Analyzing data using linear models

Stephanie van den Berg

Versie 0.1
(April 17, 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

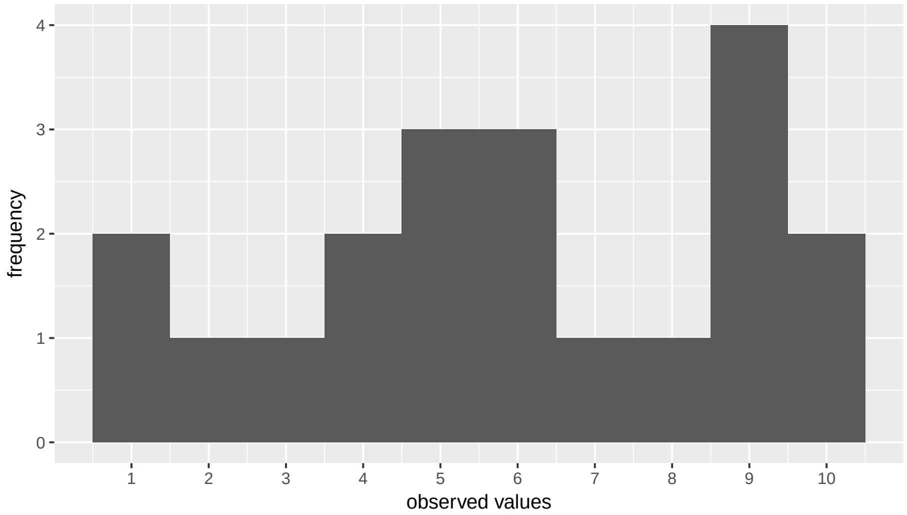


Figure 1.1: A frequency distribution

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

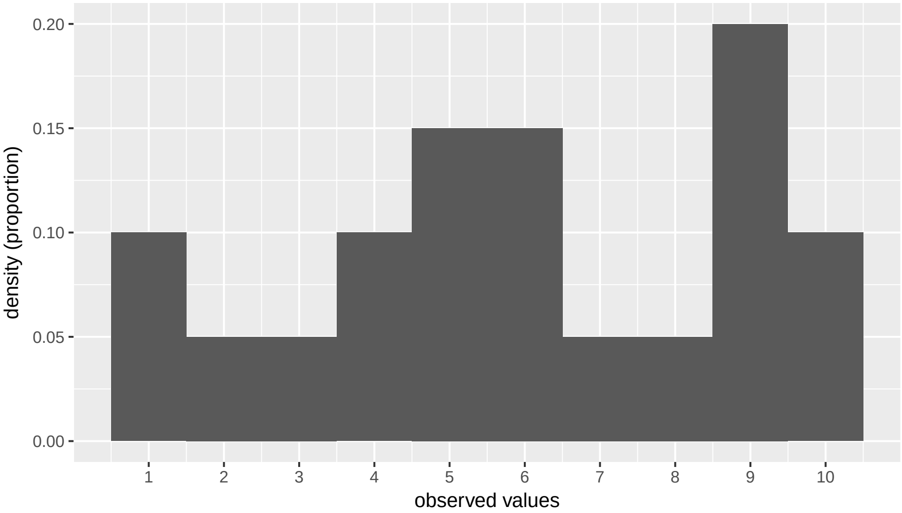


Figure 1.2: A histogram

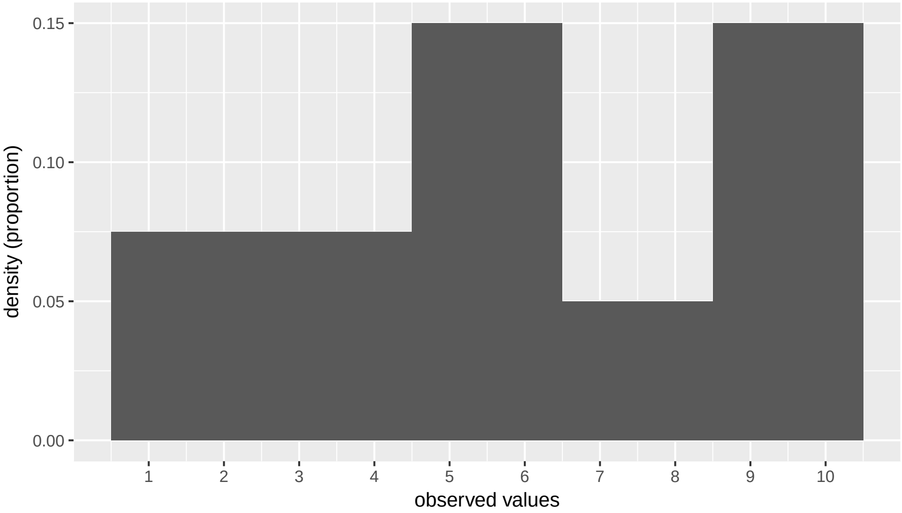


Figure 1.3: A histogram

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 FULYA

2.1 Linear relationships

2.2 Pearson correlation

2.3 Simple regression with a continuous predictor

2.4 Predicting the dependent variable

Chapter 3

Multivariate regression

3.1 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:

