REGRESSION ANALYSIS:

A Primer for the Social Sciences

Marco R. Steenbergen

February 2016

Contents

List of Figures

List of Tables

Preface

Regression analysis remains the work horse of quantitative social science. Although more advanced statistical models have entered the scene in recent decades, there is no doubt that regression analysis retains a prominent place in consumer research, criminology, economics, political science, psychology, public policy, sociology, and other social sciences. Neither is there any doubt that a solid understanding of the linear regression model provides the best access to the advances in statistical modeling that have taken place in the social sciences.

This book aims at introducing the most important topics in linear regression analysis from a social science perspective. This means that the book focuses on variants of the linear regression model that are commonly found in the social sciences. It also means that the examples will be primarily drawn from the social sciences. Finally, the social science perspective means that the focus is mostly on application as opposed to statistical proofs.

For the applied researcher, it is important to know when one should use a particular model, what assumptions are being made, and how one should interpret the results. These three topics form the core of the book. Applied researchers will also be interested in how they can carry out regression analysis on their own. Although the primary focus of this book is not on statistical programming, readers will find R code blocks throughout the text. I decided to focus on R not only because this software is free, but also because it is rapidly becoming the statistical programming platform of choice in the social sciences. R can be downloaded from the Comprehensive R Archive Network for Linux, OS X, and Windows.

Prerequisites

To take full advantage of this book, the reader should be familiar with basic algebra, including exponents, logarithms, linear, and quadratic equations. She should also have completed a first course in statistics, covering descriptive statistics, probability theory, and hypothesis testing. Without this background, the material will be difficult to follow.

In certain parts, this book relies on calculus and matrix algebra. Prior knowledge of these topics is not assumed, as they are covered in the appendix. While calculus and matrix algebra facilitate the understanding of certain concepts, readers should be able to follow the book without them.

Organization

The book is organized into three parts. In Part I, key concepts of regression analysis are introduced in the context of the simple linear regression model. This model allows for a single predictor only. In Part 2, the model is extended to include multiple predictors. This results in the so-called multiple regression model. In this part, we also discuss widely used extensions of the linear regression model such as polynomial regression, categorical predictors, and interactions. In Part 3, we discuss regression assumptions and diagnostics. Here, we also extend the model to time series and panel data. There are three sets of appendices, covering differentiation and optimization, matrix algebra, and regression proofs. I chose to remove the proofs from the main text so that the flow would not be interrupted too much. Although this makes it easier to skip the proofs altogether, I encourage readers to take a look at them as they help to deepen one's understanding of regression analysis.

Acknowledgements

This book contains several animations that were programmed by Julius Mattern and Christian Müller. Without their help, this book would not have been possible. Kushtrim Veseli provided extensive feedback on early drafts of chapters,

for which I owe him a debt of gratitude. Finally, I also would like the thank the many statisticians and programmers who have contributed to the large variety of R packages that allow the analyses in this book. Thanks to their invaluable contributions, R has become the powerful statistical programming platform that it now is.

Zurich, Switzerland—MRS

Part I

Simple Regression Analysis

Chapter 1

Regression as a Descriptive Tool

In simple linear regression analysis, we predict a continuous dependent variable with a single predictor. The method can be construed as a descriptive tool, which aims at clarifying the relationship between two variables in a data set. However, it can also be thought of as a model that represents a set of hypotheses or an entire theory for some outcome. In this chapter, we discuss regression as a descriptive tool, in the way one would do in a course on descriptive statistics. In the next chapter, we will then elaborate on regression as a statistical model.

1.1 A First Example

Consider the data in Table 1.1. These are Labour vote and seat shares (in percentages) in Great Britain in all elections since World War II. Given that Britain is a democracy, we would expect there to be a relationship between the Labour vote and seat shares. The question is what this relationship is. More specifically, if we were to look for a linear relationship between vote and seat shares, what would the linear equation look like? To answer this question, we rely on simple linear regression.

To perform linear regression analysis, we begin by turning the tabular data into a scatter plot. We place the Labour vote shares on the horizontal axis and

Table 1.1: Labour Vote and Seat Shares Since WWII

Date	Vote Share	Seat Share
1945-07-08	47.7	61.4
1950-02-23	46.1	50.4
1951-10-25	48.8	47.2
1955-05-26	46.4	44.0
1959-10-08	43.8	41.0
1964-10-15	44.1	50.3
1966-03-31	48.0	57.8
1970-06-18	43.1	45.7
1974-02-28	37.2	47.4
1974-10-10	39.2	50.2
1979-05-03	36.9	42.4
1983-06-09	27.6	32.2
1987-06-11	30.8	35.2
1992-04-09	34.4	41.6
1997-05-01	43.2	63.4
2001-06-07	40.7	62.5
2005-05-05	35.2	55.2
2010-05-06	29.0	39.7
2015-05-07	30.4	35.7

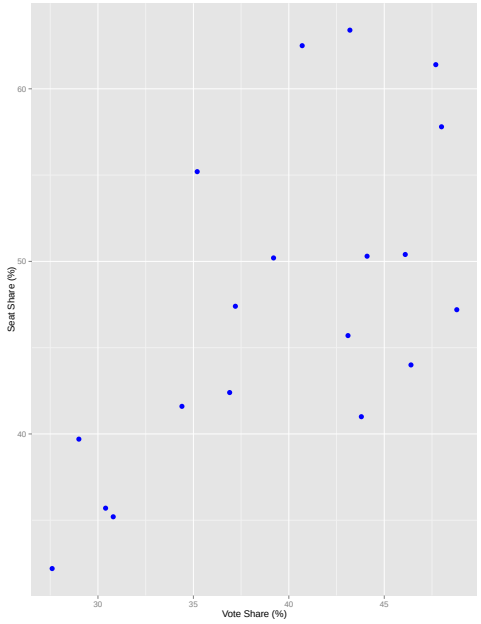
Note: Vote and seat shares are in percentages.

the seat shares on the vertical axis. We then depict each combination of a vote and seat share as a coordinate in a 2-dimensional Cartesian axis system. The result is found in Figure 1.1.

The scatter plot reveals a positive relationship between the vote and seat shares for Labour that, at first glance, looks to be roughly linear. As the Labour vote share increases, the seat share tends to increase as well. Since the points in the scatter plot do not lie on a straight line, we know that the relationship between vote and seat share is not perfect. Indeed, the Pearson product moment correlation between the two variables is 0.639, which is strong but certainly not perfect.

Useful as the scatter plot is, it carries an important limitation. It does not

Figure 1.1: Labour Vote and Seat Shares



Note: The blue points correspond to the pairs of vote and seat shares shown in Table 1.1.

tell us precisely what share of the seats Labour is expected to receive for a given vote share. This is where linear regression analysis enters the picture. In linear regression analysis, we find the line that best fits the cloud of data points in Figure 1.1. This line takes the form of

$$\widehat{\text{Seat Share}} = a + b \cdot \text{Vote Share}$$

Here $\widehat{\text{Seat Share}}$ is the predicted Labour seat share based on Vote Share, a is the **intercept**, and b is the **slope**. We say that we regress the Labour seat share onto the Labour vote share. For the data in Figure 1.1, the **regression line** satisfies the following equation:

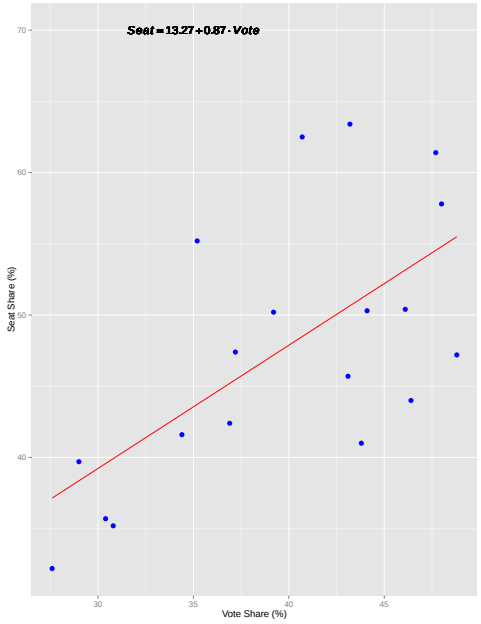
$$\widehat{\text{Seat Share}} = 13.27 + 0.87 \cdot \text{Vote Share}$$

This means that the intercept is equal to 13.27. We can think of this as the seat share that Labour gets regardless of its electoral performance. The slope is 0.87. This means that for each additional percent of the vote share, Labour is expected to receive about eight-tenth of a percent extra of the seat share. The scatter plot with the regression line in the color red is shown in Figure 1.2.

It is important to realize that the regression line gives a prediction that may deviate from the actual observed value of the dependent variable. In fact, unless the predictor and dependent variable are correlated perfectly, there will be discrepancies between the observed and predicted values of the dependent variable. We call such discrepancies the **residuals** of the regression. The smaller the vertical distance of an observed value (the blue points in Figure 1.2 to the regression line, the smaller the residual is. In Figure 1.2, we see hat some of the residuals are small, while others are quite large. You should also keep in mind that the regression line is the best fitting line, which means that we have already drawn it in such a manner that the residuals are minimized in some way. Thus, even though the line represents, in some sense, an optimum this does not mean that the actual and predicted Labour vote shares coincide.

Table 1.2 shows the predicted values and residuals for the Labour vote share in each election. Where the residuals are negative, the predicted seat share

Figure 1.2: Regressing the Labour Seat Share



Note: The red line is the regression line.

Table 1.2: Preditcions and Residuals for the Labour Seat Share

Date	Seat Share	Prediction	Residual
1945-07-08	61.4	54.5	6.9
1950-02-23	50.4	53.2	-2.8
1951-10-25	47.2	55.5	-8.3
1955-05-26	44.0	53.4	9.4
1959-10-08	41.0	51.2	-10.2
1964-10-15	50.3	51.4	-1.1
1966-03-31	57.8	54.8	3.0
1970-06-18	45.7	50.6	-4.9
1974-02-28	47.4	45.5	1.9
1974-10-10	50.2	47.2	3.0
1979-05-03	42.4	45.2	-2.8
1983-06-09	32.2	37.1	-4.9
1987-06-11	35.2	39.9	-4.7
1992-04-09	41.6	43.0	-1.4
1997-05-01	63.4	50.6	12.8
2001-06-07	62.5	48.5	14.0
2005-05-05	55.2	43.7	11.5
2010-05-06	39.7	48.4	1.3
2015-05-07	35.7	39.6	-3.9
Mean	47.5	47.5	0.0

Note: Table entries are percentages.

exceeds the actual vote share. This happens, for example, in the 1959 elections when the Labour seat share was predicted to be 51.2 percent, whereas the actual seat share was only 41.0 percent. Where the residuals are positive, the predicted seat share falls short off the actual seat seat share. This happened, for example, in 2001. In this election, the predicted Labour seat share was 48.5 percent—shy of a majority—whereas the actual seat share was 62.5 percent. Looking at the residuals in greater detail, we notice that they add to 0. Thus negative and positive discrepancies cancel each other. Put differently, on the average, the predicted Labour seat shares are correct. This can be seen in the last row of Table 1.2, which is labeled "Mean."

1.2 Inside Regression Analysis

Now that we have seen an example of how simple regression analysis operates, let us develop some general notation. Every regression analysis starts with a **predictor**, i.e., a variable that is used to predict the dependent variable. Depending on the literature, this variable may also be referred to as the regressor or the independent variable. It can be continuous, as is the case with the Labour vote share. It may, however, also be discrete (although we shall explore the topic of discrete predictors only much later). To allow for a more fine-grained distinction, continuous predictors are sometimes called **covariates**, whereas discrete predictors are sometimes called **factors**.

Every (simple) regression analysis also has a **dependent variable**, which is the characteristic that is being predicted. Depending on the literature, this may also be known as the regressand, outcome or response variable. Whereas the predictor can be both continuous and discrete, the dependent variable is expected to be *continuous*. Seat share is an example of a continuous dependent variable.

In the sample, the dependent and predictor variables are connected through the so-called **sample regression function**:

Equation 1.1

$$\hat{y}_i = a + b \cdot x_i$$

Here $\hat{y}_i$ is the **prediction**, i.e., the value that we would expect Y to take given a value $X = x$ and the values of the intercept and slope.

The predictions are related to the actual values of the dependent variable through the **sample regression model**:

Equation 1.2

$$\begin{aligned} y_i &= \hat{y}_i + e_i \\ &= a + b \cdot x_i + e_i \end{aligned}$$

That is, dependent = prediction plus residual, or, equivalently, $e_i = y_i - \hat{y}_i$. Here e_i is the residual, which is nothing more than a term that makes up the discrepancy between two known quantities: the observed value of the dependent variable and the prediction.

The residuals have the following important property, which we already saw in the analysis of the Labour seat shares (see Table 1.2):

Equation 1.3

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = 0$$

where n denotes the **sample size**. Thus, the residuals average to 0, meaning that, on average, the predictions recover the dependent variable.

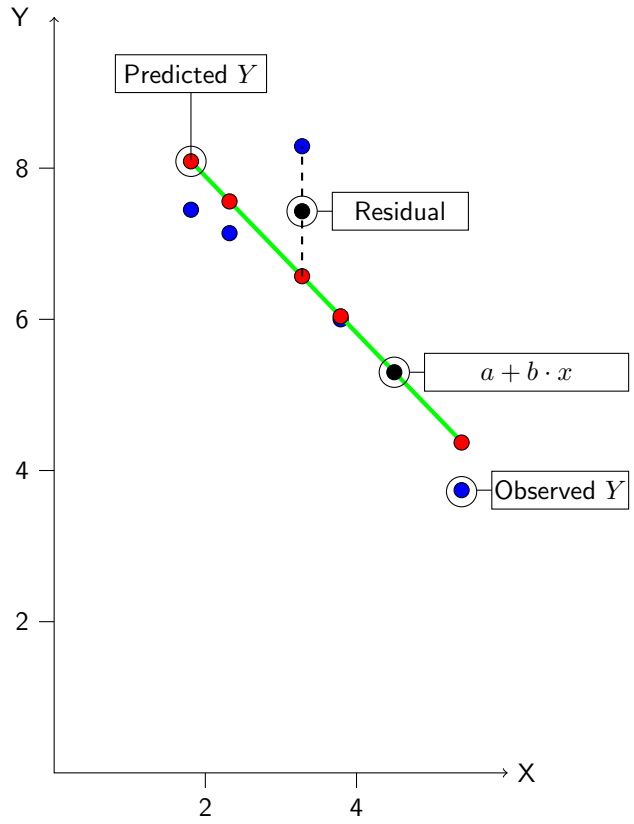
Figure 1.3 shows the regression concepts we have encountered so far, including the observed dependent variable, the prediction, the sample regression function, and the residual.

1.3 Regression Fit

Let us revisit the data from Table 1.1 and consider some alternative forms of the regression line. One possible form is that there is no relationship between the vote and seat shares for Labour. Of course, this would be bad news for democracy but as a theoretical possibility it is worth exploring. In this case, $b = 0$ and

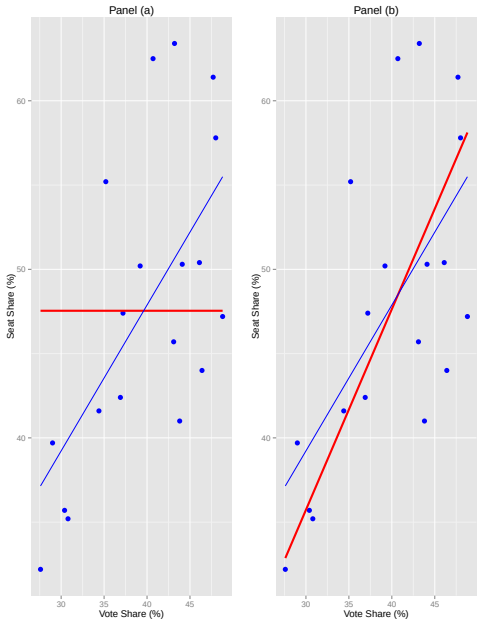
$$\widehat{\text{Seat Share}} = a + 0 \cdot \text{Vote Share} = a$$

Figure 1.3: Anatomy of a Sample Regression



Note: The plot indicates the regression line, the observed and predicted values of the dependent variable, and the residuals.

Figure 1.4: Alternative Specifications of Labour Sear Shares



Note: In panel (a), the slope coeficient is set equal to 0; in panel (b), the intercept is 0.

This scenario produces a flat regression line, as is shown in Panel (a) of Figure 1.4. Here, $a = 47.5$. This is the predicted seat share, which remains constant regardless of Labour's electoral performance. For purposes of comparison, the fitted regression line is shown as well.

A second possible scenario is that the intercept is 0. This is called **regression through the origin** and implies that the predicted Labour seat share is 0 when the Labour vote share equals 0 (which, of course, never happened). With $a = 0$,

$$\widehat{\text{Seat Share}} = 0 + b \cdot \text{Vote Share} = b \cdot \text{Vote Share}$$

The empirical estimate of the slope is $b = 1.19$. This scenario is depicted in Panel (b) of Figure 1.4.

Each of these models presents an alternative idea about the relationship between votes and seats in Britain. Importantly, none of these alternative no-

tions fit the data as well as the regression equation with which we started out: $\widehat{\text{Seat Share}} = a + b \cdot \text{Vote Share}$. How do we know this? As a criterion, we can use the sum of the squared residuals:

Equation 1.4

$$SSE = \sum_{i=1}^{n} e_i^2$$

Here SSE is known as the **sum of squared errors**. It has this name because the residuals can be viewed as prediction errors. The smaller the SSE, the smaller the prediction errors, and the better the fit. Indeed, if all of the observed values of the dependent variable are on the regression line, then the SSE is 0 and we have a perfect fit.

The SSEs for the regressions in Figures 1.2 and 1.4 are shown in Table 1.3, along with key characteristics of those regressions. It is clear that the smallest SSE is found for the original regression of Figure 1.2. The remaining specifications all have larger—sometimes much larger—SSEs. This lends credence to our earlier claim that regression analysis selects the *best* fitting line.

The reason that the models from Figure 1.4 fit so poorly is easily understood when we contrast the estimates to the implied constraint. The model in panel (a) stipulates $b = 0$, but the value we obtained in Figure 1.2 was 0.87, quite a ways removed from 0. The model in panel (b) stipulates $a = 0$, but this is again far removed from the value we obtained in Figure 1.2, which was 13.27. When restrictions like $a = 0$ and $b = 0$ are false, i.e., they do not correspond to the data, then this is automatically translated into an inferior fit.

Table 1.3: Regression Specifications and Fit

Specification	a	b	SSE
Figure 1.2	Estimated	Estimated	926.35
Figure 1.4(a)	Estimated	Fixed at 0	1564.00
Figure 1.4(b)	Fixed at 0	Estimated	1019.00

In the practice of social research, the SSE is often converted into the **coefficient of determination**, which is also known colloquially as the **R-squared**:

Equation 1.5

$$R^2 = 1 - \frac{SSE}{SST}$$

Here

Equation 1.6

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

is the **sum of squares total**, which measures the variation in the dependent variable. Like the SSE, the coefficient of determination has a clear lower bound of 0. Unlike the SSE, which does not have a clear upper-bound, the coefficient of determination is bounded from above at 1. This upper-bound is reached if $SSE = 0$ and indicates that the regression fits the data perfectly. We say that the regression accounts for 100 percent of the variance in Y. The lower-bound is reached when the regression line is flat, as in Panel (a) in Figure 1.4. In this case, $SST = SSE$ and $R^2 = 0$. We say that the regression accounts for none of the variance in Y.

In the regression in Figures 1.2, the sample variance in Labour seat shares is 86.89. Since $SST = (n-1) \cdot s_Y^2$, it follows that $SST = 1564.026$. Applying Equation 1.5 yields an R-squared value of 0.408 for the fitted model. Thus, we can say that the Labour vote share explains around 40.8 percent of the variance in the Labour seat share. Earlier, we saw that the Pearson product-moment correlation between the seat and vote shares was $r = 0.639$. If we square this correlation we obtain the R-squared of 0.408 exactly. Thus, in the simple regression model, the R-squared is simply equal to the square of the Pearson product-moment correlation, as is shown in Appendix C:

Equation 1.7

In simple regression analysis,

$$R^2 = r^2$$

1.4 Conclusion

In this chapter, we have looked at simple regression analysis as a descriptive tool. We have seen that the regression line is the best fitting line to a set of data points in a scatter plot and that this line can be used to make predictions about a dependent variable. This is not the only way to look at simple regression analysis, however. It is also possible to view this type of analysis as a statistical model that pertains to a population and is estimated using sample data. We shall introduce this idea in the next chapter.

Chapter 2

Regression as a Model

In this chapter, we look at regression analysis from a new perspective, namely that of statistical modeling. After a brief general introduction to (statistical) modeling, we derive simple regression analysis as one particular instance of a statistical model. We also state the assumptions of the simple regression model, which play an important role in statistical inference, as we shall see in Chapter 3.

2.1 Models in the Social Sciences

Models play a central role in the social sciences, especially in quantitative approaches. To understand what they are, we invoke one of several definitions that the *Oxford English Dictionary* offers:

> A simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.

Several features stand out in this definition. First and foremost, all models are simplifications. Just like a model air plane lacks many features that a real air plane possesses, a social scientific model is no carbon copy of social reality. Nor

is it meant to be, for simplification is the whole intent and purpose of models, as we shall see.

A second key feature is that models are used to enhance our empirical understanding of phenomena and to help us make predictions. We formulate a model to enhance our grasp of phenomena such as democracy, conflict, and distribution and, to a lesser extent, to make future predictions for those phenomena. In this sense, all models are tools used to enhance insight.

A third feature mentioned by the *Oxford English Dictionary* is that models are typically stated in mathematical terms. This is certainly true of statistical models such as the linear regression model. Mathematics is not the only modeling language, however. Especially in the social sciences, all modeling starts and ends with verbal representations. That is, we use natural language to state the model before rendering it in mathematical terms. Once we have estimated the model—a topic we shall discuss in Chapter 3—we return to natural language to offer interpretations and derive implications.

The idea that models are simplifications requires some further attention. It is an idea that stems from neo-positivist philosophy of science. Bryman (1988) calls this the *doctrine of elementarism*. Here, the crucial assumption is that the best pathway toward knowledge and understanding is to break down a phenomenon into its parts and to study the relationships between those parts. The key to the whole exercise is to retain those parts that are crucial to the phenomenon and to discard everything else. Indeed, in the spirit of simplification, the modeler would like to retain as few parts as is necessary to grasp essential features of the phenomenon under study. All of this stands in sharp contrast with holistic approaches, which aim at grasping a phenomenon in its full complexity.

At first sight, the holistic approach would seem far superior to the doctrine of elementarism. Why should one obtain only a partial understanding of a phenomenon instead of pursuing the complete truth? However, there are both pragmatic and philosophical arguments that would cause one to favor elementarism over holism. Pragmatically speaking, it is obviously tremendously complicated to understand a social phenomenon in all of its facets. Those that claim to be holistic are frequently criticized for not accomplishing this goal.

Indeed, from a psychological perspective (e.g., bounded rationality) it is even questionable that our brains are designed for holistic understanding. From a philosophical perspective, it may also be the case that elementarism suffices. Think, for example, of an engineer who is testing the aerodynamic properties of a new aircraft design. For this purpose, a scale model typically is good enough. This model does not have functioning engines, control surfaces, or avionics, nor does it contain passenger seats. None of these features are essential for modeling the aerodynamics of the aircraft. In the social sciences, we may also not need to know everything about, for example, societies to understand how party systems form. It may suffice to know something about the social cleavages in a society. Indeed, many theories in the social sciences are in essence reductions of phenomena to core elements. This is true even of grand theories such as Marxism, which singles out property relations as the core element for understanding economic and social development.

When we take the doctrine of elementarism seriously, then it follows that we cannot expect models to tell us "the truth, the whole truth, and nothing but the truth." Due to their simplifying logic, it is clear that models cannot deliver the whole truth. Nor is this the intent, so to hold models to this standard is both nonsensical and unfair.

What, then, can models deliver? By what standards should they be judged? Gilchrist (1984) formulates two such standards:

1. *Alitheia*, i.e., to make unhidden what might otherwise remain hidden. By this criterion, a model has value when it uncovers aspects that would otherwise be obscured by other, less fundamental, features of the problem one is researching.

2. *Adeaquatio intellectus*, i.e., to provide insight. This broader than *alitheia* because insight may be obtained even when hidden features remain hidden. The critical question here is: Do we learn something from the model?

It is perfectly valid to apply these utilitarian criteria to models because they are central to the very activity of modeling. That is, we build models to gain insight and/or make thinks unhidden. Some models may do a better job at that than

others. However, no model can deliver the whole truth. To cite the statistician G.E.P. Box, "all models are wrong, but some are useful."

2.2 Verbal and Mathematical Models

To understand models and modeling languages, let us consider a simple example. Imagine we are interested in foreign direct investment (FDI) in sub-Saharan Africa. The units of analysis are years. In each year we record the per capita FDI in the region. Our goal is to explain the per capita FDI level in each year.

Imagine we believe that per capita FDI depends on one thing and one thing only, to wit the level of political turmoil in the region. Of course, it is unlikely that turmoil is the only factor driving FDI but, for the sake of argument, let us say that we are wiling to make that assumption. We could now formulate the following verbal model: per capita FDI in sub-Saharan Africa depends on the level of political turmoil in the region. We could even lend a specific direction to this model by saying that increased levels of turmoil tend to decrease FDI. The model itself may derive from a larger theory about risk aversion in investors, which causes them to reduce investments in politically unstable regions.

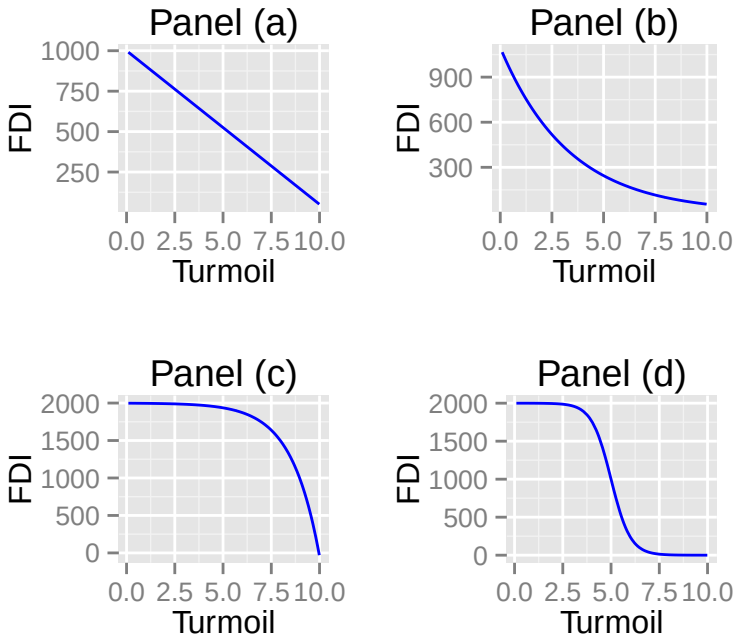
What the verbal model does is to use natural language to express an idea about the drivers of FDI. We have reduced FDI to what we consider to be its core: political turmoil. In the process, we have left out myriad other factors such as the economic state of the region and political corruption. Without realizing it, we have applied the doctrine of elementarism.

The major drawback of verbal models is that they lack precision. A statement like "increased levels of political turmoil tend to decrease FDI" immediately raises a number of questions. How much does FDI decrease? And what kind of increase in political turmoil does it take to bring about this decrease? We can add precision by turning the verbal model into a mathematical model. The language of mathematics adds precision. In addition, it makes available a whole new set of operators.

In its most general form, a mathematical model of FDI may be written as

$$h(y, \theta) = f(x, \beta)$$

Figure 2.1: Four Different Mathematical Models of FDI



Note: A linear (panel (a)), logarithmic (panel (b)), exponential (panel (c)), and logistic (panel (d)) model of the relationship between FDI and political turmoil.

Here y stands for FDI and x stands for political turmoil. The symbols f and h stand for two different functions. The symbols β and θ represent different (sets of) parameters that influence the form of the two functions.

The generic model can be filled in by specifying the nature of f and h and by defining β and θ. For example, we could theorize that FDI is a linear function of political turmoil (see Panel (a) of Figure 2.1):

$$y = \beta_0 + \beta_1 x$$

Here $h(y, \theta) = y$ (the so-called, identity link), $f(x, \beta)$ is a linear function, and β consists of an intercept (β_0) and a slope (β_1). Alternatively, we could specify

the following model, which is depicted in Panel (b) of Figure 2.1:

$$y = e^{\beta_0 + \beta_1 x}$$

Here $h(y, \theta) = y$, $f(x, \beta)$ is an exponential function, and β again has two elements, β_0 and β_1. An equivalent formulation of this model is

$$\ln y = \beta_0 + \beta_1 x$$

Here $h(y, \theta)$ is a logarithmic function, $f(x, \beta)$ is a linear function, and β consists of a slope and intercept. As displayed here, increases in political turmoil have a particularly severe effect when the initial turmoil level is low; when there is a lot of turmoil to begin with, a further increase does not have as much of an effect. The reverse of this pattern is demonstrated in Panel (c) of Figure 2.1. Here, increases in turmoil have little impact when the initial level of turmoil is low. But when turmoil is already high, a further increase has a large effect. The mathematical specification is

$$y = \beta_0 + \beta_1^x$$

Finally, consider

$$y = \frac{\beta_0}{1 + \exp(\beta_1(x - \beta_2))},$$

which is depicted in Panel (d) of Figure 2.1.[1] Here $h(y, \theta) = y$, $f(x, \beta)$ is the logistic function, and β consists of three parameters, to wit β_0, β_1, and β_2. Now FDI is relatively irresponsive to political turmoil when the initial level of turmoil is low. It is also relatively irresponsive to turmoil when there is already a lot of it. However, at some level of turmoil—around 5 in the graph—a further increase produces a rapid decline in FDI. Of course, the specifications shown here do not even begin to scratch the surface of how the relationship between FDI and political turmoil may be modeled. What all of these models have in common, however, is that they specify precisely how FDI and political turmoil are related.

[1] The expressions $\exp q$ and e^q mean the same thing.

One of the things mathematical models allow us to do is to quantify the effect of some variable x on another variable y, e.g., the effect of political turmoil on FDI. One of the easiest ways to do this is by computing **marginal effects**. These are defined as the change in an outcome such as y relative to an infinitesimally small change in x. We can think of this as the instantaneous rate of change. Mathematically, the marginal effect for the outcome y is given by

Equation 2.1

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{\partial y}{\partial x}$$

Here Δx denotes the change in x, which we let go to 0. Further, Δy is the change in y. Finally, $\partial y / \partial x$ is the first partial derivative of y with respect to x (see Appendix A). For Panel (a) in Figure 2.1, the marginal effect is

$$\frac{\partial y}{\partial x} = \beta_1$$

We see that this effect does not depend on x: a very small increase in political turmoil always produces the same change in per capita FDI regardless of the starting point at which the increase is introduced.

This is certainly not true for the remaining models depicted in Figure 2.1. For the model in Panel (b), for example,

$$\frac{\partial y}{\partial x} = \beta_1 e^{\beta_0 + \beta_1 x}$$

Here, we see that the marginal effect is a function of x, so that the impact of a small change in x depends on the starting point for x. For Panel (c), the marginal effect is

$$\frac{\partial y}{\partial x} = \beta_1^x \ln \beta_1$$

and for Panel (d) it is

$$\frac{\partial y}{\partial x} = -\beta_1 \frac{\beta_0 \exp(\beta_1(x - \beta_2))}{(1 + \exp(\beta_1(x - \beta_2)))^2}$$

All of these marginal effects clearly depend on the initial level of x, e.g., political turmoil. We shall be using marginal effects like these frequently in the interpretation of linear regression models.

An alternative way of quantifying an effect is by using the **discrete change**. Here, we change the predictor by δ units, i.e., $\Delta x = \delta$. We now define the discrete change as

Equation 2.2

$$\Delta y = f(x + \delta, \beta) - f(x, \beta)$$

For example, for the model in Panel (a) of Figure 2.1, the discrete change is

$$\Delta y = [\beta_0 + \beta_1(x + \delta)] - [\beta_0 + \beta_1 x] = \beta_1 \delta$$

We typically use discrete changes when it makes no sense to assume an infinitesimally small change in x, for example, when x is discrete.

2.3 Statistical Models

We have spent some time on mathematical models because they are intimately related to statistical models. Indeed, all statistical models are mathematical models in that they use the language of mathematics to represent their contents. Statistical models, however, have some unique features that set them apart. First, statistical models are models of a **data generating process** (DGP) and, in this sense, they are intrinsically empirical in their focus. When we conduct statistical research, we collect data. The data generating process is the process we believe to have generated this data. A statistical model formalizes the DGP that is being theorized.

Second, statistical models are inherently **stochastic**. They explicitly recognize the role that uncertainty plays in producing data and make this an essential feature of the model **specification**. More specifically, statistical models include one or more **error terms**. These are unobserved random variables that are included to capture (1) omitted predictor variables; (2) measurement error in the dependent variable; and (3) idiosyncratic variation. In a sense, the error term acknowledges the limitations of the doctrine of elementarism. It captures all those aspects of the DGP that are not captured through the elements that we have singled out as partial explanations of the dependent variable. If we call the latter part the fixed component of the model, then we can say that the model is made up of both fixed and stochastic components. The presence of a stochastic term driving the dependent variable means that it, too, is a random variable.

Having defined the key distinguishing characteristics of statistical models, we can now think more specifically about the elements that make up a model. Following the literature on generalized linear models (McCullagh and Nelder, 1983), we can identify the following elements:

1. **Distribution:** We specify a particular probability distribution, i.e., a probability density or mass function, for the dependent variable.

2. **Outcome:** By outcomes, we mean one or more parameters of the distribution that are explicitly modeled as a function of a set of predictors. Not all parameters have to be turned into outcomes of predictors. However, every model has at least one parameter that is turned into an outcome.

3. **Linear Predictor:** A function of the predictors that is linear in the parameters, i.e., a function with parameters that act as multiplicative weights of the predictors and where the weighted predictors are summed to form a linear composite. With only a single predictor, this takes the form of

Equation 2.3

$$\eta_i = \beta_0 + \beta_1 x_i$$

4. **Link function:** A function that links the linear predictor to an outcome of interest.

These elements help to formulate a large variety of statistical models. We now consider a specific variant of statistical models, to wit the simple linear regression model.

2.4 The Simple Linear Regression Model

2.4.1 Basics

Consider a population with a random variable Y that is continuous and unbounded (i.e., its support is the real number line). We postulate that

Equation 2.4

$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

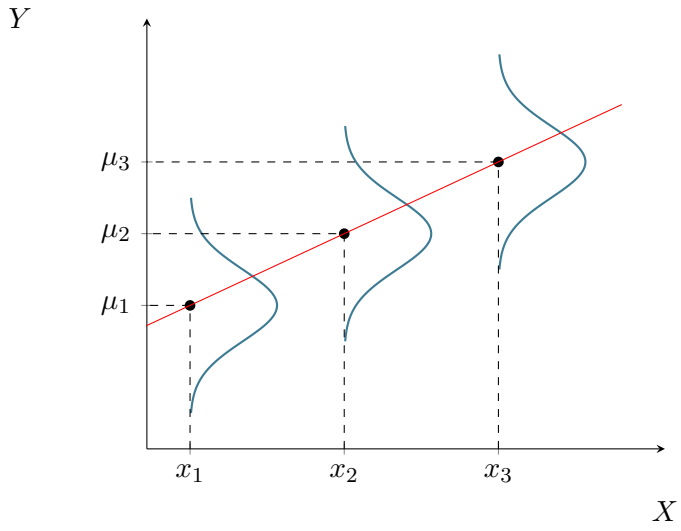
That is, for each population unit i, the dependent variable is normally distributed. The standard deviation of the normal distribution does not vary across units, a condition that is called **homoskedasticity**. By contrast, the mean of the dependent variable fluctuates across units, whence the subscript i on μ. As such, it constitutes an outcome in the sense of the previous section. We model this outcome using Equation 2.3, so that

Equation 2.5

$$\begin{aligned} \mu_i &= \eta_i \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

Since the mean is identical to the linear predictor, we say that the link function is the *identity link*. Taken together, Equations 2.4 and 2.5 constitute the **population regression model**. Equation 2.5 is known as the **population regression function**. This function expresses the conditional expectation of Y

Figure 2.2: The Population Regression Model



Note: The red line is the regression line. The black dots on the line are predictions based on different values of X. The dark blue normal distributions centered about the black dots reflect the uncertainty in the predictions. The more variance, the wider the distributions, and the greater the uncertainty.

given a particular value of the predictor: $\mu_i = E[y_i|x_i]$. As such, it is what we would expect to see given the value x and the linear predictor $\beta_0 + \beta_1 x_i$.

Figure 2.2 illustrates the model. The red line represents the population regression function, $\mu_i = E[y_i|x_i] = \beta_0 + \beta_1 x_i$. This line gives the values that we expect to observe for the dependent variable, given a particular value of the predictor. As such it gives the conditional expectation of Y. The three black dots on the red regression line represent three different expected values of the dependent variable. Specifically, for $X = x_1$, we obtain a prediction $\mu_1 = E[y|x_1] = \beta_0 + \beta_1 x_1$, for $X = x_2$ we obtain $\mu_2 = E[y|x_2] = \beta_0 + \beta_1 x_2$, and for $X = x_3$ we obtain $\mu_3 = E[y|x_3] = \beta_0 + \beta_1 x_3$. About the population regression function, we have drawn normal distributions of the type $\mathcal{N}(\mu_i, \sigma)$. These give the distribution of the observed dependent variable around the regression line. For example, at $X = x_1$ we expect a value of μ_1 for the dependent variable. But some of the actual values y are larger than μ_1; these fall in the left tail

of the normal distribution that is drawn at $X = x_1$. Other observed values of Y are smaller than μ_1; these fall in the right tail of the normal distribution. In general, the wider the dispersion of the normal distribution, the more the observed values of Y tend to deviate from the conditional expectations, μ_i, and the less predictable the dependent variable is. Figure 2.2 also clearly shows that, while μ_i shifts depending on the value of X, the standard deviation of the normal distribution remains constant; the spread of the normal distribution does not vary. This is the homoskedasticity assumption.

The model as we have derived it, relies heavily on the terminology and notation of the generalized linear modeling literature (McCullagh and Nelder, 1983). Econometricians (e.g., Greene, 2011) typically use a different notation for the model. In this notation, the error term makes an explicit appearance:

Equation 2.6

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma) \end{aligned}$$

In this parametrization, ε is the error term we discussed earlier. We assume this error term to be normally distributed with a mean of 0 and a standard deviation of σ. Due to these assumptions, it follows that $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$, which is identical to what Equations 2.4 and 2.5 state. That the mean of Y reduces to $\beta_0 + \beta_1 x_i$ is easily shown:

$$\begin{aligned} E[y_i] &= E[\beta_0 + \beta_1 x_i + \varepsilon_i] \\ &= E[\beta_0] + E[\beta_1 x_i] + \underbrace{E[\varepsilon_i]}_{= 0 \text{ by assumption}} \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

We see that the econometric and generalized modeling views of the linear regression model are in the end indistinguishable. It is good to know both views. The econometric parametrization is commonly found in the social sciences and

we shall rely heavily on it in this textbook as well. The generalized linear modeling parametrization, on the other hand, is useful because statistical programs such as R use it in their syntax for statistical commands. We shall see this, for example, in Chapter 3.

2.4.2 Relationship with the Sample Regression Model

How does the population regression model relate to the model shown in Equation 1.2? Stated in the simplest way, Equations 2.4-2.6 pertain to the population, whereas Equation 1.2 pertains to a sample. Going below the surface, the quantity a in Equation 1.2 is the estimator of the parameter β_0 in Equations 2.5-2.6, whereas the quantity b in Equation 1.2 is the estimator of the parameter β_1 in Equations 2.5-2.6. Similarly, the residuals in the sample regression model may be viewed as estimators of sorts the error term in Equation 2.6, although we should keep in mind that they do not behave quite the same way. How these estimators are obtained is the subject of the next chapter.

2.4.3 Interpretation

The population regression function may be interpreted in a number of different ways. One approach is to compute the marginal effect for the outcome μ:

Equation 2.7

$$\frac{d\mu}{dx} = \frac{\partial(\beta_0 + \beta_1 x)}{\partial x} = \beta_1$$

This may be interpreted as the instantaneous rate of change in the conditional expectation of Y. We see this rate is constant. Alternatively, it is possible to compute the discrete change in μ due to a change of δ units in X:

Equation 2.8

$$\Delta\mu|\Delta x = \delta = [\beta_0 + \beta_1(x + \delta)] - [\beta_0 + \beta_1 x] = \beta_1\delta$$

Usually, we set $\delta = 1$. We can then say the following:

> For a unit increase in X, Y is *expected* to change by β_1 units.

If $\beta_1 = 0$, then we do not expect the dependent variable to change at all—there is no linear relationship between Y and X. If $\beta_1 < 0$, then we expect the dependent variable to decrease for a unit increase in X. Finally, if $\beta_1 > 0$, then we expect the dependent variable to increase for a unit increase in X.

2.4.4 Assumptions

The population regression model of Equations 2.3-2.5 is actually somewhat incomplete. Usually, we add a number of assumptions, whose purpose will become apparent in Chapter 3. It is important to understand these assumptions, as they are part of the model and are usually not innocuous. Indeed, most of the third part of this book pertains to the question what to do when these assumptions fail.

To create some order into the assumptions, we divide them into three rubrics: (1) assumptions about the predictors; (2) assumptions about the error terms; and (3) assumptions about the relationship between the predictors and the error terms. We shall utilize this organizational scheme throughout the book, adding to it when this becomes necessary as we extend the regression model.

Assumptions about the Predictors The linear regression model does not make many assumptions about the predictors. As we have already seen, for example, these can be both continuous and discrete. In addition, they can be transformed in any admissible way. For example, $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is a perfectly legitimate regression model. While this model is no longer linear in the predictors (the predictor is now squared), it is still linear in the parameters (the parameters occur as multiplicative weights and the various terms—constant and quadratic x—are summed together). Thus, the linear regression model is extremely flexible when it comes to the predictor side of things. This one reason why the model remains a powerful tool of quantitative social research.

There is one aspect of the predictors, however, where we usually assume away flexibility. This is the assumption that the only stochastic variable in the

model is the dependent variable. The predictors are generally not considered random variables or, put differently, we assume them to be fixed.

Assumption 2.1

The values x of the predictor are assumed to be fixed in repeated samples.

What does it mean for the values of a predictor to be fixed in repeated sampling? It means that the values of the predictor are under the complete control of the researcher. The researcher determines which values of the predictor occur and with what relative frequency. Put differently, "nature" and its whims play no role in generating the values x. Consequently, the only source of random variation in the dependent variable is the error term.

Stating that the values of the predictor are fixed is a convenient assumption. We relied on it already once when we demonstrated that $E[y_i|x_i] = \beta_0 + \beta_1 x_i$.[2] But is the assumption also reasonable? It depends on the nature of the research that is being conducted. In experimental research, it is indeed the case that the researcher has complete control over the values of the predictor. She decides how many values of the predictor are realized, what these values are, and how often they occur in the experiment. Moreover, the same values and relative frequencies will emerge in each iteration of the experiment, no matter how frequently it is being conducted, as long as the experimental protocol remains unchanged. This is tantamount to saying that the values x are fixed in repeated sampling. While experiments are still somewhat rare outside of psychology, they have gained popularity everywhere in the social sciences. Thus, Assumption 2.1 is applicable to many social scientific studies.

Most social research remains non-experimental, however, and here Assumption 2.1 is much less plausible. Measures of the predictors are usually collected at the same time as those of the dependent variable, using surveys and other methods. To say that some of these measures—those of the predictor—are fixed, whereas others—those of the dependent variable—are not, would seem,

[2] Otherwise, we would have had to write $E[y_i] = \beta_0 + \beta_1 E[x_i]$. With x_i being fixed, we can treat it as a constant, so that $E[x_i] = x_i$.

at best, heroic.

Assumptions about the Error Terms We generally make several assumptions about the error term. One assumption that we have already seen is the following.

Assumption 2.2

The error terms are normally distributed.

Due to the normality assumption, we argue that the error terms follow a distribution that is symmetrical around the mean, as well as bell-shaped. Thus, the probability mass for negative errors equals the probability mass for positive errors. Further, large errors are less common than smaller errors. The corollary of Assumption 2.2 s that the dependent variable is normally distributed.

The normality assumption is a useful first approximation of error processes. In many cases, it also has a great deal of face validity. Remember that one of the ingredients of the error terms is measurement error. It is usually reasonable to assume that measurement errors are symmetric and that really large errors are less likely than smaller ones. But there are exceptions to this rule. For example, if some responses on the dependent variable are more socially desirable than others, then you would not expect measurement errors to be symmetrically distributed. More people would err on the side of the socially desirable response than on the side of the less desirable response. Under these circumstances, you would want to avoid a symmetry assumption. In other cases, the symmetry assumption may be reasonable, yet normality does not seem to be the most appropriate assumption. In the political agendas literature, for example, changes in policy priorities tend to be smaller than normality would imply (e.g., Baumgartner et al., 2009). There is excess kurtosis that, by definition, is inconsistent with normality. In Part 3, we shall briefly look into the topic of violations of the normality assumption.

A second assumption about the error term that we have already seen concerns its mean.

Assumption 2.3

$$E[\varepsilon_i] = 0$$

This assumption implies that there is no systematic direction to the error terms. Over all units, negative and positive errors cancel each other so that $\mu_i = \beta_0 + \beta_1 x_i$.

This assumption is often reasonable, but there are circumstances under which it is not valid. One example is again the presence of a social desirability bias. When there is a systematic tendency to err in a particular direction, then there is no reason to believe that the errors average to zero. Another example are stochastic frontiers in economics, where production processes may suffer from random shocks (e.g., due to the weather) but may also have unobserved built-in inefficiencies that cause the predicted production to be too optimistic (e.g., Kumbhakar and Lovell, 2003).

The third assumption about the errors, we have also seen already: the errors are supposed to be homoskedastic.

Assumption 2.4

$$Var[\varepsilon_i] = \sigma^2$$

That is, the values of the dependent variable are equally predictable (or unpredictable) for all units. More specifically, the variance around the regression line is not a function of the predictors.

Like the other assumptions we have made so far, the homoskedasticity assumption is usually a reasonable starting point. However, there are situations were it is a priori implausible. One such situation arises when there are learning/expertise effects. Imagine we are interested in modeling how quickly street level bureaucrats (e.g., police officers) dispense with their tasks. We have one bureaucrat who is new at the job and another one who has done it for 10 years. It seems likely that the seasoned bureaucrat is much easier to predict, i.e., her variance is smaller, than the novice. To assume, then, that the variance is equal

for all bureaucrats is implausible.

Another situation where homoskedasticity is implausible is when actors have different amounts of leeway for discretionary behavior. Take, for example, consumer spending on luxury goods. Spending levels on these goods may be much more predictable for those with small incomes than for those with large incomes. Those who earn little will have little discretionary spending power, which means that there is little room for spending on luxury goods. This translates into a small variance around the regression line when x (income) is low. Those who earn a lot, also can afford a great deal of discretionary spending. But whether they use this to purchase luxury items depends on unobserved factors such as the utility of luxury goods, which end up in the error term and which can vary dramatically. Thus, we would expect a wide variation in luxury spending, with some high income people spending nothing on these items and others spending a lot. Again, to assume that all income groups display the same variation in the dependent variable is a priori implausible, so that homoskedasticity may not be the right starting point.

There is yet a fourth assumption that we frequently make about the error terms. This one, we have not yet seen and it states that the errors are uncorrelated with each other. We say that there is **no autocorrelation**.

Assumption 2.5

$$Cov[\varepsilon_i, \varepsilon_j] = 0$$

for $i \neq j$.

In Chapter 3, we shall see why we need this assumption. For now, we ask again the question whether it is reasonable. This depends on the research design that is being used. With **cross-sectional data**, we observe a sample of n units at a single point in time. For such data, the assumption of no autocorrelation is frequently reasonable, although this depends on the sampling design. Under *simple random sampling*, the observations are independent, which also means that it is reasonable to assume that $Cov[\varepsilon_i, \varepsilon_j] = 0$ for $i \neq j$. Especially in survey research, however, cross-sectional data are obtained frequently via cluster

sampling and, in this case, the assumption of no auto-correlation is valid only in a limited sense. Specifically, the errors of units from different clusters may be assumed to be uncorrelated, but the errors of units from the same cluster are typically correlated.

With **time series data** the assumption of no autocorrelation is violated almost by definition. Such data, which will be discussed in greater detail in Part 3, consist of successive time points such as days, weeks, months, quarters, or years. The subscript i on ε_i thus references a particular time point. If we now take the errors at time points i and j, they are almost always correlated, especially when j is immediately adjacent to i. The reason is quite simple: a random shock that occurs at time point i usually will continue to be felt at time point j. Take, for example, the financial crisis of 2008. The Lehman Brothers bankruptcy in September, 2008 constituted a shock to the financial markets that did not immediately dissipate. It continued to be felt in October—and far beyond—affecting the stock markets and other aspects of the world economy. The "staying power" of shocks causes the assumption of no autocorrelation to be violated in time series data.

Assumptions about the Relationship between the Predictor and the Errors The last assumption that we should discuss concerns the relationship between X and ε. In regression analysis, we typically assume **exogeneity** of the predictor (Engle, Hendry and Richard, 1983). Stated mathematically,

Assumption 2.6

$$E[\varepsilon_i|x_i] = 0$$

This means that the errors are independent from the predictor.[3]

Innocuous as Assumption 2.6 may look at first sight, it has far reaching

[3]Assumption 2.6 states the definition of *strict* exogeneity. In many cases, we require only *weak* exogeneity, which can be stated as $E[\varepsilon_i x_i] = 0$. This amounts to a lack of correlation between the error term and the predictor, which is a subset of independence. Also note that the law of iterative expectations implies that $E[\varepsilon_i] = 0$—Assumption 2.3—if Assumption 2.6 is satisfied.

implications. Specifically, the assumption implies:

1. Any and all omitted predictors are unrelated to the one included in the model.

2. The functional form is correctly specified.

3. There are no feedback loops between Y and X.

Taken together, the implications mean there are no **specification errors**: the model specifies the correct DGP—it is *correctly specified*.

It is a rather strong statement to assume away specification errors, especially in the case of the simple regression model, where everything hinges on a single predictor variable. Just on its face, it would seem that no single predictor can do justice to the DGP that underlies such complex social phenomena like crime, electoral performance, organizational behavior, and mental illness, to name just a few disparate examples. Let us consider a few scenarios of what can go wrong with Assumption 2.6 in the practice of social research.

To do this, we revisit the earlier example of political turmoil and FDI in sub-Sharan Africa. In this example, we stipulated to have time series data for a number of years. Denoting each year by i, the linear regression model can be formulated as a stochastic version of Panel (a) in Figure 2.1:

$$\text{FDI}_i = \beta_0 + \beta_1 \text{Turmoil}_i + \varepsilon_i$$

As per assumption 2.6, $E[\varepsilon_i|\text{Turmoil}_i] = 0$. But is this realistic? First, consider the notion that omitted predictors are unrelated to the predictor. One of the predictors missing from the model is economic growth. Because we do not explicitly include the predictor in the model, it becomes a part of the error term. Assumption 2.6 implies that we assume economic growth to be unrelated to political turmoil, but this seems odd. After all, it is plausible that growth suffers when turmoil increases, implying a clear relationship rather than the absence of one.

