Chapter 9

Quantitative data analysis: multivariate analysis

In this chapter, we will slowly shift to the more applied side of statistics and research methodology, in the sense that we will discuss multivariate models that are often used in applied research. The model that is central in this chapter, the *regression model*, tries to give us an answer to the question whether, in a certain sample, the scores on a dependent variable (an 'effect') can be predicted from the scores on one or more independent variables (the 'causes'). This means that in this chapter, we have arrived at *multivariate analysis*.

In previous chapters, we have so far discussed univariate (one-variable) analyses (analysis of the mean or median), or bi-variate analyses (analysis of correlations or odds ratios). In this chapter, we discuss two techniques with which the impact on one dependent variable of *several* independent variables is simultaneously analysed. When we use a regression model, we have several independent variables that are of interval measurement level or higher, and the dependent variable is also interval level. When we use an analysis of variance model, we have several independent variables that are of nominal measurement level, with the dependent variable at interval level

In the following, we will first expand and philosophize a little on the term 'model'. What are models, and what do we use them for in research practice? Next, in section 9.2 we will introduce the regression model, in its simple, bivariate form and its multivariate form, as well as touch upon a number of more complicated, multivariate varieties. Such models are hugely important: many of the phenomena that we study in empirical legal research are complex. In general, more than one variable or factor operates on a certain outcome, and we must then also study these phenomena in a multivariate fashion, that is, by taking into account simultaneously all factors that may be relevant.

These multivariate models come with tests and possibilities to assess model fit, i.e. measures that describe how well the model describes the data. We describe these briefly as well. We illustrate both techniques with short examples.

9.1 Models

In daily life, we frequently use models. When we buy for instance a new coffee machine, the box in which it is wrapped generally contains a drawing of the machine that indicates where essential buttons are, how to replace parts etc. That drawing does not depict every detail, colour, or curve of the machine. The drawing serves to outline the basic properties of the machine and to point out the vital buttons. It is made such that you can find your way, that you know how to operate it. As such it is a summary of the machine's appearance. It is an abstraction, in which the essentials have been maintained and nonessentials left out. This has been done so that it is easier to understand the functionality of the machine, to focus on the functions that you as a user are interested in.

For the same purpose that we use models in daily life, we also use them in academic research. Models can be models of scale, giving an overview of a lab machine. Other common models are computer simulation models. Models can also be simple drawings that outline the relations between variables; a number of examples of such conceptual models will be given below. All models have in common that they are an abstraction from reality, in which core elements have been retained. They also have a similar purpose: to facilitate understanding of the functioning or structure of the object they represent.

From the above, we can deduce a number of properties of models. First, models must have a similar structure to the object they are representing. Second, models should enable us to observe and possibly manipulate the object they represent, without interfering with the object itself however. Computer simulations of disease prevalence for instance enable us to model what happens to the spreading of the disease if half of the population is vaccinated, or 25% of the population is vaccinated, etc. We don't need to try this out ourselves: we would not want to! If we feed the right parameters into the model, the model will give us an estimate, a prediction of what will happen. Models can be physical (such as manikins) and virtual (computer models, or even mental representations).

Third, and this is exactly the reason that we work with them, it is much easier to observe the model than to observe reality. That is the crux, and what makes models so eminently useful. Reality is often complex, and hard to get an overview of its complicated intricate entirety. Models are exactly the reverse: they are simple and you can look at them from all angles, roll them around in your hands, put them upside down, whatever you wish. Without needing to worry about that complex and multifaceted reality out there, where many things are interrelated and happen at the same time, with a model we can simply manipulate one aspect – and see what happens, and next manipulate another aspect – and see what happens next. We don't need to worry about all those annoying confounders; we control them.

In summary, a model of reality is a representation of that reality that is firstly independent of it, secondly better observable, and thirdly has a similar structure.

9.1.1 Models with dependent and independent variables

In scientific research we want to – in the end – find the causes of the phenomena we are studying. We try to discover with what other factor our phenomenon of interest correlates, and – importantly – what it does not correlate with. We try to prevent our perception being muddled by any spurious associations. We try to get a grip on the phenomenon, we try to grasp what is going on and what is driving the trend we are seeing, the increase in litigation, the reason why divorce procedures take increasingly long, or whatever we are studying.

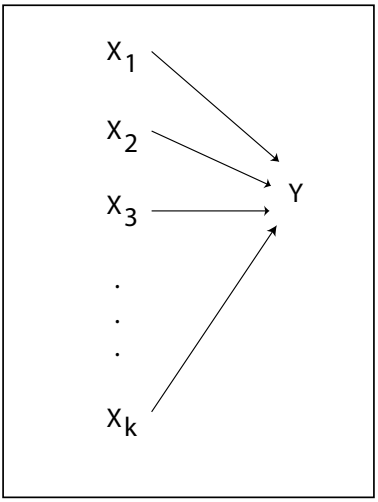
In these examples, there is each time one central property we are interested in: duration of procedures, number of litigation cases, trust in the criminal justice system. As discussed previously in section 5.2, we call such a central variable, which we in a sense want to find the drivers or causes of, the *dependent variable*. The dependent variable is the variable we are interested in, the variable we want to explain, understand or predict from other variables, the variable for which we want to know what the mechanisms are that drives it. We write the dependent variable as 'Y'. Given that the dependent variable is the variable we want to find the causes of, the so-called *independent variable* is then that cause. Defined more precisely, independent variables are variables that are not predicted by other variables, they are the external forces that exert force on the dependent variable. The dependent variable is also often referred to as the *output variable*, the *endogenous variable*, or also simply the outcome. Independent variables are also referred to as *predictors*, *input variables*, *exogenous variables*, *explanatory variables* or, even a bit confusingly, *covariates*. We denote the independent variable by 'X'.

Many of the phenomena we focus on in empirical legal studies are multi- rather than monocausal: often, more than one 'cause' or factor is relevant in explaining what is happening with a dependent variable. So, we may have more than one independent variable, with the various independent variables written as X_1, X_2, X_3 etc.

As an example we again take the theory of procedural justice (Tyler, 1990, 2008), which postulates that for citizens to perceive the outcomes of legal procedures as fair, not just the outcome plays a role, but also and perhaps predominantly the trajectory along which that outcome was arrived at. Procedural justness comprises, according to this theory, several dimensions or variables. A first is whether there were opportunities for participation by citizens, also referred to as 'voice'. A second is the degree to which participants were treated with dignity and respect. As a third factor neutrality is often postulated, that is, whether the forum that proclaimed the outcome is regarded as neutral. Other factors may be incorporated depending on the particular situation investigated. A schematic representation of such a model is given in Figure 9.1.

More complex models may also contain other variables, which play a role 'in between' the independent and dependent variable. Such 'in-between' variables are called *mediating* when they simply channel the association between the exogenous and dependent variable. Mediating variables are no more than a transmitter, they are the delivery man that brings the parcel that your aunt dispatched to your doorstep: without the delivery man the parcel would not get from your aunt to you. An often-used example is the impact of lack of sleep on task performance. When you do not sleep well, your alertness is decreased, and it is through this decreased alertness that task performance is affected.

Figure 9.1: Visual representation of model with predictors and dependent variable



'In-between' variables sometimes do more than simply channel the impact of the independent variable on the dependent one: they may also change, amplify or suppress an effect of a certain X on Y. For instance, painkillers can be safely used at a certain dosage. However, these painkillers may have adverse side-effects in persons with a certain medical condition. That medical condition is then a moderator: it changes the effect of the painkiller. Such variables are called *moderator variables*, and these effects are more broadly referred to as interaction effects. We return to this in topic in section 9.3.3.

9.1.2 Model building

In practice, models have to be built. While a manikin can be a simplified replica of a physical entity, a model in social research practice is usually conceptual. Using data, and reasoning from theory, we try to explain some phenomenon by building a conceptual structure. We presume for instance that divorce procedures of spouses with underage children last longer than divorce procedures of spouses without children. And we also presume that divorces of spouses that have disputes about division of matrimonial property take more time than simple divorces without any financial disputes. Our model is then, very simply written:

$$\text{disposition time}_{\text{divorce procedures}} = f(\text{underage children} + \text{financial disputes}),$$

which reads like: disposition time is a function of the variables 'Underage children' and 'Financial disputes'. We can write it in a more compact way as:

$$\text{DT}_\text{d} = f(\text{U} + \text{F}).$$

We know that disposition time will not be approximated by simple addition of these two independent variables U and F. In fact we expect that the presence of underage children (a score of '1' on the variable U) will increase DT and that financial disputes (a score of '1' on the variable F) will also increase DT. We can express this by inserting weights b_U and b_F in the formula:

$$DT_d = b_U * U + b_F * F,$$

written more generically as:

$$DT_d = b_1U + b_2F.$$

We are expecting that b_1 and b_2 will both be positive: both underage children and financial disputes are expected to increase disposition time. Before we start estimating our model, we note that – even if we are in the most favourable situation without financial disputes between spouses (a score of '0' on F) and no underage children (a score of '0' on U) – disposition time cannot be zero. There will always be a lag between the filing for divorce and the day the case is legally closed. We know that minimum disposition time is 30 days, with most cases taking 60 days or more. We accommodate for such a start 'tariff' by inserting a so-called *intercept*. The formula then becomes:

$$DT_d = a + b_1U + b_2F.$$

The value of the intercept will have to be estimated, and it will likely not be lower than 30 days. The coefficients b_1 and b_2 are called *weights*. Both the coefficient a and the b_j are *parameters*.

Now, we may try to find out whether this model actually does describe our data well. Are we able to predict time to conclusion of cases using these two variables: 'Underage children' and 'Financial disputes'? Or are our predictions very much off the mark?

