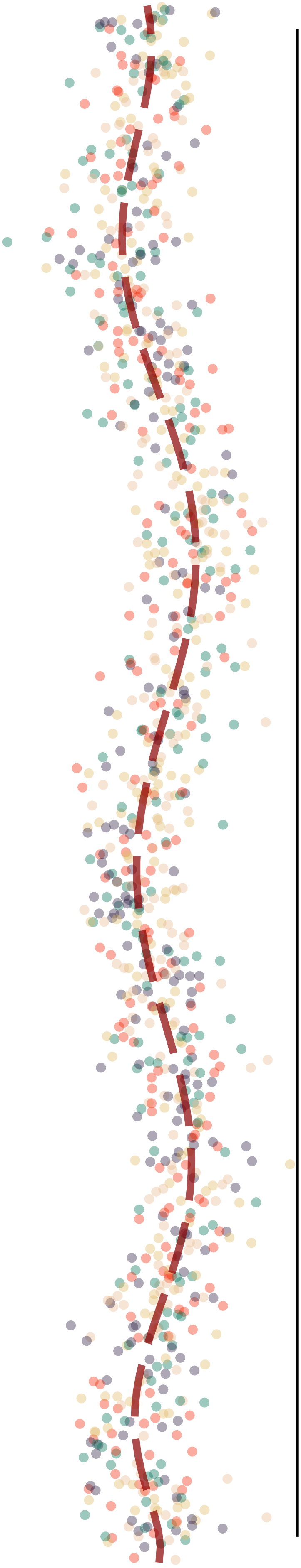


EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Palacký University Olomouc
Faculty of Arts

Linear statistical models in psychology

Daniel Dostál

Olomouc
2021

This book is published as a part of the project Establishment of a Doctoral Program for Innovative Research in Work and Organizational Psychology (CZ.02.2.69/0.0/0.0/16_018/0002313).

Reviewers:
Mgr. Vladimír Matlach, Ph.D.
Mgr. Ondřej Vencálek, Ph.D

1st edition

DOI: 10.5507/ff.21.24460345
ISBN 978-80-244-6034-5 (online: iPDF)

Contents

Introduction

Introductory statistics courses equip students with a variety of tools for describing data and null hypothesis significance testing. Measures of location, variability and other characteristics, correlation coefficients, as well as parametric and non-parametric bivariate tests serve as useful instruments in countless situations. However, when designing more advanced experiments, large-scale questionnaire surveys, and other sophisticated studies, the shortcomings of the above procedures are often felt. They describe the behavior of variables separately, and they can only capture statistical relationships at the level of pairs of variables.

In this textbook, we present a more advanced view of quantitative data analysis that allows us to describe diverse relationships within groups of variables. The purpose of this textbook is to introduce students to the world of statistical modeling, particularly to linear regression models.

Regression models of various kinds are the central research tool in psychology and almost all other empirical sciences. Without knowledge of linear regression, it is difficult to publish research results and, especially for students considering a career in academia, a good knowledge of the topic is essential.

This textbook expands on the topics covered in the basic statistics courses. Thus, for a good understanding, the reader should be familiar with concepts such as probability distribution, mean, variance, statistical estimation, hypothesis tests, and p-values. On the other hand, the aim of the author of this textbook is to make a rather complicated subject accessible to those students who have heard the concepts mentioned but whose deep knowledge has never been acquired or has been lost long ago. As a result of this effort, the author often had to choose between mathematical precision and clarity. If the topic of statistical modeling caught your interest, compare the knowledge gained with other more advanced texts that correct many of the simplifications and inaccuracies in this textbook.

In the following chapters, various procedures will be demonstrated on several data sets. Most of them can be downloaded in an MS Excel file at

dostal.vyzkum-psychologie.cz/soubory/data_linear_models.xlsx

I would like to thank my teacher and friend Ondřej Vencálek for reading the text and, where my statements deviated from mathematical theory in a particularly ignoble way, for stepping in and making me rewrite the passages in question.

author

1 Statistical model

All models are wrong,
but some are useful.

The famous statement by George Box, the eminent British statistician and incidentally the son-in-law of Ronald Fisher, illustrates well the essence of what we mean by the word "model". What are models for if none of them are correct? And what do we actually mean by the word model?

Reality is infinitely complex. It is so complex that we can never really explore the deepest laws of its functioning. The truth of what laws govern nature, including human behavior and mind, will forever remain a mystery for humankind.

If we can never understand reality in its entirety, how is it possible that many things that come from the workshop of man simply *work*? How is it that we can fly on holiday in an airplane when no-one has managed to work out exactly what rules does airflow in turbines of airplane engines follow? How is it that a skilled therapist can rid us of our phobia of elevators when they have no idea what exactly is going on in our brain and what was the phobia caused by?

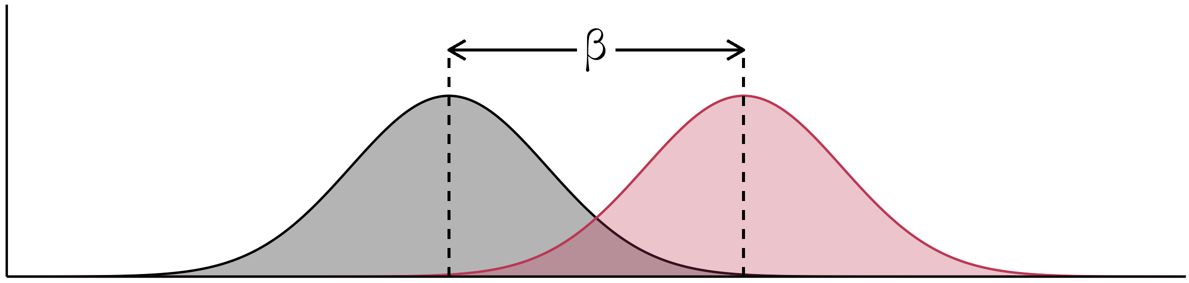
We do not need to know the whole truth with all the details in order to influence reality and to put scientific knowledge into practice. What we need is a clever simplification of reality that leaves out the less important aspects while remaining accurate enough to resemble how reality works. Strictly speaking, this simplification is *wrong* because it is not equal with reality. On the other hand, this simplification can be extremely *useful*. From now on, we will refer to this simplification of reality using the word **model**.

In the primary statistics courses, we learned a trick to simplify reality. Although we assume that the world follows cause and effect rules and is strictly deterministic, it often pays to ignore many factors and declare that chance plays a role. The phenomena we label as random can be described using random variables. If we follow this approach when constructing our model, we refer to it as the **statistical model**.

Statistical models offer a wide range of advantages. In particular, we can compare them with the world around us easily. We can make observations (measurements) of the studied phenomenon and compare what we see with what we should see according to our model. We can therefore easily assess to what extent our model fits or contradicts reality.

But the story does not end here. We can use statistical models to find the answers to our questions. We can incorporate some (free) **parameters** into our model. Since the statistical model is described by mathematical equations, this parameter is a number. However, it is not a specific number, but an unknown value with a magnitude that can only be estimated by our observation.

Figure 1: Simple statistical model



Let us illustrate a particular statistical model that has free parameters by a following example. We want to investigate whether a particular memory training has an effect on performance in a memory test. For this purpose, we approach a group of volunteers, randomly dividing them into two groups – an experimental group and a control group. We then expose the experimental group to memory training, while the control group is just talked to instead of being trained. Finally, both groups take the memory test.

The model could look like the following: Let us describe the experimental group members' performance in a memory test by a random variable A. Let us suppose this random variable has normal distribution. Let us describe the performance of the control group by a random variable B. This random variable has normal distribution as well, and furthermore, let us suppose that both random variables have the same variance. The expected values of random variables A and B differ by some value, let us call it β. This difference β is a free parameter of the model. If $\beta > 0$, then people exposed to memory training in fact score higher than people not exposed to it. If β is zero or less than zero, then this is not the case. Our model is shown in Figure 1.

Note that this is indeed just a model, which of course differs from reality. The deterministic conception of the world claims that randomness does not exist, but we are talking about random variables. We are saying that A and B both have normal distributions, which is probably not quite true even if we allow for the existence of some random variables. We claim that they both have exactly the same variance, which is almost certainly not true. Obviously, this model is wrong, simply because it is just a model. On the other hand, we can expect that in a given situation this model is very useful.

By observing the realizations of the random variables A and B, we can start estimating the size of the parameter β. Moreover, since this is a statistical model, we can test the validity of our hypotheses concerning the size of the parameter β.

We have been through this model many times before; we just probably did not realize that it was a statistical model. We created it each time we performed a t-test for two independent samples.

1.1 Linear statistical models

There are countless statistical models, and we could invent more on the fly to describe any situation. From this diverse array, we will choose only a relatively narrow group of models called linear models. Moreover, we will again choose only a small subset of linear models that are similar in their structure.

We will only consider models that describe the behavior of a single dependent random variable Y which we assume to be influenced by one or more factors X_1, X_2, $\ldots$, X_k. The word *linear* means that we assume that the value of the random variable Y corresponds to a **linear combination** of the factors X_1, X_2, $\ldots$, X_k. When a mathematician uses the term linear combination, they are referring to a simple summation, but they assign different weights to the individual addends. We will refer to these weights β_1, β_2, $\ldots$, β_k. Each of the models we will be discussing in the following pages will have the form[1]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

In this textbook, we will always note the dependent variable by Y and refer to it as the dependent or possibly the explained or the predicted variable. We will refer to the independent variables X_1 to X_k as **factors**, **regressors**, or **predictors**.

By requiring our model to always follow this structure, we have severely limited our options. On the other hand, most of the research problems we encounter in psychology students' theses and dissertations (and, after all, in most research articles) fit into this structure easily. It is common to examine some quantity Y and ask how it is affected by some factors. For example, if we compose a thesis titled *Sleep Disorders in Preschool Children* and decide to use such a linear model, we can expect the dependent variable to be sleep quality (operationalized, for example, as the score of some questionnaire that measures it). The X variables would then include, for example, the number of minutes the child spends in front of the TV or playing computer games in the evening, their gender, age, as well as information on whether the child was assigned to a group that performs some kind of relaxation activity before falling asleep.

[1] We will see this relationship in many various modifications on the pages of this textbook countless times. More advanced readers should therefore be advised that this notation is not entirely correct, and it is used here deliberately in order to make the text more readable. In fact, the relationship only holds "on average", so the left-hand side of the equation should be the expected (mean) value of the random variable Y, i.e., $\mathbf{E}(Y)$, or we should use a symbol indicating only an approximate equality instead of the equal sign.

The fact that the model is linear may sometimes not be apparent at first sight. For example, we would probably not say at first glance that model

$$Y = \beta_0 + \beta_1 \sqrt{Z_1 Z_2} + \beta_2 \left(\frac{4}{Z_3} + 3 \right) + \beta_3 \sin(2\pi Z_4)$$

belongs to the family of linear models. However, in fact, we can imagine that $X_1 = \sqrt{Z_1 Z_2}$, $X_2 = \frac{4}{Z_3} + 3$ and $X_3 = \sin(2\pi Z_4)$, and again we get the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

In contrast, for example, the model

$$Y = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}$$

cannot be considered linear, since even with all effort we cannot convert it into a linear combination of some regressors.

2 Parameters of the model and their estimation

As mentioned above, statistical models contain some parameters with unknown values. Using the example of a model called **simple regression**, let us see what role these parameters play and how we estimate their sizes. The simple regression is a model that has only a single regressor X along with a single dependent variable Y. We could describe it with the following equation:

$$Y = \beta_0 + \beta_1 \cdot X$$

To get a better idea, let us proceed with a practical example. For example, let us try to describe how many points a student scores on a written exam in cognitive psychology (variable Y) depending on how many hours they have spent studying for it (X). Therefore:

$$\text{(number of points)} = \beta_0 + \beta_1 \cdot \text{(number of hours spent studying)}$$

To determine the values of β_0 and β_1 weights, we need to ask at least two students how long they studied for and how they scored on the test. We asked Agatha, who scored 42 points and had studied for 16 hours, and Otto, who scored 30 points and had studied for 10 hours.

Agatha: $42 = \beta_0 + \beta_1 \cdot 16$

Otto: $30 = \beta_0 + \beta_1 \cdot 10$

The data we obtained are consistent with what common sense tells us – Otto, who spent six hours less time revising for the test than Agatha, actually scored lower. The weight β_1 is obviously a positive number indicating how many points each hour of studying will on average give us. The problem can be solved as a system of equations with two unknowns (β_0 and β_1). We can easily find that the value of β_1 is 2 and β_0 is 10. Therefore,

Agatha: $42 = 10 + 2 \cdot 16$

Otto: $30 = 10 + 2 \cdot 10$

We can conclude our mini research by claiming that each hour of studying for a cognitive psychology test leads to a gain of two points. We also intuitively understand the role of the coefficient β_0. The model predicts that if we do not even open the textbook ($X = 0$), we will score 10 points on average ($\beta_0 = 10$).

If we were conducting real research, we would probably not rely on data from only two respondents. Let us include Ursula in our sample. She did not put much effort into preparation for the test, she only studied for two hours, but she probably managed to copy some of the results from Agatha, or she is extremely talented, it is hard to say. She managed to score 24 points.

Ursula: $24 = \beta_0 + \beta_1 \cdot 2$

Unfortunately, expanding the research sample reveals an unpleasant fact – the model fails to bring any results.

$$\text{Ursula:} \qquad 24 \neq 10 + 2 \cdot 2$$

What more, there are no values of β_0 and β_1 that are suitable for all three participants in our research. The solution is to extend our original model with one more term, which we call **residual** and note it ϵ (epsilon)[2]. We can think of it as the size of the error that the model generates for a given individual. Thus, the equation describing our model with the residual would take the form:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

If we insist that $\beta_0 = 10$ and $\beta_1 = 2$, then the residuals of each observation would take the following values:

$$\begin{aligned} \text{Agatha:} &\qquad 42 = 10 + 2 \cdot 16 + 0 \\ \text{Otto:} &\qquad 30 = 10 + 2 \cdot 10 + 0 \\ \text{Ursula:} &\qquad 24 = 10 + 2 \cdot 2 + 10 \end{aligned}$$

This is obviously not a very fair solution. Our model perfectly accommodates the results of Agata and Otto, where the error (residual) is zero, but in the case of Ursula it is wrong by ten points. We could, of course, choose a different pair of numbers β_0 and β_1, and get a different set of residuals. Some solutions will be better and other worse. We consider the optimal solution to be the one that produces residuals as close to zero as possible.

In order to compare the solutions, we need to develop an indicator (so-called minimization criterion) that converts the obtained set of residuals into a single number that reflects the quality of the solution. It turns out that the best minimization criterion is not a simple sum of the residuals (or their absolute values), but a sum of their squares. We will call this criterion **residual sum of squares** (RSS) and its formula takes the following form:

$$\text{RSS} = \sum_{i=1}^{n} \epsilon_i^2$$

The letter n indicates the sample size. In our case, n equals three and RSS equals 100 (since $0^2 + 0^2 + 10^2 = 100$).

[2] Note a minor terminological inaccuracy – in this text, we will not distinguish between a random component and a residue, even though they are two related, however, not identical, concepts.

2.1 Least squares method

Searching for parameter values that minimize RSS by trial and error would be tedious and we would probably never achieve an accurate result. However, the calculation can be done using the **least squares method**. The least squares method is famous for its elegance and versatility and was used by Carl Friedrich Gauss in his calculations as early as 1795. To calculate it, we need to know basic matrix and vector calculus. The parameter estimation procedure using the least squares method can be described by the following brief formula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of parameter estimates, $\mathbf{X}$ is a design matrix containing a column of ones and the values of all regressors in the other columns, $\mathbf{Y}$ is a vector of values of the dependent variable. The operators $'$ and $^{-1}$ stand for transpose and matrix inverse, respectively. In our example, the individual elements of the equation would have the following values:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} 42 \\ 30 \\ 24 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 16 \\ 1 & 10 \\ 1 & 2 \end{pmatrix}$$

Note that instead of β_0 and β_1 we use the symbols $\hat{\beta}_0$ and $\hat{\beta}_1$. **We will use the hat symbol whenever we want to express that we are referring to an estimate, not an exact value.** If we repeated our calculation on triples of students other than Otto, Agatha, and Ursula, we would arrive at different estimates of $\hat{\boldsymbol{\beta}}$, which would be distributed around the actual unknown values of $\boldsymbol{\beta}$. Estimates of $\hat{\boldsymbol{\beta}}$ are random variables (statistics), and what more, they are the minimum-variance unbiased estimates of the parameters $\boldsymbol{\beta}$. Moreover, they become increasingly more accurate the more observations we have.

In our case, after substituting into the equation utilizing the least squares method, we obtain the values $\hat{\beta}_0 = 20.27$ and $\hat{\beta}_1 = 1.26$.[3] After the substitution, we find the following values of the residuals:

$$\begin{aligned} \text{Agatha:} \quad & 42 = 20.27 + 1.26 \cdot 16 + 1.62 \\ \text{Otto:} \quad & 30 = 20.27 + 1.26 \cdot 10 - 2.84 \\ \text{Ursula:} \quad & 24 = 20.27 + 1.26 \cdot 2 + 1.22 \end{aligned}$$

[3] Let us add that in this text we use the same symbol $\hat{\beta}$ to denote the random variable (estimator) and the value of its realization (estimate). Readers can easily distinguish whether it is a number or a statistic by its context.

The residual sum of squares is now 12.16 ($1.62^2 + (-2.84)^2 + 1.22^2$) and as expected it has dropped significantly from the original value of 100. As evident in the above-mentioned properties of the least squares method, there is no pair of values $\hat{\beta}_0$ and $\hat{\beta}_1$ that finds a lower value of RSS than 12.16 for our data.

2.2 Unstandardized regression coefficients

The parameters β_0, β_1 to β_k are referred to as *standardized regression coefficients*, or less precisely as regression weights. To understand the results of the linear model, it is crucial to understand their meaning thoroughly.

Each of the parameters β_1 to β_k belongs to one regressor (X_1 to X_k). The value of the unstandardized regression coefficient shows us **how much the size of the dependent variable *Y* changes on average when the value of the respective regressor increases by one** (while other regressors remain unchanged). In our example, we arrived at the result that $\hat{\beta}_1 = 1.26$ which can be translated into a more understandable statement *for each hour of preparation, students receive on average approximately one and a quarter points more*. If we found a weight of a regressor is equal to zero, it would mean that regardless of the change in the value of the regressor, the average value of the dependent variable does not change (i.e., the regressor has no effect).

The parameter β_0 does not belong to any regressor (or rather it belongs to that mysterious column of ones in the design matrix). We refer to this parameter as the **absolute term**, **constant** or **intercept**. **The intercept tells us what value the dependent variable *Y* will take on average if all regressors are equal to zero.** Thus, in our example where $\hat{\beta}_0 = 20.27$, this would mean that a completely untaught student who has spent zero hours revising for the test would score an average of just over 20 on the test.

Understanding the regression coefficients allows us to answer questions like *how many points are we likely to get if we only have the night before the exam to prepare for it, say 6 hours from 10pm to 4am*. The answer would be 28 points. More precisely, $20.27 + 1.26 \cdot 6 = 27.81$ points. This sounds fairly optimistic if the threshold for passing the exam is, say, 25 points and we have confidence in our model which, as we will see below, is not particularly appropriate in this case. So, we better start studying earlier rather than pulling an all-nighter the night before the test.

2.3 Standardized regression coefficients

In some circumstances, the unstandardized regression coefficients alone may not provide easy-to-understand information to the reader. Imagine, for example, that you are conducting marketing research to map how satisfied users are with the quality of their internet connection. You collect data including a satisfaction rating from each user – a number between one (extremely dissatisfied) and ten (extremely satisfied), also how much they pay for the service monthly, how many connectivity outages longer than 5 minutes occur each month, and finally, how fast the connection is. Moreover, the connection speed measured in kilobits per second (Kbps) was converted to a decadic logarithm scale (i.e., 1 Kbps was coded as 0, 10 Kbps as 1, 100 Kbps as 2, etc.). The resulting unstandardized regression coefficients might look something like this:

Regressor	$\hat{\beta}$
Intercept	3.950
Price (€ per month)	-0.050
Outages (occurrences per month)	-0.140
Speed ($\log_{10}$ Kpbs)	0.862

Probably even if we tried our best, we would not be able to use the table to determine which factor has an considerable effect on customer satisfaction and which does not. We can draw conclusions such as *"for every euro paid per month, satisfaction decreases by 0.05 points"* or *"with every outage, customer satisfaction decreases by 0.14 points"*, but we can hardly say whether these effects are major or minor. Not to mention that interpretating the 0.862 weight of the *connection speed* regressor will be quite difficult.

This is exactly the situation in which we use standardized regression coefficients. Standardized regression coefficients are obtained by exactly the same procedure as non-standardized coefficients, except for one difference: before performing the calculation, we convert the dependent variable as well as each regressor into z-score form. The z-score transformation means that we subtract the arithmetic mean from each value of a given column of the data matrix and then divide each value by the sample standard deviation of that column. Thus, if the average customer satisfaction rate is 5.5 and the sample standard deviation of satisfaction rate is 2.5, then if someone responded that their satisfaction rate is 4, the value of their satisfaction z-score would be -0.6 (since $(4-5.5)/2.5 = -0.6$).

The advantage of the z-score transformation is that all variables are projected onto the same scale. So, it does not matter if we measured the price in euros or cents, for example, we always get the same result. The unit of all regressors becomes one standard deviation. **The standardized regression coefficient expressed by how many standard deviations will the value of the dependent variable Y increase on average if the value of the respective regressor increases by one standard deviation.** In this text, the standardized regression coefficients will be referred to by the β^* symbol.

We can also calculate the standardized coefficients from the unstandardized ones by multiplying the β weight by the standard deviation of the regressor and dividing by the standard deviation of the dependent variable. Thus

$$\hat{\beta}^* = \frac{\hat{\sigma}_X}{\hat{\sigma}_Y}\hat{\beta}$$

The following table reports the results of our model including standardized regression coefficients:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
Intercept	3.950	
Price (€ per month)	−0.050	−0.172
Outages (occurrences per month)	−0.140	−0.463
Speed ($\log_{10}$ Kpbs)	0.862	0.351

Standardized regression coefficients can be used to compare the effect size of individual regressors, even across different models. Here we can say that the number of internet outages has the strongest effect on customer satisfaction and the price of the service plays the least significant role. We could also state, for example, that *if the number of outages increases by one standard deviation, customer satisfaction rate drops on average by nearly half a standard deviation.*

Note that the standardized regression coefficient for the absolute term is missing from the table and we will probably see the omitted field in the output of any type of statistical software. This is a consequence of the fact that β_0^* always equals 0. That is to say, if we have observed the average values of all regressors (i.e., the z-scores of all regressors equal 0), then we expect the value of the z-score transformed dependent variable to be average as well (i.e., 0).

Since standardized regression coefficients almost always come out in the range $[-1; 1]$, they can be interpreted in a similar way to the Pearson correlation coefficient. Moreover, if a given regressor is perfectly uncorrelated with all other regressors within the model, then the standardized regression coefficient exactly equals the value of the Pearson correlation coefficient. If we were talking about a simple regression, then always $\hat{\beta}_1^* = r_{X,Y}$, since the single regressor in the model naturally cannot correlate with any other regressor.

Standardized regression coefficients are quite popular in psychology, however, there are some fields where this kind of indicator is practically never used. For this reason, regression coefficients tend to be labeled differently in different fields – while texts by statisticians denote non-standardized and standardized coefficients (as does this textbook) β and β^*, in psychology texts we most often see b and β. As clearly illustrated here:

Regression coefficient	In statistics	In psychology
Unstandardized	β	b
Standardized	β^*	β

It should be noted that standardized regression coefficients are not useful in every scenario. If we decide which set of coefficients to present, then a simple rule of thumb applies: present those coefficients that, for a given model, will help readers understand the results. If in a given case both types of coefficients fulfill this role, then it is probably the best decision to present both sets of coefficients.

2.4 Graphical representation of simple regression

The advantage of simple regression is that we can represent it graphically. By plotting the measured values on a scatter plot, where the x and y axes correspond to the random variables X and Y, we can display the fitted model as a *regression line*. See the Figure 2. This is a graphical representation of the test results of Agatha, Otto and Ursula and five other classmates of theirs. The estimate of the parameter β_0 equals 22.00 and the estimate of the parameter β_1 equals 1.46.

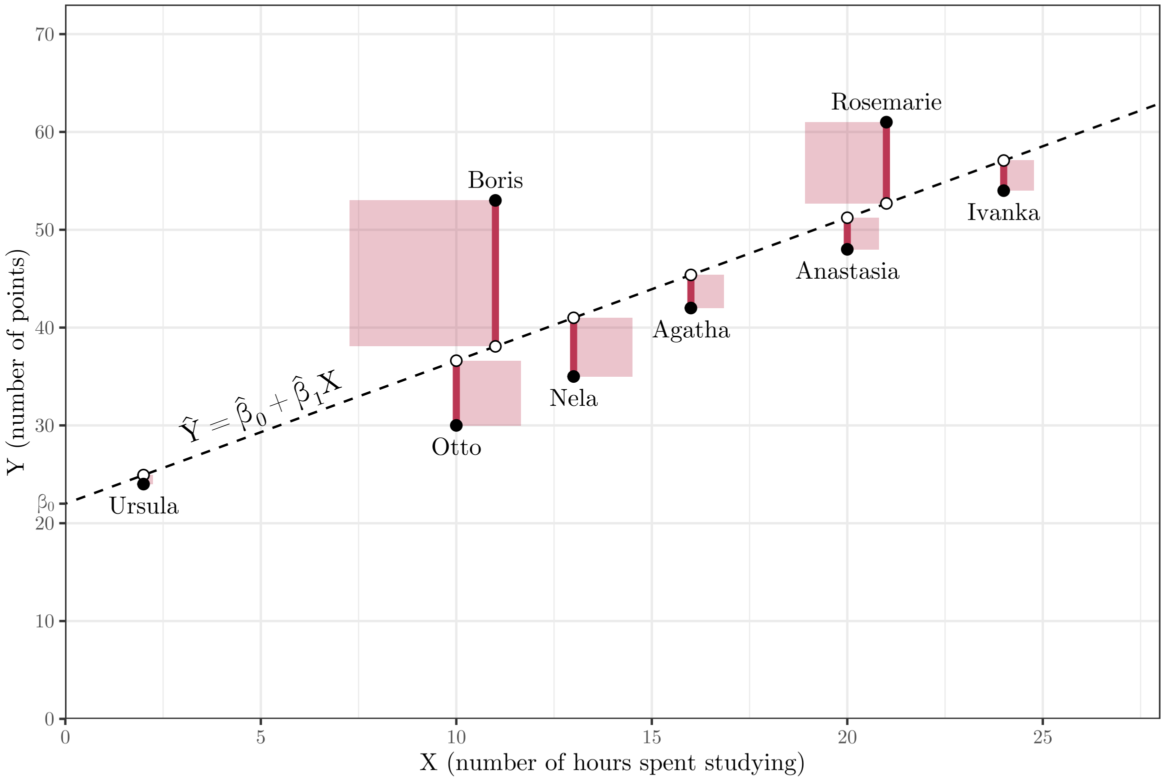
The filled points show the observed values of the dependent and independent random variable. Blank points indicate the **expected values** of the variable Y using our model (labeled $\hat{Y}$). We also refer to them as **fitted values** or **predicted values** of the observed variable. The expected values lie on the regression line defined by the equation of our model, i.e., $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. The differences between the observed and expected values of the dependent variable are the residuals we are already familiar with (illustrated by red lines in the graph). The red areas then represent the squared residuals. Thus, the sum of the red areas is RSS, which we seek to minimize with the least squares method.

