and Day3: 7.

4. $N = 4$ and $k = 3$

5.
$$\begin{aligned} F_r &= \left[\frac{12}{4 \times 3(3+1)} \times (7^2 + 10^2 + 7^2)\right] - 3 \times 4(3+1) \\ &= \left[\frac{12}{48} \times 198\right] - 48 = 1.50 \end{aligned}$$

6.
```
NPAR TESTS
/FRIEDMAN=  Mood_1   Mood_2    Mood_3
/METHOD=Exact.
```

7. We found no significant effect of day of the week on mood, $F_r = 1.50, p = 0.65$, so the null-hypothesis of equal mood during the week is not rejected. Note however that the sample size

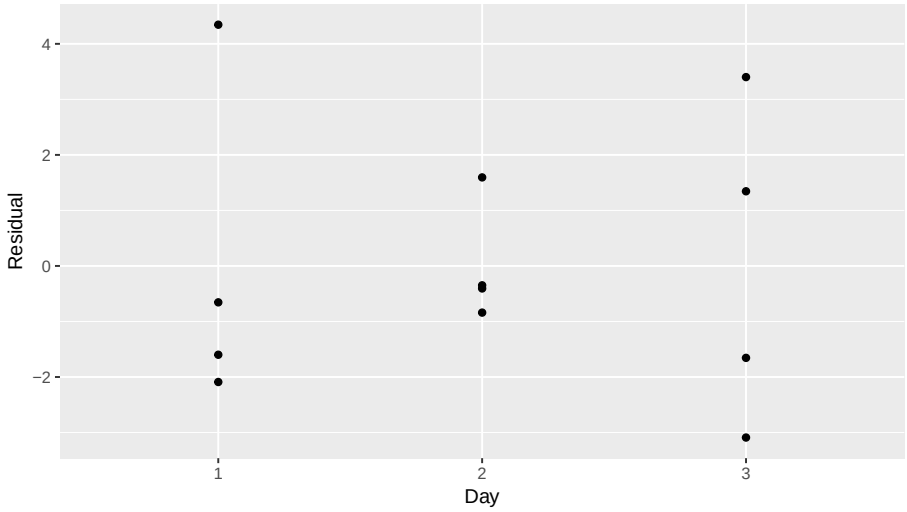


Figure 12.12: Residual plot after a linear mixed model analysis.

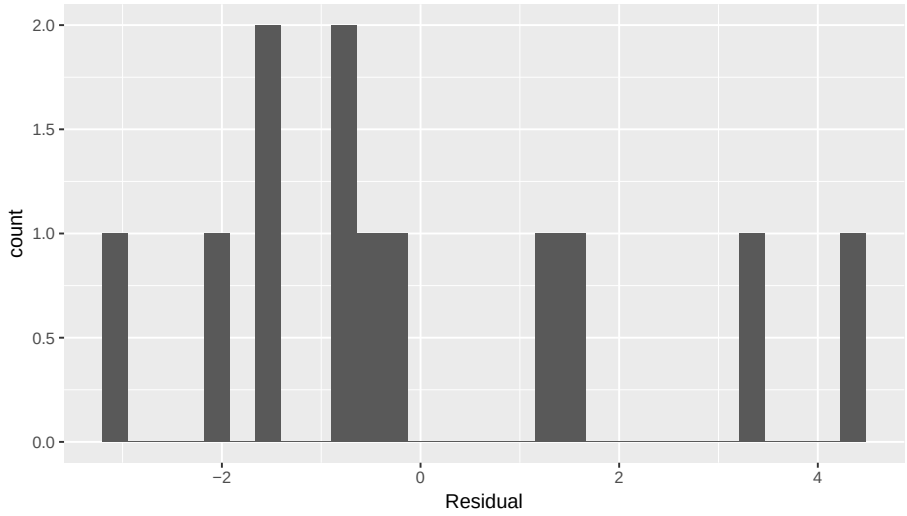


Figure 12.13: Residual plot after a linear mixed model analysis.

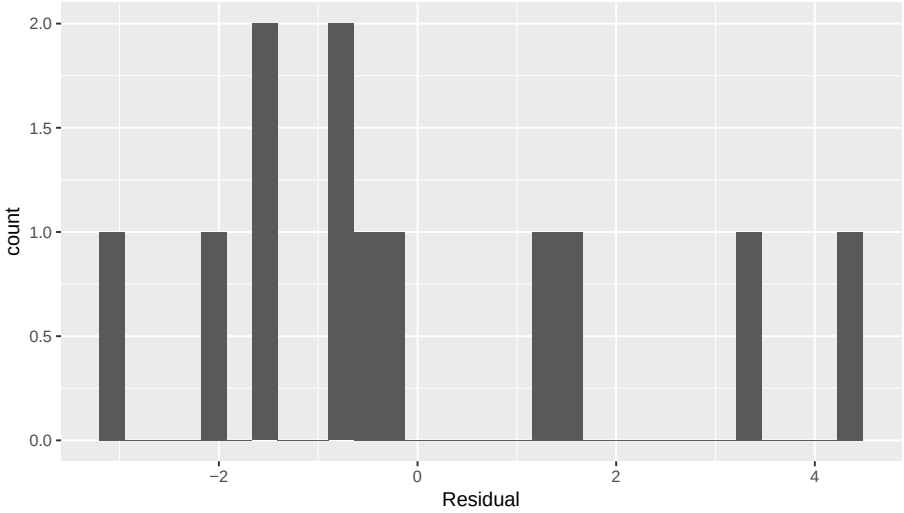


Figure 12.14: Histogram of residuals after a linear mixed model analysis.

was extremely small (12 data points), so even if there is a real relationship between mood and day of the week, there was little chance to find evidence of that in this data set.

8. The highest column sum of the ranks was found for day 2, which was Wednesday. So in this data set we saw that the four individuals generally showed their personal highest mood score on Wednesday. Actually, 2 persons out of 4 showed their highest score (rank 3) on Wednesday (ID=2 and ID=3).

9. The plots suggests that the variance of the residuals is very small for the second day, compared to the other two days. The distribution is also hardly normal. But it is hard to tell whether the assumptions are reasonable, since there are so few data points. It would therefore be safest to report a Friedman test.

10. A Wilcoxon test can only be performed on two measures, say Monday and Wednesday data, or Monday and Friday data. You could not test the null-hypothesis of the same moods on three days with a Wilcoxon test.

Chapter 13

Generalized linear models part I: logistic regression

13.1 Introduction

In previous chapters we were introduced to the linear model, with its basic form

$$y = b_0 + b_1 X_1 + \ldots + b_n X_n + e \tag{13.1}$$
$$e \sim N(0, \sigma_e^2) \tag{13.2}$$

Two basic assumptions of this model are the linearity in the parameters, and the normally distributed residual e. Linearity in the parameters means that the effects of intercept and the independent variables $X_1, X_2, \ldots X_n$ are additive: the assumption is that you can sum these effects to come to a predicted value for y. So that is also true when we include interaction effects to account for moderation effects,

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e \tag{13.3}$$
$$e \sim N(0, \sigma_e^2) \tag{13.4}$$

or when we use a quadratic term to account for other types of nonlinearity in the data:

$$y = b_0 + b_1 X_1 + b_2 X_1 X_1 + e \tag{13.5}$$
$$e \sim N(0, \sigma_e^2) \tag{13.6}$$

In all these models, the assumption is that the effects of the parameters can be added to one another.

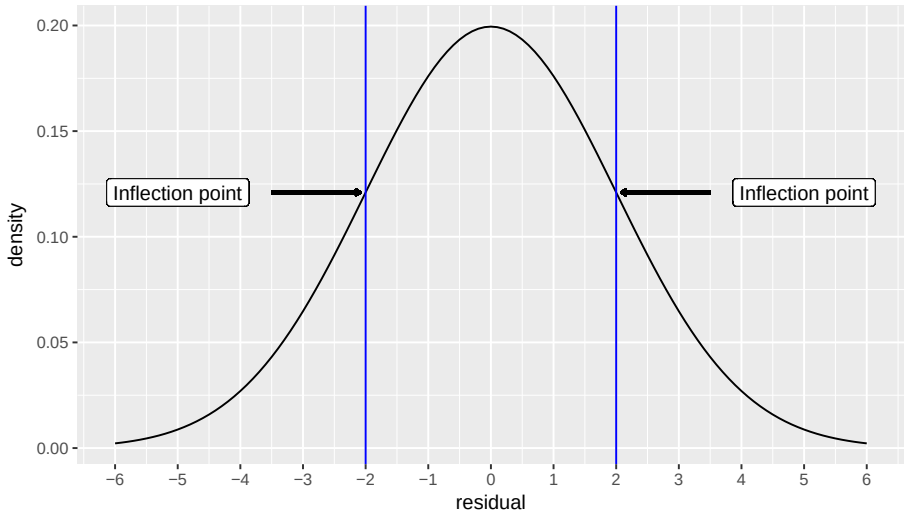


Figure 13.1: Density function of the normal distribution, with mean 0 and variance 4 (standard deviation 2). Inflection points are positioned at residual values of minus 1 standard deviation and plus 1 standard deviation.

The other major assumption of linear (mixed) models is the normal distribution of the residuals. As we have seen in for instance the previous chapter, sometimes the residuals are not normally distributed. Remember that with a normal distribution $N(0, \sigma^2)$, in principle all values between $-\infty$ and $+\infty$ are possible, but they tend to concentrate around the value of 0, in the shape of the bell-curve. Figure 13.1 shows the normal distribution $N(0, \sigma^2 = 4)$: it is centered around 0 and has variance 4. Note that the inflection point, that is the point where the decrease in density tends to decelerate, is exactly at the values -2 and +2. These are equal to the square root of the variance, which is the standard deviation, $+\sigma$ and $-\sigma$.

A normal distribution is suitable for continuous data: for example a variable that can take all possible values between -1 and 0. For many variables this is not true. Think for example of temperature measures: if the thermometer gives degrees centigrade with a precision of only 1 decimal, we can never have values of say 10.07 or -56.789. Our data will in fact be *discrete*, showing rounded values like 10.1, 10.2, 10.3, but no values in between.

Nevertheless, the normal distribution can still be used in many such cases. Take for instance a data set where the temperature in Amsterdam in summer was predicted on the basis of a linear model. Fig 13.2 shows the distribution of the residuals of that model:

The temperature measures were discrete with a precicsion of one tenth of a degree centigrade, but the distribution seems well approximated by a normal curve.

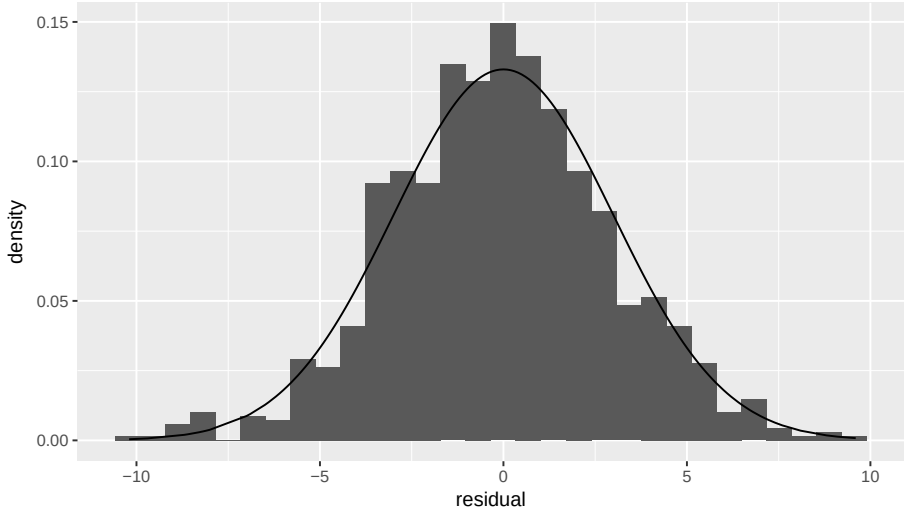


Figure 13.2: Even if residuals are really discrete, the normal distribution can be a good approximation of their distribution.