Second, consider the idea that the functional form is correctly specified. This means that the relationship between FDI and political turmoil is indeed

linear, not just in the parameters but also in the predictor. But what if the relationship is more like that shown in Panel (b) of Figure 2.1? Then we should have included a term of the form $\exp(\text{Turmoil})$. The fact that we did not, means that this term now ends up in ε. Since the omitted predictor is a function of the predictor in the model, it is actually impossible to satisfy Assumption 2.6: the central tendency of the errors is a function of turmoil and it is ludicrous to assume otherwise.

Finally, consider the idea that there shall be no feedback mechanism from Y to X. In our example, this means that changes in FDI should not produce changes in political turmoil. But one could easily imagine a situation in which there is feedback. As FDI deteriorates, for example, one could imagine that this has a negative effect on the economy, which in turn may create political discontent. If this mechanism holds, then turmoil is a function of FDI. However, since FDI is driven in part by an error term, *ipso facto* turmoil is also driven by the error term. In this case, there is a possible correlation between turmoil and ε, which would invalidate Assumption 2.6.

2.5 Conclusion

In this chapter, we developed a modeler's perspective on simple regression analysis. Key to this perspective is to view regression as a statement about a data generating process. In this statement, we specify the dependent variable and the predictor. We also develop an explicit function that relates the predictor to the dependent variable. All of this can be viewed as a formalization of a verbal theory of some social or political phenomenon. Key, too, is that we explicitly allow for uncertainty in this formalization by incorporating a so-called error term.

We have also seen that the simple regression model typically comes with a large number of assumptions. In this chapter, we identified as many as six of them. While these assumptions sometimes are written off as mere technicalities of the regression analysis, we have seen that they are actually empirical statements. More specifically, the assumptions reflect suppositions about the data generating process. These suppositions may be wrong and this will impact the veracity of the regression results. Why this is so will become apparent in the

next chapter, which deals with statistical inferences about the linear regression model.

Chapter 3

Statistical Inference in Simple Regression

In Chapter 2, we saw that the sample regression model can be viewed as the estimated counter-part of the population regression model. More specifically, we argued hat the intercept a of the sample regression model serves as an estimator of the parameter β_0 in the population regression model, and that the slope b of the sample regression model serves as an estimator of the parameter β_1 in the population regression model. In this chapter, we now describe in greater detail what these estimators look like and what their properties are. We introduce three perspectives on estimation: ordinary least squares, methods of moments estimation, and maximum likelihood. We derive the standard errors and confidence intervals of the estimators, as well as sample regression function. Finally, we discuss the topic of hypothesis testing.

3.1 The Estimation Problem

The problem with the population regression model is that it contains several unknowns, to wit β_0, β_1, and σ^2. These are the so-called parameters of the model and they will have to be estimated. Estimation means that we produce an educated "guess" of the parameters based on the information in the sample. Statistical estimation theory is concerned with structured principles for finding

estimators that possess certain desirable qualities. As a matter of definition, the **estimator** is a rule to derive the value of the parameter from observed data in the sample. We speak of an **estimate** when we reference the specific value of the estimator in the sample.

To focus our thoughts let us designate the generic parameter as θ. The regression coefficients and variance are all instances of this generic parameter. We use the symbol $\hat{\theta}$ to reference the estimator of θ. We want this to be a function of the sample data only; there should be no remaining unknowns driving $\hat{\theta}$. The first question that we should ask is how we should use the data so that they can shed light on the parameter. Here, we can rely on various estimation principles such as least squares, method of moments estimation, and maximum likelihood, which often—but not always—produce the same estimators, i.e., the same formula. A second question we should ask is why we should trust the results these estimators produce. This gets into the properties of estimators.

In general, we would like our estimators to possess certain properties. These include unbiasedness, efficiency, consistency, and a known sampling distribution. We say that $\hat{\theta}$ is unbiased when, in expectation, it recovers θ: $E[\hat{\theta}] = \theta$. We say that $\hat{\theta}$ is efficient when it is unbiased and has the smallest possible variance of any unbiased estimator. We speak of consistency, if $\hat{\theta}$ converges in probability to θ as the sample size goes to infinity. This means that, in very large samples, the estimator almost certainly comes arbitrarily close to the true value of the parameter. Finally, if $\hat{\theta}$ converges to a known distribution it will be possible to derive confidence limits and to perform hypothesis tests. We now show three estimation procedures for the simple linear regression model, which can be shown to have these desirable properties when certain assumptions are met.

3.2 Least Squares Estimation

3.2.1 The General Principle

Least squares estimation is a general estimation principle that works extremely well for linear models. Although properties of this estimator have to be proved on an ad hoc basis, least squares estimators are widely employed in statistics,

including linear regression analysis. One of the reasons for this widespread use is that we have to make relatively few assumptions, especially when compared to maximum likelihood or even the method of moments.

Imagine, we are interested in estimating the kth moment about the origin: $\mu_k = E[Y^k]$. The least squares estimation principle states that we should pick an estimator such that

Equation 3.1: Least Squares Criterion

$$S = \sum_{i=1}^{n}(y_i^k - \mu_k)^2$$

is being minimized (Kmenta, 1997). This amounts to selecting the best estimator, in the sense of creating the smallest distances to the data values y^k. Since there is no weighting, we sometimes call this **ordinary least squares** (OLS) to contrast it with *weighted* least squares, a topic we shall encounter in Part III of this book.

A simple example can illustrate the principle. Imagine that we are interested in estimating the mean. This is $\mu_1 = E[Y]$, so that $k = 1$ and the least squares criterion may be formulated as

$$S = \sum_{i=1}^{n}(y_i - \mu_1)^2$$

To compute the minimum of S, we start by taking its first derivative:[1]

$$\frac{dS}{d\mu_1} = -2\sum_{i=1}^{n}(y_i - \mu_1)$$

This gives the slope of the tangent line of S. We now set the first derivative

[1] See Appendix A.

equal to zero because, at a minimum, the slope of the tangent is zero:

$$-2\sum_{i=1}^{n}(y_i - \mu_1) \quad = \quad 0$$

The last thing to do is to solve for μ_1. With simple algebra we can show that $\sum_i y_i = n\mu_1$, so that $\hat{\mu}_1 = \sum_i y_i/n = \bar{y}$. This is the least squares estimator. It is a proper estimator because the right-hand side is a function of the data only,[2] The estimator that we have derived here has desirable properties. It is unbiased, efficient, consistent, and asymptotically normally distributed (see Kmenta, 1997).

As a second example, consider the estimation of the variance σ^2. From mathematical statistics, we know that $\sigma^2 = E[Y^2] - (E[Y])^2 = \mu_2 - \mu_1^2$. We have already derived the least squares estimator of μ_1, to wit $\bar{y}$. We now need to obtain the least squares estimator of μ_2. As per Equation 3.1, this estimator can be found by setting $k = 2$ and by minimizing

$$S \quad = \quad \sum_{i=1}^{n}(y_i^2 - \mu_2)^2$$

The first derivative of S with respect to μ_2 is $-2\sum_i(y_i^2 - \mu_2)$. Setting this to zero and solving for μ_2 yields $\hat{\mu}_2 = \sum_i y_i^2/n$. Consequently,

$$\hat{\sigma}^2 \quad = \quad \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{\sum_i y_i^2}{n} - \bar{y}^2 = \frac{\sum_i (y_i - \bar{y})^2}{n}$$

This estimator has far less desirable properties. For example, it has a bias of $-\sigma^2/n$. This is what it means when we say that desirable properties of least squares cannot be generalized but have to be demonstrated on an ad hoc basis.

3.2.2 Application to Simple Regression Analysis

How are the ideas that we have developed so far relevant for the regression model? Remember that in the simple regression model the first raw moment

[2] It is a minimum because the 2nd derivative test comes out positive (see Appendix A).

varies across observations and is given by $\mu_{i1} = \beta_0 + \beta_1 x_i$. Thus, the OLS criterion may be written as:

Equation 3.2: OLS for the Simple Regression Model

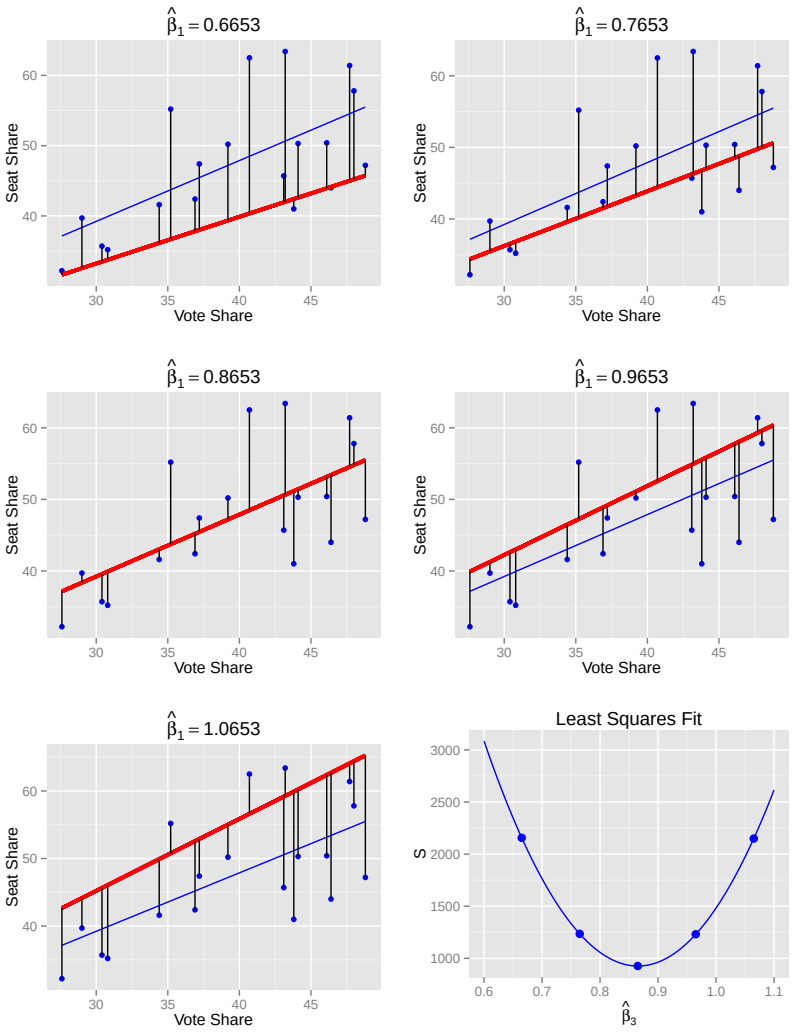
$$\begin{aligned} S &= \sum_{i=1}^{n}(y_i - \mu_{i1})^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\ \hat{\beta}_0, \hat{\beta}_1 &= \underset{\beta_0, \beta_1 \in \mathbb{R}}{\operatorname{argmin}} \; S \end{aligned}$$

Here, argmin stands for argument of the minimum, which means the set of values for β_0 and β_1 at which S is being minimized. Note that the values of these parameters are not constrained in any way: they can take on any value on the real number line ($\mathbb{R}$).

We can obtain analytic solutions for β_0 and β_1 and shall do so in a moment. Before doing so, however, it may be useful to illustrate how least squares estimation operates. Essentially, what we do is to select estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, which minimize $\hat{S} = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_i e_i^2 = SSE$. Figure 3.1 shows this process for the Labour seat data in Table 1.1. Here, we assume that we know $\hat{\beta}_0$ and we try different values for the slope. The blue line always panel represents the OLS regression line from Figure 1.2. The red line represents the regression line with the OLS estimator for the intercept and a trial value for the slope. The figure cycles through different values of the slope. When it selects the OLS estimate $\hat{\beta}_1 = 0.87$, then the blue and red regression lines coincide. When we now look at the last panel, we see that S, displayed on the vertical axis, reaches its minimum when $\hat{\beta}_1 = 0.87$. For all other values of the slope, we observe that S is higher. So picking the OLS estimator of the slope optimizes the least squares criterion.

Normally, we would not find the estimates by trial-and-error but use analytic methods. We would do what we did before, which is to take the derivative, set it to zero, and solve for the parameter. The only complication here is that we have two parameters, so we take the *partial* derivatives and set both of them

Figure 3.1: The Logic of Least Squares Estimation



Note: Five different choices of $\hat{\beta}_1$ and their implications for the OLS fit criterion S. The first five panels select different values of the slope estimate, whereas the last panel shows the OLS fit criterion. In the first five panels the observed data are depicted as blue dots. The OLS regression line is displayed in blue as well. The regression line implied by the choice of $\hat{\beta}_1$ is shown in red. The residuals are shown through the black lines. We see that these tend to be the smallest in the third panel, which selects the actual OLS estimate of the slope. The last panel shows how the OLS fit criterion is minimized when we select $\hat{\beta}_1 = 0.87$. The blue dots in this graph correspond to the values of $\hat{\beta}_1$ selected in the first five panels.

equal to zero.[3] It is easy to demonstrate that the partial derivatives are given by:

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\\ \frac{\partial S}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\, x_i\end{aligned}$$

When we set the first partial derivative to 0, then we get $\sum_i y_i - n\beta_0 - \beta_1 \sum_i x_i = 0$ or $n\bar{y} - n\beta_0 - \beta_1 n\bar{x} = 0$. Adding $n\beta_0$ to both sides, dividing by n, and substituting the estimator for β_1 we obtain

Equation 3.3: OLS Estimator of the Intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y}$ and $\bar{x}$ are the sample means of the dependent variable and the predictor, respectively. This is what we called a in the previous two chapters. Setting the second partial derivative to 0, we get $\sum_i x_i y_i - \beta_0 \sum_i x_i - \beta_1 \sum_i x_i^2 = 0$. Substituting the OLS estimator of the intercept, this becomes $\sum_i x_i y_i - \hat{\beta}_0 \sum_i x_i - \beta_1 \sum_i x_i^2 = \sum_i x_i y_i - (\bar{y} - \beta_1 \bar{x}) n\bar{x} - \beta_1 \sum_i x_i^2 = \sum_i x_i y_i - n\bar{x}\bar{y} - \beta_1(\sum_i x_i^2 - n\bar{x})^2 = 0$. Solving for β_1, we get

Equation 3.4: OLS Estimator of the Slope

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{s_{XY}}{s_X^2}$$

Here s_{XY} is the sample covariance between X and Y and s_X^2 is the sample variance of X. This is what we called b in the previous two chapters.

Let us illustrate the OLS estimators again using the Labour vote and seat share data from Table 1.1. For these data, $s_{XY} = 40.94$ and $x_X^2 = 47.31$. Hence, $\hat{\beta}_1 = 40.94/47.31 = 0.87$. This is the slope estimate that we used to

[3] See Appendix A for a discussion of partial derivatives.

draw the regression line in Figure 1.2. It is also easily shown that $\bar{y} = 47.54$ and $\bar{x} = 39.61$. Thus, applying the formula for $\hat{\beta}_0$, we obtain an OLS intercept estimate of $47.54 - 0.87 \cdot 39.61 = 13.27$. This is the intercept estimate that we used to draw the regression line in Figure 1.2.

The OLS slope estimator has several features. First, if $s_{XY} = 0$, then the slope is equal to 0. Thus, a lack of covariance or correlation between the dependent variable and the predictor causes the regression line to be flat. Second, the slope estimator is not defined if $s_X^2 = 0$, as we would be dividing by zero. We know that $x_X^2 = 0$ when X is constant in the sample. The second property thus means that we cannot explain a variable (Y) with a constant. The intercept estimator also has an important feature. Rearranging terms, we see that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. This means that the regression line goes through $(\bar{x}, \bar{y})$.

3.3 Method of Moments Estimation

A second approach to estimating the linear regression model is **method of moments** or MM. The advantage of this approach is that it guarantees consistent estimators and requires nothing more than a reliance on the law of large numbers. The approach is used widely in statistics, in particular in the context of models for heteroskedastic, time series, and panel data (see Part III). As such, it is useful to introduce its logic in this chapter.

3.3.1 General Principle

The object of method of moment estimators is to minimize the difference between population and sample moments. The vehicle for accomplishing this goal is the so-called moment condition, which may be written as

Equation 3.5: Moment Condition

$$m(\theta) \quad = \quad E[f(y_i, \theta)] = 0$$

Here θ is a generic parameter and $f(.)$ is some function, which in expectation

is equal to 0. We now formulate the sample moment equivalent of the moment condition, which is known as the sample moment condition:

Equation 3.6: Sample Moment Condition

$$\bar{m}(\theta) = \frac{1}{n}\sum_i f(y_i, \theta) = 0$$

The law of large numbers states that, for large values of n, $\bar{m}(\theta) = m(\theta)$. Thus, the sample moment condition can serve as a basis for selecting a consistent estimator of θ. We do this by selecting an estimator such that Equation 3.6 is satisfied.

In general, it is possible that we have multiple moment conditions. When the number of moment conditions is exactly equal to the number of parameters, then we are dealing with classical method of moments estimation. When the number of moment conditions exceeds the number of parameters, then we are in the domain of *generalized method of moments* (GMM) estimation.

As an example of method of moments estimation, let us consider the problem of estimating the population mean for some distribution. We formulate the following moment condition:

$$E[y] - \mu = 0$$

This is a trivial restatement of the definition of the population mean: $\mu = E[y]$. The sample moment condition is

$$\frac{1}{n}\sum_i y_i - \mu = 0$$

Adding μ to both sides of the equation yields $\hat{\mu} = \sum_i y_i/n = \bar{y}$. Thus, the sample mean is the method of moments estimator of μ. This is an example of classical methods of moments estimation because the number of moment conditions is identical to the number of parameters, in this case one.

3.3.2 Application to Simple Regression Analysis

Moment conditions often follow automatically from the model that we formulate. This is true, too, of the linear regression model. This model entails the following important moment conditions:

1. $E[\varepsilon_i] = 0$
2. $E[\varepsilon_i x_i] = 0$

These conditions all involve moments (specifically, means, variances, and covariances) and derive directly from the model assumptions (Assumptions 2.3 and 2.6).

Using the definition of the error term, the moment conditions may be written as:

$$\begin{aligned} E[\varepsilon_i] &= E[y_i - \beta_0 - \beta_1 x_i] = 0 \\ E[\varepsilon_i x_i] &= E[(y_i - \beta_0 - \beta_1 x_i) x_i] = 0 \end{aligned}$$

The sample moment conditions are

$$\begin{aligned} \frac{1}{n}\sum_i (y_i - \beta_0 - \beta_1 x_i) &= 0 \\ \frac{1}{n}\sum_i (y_i - \beta_0 - \beta_1 x_i) x_i &= 0 \end{aligned}$$

The first condition produces the estimator of the intercept. It may be written as $\sum_i y_i/n - \beta_0 - \beta_1 \sum_i x_i/n = 0$, which is equal to $\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$. Substituting the estimator for β_1 and rearranging terms, we obtain

Equation 3.7: MM Estimator of the Intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

It is easily verified that this estimator is identical to the OLS estimator. The estimator for β_1 follows from the second sample moment condition. This may

be written as $\sum_i x_i y_i/n - \beta_0 \sum_i x_i/n - \beta_1 \sum_i x_i^2/n = 0$ or $\sum_i x_i y_i/n - \beta_0 \bar{x} - \beta_1 \sum_i x_i^2/n = 0$. Substituting the expression for β_0 from the first sample moment condition, we get $\sum_i x_i y_i/n - (\bar{y} - \beta_1 \bar{x})\bar{x} - \beta_1 \sum_1 x_i^2/n = 0$, which is equal to $\sum_i x_i y_i/n - \bar{x}\bar{y} - \beta_1(\sum_i x_i^2/n - \bar{x}^2) = 0$. The first term on the left-hand side is equal to s_{XY}, whereas the second term is equal to $\beta_1 s_X^2$, so that we can also write the second sample moment condition as $s_{XY} - \beta_1 s_X^2 = 0$. Rearranging terms and solving for β_1 then yields:

Equation 3.8: MM Estimator of the Slope

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

Once more, this is identical to the OLS estimator that we derived earlier.

3.4 Maximum Likelihood Estimation

As a class, (ordinary) least squares estimators are not automatically associated with desirable properties. Method of moment estimators are consistent but do not automatically possess other desirable qualities such as efficiency. Are there estimators, which we know to possess a broad class of desirable properties? The answer is affirmative and one such estimator is the **maximum likelihood** estimator (MLE). Under certain so-called regularity conditions, MLEs are known to be consistent, asymptotically efficient, and asymptotically normally distributed. Note that these are all asymptotic properties; in finite samples, the properties of MLEs may not be so desirable. These asymptotic results come at a price, however, for MLEs require that we exploit Assumption 2.2: the errors are normally distributed. In OLS, we did not have to invoke this assumption, but in MLE we cannot avoid it. In addition, we assume that the sample observations are statistically independent. In sum, we are assuming that the errors and, by extension, the values of the dependent variable are *normally and independently distributed* (n.i.d).

3.4.1 The General Principle

The intuition behind maximum likelihood estimation is actually quite simple. Imagine, we knew the entire distribution of a random variable, including the parameter values. Then it could be easily seen that sampling from the distribution is differentially likely to produce certain samples. A simple example can show this. Assume that Y follows a binomial distribution with parameter $\pi = .5$. An application would be the flipping of a fair coin. Say that we sample $n = 2$ observations, e.g., we flip the coin twice. What is more likely, that we observe one instance of heads or two instances of heads? Evaluating the binomial probability mass function for $y = 1$ yields a probability of .5. On the other hand, evaluating it for $y = 2$ yields a probability that is only half that size. Thus, we would say that obtaining one instance of heads is more likely than obtaining two instances.

In this example, we assumed that we know the parameters and want to derive implications for the data that we are likely to observe. In the estimation problem, this is reversed: we know the data and seek to draw inferences about the parameters. But we can reverse the logic and ask which parameters are most likely to have given rise to the data at hand. This is the idea of maximizing the likelihood. For example, in the case of the binomial distribution, a value of .5 of the parameter π is most likely to have given rise to one instance of heads in two trials.

One of the main contributors to maximum likelihood, Sir Ronald Fisher, explained the idea in the following way:

> The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed. (Fisher, 1922, 310).

The parameter value(s) that maximize the likelihood are the ones that we select as our estimates.

From Fisher, it follows that the first thing that we have to do is to characterize the likelihood of the data. Our data consist of observations $y_1, y_2, \cdots, y_n$

that are the realized values of the random variable Y. The likelihood is defined as the joint distribution over these observations, which can be computed only if the probability density or mass function of Y is known. If we have this information, then the likelihood is given by

$$\mathcal{L}(y_1, \cdots, y_n|\theta) \equiv f(y_1, \cdots, y_n|\theta)$$

where f is the probability mass or density function. We have assumed here that the same mass or density applies to all observations, i.e., they are identically distributed. If we can also assume that the observations are independent, as would be the case under simple random sampling, then we can simplify the likelihood. From probability theory, we know that the joint distribution is equal to the product of the marginal distributions if (and only if) the random variables are statistically independent. Under this assumption,

Equation 3.9: Likelihood Function

$$\mathcal{L}(y_1, \cdots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Here, $\prod$ is the product operator, which means that the terms following it are being multiplied.

The maximum likelihood estimator of θ is the value that maximizes the likelihood of the data:

Equation 3.10: Maximum Likelihood Estimator

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(y_1, \cdots, y_n|\theta)$$

Here Θ is the so-called parameter space, i.e., the set of feasible values of the parameter. For simple problems, finding the MLE is a simple process of taking the first derivative, setting it to zero, and solving for θ.

In practice, it is usually not the likelihood that is being maximized but the log of the likelihood. We do this for reasons of convenience, since taking the logarithm will turn the product operator into a sum operator, which is easier to process. Consequently, we can write

Equation 3.11: Maximizing the Log-Likelihood

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \ell \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \ln\mathcal{L} \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \sum_{i=1}^{n}\ln f(y_i|\theta)\end{aligned}$$

where $\ell = \ln\mathcal{L}$ is the so-called *log-likelihood function*.

As an example, let us consider the normal distribution. Imagine that our data consist of n independent draws from the normal distribution $\mathcal{N}(\mu,\sigma)$. We know that the normal probability density function is

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

We now take the natural logarithm of this function in order to produce the contribution to the log-likelihood of a single sample unit:

$$\ell_i = \ln f(y_i|\mu,\sigma) = -\ln\sigma - .5\ln(2\pi) - .5\frac{(y_i-\mu)^2}{\sigma^2}$$

Summing the individual contributions to the log-likelihood, we become

$$\ell = \sum_{i=1}^{n}\ell_i = -n\ln\sigma - .5n\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

The MLEs are now those values of μ and σ for which ℓ reaches its maximum. To find them, we start by taking the partial derivatives of ℓ with respect to μ

and σ, respectively:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{\sigma^3}\end{aligned}$$

We now set the first partial derivative equal to zero and solve for μ. It is easily shown that this produces

$$\hat{\mu} = \bar{y}$$

Setting the second partial derivative equal to zero and solving for σ yields

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

Evaluation of the second derivatives demonstrates that these two estimators indeed maximize ℓ and are thus proper MLEs.

Comparison of the MLEs with the OLS estimators that we derived earlier reveals that they are identical in the case of the normal distribution. This happens frequently in statistical estimation. Still, there is a benefit to showing this equivalence. MLEs are consistent, asymptotically efficient, and normally distributed. Having shown the equivalence of the OLS estimators to the MLEs, this means we can now assume, for example, that $\bar{y}$ is a consistent estimator.

3.4.2 Application to Simple Regression Analysis

Let us apply maximum likelihood to the simple regression model. According to Equation 2.4, $y_i \sim \mathcal{N}(\mu_i, \sigma)$, so that

$$\ell_i = -\ln\sigma - .5\ln(2\pi) - \frac{1}{2\sigma^2}(y_i - \mu_i)^2$$

From equation 2.5, we know that $\mu_i = \beta_0 + \beta_1 x_i$, so that we may also write

$$\ell_i = -\ln\sigma - .5\ln(2\pi) - \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2$$

Aggregating over all sample units produces the following estimation criterion:

Equation 3.12: ML Estimation of the Simple Regression Model

$$\begin{aligned} \ell &= -n \ln \sigma - .5n \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\ \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma} &= \underset{\beta_0, \beta_1 \in \mathbb{R}, \sigma > 0}{\operatorname{argmax}} \ \ell \end{aligned}$$

We proceed in the usual manner by taking the first partial derivatives of the log-likelihood with respect to the parameters:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

We see that, apart from a multiplier, the partial derivative of ℓ with respect to β_0 is identical to the partial derivative of S with respect to β_0, which we analyzed earlier. It is no surprise then that the MLE of β_0 is identical to the OLS estimator of that parameter:

Equation 3.13: MLE of the Intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Similarly, the partial derivative of ℓ with respect to β_1 is, short of a multiplier, identical to the partial derivative of S with respect to β_1. Here, too, then the MLE is identical to the OLS estimator:

Equation 3.14: MLE of the Slope

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{s_{XY}}{x_X^2}$$

Turning to the remaining partial derivative, we can derive the MLE for σ by setting this derivative equal to 0. Doing this produces $n\sigma^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. Substituting the MLEs for the regression coefficients, this becomes $n\sigma^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. From the definition of the SSE, we know that the right-hand side is equal to the sum of the squared residuals. If we now further divide both sides by n and take the square root, we obtain the following estimator:

Equation 3.15: MLE of the Standard Deviation

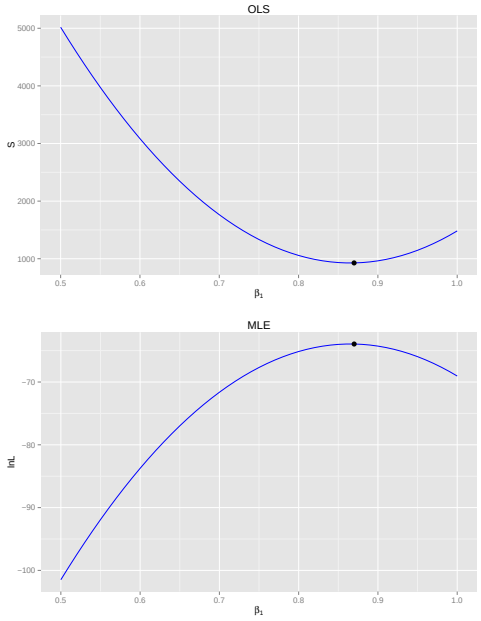
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} = \sqrt{\frac{SSE}{n}}$$

That OLS and MLE produce identical results can be easily shown for the data from Table 1.1. Imagine that we knew $\hat{\beta}_0$ and $\hat{\sigma}$, then the only remaining parameter to estimate is β_1. Let us vary the values of this parameter between 0.5 and 1.0. Figure 3.2 shows the OLS and ML fit functions, which clearly reach their minimum and maximum, respectively, at $\hat{\beta}_1 = 0.87$.

3.5 Properties of the Estimators

Now that we have derived the estimators of the regression parameters, the next question we should ask is why we should trust them. What are their properties? We now consider this question.

Figure 3.2: A Comparison of OLS and ML



Note: Based on the data from Table 1.1.

3.5.1 Regression Coefficients

Finite Sample Properties

Finite sample properties refer to those properties of estimators that hold when the sample size is assumed to be finite. This includes situations in which the sample size is small, e.g., less than 25. We contrast finite sample properties with asymptotic properties, which hold when the sample size approaches infinity.

In regression analysis, the finite sample properties of the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are covered by the **Gauss-Markov theorem**:

Theorem 3.1: Gauss-Markov Theorem

Assuming that Assumptions 2.3-2.6 hold true, the OLS estimators of β_0 and β_1 are BLUE: best linear unbiased estimators.

Since the MLEs of the regression coefficients are identical to the OLS estimators, the theorem of course also applies to them.

We shall postpone the proof of the theorem until Chapter 5. For now it is sufficient to understand both the importance and the limitations of the theorem. The theorem should be read in terms of consecutive sets. Specifically, within the set of conceivable estimators of the regression coefficients, there resides a subset of so-called *linear* estimators. They are called this because they are linear functions of the data for the dependent variable. Within the subset of linear estimators, there is a subset that is *unbiased* so that $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$. Finally, within the subset of linear unbiased estimators, there are estimators that are *best* in the sense of having the smallest variance. From the Gauss-Markov theorem, we thus know that the OLS/ML estimators of the regression coefficients are unbiased and efficient. These are desirable properties, indeed. On the average, our estimator gets it right and if it gets it wrong, then the errors are smaller than those obtained from other linear unbiased estimators.

Nice as all of this may sound, you should keep in mind that the Gauss-Markov theorem is contingent on a number of assumptions. First, we assume that the predictor and the errors are uncorrelated. We already discussed that

this is a hefty assumption and if it fails, the OLS/ML estimators are no longer unbiased. Second, we assume that the errors are zero in expectation; if this assumption fails, then the OLS/ML estimator of β_0 is no longer unbiased. Third, we assume homoskedasticity and the absence of autocorrelation. If these two assumptions fail, then the OLS/ML estimators are still unbiased but no longer "best." Specifically, it will be possible to find other unbiased estimators with an even smaller variance.

We should take notice of another limitation of the Gauss-Markov theorem. The OLS/ML estimators may be linear unbiased estimators with the smallest variance. This does not mean, however, that there are not other estimators that are biased but have a smaller variance than OLS/MLE. We shall take advantage of this in Chapter 10.

Asymptotic Properties

The asymptotic properties follow from the fact that the estimators of the regression coefficients are MLEs (and MMEs). Assuming that the model is correct—i.e., the various regression assumptions hold—we thus know that the OLS estimators are consistent: $\operatorname{plim}_{n\to\infty} \hat{\beta}_{0n} = \beta_0$ and $\operatorname{plim}_{n\to\infty} \hat{\beta}_{1n} = \beta_1$. In addition, we know that the asymptotic sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal. Later in this chapter, we shall use this property for purposes of hypothesis testing.

3.5.2 Error Variance

Finite Sample Properties

From Equation 3.12, we know that the MLE of σ^2 is given by $\hat{\sigma}^2 = n^{-1}\sum_i e_i^2$. However, this estimator is biased. Specifically,

$$E[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2 \neq \sigma^2$$

This produces a bias of $-2\sigma^2/n$, which means that the error variance tends to be underestimated. We shall leave the proof of this result for Chapter 5, but the intuition for the bias is quite simple. The formula for $\hat{\sigma}^2$ assumes that we

know the values of $\hat{\beta}_0$ and $\hat{\beta}_1$. In reality, these values are being estimated, which introduces sampling variation. The uncertainty associated with the estimation of the two regression coefficients is not taken into consideration in the formula for $\hat{\sigma}^2$, thus resulting in an underestimate of the total uncertainty in the regression.

An unbiased estimator can be derived quite easily. All we have to do is to multiply $\hat{\sigma}^2$ by a factor of $n/(n-2)$. This yields

Equation 3.16: Unbiased Estimator of the Error Variance

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2},$$

which is known as the **mean squares due to error** (MSE). Although this estimator is unbiased, it has a greater sampling variance than $\hat{\sigma}^2$.

We can illustrate the computation of the MSE using the data from Table 1.2. In Chapter 1, we saw that $SSE = 926.35$ (see Table 1.3). The sample size is $n = 19$. Consequently, $s^2 = 926.35/(19-2) = 54.49$ and $s = \sqrt{56.79} = 7.38$.

Asymptotic Properties

Both $\hat{\sigma}^2$ and s^2 are consistent estimators. As long as the true variance around the regression line is not too small, then these estimators are asymptotically approximately normally distributed. However, we typically do not rely on this asymptotic distribution for purposes of constructing confidence intervals (see Section 3.6).

3.6 Standard Errors

3.6.1 Regression Coefficients

The coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates and hence, they display sampling fluctuation. Thus, we would want to know their variance. Postponing the proof

until Chapter 5, it is possible to demonstrate that

$$\begin{aligned} V[\hat{\beta}_0] &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right) \\ V[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \end{aligned}$$

provided that the errors are homoskedastic and display no autocorrelation. Of course, these results depend on the variance around the population regression line, which is unknown. Substituting the unbiased estimator s^2 for σ^2, the *estimated* variances of the regression coefficients are:

Equation 3.17: Estimated Variances of the Regression Coefficients

$$\begin{aligned} \hat{V}[\hat{\beta}_0] &= s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right) \\ \hat{V}[\hat{\beta}_1] &= \frac{s^2}{\sum_i (x_i - \bar{x})^2} \end{aligned}$$

Notice the hat on top of the V, which indicates that we are dealing with an estimated variance. The standard errors are equal to the square roots of the estimated variances:

$$\begin{aligned} \widehat{SE}[\hat{\beta}_0] &= \sqrt{\hat{V}[\hat{\beta}_0]} \\ \widehat{SE}[\hat{\beta}_1] &= \sqrt{\hat{V}[\hat{\beta}_1]} \end{aligned}$$

We can apply these formulas to the Labour vote and seat share data from Table 1.1. Earlier, we saw that $n = 19$, $\bar{x} = 39.61$, $s_X^2 = 47.31$, and $s^2 = 54.49$. From this, it follows that $\sum_i (x_i - \bar{x})^2 = (n-1)s_X^2 = 851.66$. We now can compute the estimated variances of the intercept and slope estimators as

$$\begin{aligned} \hat{V}[\hat{\beta}_0] &= 54.49 \left(\frac{1}{19} + \frac{39.61^2}{851.66} \right) = 103.26 \\ \hat{V}[\hat{\beta}_1] &= \frac{54.49}{851.66} = 0.06 \end{aligned}$$

Consequently, $\widehat{SE}[\hat{\beta}_0] = \sqrt{103.26} = 10.16$ and $\widehat{SE}[\hat{\beta}_1] = \sqrt{0.06} = 0.25$.

One aspect that we have not yet considered is that the estimators of the intercept and slope are not independent. It can be demonstrated that

Equation 3.18: Covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\widehat{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{s^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = -\bar{x}\hat{V}[\hat{\beta}_1]$$

For the data from Table 1.1, this means that the estimated covariance between the intercept and slope estimators is $-39.61 \cdot 0.06 = -2.53$. We shall need this covariance to compute the confidence interval around the regression line.

3.6.2 Error Variance

For the error variance, it can be demonstrated (e.g., Kmenta, 1997) that

Equation 3.19: Variance of the MSE

$$V[s^2] = \frac{2\sigma^4}{n-2}$$

This result again depends on the assumptions that the errors are homoskedastic and uncorrelated.

3.6.3 Predicted Values

Finally, consider the standard error of the predicted values. As is shown in Appendix C.1, the variance of $\hat{y}_i$ is given by

$$V[\hat{y}_i] = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{2}(x_i - \bar{x})^2}\right)$$

Substituting the unbiased estimator of the variance we get

Equation 3.20: Estimated Variance of the Predicted Values

$$\hat{V}[\hat{y}_i] = s^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{2} (x_i - \bar{x})^2} \right)$$

The square root of Equation 3.16 gives the standard error of the predicted values. Clearly, this is smallest when we select $x_i = \bar{x}$.

3.7 Confidence Intervals

3.7.1 Regression Coefficients

The formulas that we derived in Equations 3.3-3.4, 3.7-3.8, and 3.13-3.14 produce point estimates of the regression coefficients. These estimates are based solely on the information in our sample and do not at all consider the fact that, due to sampling fluctuation, other estimates would have been obtained in different samples. To take the sampling fluctuation into account, it is useful to construct confidence intervals, i.e., to produce interval estimates.

Let the chosen level of the confidence interval be $1 - \alpha$. Then it can be demonstrated (see Chapter 5) that the confidence intervals for the intercept and slope are

Equation 3.21: Confidence Intervals Regression Coefficients

$$\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}} \widehat{SE}[\hat{\beta}_0] \leq \beta_0 \leq \hat{\beta}_0 + t_{n-2,\frac{\alpha}{2}} \widehat{SE}[\hat{\beta}_0]$$
$$\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}} \widehat{SE}[\hat{\beta}_1] \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \widehat{SE}[\hat{\beta}_1]$$

Here $t_{n-2,\frac{\alpha}{2}}$ is a value of a t-distribution with $n-2$ degrees of freedom such that a probability of $\frac{\alpha}{2}$ remains in each tail.

For the Labour vote and seat share data from Table 1.1, we need to reference the t_{19-2}-distribution. Imagine that we want to obtain the 95% confidence interval. Then $t_{n-2,\frac{\alpha}{2}} = 2.11$. Using the estimates and their standard errors

that we derived earlier, we obtain the following confidence intervals:

$$13.27 - 2.11 \cdot 10.17 \leq \beta_0 \leq 13.27 + 2.11 \cdot 10.17$$
$$0.87 - 2.11 \cdot 0.25 \leq \beta_1 \leq 0.87 + 2.11 \cdot 0.25,$$

or $-8.17 \leq \beta_0 \leq 34.71$ and $0.33 \leq \beta_1 \leq 1.40$. The correct interpretation of these confidence intervals is that, if they would be computed time and again in different samples of the same size, 95 percent of the confidence intervals would include the true intercept and slope, respectively.

3.7.2 Error Variance

It is relatively rare that we compute a confidence interval for the error variance but, for the sake of completeness, we discuss the formula here. The starting point is the well-known result from mathematical statistics that

$$\frac{(n-2)s^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$$

(e.g., Kmenta, 1997). If the chosen confidence interval is $1 - \alpha$, then

$$\Pr\left[\chi^2_{n-2,1-\frac{\alpha}{2}} \leq \frac{SSE}{\sigma^2} \leq \chi^2_{n-2,\frac{\alpha}{2}}\right] = 1 - \alpha$$

Rearranging terms now yields

Equation 3.22: Confidence Interval Regression Variance

$$\frac{SSE}{\chi^2_{n-2,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{SSE}{\chi^2_{n-2,\frac{\alpha}{2}}}$$

For the data in Table 1.1, we can compute a 95% confidence interval by setting $\alpha = .05$. It can be easily demonstrated that $\chi^2_{n-2,1-\frac{\alpha}{2}} = \chi^2_{17,.975} = 30.19$ and $\chi^2_{n-2,\frac{\alpha}{2}} = \chi_{17,.025} = 7.56$. We already saw that $SSE = 54.49$. It is then easily shown that $1.80 \leq \sigma^2 \leq 7.20$.

3.7.3 Conditional Expectation Function

When depicting the regression line, it is useful to depict its confidence interval as well. It can be demonstrated (see Chapter 5) that

$$\frac{\hat{y}_i - (\beta_0 + \beta_1 x_i)}{\widehat{SE}[\hat{y}_i]} \sim t_{n-2}$$

From this, we can immediately derive a confidence interval for $\beta_0 + \beta_1 x_i = E[y_i]$:

Equation 3.22: Confidence Interval Around the Regression Line

$$\hat{y}_i - t_{n-2,\frac{\alpha}{2}} \widehat{SE}[\hat{y}_i] \leq E[y_i] \leq \hat{y}_i + t_{n-2,\frac{\alpha}{2}} \widehat{SE}[\hat{y}_i]$$

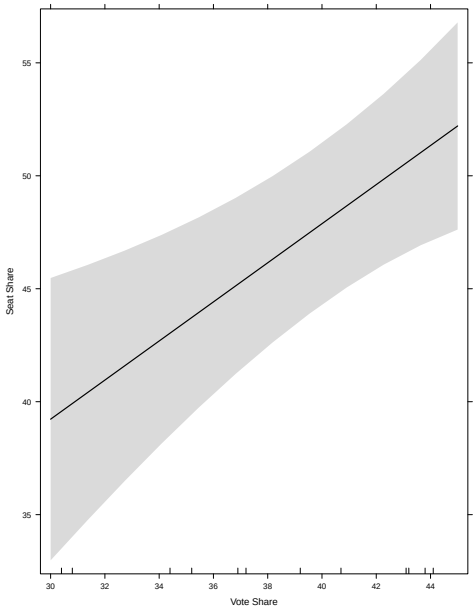
Figure 3.3 illustrates the confidence band for the regression of Labour seat shares on vote shares. The confidence band is at its smallest when we set the Labour vote share to its mean value of 39.6 percent, since this is where the standard error of $\hat{y}_i$ is minimized. The further we move away from the mean vote share in either direction, the wider the confidence interval becomes. This means that we are less confident in our predictions when the Labour vote share is assumed to be atypically small or large.

3.8 Testing Simple Hypotheses

Until now we have taken an in-depth look at the estimation of the simple regression model. There is another aspect of statistical inference, however, and this is hypothesis testing. In simple regression analysis, the most common statistical tests pertain to the slope coefficient. Under some conditions, it may also be interesting to test hypotheses about the intercept, but this is less common.[4] For the remainder of this section, we consider only tests of hypotheses about the slope. The procedure for testing hypotheses about the intercept is the same, however.

[4] In Chapter 8, we shall encounter one context in which hypothesis tests of intercepts are useful.

Figure 3.3: Labour Seat Share Regression with Confidence Interval



Note: Based on the data from Table 1.1. The black line represents the regression line, on which the predicted values lay. The gray area around the regression line represents the 95% confidence interval.

If our interest is in testing a hypothesis about a single parameter such as the slope coefficient, then we are testing a simple hypothesis. Typically, we formulate the null hypothesis in terms of a particular parameter value. The alternative hypothesis then encompasses all other values that the parameter can take on. In the case of the slope coefficient, we can thus formulate

$$H_0 : \beta_1 = q$$
$$H_A : \beta_1 \neq q$$

Here q is the hypothesized value of the slope coefficient in the population. Very frequently, $q = 0$ so that the null hypothesis implies that the predictor has no effect on the dependent variable in the population. A test of this hypothesis is known as a *significance test*. Significance tests are the default tests performed in R.

How would we go about testing H_0? We should consider two different pieces of information. First, we have an estimate of the slope and can compare this to the hypothesized value. If $\hat{\beta}_1 - q$ is large, regardless of the direction, then this is a priori evidence against the null hypothesis. Second, we need to take into consideration sampling fluctuation. What may look like an abnormally large discrepancy between the estimate and the hypothesized value may, in fact, be a fairly common occurrence when the null hypothesis is true. In that case, we should probably not reject the null hypothesis. However, if the discrepancy is indeed extreme, meaning that it occurs infrequently when the null hypothesis is true, then we may take this as evidence against the null.

The various considerations imply that we compute our test statistic as

Equation 3.23: Test Statistic for the Slope

Let $H_0 : \beta_1 = q$. Then

$$t = \frac{\hat{\beta}_1 - q}{\widehat{SE}[\hat{\beta}_1]}$$

To compute the p-value that is associated with the test statistic, we recognize that it follows a t_{n-2}-distribution. Thus, the p-value is equal to the probability

of obtaining a test statistic as large as $|t|$ or even larger in the t-distribution when the null hypothesis is true.

To illustrate the test procedure, let us revisit the Labour vote and sear share data from Table 1.1. We first test $H_0 : \beta_1 = 0$, which amounts to hypothesizing that there is no relationship between the seat share that Labour receives and its electoral performance measured in vote share. Imagine that we set our Type-I error rate to .10; for a small sample like we have, this is a reasonable decision. We have seen that $\hat{\beta}_1 = 0.87$ and $\widehat{SE}[\hat{\beta}_1] = 0.25$. Hence, $t = 0.87/0.25 = 3.42$. When we now compute $\Pr > |t|$ using a t_{19-2}-distribution, we obtain a p-value of .00, which is far smaller than the Type-I error rate and leads us to reject H_0. We have statistically reliable evidence that there is a relationship between the Labour vote share and its seat share in the House of Commons.

Now consider a different null hypothesis: $H_0 : \beta_1 = 1$. This hypothesis implies that a vote increase of one percent translates into a seat increase of one percent. The test statistic is now computed as $t = (0.87 - 1.00)/0.25 = -0.53$. The p-value is now $\Pr > |-0.53|$; using again the t_{17}-distribution, this yields 0.60. Since this exceeds the Type-I error rate, we fail to reject the null hypothesis.

The procedure outlined here concern so-called two-sided tests. However, they can be modified easily to accommodate one-side tests. For example, a reasonable set of hypotheses for the Labour seat and vote share data is

$$\begin{aligned} H_0 : &\quad \beta_1 \leq 0 \\ H_A : &\quad \beta_1 > 0 \end{aligned}$$

The alternative hypothesis reflects the belief that there is a positive linear relationship between Labour vote and seat shares. The null hypothesis states that there is either no relationship or a negative relationship. We would now compute the test statistic as before. However, the p-value is now defined as $\Pr > t$, since only positive test statistics are evidence against the null hypothesis. We can now take the earlier p-value from the two-sided test and cut it in half. This produces $p = .00$, so that H_0 is rejected.

3.9 Statistical Inference Using R

So far, we have done a lot of computations by hand. There is no reason to do this, however, as computers are much faster and adept at performing these computations. Regression analysis is so ubiquitous that it is a standard part of all statistical software programs. R alone contains several routines for performing regression analysis. Here, we shall focus on the `lm` command, which stands for linear models.[5] The basic syntax for simple regression analysis is

```
object <- lm(y ~ x, data = df)
summary(object)
```

Here, `object` is the name of the object that is to store the regression results, `y` is the name of the dependent variable, `x` is the name of the predictor, and `df` is the name of the data frame that contains `y` and `x`. If we run only the first line of the program, then `object` is created but no results are shown on the computer display. The second line takes care of this; after it has been run, a summary of the regression results (see Figure 3.4) is obtained.

The `lm` command can be adjusted in a number of different ways. For example, `lm(y ~ 0 + x, data = df)` performs a regression through the origin. And `lm(y ~ 1, data = df)` fits an intercept only.

Let us apply the command to the Labour vote and seat share data from Table 1.1. Assume that we created a data frame called `labour`, which contains `seat` as the dependent variable and `vote` as the predictor. We execute the following command: `seat.fit <- lm(seat ~ vote, data = labour)`. The object `seat.fit` now contains the regression results and is summarized in Figure 3.4. We observe that the OLS/ML estimator of the intercept is 13.2670, whereas the estimator for the slope coefficient associated with `vote` is 0.8653 (see the box marked (B)). The estimated standard errors for the intercept and slope are 10.1615 and 0.2529, respectively (see the box marked (C)). The test statistic for $H_0 : \beta_0 = 0$ is 1.306, whereas the test statistic for $H_0 : \beta_1 = 0$ is 3.421

[5] Other commands include `glm` and `ols`. The `lm` package, however, is used more commonly for regression analysis and integrates well with a number of other packages that will prove useful in subsequent chapters.

Figure 3.4: R Regression Output for the Labour Seat Share Data

```
Residuals:                                                    (A)
    Min      1Q  Median      3Q     Max
-10.167  -4.790  -1.433   3.006  14.015

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.2670    10.1615   1.306  0.20908
vote          0.8653     0.2529   3.421  0.00326 **
---             (B)        (C)     (D)     (E)   (F)
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.382 on 17 degrees of freedom      (G)
Multiple R-squared:  0.4077,    (H)  Adjusted R-squared:  0.3729
F-statistic:  11.7 on 1 and 17 DF,  p-value: 0.003257
```

Note: Output from the `lm` command based on the data from Table 1.1. (A) = summary of the residuals; (B) = OLS/ML estimates of the regression coefficients; (C) = estimated standard errors of the regression coefficients; (D) = t-statistics for the null hypothesis that a regression coefficient is 0 in the population; (E) p-values for the t-statistics; (F) = a graphical indication of the level of significance; (G) = $\sqrt{MSE} = \sqrt{s^2} = s$ plus an indication of the degrees of freedom; and (H) = the coefficient of determination.

(see the box marked (D)). For a two-sided test, the p-value of the test statistic for the intercept is 0.20908; for the test statistic associated with the slope, the p-value is 0.00326 (see the box marked (E)). This means, that the intercept is not statistically significant. The slope is statistically significant at the .01 level (see the box marked (F)). This means that we would be able to reject the null hypothesis with a Type-I error rate as low as 1 in 100. The spread around the regression line is given by the "residual standard error," which is the square root of the MSE. In our case, this is 7.382. The degrees of freedom are 17 (see the box marked (G)). The coefficient of determination or "multiple R-squared" is 0.4077 (see the box marked (H)). Within rounding error, all of these estimates are identical to what we computed by hand in this and earlier chapters. Of final relevance in the output is the information about the residuals (the box marked (A)). For example, the fact that the median of the residuals is not zero calls into question the normality assumption: in a normal distribution, the mean and median are identical. While the output contains other information, this will become relevant only after our discussion of multiple regression analysis.

Once we have run the `lm` command, other information can be obtained quite easily. For example, you will notice that R does not report the confidence intervals for the regression coefficients. However, this can be rectified quite easily by issuing the following command:

```
confint(object, level = #)
```

Here, `level` is set to the desired confidence level. For our data, the 95% confidence interval for the intercept is $[-8.17, 34.71]$, whereas it is $[0.33, 1.40]$ for the slope.

It is also quite easy to obtain fitted values ($\hat{y}_i$) and residuals (e_i) for the regression model, as can be seen below:

```
seat.res <- residuals(seat.fit)
seat.pred <- fitted.values(seat.fit)
```

The object `seat.res` now contains the residuals and `seat.pred` contains the predicted values.

