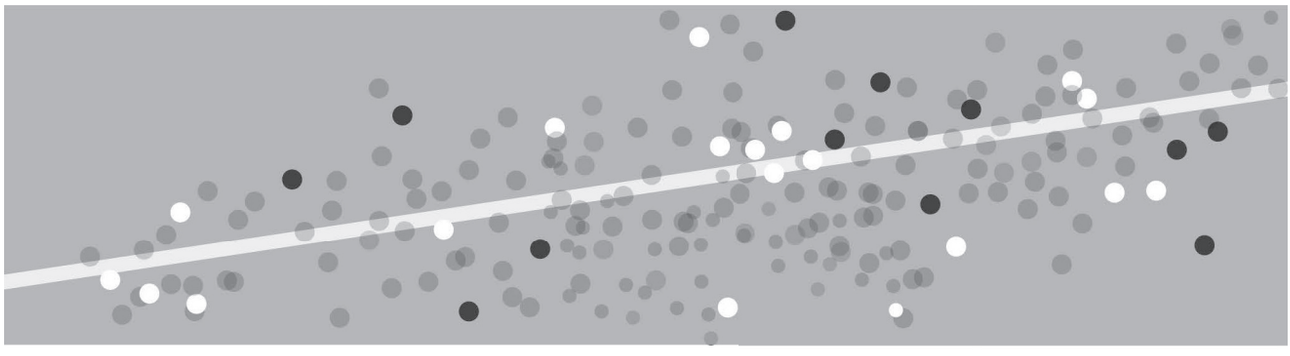
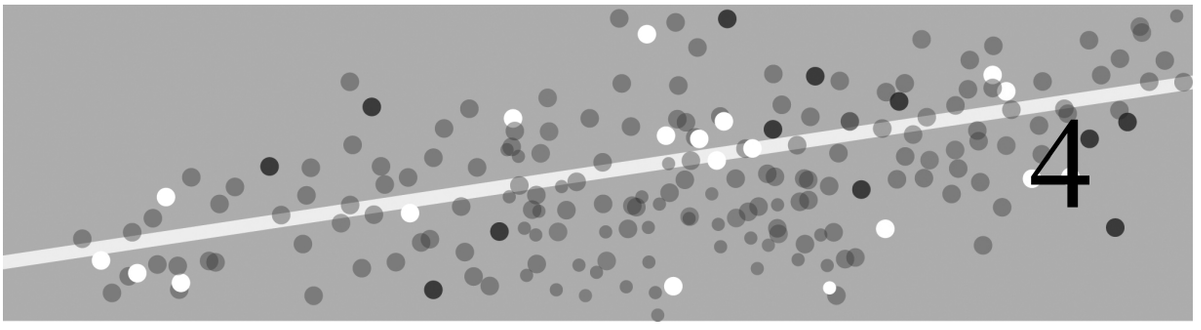
The SAGE Handbook of Regression Analysis and Causal Inference



Edited by
Henning Best and
Christof Wolf

SAGE reference

Los Angeles | London | New Delhi
Singapore | Washington DC



4 Linear regression

Christof Wolf and Henning Best

INTRODUCTION

In this chapter we first present the basic idea of linear regression and give a non-technical introduction. Next we cover the statistical basis of this method and discuss estimation, testing and interpretation of regression results. The third section is devoted to the presentation of an example analysis. In closing, we first discuss issues related to the causal interpretation of OLS regression coefficients and then mention some general problems encountered in linear regression and recommend further reading.

In science we often are interested in studying hypotheses of the form 'the more X, the more/less Y', for example 'the higher the education of a person is, the more willing s/he is to accept immigrants'. Thus, we assume that acceptance of immigrants is partly determined by education, or more technically that the acceptance of immigrants is a function of education. We can express this idea mathematically as

$$\text{Acceptance of Immigrants} = f(\text{Education}) \quad \text{or} \quad y = f(x).$$

If we choose a linear function $f(\cdot)$ to link y with x_1 the result is a linear regression model. Alternative link functions result in other regression models; some of which are discussed in Chapters 8 and 9 of this volume. Expressed as linear model, we get

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{4.1}$$

where y is referred to as the dependent (endogenous) variable and the x as an independent (exogenous) variable or predictor. This equation can be seen as a specification of our hypothesis 'the higher the education of a person is, the more willing s/he is to accept immigrants'. The specification consists of the assumption that we have a linear effect of education on accepting immigrants of size β_1. This means we assume that if education increases by one unit the acceptance of immigrants changes by β_1 units. It is also important to note that if we specify a linear effect as in equation (4.1) we assume that the effect of education is the same for any given level of education, that is, the effect is constant throughout the range of x.

Equation (4.1) contains an element that has not yet been introduced. The term ε is referred to as an error term or residual. It is equal to the difference between the observed values of y and the

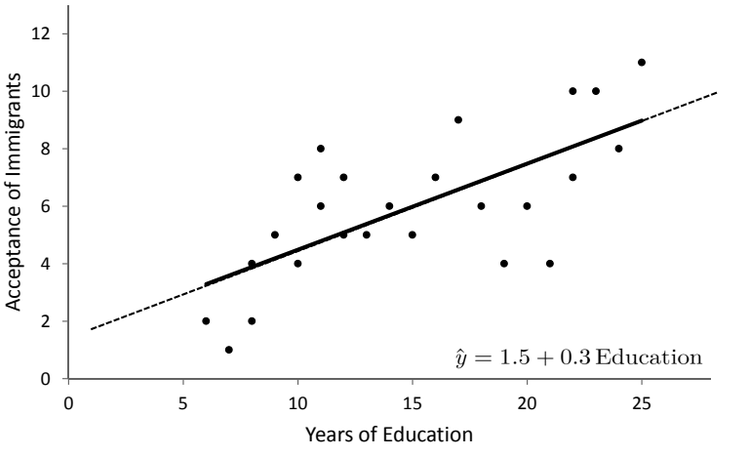


Figure 4.1 Scatter plot with linear fit line

values predicted by the independent variable(s). Subtracting ε from both sides of equation (4.1) yields

$$y - \varepsilon = \hat{y} = \beta_0 + \beta_1 x, \tag{4.2}$$

where $\hat{y}$ denotes the predicted values of y.

Figure 4.1 illustrates the relationship between education and acceptance of immigrants for 25 people, where education is measured in years of full-time education and acceptance is measured on an 11-point scale ranging from 1 (no acceptance) to 11 (full acceptance). Each dot in this scatter plot represents one person in this property space. The line drawn in Figure 4.1 represents the linear relationship between education and acceptance; it follows the equation $\hat{y} = 1.5 + 0.3x$. Here, $\beta_1 = 0.3$ implies that a person with one more year of education on average scores 0.3 points higher on the acceptance of immigration scale. Figure 4.1 also illustrates why β_1 is referred to as the slope: it determines how shallow or steep the regression line is. β_0, in this example equal to 0.5, is called the intercept, since it equals the value of y at which the regression line 'intercepts' or crosses the y-axis. In our little example the intercept of 0.5 can be interpreted as the predicted value for a person with zero years of education. The problem with this interpretation is that we have no data for this range of values of the independent variable and thus we should abstain from making any predictions. Of course, this is also true for the other end of the distribution. For example, we would predict a value of $\hat{y}\,(x = 50) = 16.5$ for someone with 50 years of education, which is an impossible value given the current measurement of acceptance with a maximum value of 11. In general, we should restrict our analysis and interpretation of results to those areas of the property space for which we have empirical data. This range is implied by the solid regression line.