Also, note the location of β_0 – it marks the point where the regression line intersects the y axis. The parameter β_1 indicates the slope of the regression line. If it is positive, the line increases, if it is negative, it decreases. If the variable Y was independent of X, then the regression line would be horizontal[4].

The residuals of linear models have a property that has been already discussed in the context of the arithmetic mean in basic statistics courses. The overall magnitude of the residuals for observations above the regression line is the same as for observations below the regression line except for their sign. Thus, the sum of all residuals is always zero.

[4] The relationship between the value of the parameter β_1 and the angle between the regression line and the x axis (labeled α) can be expressed as $\beta_1 = \tan(\alpha)$.

Figure 2: Graphical representation of simple regression



3 Model quality indicators

At the beginning we discussed that statistical models try to propose the most accurate simplification of reality. When we estimate the parameters of the model, we can ask how accurately our model replicates reality. For example, in the previous chapter, we worked with a model that supposes that, with a certain amount of simplification, the number of points students receive on a test is determined solely by the number of hours they spent studying, and that this relationship is linear (i.e., for every hour of studying, we are given β_1 points). It is probably reasonable to assume that the amount of time spent studying is one of the main factors that affects test scores, but have we not oversimplified reality? To what extent does our model correspond to reality? We will use indicators of model quality in our search for an answer to this question.

We have already been introduced to the residual sum of squares, which somehow quantifies how major the errors of models are; how much it differs from reality. Under certain circumstances, therefore, RSS could be used as a measure for model accuracy. However, we do not usually use it in this way for two reasons. RSS depends on the number of observations (n), so if we have many observations, then we usually get a higher RSS than if we have only a few of them. The second property that may bother us is that RSS depends on the scale of the dependent variable.

A more useful indicator is **residual variance**. It simply represents the sample variance of the residuals of our model. Compared to the usual procedure for calculating variance, the only difference here is that the number of degrees of freedom is not $n-1$, as we are used to, but generally $n-p$, where p is the number of estimated parameters. Therefore

$$S_\epsilon^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{\text{RSS}}{n-p}$$

In a simple regression, we use $n-2$ degrees of freedom for the calculation since we have to estimate two parameters (β_0, β_1). Sometimes we also encounter the square root of the residual variance S_ϵ, the **residual standard deviation** (also called the residual standard error).

In many cases, the residual variance or standard deviation is a useful indicator of model quality. However, it is not suitable, for example, for comparing different models from different studies, as it depends on the unit of measurement. This issue can be overcome by far by the most popular indicator of model quality, which is the **coefficient of determination** R^2. The coefficient of determination can be calculated in several ways. For example, it can be derived from RSS by standardizing it and subtracting it from one:

$$R^2 = 1 - \frac{\text{RSS}}{\text{SS}_Y}$$

where SS_Y is the sum of squares of the dependent variable, i.e., $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Another way of understanding the coefficient of determination is the square of the Pearson correlation coefficient between the dependent variable Y and its fitted (i.e., estimated) values $\hat{Y}$. The coefficient of determination takes any value between zero and one.

The coefficient of determination is favorable for its universal use – it can be used to compare models regardless of their scale or sample size. Moreover, it can be interpreted very intuitively as **the percentage of explained variance of the dependent variable**. Therefore, if R^2 equals 1.0, we can perfectly predict values of the Y variable from the values of the X variables. If R^2 equals 0.5, it means that we can explain 50 % of the variance of Y. The remaining 50 % then accounts for the factors we did not include in our model, along with the measurement error of the Y variable.

Despite its excellent properties, R^2 has one weakness. If we add another regressor to the model, the value of R^2 increases. Even if the new regressor is completely meaningless, the percentage of explained variance never decreases. In the most extreme case, it may stay the same, but in practice it will always increase due to random variation – the fewer observations we have, the larger the random variation, and each additional (albeit meaningless) regressor explains that much more of the variance. So, if we include a large number of regressors in the model and have a relatively small set of observations, the coefficient of determination starts to rise to unrealistically high values[5].

This problem can be overcome by adjusted coefficient of determination $R^2_{adj.}$ which takes into account the number of estimated parameters:

$$R^2_{adj.} = 1 - \frac{S^2_\epsilon}{S^2_Y} = 1 - \frac{\mathrm{RSS}}{\mathrm{SS}_Y} \cdot \frac{n-1}{n-p} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p}$$

However, values obtained using this method do not have such a straightforward interpretation (in the extreme case it may even be negative), so it is convenient to present it together with the original R^2. When presenting a linear model, we generally always use the R^2 indicator. If the reader might suspect that there is an overestimation of R^2 due to the high number of estimated parameters, then it is appropriate to present R^2 as well as $R^2_{adj.}$.

Let us show an example of calculating model quality indicators on the data from Agatha, Otto and their classmates. Table 1 contains the values that were used to construct the plot in Figure 2.

[5] In an extreme case where the number of estimated parameters is the same as the number of observations ($n = p$), R^2 is always equal to 100 % regardless of which dependent or independent variables have been selected.

Table 1: Values of observed variables, predictions and residuals

Student	hours studying X	points Y	prediction $\hat{Y}$	residual ϵ	residual squared ϵ^2
Agatha	16	42	45.38	−3.38	11.46
Otto	10	30	36.62	−6.62	43.76
Ursula	2	24	24.92	−0.92	0.85
Boris	11	53	38.08	14.92	222.70
Ivanka	24	54	57.08	−3.08	9.47
Anastasia	20	48	51.23	−3.23	10.44
Nela	13	35	41.00	−6.00	36.00
Rosemarie	21	61	52.69	8.31	69.02

The least squares method applied to data from eight students gives us parameter estimates of $\hat{\beta}_0 = 22.00$ and $\hat{\beta}_1 = 1.46$. For each student, we substitute value X_i into the equation $\hat{Y}_i = 22.00 + 1.46 \cdot X_i$ to obtain the fitted values, i.e., the predicted values of the dependent variable. The residuals are calculated as the differences between the actual and fitted values of the dependent variable, i.e., $\epsilon_i = Y_i - \hat{Y}_i$. RSS is then obtained as the sum of the squares of the residuals:

$$\text{RSS} = 11.46 + 43.76 + 0.85 + 222.70 + 9.47 + 10.44 + 36.00 + 69.02 = 403.69$$

It is difficult to judge by eye whether a value of over four hundred is a lot or a little. The estimate of the residual variance S^2_ϵ is slightly more informative:

$$S^2_\epsilon = \frac{\text{RSS}}{n-p} = \frac{403.69}{8-2} = 67.28$$

and in particular the residual standard deviation S_ϵ:

$$S_\epsilon = \sqrt{67.28} = 8.20$$

The standard deviation of the residuals is therefore just over eight points. This already helps us to get a picture – if we predict how someone will perform according to our model, we can expect errors of, say, 5 or 10 points, but we can practically rule out the possibility that the model would be wrong by, say, 30 or 50 points.

In order to decide more accurately whether the 8.2 points are a lot or a little, we need to examine the variance of the dependent variable Y. We calculate the average number of points $\bar{Y} = 43.38$ and use it to determine the sum of squares $\text{SS}_Y = \sum_{i=1}^{8}(Y_i - 43.38)^2 = 1163.88$. After dividing by $n - 1$ degrees of freedom, we obtain the sample variance of the variable Y, $S^2_Y = 166.27$, or its standard deviation of $\sqrt{166.27} = 12.89$ points. This can be interpreted as if we did not know how long each student had been studying, and thus we expected an average score for each student (43.38), our predictions would have a standard deviation of 12.89. If we use the information about the time spent studying to make the prediction, the standard deviation of the errors drops to 8.20 points, which is a reasonably satisfactory result.

The most informative indicator is nonetheless the coefficient of determination R^2:

$$R^2 = 1 - \frac{403.69}{1163.88} = 0.65$$

Thus, we can say that we were able to explain (describe, predict) 65 % of the variance in the number of points students receive on the exam using information about their time spent studying. In contrast, error variance accounts for the remaining 35 %. This covers all the influences that we did not include in the model (e.g., prior knowledge, learning ability, test inaccuracy, copying...).

If we suspect that our result is overestimated due to the small number of observations ($n = 8$) relative to the number of estimated parameters ($p = 2$), we calculate the adjusted coefficient of determination $R^2_{adj.}$:

$$R^2_{adj.} = 1 - \frac{67.28}{166.27} = 0.60$$

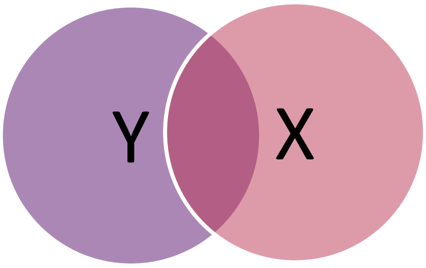
It can be seen that there has been some change and we should probably point out to the reader the possible influence of the limited sample size.

The question is what value of R^2 we can be satisfied with and which one we should consider insufficient. Unfortunately, the answer is not straightforward and depends on the purpose of our model. For example, if you come up with a shocking claim that in adult population there is a parasite spreading widely and it lowers the IQ of its host, then even a model whose R^2 is barely 3 % will make the newspaper headlines. On the other hand, if you try to argue that the results of the school leaving exams are a good indicator of a student's abilities and that they can predict their academic success at university, then you will not find R^2 values below 30 % satisfactory. Finally, if the model covers physics or some of the technical disciplines, then models with R^2 below 99.9 % will not be worth attention at all (let us add that in this case we will probably reach for another indicator of accuracy, probably the familiar residual standard deviation).

4 Models with multiple regressors

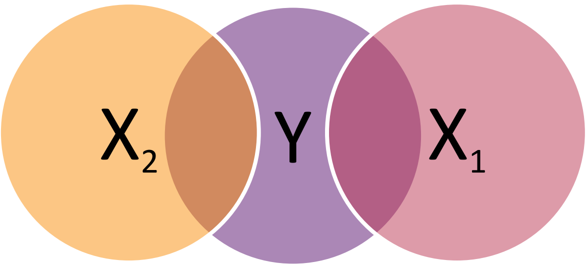
Everything we have learned in the previous pages applies not only to single regression, but also to models with multiple regressors. However, unlike simple regression, more complex models cannot usually be represented by a regression line in a scatter plot, so we need to understand the relationships between the variables using only the values provided by statistical software. It is therefore useful to be able to imagine what the relationships between the independent variables might look like and how this will be reflected in their regression weights.

Imagine the variance of the dependent variable Y and the independent variable X as two circles.



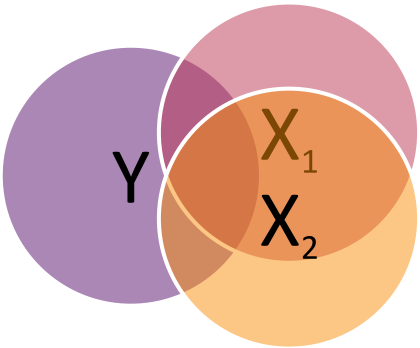
The two variables share part of the variance, therefore the circles overlap. Using the X variable we could therefore explain part of the variance of the Y variable. In this case, the overlap is 25 % of the area of the circle, so $R^2 = 0.25$. Since we know that in the case of a simple regression, R^2 is equal to the square of the Pearson correlation coefficient between the variables X and Y, or the square of the standardized regression coefficient β^*, since $r_{XY} = \beta^* = \sqrt{R^2} = 0.5$.

Let us add another regressor to the model, X_2.



The figure shows that this regressor also shares 25 % of the variance with the Y variable. Regressors X_1 and X_2 are not correlated (share no variance) and thus their circles do not overlap. Even in this case, we can easily estimate the values of the coefficient of determination and the regression weights. If regressor X_1 explains 25 % of the variance and regressor X_2 explains another 25 % of the variance, then $R^2 = 0.5$ since the model can explain a total of a half of the variance of the dependent variable. Let us add that the correspondence between the standardized regression weights and Pearson correlation coefficients between the regressors and the dependent variable also applies here. Furthermore, in the case of uncorrelated regressors, it is also true that the sum of the squares of the standardized regression coefficients is equal to the coefficient of determination: $0.5^2 + 0.5^2 = 0.25 + 0.25 = 0.5$.

In reality, however, we most often encounter a third case, where regressors share part of the variance not only with the dependent variable but also among themselves:



Again, X_1 overlaps Y by one quarter and the same is true for X_2. This time, however, the two regressors are correlated and the part of the variance that can be explained by X_1 can be explained as well with X_2. The coefficient of determination will obviously be lower than the original 50 %. The figure also shows that by adding more regressors, R^2 cannot decrease, while it can increase. The question is which regressor is given what regression weight, i.e., which regressor will be used to describe the variance shared by all three variables. We cannot use our circle metaphor here – the least squares method will find the answer – and we would hardly find a reasonable rule that could be described in words[6].

The message we should take from the previous paragraphs is that adding another regressor will usually change the values of all the other regressors' weighs in any possible direction. Usually, this change is desirable – adding another regressor will reduce some of the error variance and give us a clearer view of what role the independent variables actually play. Let us illustrate this with an example.

[6] Our circle metaphor suggests that R^2 never exceeds the sum of the squared coefficients β^*, which is the sum of the variances that explain each regressor separately. However, this is not true. There is a situation where the individual regressors are weakly correlated with the dependent variable, but once they are put into the model together, they explain an unexpectedly large amount of variance. This is, incidentally, a situation where we may see coefficients of $beta^*$ exceeding 1. In academic papers, scholars use the term *suppressor variable*.

Let us test a hypothesis that the intense feeling of stage fright that accompanies a person's public performance is related to their self-esteem. We could conduct a small research project for this purpose. We let 20 students present their papers in front of a full auditorium and use a questionnaire to measure how much stage fright they experienced. We will also have them complete several questionnaires before the experiment, including the Rosenberg's Self-Esteem Scale (RSS) and the Neuroticism scale from the NEO-FFI questionnaire. The data obtained could look like those in the Table 2.

Table 2: Data: neuroticism, self-esteem and stage fright

Student	Neuroticism	Self-esteem	Stage fright
1	14	39	2
2	22	31	6
3	16	36	2
4	19	32	8
5	24	36	5
6	24	26	6
7	25	23	5
8	26	36	4
9	20	23	4
10	21	28	5
11	30	24	7
12	18	35	5
13	33	22	6
14	24	31	6
15	23	28	5
16	14	29	1
17	30	25	5
18	28	23	10
19	16	37	2
20	21	31	2

If we focus only on the relationship between self-esteem and experienced stage fright, we are likely to be pleased with the results. After running a simple regression, we find the following regression weights:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	10.90	
Self-esteem	-0.21	-0.50

The value of the standardized regression coefficient (and hence the Pearson correlation coefficient) is a satisfying -0.50. We can conclude that this is a moderate to strong relationship and that the Rosenberg test score and the rating of experienced stage fright share 25 % of the variance.

If we decide to examine separately the relationship between the neuroticism score and the ratings of stage fright, we find that a close relationship is obvious here as well:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	-1.16	
Neuroticism	0.27	0.64

The correlation coefficient between the two variables is 0.64, and thus the entire 41 % of the variance of stage fright can be described using neuroticism. Thus, we might believe that we have found two key factors that are related to the stage fright the participants experienced. However, we have omitted in our considerations the fact that the constructs of self-esteem and neuroticism overlap to a large extent. According to our data, the correlation between the two constructs is -0.65. Thus, it cannot be ruled out that both variables explain the same variance in the stage fright variable. Results of multiple regression confirm this concern:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	1.50	
Neuroticism	0.23	0.55
Self-esteem	-0.06	-0.15

The two regressors together can explain 42 % of the variance, just one percent more than neuroticism alone. The weight of self-esteem is close to zero, while the weight of neuroticism remains high. The intensity of the stage fright the participants experienced is dependent on the individual's neuroticism, while self-esteem has only little impact.

The example might give the reader the impression that multiple regression is a guaranteed way to ruin promising results. However, the opposite is true – multiple regression helps us identify relevant relationships and reinforces our belief that the observed correlation is not an artifact due to a third factor. It is highly desirable to include regressors such as age of probands and their gender in the model to remove their unwanted influence. To these regressors that we include in the model to avoid bias, but their influence is not the focus of our interest we refer to as **covariates**.

4.1 Nominal regressors

So far, we have only considered cases where the independent variables X are quantitative. Quite often, however, we need to include in our considerations regressors measured on a nominal scale, such as the gender of the proband, their assignment to one of the experimental or control groups, or for example their nationality. We can easily address this issue in the context of linear models.

To begin with, let us imagine the simplest situation where our model contains a single independent variable X that is dichotomous. For example, this would be the case where we are investigating the effect of tetrahydrocannabinol (THC) intoxication on human reaction time. The design could look like this: we randomly divide volunteers into an experimental group and a control group. The experimental group will be given a dose of THC while the control group will be given a placebo. We then use a computer test to measure reaction times of both groups. The data matrix might look something like this:

Table 3: Data: reaction time and THC

Proband	THC	RT [ms]
1	0	523
2	0	603
3	0	669
4	0	500
5	0	662
6	0	643
7	1	657
8	1	784
9	1	504
10	1	802
11	1	561
12	1	514

We coded which group an individual belongs to using ones and zeros, where one indicates the experimental (intoxicated) group. If we want to build a model that describes the impact of the experimental condition, we surprisingly end up with exactly the same equation we used in the simple regression:

$$Y = \beta_0 + \beta_1 X$$

The X variable in this case is called **dummy variable** and it denotes individuals who belong to the experimental group with the number one. The computation of the coefficients of $\hat{\beta}$ and all other procedures remain exactly the same as in the simple regression. The least squares method provides us with these estimates:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	600	
THC	37	0.19

Not only has the calculation procedure not changed, but the interpretation of the regression weights also remains the same. The parameter β_0 shows what value the dependent variable has on average when all regressors are equal to zero. If our dummy variable is equal zero, it means that we are talking about an individual from the control group. Thus, a result $\hat{\beta}_0 = 600$ means that the average reaction time of people in the control group is equal to 600 milliseconds. The β_1 parameter generally indicates how many points on average the value of the dependent variable increases when the regressor value increases by one. In our case, the fact that the value of the regressor increases by one does not mean anything other than that the individual moves from the control group to the THC-intoxicated group. Thus, the averages of the two groups differ by 37 ms, from which we can easily deduce that probands from the experimental group had a reaction time equal to 637 milliseconds on average.

Interpreting the β^* coefficient is a bit more difficult. Again, it is the number of standard deviations by which Y increases when X increases by one standard deviation. But the standard deviation of a dichotomous variable is not a very meaningful concept. Therefore, the coefficient β^* cannot be interpreted very sensibly; however, we can use it as an indicator of the effect size, perhaps to compare the effects of multiple variables X with each other.

The situation becomes a bit more complicated when the nominal regressor has more than two levels. What if we wanted to extend our experiment to other addictive substances and include a third group intoxicated with a dose of ethanol? Our first idea might be to label this alcohol group with 2 in the second column of the table. But the results obtained in this way would be meaningless. It would mean that we are assuming the existence of nothing-THC-ethanol continuum, where ethanol is twice as high as THC. But in reality, these are two qualitatively different aspects that we cannot project onto one axis.

The solution is to introduce one dummy variable for THC and another dummy variable for ethanol. This solution can be seen in the Table 4.

Table 4: Data: reaction time, THC and ethanol

Proband	THC	Ethanol	RT
1	0	0	523
2	0	0	603
3	0	0	669
4	0	0	500
5	0	0	662
6	0	0	643
7	1	0	657
8	1	0	784
9	1	0	504
10	1	0	802
11	1	0	561
12	1	0	514
13	0	1	677
14	0	1	790
15	0	1	778
16	0	1	723
17	0	1	646
18	0	1	682

The parameter estimates are again obtained without changing the procedure using the least squares method. The data above lead us to the following results:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	600	
THC	37	0.19
ethanol	116	0.56

Of course, adding another group did not affect the average in the previous two groups – still the average reaction time in the control group is 600 ms and in the THC-intoxicated group 600+37 = 637 ms. In addition, however, we obtain the information that our ethanol intoxicated probands are on average 116 ms slower compared to the control group, so their average reaction time is 716 ms.

The use of dummy variables can sometimes be somewhat counter-intuitive. Let us therefore point out a few facts:

- Although we are working with three groups, we have only two dummy variables. We would find the same relationship for higher numbers; for example, working with a nominal variable of ten levels, we would include 9 dummy variables in the model. The tenth, omitted, variable is called **reference group** and can be chosen as desired.

- The coefficient $\hat{\beta}_0$ (intercept) is the average value of the variable Y in the reference group.
- The coefficients $\hat{\beta}_1, \hat{\beta}_2, ...$ show how much each group differs from the reference group. If we want to compare the groups with each other, we can specify a different factor level as the reference group.
- If we choose a different reference group, the overall accuracy of the model (RSS, R^2) does not change, but of course the regression coefficients of each group change. For example, if we choose THC-intoxicated probands as the reference group, we obtain results that, upon closer examination, lead us to the same group means:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	637	
ethanol	79	0.38
control	−37	−0.18

If we include a set of dummy variables in the model, then we are usually no longer talking about regression, but about a **general linear model**. Of course, we can include more than one categorical regressor in the model and they can also be freely combined with continuous regressors.

Dummy coding is particularly useful when we have a single control group against which we compare several experimental groups, or when we are working with a dichotomous variable. If we need to compare several groups, none of which is in any way privileged to be considered a reference group, this method is not very elegant.

Note also that there is a number of other ways of coding the nominal variable that lead to different interpretations of the regression parameters. In general, however, they do not change the overall accuracy of the model.

4.2 Interactions of regressors

The relationship between the independent variables X and the dependent variable Y is sometimes more complicated than to be described as a weighted sum. A trivial example is coffee sweetening. If you stir the coffee, it will not be sweet all of a sudden. If you put sugar in it, the taste will not change either, because the sugar sits at the bottom of the cup and does not dissolve. So, we would say that neither action affects the taste of the coffee. But if you do both – put sugar in the coffee and stir it – you get an incomparably greater effect. The interaction can also work in the opposite way. If we stick to simple culinary examples, we can model how a pancake tastes depending on what ingredients we put on it. For example, marmalade or Nutella will definitely enhance the tastiness of the pancake. Similarly, cheese or ham will have a positive effect. However, if you add all these ingredients at the same time, the effects will not add up, but on the contrary, the tastiness of the resulting product will definitely be in the negative numbers.

Both quantitative and nominal variables can enter the interaction. For now, let us ignore nominal variables with more than two levels, since it these instances the situation is a bit more complicated. In all other cases we can include the effect of the interaction between X_1 and X_2 by adding **interaction term** to the regression equation. A model with k regressors, where the first two regressors interact, would be described by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \cdots + \beta_k X_k$$

The X_1X_2 variable is in fact the value of the X_1 regressor multiplied by the value of the X_2 regressor. The β_{12} coefficient then represents the weight of this newly created regressor. Thus, the actually including the interaction into our model is technically very straightforward. The greater challenge is to interpret the interaction term correctly. Let us demonstrate how to work with the interaction on the following exercise.

Rumors are spreading about a professor – he is not at all objective in his examinations, but grades students on the basis of only one criterion, which is the length of their skirts. The students decided to verify this suspicion with a linear model. They simply measured the length of skirts or trousers of anyone who took the exam and wrote the value in centimeters in a table. They also made a note of whether the person was a man or a woman, and what grade they were given. Table 5 contains the students' records at the end of the examination period. For easier mathematical processing, the grades were converted to numbers from the best grade A, labeled 1, to the worst grade F, labeled 6.

The idea that grading can be influenced by two factors (the gender of the student and the length of the skirt or trousers) seems correct at first sight in this context. Therefore, let us test the students' assumption using a model with these two regressors

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot \text{gender} \ + \ \beta_2 \cdot \text{length}$$

which leads to the following result:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	3.001	
gender	0.045	0.015
length	0.006	0.092
R^2	1 %	

Table 5: Records of students' grades and the length of their skirts or trousers

Student	Grade	Gender	Length
1	B (2)	1	104
2	D (4)	0	84
3	B (2)	1	86
4	B (2)	0	48
5	E (5)	1	92
6	A (1)	1	90
7	C (3)	1	95
8	D (4)	0	88
9	E (5)	0	59
10	A (1)	1	101
11	C (3)	0	38
12	E (5)	0	76
13	E (5)	0	82
14	A (1)	0	28
15	B (2)	0	29
16	D (4)	1	59
17	C (3)	1	100
18	A (1)	0	41
19	C (3)	0	48
20	B (2)	0	48
21	E (5)	0	79
22	A (1)	0	33
23	F (6)	0	70
24	C (3)	0	40
25	E (5)	1	59
26	D (4)	0	80
27	B (2)	0	21
28	D (4)	0	70
29	C (3)	1	99
30	E (5)	1	52
31	C (3)	0	48
32	E (5)	0	89
33	E (5)	0	83
34	C (3)	1	79
35	F (6)	1	30
36	B (2)	0	39
37	B (2)	1	102
38	B (2)	1	100
39	C (3)	0	46
40	E (5)	0	92
41	E (5)	1	91
42	C (3)	0	88
43	C (3)	1	95
44	C (3)	1	80
45	F (6)	1	41
46	B (2)	0	41
47	F (6)	1	52

Men are coded 1, women 0. The length of the skirt or trousers is in centimeters.