As models seldom explain all 'behaviour' of the dependent variable on the first try, models are generally built step by step, in an iterative process, to try to improve upon model performance. Researchers may do so themselves. Also, one researcher may start with a model, after which the next researcher adds on to it, specifying certain relations, adding new ones, inserting better measures of certain variables, deleting or replacing old ones. In our example, for instance, it may be that the model misses out on whether a mediator was involved in the divorce process. Also, it may be the case that procedures are concluded faster in certain courts: they may be faster in courts that serve rural areas. But when divorces are more acrimonious they may take longer. Or divorces that are filed by both partners instead of unilaterally may conclude faster. Or it may be that the specific combination of underage children and high conflict is particularly obstructive (an interaction effect).

For our example we could indeed try to improve our predictions by adding such additional characteristics of court cases. We may for instance add whether a mediator ('M') was present:

$$DT_d = a + b_1U + b_2F + b_3M,$$

and next add a variable that captures the degree of spousal conflict, such as the Glasl escalation ladder (Glasl, 1977). The model then becomes:

$$DT_d = a + b_1U + b_2F + b_3M + b_4G.$$

With each variable that we add, we make the model more complicated. We do not only include more variables, but with each variable we add we need to estimate an additional parameter b_j. In general, models with more variables will better predict the scores on the dependent variable. That is logical. If we add variables, we have more information, so it is likely that our prediction improves (that it is to say: it cannot get worse). But, as models get more complicated, we in a sense pay a price for that: our model is then also less easy to understand, to grasp conceptually, and to manipulate. Of course, we could choose to work with simple, easy-to-handle models that describe reality in a few 'sound bites'. The price we pay then, conversely, is that they will describe reality less well. In that sense, there is always a trade-off. We prefer our models to describe the data as well as possible. At the same time, we want them to be as simple as possible – the latter is called the principle of *parsimony*.

The two 'bite' each other. For perfect prediction, we might want to add all the characteristics we can find. But for understanding reality, models would rather be simple. The art of model fitting is therefore to find the best balance between simplicity and prediction. We want to find the best-fitting model with the fewest variables. If we can choose between two models that predict equally well, and one is sparser than the other, we would choose the sparse, simple one. In the following section, we will discuss this a bit more.

9.1.3 Model fit

In general, simpler models are easier to work with, easier to understand and conceptualize, easier to test. However, they have less explanatory power: they are simply less well able to describe the complicated world. How well a model describes reality is often expressed in a number that represents how close the values of the dependent variable as predicted by the model are to the real, observed values. If the dependent variable as predicted by the model approaches the observed dependent variable very neatly, we say that the *model fit* – also *goodness of fit* – is high. The model that we built is then able to predict what the scores on Y are. The model then, as we say, 'fits' the data. This is what we want. We have then a model that performs well.

Complex models generally describe reality better. As we said, if we include more factors that might play a role, i.e. more variables, it is, as we said, only logical that will become better able to capture what is happening and thus to describe reality better. We then have the possibility to model lots of causal relationships, the manner in which variables mutually influence each other, or influence the dependent variable through loops, mediate and moderate, so that it is not surprising that we are then better equipped to describe all manners of behaviour of the dependent variable.

Complex models have disadvantages too. First, as already mentioned, they may become too complex to conceptualize, too complex to understand, too intricate to be of practical use. But in addition, they may also become untestable. That is because, for

a very complex model, we would really need a large sample size, a lot of respondents or objects that we studied, to be able to investigate whether the model 'holds'. For the models that we discuss in this chapter, there are two general rules of thumb. The first one harks back to what we discussed in the chapter on testing (chapter 8), where we stated that for statistical testing in general sample sizes from 100 become workable. This does not mean that a sample size of 90 is useless, and as we will see, multivariate analysis can even be conducted (although with limitations) with datasets with an N a little over 60 (see section 10.4). The second rule is also best used as a rule of thumb. It states that for estimating the multivariate models we discuss in this chapter, sample size should be approximately 10 times the number of variables one wants to use.

So if we wanted to test a model with 12 variables (which is likely a quite complicated model), we would need a sample size of at least 120. It is not that the model will immediately break down if too many variables are stuffed into the model for the sample size. The software with which one runs computations will generally produce results. However, extremely complex models may also become trivial in the sense that while they might predict the behaviour of the dependent variable pretty well, they capitalize on small differences, which are of little conceptual or practical use. As such, they do not offer a lot of insight into the mechanisms we are interested in; they distract from the core issues. Formulated differently, such very complex models are not really models any more: models are supposed to be an abstraction from reality, simplified for the purposes of manipulation and easier understanding. Very complex models do not do that anymore, so they may deserve the qualification 'model' no longer.

We reiterate that we prefer models to be simple, or 'parsimonious'. It is not difficult to attain a good model fit if you use a lot variables. The art of model building is to arrive at a good fit using just a few variables, to get to the core of things, the crux of the mechanisms we are studying.

We will return to the issue of model fit in section 9.2.4, where we introduce measures that capture model fit. We will then also discuss a particular measure that not only takes into account how well the dependent variable is predicted by the model, but also gives a 'penalty' for the complexity of the model.

It should be noted that the fact that a model fits the data well does not mean that we have discovered the mechanism that drives the dependent variable, that we have discovered the causal mechanism. The Ancient Greeks believed that malaria was caused by living at low altitudes – they had observed that in the highlands people never contracted malaria. Putting these variables in a model gives a good fit: we can very well predict whether someone will catch malaria from a variable that measures whether someone lives at sea level on the Mediterranean or in the Alps. But the real mechanism is not captured. One variable is omitted: the presence of mosquitos and more precisely the presence of mosquitos that carry the malaria parasite. We can predict the incidence of malaria with a variable that correlates with the true causal agent. This we call *model misspecification*.

Models on observational data as we discussed here (malaria, disposition time) can therefore never prove causality. When models work well, when they help us to understand the world around us, we use them. If we find that they are not helpful anymore, we discard them. Models are not true or false, or just or discriminating. Models are no more than a tool for understanding the world around us.

9.2 Regression model

In this section we will discuss the regression model more formally. The regression model is a model just like the models we discussed above. It is a model in the sense that a mathematical formula is used in which one or more independent variables X_1, $X_1 \ldots X_k$ play a role, and that with these a dependent variable Y is predicted. Regression analysis can be carried out with just one predictor variable, or with several; the latter is the usual situation.

Regression models are used very often in applied social research, which empirical legal studies is. That is because many of the questions we ask centre around causality. The regression model, if specified adequately, is a suitable tool for investigating questions of causality. Regression analysis, or 'regression' for short, has been called the workhorse of social science.

Estimations will tell you what the best predictors out of a set X_1, $X_1 \ldots X_k$ are of a phenomenon of interest Y. If the model fits well – which is when X_1, $X_1 \ldots X_k$ contribute significantly to predicting Y – then it can be used to predict future Ys. It can also point us towards interventions. If we were to find, for instance, in a model in which we predict disposition time of divorce cases using a large number of input variables, that the use of mediation predicts a shorter disposition time, this might help courts decide to offer such mediation services (although it would be wise to evaluate their effect using an RCT).

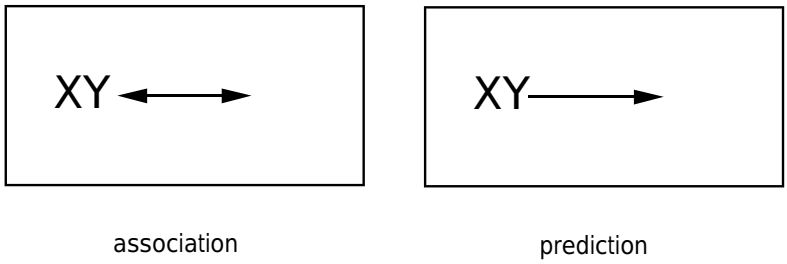
9.2.1 When to employ regression analysis

Regression analysis is a feasible technique if one wants to predict one variable Y from one or more independent variables X. The X_k are regarded as the causes or determinants in the process that the regression model attempts to capture, which means that they function as the independent variables. The Y is seen as the 'effect', which means that it fulfils the role of dependent variable.

Regular, that is, linear, regression analysis can be performed if all these variables meet some requirements. Regression analysis is feasible if the dependent variable is at least at interval measurement level and if it is not too weirdly distributed. If Y were to be for instance very skewed, then regular regression analysis is not possible: the conclusions from the analysis are then untrustworthy. The X are supposed to be interval level variables too, even though some accommodations for this are possible in practice (see the example discussed in section 10.4).

We introduced the word 'linear' in the above paragraph. By that we indicate that it is assumed that the association between the variables in the model is linear. If the association is curvilinear (for instance if Y is high for average values of X, but low for extreme values of X), the regression analysis will not be able to describe that relation, as it can only fit linear relations. The model is then misspecified. In such cases, it is possible to transform the independent or dependent variable: especially when variables have a skewed distribution, some researchers regress the X variables on a logged version of the Y variable. At times one also encounters squared versions of the input variables. These are all attempts to 'fold' non-linear relations into the model.

Figure 9.2: Visual representation of association between X and Y, and prediction of Y from X



If the Y variable is not interval level, a different kind of regression analysis must be performed. For instance, if Y is a dichotomous variable (meaning it has values 0 and 1 only), a so-called logistic regression analysis should be performed. If Y is an ordinal variable, linear regression is unsuitable (even though it must be said that if deviations from linearity are not too large, most researchers simply perform regular regression analysis on the rank-ordered scores of the ordinal variable). If Y has yet other properties, other varieties are possible (see for an overview Bijleveld et al., 2015, chapter 2). We will discuss none of these in detail; for our purposes it is only important to know that it must always be checked whether Y has the properties it should have. See also section 9.2.6 below.