But let's look at an example where the discreteness is more prominent. In Figure 13.3 we see the residuals of an analysis of exam results. Students had to do an asssignment that had to meet 4 criteria: 1) originality, 2) language, 3) structure, and 4) literature review. Each criterion was scored as either fulfilled (1) or not fulfilled (0). The score for the assignment was given on the basis of *the number of criteria* that were met, so the scores could be 0, 1, 2, 3 or 4. The score was predicted on the basis of the average exam score on previous assignments using a linear model.

Figure 13.3 shows that the residuals are very discrete, and that the continous normal distribution is a very bad approximation of the histogram. We often see this phenomenon when our data consists of *counts* with a limited maximum number.

An even more extreme case we observe when our dependent variable consists of whether or not students passed the assignment: only those assignments that fulfilled all 4 criteria are regarded as sufficient. If we score all students with a sufficient assignment as passed (1) and all students with an insufficient assignment as failed (0) and we predict this again by the average exam score on previous assignments using a linear model, we get the residuals displayed in Figure 13.4.

Here it is definitely evident that a normal approximation of the residuals will not do. When the dependent variable has only 2 possible values, a linear model will never work because the residuals can never have a distribution that is even remotely looking normal.

In the coming two chapters we will discuss how generalized linear models

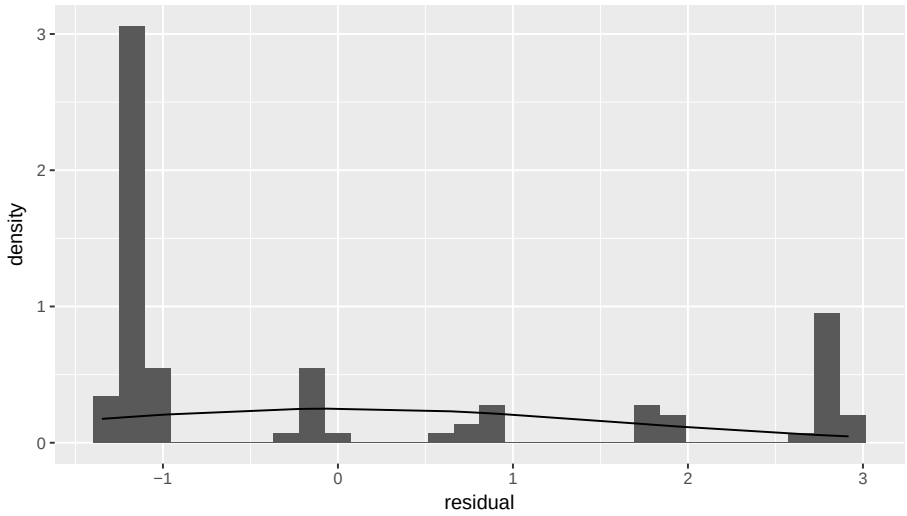


Figure 13.3: Count data example where the normal distribution is not a good approximation of the distribution of the residuals.

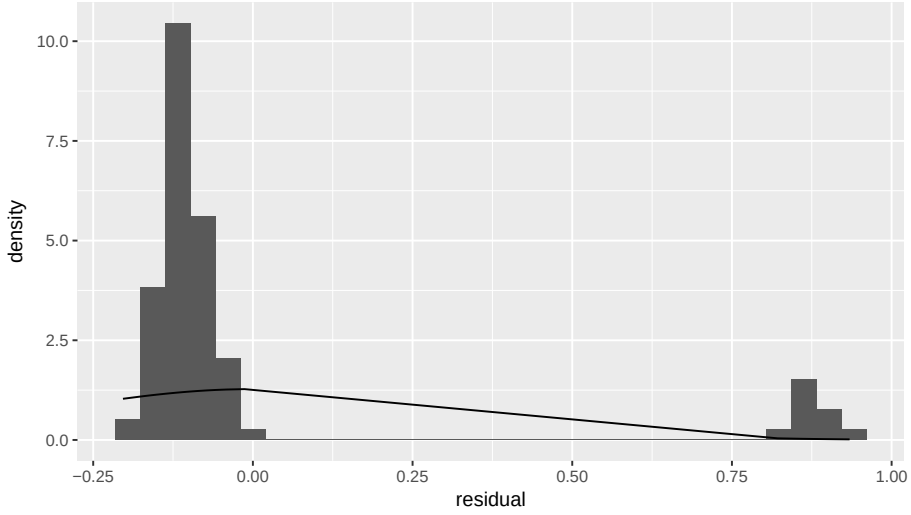


Figure 13.4: Dichotomous data example where the normal distribution is not a good approximation of the distribution of the residuals.

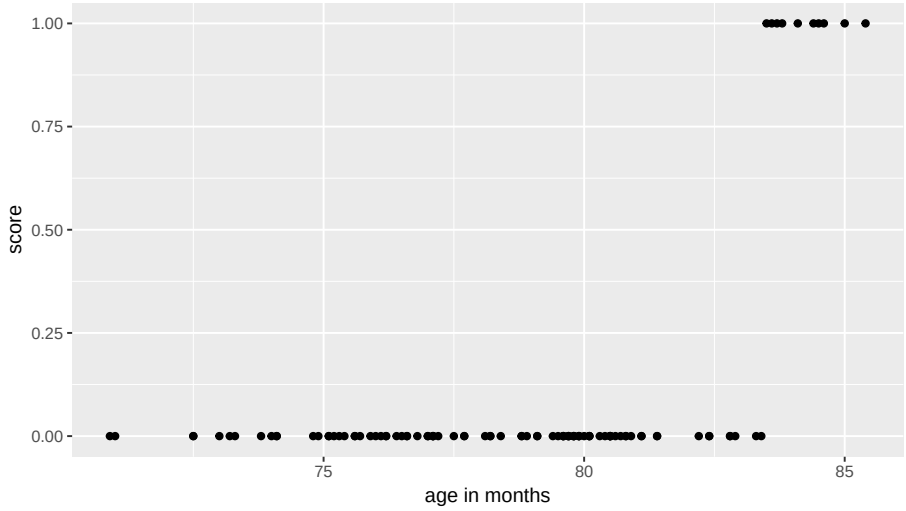


Figure 13.5: Data example: Exam outcome (score) as a function of age, where 1 means pass and 0 means fail.

can be used to analyze data sets where the assumption of normally distributed residuals is not tenable. First we discuss the case where the dependent variable has only 2 possible values (dichotomous dependent variables like yes/no or pass/fail, heads/tails, 1/0). In the next chapter, we will discuss the case where the dependent variable consists of counts $(1, 2, 3, 4, \ldots)$.

13.2 Logistic regression

Imagine that we analyze results on an exam for third grade children. These children are usually either 6 or 7 years old, dependending on what month they were born in. The exam is on February 1st. A researcher wants to know whether the age of the child can explain why some children pass the test and others fail. She computes the age of the child in months. Each child that passes the exam gets a score 1 and all the others get a score 0. Figure 13.5 plots the data.

She wants to use the following linear model:

$$score = b_0 + b_1 age + e \tag{13.7}$$

$$e \sim N(0, \sigma_e^2) \tag{13.8}$$

Figure 13.6 shows the estimated regression line and Figure 13.7 shows the distribution of the residuals as a function of age.

Clearly a linear model is not appropriate. Here, the assumption that the dependent variable, score in this case, is scattered randomly around the predicted value with a normal distribution is not reasonable. The main problem is

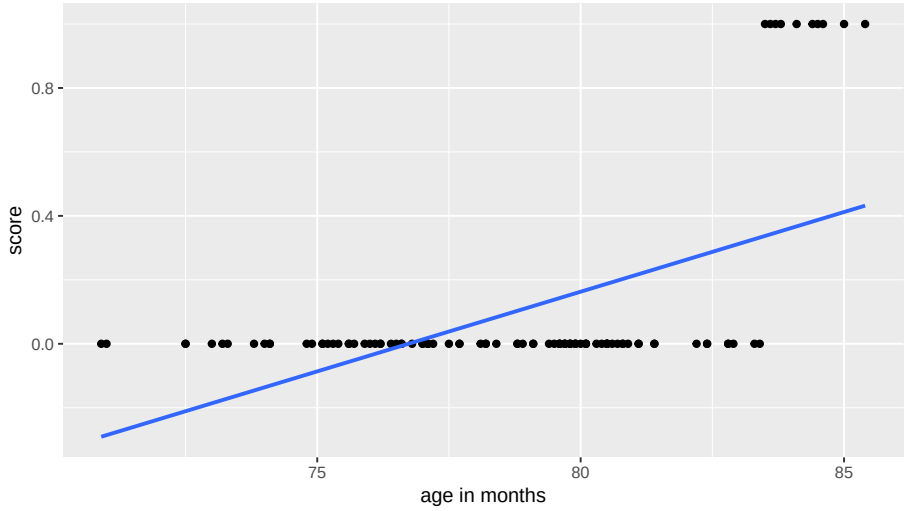


Figure 13.6: Example exam data with a linear regression line.

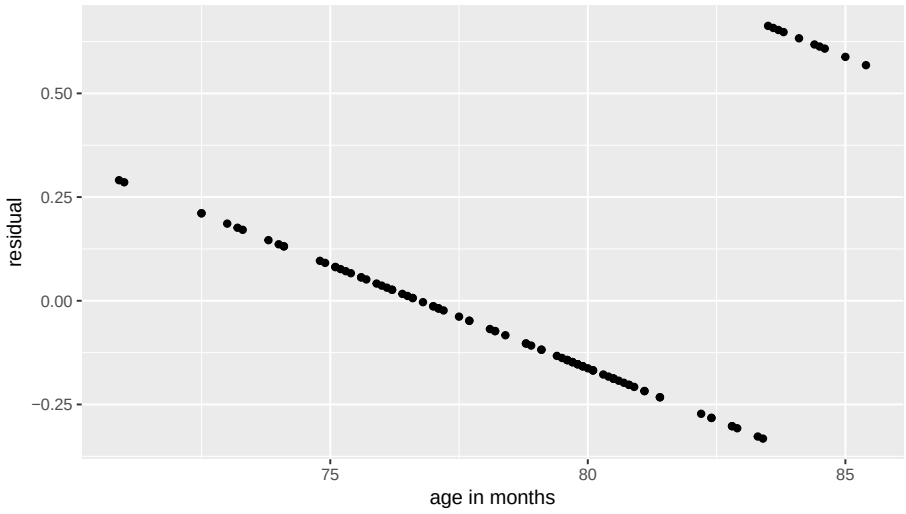


Figure 13.7: Residuals as a function of age, after a linear regression analysis of the exam data.

that the dependent variable score can only have 2 values: 0 and 1. When we have a dependent variable that is categorical, so not continuous, we generally use *logistic regression*. In this chapter we cover the case when the dependent variable takes binary values, like 0 and 1.