We can easily interpret the estimated weights of both regressors. Men, labeled with one, receive a 0.045 grade (i.e., about one-twentieth) higher (i.e., worse) than the reference group of women. For every centimeter of skirt length, the grade increases (worsens) by 0.006 degrees (i.e., about one hundred and seventy-seventh). At first glance, it is obvious that we are talking about virtually zero effect. The regressors explain approximately 1 % of the variance of the observed variable.

Was it really a false accusation and do we owe the accused professor an apology? Before we accept that conclusion, let us think a little more closely about what our model says. A graphical representation of the model as a regression line will help. Since we know that the regressor *gender* takes the value 0 for all women and 1 for all men, we can look at what the model says about men and women separately simply by substituting the gender regressor with a corresponding value (0 or 1).

In the case of women (gender = 0):

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot \text{gender} \ + \ \beta_2 \cdot \text{length}$$

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot 0 \ + \ \beta_2 \cdot \text{length}$$

$$\text{grade} = \beta_0 \ + \ \beta_2 \cdot \text{length}$$

$$\text{grade} = 3.001 \ + \ 0.006 \cdot \text{length}$$

In the case of men (gender = 1):

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot \text{gender} \ + \ \beta_2 \cdot \text{length}$$

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot 1 \ + \ \beta_2 \cdot \text{length}$$

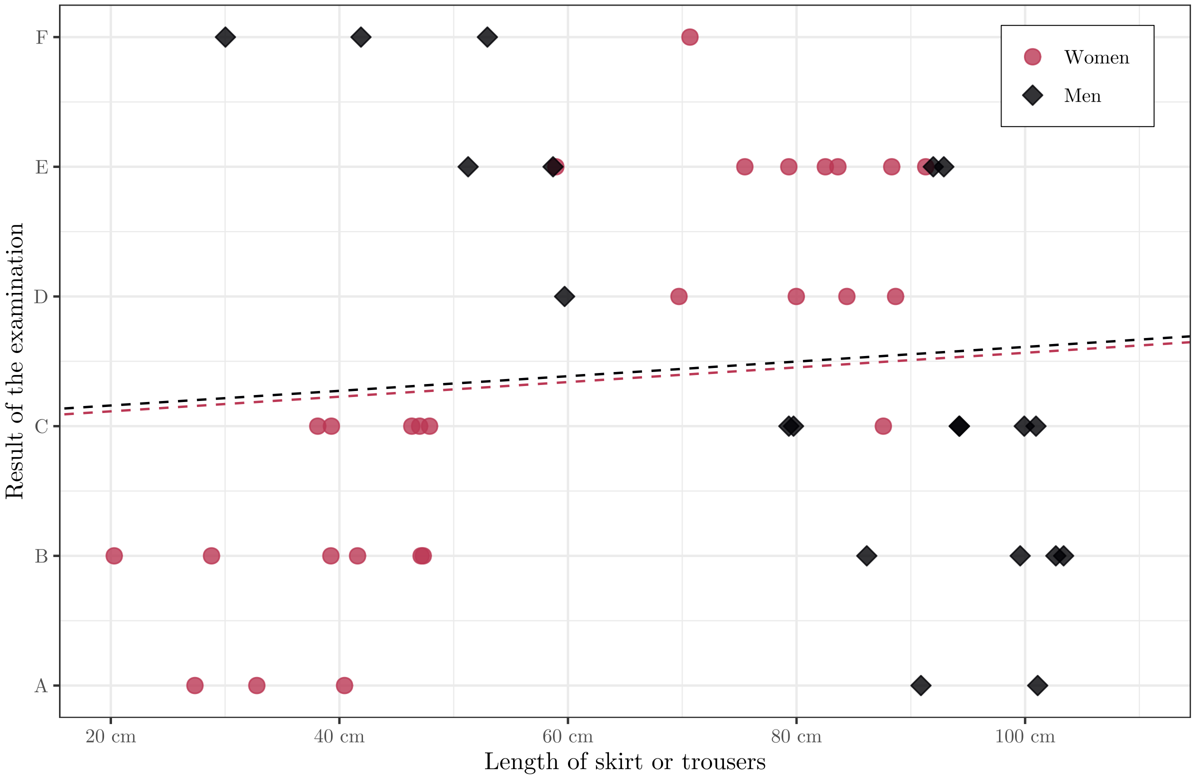
$$\text{grade} = (\beta_0 + \beta_1) \ + \ \beta_2 \cdot \text{length}$$

$$\text{grade} = (3.001 + 0.045) \ + \ 0.006 \cdot \text{length}$$

$$\text{grade} = 3.046 \ + \ 0.006 \cdot \text{length}$$

Both regression lines have the same slope (0.006) and differ only in the vertical displacement (the intercepts are equal to 3.001 and 3.046). The similarity of the two lines can be seen in Figure 3.

Figure 3: Inadequate model without an interaction term



Many readers are probably already aware of the mistake we made in the specification of the model. The length of skirt or trousers has a somewhat different impact depending on whether we are talking about a woman or a man. To allow our model to describe reality more accurately, we need to design it in a way that the weight of the *length* can have different values depending on the value of the *gender*. And this is exactly what we get by adding the interaction term *gender* × *length*.

As mentioned above, in order to include the interaction, we extend the data table by one more column, where the *gender* values are multiplied by *length* values (the first few rows are given in Table 6). The interaction term is equal to zero for women and to the value of the *length* for men since the *gender* takes values of 0 and 1 only. The procedure would be the same for the two quantitative regressors as well.

Table 6: An example of data for a model with an interaction term

Student	Grade	Gender	Length	Interaction
1	B (2)	1	104	104
2	D (4)	0	84	0
3	B (2)	1	86	86
4	B (2)	0	48	0
5	E (5)	1	92	92
6	A (1)	1	90	90
7	C (3)	1	95	95
8	D (4)	0	88	0
9	E (5)	0	59	0
10	A (1)	1	101	101
⋮	⋮	⋮	⋮	⋮

Our model with interaction takes the following shape:

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot \text{gender} \ + \ \beta_2 \cdot \text{length} \ + \ \beta_3 \cdot \text{gender} \cdot \text{length}$$

and after recalculation it leads to these values of parameter estimates:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	0.260	
gender	7.759	2.539
length	0.052	0.853
interaction	−0.109	−3.038
R^2	63 %	

Before discussing the drastic increase in the explained variance of the dependent variable, let us consider what the individual unstandardized coefficients of the model with interaction imply. Again, it helps to examine the model separately for men and women.

In the case of women (gender = 0):

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot \text{gender} \ + \ \beta_2 \cdot \text{length} \ + \ \beta_3 \cdot \text{gender} \cdot \text{length}$$

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot 0 \ + \ \beta_2 \cdot \text{length} \ + \ \beta_3 \cdot 0 \cdot \text{length}$$

$$\text{grade} = \beta_0 \ + \ \beta_2 \cdot \text{length}$$

$$\text{grade} = 0.260 \ + \ 0.052 \cdot \text{length}$$

In the case of men (gender = 1):

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot \text{gender} \ + \ \beta_2 \cdot \text{length} \ + \ \beta_3 \cdot \text{gender} \cdot \text{length}$$

$$\text{grade} = \beta_0 \ + \ \beta_1 \cdot 1 \ + \ \beta_2 \cdot \text{length} \ + \ \beta_3 \cdot 1 \cdot \text{length}$$

$$\text{grade} = (\beta_0 + \beta_1) \ + \ (\beta_2 + \beta_3) \cdot \text{length}$$

$$\text{grade} = (0.260 + 7.759) \ + \ (0.052 - 0.109) \cdot \text{length}$$

$$\text{grade} = 8.020 \ - \ 0.056 \cdot \text{length}$$

The β_0 and β_2 parameters can be interpreted as the intercept and slope of the regression line in the reference group of women. The β_1 and β_3 parameters tell how much the origin and slope differ in the group of men compared to the reference group of women. Note that β_1 and β_3 are not the values of the intercept and slope of the regression line for men! The origin for men is $\beta_0 + \beta_1$ and the slope of the regression line is $\beta_2 + \beta_3$. This is probably the most common mistake students make when working with interactions.

Let us compare the regression lines in Figure 3 with the lines in Figure 4. As one can observe, for every 20 centimeters of skirt length, women's grades on average worsen by approximately one level (more precisely by $0.052 \cdot 20 = 1.045$ levels), while for men every 20 centimeters of trousers length improves the grade by approximately one level ($-0.056 \cdot 20 = -1.125$ levels).

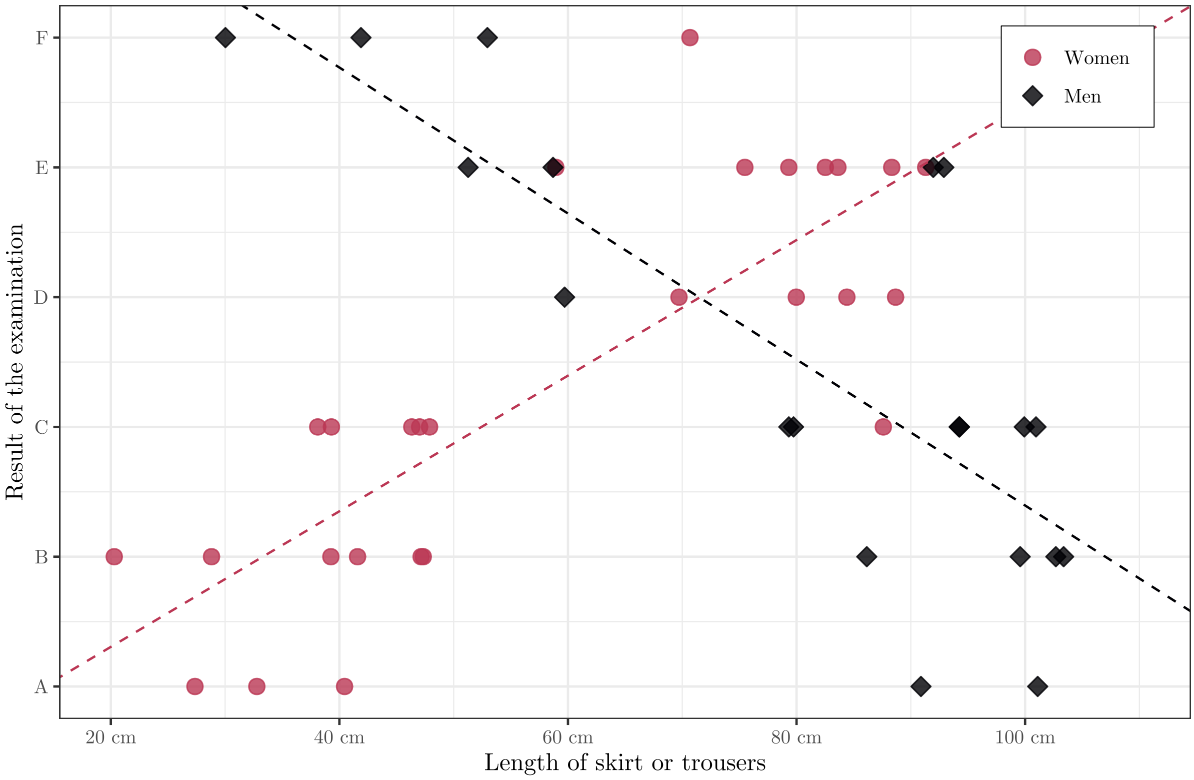
Was the students' concern justified? In this case, $R^2 = 63$ % is extremely high. The result basically proposes that almost two-thirds of the variance in grades depends on whether the examinee is a man or a woman and what they are wearing. The remaining one-third of the variance is then crammed with all the other influences, such as knowledge, which should be the main factor.

As mentioned several times before, any type of variable can enter the interaction. If we wanted to examine the interaction of a nominal variable with more than two levels, we would have to create an interaction term for each dummy variable that we derived from the original nominal variable. Thus, to calculate the interaction between a nominal variable with three levels and a nominal variable with four levels, we would need to create 6 ($= (3 - 1) \cdot (4 - 1)$) interaction terms.

Higher order interactions can be created analogously – to multiply more than two variables together. However, in the case of two variables, interpretation is often difficult already, and in the case of three or more variables, it is usually almost impossible.

When analyzing interactions, it makes no sense to include standardized regression coefficients β^* in our considerations, as they provide virtually no relevant information.

Figure 4: Model with interaction term



4.2.1 Centering of regressors

When describing the results of the previous example, the interpretation of the intercept β_0 and the weight of the regressor *gender* was avoided. Both weights have their interpretations, although, as we will see, in our case they are not very meaningful. The intercept β_0 informs us what value the dependent variable keeps when all regressors are zero. Thus, it is the average score of a woman ($gender = 0$) whose skirt is 0 cm long. This scenario is still conceivable with a little imagination, but the value 0.26 itself is no longer meaningful (it is a grade three-quarters of a level better than A).

The weight β_1 describes an even more bizarre situation: it is the difference between an average grade of a woman without a skirt ($length = 0$ cm) and a man without trousers. The value between a 7 and an 8 ($\beta_1 = 7.759$) is again meaningless, given that the grades have only 6 levels.

We can come across similar meaningless weights quite often when working with linear models. And it does not have to be only models with an interaction term, although this situation is typical for them. Let us imagine that we predict the wage using IQ of an individual. The intercept will then be the average wage of a non-existent person who has an IQ of zero. When we include age in the model, the intercept will represent the salary of a newborn baby, and so on.

All of the situations described above are correct from a mathematical point of view, but it is hard to deny that they contradict common sense. To avoid such absurdities, it is worth to master *centering of regressors.*

Centering the regressor means that we calculate its arithmetic mean and subtract it from the value of the regressor in every observation. If we wanted to center the length of the skirt or trousers in the previous example, we would calculate the average length, which equals 67.979 cm, and subtract this value from all lengths. So, the variable *length* no longer gives the absolute length, but how much the length differs from the average length of the skirt or trousers in our sample. The new form of the data matrix is shown in the Table 7. The estimates of the regression weights of the model change as follows:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	3.812	
gender	0.384	0.126
length	0.052	0.853
interaction	−0.109	−1.135
R^2	63 %	

Table 7: Example of a centered regressor

Student	Grade	Gender	Length	Interaction
1	B (2)	1	36.021	36.021
2	D (4)	0	16.021	0
3	B (2)	1	18.021	18.021
4	B (2)	0	−19.979	0
5	E (5)	1	24.021	24.021
6	A (1)	1	22.021	22.021
7	C (3)	1	27.021	27.021
8	D (4)	0	20.021	0
9	E (5)	0	−8.979	0
10	A (1)	1	33.021	33.021
⋮	⋮	⋮	⋮	⋮

Centering the regressor does not change the accuracy of the model or the predictions provided by the model. The slopes of the regression lines remain the same. Only the value and interpretation of the β_0 and β_1 weights have changed. The intercept still indicates, on average, what value of the dependent variable we get if all regressors are equal to zero. But since the regressor *length* is centered, a zero value means that we are talking about an average skirt length. The weight β_1 then quantifies the difference in the expected grade between a man and a woman who have average-length trousers or skirt, respectively.

If we center the *IQ* regressor, the intercept will refer to a person with average intelligence, if the *age* regressor, then to an average aged person, etc. We do not have to use the arithmetic mean, but any other constant (although this is not centering in the true sense of the word). For example, we could have used the number 70 instead of the average 67.979 cm, or even centered men and women separately by their group averages. None of these steps would change the results the model gives us. It will, however, change the values and interpretations of the regression coefficients involved.

There is usually no point in centering dichotomous variables. If our model involves the interaction of two quantitative variables, centering them is almost always a necessity if we are looking for understandable results.

4.3 Exploring non-linear relations

Common knowledge, we are all familiar with from basic statistics courses, says that the Pearson correlation coefficient describes only the *linear* relationship between the variables of interest. The regression lines we have worked with in previous chapters also have this property. In practice, however, we often encounter situations where there is different than a linear relationship between the variables of interest.

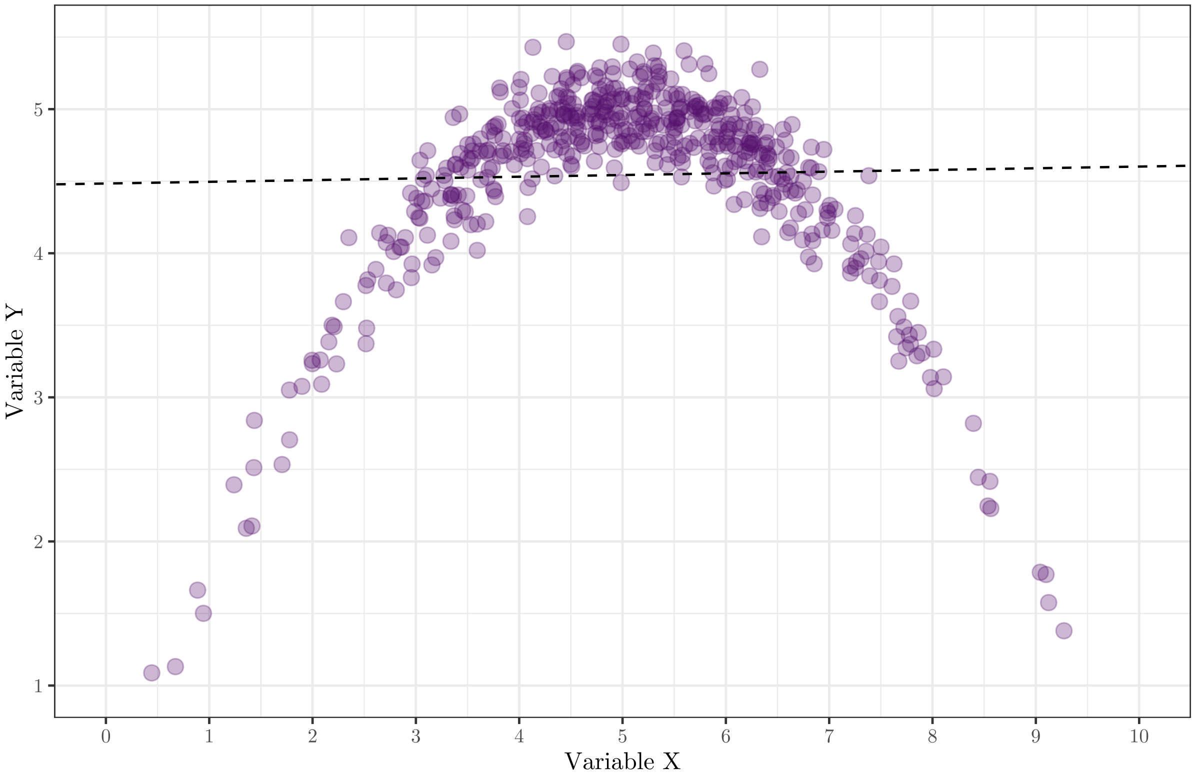
Let us imagine that we are working with two variables from the scatter plot in Figure 5. There is obviously a close relationship between these variables. However, if we try to describe this relationship using a simple regression model, we fail to detect it.

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	4.484	
variable X	0.012	0.026
R^2	0 %	

In fact, the linear model can describe a wide range of non-linear relationships, but we have to adjust the regressors accordingly[7]. In psychology, when we talk about a nonlinear relationship, in the vast majority of cases we mean a quadratic relationship. This assumes that the dependence is parabolic – it resembles the letter U (or ∩), or some part of it.

[7] It is a fairly common belief that the word *linear* in the name of the model is somehow related to the type of relationship between X and Y. However, this is a misunderstanding – the word linear means that the dependent variable is modeled as a linear combination (i.e., a weighted sum) of regressors, with the model parameters as weights.

Figure 5: Data incorrectly fitted with a line



To describe a quadratic dependence, we need to add a given regressor in the second power to the model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

In practice, we make this change by adding a new column to the data table that contains the squared values of the original regressor. Table 8 contains a few rows of the data matrix modified in this way. The next steps of the calculation are identical to the procedure we are already familiar with. The data from Figure 5 will lead us to the following parameter estimates.

Regressor	$\hat{\beta}$	$\hat{\beta}^*$
(intercept)	-0.012	
variable X	2.004	4.371
variable X^2	-0.200	-4.449
R^2	92 %	

Figure 6: Data fitted with a quadratic function

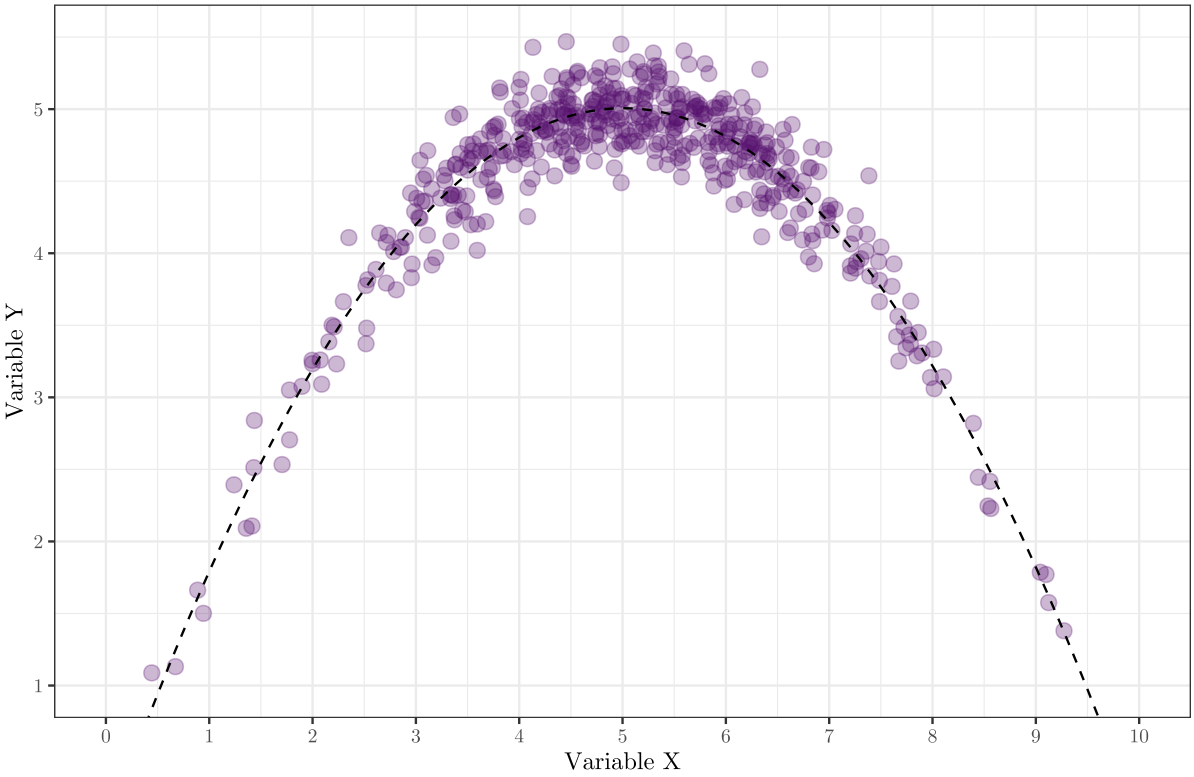


Table 8: Example of data for quadratic dependence modelling

Y	X	X^2
13.2	1.5	2.25
21.2	3.0	9.00
21.4	3.2	10.24
23.6	4.0	16.00
24.4	4.2	17.64
25.3	4.8	23.04
24.8	5.5	30.25
24.3	5.7	32.49
23.7	6.0	36.00
16.9	7.8	60.84
10.4	8.8	77.44
⋮	⋮	⋮

Fitting the data with a parabola gives a satisfactory result in our case. It is not only the high value of the coefficient of determination, but also the plot in Figure 6, which shows our estimated parabola, that give us the confidence.

When we model the effect of a regressor as quadratic, the regression weights lose their elegant interpretations. The standardized regression coefficients do not help us interpret them at all (note that in our example they came out at absurdly high values), and even the non-standardized ones provide only limited insight. The coefficient β of the regressor X^2 indicates whether the curve is U-shaped (positive values) or ∩-shaped (negative values). **If the weight of the quadratic regressor is close to zero, it means that the dependence is linear and it was unnecessary to include the regressor in the model.** The coefficient of the regressor X provides even less relevant information. If its value is close to zero, it means that the parabola has its peak (or bottom) horizontally located near 0. Values different from zero indicate a shift to the left or right (if both coefficients have the same sign, to the left, otherwise to the right). But this information is usually of no use.

4.3.1 Other non-linear relationships

In psychology, we are usually satisfied with linear relationships, rarely applying quadratic ones. However, linear models in fact allow us to model any other relationship that we are able to describe as a weighted sum of some regressors or combinations of regressors. For example, we could add an X^2 regressor to the equation along with an X^3 or even an X^4 regressor, and what is more, we can model exponential dependence, absolute value functions, goniometric functions, etc.

As an example of an interesting use of this range of possibilities, let us mention the modeling of a sine function curve. This kind of dependence is typically encountered when the regressor X denotes time and the dependent variable Y indicates the values of an event that took place at different times. The sinusoid indicates that the values of the Y variable rise and fall repeatedly, and that this pattern always has the same length corresponding to a certain *period*. Assuming that we know this period, we need to determine the values of two other parameters: the amplitude (the amount of fluctuation) and the horizontal displacement (i.e., the time at which the first observed cycle begins).