9.2.2 Simple regression analysis

We will discuss simple regression analysis in somewhat more technical detail, because the technique can easily be illustrated for this simple case, with just one X variable to predict Y. We then speak of 'simple regression analysis' or 'simple linear regression'. We are then essentially carrying out a *bivariate analysis*.

When we investigate the association between X and Y, we are interested in the extent to which they have something in common. We can do so through a correlation coefficient. When we carry out such a correlational analysis, X and Y are interchangeable; the analysis is symmetric. However, when we carry out a regression type of analysis, the two variables have different roles: one is modelled as the consequence of the other. The roles of X and Y are therefore not interchangeable. See Figure 9.2.

As said, the analysis we are conducting is directional. We want to predict Y, not X. Our question is not about the strength of the association between X and Y, or in the sign of that association (positive or negative), we want to know whether information on X helps us to conclude something about Y. Obviously, correlation is important. If X and Y are correlated, then it is likely that information on X also tells us something about Y. If we know that someone has shoe size 46, then that person is likely tall. The two are related, which means one can be used to predict the other. If we denote the Y predicted by the regression model as $\hat{Y}$, then the art of regression analysis is to find a

$\hat{Y}$, a predicted Y, on the basis of the regression model, that approaches the observed Y closely.

In order to predict Y from X, we construct the following linear model:

$$Y = a + b\,X,$$

which we use as follows for predicting the individual scores:

$$Y_i = a + b\,X_i + e_i,$$

where Y_i is the score from respondent i, and a is the *intercept* that we introduced before (see section 9.1.1), added to which is the same respondent's score on the predictor variable, X_1, weighted by its *regression coefficient* referred to as b. The $a + bX_i$ are what is predicted for Y on the basis of the model. As that prediction will not always be perfect, a so-called *error term* written as e_i is defined as the difference between the actual, observed Y, and $\hat{Y}$, the predicted Y:

$$Y_i - \hat{Y}_i = e_i.$$

The error term is actually best thought of not as an 'error' in the sense of a mistake, but as a prediction error. If we attempt to predict Y from X, that prediction will not be perfect: in some instances $\hat{Y}$ will be a little too low, in other instances it may be a little too high. The difference between Y and $\hat{Y}$ is our prediction error.[1]

Let us illustrate all this with a simple example. Suppose that I want to predict someone's height from his or her shoe size. I myself am 170 cm tall and I wear a size 39. My brother is 191 cm tall and wears a size 46. If we were to concoct the following regression model – and this is a home industry example – we would be able to predict my and my brother's shoe sizes:

$$\text{height} = 53 + 3 \times \text{shoe size},$$

which means nothing more than: if I wanted to predict someone's height from his or her shoe size, I would have to start from 53 cm, and add 3 cm for every shoe size up. For myself this would become $53 + 3 \times 39 = 170$, i.e. a perfect prediction (so that $e_{\text{Catrien}} = 0$). For my brother the formula would be $53 + 3 \times 46 = 191$, i.e. a perfect prediction again, with the corresponding error term again equal to zero.

In reality, predictions will hardly ever be perfect. We illustrate this with a small real dataset, which contains shoe sizes and heights for four more people. See Table 9.1.

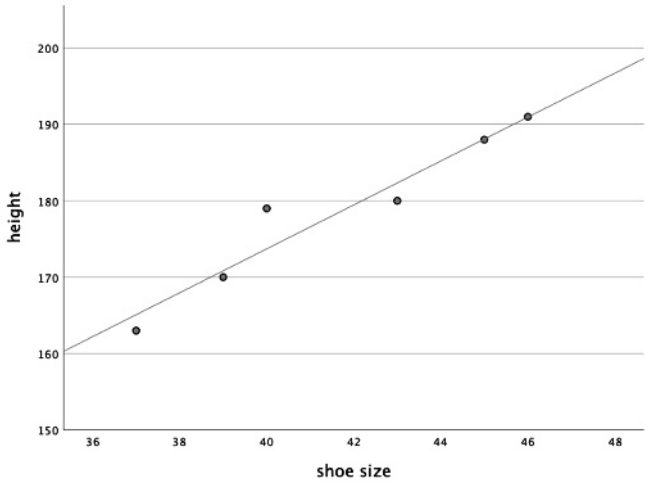
We see that in this dataset there is a fairly strong correlation between shoe size and height: another male has large feet (shoe size 44) and there is a shorter male with shoe size 43, but we also have a fairly tall female (179 cm) with shoe size 40, and a shorter female with shoe size 37. For these other sample members, the regression model that we just thought up will not predict perfectly. For instance, for the third respondent, clearly $53 + 3 \times 40 \neq 179$, and for the relatively shorter male, $53 + 3 \times 43 \neq 180$.

[1] The parameters a and b are computed as $b = r_{XY}\frac{s_Y}{s_X}$, with r_{XY} the correlation coefficient between X and Y, s_Y the standard deviation of Y, and s_X the standard deviation of X, and as $a = \bar{Y} - \bar{X}$, with $\bar{Y}$ and $\bar{X}$ the means of the variables X and Y.

Table 9.1: Dataset with shoe size, height and gender

respondent	shoe size	height	gender
1	39	170	female
2	46	191	male
3	40	179	female
4	45	188	male
5	43	180	male
6	37	163	female

Figure 9.3: Scatterplot of shoe size against height



However, the correlation coefficient between height and shoe size is very high (r = .965), so while we will not get a perfect prediction, we do get a pretty nice prediction.

We illustrate this in Figure 9.3, which contains a scatterplot of shoe size against height for this dataset. We see that the dots are aligned, and we are able to fit a straight line through the dots.

If we were to estimate a regression model using a statistical software package, we would arrive at:

$$\text{height} = 58.763 + 2.874 \times \text{shoe size}.$$

The value 58.763 is the intercept a, and the regression weight b equals 2.874. The intercept can be thought of as the predicted height for someone with a hypothetical

shoe size 0: if we were to expand the x-axis to the left, the regression line would cross the y-axis at 58.763. Someone with shoe size 1 has an estimated height of 61.637 cm, someone with shoe size 2 has an estimated height of 64.511 cm, and so forth. The regression weight reflects the steepness of the regression line. All predicted values fall exactly on the regression line; the observed values are in the scatterplot a little above or a little below the line. The closer the observed values are to the line, the better the line describes the true scores, and the more successful the prediction.

My predicted height is now 170.849, quite close to my true height. But the predicted height of the tall female is actually 173.723: she is therefore at 179 cm slightly tall for her shoe size – or conversely she has relatively small feet. The picture shows this too: her observed value lies indeed quite a bit above the fitted regression line in Figure 9.3 and her predicted value, on the regression line, is quite far from her observed value. That distance, the prediction error, can be envisaged as a vertical line from the observed value to the regression line. If we were to draw all those lines for the six sample members, we would see that they are all quite short lines.[2]

At this point we remark that both the intercept and the regression weight are *scale-dependent*. They depend on the scale in which we measured the variables. If we were to for instance measure height not in centimeters but in meters, the intercept would be 0.588 and the regression weight b would equal .029. And had we not used EU shoe sizes but UK shoe sizes (which are EU size minus 33), the intercept would change to 153.595 (the regression weight will not change because going from EU to UK shoe sizes does not entail a change of scale but only a shift along the same scale). The prediction is just as good, as it does not matter for the predictability of height whether we measure height or shoe size in one scale or the other: shoe size predicts just as nicely, and the line will fit just as nicely in the scatterplot.

This has one important implication. Give that the weights and intercept are scale-dependent, we cannot infer the importance of an independent variable from its absolute value. The scale-dependent regression weights are referred to as *unstandardized*. Their value will change if we use a different metric for the variables, such as UK shoe sizes, or inches or feet for height. It is possible to standardize the weights. For that, one standardizes the X and Y, which means that a will be zero. This has the advantage that, if we had more than one variable in our model, we could gauge the relative importance of the variables from the values of the standardized regression weights. The standardized b is generally written as β. Obviously, the disadvantage is that we can less easily translate the impact of the independent variables on the dependent variable.

To assess the significance of a variable in predicting a dependent variable, we use a significance test. For the example we just used, shoe size (even in this tiny example) is a highly significant predictor of height ($p = .002$). We will return to this in section 9.2.5.

Now, when does a regression analysis perform well? As said, we judge this by how close the predicted Y, $\hat{Y}$, are to the observed Y. Ideally, we would like the differences between predicted and observed values of Y to be zero, a situation of perfect

[2] By computing a and b with the formulas given in the previous footnote, the error terms always add up to zero so that the average predicted Y is always equal to the average observed Y. These formulas also ensure that a best fitting line is found, a line that minimizes the sum of the squared error terms $\sum_{i=1}^{N} e_i^2$. Formulated more wordily: using these formulas ensures that we will always find the best possible solution. The manner in which a and b are computed is called 'least squares' or 'ordinary least squares' (OLS).

predictability that is in ordinary social research practice hardly ever achieved (and if it is, is cause for suspicion and careful scrutiny of the data). So, if the error terms e_i are large – the distances between the observed values and predicted values on the regression line – prediction is not going so well. If they are small, prediction is going well. We will discuss measures that express predictability below in section 9.2.4.

9.2.3 Multiple regression analysis

Simple regression analysis is indeed a fairly 'simple' technique. If we were to employ a regression model with just one predictor, we would assume that the variable Y is under just one other external influence. For many practical situations, that is too simplistic. In empirical legal studies, we know almost for a fact that most processes are not monocausal. Mostly, various factors play a role in the processes we are interested in. So, it would seem logical to employ not one but several independent variables.