$$y = b_0 + b_1 x \tag{3.1}$$

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:

$$e = \hat{y} - y \tag{3.2}$$

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:

$$y = b_0 + b_1 x + e \tag{3.3}$$

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

$$\hat{y} = b_0 + b_1 x \tag{3.4}$$

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

$$e \sim N(0, \sigma^2) \tag{3.5}$$

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.

$$Var(y) = Var(\hat{y}) + Var(e) = Var(b_0 + b_1 x) + \sigma^2 \tag{3.6}$$

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:

$$Var_{explained} = \frac{Var(b_0 + b_1 x)}{Var(y)} = \frac{Var(b_0 + b_1 x)}{Var(b_0 + b_1 x) + \sigma^2} \tag{3.7}$$

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.

3.2 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:

$$weight = 107.7 + 0.71 \times volume + e \tag{3.8}$$

$$e \sim N(0, 15362) \tag{3.9}$$

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:

$$weight = 22.4 + 0.71 \times volume + 0.5 \times area + e \tag{3.10}$$

$$e \sim N(0, 6031) \tag{3.11}$$

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.*

3.3 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^2_e. 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 $(100 - 80)/100 = 0.20$.

$$R^2 = \sigma^2_{explained}/\sigma^2_y = (1 - \sigma^2_{unexplained})/\sigma^2_y = (1 - \sigma^2_e)/\sigma^2_y \tag{3.12}$$

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

$$\widehat{\sigma^2_y} = \frac{\Sigma(y - \bar{y})^2}{n - 1} \tag{3.13}$$

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

$$\widehat{\sigma^2_e} = \frac{\Sigma e^2}{n-1} \tag{3.14}$$

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

$$\begin{aligned}\widehat{R^2} &= \frac{\frac{\Sigma(y-\bar{y})^2}{n-1} - \frac{\Sigma e^2}{n-1}}{\frac{\Sigma(y-\bar{y})^2}{n-1}} \\ &= \frac{\Sigma(y-\bar{y})^2 - \Sigma e^2}{\Sigma(y-\bar{y})^2} = 1 - \frac{SSE}{SST} \end{aligned} \tag{3.15}$$

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.

3.4 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:

$$weight = b_0 + b_1 \times volume + e \tag{3.16}$$

$$weight = b_0 + b_1 \times area + e \tag{3.17}$$

$$weight = b_0 + b_1 \times volume + b_1 \times area + e \tag{3.18}$$

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 $round)cor(allbacks$area, 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 3.1 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 3.1: 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

$$age = 11.4 + 0.82 \times weight + e \tag{3.19}$$

$$e \sim N(0, 200) \tag{3.20}$$

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:

$$age = 20.6 + 0.11 \times heart + e \tag{3.21}$$
$$e \sim N(0, 307) \tag{3.22}$$

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:

$$age = 10.3 + 0.99 \times weight - 0.03 \times heart + e \tag{3.23}$$
$$e \sim N(0, 204) \tag{3.24}$$

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!

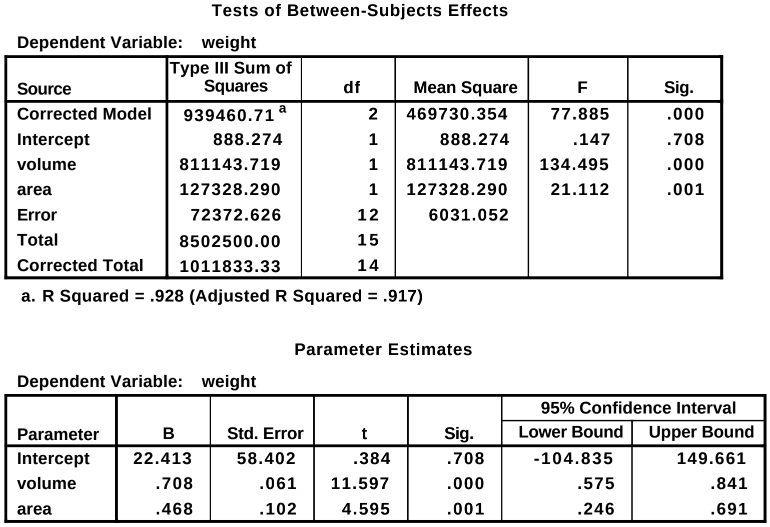
When two predictor variabels are perfectly correlated, either 1 or -1, estimation is no longer possible, the software stops and you get a warning. We call such a situation *multiple collinearity.* But also if the correlation is close to 1 or -1, you should be very careful interpeting the regression parameters. 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.