In a real application we would usually assume that the phenomenon of interest is affected by more than one factor. In the case of opinions towards immigrants such additional factors could be age, sex, employment status, income, etc. The idea that acceptance of immigrants is affected by all these factors again can be expressed in a linear model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \tag{4.3}$$

In this equation we have refined our hypothesis that 'the higher the education of a person is, the more willing s/he is to accept immigrants' in an important way. Let us assume that

education is represented by x_1 and consequently the effect of education is β_1. As in the earlier example, we assume that if education increases by one unit the acceptance of immigrants changes by β_1 units. But equation (4.3) extends our hypothesis by stating that the effect of education on acceptance of immigrants is estimated while 'holding all other variables constant'. This means that equation (4.3) allows us to observe the effect of education controlling for third variables such as age and sex. The same in turn is true for the effects of all other independent variables x_j in equation (4.3). Their effects are estimated under the assumption that all other independent variables are held constant. Thus, linear regression allows us to estimate the effect of an independent variable on a dependent one as if the units of analysis did not differ with respect to other characteristics contained in the model. For social science applications this is an enormous advantage because, unlike other sciences, we often cannot experimentally manipulate the variables we want to study.[1]

Now assume that instead of analyzing the effect of education we are interested in analyzing the effect of church membership on the sentiment towards immigrants. A variable like this with only two levels, member and non-member, is referred to as binary variable. In linear regression analysis the effects of such binary variables can be modeled straightforwardly. All we have to do is to decide how to code this variable. The standard approach is dummy coding, that is, assigning one of the two categories the value 0 (this category serves as the so-called 'reference category'), and the other the value 1. For example, we could create a variable having the value 0 for non-members and the value 1 for members of churches. Let us denote this variable by D_c and insert it in the regression equation. This gives

$$\hat{y} = \beta_0 + \beta_1 D_c. \tag{4.4}$$

To understand what this means we look at this model for non-members only, $D_c = 0$. Then equation (4.4) reduces to

$$\hat{y}(D_c = 0) = \beta_0.$$

For members, $D_c = 1$, equation (4.4) yields

$$\hat{y}(D_c = 1) = \beta_0 + \beta_1.$$

Thus, in equation (4.4) the intercept (β_0) is identical to the expectation of $\hat{y}$ for non-members, while the slope (β_1) equals the expectation of the difference between members and non-members with respect to $\hat{y}$. Figure 4.2 illustrates a regression result for a dummy variable. In this example non-members average 4.6 points and members 7.0 on the immigration acceptance scale (indicated by the vertical bars). From these figures we obtain the regression results

$$\hat{y} = 4.6 + (7 - 4.6)\ \text{Member} = 4.6 + 2.4\,\text{Member}.$$

The inclusion of binary predictors in regression models can easily be extended to categorical variables with several categories, for example marital status or nationality. In this case we need more than one dummy variable to represent these effects. More precisely, we need one variable fewer than we have groups. Assume we want to distinguish between single, married, divorced/separated and widowed persons. Then we would need three dummy variables.[2] One of the categories will be the reference category, that is, the category relative to which all the differences are expressed. If we take 'single' as the reference category, the regression of y on marital status will give us

$$\hat{y} = \beta_0 + \beta_1 D_m + \beta_2 D_d + \beta_3 D_w.$$

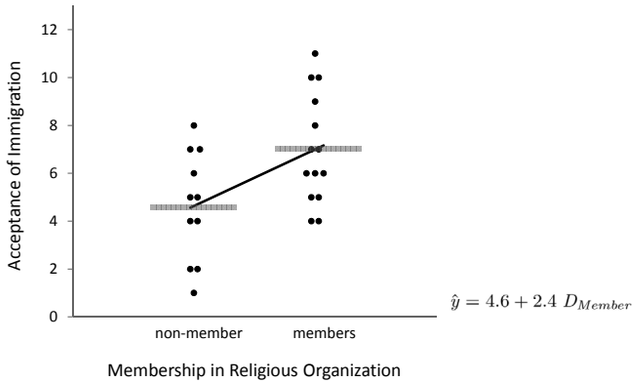


Figure 4.2 Scatter plot for binary predictor

Here β_0 is the expectation of $\hat{y}$ for single persons, the reference category, while β_1, β_2 and β_3 are the differences in expectation of married, divorced and widowed people, respectively.

Using dummies to represent a variable does not have to be restricted to non-metric variables. It can also be a means to test whether the relationship of an independent variable to the dependent variable is non-linear. Take again the example of the effect of education on sentiments towards immigrants. We may have doubts as to whether years of education are linearly related to our variable of interest. In this case we might group years of education into 3-year bands, for example under 8 years, 8 to 10 years, 11 to 12 years, 13 to 15 years, 16 to 18 years, 19 to 21 years, 22 to 24 years, 25 years and over. Inspecting the regression slopes of the respective dummy variables gives us an indication whether or not the assumption of a linear relationship between education and attitudes towards immigrants is warranted.

MATHEMATICAL FOUNDATIONS

The model

In the previous section we have already introduced the general model of multiple linear regression analysis as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

Using the summation sign, this expression can be written more compactly as

$$y = \sum_{j=0}^{k} \beta_j x_j + \varepsilon,$$

with $x_0 = 1$. If we use matrix notation this can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{4.5}$$

which is identical to

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The predicted values $\hat{\boldsymbol{y}}$ can be written as $\boldsymbol{X\beta}$.

Identifying the regression coefficients

Once we have specified a regression model of the kind presented in equation (4.5), the next step is to identify the regression coefficients, that is, the β_j. Think back to the bivariate case for a moment and look again at Figure 4.1. We can easily imagine different lines representing the cloud of points in the scatter plot; the question is which one of these is the best. Obviously the line we want (i.e. the regression coefficients sought) should minimize the difference between observed and predicted values of the dependent variable. However, there are different ways to 'minimize' this difference. The most common approach is to minimize the *sum of the squared differences* between observed and predicted values (i.e. errors). That is, the regression coefficients are found by minimizing

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2 . \tag{4.6}$$

Because this method minimizes the sum of squared errors it is usually referred to as *ordinary least squares* (OLS) regression. We can now find values of β_j that minimize equation (4.6) by partially differentiating (4.6) for each β_j, setting the resulting equation equal to zero and solving for β_j. To illustrate how this works, let us explain this procedure for β_1 in more detail. Because equation (4.6) is a composition of two functions we have to obey the chain rule, that is, we have to differentiate the outer and inner part and multiply the result. The derivative of $\sum(\cdot)^2$ equals $2\sum(\cdot)$; the derivative of $(\boldsymbol{y} - \boldsymbol{X\beta})$ for β_1 equals $-x_{i1}$. Multiplying both results and setting the equation to 0 yields

$$2 \sum_{i=1}^{n} -x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) = 0$$

or

$$-2 \sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) = 0. \tag{4.7}$$

We can simplify this expression by dividing both sides of the equation by -2, giving

$$\sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) = 0. \tag{4.8}$$

To complete the exercise, we have to repeat the differentiation for the other regression coefficients. The resulting system of equations is (cf. Wooldridge, 2009, p. 800)

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) &= 0 \\
&\vdots \\
\sum_{i=1}^{n} x_{ik}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) &= 0,
\end{aligned} \tag{4.9}$$

where the first equation results from differentiation for β_0, the second for β_1, etc. In matrix notation this can be written as

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{4.10}$$

Multiplying and rearranging terms gives

$$(\boldsymbol{X}'\boldsymbol{X}')\,\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}. \tag{4.11}$$

Assuming $(X'X)$ has full rank, that is, none of the independent variables is a perfect linear combination of other independent variables, we can left-multiply both sides by the inverse matrix $(X'X)^{-1}$, resulting in

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X}')^{-1}\,\boldsymbol{X}'\boldsymbol{y}. \tag{4.12}$$

This equation gives regression coefficients that minimize the sum of squared errors. Thus, unlike maximum likelihood estimation used in logistic or probit regression, there exists a closed-form solution for finding regression coefficients in OLS regression (for more on OLS and maximum likelihood estimation see Chapter 2 of this volume).

Assessing model fit

As we have seen in the previous subsection, it is always possible to solve a linear regression problem using the OLS principle. And each model estimated with this principle minimizes the squared difference between observed and estimated values of the dependent variable. However, this does not imply that every regression model fits the data equally well. On the contrary, some models will have very poor fit while others will fit the data better. The degree of fit obviously depends on the degree to which the predicted values for the dependent variable $\hat{y}$ are similar to the observed values of y. Or, to put it slightly differently, the more differences in the observed variables a model can account for, the better its fit.

To operationalize this idea of lesser or better fit we make use of a basic concept from the analysis of variance, namely that the total variation of the dependent variable can be partitioned into a part explained by the regression model and a part not explained by the model. Put more formally,

$$\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2$$

or

$$\text{TSS} = \text{ESS} + \text{RSS},$$

where TSS, ESS and RSS stand for total sum of squares, explained sum of squares and residual sum of squares, respectively. We can now look at the ratio of explained variation relative to the total variation,

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}. \tag{4.13}$$

The ratio R^2, also called the coefficient of determination, can vary between 0 and 1 and reflects the proportion of variance explained by the regression model.

One problem with this measure of fit is that if we add more variables to a given model R^2 can only increase, even though the variables we add may be irrelevant with respect to the dependent variable of interest. This happens because R^2 and consequently also changes in R^2 can only be positive. Therefore, independent variables unrelated to the dependent variable can by chance produce an increase in R^2. To correct for this tendency an adjusted version of R^2 has been proposed which is given by

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

with n denoting the number of cases and k the number of independent variables. In contrast to R^2, R^2_{adj} can decrease when adding new variables to an equation that are irrelevant with respect to the dependent variable.