Figure 3.3 was generated using the package `effects`, which is an extremely useful add-on to `lm` (Fox, 2003). This program is not a part of the standard R installation and, thus, has to be downloaded first. Further, it will have to be loaded into memory in order to use it. For Figure 3.3, the following syntax was used:

```
library(effects)
plot(effect("vote", seat.fit), main = "",
    xlab = "Vote Share", ylab = "Seat Share")
```

3.10 Conclusion

In this chapter, we explored statistical inference of the simple regression model. We derived estimators for the various parameters of the model, to wit β_0, β_1, and σ. We also demonstrated that, under certain conditions, the estimators

of the intercept and slope coefficients are unbiased, efficient, consistent, and asymptotically normally distributed. We discussed both point and interval estimation and we demonstrated test procedures for simple hypotheses. Finally, we showed how all of this can be implemented in R.

We have done about as much as we can with the simple regression model. The model is inherently limited in that it allows for a single predictor only. The next logical step is to expand the linear regression model so that it can include multiple predictors. This brings us to the topic of *multiple* regression analysis, which is taken up in the next part of this book.

Part II

Multiple Regression Analysis

Chapter 4

The Multiple Regression Model

The multiple regression model is a linear regression model that allows for multiple predictors.[1] When we argued earlier that the regression model remains the work horse of quantitative social science, we really had the multiple regression model in mind. The fact that multiple predictors are allowed greatly expands the versatility of regression analysis. In this part of the book, we shall see several examples of this, for example, when we create polynomial regression models, include discrete predictors, or model interaction effects.

4.1 The Population Regression Model

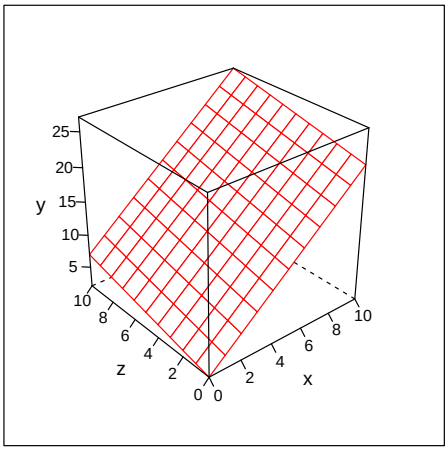
4.1.1 Scalar Notation

The simplest extension of the simple regression model is to add a second predictor. For example, if we believe that both X and Z influence Y, then we could specify the following population regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

[1] The model is sometimes referenced as the multivariate regression model, but this is a mistake. The term multivariate regression has a specific meaning in statistics: it pertains to regression models with multiple dependent variables. Such models will not be considered in this book.

Figure 4.1: Regression Plane in a Model with Two Predictors



Note: For each pair (x_i, z_i), the regression plane shows μ_i. In this graph, X and Z are assumed to be uncorrelated.

This is identical to the simple regression model, except for the inclusion of the values z_i and their associated parameter β_2. The model makes explicit that Z is a predictor of Y. Less obvious from the equation, but still an essential aspect of the model, is that X and Z can be correlated. In the estimation of the effect of X, any overlap between Y and Z and between X and Z is taken into consideration, as we shall see in Chapter 5. This means that β_1 can be viewed as a **partial slope**, i.e., a slope which is net of the effect of Z.

Corresponding to the population regression model is a population regression function or conditional expectation function of the form

$$E[y_i] = \beta_0 + \beta_1 x_i + \beta_2 z_i = \mu_i$$

This function constitutes a plane in a 3-dimensional space, as is illustrated in Figure 4.1. Thus, we see that the multiple regression equivalent of the regression line is a plane.

The expected values may be compared to the actual values of Y. In most cases, such a comparison will show a discrepancy, which is absorbed into an

error term:

$$\begin{aligned} \varepsilon_i &= y_i - \mu_i \\ &= y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i \end{aligned}$$

As was the case in simple regression analysis, the errors are attributed to omitted predictors, measurement error in the dependent variable, and the inherent unpredictability of human behavior.

The inclusion of two predictors is a vast improvement over the simple regression model. Still, we would want to include even more predictors (although such inclusion should always be weighted against the goal of simplification that is characteristic of modeling). Fortunately, it is easy to extend the model. To do so, let us designate the predictors as $X_1, X_2, \cdots, X_K$, where K is an arbitrary integer. The approach of indicating predictors as X_k (where $k = 1, 2, \cdots, K$) is more flexible than using different letters from the alphabet, which, in English, would limit us to 26 predictors. With this notational convention in place, the multiple population regression model can now be written as:

Equation 4.1: Multiple Population Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

The multiple conditional expectation function is given by

Equation 4.2: Multiple Population Regression Function

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} \end{aligned}$$

For the errors, we have

Equation 4.3: Errors

$$\begin{aligned} \varepsilon_i &= y_i - \mu_i \\ &= y_i - \beta_0 - \sum_{k=1}^{K} \beta_k x_{ik} \end{aligned}$$

It is conventional to add a number of regression assumptions to the model, but we shall hold off on this until section 4.2.

4.1.2 Matrix Notation

The multiple regression model can be written more compactly when we adopt matrix notation.[2] In addition to simplification, the use of matrix notation also allows for a useful geometric interpretation of the model. For these reasons, even applied users typically formulate the regression model in matrix form.

Matrix formulation of the multiple regression model starts with the recognition that Equation 4.1 specifies a system of n equations, one for each sampling unit. Specifically,

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_K x_{1K} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_K x_{2K} + \varepsilon_2 \\ &\vdots \quad\quad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nK} + \varepsilon_n \end{aligned}$$

Like any other system of equations, this system can be represented in matrix terms. First, we collect the different elements of the regression model into vectors and matrices, making sure that similar elements are retained in the same matrix/vector. Next, we bring those vectors and matrices together in an equation that adequately captures the full system.

[2] This notation is introduced in Appendix B, which you should master before proceeding with the current section.

Let us define three vectors. First, $\mathbf{y}$ is a vector of n elements containing the data on the dependent variable:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Second, $\boldsymbol{\varepsilon}$ is a vector of n elements containing the errors:

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Third, $\boldsymbol{\beta}$ is a vector of $K+1$ regression coefficients:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_K \end{pmatrix}$$

In addition, we define a $n \times (K+1)$ matrix $\mathbf{X}$ that contains a constant as well as the data on the predictors:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{pmatrix}$$

The first column in $\mathbf{X}$ contains the constant, which is 1 for all units and has β_0 as its associated parameter. The second column contains the data on the first predictor, which has β_1 as its associated parameter. The third column contains the data on the second predictor, which has an associated parameter of β_2, etc.

We have now captured all of the information in the system of equations that

Table 4.1: FDI in Four African Countries in 2012

Country	i	FDI Per Capita	GDP Per Capita	Political Stabiliy
Botswana	1	73.39	7254.56	88.2
Equatorial Guinea	2	2736.67	22404.76	53.1
Nigeria	3	42.06	2742. 22	3.3
Uganda	4	33.16	551.38	19.4

Note: Data from the World Bank. FDI and GDP are measured in constant USD. Stability is a percentile rank.

is Equation 4.1. The only thing that is left to do is to place these components together in an equation, to wit

Equation 4.4: Regression Model in Matrix Form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

This is one of the most important equations in this book, one that we shall work with extensively throughout.

Because Equation 4.4 is so important, it may be useful to illustrate its operation. For this, we consider the data in Table 4.1. This data lists the per capita foreign direct investment (FDI), per capita GDP, and political stability values for four African countries in 2012. If we treat per capita FDI as the dependent variable, then we can formulate the following model

$$\text{FDI}_i = \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{Stability} + \varepsilon_i$$

Collecting the data on the dependent variable we get $\mathbf{y}^\top = (y_1\ y_2\ y_3\ y_4) = (73.39\ 2736.67\ 42.06\ 33.16)$. Here the $\top$ symbol denotes the transpose, i.e., the operation that turns the column vector $\mathbf{y}$ into a row vector. In a similar vein, we can define $\boldsymbol{\varepsilon}^\top = (\varepsilon_1\ \varepsilon_2\ \varepsilon_3\ \varepsilon_4)$, which is the vector of errors for Botswana, Equatorial Guinea, Nigeria, and Uganda (in that order). The vector $\boldsymbol{\beta}^\top =$

$(\beta_0\ \beta_1\ \beta_2)$ is the vector of regression coefficients. Finally, we define

$$\mathbf{X} = \begin{pmatrix} 1 & 7254.56 & 88.2 \\ 1 & 22404.76 & 53.1 \\ 1 & 2742.22 & 3.3 \\ 1 & 551.38 & 19.4 \end{pmatrix}$$

Here, the first column contains the constant, the second column contains per capita GDP, and the third column contains political stability. Alternatively, the first row contains the data on the predictors for Botswana, the second row for Equatorial Guinea, the third row for Nigeria, and the fourth row for Uganda.

Equation 4.4 states that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ recovers four regression equations, one for each of the four countries in the data. To see if this is the case we begin by expanding $\mathbf{X}\boldsymbol{\beta}$:

$$\begin{pmatrix} 1 & 7254.56 & 88.2 \\ 1 & 22404.76 & 53.1 \\ 1 & 2742.22 & 3.3 \\ 1 & 551.38 & 19.4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 \cdot \beta_0 + 7254.56 \cdot \beta_1 + 88.2 \cdot \beta_2 \\ 1 \cdot \beta_0 + 22404.76 \cdot \beta_1 + 53.1 \cdot \beta_2 \\ 1 \cdot \beta_0 + 2742.22 \cdot \beta_1 + 3.3 \cdot \beta_2 \\ 1 \cdot \beta_0 + 551.38 \cdot \beta_1 + 19.4 \cdot \beta_2 \end{pmatrix}$$

These are the linear predictors for Botswana, Equatorial Guinea, Nigeria, and Uganda, respectively. To these, we add the vector $\boldsymbol{\varepsilon}$:

$$\begin{pmatrix} 1 \cdot \beta_0 + 7254.56 \cdot \beta_1 + 88.2 \cdot \beta_2 \\ 1 \cdot \beta_0 + 22404.76 \cdot \beta_1 + 53.1 \cdot \beta_2 \\ 1 \cdot \beta_0 + 2742.22 \cdot \beta_1 + 3.3 \cdot \beta_2 \\ 1 \cdot \beta_0 + 551.38 \cdot \beta_1 + 19.4 \cdot \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix} =$$

$$\begin{pmatrix} 1 \cdot \beta_0 + 7254.56 \cdot \beta_1 + 88.2 \cdot \beta_2 + \varepsilon_1 \\ 1 \cdot \beta_0 + 22404.76 \cdot \beta_1 + 53.1 \cdot \beta_2 + \varepsilon_2 \\ 1 \cdot \beta_0 + 2742.22 \cdot \beta_1 + 3.3 \cdot \beta_2 + \varepsilon_3 \\ 1 \cdot \beta_0 + 551.38 \cdot \beta_1 + 19.4 \cdot \beta_2 + \varepsilon_4 \end{pmatrix}$$

Equating this to the vector $\mathbf{y}$ yields the following equations

$$\begin{aligned} 73.39 &= \beta_0 + \beta_1 \cdot 7254.56 + \beta_2 \cdot 88.2 + \varepsilon_1 \\ 2736.67 &= \beta_0 + \beta_1 \cdot 22404.76 + \beta_2 \cdot 53.1 + \varepsilon_2 \\ 42.06 &= \beta_0 + \beta_1 \cdot 2742.22 + \beta_2 \cdot 3.3 + \varepsilon_3 \\ 33.16 &= \beta_0 + \beta_1 \cdot 551.38 + \beta_2 \cdot 19.4 + \varepsilon_4 \end{aligned}$$

The first regression equation pertains to Botswana, the second equation to Equatorial Guinea, the third equation to Nigeria, and the fourth equation to Uganda. The formula $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ indeed recovers the regression model for each of the sample units.

Having demonstrated the basics of the matrix notation of the multiple regression model. we can derive two further results that are of considerable importance. First, the conditional expectation function is given by

Equation 4.5: Regression Function in Matrix Form

$$\boldsymbol{\mu} = E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\mu}^\top = (\mu_1 \ \mu_2 \cdots \mu_n)$ is a vector of expected values of the dependent variable derived from all of the predictors in the model. Second, the errors are equal to

Equation 4.6: Errors in Matrix Form

$$\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{\mu} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

4.2 Regression Assumptions

It is customary to add a number of assumptions to the multiple regression model. The basic assumptions do not fundamentally differ from those of the simple regression model. The major difference is that we make one additional assumption in the multiple regression model regarding the predictors. Using

matrix notation, we can also write the assumptions a bit more compactly than we did in Chapter 2.

As we did in Chapter 2, we divide the assumptions into three blocks: (1) assumptions about the predictors; (2) assumptions about the errors; and (3) assumptions about the relationship between the predictors and the errors. In terms of the predictors we stipulate the following:

Assumption 4.1

1. $\mathbf{X}$ is fixed
2. $\mathbf{X}$ is full rank

Assumption 4.1.1 is a generalization of Assumption 2.1 to multiple predictors. As such, we do not have to dwell on its meaning. Assumption 4.1.2, however, is new. The matrix $\mathbf{X}$ is full rank as long as $K + 1 \leq n$ and there is no perfect **multicollinearity**. Under perfect multicollinearity, one or more predictors in the model are perfect linear functions of the remaining predictors. Effectively, this means the data on these predictors contain no unique information, which makes it impossible to estimate their effect. The topic of multicollinearity is discussed in greater detail in Chapter 10.

The various assumptions about the error terms can be summarized as follows:

Assumption 4.2

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

This assumption has four parts. First, the elements of the vector $\boldsymbol{\varepsilon}$ are assumed to be draws from a normal distribution (cf. Assumption 2.2). Second, this normal distribution has a mean of 0 for each of the error terms: $E[\varepsilon_i] = 0$, just as we stated in Assumption 2.3. Third, the errors are homoskedastic (cf. Assumption 2.4) and display no autocorrelation (cf. Assumption 2.5). The last two assumptions are implied by the specification of the variance-covariance

matrix of the normal distribution, which is $\sigma^2\mathbf{I}$. The customary distribution that we impose on $\boldsymbol{\varepsilon}$ has a covariance matrix of

$$\begin{aligned}\boldsymbol{\Omega} &= E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] \\ &= \begin{pmatrix} Var[\varepsilon_1] & Cov[\varepsilon_1,\varepsilon_2] & \cdots & Cov[\varepsilon_1,\varepsilon_n] \\ Cov[\varepsilon_2,\varepsilon_1] & Var[\varepsilon_2] & \cdots & Cov[\varepsilon_2,\varepsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\varepsilon_n,\varepsilon_1] & Cov[\varepsilon_n,\varepsilon_2] & \cdots & Var[\varepsilon_n] \end{pmatrix}\end{aligned}$$

By equating this to $\sigma^2\mathbf{I}$, we obtain

$$\begin{aligned}\boldsymbol{\Omega} &= \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \\ &= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}\end{aligned}$$

Consequently, $Var[\varepsilon_1] = Var[\varepsilon_2] = \cdots = Var[\varepsilon_n] = \sigma^2$, which amounts to saying there is homoskedasticity. Further, all of the off-diagonal elements, which correspond to the covariance between the errors of different units, are zero. Hence, none of the errors covary, which is the same as saying there is no autocorrelation.

The final regression assumption that we make concerns the relationship between the errors and the predictors:

Assumption 4.3

$$E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$$

This assumption generalizes Assumption 2.6. As such, it carries the same impli-

cations: (1) omitted predictors are unrelated to the predictors in the model; (2) the functional form has been correctly specified; and (3) there is no reciprocal relationship between the dependent variable and any of the predictors.

4.3 The Sample Regression Model

While our primary interest lies with the population regression model, we cannot study it directly because the regression coefficients and errors are unknown. Hence, we rely on the sample regression model, which serves as an estimator of the population regression model. We now consider this model in greater detail.

4.3.1 Scalar Notation

Consider again the population regression model with two predictors, X and Z. In order to study this model, we estimate the parameters using the information in our sample. We can then formulate the sample regression model as

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + e_i$$

Here $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the OLS/MM/ML estimators of the regression coefficients, and e_i is the residual. The sample regression function gives the fitted or predicted values in the sample:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i$$

Consequently,

$$e_i = y_i - \hat{y}_i$$

Compared to the population regression function, there are no unknowns in the sample regression function: we have determinate values for all of its ingredients.

These ideas can be easily generalized to K predictors. Specifically, the general form of the population regression model is

Equation 4.7: Sample Regression Model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_K x_{iK} + e_i$$

The sample regression function is

Equation 4.8: Sample Regression Function

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_{ik}$$

Finally, the residuals are defined as

Equation 4.9: Residuals

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_k x_{ik},$$

where $\sum_i e_i = 0$. All of these quantities are either measured or estimated, so that there are no unknowns.

4.3.2 Matrix Notation

To formulate the sample regression model in matrix form, we define three new vectors: (1) $\hat{\boldsymbol{\beta}}^\top = (\hat{\beta}_0\ \hat{\beta}_1 \cdots \hat{\beta}_K)$ is a vector of estimated regression coefficients; (2) $\mathbf{e}^\top = (e_1\ e_2 \cdots e_n)$ is a vector of residuals; and (3) $\hat{\mathbf{y}}^\top = (\hat{y}_1\ \hat{y}_2 \cdots \hat{y}_n)$ is a vector of fitted (predicted) values. The sample regression model may then be defined as

Equation 4.10: Sample Regression in Matrix Form

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e} \\ &= \hat{\mathbf{y}} + \mathbf{e} \end{aligned}$$

As we shall see in the next chapter, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Consequently, the fitted values are given by

Equation 4.11: Fitted Values

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top}_{\mathbf{H}}\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

Here, $\mathbf{H}$ is a $n \times n$ symmetric matrix that is known as the **hat matrix**. This matrix maps the observed onto the fitted values. It is an idempotent matrix with the following shape

$$\mathbf{H} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix}$$

The sum of the diagonal elements of $\mathbf{H}$—the so-called *trace* of the matrix—is equal to $K+1$, i.e., the number of predictors plus the constant. The off-diagonal elements are bounded between -.5 and .5.

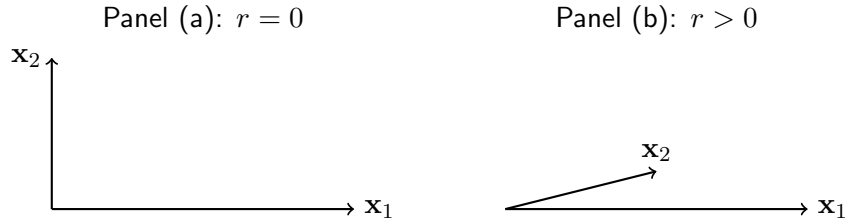
As is shown in Appendix C.2, the residuals may also be expressed in terms of the hat matrix. Specifically,

Equation 4.12: Residuals

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$

This vector is uncorrelated with the vector of fitted values.

Figure 4.2: Geometric Representation of Predictors as Vectors



Note: The length of each vector is proportional to the standard deviation of each variable. The correlation between the vectors is a function of their correlation. In the left panel, the two predictors are uncorrelated. In the right panel, they are correlated positively.

4.4 Vector Geometry

The matrix representation of the multiple regression model allows us to depict that model in a vector space. Any matrix can be represented in terms of vectors. In regression analysis, it is useful to depict the predictors as vectors in a K-space, where, in keeping with the notation so far, K denotes the number of predictors in the model. Doing so adds some interesting insights about regression analysis.

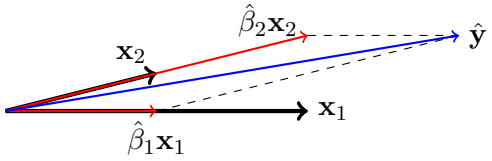
We illustrate the vector geometry of multiple regression analysis by considering a model with two predictors. We simplify the model by assuming that all variables have been centered about their sample means. This allows us to drop the constant from the model. We now have the following sample regression function

$$\hat{\mathbf{y}} = \hat{\beta}_1 \mathbf{x}_1 + \hat{\beta}_2 \mathbf{x}_2$$

The absence of the intercept immediately reveals that everything is in mean deviation form.

With this setup, we can depict the predictors in two dimensions, as is done in Figure 4.2. Here, $\mathbf{x}_1$ and $\mathbf{x}_2$ each are depicted as a vector. The length of each vector—the so-called vector norm—is equal to the square root of $n - 1$ times the variance of the predictor. The vector angle is a function of the correlation between the two vectors. Specifically, if α is the angle between the two vectors, then $\cos \alpha = r$ is the correlation. In the left panel of Figure 4.2, we see that the vectors for the two predictors are orthogonal, which means that the correlation is

Figure 4.3: Vector Representation of the Regression Plane



Note: The black vectors represent the predictors. The red arrows represent the predictors weighted by their partial slope coefficients. The blue vector of predicted values runs between the corners of the parallelogram that can be created from the weighted predictors.

0. In the right panel, we see that the vectors are not orthogonal but at an acute angle. This means that they are positively correlated. Note that Assumption 4.1 precludes the two predictors from being perfectly correlated. In vector geometric terms, this means that the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$ cannot be 0.

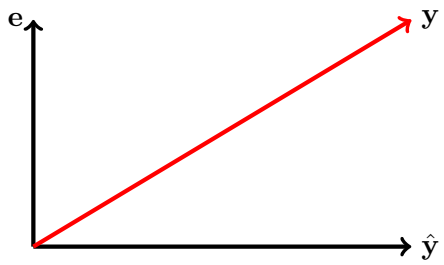
We would now like to represent the predicted values as a vector as well. Let us take Panel (b) in Figure 4.2 as the starting point. We begin by multiplying the vectors by their respective regression coefficients. Imagine that $\hat{\beta}_1 = .5$ and $\hat{\beta}_2 = 2$. The weighted vectors are now represented by the red arrows in Figure 4.3. These can be used to create a parallelogram, which is the regression plane. The vector between the lower left and upper right corner of the parallelogram is the vector of predicted values. It is the resultant of the weighted predictor vectors.

The vector of actual values of the dependent variable is the resultant of the vectors of predicted values and residuals. As we show in Appendix C.2, these are uncorrelated and thus orthogonal. Figure 4.4 shows how these vectors are combined.

4.5 Interpretation

The elements $\beta_1 \cdots \beta_K$ of the vector $\boldsymbol{\beta}$ are known as the partial slopes. They can be interpreted using a **ceteris paribus** assumption. This means that we hold everything else constant when we interpret the effect of a particular predictor on the dependent variable; only the predictor of interest is allowed to change.

Figure 4.4: Vector Representation of the Predicted Values and Residuals



Note: The vector $\mathbf{y}$ is the resultant of $\hat{\mathbf{y}}$ and $\mathbf{e}$.

In terms of the marginal effect, the *ceteris paribus* assumption implies that we take the *partial* derivative of the population regression function with respect to the predictor of interest:

Equation 4.13: Marginal Effect

$$\frac{\partial \mu}{\partial x_j} = \frac{\partial(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots \beta_K x_K)}{\partial x_j} = \beta_j$$

In terms of discrete changes, we do the same. For example, if we increase X_j by one unit, we leave all of the remaining predictors constant. Let μ^1 and μ^2 be the expected values of Y for $X_j = x_j$ and $X_j = x_j + 1$, respectively. Then, *ceteris paribus*,

$$\begin{aligned} \mu^2 &= \beta_0 + \beta_1 x_1 + \cdots + \beta_j (x_j + 1) + \cdots + \beta_K x_K \\ \mu^1 &= \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_K x_K \end{aligned}$$

It is now easily verified that $\Delta\mu = \mu^2 - \mu^1 = \beta_j$. We say that for a unit increase in X_j, the dependent variable is expected to change by β_j units *cetris paribus*. If instead we apply a change of δ units to X_j, then the discrete change is

Equation 4.14: Discrete Change

Ceteris paribus,

$$\Delta\mu|\delta x_j = \beta_j\delta$$

Note that the *ceteris paribus* assumption means that we keep constant predictors even if they are correlated with the predictor of interest. This is, of course, somewhat artificial. After all, a change in one predictor may be associated with simultaneous changes in other predictors; indeed, in the real world, it almost always is. Without the simplifying device of *ceteris paribus*, however, it would be quite difficult to interpret the effect of single predictors and this is why we employ it.

As an illustration consider the regression of per capita FDI on per capita GDP and political stability in Africa. In Table 4.1, we showed the data for four countries. The full sample consists of 50 countries and is shown in Table 4.2.[3] For this sample, we find

$$\widehat{\text{FDI}}_i = -116.27 + 0.08 \cdot \text{GDP}_i + 1.91 \cdot \text{Stability}$$

We can give the following interpretation to the coefficient for per capita GDP: for each additional dollar of per capita GDP, per capita FDI is expected to increase by 8 dollar cents, holding political stability constant. The effect of political stability can be interpreted as follows: for an increase of one percentile in political stability, per capita FDI is expected to increase by one dollar and 91 cents, holding per capita GDP constant.

4.6 Assessing the Importance of Predictors

One issue that frequently arises in multiple regression analysis is an assessment of the (relative) importance of different predictors. How much does a particular predictor "matter?" Which predictor matters the most? These are common

[3] The variable corruption will not be used in this analysis and will make its first appearance in Chapter 5.

Table 4.2: FDI in Africa in 2012

Country	FDI	GDP	Stability	Corruption	Country	FDI	GDP	Stability	Corruption
Algeria	72.06	9813.92	10.0	38.3	Madagascar	36.44	444.95	27.0	34.4
Benin	28.01	750.51	57.8	19.1	Malawi	8.14	266.59	45.0	39.7
Botswana	73.31	7254.56	88.2	78.9	Mali	26.79	696.18	3.8	25.4
Burkina Faso	20.01	651.65	26.5	37.3	Mauritania	365.13	1042.82	15.6	32.1
Burundi	0.06	251.01	5.7	0.1	Mauritius	456.19	8861.83	79.1	67.0
Cameroon	24.23	1219.93	27.5	5.7	Morocco	87.39	2948.96	32.2	41.1
Cape Verde	149.80	3554.41	72.0	74.2	Mozambique	223.58	593.29	58.3	32.5
CAR	15.73	479.47	5.2	19.6	Namibia	486.14	5770.31	78.2	66.0
Chad	27.54	1035.26	17.1	6.2	Niger	48.70	385.34	14.7	29.2
Comoros	14.46	767.21	35.1	26.3	Nigeria	42.06	2742.22	3.3	11.0
Côte D'Ivoire	16.23	1365.87	11.4	20.6	Rep. Congo	635.90	3153.74	30.8	9.6
Dem. Rep. Congo	44.01	446.03	2.8	4.3	Rwanda	13.95	630.11	39.3	72.7
Djibouti	127.96	1574.63	51.7	45.0	Sao Tome	56.32	1399.95	46.9	43.5
Egypt	34.66	3256.02	7.6	33.0	Senegal	21.73	1023.29	41.2	48.3
Equatorial Guinea	2736.67	22404.76	53.1	0.5	Seychelles	1886.78	11689.29	70.1	67.5
Ethiopia	3.04	472.16	7.1	31.1	Sierra Leone	24.10	590.32	37.4	18.2
Gabon	426.32	10929.88	56.9	35.9	South Africa	88.49	7313.98	43.1	53.1
Gambia	18.72	509.39	44.5	29.7	Sudan	62.18	1697.85	2.1	1.0
Ghana	129.88	1645.52	50.2	55.5	Swaziland	72.88	3289.72	33.6	46.9
Guinea	52.88	493.49	10.9	12.9	Tanzania	37.66	591.18	46.4	23.9
Guinea-Bissau	7.69	576.39	18.0	9.1	Togo	14.12	589.46	34.1	16.3
Kenya	5.99	1165.75	10.4	12.0	Tunesia	144.21	4197.51	22.3	53.6
Lesotho	35.91	1134.85	55.9	61.7	Uganda	33.16	551.38	19.4	17.2
Liberia	154.30	413.76	31.8	34.0	Zambia	123.02	1771.89	65.4	45.9
Libya	231.53	13302.79	6.6	2.9	Zimbabwe	29.11	908.78	21.8	5.3

Note: Data from the World Bank. FDI and GDP are measured in constant USD per capita. Stability is a percentile rank. The same is true of corruption control.

questions among practitioners of regression analysis. There are several ways to answer them. Following Achen (1982), we distinguish between theoretical, level, and dispersion importance.

4.6.1 Theoretical Importance

The theoretical importance of a predictor is simply its potential effect, i.e., the change in the outcome that is expected to result from a change in the predictor (*ceteris paribus*). This can be ascertained via the marginal effect and the discrete change defined in the previous section. In addition to reporting these statistics, researchers are expected to provide a qualitative judgment as to whether the effect is large or small. One way to do this with discrete changes is to look at the maximum effect that would arise if the predictor is moved across the full extent of the scale. In the FDI example, the maximum change on political stability is 100 units. An increase of this size would produce a change in the expected per capita FDI of 191 dollars. That covers around 7 percent of the empirical range of per capita FDI, which would seem moderately important. Note that maximum effects represent extreme changes in a predictor, which may be unrealistic. This lack of realism is frequently criticized. To avoid this criticism, you could try a different range of values for the predictor variable, e.g. 5th to 95th percentile or 1st to 3rd quartile.

The great advantage of assessing theoretical importance is that it can be done based on information that is directly available from the estimation. A major limitation, however, is that it is not possible to compare the coefficients across predictors. The reason is that the estimates of the partial slopes are driven in part by the scale of the predictors. In most applications, different predictors are on different scales, so that this fact alone accounts for differences in the slopes.[4] To enable direct comparisons across regression coefficients, it may be useful to consider using standardized regression coefficients.

[4]Indeed, one can change the slope coefficients by altering the measurement scale of the predictors. With the FDI regression, for example, we could have measured GDP per capita not in dollars but in hundreds of dollars. Now the slope coefficient associated with GDP is 8.14, but this does not mean that GDP suddenly has undergone a 100-fold increase in its effect. It simply means that a unit increase now corresponds to 100 dollars, which is 100 times larger than a unit increase on the dollar scale.

4.6.2 Level Importance

A second criterion for judging the importance of predictors is level importance. Here, the focus is on the impact of a predictor on the level of the dependent variable as measured by the mean. Thus, we care about central tendency. The level importance of a predictor X_j is defined as

Equation 4.15: Level Importance

$$LI_j = \hat{\beta}_j \bar{x}_j$$

(Achen, 1982). Level importance has the nice property that $\bar{y} = \hat{\beta}_0 + \sum_k \hat{\beta}_k \bar{x}_k = \hat{\beta}_0 + \sum_k LI_k$.[5] However, comparisons between the level importances of different predictors are difficult to make when there are scale differences between those predictors. This may be the reason why this criterion is used relatively infrequently in the social sciences.

For the FDI regression, we have a mean GDP of 2972.40, resulting in $LI_{GDP} = 0.08 \cdot 2972.40 = 241.93$. The mean of political stability is 33.50, so that $LI_{Stability} = 1.91 \cdot 33.50 = 63.85$. If we add the two level importances, we obtain 301.78. Adding $\hat{\beta}_0 = -116.27$, we get 189.51, which is the mean level of per capita FDI in the sample.

4.6.3 Dispersion Importance

A third way of judging the importance of a predictor is to consider dispersion importance, which focuses on the dispersion rather than the mean of the dependent variable. Here, the importance of a predictor is judged by the size of its **standardized regression coefficient**:

Equation 4.16: Standardized Regression Coefficient

$$\hat{\beta}_j^s = \hat{\beta}_j \frac{s_{x_j}}{s_y}$$

[5] This property follows from the OLS/ML estimator of the constant, which is $\hat{\beta}_0 = \bar{y} - \sum_k \hat{\beta}_k \bar{x}_k$ (see Chapter 5). Rearranging terms yields the expression for the mean of Y.

This may be seen as a partial slope coefficient with the measurement units of the dependent and predictor variables removed.[6] This criterion focuses on the question of what explains the variance in the dependent variable. The standardized coefficient is the square root of the portion of the variance of the dependent variable that is explained by the predictor. An advantage of this approach is that standardized coefficients can be compared across predictors.

For the FDI regression example, we have $s_{FDI} = 467.23$, $s_{GDP} = 4321.01$, and $s_{Stability} = 23.13$. Thus, $\hat{\beta}^s_{GDP} = 0.08 \cdot (4321.01/467.23) = 0.75$ and $\hat{\beta}^s_{Stability} = 1.91 \cdot (23.13/467.23) = 0.09$. In terms of dispersion importance, GDP is a more powerful predictor than political stability.

Dispersion importance may seem like the natural choice as an importance criterion, but it is not without limitations. First, hypothesis tests for standardized coefficients are difficult to obtain due to their complex sampling distributions. Second, summing the standardized coefficients across all predictors does *not* produce a measure of the total variance in the dependent variable, unless the predictors are uncorrelated. Despite these limitations, the use of standardized regression coefficients remains a common practice in certain fields such as psychology.

4.6.4 Sequential Contributions to R^2

To overcome the second problem with traditional dispersion importance metrics, we can evaluate the sequential R^2 contributions of predictors in a model. This ensures that the sum of the individual contributions recovers the actual R^2. This sounds easier than it is. When the predictors are correlated, then the sequence through which we add to the model matters a great deal.[7] In the FDI regression, for example, entering political stability as the first predictor yields a contribution to the R^2 of 0.11. When we add it as the second predictor, however, then the contribution is only 0.01. Clearly, we should take sequence into account when assessing the importance of political stability for FDI. With just two predictors

[6] It can be easily shown that $\hat{\beta}^s_j$ is the OLS/MM/ML estimator after we apply the z-transformation to Y and to X_j,i.e., after both variables have been *standardized*.

[7] The reason why this is so will become apparent in Chapter 9 when we discuss different sums-of-squares.

in the model, this is quite easy. When there are many predictors, however, the number of distinctive sequences can become very large. We then need to rely on the computer to perform this task for us.

The R package `relaimpo` will do the heavy lifting for us (Grömping, 2006). It implements a variety of relative importance metrics, but we shall limit ourselves to the so-called LMG metric (Lindeman, Merenda and Gold, 1980). Although, the computational details are complex, this may be viewed as a an average of the R^2 contributions of a predictor, where we average across all permutations for entering this predictor. The R syntax is

```
library(relaimp)
calc.relimp(object, type = "lmg")
```

Here `object` is the name of the object containing the regression results. When we apply this command to the FDI regression, we see that the relative importance for political stability is 0.06 (the average of .11 and .01), whereas it is 0.056 for GDP. The total R^2 is 0.62, so it is obvious that the LMG metric sums to the coefficient of determination. Based on these results, there can be little doubt that GDP is a more important predictor (in terms of dispersion importance) than political stability, although this in no way means that the latter variable is unimportant for FDI.

4.7 Conclusion

In this chapter, we introduced the multiple regression model. We discussed how multiple predictors can be accommodated within this model, howe their effects can be interpreted, and how their importance may be assessed. We have not yet discussed how to estimate the multiple regression model and how to test hypotheses about it. This will be the topic of the next chapter.

Chapter 5

Statistical Inference in Multiple Regression

In the previous chapter, we introduced the multiple regression model. We also previewed the estimator of the partial slope coefficients. It is now time to derive this estimator and also consider other aspects of statistical inference of the multiple regression model. In parallel to Chapter 3, we begin by deriving the OLS, MM, and ML estimators. We then discuss the properties of these estimators. We conclude by discussing the topic of hypothesis testing.

5.1 Ordinary Least Squares

In Chapter 3, we introduced the least squares criterion and used it to estimate the regression coefficients of the simple regression model. The same criterion can be used to estimate the constant and partial slopes of the multiple regression model. We start by doing this in scalar notation for a simple model. As we shall see, this produces a complex equation. We then switch to matrix notation, which results in a less complex expression that applies to any linear regression model.

5.1.1 Scalar Notation

Consider again the multiple regression model that we introduced at the start of Chapter 4:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

For this model, the least squares criterion is

$$S = \sum_{i=1}^{n} (y_i - \mu_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

We seek to minimize this criterion with respect to β_0, β_1, and β_2 simultaneously. In operational terms, this means that we take the partial derivatives of S with respect to these parameters and set them equal to zero:

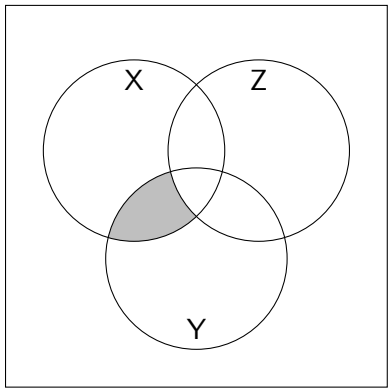
$$\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i) \\
\frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i) x_i \\
\frac{\partial S}{\partial \beta_2} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i) z_i
\end{aligned}$$

Solving these equations for the three unknown parameters yields:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z} \\
\hat{\beta}_1 &= \frac{S_{XZ} \cdot S_{YZ} - S_{XY} \cdot S_Z^2}{S_{XZ}^2 - S_X^2 \cdot S_Z^2} \\
\hat{\beta}_2 &= \frac{S_{XZ} \cdot S_{XY} - S_{YZ} \cdot S_X^2}{S_{XZ}^2 - S_X^2 \cdot S_Z^2}
\end{aligned}$$

Here, S_{XY} is $n-1$ times the covariance between X and Y, S_{XZ} is $n-1$ times the covariance between X and Z, S_{YZ} is $n-1$ times the covariance between Y and Z, S_X^2 is $n-1$ times the variance of X, and S_Z^2 is $n-1$ times the variance of Z. Thus, the OLS estimator not only takes into consideration the relationships between both predictors and the dependent variable, but also the relationship between the two predictors.

Figure 5.1: OLS in a Regression with Two Predictors



Note: Each circle represents a variable. The focus is on determining the effect of variable X. This is captured by the gray area, which reflects the overlap between the unique part of X and the part of Y that has not already been explained by Z.

The conceptual representation of this result can be found in Figure 5.1. Here, the circles represent the variables in the regression model. More precisely, each circle captures the variance in each variable. (That the circles are of equal size is a mere coincidence.) Overlaps between the circles represent shared variance, i.e., covariance.

We now wish to use this information about variances and covariances to estimate the regression coefficients. Without loss of generality, imagine our interest is in estimating β_1—the effect of X on Y. For this estimate, we consider only the information contained in the gray area. This area has two important properties. First, it lies outside the intersection between Y and Z. This means that we are only considering that part of the variance in Y that has not already been explained by Z. Second, the area lies outside the intersection between X and Z. As such, we only consider that part of X, which is not redundant with Z. In this manner, the estimated effect of X on Y is assured to capture the unique contribution of X.

We shall revisit the idea of capturing only the unique contribution of a predictor momentarily. But first it is time to generalize the idea of OLS estimation to regression models of all sizes.

5.1.2 Matrix Notation

The problem with generalizing the OLS estimator to models with potentially many predictors is that even the expressions for models with two predictors are already very complex. To simplify things, we return to matrix notation of the linear regression model. Using the definition of the inner product, it can be shown that the least squares criterion is identical to

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, as we saw in the previous chapter. Expansion of S gives

$$S = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$$

(see Appendix C.2).

We now proceed in the usual manner: we take the first partial derivatives, set these to zero, and solve for the parameters. It can be shown that

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = 2\left(\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top \mathbf{y}\right)$$

(see Appendix C.2). This is a $(K + 1) \times 1$ vector of partial derivatives, which we set equal to a vector $\mathbf{0}$ of equal length, which consists entirely of 0s:

$$2\left(\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top \mathbf{y}\right) = \mathbf{0}$$

Multiplying both sides by $1/2$ and rearranging terms, this produces the following so-called *normal equations*:

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}$$

This is a system of equations that can be solved quite easily as long as the inverse of $\mathbf{X}^\top \mathbf{X}$ exists. In this case, premultiplication of both sides of the normal equations by $(\mathbf{X}^\top \mathbf{X})^{-1}$ yields

Equation 5.1: The OLS Estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

(see Appendix C.2). It is possible to invert $\mathbf{X}^\top \mathbf{X}$ as long as Assumption 4.1.2 is satisfied, i.e., there is no perfect multicollinearity or another problem that causes $\mathbf{X}$ not to be full rank.

Equation 5.1 applies to any and all multiple regression models, regardless of how many predictors they contain. It consists of two parts. The part that is inverted, $\mathbf{X}^\top \mathbf{X}$, is a matrix of sums of squares and cross-products of the predictors. This means that the diagonal elements capture the sums of the squared predictors, whereas the off-diagonal elements capture the sums of the products of predictors with each other. On the whole, one can say that $\mathbf{X}^\top \mathbf{X}$ captures information about the predictors and their relationships. The second part, $\mathbf{X}^\top \mathbf{y}$, contains cross-products between the predictors and the dependent variable. As such, this part captures information about the relationships between the predictors and the dependent variable. Thus, we recognize the elements we saw earlier in our derivation of the OLS estimator of a model with two predictors. I hope you agree, however, that Equation 5.1 is much more elegant.

As an illustration of Equation 5.1 let us consider regression through the origin with a single predictor: $y_i = \beta_1 x_i + \varepsilon_i$. In this model,

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Notice that the column of ones has disappeared because there is no constant in the model. It can be easily shown that $\mathbf{X}^\top \mathbf{X} = \sum_i x_i^2$ and that $(\mathbf{X}^\top \mathbf{X})^{-1} = 1/\sum_i x_i^2$. We also can show that $\mathbf{X}^\top \mathbf{y} = \sum_i x_i y_i$. Thus, the OLS estimator of the slope is equal to

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

5.1.3 A Conceptual Look at OLS

So far, we have provided a mathematical discussion of OLS. But what exactly happens during an OLS estimation? To answer this question, we take a more conceptual look at OLS.

For the sake of simplicity, we revisit the model with two predictors: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$. Without loss of generality, we focus on the estimation of β_1. Our goal is to ascertain the effect of X on Y net of the effect of Z. As was visualized in Figure 5.1, there are two kinds of relationships involving Z that we want to purge from the estimate of β_1: (1) the relationship between Z and Y and (2) the relationship between Z and X. We accomplish this by proceeding in the following manner:

1. Regress Y on Z and save the residuals U. We can think of these residuals as the values of Y after the effect of Z has been removed.

2. Regress X on Z and save the residuals V. We can think of these residuals as the values of X that remain after their overlap with Z has been removed.

3. Regress U on V. The resulting slope coefficient is the estimate of β_1.

Step 3 is key. Since the values u only contain the part of y that has not already been explained by z, and since the values v only contain that part of x that is not redundant with z, a regression of u onto v is no longer marred by the presence of z. Thus, the resulting slope coefficient has to be the partial slope coefficient associated with X. We could estimate β_2 analogously, except that the regressions in steps (1) and (2) are now on X.

Let us illustrate the workings of OLS by considering once more the data on per capita FDI. To keep things simple, we consider only the data from Western Africa, which are shown in Table 5.1. We wish to estimate the effect of political stability on FDI. In a regression of FDI on GDP and stability, we find that the estimated partial slope coefficient for stability is -0.01. We can recover this estimate by first regressing on FDI on GDP; this yields

$$\widehat{FDI} = 44.20 + 0.02 \cdot GDP$$

Table 5.1: Foreign Direct Investment in West Africa

Country	FDI	GDP	Stability	u	v
Benin	28.01	750.51	57.8	-34.38	30.32
Burkina Faso	20.01	651.65	26.5	-40.00	-0.32
Cabo Verde	149.80	3554.41	72.0	19.43	25.78
Côte d'Ivoire	16.23	1365.87	11.4	-61.08	-20.19
Gambia	18.72	509.39	44.5	-37.84	18.64
Ghana	129.88	1645.52	50.2	45.78	16.74
Guinea	52.87	493.49	10.9	-3.30	-14.86
Guinea-Bissau	7.69	576.39	18.0	-50.49	-8.31
Liberia	154.30	413.76	31.8	100.07	6.57
Mali	26.79	696.18	3.8	-34.29	-23.31
Mauritania	365.13	1042.82	15.6	295.65	-13.83
Niger	48.70	385.34	14.7	-4.85	-10.34
Nigeria	42.06	2742.22	3.3	-68.63	-37.49
Senegal	21.73	1023.29	41.2	-47.28	11.90
Sierra Leone	24.10	590.32	37.4	-34.41	10.99
Togo	14.12	589.46	34.1	-44.38	7.70

Note: FDI and GDP are in constant dollars and per capita. u is the residual from the regression of FDI on GDP. v is the residual from the regression of stability on GDP.

The residuals from this regression are shown in the column titled u in Table 5.1. Next, we regress stability on GDP, which yields

$$\widehat{Stability} = 22.46 + 0.01 \cdot GDP$$

The residuals from this regression are called v in Table 5.1. Finally we regress u on v:

$$\hat{u} = 0.00 - 0.10 \cdot v$$

In the last step we indeed recover the OLS estimator of the effect of political stability. What we have accomplished here in three steps, Equation 5.1 does in one step. And it does not do this for just one predictor, but for all predictors in the model.

5.2 Method of Moments Estimation

We can also estimate the multiple regression model using the method of moments. Key here is Assumption 4.3, which produces moment conditions of the type

$$\begin{aligned}\mathbf{m}(\boldsymbol{\beta}) &= E\left[\mathbf{X}^\top \boldsymbol{\varepsilon}\right] \\ &= E\left[\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\ &= E\left[\mathbf{X}^\top \mathbf{y}\right] - E\left[\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\right] \\ &= \mathbf{0}\end{aligned}$$

The corresponding sample moment condition may be written as

$$\bar{\mathbf{m}} = \frac{1}{n}\sum_i \mathbf{x}_i^\top \mathbf{y}_i - \frac{1}{n}\sum_i \mathbf{x}_i^\top \mathbf{x}_i \boldsymbol{\beta} = \mathbf{0}$$

Rearranging terms, we get $(1/n)\sum_i \mathbf{x}_i^\top \mathbf{y}_i = (1/n)\sum_i \mathbf{x}_i^\top \mathbf{x}\boldsymbol{\beta}$. If we invert $(1/n)\sum_i \mathbf{x}_i^\top \mathbf{x}_i$ and pre-multiply both sides of the equation with this inverse, then we become the method of moments estimator of $\boldsymbol{\beta}$:[1]

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_i \mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\left(\frac{1}{n}\sum_i \mathbf{x}_i^\top \mathbf{y}_i\right) \\ &= n\left(\sum_i \mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\frac{1}{n}\left(\sum_i \mathbf{x}_i^\top \mathbf{y}_i\right),\end{aligned}$$

which can be simplified to

Equation 5.2: MM Estimator of the Regression Coefficients

$$\hat{\boldsymbol{\beta}} = \left(\sum_i \mathbf{x}_i^\top \mathbf{x}_i\right)^{-1}\sum_i \mathbf{x}_i^\top \mathbf{y}_i$$

[1] This requires that assumption 4.1.2 holds.

This is an alternative way of expressing Equation 5.1.

5.3 Maximum Likelihood Estimation

A third estimation method is maximum likelihood. Here, we start with the assumption that the dependent variable is normally distributed with a mean of μ_i and a variance of σ^2. As we have seen before, it then follows,

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\right\}$$

Define the parameters as consisting of $\boldsymbol{\beta}$ and σ^2, then the likelihood function for a single sample unit may be written as:[2]

$$\ell_i = \ln f(y_i) = -.5\ln(2\pi) - .5\ln\sigma^2 - \frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}$$

Aggregating over the entire sample, this yields a log-likelihood of

$$\ell = -.5n\ln(2\pi) - .5n\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$$

We now recognize that $\sum_i (y_i - \mu_i)^2 = \sum_i \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Consequently,

$$\ell = -.5n\ln(2\pi) - .5n\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

We seek estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ that optimize the log-likelihood. We find those estimators by taking the partial derivatives of the log-likelihood and setting the results to zero. Using the results from Appendix C.2, we have

$$\begin{aligned}
\frac{\partial\ell}{\partial\boldsymbol{\beta}} &= -\frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top\mathbf{y}\right) = \mathbf{0} \\
\frac{\partial\ell}{\partial\sigma^2} &= -\frac{1}{\sigma^2}\left(\frac{n}{2} - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) = 0
\end{aligned}$$

[2] In estimating the simple regression model, we focused on σ instead of σ^2. Both choices are valid, but a focus on the variance is more common.

These are the first-order conditions for the maximum likelihood estimators.

The first-order condition for the regression coefficients may be written as the normal equations $\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}$. As long as Assumption 4.1.2 holds true, then $\mathbf{X}^\top\mathbf{X}$ can be inverted and we can multiply both sides of the normal equations by this inverse. This produces the estimator of Equation 5.1, which is thus also the MLE.

With $\hat{\boldsymbol{\beta}}$ defined, the MLE for σ^2 can be found by evaluating $\frac{n}{2} - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{n}{2} - \frac{1}{2\sigma^2}\mathbf{e}^\top\mathbf{e} = 0$. This results in the following estimator:

Equation 5.3: MLE of the Variance

$$\hat{\sigma}^2 = \frac{\mathbf{e}^\top\mathbf{e}}{n} = \frac{SSE}{n}$$

This is the multiple regression analogue of the (biased) estimator that we derived in Chapter 3 for the simple regression model.

5.4 Properties of the Estimators

5.4.1 Regression Coefficients

The finite sample properties of $\hat{\boldsymbol{\beta}}$ derive from the *Gauss-Markov Theorem*, which we already discussed in Chapter 3. Specifically, under Assumptions 4.2-4.3, it can be shown that $\hat{\boldsymbol{\beta}}$ is BLUE, i.e., the best linear unbiased estimator. A proof of this result is offered in Appendix C.2.

Because $\hat{\boldsymbol{\beta}}$ is also the MLE, several asymptotic properties follow. Specifically, assuming that Assumptions 4.2-4.3 hold true and other regularity conditions have been satisfied, $\hat{\boldsymbol{\beta}}$ can be shown to be consistent, asymptotically efficient, and asymptotically normally distributed.

5.4.2 Error Variance

The estimator in Equation 5.3 is asymptotically unbiased. In small samples, however, it displays a negative bias. Specifically,

$$E[\hat{\sigma}^2] = \frac{n-K-1}{n}\sigma^2$$

(see Appendix C.2). Asymptotically, $(n-K-1)/n \to 1$ and $E[\hat{\sigma}^2] = \sigma^2$. In small samples, $(n-K-1)/n < 1$ and $E[\hat{\sigma}^2] < \sigma^2$; the error variance is systematically underestimated.

To overcome the finite sample bias in the estimator of the error variance, we apply the following correction.

Equation 5.4: Unbiased Estimator of the Variance

$$s^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n-K-1}$$

This estimator corrects for the degrees of freedom lost in estimating the regression coefficients.

5.5 Standard Errors and Confidence Intervals

5.5.1 Regression Coefficients

When Assumption 4.2 holds true, and the errors are homoskedastic and not correlated, then it can be shown that the **variance-covariance matrix of the estimators** (VCE) is

Equation 5.5: VCE of $\hat{\boldsymbol{\beta}}$

$$\mathbf{V}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

(see Appendix C.2 for a proof). This is a $(K+1) \times (K+1)$ matrix of variances and covariances between the estimators of the regression coefficients. Specifi-

cally,

$$\mathbf{V}[\hat{\boldsymbol{\beta}}] = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0, \hat{\beta}_K) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_K) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_K, \hat{\beta}_0) & Cov(\hat{\beta}_K, \hat{\beta}_1) & \cdots & Var(\hat{\beta}_K) \end{pmatrix}$$

The diagonal elements are the squared standard errors of the OLS estimators of the regression coefficients. These can be used to test hypotheses about individual coefficients, as we shall see in Section 5.6. The off-diagonal elements are the covariances between the OLS estimators for different regression coefficients. In general, these elements will be zero only if we, as researchers, have full control over the values of the predictors, as in an experiment. In that situation, we can ensure that the predictors are uncorrelated and through this action, the regression coefficients also will be uncorrelated. When the predictors are not manipulated but simply observed, then they generally will be correlated to some degree. This, in turn, will cause the regression coefficients to be correlated and the off-diagonal elements of the VCE to be non-zero.