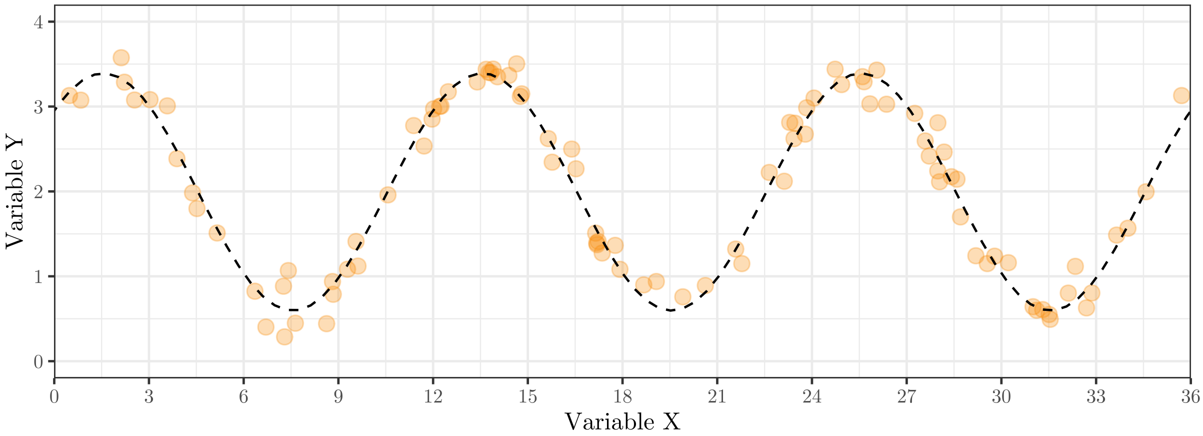
The following formula can be used to describe this kind of dependency:

$$Y = \beta_0 + \beta_1 \cdot \sin\left(\frac{2\pi}{s}X\right) + \beta_2 \cdot \cos\left(\frac{2\pi}{s}X\right)$$

where s is the length of the period – so if the values of the regressor X were defined in months and the cycle was defined by a period of one year, then $s = 12$. A suitable use of this model is presented in Figure 7.

Note that the estimated regression weights cannot be directly interpreted as amplitude and displacement. However, we can calculate these values: amplitude $\rho = \sqrt{\beta_1^2 + \beta_2^2}$ and displacement $\theta = \arcsin\frac{\beta_1}{\rho} = \arccos\frac{\beta_2}{\rho}$. The model can then be redescribed into a more understandable form $Y = \beta_0 + \rho\cos(\frac{2\pi}{s}X - \theta)$.

Figure 7: Data fitted with the sine function



5 Tests of statistical significance

Basic statistics courses have taught us that any descriptive statistic is only an estimate of the parameter of interest, not its exact value. For example, when in a previous chapter we calculated that the average reaction time of six randomly selected subjects was 600 ms, this does not mean that any random people in fact have an average reaction time of exactly 600 ms. If we were to repeat the experiment on different examinees, we might get a value of 558 or 613, for example. The arithmetic mean is a statistic – an estimator designed to approximate as accurately as possible the true value of the parameter of interest.

In linear model framework, all regression weights $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ act as estimators. Although their values vary among different groups of the subjects, we know that they fluctuate around some unknown but invariant values $\beta_0, \beta_1, ..., \beta_k$. If we had an infinitely large set of observations, our estimates would be perfectly accurate[8].

Similarly, we can imagine that the coefficient of determination R^2 is an estimate of a parameter which we could label ρ^2 and whose value would again only be known if we were working with a hypothetical infinite sample. Similarly, the residual variance S_ϵ^2 mentioned earlier is an estimate of the true, though unknown to us, value of σ_ϵ^2.

As we are used to from basic statistics courses, we can formulate hypotheses about the parameters and test their validity with statistical tests. We can also apply previously acquired knowledge about test statistics, p-values, null or alternative hypotheses in the context of linear models. The only noticeable difference is that this time we will not have to learn dozens of different statistical tests, but we will suffice with a single statistical test to evaluate any hypothesis: the submodel test.

5.1 Submodel test

A submodel is a statistical model that was created from an original model by dropping one or more regressors or by constraining the weights of the regressors. Let us now consider only the first of these possibilities, since the topic of models with constraints is not covered in this textbook. For example, consider a linear model with three regressors defined by the equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

[8] Non-standardized regression weights have a number of useful properties as estimators. They are the minimum-variance unbiased estimators of the β parameters. As the sample size increases, they reduce their variance and thus become more accurate. Therefore, we can call them weakly consistent.

We can define for example the following submodels to this model:

$$
\begin{aligned}
Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 \\
Y &= \beta_0 + \beta_1 X_1 \\
Y &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3
\end{aligned}
$$

or possibly a submodel without any regressors at all:

$$Y = \beta_0$$

For any model or submodel, we can see how closely it fits the observed data. For this purpose, we can use the coefficient of determination R^2 derived from RSS. We can expect that if we remove one or more regressors from the model, then R^2 will decrease (RSS will increase) as the model loses its accuracy. The more significant regressors were removed, the more noticeable is the decrease in R^2. Conversely, if R^2 remains unchanged after removing the regressors, this means that the removed regressors are redundant in the model – their regression weights are zero. This is the essence of the submodel test. **The submodel test evaluates the validity of the null hypothesis $\rho^2_{model} = \rho^2_{submodel}$ which is equivalent to the statement that the weights of the removed regressors were all zero.**

The test statistic that will help us decide the validity of this hypothesis can be derived from three properties that RSS has. Let us add that these properties are valid only under a few assumptions which we will address in section 7, for now let us consider them met.

- If we estimate p regression weights using n observations, then the $\frac{\text{RSS}}{\sigma^2_\epsilon}$ statistic has a chi-square distribution with $n - p$ degrees of freedom. We write $\frac{\text{RSS}}{\sigma^2_\epsilon} \sim \chi^2_{n-p}$.
- If the true value of the regression weights of the regressors we removed from the model is zero (i.e., the null hypothesis holds), then $\frac{\text{RSS}_{submodel} - \text{RSS}_{model}}{\sigma^2_\epsilon} \sim \chi^2_h$, where h is the number of parameters removed, i.e., the difference in degrees of freedom of the model and submodel.
- The statistics RSS_{model} and $\text{RSS}_{submodel} - \text{RSS}_{model}$ are independent of each other.

The Fisher probability distribution is defined as the proportion of two independent random variables with χ^2 distributions each of them divided by its degrees of freedom. Therefore, we can construct a random variable

$$F = \frac{\frac{\text{RSS}_{submodel} - \text{RSS}_{model}}{df_{submodel} - df_{model}}}{\frac{\text{RSS}_{model}}{df_{model}}}$$

where *df* denotes the number of degrees of freedom ($df = n - p$); which has under the null hypothesis Fisher distribution with degrees of freedom corresponding to the denominators of the two fractions, i.e., $df_{submodel} - df_{model}$ and df_{model}. Many readers prefer R^2 to RSC in the computation, so let us introduce an equivalent shape of the F statistic based on the coefficient of determination:

$$F = \frac{\frac{R^2_{model} - R^2_{submodel}}{df_{podmodel} - df_{model}}}{\frac{1 - R^2_{model}}{df_{model}}}$$

Of course, both versions of the formula lead to identical results. We can simplify the statistics even further by noting that the number of degrees of freedom of the model and submodel differ by the number of dropped parameters in the submodel. Thus, if we created the submodel by omitting a single parameter, then $df_{submodel} - df_{model}$ equals one, if we left out two parameters, then two, and so on. If we label the number of dropped parameters h and the difference of the coefficients of determination of the model and submodel by ΔR^2, then the F statistic for the submodel test can be written in the form:

$$F = \frac{\frac{\Delta R^2}{h}}{\frac{1 - R^2_{model}}{n-p}} \sim F_{h,\, n-p}$$

In general, we can say that **the submodel test evaluates the validity of the null hypothesis that removing one or more regressors does not decrease the percentage of explained variance of the dependent variable**.

How do we use the submodel test in practice? We typically test the following submodels:

- A submodel created by removing a single regressor with weight β_j. The null hypothesis states that the model explains the same amount of variance whether or not it contains a given regressor. This hypothesis is equivalent to the claim that a given regression coefficient is equal to zero, i.e., $H_0 : \beta_j = 0$.
- A submodel created by removing all regressors (i.e., $Y = \beta_0$). Since the submodel without regressors explains no variance, this test evaluates the validity of a hypothesis that our model explains a non-zero amount of variance at all. Thus, the null hypothesis could be defined as $H_0 : \rho^2 = 0$.
- A submodel created by removing all dummy variables belonging to one nominal regressor. This test will allow us to test a null hypothesis that the nominal regressor does not improve the accuracy of the model and therefore has no relationship with the dependent variable Y.

5.2 Wald statistic and confidence interval for regression weight

If we are asking ourselves which regressors in our model have a statistically significant effect (i.e., their regression coefficients differ from zero), we can imagine that a computer program runs through all the regressors in the model always removing a given regressor from the model, calculating R^2 for that submodel, then returning the regressor back to the model, removing another, etc. In fact, the calculation is not conducted in this way. We use the Wald statistic instead. Wald statistic provides identical p-values to the submodel test but is computationally less demanding and has several other useful properties.

To construct the Wald statistic, we first estimate the covariance matrix of the regression weights. Recall that the individual estimates of $\hat{\beta}_j$ are random variables and thus have variances. Moreover, in general they are not independent, therefore for any two coefficient estimates we can estimate their covariance. The covariance matrix contains precisely this information – it is a square matrix with estimates of the variances of each regression coefficient on the diagonal (labeled S_j^2) and estimates of the covariances off-diagonal (labeled S_{ij}):

$$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} S_0^2 & S_{0,1} & \cdots & S_{0,k} \\ S_{1,0} & S_1^2 & \cdots & S_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k,0} & S_{k,1} & \cdots & S_k^2 \end{pmatrix}$$

Estimating the covariance matrix is relatively easy, but again we have to consult matrix algebra:

$$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = S_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$$

When calculating the Wald statistic, we suffice with the diagonal elements of this matrix, i.e., the variances of the estimates of the regression coefficients. The Wald statistic T is the ratio of the estimate of the regression weight β_j to its standard deviation S_j. Under the null hypothesis $H_0 : \beta_j = 0$, the Wald statistic has a Student distribution with $n - p$ degrees of freedom:

$$T = \frac{\hat{\beta}_j}{\sqrt{S_j^2}} \sim t_{n-p}$$

When testing the significance of groups of regressors, the Wald statistic cannot be used in this shape; however, when testing individual regressors, it is the go-to test. Not surprisingly, statistical software usually provides the standard deviation (S_j) and the value of the Wald statistic t to the estimate of each regression weight.

Again, let us emphasize that the Wald statistic leads to the same results as the corresponding submodel test. We can even notice the relationship between the two statistics: $t^2 = F$. This corresponds to the relationship between the Student and Fisher distributions:

$$t^2_{n-p} = F_{1,n-p}$$

Knowing the standard deviation of the regression weight estimate (the so-called standard error) also has the advantage of allowing us to generate **confidence intervals** for individual regression weights. The estimates of the regression coefficients $\hat{\beta}_j \sim N(\beta_j, \sigma_j^2)$ and the estimate of the variance of the coefficient $S_j^2 \sim \frac{\sigma_j^2}{n-p}\chi^2_{n-p}$ are independent. We can therefore construct a confidence interval for the regression weights analogous to the confidence interval for the mean that we know from the basic courses:

$$I_{1-\alpha} = \hat{\beta}_j \pm S_j \cdot t_{n-p;1-\frac{\alpha}{2}}$$

where $t_{n-p;1-\frac{\alpha}{2}}$ is the corresponding quantile of the Student's t distribution (α is usually set to 5 %). The use of interval estimates is a good practice when reporting results, as it allows the reader to gain a clearer picture of the plausibility of the results than the p-value alone.[9]

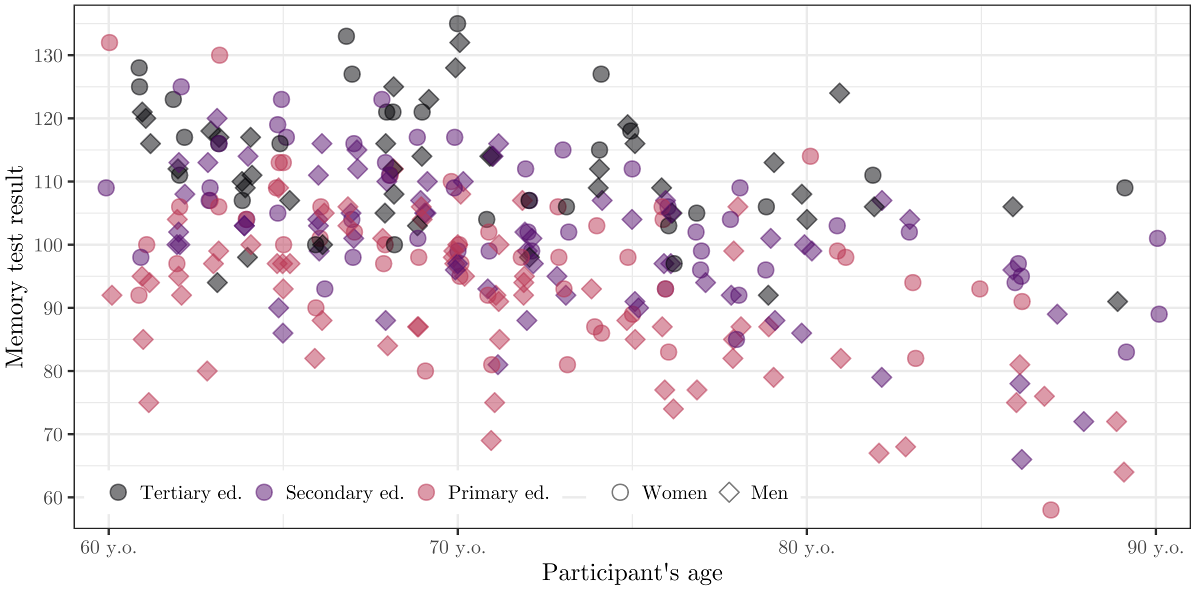
5.3 Equivalence to bivariate tests

To gain a better idea of the logic of null hypothesis tests, let us demonstrate their use on simulated data. Imagine a situation when the data obtained from a large population of elderly people is used to model the decline in memory skills over time. For each participant, we record their gender, age, highest level of their education, and memory test score. The data is shown in Figure 8. To be able to interpret the intercepts of our models, we subtract 60 from the variable *age*, so that zero corresponds to sixty years of age.

Before proceeding to test the null hypotheses in the context of a model with many regressors, let us demonstrate a fact that may be surprising to many readers: a lot of the bivariate tests (i.e., tests of pairs of variables) that we know from basic statistics courses are actually special cases of the submodel test. We will find that t-tests, analysis of variance, or Pearson correlation coefficient tests are instances of the same test if we look at them through the lens of linear models.

[9] Confidence intervals can be created not only for individual regression weights, but also for pairs or larger groups of regressors together. Then we are referring to confidence ellipses or multidimensional ellipsoids. Since we do not usually use this procedure in psychology, curious readers should consult more advanced statistical literature.

Figure 8: Visualisation of data



T-test for two independent samples

Let us suppose we are asking whether men and women in a given population segment receive equal scores on average in the memory test, ignoring the education and age aspect. Under normal circumstances, we perform a t-test for two independent samples. In our particular case, it yields a result of $t(298) = -3.708, p < 0.001$. But the same hypothesis can be tested using this model:

$$Y = \beta_0 + \beta_1 \cdot \text{gender}$$

The Wald statistic of the coefficient β_1 exactly matches the t statistic above, and neither the degrees of freedom nor the p-value change:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	104.383		t(298) = 90.561	< 0.001
gender	−5.644	−0.210	t(298) = −3.708	< 0.001
R^2	4.4 %		F(1; 298) = 13.749	< 0.001

The standardized coefficient β_1^* corresponds to the value of the point-biserial correlation coefficient between the gender variable and test performance. The effect size measure, Cohen's d, would be obtained as the proportion of the difference between the two groups (β_1) by the residual standard deviation $\hat{\sigma}_\epsilon$.

Test of Pearson correlation coefficient

Similarly to how we tested the significance of the *gender* regressor in isolation, we can test whether there is a linear relationship between the age of the examinees and their performance on the memory test. If we ignore the influence of other factors – gender and education – we would probably use Pearson's correlation coefficient and a statistical test to check whether or not the correlation is zero. We would find values of $r = -0.427, t(298) = -8.148, p < 0.001$. A test of the regressor *age* in the corresponding model leads us to exactly the same result

$$Y = \beta_0 + \beta_1 \cdot \text{age}$$

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	110.050		t(298) = 84.929	< 0.001
Age	−0.782	−0.427	t(298) = −8.148	< 0.001
R^2	18.2 %		F(1; 298) = 66.385	< 0.001

Moreover, the value of the correlation coefficient corresponds exactly to the standardized weight $\hat{\beta}_1^*$ and the coefficient of determination R^2 corresponds to its squared value.

Analysis of variance

If we were to ask whether subjects with different levels of education receive different scores on average, and we did not take gender or age into account, an one-way analysis of variance would be an adequate choice. The test would lead us to the result $F(2; 297) = 59.741$, $p < 0.001$. This test can also be implemented within the linear model

$$Y = \beta_0 + \beta_1 \cdot \text{SecEdu} + \beta_2 \cdot \text{PrimEdu}$$

where *SecEdu* and *PrimEdu* are the dummy variables of the education regressor. The analysis of variance corresponds to the test of the whole group of dummy variables associated with the variable *education*, or the significance test of the entire model. Of course, it does not matter which group we choose as the reference group.

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	112.536		t(297) = 82.856	< 0.001
SecEdu	−10.766	−0.392	t(297) = −6.246	< 0.001
PrimEdu	−18.646	−0.685	t(297) = −10.905	< 0.001
R^2	28.7 %		F(2; 297) = 59.741	< 0.001

In the context of analysis of variance, the LSD test (*Least Significant Difference*) is often mentioned among post-hoc tests. This is a test that compares all pairs of groups with each other, but unlike, for example, the Scheffé and Tukey tests, the LSD test does not include correction for multiple testing. The results of the LSD test correspond to the results of the tests of statistical significance of the individual indicator variables.

One-sample t-test

The hypothesis that the average memory test score is different from 100 could be tested using a one-sample t-test. In our case, it would lead to the result $t(299) = 1.492, p = 0.137$. Here again, the test can be performed within the linear model by comparing the value of the absolute term with a specified value of 100. This can be done using Wald statistic in a more general form

$$T = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{S_j^2}} = \frac{\hat{\beta}_j - 100}{\sqrt{S_j^2}} \sim t_{n-p}$$

However, probably the most comfortable way is to transform the dependent variable by subtracting 100 and test the hypothesis that $\beta_0 = 0$, within the model

$$(Y - 100) = \beta_0$$

In this case we again get an identical result.

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	1.147		t(299) = 1.492	0.137

5.4 Example of using null hypothesis tests

In practice, of course, we would never test every hypothesis with a separate model, as this would cost us the benefits that statistical modelling offers. The correct way is to use a single overall model which we use for each test. For example, it could look like this:

$$Y = \beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu}$$

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	123.901		t(295) = 81.034	< 0.001
Age	−0.755	−0.412	t(295) = −9.994	< 0.001
Gender	−5.995	−0.223	t(295) = −5.436	< 0.001
Education			F(2; 295) = 81.977	< 0.001
SecEdu	−9.488	−0.346	t(295) = −6.541	< 0.001
PrimEdu	−18167	−0.668	t(295) = −12.677	< 0.001
R^2	50.4 %		F(4; 295) = 74.821	< 0.001

The p-values of each regressor relate to the null hypotheses $H_0 : \beta_j = 0$. In the case of age and gender, their significance is obvious, but in the case of the indicator variables *PrimEdu* and *SecEdu*, the test refers to the null hypothesis that the given group does not differ from the reference group, which in our case is *TerEdu*. If we wanted to test the statistical significance of the difference between *PrimEdu* and *SecEdu*, the simplest way would be to select some other reference group and run the test again. The F statistic on the line *Education* is the result of a test of the two dummy variables (i.e., *PrimEdu* and *SecEdu* together).

The statistical test in the last row of the table applies to the model as a whole and evaluates whether the amount of explained variance (ρ^2) differs from zero.

Interaction tests

The model above assumes that the effect of age will be the same for men and women, and for people of any education. Thus, whichever group we are talking about, each year costs approximately three-quarters of a point on average. At the same time, however, it is true that a man of any age has, on average, 6 points lower score than a woman of the same age, and that there is a difference of about 18 points between a college graduate and a person with primary education (assuming both are the same age and gender). Does this correspond to reality? It would be quite reasonable to hypothesize that the rate of decline in memory skills (i.e., the age effect) varies across groups.

The significance of the interaction of the regressors *gender* and *age* is represented in the model by a single regressor which is the product of the two regressors. To test the significance of the interaction, we could use the corresponding Wald statistic or a submodel test, where we omit the interaction term (marked in red) from the model:

$$Y = \beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} + \beta_5 \cdot \text{gender} \cdot \text{age}$$

The regression weight $\hat{\beta}_5$ is almost exactly zero and its effect is not statistically significant, $t(294) = 0.045, p = 0.964$. Thus, we find no evidence that the decline in memory competence is differently steep for men and women.

If a nominal regressor with multiple levels enters the interaction, we test its significance using a submodel test, dropping all interaction terms. If we ask whether the rate of memory decline varies across groups by education, we will represent this with the following model and submodel:

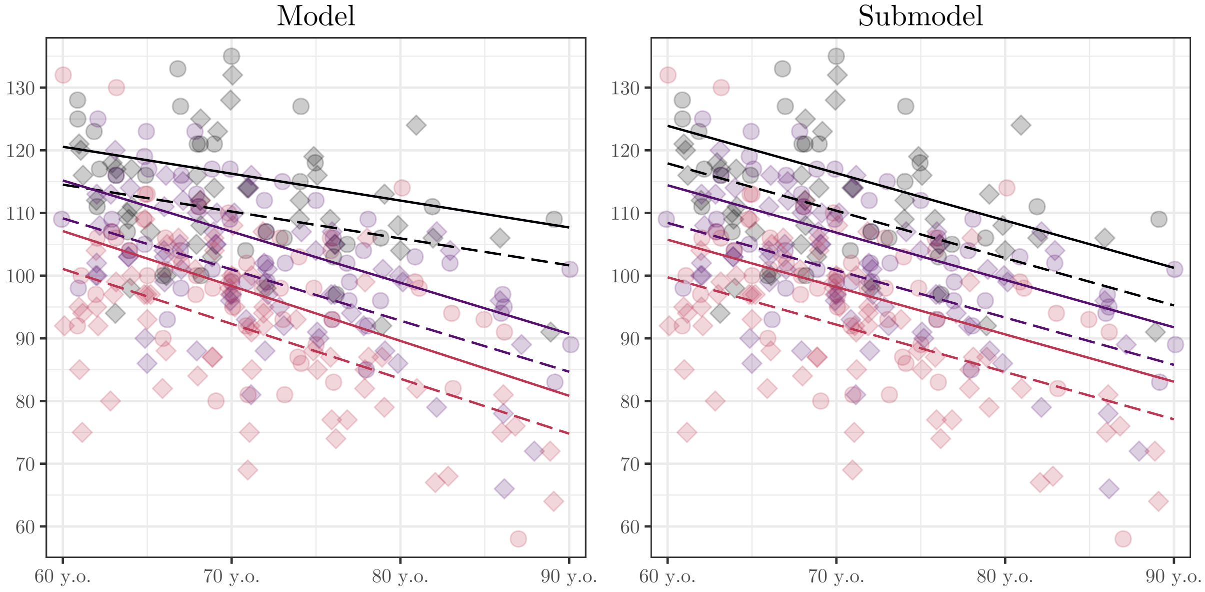
$$\begin{aligned} Y = \beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} \\ + \beta_5 \cdot \text{SecEdu} \cdot \text{age} + \beta_6 \cdot \text{PrimEdu} \cdot \text{age} \end{aligned}$$

The difference between the model and the corresponding submodel (i.e., without the terms marked in red) is shown in Figure 9. The statistical test yields the following results:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	120.560	0.000	t(293) = 57.139	< 0.001
Age	−0.429	−0.234	t(293) = −2.657	0.008
Gender	−6.040	−0.225	t(293) = −5.505	< 0.001
Education			F(2; 293) = 14.515	< 0.001
SecEdu	−5.380	−0.196	t(293) = −2.039	0.042
PrimEdu	−13.452	−0.494	t(293) = −5.218	< 0.001
Education × Age			F(2; 293) = 2.661	0.072
SecEdu × Age	−0.386	−0.218	t(293) = −1.926	0.055
PrimEdu × Age	−0.447	−0.237	t(293) = −2.212	0.028
R^2	51.2 %		F(6; 293) = 51.329	< 0.001

Although in Figure 9 we can clearly see that, at least in our sample, the decline in memory skills is slightly slower in people with tertiary education than in the other two groups, the difference is not significant ($p = 0.072$). Thus, we find no support for our hypothesis claiming that the rate of decline in memory competence is related to an individual's level of education. Somewhat paradoxical may be the partial results – the difference in the effect of age between people with a tertiary education and people with a primary education is significant ($p = 0.028$) and between people with a tertiary education and people with a secondary education is borderline statistical significant ($p = 0.055$). The reason for this is that pairwise comparisons are not corrected in any way, whereas the omnibus nominal regressor test accounts for the fact that we are examining multiple variables at once. We should therefore be cautious when interpreting the significance of the interaction term *PrimEdu* × *age* as it does not provide solid evidence.

Figure 9: Model with and without gender × age interaction



As in Figure 8, the colors distinguish the groups by education and the shape of the marker distinguishes men and women. The plotted lines then correspond to the regression lines of each group by education and gender (women are marked with a solid line, men with a dashed line). The same applies to the lines and curves in the following figures.