3.5 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 3.1: SPSS output of a linear model (multiple regression) for predicting the weight of books.

Figure 3.1 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:

$$weight = 22.413 + 0.708 \times volume + 0.468 \times area + e \tag{3.25}$$

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 3.1 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 $(1011833 - 72372.626)/1011833 = 0.928$.

3.6 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).

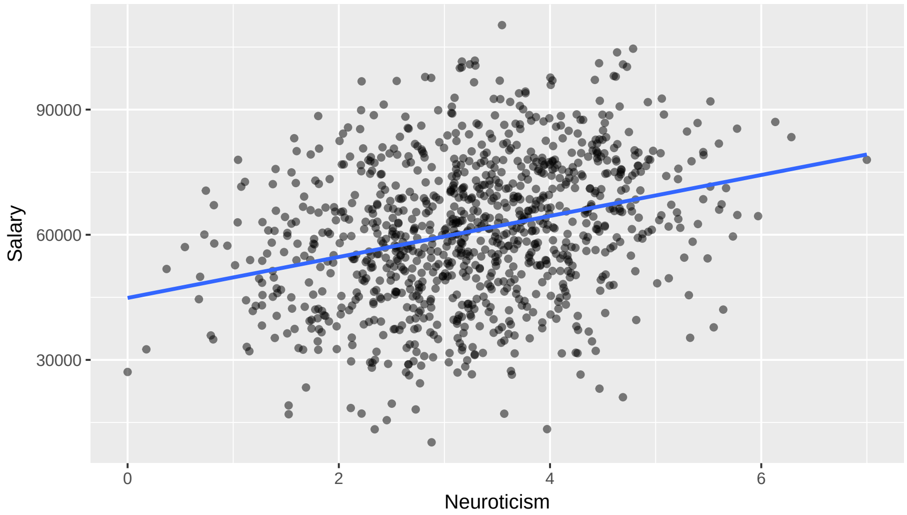
We carry out a simple regression, with salary as our dependent variable and Neuroticism as our independent variable. We then find the following regression equation:

$$salary = 44857 + 4912 \times Neuroticism + e \tag{3.26}$$

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:

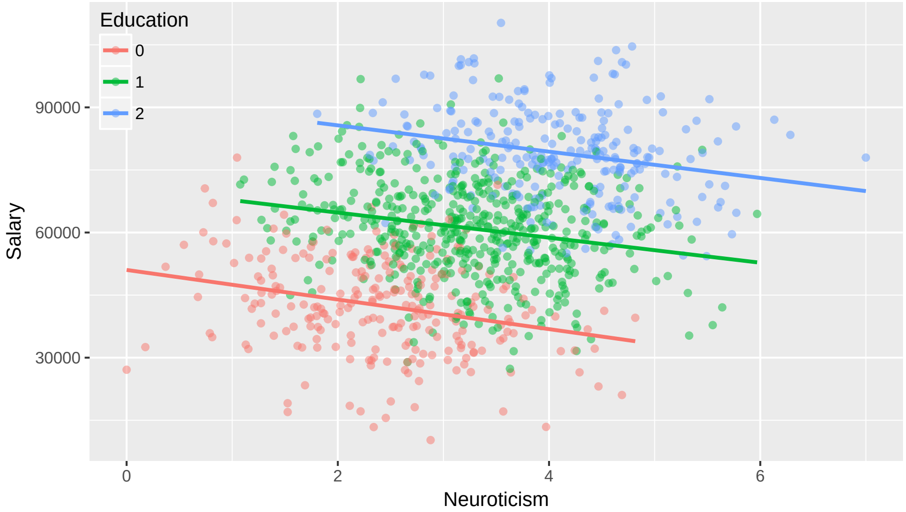
$$salary = 50249 - 3176 \times Neuroticism + 20979 \times Education + e \qquad (3.27)$$

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.

3.7 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 3.2: what would your linear equation look like?

$$\ldots\ldots = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + e \tag{3.28}$$

Table 3.2: 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 3.2 and 3.3, 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?