The off-diagonal elements serve an important diagnostic function, as we shall see in Chapter 10. Specifically, if the estimators of different coefficients are strongly correlated, then this suggests that it may be difficult to establish the partial effects of the associated predictors. Moreover, a high correlation between regression estimates has implications for hypothesis testing, as we shall see later in this chapter.

A problem with Equation 5.5 is that σ^2 is generally unknown. An unbiased estimator of the *estimated* VCE is given by

$$\hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}] = s^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

The diagonal elements of this matrix are the squared *estimated* standard errors of the regression coefficients.

As an illustration let us consider the regression model for the data in Table

4.2. As a reminder, the model is given by

$$\text{FDI}_i = \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{Polstab}_i + \varepsilon_i$$

where Polstab is political stability. For this model, the VCE is given by

$$\hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}] = \begin{pmatrix} 5549.3188 & & \\ -0.1078 & 0.0001 & \\ -104.3894 & -0.0061 & 3.6548 \end{pmatrix}$$

where we have chosen to depict only the lower diagonal since the matrix is symmetric. The interpretation of the various elements is as follows:

- $\widehat{Var}[\hat{\beta}_0] = 5549.3188$, which means that $\widehat{SE}[\hat{\beta}_0] = \sqrt{5549.3188} = 74.4938$.
- $\widehat{Var}[\hat{\beta}_1] = 0.0001$, which means that $\widehat{SE}[\hat{\beta}_1] = \sqrt{0.0001} = 0.0102$.
- $\widehat{Var}[\hat{\beta}_2] = 3.6548$, which means that $\widehat{SE}[\hat{\beta}_2] = \sqrt{3.6548} = 1.9118$.
- $\widehat{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -0.1078$, which means that $\widehat{Cor}[\hat{\beta}_0, \hat{\beta}_1] = -0.1078/(74.4938 \cdot 0.0102) = -0.1414$.
- $\widehat{Cov}[\hat{\beta}_0, \hat{\beta}_2] = -104.3894$, which means that $\widehat{Cor}[\hat{\beta}_0, \hat{\beta}_2] = -104.3894/(74.4938 \cdot 1.9118) = -0.7330$.
- $\widehat{Cov}[\hat{\beta}_1, \hat{\beta}_2] = -0.0061$, which means that $\widehat{Cor}[\hat{\beta}_1, \hat{beta}_2] = -0.0061/(0.0102 \cdot 1.9118) = -0.3105$.

The standard errors for the constant and the partial slope for political stability are quite sizable. The partial slope for stability also shows a sizable correlation with the constant. On the other hand, the correlation between the two partial slopes seems rather modest. This suggests no great difficulties in separating out the effects if these two predictors.

With the help of the standard errors, we can compute confidence intervals for the regression coefficients. Analogous to Equation 3.21, we define the

confidence interval at level $1-\alpha$ as

$$\hat{\beta}_k - t_{n-K-1,\frac{\alpha}{2}}\widehat{SE}[\hat{\beta}_k] \leq \beta_k \leq \hat{\beta}_k + t_{n-K-1,\frac{\alpha}{2}}\widehat{SE}[\hat{\beta}_k]$$

The only change relative to Equation 3.16 is in the degrees of freedom, $n-K-1$ instead of $n-2$.

As an example, let us focus on the data in Table 4.2 and compute the confidence interval for the regression coefficient associated with per capita GDP. We have seen that the regression coefficient for this predictor is $\hat{\beta}_1 = 0.08$, whereas its estimated standard error is $\widehat{SE}[\hat{\beta}_1] = 0.01$. Let us compute the 90% confidence interval, so that $\alpha = 0.10$. With $n = 50$, we find that $t_{n-K-1,\alpha/2} = t_{47,05} = -1.68$. Thus,

$$0.08 - 1.68 \cdot 0.01 \leq \beta_1 \leq 0.08 + 1.68 \cdot 0.01,$$

which produces a confidence interval that runs from 0.06 to 0.10.

5.5.2 Predicted Values

We have seen that we can obtain predicted values of Y through $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ in matrix notation, or $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}$ in scalar notation. Since the predicted values are a function of the estimators, they are subject to sampling fluctuation. We can compute the sampling variance in the predicted values using Equation 5.5:

Equation 5.6: Variance of the Predicted Values

$$\begin{aligned} V[\hat{y}_i] &= \sum_k (x_{ik} - \bar{x}_k)^2 Var[\hat{\beta}_k] + \\ & \quad 2\sum_{j<k}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)Cov(\hat{\beta}_j, \hat{\beta}_k) + \\ & \quad \frac{\sigma^2}{n} \end{aligned}$$

(see Appendix C.2 for a derivation). This can be estimated by substituting s^2 for σ^2, as well as the elements of the estimated VCE.

Having defined the sampling variance, we can now also define a confidence interval for the predicted values, analogous to Equation 3.18. Specifically, at level $1-\alpha$, the confidence interval for the predicted values is:

$$\hat{y}_i - t_{n-K-1,\frac{\alpha}{2}}\sqrt{\hat{V}[\hat{y}_i]} \leq E[y_i] \leq \hat{y}_i + t_{n-K-1,\frac{\alpha}{2}}\sqrt{\hat{V}[\hat{y}_i]}$$

The change relative to Equation 3.18 is in the degrees of freedom, which have been adjusted to accommodate more than a single predictor.

We illustrate this using the data in Table 4.2 once more. Imagine, we would like to predict the per capita FDI when per capita GDP is $1,000 and political stability is 50, i.e., 50 percent of the countries in the world are more stable than our hypothetical case. The regression model yields

$$\widehat{\text{FDI}}_i = -116.27 + 0.08\text{GDP}_i + 1.91\text{Polstab}_i$$

At the hypothesized values of the predictors, we expect per capita FDI to be $ 60.41. The sampling variance can be computed as follows:

$$\begin{aligned}\hat{V}[\hat{y}_i] &= (1000-2936.25)^2 \cdot 0.0001 + (50-32.88)^2 \cdot 3.6548 + \\ &\quad 2(1000-2936.25)\cdot(50-32.88)\cdot -0.0061 + \frac{4069328}{50} \\ &= 82244.59\end{aligned}$$

To compute the 95% confidence interval, we set $t_{n-K-1,\alpha/2} = t_{13,0.025} = -2.01$. This produces a confidence interval of

$$60.41 - 2.01 \cdot \sqrt{82244.59} \leq E[y_i] \leq 60.41 + 2.01 \cdot \sqrt{82244.59},$$

which runs from -516.52 to 637.34. Given the length of the confidence interval, our prediction is quite imprecise.

5.6 Analysis of Variance

In Chapter 1, we encountered the concepts of SSE and SST. These concepts also play an important role in multiple regression analysis. They are part of the **analysis of variance** (ANOVA) of the regression model, which is crucial for hypothesis testing, as we shall see in the next section. As we already saw in Chapter 1, the ANOVA elements of SSE and SST also define the coefficient of determination, a topic we shall revisit in Chapter 6. Hence, it is useful to dwell a little on the ANOVA and to contemplate the relationship between SST and SSE.

In scalar notation, $SST = \sum_i (y_i - \bar{y})^2$.[3] We know that $y_i = \hat{y}_i + e_i$, so that we can also write $SST = \sum_i [(\hat{y}_i + e_i) - \bar{y}]^2$. The following result can then be demonstrated (see Appendix C.2):

Equation 5.7: ANOVA

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} e_i^2}_{SSE}$$

This decomposition of the variation in the dependent variable is the ANOVA. We see that the total variation or SST can be broken down into two parts. The first part is the **sum of squares regression** (SSR), which captures the portion of the variance due to the regression. The second part is the SSE, which captures the portion of the variance due to the errors. We can say that the total sample variation in the dependent variable breaks down into two pieces: (1) a piece, SSR, that is explained by the model and (2) a piece, SSE, that is unexplained by the model. Thus, total variation equals explained variation plus unexplained variation.

It is useful to dwell a bit longer on the meaning of the SSR. This term compares two types of predictions of the dependent variable. One type of prediction is $\hat{y}_i$, which is the prediction of the dependent variable that we obtain

[3] Appendix C derives the results in matrix form.

by considering information about the predictors. The second type of prediction is $\bar{y}$, which is the best prediction of the dependent variable if we ignore information about the predictors. That is to say, if I do not know anything about a sampling unit, then predicting that a particular unit is at the mean is the best guess I can make in the sense of minimizing the squared prediction errors. The SSR deviates from zero to the extent that the two types of predictions differ. If $SSR = 0$, this means that knowledge about the predictors produces exactly the same predictions as ignoring that knowledge. In this case, we would say that the regression contributes nothing to explaining the variation in Y.

Consider again the data from Table 4.2. For these data, $SST = 10,696,984$ and $SSE = 4,069,328$. This leaves $SSR = 6,627,656$. We see that most of the observed variation in per capita GDP is explained variation. Not all of it is explained, however, since $SSE \neq 0$. In Chapter 6, we shall see how the SSR can be parlayed into a measure of explained variation.

5.7 Hypothesis Testing

In a multiple regression model of the variety $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i$, we can perform two kinds of hypothesis test. The first kind is similar to what we discussed in Chapter 3: we can test a hypothesis about the partial slope coefficient that is associated with a particular predictor. Since only a single parameter is involved, we refer to this as a test of a simple hypothesis. The second kind of hypothesis we can test considers a number of partial slope coefficients simultaneously. We call this a test of a joint hypothesis. In the most extreme case, this hypothesis pertains to the totality of slope coefficients. We need different procedures for each of the hypothesis types.

5.7.1 Testing Simple Hypotheses

Consider the predictor X_k. For this predictor, we can test the null hypothesis that the associated regression coefficient is equal to some value q:

$$H_0 : \quad \beta_k = q$$

To test this hypothesis, we can use the same approach that we used in simple regression analysis (see Equation 3.23). That is, we compute the test statistic

$$\frac{\hat{\beta}_k - q}{\widehat{SE}[\hat{\beta}_k]} \sim t_{n-K-1}$$

By referencing the t-distribution, we can compute a p-value, which can then be used to decided the fate of the null hypothesis. Typically, we set $q = 0$ so that the null hypothesis states that the predictor has no effect in the population.

As an example, consider the regression of per capita FDI onto per capita GDP and political stability (see Table 4.2). We seek to test the null hypothesis that political stability has no effect on per capita GDP in the population: $\beta_2 = 0$. We have seen that $\hat{\beta}_2 = 1.91$ and that the standard error is 1.91. The test statistic is 0.997. When referred to a t-distribution with 47 degrees of freedom, we obtain $p = 0.324$. This is much greater than any conventional significance level, so that we fail to reject the null hypothesis. We conclude that political stability is not a significant predictor of FDI.

5.7.2 Testing Joint Hypotheses: Introducing the F-Test

Often, we want to test multiple predictors. In this context, a somewhat radical hypothesis would be that all of the predictors exert a null effect in the linear model. Here the null hypothesis is

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

This amounts to saying that the model, $E[y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$ is reducible to $E[y_i] = \beta_0$, without this resulting in a worse fit to the data. The alternative hypothesis is that at least one of the βs is non-zero in the population.

At first sight, it would seem that the joint hypothesis $\beta_1 = \beta_2 = \cdots = \beta_K = 0$ can be tested quite easily: just perform K simple hypothesis tests and say that the null hypothesis fails to be rejected when none of the tests come out as statistically significant. This approach is highly problematic, however, not least because the null hypothesis may be false even when none of the individual

coefficients are statistically significant. This happens, for example, when the predictors are highly correlated (see Appendix C.2).

We need to take a different approach to testing the joint hypothesis and this takes the form of the F-test. The intuition behind this test is straightforward and is based on the ANOVA shown earlier. Imagine that the null hypothesis is true. Then it can be shown that the SSR and the SSE, both divided by their degrees of freedom, are unbiased estimators of σ^2—in expectation, they yield the same result (see Appendix C.2). In Chapter 3, we already saw that SSE divided by its degrees of freedom is known as the MSE. Analogously, SSR divided by its degrees of freedom is called MSR. If the null hypothesis is false, then MSR is greater than MSE in expectation (see Appendix C.2).

In light of these considerations, it would make sense to use the ratio of MSR and MSE as a test statistic. Under the null hypothesis that the predictors have no effect in the linear model, this ratio should be approximately 1. Values greater than 1, would suggest that the null hypothesis is false.

To compute a p-value, we need to find the sampling distribution of the statistic MSR/MSE. For a normally and independently distributed (n.i.d.) dependent variable, it can be shown that

Equation 5.8: The F-Test Statistic

$$\frac{MSR}{MSE} \sim \mathcal{F}[K, n-K-1]$$

(see Appendix C.2). Hence, the ratio MSR/MSE is called the F-test statistic.

For the data from Table 4.2, we established earlier that $SSE = 4,069,328$ and $SSR = 6,627,656$. With a sample size of $n = 50$ and two predictors, this means that $MSE = 4,069,328/47 = 86,581.45$ and $MSR = 6,627,656/2 = 3,313,828$. Hence, $F = 3,313,828/86,581.45 = 38.27$. This obviously much larger than 1. When we refer this test statistic to the $\mathcal{F}[2, 47]$-distribution, we obtain $p = 0.000$. Thus, we conclude that we reject the null hypothesis that both per capita GDP and political stability have null effects in the population regression function.

At this point, it is useful to point to one common misunderstanding about

the F-test. Sometimes, a small value of F is interpreted as proof that the predictors have no effect on the dependent variable. However, this inference goes well beyond the information contained in the F-test. All it can tell us that the predictors have no effect given the functional form of the model that we have estimated. With a different specification, however, it may well be that the F-test becomes significant. For example, it is possible that a set of variables is highly important but that this becomes visible only when we interact them with other variables. Thus, the F-test should always be understood in terms of the model specification; to draw broader inferences about predictors is dangerous.

5.7.3 Testing Subsets of Predictors: Expanding the F-Test

In the previous section, we saw how the F-test can be used to test a hypothesis about the entire set of predictors. We can also use it to test a subset of the predictors, however. To determine our thoughts, imagine that we estimate the following FDI model:

$$\text{FDI}_i \;=\; \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{Stability}_i + \beta_3 \text{Corruption}_i + \varepsilon_i$$

An economist argues that we do not need the two political variables. She believes that a model without them will fit the data just as well as one with. The corresponding null hypothesis is $H_0 : \beta_2 = \beta_3 = 0$. This hypothesis pertains to a subset of the partial regression coefficients. How would we test it?

Two-Step Approach There are two approaches we can take, both resulting in asymptotically equivalent F-test statistics. Conceptually, the easiest approach is to estimate two models: (1) the model as specified and (2) a model that omits the political variables. Call these models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. We assume—this is important—that the sample size remains constant across the two models. For each, we can record their SSR. If the economist is correct, then $SSR_2 \approx SSR_1$. However, if the two political predictor variables significantly contribute to the model fit, then $SSR_1 > SSR_2$. Thus, we compare the two SSRs as the basis for the hypothesis test.

In the general case, we have a set of predictors that can be partitioned into two subsets: $\mathbf{X}_1$, which includes the constant, and $\mathbf{X}_2$. We write the regression model as:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

We formulate the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. We first estimate the full model, including $\mathbf{X}_2$, and record SSR_1. If $\mathbf{X}_1$ and $\mathbf{X}_2$ contain K and M predictors, respectively, then the degrees of freedom associated with the full model are $K + M$. In a next step, we estimate

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

which is what remains after substituting $\boldsymbol{\beta}_2 = \mathbf{0}$ into the full model. We now again record SSR, which now has K associated degrees of freedom. The test statistic is now defined as

Equation 5.9: Testing a Subset of Predictors

$$\frac{(SSR_1 - SSR_2)/M}{SSE_1/(n-K-M-1)} \sim \mathcal{F}[M, n-K-M-1]$$

Here SSE_1 is the error sum of squares associated with the full model.

In our example, $\mathbf{X}_1$ consists of the constant and GDP, whereas $\mathbf{X}_2$ contains political stability and corruption. When we estimate this model for the 2012 African FDI data, we obtain $SSE_1 = 3,708,867$ and $SSR_1 = 6,988,118$. Under the null hypothesis, we can drop political stability and corruption from the model. When we estimate this reduced model, we obtain $SSR_2 = 6,541,616$. The test statistic is now

$$\frac{(6,988,118 - 6,541,616)/2}{3,708,867/46} = 2.77$$

When referred to a $\mathcal{F}[2, 46]$ distribution, we obtain $p = 0.073$. Our decision about the null hypothesis now depends on the Type-I error rate. At $\alpha = 0.05$, we fail to reject the null hypothesis and would be inclined to agree with the

economist. At $\alpha = 0.10$, we would reject the null hypothesis and be inclined to argue that the political variables matter. Given the small sample size, I would be inclined to select $\alpha = 0.10$ in order to maximize statistical power. More conservative statistical minds, however, would probably balk at that choice and insist on using $\alpha = 0.05$.

Single Step Approach In order to test a hypothesis about $\boldsymbol{\beta}_2$, it actually suffices to estimate the full model. In what is known as the Wald test, we can compare the estimate of $\boldsymbol{\beta}_2$ to the hypothesized values and decide the faith of the null hypothesis in this way.

To understand this approach, we begin by formulating the concept of a **linear constraint**. A linear constraint is a linear function involving one or more parameters of the regression model. For instance, the hypothesis $\beta_2 = 0$ is a linear constraint because it equates a single parameter (β_2) to a particular value (0). In our earlier example, we had $\beta_2 = \beta_3 = 0$. This actually corresponds to two linear constraints, one for β_2 and one for β_3.

Any set of linear constraints can be formulated as a systems of equations:

Equation 5.10: Linear Constraints

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Here $\boldsymbol{\beta}$ is the familiar vector of partial regression coefficients, $\mathbf{r}$ is a $Q \times 1$ vector of values, and $\mathbf{R}$ is a $Q \times (K+1)$ matrix that defines which parameters in $\boldsymbol{\beta}$ will be constrained and how. The total number of constraints is given by Q. In our example, $\boldsymbol{\beta}^\top = (\beta_0 \ \beta_1 \ \beta_2 \ \beta_3)$. If we define $\mathbf{r}^\top = (0 \ 0)$ and

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

then $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ selects the parameters β_2 and β_3 and sets both equal to 0.

Wald's insight was to compute $\mathbf{R}\hat{\boldsymbol{\beta}}$ and to compare this to $\mathbf{r}$. If the constraints are valid, then the estimates should be close to the hypothesized values, so that $\mathbf{R}\hat{\boldsymbol{\beta}} \approx \mathbf{r}$ or $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} \approx \mathbf{0}$. On the other hand, if the constraints are false,

then $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ would deviate from $\mathbf{0}$, possibly considerably so.

To turn this idea into a test, we need to add two further considerations. First, the sign of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ does not matter for the faith of the constraint. Only the magnitude matters. To get rid of signs, we can square the discrepancies between the estimates and the hypothesized values. Second, we need to take into account sampling fluctuation. The discrepancies $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ are based on point estimates, but these fluctuate from sample to sample. To take this into account, it becomes necessary to incorporate sampling fluctuation into the test statistic. The net result of these considerations is that the test statistic may be formulated as

Equation 5.11: Testing a Subset of Predictors Redux

$$\frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[s^2 \mathbf{R} \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{R}^\top \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{Q} \sim \mathcal{F}[Q, n - K - 1]$$

Here $s^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is the estimated VCE. The pre-multiplication by $\mathbf{R}$ and post-multiplication by $\mathbf{R}^\top$ ensures that the appropriate elements—those pertaining to the parameters that are being constrained—are selected. The fact that the term $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ appears twice is the matrix equivalent of squaring the values.

Let us apply this approach to the FDI example from before. It turns out that $\hat{\boldsymbol{\beta}} = (-56.92 \; 0.08 \; 5.70 \; -5.56)$. Hence,

$$\begin{aligned}
\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} &= \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -56.92 \\ 0.08 \\ 5.70 \\ -5.56 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 5.70 \\ -5.56 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 5.70 \\ -5.56 \end{pmatrix}
\end{aligned}$$

The estimated VCE for the full regression model is given by

$$s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \begin{pmatrix} 5955.644 & -0.140 & -46.889 & -73.814 \\ -0.140 & 0.000 & -0.008 & 0.004 \\ -46.889 & -0.008 & 6.617 & -4.714 \\ -73.814 & 0.004 & -4.714 & 6.915 \end{pmatrix}$$

Pre-multiplication by $\mathbf{R}$ and post-multiplication by $\mathbf{R}^\top$ yields

$$s^2\mathbf{R}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{R}^\top = \begin{pmatrix} 6.617 & -4.714 \\ -4.714 & 6.915 \end{pmatrix}$$

This is the part of the estimated VCE that corresponds to $\hat{\beta}_2$ and $\hat{\beta}_3$, the two parameters that are being constrained. The numerator of the test statistic is now

$$\begin{pmatrix} 5.70 & -5.56 \end{pmatrix}\begin{pmatrix} 6.617 & -4.714 \\ -4.714 & 6.915 \end{pmatrix}^{-1}\begin{pmatrix} 5.70 \\ -5.56 \end{pmatrix} = 5.54$$

Dividing this by $Q = 2$ yields a test statistic of 2.77, which is the same as we derived earlier in the two-step approach. The p-value is again 0.073, so that the faith of the null hypothesis depends once more on the choice of the Type-I error rate.

The computations involved in the one-step approach may seem quite difficult. Fortunately, most statistical programs have automated procedures for this approach. The same is often true for the simpler two-step procedure. We shall discuss the relevant R routines in section 5.9.

5.8 The Conditional Expectation Function

So far, we have said a great deal about inference for the predictors. But what about the conditional expectation function, which brings together all of those predictors? What inferences can we draw about it?

Our estimator of the conditional expectation function is $\hat{y}_i = \hat{\beta}_0 + \sum_k \hat{\beta}_k x_{ik}$.

Provided that Assumption 4.3 holds, this is an unbiased estimator: $E[\hat{y}_i] = \beta_0 + \sum_k \beta_k x_{ik} = E[y_i]$. Its variance is given by

Equation 5.12: Variance of the Fitted Values

$$\begin{aligned}\mathsf{Var}(\hat{y}_i) &= \sum_k (x_{ik} - \bar{x}_k)^2 \mathsf{Var}(\hat{\beta}_k) + \\ & \quad 2\sum_{j<k} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\mathsf{Cov}(\hat{\beta}_j, \hat{\beta}_k) + \\ & \quad \frac{\sigma^2}{n}\end{aligned}$$

(see Appendix C.2). We see that this variance reduces to σ^2/n when all of the predictors are set to their sample means. Because σ^2 is unknown, we substitute the unbiased estimator s^2. This results in the estimated sampling variance of the fitted values.

The confidence interval can now be obtained analogous to Equation 3.22:

Equation 5.13: Confidence Interval Around the Regression Line

$$\hat{y}_i - t_{n-K-1,\frac{\alpha}{2}}\widehat{SE}[\hat{y}_i] \leq E[y_i] \leq \hat{y}_i + t_{n-K-1,\frac{\alpha}{2}}\widehat{SE}[\hat{y}_i]$$

This differs from the earlier result only in the adjustment of the degrees of freedom.

As an example, let us return to the data in Table 4.2. Imagine, we want to predict the per capita FDI when the per capita GDP is 2000 US Dollars and political stability is at 50. With $n = 50$, the sample mean of per capita GDP is 2972.40, while that of political stability is 33.50. The regression coefficients are $\hat{\beta}_0 = -116.27$, $\hat{\beta}_{GDP} = 0.08$, and $\hat{\beta}_{Stable} = 1.91$. Hence, $\hat{y} = -116.27 + 0.08 \cdot 2000 + 1.91 \cdot 50 = 141.80$. The estimated variance of this prediction requires that we obtain the variances and covariances of the estimates: $\mathsf{Var}(\hat{\beta}_{GDP}) = 0.0001$, $\mathsf{Var}(\hat{\beta}_{Stable}) = 3.6548$, and $\mathsf{Cov}(\hat{\beta}_{GDP}, \hat{\beta}_{Stable}) = -0.0061$. Further, we know

that $s^2 = 86,581$. Consequently,

$$\begin{aligned}\widehat{\text{Var}}(\hat{y}) &= (2000 - 2972.40)^2 \cdot 0.0001 + (50 - 33.50)^2 \cdot 3.6548 + \\ &\quad 2 \cdot (2000 - 2972.40) \cdot (50 - 33.50) \cdot -0.0061 + \frac{86581}{50} \\ &= 3020.34\end{aligned}$$

To obtain the 95% confidence interval, the critical values of the t-distribution with 47 degrees of freedom are ± 2.012. The confidence interval for $E[y]$ is thus

$$141.80 - 2.012 \cdot \sqrt{3020.34} \leq E[y] \leq 141.80 + 2.012 \cdot \sqrt{3020.34}$$

This is is approximately equal to $31.24 \leq E[y] \leq 252.36$.

5.9 Multiple Regression in R

In this chapter, we have gone through a number of sometimes intricate procedures for drawing inferences about the multiple regression model. Fortunately, many of these procedures are automated in R. In this section, we discuss how R can be used to perform estimations and hypothesis tests. We show this in the context of the data in Table 4.2. Specifically, we shall estimate the following model:

$$\text{FDI}_i = \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{Stable}_i + \beta_3 \text{Corrupt}_i + \varepsilon_i$$

We assume that the data are contained in the data frame `africa`, which contains the variables `fdipc` (FDI), `gdppc` (GDP), `polstab` (Stable), and `corrupt` (Corrupt).

5.9.1 Model Estimation

To estimate the model, we issue the following syntax:

```
fdi.fit <- lm(fdipc ~ gdppc + polstab + corrupt,
data = africa)
```

Figure 5.2: R Output for a Multiple Regression Model

```
Residuals:
    Min      1Q  Median      3Q     Max
-775.73  -75.61   -7.50   81.90 1003.45

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -56.918884  77.172818  -0.738   0.4645
gdppc         0.078383   0.009978   7.856 4.83e-10 ***
polstab       5.696335   2.572432   2.214   0.0318 *
corrupt      -5.560058   2.629611  -2.114   0.0399 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 283.9 on 46 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6307
F-statistic: 28.89 on 3 and 46 DF,  p-value: 1.181e-10
```

Note: The part in the red box gives the F-test for $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Here, we have stored the estimation results into the object `fdi.fit`. We see that the syntax for the multiple regression model is quite intuitive: we place the dependent variable before the tilde symbol and the predictors after, connecting the latter with plus signs. As is always the case in R, the estimation results do not automatically appear on the screen; to view them, we need to issue the command `summary(fdi.fit)`. This produces the results in Figure 5.2.

Most of this figure was already discussed in the context of Figure 3.4. The part that we did not discuss before is the F-test, which is shown in the red box. By default, R shows the test statistic, degrees of freedom, and p-value for the null hypothesis that all of the predictors have null effects in the population. In our case, this means that we test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. For this hypothesis, we obtain $F = 28.89$. The degrees of freedom associated with the test statistic are $K = 3$ and $n - K - 1 = 46$, respectively. When referred to a $\mathcal{F}[3, 46]$ distribution, we obtain $p = 0.000$. For any conventional Type-I error rate, the null hypothesis is rejected.

Figure 5.3: ANOVA Table

```
Analysis of Variance Table

Response: fdipc
           Df  Sum Sq Mean Sq F value    Pr(>F)
gdppc       1 6541616 6541616 81.1338 1.014e-11 ***
polstab     1   86040   86040  1.0671   0.30700
corrupt     1  360462  360462  4.4707   0.03993 *
Residuals  46 3708867   80628
---         (A)  (B)     (C)
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note: (A) degrees of freedom; (B) sums of squares; (C) mean squares. In addition, F-test statistics and p-values are shown for the individual predictors.

5.9.2 ANOVA

Once we have estimated the linear model, it is also easy to obtain the ANOVA. This can be done using the `anova` command:

```
anova(fdi.fit)
```

This produces the results shown in Figure 5.3. The output shown here deviates a bit from our discussion of ANOVA. R does not report a single sum of squares for the predictors but shows individual sums of squares for each predictor. We can derive the SSR as we defined it simply by adding the individual components.

Looking at the Figure in greater detail, panel (A) shows the degrees of freedom associated with the different sums of squares. For each of the predictors we have one degree of freedom, so that the total degrees of freedom associated with the predictors is 3. For the residuals, we have 46 degrees of freedom. Adding all of the degrees of freedom together, we get 49; this is equal to $n-1$. Panel (B) shows the sums of squares. For the residuals, we get $3,708,867$; this is SSE. If we add the sums of squares of the individual predictors, then we obtain SSR. Thus, $SSR = 6,541,616 + 86,040 + 360,462 = 6,988,118$. In panel (C), the sums of squares are divided by their degrees of freedoms in order to obtain mean squares. For the residuals, this produces 80,628; this is MSE or s^2. To obtain the MSR, we can add the individual mean squares of the

predictors and then divide the result by 3. This yields $MSR = 2,329,373$. The division of MSR by MSE yields the F-statistic that is reported as part of the regression output.

Notice that the R ANOVA table also shows F-tests for the individual predictors. In general, I recommend against relying on these in lieu of the t-tests shown in the regression output. The problem with the F-statistics shown here is that they depend on the order in which the predictors are specified in the `lm` command: GDP before political stability, before corruption. With another order, the F-statistics would have come out differently (also see the discussions in Chapter 4.6.4 and Chapter 8).

5.9.3 F-Tests for Subsets of Predictors

Imagine that we are interested in testing the null hypothesis $\beta_2 = \beta_3 = 0$. As we have seen, this hypothesis can be tested using either a two- or one-step approach. Both approaches are available in R, although they require that add-on libraries are installed first.

Two-Step Approach To implement the two-step approach, we need the library `lmtest`. We also need to estimate the model that omits political stability and corruption control. Finally, we need to implement the F-test. All of this can be accomplished using the following syntax:

```
library(lmtest)
reduced.fit <- lm(fdipc ~ gdppc, data = africa)
waldtest(fdi.fit, reduced.fit)
```

In the first line, the library `lmtest` is loaded. In the second line, we estimate the model under the null hypothesis. Again, it is essential that the sample size remains the same as in the full model. The third line implements the F-test. It takes two arguments: the full model, which we stored in `fdi.fit`, and the model under the null hypothesis, which we stored into `reduced.fit`. The test results are shown in Figure 5.4.

Figure 5.4: F-Test on a Subset of Predictors Using a Two-Step Approach

```
Wald test

Model 1: fdipc ~ gdppc + polstab + corrupt
Model 2: fdipc ~ gdppc
  Res.Df Df      F Pr(>F)
1     46
2     48 -2 2.7689 0.0732 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: Obtained using `waldtest` in the `lmtest` library.

The output in Figure 5.4 repeats the two model specifications. It also shows the degrees of freedom associated with SSE, namely 46 for the complete and 48 for the reduced model. The key information is contained in the last three columns. The column labeled "Df" subtracts the degrees of freedom: 46 - 48 = -2. The absolute value of this number corresponds to the number of restrictions, Q. The column labeled "F" shows the F-statistic for the null hypothesis, in this case 2.7689. Finally, the column labeled "Pr(>F)" gives the p-value, which is the same as we computed by hand earlier.

One-Step Approach For the one-step approach, we need the library `aod` and its associated `wald.test` function.[4] The syntax for this command is more complicated, but still a lot less involved than the computations implied by Equation 5.11.

```
library(aod)
wald.test(vcov(fdi.fit), coef(fdi.fit), Terms = 3:4,
df=df.residual(fdi.fit))
```

The command requires only the estimation results from the full model. It has four ingredients. First, it requires the VCE, which we provide here by writing `vcov(fdi.fit)`. Next, it needs the parameter estimates (`coef(fdi.fit)`).

[4] Notice the presence of the period between wald and test. This is a distinguishing characteristic from the `waldtest` function that we just described.

Figure 5.5: F-Test on a Subset of Predictors Using a One-Step Approach

```
Wald test:
----------

Chi-squared test:
X2 = 5.5, df = 2, P(> X2) = 0.063

F test:
W = 2.8, df1 = 2, df2 = 46, P(> W) = 0.073
```

Note: Obtained using `wald.test` in the `aod` library. The key results are indicated in the red box.

Third, it needs to know which parameters are being restricted under the null hypothesis, so that it can set up the matrix $\mathbf{R}$. This is done through the `Terms` option. R indexes the parameters as follows: $1 = \beta_0$, $2 = \beta_1$, $3 = \beta_2$, and $4 = \beta_3$. We want to restrict β_2 and β_3, which involve the third and fourth index. Hence, we set `Terms = 3:4` (the colon may be read as "through"). The final ingredient is the degrees of freedom, so that R can compute Q. These are obtained by issuing `df = df.residual(fdi.fit)`.

The results are shown in the red box in Figure 5.5. We again obtain an F-statistic of roughly 2.8. When referred to a $\mathcal{F}$ distribution with 2 and 46 degrees of freedom, we obtain $p = 0.073$. This is identical to what the two-step approach showed.

5.10 Reporting Multiple Regression Results

We have gone through a large number of statistical results about the multiple regression model that one could possibly be report. But what is the absolute minimum that should be reported? There is no universal standard for regression tables in political science. Most journals, however, require something akin to Table 5.2.[5]

The table has several important features. First, it clearly indicates the

[5]This table was generated using the `stargazer` package in R. This makes it possible to directly turn an estimation object such as [fdi.fit] into a table in ASCII, HTML, or Latex format.

Table 5.2: Example of a Publishable Regression Table

	Dependent variable:
	Per Capita FDI
Per Capita GDP	0.08^{***}
	(0.01)
Political Stability	5.70^{**}
	(2.57)
Corruption Control	-5.56^{**}
	(2.63)
Constant	-56.92
	(77.17)
Observations	50
Note:	*p<0.1; **p<0.05; ***p<0.01

sample size. This is an absolute must when reporting regression results. Second, it contains plane language variable names, both for the dependent variable and the predictors. It is an absolute no go to use the names of the variables as they appear in the data frame, since these are typically extremely cryptic and, as such, difficult to understand for anyone other than the author. Third, the table reports both the parameter estimates and the standard errors. It is never sufficient to report only the parameter estimates. The reader should be able to get a sense of the precision of those estimates and the easiest way to communicate this is via the standard errors. As per convention, these appear here in parentheses. Fourth, there is an indication of the significance of the predictors using the ubiquitous star notation. This notation is clearly explained in the footnote. The only thing one might add here is that the p-values are for two-sided tests.[6]

One could add to the table. Most commonly, political scientists would add a measure of fit. Since we haven't discussed this topic yet in the context of multiple regression models, I have chosen to refrain from doing so. In the next chapter, we shall see an example that adds this information.

[6]Some journals are now beginning to shun the star notation and accept only estimates and standard errors.

It is useful to comment on the precision of the estimates and standard errors. Here, I have opted to use a precision of two decimal places because this fits nicely with the dependent variable, which is measured in dollars and cents. In general, the precision should not be less than this. However, it should also not be more than three decimal places. If we do more than that it amounts to little more than feigned precision.

In this connection, it is useful to present a tip about how one can increase the size of a coefficient associated with a covariate. Imagine that the covariate in question has a partial regression coefficient of 0.0003. If we can only report three decimal places, then the estimate would have to be reported as 0.000 with proper rounding. This can look strange, especially if the effect is statistically significant. In that case, we would see a null effect that is nevertheless statistically significant. To change this, we can play a simple trick. If we transform the original covariate by dividing it by 10,000, then a specification of the regression model in terms of this new predictor will yield a regression coefficient of $10,000 \cdot 0.0003 = 3$. Many variables can be meaningfully transformed in this manner. For example, instead of measuring income in dollars, we might measure it in hundreds or thousands of dollars. Instead of measuring age in years, we might measure it in decades, etc.

5.11 Conclusions

In this chapter, we showed how we can draw inferences about the linear regression model. With this knowledge, we can already do a great deal with the regression model. However, we can do more still and this is the topic of the remaining chapters of Part II. A first step in this direction is to assess model fit and outline procedures for comparing different models.

Chapter 6

Model Fit and Comparison

One of the enduring questions of empirical social science concerns model selection. In Chapter 2, we saw that all models are simplifications; they are approximations of a data generating process. But how good of an approximation are they? And is one model better than another? These are important questions that take us well beyond the realm of null hypothesis testing. In this chapter, we provide some basic insights from the econometric and statistical literature.

6.1 Model Fit

6.1.1 The Coefficient of Determination

In Chapter 1, we introduced the coefficient of determination or R^2. This coefficient can be used for the multiple regression model, as well. In practice, however, it is often adjusted in that context, in order to take stock of the number of predictors relative to the sample size.

R-Squared The basic definition of the coefficient of determination is no different in the context of the multiple compared to the simple regression model. Thus, Equation 1.5 is once more appropriate. It may be useful, however, to derive it from a new logic, namely that of the *proportional reduction in error*.

From the perspective of prediction, we can generally identify two different approaches. In the first, we try to make a prediction without considering any logic we may have about predictors. For example, we may try to predict Burundi's per capita FDI without bringing in information about this country's per capita GDP or political stability. Let us call this an unconditional prediction because we do not condition on any predictors. A second approach is to take advantage of the predictors, in hopes that it will improve our predictions. In this case, then, we would predict Burundi's FDI income from its values on per capita GDP and political stability. Let us call this a conditional prediction because we condition on a set of predictors.

Both approaches may result in prediction errors. Let E_1 be a measure of the predictive error that arises under the unconditional approach, whereas E_2 is a measure of this error under the conditional approach. Proportional reduction in error measures now compare the size of the two errors. Specifically, we define the proportional reduction in error as

$$PRE = \frac{E_1 - E_2}{E_1}$$

Logically speaking, our predictions can never be worse using information about predictors than ignoring this information. In the worst case, the two predictive errors are identical. Thus, $0 \leq E_2 \leq E_1$. Consequently, $0 \leq PRE \leq 1$. If the predictors produce perfect predictions, then $E_2 = 0$ and $PRE = 1$. If the predictors do not help at all with the prediction, then $E_2 = E_1$, and $PRE = 0$.

The coefficient of determination is a PRE measure.[1] Here, the unconditional prediction is $\bar{y}$. In terms of minimizing the squared prediction errors, this is the best prediction of a sample unit's value on the dependent variable if we do not know anything else about that unit. We define E_1 in terms of this squared prediction error:

$$E_1 = \sum_i (y_i - \bar{y})^2 = SST$$

The conditional prediction is $\hat{y}_i = \hat{\beta}_0 + \sum_k \hat{\beta}_k x_{ik}$. We define E_2 in terms of

[1] It is certainly not the only one. Another example is Goodman and Kruskal's λ.

the residuals:

$$E_2 = \sum_i e_i^2 = SSE$$

The PRE is now

$$PRE = \frac{SST - SSE}{SSE} = 1 - \frac{SSE}{SST} = R^2$$

This is identical to Equation 1.5.

From the analysis of variance (ANOVA), we know that $SST = SSR + SSE$, so that the PRE may also be written as:

$$PRE = \frac{SSR + SSE - SSE}{SST} = \frac{SSR}{SST} = R^2$$

Consequently,

Equation 6.1: Coefficient of Determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

This means that the coefficient of determination can be computed directly from the components of the ANOVA table.

Let us return to the data from Table 4.2 and the regression of per capita FDI on per capita GDP and political stability. Here, we find that $SSE = 4,069,328$ and $SSR = 6,627,656$, so that $SST = 10,696,984$. If we divide 6,627,656 by 10,696,984, we obtain $R^2 \approx 0.620$. We explain around 62 percent of the variance in per capita FDI.

Adjusted R-Squared A major drawback of the coefficient of determination is that it never decreases when we add new predictors, even when those predictors are utterly useless for the prediction of the dependent variable. Moreover, the coefficient of determination does not at all consider the number of predictors that we include in the model relative to the sample size. Linking the size of the

model, in terms of the number of unknown parameters it contains, to the sample size seems sound statistical practice. Indeed, Lehmann (1990, , pp. 160-161) attributes the following quote to Sir Ronald Fisher: "More or less elaborate forms [of the model] will be suitable according to the volume of the data." This would suggest that the quality of a model—in the sense of its fit—should take into account both the sample size and the model complexity.

These considerations have caused the development of the so-called adjusted R^2:

Equation 6.2: Adjusted Coefficient of Determination

$$\bar{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)} = 1 - \frac{MSE}{MST}$$

Here MST is the mean square of the total, which essentially is the sample variance of the dependent variable.

Asymptotically, $\bar{R}^2 \to R^2$ for a finite K. In small samples, however, the adjusted will be less than the regular R-squared. Hence, $\bar{R}^2 \leq R^2$. A further property of the adjusted R-squared is that it, unlike R^2, can take on negative values.

In general, I recommend reporting the adjusted R-squared for the multiple regression model. This is part of the standard output of almost all statistical packages, including R. Of course, the coefficient can also be computed by hand, either from the ANOVA table or from the regular R-squared.

Consider the earlier regression of per capita FDI on per capita GDP and political stability. For this regression, $n = 50$. We have also seen that $SST = 10,696,984$ and $SSE = 4,069,328$. Hence,

$$\bar{R}^2 = \frac{4069328/(50-2-1)}{10696984/(50-1)} \approx 0.603$$

Thus, the adjusted R-squared is only slightly below the regular R-squared that we computed earlier.

It is also possible to convert the regular R-squared directly into an adjusted

R-squared. It is easily demonstrated that Equation 5.2 can also be written as

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-K-1}$$

(see Appendix C.2). In our case, $(n-1)/(n-K-1) = 1.043$. We have already seen that $R^2 = 0.620$. Hence, $\bar{R}^2 = 1 - (1 - 0.620) \cdot 1.043 = 0.603$.

The Relationship between R^2 and the F-Statistic It is sometimes argued that the F-statistic does not measure model fit. But this is not actually true because there is a simple mathematical relationship between the statistic and the coefficient of determination:

$$F = \frac{R^2}{1-R^2} \frac{n-K-1}{K}$$

(see Appendix C.2). We see that $F = 0$ when $R^2 = 0$. We also see that $F \to \infty$ when $R^2 = 1$ (assuming finite n). Thus, there is a clear relationship between model fit, as measured by the coefficient of determination, and the F-test.

6.1.2 The Root Mean Squared Error

Next to the ubiquitous coefficient of determination, it is also possible to assess model fit using the **root mean squared error** or residual standard error. This is simply the square root of the mean squared error:

Equation 6.3: Root Mean Squared Error

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_1 e_i^2}{n-K-1}}$$

This is measured on the same scale as the dependent variable. A value of 0 means that there are no prediction errors: the model fits the data perfectly. The upper-bound of the RMSE depends on the scale of the dependent variable.

In the FDI example, we saw that $SSE = 4,069,328$. With 50 observations

and two predictors, this means that $MSE = 86,581.45$. Taking the square root yields $RMSE = 294.25$. We can judge the size of this error by comparing it to the empirical range of the dependent variable, which is 2736.61. The RMSE covers around 11 percent of the range, which can be considered small. Thus, we would conclude that the model fits the data quite well.

At the start of the 1990s, there was considerable debate about the relative merits of the RMSE versus the R-squared (Achen, 1990; King, 1990; Lewis-Beck and Skalaban, 1990). The R-squared has some well-known disadvantages, including the lack of a known sampling distribution. Critics argued that everything we need to know about model fit can be gleaned from the RMSE. By now, the debate has subsided. Most political scientists continue to report R-squared values in their papers, whereas a minority opts for the RMSE. It is really a matter of taste which of these fit measures will be reported and, in case of doubt, you can simply report both.

6.1.3 Reporting Regression Results with Fit Statistics

In Chapter 5, we discussed how one should report regression results. At that time, we did not include any information about model fit. Generally, political science journals want to see such information. One way to present it is shown in Table 6.1, which pertains to the regression of per capita FDI on per capita GDP and political stability. Here the fit information—both the RMSE and the adjusted-R^2—is provided inside the table. An alternative is to provide this information in the note below the table, although this is less common.[2]

6.2 Model Comparison

So far, we have generally considered only one model for the data. This corresponds to a world in which there is only one theory for the data. This theory is one that we have chosen because we think it is the best approximation of the DGP.

[2]Like Table 5.2, Table 6.1 was generated using the `stargazer` library.

Table 6.1: A Publishable Regression Table with Fit Measures

	Dependent variable:
	Per Capita FDI
Per Capita GDP	0.08^{***}
	(0.01)
Political Stability	1.91
	(1.91)
Corruption Control	-116.27
	(74.49)
Observations	50
Adjusted R^2	0.60
Residual Std. Error	294.25 (df = 47)
Note:	*p<0.1; **p<0.05; ***p<0.01

This account of science is quite simplistic, however. In most cases, we do not know a priori what the best model is. What we have instead is a number of competing hypotheses, each providing a different account of the DGP. An important goal of our research is to sort out which model in a set of models is the best.

Indeed, one could even argue with Chamberlin (1965) that the idea of entertaining multiple models is extremely healthy for science. Rather than settling on a particular model a priori, which we then try to "prove" with all our might, we might consider a number of partially competing and partially overlapping models. Each model reveals different elements of the data. Some may be simple, others may be complex. In the end, we wish to ascertain which model provides the best account of the data, keeping in mind that one of the purposes of modeling is to keep things simple when at all possible. We thus have a rather open-minded attitude, entertaining multiple accounts of the data all at once.

Laudable as this open-mindedness may sound, how is one to decide in the end which model is best? In this section, we shall outline a number of procedures that have been proposed over the course of several decades of methodological research. A particularly promising approach is the use of the **Akaike Infor-**

mation Criterion, which we shall discuss in considerable detail as it offers an elegant solution to the task of selecting a model.

6.2.1 Nested versus Non-nested Models

One distinction that figures prominently in the literature is that between nested and non-nested models. A model is nested inside another model when it may be viewed as a subset of the latter. When this is not the case, then we say that the models are non-nested.

To determine our thoughts, let us consider the data reported in Chirot and Ragin (1975) and replicated in Table 6.2. These data pertain to the determinants of the intensity (I) of peasant revolt in Romania in 1907. Those determinants can be divided into two sets. The first set consists of the commercialization of agriculture (C) and traditionalism (T). One could call these transitional society predictors. The second set consists of the strength of the middle peasantry (M) and inequality of land ownership (G). We call these the structural predictors.

Imagine that we formulate the following two regression models for the data in Table 6.2:

$$\begin{aligned} &\text{Model I:} \quad I_i = \beta_0 + \beta_1 M_i + \beta_2 G_i + \varepsilon_i \\ &\text{Model II:} \quad I_i = \beta_0 + \beta_1 M_i + \beta_2 G_i + \beta_3 C_i + \beta_4 T_i + \varepsilon_i \end{aligned}$$

This is an example of **nested models**: Model I is a subset of Model II, which comes about by setting β_3 and β_4 equal to 0. Now consider the following setup, which actually corresponds to the way in which Chirot and Ragin (1975) approach their data:

$$\begin{aligned} &\text{Model I:} \quad I_i = \beta_0 + \beta_1 M_i + \beta_2 G_i + \varepsilon_i \\ &\text{Model II:} \quad I_i = \beta_0 + \beta_1 C_i + \beta_2 T_i + \varepsilon_i \end{aligned}$$

These two models are not nested: Model I is not a subset of Model II or vice versa.

More generally, consider the models $\mathcal{M}_1$ and $\mathcal{M}_2$. When $\mathcal{M}_1 : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and $\mathcal{M}_2 : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, then we say that $\mathcal{M}_1$ is nested inside $\mathcal{M}_2$.

Table 6.2: Peasant Revolt in Romania in 1907

County	C	T	M	G	I	County	C	T	M	G	I
Northern Romania:						Dambovita	28.6	84.2	2.9	0.58	-1.19
Bacau	13.8	86.2	6.2	0.60	-1.37	Ialomita	36.5	78.1	4.3	0.72	-1.19
Botosani	20.4	86.7	2.9	0.72	0.67	Ilfov	40.9	84.4	2.3	0.64	-0.84
Covurlui	27.6	79.3	16.9	0.66	1.90	Muscel	6.8	76.3	3.6	0.58	-1.36
Dorohoi	18.6	90.1	3.4	0.74	-0.13	Olt	41.9	89.7	6.6	0.66	2.83
Falciu	17.2	84.5	9.0	0.70	-0.84	Prahova	25.4	83.2	2.5	0.68	-1.02
Iasi	21.5	81.5	5.2	0.60	0.13	Ramnicu-Sarat	30.5	80.2	4.1	0.76	1.59
Neamtu	11.6	82.6	5.1	0.52	-0.48	Teleorman	48.2	91.0	4.2	0.70	4.33
Putna	20.4	82.4	6.5	0.64	-0.84	Vlasca	46.0	90.5	3.7	0.68	3.80
Roman	19.5	87.5	4.8	0.68	-0.22	*South West Romania:*					
Suceava	8.9	85.6	9.5	0.58	-0.75	Dolj	45.1	85.5	5.1	0.64	3.80
Tecuci	25.8	82.2	10.9	0.68	-0.93	Gorj	12.5	83.8	7.2	0.50	-1.72
Tutova	24.1	83.5	8.4	0.74	-1.55	Mehedinti	39.3	85.6	4.9	0.60	0.84
Vaslui	22.0	88.3	6.2	0.70	-0.48	Romanati	47.7	87.6	5.2	0.58	2.61
South Central Romania:						Valcea	15,2	87.3	10.8	0.42	-0.84
Arges	24.2	84.9	6.1	0.62	-1.55	*Eastern Romania:*					
Braila	30.6	76.1	1.3	0.76	-0.48	Constanta	11.7	82.3	81.7	0.42	-1.81
Buzau	33.9	86.5	5.8	0.70	-1.10	Tulcea	25.6	80.1	68.4	0.26	-1.81

Note: Source is Chirot and Ragin (1975). C = commercialization of agriculture (% of arable land devoted to wheat); T = traditionalism (% of rural population that is illiterate); M = middle peasant strength (% of land owned in units of 7-50 hectares); G = gini coefficient of inequality of land ownership; I = intensity of the revolt, which is the sum of the standardized values of spread of rebellion and violence.

That is, we obtain $\mathcal{M}_1$ by stipulating $\boldsymbol{\beta}_2 = \mathbf{0}$. However, if $\mathcal{M}_1 : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and $\mathcal{M}_2 : \mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, then the two models are not nested. This is true even though $\mathbf{X}_1$ and $\mathbf{X}_2$ may share some common variables in the case of non-nested models.

The distinction between nested and non-nested models is relevant only for model comparison approaches that take the form of hypothesis tests. It is actually irrelevant for the Akaike Information Criterion. This is just one reason why this criterion is so powerful.

6.2.2 Model Comparison Through Hypothesis Testing

Nested Models When we consider the general formulation of nested models, it is clear that we can derive $\mathcal{M}_1$ from $\mathcal{M}_2$ by setting $\boldsymbol{\beta}_2 = \mathbf{0}$. If we formulate this restriction as a null hypothesis, i.e., $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, then we can use an F-test to perform the model comparison. If we can reject the null hypothesis, then we would decide in favor of $\mathcal{M}_2$. If we fail to reject the null hypothesis, then we would decide in favor of $\mathcal{M}_1$ by virtue of its greater parsimony. We would conclude that we do not need the additional predictors in $\mathbf{X}_2$, as they add nothing to the model fit.