13.2.1 Bernoulli distribution

Rather than using a normal distribution, we could try a Bernoulli distributiuon. The Bernoulli distribution is the distribution of a coin flip. For example, if the probability of heads is 0.1, we can expect that if we flip the coin, on average we expect to see 0.1 times heads and 0.9 times tails. Our best bet then is that the outcome is tails. However, if we actually flip the coin, we might see heads anyway. There is some randomness to be expected. Let y be the outcome of a coin flip: heads or tails. If we have a Bernoulli distribution for variable y with probability p for heads, we *expect* to see heads p times, but we actually *observe* heads or tails.

$$y \sim Bern(n, p) \tag{13.9}$$

The same is true for the normal distribution in the linear model case: we *expect* that the observed value of y is exactly equal to its predicted value ($b_0 + b_1 X$), but we always *observe* that it is different.

$$y \sim N(\mu = b_0 + b_1 X, \sigma_e^2) \tag{13.10}$$

In our example, the pass rate could also be conceived as the outcome of a coin flip: pass instead of heads and fail instead of tails. So would it be an idea to predict the *probability* of success on the basis of age? And then for every predicted probability, we allow for the fact that actually the observed success can differ. Our linear model could then look like this:

$$p_i = b_0 + b_1 age_i \tag{13.11}$$

$$score_i \sim Bern(p_i) \tag{13.12}$$

So for each child i, we predict the probability of success, p_i, on the basis of her/his age. Next, the randomness in the data comes from the fact that a probability is only a probability, so that the observed success of a child $score_i$, is like a coin toss with probability of p_i for success.

For example, suppose that we have a child with an age of 80 months, and we have $b_0 = -3.8$ and $b_1 = 0.05$. Then the predicted probability p_i is equal to $-3.8 + 0.05 \times 80 = 0.20$. The best bet for such a child would be that it fails the exam. But 0.20 is only a probability, so by chance the child could pass the exam. This model also means that if we would have 100 children of age 80 months, we would *expect* that 20 of these children would pass the test and 80 would fail. But we can't make predictions for one individual alone: we don't know which child exactly will pass and which child won't. Note that this is similar to the

normally distributed residual in the linear model: in the linear model we expect a child to have a certain value for y, but we know that there will be a deviation from this predicted value: the residual. For a whole group of children with the same predicted value for y, we know that the whole group will show residuals that have a normal distribution. But we're not sure what the residual will be for each individual child.

Unfortunately, this model for probabilities is not very helpful. If we use a linear model for the probability, this means that we can predict probability values of less than 0 and more than 1, and this is not possible for probabilities. If we use the above values of $b_0 = -3.8$ and $b_1 = 0.05$, we predict a probability of -.3 for a child of 70 months and a probability of 1.2 for a child of 100 months. Those values are meaningless!

13.2.2 Odds and logodds

Instead of predicting probabilities, we could predict *odds*. The nice property of odds is that they can have very large values, much larger than 1.

What are odds again? Odds are a different way of talking about probability. Suppose the probability of winning the lottery is 1%. Then the probability of loosing is 99%. This is equal to saying that the odds of winning against loosing are 1 to 99, or $1 : 99$, because the probability of success is 99 times smaller than the probability of loosing.

As another example, suppose the probability of being alive tomorrow is equal to 0.9999. Then the probability of not being alive tomorrow is $1 - 0.9999 = 0.0001$. Then the probability of being alive tomorrow is $0.9999/0.0001 = 9999$ times larger than the the probability of not being alive. Therefore the odds of being alive tomorrow against being dead is 9999 to 1 (9999:1).

If we have a slightly biased coin, the probability of heads might be 0.6. The probability of tails is then 0.4. Then the probability of heads is then 1.5 times larger than the probability of heads (0.6/0.4=1.5). So the odds of heads against tails is then 1.5 to 1. For the sake of clarity, odds are often multiplied by a constant to get integers, so we can also say the odds of heads aganst tails are 3 to 2. Similarly, if the probablity of heads were 0.61, the odds of heads against tails would be 0.61 to 0.39, which can be modified into 61 to 39.

Now that we know how to go from probability statements to statements about odds, how do we go from odds to probability? If someone says the odds of heads against tails is 10 to 1, this means that for every 10 heads, there will be 1 tails. In other words, if there were 11 coin tosses, 10 would be heads and 1 would be tails. We can therefore transform odds back to probabilities by noting that 10 out of 11 toin tosses is heads, so $10/11 = 0.91$, and 1 out of 11 is tails, so $1/11 = 0.09$.

If someones says the odds of winning a gold medal at the Olympics is a thousand to one (1000:1), this means that if there were $1000 + 1 = 1001$ opportunities, there would be a gold medal in 1000 cases and failure in only one. This corresponds to a probability of 1000/1001 for winning and 1/1001 for failure.

As a last example, if at the horse races, the odds of Bruno winning against Sacha are four to five (4:5), this means that for every 4 winnings by Bruno, there would be 5 winnings by Sacha. So out of a total of 9 winnings, 4 will be by Bruno and 5 will be by Sacha. The probability of Bruno outrunning Sacha is then $4/9 = 0.44$.

If we would summarize the odds by doing the division, we have just one number. For example, if the odds are 4 to 5 (4:5), the odds are $4/5 = 0.8$, and if the odds are a thousand to one (1000:1), then we can also say the odds are 1000. Odds, unlike probabilities, can have values that are larger than 1.

However, note that odds can never be negative: a very small odds is one to a thousand (1:1000). This can be summarized as an odds of 0.000999001, but that is still larger than 0. In summary: probabilties range from 0 to 1, and odds from 0 to infinity.

Because odds can never be negative, mathematicians have proposed to use the *natural logarithm*[1] of the odds as the preferred transformation of probabilities. For example, suppose we have a probability of heads of 0.42. This can be transformed into an odds by noting that in 100 coin tosses, we would expect 42 times heads and 58 times tails. So the odds are 42:58, which is equal to $\frac{42}{58} = 0.7241379$. The *natural* logarithm of 0.7241379 equals -0.3227734 (use the *ln* button on your calculator!). If we have a value between 0 and 1 and we take the logarithm of that value, we always get a value smaller than 0. In short: a probability is never negative, but the corresponding logarithm of the odds can be negative.

Figure 13.8 shows the relationship between a probability (with values between 0 and 1) and the natural logarithm of the corresponding odds (the *logodds*). The result is a mirrored S-shaped curve on its side. For large probabilities close to one, the equivalent logodds becomes infinitely positive, and for very small probabilities close to zero, the equivalent logodds becomes infinitely negative. A logodds of 0 is equal to a probability of 0.5. If a logodds is larger than 0, it means the probability is larger than 0.5, and if a logodds is smaller than 0 (negative), the probability is smaller than 0.5.

In summary, if we use a linear model to predict probabilities, we have the problem of predicted probabilities smaller than 0 and larger than 1 that are meaningless. If we use a linear model to predict odds we have the problem of predicted odds smaller than 0 that are meaningless: they are impossible! If on the other hand we use a linear model to predict *the natural logarithm of odds* (logodds), we have no problem whatsoever. We therefore propose to use a linear model to predict *logodds*: the natural logarithm of the odds that correspond to

[1]The natural logarithm of a number is its logarithm to the base of the constant e, where e is approximately equal to 2.7. The natural logarithm of x is generally written as lnx or $log^e x$. The natural logarithm of x is the power to which e needs to be raised to equal x. For example, $ln(2)$ is 0.69, because $e^{0.69} = 2$, and $ln(0.2) = -1.6$ because $e^{-1.6} = 0.2$. The natural logarithm of e itself, $ln(e)$, is 1, because $e^1 = e$, while the natural logarithm of 1, $ln(1)$, is 0, since $e^0 = 1$.

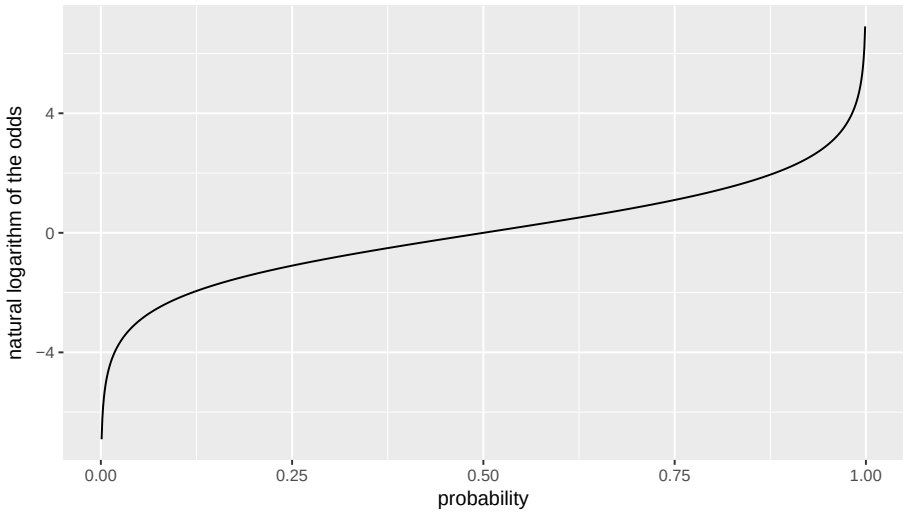


Figure 13.8: The relationship between a probability and the natural logarithm of the corresponding odds.

a particular probability.

Returning back to our example of the children passing the exam, suppose we have the following linear equation for the relationship between age and the logarithm of the odds of passing the exam

$$logodds = -3.82 + 0.05age,$$

This equation predicts that a child aged 70 months has a logodds of $-3.82 + 0.05 \times 70 = -0.34$. In order to transform that logodds back to a probability, we first have to take the exponential of the logodds[2] to get the odds:

$$odds = exp(logodds) = e^{logodds} = e^{-0.34} = 0.71$$

An odds of 0.71 means that the odds of passing the exam is 0.71 to 1 (0.71:1). So out of $1 + 0.71 = 1.71$ times, we expect 0.71 successes and 1 failure. The probability of success is therefore $\frac{0.71}{1+0.71} = 0.42$. Thus, based on this equation, the expected probability of passing the exam for a child of 70 months equals 0.42.

[2]If we know $ln(x) = 60$, we have to infer that x equals e^{60}, because $ln(e^{60}) = 60$ by definition of the natural logarithm, see previous footnote. Therefore, if we know that $ln(x) = c$, we know that x equals e^c. The exponent of c, e^c, is often written as $exp(c)$. So if we know that the logarithm of the odds equals c, $logodds = ln(oddsratio) = c$, then the odds is equal to $exp(c)$.

If you find that easier, you can also memorize the following formula for the relationship between a logodds of x and the corresponding probability:

$$p_x = \frac{exp(x)}{1 + exp(x)} \tag{13.13}$$

Thus, if you have a logodds x of -0.34, the odds equals $exp(-0.34) = 0.71$, and the corresponding probability is $\frac{0.71}{1+0.71} = 0.42$.

13.2.3 Exercises

From probability to logodds:

Given: In the Netherlands, 51% of the inhabitants is female.

1. If we randomly pick someone from this Dutch population, what is the probability that that that person is female?
2. If we randomly pick someone from this Dutch population, what are the odds that that that person is female over being male? (:)
3. If we randomly pick someone from this Dutch population, what are the odds that that that person is male over being female? (:)
4. What is the odds of randomly picking an inhabitant that is female, expressed as one number?
5. What is the odds of randomly picking an inhabitant that is male, expressed as one number?
6. What is the logodds of randomly picking an inhabitant that is female?
7. What is the logodds of randomly picking an inhabitant that is male?