If using a regression approach to model data we would often start with a basic model to which we would add more variables step by step each time making it slightly more complex. For example, if we were interested in the effect political orientation has on attitudes towards immigrants we could first estimate a model containing only socio-demographic variables. This first model will give a fit of R^2_1. In a second model we then add political orientation to the independent variables and estimate a model which gives us R^2_2. The difference between the two measures of fit, $R^2_2 - R^2_1$, then indicates the effect of political orientation on attitudes towards immigrants net of the socio-demographic variables controlled for in the model. This strategy is particularly useful if we want to estimate the effect of several variables, an approach we can also use to determine the impact of a set of dummy variables on the explained variance.

If the models we are interested in are not nested, R^2 should not be used for comparisons. If we want to compare the same model in different populations (e.g. men and women or Switzerland and Germany), then we can apply the Chow test presented below on page 66.

Before closing this section, we would like to add a final word on the size of R^2. A question often asked is how large R^2 should be. The answer to this question depends on the purpose of our research. If we are aiming to explain a certain variable, such as attitudes towards immigrants, then we would like to maximize the R^2 of our model. If, on the other hand, we are interested in the effect of political orientation and class position on the attitudes of immigrants then we would not care about the overall fit of our model so much but focus on the effect sizes for the variables we are interested in.

Statistical inferences of regression results

Usually a regression model is estimated on the basis of data from a (random) sample of the target population of interest. For example, the short empirical analysis we present in the next section of this chapter is run on data from a Swiss and German sample of the European Social Survey. When we estimate the regression models we are not so much interested in studying the sample as such, but instead aim to learn something about the population from which the sample was drawn. So, in our illustrative analysis presented below, we aim to gain better knowledge of certain attitudes of the adult populations of Switzerland and Germany. As in other areas of statistics, we can apply technics of statistical inference to draw these conclusions for the populations based on data from random samples. For simplicity of the presentation the tests we discuss in this section assume that the data come from a simple random sample. In most cases this will be an oversimplification because our data typically stem from multistage (stratified) samples. If this is the case we should use the appropriate adaptations of the tests we present here. Some of these tests will be presented in Chapter 11 of this volume which discusses regression analysis for data from 'complex' samples. However, the logic of statistical inference remains unchanged.

To indicate that a regression model is estimated with sample data the standard notation is slightly modified by adding a circumflex (ˆ or 'hat') to the regression coefficients estimated by the model. The regression model then becomes

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \varepsilon = \sum_{j=0}^{k} \hat{\beta}_j x_j + \varepsilon \tag{4.14}$$

or, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}. \tag{4.14'}$$

The $\hat{\beta}_j$ are estimates of the β_j computed from data from a sample. Related to such a model we can make two different types of inferences: first, inferences about the model itself or a comparison between different models; and second, inferences about one or two regression coefficients. We begin by discussing this second type of inference about regression coefficients.

Inferences about one regression coefficient

A first question we may want to ask is whether it can be assumed with some reasonable level of certainty that a regression coefficient in the population (β_j) is equal to or different from some value a. The decision between the statistical hypotheses[3]

$$H_0\colon \beta_j = a,$$
$$H_1\colon \beta_j \neq a$$

is made based on the following test statistic:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a}{s_{\hat{\beta}_j}}, \tag{4.15}$$

with $s_{\hat{\beta}_j}$ as standard error of the regression coefficient. The standard error is given by

$$s_{\hat{\beta}_j} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k-1)}{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2\left(1 - R_j^2\right)}}, \tag{4.16}$$

with R_j^2 denoting the amount of variance explained in x_j by the other independent variables (cf. Wooldridge, 2009, p. 89). If the assumptions behind OLS analysis are met, the t-statistic follows a t-distribution with $n - k - 1$ degrees of freedom. Based on this test statistic, we can test $\hat{\beta}_j$ against any value a. Standard software usually gives results for the two-sided test for $a = 0$, that is, the hypotheses

$$H_0\colon \beta_j = 0,$$
$$H_1\colon \beta_j \neq 0.$$

The test tells us if we can assume with a given degree of certainty that the null hypothesis (H_0) can be rejected, meaning that we can assume that x_j has an influence on y in the target population. The degree of certainty we adopt is a convention which is often set to 95% or 99% in social science applications. However, depending on the research question, different levels of certainty will make sense.

Inferences about the relative size of two regression coefficients from the same population

Sometimes we may be interested in testing whether the effect of one variable is stronger than that of another variable from the same model. For example, we could ask if the effect of education (β_1) on acceptance of immigrants is stronger than the effect of age (β_2). The statistical hypotheses are

$$H_0\colon \beta_1 \leq \beta_2,$$
$$H_1\colon \beta_1 > \beta_2$$

and the test statistic is

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{s^2_{\hat{\beta}_1} + s^2_{\hat{\beta}_2} - 2s_{\hat{\beta}_1\hat{\beta}_2}}}, \tag{4.17}$$

with $s_{\hat{\beta}_1\hat{\beta}_2}$ being the covariance between the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$; t has $df = n - k - 1$ degrees of freedom.

Inferences about the relative size of a regression coefficient in two different populations

In other situations we may be interested in learning whether a predictor has the same effect in different populations. We could be interested, for example, in testing whether the effect of education on acceptance of immigrants is the same in Switzerland and Germany. The statistical hypotheses would be

$$H_0\text{: } \beta_1|\text{Switzerland} = \beta_1|\text{Germany},$$
$$H_1\text{: } \beta_1|\text{Switzerland} \neq \beta_1|\text{Germany}.$$

One way to answer this question is to combine the samples of interest into one data set and add an indicator variable to the data set taking the value 0 for data from the first sample and 1 for data from the second sample. Finally, we would have to create an interaction term between the data set indicator and the independent variable we want to test. The resulting model can be expressed by the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_d D + \hat{\beta}_{1d} x_1 D + \sum_{j=2}^{k} \hat{\beta}_j x_j + \varepsilon. \tag{4.18}$$

Based on this equation, the statistical hypotheses are modified to

$$H_0\text{: } \beta_1|\text{Switzerland} - \beta_1|\text{Germany} = \beta_{1d} = 0,$$
$$H_1\text{: } \beta_1|\text{Germany} - \beta_1|\text{Germany} = \beta_{1d} \neq 0.$$

As can be seen, from this expression for the statistical hypotheses, the original question has been transformed into one asking if a regression coefficient is different from zero. This question can easily be answered by the test introduced above (see equation (4.15)) which allows us to test whether β_{1d} is significantly different from zero. If so, we would have to conclude with a given level of confidence that the effect of x_1 is different in the two populations.

Inferences about an entire model

Having covered some tests concerning regression coefficients, we now turn to testing entire models. The question we ask is whether the regression model we estimate can be expected with reasonable certainty to explain at least some of the variation of the dependent variable we study in the population. The statistical hypotheses can be formulated as

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$
$$H_1\text{: at least one } \beta_j \neq 0.$$

This pair of hypotheses can be tested by the statistic

$$F = \frac{\sum(\hat{y} - \bar{y})^2 / k}{\sum(y - \hat{y})^2 / (n - k - 1)} = \frac{\text{ESS}/k}{\text{RSS}/(n - k - 1)} = \frac{\text{EMS}}{\text{RMS}}, \tag{4.19}$$

which follows an F-distribution with $df_1 = k$ and $df_2 = n - k - 1$. This statistic is well known from the analysis of variance framework, in which the numerator is also known as explained mean squares (EMS) and the denominator as residual mean squares (RMS). If the F statistic is significant then we can conclude with the given level of certainty that at least one independent variable included in the model is (linearly) related to the dependent variable in the population of interest.

Inferences about two nested models

Often we take a stepwise approach to modeling a dependent variable with regression analysis. As mentioned before, we may, for example, first want to see how much attitudes towards immigrants depend on socio-demographic variables and then add political orientation in a second step. But we could also take the reverse route and first explore the effect of political orientation on attitudes towards immigrants and only then control for socio-demographics. In either analysis we may want to know if the additional variables added in the second step improve the model fit significantly. If we want to test the difference of such 'nested' models, that is, models where the parameters of the first model are a true subset of the parameters of the second model, we can use the following F-distributed test (Fox, 2008, p. 201):

$$F = \frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(k_2 - k_1)}{\mathrm{RSS}_2/(n - k_2)}, \tag{4.20}$$

with RSS_1 and RSS_2 referring to the residual sum of squares for model 1 and model 2, respectively, where model 1 is nested in model 2, so that $k_1 < k_2$. If the F-statistic is not significant then model 2, the model with more variables, does not predict the dependent variable better than model 1, the simpler model. If the F-statistic is significant we can be reasonably certain that model 2 fits the data better than model 1.