The previous chapter provides all of the necessary tools to perform model evaluation in this manner. Using either a one- or two-step approach, we can derive the F-statistic that allows us to test the null hypothesis. In the two-step approach this statistic is equal to

$$\frac{(SSE_1 - SSE_2)/K_2}{SSE_2/(n - K1 - K2)} \quad \sim \quad \mathcal{F}[K_2, n - K_1 - K_2]$$

Here, K_1 is the number of predictors in $\mathbf{X}_1$, including the constant, and K_2 is the number of predictors in $\mathbf{X}_2$. The one-step approach obviously gives the same result.

Let us apply the approach to the data in Table 6.2. We define $\mathcal{M}_1 : I_i = \beta_0 + \beta_1 M_i + \beta_2 G_i + \varepsilon_i$ and $\mathcal{M}_2 : I_1 = \beta_0 + \beta_1 M_i + \beta_2 G_i + \beta_3 C_i + \beta_4 T_i + \varepsilon_i$. We favor $\mathcal{M}_2$ over $\mathcal{M}_1$ if we can reject $H_0 : \beta_3 = \beta_4 = 0$. Using the R procedures outlined in the previous chapter, we obtain $F = 15.537$. When referred to a

$\mathcal{F}[2, 27]$, we obtain $p = 0.000$. Hence, we reject H_0 and decide in favor of the more complex model.

Non-Nested Models Non-nested models are more difficult to evaluate using hypothesis testing, but a useful procedure—the so-called J-test—was outlined by Davidson and MacKinnon (1981). Imagine we have two competing models:

$$\begin{aligned} \mathcal{M}_1 \quad & \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \\ \mathcal{M}_2 \quad & \mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \end{aligned}$$

The fundamental logic of the J-test is to bring these two models together into a single model:

$$\mathbf{y} \quad = \quad (1 - \alpha)\mathbf{X}_1\boldsymbol{\beta}_1 + \alpha\mathbf{X}_2 + \boldsymbol{\varepsilon}$$

If $\alpha = 0$, then this equation obviously reduces to $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, which corresponds to $\mathcal{M}_1$. On the other hand, if $\alpha = 1$, then we obtain $\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, which is $\mathcal{M}_2$. We can now turn the decision between the models into a test of $H_0 : \alpha = 0$. If we fail to reject this hypothesis, than we settle on $\mathcal{M}_1$. On the other hand, if we reject this hypothesis, then we opt for $\mathcal{M}_2$.

The problem with this setup is that we do not have sufficient information to estimate α; this parameter is not identified. We can identify α by substituting the unbiased OLS estimator for $\boldsymbol{\beta}_2$, which is exactly what Davidson and MacKinnon (1981) proposed. Hence,

Equation 6.4: J-Test

1. Estimate $\mathbf{y} = (1 - \alpha)\mathbf{X}_1\boldsymbol{\beta}_1 + \alpha\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{\varepsilon}^*$
2. Test $H_0 : \alpha = 0$ using

$$\frac{\hat{\alpha}}{\widehat{SE}[\hat{\alpha}]} \sim \mathcal{N}(0, 1)$$

Here, we should keep in mind that the standard normal sampling distribution

Figure 6.1: J-Test for Two Models of Romanian Peasant Rebellion

```
J test

Model 1: I ~ M + G
Model 2: I ~ C + T
                Estimate Std. Error t value  Pr(>|t|)
M1 + fitted(M2)  0.96329    0.16969  5.6767 4.384e-06 ***
M2 + fitted(M1)  0.28224    0.40568  0.6957    0.4923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: Obtained using `jtest` in the `lmtest` library. Model 1 is the structural model, whereas Model 2 is the transitional society model of peasant rebellion.

requires that the sample size is not too small.

In R, the procedure is automated in the `lm` package. We start by estimating the two models. Let `m1.fit` and `m2.fit` contain the estimation results of $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Now, the J-test can be obtained using

```
library(lmtest)
jtest(m1.fit, m2.fit)
```

Let us apply the procedure to the data in Table 6.2. Chirot and Ragin (1975) propose two models: (1) a structural model that includes the predictors M and G, and (2) a transitional society model that, in its simplest form, includes the predictors C and T. Let $\mathcal{M}_1$ and $\mathcal{M})_2$ denote the structural and transitional society models, respectively. When we apply the `jtest` command, we obtain the results shown in Figure 6.1. Let us focus on the first result, which is labeled `M1 + fitted(M2)`. The J-test here corresponds to

$$I_i \quad = \quad (1-\alpha)\left(\beta_0 + \beta_1 M_i + \beta_2 G_i\right) + \alpha\left(\hat{\beta}_3 C_i + \hat{\beta}_4 T_i\right) + \varepsilon_i^*$$

(We see that the adjective "fitted" means that we use the OLS estimates.) For this setup, $\hat{\alpha} = 0.96$. Since this estimate is close to one, it would suggest that the weight of the evidence favors the transitional society model. A formal test of the null hypothesis $\alpha = 0$ confirms this. The test statistic is 5.68 and has an associated p-value of 0.00. Hence, there is clear support for the transitional

societies model.

The problem here is that it is quite arbitrary whether we select $\mathcal{M}_1$ or $\mathcal{M}_2$ as the fitted model. What if we reverse the order? Then we can set up the J-test in terms of the following model:

$$I_i = (1-\alpha)(\beta_0 + \beta_3 C_i + \beta_4 T_i) + \alpha\left(\hat{\beta}_1 M_i + \hat{\beta}_2 G_i\right) + \varepsilon_i^*$$

This model corresponds to the second result in Figure 6.1, which is labeled `M2 + fitted(M1)`. The estimate for α in this setup is 0.28, which is not statistically significant by any stretch of the imagination ($p = 0.49$). Thus, we conclude against the structural and in favor of the transitional societies model.

When we take into account the possibility of reversing the nature of the fitted model, then we obtain four logical possibilities:

1. The coefficient α is not significant for either `M1 + fitted(M2)` or `M2 + fitted(M1)`. In this case, we would conclude that neither model is useful for the data at hand.

2. The coefficient α is not significant for `M1 + fitted(M2)`, but it is for `M2 + fitted(M1)`. In this case, the J-test favors $\mathcal{M}_1$.

3. The coefficient α is significant for `M1 + fitted(M2)`, but it is not for `M2 + fitted(M1)`. The J-test now favors $\mathcal{M}_2$. This is the scenario that we encountered for the Chirot and Ragin (1975) data.

4. The coefficient α is significant for `M1 + fitted(M2)` and for `M2 + fitted(M1)`. Both models are now accepted; there is not enough information in the data to discriminate between the models.

In light of these different scenarios, it is essential that we consider both `M1 + fitted(M2)` and `M2 + fitted(M1)` when conducting the J-test in R.

Hypothesis Testing Approaches: An Assessment Hypothesis testing approaches to model evaluation are widespread in the social sciences. Nevertheless, they suffer from a number of weaknesses. To our minds, one of the more significant problems is that they tend to produce binary decisions: either $\mathcal{M}_1$ or

$\mathcal{M}_2$ is correct. Sometimes things are much less clear cut and there is support for both models, but in different degrees. Those subtleties cannot be captured in hypothesis tests.

Another disadvantage is that the procedures are quite different for evaluating nested versus non-nested models. Although there are methods for evaluating non-nested models, the literature on hypothesis testing in this area is much less developed than it is for nested models. For example, the J-test can handle only two models, which is a significant limit. For all of these reasons, it is useful to look for alternatives to hypothesis testing. The Akaike Information Criterion is one such alternative.

6.3 The Akaike Information Criterion

The Akaike Information Criterion (AIC) is a numeric measure of the quality of an estimated model. The criterion is connected to the Kullback-Leibler Information, a criterion that allows us to determine how close a model is to the truth (see Appendix C.3). However, the AIC can be used to evaluate competing models without us having to know what the true model is. This is a major benefit because we generally do not know the true model. Other benefits of the AIC include that it can be used to cover any number of models, which may or may not be nested. Moreover, the AIC can be parlayed into a set of weights that give a fine-grained metric for evaluating different models. Thus, binary decisions can be avoided.

6.3.1 Defining the AIC

To obtain the AIC of a model, we need two ingredients: (1) the log-likelihood, ℓ of the model and (2) the number of estimated parameters, K. The AIC is now defined as

Equation 6.5: Akaike's Information Criterion

$$AIC_j = -2\ell_j + 2K_j$$

Table 6.3: AIC Example with Hypothetical Data

Model	ℓ	K	AIC
I	-1426.577	5	2863.054
II	-1426.635	4	2861.270
III	-2655.705	2	5315.410

Note: Model I contains three predictors; the remaining parameters are the constant and error variance. Model II contains two predictors. Finally, Model III contains no predictors.

Here, the subscript j identifies the jth model that we estimate. The derivation of Equation 6.5 is sketched in Appendix C.3.

One way of looking at the AIC is that it employs two criteria for assessing the quality of a model. First, it considers model fit through $-2\ell_j$. The smaller this value is, the better the fit. Next, it considers parsimony through $2K$. The smaller this term is, the more parsimonious our model is. We would like to optimize both fit and parsimony (see Chapter 2). In line with this principle, then, we would like AIC to be as small as possible.

How do we do this in practice? Imagine that we estimate M different models with the same data, ensuring that the sample size remains constant throughout. We now obtain an equal number of AIC values. We now favor the model that yields the smallest AIC. This model is the best in the set of models that we estimated.

The use of Equation 6.5 is illustrated in Table 6.3. Here we postulate three regression models. In reversed order, Model III contains no predictors. The only parameters estimated in this model are β_0 and σ^2, whence $K = 2$. The (hypothetical) log-likelihood for this model is -2655.705; this is the value of the log-likelihood function at the maximum likelihood estimates of β_0 and σ^2. Hence,

$$AIC_{III} = -2 \cdot (-2655.705) + 2 \cdot 2 = 5315.410$$

Model II adds two predictor variables, so that $K = 4$. At the maximum likelihood estimates, the value of the log-likelihood function for this model is -1426.635. Performing an analogous computation to the one we just saw for model III, we obtain $AIC_{II} = 2861.270$. Finally, Model I adds yet another predictor, bringing the total to three predictors and $K = 5$. With a log-likelihood function of -1426.527, we obtain $AIC_I = 2863.054$. We see that the smallest AIC-value is obtained for Model II. Hence, this is our preferred model among the three models that we estimated.

It is worthwhile to dwell a little longer on these calculations and conclusions. First it is important to stress the qualifier "among the three models that we estimated." We do not know if Model II is the best model in general. But for the data at hand and among the models that we estimated, it is the best.

Second, the computations reveal how AIC considers both fit and parsimony. Model I fits the data better than Model II, witness the smaller value of its log-likelihood function. However, the improvement in fit is so small that it does not warrant the sacrifice in parsimony, i.e., the extra parameter that needs to be estimated.

Third, if we compare AIC_I and AIC_{II}, they are not worlds apart. Indeed, the differences between these AIC values is rather small, especially when compared to their difference to AIC_{III}. This suggests that we may favor Model II over Model I, but certainly not to such an extent that we would conclude there is nothing to Model I. Later in this chapter, we shall see how one can quantify the likelihoods of different models so that we can determine the comparative plausibility of each.

Equation 6.5 is appropriate for large samples. More specifically, it may be used when $n/K > 40$, i.e., we have over 40 observations for each parameter that we estimate (Burnham and Anderson, 2004). If $/K \leq 40$, then we should use the small sample version of Akaike's Information Criterion:

Equation 6.6: Small Sample Version of Akaike's Information Criterion

$$AIC_j^c = -2\ell_j + 2K_j + \frac{2K_j(K_j+1)}{n-K_j-1}$$

Here, the superscript "c" stands for corrected. We can think of the last term in Equation 6.6 as an extra penalty for over-fitting in small samples. By this, we mean that we are fitting more parameters than is necessary or warranted. For finite K_j, this penalty term obviously goes to zero when n goes to infinity. We then again obtain Equation 6.5.

Let us apply Equation 6.6 by revisiting the two models proposed by Chirot and Ragin (1975). Each model estimates four parameters: three regression coefficients and an error variance. With $n = 32$, we have $n/K = 8$, so that we should clearly be using Equation 6.6. For each model, the log-likelihood at the maximum likelihood estimates is defined as

$$\ell_j \quad = \quad -n \ln \hat{\sigma}_j - .5n \ln(2\pi) - \frac{SSE_j}{2\hat{\sigma}_j^2}$$

For the structural model, we have: $n = 32$, $SSE = 88.032$, and $\hat{\sigma}^2 = SSE/n = 2.751$. Substitution into the log-likelihood function yields $\ell_{struc} = -61.667$. For the transitional society model, we have: $n = 32$, $SSE = 41.640$, and $\hat{\sigma}^2 = SSE/n = 1.301$. Substitution yields $\ell_{trans} = -49.619$. For the sake of this example, let us estimate a third model that combines the predictors of the structural and transitional society models. For this model, $n = 32$, $SSE = 40.928$, and $\hat{\sigma}^2 = SSE/n = 1.279$. This yields $\ell_{combi} = -49.343$.

With the preparatory work out of the way, we can now perform the computation of AIC^c; this is shown in Table 6.4. For example, the computation of the AIC of the combined model is given by

$$AIC_{combi} \quad = \quad -2 \cdot (-49.343) + 2 \cdot 6 + \frac{2 \cdot 6 \cdot (6+1)}{32-6-1} = 114.047$$

From the results, it is clear that the transitional society model is the best of the

Table 6.4: AIC^c for Three Models of Romanian Peasant Rebellion

Model	ℓ	K	$\frac{2K(K+1)}{n-K-1}$	AIC^c	AIC
Structural	-61.597	4	1.481	132.677	131.195
Transitional Society	-49.619	4	1.481	108.720	107.239
Combined	-49.343	6	3.360	114.047	110.687

Note: Based on the data in Table 6.2.

three models shown here, since it has the smallest AIC^c.

For comparison, the last column in Table 6.4 shows the normal (large sample) AIC values. We see that the distance between the corrected AICs of the transitional society and combined models is much larger than that between the regular AICs. The regular AIC already recognizes that the combined model does not fit the data much better than the transitional society model, so that the addition of 2 extra parameters is not really worthwhile. The corrected AIC penalizes the inclusion of these parameters in the combined model even more because it recognizes that 2 additional parameters on a sample size of 32 places a lot of extra demand on relatively scarce data.

6.3.2 AIC in R

We do not actually have to compute the log-likelihood functions by hand in R. We can extract them, as well as the number of estimated parameters, by using the following syntax:

```
logLik(trans.fit)
```

where `trans.fit` contains the estimation results from the transitional society model. This generates the output shown in Figure 6.2. We see that the log-likelihood for the model is -49.61934 and that it contains 4 parameters (since df $= 4$).

The information shown in Figure 6.2 allows us to compute the AIC by hand. But we do not actually have to do this either, since R can directly return AIC or AIC^c. To obtain AIC, we can issue the following command:

Figure 6.2: Extracting the Log-Likelihood from a Regression Object in R

```
'log Lik.' -49.61934 (df=4)
```

Note: df indicates the number of estimated parameters in the model.

```
AIC(trans.fit)
```

To obtain AIC^c, we employ the following function:

```
AICc <- function(m) {
  lnL <-logLik(m)[1]
  n <- attr(logLik(m), "nobs")
  K <- attr(logLik(m), "df")
  aic.c <- -2*lnL + 2*K + (2*K*(K+1))/(n-K-1)
  return(aic.c)
}
```

Application of the two commands yields the results shown in Figure 6.3.

6.3.3 Delta Values, Model Likelihoods, and Akaike Weights

A limitation of the AIC values is that they are influenced strongly by the sample sizes. This can easily result in misunderstandings. For example, imagine that $AIC_1 = 108200$ and $AIC_2 = 108210$. One might be inclined to say that the AIC difference between these two models is so small that they are, for all intents and purposes, equally good. After all, what is 10 points difference on a magnitude of over 100 thousand? The magnitudes, however, may be so large because the sample size is large. Remember that one of the components of the AIC is the log-likelihood, and this has several terms in n in the regression model. Hence, large sample sizes generate large negative values of the log-likelihood, which in turn produce large positive values of AIC (due to the multiplication by -2). Without adjusting for these effects, evaluating differences in AIC values can be misleading.

Figure 6.3: Extracting AIC and AIC^c from a Regression Object in R

```
> AIC(trans.fit)
[1] 107.2387
> AICc(trans.fit)
[1] 108.7202
```

Note: The first value is for AIC and the second value for AIC^c.

Delta Values Delta values of the AIC are obtained by subtracting the minimum AIC value in a set from each of the model AICs.

Equation 6.7: The Delta Value of a Model

$$\Delta_j = AIC_j - AIC_{\text{Min}}$$

By subtracting the minimum AIC, we remove those components that are heavily influenced by the sample size.

Let us return to the hypothetical example from Table 6.3. Here, we saw that Model II achieved the lowest AIC value. Hence, $AIC_{\text{Min}} = AIC_{II} = 2861.270$. We now subtract this value from the AICs of all of the models. The result is shown in the third column of Table 6.5.

The delta values can be interpreted directly. Burnham and Anderson (2004) provide the following rules of thumb:

$\Delta_j \leq 2$	Substantial support
$4 \leq \Delta_j \leq 7$	Weak support
$\Delta_j > 10$	No support

By these guidelines, both Model I and II receive substantial support, whereas Model III receives no support.

Model Likelihoods The likelihood or evidence of the model given the data is equal to

Table 6.5: Delta Values, Model Likelihoods, and Akaike Weights with Hypothetical Data

Model	AIC	Δ	$\mathcal{L}$	w
I	2863.054	1.784	0.410	0.291
II	2861.270	0.000	1.000	0.709
III	5315.410	2454.140	0.000	0.000

Note: Based on the results from Table 6.3.

Equation 6.8: The Likelihood of a Model

$$\mathcal{L}(M_j|Data) = \exp(-.5\Delta_j)$$

By definition, the likelihood of the best model is equal to 1. The likelihoods of the models from Table 6.3 are shown in the fourth column of Table 6.5.

Based on the model likelihoods, we can define the **evidence ratios**. These are the relative likelihoods of two models (Burnham and Anderson, 2004; Wagenmakers and Farrell, 2004). In our case, we can define three unique evidence ratios:

$$\begin{aligned}
\frac{\mathcal{L}(M_I|Data)}{\mathcal{L}(M_{II}|Data)} &= \frac{1.000}{0.410} = 2.440 \\
\frac{\mathcal{L}(M_I|Data)}{\mathcal{L}(M_{III}|Data)} &= \frac{1.000}{0.000} \rightarrow \infty \\
\frac{\mathcal{L}(M_{II}|Data)}{\mathcal{L}(M_{III}|Data)} &= \frac{0.410}{0.000} \rightarrow \infty
\end{aligned}$$

The evidence ratios, too, can be used to shed light on the models. For example, the evidence ratio of Model I relative to Model II is 2.44. This means that Model I is 2.44 more likely than Model II given the data. When the evidence ratio is very large, as in the comparisons between Model I and Model II, on one hand, and Model III, on the other, then we may conclude that the model in the denominator is very poor for the data at hand. Note that the evidence ratios depend only on the likelihoods of the two models that are being compared; other

models play no role.

Akaike Weights The Akaike weight may (heuristically) be viewed as a measure of the probability that some model is the best Burnham and Anderson (2004); Wagenmakers and Farrell (2004). It can be obtained directly from the model likelihood:

Equation 6.9: The Akaike Weight

$$w_j = \frac{\mathcal{L}(M_j|Data)}{\sum_{k=1}^{K} \mathcal{L}(M_k|Data)}$$

Here, we assume that we have estimated a total of $K > 1$ different models.

The denominator is equal to the sum of the likelihoods of all of the models that we have estimated. In Table 6.5, this is 1.410. We obtain the weights by dividing each likelihood by this sum. The results are shown in the last column of Table 6.5. We see that there is a probability of 0.291 that Model I is the best model, whereas the probability that Model II is best is 0.709. There is essentially no chance that Model III is the best model. We now obtain a relatively nuanced view of the quality of the different models. We can readily dismiss Model III. We also know that chances are that Model II is the best model. However, there remains a reasonably large probability that Model I is the best model.

Taking Advantage of R The work that we have done here by hand is automated in the `qpcR` library in R. I illustrate this using the corrected AIC values from Table 6.4. We make these the elements of a vector `aic`, which is then fed into the `akaike.weights` command:

```
library(qpcR)
aic <- c(132.6765, 108.7202, 114.0466)
akaike.weights(aic)
```

The output is shown in Figure 6.4. We see that $\Delta_{struc} = 23.956$, $\Delta_{trans} = 0.000$, and $\Delta_{combi} = 5.326$. This produces model likelihoods of $\mathcal{L}_{struc} =$

Figure 6.4: Delta Values, Model Likelihoods, and Akaike Weights in R

```
$deltaAIC
[1] 23.956310  0.000000  5.326439

$rel.LL
[1] 6.279908e-06 1.000000e+00 6.972337e-02

$weights
[1] 5.870556e-06 9.348156e-01 6.517850e-02
```

Note: The input data are from Table 6.4.

0.000, $\mathcal{L}_{trans} = 1.000$, and $\mathcal{L}_{combi} = 0.070$. Consequently, $w_{struc} = 0.000$, $w_{trans} = 0.935$, and $w_{combi} = 0.065$. In the set of models that I estimated, then, the structural model has effectively no chance of being the best, whereas the transitional society model has a 0.935 probability of being the best. This leaves a relatively small probability of 0.065 that the combined model is the best. In this case, then, there is overwhelming evidence for one particular model specification.

6.4 The Bayesian Information Criterion

Several alternatives to Akaike's Information Criterion exist. The best known of these is the Bayesian Information Criterion or BIC:

Equation 6.10: The Bayesian Information Criterion

$$BIC_j = -2\ell_j + K_j \ln n$$

Hence, the BIC differs from AIC in the second term: $K_j \ln n$ in lieu of $2K_j$. The second term in BIC exceeds that of AIC when $n > 8$. This means that BIC tends to penalize more than AIC for lack of parsimony (although not necessarily more than AIC^c). Many statisticians like this and consequently favor BIC (but see Burnham and Anderson, 2004).

Table 6.6 shows the BIC values for the three models estimated for the

Table 6.6: BIC for Three Models of Romanian Peasant Rebellion

Model	BIC	AIC	AIC^c
Structural	137.058	131.195	132.677
Transitional Society	113.102	107.239	108.720
Combined	119.481	110.687	114.047

Note: Based on the data in Table 6.2.

Chirot and Ragin (1975) data, along with AIC and AIC^c.[3] Like AIC and AIC^c, the BIC favors the transitional society model. The difference between the combined and transitional society models are somewhat larger using BIC than using AIC^c, but overall the differences are small.

6.5 Conclusion

In this chapter, we have spent considerable time on the topics of model fit and, especially, model comparison. Comparing different models is crucial because we never know the true model and usually can think of a number of different specifications for the same data. It is also essential because our estimates for a predictor do not just depend on the data but also on the other variables that we add to the model. Unfortunately, a careful consideration of different model specifications remains all too rare in the social sciences. There is no reason for this, since there are powerful tools for engaging in model comparison. Akaike's information criterion is one of those tools, which I hope is now on your radar when you start thinking about building models for your own data.

[3] The BIC values were obtained using R's `BIC` command, which works identical to the `AIC` command.

Chapter 7

Non-Linear Models

Scholars sometimes refer to the multiple regression model as the *linear* regression model. This is a bit of a misnomer, since it suggests that we would not be able to formulate non-linear regression models. In fact, the linearity of the regression model pertains only to the model parameters: these enter the model as multiplicative weights, not as powers. The multiple regression model does *not* assume linearity in the variables. Consequently, it is possible to construct a variety of non-linear models. These greatly increase the versatility of multiple regression analysis.

In this chapter, we consider three types of non-linear model: polynomial, logarithmic, and reciprocal regression. the differences between these models are summarized in Table 7.1. All of these models replace the straight regression line or plane with some form of curved geometry. The models differ in two respects. First, is the curved relationship between the dependent variable and a predictor monotonic or not? Second, is this relationship in some form bounded?

Table 7.1: Three Types of Non-Linear Regression Analysis

Model	Monotonic	Bounded
Polynomial Regression	No	No
Logarithmic Regression	Yes	No
Reciprocal Regression	Yes	Yes

By contemplating these questions, you can choose an appropriate non-linear regression model.

7.1 The Polynomial Regression Model

7.1.1 What Is Polynomial Regression?

In a polynomial regression model, we include a predictor as well as as increasing powers of that predictor. That is,

Equation 7.1: The Polynomial Regression Model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdot + \beta_P x_i^P + \varepsilon_i \\ &= \beta_0 + \sum_{p=1}^{P} \beta_p x_i^p + \varepsilon_i \end{aligned}$$

This is the so-called Pth-order polynomial regression model. Since it is linear in the parameters it remains a linear regression model and can be estimated using OLS.

As an example, consider the determinants of per capita foreign direct investment in Africa. So far, we have considered relatively simple models, which contain only per capita GDP, political stability, and in some cases corruption control (see Chapters 4-5). Now we are interested in a recipient's country level of democracy. We stipulate a non-linear effect for democracy. The argument is that investors are risk averse. They know what they can expect from a democracy. They also know this for an autocracy. What they consider risky are hybrid regimes that are neither full democracies nor full autocracies and, as such, ar ill defined. We assume that risk perceptions bear a relationship on the amount that is invested. With this theoretical argument, we expect a U-shaped relationship between a country's level of democracy and per capita FDI. As one moves from low to middling levels of democracy, per capita FDI first declines. This is the move from clear autocracies to hybrid regimes. At some point, however, further

increases in democracy result in higher levels of per capita FDI. This is the move from hybrid regimes to full democracies. Since the relationship between democracy and per capita FDI is not expected to be monotonic, a polynomial regression model is in order. In this case. we need a polynomial of order 2 to capture the shifting sign of the relationship between democracy and per capita FDI. The estimated model is thus

$$\text{FDI}_i = \beta_0 + \beta_1 \text{demo}_i + \beta_2 \text{demo}_i^2 + \varepsilon_i$$

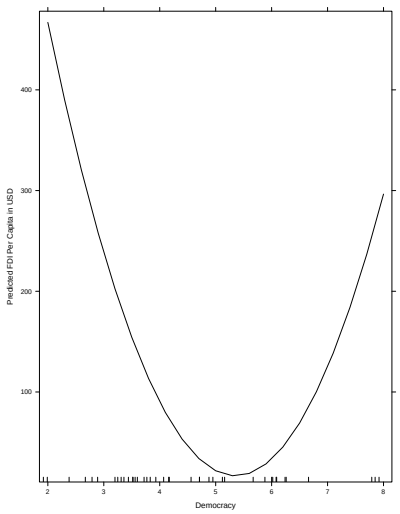
This model can be estimated in R using the following syntax:

```
fdi.fit <- lm(fdipc ~ poly(democ, degree=2, raw=TRUE),
data=africa)
```

Much of this command will look familiar (see Chapter 5). Indeed, the only new element is `poly`, which stands for polynomial. This option has several important arguments. First, we should specify the variable for which the polynomial terms should be defined, in this case `democ`, which stands for democracy. Next, one should define the order of the polynomial. In this case, we want a 2nd-order polynomial, which means that we specify `degree=2`. Finally, we add `raw=TRUE`, which prevents R from trying to transform the polynomial terms in such a way that they are no longer correlated with each other. With this command, R will now include both the linear and quadratic terms of democracy. A graphical display of the regression line can be found in Figure 7.1. The graph indeed shows the expected curvilinear effect. This graph was obtained using the `effects` library:

```
library(effects)
plot(Effect("democ", fdi.fit, se=FALSE),
xlab="Democracy", ylab="Predicted␣FDI␣Per
Capita␣in␣USD", main="")
```

Figure 7.1: Democracy and Foreign Direct Investment in Africa



Note: The OLS estimate for the linear effect of democracy is $\hat{\beta}_1 = -428.40$; the OLS estimate for the quadratic effect is $\hat{\beta}_2 = 40.00$.

7.1.2 Interpretation

Marginal Effects In our example, the predictor variable is continuous in nature. Hence, interpretation can proceed using marginal effects. For the polynomial regression model in Equation 7.1,

$$\begin{aligned} \frac{d\mu}{dx} &= \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + \cdots + P\beta_P x^{P-1} \\ &= \sum_{p=1}^{P} p\beta_p x^{p-1} \end{aligned}$$

This is the instantaneous rate of change, which depends on the value of the predictor X. This is also clearly visible in Figure 7.1.

Let us illustrate this for two different values of democracy. The estimated marginal effect for democracy is $\hat{\beta}_1 + 2\hat{\beta}_2 x$. First, consider a level of democracy of 3. With $\hat{\beta}_1 = -428.20$ and $\hat{\beta}_2 = 40.00$ (see Figure 7.1), the marginal effect is $-428.20 + 2 \cdot 40.00 \cdot 3 = -188.40$. At this level of democracy, then, per capita FDI is on a negative trajectory. Next, consider a level of democracy of 6. Now

the marginal effect is $-428.20 + 2 \cdot 40 = 51.60$. At this level of democracy, per capita FDI is on an upward trajectory.

One of the uses of marginal effects is to compute the extremes of the regression line for a given predictor, while assuming that all else remains constant. These can be obtained by setting the marginal effect for the predictor X equal to 0. Let us call the solution x_0. Three kinds of extremes can now be identified (also see Appendix A.4):

1. *Minimum:* If the marginal effect is negative to the left of x_0 and positive to the right, then x_0 is a minimum.

2. *Maximum:* If the marginal effect is positive to the left of x_0 and negative to the right, then x_0 is a maximum.

3. *Inflection Point:* If the marginal effect has the same sign to the left and the right of x_0, then x_0 is an inflection point.

Minimums and maximums constitute tipping points, in that the nature of the relationship between the dependent variable and the predictor changes signs at x_0.

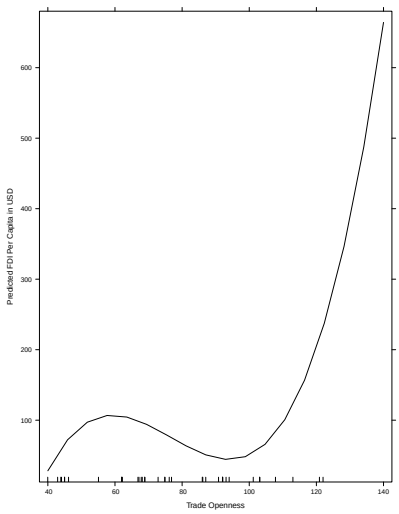
Consider, for example, our regression model of per capita FDI and democracy. We set $X = x_0$, set the marginal effect to 0, and solve for x_0:

$$-428.20 + 2 \cdot 40.00 \cdot x_0 \quad = \quad 0$$

This yields $x_0 = 5.36$; at this level of democracy, which roughly corresponds to Uganda, the marginal effect is 0. It is a minimum, as Figure 7.1 clearly reveals: to the left of 5.36, the marginal effect is negative, whereas it turns positive to the right of this value. Thus, we have a tipping point, just as the theoretical argument about risk aversion implied.

Tipping points sometimes occur outside of the empirical range of a predictor. For example, we could regress per capita FDI on political stability, and political stability squared. Let $\hat{\beta}_1$ be the estimate for political stability; in a model without other predictors, this is 8.10. Let $\hat{\beta}_2$ be the estimate for political stability squared, which is -.04. The tipping point is now $\hat{\beta}_1/(-2\hat{\beta}_2) = 95.73$.

Figure 7.2: Trade Openness and Foreign Direct Investment in Africa



Note: The OLS estimate for the linear effect of trade openness is $\hat{\beta}_1 = 49.885$; the OLS estimate for the quadratic effect is $\hat{\beta}_2 = -0.686$; the OLS estimate for the cubic effect is $\hat{\beta}_3 = 0.003$.

Although this a feasible value of the political stability scale, which ranges from 0 to 100, there are no African countries with this high of a stability score. Thus, the tipping point is never realized in the sample. It is important to check for this issue to prevent problematic interpretations. A graphical display of the full effect of the predictor will help a lot here. If you do not see a sign reversal in this plot, then you know that the tipping point is not realized in the sample.

Now let us consider a more complex example. Imagine that we regress per capita FDI onto trade openness, using a polynomial of order 3. The fitted regression line is shown in Figure 7.2. The estimated marginal effect equation is

$$\begin{aligned} \frac{\partial \hat{\mu}}{\partial \text{Open}} &= \hat{\beta}_1 + 2\hat{\beta}_2 \text{Open} + 3\hat{\beta}_3 \text{Open}^2 \\ &= 49.885 - 1.373 \text{Open} + 0.009 \text{Open}^2 \end{aligned}$$

This is a quadratic equation, which is a bit more difficult to solve. Fortunately,

R can help with this process via the `rootSolve` library:

```
library(rootSolve)
fdi.fit <- lm(fdipc ~ poly(openness, degree=3,
raw=TRUE), data=africa)
marg.eff <- function (x) {coef(fdi.fit)[2] +
2*coef(fdi.fit)[3]*x + 3*coef(fdi.fit)[4]*x^2}
roots <- uniroot.all(marg.eff, c(27, 158))
roots
```

The first line loads the `rootSolve` library. The second and third lines fit the 3rd order polynomial regression with trade openness. The fourth and fifth lines define the marginal effects function, calling the OLS estimates from the regression object `fdi.fit`. The sixth line finds the roots of the marginal effects function and stores them in the object `roots` object. Important here is that we set the range of the predictor. For our data, trade openness ranges between 27 and 158. The solutions that `uniroot.all` generates fall within this range. The final line displays the roots. In our example, the first root is a value of trade openness of 59.23, which constitutes a local maximum. The second root is a value of trade openness of 94.01; this is a local minimum.

Discrete Changes An alternative approach to interpretation is to compute the discrete change. Consider a model in which the conditional mean depends on the P polynomial terms of the predictor X. We now let X move from x to $x + \Delta$, while all other predictors remain constant. The discrete change is then given by

$$\Delta\mu = \sum_{p=1}^{P} \beta_p (x + \Delta)^p - \sum_{p=1}^{P} \beta_p x^p$$

Consider again the model where we predict per capita FDI from democracy and its square (Figure 7.1). Imagine that we move democracy from 4.12 (the sample median) to 5.12. The discrete change in the predicted per capita FDI would then be $(-428.40 \cdot 5.12 + 40.00 \cdot 5.12^2) - (-428.40 \cdot 4.12 + 40.00 \cdot 4.12^2) =$

-58.80 dollars. This is the discrete change due to a unit increase in democracy, starting at the median. The latter qualification is important because the starting point matters for polynomial regression models of order 2 and above.

7.1.3 Testing Hypotheses

In a polynomial regression, we can perform two kinds of tests that should be clearly distinguished. The first is a test for the overall significance of a predictor. This involves all of the polynomial terms for that predictor. The second is a test of the significance of a particular polynomial term; this involves only the relevant term.

Let us illustrate the differences by looking once more at the regression shown in Figure 7.1. One question we can ask is whether democracy is at all a statistically significant predictor of per capita FDI. If it is not, then neither the linear nor the quadratic term of democracy is different from zero. Thus, we formulate the following null hypothesis: $H_0 : \beta_1 = \beta_2 = 0$. We can test this hypothesis with the Wald test procedure introduced in Chapter 5. In our case, we obtain $F = 2.9$, which yields $p = 0.069$ when referred to a $\mathcal{F}[2, 41]$ distribution. Given the small sample size, I would be inclined to set the Type-I error rate to 0.10, which means that we would conclude that democracy is statistically significant.

We can now ask whether we really need the quadratic term for democracy. To answer this question, we test the null hypothesis $H_0 : \beta_2 = 0$. This is more limited then what we tested a moment ago. This hypothesis can be tested using a t-test. From the standard R regression output, we obtain $t = 2.123$ and $p = 0.040$. With a p-value this low, we reject the null hypothesis and conclude that inclusion of the quadratic democracy terms makes statistical sense.

7.1.4 Settling on an Order of the Polynomial

When one performs a polynomial regression analysis, an important question is how many polynomial terms should be fitted. In Figure 7.1, we stopped at a polynomial of the second order, but why did not we add a cubic term or terms of an even higher order? In principle, one could fit a polynomial of order $n - 1$, at least if there are no other predictors. Such a model would fit the data perfectly.

Table 7.2: Selecting the Order of the Polynomial for Democracy

Order	$\bar{R}^2$	$\Delta\bar{R}^2$	AIC^c	ΔAIC^c
1	0.003		661.487	
2	0.080	0.077	659.325	-2.162
3	0.087	0.008	660.425	1.100
4	0.078	-0.001	662.429	2.005

Notes: $\Delta\bar{R}^2$ and ΔAIC^c are the changes in the adjusted R-squared and corrected AIC of the next lower order polynomial, respectively.

In constructing a polynomial regression model, theory should be the first and foremost consideration. If our theory foresees a tipping point, then a polynomial of at least the second order should be specified. If one foresees an inflection point, then the order of the polynomial should be at least three. One can add more complexity if this makes theoretical sense and if the benefits outweigh the loss of parsimony.

That said, statisticians sometimes take a more empiricist approach and let the data speak to the order of the polynomial. This makes a lot of sense when there is a lot of data, part of which can be set aside to learn and another part to cross-validate what has been learned. Indeed, this is a very common approach in the analysis of "big data."

An empiricist approach typically starts by estimating a polynomial of order 1. In a next step, a quadratic term is added, followed by a cubic term, etc. At each juncture, some fit criterion is evaluated. Oftentimes, this is the adjusted R-squared, but one could also use AIC. As long as the fit criterion improves, we continue to add polynomial terms. Once it begins to deteriorate, we stop.

Table 7.2 illustrates the process for the level of democracy as a predictor of per capita FDI. We start with a polynomial of order 1. When we add a quadratic term, both the adjusted R-squared and AIC^c improve.[1] Adding the cubic term, we still see an improvement of the adjusted R-squared. AIC^c, on

[1] We use AIC^c because the sample size is small in this example.

the other hand, is already beginning to deteriorate. When we add a quartic term as well, then both the adjusted R-squared and AIC^c worsen. My own rule of thumb is that an increase in the polynomial is warranted if *both* the adjusted R-squared and Akaike's information criterion improve as a result. By this standard, we should settle for a polynomial of order 2.

7.2 Logarithmic Models

Logarithmic models involve logarithms on one or both sides of the population regression function. The underlying model is not linear in the parameters, but by taking logarithms it is linearized. These models come in several forms and have quite useful applications in the social sciences.

7.2.1 Log-Linear Models

The log-linear regression model, which is also known as the log-log or double-log model, is given by

$$\begin{aligned} y_i &= e^{\beta_0} x_{i1}^{\beta_1} x_{i2}^{\beta_2} \cdots x_{iK}^{\beta_K} e^{\varepsilon_i} \\ &= \alpha \prod_{k=1}^{K} x_{ik}^{\beta_k} \exp(\varepsilon_i) \end{aligned}$$

where $\alpha = \exp \beta_0$. In this form, the model is not linear in the parameters, which appear as powers. The model can be linearized, however, by taking the natural logarithm of both sides of the equation:

Equation 7.2: The Log-Linear Model

$$\ln y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{ik} + \varepsilon_i$$

This model is linear in the parameters and can be estimated using OLS. Since the linearity comes about by taking logarithms, we call this the log-linear model.

As an example, consider the regression of per capita FDI on per capita GDP. Many economists would specify this as a log-linear model:

$$\ln \text{FDI}_i = \beta_0 + \beta_1 \ln \text{GDP}_i + \varepsilon_i$$

One would choose a log-linear model in this context to transform quantities in the range from 0 to positive infinity (such as per capita GDP) into quantities that are not bounded. Specifying a log-linear model in R is extremely easy:

```
fdi.fit <- lm(log(fdipc) ~ log(gdppc), data = africa)
```

All one has to do, then, is to apply the log-function (which stands for the natural logarithm) to the left- and right-hand sides of the tilde. The result is shown in Figure 7.3.[2] We observe a monotonic relationship between per capita GDP and FDI. However, the rate of change is not constant, as the regression line is concave.

The regression of log-per capita FDI on log-per capita GDP yields a partial slope of 0.98. How do we interpret this? For the jth predictor in a log-linear model, it can be shown that

$$\beta_j = \frac{\partial y / y}{\partial x_j / x_j}$$

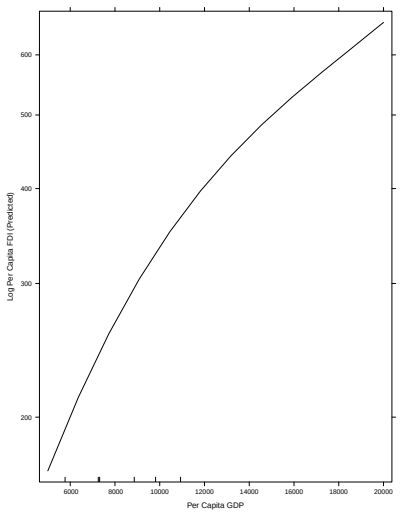
(see Appendix C.4). This is an **elasticity**: it is the percentage change that we can expect in the mean for a one percent increase in the predictor, *ceteris paribus*. In our case, we have an estimate of 0.98, which means that a one percent increase in per capita GDP is expected to increase per capita FDI by 0.98 percent. This is practically a 1-to-1 change.

7.2.2 Semi-Log Models

In the log-linear model, we wind up taking the logarithm of both the dependent and predictor variables. There also exist models in which the logarithm is applied

[2] This was created using the `effects` library. The major change compared to the regular syntax is the inclusion of `transformation=list(link=log, inverse=exp)` inside of the `Effect` command.

Figure 7.3: GDP and Foreign Direct Investment in Africa



Note: Log-linear model with $\hat{\beta}_0 = -3.22$ and $\hat{\beta}_1 = 0.98$.

to only one of these variables. These are known as semi-log models and they come in two variants: log-lin and lin-log models.

Log-Lin Models Consider the following model, which is non-linear in the parameters:

$$y_i = \exp(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

We can linearize this model by taking the natural logarithm of the dependent variable:

Equation 7.3: A Log-Lin Model

$$\ln y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Since the left-hand side is a logarithm, while the right-hand side is a linear function, we call this the log-lin model.

A major application of log-lin models is in modeling growth. In fact, if we

Table 7.3: Indian Population Data 1901-2011

Year	t	Population	Year	t	Population
1901	0	23,83,96,327	1961	6	43,92,34,771
1911	1	25,20,93,390	1971	7	54,81,59,652
1921	2	25,13,21,213	1981	8	68,33,29,097
1931	3	27,89,77,238	1991	9	84,64,21,039
1941	4	31,86,60,580	2001	10	1,02,87,37,436
1951	5	36,10,88,090	2011	11	1,21,08,54,977

Notes: Data can be found on the Indian Census.

substitute time (t) for x, then Equation 7.3 gives the exponential growth model $\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t$. The interpretation of β_1 in this model is as follows (see Appendix C.4):

$$\beta_1 \quad = \quad \frac{\partial y/y}{\partial t}$$

This gives the relative change in Y for an absolute change in t and may be interpreted as the average rate of growth.

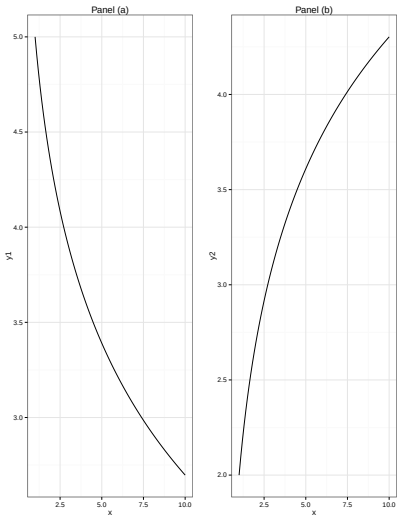
As an example, consider the data in Table 7.3; these are decennial census data showing the population of India. We estimate the following exponential growth curve model:

$$\ln \text{Pop} \quad = \quad \beta_0 + \beta_1 t + \varepsilon$$

We start at $t = 0$. Every unit increase in t corresponds to a decade in real time The OLS estimate of the intercept is 19.069, whereas the estimate of the slope is 0.159. This means that adding another decade is expected to increase the population by roughly 15.9 percent.

Lin-Log Models In the lin-log model, the left-hand side of the model is the (untransformed) dependent variable, while the right-hand side is a linear function of the logarithm of the predictor(s). For example,

Figure 7.4: Two Lin-Log Regression Functions



Note: The slope is negative in Panel (a) and positive in Panel (b).

Equation 7.4: A Lin-Log Model

$$y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$$

This model is particularly useful for modeling marginally declining rates. If $\beta_1 < 0$, then Y will decrease at a decreasing rate, producing a convex relationship (see Panel (a) of Figure 7.4). If $\beta_1 > 0$, then Y will increase at a decreasing rate, producing a concave relationship (see Panel (b) of Figure 7.4). Marginally declining utility functions may, for example, be captured using a lin-log specification.

The interpretation of β_1 is the expected absolute change in the dependent variable for a relative change in the predictor:

$$\beta_1 = \frac{\partial y}{\partial x/x}$$

(see Appendix C.4). For example, consider the model GNP $= \beta_0 + \beta_1 \ln \mathrm{M} + \varepsilon$, where M is the money supply and GNP is measured in billions of dollars. Imagine

that $\hat{\beta}_1 = 1000$. Then the interpretation is that a 1 percent increase in money supply boosts GNP by 1000/100 = 10 billion.

7.3 Reciprocal Models

Reciprocal regression models come in different varieties. The simplest model, however, is of the following form:

Equation 7.5: A Reciprocal Regression Model

$$y_i = \beta_0 + \frac{\beta_1}{x_i} + \varepsilon_i$$

The model derives its name from the fact that the reciprocal of x is used on the right-hand side. This allows us to put some curvature on the regression line, but it also forces Y to be bound (by β_0).[3] Figure 7.5 shows two examples of the reciprocal model. Panel (a) assumes $\beta_1 < 0$ and results in bounding from above, whereas panel (b) assumes $\beta_1 > 0$ so that bounding is from below. A drawback of the reciprocal model is that there is no straightforward interpretation of the elasticities.

As an example, imagine that we believe there to be a reciprocal relationship between per capita FDI and democracy on the African continent. We can model this in R by invoking the following syntax:

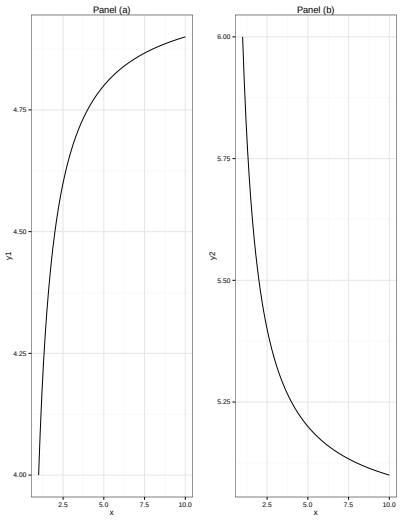
```
fdi.fit <- lm(fdipc ~ I(1/democ), data = africa)
```

You'll notice a new element in the syntax: the symbol `I`. This protects the integrity of the expression that follows in parentheses.[4] The estimates are $\hat{\beta}_0 = -128.90$ and $\hat{\beta}_1 = 1112.80$. Thus, the relationship looks like the one shown in panel (b) of Figure 7.5, with a lower asymptote of -128.90. This means that FDI tends to decrease with democracy.

[3] As $x \to \infty$, $\mu \to \beta_0$, assuming a finite β_1.

[4] Without this, R would only read the 1 in `1/democ`, which causes it to fit a constant only. That is not what we want.

Figure 7.5: Two Reciprocal Regression Functions



Note: The slope is negative in Panel (a) and positive in Panel (b).

Parenthetically, we have now investigated two different models of the relationship between democracy and per capita FDI: a polynomial and a reciprocal regression model. Which one is better? The corrected AIC for the polynomial model was 659.325. For the reciprocal model, it is 658.329. Thus, the evidence favors the reciprocal model, but only slightly: the Akaike weight for the polynomial model is 0.38, whereas it is 0.62 for the reciprocal model.

7.4 Conclusions

In this chapter, we have discussed how the linear regression model can be specified so as to capture non-linear relationships. The principle that makes this possible is that the linearity assumption in regression analysis references only the parameters, not the predictors. By transforming the predictors and, on occasion, the dependent variable, a wide variety of complex relationships can be modeled. Here, we have approached these transformations from a theoretical perspective. One can also approach this from an empirical perspective, a topic that we shall discuss in Chapter X.

Chapter 8

Factors

The multiple regression model assumes that the dependent variable is continuous. This follows from the assumption that the errors are normally distributed. The model makes no assumptions about the measurement level (or, for that matter, the distribution) of the predictors. Thus, the predictors can be continuous in nature (interval and ratio scales), but they can also be discrete (nominal and ordinal scales). From a statistical perspective, the measurement level of the predictors is inconsequential.

From a substantive perspective, however, discrete predictors or factors create problems with interpretation. To see this, consider again the literal interpretation of the partial slope coefficient: β_k is the expected change in Y for a unit change in x_k, while holding all else equal. But what does a "unit change" mean when the predictor is discrete? Since unit differences do not imply the same distance for such variables—they are used solely to establish differences or rank-orderings—it would seem strange to speak of a "unit change" in the first place.

The ambiguities of interpreting the effects of factors necessitate a special approach to incorporating them into the regression model. This approach hinges on modeling effects as shifts in the intercept. The vehicle for creating those shifts is to create a series of **dummy variables**, i.e., 0-1 or Boolean variables.

8.1 Factors With Two Levels

8.1.1 Specification

The simplest case of a factor arises when it has two levels, i.e., two distinctive values. To illustrate this scenario, consider once more the African data on per capita FDI. We want to predict this variable based on two variables: per capita GDP and regime status, i.e., whether the country is a democracy or not.[1] The latter variable is a factor with two levels: Non-Democratic and Democratic. Our strategy is to transform this variable into the following dummy:

$$D = \begin{cases} 1 & \text{if a democracy} \\ 0 & \text{if a non-democracy} \end{cases}$$

The category that receives the value of 0 is known as the *baseline category*. This serves as the reference point for the other category. We now estimate the following model

$$\text{FDI}_i = \beta_0 + \beta_1 D_i + \beta_2 \text{GDP}_i + \varepsilon_i$$

Thus, instead of entering democratic status directly into the regression model, we enter the dummy variable. Specified in this way, the model is known as the **analysis of covariance** model. It can be estimated in the usual manner, using OLS, maximum likelihood, or method of moments.

8.1.2 Interpretation

How do we interpret this model? The easiest approach is to write out the conditional expectation function for democracies and non-democracies. The conditional expectation function is

$$\mu_i = \beta_0 + \beta_1 D_i + \beta_2 \text{GDP}_i$$

[1] The distinction is based on the Economist Intelligence Unit's cutoff of 6.00 on their democracy score.

Table 8.1: A Regression with a 2-Level Factor

Group	CEF
Non-Democracies	$\mu_i = \beta_0 + \beta_2 \text{GDP}_i$
Democracies	$\mu_i = (\beta_0 + \beta_1) + \beta_2 \text{GDP}_i$
Difference	β_1

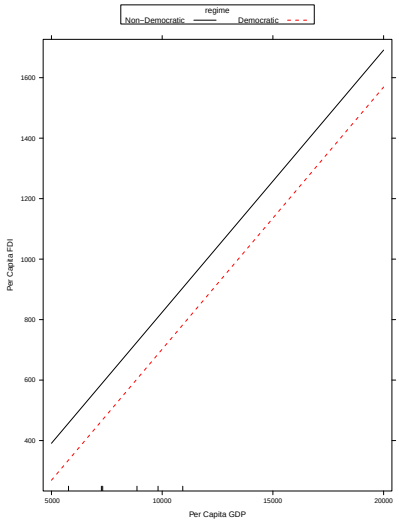
Notes: CEF=conditional expectation function.

We can apply this equation to non-democracies by substituting 0 for D. This yields $\mu_i = \beta_0 + \beta_1 \cdot 0 + \beta_2 \text{GDP}_i = \beta_0 + \beta_2 \text{GDP}_i$. We can apply the conditional expectation function to democracies by substituting 1 for D: $\mu_i = \beta_0 + \beta_1 \cdot 1 + \beta_2 \text{GDP}_i = (\beta_0 + \beta_1) + \beta_2 \text{GDP}_i$ (see Table 8.1). We see that the difference in the population mean for a democracy and a non-democracy with the same per capita GDP is equal to β_1. This is a shift in the intercept. For non-democracies, the intercept is β_0; for democracies it is $\beta_0 + \beta_1$. The difference between these intercepts is the democracy effect. If $\beta_1 < 0$, then democracies are expected to have a lower per capita FDI than non-democracies, all else equal. If $\beta_1 > 0$, then democracies are expected to have a higher per capita FDI than non-democracies, again all else equal. Finally, $\beta_1 = 0$ means that the expected per capita FDI is identical for democracies and non-democracies, *ceteris paribus*.

The example illustrates why we call $D = 0$ the baseline category. This category is absorbed into the intercept, which is commonly interpreted as the baseline of the regression. The example also shows why we consider the effect of factors in terms of shifts in the intercept. After all, β_1 shifts where the regression line crosses the y-axis.

Graphically, a shift in intercept implies that we obtain two parallel regression lines (Figure 8.1). The lines are parallel since the slope coefficients associated with per capita GDP are identical for democracies and non-democracies. The distance between the lines depends on the size of the coefficient associated with D. In our example, this is estimated at about -122.22 US dollars.

Figure 8.1: Regime Status, GDP, and FDI in Africa



Note: The intercept shift is $-122.22 when comparing democracies to non-democracies.

8.1.3 Implementation in R

R makes it extremely easy to incorporate factors into regression models. As long as a predictor is declared as a factor, the creation of the dummy variable is automatic.[2] Here, R treats the first level as the baseline. All the user now has to do is to provide the usual regression syntax. For example,