Answers:

1. 0.51
2. 51 to 49 (51:49).
3. 49:51.
4. 51/49=1.04
5. 49/51=0.96
6. ln(51/49)= ln(1.04)=0.04
7. ln(49/51)= ln(0.96)=-0.04

From logoddss to probabilities:

Given: In the Netherlands, 51% of the inhabitants are female. Females tend to get older than males, so if we predict sex by age, we should expect a higher probability of a female for older ages. Suppose we have the following linear model for the relationship between age (in years) and the logodds of being female:

$$logodds_{female} = -0.01 + 0.01 \times age,$$

1. What is the predicted logodds of being female for a person of age 20?
2. What is the predicted logodds of being female for a person of age 90?
3. What is the predicted odds of being female for a person of age 20?
4. What is the predicted odds of being female for a person of age 90?
5. What are the predicted odds of being female for a person of age 20?
6. What are the predicted odds of being female against being male for a person of age 90?
7. What is the predicted probability of being female against being male for a person of age 20?
8. What is the predicted probability of being female for a person of age 90?
9. What is the predicted probability of being MALE for a person of age 90?

Answers:

1. $-0.01 + 0.01 \times 20 = 0.19$
2. $-0.01 + 0.01 \times 90 = 0.89$
3. $exp(0.19) = 1.21$
4. $exp(0.89) = 2.44$
5. 1.21 to 1, or 1.21:1
6. 2.44 to 1, or 2.44:1
7. 1.21/ (1.21 + 1)= 0.55
8. 2.44 / (2.44 + 1)= 0.71
9. 1 - 0.71 = 0.29

A big data analyst constructs a model that predicts whether an account on Twitter belongs to either a real person or organisation, or to a bot.

1. For one account, a user of this model finds an logodds of 4.5 that the account belongs to a bot. What is the corresponding probability that the twitter account belongs to a bot? Give the calculation.

2. For a short tweet with only a hyperlink, the probability that it comes from a bot is only 10%. What is the logodds that corresponds to this probability? Give the calculation.

Answers:

1. The logodds is 4.5, so the oddsratio is exp(4.5)=90.0. The odds of being a bot is then 90:1. The probability of being a bot is 90/ (90+1)= 0.99

2. Out of 100 tweets with only a hyperlink, 10 are by bots and 90 are by real persons or organisations. So the odds of coming from a bot are 10:90. The odds is therefore 10/90 = 0.11. When we take the natural logarithm of this odds, we get the logodd: ln(0.11) = -2.21.

13.2.4 Logistic link function

In previous pages we have seen that logodds have the nice property of having meaningful values between $-\infty$ and $+\infty$. This makes them suitable for linear models. In essence, our linear model for our exam data in children might then look like this:

$$logodds_{pass} = b_0 + b_1 age \tag{13.14}$$
$$y \sim Bern(p_{pass}) \tag{13.15}$$

Note that we can write the odds as $p/(1-p)$, p is a probability (or a proportion). So the logodds that corresponds to the probability of passing the exam, p_{pass}, can be written as $ln\frac{p_{pass}}{1-p_{pass}}$, so that we have

$$ln\frac{p_{pass}}{1-p_{pass}} = b_0 + b_1 age \tag{13.16}$$
$$y \sim Bern(p_{pass}) \tag{13.17}$$

Note that we do not have a residual anymore: the randomness around the predicted values is no longer modelled using a residual e that is normally distributed, but is now modelled by a y-variable with a Bernoulli distribution. Also note the strange relationship between the probability parameter p_{pass} for the Bernoulli distribition, and the dependent variable for the linear equation $b_0 + b_1 age$. The linear model predicts the logodds, but for the Bernoulli distribution, we use the probability. But it turns out that this model is very flexible and useful in many real-life problems. This model is often called a *logit* model: one often writes that the *logit of the probability* is predicted by a linear model.

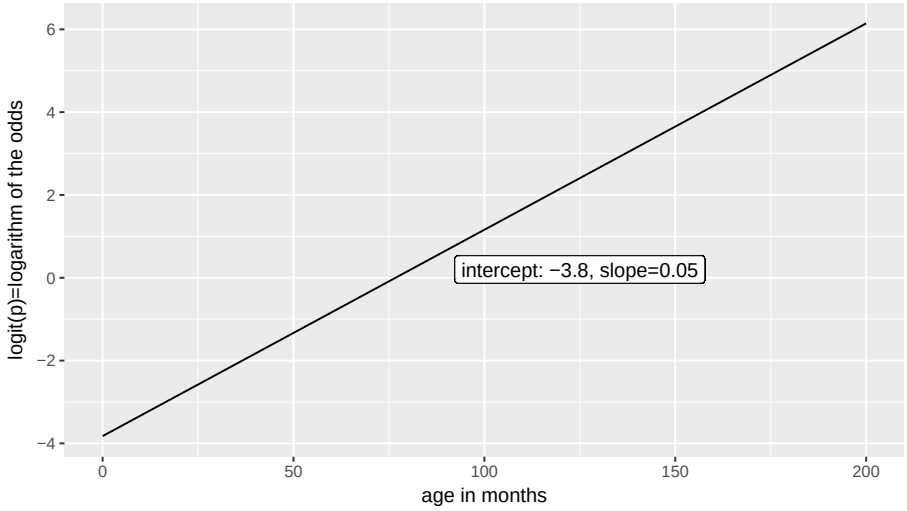


Figure 13.9: Example of a linear model for the logit of probabilities of passing an exam.

$$logit(p_{pass}) = b_0 + b_1 age \tag{13.18}$$

$$y \sim Bern(p_{pass}) \tag{13.19}$$

In essence, the logit function transforms a p-value into a logodds:

$$logit(p) = ln(\frac{p}{1-p}) \tag{13.20}$$

So what does it look like, a linear model for logodds (or logits of probabilities)?

In Figure 13.9 we show a hypothetical example of a linear model for the logit of probabilities of passing an exam. These logits or logodds are predicted by age using a straight, linear regression line:

When we take all these predicted logodds and convert them back to probabilities, we obtain the plot in Figure 13.10. Note the change in the scale of the vertical axis, the rest of the plot is the same as in Figure 13.9.

Here again we see the S-shape relationship between probabilities and the logodds. We see that our model predicts probabilities close to 0 for very young ages, and probabilities close to 1 for very old ages. There is a clear positive effect of age on the probability of passing the exam. But note that the relationship is not linear on the scale of the probabilities: it is linear on the scale of the logit of the probabilities, see Figure 13.9!

The curvilinear shape we see in Figure 13.10 is called a *logistic* curve. It is based on the logistic function: here p is a logistic function of age (and note the similarity with Equation 13.13):

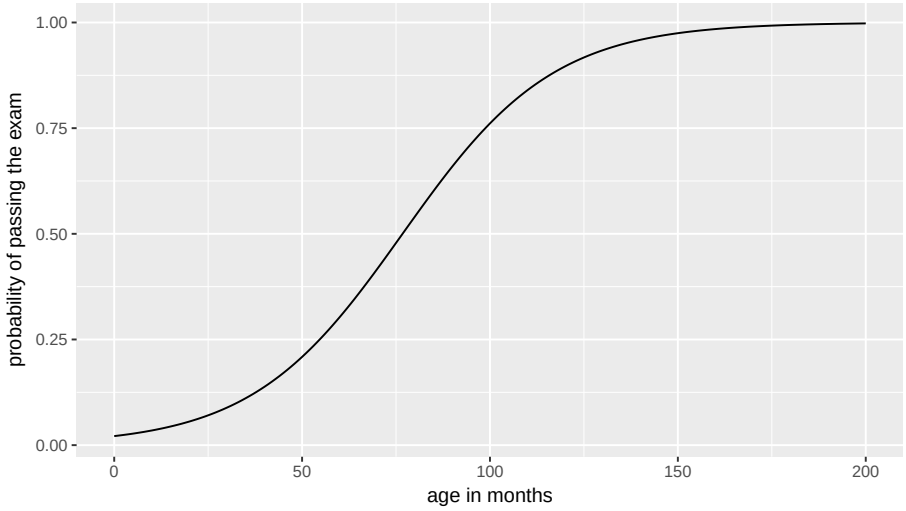


Figure 13.10: Example with logodds transformed into probabilties (vertical axis).

$$p = logistic(b_0 + b_1 age) = \frac{exp(b_0 + b_1 age)}{1 + exp(b_0 + b_1 age)} \tag{13.21}$$

In summary, if we go from logodds to probabilties, we use the logistic function, $logistic(x) = \frac{exp(x)}{1+exp(x)}$. If we go from probabilities to logodds, we use the logit function, $logit(p) = ln\frac{p}{1-p}$. The logistic regression model is a generalized linear model with a logit link function, because the linear equation $b_0 + b_1 X$ predicts the logit of a probability. It is also often said that we're dealing with a logistic link function, because the linear equation gives a value that we have to subject to the logistic function to get the probability. Both terms, logit link function and logistic link function can be used.

If we go back to our data on the third-grade children that either passed or failed the exam, we see that this curve gives a description of our data, see Figure 13.11. The model predicts that around the age of 75 months, the probability of passing the exam is around 0.50. We indeed see in Figure 13.11 that some children pass the exam (score=1) and some don't (score=0). On the basis of this analysis there seems to be a positive relationship between age in third-grade children and the probability of passing the exam in this sample.

What we have done here is a *logistic regression* of passing the exam on age. It is called logistic because the curve in Figure 13.11 has a logistic shape. Logistic regression is one specific form of a *generalized linear model*. Here we have applied a generalized linear model with a so-called *logit link function*: instead of modelling dependent variable y directly, we have modelled *the logit of the probabilities of obtaining a y-value of 1*. There are many other link functions

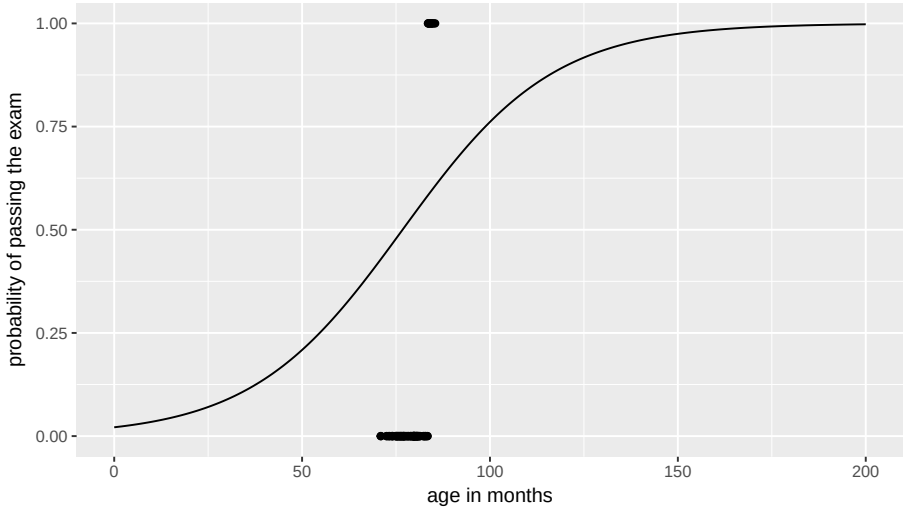


Figure 13.11: Transformed regression line and raw data points.

possible. One of them we will see in the section on generalized linear models for count data. But first, let's see how logistic regression can be performed in SPSS, and how we should interpret the output.