Testing the quadratic relationship

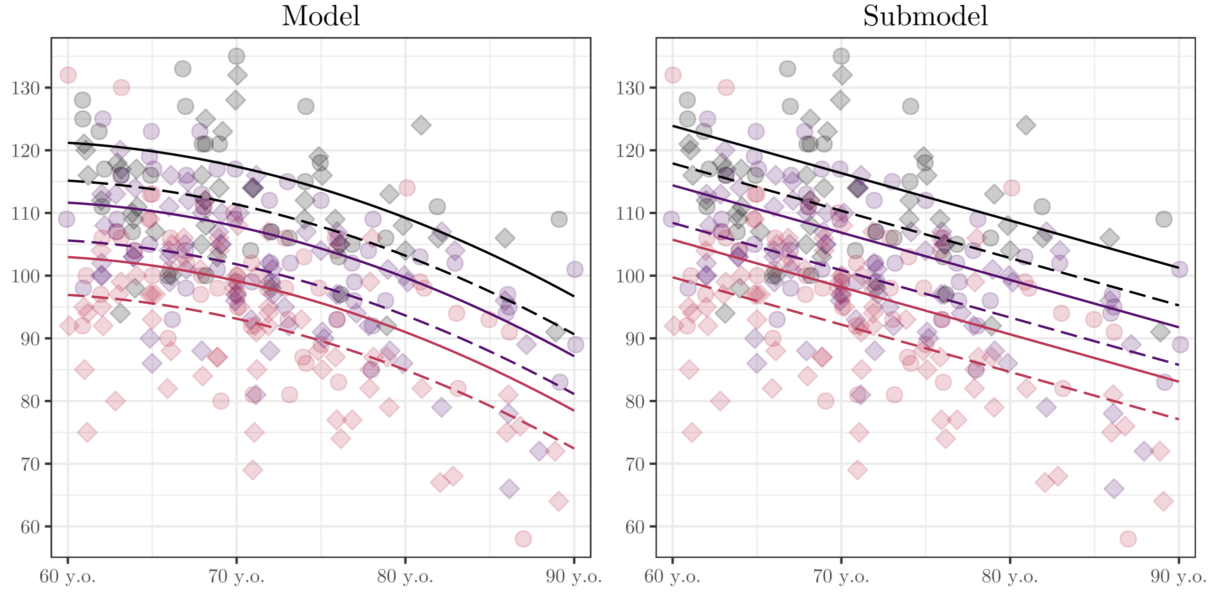
It would also be reasonable to assume that the decline does not occur at a steady rate but is accelerating over time. We can model this relationship as quadratic. We would test the hypothesis using the following model and its submodel:

$$Y = \beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} + \color{red}{\beta_5 \cdot \text{age}^2}$$

The difference between the model and the submodel is shown in Figure 10 and the statistical significance of the difference in the following table:

Regressor	$\hat{\beta}$	$\hat{\beta}^*$	Statistic	p
(intercept)	121.192		t(294) = 64.434	< 0.001
Age	−0.161	−0.088	t(294) = −0.628	0.530
Age^2	−0.022	−0.339	t(294) = −2.434	0.016
Gender	−6.039	−0.225	t(294) = −5.520	< 0.001
Education			F(2; 294) = 83.879	< 0.001
SecEdu	−9.545	−0.348	t(294) = −6.634	< 0.001
PrimEdu	−18.232	−0.670	t(294) = −12.826	< 0.001
R^2	51.3 %		F(5; 294) = 61.040	< 0.001

Figure 10: Model with quadratic and linear effect of age



Since this is a single regressor test, we can use the Wald statistic, $t(294) = -2.434$, $p = 0.016$, and conclude that our hypothesis of a quadratic effect of age can be accepted.

The interpretation of the statistical test of the *age* regressor is very unintuitive. It is not a test of the hypothesis that *age* affects memory skills, but of the rather uninteresting hypothesis that the peak of the age effect parabola is located above zero (i.e., at 60 years). We have already addressed this issue in Chapter 4.3. Although the regressor *age* forms an integral part of the model, there is no point in interpreting its value and statistical significance. If we wanted to test statistical significance of age, assuming that *age* can have a quadratic effect, we would have to create a submodel without both *age* and *age*^{2} regressors:

$$Y = \beta_0 + \beta_1 \cdot \text{gender} + \color{red}{\beta_2 \cdot \text{age}} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} + \color{red}{\beta_5 \cdot \text{age}^2}$$

The corresponding test confirms a statistically significant effect of age, $F(2; 294) = 53.739$, $p < 0.001$.

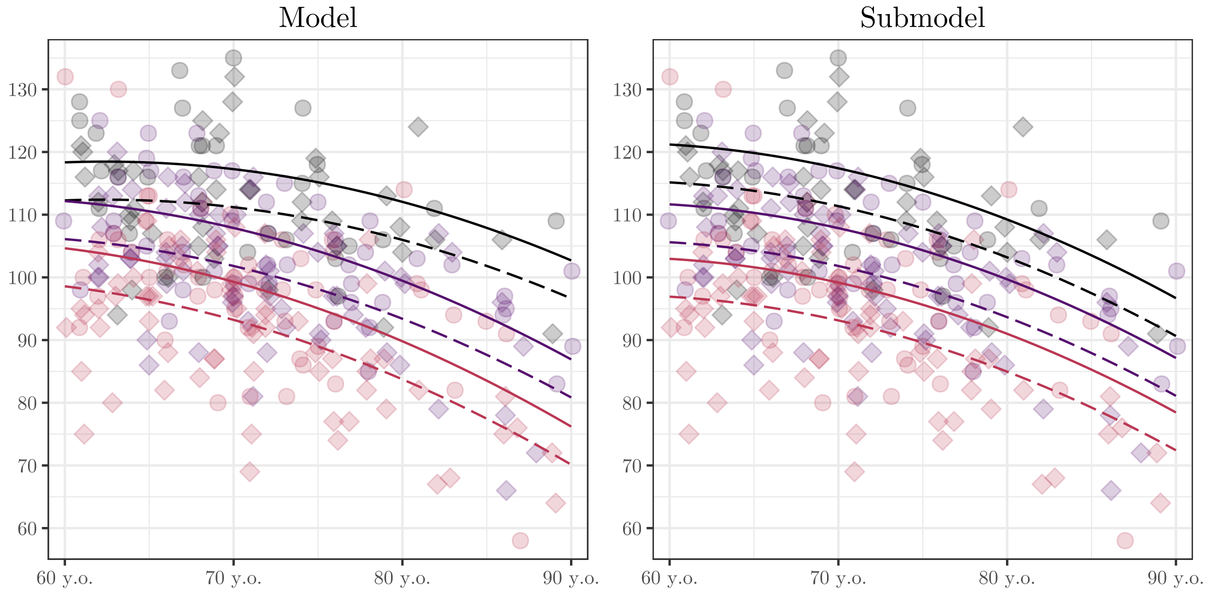
The situation becomes even more complicated if we include the quadratic effect of age in the model and assume that this effect varies for different groups by education. If we were to test this model and submodel

$$\begin{aligned} Y = \beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} \\ + \color{red}{\beta_5 \cdot \text{SecEdu} \cdot \text{age}} + \color{red}{\beta_6 \cdot \text{PrimEdu} \cdot \text{age}} + \beta_7 \cdot \text{age}^2 \end{aligned}$$

we assume that the age curve is equally curved for all groups, but in addition to shifting up and down, it can also shift left and right for individual groups. Thus, we again test the not very interesting hypothesis of whether the peaks of each parabola are localized over the same value of the x axis or not.

The result is not significant, $F(2;292) = 2.303, p = 0.102$, although the graphical representation in Figure 11 can be interpreted relatively well.

Figure 11: Model with quadratic effect of age with and without interaction term



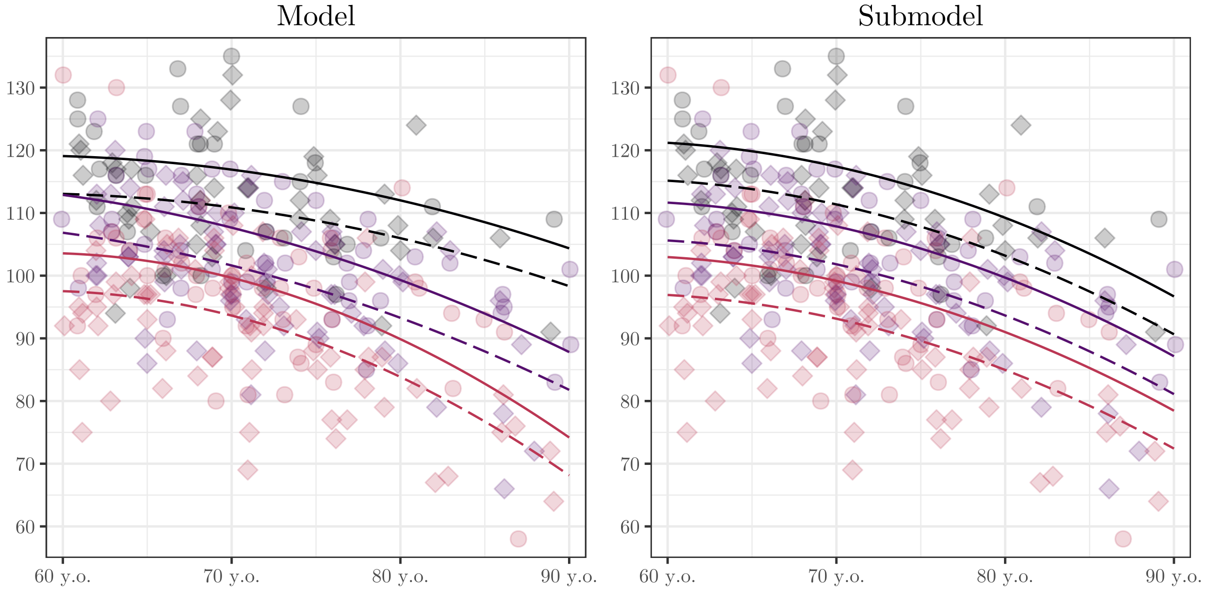
It would probably make more sense to model the dependent variable using a model with the interaction *education* × *age*, but also *education* × *age*^{2}:

$$\begin{aligned} Y = \beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} + {\color{red}\beta_5 \cdot \text{SecEdu} \cdot \text{age}} \\ + {\color{red}\beta_6 \cdot \text{PrimEdu} \cdot \text{age}} + \beta_7 \cdot \text{age}^2 + {\color{red}\beta_8 \cdot \text{SecEdu} \cdot \text{age}^2} + {\color{red}\beta_9 \cdot \text{PrimEdu} \cdot \text{age}^2} \end{aligned}$$

A test of the submodel without the red terms tests the hypothesis that the curve of memory skills decline has a different shape depending on education (see Figure 12). In this case, we also found no evidence to reject the null hypothesis, $F(4;290) = 1.300$, $p = 0.270$, which is not surprising since (as we can see by comparing Figures 11 and 12) our model provides almost the same solution as the previous one, while using four parameters to describe the role of age instead of the original two.

Even though we have not shown a differential effect of age in various groups, admitting this possibility would change the way we test the general hypothesis that the *age* regressor affects memory test performance. We would compare the above model with a submodel that lacks all regressors (either interaction terms or main effects) containing the *age* regressor (to any power):

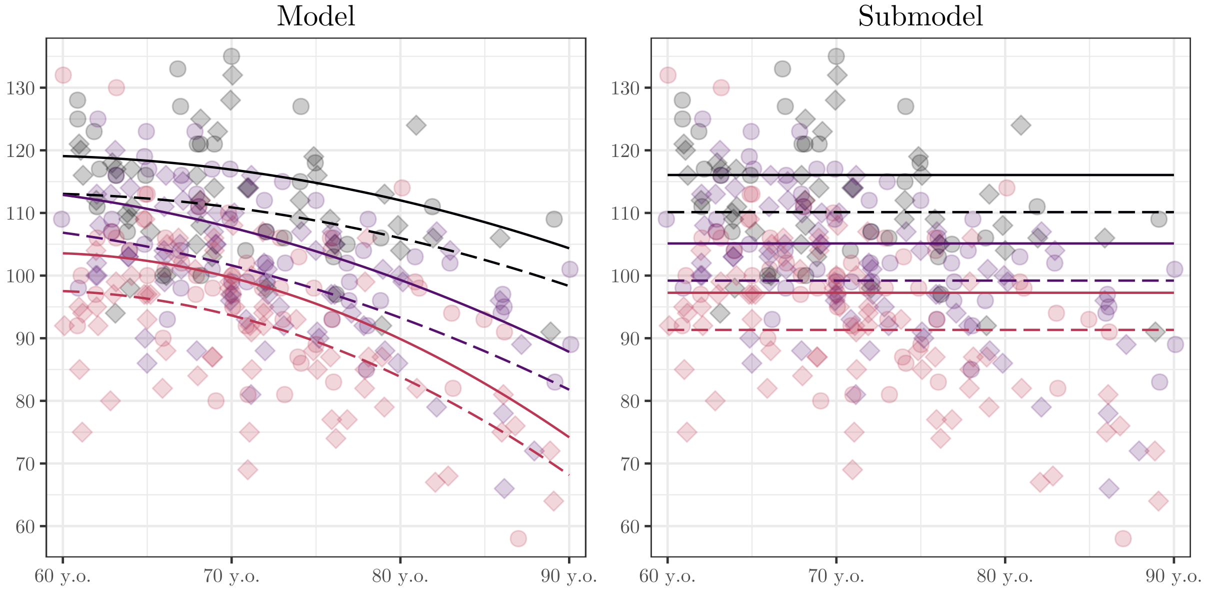
Figure 12: Model with quadratic effect of age with and without interactions



$$Y = \beta_0 + \beta_1 \cdot \text{gender} + \color{red}{\beta_2 \cdot \text{age}} + \beta_3 \cdot \text{SecEdu} + \beta_4 \cdot \text{PrimEdu} + \color{red}{\beta_5 \cdot \text{SecEdu} \cdot \text{age}}$$
$$+ \color{red}{\beta_6 \cdot \text{PrimEdu} \cdot \text{age}} + \color{red}{\beta_7 \cdot \text{age}^2} + \color{red}{\beta_8 \cdot \text{SecEdu} \cdot \text{age}^2} + \color{red}{\beta_9 \cdot \text{PrimEdu} \cdot \text{age}^2}$$

The result is again statistically significant, $F(6; 290) = 18.853, p < 0.001$. The difference between the model and the submodel can be seen in Figure 13.

Figure 13: Model with and without quadratic effect of age including interactions



6 Predictions and interval estimates

Consider once again that all the results we obtain during the calculations within the linear model are realizations of random variables. They are therefore not invariant "true" values – if we lost our data matrix and run the research again, even though we would follow all the steps and the actual relationship of the phenomena under study would remain unchanged, we would get slightly different results. We know from basic statistics courses that it is often useful to give a confidence interval instead of a point estimate (although often very precise, it is always at least slightly different from the true value), which tells us within what range it is reasonable to expect the true value of the parameter being estimated. In Chapter 5.2 we have already described the interval estimates of the individual coefficients of β. In this chapter, the possibility of constructing confidence intervals for the entire regression line as well as prediction intervals for new data points are explored.

6.1 Point-wise and simultaneous confidence bands

In Chapter 2.2 we asked how many points Otto, Agatha, Ursula or some other classmate would get on their exam if they studied for 6 hours. Let us say we use the data from eight students in Table 1 on page 20 to answer this time. We have already estimated the regression coefficients, therefore we can formulate our model as

$$Y = 22.00 + 1.462 \cdot X$$

where Y is the number of points earned and X is the number of hours spent studying. The point estimate $\hat{Y}$ assuming $X = 6$ is $22.00 + 1.462 \cdot 6 = 30.769$. Regardless of the number of regressors, we could express the prescription for $\hat{Y}$ as the product of two vectors $\hat{Y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a vector of regression weight estimates $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ and $\mathbf{x}$ is the vector of the regressor values used for the prediction. This vector always starts with 1 which belongs to the absolute term. In our case $\mathbf{x} = (1, 6)'$.

Thus, our model indicates that people who spend 6 hours studying for the exam will score about 31 points on average. In reality, however, this claim relies on point estimates of $\hat{\boldsymbol{\beta}}$ which themselves may not correspond to reality as they are subject to random variability. If we wanted to be precise, we could define a confidence interval that covers the expected value of $\hat{Y}$ for a given $\mathbf{x}$ with probability $1 - \alpha$ (typically 95 %).

This task is relatively simple, since we know that the random variable $\hat{Y}$ is the sum of the random variables $\hat{\boldsymbol{\beta}}$ multiplied by the given constants $\mathbf{x}$. We also know that the estimates of the individual coefficients of $\hat{\beta}_j$ have a normal distribution $N(\beta_j, \sigma_j^2)$ when the assumptions are satisfied, and notably that we can estimate the variances and covariances of the individual regression weights using the relation $\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}}) = S_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$. In our case

$$\widehat{\mathbf{VAR}}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = 67.282\cdot\begin{pmatrix}8 & 117\\ 117 & 2067\end{pmatrix}^{-1} = 67.282\cdot\begin{pmatrix}0.726 & -0.041\\ -0.041 & 0.003\end{pmatrix} = \begin{pmatrix}48.849 & -2.765\\ -2.765 & 0.189\end{pmatrix}$$

Therefore, we know that $\widehat{\mathbf{VAR}}(\hat{\beta}_0) = 48.849$, $\widehat{\mathbf{VAR}}(\hat{\beta}_1) = 0.189$ and $\widehat{\mathbf{COV}}(\hat{\beta}_0, \hat{\beta}_0) = -2.765$. Hence, if we are interested in the variance of $\hat{Y}$, then with knowledge from basic statistics courses we will derive the relationship:

$$\begin{aligned}\widehat{\mathbf{VAR}}(\hat{Y}) = \widehat{\mathbf{VAR}}(1\cdot\hat{\beta}_0) + \widehat{\mathbf{VAR}}(6\cdot\hat{\beta}_1) + 2\cdot\widehat{\mathbf{COV}}(1\cdot\hat{\beta}_0, 6\cdot\hat{\beta}_1) = \\ 1^2\cdot\widehat{\mathbf{VAR}}(\hat{\beta}_0) + 6^2\cdot\widehat{\mathbf{VAR}}(\hat{\beta}_1) + 1\cdot 6\cdot 2\cdot\widehat{\mathbf{COV}}(\hat{\beta}_0, \hat{\beta}_1) = \\ 1\cdot 48.849 + 36\cdot 0.189 + 12\cdot(-2.765) = 22.475\end{aligned}$$

The exact same procedure would be much easier to write in matrix formula as $\widehat{\mathbf{VAR}}(\hat{Y}) = S_\epsilon^2\cdot\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$.

Since the random variable $\hat{Y}$ is produced as a weighted sum of random variables with a normal distribution, it also has a normal distribution. Similar to the construction of the confidence interval for the expected value in the basic courses, this is a situation where we have the variance in the form of a point estimate, and we do not know the true value of the parameter. We will therefore use a t-distribution with $n-p$ degrees of freedom instead of a normal distribution, where p is the number of estimated parameters (i.e., 2 in this case).

The confidence interval for the expected value of the random variable $\hat{Y}$ for the given values of the regressors $\mathbf{x}$ can be formulated as

$$I_{1-\alpha} = \mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{n-p,(1-\frac{\alpha}{2})}\cdot S_\epsilon\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

where the expression $t_{n-p,(1-\frac{\alpha}{2})}$ denotes the corresponding quantile of the random variable with Student's t-distribution. If we plug in the values from our example and set the α level to the usual 5%, we get the following result:

$$I_{95\%} = 30.769 \pm 2.447\cdot 4.741 = 30.769 \pm 11.600 = (19.169; 42.369)$$

Now, we can say that if we have an infinitely large set of students who spent exactly 6 hours studying for the test, then their average score will be equal to some number we can expect to lie within the interval $(19.169; 42.369)$.

In practice, it might be useful to calculate the confidence interval for all possible number of hours spent studying, thus creating a confidence interval for each point on the regression line. This solution is indeed used and can be found in literature under the name **point-wise confidence band around the regression function**. However, a certain limitation of this procedure stems from the fact that although each of the individual confidence intervals created has a specified confidence level and we cannot claim that the band created by this procedure will cover entire regression line in $1 - \alpha$ cases. Once again, we run into the multiple testing problem – if each of the intervals created has a 5 % probability of not covering the value of interest, then the probability that at least one fails is of course noticeably higher.

The solution may be to use a multiple testing correction. Since we need an *infinite* number of intervals to create a confidence band, we can use Scheffé's approach which is more parsimonious than Bonferroni's approach when creating a large number of intervals. It ensures that the confidence band covers the entire regression line or curve with the required level of confidence. The principle of Scheffé's theorem will not be presented here, but the solution itself will be given. The Scheffé's confidence band is defined as

$$I_{1-\alpha} = \mathbf{x}'\hat{\boldsymbol{\beta}} \pm S_\epsilon \sqrt{p \cdot F_{p,n-p,(1-\alpha)} \cdot \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

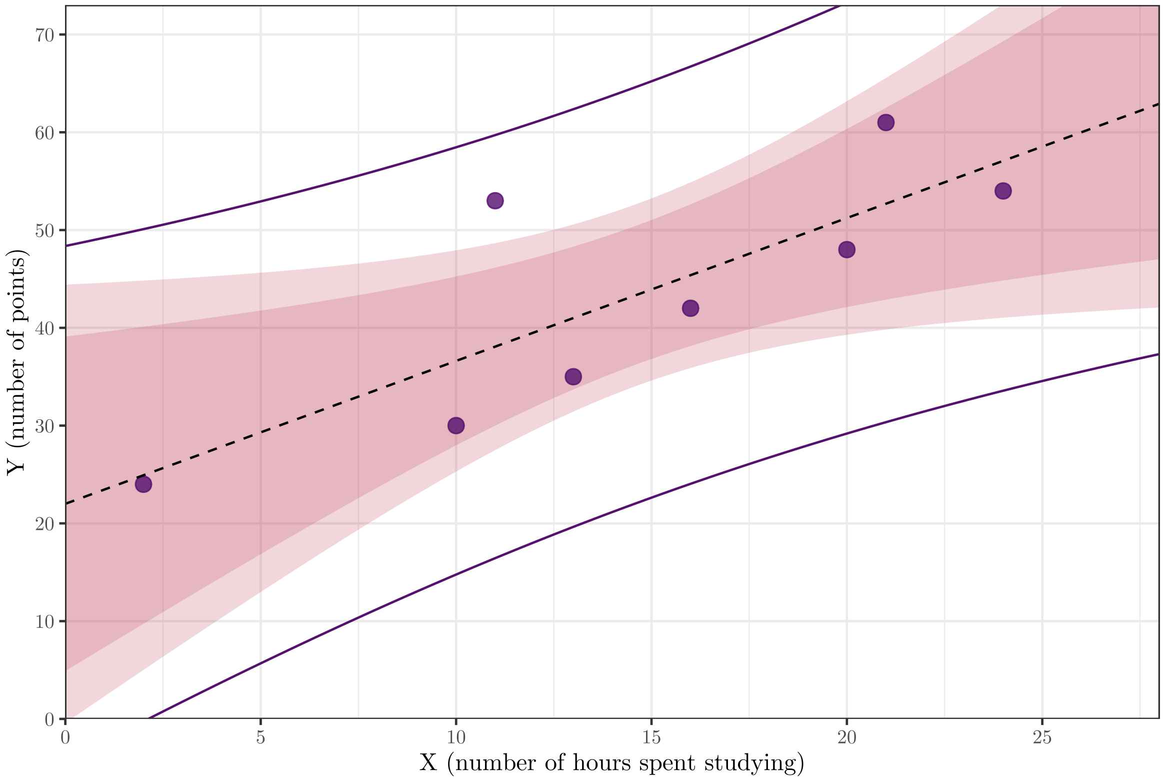
where $F_{p,n-p,(1-\alpha)}$ denotes the corresponding quantile of the random variable with Fisher distribution. We refer to this region as **simultaneous confidence band for the regression function**.

6.2 Prediction interval

In addition to finding intervals for the expected value of $\hat{Y}$, we can also create prediction intervals. A prediction interval describes the behavior of future measurements – it is the interval into which any future observation with the values of the regressors $\mathbf{x}$ will fall with probability $1 - \alpha$. We construct the prediction interval similarly as the confidence interval, keeping in mind the difference that we have two sources of imprecision instead of one in prediction. When we estimate the variance of the random variable under study, we include, as in the previous case, the variance arising from the uncertainty of the estimation of the regression weights β, but we also add the variance of the individual values around the regression line, which is characterized by the residual variance S_ϵ^2:

$$P_{1-\alpha} = \mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{n-p,(1-\frac{\alpha}{2})} \cdot S_\epsilon \sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

Figure 14: Confidence and prediction bands for the regression line



The dark area shows the point-wise band and the light area shows the simultaneous band for the regression function. The purple curves define the prediction band. The bands reach their narrowest points right above the mean value of the regressor.

If we know that the expected observations will be measured with a different accuracy than the existing ones, then we can replace the number 1 in the formula with a number that expresses how many times more or less accurate the new measurements are (more precisely, how many times more or less variance they have). Again, we can calculate an interval for any vector $\mathbf{x}$, creating a band into which any future observation will fall with a specified probability.

Returning to the question of how many points we will get if we study exactly 6 hours for the exam, the best answer is the interval $(7.59, 53.95)$, since we can expect such a score with 95 % probability. Obviously, with estimates made based on only eight observations, our model is quite unreliable. A graphical representation of the prediction bands as well as point-wise and simultaneous confidence bands for the regression line are given in Figure 14.

7 Assumptions of the linear models

From basic statistics courses we already know that every statistical test is bound by certain assumptions we make about the random variables we are describing. Tests within linear models are no exception, and the least squares method itself has certain conditions of use. If we understand the assumptions of linear models well, we will also know in detail the assumptions for calculating t-tests, analysis of variance or correlation coefficient tests, since, as we already know, these are special cases of linear models.

Traditionally five conditions of linear models are mentioned. To those we add a brief discussion of the appropriate sample size and also a sort of null condition that prescribes that the dependent variable must be a quantitative (metric) variable and the regressors must be quantitative or alternative variables (or nominal, converted to alternative dummy variables). Situations where we work with other variables can also be modelled, but then we would have to use more advanced *nonlinear* statistical models.

7.1 Correctly specified model

Each model can describe only those relationships which it is designed for. So, for example, a simple regression can only describe linear relationships, but it would fail to describe a dependency in the form of a curve (this is illustrated by Figures 5 and 6 in Chapter 4.3). If we use an inappropriate model, we will easily overlook the existing relationship, and be mistaken that the regressors in question are not related to the dependent variable.