```
lm(y ~ f+x, data=object)
```

runs a regression of `y` on the factor `f` and the covariate `x` using the data in `object`.

The sample output is shown in Figure 8.2. The entry for `(Intercept)` corresponds to $\hat{\beta}_0$ in our earlier specification

$$\mu_i = \beta_0 + \beta_1 D_i + \beta_2 \text{GDP}_i$$

[2]To check if the variable `x` is a factor, you can simply type `is.factor(x)`. If the command returns `TRUE`, then the variable is a factor.

Figure 8.2: R Output With a Factor With Two Levels

```
Call:
lm(formula = fdipc ~ regime + gdppc, data = africa)

Residuals:
    Min      1Q  Median      3Q     Max
-736.25  -54.53   11.99   98.69  837.05

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.230e+01  4.774e+01  -0.886    0.381
regimeDemocratic -1.222e+02  8.515e+01  -1.435    0.159
gdppc             8.668e-02  9.156e-03   9.466 7.16e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 242.7 on 41 degrees of freedom
Multiple R-squared:  0.6863,    Adjusted R-squared:  0.6709
F-statistic: 44.84 on 2 and 41 DF,  p-value: 4.785e-11
```

Note: The entry `regimeDemocratic` corresponds to D in the regression specification shown earlier.

The entry for `regimeDemocratic` corresponds to $\hat{\beta}_2$. Finally, the entry for gdppc corresponds to $\hat{\beta}_2$.

When reporting regression results with factors, it is best to avoid cryptic names such as D or regimeDemocratic. There exist several methods of labeling dummy variables in regression tables:

- Use the name of the non-baseline category, e.g., democracy or democratic regime.
- Use the name of the variable and indicate in parentheses what a value of 1 means. For example, regime type (1=democracy).

The second method is illustrated in Table 8.2; the first method will be illustrated later in this chapter.

8.1.4 Hypothesis Testing

Testing for Group Differences In our example, we would like to know whether the difference in intercepts between democratic and non-democratic regimes is statistically significant. Under the null hypothesis $H_0 : \beta_1 = 0$, the difference between the two regime types vanishes: the intercept is now β_1 for both democracies and non-democracies (take a look once more at Table 8.1). Thus, we

Table 8.2: Reporting Factors in Published Research I

	Dependent variable:
	Per Capita FDI
Regime Type (1=Democracy)	−122.22
	(85.15)
Per Capita GDP	0.09***
	(0.01)
Constant	−42.30
	(47.74)
Observations	44
Adjusted R^2	0.67
Note:	*p<0.1; **p<0.05; ***p<0.01

can establish whether statistically significant differences in per capita FDI exist by testing $\beta_1 = 0$.

The decision regarding the null hypothesis can be derived directly from the output shown in Figure 8.2. Here, we see that the t-value for D is -1.435. The associated p-value is 0.159 and exceeds any conventional Type-I error rate. Thus, we fail to reject the null hypothesis and conclude that no statistically significant differences exist between democracies and non-democracies in terms of the expected level of FDI.

With an eye on factors with multiple levels, we could also have performed an F-test, in the manner described in Chapter 5 and involving only `regimeDemocratic`. This yields $F = 2.1$ and $p = 0.16$, so that we again fail to reject the null hypothesis. Note that $F = t^2$, which is always true when we test a single parameter.

Testing Intercepts A second question we would like to answer is whether the intercept is significantly different from 0 at a particular level of the factor. In our case, we may want to know if the intercept is significantly different from 0 for non-democracies. We may also want to know if is statistically significant for

democracies.

To test the significance of the intercept for non-democracies, our baseline category, is simple enough. All we need to do is to assess whether we can reject $H_0 : \beta_0 = 0$. After all, β_0 is the intercept for non-democracies (see Table 8.1). Decisions about this null hypothesis can be based directly on the R output from Figure 8.2. We observe that the t-test statistic associated with the intercept is -0.886. The associated p-value is 0.381, which is way too large to reject the null hypothesis. We conclude that the intercept for non-democracies is not statistically significant at conventional levels, i.e., using conventional Type-I error rates.

Testing the significance of the intercept for democracies is considerably more complicated. The intercept here is equal to $\beta_0 + \beta_1$. We say that the intercept is statistically significant if we can reject $H_0 : \beta_0 + \beta_1 = 0$. We can test this hypothesis using a t-test:

$$t \quad = \quad \frac{\hat{\beta}_0 + \hat{\beta}_1}{\widehat{SE}[\hat{\beta}_0 + \hat{\beta}_1]}$$

Here,

$$\widehat{SE}[\hat{\beta}_0 + \hat{\beta}_1] \quad = \quad \sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_0) + \widehat{\mathsf{Var}}(\hat{\beta}_1 + 2\widehat{\mathsf{Cov}}(\hat{\beta}_0, \hat{\beta}_1)}$$

The ingredients for this test statistic are only partially found in the R output. Computing the test statistic in this way can thus be quite cumbersome.

Equivalently, we can use the Wald test procedure outlined in Chapter 5.7. Our null hypothesis may be written in the form of

$$\mathbf{R}\boldsymbol{\beta} \quad = \quad \mathbf{r}$$

Specifically, let $\mathbf{R} = (1\ 1\ 0)$ and $\mathbf{r} = 0$, then

$$\begin{pmatrix} 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad = \quad 0$$

Figure 8.3: Testing the Significance of the Intercept in Democracies

```
         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = fdipc ~ regime + gdppc, data = africa)

Linear Hypotheses:
      Estimate Std. Error t value Pr(>|t|)
1 == 0  -164.52      80.17  -2.052   0.0466 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Note: Based on the regression results shown in Figure 8.2.

is identical to $\beta_0+\beta_1 = 0$. We can now use equation 5.11 to test this constraint.

The second approach is actually the easier one to implement in R. To do this one needs the `multcomp` library:

```
library(multcomp)
R <- matrix(c(1, 1, 0), nrow=1)
summary(glht(dem.fit, linfct = R))
```

Here `dem.fit` is the regression object. The results are shown in Figure 8.3. The column labeled "estimate" shows $\hat{\beta}_0 + \hat{\beta}_1$, which is -164.52. The column labeled "Std. Error" shows $\widehat{SE}[\hat{\beta}_0 + \hat{\beta}_1]$, which in our case is 80.17. The ratio of these two quantities is the t-statistic, which is -2.052. The associated p-value is 0.047. Thus, we would conclude that the intercept for democracies is statistically significant at the .05-level: using a Type-I error rate of .05, we have to reject the null hypothesis $\beta_0 + \beta_1 = 0$.

The result is a bit surprising. We showed that the intercept for non-democracies is not significant. We also showed there is a non-significant difference between the intercepts of democracies and non-democracies. Yet, the intercept for democracies is statistically significant. This can happen on occasion and has to do with the fact that we combine the estimates of two parameters, which are not independent from each other.

Choosing the Baseline: Does It Matter?

At this point, you may wonder how much the results depend on which level of the factor was designated as the baseline. We chose non-democracies, but what would have happened had we chosen democracies? In this case, the sample regression function would have been

$$\widehat{FDI}_i = -164.52 + 122.22 \cdot \text{Non-Democracy}_i + 0.09 \cdot \text{GDP}_i$$

When we use non-democracies as the baseline, the sample regression function is

$$\widehat{FDI}_i = -42.30 - 122.22 \cdot \text{Democracy}_i + 0.09 \cdot \text{GDP}_i$$

Let us compare these two equations. First, which level of regime type is designated the baseline has no effect whatsoever on the estimate for per capita GDP. Second, the magnitude of the coefficient associated with the non-baseline category is the same across the two equations. The difference in the intercepts for democracies and non-democracies is 122.22 in absolute value, regardless of which level is used as the baseline. Third, the intercept changes across the two setups. When democracies are the baseline, then the intercept is -164.52; when non-democracies are the baseline, then the intercept is -42.30. This should be the case because the intercept absorbs the baseline, which is different in the two setups. However, if we reconstruct the intercept for the two regime types it is identical across the two sample regression functions: -164.52 for democracies and -42.30 for non-democracies. In sum, which level is designated the baseline is without consequence.

8.2 Factors With More Than Two Levels

8.2.1 Specification

How do we proceed when a factor has more than two levels? In this case, we create multiple dummy variables. Specifically, if the original factor has M levels, then we create $M - 1$ dummy variables. The approach can be summarized in

three steps.

1. *Designate a Baseline:* This is the level to which the other levels of the factor will be compared. The effect of the baseline is absorbed into the intercept. Which level is designated as the baseline is arbitrary. R uses the first level for this purpose.

2. *Generate Dummies for the Remaining Levels:* Since the baseline already has a parameter associated with it, namely β_0, we should not introduce a separate dummy for this level. The remaining $M - 1$ levels, however, require their own dummy variables.

3. *Estimate the Model:* Estimate a regression model containing the $M - 1$ dummies and any covariates one wishes to include.

Note that R automates these steps for us as long as the variable of interest has been declared as a factor.

As an example, consider again FDI in Africa. Certain organizations have divided the African continent in investment regions. RisCura, for example, employs the classification scheme shown in Table 8.3. This is based on the investment risks and opportunities in different parts of Africa. We now wish to include the region variable in our regression model along with GDP.[3] Designating Central Africa as the baseline, we would estimate the following model:

$$\begin{aligned} \text{FDI}_i &= \beta_0 + \beta_1 \text{EA}_i + \beta_2 \text{ES}_i + \beta_3 \text{FWA}_i + \beta_4 \text{M}_i + \beta_5 \text{N}_i + \\ &\quad \beta_6 \text{OWA}_i + \beta_7 \text{SA1}_i + \beta_8 \text{SA2}_i + \beta_9 \text{GDP}_i + \varepsilon_i \end{aligned}$$

The terms EA, ES, FWA, M, N, OWA, SA1, and SA2 are all dummy variables. They take on the value 1 if a particular country resides in a particular region and 0 otherwise. For example, Burundi is part of East Africa. For Burundi, then, EA is 1, whereas ES, FWA, M, N, OWA, SA1, and SA2 are all 0, since it does not belong to those regions. As another example, Sierra Leone is part of other West Africa. Hence it scores 1 on OWA and 0 on EA, ES, FWA, M,

[3] Parenthetically, we run the risk of over-fitting here because GDP may already have informed the regional divisions in Table 8.3.

Table 8.3: African Investment Regions

Code	Region	Abbreviation	Countries
1	Central Africa	CA	Cameroon, Central African Republic, Chad, Democratic Republic of the Congo, Equatorial Guinea, Gabon, Republic of the Congo
2	East Africa	EA	Burundi, Ethiopia, Kenya, Rwanda, Tanzania, Uganda
3	Egypt & Sudan	ES	Egypt, Sudan
4	Francophone West Africa	FWA	Benin, Burkina Faso, Cape Verde, Guinea, Ivory Coast, Mali, Niger, Senegal
5	Maghreb	M	Algeria, Mauritania, Morocco, Tunisia
6	Nigeria	N	Nigeria
7	Other West Africa	OWA	Gambia, Ghana, Liberia, Sierra Leone
8	South Africa	SA1	Lesotho, South Africa, Swaziland
9	Southern Africa	SA2	Botswana, Comoros, Madagascar, Malawi, Mauritius, Mozambique, Namibia, Zambia, Zimbabwe

Notes: Based on RisCura.

N, SA1, and SA2. A country situated in the baseline group, i.e., East Africa, scores 0 on all of the dummies.

The model can be estimated using OLS. In R one only has to include region and GDP as predictors of FDI. As long as region has been declared as a factor, R will set up the model as we have specified it here. The estimation results can be found in Figure 8.4. Later, we shall discuss how this can be presented more neatly in tabular form.

8.2.2 Why Cannot We Include M Dummies?

One question has not been answered until now. Why is it that we include one fewer dummy variables than there are levels of the factor variable? Put

Figure 8.4: R Output With a Factor With Multiple Levels

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     76.00177  111.11341   0.684   0.4986
regionEast Africa             -112.33080  148.26148  -0.758   0.4539
regionEgypt & Sudan           -238.52806  202.49007  -1.178   0.2470
regionFrancophone West Africa -125.45070  137.36743  -0.913   0.3675
regionMaghreb                 -292.10930  156.62201  -1.865   0.0708 .
regionNigeria                 -267.48138  268.06623  -0.998   0.3254
regionOther West Africa        -61.51045  164.19640  -0.375   0.7103
regionSouth Africa            -343.47216  172.88339  -1.987   0.0551 .
regionSouthern Africa         -166.91786  128.67329  -1.297   0.2033
gdppc                            0.08516    0.01041   8.182 1.52e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: The entry `regionEast Africa` corresponds to EA in the regression specification shown earlier. Likewise, `regionEgypt & Sudan` corresponds to ES, etc.

differently, why cannot we include M dummy variables?

The answer lies in Assumption 4.1, namely that the matrix of predictors $\mathbf{X}$ has to be full rank. This cannot be true if we include both a constant and M dummy variables. To see this consider the following fragment of $\mathbf{X}$, which shows data from nine countries from an equal number of regions:

$$\mathbf{X} = \begin{bmatrix} \underline{Const} & \underline{CA} & \underline{EA} & \underline{ES} & \underline{FWA} & \underline{M} & \underline{N} & \underline{OWA} & \underline{SA1} & \underline{SA2} & \underline{GDP} \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1219.93 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 251.01 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 3256.02 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 750.51 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 9813.92 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2742.22 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 509.39 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1134.85 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 7254.56 \end{bmatrix}$$

Here Const is the constant and CA is a newly formed dummy that takes on the value 1 for Central African countries and 0 elsewhere. If we now consider the first 10 columns, we see that

$$Const = CA + EA + ES + FWA + M + N + OWA + SA1 + SA2$$

We have a linear dependency between the constant and the 9 regional dummies we have created, i.e., we have perfect multicollinearity. The existence of such collinearity means that $\mathbf{X}$ cannot be full-rank.

To solve the problems we can pursue two strategies. Either we drop the

constant and create M dummy variables or we drop one of the dummies, for example, CA. In practice, the second strategy is far more common. This is the one we shall pursue throughout this book.

8.2.3 Interpretation

How do we interpret the results from Figure 8.4? It is easiest to do this by developing scenarios for each of the regions. Take for example Cameroon, which is located in Central Africa. For this country, we may substitute 0s for all of the dummy variables, so that the conditional expectation function is:

$$\begin{aligned}
\mu_i &= \beta_0 + \beta_1 \mathsf{EA}_i + \beta_2 \mathsf{ES}_i + \beta_3 \mathsf{FWA}_i + \beta_4 \mathsf{M}_i + \beta_5 \mathsf{N}_i + \\
&\quad \beta_6 \mathsf{OWA}_i + \beta_7 \mathsf{SA1}_i + \beta_8 \mathsf{SA2}_i + \beta_9 \mathsf{GDP}_i \\
&= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0 + \beta_5 \cdot 0 + \\
&\quad \beta_6 \cdot 0 + \beta_7 \cdot 0 + \beta_8 \cdot 0 + \beta_9 \mathsf{GDP}_i \\
&= \beta_0 + \beta_9 \mathsf{GDP}_i
\end{aligned}$$

Now consider, for example, Lesotho. This country is located in the South Africa region so that SA1 equals 1 and all of the other dummy variables are 0. Substitution in the conditional expectation function now yields:

$$\begin{aligned}
\mu_i &= \beta_0 + \beta_1 \mathsf{EA}_i + \beta_2 \mathsf{ES}_i + \beta_3 \mathsf{FWA}_i + \beta_4 \mathsf{M}_i + \beta_5 \mathsf{N}_i + \\
&\quad \beta_6 \mathsf{OWA}_i + \beta_7 \mathsf{SA1}_i + \beta_8 \mathsf{SA2}_i + \beta_9 \mathsf{GDP}_i \\
&= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \beta_4 \cdot 0 + \beta_5 \cdot 0 + \\
&\quad \beta_6 \cdot 0 + \beta_7 \cdot 1 + \beta_8 \cdot 0 + \beta_9 \mathsf{GDP}_i \\
&= \beta_0 + \beta_7 + \beta_9 \mathsf{GDP}_i
\end{aligned}$$

If we repeat this process for all of the regions, then we obtain the results shown in Table 8.4. We observe again that the effect of per capita GDP remains constant across the regions. The intercepts, however, move around. For example, the estimated intercept for Central Africa is 76.00, whereas it is -267.47 in South Africa. The shifts in intercepts are visualized in Figure 8.5.

Table 8.4: A Regression with a M-Level Factor

Country is (in)	CEF	SRF
Central Africa	$\mu_i = \beta_0 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 + 0.09\text{GDP}_i$
East Africa	$\mu_i = \beta_0 + \beta_1 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 112.33 + 0.09\text{GDP}_i = -36.33 + 0.09\text{GDP}_i$
Egypt/Sudan	$\mu_i = \beta_0 + \beta_2 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 238.53 + 0.09\text{GDP}_i = -162.53 + 0.09\text{GDP}_i$
Francophone West Africa	$\mu_i = \beta_0 + \beta_3 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 125.45 + 0.09\text{GDP}_i = -49.45 + 0.09\text{GDP}_i$
Maghreb	$\mu_i = \beta_0 + \beta_4 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 292.11 + 0.09\text{GDP}_i = -216.11 + 0.09\text{GDP}_i$
Nigeria	$\mu_i = \beta_0 + \beta_5 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 267.48 + 0.09\text{GDP}_i = -191.48 + 0.09\text{GDP}_i$
Other West Africa	$\mu_i = \beta_0 + \beta_6 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 61.51 + 0.09\text{GDP}_i = 14.49 + 0.09\text{GDP}_i$
South Africa	$\mu_i = \beta_0 + \beta_7 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 343.47 + 0.09\text{GDP}_i = -267.47\text{GDP}_i$
Southern Africa	$\mu_i = \beta_0 + \beta_8 + \beta_9 \text{GDP}_i$	$\widehat{\text{FDI}}_i = 76.00 - 166.92 + 0.09\text{GDP}_i = -90.92 + 0.09\text{GDP}_i$

Notes: CEF=conditional expectation function; SRF=sample regression function (based on the estimates in Figure 8.4).

When we inspect Table 8.4, then we observe the following:

Coefficient	Intercept difference between
β_1	East and Central Africa
β_2	Egypt/Sudan and Central Africa
β_3	Francophone West and Central Africa
β_4	Maghreb and Central Africa
β_5	Nigeria and Central Africa
β_6	Other Western and Central Africa
β_7	South and Central Africa
β_8	Southern and Central Africa

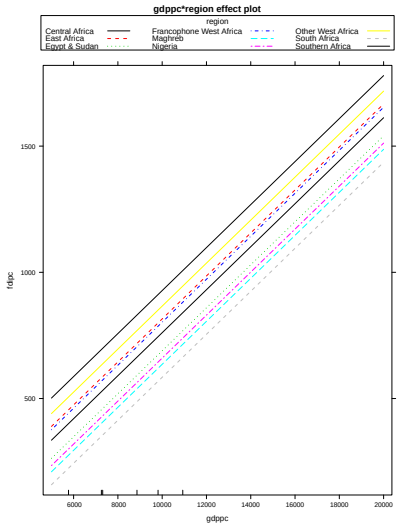
Thus, the coefficients β_1 through β_8 measure differences in the intercept between non-baseline levels and the baseline. In this light, the estimates from Figure 8.4 suggest that, for example, the intercept in South African countries is -343.47 points (or dollars) lower than the intercept for Central African countries. Put differently, holding per capita GDP constant, we predict per capita FDI to be 343.47 dollars less in Southern than in Central African countries.

Less obvious from Table 8.4 are the differences between non-baseline levels. However, they can be easily derived from the table. Imagine, for example, that we wish to compare the intercepts for countries in South and Southern Africa. From table 8.4, we know that the conditional expectation function for South African countries is $\mu = \beta_0 + \beta_7 + \beta_9 \text{GDP}$. For Southern African countries, this equation is $\mu = \beta_0 + \beta_8 + \beta_9 \text{GDP}$. If we now assume that per capita GDP is the same across countries from each region, then

$$\begin{aligned} \Delta\mu &= \mu|\text{SA2} - \mu|\text{SA1} \\ &= (\beta_0 + \beta_8 + \beta_9 \text{GDP}) - \\ &\quad (\beta_0 + \beta_7 + \beta_9 \text{GDP}) \\ &= \beta_8 - \beta_7 \end{aligned}$$

where $\mu|$SA2 is the conditional expectation function given that the country is located in Southern Africa and a similar meaning adheres to $\mu|$SA1. The

Figure 8.5: Region, GDP, and FDI in Africa



Note: Notice the intercept shifts across the regions.

quantity $\beta_8 - \beta_7$ is is a difference in intercepts. When GDP is assumed to remain constant, it can be interpreted as the discrepancy in the FDI expect between Southern and South African countries. Based on the results from Figure 8.4, our estimate of this discrepancy is $-166.92 - (-343.47)) = 176.55$: holding GDP constant, we expect per capita FDI to be 176.55 dollars higher in Southern than in South African countries. Obviously, other differences can be derived and interpreted analogously.

8.2.4 Hypothesis Testing

Testing for Group Differences Across the Board With M-level factors, the very first hypothesis to test is the null hypothesis that there are no differences across any of the levels. The null hypothesis implies that the coefficients associated with the $M - 1$ dummy variables are all equal to 0. In this case, there is a common intercept that applies to all levels of the factor.

In our example, we can formulate $H_0 : \beta_1 = \beta_2 = \beta 3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. Under this hypothesis, the conditional expectation function

for each and every region of Africa reduces to $\mu_i = \beta_0 + \beta_9 \text{GDP}_i$. It is easily verified this is the case by substituting the hypothesized values of the coefficients associated with the dummies in the conditional expectation functions shown in the second column of Table 8.4. We can test the null hypothesis with the F-test approach outlined in Chapter 5. When we use the `wald.test` or `waldtest` function in R we obtain the following result: $F = 0.86$ and $p = 0.56$ (based on the $\mathcal{F}[8, 34]$ distribution). The conclusion is clear: the null hypothesis cannot be rejected and we may proceed with a model with a single intercept for all African regions.

Normally speaking, we would end our explorations of regional differences in Africa at this point, at least with the current operationalization of regions. For the sake of completeness, we shall proceed with the next step, which is to look at more specific differences.

Testing for Specific Group Differences Imagine we could have rejected the null hypothesis. This does not mean that all intercepts are different from each other. We know, however, that at least some are different from each other in the population. If we want to explore which ones, then we need to look into comparisons between specific levels of the factor.

Comparisons with the baseline are the easiest. Imagine that we want to know if the intercept in South Africa is different from Central Africa. This amounts to testing $H_0 : \beta_7 = 0$. If we fail to reject this hypothesis, then the conditional expectation functions for the two regions are indistinguishable (see once more Table 8.4). All the information for this hypothesis test is provided in the R output shown in Figure 8.4. If we go to the entry `regionSouth Africa`, which corresponds to β_7, we find $t = -1.987$ and $p = 0.055$. Assuming a Type-I error rate of 0.10, we would be inclined to reject the null hypothesis and say that the intercept in the region of South Africa is different than that of Central Africa. Momentarily, we shall revisit this conclusion, but for now it is important to understand simply how we derived the relevant information from the R output.

Comparisons between two non-baseline values are more complicated. Let us revisit the comparison between the Southern and South African regions. Earlier,

we saw that the difference in the intercepts is equal to $\beta_8 - \beta_7$. If we want to say that this difference is statistically significant at some level, then we need to be able to reject $H_0 : \beta_8 - \beta_7 = 0$ at that level. If we fail to reject the null hypothesis, then we have to assume that $\beta_8 - \beta_7 = 0$, which is the same as saying $\beta_7 = \beta_8$, i.e., the two intercepts are the same.

We can test this hypothesis using the `multcomp` library. We once more use the linear equation $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. In this case, we define $\mathbf{r} = 0$ and

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{pmatrix}$$

The R syntax is now

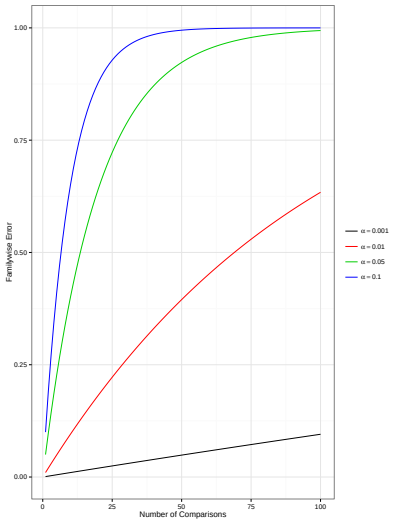
```
library(multcomp)
R <- matrix(c(0, 0, 0, 0, 0, 0, 0, -1, 1, 0), nrow=1)
summary(glht(region.fit, linfct = R))
```

Here `region.fit` is the regression object that we created. We find an estimate of $\beta_8 - \beta_7$ of 176.6 with a standard error of 166.4. This yields $t = 1.061$ and $p = 0.296$. Thus, we fail to reject the null hypothesis and conclude that the intercepts for the two regions are indistinguishable in the population.

The `multcomp` library offers an easy way of creating all possible comparisons between African regions. Before we embark on this subject, however, we need to revisit the Type-I error rate when we conduct many different tests. With 9 regions, it is possible to perform $.5 \cdot 9 \cdot (9 - 1) = 36$ comparisons.[4] Each involves a null hypothesis that we test. The problem with testing this many null hypotheses is that the actual Type-I error will exceed the nominal Type-I error, α, sometimes by a lot. The nominal Type-I error rate is what we think we have set, say 0.10 or 0.05. The actual, or what is known as the *familywise*, Type-I error rate refers to a set (family) of inferences. It is defined as the probability of making at least one Type-I error in the family. This probability is given by $1-(1-\alpha)^q$, where q is the number of inferences in the family. Figure 8.6 shows the familywise error rate for different values of α and q. The take home message

[4] In general, with M factor levels we can make $.5M(M-1)$ comparisons.

Figure 8.6: Familywise Error Rates



Note: The vertical axis shows the familywise error rate, whereas the horizontal axis shows the number of comparisons. Especially at higher nominal Type-I error rates, the familywise error exceeds 0.25 or even 0.50 very quickly.

of this figure is that, with many comparisons, it becomes quickly improbable not to reject null hypotheses incorrectly and to conclude erroneously that a significant difference in intercepts exists.

Most statisticians would argue that we need to take some precaution to avoid ridiculously high familywise error rates. There is no real consensus on the nature of this precaution and it would lead us too far astray to discuss all of the many possibilities. I discuss here the Holm-Bonferroni procedure, which is a more powerful variant of the Bonferroni adjustment (Holm, 1979). The Bonferroni adjustment simply consists of dividing the nominal Type-I error rate by the number of comparisons, i.e., α/q. This is often too conservative in that it fails to reject the null hypothesis even when it is false. In Holm-Bonferroni, we apply the adjustment sequentially, starting with the comparison whose p-value is the smallest. This can be shown to improve the statistical power of the test considerably.

If this all sounds extremely complicated, do not despair. R's `multcomp`

library will automate the process for us. The syntax is

```
library(multcomp)
region.ht <- glht(region.fit,
linfct=mcp(region="Tukey"))
summary(region.ht, test=adjusted("holm"))
```

The second and third lines invoke the multiple comparison procedure (`mcp`) for the variable region. The option Tukey means that we seek to make all possible comparisons between intercepts.[5] The fourth line results in the adjustment of the p-values according to the Holm-Bonferroni procedure.

The results are shown in Figure 8.7. Looking at the last column, we observe that the adjusted p-values are uniformly 1 for all of the 36 comparisons we are making. This means that not a single pair of regions displays a statistically significant difference in the intercept, a finding that is entirely consistent with the F-test shown earlier. Also note that the standard R regression output, shown in Figure 8.4, shows a limited number of comparisons, to wit those involving the baseline. The p-values reported in this output have not been adjusted. Consequently, I would not rely on these too much and prefer using p-values that have been corrected for multiple comparisons.

Testing Intercepts One topic in the discussion of regression results remains: hypothesis tests of region-specific intercepts. This is easiest for the baseline. If we want to know if the intercept is statistically significant for Central Africa, then all we need to do is to consult the regression output from Figure 8.4. We see that the t-statistic for the intercept is quite small, at 0.684. With a p-value of 0.499 we fail to reject $H_0 : \beta_0 = 0$; the intercept in Central Africa is not significantly different from 0.

Since we have already shown that the intercepts of the remaining regions do not differ significantly from the intercept in Central Africa, we could end the process of testing intercepts here. If we want to continue, then we proceed

[5] Technically speaking, the option means that we compare all means. This is actually what we do when we compare the conditional expectation functions, as long as we assume that per capita GDP is always the same.

Figure 8.7: Multiple Comparisons Across African Regions

	Estimate	Std. Error	t value	Pr(>|t|)
East Africa - Central Africa == 0	-112.33	148.26	-0.758	1
Egypt & Sudan - Central Africa == 0	-238.53	202.49	-1.178	1
Francophone West Africa - Central Africa == 0	-125.45	137.37	-0.913	1
Maghreb - Central Africa == 0	-292.11	156.62	-1.865	1
Nigeria - Central Africa == 0	-267.48	268.07	-0.998	1
Other West Africa - Central Africa == 0	-61.51	164.20	-0.375	1
South Africa - Central Africa == 0	-343.47	172.88	-1.987	1
Southern Africa - Central Africa == 0	-166.92	128.67	-1.297	1
Egypt & Sudan - East Africa == 0	-126.20	204.34	-0.618	1
Francophone West Africa - East Africa == 0	-13.12	134.65	-0.097	1
Maghreb - East Africa == 0	-179.78	165.83	-1.084	1
Nigeria - East Africa == 0	-155.15	270.01	-0.575	1
Other West Africa - East Africa == 0	50.82	160.82	0.316	1
South Africa - East Africa == 0	-231.14	179.48	-1.288	1
Southern Africa - East Africa == 0	-54.59	133.56	-0.409	1
Francophone West Africa - Egypt & Sudan == 0	113.08	197.46	0.573	1
Maghreb - Egypt & Sudan == 0	-53.58	216.78	-0.247	1
Nigeria - Egypt & Sudan == 0	-28.95	305.13	-0.095	1
Other West Africa - Egypt & Sudan == 0	177.02	216.47	0.818	1
South Africa - Egypt & Sudan == 0	-104.94	227.91	-0.460	1
Southern Africa - Egypt & Sudan == 0	71.61	194.82	0.368	1
Maghreb - Francophone West Africa == 0	-166.66	156.58	-1.064	1
Nigeria - Francophone West Africa == 0	-142.03	264.79	-0.536	1
Other West Africa - Francophone West Africa == 0	63.94	152.60	0.419	1
South Africa - Francophone West Africa == 0	-218.02	171.16	-1.274	1
Southern Africa - Francophone West Africa == 0	-41.47	122.57	-0.338	1
Nigeria - Maghreb == 0	24.63	279.14	0.088	1
Other West Africa - Maghreb == 0	230.60	180.35	1.279	1
South Africa - Maghreb == 0	-51.36	190.38	-0.270	1
Southern Africa - Maghreb == 0	125.19	150.57	0.831	1
Other West Africa - Nigeria == 0	205.97	279.28	0.738	1
South Africa - Nigeria == 0	-75.99	287.93	-0.264	1
Southern Africa - Nigeria == 0	100.56	262.62	0.383	1
South Africa - Other West Africa == 0	-281.96	193.03	-1.461	1
Southern Africa - Other West Africa == 0	-105.41	151.40	-0.696	1
Southern Africa - South Africa == 0	176.55	166.38	1.061	1

(Adjusted p values reported -- holm method)

Note: The p-values have been adjusted using Holm-Bonferroni.

similarly to the two-level case. For example, if we want to look at the intercept in Southern Africa, then Table 8.4 tells us that this is $\beta_0 + \beta_8$. The intercept is significantly different from 0 if we can reject $H_0 : \beta_0 + \beta_8 = 0$. Relying once more on the equation $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, we can set $\mathbf{r} = 0$ and $\mathbf{R} = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0)$; this yields $\beta_0 + \beta_8 = 0$. We can now use R's `multcomp` library:

```
library(multcomp)
R <- matrix(c(1, 0, 0, 0, 0, 0, 0, 0, 1, 0), nrow=1)
summary(glht(region.fit, linfct=R))
```

We obtain an estimate of -90.92 (cf. Table 8.4) with an estimated standard error of 88.57. The t-statistic is -1.026 and yields $p = 0.312$. Hence, we fail to reject the null hypothesis and conclude that the intercept for Southern Africa is not reliably different from 0.

8.2.5 Reporting Regressions with Factors

In this section, we have discussed a large many topics. How would one normally report the many statistical results and tests that have made their appearance? Let us start with the regression output. In Table 8.2, I showed one way of formatting this output. Now let us consider a second way, which is shown in Table 8.5. Here the factor levels are indicated by their proper names, with the exception of the baseline, which is captured through the constant. Any regression analysis involving a M-level factor should at least show the regression estimates so that a table like Table 8.5 is indispensable.

What should also be reported is the F-test across the factor levels. One could do this in the note below the table, but I would personally opt for inclusion in the text so that it does not elude the reader. For example, I might write the following:

Table 8.5: Reporting Factors in Published Research II

	Dependent variable:
	Per Capita FDI
East Africa	−112.33
	(148.26)
Egypt and Sudan	−238.53
	(202.49)
Francophone West Africa	−125.45
	(137.37)
Maghreb	−292.11*
	(156.62)
Nigeria	−267.48
	(268.07)
Other West Africa	−61.51
	(164.20)
South Africa	−343.47*
	(172.88)
Southern Africa	−166.92
	(128.67)
Per Capita GDP	0.09***
	(0.01)
Constant	76.00
	(111.11)
Observations	44
Adjusted R^2	0.65
Note:	*p<0.1; **p<0.05; ***p<0.01

> We fail to reject the null hypothesis that there are no differences across the regions after controlling for per capita GDP: $F[8, 34] = 0.86$, *ns*.

Here, the symbol *ns* stands for "not significant." Note that it is important to report the degrees of freedom associated with the F-test statistic. Had the F-test been significant then we would have indicated this in lieu of *ns*. For example, one might write $p < .05$ if the p-value is less than .05; this tells the reader that the test is significant at the .05-level (and hence also at the .10-level, but not at the .01-level).

Other results that we have discussed should be discussed only in as far as they are important for the theoretical argument. If specific regional differences are important, for example, one can discuss those in the text. For example, if the difference between Nigeria and Egypt/Sudan is particularly important, then I might write:

> There does not appear to be a statistically significant difference between Nigeria and Egypt/Sudan: $t[34] = -0.095$, ns.

If many comparisons are important, then it might prove useful to include them in a table. At that point, it also becomes important to provide details about the manner in which the p-values have been adjusted.

8.3 Multiple Factors in a Regression Model

So far, we have focused on regression models that include a single factor. However, one can easily combine factors in a single regression model. The only thing that is affected by this is the interpretation.

To determine our thoughts, imagine that we seek to predict per capita FDI with per capita GDP and two factors. The first of these factors is the distinction between democracies and non-democracies that we used earlier. The second factor distinguishes Sub-Saharan Africa from North Africa. As per the United Nations Statistics division, we assume that Algeria, Egypt, Libya, Morocco, Sudan, and Tunisia belong to North Africa; all other countries belong to

Table 8.6: A Regression with a Two Factors

Group	Sub	Dem	CEF
North-African non-democracies	0	0	$\mu_i = \beta_0 + \beta_3 \text{GDP}_i$
North-African democracies	0	1	$\mu_i = \beta_0 + \beta_2 + \beta_3 \text{GDP}_i$
Sub-Saharan non-democracies	1	0	$\mu_i = \beta_0 + \beta_1 + \beta_3 \text{GDP}_i$
Sub-Saharan democracies	1	1	$\mu_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 \text{GDP}_i$

Notes: CEF=conditional expectation function.

Sub-Saharan Africa. We now formulate the following conditional expectation function:

$$\mu_i \quad = \quad \beta_0 + \beta_1 \text{Sub}_i + \beta_2 \text{Dem}_i + \beta_3 \text{GDP}_i$$

Here Sub is a dummy that takes on the value 1 if the country is located in Sub-Saharan Africa and 0 otherwise. Similarly, Dem takes on the value 1 if a country is democratic and 0 otherwise. We can estimate this model in the usual manner.

For the interpretation it is once more useful to consider different groupings of the cases, as we have done in Table 8.6. For example, the equation for North-African non-democracies is obtained by substituting a 0 for both Sub and Dem in the conditional expectation function. Similarly, the equation for Sub-Saharan democracies comes about by substituting a 1 for Sub and a 1 for Dem.

We now see that the intercept, β_0, applies to a particular intersection of the two factors, namely between North-African countries and non-democracies. It should be interpreted as such. This is different from what we had before: where β_0 applied to just one baseline in the discussion so far, it now covers the combination of two baselines.

For purposes of interpretation, let us also consider the effect of being a democracy. For the North-African countries, the shift in intercept that is attributable to being a democracy is β_2. We know this because the intercept for North-African democracies is $\beta_0 + \beta_2$, whereas it is β_0 for North-African non-democracies. Moving to Sub-Saharan Africa, we see that the

shift in the intercept that is attributable to being a democracy is again β_2: $(\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_1) = \beta_2$. Hence, β_2 can be interpreted unambiguously as the effect of being a democracy.

In a similar vein, we can consider the effect of being in the Sub-Sahara. In non-democracies, the shift attributable to being a Sub-Saharan country is β_1. After all, the intercept in Sub-Saharan non-democracies is $\beta_0 + \beta_1$, whereas it is β_0 in North-African non-democracies. The same difference obtains for democracies: $(\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_2) = \beta_1$. Hence, β_1 is the effect of being in Sub-Saharan Africa.

We see that the interpretation becomes a bit more complex when we have multiple factors because there are now combinations of those factors that we have to consider. However, as long as you apply a strategy like the one shown in Table 8.6, there is not much room for confusion.

8.4 When To Use Dummy Variables

Many predictors in the social sciences are factors. For example, when we use a feeling thermometer to predict a political attitude, then that predictor is not really continuous. Rather, thermometer scores are typically recorded from 0 to 100, in steps of 1. Does this mean that we should create 100 dummy variables to absorb the effect of the feeling thermometer? The answer is no.

We use dummy variables to avoid making a questionable assumption that the levels of a factor are equal interval (i.e., equidistant). However, as the number of levels increases, this assumption becomes less and less worrisome. The intervals between adjacent levels become ever smaller and in the limit they converge to a common size of 0 (if the number of levels goes to infinity). Not only is it unnecessary to use dummies under these circumstances, it may in fact be ill-advised. If we enter the feeling thermometer as a covariate, we need to estimate only one parameter. If we enter it through a dummy specification, then we suddenly need 100 parameters. This consumes many more degrees of freedom, inflates standard errors, and generally makes interpretation way too difficult.

So when should we use dummy variables? There are two situations were we

cannot avoid them.

1. *Nominal Predictors:* If a predictor is measured on a nominal scale, then it should be entered into the regression model through a series of dummy variables. This is true even when the number of levels is large. Thus, predictors such as region, gender, race, and country should always be "dummied up."

2. *Ordinal Predictors with Few Categories:* If a predictor is measured on an ordinal scale and this scale is crude, i.e., has few categories, then the predictor should probably be entered as a set of dummies. Practice here is less consistent. Some scholars never use dummy variables for ordinal predictors, while others uses dummies when the number of levels falls short of some threshold. There is no consensus on what the size of this threshold is. Some set it at 5, others at 7, and still others at 10. Personally, I use a cutoff of 7 levels, although this is no more than a rule of thumb.

8.5 Conclusions

Once we recognize that the linear regression model makes no assumptions whatsoever about the measurement level of the predictors, its versatility is even more obvious. Still, it behooves us to treat factors different from covariates. We do this not because of some statistical rationale but only to avoid problematic interpretations. This chapter has shown how one can use dummy variables to aid with the interpretation of factors. In the next chapter, we shall continue our discussion of dummy variables, showing how they can be used to expand regression analysis to allow for heterogeneous effects.

Chapter 9

Interaction Effects

A **moderator** is a variable that influences the relationship between two other variables, specifically, that between a predictor and the dependent variable (see Figure 9.1). By bringing moderators into the regression model, we can engage in *condition-seeking* (Greenwald et al., 1986). This means that we can ascertain under what conditions, i.e., values of the moderator, the relationship between a predictor and dependent variable exists, reverses signs, or gains in strength.

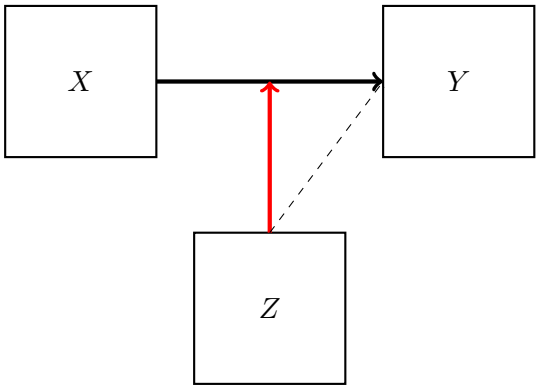
The vehicle that allows us to engage in condition-seeking is the **interaction**. This is simply the product of a moderator and a predictor variable. It allows us to capture the impact of a particular combination of values of the predictor and the moderator that goes over and beyond their additive effects. In this sense, interactions allow us to model non-additive relationships.

In this section, we discuss different types of interactions and their interpretations. We discuss some useful applications of interactions and address some pitfalls.

9.1 Interactions Between Factors

Let us return to our models explaining per capita FDI in Africa. Imagine that we have at our disposal a continuous indicator of democracy, which we include only as a linear term to keep things simple. Further, we have crude indicators of wealth and trade openness. Specifically, in terms of wealth we only know

Figure 9.1: The Role of a Moderator Variable



Note: Z is the moderator variable. When Z has a direct effect on Y (dashed line) we call it a quasi-moderator. When this effect is absent, then Z is a pure moderator.

whether a country's per capita GDP is above or below the African median. If it is above, we say that the country is rich; otherwise we say it is poor. For trade openness, we also only know whether a country is above or below the African median. If it is above, we say that the country's economy is open; otherwise, we say it is closed. The model so far is typical of the models that we discussed in the previous chapter. It may be written as

$$\text{FDI}_i = \beta_0 + \beta_1 \text{Open}_i + \beta_2 \text{Rich}_i + \beta_3 \text{Democ}_i + \varepsilon_i,$$

where Rich and Open are two dummy variables. But now we add the wrinkle that the effect of openness varies depending on a country's wealth. For example, we may believe that openness matters the most to potential investors when the country is rich. How would we capture this in a regression model?

9.1.1 The Interaction Term

The answer to the previous question is that we create the multiplicative term Rich $\times$ Open and add this to the regression model:

$$\begin{aligned} \text{FDI}_i = {} & \beta_0 + \beta_1 \text{Open}_i + \beta_2 \text{Rich}_i + \beta_3 \text{Democ}_i + \\ & \beta_4 \text{Rich}_i \times \text{Open}_i + \varepsilon_i \end{aligned}$$

We call this multiplicative term the interaction. The constituent terms, Open and Rich, are known as the statistical main effects.

Where does the multiplicative term come from? To answer this question let us revisit the model without interactions. In this model, we argued that the effect of openness may vary with wealth. Let us focus on the first part of this theoretical expectation, i.e., the effect of openness varies. We can capture this by writing

$$\text{FDI}_i = \beta_0 + \alpha_{1i}\text{Open}_i + \beta_2\text{Rich}_i + \beta_3\text{Democ}_i + \varepsilon_i$$

This model is identical to the original model, except that I have added a subscript i to the parameter for trade openness: β_1 has become α_{1i}. Adding the subscript implies that the effect of openness varies across countries. In other words, the parameter for openness has become a variable.

Like any other variable, α_{1i} can be modeled. Our theoretical expectation is that variation in the effect of openness can be accounted for through country wealth. This suggests the following model:

$$\alpha_{i1} = \beta_1 + \beta_4\text{Rich}_i$$

This is a regression equation, except that the left-hand side is not your usual variable and the right-hand side does not include an error term.[1]

We can now take the model for α_{i1} and substitute it into the equation for FDI:

$$\begin{aligned}\text{FDI}_i &= \beta_0 + (\beta_1 + \beta_4\text{Rich}_i)\,\text{Open}_i + \beta_2\text{Rich}_i + \\ &\quad \beta_3\text{Democ}_i + \varepsilon_i \\ &= \beta_0 + \beta_1\text{Open}_i + \beta_2\text{Rich}_i + \beta_3\text{Democ}_i + \\ &\quad \beta_4\text{Rich}_i \times \text{Open}_i + \varepsilon_i\end{aligned}$$

This model can be estimated quite simply using OLS. As we shall see in the

[1] The error term would not be identified, which means that we do not have sufficient information to estimate its characteristics (e.g., variance).

next section, R makes this especially straightforward.

9.1.2 Using R

The model that we specified can be estimated quite easily in R. The key is that we specify the variables Rich and Open as factors. Beyond that, the following syntax suffices:

```
lm(fdipc ~ rich*open+democ, data=africa)
```

The multiplication sign between the two dummies causes R to create the interaction term. In addition, the two main effects for the dummies are included in the model. The estimation results can be seen in Table 9.1.

9.1.3 Interpretation

Now that we have seen the derivation and estimation of a model involving an interaction between two factors, how do we go about interpretation? It is easiest to characterize the conditional expectation function at different values of the factors and then draw comparisons. The necessary computations are shown in Table 9.2. For example, the CEF for rich, open economies is obtained by substituting ones for Rich and Open:

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1 + \beta_3 \text{Democ}_i + \beta_4 \cdot 1 \times 1 \\ &= \beta_0 + \beta_1 + \beta_2 + \beta_4 + \beta_3 \text{Democ}_i \end{aligned}$$

We can now look at the effect of trade openness in poor and in rich countries. Holding the level of democracy constant, the difference in the CEFs of open and closed poor economies is

$$(\beta_0 + \beta_1 + \beta_3 \text{Democ}_i) - (\beta_0 + \beta_3 \text{Democ}_i) = \beta_1$$

The same difference for rich economies is

$$(\beta_0 + \beta_1 + \beta_4 + \beta_3 \text{Democ}_i) - (\beta_0 + \beta_2 + \beta_3 \text{Democ}_i) = \beta_1 + \beta_4$$

Table 9.1: Example of an Interaction Between Two Factors

	Dependent variable:
	Per Capita FDI
Rich	21.87
	(176.40)
Open	−27.91
	(177.07)
Democracy	−76.67**
	(36.05)
Rich × Open	440.37*
	(253.60)
Constant	359.09*
	(180.84)
Observations	44
Adjusted R^2	0.17
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9.2: The Interpretation of Dummy Interactions

Group	Open	Rich	CEF
Poor closed economies	0	0	$\mu_i = \beta_0 + \beta_3 \text{Democ}_i$
Rich closed economies	0	1	$\mu_i = \beta_0 + \beta_2 + \beta_3 \text{Democ}_i$
Poor open economies	1	0	$\mu_i = \beta_0 + \beta_1 + \beta_3 \text{Democ}_i$
Rich open economies	1	1	$\mu_i = \beta_0 + \beta_1 + \beta_2 + \beta_4 + \beta_3 \text{Democ}_i$

Notes: CEF = conditional expectation function.

The difference in the differences is thus

$$(\beta_1 + \beta_4) - \beta_1 \quad = \quad \beta_4$$

Let us now bring in the estimates from Table 9.1. The estimate associated with trade openness is -27.91; this is $\hat{\beta}_1$. This is *not* the effect of openness in general. Instead, it is the effect of openness for poor African economies. The estimate associated with the interaction is 440.37; this is $\hat{\beta}_4$. This is the difference in the effect of openness for rich and poor countries: the effect of openness is 440.37 points (in this case, U.S. dollars) higher in rich than in poor countries. This means that the effect of openness in rich countries is equal to $-27.91 + 440.37 = 412.46$; this is $\hat{\beta}_1 + \hat{\beta}_4$.

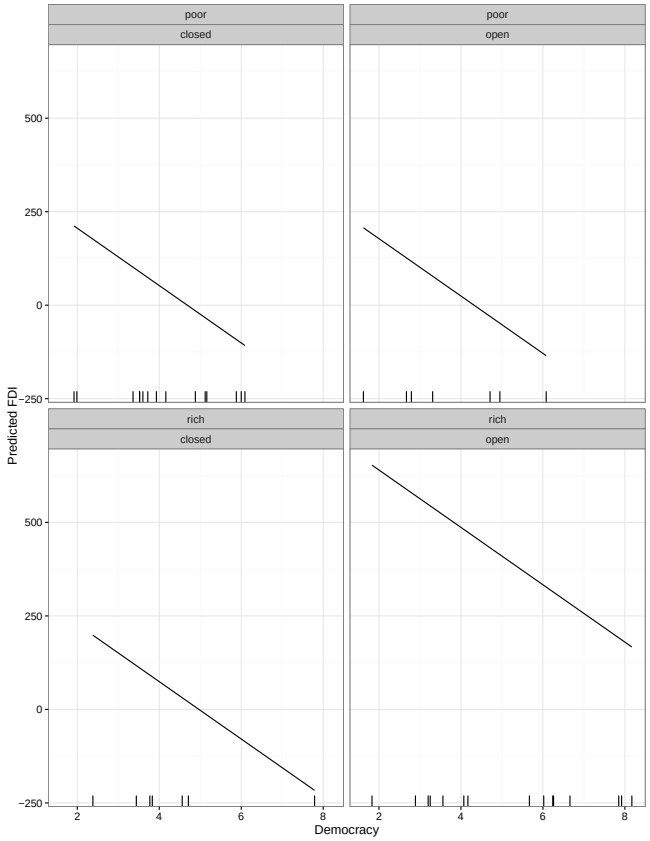
The complete set of results is depicted in Figure 9.2. Here, we have drawn the regression line for democracy and per capita FDI for different combinations of wealth and trade openness. The big jump in the intercept that arises for rich, open economies is due to the large interaction.