The model we then use becomes correspondingly more complicated:

$$\hat{Y}_i = a + b_1\ X_{1i} + b_2 X_{2i} + b_3 X_{3i} + \cdots + b_k X_{ki}.$$

Note that, as we use k different independent variables to predict the dependent variable Y, we also have as many b-coefficients: one for each X variable. Still, we have only one intercept, the in a sense 'starting value', to which multiples of the respective X variables must be added to arrive at an estimated Y. Again, we arrive at a predicted Y value, $\hat{Y}$, and we have differences between the observed Y and $\hat{Y}$:

$$Y_i - \hat{Y}_i = e_i.$$

Also, just like the simple regression case, we hope that the e_i are small. For fitting the model, we compute an a, and we compute several b, just as many as we have independent variables. Very few researchers still compute the values of these parameters by hand[3], and software packages like SPSS, Stata or freeware like R are generally used to do so.

We will now discuss an important property of the multiple regression model. For that, it may be instructive to look at the formulas in the footnotes given, without needing to grasp them fully. It is important to note that the formula for β_1 – the standardized regression coefficient of b_1 – contains the correlation coefficient between Y and X_1, but

[3] We give the example for the case when there are two independent variables, X_1 and X_2. We give for reasons of simplicity the standardized regression weights β_1 and β_2:

$$\beta_1 = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}},$$

and

$$\beta_2 = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{1 - r^2_{X_1X_2}}.$$

As explained above, when a regression weight is standardized, this means that the regression weight takes into account that the original variables have been standardized. The scale or unit in which they were measured therefore does not play a role anymore; they are the regression weights we would obtain if we worked with standardized variables only.

it also includes r_{YX_2} as well as $r_{X_1X_2}$. And the formula for β_2 contains the correlation coefficient between Y and X_2, but it also has r_{YX_1} as well as $r_{X_1X_2}$. Clearly, in computing the weight for estimating Y from one predictor X_1, the other variable X_2 plays a role as well – it is taken along in the formula – as does the correlation coefficient between the two predictors. Apparently, the regression coefficient for one variable X_1 is not computed in isolation. The association between the other predictor and Y, as well as the intercorrelation between the two predictors, is taken along as well. Why is this?

It is because the regression weight does not so much tell you whether some predictor variable X_k and Y are associated; rather, it tells you how much a variable X_k contributes to predicting Y, given the other variables in the model – in other words, how much a variable contributes to the prediction of Y over and above the other variables. This is a very important property of multiple regression and is exactly what makes it such a powerful technique in theory building and the search for causal relations.

To illustrate this, let us imagine the following situation. We are predicting Y from X_1. As we have only one predictor, we are performing a simple linear regression. To this model, we now add a second predictor, X_2; this predictor is however identical to X_1. As X_2 is identical to X_1, adding X_2 to the model does not improve the prediction. X_2 does not supply us with new information; all there was to know is already contained in X_1. So, essentially, to improve the prediction of Y, adding the variable X_2 is of no use. One could say, given that X_1 is already there in the equation, X_2 might just as well receive a regression weight of 0, as it cannot add to the prediction of Y. All that X_2 has, X_1 had already.

This may seem paradoxical, because, as X_2 is equal to X_1, it correlates just as strongly as X_1 does with Y, so you would expect it to play a similar role in the regression equation! However, and this is what is really important here, the regression coefficients do not tell you how strongly variables are correlated with Y, but how much each variable adds to the prediction of Y, *given the other variables in the model*.[4]

It is important to always keep in mind that the regression weights have meaning only relative to the other variables in the model. That is why in reporting on regression weights the formulation should always run like: 'Given the variables A and B, the

[4]Suppose that we have two predictor variables that are fairly similar. Let us assume that they correlate really strongly, so $r_{X_1X_2}$ =.9. We also assume that both are equally strongly related to Y, such as r_{YX_1} = .8, r_{YX_2} = .8. If we would estimate two univariate models, once for X_1 predicting Y and once for X_2 predicting Y, the standardized regression weights β would in both cases equal .8. But what happens to the weights in the multivariate situation when we perform multiple regression? Now, however, let us look at the regression weight for X_1 in the multivariate situation, i.e. after we have added X_2 to the regression equation. Then:

$$\beta_1 = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1\text{-}r^2_{X_1X_2}} = \frac{.8 - .8 \times .9}{1 - .9^2} = .421.$$

The weight β_1 is now much lower than we would expect on the basis of the bivariate model (and β_2 is exactly similar). However, this is not strange if we realize that the weight tells us how important X_1 is *given the other variables in the model*. Given that X_2 is already present, X_1 adds less, is less useful than it would be if it were on its own. In the univariate case, the regression coefficient is .8; in the multivariate case it becomes .42. This can be understood by realizing that the *unique* contribution of X_1 to the prediction is less in the multivariate than in the univariate case.

Suppose now that the r_{X_1Y} and r_{X_2Y} are exactly the same, but that the correlation between X_1 and X_2 is not .9, but .3. Then β_1 becomes .615! The association between Y and X_1 is unchanged, but because there is less overlap between X_1 and X_2, the unique contribution of each variable in the prediction of Y is larger.

variable C contributes so-and-so much to ...'. If there is strong intercorrelation between independent variables, such as in our example just now, this is referred to as *multicollinearity*. Multicollinearity is undesirable. It makes it harder to interpret the regression coefficients. In extreme situations, it can make the model unstable. Researchers should prevent multicollinearity by first checking whether the predictors they want to include do not overlap too much. Correlations above .8 are generally considered too high for variables to be included in one regression equation, and specific tests for multicollinearity (such as the so-called Variance Inflation Factor or VIF which tells you how much each independent variable overlaps with the other variables in the model) are included in most software packages. Such tests are also necessary from the viewpoint of parsimony. As stated above, we want to build sparse models, with good predictions for a small number of independent variables. It then does not make sense to build a model with lots of variables that essentially measure the same.

9.2.4 Model fit

We already introduced the term model fit, and used it mainly conceptually. But how to express the fit of a model? A commonly used statistic to express model fit is the correlation coefficient between the observed Y and the predicted Y, $\hat{Y}$. It is only to be expected that, if a model is able to predict the dependent variable well, the correlation between Y and $\hat{Y}$ will be high. This correlation coefficient is denoted with a capital R:

$$R = r_{Y\hat{Y}}.$$

As a prediction gets better, the correlation between the observed and predicted Y will also be higher. Therefore, R indeed reflects the quality of the prediction. So this is all quite logical and intuitive. Mostly, however, instead of R, its squared value R^2 is used, and is referred to as the *percentage of explained variance*. If R^2 is for instance 0.64, it is said that the regression model is able to explain 64% of the variance of Y. R^2 is regarded as a 'goodness of fit' measure.

It is hard to say what acceptable values for R^2 are. Values around 0.80 are actually exceptional in most empirical studies. Depending on the research topic, some researchers are even content with values around 0.05. Much depends on the extent to which the phenomenon under investigation *is* predictable. Some processes are more accidental, others follow certain regular rules more. In the first case, we might be happy with an R^2 of around 0.10; in the second case we would expect higher values.

Many researchers prefer to use, instead of R^2, a slightly different statistic, which not only expresses how much of the variation of Y can be explained by the model, but penalizes models that employ too many variables. It is called the adjusted R^2, and written as R^2_{adj}.[5] There are several reasons why one would want to express model fit

[5]The formula for the adjusted R^2 is:

$$R^2_{adj} = 1 - (1 - R^2) \times \frac{N-1}{N-k-1}.$$

which shows that as k, the number of independent variables, increases, R^2_{adj} decreases. The formula also shows that this is particularly the case if N is relatively small, so in small samples. The impact on R^2_{adj} of including many variables is then bigger.

using R^2_{adj} rather than through R^2. A first is that an R^2 of, say, 0.65 in a model with six variables is more noteworthy than an R^2 of 0.65 in a model with 12 variables: R^2 automatically increases when more variables are added to the model, and R^2_{adj} adjusts for that. A second reason is that we aim to build parsimonious models. A third is that models with many variables may capitalize on accidental and idiosyncratic patterns in the data, which will be difficult to replicate; we call such models *overfitted*.

9.2.5 Tests

Just like we can test whether a mean in a sample is highly unlikely under a certain null hypothesis, so we can test whether finding a regression coefficient b or standardized coefficient β with a certain value is likely or unlikely under the null hypothesis:

$$H_0 : \beta = 0,$$

with the alternative hypothesis:

$$H_1 : \beta \neq 0.$$

Note that the β in the hypothesis refers to the hypothetical distribution of the unstandardized b (or standardized β) in the population. It may be slightly confusing that we use one and the same symbol for two different entities, but because when we are testing we are referring to unknown populations, we always use Greek symbols. In the null hypothesis, it is stated that the b or β as observed stems from a population where it equals zero (meaning that it does not at all contribute to the prediction of Y). The significance test will then tell us how likely it is to find the observed regression weight if that were true. If that likelihood is very small, we reject the null hypothesis. We thereby state that the particular variable X_k contributes significantly to the prediction of Y.

For testing whether that is so, it does not matter whether b or its standardized version is investigated. In fact it would be very strange if the particular scale in which a variable is measured would make a difference for how much that variable could aid to predicting a certain outcome. If a regression weight differs significantly from 0, we state that the variable adds significantly to the prediction of Y. It should be noted that, just as in the case where we are testing means, regression coefficients will turn up as significant sooner when samples are large, and will turn up as significant less easily when samples are small.

Other tests are possible when we conduct regression analysis. For instance, we can test whether the R^2 from one model is significantly higher than the R^2 from another model. We can therefore test whether a model with a few extra variables leads to a significantly better prediction of Y. Such tests enable us to compare models with a statistical criterion. Relatedly, most statistical software packages test – before carrying out regression analysis – whether the model performs better than the so-called null model, a model with only an intercept. Only when this is the case, does it make sense to proceed and interpret the regression coefficients.