13.3 Logistic regression in SPSS

Imagine a data set on travellers from Amsterdam to Paris. From 1000 travellers, randomly sampled in 2017, we know whether they took the train to Paris, or whether they used other means of transportation. Of these travellers, we know their age, sex, yearly income, and whether they are travelling for business or not.

Part of the data are displayed in Table 13.1. A score of 1 on the variable **train** means they took the train, a score of 0 means they did not.

Table 13.1: Taking the train to Paris data.

train	age	sex_male	income	business
1	35.12	1	7544.00	1
1	66.66	1	7096.00	0
0	42.77	1	29261.00	1
0	72.63	0	24977.00	0
1	76.25	0	876.00	1
0	19.87	1	126943.00	1

Suppose we want to know what kind of people are more likely to take the

train to Paris. We can use a logistic regression analysis to predict whether people take the train or not, on the basis of their age, sex, income, and main purpose of the trip.

Let's first see whether income predicts the probability of taking the train. The syntax for such a model involves the GENLIN procedure, which stands for GENeralized LINear model.

```
GENLIN train (REFERENCE=FIRST) WITH income
  /MODEL income
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES SOLUTION.
```

Note the similary with the GLM and MIXED procedures: start with the dependent variable (**train** in this case, with only two possible values) and then after the WITH word the variables that you'd like to treat quantitatively, here **income**. Under the MODEL subcommand we specify the model, here only a main effect of **income**. But further we have to specify that we want to use the Bernoulli distribution and a logit link function. So LINK=LOGIT, but why a binomial distribution? Well, a Bernoulli distribution (one coin flip) is only a special case of the Binomial distribution (the distribution of several coin flips). So here we use a binomial distribution for one coin flip, which is equivalent to a Bernoulli distribution. The last line indicates what type of output we want to see: case processing statistics, descriptives and the solution in terms of parameter estimates.

One very important part of the syntax is the (REFERENCE = FIRST) statement for the dependent variable. The default SPSS syntax uses (REFERENCE = LAST), so that's what you get when you do not specify this part. (REFERENCE = LAST) means that the reference category of the train variable is the last value. Since there are only two values, 0 an 1, the last value is equal to 1. In that case, SPSS will derive a model that predicts the logoddss for NOT taking the train, since it estimates the effect of income on the dependent variable *relative to taking the train*. In our case, it makes more sense to derive a model for the logoddss of taking the train. We want to predict logodddsratios for taking the train, so we need to specify that our first value, 0, is our reference category: (REFERENCE = FIRST).

In Figure 13.12 we see the parameter estimates from this generalized linear model run on the train data.

The parameter estimates table from a GENLIN analysis looks very much like that of the ordinary linear model and the linear mixed model. The only difference is that we no longer see t-statistics, but Wald Chi-Square statistics. This is because with logistic models, the ratio B/SE does not have a t-distribution. In ordinary linear models, the ratio B/SE has a t-distribution because in linear models, the variance of the residuals, σ_e^2, has to be estimated. If the residual variance was known, B/SE would have a standard normal distribution. In logistic models, there is no σ_e^2 that needs to be estimated, so the ratio B/SE

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	90.017	32.5180	26.283	153.751	7.663	1	.006
income	-.008	.0030	-.014	-.002	7.541	1	.006
(Scale)	1[a]						

Dependent Variable: train
Model: (Intercept), income

a. Fixed at the displayed value.

Figure 13.12: SPSS output of a generalized linear model for predicting taking the train from income.

has a standard normal distribution[3]. One could therefore calculate a Z-statistic $Z = B/SE$ and see whether that value is smaller than 1.96 or larger than 1.96, if you want to test with a Type I error rate of 0.05. SPSS has chosen to not compute such a Z-statisic, but to compute a chi-square statistic $X^2 = B^2/SE^2$. This chi-square or X^2-statistic has a χ^2 distribution with 1 degree of freedom. Both approaches, computing Z or X^2, are equivalent.

The interpretation of the B-parameters is very similar to other linear models. Note that we have the following equation for the logistic model:

$$\begin{aligned} logit(p_{train}) &= b_0 + b_1 income \\ train &\sim Bern(p_{train}) \end{aligned} \tag{13.22}$$

If we fill in the values from the SPSS output, we get

$$\begin{aligned} logit(p_{train}) &= 90.017 - 0.008 \times income \\ train &\sim Bern(p_{train}) \end{aligned} \tag{13.23}$$

We can interpret these results by making some predictions. Imagine a traveller with a yearly income of 11,000 Euros. Then the predicted logodds equals $90.017 - 0.008 \times 11000 = 2.017$. When we transform this back to a probability, we get $\frac{exp(2.017)}{1+exp(2.017)} = 0.542$. So this model predicts that for people with a yearly

[3]This is the reason why you see (scale) equal to constant 1 in the SPSS output, right under the parameter for **income**. In the logistic model, the variance (scale) is fixed (assumed known).

income of 11,000, about 52% of them take the train (if they travel at all, that is!).

Now imagine a traveller with a yearly income of 100,000. Then the predicted logodds equals $6.752 - 0.001 \times 100000 = -709.983$. When we transform this back to a probability, we get $\frac{exp(-709.983)}{1+exp(-709.983)} = 0$. So this model predicts that for people with a yearly income of 100,000, close to none of them take the train. Going from 11,000 to 100,000 is a big difference. But the change in probabilities is also huge: it goes down from 0.52 to 0.

We found a difference in this sample of 1000 travellers, but is there also a difference in the entire population of travellers between Amsterdam and Paris? The SPSS table shows us that the effect of income, -0.008, is statistically significant, $X^2(1) = 7.541, p < 0.01$. We can therefore reject the null-hypothesis that income is not related to whether people take the train or not.

Note that similar to other linear models, the intercept can be interpreted as the predicted logodds for people that have values 0 for all other variables in the model. Therefore, 90.017 means in this case that the predicted logodds for people with zero income equals 90.017. This is equivalent to a probability of very close to 1.

13.3.1 Exercises

Using the train data, we try to predict whether people take the train or not by their purpose of their trip: business or not.

1. What does the SPSS syntax look like? Note the data in Table 13.1.
2. Suppose the results look like those in Figure 13.13. What is the predicted probability of taking the train for people that travel for business? Provide the calculations.
3. Suppose the results look like those in Figure 13.13. What is the predicted probability of taking the train for people that travel NOT for business? Provide the calculations.
4. Suppose the results look like those in Figure 13.14. What is the predicted probability of taking the train for people that travel for business? Provide the calculations.
5. Suppose the results look like those in Figure 13.14. What is the predicted probability of taking the train for people that travel NOT for business? Provide the calculations.
6. On the basis of this SPSS output, do business travellers tend to take the train more or less often than non-business travellers? Motivate your answer.
7. Suppose in SPSS output for logistic regression, you find an intercept value of 0.5 with a standard error of 0.1. There is a corresponding Wald chi-square value of 25. Explain where this Wald chi-square value comes from.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.155	.1196	-1.389	-.921	93.321	1	.000
business	-.050	.1531	-.351	.250	.108	1	.742
(Scale)	1[a]						

Dependent Variable: train
Model: (Intercept), business

a. Fixed at the displayed value.

Figure 13.13: SPSS output of a generalized linear model for predicting taking the train from purpose of the trip.

8. Suppose we have the data on coin flips in following table:

ID	Heads	weight	type
1	0	2.7831226	5cents
2	1	0.8058492	10cents
3	1	3.1401581	1Euro
4	1	1.0156831	10cents
5	1	4.4503490	1Euro

If we want to predict the outcome of the coin flip, on the basis of the type of coin, should we use a linear model, a linear mixed model, or a generalized linear model? Motivate your answer.

If we want to predict the weight of the coin, on the basis of the type of the coin, should we use a linear model, a linear mixed model, or a generalized linear model? Motivate your answer.

Answers:

1. It could look like this (using WITH, treating the independent variable as quantitative):

```
GENLIN train (REFERENCE=FIRST) WITH business
  /MODEL business
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES   SOLUTION.
```

or like this (using BY, treating the independent variable as qualitative)

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.205	.0957	-1.393	-1.018	158.757	1	.000
[business=.00]	.050	.1531	-.250	.351	.108	1	.742
[business=1.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: train
Model: (Intercept), business

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 13.14: SPSS output of a generalized linear model for predicting taking the train from purpose of the trip.

```
GENLIN train (REFERENCE=FIRST) BY business
  /MODEL business
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES   SOLUTION.
```

2. People that travel for business score 1 on the business variable. So the predicted logodds for those people is $-1.155 - 0.050 \times 1 = -1.205$. The odds is the $exp(-1.205) = 0.299692$. So the odds of going by train are 0.30 to 1. This is equivalent to 3 to 10. So suppose we have 13 trips, 3 are by train and 10 are not by train. So the probability of a trip being by train equals $3/13 = 0.23$.

3. People that travel NOT for business score 0 on the business variable. So the predicted logodds for those people is $-1.155 - 0.050 \times 0 = -1.155$. The odds is the $exp(-1.155) = 0.3150575$. So the odds of going by train are 0.32 to 1. This is equivalent to 32 to 100. So suppose we have 132 trips, 32 are by train and 100 are not by train. So the probability of a trip being by train equals $32/132 = 0.24$.

4.

5.

6. If we want to predict the outcome of the coin flip, on the basis of the type of coin, we should use a generalized linear model, because the dependent variable is dichotomous (has only 2 values), so the residuals can never have a normal distribution.

   If we want to predict the weight of the coin, on the basis of the type

of the coin, we should use a linear model, because the dependent variable is continuous.

Chapter 14

Generalized linear models for count data: Poisson regression

14.1 Poisson regression

Count data are inherently discrete, and often when using linear models, we see non-normal distributions of residuals. Let's go back to the beginning of this chapter, where we discussed a data set on the scores that a group of students got for an assignment. There were four criteria, and the score consisted of the number of criteria that were met for each student's assignment. Figure 13.3 showed that after an ordinary linear model analysis, the residuals did not look normal at all.

Table 14.1 shows part of the data that were analysed. Similar to logistic regression, perhaps we can find a distribution other than the normal distribution that is more suitable for this kind of data? For dichotomous data (1/0) we found the Bernoulli distribution very useful. For count data, the traditional distrbitution is the Poisson distribution.

Table 14.1: Scores on an assignment.

ID	score	previous
1	0	0.41
2	2	-0.47
3	4	0.07
4	0	-0.50
5	2	-0.83
6	3	0.17

The normal distribution has two parameters, the mean and the variance. The

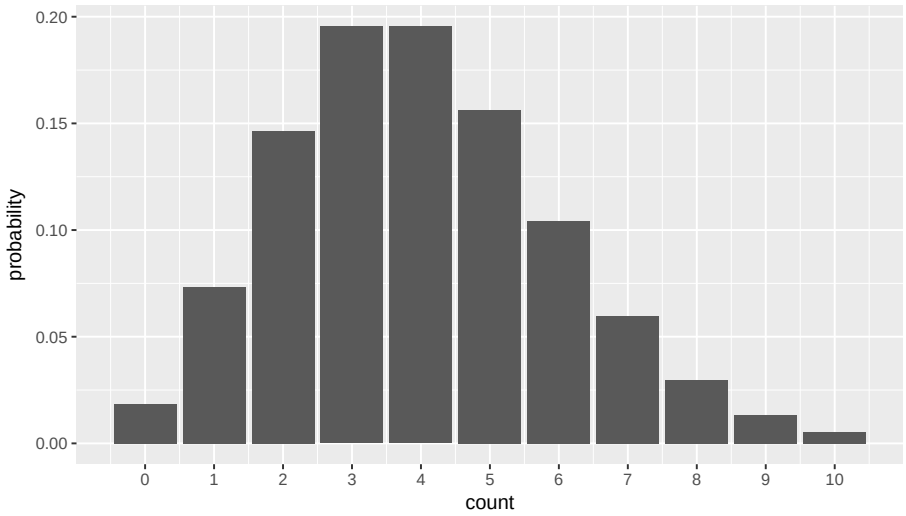


Figure 14.1: Count data example where the normal distribution is not a good approximation of the distribution of the residuals.