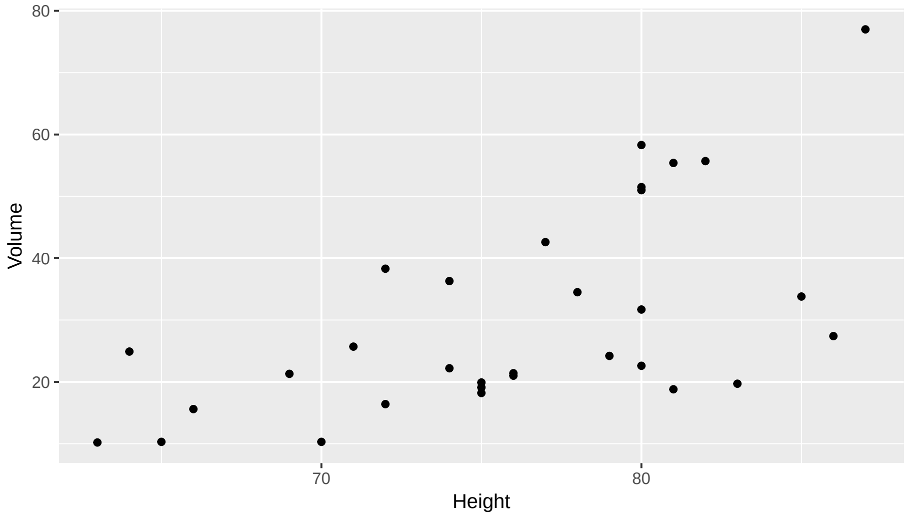


Figure 3.2: A scatterplot for the relationship between height and volume of a tree.

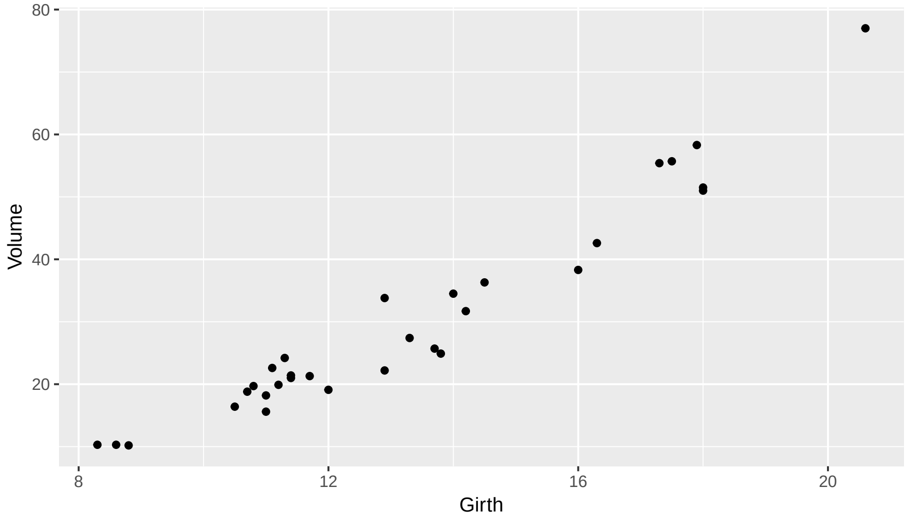


Figure 3.3: A scatterplot for the relationship between girth and volume of a tree.

Chapter 4

Inference

In the previous chapters on simple and multiple regression we have seen 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 or more other variables, the independent variables. Sometimes we are indeed interested in the relationship between 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 they 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 and determines from log files the temperature in the factory during production for each measured bottle. The researcher might find a small effect of temperature (t) on the average volume of beer in the produced bottles, say the linear equation might be $volume = 31.7225839 - 0.0879535 \times t + e$, but the question is whether this effect of temperature is present in *all* bottles.

In other words, we might have data on a sample of bottles, but we might really be interested to know whether there is an effect *had we been able to measure the volume in all bottles.*