Tests are used for model building: we may want to build a sparse model, a parsimonious model, and retain only variables that contribute significantly to the prediction of

Y (see also section 9.1.2). What most researchers do is first fit a model with all predictors that would seem relevant from theory or earlier research. They then remove those predictors that do not contribute significantly to the prediction of Y, and retain only the significant predictors. Most software packages offer stepwise procedures that either remove variables one by one until only significant predictors are retained, or that build the model by adding predictors until the fit of the model does not improve significantly anymore.

9.2.6 The workhorse of multivariate analysis

Multiple regression is often named the 'workhorse' of (linear) multivariate analysis, that is, analysis of the relations between a number of variables simultaneously. It is such an immensely powerful and handy technique because it can assess the effects of a large number of independent variables on a certain dependent variable that one aims to predict. In doing so, and this is likely its most attractive property, it returns the importance of each variable *given the other variables*. Importance of a variable is therefore *conditional*, *net of* the other variables. This means that researchers obtain ground to make statements such as 'taking into account the impact of A, B and C, the factor D still significantly contributes to predicting Y', or: 'the impact of D on Y cannot be attributed to confounders A, B or C'. This is why multiple regression a very important tool in theory building, in testing models to explain the association between variables.

The linear multiple regression model that we discussed is but one variety in a family of regression models. Depending on the type of variable that one wants to build a prediction model for, varieties of multiple regression exist. A very commonly used other type of regression model from that family is logistic regression, an adaptation of the multiple regression model where the dependent variable is not a continuous variable but a dichotomous variable (it has only two categories, such as 'yes' and 'no', or 'pass' and 'fail'). Other types exist, such as Poisson regression (for predicting very sparse events that follow a certain distribution, such as convictions). An important variant is Cox regression (for predicting variables that reflect the duration to a certain event, such as time to settlement of a claim); we encountered the univariate version of Cox regression in chapter 6 as survival analysis.

We will end this section with a small fictitious illustration of multiple regression. We investigate possible gender discrimination at universities. It is well known, and fairly universal, that female academics earn less than male academics. The issue at stake is whether these lower earnings are due to existing differences between male and female academics, or due to discrimination. A first explanation for this disparity due to existing differences may be that women are less often associate and full professor than men. Another explanation may be that women, because they more often work part-time, publish fewer articles. Then, women academics may be on average younger, which might also explain part of the salary difference. It has also been noted that some disciplines are more gendered than others (with for instance few women in beta sciences and more in the 'soft', alpha and gamma sciences). It has in addition been suggested that women with underage children perform more household chores, so that they have fewer hours in the evenings and weekends to work on their careers. All these

factors are possible confounders: they co-vary with gender and they may be the real explanation for, the driving force behind, the difference in hourly wages.

One way to solve this conundrum is to carry out a multiple regression analysis. By including all these possible confounders in a regression model, together with the variable gender, one will be able to estimate the impact of gender on hourly wages *net of* rank, number of articles published, age, discipline, caring for underage children, working part-time.

Now let us assume that we have a dataset at our disposal with data on 2,000 academics, men and women, and that we have data on all these variables. We run three analyses: one with only gender, one with gender and academic variables added (rank, discipline), and a third where we add the remaining variables to the model (number of articles published, age, caring for underage children, working part-time). It should be noted that some variables will overlap: age will be strongly associated with caring for underage children or not, as will be rank. There is therefore bound to be some multicollinearity (on top of overlap of variables with gender). But that need not bother us much: we are interested to filter out as many alternative explanations for the lower salaries of female academics. Because the sample size is so large, we will not inspect the R^2_{adj} (it is likely to be barely affected by the inclusion of more predictor variables); instead we will focus on the R^2. The fictitious analysis results are in Table 9.2.

Table 9.2: Unstandardized regression weights of gender and hourly wage in euros

Variable	Model I	Model II	Model III
Gender (0=male, 1=female)	-.28**	-.18**	-.13*
Rank (asst. ref)			
associate		.20**	.17**
professor		.30**	.25**
Discipline (beta. ref)			
alpha		-.05	-.04
gamma		-.02	-.01
No. yearly articles		.02	.01
Age			.03*
Underage children			-.06*
Part-time			-.02*
R^2	.25	.39	.47
** $p<.01$, * $p<.05$			

From these analyses, we can conclude the following. In Model I, we predicted salaries of male and female academic staff from gender only. The R^2 equals 0.25, showing that 25% of the variation in hourly wages can be predicted from an academician being male or female. This is in an absolute sense not very high, but even so remarkable: a regression model with just one variable predicting that well is rare. The

regression coefficient is negative, meaning that those with higher scores on the variable gender (i.e. females) are predicted to earn less per hour. As the b is unstandardized, we may infer that for every euro that a man earns, a woman earns 72 eurocents.

Model II includes two important additional predictor variables, namely an academician's rank (likely very important as rank dictates the bandwidth of salaries) and the discipline in which people are employed. Rank is an ordered variable with assistant professor – the reference category here – the rank where salaries are likely lowest, followed by associate professor and at the top full professor. Discipline is a nominal variable (with three categories: beta, alpha and gamma). Both variables are entered as so-called dummy variables. For rank we construct two variables: the first dummy whether someone is associate professor or not and the second dummy whether someone is a full professor or not. Someone who scores '1' on the first dummy variable is an associate professor; someone who scores '1' on the second dummy variable is a full professor. If someone scores '0' on both dummy variables then that person is an assistant professor. The same trick is used for the nominal variable discipline: two dummy variables are created, one for working in an alpha discipline or not and the second for working in a gamma discipline or not. We see that indeed these two predictors are quite important: the R^2 increases to 0.39. We also see how associate professors are predicted to earn 20 cents more per euro than assistant professors, and full professors even 30 cents more than assistant professors. Discipline is not significant. We see how the regression coefficient for gender is lower than in Model I: over and above effects for rank and discipline, women now earn 18 cents per euro less than men. This means that some of the 28 cents difference that we saw in Model I was attributable to women working in lower ranks and/or in disciplines where salaries are generally lower.

In Model III, four more variables are added. We see that the R^2 again increases substantially. In this final model, we are able to explain 47% or almost half of the variation in salary with our regression model. The number of yearly articles published is not significant. Age is significant and the model shows that every year adds 3 cents to people's salary. Having underage children is also significant, with those caring for underage children earning 6 cents less than those without such care tasks. Note that both men and women could care for underage children and would then be similarly affected in their salaries. Lastly, working part-time also adds significantly to the prediction, and leads to a decrease by 2 cents in predicted salary. We note how the regression weight for gender has decreased even further: it is now -.13. We summarize this by saying that over and above the impact of rank, discipline, number of articles, age, childcare and working part-time, an effect of gender remains. The 13 cents less that women receive are on top of their rank, discipline, number of articles, etc.; the effect of gender is what remains after we have taken all these other factors into account.

Does this analysis prove that women are discriminated against? It does not. It may be that other confounders still play a role, factors that co-vary both with gender and salary, that we did not incorporate in our model. If we have indeed missed important variables, we have not specified a correct model (this is also referred to as omitted variables bias), and we should continue our search and extend and improve our model. What we can be sure of is that the 13 cents-effect that remains in this analysis cannot be attributed to any of the common explanations for the pay gap between men and women we included in our model.

9.3 Analysis of variance model

In an analysis of variance model, we also attempt to predict the scores on a dependent variable Y from the scores on one or more independent variables X. Here, however, the X are not interval variables, but nominal. This means that for envisaging the model, the same conceptual representation as that for regression analysis (see Figure 9.1) can be used, but the X variables are a different kind of variable, they have a different, nominal measurement level.

Analysis of variance is in that sense a fairly straightforward analogue of regression analysis. Very often, in addition to testing whether the independent variables contribute significantly to the prediction of Y, the impact of so-called interaction effects is assessed, the impact of specific combinations of categories of the independent variables. Analysis of variance has numerous other attractive options too, like the possibility to analyse particular designs, such as designs where respondents have been observed repeatedly, or designs where respondents have for instance been offered tasks to complete in a systematic manner.

The notation that we employ in analysis of variance is slightly different from the notation in regression analysis – independent variables are in many textbooks not written as X_1, X_2, X_3, ... anymore but as A, B, C – and the model specification appears different from what we were used to. Also, the format in which results are reported differs.

Analysis of variance is very widely used, particularly in evaluation research, such as for assessing the impact of a certain therapy. In research using so-called 'vignettes', it is the standard analysis technique: in vignette studies respondents are offered – usually in a revolving or randomized way – different alternative scenarios and asked for their assessment of a dependent variable embedded in that scenario. An example is where researchers artificially construct court files, in which certain aspects are randomized such as the sex of the offender or the type of offence – all nominal variables. See also section 5.7. Analysis of variance detects what characteristics of these court cases and the respondents best predict answers on the vignette's dependent variable.

9.3.1 When to employ analysis of variance

Analysis of variance – often abbreviated as AN(alysis)O(f)VA(riance) – is a suitable analysis technique whenever we are interested in finding out the impact of one or more (combinations of) nominal independent variables on the scores on a dependent variable. The categories of the independent variables are treated as nominal, unordered categories. So if the independent variable were to be disciplinary background, ANOVA would investigate the impact of each of the separate categories (hard sciences, social sciences, law, humanities) on the dependent variable Y.

This means that when an independent variable is an interval level variable, it is not efficient to use ANOVA. ANOVA would treat each value of the interval level variable as a separate category. If we had for instance an independent variable 'age', ANOVA would treat each and every category (18, 19, 20, ... 65) as a separate category. This would not only make inspection and interpretation quite cumbersome, but there is also loss of information, as the technique does not reckon with the fact that the categories are

ordered and equidistant. So, when an independent variable is an interval level variable, it is more efficient to employ a model (such as the regression model) that treats the independent variable at the measurement level it has.