Bernoulli distribution has only 1 parameter (the probability), and the Poisson distribution has also only 1 parameter, lambda or λ. λ is a parameter that indicates tendency. Figure 14.1 shows a Poisson distribution with a tendency of 4.

What we see is that many values center around the tendency parameter value of 4 (therefore we call it a tendency parameter)! We see only discrete values, and no values below 0. We see a few values higher than 10. If we take the mean of the distribution, we will find a value of 4. If we would compute the variance of the distribution we would find also find 4! In general, if we have a Poisson distribution with a tendency parameter $\lambda = 4$, we know that both the mean and the variance will be equal to λ.

A Poisson model could be suitable for our data: a linear equation could predict the parameter λ and then the actual data show a Poisson distribution.

$$\lambda = b_0 + b_1 X \tag{14.1}$$

$$y \sim Poisson(\lambda) \tag{14.2}$$

However, because of the additivity assumption, the equation $b_0 + b_1 X$ leads to negative values. A negative value for λ is not logical, because we then have a tendency to observe data like -2 and -4 in our data, which is contrary to the having count data, which consists of non-negative integers. A Poisson distribution always shows integers of at least 0, so one or way or another we have to make sure that we always have a *lambda* of at least 0.

Remember that we saw the reverse problem with logistic regression: there

we wanted to have negative values for our dependent variable logoddsratio, so therefore we used the logarithm. Here we want to have positive values for our dependent variable, so we can use the inverse of the logarithm function: the exponential. Then we have the following model:

$$\lambda = exp(b_0 + b_1 X) = e^{b_0 + b_1 X} \tag{14.3}$$

$$y \sim Poisson(\lambda) \tag{14.4}$$

This is a generalized linear model, now with a Poisson distribution and an exponential link function. The exponential function makes any value positive, for instance $exp(0) = 1$ and $exp(-100) = 0$.

Let's analyze the assignment data with this generalized linear model. Our dependent variable is the number of criteria met for the assignment (a number between 0 and 4), and the independent variable is previous, which is a standardized mean of a number of previous assignments. We expect that the mean score on previous assignments is associated with a higher score on the present assignment. When we run the analysis, the result is as follows:

$$\lambda = exp(0.1576782 - 0.0548685 \times previous) \tag{14.5}$$

$$score \sim Poisson(\lambda) \tag{14.6}$$

What does it mean? Well, similar to logistic regression, we can understand such equations by making some predictions for interesting values of the independent variable. For instance, a value of 0 for **previous** means an average grade on previous advanced that is around the mean value. So if we choose **previous**=0, then we have the prediction for an average student. If we fill in that value, we get the equation $\lambda = exp(0.1576782 - 0.0548685 \times 0) = exp(0.1576782) = 1.17$. Thus, for an avarage student, we expect to see a score of 1.17. A Poisson distribution with $\lambda = 1.17$ is depicted in Figure 14.2.

Another interesting value of **previous** might be -2. That represents a student with generally very low grades. Because the average grades were standardized, only about 2.5% of the students has lower average grade than -2. If we fill in that value, we get: $\lambda = exp(exp(0.1576782 - 0.0548685 \times -2) = 0.85$. A Poisson distribution with $\lambda = 1.31$ is depicted in Figure 14.3.

The last value of **previous** for which we calculate λ is +2, representing a high-performing student. We then get $\lambda = exp(0.1576782 - 0.0548685 \times 2) = 1.6$. A Poisson distribution with $\lambda = 1.6$ is depicted in Figure 14.4.

If we superimpose these figures, we obtain Figure **??**, where we see that the higher the average score on previous assignments, the higher is the expected score on the present assignment.

We found that in this data set, previous high marks for assignments predicted a higher mark for the present assignment. In the next section we see how to perform the analysis in SPSS, and check whether there is also a relationship in the population of students.

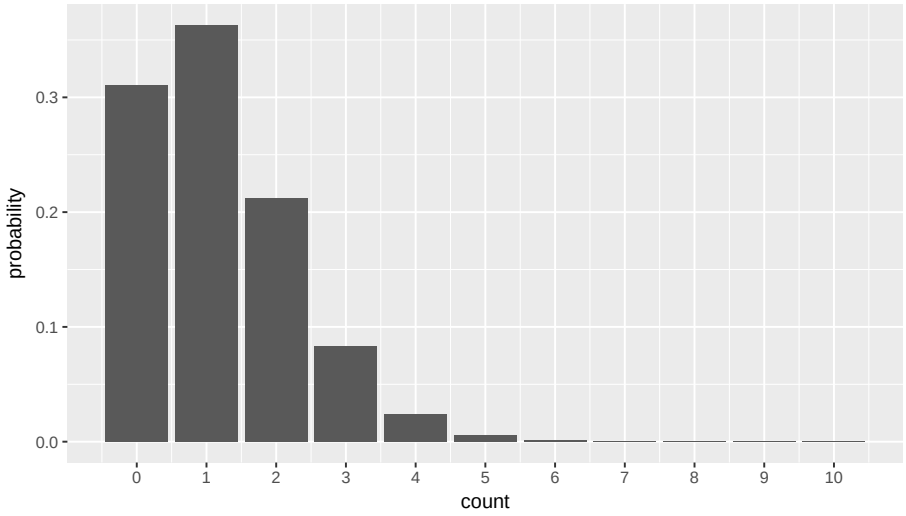


Figure 14.2: Poisson distribution with lambda=1.17.

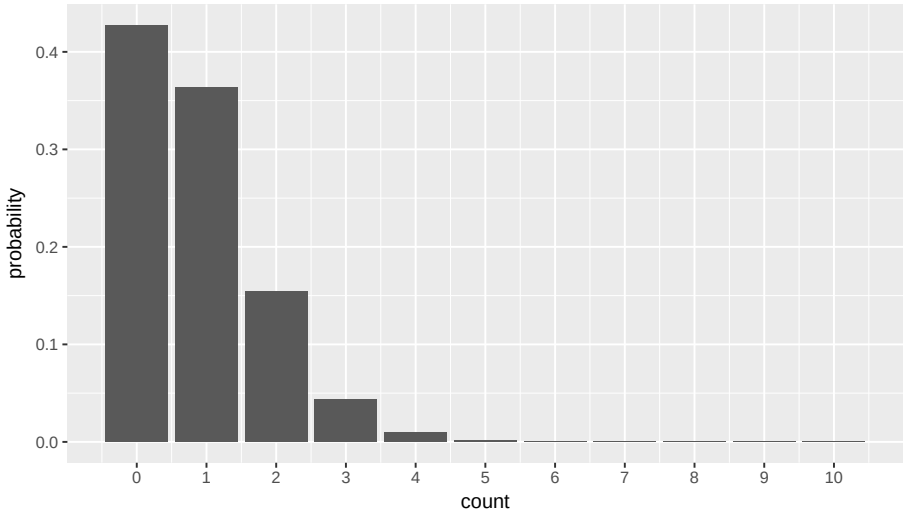


Figure 14.3: Poisson distribution with lambda=0.85.

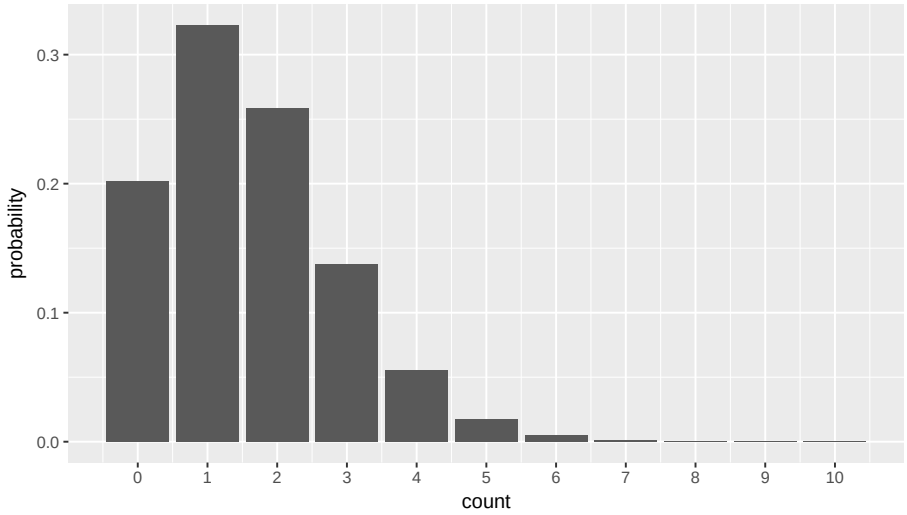


Figure 14.4: Poisson distribution with lambda=1.60.

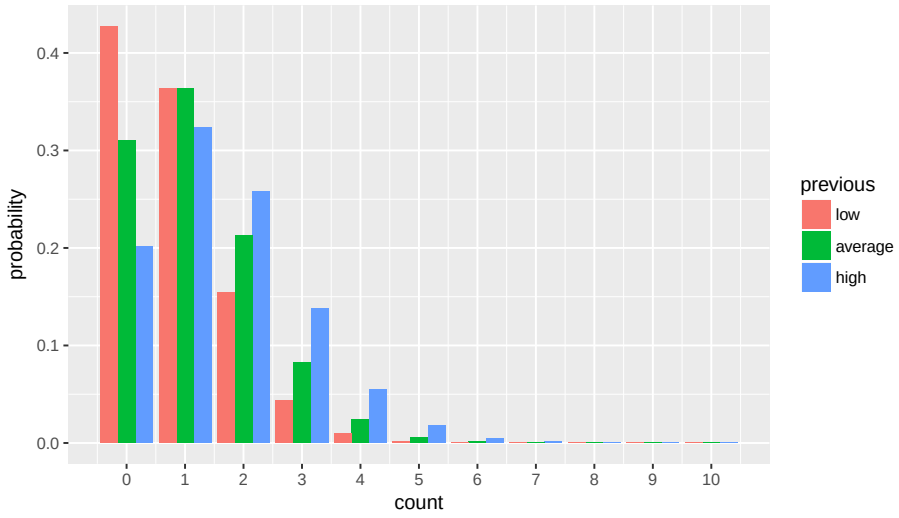


Figure 14.5: Three different Poisson distributions with lambdas 0.85, 1.17, and 1.60, for three different kinds of students.

14.2 Poisson regression in SPSS

Poisson regression is form of a generalized model analysis, similar to logistic regression. However, instead of using a Bernoulli distribution we use Poisson distribution. For a quantitative predictor like the variable **previous**, the syntax is as follows.

```
GENLIN scores WITH previous
  /MODEL previous
 DISTRIBUTION=POISSON LINK=LOG
  /PRINT CPS DESCRIPTIVES SOLUTION.
```

The output with parameter values is shown in Figure 14.6.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.158	.0925	-.024	.339	2.908	1	.088
previous	-.055	.0899	-.231	.121	.372	1	.542
(Scale)	1[a]						

Dependent Variable: score
Model: (Intercept), previous

a. Fixed at the displayed value.