Inferences about two models for different populations

There may be cases in which we are interested in knowing whether the same model holds for different populations. Thus, we may want to know whether a given model intended to explain attitudes towards immigrants leads to identical conclusions for Switzerland and Germany. In this situation, instead of comparing two slopes we will have to compare the overall fit of the two models. The relevant test statistic, also known as the Chow statistic (see Wooldridge, 2009, p. 245), again follows an F-distribution and is defined by

$$F = \frac{\big(\mathrm{RSS}_p - (\mathrm{RSS}_1 + \mathrm{RSS}_2)\big)/(k+1)}{(\mathrm{RSS}_1 + \mathrm{RSS}_2)/\big(n - 2(k+1)\big)}, \quad df_1 = k+1, \quad df_2 = n - 2(k+1), \tag{4.21}$$

where RSS is the residual sum of squares of a pooled (RSS_p) analysis and the separate analyses (RSS_1 and RSS_2). A significant test result implies that the regression models in the two groups are not identical, that at least one slope or the intercept differs between the two populations.[4]

In closing this subsection on significance tests, we would like to remind readers that these tests only indicate whether a certain hypothesis holds or does not hold with a specified, predefined level of certainty. Even more importantly, statistical significance does not tell us anything about the substantive significance of an effect. If our samples are large, even very small effects will become statistically significant but often they would be not very meaningful from a substantive viewpoint. Take, for example, a literacy test with mean 250 and standard deviation 50 points. Assume that males score 3 points higher than females, and that this difference is statistically

significant. Should we conclude that the gender difference is important and should we advise policy-makers to act on this? Most likely this would not be very sound advice. Given that the difference between males and females amounts to less than a tenth of a standard deviation of the literacy measure we should probably focus our attention on other factors, perhaps education. The bottom line of this is that substantive importance of regression results has to be judged based on substantive criteria.

Assumptions in ordinary least squares regression

Ideally OLS regression estimators are best linear unbiased estimates (BLUE). This is the case if the data meet the assumptions on which this method is based. Analysts should be aware of these assumptions and test whether they apply. We will briefly describe the most important assumptions underlying OLS regression; for a fuller discussion of this issue, see Berry (1993) or the next chapter of this volume:

- The dependent variable has to be metric; the independent variables may be metric or coded as dummy variables or other contrasts.
- If we want to draw inferences from our data it must come from a random sample of the population of interest.
- The independent variables have to be measured without measurement error.
- None of the independent variables must be a constant or a linear combination of the other independent variables; that is, there should be no perfect multicollinearity. Technically this means the matrix X must have full rank.
- The error terms (residuals) must follow a normal distribution.
- For each value of the independent variables the variance of the error term has to be identical, $\text{var}(\varepsilon|x) = \text{const.}$; this is also referred to as a situation of homoscedasticity.
- For each combination of independent variables the expectation of the error term has to be zero, $E(\varepsilon|x) = 0$. This assumption implies that no independent variable is correlated with the error term – a situation described in econometrics as strict exogeneity.
- The aforementioned assumption implies that the model is correctly specified, that is, all relevant variables are included in the model and the model does not contain irrelevant variables. In addition, the parametrization of the model has to be correct, that is, in the given operationalization and parametrization the independent variables have to be linearly associated with the dependent variable.

The OLS estimates of the regression coefficients and their standard errors are BLUE if these assumptions are met. However, in real-world applications of OLS the assumptions listed above will only be met to a certain degree, with the effect that the OLS estimates will deviate from the ideal of being unbiased and efficient (have minimum variance). To assess the quality of a regression model it is important to be aware of the consequences of deviations from the assumptions. Multicollinearity, heteroscedasticity and non-normal residuals lead to biased standard errors of regression estimates which lead to incorrect significance tests and confidence intervals. The estimates of the regression coefficients (intercept and slopes), however, remain unbiased. Deviations from the other assumptions have an even stronger effect on results. In this case not only the standard errors but also the regression coefficients are biased.

We would like to briefly show why this happens in the case of misspecification. Let us assume the true model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \beta_m x_m + \varepsilon.$$

Now let us assume we were not aware of the factor x_m and we specify the model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon^*,$$

without x_m. The error term of this model will be identical with the error term of the true model plus the variable x_m, that is, $\varepsilon* = \beta_m x_m + \varepsilon$. If x_m is correlated with at least one of the other independent variables – and this will be the case in almost all real situations – the error term of the misspecified model is correlated with independent variables. Thus, the assumption of strict exogeneity is violated and the estimates for the regression coefficients will be biased. If we inspect equation (4.12) we see why this is the case: the estimation of regression coefficients takes the correlations between the independent variables into account. If important independent variables which are both related to the dependent and the independent variables are left out of the model the estimates are biased – a bias also referred to as *omitted variable bias*. The only way to avoid misspecification of regression models is to root them in a sound theoretical foundation and use adequate operationalizations of the concepts of interest.

Another often encountered problem is unreliable measurement of independent variables. Measurement error of independent variables, be it systematic or random measurement error, leads to biased estimates of the regression coefficients and their standard errors (Cohen et al., 2003, p. 119). The larger the measurement error of a variable x_k, the more the regression coefficient β_k underestimates the true effect of x_k on y, an effect also known as attenuation. Therefore, we should strive to improve our measurement instruments and scaling techniques. If we have several indicators for the concept of interest we could consider using structural equation modeling instead of OLS regression (see Kline, 2010). If we only have a single indicator for a concept, we might be able to obtain a reliability estimate through the web-based Survey Quality Prediction program maintained by the Research and Expertise Centre for Survey Methodology at Universitat Pompeu Fabra under the guidance of Willem Saris (see `http://sqp.upf.edu/`). The estimated reliability could be used to estimate the amount of attenuation of regression coefficients.

The next chapter of this volume presents a much more comprehensive discussion of the assumptions underlying OLS regression as well as ways to test to what extent they are met.

Interpretation of regression results

Once we have established that a regression model does 'explain' the dependent variable at least partly (i.e. is statistically significant), we still are faced with interpreting the regression results from a substantive point of view. Let us first focus on interpreting the regression coefficients or slopes β_j (these coefficients are often referred to as unstandardized regression coefficients, in contrast to standardized coefficients which we will discuss in the next section). Frequently one reads that these coefficients indicate the unit change in the dependent variable if the independent variable is increased by one unit. So if $\beta_1 = 0.5$ a one unit increase in x_1 would result in an increase of half a unit of y. In most practical instances this interpretation will be incorrect. In particular, if we use cross-sectional data to estimate a regression model we should abstain from interpreting the results in a dynamic way. A correct interpretation would be that the expectation for y is 0.5 units higher for those with $x_1 = a + 1$ compared to those with $x_1 = a$. Additionally, if β_j is estimated in a regression model with more than one independent variable then this coefficient is conditional on the other predictors. In other words, the estimate is an attempt to model a situation in which the other independent variables are held constant. If we can assume that all relevant variables are included in the model and parametrized correctly, x_i is conditionally uncorrelated with ε and β_i can be interpreted as causal effect (for a more thorough discussion see below).

Let us have a closer look at the following model (numbers in parentheses are standard errors of estimates):

$$\text{Control Immigration} = 4 + \underset{(0.01)}{0.05}\,\text{Age} - \underset{(0.5)}{0.5}\,\text{Female} - \underset{(0.00004)}{0.0001}\,\text{Income} + \varepsilon, \quad R^2 = 0.04. \quad (4.22)$$

Both age and income have significant effects on the attitudes towards immigration (the coefficients are more than twice their standard errors). In contrast, being female rather than male has no significant effect. For each age group the expected value on the attitude scale is 0.05 points higher than for the group one year younger, irrespective of sex and income. Similarly, each additional dollar decreases the opposition towards uncontrolled immigration by a small amount (0.0001 points) controlling for age and sex. This interpretation draws attention to three crucial characteristics of multiple regression. First, the regression coefficients reflect the estimated effect of one variable, controlling for all other variables in the model. In our case this means that the effect of age is estimated by taking sex and income into account. Second, only linear effects are modeled and correctly reflected in regression estimates. In our example this means we assume that there is the same difference in attitudes towards immigration between a 21- and a 20-year-old person as between an 81- and an 80-year-old person, namely 0.05 units. If we had reason to believe that the relationship between a predictor and a dependent variable is non-linear we could still model this in the framework of linear regression. However, we would have to transform the variable in question in such a way that the regression model is linear with respect to the transformed variable. For example, if we assume that the increase in opposition to immigration gets smaller with increasing age we could use the logarithm of age instead of age in our model. Third, our model implies that the predictors' effects are additive and do not depend on each other. Again, if we had reason to believe that the effect of one independent variable depends on levels of another independent variable we could model this in the framework of linear regression by incorporating interaction effects. Because Chapter 6 exclusively discusses non-linear and non-additive effects in linear regression we do not discuss these issues here any further.