4.1 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

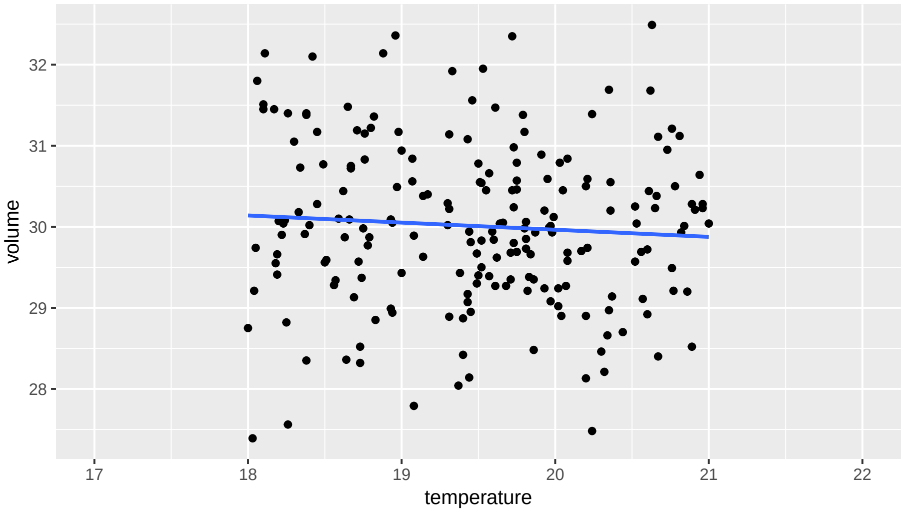


Figure 4.1: A scatterplot for the relationship between girth and volume of a tree.

scatter plot in Figure 4.2.

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 4.2.

However, in the data example above, data was only collected on 200 bottles. These bottles were randomly selected: 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 0.0012583 in the population, and we see a slope of -0.0879535 in the sample.

The discrependency here is simply the result of chance: had we selected another sample of 200 bottles, we probably would have found a different linear equation. The intercept and slope based on sample data, are the result of chance. The population intercept and slope (the true ones) are fixed, but unknown. If we want to know something about the population means, 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 intercept and slope are to the population intercept and slope?

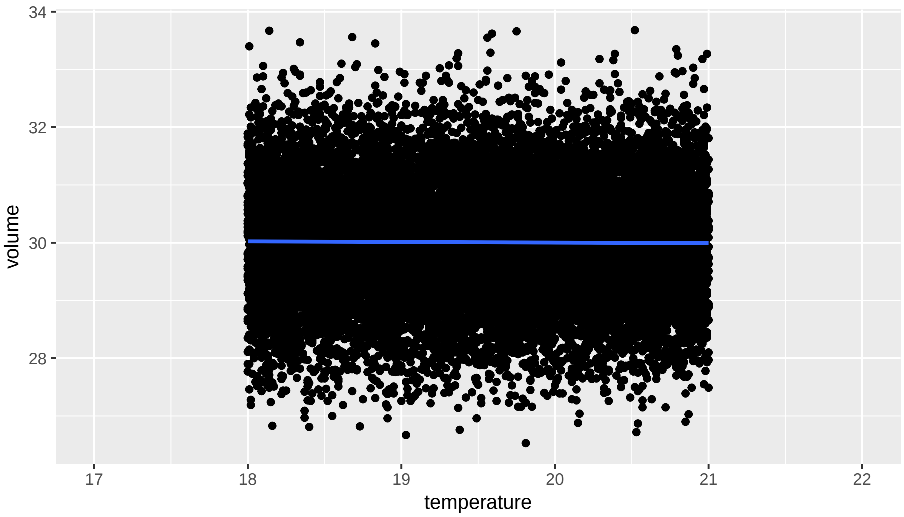


Figure 4.2: A scatterplot for the relationship between girth and volume of a tree.

4.2 Random sampling and confidence intervals

In order to know how close the intercept and slope in a sample are to their values in the population, let's do another thought experiment. Let's see what happens if we take another random sample of 200 bottlees. With random, we mean that every bottle has the same chance of being picked.

Now let's take another random sample. 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 their IQs and compute the average and the standard deviation. Now we find a mean of 97 and a standard deviation of 16.

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 100 sample intercepts and sample slopes we get the following picture:

We see a large variation in the intercepts that we find, and only a small variation in the slopes (all very close to 0).

Let's focus on the slope. Below we see the histogram if we carry out the random sampling 1000 times:

If we look at the distribution of the sample slope in Figure 4.4, we see that on average the sample slope is around 0, which is the population mean (the true intercept of all bottles). But there is variation around that mean of 0: the standard deviation of all 1000 sample means turns out to be 0.0840496.