Notwithstanding the above, in some cases researchers do want to use the ANOVA model and have interval level variables that they want to incorporate in the model. If the independent variable has many categories, it is then often re-coded into a smaller number of categories such as '18–22', '23–30', '31–40', and the re-coded variable used as a nominal variable in the analysis. One important reason to do so is that one suspects that the impact of age on the dependent variable may not be linear. It may for instance be the case that one thinks that labour market orientation is pretty low when people are still studying, that it is higher as people need to find a job after graduating, between the ages of 23 and 30, but that it declines somewhat as people are in the age group where they have settled in a job, are starting a family, etc.

Analysis of variance is a technique that already gives good results with small sample sizes. Formally, it is assumed that the dependent variable Y is normally distributed. However, it has been shown that even with fairly substantive deviations from normality, the technique still performs well. It is, we say, *robust* against such deviations from normality. Analysis of variance is easy to interpret as well. For all these reasons, it is widely used. Some of the possibilities it has (like incorporating interaction effects) are also possible to include in other analysis techniques but are then less easily implemented.

9.3.2 Analysis of variance proper

In analysis of variance, we attempt to investigate whether the scores on an interval dependent variable Y can be predicted from one or more categorical variables. As said, the independent variables that we have so far named X generally appear under a different name. Variables are named 'A' and 'B', etc. – although they may be named 'Gender' or 'Type of Claim' too.

Starting with the simple example where there is just one independent variable A, its categories are denoted by α. If A has two categories, we have α_1 and α_2; if it has three categories there is α_1, α_2 and α_3. The analysis of variance model is now construed as follows:

$$Y_{ij} = \mu + \alpha_j + e_{ij},$$

in which Y_{ij} is the score on the variable Y for the i-th person in the j-th category of A, and α_j is the 'effect' of the j-th category of the independent variable. The term μ is an overall average, a sort of starting point for the building of the model, just like the coefficient 'a' in regression analysis. So, every respondent starts from the same μ. If a respondent is in category 1 of the independent variable A, then for that respondent α_1 applies, if a respondent has scored category 2, α_2 applies, etc. If the total number of respondents studied is N, then in each of the k categories of A we have n respondents (and, assuming for simplicity equal cell sizes: $N = n \times k$). In ANOVA, α is computed such that the sum of all the α's over the respective k categories or so-called 'levels' of A is zero:

$$\sum_{j=1}^{k} \alpha_j = 0.$$

Obviously, not all Y scores can be predicted perfectly in this way, so also for ANOVA, just like we did in regression analysis, we need an error term to capture the difference between the observed and the predicted Y scores. The predicted Y is:

$$\hat{Y}_{ij} = \mu + \alpha_j.$$

The error terms are defined as the difference between the predicted and observed Y, just like in regression analysis:

$$e_{ij} = Y_{ij} - \hat{Y}_{ij}.$$

And, just like in the regression model, here too the error terms add up to zero, within each cell and over all levels of A:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} e_{ij} = 0.$$

How do we determine the error term in practice? Suppose that we see that Y_{13} = 10, and we have μ = 7, and α_1 = 2, then given that:

$$\hat{Y}_{13} = \mu + \alpha_1 = \ 7 + 2 = 9.$$

we know that

$$e_{13} = Y_{13} - \hat{Y}_{13} = 10 - 9 = 1.$$

In words, given that μ equals 7, and α_1 equals 2, the error term e_{13} equals 1. If we had a value of 6 for Y_{14} we would conclude that e_{14} equals -3: 7 + 2 + (-3) = 6. The estimate of Y_{13} is slightly too low and the error term is 1; the estimate of Y_{14} is too high and the error term is -3. In this manner all observations are decomposed into a part constituting the mean score (μ), an effect α_1 attributable to the fact that someone is in the first category of A (A_1), and personal deviations from the ensuing estimates captured by the error terms.

So, once we have values for μ, for the respective α_j's and for all the error terms e_{ij}, how can we determine whether the independent variable A contributes significantly to the prediction of Y? Clearly, if the respective α_j's are very large, this would more likely be the case: then, it matters a lot whether someone is in one or the other category of A. And if the α_j's are all very close to zero, it is less likely that we will find a significant impact of A, as then the predicted scores are quite close to the mean μ. We will explain conceptually the reasoning that is used to determine whether an effect is considered significant. We do this with an example.

Suppose that I have developed a calculus course for those who become fearful when presented with numbers. Before the course, the participants scored on average a '5' for calculus. After the course, they score on average a '7'. Not all had the same scores,

Figure 9.4: ANOVA example: improvement in calculus before and after course

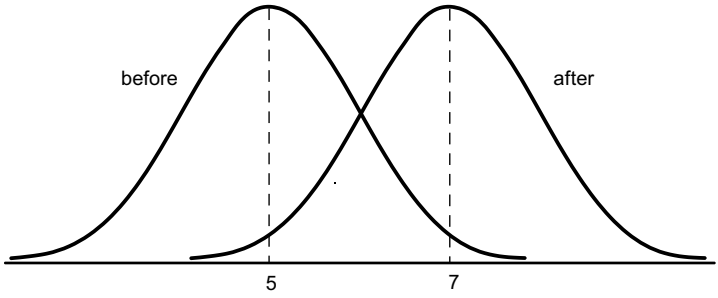
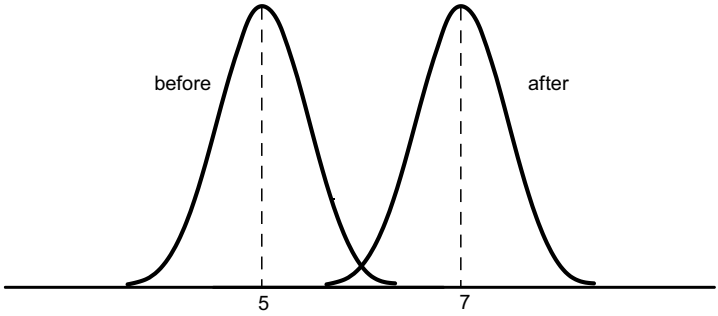


Figure 9.5: ANOVA example: calculus improvement with less variability



there was some variability, but we can say that the average improvement was 2. This situation is depicted in Figure 9.4.

The figure shows how indeed the average scores improved, from an average 5 to an average 7. The figure also shows that there is quite some variability: some respondents still had – after finishing the course – low scores, whereas some already had a '7' before the course started. Compare for that matter this figure with Figure 9.5. Here a situation with much less variability is depicted. Respondents here had scores that clung much more around the mean: almost all respondents had scores really close to 5 before the training, and almost all had scores around 7 after the course. The scores fluctuate much less.

Now if you were to ask someone to tell you which result s/he would find more meaningful – the one in Figure 9.4 or the one in Figure 9.5 – almost everyone would tell you – without any formulas or complicated reasoning – that Figure 9.5 is the stronger result. This is remarkable, because the net difference is the same in both examples: in both, people improved 2 points on average. However, the last figure is definitely the stronger one. Why is this so? One way of saying this is by noting that, while in the first figure some people already had good marks before the training and quite a few still had bad marks after the training, in the second figure, almost all respondents improved. In

the second example, the improvement was so to speak 'en bloc' – almost all had bad marks before the training and almost all had improved afterwards. Formulated more mathematically, we find the second figure indicative of a stronger result because – even though the net average improvement is the same – it is much larger gauged against the smaller variability in the scores.

What analysis of variance does is nothing more than give you the same intuitively appealing conclusion – except for the fact that it uses formulas and gives you exact significance of results. What analysis of variance does is to evaluate the differences between a number of 'conditions' (before and after training) against random variability in the data (the e_{ij}). If we were to sum all the α's and the result turns out much larger than the sum of the e_{ij} (a situation that we clearly see in Figure 9.5), we believe the impact of A is telling, 'significant'. Conversely, if the α's and the e_{ij} are of similar magnitude, we would say that the impact of A is in fact similar to differences you might expect just by chance. This is what you see in Figure 9.4.

How then exactly do computations go? We suggested just now that one could compare the sum of the α_j with the sum of the e_{ij}, but of course both sum to zero (see above). To solve for this, in ANOVA, for each respondent his or her α_j is squared, and all squared terms summed. Recall that each respondent's α_j is the part of his or her score that is attributable to him or her being in a particular category of the independent variable A. The result is called 'SS_A', the sum of squares (SS) of all terms pertaining to the independent variable A, i.e. the variation in the Y scores that is attributable to the categories of the independent variable A:

$$n \sum_{j=1}^{k} \alpha_j^2 = SS_A.$$

Next, for each respondent his or her error term is squared and all these squared terms again summed. Recall that each respondent's e_{ij} is the part of his score that is attributable to chance. The result of adding up the squared terms is called 'SS_{error}', the sum of squared error terms, i.e. the variation in the Y scores that is attributable to chance:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} e_{ij}^2 = SS_{error}.$$

The idea is that if SS_A is much larger than SS_{error}, we judge the impact of A as important. If, however, SS_A is comparable to SS_{error} or even smaller, we judge the effect of A not to be much larger than chance and therefore as not very telling.

However, before this comparison can be made, first both SS_A and SS_{error} have to be divided by the so-called 'degrees of freedom', terms that take into account the fact that SS_A will be larger if we have more levels of A, and that SS_{error} will similarly be larger if we observe more respondents. It would be an unfair comparison if we did not correct for that in some way: we might then find that with exactly the same variability, and exactly the same mean difference, one test turns out to be significant while the other

does not – only because of having more or fewer respondents. Correcting for such issues, we compute the so-called mean sums of squares (MS) for both A:

$$\frac{\mathrm{SS_A}}{\mathrm{k}-1} = \mathrm{MS_A},$$

and for the error terms:

$$\frac{\mathrm{SS_{error}}}{\mathrm{k}(\mathrm{n}-1)} = \mathrm{MS_{error}}.$$

Once these terms $\mathrm{MS_A}$ and $\mathrm{MS_{error}}$ have been computed, we have arrived at the point where we can finally do statistical testing. This is done by computing the ratio of these terms, just like we intuitively did in the figures above:

$$\frac{\mathrm{MS_A}}{\mathrm{MS_{error}}} = F_{\mathrm{A}}.$$

So, analysis of variance is exactly what its name says: analysis of the variance in the data, and more precisely an evaluation of the variability of the scores due to the independent variable A ($\mathrm{MS_A}$) against variation due to chance ($\mathrm{MS_{error}}$). To get exact significance statements, ANOVA tests the ratio of these terms by employing a statistical distribution, the F-distribution.