The interpretation of the effects of 0–1 coded binary variables is similar to the interpretation of effects of continuous variables. The coefficient reported above for being female implies that the conditional expected value of y is 0.8 units higher for females than for males, controlling for age and income. But as we have seen, this difference is not statistically significant.

The interpretation of regression results can often be facilitated by changing the scale of the independent variable. Assume we had measured income not in dollars but in tens of thousands of dollars. Then the regression coefficient would change from 0.0001 to 1, implying that an income difference of $10,000 is associated with an expected difference of one unit on the dependent variable, controlling for age and sex. Another example would be age. If we divide age in years by 10, thus measuring age in decades, the above age effect would change from 0.05 to 0.5, indicating that people who are 10 years apart are expected to be half a scale point apart on the immigration scale.

How can we interpret the intercept β_0? This is the expectation for the dependent variable if all of the independent variables x_j are zero. For the above model we could say that for men (female $= 0$) who are zero years old and who have zero income we expect a value of 4 on the attitude scale. Here and in most other cases this information is of no interest. It could even be misleading because $x_1 \cdots x_k = 0$ most often lies outside of the window we observe. Here, for example, we would assume that the observations were restricted to the adult population. Also, it is safe to assume that newborns do not have attitude towards immigrants. One way to avoid

misleading interpretations of the intercept is to center all (metric) variables on their mean (or alternatively on some other meaningful value). Then the intercept reflects the expectation for the 'average' person, a figure which might be of substantive interest.

Standardized regression coefficients

The size of the regression coefficients we have reported and interpreted so far depends on the units used to measure the independent and dependent variable. As long as variables have 'natural' or intuitive measurement units (e.g. age in years or income in dollars) the interpretation of coefficients is straightforward. However, many measures in the social sciences are based on arbitrary units derived from answers to rating scales of various types and lengths. Because of the arbitrariness of the units of such measures, their regression coefficients are not very informative. A related problem arises if we are interested in the relative effect of variables measured on different scales. Reconsider the example given above where we found that the effect of age was 0.05 and the effect of income was 0.0001 (see equation (4.22)). Do these coefficients imply that attitudes towards immigrants are more affected by age than by income? Obviously not, because – as we have seen – the size of the unstandardized regression coefficient depends on the units used to measure the variables. As mentioned above, the slope for income would have been 1 if we had measured income in tens of thousands of dollars.

Arbitrary units of measurement and the assessment of relative importance of predictors are typically addressed by interpreting standardized regression coefficients. These coefficients are computed by multiplying the unstandardized coefficient by the ratio of the standard deviations of the independent and dependent variable,[5] that is,

$$\beta_j^s = \beta_j \frac{\sigma_{x_j}}{\sigma_y}. \tag{4.23}$$

Expressed in standard deviations of y, these standardized coefficients tell us how much two groups are expected to differ with respect to y if they differ by one standard deviation on x_j. In this parametrization all effects are expressed in standard deviations of y and x_j. An increase of one standard deviation of x_j results in a change of y by β_j^s standard deviations.

In the social sciences it has long been common practice to report and interpret only standardized regression coefficients. However, their use has been criticized for several reasons (cf. Bring, 1994). One problem is that standardized regression coefficients reflect not only effect sizes but also variation of the variables. Therefore, standardized coefficients may vary between samples or populations just because the variables of interest have different variances, even though the effects of x_j on y are identical. Assume we want to compare the effect of income on attitudes for men and women. If the variation of income differs between men and women this will influence the standardized coefficients and so it will be impossible to know if the difference in standardized coefficients is due to differences in the effect of income or the variation of income in the two groups. Consequently, we should use unstandardized measures in comparative analysis.

But even the comparison of standardized coefficients within the same model has been criticized. To see why, let us inspect the following example. Let us assume that

$$\text{Control Immigration} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Income} + \varepsilon$$

is the model of interest. As we have seen above, β_1 reflects the effect of age, holding income constant. According to equation (4.23) the standardized effect of age is determined by multiplying β_1 by the standard deviation of age (σ_{x_1}). This may be seen, as Bring (1994) has claimed,

as inconsistent because β_1 is an estimate conditional on other variables (controlling for ...) in the model while σ_{x_1} is the unconditional standard deviation referring to the entire population unadjusted for other measures. In essence, the slope and standard deviation refer to different populations. As a solution Bring (1994, p. 211) suggests using the partial standard deviation of x_j averaged over the groups formed by the independent variables.

A further critique of standardized coefficients is that they only reflect the relative contribution of an independent variable to R^2, the explained variance, if the independent variables are all uncorrelated. In this very limited case R^2 is identical to the sum of squared correlation coefficients between independent variables and dependent variable which are in this special case identical to the standardized coefficients. Therefore, only if all independent variables are unrelated to each other do the standardized regression coefficients reflect the relative contribution of each variable to the explained variance. This special case, however, never occurs when working with real data. And if it did occur we would not need to use multiple regression because there would be no need to 'control' the effects of independent variables for other independent variables. With correlated independent variables R^2 can be decomposed as

$$R^2 = \sum_{j=1}^{p} \beta_j^{s\,2} + 2 \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_j^s \beta_k^s \rho_{jk}, \tag{4.24}$$

with β_j^s and β_k^s representing the standardized effect of x_j and x_k, respectively, and ρ_{jk} indicating the correlation between x_j and x_k (cf. Grömping, 2007, p. 140). Bring (1994) has suggested measuring the relative importance of independent variables by multiplying the correlation between independent and dependent variable by the unstandardized regression coefficient. This measure has the advantage of summing to R^2 over all independent variables. That is, R^2 can be partitioned as

$$R^2 = \sum_{j=1}^{k} \beta_j \rho_{jy}. \tag{4.25}$$

Though the products $\beta_j \rho_{jy}$ sum to the explained variance and thus can be considered as indicating relative importance of the predictors, there is a problem with this interpretation when the signs of the two factors differ. In this case the product has a negative sign implying that the independent variable contributes not to the explained but rather to the unexplained variance.

In the meantime several measures to reflect relative importance of predictors have been proposed in an attempt to overcome the shortcomings of standardized coefficients. Chao et al. (2008) compare six of these proposals to capture the relative importance of predictors in multiple linear regression. They base their comparison on three criteria: (a) (squared) coefficients of relative importance should sum to R^2; (b) coefficients of relative importance should never be negative; (c) coefficients of relative importance should not depend on the order in which predictors are entered into a regression equation. Only two of the six measures examined by Chao et al. (2008) meet all three criteria: a proposal by Budescu (1993) and one by Johnson (2000). Because the coefficient proposed by the former is very cumbersome to compute and because both approaches result in very similar estimates, Chao et al. (2008) recommend using the method introduced by Johnson (2000). Therefore, we will only discuss this latter measure briefly.

Suppose we have a regression model with k predictors. Using principal components analysis, Johnson (2000) suggests extracting k orthogonal factors z_m and rotating them so as to minimize the sum of squared differences between the rotated factors and the variables x_j. Then the dependent variable of interest is regressed on the rotated factors z_m. Because the factors are orthogonal

the squared standardized coefficients $\beta^s_{z_m}$ sum to R^2. In a final step we have to determine the importance of the original variables x_j by

$$\beta^{\dagger}_{x_j} = \sum_{m=1}^{k} \lambda_{j_m} \beta^s_{z_m},$$

with λ_m denoting the correlation or loading between x_j and z_m.

A more readily available alternative to determine relative importance of predictors is the t-value of the typically reported two-sided test of the slopes (see equation (4.15)). This test statistic can also be represented by

$$t_1 = \sqrt{\frac{R^2_{1,2,3,\ldots,k} - R^2_{2,3,\ldots,k}}{(1 - R^2_{1,2,3,\ldots,k})/(n-k-1)}}, \tag{4.26}$$

implying that the t-value is a direct function of the increase in R^2 produced by entering the variable of interest into a model containing all other independent variables (for more details, see Bring, 1994, p. 213). Hence, comparing t-values of the same model allows us to order the independent variables with respect to their importance relative to the dependent variables.