9.1.4 Hypothesis Testing

Several hypotheses can be tested for the model with factor interactions. First, we can test if there is a non-zero interaction effect. Next, we can test the intercepts for different combinations of wealth and openness. Finally, we can test for differences between the intercepts.

Testing the Interaction Imagine that $\beta_4 = 0$. Then Table 9.2 would show that the effect of trade openness is β_1, regardless of whether a country is rich

Figure 9.2: Democracy, Wealth, Trade Openness and FDI



Note: Based on the estimates from Table 9.1.

or poor. In other words, we now obtain the same effect of openness across all countries. There is no interaction, i.e., the effect of openness is not moderated by wealth.

Establishing that there is a moderator effect from wealth thus amounts to testing and rejecting $H_0 : \beta_4 = 0$. The test is a simple t-test of the partial slope associated with the interaction term. Consulting Table 9.1, we see that this test yields $p < .10$. The conclusion now depends on the Type-I error rate that one has set. If this is .05, then we would have to conclude that we fail to reject H_0; there is no evidence of an interaction between wealth and openness. If the Type-I error rate was set at .10, which makes sense for a sample this small, then we would reject H_0 and conclude there is an interaction effect.[2]

Testing Intercepts Testing intercepts in a model with interacting dummy variables is no different than what we described for factor variables in the previous chapter. For instance, consider the question of whether the intercept for rich, open economies is statistically different from 0. This amounts to testing $H_0 : \beta_0 + \beta_1 + \beta_2 + \beta_4 = 0$. We can use the `multcomp` procedure in R to accomplish this task. In our case, we obtain $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4 = 793.4$. This has a standard error of 211.7. Consequently, the test statistic under the null hypothesis is 3.75. When referred to a t-distribution, we obtain $p = 0.001$. Hence, we reject the null hypothesis and conclude that the intercept for rich, open economies is significantly different from 0.

Testing Differences Between Intercepts The same procedure may be used to test differences between intercepts. For example, we may wish to ascertain whether FDI is different for open and closed economies when the countries involved are wealthy. We have seen that the difference is $\beta_1 + \beta_4$. The `multcomp` procedure shows an estimate of 412.5, which is statistically significant at the .05-level. Thus, we can reject the null hypothesis that there is no difference between open and closed *wealthy* economies in terms of the expected level of FDI.

[2] The reason to increase the Type-I error rate is to obtain reasonable statistical power despite having few observations.

9.1.5 Interaction Effects are Symmetric

In the discussion so far, we have assumed that wealth moderates the effect of trade openness. But could we turn this around and argue that openness moderates the effect of wealth? The answer is yes and to see this we consider again the breakdown from Table 9.2. Considering closed economies (and holding democracy constant) the difference in the conditional expectation function due to wealth is β_2. Repeating this computation for open economies, we obtain $\beta_2 + \beta_4$. The difference in these effects is β_4, i.e., the regression coefficient that is associated with the interaction term. Thus, the interaction effect cannot distinguish between the moderating effect of openness on wealth and the moderating effect of wealth on openness. This is what we mean when we say that the interaction is symmetric.

9.2 Interactions Between Factors and Covariates

In addition to building interactions between factors, it is possible to construct such terms between factors and covariates. TO illustrate this, let us revisit the FDI model from the previous section. This time, however, per capita GDP is measured on a continuous scale. As such, we treat it as a covariate. We assume that the effect of this covariate depends on trade openness, which we continue to treat as a factor. Thus, the model that we propose is the following:

$$\begin{aligned} \mathsf{FDI}_i &= \beta_0 + \beta_1 \mathsf{Open}_i + \beta_2 \mathsf{GDP}_i + \beta_3 \mathsf{Democ}_i + \\ &\quad \beta_4 \mathsf{Open}_i \times \mathsf{GDP}_i + \varepsilon_i \end{aligned}$$

This model can be estimated using OLS. The estimation results are shown in Table 9.3.

9.2.1 Interpretation

How do we interpret this regression model? It is easiest to do this by deriving the **simple slope** equation. This shows the marginal effect of a predictor, in this case GDP, at different values of the moderator, in this case trade openness.

Table 9.3: Example of an Interaction Between a Factor and a Covariate

	Dependent variable:
	Per Capita FDI
Democracy	-40.87^{**}
	(17.21)
GDP per capita	0.01
	(0.02)
Open Economy	-94.69
	(72.20)
Open Economy $\times$ GDP	0.09^{***}
	(0.02)
Constant	195.36^{**}
	(84.32)
Observations	44
Adjusted R^2	0.80
Note:	*p<0.1; **p<0.05; ***p<0.01

Consider the model $\mu = \beta_1 x + \beta_2 z + \beta_3 x \times z + \mathbf{x}_o^\top \boldsymbol{\beta}$, where $\mathbf{x}_o$ includes the constant and any predictors other than X and Z. Then mathematically, the simple slopes are

Equation 9.1: Simple Slope

$$\begin{aligned} \frac{\partial \mu}{\partial x} &= \beta_1 + \beta_3 z \\ \frac{\partial \mu}{\partial z} &= \beta_2 + \beta_3 x \end{aligned}$$

Note that we have two simple slope equations due to the symmetry of the interaction, which allows us to treat Z as the moderator of X and vice versa.

In our case, the simple slope for GDP is given by

$$\frac{\partial \mu}{\partial \text{GDP}} = \beta_2 + \beta_4 \text{Open}$$

In a closed economy, Open is equal to 0 so that the simple slope equation may be written as $\beta_2 + \beta_4 \cdot 0 = \beta_2$. In an open economy, Open is equal to 1 so that the simple slope equation may be written as $\beta_2 + \beta_4 \cdot 1 = \beta_2 + \beta_4$ (see Table 9.4). Thus, β_4 gives the difference in the GDP slope for open as compared to closed economies.

The implication of this analysis is that β_2 should not be interpreted as the unconditional effect of GDP. In fact, it is a conditional effect, namely the effect of GDP provided that the economy is closed. To arrive at the effect of GDP when the economy is open, we have to add β_4 to β_2.

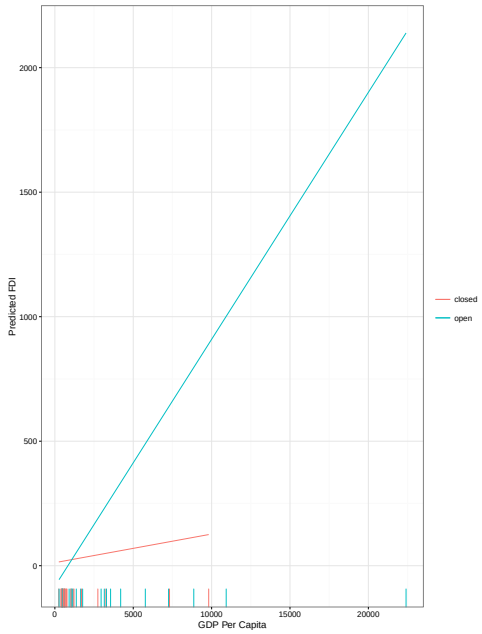
Another implication is that we now obtain non-parallel regression lines for

Table 9.4: Simple Slope Equations

Economy	Equation	Estimate
Closed	β_2	0.01
Open	$\beta_2 + \beta_4$	195.37

Notes: Estimates based on Table 9.3.

Figure 9.3: Simple Slopes for GDP by Trade Openness



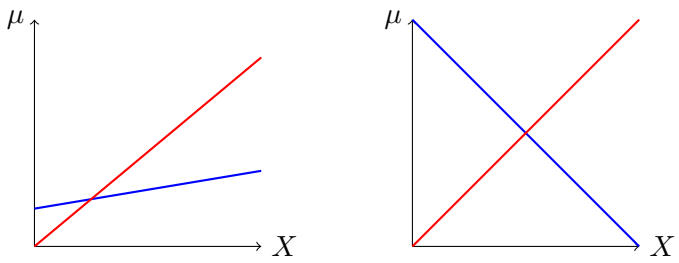
Note: Based on the estimates from Table 9.3. Confidence intervals have been omitted.

GDP, as is illustrated in Figure 9.3. For closed economies, the simple slope is much shallower than it is for open economies. This means that the effect of GDP on FDI is much stronger in open than in closed economies.

The pattern that we observe in Figure 9.3 is an example of a so-called **ordinal** interaction. In these interactions, the magnitude of the simple slope changes but the direction remains always the same. In Figure 9.3, the simple slope for GDP is always positive. What changes is the magnitude of this slope. It is also possible to obtain a so-called **dis-ordinal** interaction. In those interactions, the distinguishing characteristic is that the direction of the simple slope changes across values of the moderator. (The magnitude may change as well.) The distinction is illustrated in Figure 9.4.

When looking at the right-hand panel of Figure 9.4, it becomes apparent why it may be crucial to engage in condition-seeking. If we were to run a regression on X alone, we would conclude that there is no effect of this predictor because

Figure 9.4: Ordinal and Dis-Ordinal Interactions



(a) Ordinal (b) Dis-Ordinal

Note: Based on hypothetical data. Red line: $Z = 1$; blue line: $Z = 0$.

the red and blue lines cancel each other. However, this conclusion would be erroneous because, in fact, there are clear effects of X in each of the sub-groups formed by the moderator. In one sub-group, the effect is positive and in the other it is negative. How X plays into the dependent variable, then, depends on the specific conditions described by the moderator variable.

9.2.2 Hypothesis Testing

The first order of business is to check if the interaction between the factor and covariate is statistically significant. Next, one can check the extent to which the simple slopes are statistically different from zero.

Testing the Interaction We have seen that β_4 constitutes the difference in the simple slope of GDP in open versus closed economies. Hence, the null hypothesis $\beta_4 = 0$ implies that the effect of GDP does not depend on trade openness—there is no interaction. It is simple to test this hypothesis. All we have to do is look at the t-statistic that is associated with the interaction. This is 4.535 and the p-value is 0.000. Thus, we can confidently conclude that the interaction is statistically significant; the effect of GDP is moderated by trade openness.

Testing the significance of an interaction is not always this straightforward. Imagine, for example, that we argue that the effect of GDP varies by region. If

we omit all other predictors and use the regional division that we introduced in Chapter 8, then the conditional expectation function is[3]

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1 \text{EA}_i + \beta_2 \text{ES}_i + \beta_3 \text{FWA}_i + \beta_4 \text{M}_i + \beta_5 \text{OWA}_i + \\ &\quad \beta_6 \text{SA1}_i + \beta_7 \text{SA2}_i + \beta_8 \text{GDP}_i + \beta_9 \text{EA}_i \times \text{GDP}_i + \\ &\quad \beta_{10} \text{ES}_i \times \text{GDP}_i + \beta_{11} \text{FWA}_i \times \text{GDP}_i + \beta_{12} \text{M}_i \times \text{GDP}_i + \\ &\quad \beta_{13} \text{OWA}_i \times \text{GDP}_i + \beta_{14} \text{SA1}_i \times \text{GDP}_i + \beta_{15} \text{SA2}_i \times \text{GDP}_i\end{aligned}$$

The simple slope equation for GDP is now

$$\begin{aligned}\frac{\partial \mu_i}{\partial \text{GDP}_i} &= \beta_8 + \beta_9 \text{EA}_i + \beta_{10} \text{ES}_i + \beta_{11} \text{FWA}_i + \beta_{12} \text{M}_i + \beta_{13} \text{OWA}_i + \\ &\quad \beta_{14} \text{SA1}_i + \beta_{15} \text{SA2}_i\end{aligned}$$

This reduces to a constant effect of β_8 if and only if $\beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$. If we state this as the null hypothesis, then we should use a Wald test to test it. In this case, we obtain $F = 1.2$, which yields $p = 0.34$. Thus, we conclude that the effect of GDP is not moderated by region.

Testing the Significance of a Simple Slope Figure 9.3 shows the partial slopes for GDP by trade openness. We may ask for each of these slopes whether it is statistically significant. For closed economies, this question can be answered quite easily. In this case, the simple slope for GDP is simply β_2 (see Table 9.4). We can pose $H_0 : \beta_2 = 0$ and test this using the t-test. All of the information is automatically provided in the R output: the estimate is 0.01 and has a standard error of 0.02, so that the test statistic is 0.659, which yields $p = 0.514$. With a p-value this high, we cannot reject the null hypothesis. Thus, we conclude that there is no effect of GDP on FDI in closed economies.

If we want to test the significance of the slope for GDP in open economies, then things become slightly more complex. Table 9.4 shows that the simple slope is now $\beta_2 + \beta_4$. An insignificant slope thus means that we fail to reject

[3]We drop Nigeria from the estimation because a separate slope cannot be estimated. To do this, we would need at least two data points but the region of Nigeria includes only one observation.

$H_0 : \beta_2 + \beta_4 = 0$. We can use the `multcomp` library to test this hypothesis. We find that the the simple slope is approximately 0.10, with an estimated standard error of 0.01. Under the null hypothesis, the test statistic is 12.18 and has a p-value of 0.000. Thus, for any customary Type-I error rate, the null hypothesis must be rejected. We conclude that there is a significant relationship between GDP and FDI in open economies.

9.3 Interactions Between Covariates

In our most recent explorations, we treated GDP as a covariate and trade openness as a factor. However, it is possible to obtain a continuous measure of trade openness. If we use this, then, the interaction between the two variables simply becomes one of two covariates. Consequently we would be estimating

$$\begin{aligned} \text{FDI}_i &= \beta_0 + \beta_1 \text{Openness}_i + \beta_2 \text{GDP}_i + \beta_3 \text{Democ}_i + \\ &\quad \beta_4 \text{Openness}_i \times \text{GDP}_i \end{aligned}$$

Again, this model can be estimated using OLS; the estimates are displayed in Table 9.5.

9.3.1 Interpretation

For the interpretation, we rely once more on the simple slope equation. For example, if we are interested in the effect of per capita GDP, we can compute the simple slope as

$$\frac{\partial \mu_i}{\partial \text{GDP}_i} = \beta_2 + \beta_4 \text{Openness}_i$$

This looks like any other simple slope equation, except that trade openness is now measured on a continuum.

We can now proceed in a number of different ways. A common approach is to depict the simple slope as a function of all observed values of the moderator.

Table 9.5: Example of an Interaction Between Two Covariates

	Dependent variable:
	Per Capita FDI
Democracy	-16.91
	(10.15)
GDP per capita	-0.08^{***}
	(0.01)
Trade Openness	0.23
	(0.63)
Openness $\times$ GDP	0.001^{***}
	(0.0001)
Constant	97.38
	(58.11)
Observations	44
Adjusted R^2	0.94
Note:	*p<0.1; **p<0.05; ***p<0.01

In our case, this would mean that we obtain the estimator

$$\hat{\beta}_2 + \hat{\beta}_4 \text{Openness}$$

and its estimated variance

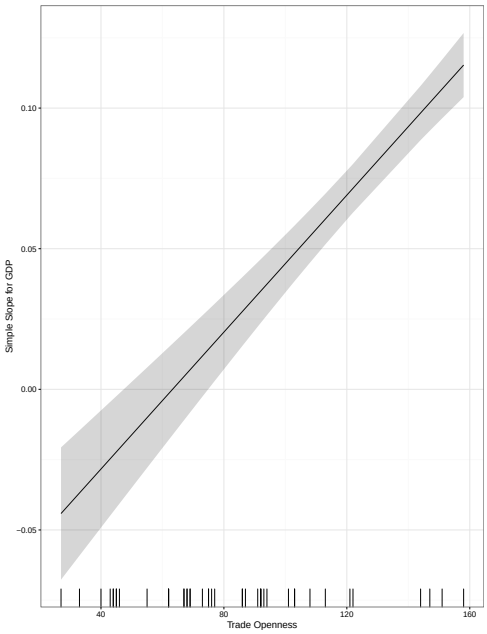
$$\begin{aligned}\widehat{\text{Var}}[\hat{\beta}_2 + \hat{\beta}_4 \text{Openness}] &= \widehat{\text{Var}}[\hat{\beta}_2] + \text{Openness}^2 \cdot \widehat{\text{Var}}[\hat{\beta}_4] + \\ &\quad 2 \cdot \text{Openness} \cdot \widehat{\text{Cov}}[\hat{\beta}_2, \hat{\beta}_4]\end{aligned}$$

This then allows us to generate a confidence interval that can be shown along with the simple slope. The necessary computations with a 95% confidence interval are shown in the syntax below, which assumes that the data are stored in `africa` and the estimation results in `fit`:

```
library(dplyr)
library(ggplot2)
africa <- mutate(africa, simple.slope=coef(fit)[3]+
coef(fit)[5]*openness)
africa <- mutate(africa, var.slope=vcov(fit)[3,3]+
openness^2*vcov(fit)[5,5]+2*openness*vcov[3,5])
africa <- mutate(africa, lb.slope=simple.slope-
qt(.975,df.residual(fit))*sqrt(var.slope))
africa <- mutate(africa, ub.slope=simple.slope+
qt(.975,df.residual(fit))*sqrt(var.slope))
ggplot(africa, aes(x=openness, y=simple.slope))+
geom_line()+
geom_ribbon(africa, aes(ymin=lb.slope, ymax=ub.slope),
alpha=.2)+
geom_rug(sides="b")+
xlab("Trade Openness")+
ylab("Slope for GDP")+
theme_bw()
```

The resulting graph is shown in Figure 9.5. It shows how the effect of a country's

Figure 9.5: Simple Slope for GDP as a Function of Trade Openness



Note: Based on the estimates from Table 9.5. 95% confidence interval is shown.

wealth increases with trade openness. It also shows that the slope for GDP is sometimes statistically significant and negative and at other times is significant and positive.[4] Note that the horizontal axis shows the values of the moderator, whereas the vertical axis shows the values of the simple slope.

A second approach is to select specific values of the moderator and to evaluate the simple slope at those values. Oftentimes, researchers select the minimum, maximum, and arithmetic mean of the moderator, but other choices may be more useful depending on your needs. For example, one could use the 25th, 50th, and 75th percentiles of the moderator. These values are taken from the estimation sample, i.e., the sample that produced the regression estimates.

Let us use the conventional values of the moderator to interpret the results from our most recent FDI model. In the estimation sample, trade openness ranges from 27.00 to 158.00, with a mean of 81.27. We can use the `multcomp`

[4]Significance is judged by whether the value of 0 is included in the 95% confidence interval at a particular value of the moderator.

library to perform the necessary computations. For example, at the minimum

```
library(multcomp)
R <- matrix(c(0,0,1,0,27), nrow=1)
summary(glht(fit, linfct=R))
```

If we perform this computation for each of the reference values of trade openness, we obtain the following results:

Openness	Slope	SE	t	p
Minimum	-0.044	0.012	-3.799	0.000
Mean	0.022	0.006	3.408	0.002
Maximum	0.115	0.006	20.450	0.000

Hence, we observe that GDP has a statistically significant negative effect when openness is at its minimum. It has a statistically significant positive effect when openness is at its mean or maximum.

9.3.2 Hypothesis Testing

Testing the Interaction When dealing with an interaction between two covariates, the very first test one should perform concerns the significance of the interaction term. Looking at the simple slope equation, the effect of GDP is rendered constant when $\beta_4 = 0$. When we formulate this as the null hypothesis, then a t-test suffices to determine its fate. Our estimate of β_4 is 0.001, with an estimated standard error of 0.0001. Dividing the estimate by the standard error, the test statistic under the null hypothesis is 10.997. This yields $p = 0.000$, so that for any customary Type-I error rate the conclusion will have to be that the null hypothesis is rejected. Thus, we conclude there is a significant interaction between trade openness and per capita GDP.

The Johnson-Neyman Technique Once we have established that there is a significant interaction effect, then we would like to ascertain for what values of the moderator the simple slope is statistically significant. For this purpose, one

can use the technique proposed by Johnson and Neyman (1936). The starting point is that the simple slope for GDP is not statistically significant when the test statistic lies between the critical values of the t-distribution. This means that significance starts at $\pm t_{Crit}$, which allows us to write

$$\pm t_{Crit} = \frac{\hat{\beta}_2 + \hat{\beta}_4 \text{Openness}}{\left[\hat{V}[\hat{\beta}_2 + \hat{\beta}_4 \text{Openness}]\right]^{.5}}$$

Rearranging terms and squaring then yields the following equation:

$$t^2_{Crit}\hat{V}[\hat{\beta}_2 + \hat{\beta}_4 \text{Openness}] - \left(\hat{\beta}_2 + \hat{\beta}_4 \text{Openness}\right)^2 = 0$$

This is a quadratic equation, which can be solved using standard algebraic methods.

The library `rockchalk` in R can be used to perform the necessary computations.

```
library(rockchalk)
simple <- plotSlopes(fit, modx = "openness",
plotx = "gdppc")
jn <- testSlopes(simple)
plot(jn)
```

The first command causes the computation (and depiction) of the simple slopes for GDP. The second command causes the computation of the values of trade openness beyond which the simple slopes for GDP are statistically significant at the .05-level. The last command produces a graphical display of the results, should one desire this. For our data, we find that the simple slope for GDP is statistically significant for values of openness below 47.39 and for values above 74.87. In between those boundaries, there is no evidence that GDP exerts a statistically significant effect on FDI.

The Johnson-Neyman technique is not well known in political science. Nevertheless, it provides a potent method for establishing where the simple slope is significant and where it is not. In our example, we find that the boundaries are

within the observed range of openness. When both boundaries lie beyond this range, then we know that the simple slopes are always significant. When one boundary lies beyond the observed range, then we know that the simple slopes on this side of the moderator distribution are all statistically significant.

9.3.3 To Center or Not to Center, That Is the Question

Interactions between covariates may induce severe multicollinearity. One consequence of this collinearity is that the interaction term and/or its components fail to achieve statistical significance due to inflated standard errors (see Chapter 10). More specifically, the statistical power could be reduced.

Where does this multicollinearity come from? Aiken and West (1991) demonstrate the following results for an interaction between the covariates X and Z:

$$\begin{aligned}\sigma_{X\times Z,X} &= E\left[(x^d)^2 z^d\right] + \sigma_X^2\mu_Z + \sigma_{X,Z}\mu_X \\ \sigma_{X\times Z,Z} &= E\left[(z^d)^2 x^d\right] + \sigma_Z^2\mu_X + \sigma_{X,Z}\mu_Z,\end{aligned}$$

where $x^d = x - \mu_X$ and $z^d = z - \mu_Z$ (see Appendix C.5 for a proof). Under multivariate normality among the predictors, these expressions simplify to

$$\begin{aligned}\sigma_{X\times Z,X} &= \sigma_X^2\mu_Z + \sigma_{X,Z}\mu_X \\ \sigma_{X\times Z,Z} &= \sigma_Z^2\mu_X + \sigma_{X,Z}\mu_Z\end{aligned}$$

When we peruse these equations, we notice that the means of X and Z play a prominent role in the covariance between the interaction and its constituent terms. Since these means are unlikely to be zero, it is possible that the covariance is quite large, thus causing a problem with multicollinearity.

Can we solve this problem? Aiken and West (1991) believe it can be and that the solution is surprisingly easy. If we center the predictors before we create the interaction term, then we eliminate most of the collinearity in on-normal cases and all of it in the multivariate normal case. This is easily demonstrated for the normal case. Using the earlier definitions of x^d and z^d, the centered interaction

is $x^d \times z^d$. If we now evaluate the covariances between the interaction and its constituent terms, we get

$$\begin{aligned} \sigma_{X^d \times Z^d, X^d} &= \sigma^2_{X^d}\mu_{Z^d} + \sigma_{X^d,Z^d}\mu_{X^d} = 0 \\ \sigma_{X^d \times Z^d, Z^d} &= \sigma^2_{Z^d}\mu_{X^d} + \sigma_{X^d,Z^d}\mu_Z]Z^d = 0 \end{aligned}$$

These expressions hold because $\mu_{X^d} = \mu_{Z^d} = 0$. We see that all of the covariance—and hence collinearity—disappears. It can be demonstrated (see Appendix C.5) that centering does not alter the estimate of the interaction.

The presentation above assumes that we know the population means. This assumption is not essential, however. We can center about the sample means of X and Z to create the interaction. After all, $E[x - \bar{x}] = E[z - \bar{z}] = 0$. If you center about the sample means, however, you should make sure to use the means from the estimation sample. In this way, you can be sure that the means of the centered variables are zero.

It is easy to implement centering in R by using the pequod library. The syntax for our model is

```
library(pequod)
fit <- lmres(fdipc~democ+openness*gdppc,
data=africa, centered=c("openness","gdppc"))
summary(fit)
```

This yields the estimates shown in Table 9.6. Looking at those estimates, we observe that the partial regression coefficient for trade openness is now statistically significant, as is the constant. This was not the case in Table 9.5 due to collinearity problems. We also observe that the partial regression coefficient for the interaction is identical to what we observed in Table 9.5; the same is true for its standard error. The coefficients for GDP and trade openness are different from those reported in Table 9.5. The constant has also changed. These are natural consequences of the centering of GDP and trade openness.

We can repeat the earlier exercise of looking at the simple slope at different values of the moderator. The minimum of the centered trade openness variable

Table 9.6: Example of Centering with Interactions

	Dependent variable:
	Per Capita FDI
Democracy	−16.91
	(10.15)
GDP per capita	0.02***
	(0.01)
Trade Openness	3.53***
	(0.55)
Openness × GDP	0.001***
	(0.0001)
Constant	175.20***
	(50.51)
Observations	44
Adjusted R^2	0.94
Note:	*p<0.1; **p<0.05; ***p<0.01

is -54.27, while the maximum is 76.73, and the mean is zero. This produces

Openness^d	Slope
Minimum	-0.044
Mean $= 0$	0.022
Maximum	0.115

We see that these simple slope estimates are identical to what we derived on the basis of Table 9.5. Centering has changed nothing to the partial effects of GDP.

While the present example shows a case of improved significance in an interactive model after centering, there is considerable skepticism among political methodologists that the procedure works (see, for example, Brambor, Clark and Golder, 2006; Kam and Franzese, 2007). The argument is that multivariate normality almost never holds for real data. More fundamentally, multicollinearity is a problem of insufficient data (see Chapter 10) and centering does not contribute any new information that should help with the estimation.

Even if one shares these doubts, there is an advantage to centering on which almost everyone can agree: it aids in the interpretation of the partial regression coefficients of the terms that constitute the interaction. To see this, let us contrast the simple slope equations that derive from the uncentered and centered analysis. In the uncentered analysis, we have seen that

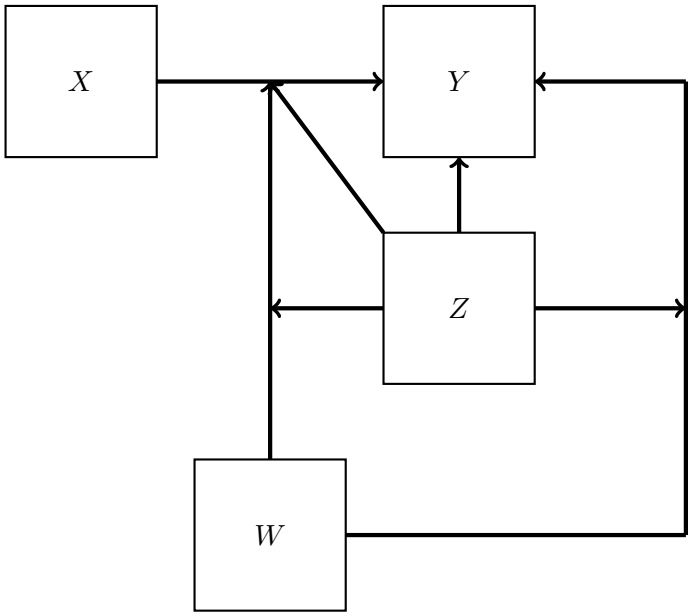
$$\frac{\partial \mu}{\partial \text{GDP}} = \beta_2 + \beta_4 \text{Openness}$$

This reduces to β_2—the partial slope of GDP—if and only if Openness is zero. But we have seen that Openness ranges between 27 and 158, so that it is never zero. The upshot is that we cannot interpret β_2. In the centered analysis, the simple slope is given by

$$\frac{\partial \mu}{\partial \text{GDP}^d} = \beta_2 + \beta_4 \text{Openness}^d$$

This reduces to β_2 if $\text{Openness}^d = 0$, which happens when Openness is at its mean value. The partial slope of GDP can thus be meaningfully interpreted.

Figure 9.6: Depiction of a Three-Way Interaction



Note: W and Z are moderators.

9.4 Higher-Order Interactions

The interactions that we have considered so-far are so-called two-way interactions because they involve two variables. It is entirely possible to create three-way, four-way and, in general, multi-way interactions should there be a theoretical reason to do so. One simply needs to be aware that it may require a large sample size to achieve some modicum of statistical power to detect such interactions.

Consider Figure 9.6, which illustrates a three-way interaction, as well as all possible two-way interactions, and main effects. The main effects are from W to Y, from X to Y, and from Z to Y. The main effect from W on Y is moderated by Z, as is the main effect of X on Y. The moderation of the relationship between X and Y via W, however, is itself also moderated (by Z). Thus, we are now exploring conditions that are themselves conditional.

The corresponding regression model can be derived as follows. We start

with a model of the dependent variable where we let the effects of X and W vary across units:

$$y_i = \beta_0 + \alpha_{1i} w_i + \alpha_{2i} x_i + \beta_3 z_i + \varepsilon_i$$

We now model α_{1i} and α_{2i}:

$$\begin{aligned} \alpha_{1i} &= \beta_1 + \beta_5 z_i \\ \alpha_{2i} &= \gamma_{1i} + \gamma_{2i} w_i \end{aligned}$$

(If the equation includes Z, then it is written in terms of βs; otherwise, it is written in terms of γ_is.) Substitution yields

$$\begin{aligned} y_i = \beta_0 &+ \beta_1 w_i + \beta_3 z_i + \beta_5 w_i \times z_i + \\ &\gamma_{1i} x_i + \gamma_{2i} w_i \times x_i + \varepsilon_i \end{aligned}$$

We now model the γ_is:

$$\begin{aligned} \gamma_{1i} &= \beta_2 + \beta_6 z_i \\ \gamma_{2i} &= \beta_4 + \beta_7 z_i \end{aligned}$$

Substitution now yields

$$\begin{aligned} y_i = \beta_0 &+ \beta_1 w_i + \beta_2 x_i + \beta_3 z_i + \\ &\beta_4 w_i \times x_i + \beta_5 w_i \times z_i + \beta_6 x_i \times z_i + \\ &\beta_7 w_i \times x_i \times z_i + \varepsilon_i \end{aligned}$$

We see that the full model includes all main effects, all two-way interactions, and a three-way interaction.

The interpretation is as always in terms of the simple slope. For example, the simple slope equation for X is

$$\frac{\partial \mu}{\partial x} = \beta_2 + \beta_4 w_i + \beta z_i + \beta_7 w_i \times z_i$$

The presence of a two-way interaction in this equation shows that the moderating effect of W is itself moderated by Z.

As an example, let us consider exit poll data from the 2008 U.S. presidential elections. Here, we have aggregated the individual survey responses to precinct-level estimates of the following attributes: (1) the percentage of the Obama vote share (dependent variable); (2) the proportion of whites; (3) the proportion of women; and (4) the proportion of voters with a BA or higher degree. We center the last three variables and enter them into the following regression model:

$$\begin{aligned} \text{Obama}_i &= \beta_0 + \beta_1 \text{White}_i + \beta_2 \text{Female}_i + \beta_3 \text{BA}_i + \\ &\quad \beta_4 \text{White}_i \times \text{Female}_i + \beta_5 \text{White}_i \times \text{BA}_i + \\ &\quad \beta_6 \text{Female}_i \times \text{BA}_i + \beta_7 \text{White}_i \times \text{Female}_i \times \text{BA}_i + \\ &\quad \varepsilon_i \end{aligned}$$

The OLS estimates are shown in Table 9.7. For the interpretation, we look at the effect of the proportion of BAs, setting the proportions of whites and women to ± 1 standard deviations about the mean:

Female	White	Slope	SE	t	p
$-1SD$	$-1SD$	17.53	9.16	1.91	0.057
$-1SD$	$+1SD$	17.43	10.06	1.73	0.084
$+1SD$	$-1SD$	23.66	10.50	2.25	0.025
$+1SD$	$+1SD$	49.46	9.63	5.14	0.000

The results show that a unit increase in the proportion of BAs (i.e., going from no to all BAs in a precinct) does not have a significant effect (at the .05-level) when the proportion of females is comparatively low. It does have a significant effect when the proportion of females is comparatively high. This is true regardless of whether the proportion of whites is low or high. The strongest effect, however, is attained when the proportions of women and whites are comparatively high. This is clear evidence of a three-way interaction between the proportion of BAs, whites, and women in the precinct.

You will have noticed that interpreting three-way interactions is a bit more

Table 9.7: Example of a Model With a Three-Way Interaction

	Dependent variable:
	Obama Vote Share
Prop. White	-46.47^{***}
	(3.80)
Prop. Female	22.19^{*}
	(11.64)
Prop. BA	27.02^{***}
	(4.96)
Prop. White $\times$ Prop. Female	50.93
	(41.90)
Prop. White $\times$ Prop. BA	22.47
	(18.81)
Prop. Female $\times$ Prop. BA	104.77^{*}
	(59.36)
Prop. White $\times$ Prop. Female $\times$ Prop. BA	248.49^{*}
	(144.92)
Constant	62.58^{***}
	(0.96)
Observations	300
Adjusted R^2	0.43
Note:	*p<0.1; **p<0.05; ***p<0.01

involved than interpreting two-way interactions, since it becomes relevant to consider combinations of the values of the moderators. The problem is compounded when we add four-way and even higher order interactions. Therefore, the inclusion of such terms is recommended only if (1) there is good theory to suggest their inclusion and (2) the sample size is sufficiently large to attain reasonable statistical power.

9.5 Important Applications of Interactions

9.5.1 The Two-Way ANOVA Model

If an interactive model includes only factors and their interactions, then the model is known as the ANOVA or analysis-of-variance model. Such models are often used to analyze randomized experiments. Although we shall revisit randomized experiments in Chapter 13, we can already introduce the basic ideas and analytic tools in this chapter.

Example In political surveys, respondents are frequently asked to provide ideological ratings of political candidates. Ostensibly, responses to these questions are based on factual knowledge about the candidate. However, a wealth of evidence suggests that prior questions may influence responses as well. Particularly relevant in this regard would be ideological rating questions of other candidates.

Imagine, we conduct the following experiment.[5] We are interested in the ideological ratings of Jimmy Carter on a 9-point ideological scale that runs from 1=conservative to 9=liberal. We consider two experimental conditions, constituting two factors. The first factor concerns the ideology of other candidates that are presented along with Carter. Here, there are four groups: (1) Carter is presented along with three liberal candidates (Jerry Brown, Ted Kennedy, and George McGovern); (2) there are two liberal candidates (i.e., the group consists of Ted Kennedy, George McGovern, and Ronald Reagan); (3) there is one liberal candidate (i.e., the group consists of Ted Kennedy, Ronald Reagan, and Gerald

[5]This is an expansion on the design described by Brown and Melamed (1990, p. 20). Their design includes only the anchoring condition; I have added the accuracy condition.

Table 9.8: Example of a Factorial Design

	Anchor (# Liberals)			
Motivation	3	2	1	0
Control	5,3,3,3	3,6,2,5	6,5,6,5	6,5,6,7
Accuracy	5,5,7,4	6,4,5,6	5,5,6,6	3,7,5,4

Notes: Hypothetical data. Table entries are ideological ratings of Carter.

Ford); and (4) there are no liberal candidates (i.e., the group consists of Ronald Reagan, Gerald Ford, and John Connoly). The expectation is that Carter is rated more conservatively the more liberals are mentioned with him. A second factor concerns motivation. Half of the participants do not receive special instructions (control). The other half, however, are told that they have to justify their rating of Carter after the experiment. It is believed that this will prompt an accuracy motivation, which makes participants less susceptible to anchoring on the ideology of the candidates mentioned along with Carter. Participants are randomly assigned to the resulting eight groups/cells. This means that a random number generator is used to determine to which group the participant is assigned. The importance of this will be discussed in much greater detail in Chapter 13. Hypothetical data from this experiment are shown in Table 9.8.

Before we analyze this experiment, a couple of terminological conventions are worth conveying. First, an experiment that uses random assignment and fully crosses two factors is known as a factorial design. Second, an experimental design in which each cell has the same number of participants is called balanced. In general, when designing experiments researchers always strive for balance.

Model and Results If we set the control motivation and 3-liberals condition as the baselines, then one way to write the two-way ANOVA model is

$$\begin{aligned} \text{Carter}_i &= \beta_0 + \beta_1 \text{Accuracy}_i + \beta_2 \text{Lib2}_i + \beta_3 \text{Lib1}_i + \beta_4 \text{Lib0}_i + \\ &\quad \beta_5 \text{Lib2}_i \times \text{Accuracy}_i + \beta_6 \text{Lib1}_i \times \text{Accuracy}_i + \\ &\quad \beta_7 \text{Lib0}_i \times \text{Accuracy}_i + \varepsilon_i \end{aligned}$$

Table 9.9: Two-Way ANOVA Results

	df	Sum Sq	F	p
Motivation	1	1.53	1.105	0.304
Anchor	3	7.34	1.767	0.180
Motivation × Anchor	3	10.84	2.609	0.075
Residuals	24	33.25		

Notes: Based on Table 9.8.

(Jobson (1991) calls this the base cell representation of the ANOVA model.) Although one can show the parameter estimates for this model, it is customary to show an analysis of variance table, as is done in Table 9.9.[6] Such a table shows how the variance is distributed over the statistical main effects, the interaction, and the residuals.

What can we conclude from Table 9.9? We observe that neither of the statistical main effects is statistically significant. The interaction, however, is significant at the .10-level. It suggests that the effect of the anchor depends on the motivation.

This can also be observed from the predicted means. The computational formulas and estimates for these means are shown in Table 9.10. In the motivational control group, it appears that the anchoring effects are quite dramatic. For example, when Carter is included in a group of three liberals, then his ideological rating is on the average 3.5, i.e., toward the conservative end of the scale. However, when he is included in a group of zero liberals, the average ideological placement shoots up to 6.0, toward the liberal end of the scale. The fluctuations in the means is much more modest in the accuracy condition. Here, the average placement hovers around 5.00, which is right at the center of the ideological scale.

Types of Sums of Squares The sums of squares listed in Table 9.9 are so-called Type-I sums of squares. There are also so-called Type-II and Type-III sums of squares. These terms were introduced by the SAS Institute to distinguish between different methods of computing the effects of factors and their

[6] This table was obtained using R's `aov` function.

Table 9.10: Predicted Means for the Experiment

Motivation	Anchor	Formula	Estimate
Control	3 Liberals	$\hat{\mu} = \hat{\beta}_0$	3.50
Control	2 Liberals	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_2$	4.00
Control	1 Liberal	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_3$	5.50
Control	0 Liberals	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_4$	6.00
Accuracy	3 Liberals	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1$	5.25
Accuracy	2 Liberals	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_5$	5.25
Accuracy	1 Liberal	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_6$	5.50
Accuracy	0 Liberals	$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_4 + \hat{\beta}_7$	4.75

Notes: Table entries are based on Table 9.8.

interactions. In a balanced design, they are indistinguishable. In an unbalanced design, however, there can be marked differences among them. Since unbalanced data are not uncommon in the practice of political research, it is important to know the differences among the sums of squares and when to use a particular type.

Let's write the ANOVA model in the following schematic manner

$$y \quad = \quad A + B + AB + \varepsilon,$$

where the first two terms capture the statistical main effects of factors A and B, respectively, the third term captures the interaction, and the last term reflects experimental error. With Type-I, the sums of squares (SS) are derived sequentially. Thus:

$$\begin{aligned} SS(A) \\ SS(B|A) \\ SS(AB|A,B) \end{aligned}$$

Since A is the first term in the model, its sums of squares are computed unconditionally. Thus A is free to account for as much variance in Y as it can. This is different for B, which enters as the second term. It can only account for what has not already been explained by A. On top of that, only the part of B that

is non-redundant with A can be doing the explaining. Finally, the interaction comes into play only after we have controlled for the statistical main effects of A and B. If there is still something to be explained, then the interaction can give it a try.

The problem with Type-I sums of squares is that order in which we include terms into a model is generally arbitrary. Any permutation of the terms A, B, and AB is legitimate, but changing the order can have dramatic effects. In fact, we saw this when we looked at relative importance in Chapter 5. Due to the arbitrariness, Type-I sums of squares are usually not what we want.

With Type-II sums of squares, we use the following computations:

$$SS(A|B)$$
$$SS(B|A)$$
$$SS(AB|A,B)$$

The statistical main effect of A is thus computed after we control for B. Similarly, the main effect of B is computed after we control for A. The sums of squares of the interaction effect, however, are computed only after the statistical main effects have been removed. This is a conservative approach for detecting interactions. Indeed, the working assumption is that the interaction is not significant. It tends to be a powerful approach for discovering statistical main effects (more so than Type-III) and eliminates the arbitrary distinction between the effects of A and B.

Finally, Type-III sums of squares assumes that the interaction is statistically significant. The sums of squares of the statistical main effects and interaction are then computed as

$$SS(A|B,AB)$$
$$SS(B|A,AB)$$
$$SS(AB|A,B)$$

i.e., we condition on the remaining factor and the interaction. This type should be avoided if the interaction is not significant, as it reduces the power to detect

Table 9.11: An Unbalanced Factorial Design

	Anchor (# Liberals)			
Motivation	3	2	1	0
Control	3,3,3	3,6,2,5	6,5,6,5	6,5,6,7
Accuracy	5,5,7,4	6,5,6	5,5,6,6	7,5,4

Notes: Hypothetical data. Table entries are ideological ratings of Carter.

statistical main effects.

In R, the `car` package allows the computation of Type-II and Type-III sums of squares; the standard `anova` function gives Type-I sums of squares. To illustrate this, imagine that we lost the data of several participants in Table 9.8, thus getting the data shown in Table 9.11. If we now run the different sums of squares we obtain the results in Table 9.12. We see that the findings regarding the statistical main effects change quite dramatically depending on what type of sums of squares we use. In this particular case, Type-III may be the best option due to the marginally significant interaction effect.

9.5.2 Difference-in-Differences

A second important application of interactions is difference-in-differences, a widely used approach for drawing causal inferences. I should state up-front that there is nothing causal about regression analysis per se. However, when the model is combined with certain assumptions that ensure causal identification, then we can draw causal inferences. By causal identification, I mean that we can uncover an unbiased estimate of the true causal effect.

But what is the true causal effect? Here, I draw from the ideas of counterfactual theories of causation. As Hume (1993) already suggested in the 18th century, one can define a causal effect as the difference in the outcome that we observe in the presence of a cause and the outcome that we would have observed in its absence. A putative cause D has a causal effect to the extent that the two outcomes deviate from each other.

The problem is that we never observe both outcomes for one and the same

Table 9.12: An Unbalanced Factorial Design

		Type-I			Type-II			Type-III		
	df	SS	F	p	SS	F	p	SS	F	p
Motivation	1	3.500	3.052	0.095	4.346	3.789	0.065	8.679	7.568	0.000
Anchor	3	10.422	3.029	0.052	10.422	3.029	0.052	19.933	5.794	0.000
Interaction	3	9.857	2.865	0.061	9.857	2.865	0.061	9.857	2.865	0.061

Notes: Type-I computed using `anova`; Type-II and -III computed using `Anova` in `car`.

unit. For example, when the cause is present, then we do not observe the outcome that would have occurred had it been absent; we only observe the outcome that did occur in the presence of the cause. Put differently, we have missing data. It might be possible, however, to impute these missing data, i.e., to develop a reasonable guess of what we would have observed in the absence of the cause. This is the intuition behind difference-in-differences.

The Approach To determine our thoughts, let us consider a famous study by Card and Krueger (1994) on the effects of a minimum wage increase on employment in the fast-food sector. This study tracked a sample of fast-food restaurants in two American states, New Jersey and Pennsylvania, at two different time points, February-March 1992 and November-December 1992. In between those two time points, New Jersey increased its minimum wage from \$4.25 to \$5.05 per hour. We call this increase the treatment, which makes New Jersey into the treated unit. Pennsylvania, which did not see a minimum wage hike, is the control unit. We capture this by assigning the value $D = 1$ to New Jersey and $D = 0$ to Pennsylvania. We also define a time dummy, T, which takes on the value 1 in the post-treatment period (November-December) and 0 in the pre-treatment period (February-March).

We now counterfactually define the causal effect of the treatment for New Jersey:

$$\alpha \quad = \quad E[Y|D = 1, T = 1, \text{NJ}] - E[Y|D = 0, T = 1, \text{NJ}]$$

The causal inference literature calls this the average treatment effect of the treated (ATET). Here, Y is the full-time employment (FTE) in a fast-food restaurant. The first term on the right hand side is estimable, since this is the average FTE after the minimum wage hike. The second term, by contrast, is unknown, since this is the average FTE we would have observed in New Jersey in November-December, had minimum wages not been increased.

In order to identify the causal effect, we need to impute the second term. How can we do this? We make the following assumption: absent the treatment, the trajectory of change in FTE in New Jersey would have been the same as

that in Pennsylvania. Let

$$\delta \quad = \quad E[Y|D=0,T=1,\text{PA}] - E[Y|D=0,T=0,\text{PA}]$$

This is the change in the expected FTE in Pennsylvania. The first term on the right-hand side gives the expected FTE in Pennsylvania in November-December. The second term does the same in February-March. Note that D is always 0 because Pennsylvania is never treated. We now use δ to impute the counterfactual outcome for New Jersey in November-December:

$$\hat{\mu}|D=0,T=1,\text{NJ} \quad = \quad E[Y|D=0,T=0,\text{NJ}] + \delta$$

Here, The first term on the right-hand side is the expected FTE in NJ prior to the minimum wage increase, which is estimable because we have data on FTEs in New Jersey in February and March. The left-hand side is the predicted mean FTE had New Jersey not seen a minimum wage hike.

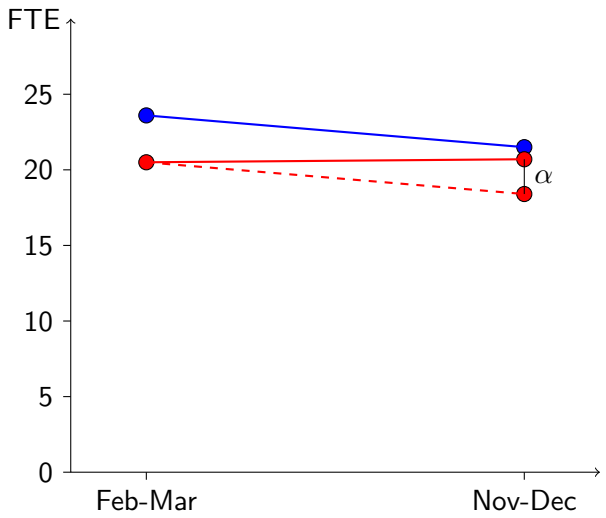
Now, it is a matter of substituting the various equations back into the formula for the treatment effect:

$$\begin{aligned}
\alpha &= E[Y|D=1,T=1,\text{NJ}] - \hat{\mu}|D=0,T=1,\text{NJ} \\
&= E[Y|D=1,T=1,\text{NJ}] - (E[Y|D=0,T=0,\text{NJ}] + \delta) \\
&= (E[Y|D=1,T=1,\text{NJ}] - E[Y|D=0,T=0,\text{NJ}]) - \\
&\quad (E[Y|D=0,T=1,\text{PA}] - E[Y|D=0,T=0,\text{PA}])
\end{aligned}$$

This can be estimated by taking the change in the mean FTE in the New Jersey sample and subtracting the change in the sample mean for Pennsylvania.

The whole idea is illustrated in Figure 9.7. Here, the blue line shows the trajectory of change in Pennsylvania. The red dashed line is the imputed trajectory of change that we would have observed in New Jersey absent the treatment. This line runs parallel to the blue line, reflecting the assumption that the trajectories of change are identical in Pennsylvania and no-treatment New Jersey. The solid red line is the actual trajectory of change in New Jersey. The difference between the end points of the solid and dashed red lines is the treatment

Figure 9.7: The Difference-in-Differences Design



Note: Based on Card and Krueger (1994). Blue = Pennsylvania; Red = New Jersey. Solid lines = observed trends; dashed line = counterfactual trend.

effect.

Obviously, the crucial point is that we draw parallel lines. This is obviously allowed only if we believe that the treated and control units are sufficiently similar. In this context, Card and Krueger (1994) spent considerable time selecting their cases and arguing why parallel trajectories of change could be assumed for New Jersey and Pennsylvania absent the treatment. Historical data on changes in FTE help here, but so does qualitative knowledge about the cases.

Model We can estimate the treatment effect by taking the difference in the differences of the sample means, as I described before. Analogously, we can formulate a regression model with S, T, and $S \times T$ as the predictors, where S equals 1 for New Jersey and 0 for Pennsylvania. Thus, the conditional expectation function is

$$\mu_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 D_i \times T_i$$

State	Period	S	T	CEF
PA	Feb-Mar	0	0	$\mu = \beta_0$
PA	Nov-Dec	0	1	$\mu = \beta_0 + \beta_1$
NJ	Feb-Mar	1	0	$\mu = \beta_0 + \beta_2$
NJ	Nov-Dec	1	1	$\mu = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Notes: CEF = conditional expectation function.

Table 9.13 shows how this function evaluates for New Jersey and Pennsylvania at different points in time. The over-time difference in means for Pennsylvania is given by

$$(\beta_0 + \beta_1) - \beta_0 \quad = \quad \beta_1$$

The over-time difference in means for New Jersey is given by

$$(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) \quad = \quad \beta_1 + \beta_3$$

The difference in the differences is

$$(\beta_1 + \beta_3) - \beta_1 \quad = \quad \beta_3$$

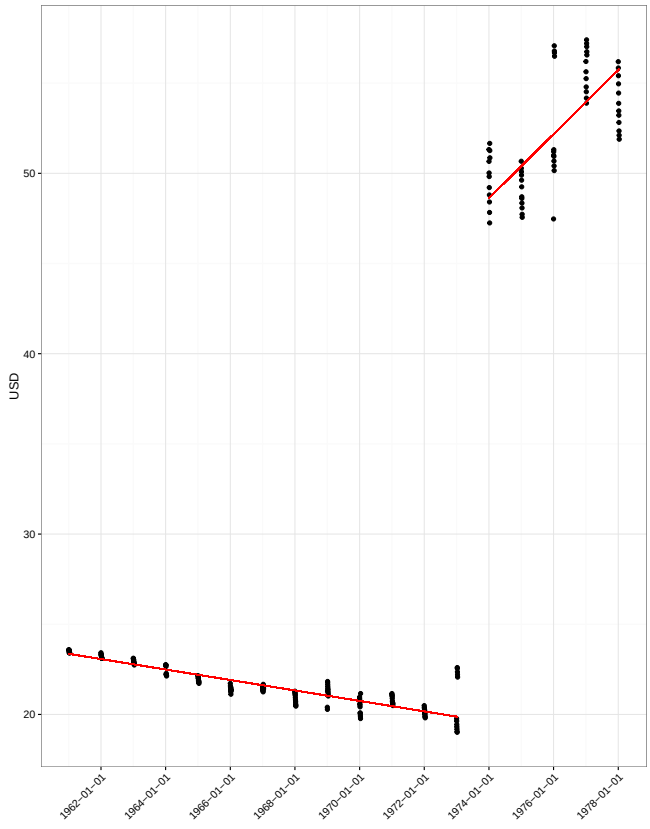
This means that $\hat{\beta}_3$ serves as an estimator of the treatment effect.