The F-distribution differs from the normal distribution, but the principle for using it in testing is the same as the principle we have been discussing all this time. The null hypothesis is that there are no differences between the levels of the independent variable (hence: *all* α_{j} are equal to zero), and using the F-distribution it can be inferred what the chances are of finding our results under this hypothesis. If that chance is very small, we declare the finding incompatible with the null hypothesis and reject it. We then switch to the alternative hypothesis, namely that there are categories of A that have an effect on the dependent variable. We then say that A adds significantly to the prediction of Y. Again, we can test at a 1%, 2%, 5% level; as usual we are the ones as researchers to decide what chance we feel comfortable with for taking the wrong decision if $\mathrm{H_0}$ is true.

Table 9.3: Example ANOVA results

source of variation	SS	df	MS	F	*sign*
A	48	2	24.0	15	1%
error	24	15	1.6		
total	72	17			

This F-value is the test statistic, just like the z-value was in the examples in chapter 8. How significant a certain F-value is, given the degrees of freedom for A and the

error terms, can be looked up in a table online or in a textbook, and statistical software packages do that for the user. In general, the larger the value of the F-statistic, the smaller the likelihood of the result under the null hypothesis. Table 9.3 gives the analysis of variance results for a fictional example. The F-statistic is here, as we see, highly significant. Although depending on sample size and characteristics of the variables, in practice, F-statistics of around 8 and higher are (almost) always significant. (Indeed, it would be highly counterintuitive if an effect due to the independent variable that is 8 times larger than chance were not regarded as significant! This means that we conclude that there are significant differences in the value of Y between the categories of the independent variable A, meaning that A contributes significantly to the prediction of the dependent variable.

Note that we only know *that* there are significant differences between the categories of A. If for instance the dependent variable were the number of typing errors that people make, and if respondents in category A_1 consumed no alcoholic drinks, those in A_2 consumed one to two alcoholic drinks, and those in A_3 consumed three or more alcoholic drinks, we cannot conclude that people who consume more alcohol make more typing mistakes. We do not know this – we only know that differences have been measured that are very unlikely to occur by chance if it were irrelevant whether someone consumes no alcohol, one to two, or three or more glasses of alcohol. The ANOVA test tells us only that – overall – there are significant differences between the categories or *levels* of the independent variable.

As said, analysis of variance or ANOVA is widely used in many areas of social research. Numerous extensions are possible to the simplest design we just discussed. However, in each design, the principle is always the same. We inspect the variability attributable to differences between categories of one (or a combination of) independent variables. We gauge that against the random fluctuation. If the differences in scores due to the independent variable (the differences between categories of the independent variable) are much larger than the fluctuation due to chance (differences within categories of the independent variable), we conclude that the independent variable has an impact on the dependent variable.

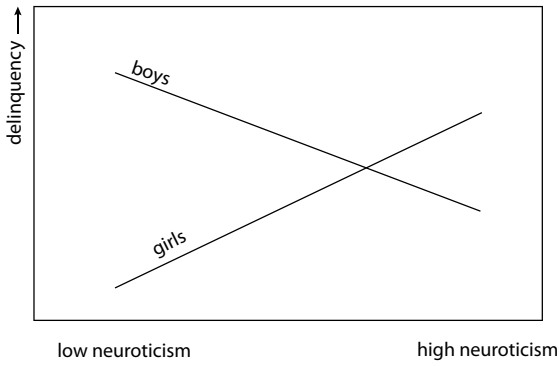
For ease of understanding, we have here always referred to the independent variable as 'A' and denoted its corresponding sums of squares as 'SS_A'. Often, the independent variable is labelled as 'Between', as it reflects the variation between categories of the independent variable. Correspondingly, the error variation is then referred to as 'Within', as this reflects the variability within the categories of A.

9.3.3 Factorial analysis of variance

Now the same can be said for analysis of variance that we said for regression analysis: a model with just one predictor is not likely to be a model that gives very good results and is likely to be too simplistic for our purposes. So, just like in regression analysis, also in ANOVA additional independent variables are often used for the prediction of Y. One example of such a model is:

$$\mathrm{Y_{ijkl}} = \mu + \alpha_\mathrm{j} + \beta_\mathrm{k} + \gamma_\mathrm{l} + \mathrm{e_{ijkl}},$$

Figure 9.6: ANOVA example: interaction effect



for the situation where we have three independent variables. This is called a *factorial* analysis of variance. In such a model, we measure the impact of several independent variables on the dependent variable. ANOVA has the option to not only include additional predictors, but to also investigate whether these variables *interact*. What is meant by that? Again we illustrate this with an example.

In research on delinquency it has been found that the psychological trait 'neuroticism' does not predict delinquency. Those who are highly, average, or below-average (all categorized into norm scores) neurotic all have approximately equal delinquency levels. However, if we disaggregate these findings for boys and girls, we find that for boys neuroticism is a *protective* factor: boys who are above-average neurotic have lower delinquency scores. For girls, on the other hand, we find the opposite: neurotic girls have increased delinquency scores, and neuroticism is in fact a risk factor for them. So, the overall result of 'no difference' was actually attributable to a levelling out of these opposite effects between boys and girls. There is heterogeneity. See Figure 9.6, which sketches the situation.

As can be inferred from this figure, there is only a slight difference between the delinquency levels of those who are highly neurotic and those who score low on neuroticism: if we were to average the scores, taking boys and girls together, we would find an almost horizontal line. There is a clear difference between delinquency levels of boys and girls: if we were to average the scores for boys and those for girls, we would see that boys have on average higher delinquency levels than girls. These are what is called main effects: the effects of the independent variables by themselves (here gender and neuroticism). In this example there is a clear main effect of gender. The main effect of neuroticism is minimal.

However, in the figure we see a so-called interaction effect of gender and neuroticism. For boys, higher scores on neuroticism predict lower delinquency levels. For girls it is the other way round: for girls higher scores on neuroticism predict higher delinquency levels. This means that gender affects the way in which neuroticism impacts delinquency levels: gender alters the effect of neuroticism on delinquency. When there is such an interaction effect, it is said that one independent variable *moderates* the

effect of the other independent variable. Gender is then called a moderator variable. The factorial model with such an interaction effect is then written as:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + e_{ijk},$$

with α and β now referred to as *main effects* and the term $\alpha\beta$ referred to as the interaction effect between A and B. Such interaction effects are often written as A $\times$ B, or even 'Neuroticism $\times$ Gender'. A hypothetical example of a table with results for this analysis of variance is given in Table 9.4.

Table 9.4: ANOVA example with interaction terms

source of variation	SS	df	MS	F	sign
neuroticism	3	2	1.5	$F_A = 0.75$	< 1
gender	10	1	10	$F_B = 5$	5%
neuroticism $\times$ gender	36	2	18	$F_{AB} = 9$	1%
error	68	34	2		
total	107	39			

The table confirms what we inferred already from the figure. There is a – significant, it turns out – main effect of gender. The effect of neuroticism is non-significant (in fact any F-value below 1 is always non-significant, as then the variation in scores of the dependent variable due to the categories of the predictor variable is less than chance variation). The interaction effect is highly significant at the 1% level.

More complicated models are possible with interaction effects between more than two variables:

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + \alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl} + \alpha\beta\gamma_{jkl} + e_{ijkl},$$

where $\alpha\beta_{jk}, \alpha\gamma_{jl}$ and $\beta\gamma_{kl}$ would be labelled '1st order interaction effects', and $\alpha\beta\gamma_{jkl}$ would be labelled the '2nd order interaction effect'. Just as we had in the univariate case a test for F_A, now we have separate tests for F_A, F_B, F_C, F_{AB}, F_{AC}, etc.

Models and computations become increasingly complex with more independent variables. The principles, however, remain the same. Each time we test differences due to (combinations of) categories of variables against chance variation.

Here too, for the analysis of variance model, the same difficulties may arise that we sketched for the regression model, and for models in general. As models become more complicated, they become harder to interpret: how should we for instance interpret a 5th order interaction effect? In addition, as more independent variables and more interaction terms are added to the model, the cells along which the squared error terms are added become smaller and the error variance (the variability within the groups formed by the combination of categories of the variables) decreases too. This inflates the F-statistics. As such – to prevent opportunistic addition and spurious inflation of

these F-statistics – it is the rule that independent variables and interaction terms be added to the model only when there are strong theoretical reasons to do so.

9.3.4 Tests

We discussed tests for the analysis of variance model. For each main and interaction effect, an F-test is carried out, which tells you what the likelihood is of finding the results if the data come from a population with no differences between the various categories of (the combinations of) the independent variables. As we said before, such tests tell you whether there are *any* differences between the (combination of the) categories of the independent variable(s). If you were interested in particular differences, say whether dosage p and dosage q of a drug differ significantly in migraine incidence, it is possible to carry out specific so-called contrast tests.

9.3.5 Model fit

When doing analysis of variance, it is not customary to present measures of model fit. Such measures are available, however. The most often used is ω^2, which is the analogue of R^2 in the regression situation. As such, ω^2 tells you how much of the variance of the dependent variable is explained by the entire ANOVA model. It is also possible to express the effects of the respective factors in a so-called η, with the η for each factor reflecting how much variance is explained by the respective factor.