As we have seen, users of multiple regression have several choices to determine the relative importance of independent variables in a regression model. Standardized coefficients β^s are readily available in most statistical software packages but may be problematic. Alternatives like the one developed by Johnson (2000) and presented above seem to better capture relative importance as contributions to R^2 but are not easy to obtain. We also do not know how stable these coefficients are between samples. We do know, however, that comparing standardized coefficients between samples or populations can lead to incorrect conclusions because they do not only depend on the effect of independent variable on the outcome. Rather, they are also affected by the variance of both variables in the two groups. Our recommendation therefore is to always report unstandardized coefficients, the t-value which indicates if a predictor is statistically significant and also reflects relative importance. Additionally, standardized coefficients may be useful but should be interpreted with caution. In the end the 'importance' of a predictor has to be determined from a substantive point of view.

MODELING ATTITUDES TOWARDS IMMIGRATION: AN EXAMPLE ANALYSIS

In this section we present a sample analysis. In contrast to many textbooks we will use real data and we will partly replicate a research paper by Green (2009). In her paper Green studies the determinants of support for different criteria to restrict immigration. Drawing on data from the first round of the European Social Survey, she studies endorsement of ascribed and acquired characteristics as criteria for granting immigration. The central individual level predictors Green studies are perceived threat and social status of host country members. In brief, she hypothesizes that those perceiving negative consequences of immigration for their own life chances (i.e. feel threatened) will oppose immigration more and be more in favor of restricting immigration. Similarly, those in lower social strata will experience more competition for jobs or affordable housing by immigrants than people in high social positions. Therefore, social status should be negatively related to accepting unconditional immigration. To test her hypotheses Green focuses on nationals who do not claim to belong to a minority group within their country (Green, 2009, p. 47).

Our replication of Green's analysis is limited in a number of ways. First, we only study attitudes with respect to immigration criteria which can be acquired (e.g. education). Second, our analysis focuses on only two countries, Switzerland and Germany, and thus, we do not study country-level predictors. Third, we only use a subset of the individual level predictors employed by Green. Our analysis is based on version 6.1 of the data from the first round of the European Social Survey.

In the following analysis we study the importance given to education, proficiency in national language, having work skills and being committed to the way of life in the host country for deciding about immigration. The answers to these four items are combined into an additive index serving as our dependent variable (see the appendix to this chapter for a detailed description of items). Higher values reflect greater importance of these criteria and can be interpreted as being in favor of higher restrictions on immigration. Perceived threat was measured by seven items which we again combine into an additive index with higher values reflecting higher levels of threat (see appendix for items). Social status was measured by education in years. To capture political orientation we follow Green and use the answers to the 11-point left–right scale. Because this measure contains a substantial number of missing values, Green categorizes this variable into left orientation (values 0 to 3), middle orientation (values 4 to 6), right orientation (values 7 to 10) and missing information on political orientation. In the following analysis we use the middle category as reference. Additionally, we control for sex and age.

Our analysis is focused on a comparison between Switzerland and Germany, two countries differing substantially with respect to immigration. In 2002, the year round 1 of the European Social Survey was carried out, 20% of the population in Switzerland did not hold Swiss citizenship, while in Germany only 9% of the population were foreigners. In the same year Switzerland welcomed over 125,000 new long-term immigrants (1.7% of its population), and Germany welcomed almost 850,000 new immigrants (amounting to 1% of the population). With these figures Germany is slightly above the European average whereas Switzerland (together with Luxembourg) is the country having the largest non-national and foreign-born population.

Table 4.1 gives an overview of the variables we will use in our analysis and their distribution in Switzerland and Germany. According to this table, Swiss people place less importance on acquired characteristics for immigration and perceive slightly less threat than Germans. Also Swiss people are on average a little older, a little less educated and a little more oriented towards the political right than Germans. To model the attitudes towards immigration in Switzerland and Germany we proceed in two steps. First, we estimate a model including only sex, age, education and political orientation. In a second step we add perceived threat to the model.

From the left panel of Table 4.2 we see that, in accordance with the social status hypothesis, higher educated people in Switzerland and Germany place less importance on acquired immigration criteria than their less educated compatriots. Although both effects are statistically significant, the effect is much larger in Germany than in Switzerland. In Germany 10 more years of schooling are associated with almost one point less on the importance scale (-0.91), while in Switzerland the same educational difference is only associated with a quarter point change (-0.26). As one might expect, older persons and persons who identify themselves as politically right-wing have more reservations towards unconditional immigration in both countries. In neither country do we observe strong sex differences, so we may conclude that attitudes of men and women with regard to immigration do not differ.

Before we inspect the results of our second model, we would like to draw readers' attention to two issues related to political orientation. As pointed out above, we followed Green in treating this variable as categorical and in using a separate category for the missing values. Because this leads to three variables and three regression coefficients in the model they do not allow us to say anything about the size of the effect political orientation has. An alternative approach would be to estimate a model without the dummies for political orientation and compare it with the

Table 4.1 Descriptive statistics

	Switzerland ($n = 1516$)				Germany ($n = 2349$)			
	min	max	mean	sd	min	max	mean	sd
Restrict immigration	0	10	6.43	1.96	0	10	7.38	1.91
Female	0	1	0.51	0.50	0	1	0.50	0.50
Age	15	103	48.90	16.95	15	93	47.39	17.37
Education	0	31	10.85	3.36	0	30	13.09	3.27
Left	0	1	0.20	0.40	0	1	0.26	0.44
Middle	0	1	0.56	0.50	0	1	0.56	0.50
Right	0	1	0.19	0.39	0	1	0.13	0.34
LR missing	0	1	0.05	0.21	0	1	0.05	0.21
Perceived threat	0.06	0.97	0.47	0.14	0.07	1	0.52	0.16

Data: European Social Survey round 1, version 6.1. Only nationals not belonging to a minority; listwise deletion of missing cases, unweighted.

Table 4.2 Model 1 to explain attitude towards immigration

	Switzerland				Germany			
	$\hat{\beta}$	$s_{\hat{\beta}}$	t	β^s	$\hat{\beta}$	$s_{\hat{\beta}}$	t	β^s
Constant	6.419	0.083	77.2	–	7.522	0.061	123.0	–
Female	–0.019	0.100	–0.2	–0.005	–0.110	0.074	–1.5	–0.029
Age[a]	0.015	0.003	5.2	0.132	0.019	0.002	8.6	0.171
Education[a]	–0.026	0.016	–1.7	–0.044	–0.091	0.011	–8.0	–0.156
Left[b]	–0.619	0.130	–4.8	–0.127	–0.759	0.091	–8.4	–0.174
Right[b]	0.596	0.132	4.5	0.119	0.486	0.109	4.5	0.087
LR missing[b]	0.169	0.239	0.7	0.018	0.015	0.178	0.1	0.002
R^2			.065				.125	
R^2_{adj}			.062				.123	
$F(df_1; df_2)$		17.61	(6; 1509)			55.63	(6; 2342)	

[a] Centered on the country-specific mean.

[b] Reference category political orientation 'middle'.

Data: European Social Survey round 1, version 6.1. Only nationals not belonging to a minority; listwise deletion of missing cases, weighted by design weight.

model displayed in Table 4.2 in terms of the increase in R^2. If we compare two such models for Germany we see that adding political orientation to the model increases the explained variance by 4.2 percentage points. By applying the F test given in equation (4.20) to the two models just estimated, we can test whether political orientation is statistically significant – which it is.

A final remark on the variable political orientation: As the reported results show, there is no statistically significant difference between those without a valid response to the left–right question and those placing themselves in the middle of the political spectrum. Thus, at least with respect to the dependent variable studied here, we find no difference between these two groups and we might come to the conclusion that substituting the missing cases on this variable by its mean and then treating this variable as continuous would make our model more parsimonious without distorting the results for left–right placement.