Figure 14.6: SPSS output of a generalized linear model for predicting assignments scores from the average of previous assignments.

We see the same values for the intercept and the effect of **previous** as in the previous section. We now also see 95% confidence intervals for these parameter values. For both, the value 0 is included in the confidence intervals, therefore we know that we cannot reject the null-hypotheses that these values are 0 in the population of students. This is also reflected by the Wald statistics. Remember that the Wald chi-square (X^2) statistic is computed by B^2/SE^2. For large enough samples, these X^2 statistics follow a χ^2 distribution with 1 degree of freedom. From that distribution we know that a value of 0.372 is not significant at the 5% level. It has an associated p-value of 0.542.

We can write:

> Scores for the assignment (1-4) for 100 students were analysed using a generalized linear model with a Poisson distribution (Poisson regression). The scores were not significantly predicted by the average score of previous assignments, $B = -0.06, X^2(1) = 0.37, p =$

0.54. Therefore we cannot reject the null-hypothesis that there is no relationship between the average of previous assignments and the score on the present assignment in the population of students.

Suppose we also have a qualitative predictor, for example degree that the students are working for. Some do the assignment for bachelor's degree (degree=1), some for a master's degree (degree=2), and some for a PhD (degree=3). The syntax would then look like:

```
GENLIN scores BY degree
  /MODEL degree
 DISTRIBUTION=POISSON LINK=LOG
  /PRINT CPS DESCRIPTIVES SOLUTION.
```

Note that only the independent variable has changed and the WITH statement is changed into BY. The output is given in Figure 14.7.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.354	.1459	.068	.640	5.878	1	.015
[degree=1.00]	-.584	.2415	-1.057	-.111	5.852	1	.016
[degree=2.00]	-.089	.2110	-.503	.325	.178	1	.673
[degree=3.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: score
Model: (Intercept), degree

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 14.7: SPSS output of a generalized linear model for predicting assignments scores from the degree that is studied for.

We see that the parameter for the degree=3 category is fixed to 0, meaning that it is used as the reference category. If we make a prediction for this group of students that is studying for a PhD degree, we have $\lambda = exp(.354 + 0) = exp(0.354) = 1.4$. For the students studying for a Master's degree we have $\lambda = exp(.354 - 0.089) = 1.3$ and for students studying for their Bachelor's degree we have $\lambda = exp(.354 - 0.584) = 0.8$. These λ-values correspond to the expected number in a Poisson distribution, so for Bachelor students we expect a score of 0.8, for Master students we expect a score of 1.3 and for Phd students a score of 1.4. Are these different scores also present in the population? We see that the effect for degree=1 is significant, $X^2(1) = 5.85, p = 0.02$, so there is a difference in score between students studying for a Bachelor's degree and students studying for a PhD. The effect for degree=2 is not significant,

$X^2(1) = 0.18, p = 0.67$, so there is no difference in assignment scores between Master students and PhD students.

Remember that for the linear model, when we wanted to compare more than two groups at the same time, we used an F-test to test for an overall difference in group means. Also for the generalized linear model, we might be interested in whether there is an overall difference in scores between Bachelor, Master and PhD students. For that we need to tweak the syntax a little bit, by stating that we also want to see an overall test printed. The PRINT statements then also needs the word SUMMARY. In other words, the syntax becomes

```
GENLIN scores BY degree
  /MODEL degree
 DISTRIBUTION=POISSON LINK=LOG
  /PRINT CPS DESCRIPTIVES SOLUTION SUMMARY.
```

We then get the relevant output in Figure 14.8. There we see a Wald Chi-Square statistic for the effect of **degree**. It has 2 degrees of freedom, since the effect for the 3 categories is coded by 2 dummy variables. So this test tells us that the null-hypothesis that the expected scores in each group of students are the same can be rejected, $X^2(2) = 6.27, p = 0.04$.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	1.844	1	.175
degree	6.271	2	.043

Dependent Variable: score
Model: (Intercept), degree

Figure 14.8: SPSS output of a generalized linear model for predicting assignments scores from the degree that is studied for.

14.3 Interaction effects in Poisson models

In the previous subsection we looked at a count variable, the number of criteria fulfilled, and we wanted to predict it from the degree that students were studying

for. Let's look at an example where we want to predict a count variable from two qualitative predictors.

In 1912, the ship Titanic sank after the collision with an iceberg. There we 2201 people on board that ship. Some of these were male, others were female. Some were passengers, others were crew, and some survived, and some did not. For the passengers there were three groups: those travelling first class, second class and third class. There were also children on board. If we focus on only the adults, suppose we want to know whether there is a relationship between the sex and the counts of people that survived the disaster. The table in 14.2 gives the counts of survivors for males and females separately.

Table 14.2: Counts of adult survivors on the Titanic.

	count
Male	338
Female	316

Let's analyse this small data set with SPSS. In SPSS we assign the value sex=1 to Females and sex=2 to Males. Our dependent variable is count, and the independent variable is sex.

```
GENLIN count BY sex
  /MODEL sex
 DISTRIBUTION=POISSON LINK=LOG
  /PRINT CPS DESCRIPTIVES SOLUTION.
```

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	5.823	.0544	5.716	5.930	11460.858	1	.000
[sex=1.00]	-.067	.0783	-.221	.086	.740	1	.390
[sex=2.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: count
Model: (Intercept), sex

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 14.9: SPSS output of a generalized linear model for predicting numbers of men and women onboard the Titanic.

From the output in Figure 14.9 we see that the expected count for females is $exp(5.823 - 0.067) = 318.3$ and the expected count for males is $exp(5.823) = 340.4$. These expected counts are close to the observed counts of males and females. The only reason that they differ from the observed is because of rounding errors (SPSS shows only the first three decimals). From the Wald statistic, we see that the difference in counts between males and females is not significant, $X^2(1) = 0.74, p = 0.39$[1].

The difference in these counts is very small. But does this tell us that women were as likely to survive as men? Note that we have only looked at those who survived. How about the people that perished: were there more men that died then women? Table 14.3 shows the counts of male survivors, female survivors, male non-survivors and female non-survivors. Then we see a different story: on the whole there were many more men than women, and a relatively small proportion of the men survived. Of the men, most of them perished: 1329 perished and only 338 survived, a survival rate of 20.3%. Of the women, most of them survived: 109 perished and 316 survived, yielding a survival rate of 74%. Does this tell us that women are much more likely than men to survive collisions with icebergs?

Table 14.3: Counts of adults on the Titanic.

sex	survived	count
Male	0	1329
Female	0	109
Male	1	338
Female	1	316

Let's first run a multivariate Poisson regression analysis including the effects of both sex and survival. The syntax is

```
GENLIN count BY sex WITH survived
  /MODEL sex survived
 DISTRIBUTION=POISSON LINK=LOG
  /PRINT CPS DESCRIPTIVES SOLUTION.
```

where we treat sex qualitatively and survival quantitatively for convenience (**survived** is alread coded as a dummy, **sex** is not).

The output is given in Figure 14.10. From the parameter values, we can calculate the predicted numbers of male (sex = 2) and female (sex = 1) that survived and perished. For female survivors we have $exp(7.04 - 1.37 - .79) = 131.63$, for female non-survivors we have $exp(7.04 - 1.37) = 290.03$, for male survivors we have $exp(7.04 - .79) = 518.01$ and for male non-survivors we have $exp(7.04) = 1141.39$.

[1]Note that a hypothesis test is a bit odd here: there is no clear population that we want to generalize the results to: there was only one Titanic disaster. Also, here we have data on the entire population of those people on board the Titanic, there is no random sample here.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	7.044	.0286	6.988	7.100	60709.658	1	.000
[sex=1.00]	-1.367	.0543	-1.473	-1.260	632.563	1	.000
[sex=2.00]	0[a]	.	.	.	.	.	.
survived	-.788	.0472	-.880	-.695	279.073	1	.000
(Scale)	1[b]						

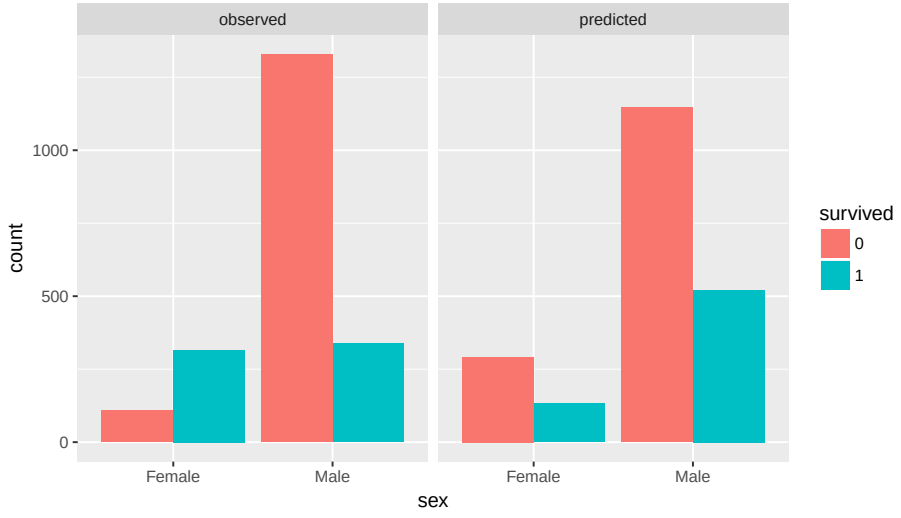
Dependent Variable: count
Model: (Intercept), sex, survived

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 14.10: SPSS output of a generalized linear model for predicting numbers of men and women that perished and survived onboard the Titanic.

These predicted numbers are displayed in Figure **??**. It also shows the observed counts. The pattern that is observed is clearly different from the one that is predicted from the generalized linear model. The linear model predicts that there are fewer survivors then non-survivors, irrespective of sex, but we observed that in females, there are more survivors then non-survivors. It seems that sex is is moderator of the effect of survival on counts.



In order to test this moderation effect, we run a new generalized linear model for counts including an interaction effect of sex by survived. This is done in SPSS syntax by changing the MODEL part by includding a sex*survived interaction:

```
GENLIN count BY sex WITH survived
  /MODEL sex survived sex*survived
 DISTRIBUTION=POISSON LINK=LOG
  /PRINT CPS DESCRIPTIVES SOLUTION.
```

The output is displayed in Figure 14.11.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	7.192	.0274	7.138	7.246	68745.825	1	.000
[sex=1.00]	-2.501	.0996	-2.696	-2.306	630.032	1	.000
[sex=2.00]	0[a]	.	.	.	.	.	.
survived	-1.369	.0609	-1.489	-1.250	505.126	1	.000
[sex=1.00] * survived	2.434	.1267	2.185	2.682	368.979	1	.000
[sex=2.00] * survived	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

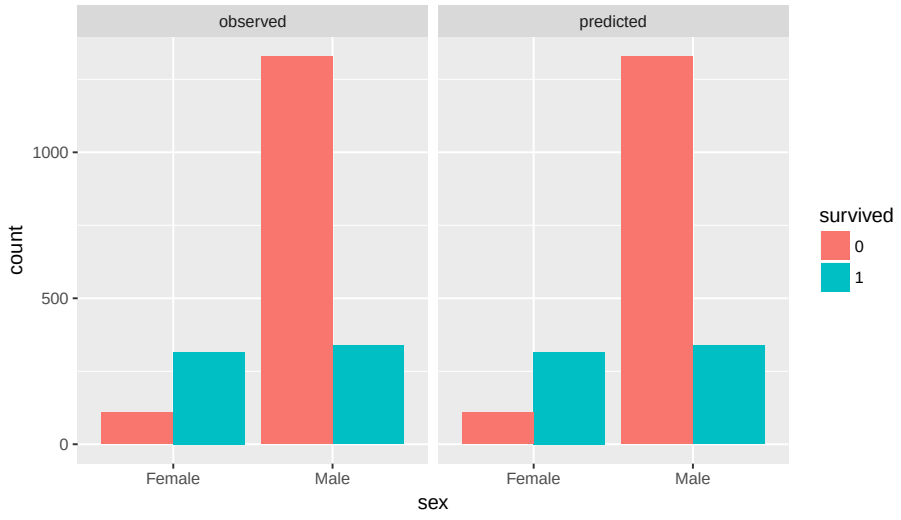
Dependent Variable: count
Model: (Intercept), sex, survived, sex * survived

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 14.11: SPSS output of a generalized linear model for predicting numbers of men and women that perished and survived onboard the Titanic.