We end this section on ANOVA with an example. Our data come from a study on workplace misbehaviour (Wesselius et al., 2023). Using vignettes, a hypothetical case was presented of an employee who sexually assaulted a colleague at a company New Year's party. The story narrated how the employee was fired by the company after complaints by the victim. Respondents were asked how 'just' they rated this outcome, on a scale from '1' to '5', with '5' indicating the respondent was in complete agreement with the decision and '1' the opposite. In the vignette, four factors had been varied: whether the perpetrator was male or female, whether the victim was male or female, whether the victim was married and had a child (an indication of the 'respectability' of the victim, following on from Nils Christie's (1986) ideal victim theory) or was single, and whether the perpetrator was a colleague or the superior of the victim. In addition to these manipulated properties, the researchers recorded the respondent's gender, age, employment status and whether the respondent knew someone who had been similarly victimized. A total of 240 respondents filled out a small questionnaire. The sample was a convenience sample.

The authors had expected that the company response (firing the perpetrator) would be rated as more just when the victim was female, the perpetrator male, the victim married and the perpetrator the victim's superior.

The authors investigated the main effects of factors they had manipulated: victim and perpetrator gender, relation between victim and perpetrator, and victim marital status. In previous exploratory analyses, they had uncovered fairly strong gendered effects, so they also included the first- and second-order interaction effects for victim, perpetrator and respondent gender. The results are in Table 9.5.

Table 9.5: ANOVA example on workplace misbehaviour

source of variation	SS	df	MS	F	sign
gender victim (v)	.934	1	.934	5.750	.132
gender perpetrator (p)	3.751	1	3.751	12.482	.153
relation	6.799	1	6.799	6.571	.011
respectability	.768	1	.768	.742	.390
categorized age respondent	14.899	5	2.980	2.880	.015
personally know victim	.554	1	.554	.535	.465
$\text{gender}_\text{v} \times \text{gender}_\text{p}$	.394	1	.394	.097	.808
$\text{gender}_\text{v} \times \text{gender}_\text{r}$	.117	1	.117	.028	.894
$\text{gender}_\text{p} \times \text{gender}_\text{r}$	.288	1	.288	.070	.835
$\text{gender}_\text{v} \times \text{gender}_\text{p} \times \text{gender}_\text{r}$	4.088	1	4.088	3.951	.048

Source: Wesselius et al. (2023)

The analysis of variance shows that three factors are significant: the business relation between the victim and perpetrator, the age of the respondent, and the second-order interaction effect between the gender of the respondent, victim and perpetrator. Of the manipulated factors, gender of the perpetrator and gender of the victim were in the anticipated direction, but the differences were clearly not large enough for the factor to emerge as significant. Respectability of the victim played only a very minor role, with differences between perceived justness for married and single victims only marginal. Personally knowing someone who had been similarly victimized also played a non-significant role.

Inspecting the significant effects, the authors report that respondents indeed evaluated the company decision as more just when the perpetrator was hierarchically superior to the victim. For age of the respondent, they inspected the category means, and found them to be almost strictly monotonically decreasing, with means as given in Table 9.6.

The authors concluded that older respondents more often disagreed with the company measure, a finding which was supported by qualitative comments that the respondents had been asked for. Obviously, ANOVA only states overall significance, but does not indicate which categories of the variable generate significant differences. From the means, it appears that there is somewhat of a watershed from age 33, and subsequent contrast tests should reveal whether that is indeed the case, that is, whether those below and above 33 indeed judge the company decision significantly differently.

The interaction effect between the genders of the perpetrator and the victim was also insignificant; this shows that specific combinations of male or female victims and perpetrators were not assessed differently by the respondents. The same goes for the other two first-order interaction effects: male and female respondents do not judge

Table 9.6: Category means for age

	mean	n
< 24	4.28	54
24 - 26	4.22	37
27 - 33	4.20	25
34 - 51	3.89	61
52 - 60	3.62	26
> 60	3.65	37
Total	3.99	240

Source: Wesselius et al. (2023)

male or female victims, or male or female perpetrators, differently. All in all, it would therefore appear that gender does not play a very large role.

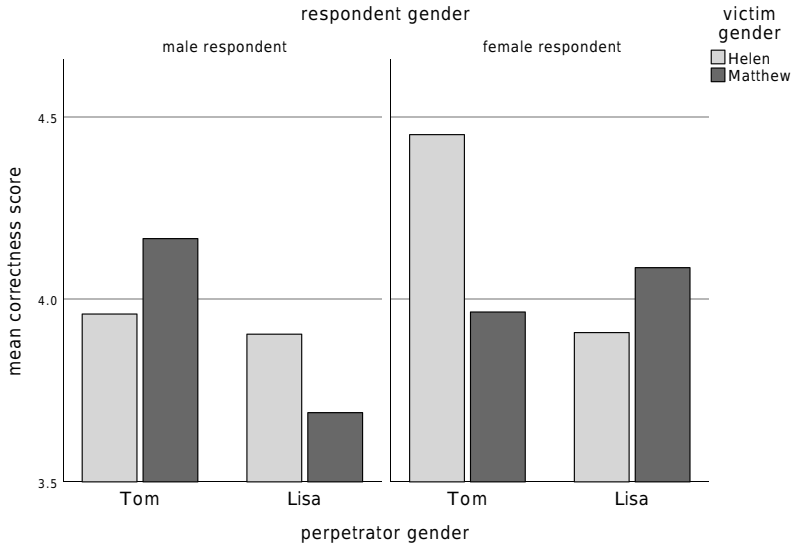
The second-order interaction effect is however significant. The authors rephrase this by saying that male and female respondents do judge the different constellations of gender of victim and perpetrator differently, and they illustrate this through a graph. In the administered vignettes the male and female victims were called Matthew and Helen respectively, and the male and female perpetrators Tom and Lisa; this is how they are indicated in the graph (Figure 9.7); note that the y-axis is shortened – it starts from 3.5.

The figure shows how female and male respondents indeed attach different evaluations to the company measure to fire the misbehaving employee. Firstly, we see that male respondents agree least with the firing of the misbehaving employee when the vignette had a female perpetrator (Lisa) sexually assaulting a male employee (Matthew). The average score is lowest here. The constellation in which male respondents agree most with firing of the culprit is where a male has sexually assaulted a male.

For female respondents, the picture is quite different. For them, there is one constellation that jumps out, and that is a male perpetrator sexually assaulting a female. This has by far the highest agreement score, and the other constellations differ only marginally in how the company measure is then evaluated. For establishing what combinations of the interacting variables differ significantly, additional contrast tests would be needed.

The authors interpret this as men and women evaluating '#MeToo'-like situations differently. Supported by qualitative remarks made by respondents, they argue that men find situations with a female accosting a male relatively innocuous. Some male respondents stated that the male victim could easily have solved the issue himself, or that if the respondent himself were to receive such sexual attentions, he would be flattered. Women on the other hand disapprove particularly of 'classical' sexual harassment contexts, in which a male harasses a female, possibly as such situations are most common and realistic for women.

Figure 9.7: ANOVA example: mean justness for different gender constellations



Source: Wesselius et al. (2023)

9.4 Wrapping up

It may seem odd, after the fairly different formulas, to reiterate that regression and analysis of variance models do essentially the same thing. Nevertheless, that is the case. The analysis of variance model can be seen as a more general case of the regression model, or the regression model as a special case of the analysis of variance model. We will not go into the mathematical details here, and simply note that it is important not to forget the similarities: in each case we are trying to predict a dependent variable from one or more independent variables. In each case the model is linear. While we discussed interaction terms for analysis of variance only, interaction terms can also be incorporated in regression models. And while we stated that regression models should be used when the independent variables are interval level or higher, it is also possible to incorporate nominal variables in regression models. The difference between the two techniques is in the manner in which the measurement levels of the independent variables are accommodated.

Both regression analysis and ANOVA are widely used techniques in quantitative empirical legal research, with regression analysis the more ubiquitous as it is so flexible

and adaptable. Many extensions of regression analysis exist (for example, the survival models we discussed in section 6.5 also exist in regression format, so that several independent variables can be used to predict the dependent time until event). Regression analysis is also used often in content analysis of qualitative material – see section 7.3, where we discussed the study by Mascini & Holvast (2023) in which regression analyses were employed to predict writing styles. ANOVA is particularly useful for the analysis of data obtained through vignette studies in which researchers systematically manipulate factors.

9.5 Further reading

A classic (almost a Bible) on analysis of variance is Kirk (1968). For regression analysis, we recommend Cohen et al. (2002).

Chapter questions

1. What three properties make models useful? (section 9.1)
2. Give a number of synonyms for independent and dependent variable (section 9.1.1)
3. What are independent and dependent variables in the theory of procedural justice? (section 9.1.1)
4. What is the difference between a mediating and a moderating variable? (section 9.1.1)
5. List all parameters in a regression model with five predictor variables (section 9.1.2)
6. What is meant by a parsimonious model? (section 9.1.3)
7. Give two reasons why a regression model needs approximately 10 times as many replications as variables (section 9.1.3)
8. Under what conditions is regression analysis feasible? (section 9.2.1)
9. What are advantages and disadvantages of standardized versus unstandardized regression weights? (section 9.2.2)
10. What is multicollinearity and what can and should be done to prevent it? (section 9.2.3)
11. Give a number of reasons why R^2_{adj} is to be preferred over R^2 (section 9.2.4)
12. What does a large difference between R^2_{adj} and R^2 signal? (section 9.2.4)
13. What ratio is assessed using the F-statistic? (section 9.3)
14. Why is it that an F-value lower than 1 is never significant? (section 9.3)

15. What is factorial analysis of variance? (section 9.3.3)
16. Give an example of an interaction effect (section 9.3.3)
17. What are the differences and similarities between regression analysis and ANOVA? (section 9.4)