7.2 Independence of the random component

Statistical tests are performed on observations that do not interact with each other, hence we can describe them as independent. In the context of linear models, we will refine this condition a bit more – we will discuss the independence of the random component. Let us imagine that we know the exact values of the parameters of β and that we are therefore able to draw the regression line (or curve) of the model without any errors. If we make observations, the measured values of Y will naturally fluctuate randomly around this regression line. The independence of the random component means that this variation will be independent for individual observations. Thus, for example, if we observe a value highly above the regression line in a certain participant, this provides us with no information about what result we will observe in all other participants[10].

[10] In some texts, this condition is referred to as *independence of residuals*. However, this is not entirely accurate, since the residuals of the model calculated as $Y_i - \hat{Y}_i$ are always at least weakly dependent. The reason for this is that the computation of an arbitrary residue relies on estimates of $\hat{\beta}$. However, these estimates were obtained using all other observations. Therefore, the exact value of any residual is

To give the reader a better idea of what is meant by independence condition, let us give some examples of circumstances in which this condition is violated. Imagine you measure the satisfaction rate of employees from several teams in a company. You might expect that if there is a negative mood among one team, its members will become infected by it and score lower on the satisfaction scale. As a result, the observations are not independent. To restore independence, we need to include information about who is in which team. We add a nominal variable *team* among the regressors. De facto, this expresses the assumption that each team has its own mood, which adds or subtracts some fixed number of points from each of its members' scores[11].

Another example of a violation of the independence condition could be when Ursula in the first example in this textbook actually copied from Agatha. In that case, the results of the two students are again somehow tied together. One more example: we are investigating an occurrence of certain communication patterns in men and women who are in a romantic relationship. The data set is made up of individuals, but we overlook the fact that some of the research participants are couples. Again, observations within these couples interact with each other.

Let us add that the independence of the random component is quite crucial and overlooking interrelated groups of observations will usually lead to false positive results on statistical tests.

7.3 Absence of collinearity

The term collinearity is defined as correlation between regressors. In general, regressors can be correlated, but only to some reasonable degree. The weights of highly correlated regressors (say, $|r| > 0.9$) are difficult to estimate and introduce large error because it is difficult to distinguish which of two nearly identical regressors plays a role in the statistical model. If it happens that some two regressors are perfectly correlated ($|r| = 1.0$), then least squares estimation cannot be conducted at all (the calculation would yield a singular matrix $\mathbf{X}'\mathbf{X}$, and thus inverse cannot be determined). Statistical software would probably warn us with an error message or automatically exclude some regressors from the calculation.

Collinearity itself is a fairly easy problem to detect – just look at the correlation matrix of the individual regressors and it is obvious which regressors are correlated in the model. But a much harder nut to crack is the **multicollinearity**. We talk about multicollinearity when there is a group of regressors in the model which can be used to

affected to some extent by each element of the data set.

[11] We could even use this elegant solution if we perform repeated measurements on each person. Of course, the rows of the data table belonging to the same person are dependent, but independence is restored if we insert the *proband* regressor into the model. However, this solution usually leads us to use so-called random factors (see Chapter 13 on mixed-effect models).

form a linear combination that is highly correlated with another regressor. Again, when correlation is equal to 1.0, parameters cannot be estimated. Typically, multicollinearity is encountered by students who, in preparing dummy variables, forget that they must omit one (reference) level of the nominal variable. If no dummy variable is omitted, the correlation between any dummy variable and the sum of the remaining dummy variables is -1.0, and the statistical software produces a warning instead of a result.

Multicollinearity cannot be detected by looking at the correlation matrix of the regressors, but we use a statistic called *variance inflation factor* (VIF) and *tolerance* to detect it. Tolerance can be calculated for any regressor. It takes form of a number between 0 and 1 that indicates how much unique variance (not shared with other regressors) a given regressor contains. It is again calculated using a linear model, but this time we choose the regressor under study as the dependent variable and the remaining regressors as the independent variables. The tolerance then corresponds to the unexplained variance, i.e., $1 - R^2$. The VIF is simply the inversion of the tolerance, i.e., $\frac{1}{\text{tolerance}}$.

The VIF indicates how many times the variance of the estimate of the weight of a given regressor has increased compared to a situation where that regressor would not be correlated with any other regressor. Thus, if the VIF is 4, the variance of the estimate is four times larger and the standard deviation as well as the width of the confidence interval for the parameter β is twice as large. A VIF value greater than 5 or 10 (i.e., less than 10–20% of the unique variance) is usually considered a problem. If we find a regressor in our model that violates this condition, we should consider whether some of the dependent variables describe almost the same thing and are thus redundant.

7.4 Normal distribution of residuals

The normal distribution condition appeared in all parametric tests in basic statistics courses. Not surprisingly, this condition is also present in linear models. What is perhaps surprising, however, is that even if neither the random variable Y nor any of the regressors have a normal distribution, this condition is not necessarily violated. We require a normal distribution only for the residuals of the model.

Therefore, after creating a model, it is useful to view a histogram of the residuals or their Q-Q plot (see below) and consider whether the shape differs too greatly from normal distribution. Normality tests can sometimes be helpful, but their use is nevertheless burdened with logical fallacy. If we are working with a small sample size, the normality test has little power and is unlikely to find a significant difference from normal distribution. If the sample size is in hundreds of observations, the power of the statistical test is strong and even a slight difference from a normal distribution will lead to a very low p-value. This is in stark contrast with the fact that the statistical tests we perform in linear models become robust to violations of the normal distribution condition as the sample size increases. Thus, if we are working with about a hundred or more observations, the model results may be relevant even though normality tests report serious violations of the condition.

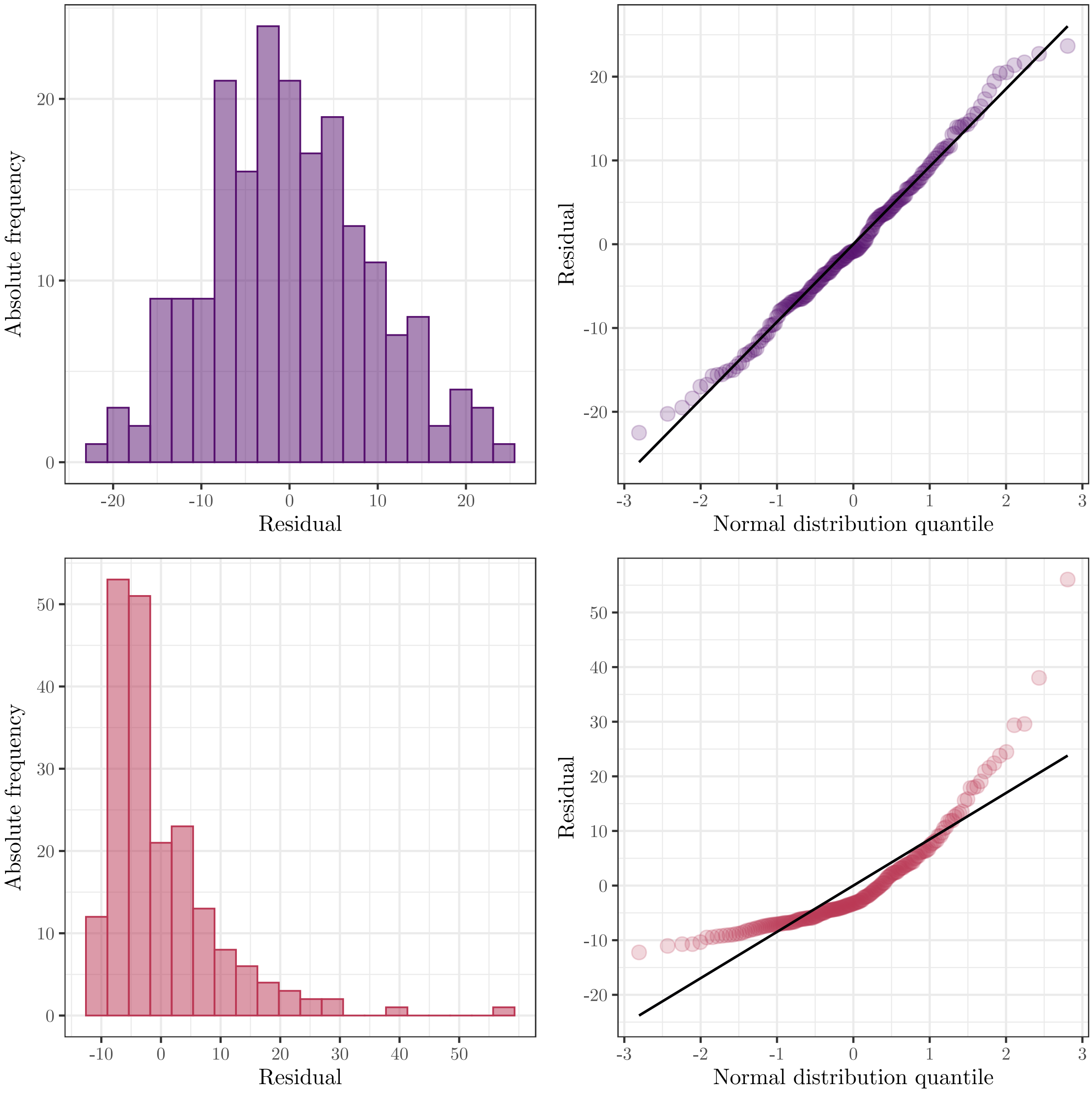
Most types of statistical software can display a histogram as well as a so-called Q-Q plot (*quantile-quantile plot*). Q-Q plot is generally used to compare probability distributions observed in two samples or to compare a sample distribution and a specified theoretical distribution. In our case, we are comparing the distribution of residuals with normal distribution. When constructing Q-Q plot, we calculate what sample quantiles correspond to each residual (thus obtaining a set of values α_1 to α_n lying between zero and one). For each of these numbers, we find the quantile value of the normalized normal distribution α_α. Finally, we create a scatter plot by plotting the residuals on one axis and the corresponding quantiles of the normal distribution ϕ_α on the other. If the distribution of the residuals resembles a normal distribution, the resulting pattern will be shaped approximately as a straight line extending from the lower left-hand corner of the graph to the upper right-hand corner. If the residuals do not follow a normal distribution, the resulting pattern will be some sort of curve. Let us add that reading Q-Q plots is not as straightforward as reading histograms and it requires more experience. For comparison, see histograms of residuals and corresponding Q-Q plots in Figure 15.

7.5 Homoscedasticity

Homoscedasticity is a term for homogeneity of residual variance. The homoscedasticity condition is satisfied when, for any regressor, the residuals have the same variance for all its levels (or values). Thus, if the regressor is gender, then we require that the residual variance is the same for men and women. If the regressor is intelligence, we require an invariant residual variance across all IQ levels from low to high. The opposite of homoscedasticity is called heteroscedasticity.

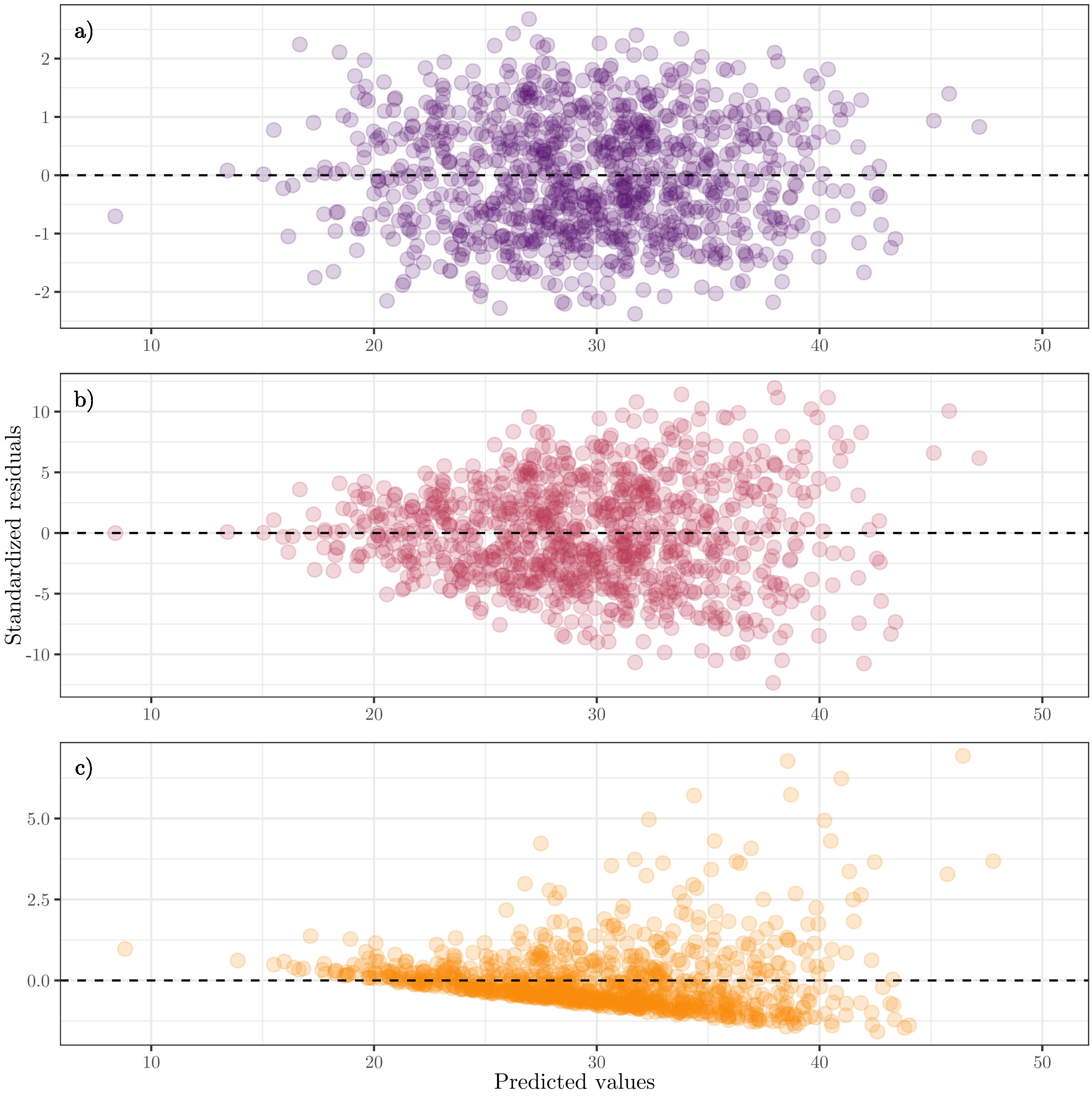
The easiest way to inspect if heteroscedasticity is present is to use a scatter plot. Plot the values of the respective regressor of each observation on the x axis and their residuals on the y axis. We would have to repeat this process for each regressor, which can be tedious for larger models, therefore we often choose the simpler approach where we plot the values of the predictions $\hat{Y}$ on the x axis and the residuals (or standardized residuals) on the y axis. Whichever approach is used, the graph should not show any meaningful shapes and should resemble figure 16a. A typical example of heteroscedasticity is figure 16b. Figure 16c is also indicative of heteroscedasticity; moreover, we can identify at a glance here that it is due to having neglected of a fundamental effect that influences the Y variable.

Figure 15: Histograms of residuals and Q-Q plots



The top two plots are based on data with a distribution close to a normal distribution. The bottom two come from a positively skewed distribution and the normality condition is therefore violated.

Figure 16: Homoscedasticity and heteroscedasticity



Unlike the normal distribution condition, the detrimental effect of heteroscedasticity does not diminish as the sample size increases. If we encounter this problem, we should reflect on whether an appropriate model was used. In addition to graphical inspection, there are statistical tests of homoscedasticity. However, their use is burdened with the same problem encountered when using the F test before selecting the appropriate t-test.

7.6 Sample size

When working with linear models, we often encounter the question about what is the lowest number of observations we need to be able to apply the learned procedures. There is no clear answer to this question, as the question itself is somewhat ambiguous.

What does adequate sample size mean? The procedures we have learned in the context of linear models are not asymptotic, and thus work for almost all small sample sizes. The least squares method can be used once the number of observations is at least as high as the number of estimated parameters. If we have more observations than there are parameters in the model, we can estimate the variance of the estimates of the regression weights, and thus perform any test of statistical significance.

However, there are two reasons why we would like to have a somewhat larger sample size than $p+1$. The first is that a small sample size is not robust to violations of residual normality condition, and a small number of data points will not even allow us to get a picture of whether or not this condition is satisfied. The second reason is statistical power. With a small sample size, statistical tests have little chance of rejecting the null hypothesis, or, if we were to calculate the confidence intervals of the regression weights, we would find that they are so wide that they communicate essentially nothing at all.

For this reason, we come across some rules of thumb discussed in literature to help us determine the number of observations needed to provide meaningful estimates. With these rules the sample size is usually derived from the number of regressors in the model (labeled k). In psychology, authors most often refer to the following recommendations[12]:

- $n \geq 104 + k$ for testing a hypothesis that R^2 is different from zero,
- $n \geq 50 + 8k$ for testing individual regressors.

Let us add that this recommendation is only a rough approximation. In addition to the number of regressors, it also depends on how correlated these regressors are (high correlation significantly weakens the tests), and in particular how large the sizes of effects we are trying to detect are. The most appropriate approach would be to use power analysis procedures, but these are difficult to implement in linear models because it is not easy to

[12] Green, S. B. (1991). How many subjects does it take to conduct a regression analysis. *Multivariate behavioral research*, *26*(3), 499–510.

estimate the correlation structure of all the included variables in advance. The proposed procedure fails for very simple models (for example, for a t-test with two independent samples it requires 105 or 58 observations, which is an unreasonably high number), but also for very complex models, where it rather underestimates the required number of observations (for example, for a model with 20 regressors, 124 or 210 observations are required, which in practice would probably not be considered a satisfactory number).

The answer is that for simple models we can work with data sets of a few dozens of observations, but we need to have a reason to believe that the assumptions are met. For very complex models, we should require a sample size of several hundred observations. For other cases, the above rules of thumb may help.

8 Problematic observations detection

To have confidence in the results we have obtained using a statistical model, we usually take several steps after estimating the parameters, commonly referred to as *model diagnostics*. These steps include examining the distribution of the residuals, verifying the absence of heteroscedasticity, and calculating the VIF for each regressor. We are already familiar with these steps from the previous section.

In addition to this, it is often useful to pay attention to the presence of observations that are atypical and stand out within our model. Usually this involves examining the presence of different kinds of outlying or influential observations. There is a myriad of statistics that we use to describe behavior of individual observations. When diagnosing a model, we generally do not utilize all of them but settle for one or two indicators that we pay attention to. We will become familiar with a few of them that are used quite frequently.

8.1 Raw residuals

One of the basic lessons of parametric statistics is that outlying observations can have a noticeable effect on the plausibility of a result. Identifying an outlier when working with one or two variables is a relatively easy task. However, in the case of more complex designs, we may encounter a situation where an observation is not an outlier in any observed feature but is nevertheless in striking conflict with the model. Conversely, there may be observations that are noticeably outlying in several features, and yet are in good agreement with the model. When looking for observations that do not agree with what the model expects, it may be useful to check for the presence of outliers in the residuals. Extremely high or extremely low values can warn us about observations that are problematic in some way (for example, we made a mistake when transcribing the results into the data matrix).

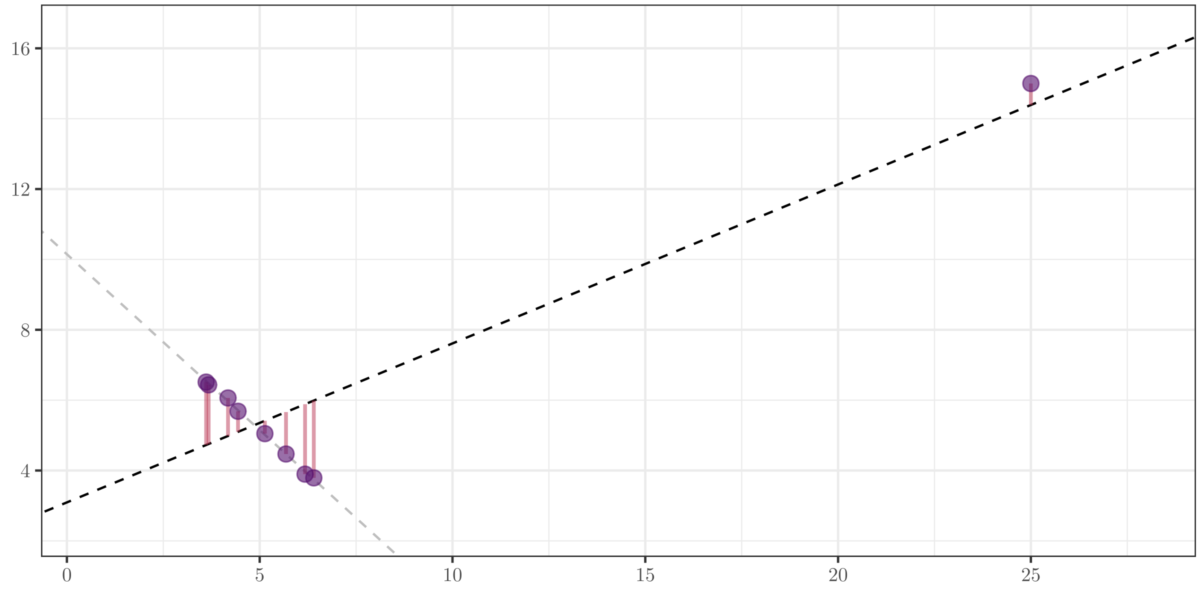
For graphical presentation, the absolute value or the square root of the absolute value is sometimes used instead of the original residuals. The one advantage of this modification is that a simple look at the highest values will suffice. Sometimes the value of the residuals divided by the estimate of the standard deviation σ_ϵ is also presented.

8.2 Deleted residuals

Despite the above-mentioned advantages, even residuals are not able to warn us about every problematic observation. Figure 17 demonstrates a situation where one outlying observation is present in a small data set and has changed the results so much that its residual is actually one of the closest to zero. Examining the residuals themselves would not help us in this case. One approach to resolve this situation is to compute *deleted residuals*. Deleted residuals are similar to raw residuals, but this time we use parameter estimates based on all data points except the one whose residual we are calculating. Thus, if we compute this statistic for the i-th observation, we firstly estimate β weights using the entire data set excluding the i-th observation, then we make a prediction for the i-th observation from the obtained weights and finally we calculate the residual.

To compare the raw residuals and the deleted residuals, we simply calculate their difference. This statistic is sometimes referred to as DFFIT in literature.

Figure 17: Influential observation



8.3 Leverages

The problem of influential observations is also solved by a statistic called **leverage**. This is a more sophisticated statistic, and to understand it we must first get acquainted with the **projection matrix**.

The projection matrix $\mathbf{H}$ (also called *hat matrix*) is a remarkable concept with a number of surprising properties[13]. Its form can be derived from two relations we already know:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad\qquad \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

[13] One of them is idempotence. If we multiply the idempotent matrix $\mathbf{H}$ by itself, then we again get the matrix $\mathbf{H}$, i.e., $\mathbf{HH} = \mathbf{H}$.

Substituting $\hat{\beta}$ from the first equation into the second equation produces the relation

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

which can be redescribed as

$$\hat{\mathbf{Y}} = \mathbf{HY} \qquad \text{where} \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The projection matrix $\mathbf{H}$ has dimension $n \times n$. As the relation above states, if we multiply the column (vector) of observations $\mathbf{Y}$ by the $\mathbf{H}$ matrix, we get the predictions $\hat{\mathbf{Y}}$. Readers familiar with the principle of matrix multiplication may notice the mechanism by which the individual predictions are computed using the $\mathbf{H}$ matrix. If we are calculating a prediction for the i-th observation, we take the i-th row of the $\mathbf{H}$ matrix and multiply its first element by Y_1, its second element by Y_2, and so on, including the i-th element which we multiply by Y_i. All the results are then summed up together. Note that the i-th element of the i-th row (i.e., any diagonal element) always indicates to what extent the value of the i-th observation (Y_i) affects its own prediction ($\hat{Y_i}$). In other words, how much a given observation is able to influence the outcome in its favor.

These values of the diagonal of the $\mathbf{H}$ matrix (we label them h_i) are referred to as *leverages*. This statistic always lies between 0 and 1 (usually much closer to zero than one), and high values indicate influential observations.

8.4 Standardized residuals

In addition to identifying the effect of observations, the h_i statistic has applications in the calculation of standardized residuals. If we wanted to standardize the residuals, we would probably consider dividing the value of each residual by the estimated standard error of the S_ϵ. This approach, while logical, overlooks one fact: although the random component of the model has the same variance across observations, the variance of individual residuals can vary considerably from observation to observation. The residuals of influential observations (typically observations with marginal values of one of the regressors) have less variance than less influential residuals. The variance of the residual ϵ_i can be estimated as

$$\widehat{\mathbf{VAR}}(\epsilon_i) = (1 - h_i)S_\epsilon^2$$

where h_i is the leverage of the observation described in the precedent paragraphs. Dividing the values of the individual residuals by the square root of the variance of their estimates yields standardized residuals. Although standardized residuals are usually not very different from residuals converted to z-scores (i.e., divided by S_ϵ), it is certainly a more accurate answer to the question of which observation outlies more than others. Standardized residuals are sometimes referred to as internally studentized residuals.

8.5 Studentized residuals

The previous discussion of residue standardization can be combined with the omission of individual observations in residue calculations. If we apply the procedure described in the section on standardizing residuals but omit the i-th element in the calculation of the i-th residual, then we refer to studentized residuals (more precisely, externally studentized).

The studentized residuals have a Student's t-distribution with $n - p - 1$ degrees of freedom, which makes them useful for performing various statistical tests.