When we plot the predicted counts from this new model with an interaction effect, we see that they are exactly equal to the counts that are actually observed in the data, see Figure **??**.



From the output we see that the interaction effect is significant, $X^2(1) = 91.82, p = 0$. If we regard this data set as a random sample of all ships that sink after collision with an icebergs, we may conclude that in such situations, sex is a significant moderator of the difference in the numbers of survivors and non-survivors. One could also say: the proportion of people that survive a disaster like this is different in females than it is in males. Here we saw a higher survival rate in women than in men.

14.4 Crosstabulation and the Pearson chi-square statistic

The data on male and female survivors and non-nonsurvivors are often tabulated in a cross-table like in Table 14.4

Table 14.4: Counts of adult survivors and non-survivors on the Titanic.

	No	Yes
Male	1329	338
Female	109	316

In the previous section these counts were analysed using a generalized linear model with a Poisson distribution and an exponential link function. We wanted to know whether there was a significant difference in the proportion of survivors for men and women. In this section we discuss an alternative method of analyzing count data. We discuss an alternative chi-square (X^2) statistic for the moderation effect of one variable of the effect of another variable.

First let's have a look at the overall survival rate. In total there we 654 people that survived and 1438 people that did not survive. Table 14.5 shows these column totals.

Table 14.5: Counts of adult survivors and non-survivors on the Titanic.

	No	Yes
Male	1329	338
Female	109	316
Total	1438	654

Looking at these total numbers of survivors and non-survivors, we can calculate the proportion of survivors overall (the survival rate) as $654/(654+1438) = 0.31$.

Table **??** shows the totals for men and women, as well as the overall total number of adults.

Suppose we only know that of the 2092 people, 1667 were men, and of all people, 654 survived. Then suppose we pick a random person from these 2092 people. What is the probability that we get a male person that survived, *given that sex and survival have nothing to do with eachother*?

Table 14.6: Counts of adult survivors and non-survivors on the Titanic.

	No	Yes	Total
Male	1329	338	1667
Female	109	316	425
Total	1438	654	2092

Well, from probability theory we know that if two events A and B are independent, the probability of observing A and B at the same time, is equal to the product of the probability of event A and the probabilty of event B.

$$Prob(AandB) = Prob(A) \times Prob(B) \tag{14.7}$$

If sex and survival are independent from eachother, then the probabilty of observing a male survivor is equal to the probability of seeing a male times the probability of seeing a survivor. The probability for survival is 0.31, as we saw earlier, and the probability of seeing a male is equal to the proportion of males in the data, which is $1667/2092 = 0.8$. Therefore, the probability of seeing a male survivor is $0.8 \times 0.31 = 0.24$. The expected number of male survivors is then that probability times the total number of people, $0.24 \times 2092 = 502.08$. Similarly we can calculate the expected number of non-surviving males, the number of surviving females, and the number of non-surviving females.

These numbers, after rounding, are displayed in Table 14.7.

Table 14.7: Expected numbers of adult survivors and non-survivors on the Titanic.

	No	Yes
Male	1155	519
Female	289	130

The expected numbers in Table 14.7 are quite different from the observed numbers in Table 14.4. Are the differences large enough to think that the two events of being male and being a survivor are NOT independent? If the expected numbers on the assumption of independence are different enough from the observed numbers, then we can reject the null-hypothesis that being male and being a survivor have nothing to do with eachother. To measure the difference between expected and observed counts, we need a test statistic. Here we use Pearson's chi-square statistic. It involves calculating the difference between the numbers in the respective cells, and standardize them by the expected number. Here's how it goes:

For each cell, we take the predicted count subtract it from the observed count. For instance, for the male survivors, we expected 519 but observed 338. The difference is therefore $338 - 519 = -181$. Then we take the square of this difference, $181^2 = 32761$. Then we divide this number by the expected number, and then we get $32761/519 = 63.1233141$. We do exactly the same thing for the

male non-survivors, the female survivors and the female non-survivors. Then we add these 4 numbers, and then we have the Pearson chi-square statistic. In formula form:

$$X^2 = \Sigma_i \frac{(O_i - E_i)^2}{E_i} \tag{14.8}$$

So for male survivors we get

$$\frac{(338 - 519)^2}{519} = 63.1233141 \tag{14.9}$$

For male non-survivors we get

$$\frac{(1329 - 1155)^2}{1155} = 26.212987 \tag{14.10}$$

For female survivors we get

$$\frac{(316 - 130)^2}{130} = 266.1230769 \tag{14.11}$$

and for female non-survivors we get

$$\frac{(109 - 289)^2}{289} = 112.1107266 \tag{14.12}$$

If we add these 4 numbers we have the chi-square statistic: $X^2 = 467.57$. Note that we only use the rounded expected numbers. Better would be to use the non-rounded numbers. Had we used the non-rounded expected numbers, we would have gotten $X^2 = 460.87$.

The Wald chi-square statistic for the sex*survived interation effect was 368.9788928, see Figure 14.11. It tests exactly the same null-hypothesis as the Pearson chi-square: that of independence, or in other words, that the numbers can be explained by only two main effects, sex and survival.

If the data set is large enough and the numbers are not too close to 0, the same conclusions will be drawn, whether from a Wald chi-square for an interaction effect in a generalized linear model, or from a crosstabulation and computing a Pearson chi-square. The advantage of the generalized linear model approach is that you can do much more with them, for instance more than two predictors, and that you make it more explicit that when computing the statistic, you take into account the main effects of the variables. You do that also for the Pearson chi-square but it is less obvious: we did that by first calculating the probability of survival and second calculating the proportion of males.

14.5 Poisson regression or logistic regression?

In the previous section we analyzed the relationship between the variable **sex** of the person onboard the Titanic, and the variable **survived**: whether or not a person survived the shipwreck. We found a relationship between these

two variables by studying the crosstabulation of the counts, and testing that relationship using a Pearson chi-square statistic. In the section before that, we saw that this relationship could also be tested by applying a Poisson regression model and looking at the sex by survived interaction effect. These methods are equivalent.

There is yet a third way to analyze the sex and survived variables. Remember that in the previous chapter we discussed logistic regression. In logistic regression, a dichotomous variable (a variable with only two values, say 0 and 1) is the dependent variable, with one or more quantitaive or qualitative independent variables. Both sex and survived are dichotomous variables: male and female, and survived yes or survived no. In prinicple therefore, we could do a logistic regression: for example predicting whether a person is a male or female, on the basis of whether they survived or not, or the other way around, predicting whether people survive or not, on the basis of whether a person is a women or a man.

What variable is used here as your dependent variable, depends on your research question. If your question is whether females are more likely to survive than men, perhaps because of their body fat composition, or perhaps because of male chivalry, then the most logical choice is to take survival as the dependent variable and sex as the independent variable.

The syntax for logistic regression then looks like

```
GENLIN survived (REFERENCE=FIRST) BY sex
  /MODEL business
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES   SOLUTION.
```

Note however that the data is the wrong format. For the Poisson regression, the data were there in the form of what we see in Table 14.4. However, for a logistic regression, we need the data in the format like in Table 14.8. For every person onboard the ship, we have to know their sex and their survival status.

Table 14.8: Individual data of adult survivors and non-survivors on the Titanic.

ID	sex	survived
238	Male	0
1302	Male	0
1274	Male	0
1303	Male	0
1798	Female	1
1337	Female	0
20	Male	0
485	Male	0
1389	Female	0
1072	Male	0

We use BY to treat the sex variable as qualitative. We use (REFERENCE =

FIRST) because we want to predict whether people survive (survive=1). Then our reference category is survive=0, which is the first value. In the output in Figure 14.12 we see that sex is a significant predictor of the survival status, $B = 2.434, X^2 = 368,98, p < 0.001$. The logoddsratio for a male surviving the shipwreck is -1.37, and the logoddsratio for a female surviving the shipwreck is $-1.37 + 2.43 = 1.06$. These logoddsratios correspond to probabilities of 0.20 and 0.74, respectively. Thus, some are much more likely to survive than men.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.369	.0609	-1.489	-1.250	505.126	1	.000
[sex=1.00]	2.434	.1267	2.185	2.682	368.979	1	.000
[sex=2.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: survived
Model: (Intercept), sex

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 14.12: SPSS output of a generalized linear model for predicting numbers of men and women that perished and survived onboard the Titanic.

However, suppose you are the relative of a passenger onboard a ship that shipwrecks. After two days, there is news that a person was found. The only thing known about the person is that he or she is alive. Your relative is your niece, so you'd like to know on the basis that the person that was found lives, what is the probability that that person is a woman, cause then it could be your believed niece! You could therefore run a logistic regression on the Titanic data to see to what extent the survival of a person predicts the sex of the person. The syntax would then look like this:

```
GENLIN sex (REFERENCE=LAST) WITH survived
  /MODEL survived
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES   SOLUTION.
```

Note that we use WITH in order to treat the dummy variable survived as quantitative. We also use (REFERENCE=LAST) to indicate that we use the last (second) category of sex (2) as the reference category, because that category refers to men, because we want to predict whether a person is a female.

The output is give in Figure 14.13

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-4.934	.2141	-5.354	-4.515	531.265	1	.000
survived	2.434	.1267	2.185	2.682	368.979	1	.000
(Scale)	1[a]						

Dependent Variable: sex
Model: (Intercept), survived

a. Fixed at the displayed value.

Figure 14.13: SPSS output of a generalized linear model for predicting numbers of men and women that perished and survived onboard the Titanic.

From this output we conclude that survival is a signficant predictor of sex, $B = -2.434, X^2 = 368,98, p < 0.001$. The logoddsratio for a surviving person to be a woman is $-4.93+2.43 = -2.50$, and the logoddsratio for a non-surviving person to be a woman is -4.93. These logoddsratios correspond to probabilities of 0.08 and 0.01, respectively. Thus, if you know that there is a person that survived the Titanic, it is not very likely that it was a woman, only 8% chance. If you think this is counterintuitive, remember that even though a large proportion of the women survived the Titanic, there were many more men onboard than women.

In summary if you have count data, and one of the variables is dichotomous, you have the choice whether to use a Poisson regression model or a logistic regression. The choice depends on the research question: if your question involves *prediction* of a dichotomous variable, logistic regression is the logical choice. If you have a theory that one or more independent variable *explain* one other variable, logistic regression is the logical choice. If however your theory does not involve a natural direction or prediction of one variable, and you are simply interested in associations among variables, then Poisson regression is the obvious choice.