The standard deviation of the sample mean is called the *standard error*. Had

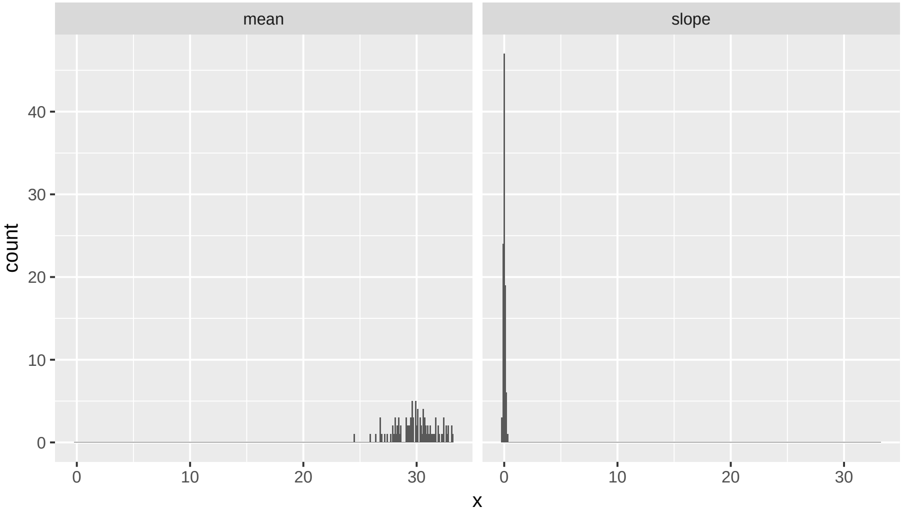


Figure 4.3: Distribution of the sample mean when population variance is 225 and sample size equals 200.

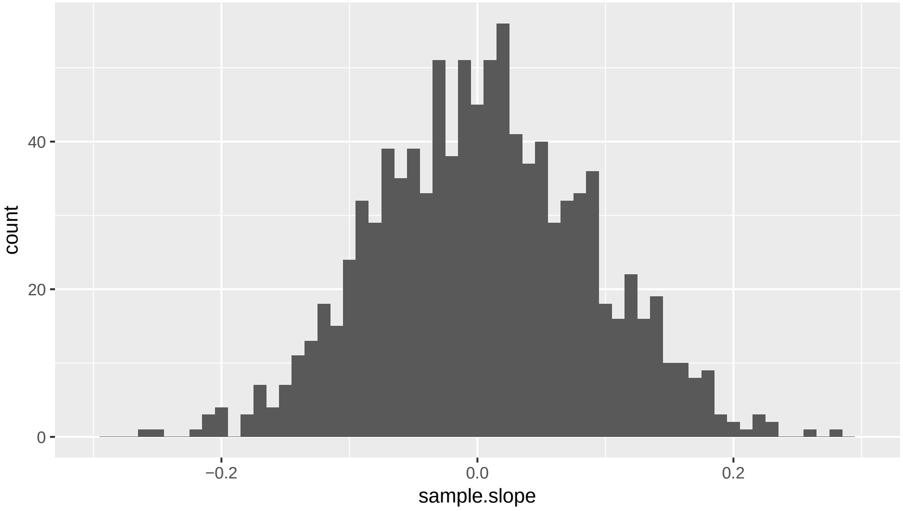


Figure 4.4: Distribution of the sample mean when population variance is 225 and sample size equals 200.

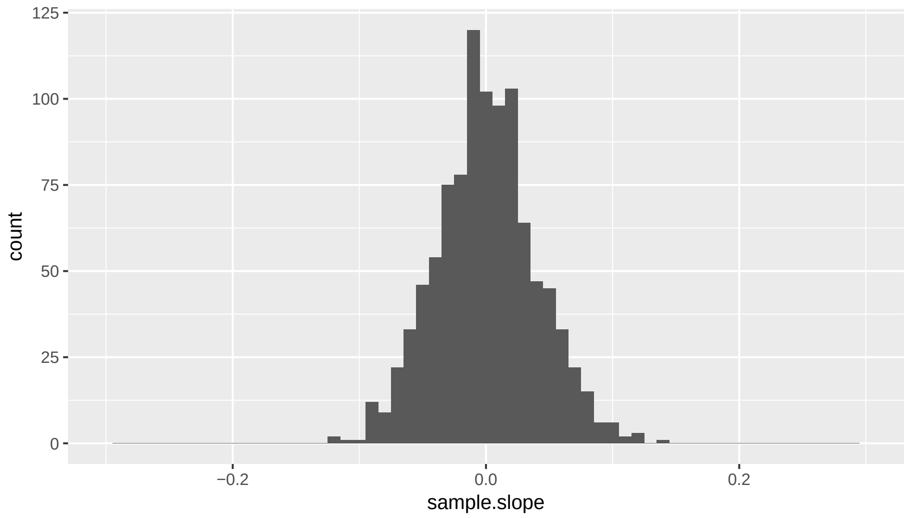


Figure 4.5: Distribution of the sample mean when population variance is 25 and sample size equals 200.

the true intercept been 110, the sample means would cluster around 110, but the standard deviation of the sample means, the standard error, would be the same.

The standard error for a sample slope represents the uncertainty about the true slope (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 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.

It turns out that the standard error for a sample slope depends on many things, but the most important factor is the sample size.