8.6 Cook's distance

Cook's distance solves a similar problem to leverage. This time, however, we evaluate not only the extent to which a given observation affects the prediction of its own value, but also the extent to which it affects the predictions of all n observations. Thus, in addition to the predictions of $\hat{Y}_i$, we need the predictions obtained when using all observations except the first (labeled $\hat{Y}_{i(1)}$), except the second ($\hat{Y}_{i(2)}$), and so on up to the last ($\hat{Y}_{i(n)}$). We then calculate the Cook's distance for the j-th observation as follows

$$D_j = \frac{1}{pS_\epsilon^2} \sum_{i=1}^{n} \left(\hat{Y}_i - \hat{Y}_{i(j)} \right)^2$$

The Cook distance has a Fisher F distribution with p and $n - p$ degrees of freedom provided the assumptions hold.

The diagnostics of influential observations can be demonstrated using the data set of Otto, Agatha and six other classmates from the example in chapter 2.4. Figure 18 shows the leverage and Cook's distance of each observation. This and other data is summarized in the table 9. As can be seen, it is Ursula who most effectively attracts the regression line to herself, providing us with valuable information about what happens when we attempt the exam without studying. At the other extreme, then, are the three students who have studied for 20 or more hours. If we use Cook's distances to evaluate who influences the model the most, we find that Boris and Rosemarie are particularly powerful, and although they do not have a significant effect on the slope of the line, they do push it up noticeably.

Figure 18: Influential observations

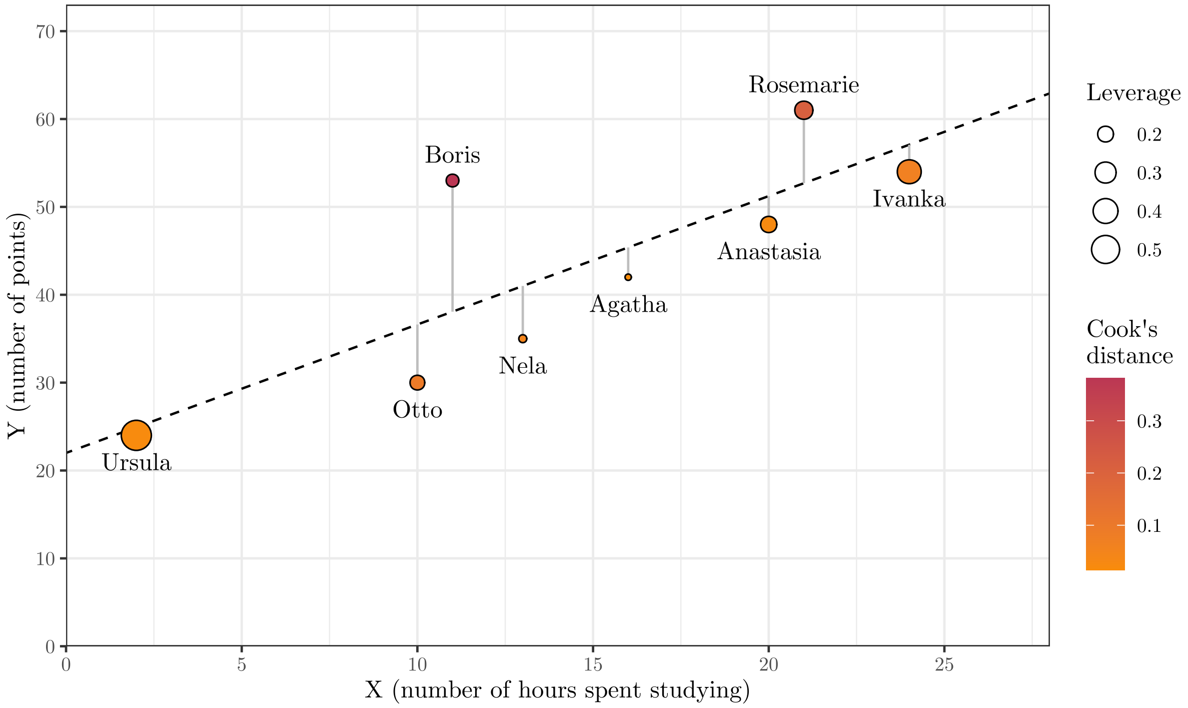


Table 9: Model diagnostics

Student	Y	X	Residual	Stand. residual	Stud. residual	Leverage	Cook's distance
Agatha	42	16	−3.38	−0.44	−0.41	0.13	0.01
Otto	30	10	−6.62	−0.89	−0.88	0.19	0.09
Ursula	24	2	−0.92	−0.17	−0.16	0.57	0.02
Boris	53	11	14.92	1.99	3.10	0.16	0.38
Ivanka	54	24	−3.08	−0.47	−0.44	0.37	0.07
Anastasia	48	20	−3.23	−0.44	−0.41	0.21	0.03
Nela	35	13	−6.00	−0.79	−0.76	0.13	0.05
Rosemarie	61	21	8.31	1.16	1.20	0.24	0.21

9 Transformations of the dependent variable

In chapter 7.4 we stated that even a major skewness of the dependent variable does not necessarily mean that the residuals of the model violate the assumption of normal distribution. This statement is true in theory, but in practice we find that in most cases when we work with a highly skewed dependent variable, this skewness will remain in residuals, and what is more, it will often be accompanied by heteroscedasticity.

If we consult a statistician about this problem, they will probably suggest leaving the world of linear models and constructing a non-linear model tailored to the particular distribution of variables. At our level of knowledge, however, we are not equipped with such tools, therefore we will have to settle for a trick that allows us to use the knowledge we already have. This trick may be a well-chosen transformation of the dependent variable Y.

By transformation, we mean that we apply a function (which is monotonic, albeit nonlinear) to the Y variable. In literature, one may come across functions such as $\sqrt{Y}$, $1/Y$ or $\log(Y)$. We then insert this adjusted dependent variable into a linear model as usual, and if we are lucky, it will already satisfy the conditions of normal distribution of residuals and homoscedasticity. If we then want to make predictions using this model, we apply our transformation inversely to the outcome (either the point estimate or the limits of the interval estimate) to return to the original units where Y was measured. For the three transformations above, these inverses are $\hat{Y}^2$, $1/\hat{Y}$ and $\exp(\hat{Y})$.

One of the main reasons why transforming the dependent variable cannot be universally recommended, and why many statisticians would label it either inelegant or even barbaric, is that most transformations make it impossible to interpret the regression coefficients we find. This is because they describe the behavior of the transformed variable, which is often far from imaginable reality (what does it mean, for example, to claim that *the square root of the number of symptoms increases by half a point*?). Often we are left with no choice but to abandon a precise description of the relationships between the variables and simply state statistical significance, supplemented where appropriate by standardized regression coefficients to illustrate the effect size.

9.1 Log-normal regression

There are transformations of the dependent variable that do not suffer from the above-mentioned flaw. It does change the way we think about the model in some way, but we do not lose the ability to interpret the coefficients obtained. Probably the most popular of these models with a transformed dependent variable is the log-normal regression.

A log-normal regression is a linear model that works with a logarithm of the dependent variable[14]:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

While the above prescription allows us to use the least squares method to obtain estimates of the parameters β, it does not indicate how to think about these numbers. Following familiar conventions, we could reason that β_1 indicates by how many points the logarithm of the variable Y increases when the regressor X_1 increases by one point. This is a true but utterly unsatisfying statement. To uncover the true meaning of the model, let us recall the rules of calculating logarithms and exponential functions and let us rewrite the model equation to a slightly different form.

Let us start with the first step, where we exponentially transform both sides of the equation. Remember that the exponential and logarithmic transformations are opposites (so-called inverse functions), and thus $\exp(\log(Y)) = Y$. Thus:

$$Y = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon)$$

We also know that the relation $\exp(A + B) = \exp(A) \cdot \exp(B)$ holds. Note that the plus sign changes to the times sign. We therefore rewrite our model as

$$Y = \exp(\beta_0) \cdot \exp(\beta_1 X_1) \cdot \exp(\beta_2 X_2) \cdot \ldots \cdot \exp(\beta_k X_k) \cdot \exp(\epsilon)$$

Finally, we apply the fact that $\exp(AB) = \exp(A)^B$:

$$Y = \exp(\beta_0) \cdot \exp(\beta_1)^{X_1} \cdot \exp(\beta_2)^{X_2} \cdot \ldots \cdot \exp(\beta_k)^{X_k} \cdot \exp(\epsilon)$$

Before we consider what this equation reveals, let us note that our model is no longer additive but multiplicative. Thus, wherever we used to ask *more by how much*, we will now ask *how many times more*. Also, in our reasoning we will no longer work with the estimates $\hat{\beta}$ themselves, but with their exponentially transformed forms $\exp(\hat{\beta})$. However, this is not a problem, since we have already estimated the values of $\hat{\beta}$ in the usual way and we can easily perform exponential transformation using a calculator or spreadsheet editor. Therefore, we can regard $\exp(\hat{\beta})$ as a specific number we know.

[14] By log in this book we mean the natural logarithm, i.e., a logarithm with base $e \approx 2.718$. If we decide to choose a different base, for example 10, we must replace the Euler number e with it in all calculations.

The interpretation of $\exp(\hat{\beta}_0)$ is analogous to the interpretation of the absolute term in the classical linear model. It is the **predicted value of the Y variable when all regressors X_1 to X_k have zero values**. This follows from the fact that any number raised to the power of is equal to one and $\exp(\hat{\beta}_0) \cdot 1$ is again equal to $\exp(\hat{\beta}_0)$. The estimates $\exp(\hat{\beta}_1)$ to $\exp(\hat{\beta}_k)$ then indicate **how many times on average the value of $\hat{Y}$ increases when the value of the regressor increases by one**.

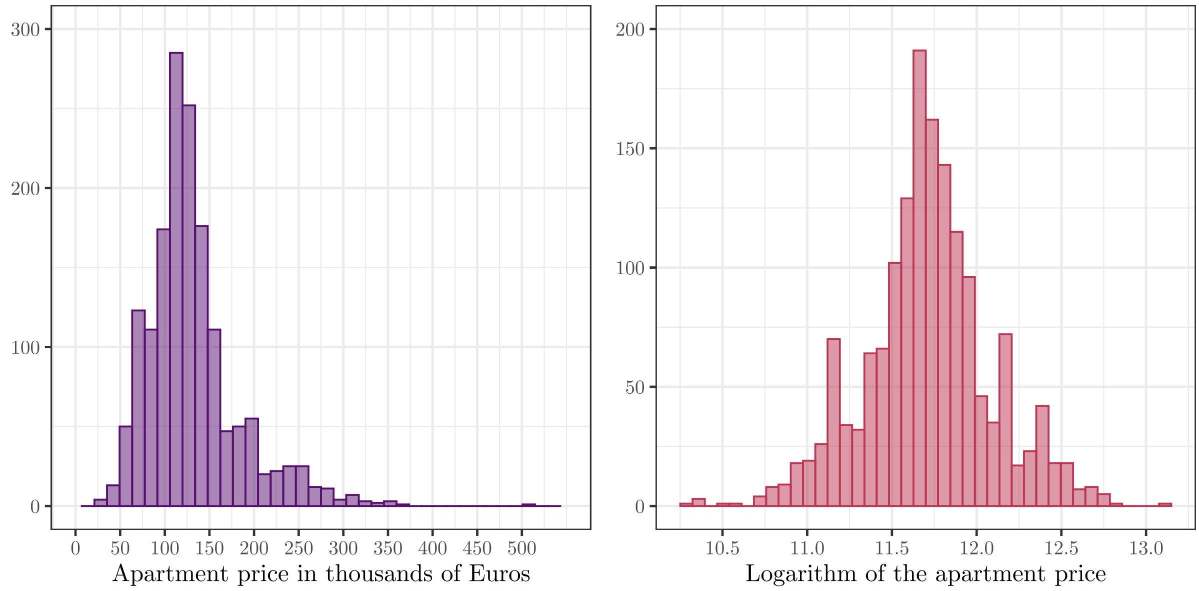
Replacing *more by how much* with *how many times more* can be confusing at first. Let us therefore demonstrate the described procedure with an example. Let us choose a topic outside the field of psychology this time.

Let us imagine that we bought an apartment in Olomouc, Czech Republic, in the summer of 2019. We can ask ourselves whether it was a good deal compared to different apartments in the same area, or whether we paid the seller more than what would be the normal price for the apartment. For this purpose, we extracted all the offers for sale of apartments in Olomouc from the website of the largest real estate seller that were available from July to November. In total, we obtained nearly 1,600 records.

The price for an apartment is determined by many factors. For the sake of simplicity, let us focus on the following: floor area, number of rooms, having separate kitchen, energy class, condition of the property and having a balcony or a terrace.

The example could be solved as an example of classical regression or as a log-normal regression. The histogram of the price variable (see Figure 19) indicates that the dependent variable does exhibit positive skewness, but not so substantial as it would invalidate the results of the statistical tests given the large sample size. Both variants of how to model the situation are justifiable. In making our decision, let us consider which of these statements makes more sense: *You pay on average €5,600 for a balcony* or *Balcony increases the price of an apartment by 5 % on average.* Let us lean towards the second option (although the first one may also make sense) and choose a log-normal regression.

Figure 19: Histogram of the apartment price and its logarithm



The values shown are actual records from a real estate website, not simulated data.

The dependent variable will be *log*(*price*). We will use the regressors as usual, but for clarity we will reduce the regressor *apartment size* by 60 and divide it by 10. The absolute term will correspond to the price of a 60-meter apartment and the new apartment size regressor will quantify how many tens of meters the size of the apartment differs from 60 meters. As a reference group, we choose a two-room open plan apartment (no separate kitchen), medium energy consumption, in good condition, without a balcony. The estimates of the regression coefficients are reported in Table 10.

The results show that if we consider a two-room open plan apartment with an area of 60 m^2, with medium energy consumption, in good condition, and without a balcony, then we can expect a price of approximately €88,485.

If we bought a three-bedroom open plan apartment of 80 m^2 in a new building that is energy highly efficient and has a balcony, then we would proceed as follows. Consider the starting price $\exp(\hat{\beta}_0)$. Assuming it is a three-bedroom apartment, let us multiply this amount by 1.17. For being energy efficient, multiply by 1.06 and for being a new building by 1.27. Furthermore, the fact that the apartment has a balcony raises the price by 5%, so we multiply it by 1.05. The last adjustment is including the floor area. Our apartment is 2 tens of meters larger than the 60 m^2 initially set. We have to multiply the price by 1.07 for every 10 extra meters. We will therefore multiply the price by $1.07^2 = 1.14$. If the apartment described was only 30 m^2, then we would multiply it by $1.07^{-3} = 1/1.07^3 = 0.81$.

Table 10: Coefficient estimates of the log-normal model

Regressor	Effect	$exp(\hat{\beta})$	$\hat{\beta}$	$t(1575)$	p-value
Intercept		€88,485	11.39	574.36	< 0.001
Floor area	7 %	1.07	0.07	25.46	< 0.001
Separate kitchen	−12 %	0.88	−0.13	−7.02	< 0.001
Has a balcony / terrace	5 %	1.05	0.05	4.80	< 0.001
1 room	−22 %	0.78	−0.24	−13.92	< 0.001
2 rooms	0 %				(ref)
3 rooms	17 %	1.17	0.16	12.78	< 0.001
4 or more rooms	18 %	1.18	0.17	6.81	< 0.001
Highly energy-efficient (A–B)	6 %	1.06	0.05	2.19	0.029
Moderately energy-efficient (C–F)	0 %				(ref)
Energy-inefficient (G)	−2 %	0.98	−0.02	−0.98	0.326
New building / project	27 %	1.27	0.24	7.68	< 0.001
Very good / renovated	18 %	1.18	0.16	9.30	< 0.001
Good	0 %				(ref)
Reconstruction needed	−5 %	0.95	−0.05	−1.38	0.168

The values in the *effect* column were computed as $(exp(\beta) - 1) \cdot 100\ \%$ and tell the percentage increase in price if the value of the regressor increases by one. The results also include the reference levels of the nominal regressors for clarity. It may seem like a paradox that apartments with separate kitchens are cheaper than open plans. This really is not an error, but a trend known from many areas in the country.

Like so the procedure is similar to what we already know, except that where we used a plus sign before, we use multiplication, and where we used multiplication, we compute power. If you find this procedure unintuitive, you can compute the prediction $log(\hat{Y})$ directly from the parameters $\hat{\beta}$ as we are used to, and then exponentially transform the result itself. Both ways will lead us to the same result.

The expected value of our 80 m^2 three-bedroom open plan apartment comes out to €$166,399$. If we are not satisfied with the point estimate, we can calculate a prediction interval. We can do this simply by calculating the prediction interval for $\log(\hat{Y})$ and exponentially transforming both bounds. For example, if we look for an 80% prediction interval, we find the values (€$132,198$, €$209,447$) after the transformation.

Before we start thinking about whether we have made a profit or a loss, let us think about one more feature of log-normal models. Our prediction is designed to capture the mean value of the variable $\log(\hat{Y})$ as accurately as possible. When we transform it, our estimate will no longer belong to the mean value (i.e., the average price), but to the median value. We can sometimes be satisfied with median value, but if we want to say how an apartment with given parameters costs *on average*, then we will have to adjust the result. It can be shown that the adjustment that turns the median value into an estimate of the expected value takes the form $\exp(\log(\hat{Y}) + \frac{S_\epsilon^2}{2})$, where S_ϵ^2 is the estimate of the residual variance of the log-normal model.

If in our case we estimated S_ϵ^2 equal to 0.032, then a fair prediction of the mean value is equal to €$169,091$. We adjust the bounds of the 80% prediction interval by analogy to (€$134,337$, €$212,836$). If the apartment we bought costed, say, 130 thousand euros, then it may warm our hearts that we bought one of the 10 % most undervalued apartments.

10 Stepwise and hierarchical regression

In previous chapters, we assumed that statistical models are built all at once – that is, we know exactly which regressors we want to include, and then we search for their weights. In this chapter, we will introduce two approaches that, although quite different from each other, have in common that they build a statistical model in steps.

10.1 Stepwise regression

Imagine a situation where we have a large number of regressors, some of which play an essential role in the model and other that do not improve our predictions at all. If we want to use our model to make predictions rather than to explore relationships between variables, we will appreciate a simple but accurate model that contains only those regressors that are useful to us. Stepwise regression can lead us to this goal. Stepwise regression can be implemented in two ways: backward selection and forward elimination.

When we use **backward elimination**, we start with the original model that contains all the regressors we have available. We then look for the regressor that has the lowest weight in the model and ask if including it was necessary. If we conclude that a given regressor does not really bring any improvement in accuracy, we will remove it from the model. We then recheck the model and again look for a regressor that could be excluded. We repeat the process until there is no regressor in the model that we have deemed redundant.

In contrast, **forward selection** starts with a model without any regressors and investigates which of the available independent variables would most improve the predictive ability of the model. If there is a variable that provides sufficient improvement, we add it to the model. We repeat the process again until we find that no other variable that could be added to the model provides further improvement.

Several criteria can be used to assess which variable provides sufficient refinement. In statistical software, this is most often the value of the F statistic of a test comparing the accuracy of a model that includes a given regressor with a submodel that does not include that regressor. If a low value is set, we will retain even those regressors that are not statistically significant in the model, while only regressors with very small p-values retain high values.

At first glance, it might seem that the backward and forward methods lead to the same result. In fact, this may not always be true. There are situations when we have multiple regressors, each of which on its own yields only a small improvement in accuracy. However, if they are in the model together, even though we have not added their interaction, they bring noticeable improvement in accuracy. Such a group would not be included in the model using the forward selection, as it adds regressors one at a time and none of the candidates would meet the entry criterion. On the other hand, the backward elimination method would retain the entire group in the model, since removing any regressor would lead to a drop in accuracy.

Stepwise regression as a model building tool is welcome in some situations, but many statisticians rightfully criticize it. The problem is that in our search for the most effective regressors, we sift through all possible candidates and easily confuse a good regressor with a false positive finding – that is, a regressor that does not actually provide noticeable improvement in accuracy but appears to do so in our sample by a stroke of luck. In conventional regressions, we would probably be wary of a situation where, say, with a total of 50 regressors, we identify three as statistically significant, since at a five percent significance level, under the null hypothesis there is on average one false positive for every twenty statistical tests. We would also probably be warned by the value of $R^2_{adj.}$, which would probably be strikingly different from the unadjusted R^2. However, if we use stepwise regression, we obtain an elegant model with three regressors and their significance is not likely to be questioned.

If, despite this tricky feature of stepwise regression, you want to use this procedure, read chapter 12 about the risks of overfitting models and ways to counteract them. In fact, cross-validation is a good remedy for all the problems mentioned.

10.2 Hierarchical regression

Similar to forward stepwise regression, in hierarchical regression we add regressors to the model sequentially. In contrast to stepwise regression, which is so-called *data driven*, hierarchical regression is *theory driven.* When deciding the order in which to add regressors in a hierarchical regression, we do not look at their statistical significance, but our decision is rationally determined by the underlying theory. Typically, we divide the regressors into several groups, which we add to the model in a predetermined order. At each step, we then examine the statistical significance of the increment of variance explained (labeled ΔR^2).

In one study, for example, we asked whether creative success (i.e., the quantity and quality of creative achievements an individual has made in their lifetime) is influenced by two lesser-known traits called *systemizing* and *empathizing* (see Simon Baron-Cohen's theory for details). In the model with the dependent variable *creative achievement*, we included the gender and especially the age of the respondents in addition to the systemizing and empathy scales, as both of these variables can play a significant role. However, we could also ask whether the possible effect of systemizing and empathizing is not simply a result of the fact that these traits correlate with the already well-known Big Five personality dimensions, which have already been shown to correlate with creativity, and thus whether it is not completely unnecessary to include systemizing and empathizing in our considerations. A hierarchical regression could take the following form.

In the first step, we compare the model without regressors and the model that contains only gender and age regressors. If the difference in model accuracy is significant, then we can say that these basic biological characteristics are related to creativity. In a second step, we compare the model with the regressors of gender and age with a model to which we add the general Big Five personality dimensions. Again, we ask whether there is a statistically significant improvement in the predictive ability of the model. If so, we can conclude that general personality dimensions are related to an individual's creative success. In addition, we would state what percentage of variance beyond the basic biological characteristics the personality profile explains (i.e., ΔR^2). The last step is the most interesting from our perspective. We compare a model containing the regressors age, gender, and the Big Five dimensions with a model to which we add the variables of interest – systemizing and empathizing. We investigate statistical significance and calculate how much the percentage of explained variance has increased this time. The hypothesis we are testing claims that empathizing and systemizing provide some relevant information about an individual's creativity beyond what we already know due to their gender, age, and general personality dimensions. In doing so, we test the incremental validity of the construct of systemizing and empathizing. If they failed this test and no longer contributed new information to the model, it would mean that these two variables are not relevant in creativity research given what we are able to describe using already well-known constructs.

In most cases, we are satisfied with just two steps, where the first group of regressors are those who we already know and are not of interest, while the second are those whose influence we want to explore. If there is a single regressor in the second group, we do not need to use hierarchical regression, as the submodel test would be identical to the Wald test. Hierarchical regression is considered to be an thorough approach when managing data.

11 Removing the influence of variables

Consider the age-old question of the extent to which human intelligence is affected by heredity. If we try to investigate this problem empirically, a naive approach will lead us to a simple research design which includes testing children of a certain age and their parents by the means of an IQ test. We would naturally find a close relationship between the two variables. But is this the answer to our question? It is not, since in addition to passing on some biological inheritance to their children (mainly through genes), parents also expose them to stimuli that can vary significantly across families and probably affect the intelligence of the child. There is a solution that would be methodologically (and ethically) problematic – to force parents to buy their children the same number of toys and books, enroll them in the same number of extracurricular activities, put them in the same schools, and provide them with the same amount of space in their rooms, etc. It would be much easier to grasp this task mathematically.

To do this, it would be necessary to map the various factors that may be involved in the formation of intelligence for each child, in addition to their IQ and the IQ of their parents, and then to remove the impact of environmental factors from the child's IQ.

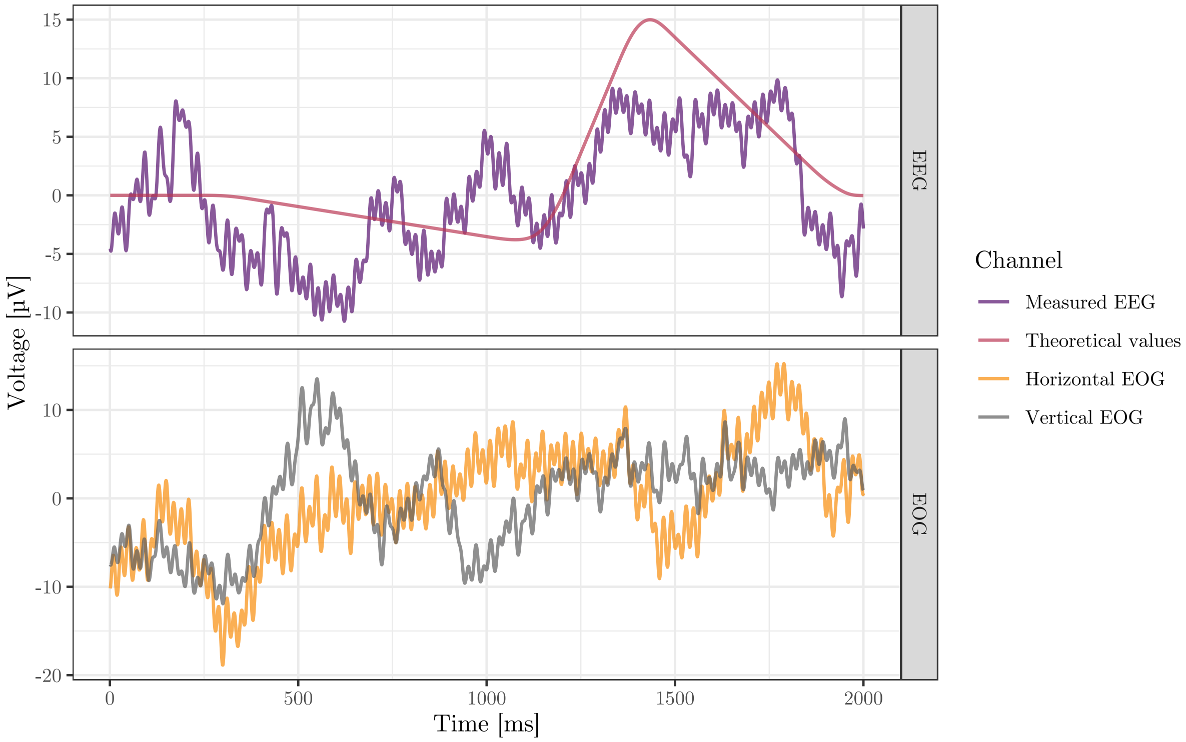
This removal is done, not surprisingly, using a linear model. We introduce the variable to be cleaned as Y and all the factors whose influence we want to control for as regressors X. Perhaps surprisingly, the new adjusted IQ values for children are the residuals of this model, since the residuals tell us how much better or worse a child does on a test than we would expect based on the environment which they grew up in.