This brings us to our second model, in which we add perceived threat as a predictor. Again, we report separate analyses for Switzerland and Germany in Table 4.3. In both countries perceived threat has a strong effect. In fact, based on the t-value and the standardized coefficient, perceived

Table 4.3 Model 2 to explain attitude towards immigration

	Switzerland				Germany			
	$\hat{\beta}$	$s_{\hat{\beta}}$	t	β^s	$\hat{\beta}$	$s_{\hat{\beta}}$	t	β^s
Constant	6.378	0.081	78.5		7.504	0.058	128.3	
Female	–0.012	0.097	–0.1	–0.003	–0.044	0.071	–0.6	–0.012
Age[a]	0.014	0.003	4.9	0.122	0.016	0.002	7.6	0.145
Education[a]	0.007	0.016	0.5	0.012	–0.033	0.012	–2.9	–0.057
Left[b]	–0.466	0.128	–3.6	–0.096	–0.555	0.088	–6.3	–0.127
Right[b]	0.558	0.129	4.3	0.112	0.323	0.105	3.1	0.058
LR missing[b]	0.017	0.234	0.1	0.002	–0.193	0.171	–1.1	–0.021
Perceived threat[a]	3.186	0.365	8.7	0.226	3.745	0.252	14.9	0.310
R^2			.110				.202	
R^2_{adj}			.106				.198	
$F(df_1; df_2)$		26.70	(7; 1508)			83.83	(7; 2341)	

[a] Centered on the country-specific mean.

[b] Reference category political orientation 'middle'.

Data: European Social Survey round 1, version 6.1. Only nationals not belonging to a minority; listwise deletion of missing cases, weighted by design weight.

threat is the most important predictor in our model in both countries. The difference between those perceiving no threat and those perceiving the maximum level of threat is 3.2 (Switzerland) and 3.7 (Germany) points on the importance scale used to measure attitudes towards immigration. Consequently, adding this perceived threat to the model substantially increases the amount of explained variance in both countries; in Switzerland by 4.5, in Germany by 7.7 points, leading to $R^2_{\text{CH}} = 0.11$ and $R^2_{\text{DE}} = 0.20$, respectively. Thus, it seems that the model fits the German data much better than the Swiss. Whether or not this difference is statistically significant can be tested with the Chow test given in equation (4.21). The result of this test clearly indicates that the model indeed does not fit the Swiss and German data equally well.

When comparing the effects of the other predictors in the model to their effects in model 1, we see that some of them are strongly affected when perceived threat is entered into the model. In particular, the regression coefficients for education and political orientation show a quite strong reduction in absolute size. This implies that some of the explained variance attributed to these variables in model 1 actually has to be attributed to perceived threat. Indeed, model 2 implies that education in Switzerland does not seem to be directly related to attitudes towards immigrants once we control for perceived threat. This implies that the social status hypothesis does not hold for Switzerland. One possible explanation for this result may be that many foreigners in Switzerland are highly qualified and so competition between the indigenous and migrant population for jobs, dwellings and so on may not be concentrated in lower status groups in Switzerland.

Finally, we may be interested in testing whether the predictors we examined differ in their effects between the two countries. To do so we applied the model given in equation (4.18). That is, we estimated a model based on a combined sample, including in the equation a dummy variable indicating the country and interaction terms for this variable with all predictors. This analysis shows that the only independent variable with significantly different effects in both countries is education. All the other variables seem to have identical effects in Switzerland and Germany.

Here we end our example analysis. If we wanted to publish the results of our analysis we should go further and test whether the assumptions of OLS regression are reasonably well

met by our data. We do not have the space to do this here, but refer readers to the next two chapters in which these assumptions, diagnostic tools and possible remedies are extensively discussed.

PROBLEMS AND REMEDIES FOR CAUSAL INFERENCE BASED ON OLS REGRESSION

In recent years, causal analysis in the social sciences has increasingly developed consensus on applying the potential outcome model, also called the counterfactual model of causality (e.g. Rosenbaum and Rubin, 1985). This framework explicates an intra-individual concept of causal analysis. The simplest setup assumes a binary causal state with treatment and control (untreated) conditions. The causal effect of the treatment D on an outcome y then can be defined as

$$\Delta_i = y_i^1 - y_i^0 , \tag{4.27}$$

with i as a person index and 0/1 denoting control and treatment state. Identification of the treatment effect is complicated by the fact that a person never will be in both states at the same time, and Δ_i hence cannot be observed. In presence of panel data, identification of the average treatment effect oftentimes is attempted using fixed-effects panel regression or related methods (see Chapter 15 of this volume). With only cross-sections being available, groups of persons which are observed under different causal states have to be compared. This, of course, imposes massive problems on the researcher if a controlled and fully randomized experiment cannot be conducted. Notably, unbiased identification of the treatment effect is possible only if the potential outcome is unconditionally or at least conditionally independent of treatment assignment:

$$(y^0, y^1) \perp\!\!\!\perp D \tag{4.28}$$

or

$$(y^0, y^1) \perp\!\!\!\perp D|z . \tag{4.29}$$

The former is a very strong assumption, and valid only in absence of selection into treatment groups, i. e. in randomized trials. The conditional independence assumption (equation 4.29) is somewhat weaker as independence only needs to hold after controlling a number of covariates z.

OLS regression has developed a bad reputation in causal inference (see e. g. Morgan and Winship, 2007, section 1.1.2). This is mostly due to the misuse of explorative regression models for causal conclusions and, more generally, the focus on "fully" explaining the variance of a dependent variable rather than identifying treatment effects of a specific manipulation (e. g. Blalock, 1964).[6] However, as proponents of the potential outcome model have pointed out, OLS regression can play a role in estimating causal effects if applied sensibly. To show why and in how far this is the case we will briefly review the main obstacles of drawing causal inference from observational data and how these are addressed by OLS regression for cross-sectional data (but see Chapter 15 of this volume for causal inference based on longitudinal data).

In our opinion, the most severe problem in using OLS regression for causal inference with non-experimental data stems from self-selection or policy endogeneity, both of which result in violating the unconditional independence assumption. For example, if we study the effect of job interview training on earnings, a self-selection bias may occur due to higher-skilled persons participating in the training classes with a higher probability. Policy endogeneity occurs when the organizer of a program targets a specific sub-group that is expected to show the strongest effects (e. g. persons with an academic education). As mentioned before, selection bias is avoided in

experimental research by randomly assigning individuals to control and treatment groups thereby ensuring that the state of treatment is the only systematic difference between the two groups; the two groups are unconditionally independent. However, if we know all relevant factors in which the "treated" and "non-treated" differ, we can adjust statistically for these factors to achieve conditional independence. In an OLS regression we can assume conditional independence provided that all selection variables are included in the regression and the parameterization of the model is correct.[7] Under these circumstances the regression coefficient of interest can be interpreted as a causal effect (cf. Angrist and Pischke, 2009, pp. 51–59; Gelman and Hill, 2007, p. 169). We must also check if there is sufficient overlap of covariates across treatment groups. This means that confounders should not be highly correlated with the treatment variable (for a more in-depth discussion see Gelman and Hill, 2007, Chapter 10.1). Furthermore, we must assume treatment homogeneity or monotonicity for a meaningful interpretation of the regression coefficient as average treatment effect (see Humphreys, 2009).

In a way, the reservations against regression may be seen as stemming from a different epistemological background of those having these reservations and a more stringent focus on research design in the potential outcomes framework – rather than from inherent statistical shortcomings of the regression approach. Applied in a careful and well-conceived way, regression can be used as a statistical tool for the estimation of treatment effects (see also Angrist and Pischke, 2009, Chapter 3; Morgan and Winship, 2007, pp. 123ff). The fundamental difference to the more direct matching approach is that the conditional independence problem is tackled by adjusting for covariates rather than by balancing, and that a different set of assumptions – such as correct parameterization – has to be met for the identification of causal effects.

That said, there will be situations in which cross-sectional OLS models are simply not suited for causal inference. In a situation with weak overlap between treatment- and control group or when the functional form of effects is unknown, propensity score matching may be the method of choice. Matching estimators have been suggested as the most direct approach to solving the problem of balancing treatment and control groups and meeting the assumption of conditional independence. As a semi-parametric model, matching rests on less assumptions than the regression approach. Additionally it allows to at least estimating a local treatment effect when overlap is weak, provided that the dataset is large enough to identify a sufficient number of matches (for a more detailed exposition of propensity score matching and its identifying assumptions see Chapter 12 of this volume). However, there may be a total lack of overlap of covariates that originates from a variable that was the basis for assigning cases to control and treatment group. Imagine we are interested in studying the long-term career effects of scholarships. Imagine further that scholarships are awarded to all students scoring in the top 80 % of an aptitude test. In this case aptitude and being granted a scholarship are perfectly related and neither OLS nor matching are suitable methods for creating conditional independence. Alternatively, we could concentrate on only those students just below and above the critical threshold for obtaining a scholarship. Taking into account measurement uncertainty we can assume that these students have equal aptitude and only differ with respect to the scholarship. This is the basic idea behind the regression discontinuity approach presented together with a more elaborate description of the example in Chapter 14 of this volume. In many cases we will also be unable to observe all relevant covariates. As we have seen in above omitting a relevant variable from the regression equation leads to biased estimates, the so-called omitted variable bias. In this case an approach known as instrumental variable regression (IV regression) may help if we have a good proxy or instrument for the omitted variable (see Chapter 13 of this volume). In the presence of panel data, we could use fixed effects regression to control for unobserved time-constant covariates and obtain unbiased estimates (see Chapter 15 of this volume).