In the above bottle example, the standard deviation of all 80,000 volumes was 1.0004831, where most of the volumes (roughly 95%) lie between 28 and 32 cl. The variance is the square of the standard deviation so the variance is 1.0009664. Now imagine that we have another population, say bottles from a different brand, where we see a much smaller variation in volumes: suppose the average volume is also 30, but the standard deviation is 0.5, so that roughly 95% of the scores lie between 29 and 31. If we then take 100 samples from this distribution of bottles from this other brand, we get the distribution in Figure **??**.

Imagine that you draw only 2 bottles from a population of bottles. Then there is quite some probabilty that by sheer luck in one sample you find one bottle with a low temperature and a small volume, and another bottle with a

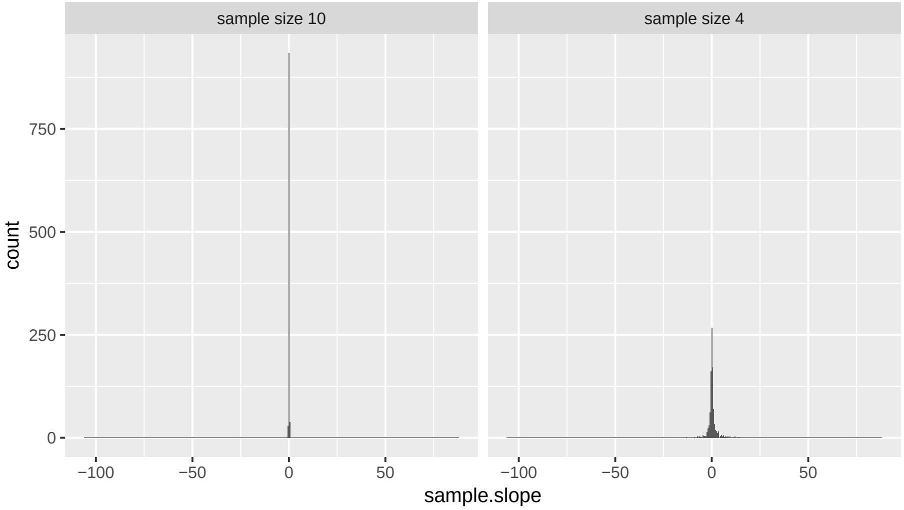


Figure 4.6: Distribution of the sample mean when population variance is 25 and sample size equals 200.

high temperature and a large volume, so that you find a sample slope that is quite large and positive. There is also an equally high probability that you get a 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 bottles, the sample slope will be 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, ergo, the standard error will be large. Now imagine that your sample size is 20. Then the probability that *all* 20 values will be around 76 or around 122 will be extremely low: there is a much bigger chance that there will be both low and high values in your sample of 20, so that the average of the 20 values will be closer to the population mean of 100.

In Figure **??** we see the distributions of the sample slope where the sample size is either 2 (left panel) or 20 (right panel). In sum: the larger the sample size, the smaller the standard error, the smaller the uncertainty about the population mean.

```
## Warning:  Removed 5 rows containing non-finite values (stat_bin).
```

So if we have a small standard error, we can be relatively certain that our sample slope is close to the population slope. Above we've done a thought experiment where we knew everything about the population intercept and slope, and we drew 1000 samples from this population. In reality, we don't know anything about the population: we only have the sample data to go on. So suppose we draw a sample of 200 from an unknown population, say IQ scores from children in Paris, and we find an average IQ of 112, we have to look at

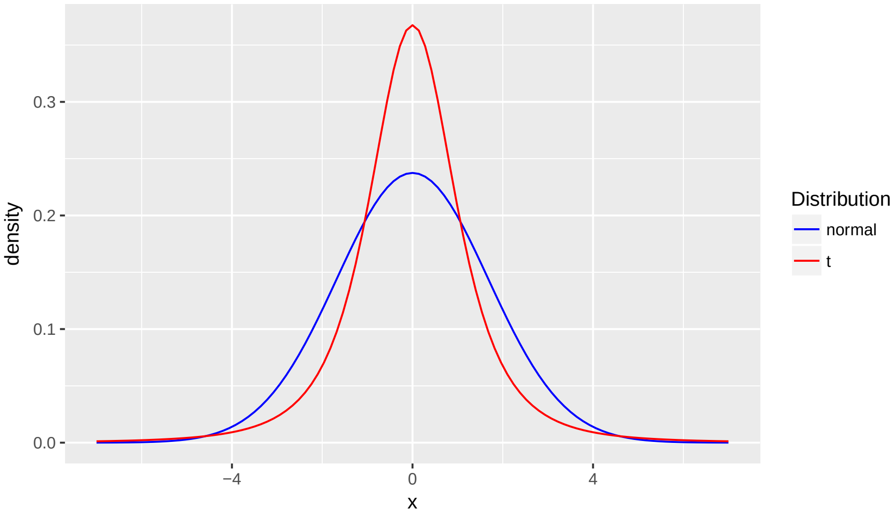


Figure 4.7: Distribution of the sample mean when population variance is 25 and sample size equals 200.

the standard error to know how close the real average in the population is to the sample mean of 112. 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).