In addition to saving the residuals thus obtained for any further calculations, we can make one cosmetic adjustment. The residuals always have a mean of 0, which is often (for example in the case of IQ) not a very meaningful value. Therefore, we can add the average value of Y (or some other meaningful value, such as 100) to each saved residual, which returns the newly obtained variable to its original level.

Another example of removing the influence of confounding variables is working with physiological data. Imagine a situation where you use an electroencephalography (EEG) to measure brain activity of a participant who is exposed to a certain stimulus. The EEG responds in a very non-specific way and instead of the studied potential we can easily measure the changes in voltage caused by, for example, motor activity. A typical source of artefacts is the activity of the oculomotor muscles, whose traces are particularly noticeable when the task does not allow the use of a fixation cross (for example, when driving a car). A simple solution is again provided by the linear model[15].

[15] If we delve deeper into this problem, we find that in practice there are a number of much more advanced and effective approaches that address this problem.

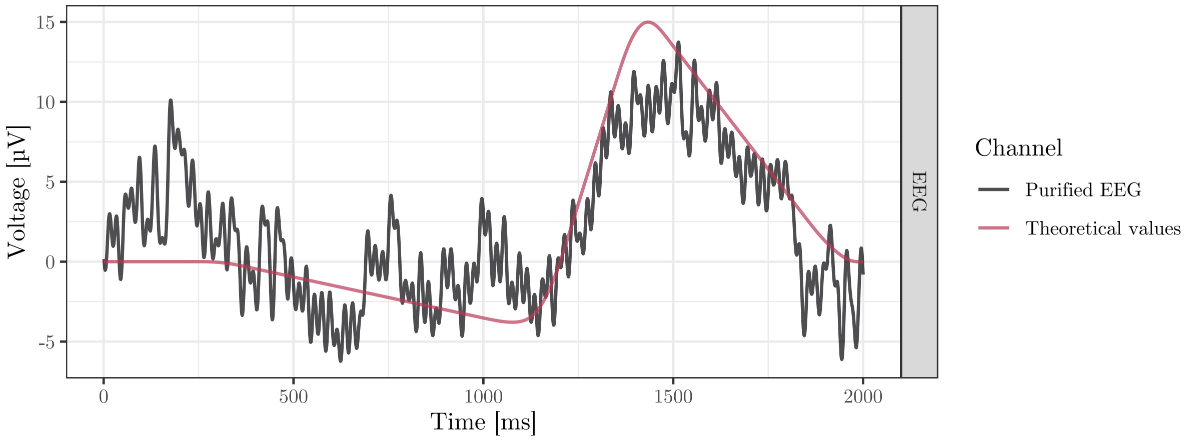
Figure 20: EEG and EOG signal



The experiment would involve an electrooculography (EOG) in addition to the EEG electrodes. If we use two pairs of EOG electrodes, we can track vertical and horizontal eye movements separately. All the measured data is illustrated in Figure 20. In addition to the data from one EEG channel, it also shows the pattern we expect to observe. We ask to what extent the EEG signal replicates our expectations. We already observe a fairly good agreement – the correlation coefficient between the EEG and the theoretical values is 0.69. At first glance, however, we see that there is a number of other artifacts present in the measured signals that may be caused by eye movement.

The EOG signals can be removed from the original EEG using a linear model. It will take the form of $\text{EEG} = \beta_0 + \beta_1 \text{EOG}_h + \beta_2 \text{EOG}_v$ and in our case it will be able to explain (remove) 24 % of the EEG variance. The residuals from this model are the EOG-adjusted EEG signal. For ease of graphical presentation, we can add some constant to these residuals (which by definition have an average of 0), such as the original average or the average of our expectation (this situation is illustrated in Figure 21). The correlation coefficient has increased to 0.82, giving even stronger support to our hypothesis.

Figure 21: EEG signal after removing the influence of EOG



11.1 Partial and semipartial correlation coefficient

We usually use the adjusted values for further calculation which introduces two new concepts. We use the term *partial correlation coefficient* to refer to the Pearson correlation coefficient calculated between two variables when we have removed the effect of a set of one or more variables from both variables before the calculation. If we had done this adjustment on only one of the two variables, then we speak of the *semipartial correlation coefficient* (which we also calculated in the EEG example above).

Statistical software usually offers the calculation of the partial correlation coefficient as a separate function and there is no need to create a linear model and manually perform all the intermediate steps. However, if for some reason we need to perform the whole procedure manually, we should take into account a small change related to the test of the statistical significance of this coefficient. The formula for calculating the test statistic T will change to the form

$$T = \frac{R_{YZ.\mathbf{X}}}{\sqrt{1 - R_{YZ.\mathbf{X}}}} \sqrt{n-k-2} \sim t_{n-k-2}$$

where k denotes the number of factors whose influence was removed (in the EEG example, $k = 2$), and $R_{YZ.\mathbf{X}}$ is the partial correlation coefficient between the variables Y and Z with the influence of the variables $\mathbf{X} = (X_1, X_2, ..., X_k)$ removed[16].

[16] Many types of statistical software, however, ignore this small difference and use the usual formula for testing significance of the Pearson correlation coefficient. Let us also add that in the specific example of EEG, the above statistical test would not be accurate, since it is not a set of independent observations, but a *time series*. It is doubtful in this context that we can assume independence of the random components of individual observations.

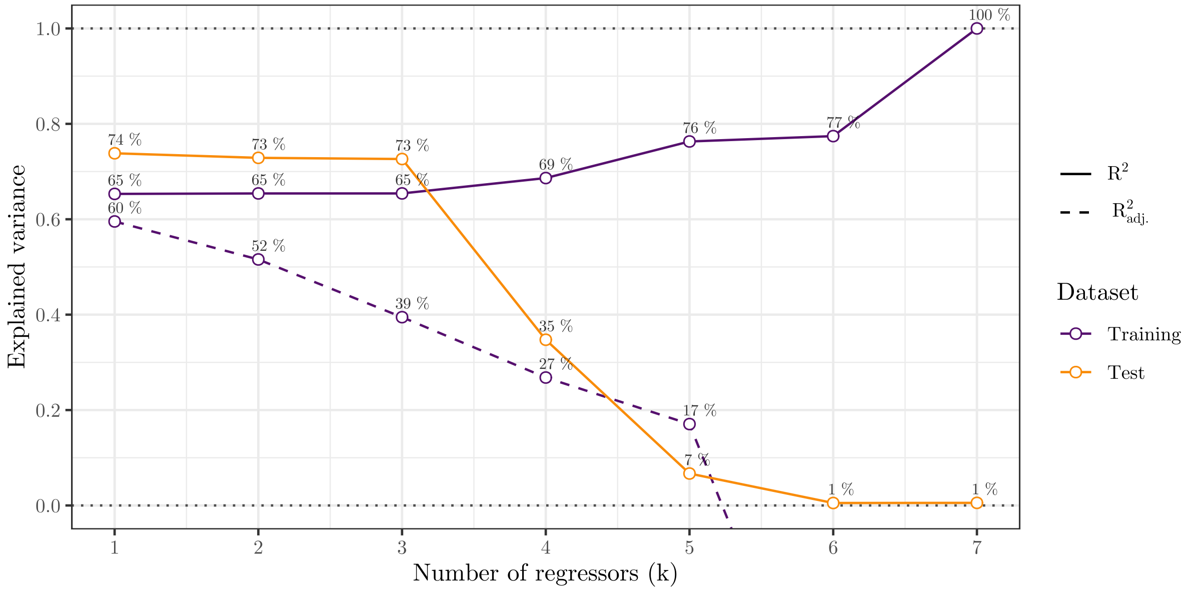
12 Overfitting and cross-validation

There is a wide range of options available when creating a model. We can decide which regressors to include, which first-order or higher-order interaction terms to add, whether to model continuous variables as linear, quadratic or higher-order powers. It is relatively easy to slip into creating a model so complex that it does not actually provide any information at all.

Let us illustrate with an example. Again, let us go back to the data from Table 1 about Agatha, Otto and their six classmates. We find that predicting the outcome of an exam using the number of hours a student had spent studying is a fairly rational procedure. But what if we abandon the assumption that the observed relationship is linear and model it as quadratic by adding the regressor *number of hours squared*? And what if this is not enough and we add the *number of hours to the power of three* regressor to model a cubic relationship? Then we could go on and add the *hours* regressor in higher and higher powers, up to k. The results will be quite compelling – as each additional regressor is added, R^2 increases. Once we reach $k = 7$, R^2 will be equal to 100 %. The model thus achieves a perfect fit to the data.

You will probably argue that for only 8 observations we have included too many parameters in the model (with $k = 7$ we estimate 8 parameters). The value of $R^2_{adj.}$ will also warn us about this, and it starts to drop precipitously if there are too many regressors in the model. The changes in the values of R^2 and $R^2_{adj.}$ as regressors are added can be seen as purple lines in Figure 22.

Figure 22: Coefficient of determination for various numbers of regressors

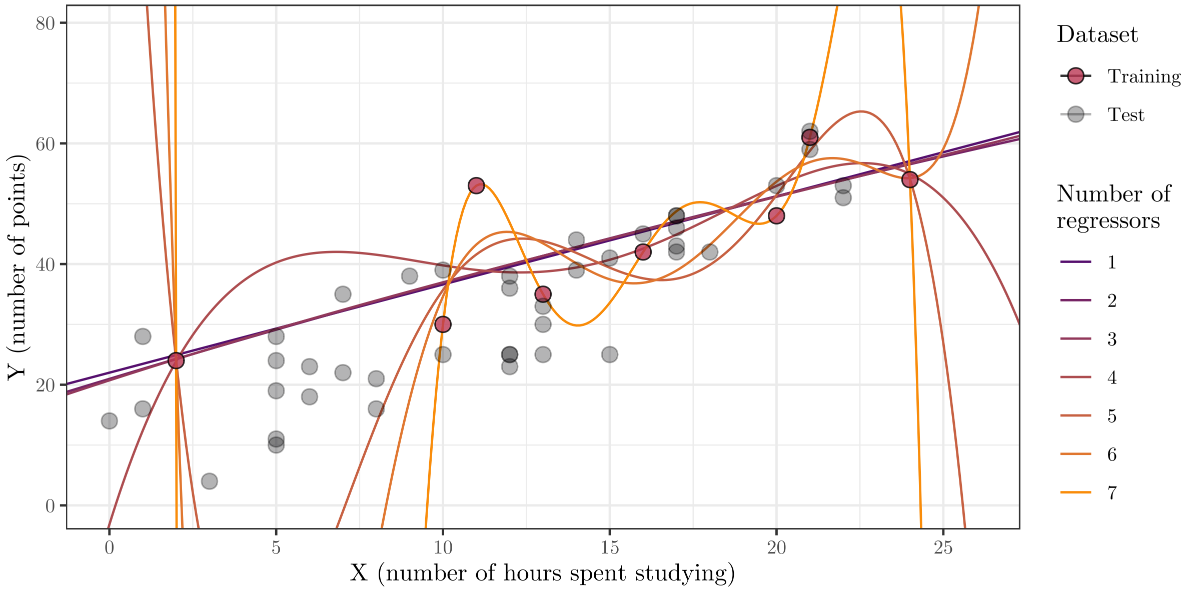


In practice, however, the situation is often much less clear. If we run through a large number of candidates when building a model (for example, similar to the method we learned in chapter about stepwise regression), we can easily produce a model that contains a few regressors (and hence $R^2_{adj.}$ remains high) and yet fits the data very closely. It is not easy for readers to question the quality of a given model, since all quantitative indicators speak in its favor, and information about how many candidates the final model was selected from usually remains secret.

The problem we face here is called overfitting. Overfitting generally means that our parameter estimates are perfectly fitted to our data but they will no longer correspond to any new data. Thus, if someone replicates our research, their result will not even remotely resemble ours.

What happens when overfitting occurs is shown in Figure 23. The red points represent the students in our data set and the purple and orange curves represent the expected results according to each model. As you can see, the linear, quadratic and cubic models behave almost identically in this case, and thus all give almost the same R^2 values. At $k = 4$, however, a strange thing starts to happen – the regression curve starts to fluctuate over a wide range of low and high values. At $k > 4$ it leaves any reasonable bounds. The curve from the model with $k = 7$ passes through all eight points without error, but at the cost of extreme fluctuations.

Figure 23: Observed values and model predictions



In our example, we can detect the presence of overfitting by a single glance at the image 23, but for more complex models this task becomes practically unsolvable. A popular and essentially the only effective solution to the overfitting problem is the **cross-validation**. Cross-validation is based on the idea that it is not advisable to validate the quality of a model on the same data which we had used to estimate its parameters. Therefore, before any computation, we split the data set into two subsets: **training** and **test**. We will use only the observations from the training set to estimate the parameters, but to validate the quality of the fit to the data we will only use the elements of the test set.

The test set is usually less numerous than the training set. In our case, we do not split the eight observations into two groups but add another 42 classmates of Agatha and Otto's to the set. In Figure 23 they are represented by gray circles. We can see that the more our model can fit the training set, the further it moves away from the test set. In Figure 22, the orange curve shows the progress of R^2 computed on the test set. At first results are very pleasing (even slightly better than on the training set, coincidentally). However, from $k = 4$ onwards it drops rapidly and for $k = 6$ and 7 it can hardly explain any variance at all.

Performing cross-validation significantly increases the trustworthiness of our results. While this procedure is not very popular in psychology, in many other fields (especially machine learning[17]) cross-validation is a necessary step, and by omitting it, the author disqualifies their results from any discussion[18].

If necessary, we can be satisfied with cross-validation as described in this textbook. However, more sophisticated cross-validation methods can be found in literature. For example, a popular method called K-fold cross-validation divides sample into K equally sized subsets. Gradually, all K subsets take turns in the role of the test set, while the remaining observations always play the role of the training set. The obtained K model accuracy estimates are then averaged.

[17] Machine learning is focused on creating highly complex models to find the right solution (prediction) in the context of challenging problems. A popular class of such models are artificial neural networks which typically contain hundreds to thousands of parameters – therefore, overfitting is a central issue in this area.

[18] The reason for the little use of cross-validation in psychology is that we usually create models to test the statistical significance of selected regressors, not for prediction purposes. If the statistical model is overfitted, the estimates of the standard deviations of the individual parameters increase dramatically, and there is little chance that any significant relationship will be found. The potential risk of false positives therefore decreases.

13 Long and wide formats, and mixed-effects models

Basic statistics courses have taught us that the usual way of formatting a data table is to assign rows to individual probands and to record the observed statistical features in the columns. This format is often referred to as *wide*. However, this is not the only way to represent data. We can grasp a number of problems much more elegantly by converting the data into a format called *long*. Let us demonstrate the difference between these formats and their usefulness in the following example.

In the context of psychophysiological research, we ask what role neuroticism plays in a stressful situation. We hypothesize that individuals scoring high on the neuroticism scale show greater stress when faced with a frustrating task compared to less neurotic individuals. We operationalize the level of stress as the heart rate measured by ECG.

We monitor each participant's heart rate with an ECG and administer three stressful tasks: the Stroop test, the arithmetic task (loudly reciting numbers in a descending sequence of 300, 293, 286, 279...) and the verbal fluency test (listing as many words as possible beginning with a given letter within a time limit). Between each of these tasks there is a relaxation phase where the participant is presented with calming stimuli. The relaxation phase is also scheduled at the beginning and the end of the experiment, so each participant goes through seven experimental conditions. The questionnaire assessing neuroticism is administered to each participant before testing begins. Sixty-two volunteers participated in the study. Table 11 shows several observations in *wide* format.

Table 11: Data table in *wide* format

		Heart rate [bpm]						
Proband	Neurot.	Relax. 1	Stroop test	Relax. 2	Aritmet. task	Relax. 3	Verbal fluency	Relax. 4
1	10	61	106	75	98	75	92	69
2	2	72	84	79	85	71	75	73
3	12	116	160	149	161	138	148	131
4	14	70	90	67	94	68	97	67
5	16	91	107	83	115	78	118	77
6	17	66	90	65	82	61	88	60
7	21	82	112	104	118	97	120	92
8	4	60	78	65	86	64	79	61
9	23	66	81	58	77	61	78	62
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

If we are looking for an answer to the question how neuroticism affects heart rate under various circumstances, we encounter difficulties in developing a statistical model. What we have here is de facto 7 dependent variables and a single regressor (neuroticism).

We would probably have to help ourselves by some averaging – perhaps by calculating the average heart rate for each proband during frustrating tasks and the average heart rate during relaxation. We subtract these two variables from each other and look for a correlation coefficient between this difference and neuroticism. However, this procedure falls far short of using all the information we have and therefore achieves only limited statistical power.

A more appropriate way might be to convert the data table to the *long* format. In this format, individual rows will represent individual heart rate observations. Thus, each proband will be represented in Table by seven rows corresponding to the seven experimental conditions. The data in *long* format is shown in the table 12. Such a table has $7 \cdot 62 = 434$ rows.

Table 12: Data table in *long* format

Proband	Neuroticism	Stage	Heart rate
1	10	Relaxation 1	61
1	10	Stroop test	106
1	10	Relaxation 2	75
1	10	Arithmetic task	98
1	10	Relaxation 3	75
1	10	Verbal fluency test	92
1	10	Relaxation 4	69
2	2	Relaxation 1	72
2	2	Stroop test	84
2	2	Relaxation 2	79
2	2	Arithmetic task	85
2	2	Relaxation 3	71
2	2	Verbal fluency test	75
2	2	Relaxation 4	73
3	12	Relaxation 1	116
3	12	Stroop test	160
⋮	⋮	⋮	⋮

A linear model that could describe the situation is now more obvious. The dependent variable is the heart rate column, and we use neuroticism, phases and their interaction as regressors.

If you suspect that we must have made some kind of mistake, since we just reformatted our data to increase the sample size from 62 to 434 rows, you are not wrong. A condition of any statistical test is the independence of individual observations. That is, the values in the first row must not be related in any way to the values in the second, third, and subsequent rows. In this case, however, we are severely violating this condition – each seven rows represent one person, and one would expect that if you had a high heart rate six times, you would have a high heart rate the seventh time as well, so there can be no question of any independence.

To remove the dependency, we need to add a categorical variable proband (62 levels in total) to the model. By doing so, we argue that all rows within each group of seven are slightly shifted towards higher numbers and or towards lower numbers. We can compensate for these displacements by using a new regressor.

Unfortunately, solving the independence problem creates another problem: the neuroticism variable and the proband variable are perfectly linearly dependent (i.e., complete multicollinearity is present). Under these circumstances, the least squares method provides no solution. Fortunately, we know a remedy for this problem as well, which is the *mixed-effects model.*

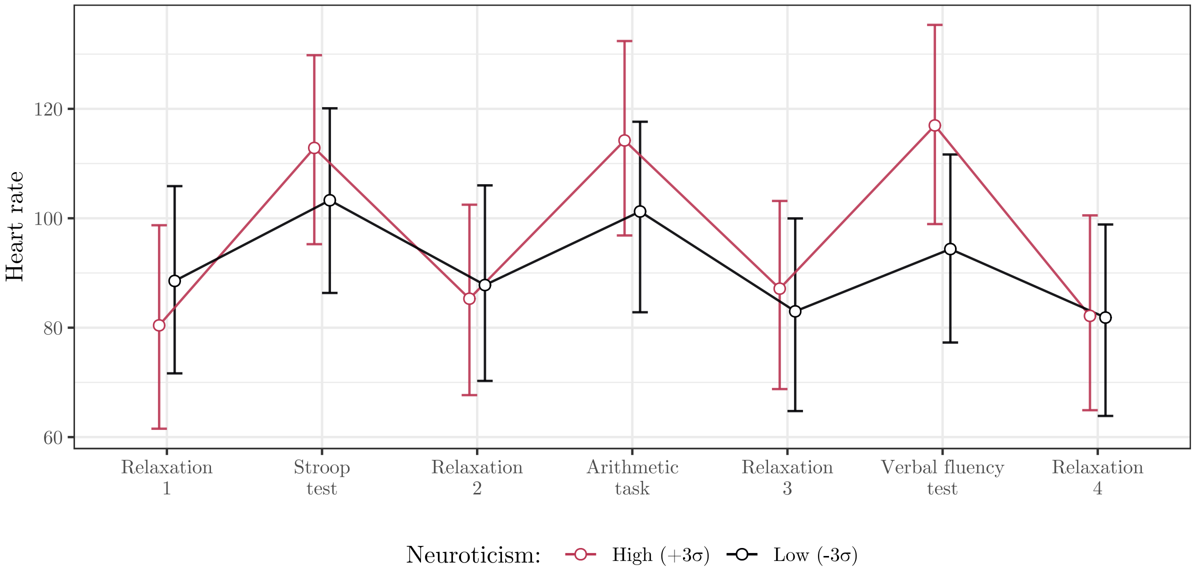
The mixed effects model contains two types of regressors: *fixed* and *random.* The former corresponds to what we learned about regressors earlier. However, the latter, random factors differ in many ways. A random factor is always nominal. **When we say a factor is random, we assume that there is a large population of levels of that factor and that the magnitudes of the regression weights of those levels have a normal distribution in that population**. However, we have only a few randomly drawn levels. In our case, we could consider the proband variable as a random regressor. There is a large population of people, but we have "drawn" only n of them. Moreover, we can expect that the heart rate diversity is normally distributed among different people.

A crucial advantage of random factors is that they are not subject to the no-collinearity condition. We can estimate the weights of random factors even when they are fully linearly dependent on any set of fixed or random factors.

When estimating a mixed-effects model, we work with two types of error variances: we estimate the residual variance already familiar to us, but we also estimate the variance of the random-effect level weights. Thus, in our case, we are not only trying to minimize the inaccuracy of the prediction (σ_ϵ^2), but also to estimate the variance of the heart rates ($\sigma_{\text{proband}}^2$) of the individual probands as low as possible.

When interpreting the results, we do not usually look at the random effect level weights, but only consider the fixed effect weights. Nevertheless, it is worth to pay attention to the estimation of the σ_{proband} parameter, as it helps us get an idea of how fixed effects sizes compare to random effects. In our case, we estimated the interindividual heart rate diversity to be approximately 7.5 beats per minute. The results suggest that the differences in heart rate between individuals scoring at opposite ends of the neuroticism spectrum are between 10 and 20 points. Thus, we can conclude that the observed effects achieve practical significance in addition to statistical significance.

Figure 24: Heart rate during each condition in probands with low and high levels of neuroticism



Mixed-effects models can be applied in more cases than our example suggests. In addition to designs with repeated measurements, the described procedure finds application in designs focused on hierarchically arranged *nested* categories. Typically, this would be the case, for example, when we observe students at several different schools and work with several classes within each of these schools. In such a case, the school and class regressors are perfectly dependent and thus not testable using the least squares method. These designs are often referred to as *hierarchical models* (not to be confused with hierarchical regression) and *nested models*.

Another appreciated feature of mixed-effects models is that they efficiently handle missing data. For example, if there were corrupted records in our heart rate example that needed to be discarded, we can use the method without modification. Whether each proband was measured seven times, or whether this number varies, does not bias the results.

List of symbols

X	Independent variable (regressor).
Y	Dependent variable.
$\hat{Y}$	Predicted value of dependent variable.
ϵ	Residual. Difference between prediction $\hat{Y}$ and the observed value Y.
β	Unstandardized regression coefficient.
β^*	Standardized regression coefficient.
β_0	Intercept (average value of Y if all regressors are equal to 0).
β_1	The slope parameter of the regression line in a simple regression.
n	Number of observations (usually sample size).
h	Number of omitted regressors by which the model differs from its submodel.
k	Number of regressors.
p	Number of estimated parameters (usually $k + 1$). Or p-value for decimal numbers.
RSS	Residual sum of squares.
R^2	Coefficient of determination (percentage of explained variance of the dependent variable).
ΔR^2	Change in R^2 after adding or excluding regressors from the model.
$R^2_{adj.}$	Adjusted coefficient of determination.
S^2_ϵ	Estimation of residual variance. Can also be written as $\hat{\sigma}^2_\epsilon$.
σ^2_ϵ	True value of the residual variance.
SS_Y	The sum of the squares of the differences Y from the mean.
F	The F statistic, or its estimate, for the submodel test.
t	Wald statistic for regressor significance test.
H_0	Null hypothesis.
α	Significance level.
df	Degrees of freedom.
D_j	Cook's distance.
$\mathbf{H}$	Projection (hat) matrix.
h_j	Leverage. The j-th diagonal element of the matrix $\mathbf{H}$.
VIF	Variance inflation factor. An indicator of the multicollinearity. Its inverse is called tolerance.
$\mathbf{x}$	Vector of specified values of X variables for prediction calculation.
$\mathbf{X}$	Design matrix. It corresponds to the matrix of all X variables, supplemented by the first column with number 1 in each row.
$\widehat{\mathbf{VAR}}(\hat{\boldsymbol{\beta}})$	Variance estimate of the regression weights variance matrix.
$I_{1-\alpha}$	In general, the confidence interval (it can be for one parameter, for a regression line and around the regression line).
$P_{1-\alpha}$	Prediction interval.

PhDr. Daniel Dostál, Ph.D.

Linear statistical models in psychology

Responsible editor Otakar Loutocký
Proofreading by Lucie Přibylová
Typesetting by Daniel Dostál

Published by Palacký University in Olomouc, Křížkovského 8, 771 47 Olomouc
vydavatelstvi.upol.cz

1st edition

Olomouc 2021

DOI: 10.5507/ff.21.24460345

ISBN 978-80-244-6034-5 (online: iPDF)

VUP 2021/0397 (online: iPDF)

Not for sale