CAVEATS AND FREQUENT ERRORS

We can only apply statistical methods and interpret their results correctly if we have at least some basic understanding of their general purpose and the assumptions on which they are built. Linear regression is no exception. One of the major questions we should ask is whether we have specified our model correctly. Were we able to include all relevant predictors? Are the predictors linearly related to the outcome variable? Are the effects of the independent variables additive or does the effect of one variable depend on the level of another variable? To answer these questions satisfactorily we first have to rely on sound theory about the substantive matter we are investigating. Second, as Chapters 5, 6 and 10 of this volume show, we can and should test the assumptions of linearity and additivity. Applying sound theory and rigorous testing of assumptions are important because – as we have seen – misspecification of a model leads to the violation of strict exogeneity, that is, violates the assumption that residuals and predictors should not be correlated. This in turn leads to biased estimates of regression coefficients and their standard errors.

Another practical issue we must look at in every analysis is the sample size. Often we would run a regression analysis with many independent variables using listwise deletion of missing values – in many statistical programs this is the default and we might not even make a deliberate decision about this choice. When our model contains many predictors or at least one predictor with many missing values, we can end up estimating our model on a relatively small, selective subsample. Therefore, we should monitor sample size on which our results are based at all stages of the analysis.

Further, we should be aware of the difference between statistical significance and substantive importance. The assertion that a given coefficient is statistically significant does not tell us anything about its substantive importance. As we have seen, it might even be problematic to rely on standardized regression coefficients for this matter. Instead, we have to make well-founded judgments based, for example, on a comparison with other effects or on the benefits/cost of (changing) the effect.

In the examples we presented in this chapter we relied on cross-sectional data. As we have pointed out, the usual interpretation of regression slopes as reflecting the changes in the dependent variable if the independent variable is increased by one unit is not valid in this situation. Instead, we should say that for someone having a value of $x_j = a + 1$ the conditional expectation for the dependent variable is β_j units higher than for someone with $x_j = a$. If, for example, the effect of one additional year of education on monthly earnings is \$50, then those having 15 years of education are expected to earn \$150 more than those with 12 years of education, all else being equal – that is, controlling for the other independent variables in the model.

We also should remind ourselves that results from cross-sectional analysis should not be used for making predictions. Suppose we decided to increase the earning potential of a person by giving him or her one more year of education. Can we hope that this person's earnings will increase by \$50? Probably not, because many factors which we might not have controlled in our model may lead to higher earnings and may also be responsible for staying in education longer, for example, cognitive abilities and endurance. In contrast to cross-sectional data analysis, these unobserved, time-constant factors can be controlled for in the framework of panel regression, a method discussed in Chapter 15 of this volume (see also Gelman and Hill, 2007, Chapter 9).

Another danger when interpreting regression results from cross-sectional surveys is what may be called the individualistic fallacy. Suppose again that we have found earnings to rise with increased education. Suppose further that we publicize this result widely and encourage people to obtain more education so as to be able to earn more. However, if everyone increases

their level of education, gains in earnings will most likely diminish strongly because the overall situation has completely changed (see Boudon, 1974, for a theoretical and empirical analysis of this phenomenon).

FURTHER READING

Linear regression is covered by almost every introductory textbook in statistics. In addition, there are countless monographs dealing with regression techniques. Therefore, it is neither easy nor particularly important to give advice on further reading. That said, we want to recommend some of the books we like and have profited from. Gelman and Hill (2007) give an excellent introduction to regression analysis. An easy-to-understand introduction to the assumptions underlying linear regression is presented by Berry (1993). Fox (2008) offers a comprehensive overview of linear regression and more general regression models, and Fox and Weisberg (2011) show how these models can be estimated with the R package. A mathematically precise and in-depth coverage of regression models can be found in Wooldridge (2009).

APPENDIX

This appendix lists the items we used to construct the indices reflecting support for restrictive immigration and perceived threat.

Criteria for immigration

Please tell me how important you think each of these things should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here. Please use this card. Firstly, how important should it be for them to:

. . . have good educational qualifications? (D10)
. . . be able to speak [country's official language(s)]? (D12)
. . . have work skills that [country] needs? (D16)
. . . be committed to the way of life in [country]? (D17)

extremely unimportant (0) . . . extremely important (10)

Perceived threat

- Average wages and salaries are generally brought down by people coming to live and work here. (D18)

  agree strongly (1) . . . disagree strongly (5)

- People who come to live and work here generally harm the economic prospects of the poor more than the rich. (D19)

  agree strongly (1) . . . disagree strongly (5)

- Using this card, would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs? (D25)

  take away jobs (0) . . . create new jobs (10)

- Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries? Please use this card. (D27)

  bad for the economy (0) . . . good for the economy (10)

- And, using this card, would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries? (D28)

  cultural life undermined (0) . . . cultural life enriched (10)

- It is better for a country if almost everyone shares the same customs and traditions. (D40)

  agree strongly (1) . . . disagree strongly (5)

- It is better for a country if there are a variety of different religions. (D41)

  agree strongly (1) . . . disagree strongly (5) (reversed)

NOTES

1 In the recent discourse on causal analysis many methods for the identification of a causal effect of x on y have been proposed (see Chapters 12–15 of this volume). Nonetheless, linear regression models may as well be used for causal inference under certain circumstances (Angrist and Pischke, 2009). Some important assumptions that need to be met are described in below and in the next below and in the next chapter.
2 Alternative coding schemes are discussed at `http://statsmodels.sourceforge.net/devel/contrasts.html`.
3 The given test statistic can also be used to test one-sided hypotheses.
4 If we wanted to test for slope differences only we could include an indicator variable for the samples in the pooled analysis (see equation (4.18)).
5 We would obtain the same result when performing a regression analysis on z-standardized variables, that is, in this case the regression coefficients β_j are standardized coefficients. Because the regression always passes through the centroid of the data $(\bar{y}, \bar{x_1}, \ldots, \bar{x_k})$ and because the centroid of z-standardized measures is zero the 'standardized' intercept is also zero.
6 In particular, controlling for irrelevant variables and those that can be considered themselves outcomes of the outcome variable of interest (i.e. endogenous variables) do not belong into a regression equation (cf. Angrist and Pischke, 2009, pp. 64–68). This is why strong theories and carefully constructed models are of paramount importance.
7 We have to assume that we specify the functional form of the relation between covariates and outcome correctly. We can, however, fulfill this assumption by including covariates and their interactions as indicator variables; an approach also dubbed the saturated regression model.

REFERENCES

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Berry, W. D. (1993). *Understanding Regression Assumptions*. Newbury Park, CA: Sage.

Blalock, H. M. (1964). *Causal Inferences in Nonexperimental Research*. Chapel Hill: The University of North Carolina Press.

Boudon, R. (1974). *Education, Opportunity and Social Inequality: Changing Prospects in Western Society*. New York: Wiley.

Bring, J. (1994). How to standardize regression coefficients. *The American Statistician*, 48(3), 209–213.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3), 542–551.

Chao, Y.-C. E., Zhao, Y., Kupper, L. L. and Nylander-French, L. A. (2008). Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. *Journal of Occupational and Environmental Hygiene*, 5(8), 519–529.

Cohen, J., Cohen, P., West, S. and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks: Sage.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Los Angeles: Sage.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Green, E. G. T. (2009). Who can enter? a multilevel analysis on public support for immigration criteria across 20 european countries. *Group Processes and Intergroup Relations*, 12(1), 41–60.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistican*, 61, 139–147.

Humphreys, M. (2009). *Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities*. Working Paper, Columbia University. Last accessed 31.03.2014: `http://www.columbia.edu/~mh2245/papers1/monotonicity7.pdf`.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1), 1–19.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39, 33–38.

Wooldridge, J. M. (2009). *Introductory Econometrics. A Modern Approach*. o.O.: South-Western.