4.2.1 *t*-distribution

Above we saw that if there is a large collection of data with a particular slope, and if you then take random samples out of this collection, each time you get 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 a slope in the sample is close to the slope in the population.

When we look at the distribution of the sample slope, for instance in Figure **??**, we notice 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 shows the difference between a *t*-distribution and a normal distribution. In this figure, the means are equal (0) and the standard deviations are equal (1.68), but the shapes are clearly different. Compared to the normal distribution, the *t*-distribution has more observed values close to the mean (the distribution is more peaked) and there are relatively more observations at the extreme ends of the distribution (heavy tails).

Actually, the shape of the distribution of sample slope depends on the size

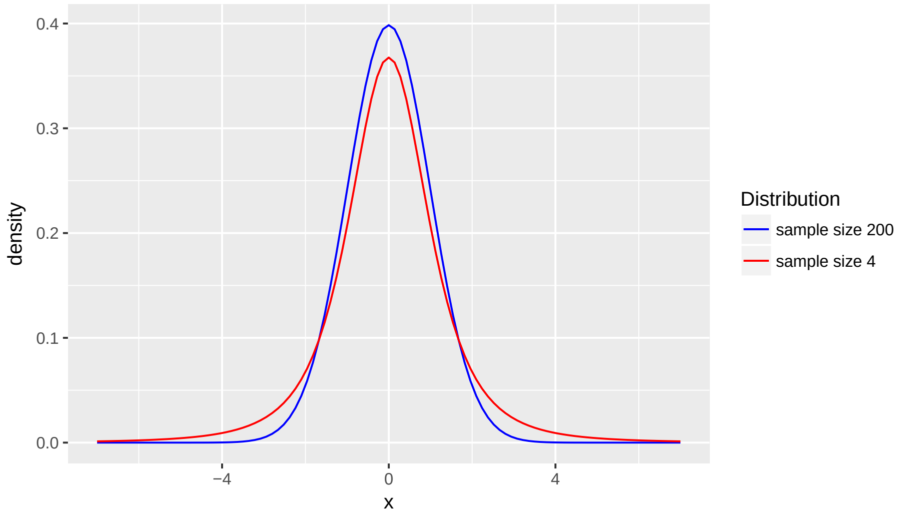


Figure 4.8: Distribution of the sample mean when population variance is 25 and sample size equals 200.

of the samples. In Figure **??** we see what the distribution would look like if all samples would be of size 4 (the red line) and what the distribution would like like if the samples would be of size 200 (the blue line). Remember: we're talking here only about the *shape* of the distribution. If sample size is large, like for instance 200 (the blue line), the shape looks extremely close to the normal distribution. In summary, when we draw many samples from a population, the shape of the distribution of sample slopes is that of a *t*-distribution. The larger the sample size, the more the shape of the distribution looks like a normal distribution.

4.2.2 Confidence intervals

From the normal distribution we know that about 95% of the observations lie between the mean $\pm 2 timesthestandarddeviationandmeanplus2timesthestandarddeviation.Ifsamplesizeislarge$

In statistics, one often uses a *confidence interval* to indicate a range of reasonable values for the population mean. Here we found a sample mean of 112. Now imagine that 112 were also the population mean. Then if we would draw many random samples of size 100, we know from the computed standard error of 1.2 that roughly 95% of the sample means would lie between $112 - 2 \times 1.2 = 109.6$ and $112 + 2 \times 1.2 = 114.4$.

Now suppose that the true population mean were not 112 but 114.4. In that case, if we draw many samples of size 100, we could reasonably find a value of 112, since 95% of the sample mean would then lie between $114.4 - 2 \times 1.2 = 112$ and $114.4 + 2 \times 1.2 = 116.8$. So even if the true population mean were 112.4, it's very possible that we could find a sample mean. We cannot neglect the possiblity that the true mean is 114.4. Similarly, we cannot neglect the possiblity that the true mean is 109.6, because if the true mean were 109.6, 95% of the sample means of size 100 would lie between $109.6 - 2 \times 1.2 = 107.2$ and $109.6 + 2 \times 1.2 = 112$. So our range of reasonable values for the population mean would be somewhere between 107.2 and 114.4. This range is referred to as the *95% confidence interval*. The 95% confidence interval can be computed by subtracting and adding twice the standard error of the mean to the sample mean.

4.3 Inference: from sample to population