When we apply the interactive model to the data collected by Card and Krueger (1994), we observe that $\hat{\beta}_3 = 2.33$, $p < .10$. Hence, there was a slight and positive effect of the minimum wage increase, which is only significant, however, at the .10-level. We can conclude that, in this case, increased minimum wages did not undermine employment in the fast-food sector.

9.5.3 Regime Change and Splines

A very useful application of dummy-continuous interactions is the modeling of regime changes and splines. We speak of a change in the regression regime when there is a break in the data such that the regression line up to a certain value of the predictor is different than that afterwards. The change can be in the slope, the intercept, or both. Splines may be viewed as a particular kind of

Figure 9.8: Oil Prices Between January 1961 and December 1978



Note: Price per barrel per month. The red lines show the regression regimes before and after the oil crisis started.

change in the regression regime, whereby the regression lines are connected at the break-point in order to create a smooth function.

Let us consider the idea of regime change for data on the monthly oil prices between January 1961 and December 1978 (see Figure 9.8). It is very clear from the data that oil prices doubled from December 1973 to January 1974 during, what is commonly known as, the oil crisis. Indeed until December 1973, oil prices where on a downward trajectory. Starting in January 1974, they were on an upward trajectory. This is emblematic of a regime change. The two regression regimes are depicted via the red regression lines in Figure 9.8.

The regression lines where constructed using the following conditional expectation function:

$$\mu_t = \beta_0 + \beta_1 D_t + \beta_2 O_t + \beta_3 D_t \times O_t$$

Here, the subscript t denotes a particular month and μ_t is the expected oil price in that month. Further, O_t is a running tally that starts at 0 in January 1961 and increases by 1 in each subsequent month. Finally, D_t takes on the value of 0 prior to January 1974 and the value 1 starting in January 1974. Given the model specification, the predicted oil prices through 1973 are given by

$$\hat{\mu}_t = \hat{\beta}_0 + \hat{\beta}_2 O_t = 23.36 - 0.02 \cdot O_t,$$

whereas the predicted oil prices from January 1974 onward are given by

$$\hat{\mu}_t = \left(\hat{\beta}_0 + \hat{\beta}_1\right) + \left(\hat{\beta}_2 + \hat{\beta}_3\right) O_t = 29.70 + 0.12 \cdot O_t$$

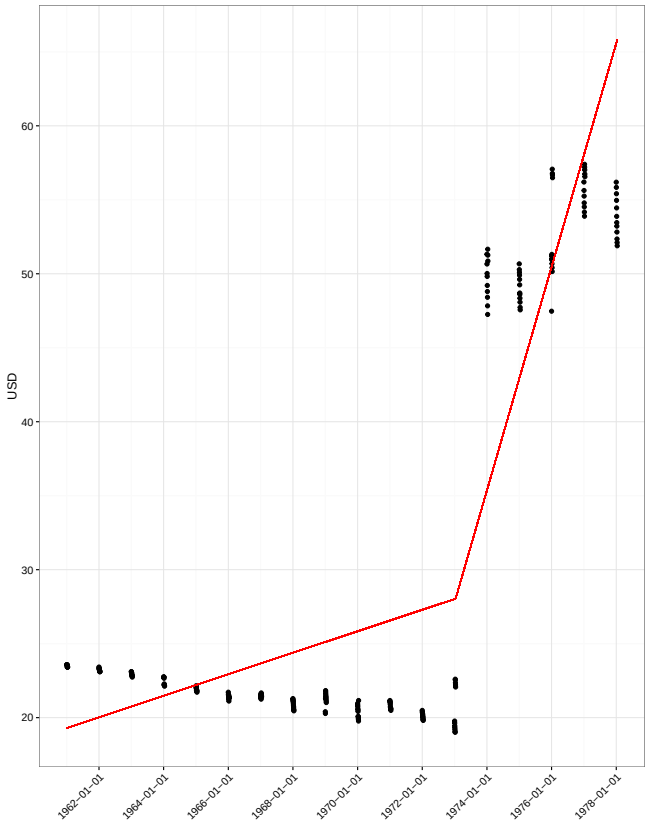
To create a spline function, we impose the restriction that the two regression lines are connected in January of 1974. The connection point, which is at $O = 156$, is called the knot. This brings about a smooth(er) transition, which is important because spline functions should be differentiable at the knot, thus requiring continuity.

We have seen that the conditional expectation function for $O < 156$ is given by $\mu_t = \beta_0 + \beta_2 O_t$. For $O \geq 156$, it is given by $\mu = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) O_t$. Forcing these equations to be identical at the knot means that

$$\begin{aligned}
\beta_0 + \beta_2 \cdot 156 = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \cdot 156 &\Leftrightarrow \\
\beta_0 + \beta_2 \cdot 156 - \beta_0 - \beta_1 - \beta_2 \cdot 156 - \beta_3 \cdot 156 = 0 &\Leftrightarrow \\
-\beta_1 - \beta_3 \cdot 156 = 0 &\Leftrightarrow \\
\beta_1 = -156\beta_3 &
\end{aligned}$$

As we can see, tying the knot amounts to imposing a linear constraint on β_1.

Figure 9.9: A Linear Spline Regression Function



Note: The knot occurs on January 1, 1974.

We can build this constraint directly into the conditional expectation function:

$$\begin{aligned} \mu_t &= \beta_0 \underbrace{-156\beta_3}_{\beta_1} D_t + \beta_2 O_t + \beta_3 D_t \times O_t \\ &= \beta_0 + \beta_2 O_t + \beta_3 D_t \left(O_t - 156\right) \end{aligned}$$

When we estimate this function for the oil price data, we find that $\hat{\beta}_0 = 19.30$, $\hat{\beta}_2 = 0.06$, and $\hat{\beta}_3 = 0.58$. This produces the spline regression function shown in Figure 9.9. We clearly observe the knot that occurs at the start of 1974.

In more general terms, if a variable x has a single knot x^*, then the linear spline regression may be written as

$$\mu_i = \beta_0 + \beta_1 x_i + \alpha_1 d_i (x_i - x^*),$$

where

$$d_i = \begin{cases} 0 & \text{if } x_i \leq x^* \\ 1 & \text{if } x_i \geq x^* \end{cases}$$

If there are multiple knots, then the model may be written as

$$\mu_i = \beta_0 + \beta_1 x_i + \sum_{m=1}^{M} \alpha_m d_{im} (x_i - x_m^*),$$

where

$$d_{im} = \begin{cases} 0 & \text{if } x_i \leq x_m^* \\ 1 & \text{if } x_i \geq x_m^* \end{cases}$$

and M is the total number of knots. One can expand this model even further by adding polynomial terms, which we shall not show here but will generally make the function smoother about the knots.

R makes it extremely easy to include splines in a regression analysis:

```
library(splines)
fit <- lm(price~bs(O, degree=1, knots=c(156)),
data=oil)
summary(fit)
```

The `bs` function generates splines, in this case for O. It does so at the knot $O = 156$ and the spline is linear (since `degree` is equal to 1). The output is shown in Figure 9.10. The first line of the output shows the estimate of the intercept. The second line shows the change in the predicted oil price between January 1961 ($O = 0$) and January 1974 ($O = 156$). Thus, to get to the predicted oil price at the knot, we simply add the estimate from the second

Figure 9.10: R Spline Regression Output

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         19.303      1.024  18.848  < 2e-16 ***
bs(idx, degree = 1, knots = c(156))1   8.737      1.639   5.329 2.51e-07 ***
bs(idx, degree = 1, knots = c(156))2  46.508      1.747  26.624  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: The predicted value at the far left of Figure 9.9 is 19.303. At the knot, it is 19.303 plus 8.737. At the far right, it is 19.303 plus 46.508.

line to the intercept. The third line shows the change in the predicted oil price between January 1961 and the end of the time series ($O = 215$). If we add this number to the intercept, we obtain the predicted oil price in December 1978.

9.6 Conclusions

In this chapter, we have paid extensive attention to interaction effects in the linear regression model. We have seen that such effects can be used to model quite complex relationships indeed. This also ends the introduction of the multiple regression model. You now know pretty much all there is to know about specifying models in such a way that they capture your theoretical ideas. The next task, then, is to return to the regression assumptions and to discuss in greater detail what can be done when they are violated.

Part III

Regression Assumptions and Diagnostics

Chapter 10

Influence, and Normality

In the previous parts of the book, we have seen how one formulates, estimates, evaluates, and interprets regression models. But how can we be sure that the results can be trusted, that they are meaningful? We have seen that the regression model makes a number of assumptions. How do these influence the credibility of the results? How do we know if these assumptions have been violated? What can we do about this? The third part of the book is dedicated to answering these questions.

In this chapter, we start by considering the problem of influence. After one runs a regression model, it is important to check the sensitivity (or, conversely, robustness) of one's results. A sensitivity analysis has several aspects, but one component is to check for the presence of influential data points.[1] This check is used to ensure that the regression results are not driven entirely by one or a few atypical observations.

We conclude the chapter by discussing the normality assumption. The presence of influential data points sometimes (but not always) implies that the normality assumption is violated. We explore the implications of such violations, methods for detecting them, and remedies.

By the end of this chapter, you will have learned the first set of diagnostic and remedial measures. Even more importantly, you will begin to understand the

[1] Other components include the sensitivity to model specification, a topic we shall take up elsewhere in this book.

crucial role that residuals play in uncovering problems with regression models.

10.1 Influential Observations

10.1.1 Defining the Problem

The concept of influence rests on two related ideas: leverage and outlier.

- **Leverage**: An observation has leverage—or is a **leverage point**—if its value on the *predictor* is atypical compared to other observations.
- **Outlier**: An observation is an outlier if its value on the *dependent variable* is atypical compared to other observations.

An observation has **influence** when it its presence or absence brings about large shifts in the regression results, especially, in the partial slope coefficients. Such influence comes about when the data point is both an outlier and has leverage. Logically,

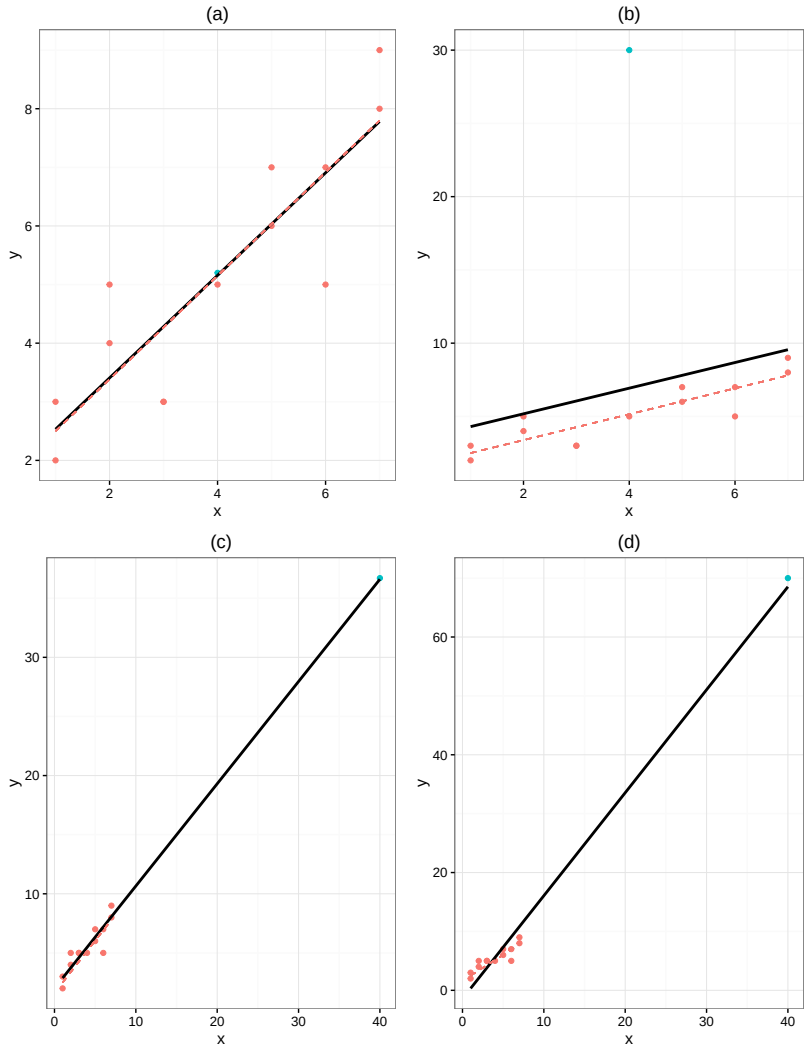
$$\text{Influence} \quad = \quad \text{Leverage} \times \text{Outlier}$$

The principle is illustrated in Panels (a)-(d) in Figure 10.1, which show the impact on the regression line of points that are leverage points, outliers, both, or neither.

Panel (a) of Figure 10.1 shows the effect on the regression line of a point that is neither an outlier nor a leverage point. This point is colored blue (all the other points are colored pink). The point is not an outlier because it is situated right around the mean of y ($\bar{y} = 5.2$). The point lacks leverage because it coincides with the mean of x ($\bar{x} = 4.0$). The black line in the plot is the regression line when we include all data points; its sample regression function is $\hat{y}_i = 1.65 + 0.88x_i$. The dashed pink line is the regression line when we drop the blue point. It, too, obliges the sample regression function $\hat{y}_i = 1.65 + 0.88x_i$. We see that the two regression lines are indistinguishable, so that the blue point lacks influence.

Panel (b) shows the effect on the regression line of a point that is an outlier but lacks leverage. This point is again colored blue in the plot. It is an outlier

Figure 10.1: Leverage Points, Outliers, and Influence



Note: The blue point is potentially anomalous. In panel (a) it is neither an outlier nor a leverage point. In panel (b) it is an outlier but not a leverage point. In panel (c) it is a leverage point but not an outlier. Finally, in panel (d), it is both an outlier and a leverage point.

because it is situated about 3 standard deviations away from the mean of y ($\bar{y} = 6.9$, $s = 7.0$). The point lacks leverage because it again coincides with the mean of x. The black line in the plot is the regression line when we include all data points. The dashed pink line is the regression line when we drop the blue data point. We see that the blue and red regression lines are parallel. Thus, the inclusion of the outlier influences the estimate of the constant, which changes from 1.65 (pink line) to 3.43 (black line). However, it exerts no effect on the slope coefficient, which is equal to 0.88 regardless of whether the blue points is included or excluded.

Panel (c) shows the effect of a leverage point that is not an outlier. The blue point is over three standard deviations removed from the mean of x ($\bar{x} = 6.6$, $s = 9.8$) and, as such, has leverage. But the point is not all that anomalous on the dependent variable. Its value on y is precisely where we would expect it to be based on the pattern in the other observations. The dashed pink regression line depicts this pattern. As you can see, the blue point falls right on this regression line, when it is extrapolated. It is no surprise, then, that adding the blue point to the analysis produces a regression line, drawn here in black, that is virtually indistinguishable from the pink line. In other words, even though the observation has leverage, it does not seem to affect the slope (or even the intercept) of the regression line.

Finally, consider panel (d). Here, the blue data point has leverage and also is an outlier. We see that the regression line is very different depending on whether this data point is included or excluded. The dashed pink regression line, which describes the remaining data points well, has a much shallower slope than the black regression line, which comes about when we add in the anomalous observation. More specifically, the sample regression function omitting the blue point is $\hat{y}_i = 1.65 + 0.88x_i$. By contrast, the sample regression function for the full data set is $\hat{y}_i = -1.42 + 1.75x_i$.

Figure 10.1 reveals the importance of leverage points and outliers. They also show that it is the combination of being an outlier and leverage point that generates influence on the slope of the regression line. Our next task is to determine how leverage points, outliers, and influential data points can be detected.

10.1.2 Diagnosing Influence

Detecting Leverage Points

Hat values are the most common diagnostic of leverage. These are the diagonal values of the hat matrix that we encountered in Chapter 4. As we saw in equation 4.11, the hat matrix is the $n \times n$ matrix

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$$

The diagonal elements of this matrix are the hat values, h_{ii}, which range between 0 and 1.[2]

Why do hat values indicate something about leverage? The answer is that the hat matrix plays a critical role in transforming the dependent variable into predicted values. As we have also seen, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. Large values on the diagonal of the hat matrix have the effect of giving a lot of weight to a particular score on the dependent variable. We also see that if the value on y is an outlier, then the product with the leverage value can be quite large. This would signify influence.

How do we judge hat values? That is, when is a hat value sufficiently large to single out an observation as a leverage point? To answer this question, it is useful to remember the following important property of hat values:

$$\sum_i h_{ii} = K + 1$$

This implies that

$$\bar{h}_{ii} = \frac{K+1}{n},$$

where $\bar{h}_{ii} < 1$ is the average hat value. This average is frequently used to evaluate the size of hat values. A common criterion is that a leverage point is defined as any point for which $h_{ii} > 2\bar{h}_{ii}$. Complementing this criterion are

[2] We will not spend any time exploring the off-diagonal elements of $\mathbf{H}$, but I should note that $-.5 < h_{ij} < .5$ for $i \neq j$.

Table 10.1: Data and Hat Values from Panel (c) of Figure 10.1

i	x	y	h
1	1.0	2.0	0.096
2	1.0	3.0	0.096
3	2.0	5.0	0.088
4	2.0	4.0	0.088
5	3.0	3.0	0.082
6	3.0	3.0	0.082
7	40.0	36.7	0.959
8	4.0	5.0	0.077
9	5.0	6.0	0.073
10	5.0	7.0	0.073
11	6.0	7.0	0.072
12	6.0	5.0	0.072
13	7.0	8.0	0.071
14	7.0	9.0	0.072

Note: $h =$ hat value.

guidelines that we should consider as a large hat value anything above .5.

We can illustrate the computation and judgment of hat values using the data that produced panel (c) of Figure 10.1; these data are shown in the second and third column of Table 10.1. In R, we obtain these hat values via the following command:

```
hatvalues(lm.object)
```

where `lm.object` contains the regression results. The average hat value in the data is .143. It is obvious that h_{77} is much greater than this (almost 7 times greater). Thus, we would identify this observation as a leverage point. None of the other observations produce large hat values, so that the 7th observation is the only leverage point that we identify.

Detecting Outliers

The detection of outliers typically involves an analysis of the residuals. Several types of residuals are relevant in this regard: raw, internally studentized, PRESS, and externally studentized residuals.

Raw Residuals Raw residuals, which are also known as response residuals (Fox and Weisberg, 2011), are the type of residuals that we first encountered in Chapter 1. They are defined as $e_i = y_i - \hat{y}_i$ or, in matrix form,

$$\begin{aligned} \mathbf{e} &= (\mathbf{I} - \mathbf{H})\,\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\,\boldsymbol{\varepsilon} \end{aligned}$$

(see Chapter 4). They have a number of important properties:

1. Their theoretical average is zero: $E[e_i] = 0$.
2. If ε_i follows a normal distribution, so does e_i.[3]
3. Let Var$(\varepsilon_i) = \sigma^2$, then Var$(e_i) = \sigma^2(1 - h_{ii})$.

These properties will prove important for diagnostic purposes.

In R, the raw residuals can be obtained by running

```
residuals(lm.object)
```

An example can be found in the fourth column of Table 10.2. This example also shows two problems with raw residuals. First, for diagnostic purposes, it may be difficult to tell if a residual is large because raw residuals do not have a standardized metric. Second, when we consider the raw residual of the 7th observation—the blue data point in panel (d) of Figure 10.1—it is not particularly large. Based on the raw residual, we would never mark this observation as an outlier, even though an inspection of its score on the dependent variable makes clear that it is. The reason that the residual for the 7th observation is so small is exactly the influence that it exerts. Through this influence, it has pulled

[3] After all, a linear function of a normal variable is itself normally distributed.

Table 10.2: Data and Residuals from Panel (d) of Figure 10.1

i	x	y	e	r	p	r^*
1	1.0	2.0	1.7	0.8	1.8	0.8
2	1.0	3.0	2.7	1.2	3.0	1.3
3	2.0	5.0	2.9	1.4	3.2	1.4
4	2.0	4.0	1.9	0.9	2.1	0.9
5	3.0	3.0	1.2	0.5	1.3	0.5
6	3.0	3.0	1.2	0.5	1.3	0.5
7	40.0	70.0	1.5	3.2	35.6	8.0
8	4.0	5.0	-0.6	-0.3	-0.6	-0.3
9	5.0	6.0	-1.3	-0.6	-1.4	-0.6
10	5.0	7.0	-0.3	-0.1	-0.3	-0.1
11	6.0	7.0	-2.1	-1.0	-2.2	-0.9
12	6.0	5.0	-4.1	-1.9	-4.4	-2.1
13	7.0	8.0	-2.8	-1.3	-3.0	-1.3
14	7.0	9.0	-1.8	-0.8	-2.0	-0.8

Note: e = raw residual; r = internally studentized residual; p = PRESS residual; and r^* = externally studentized residual.

the regression line toward itself, so much so that the prediction and actual value of the dependent variable are very close. Precisely because of these limitations, it is useful to consider some transformations of the raw residuals.

Internally Studentized Residuals To place the residuals on a bounded scale, we can transform them into so-called internally studentized residuals.[4] This is done by dividing them by their standard deviation:

[4] Fox and Weisberg (2011) call this the standardized residual, but that term is also reserved for another form of standardization, where the raw residuals are divided by $\sqrt{MSE}$. Yet another variant is the normalized residual, which divides the raw residual by the maximum likelihood estimator of σ.

Equation 10.1: Internally Studentized Residuals

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}},$$

where $s = \sqrt{MSE}$ is the unbiased estimator of σ. The scale of r is bounded by $\pm\sqrt{n-K-1}$. This allows one to judge the size of the residual.

In R, the internally studentized residuals are obtained by running

```
rstandard(lm.object)
```

This produces the residuals shown in the fifth column of Table 10.2. With 14 observations and one predictor, $-3.46 \leq r \leq 3.46$. Judged by this metric, we actually see a relatively large internally standardized residual for the 7th observation. So standardization has also helped to draw attention to the problematic data point. Why this is the case will make more sense once we have looked at PRESS residuals.

PRESS Residuals Researchers often want to know what the residuals are once a particular observation has been removed from the regression. One variant of this is PRESS, the prediction error sums of squares residual:[5]

Equation 10.2: PRESS Residuals

$$\begin{aligned} p_i &= y_i - \hat{y}_{i(i)} \\ &= \frac{e_i}{1-h_{ii}} \end{aligned}$$

Here $\hat{y}_{i(i)}$ is the prediction for the ith observation when that observation is not considered in the computation of the regression coefficients. The formula is derived in Appendix C.6.

[5] In the literature, the prediction error sum of squares is given by $\sum_i p_i^2$.

There exists a simple relationship between the internally studentized and PRESS residuals:

$$r_i = \frac{\sqrt{1-h_{ii}}}{s} p_i$$

It is no wonder, then, that the internally studentized residuals pick up the effect of the deletion of a particular data point.

To obtain PRESS residuals in R, we need to use the qpcR library. The syntax is

```
library(qpcR)
PRESS(lm.object, verbose=TRUE)
```

This produces, among other things, the PRESS residuals. If we only want the residuals, we can type `PRESS(lm.object, verbose=TRUE)$residuals`. For the data from panel (d) in Figure 10.1, the p_is are shown in the sixth column of Table 10.2. Here, we clearly see the problematic nature of the 7th data point. If we estimate the regression model without this point, the predicted value falls nearly 36 points short off the actual response. This is a clear indication that the 7th observation is an outlier.

Externally Studentized Residuals We have seen that the internally studentized residuals, through their relationship with the PRESS residuals, show what would happen when we delete a data point. However, these residuals may still mask outliers because one of their ingredients, s, is still based on the totality of the data. To overcome this problem, we can compute externally studentized residuals, which are also known as R residuals:

Equation 10.3: Externally Studentized Residuals

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_i i}} \tag{10.1}$$

Here, $s_{(i)}$ is the standard deviation of the regression after omitting the ith

observation. It is equal to the square root of

$$s^2_{(i)} = \frac{(n-K-1)s^2 - \frac{e_i^2}{1-h_{ii}}}{n-K-2}$$

These residuals follow a student's t-distribution with $n-K-2$ degrees of freedom.

In R, the computation of the externally studentized residuals proceeds using

```
rstudent(lm.object)
```

These residuals are shown in the last column of Table 10.2. With $n = 14$ and $K = 14$, the t-distribution has the following characteristics:

- Ninety-nine percent of the R residuals lie between ± 3.464.
- Ninety-five percent lie between ± -2.201.
- Ninety percent lie between ± 1.796.

Inspecting the entries in Table 10.2, we can clearly see that all residuals are situated between the limits of the ninety percent confidence interval, with one exception. This exception is the 7th observation, whose externally studentized residual is 8.00, thus exceeding the limits of even the 99% confidence interval. There is no doubt that this observation is an outlier.

Detecting Influence

Now that we know how to assess whether a point is an outlier or has leverage, the next question is how do we assess influence. Here, we need to distinguish between varieties of influence. Does a point influence the predicted values, the coefficients, efficiency, or what? We now consider a series of measures that help to answer these questions.

DFFITS One standard by which to judge influence is to assess the impact of including an observation on the predicted values of the other observations. The DFFITS measure (i.e., difference in fitted values)—also known as the Welsch-Kuh distance—does precisely this. The measure is computed ass[6]

Equation 10.4: DFFITS

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}}$$

Here

$$\hat{y}_i - \hat{y}_{i(i)} = \frac{h_{ii}e_i}{1 - h_{ii}}$$

This may also be written in terms of the externally studentized residuals:

$$\text{DFFITS}_i = r_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

(see Appendix C.6). There are as many DFFITS values as there are observations. We can think of these values as scaled differences $\hat{y}_i - \hat{y}_{i(i)}$. A large value on the DFFITS implies that dropping an observation from the estimation will dramatically change the predicted value for that observation. This is an indication of great influence. The conventional cutoff is to consider as influential those points whose absolute DFFITS are greater than 1 or 2.

In R, the DFFITS values are obtained using

```
dffits(lm.object)
```

For panel (d) in Figure 10.1, the values are shown in Table 10.3. We see that the only DFFITS value that exceeds the threshold occurs for the 7th observation. This can be considered an influential data point in terms of the predictions.

[6] Sometimes the absolute value of $\hat{y}_i - \hat{y}_{i(i)}$ is taken to define the DFFITS (see Chatterjee and Hadi, 1988).

Table 10.3: Influence Statistics for Panel (d) of Figure 10.1

i	DFFITS$_i$	$\hat{\beta}_1 - \hat{\beta}_{1(i)}$
1	0.249	-0.008
2	0.415	-0.013
3	0.437	-0.012
4	0.274	-0.008
5	0.156	-0.004
6	0.156	-0.004
7	38.398	0.945
8	-0.073	0.001
9	-0.166	0.002
10	-0.040	0.000
11	-0.263	0.001
12	-0.590	0.002
13	-0.371	-0.001
14	-0.229	-0.001

Note: Based on the data in Table 10.2

Change in Coefficients Another criterion by which we can judge influence is an observation's impact on the regression coefficients. One way to judge this is to compute the change in the parameter estimates when an observation is deleted:

Equation 10.5: Change in the Regression Coefficients

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i e_i}{1 - h_{ii}}$$

(see Kmenta, 1997). Here $\hat{\boldsymbol{\beta}}_{(i)}$ is the estimator after the ith observation has been omitted. The statistic is reported in the third column of Table 10.3. We see that the changes in the coefficients are small for all of the observations, excepting the 7th observation. Here, exclusion of the regression line causes a

big change in the regression slope.[7]

Although Equation 10.5 is very useful, another measure has gained more prominence: DFBETA. This can be seen as a scaled version of the raw differences between the estimates over the full data and the estimates obtained after dropping a single observation. Computationally,

Equation 10.5: Change in the Regression Coefficients

$$\text{DFBETA}_{ik} = \frac{r_i^* w_{ik}}{\sqrt{1-h_{ii}}\sum_i w_{ik}^2}$$

(see Chatterjee and Hadi, 1988, for a derivation). Here k denotes a particular element in the vector $\hat{\boldsymbol{\beta}}$. Further, w_{ik} is the residual that arises when x_k is regressed on all of the other predictors in the model. This formula produces a vector of DFBETAS for each predictor in the model. An element in each of these vectors indicates how much the regression coefficient for that predictor would change if the observation were deleted from the analysis. Belsley, Kuh and Welsch (1980) suggest that points are influential if $|DFBETAS_{ik}| > 2/n$.

[7] The results were obtained using `dfbeta` in R. This should not be confused with `dfbetas`, which computes DFBETA.

Appendices

Appendix A

Basics of Differentiation

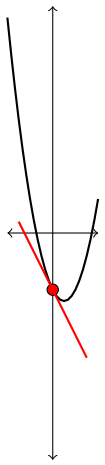
Derivatives make a common appearance in statistics. In this appendix, we provide a brief review derivatives and their use. The key is that you become familiar with the notation as well as the logic of differentiation and optimization. There is no expectation that you will become an expert at doing complex derivatives on your own. Fortunately, most of the derivatives that we need are quite simple, so that even doing them by hand will not create major headaches.

A.1 Definition

As is shown in Figure A.1, he derivative gives the slope of the tangent line at a particular point for some function. This can be used to find the extremes (minimum and maximum) of a function but also to characterize the rate of change, i.e., the sensitivity of the dependent variable to a change in the independent variable of a function at a particular point. Both uses are important for linear regression analysis. The rate of chance helps us to interpret the regression function, whereas using derivatives for the purpose of finding extremes is essential for least squares and maximum likelihood estimation.

By defining the derivative as the slope of the tangent at a particular point, we learn two things immediately. First, we know that we are dealing with a change in y relative to a change in x. Second, the change in x is infinitesimally small, so that we approach a point. Thus follows the formal definition of the derivative: for a function $y = f(x)$,

Figure A.1: The Slope of the Tangent



the derivative with respect to x is given by

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}$$

This actually is the derivative stated in Lagrange's notation. Leibniz used a different notation and indicated the derivative as dx/dy. More precisely, this is known as the *first derivative*.

We can apply the formula for $f'(x)$ to derive an expression of the derivative of any function. As an example, consider the quadratic equation $f(x) = ax^2 + bx + c$. Let Δx denote the change from x to $x + h$, where h goes to 0. Then

$$\begin{aligned}
\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} &= \lim_{h \to 0} \frac{ax^2 + 2ahx + ah^2 + bx + bh + +c}{h} - \\
&\quad \lim_{h \to 0} \frac{ax^2 + bx + c}{h} \\
&= \lim_{h \to 0} \frac{2ahx + ah^2 + bh}{h} \\
&= \lim_{h \to 0} 2ax + ah + b \\
&= 2ax + b
\end{aligned}$$

Table A.1: Useful Derivatives

Function	First Derivative
$f(x) = a$	$f'(x) = 0$
$f(x) = bx$	$f'(x) = b$
$f(x) = ax^b$	$f'(x) = abx^{b-1}$
$f(x) = e^{bx}$	$f'(x) = be^{bx}$
$f(x) = a^{bx}$	$f'(x) = ba^{bx} \ln a$
$f(x) = \ln x$	$f'(x) = 1/x$
$f(x) = log_b x$	$f'(x) = 1/(x \ln b)$

A.2 Important Derivatives

Table A.1 summarizes the derivatives for a number of important functions that we encounter in this book. We can also combine these functions. In most cases, this is done by adding terms. In this case, it is very easy to find the derivative because the derivative of a sum is equal to the sum of the derivatives. Consider again the generic quadratic function $f(x) = ax^2 + bx + c$. Using Leibniz' notation,

$$\frac{df(x)}{dx} = \frac{dax^2}{dx} + \frac{dbx}{dx} + \frac{dc}{dx}$$

From the first row of Table A.1, we know that $dc/dx = 0$. From the second row, we know that $dbx/dx = b$. Finally, from the third row we know that $dax^2/dx = 2ax^{2-1} = 2ax$. Thus, $df(x)/dx = 2ax + b$, the result that we derived before.

A.3 Higher-Order Derivatives

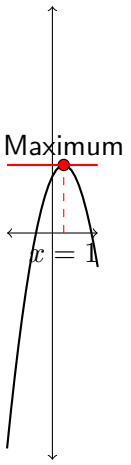
So far, we have differentiated a function once with respect to a variable. There is, however, no reason why we could not differentiate multiple times. For our purposes, it suffices to differentiate no more than two times. The so-called second derivative can be seen as the derivative of the first derivative. In Lagrange's notation,

$$f''(x) = (f'(x))'$$

In Leibniz' notation, the second derivative is written as d^2f/dx^2.

Consider again the function $f(x) = ax^2 + bx + c$. We have already seen that $f'(x) = 2ax + b$. Then, $f''(x)$ is the derivative of $f'(x)$ with respect to x. Using the

Figure A.2: Identifying a Maximum



rules developed in the previous section it is easy to show that $f''(x) = 2a$.

A.4 Function Analysis

Derivatives are extremely useful in function analysis, especially for purposes of finding the minimum or maximum of a function. The reason is simple: as is illustrated in Figure A.2, the slope of the tangent line is zero at a minimum or maximum and this, in turn, means that the first derivative is zero. Thus, we can detect extreme values such as minimums and maximums simply by taking the first derivative, setting it to zero, and solving for x.

As an example, consider the function $f(x) = -x^2 + 2x + 5$. This is the function shown in Figure A.2. Its first derivative is $f'(x) = -2x + 2$. We now set this to 0: $-2x + 2 = 0$. This is known as the *first order condition* for a minimum/maximum. When we solve for x, we obtain $x = 1$ as the point where the extreme value occurs.

In our example, the graph clearly shows that the extreme in this case is a maximum. The first order condition by itself does not provide such detailed information. All it tells us is that there is an extreme value of some kind at $x = 1$. If we want to know in greater detail what kind of extreme this is, then we have to rely on the *second order condition*. Here, we examine the second derivative at the extreme and explore its sign.

Specifically,[1]

$$\begin{array}{ll} f''(x) > 0 & \text{Minimum} \\ f''(x) < 0 & \text{Maximum} \end{array}$$

In our example, $f''(x) = -2 < 0$, so that the second order condition tells us that $x = 1$ is a maximum.

A.5 Partial Derivatives

In regression analysis, most of the time our functions involve multiple variables. In this context, we need to conceptualize derivatives as *partial* derivatives:

> The first partial derivative of $f(x_1, x_2, \cdots, x_K)$ with respect to x_k is the rate of change in the dependent variable as a result of changing x_k at a particular point, while holding everything else constant. It is written as $\partial f / \partial x_k$

Holding constant literally means that, when we take the derivative with respect to x_k, all of the other variables are treated as if they were constants.

Consider the following example: $f(x, y) = x^2 + 2xy - y^3$. When we take the partial derivative with respect to x, we may rewrite the function as $x^2 + ax - c$, where $a = 2y$ and $c = y^3$ are treated as constants. Using the rules of differentiation that we developed earlier, the first partial derivative is

$$\frac{\partial f}{\partial x} = \frac{\partial x^2}{\partial x} + \frac{\partial ax}{\partial x} - \frac{\partial c}{\partial x} = 2x + a - 0$$

Making the appropriate substitutions, this becomes $\partial f / \partial x = 2x + 2y$. When we take the partial derivative with respect to y, then the function may be rewritten as $a + by - y^3$, where $a = x^2$ and $b = 2x$ are constants. Hence,

$$\frac{\partial f}{\partial y} = \frac{\partial a}{\partial y} + \frac{\partial by}{\partial y} - \frac{\partial y^3}{\partial y} = 0 + b - 3y^2$$

Making the appropriate substitutions, we get $\partial f / \partial y = 2x - 3y^2$.

When we seek the minimum or maximum of a multivariate function, then the first order condition states that all of the partial derivatives should be simultaneously 0. For our example, this produces two extreme values: $x = 0, y = 0$ and $x = 2/3, y = -2/3$.

[1] A third scenario is that $f''(x) = 0$. This is a necessary but not a sufficient condition for an inflection point. Outside of polynomial regression, we generally do not have to worry about inflection points so that we shall skip this topic.

To identify whether the extreme values constitute a minimum, maximum, or something else, we would need to apply the second derivative test. However, this gets to be quite complicated because there are several second derivatives that we can compute and all of these are relevant for the second derivative test. Specifically, we can compute the following second derivatives:

- $\frac{\partial^2 f}{\partial x^2}$: Here we first differentiate with respect to x and we then take the derivative and differentiate it once more with respect to x.
- $\frac{\partial^2 f}{\partial x \partial y}$: Here we first differentiate with respect to x and we then take the derivative and differentiate it with respect to y.
- $\frac{\partial^2 f}{\partial y \partial x}$: Here we first differentiate with respect to y and we then take the derivative and differentiate it with respect to x.
- $\frac{\partial^2 f}{\partial y^2}$: Here we first differentiate with respect to y and we then take the derivative and differentiate it once more with respect to y.

In our example, this produces the following second derivatives: (1) $\partial^2 f/\partial x^2 = 2$; (2) $\partial^2 f/\partial x \partial y = 2$; (3) $\partial^2 f/\partial y \partial x = 2$; and (4) $\partial^2 f/\partial y^2 = -6y$.

For the second order condition, we would now have to place all of these derivatives into a matrix. We then would compute the eigenvalues of that matrix and consider their signs. However, this topic lies well beyond the scope of this course, so that we shall generally not perform second derivative tests in the multivariate case.

Appendix B

Basics of Matrix Algebra

Matrices provide a useful shorthand in multivariate statistics such as multiple linear regression analysis. A matrix can capture a great deal of information such as all of the data that we have on the predictors in a model. Mathematical operations on the information can then be performed at the level of the matrices, as opposed to their individual elements. Apart from economy of notation, matrices thus offer computational advantages. In this appendix, we develop the basic intuitions of matrix algebra. The emphasis is on notation and on those operations that are useful for understanding multiple regression analysis.

B.1 The Matrix Concept

B.1.1 Definition

For our purposes, a matrix may be defined as a rectangular array of numbers. This array is characterized by r rows and c columns. It contains $r \times c$ numbers. Accordingly, we say that the matrix is of *order* $r \times c$. It is customary to denote matrices with capital bold letters such as $\mathbf{X}$.

The most obvious example of a matrix in statistics is the data matrix. Imagine we collect data on the Conservatives, Labour, Liberal Democrats, and UKIP in Great Britain. In addition to recording the party name, we have data on the left-right position of these parties as well as their support for European integration. Support for European integration is measured on a 7-point scale that runs from "strongly oppose" to "strongly support". On this scale, the Conservatives score 2.3, Labour scores 4.8, the Liberal Democrats score 6, and the UKIP scores 1. The left right position is measured on an

11-point scale, which runs from "extreme left" to "extreme right". On this scale, the Conservatives score 7.1, Labour scores 4.5, the Liberal Democrats score 5, and UKIP scores 8.8.[1] We can collect these data in the following matrix that is of order 4×2:

$$\mathbf{D} = \begin{pmatrix} 2.3 & 7.1 \\ 4.8 & 4.5 \\ 6.0 & 5.0 \\ 1.0 & 8.8 \end{pmatrix}$$

This is now our data matrix, which may also be written as $\mathbf{D}_{4\times 2}$ to indicate its order explicitly. But data matrices are only one type of matrix that we encounter in statistics and we shall see many other examples later in this appendix.

We treat $\mathbf{D}$ as a single object—this is what brings the economy of notation—but we can always access individual elements, i.e., the individual numbers. For example, if I want to know the left-right score of UKIP, then all I need is to point to the element in the fourth row and second column. We call this element $d_{4,2}$ or, in R notation, $D[4, 2]$.

B.1.2 Types of Matrices

Scalar A scalar is a single number. We can think of this as an order 1×1 matrix.

Vectors A matrix consisting of a single row or column is called a vector. We speak of a row vector when the matrix consists of a single row. When it consists of a single column, then we call it a column vector. We generally denote vectors through lowercase boldface letters such as $\mathbf{x}$.

Returning to the previous example, it may be of interest to us to capture all of our data about UKIP. This can be done by creating the row vector

$$\mathbf{d} = \begin{pmatrix} 1.0 & 8.8 \end{pmatrix}$$

Similarly, we may be interested in creating an object containing all of our data about left-right party placements. This could be the column vector

$$\mathbf{d} = \begin{pmatrix} 7.1 \\ 4.5 \\ 5.0 \\ 8.8 \end{pmatrix}$$

[1] These data come from the 2010 Chapel Hill Expert Survey.

In this book, we follow the convention that all vectors are taken to be column vectors.

Square Matrices If a matrix has as many rows as it has columns, we say that it is square. A prime example of this is the matrix of sums of squares and cross-products, which plays an important role in least squares estimation.

Symmetric Matrices A matrix $\mathbf{A}$ is said to be symmetric if it is square and if $a_{ji} = a_{ij}$ for all $j \neq i$. For example, the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

is symmetric because $a_{21} = a_{12} = 0$, $a_{31} = a_{13} = 2$, and $a_{32} = a_{23} = 3$. A good example of a symmetric matrix is the covariance matrix.

Diagonal Matrices A diagonal matrix is a square matrix for which all of the off-diagonal elements are zero. The diagonal elements are not all zero.

The Identity Matrix The identity matrix is a special kind of diagonal matrix. Here, all of the diagonal elements are equal to 1 and the off-diagonal elements are equal to 0. This matrix is extremely important in matrix algebra, as it plays a role that is similar to that of the scalar 1 in ordinary algebra. We indicate the identity matrix as $\mathbf{I}$. Sometimes, we add a subscript to this to indicate the number of rows/columns. For example,

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Partitioned Matrices A partitioned matrix is a matrix whose elements are themselves matrices of sorts. For example, our data matrix about British political parties may be partitioned as

$$\mathbf{D} = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \end{pmatrix}$$

Here, $\mathbf{d}_1$ and $\mathbf{d}_2$ are column vectors of EU party positions and left-right party positions, respectively.

Block-Diagonal Matrices A block-diagonal matrix is a partitioned matrix such that the off-diagonal elements are matrices that consist entirely of zeros. For example,

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

is a block-diagonal matrix, as it can be written as

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 \end{pmatrix}$$

Here, $\mathbf{C}_1$ is a 2×2 matrix with 1s on the diagonal and 2s on the off-diagonal, $\mathbf{C}_2$ is a 2×2 matrix with 2s on the diagonal and 1s on the off-diagonal, and $\mathbf{0}$ is a 2×2 matrix consisting entirely of 0s. Because of this, it is often referred to as the *null matrix*.

B.2 Matrix Operations

It is possible to conduct various mathematical operations on matrices. There are similarities between these operations and the operations conducted in scalar algebra, but one cannot take the parallels too far. Some operations in matrix algebra such as transposition, for example, do not have an equivalent operation in scalar algebra. Other operations common in scalar algebra such as division do not have an equivalent in matrix algebra. In general, operations in matrix algebra do not work quite the same as those in scalar algebra.

B.2.1 Transpose of a Matrix

When we transpose a matrix, we sort of flip it on its side. What used to be rows now become columns and vice versa. If $\mathbf{A}$ is an $r \times c$ matrix, then the transpose, $\mathbf{A}^\top$, is a $c \times r$ matrix. The ith row in $\mathbf{A}$ becomes the ith column in $\mathbf{A}^\top$ and the jth column in $\mathbf{A}$ becomes the jth row in $\mathbf{A}^\top$.

To illustrate this, let us consider the 2×3 matrix

$$\mathbf{A} = \begin{pmatrix} -1 & 0 & 1 \\ 0 & 2 & 0 \end{pmatrix}$$

The transpose is now the 3×2 matrix

$$\mathbf{A}^\top = \begin{pmatrix} -1 & 0 \\ 0 & 2 \\ 1 & 0 \end{pmatrix}$$

We clearly see that the 1st row in $\mathbf{A}$ has become the 1st column in $\mathbf{A}^\top$. Similarly, the 2nd row in $\mathbf{A}$ is now the 2nd column in $\mathbf{A}^\top$, We shall use the transpose operator frequently, for example, to switch back and forth between row and column vectors.

There are no restrictions on the transpose; any matrix can be transposed. By transposing a matrix twice over, we reverse to the original matrix: $(\mathbf{A}^\top)^\top = \mathbf{A}$. If $\mathbf{A}$ is symmetric, then $\mathbf{A}^\top = \mathbf{A}$. Other properties of the transpose will be discussed with other matrix operations.

B.2.2 Matrix Addition and Subtraction

Just like two scalars can be added or subtracted, it is also possible to add and subtract matrices. Whereas any two scalars can be added or subtracted, however, not every pair of matrices can be added or subtracted. Matrix addition and subtraction are possible only if they are *conformable*. For matrix addition and subtraction the conformability condition is:

> $\mathbf{A} + \mathbf{B}$ and $\mathbf{A} - \mathbf{B}$ exist if and only if $\mathbf{A}$ and $\mathbf{B}$ are of the same order.

If conformability holds, then the matrix addition will produce a new matrix, $\mathbf{C}$ with the following properties:

> (1) $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is of the same order as $\mathbf{A}$ and $\mathbf{B}$. (2) Element $c_{ij} \equiv a_{ij} + b_{ij}$.

For matrix subtraction, the properties are:

> (1) $\mathbf{C} = \mathbf{A} - \mathbf{B}$ is of the same order as $\mathbf{A}$ and $\mathbf{B}$. (2) Elements $c_{ij} \equiv a_{ij} - b_{ij}$.

As an example, consider the following three matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \; \mathbf{B} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \; \mathbf{C} = \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}$$

The sum $\mathbf{A} + \mathbf{B}$ does not exist because the two matrices are not of the same order. By contrast, $\mathbf{A}$ and $\mathbf{C}$ are conformable for addition. When we add these two matrices,

we get

$$\begin{aligned}\mathbf{A}+\mathbf{C} &= \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{pmatrix} = \begin{pmatrix} 1+2 & 2+1 \\ 3+2 & 4+1 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 3 \\ 5 & 5 \end{pmatrix}\end{aligned}$$

The difference $\mathbf{C}-\mathbf{B}$ does not exist due to the difference in the order of these matrices. However $\mathbf{C}$ is conformable with $\mathbf{A}$ for subtraction. The result is:

$$\begin{aligned}\mathbf{C}-\mathbf{A} &= \begin{pmatrix} c_{11}-a_{11} & c_{12}-a_{12} \\ c_{21}-a_{21} & c_{22}-a_{22} \end{pmatrix} = \begin{pmatrix} 2-1 & 1-2 \\ 2-3 & 1-4 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -1 \\ -1 & -3 \end{pmatrix}\end{aligned}$$

Matrix addition and subtraction have several useful properties. One property that we shall use frequently is the following:

> The transpose of a sum (or difference) of two matrices is equal to the sum (or difference) of the transposes.

Thus, $(\mathbf{A}+\mathbf{B})^\top \equiv \mathbf{A}^\top + \mathbf{B}^\top$ and $(\mathbf{A}-\mathbf{B})^\top \equiv \mathbf{A}^\top - \mathbf{B}^\top$, where it is assumed that conformability holds.

B.2.3 Matrix Multiplication

Whereas the multiplication of scalars is relatively straightforward, the same cannot be said for matrix multiplication. One complication is that it is actually possible to define different kinds of matrix products. For our purposes, three products are particularly useful: (1) scalar products; (2) inner products; and (3) matrix products.

Scalar Product Scalar products can be obtained for any matrix. They are defined as the product of a scalar into a matrix. This results in the multiplication of every element of the matrix with the scalar. Thus, the following definition applies.

> Consider a scalar k and a $r \times c$ matrix $\mathbf{A}$. The scalar product $\mathbf{B} = k\mathbf{A}$ is a $r \times c$ matrix with elements $b_{ij} \equiv ka_{ij}$.

As an example, consider the 2×2 identity matrix and the scalar $k = 3$. Then

$$k\mathbf{I}_2 = \begin{pmatrix} 3 \cdot 1 & 3 \cdot 0 \\ 3 \cdot 0 & 3 \cdot 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$$

Inner-Product The inner-product is the product of a row vector into a column vector of equal length, which results in a scalar as the outcome. For the product to exist, the row and column vector have to have the same number of elements.

> Let $\mathbf{u}^T$ be a row vector consisting of K elements. Let $\mathbf{v}$ be a column vector, also consisting of K elements. The inner-product $\mathbf{u}^\top\mathbf{v}$ produces a scalar
>
> $$s = \sum_{i=1}^{K} u_{1i}^\top v_{i1}$$

Note that we transpose $\mathbf{u}$ because, by convention, $\mathbf{u}$ is a column vector.

What the formula for s amounts to is that we take the product of the first element of $\mathbf{u}^\top$ and the first element of $\mathbf{v}$. We then add to that the product of the second element of $\mathbf{u}^\top$ and the second element of $\mathbf{v}$ and so on and so forth. For example, let $\mathbf{u}^\top = (1\ 2\ 0)$ and let $\mathbf{v}^\top = (3\ 3\ 1)$. Then,

$$\mathbf{u}^\top\mathbf{v} = \begin{pmatrix} 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix} = 1 \cdot 3 + 2 \cdot 3 + 0 \cdot 1 = 9$$

Inner products play an important role in statistics, including the linear regression model. For example, least squares estimation is based on the minimization of an inner product of errors and the SSE is equal to the inner-product of the residuals. In addition, inner-products form the basis of matrix multiplication, a topic to which we shall turn next.

Matrix Multiplication By matrix multiplication, I mean the multiplication of one matrix into the other. There are actually several such products but for this course we really only need to concern ourselves with the multiplication of rectangular matrices.

Whereas we can multiply any two scalars, the same is not true of rectangular matrices. The matrices have to be conformable for multiplication, which places restrictions on their order.

Imagine we are interested in the product $\mathbf{AB}$. This product exists only if the number of columns in $\mathbf{A}$ is identical to the number of rows in $\mathbf{B}$. Thus, if $\mathbf{A}$ is of order $m \times \boldsymbol{p}$, then $\mathbf{B}$ has to be of order $\boldsymbol{p} \times q$.

If $\mathbf{AB}$ exists, then it produces a matrix with the following properties.

For conformable matrices, $\mathbf{A}_{m\times p}$ and $\mathbf{B}_{p\times q}$, the matrix product $\mathbf{AB}$ produces a matrix $\mathbf{C}$ of order $m \times q$ with elements c_{ij} that are equal to the inner-product of the ith row in $\mathbf{A}$ and the jth column in $\mathbf{B}$.

To illustrate these ideas, let us consider the following matrices:

$$\mathbf{A} = \begin{pmatrix} -1 & 0 & 1 \\ 1 & 2 & 3 \end{pmatrix} \; \mathbf{B} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

The matrix product $\mathbf{AB}$ does not exist: $\mathbf{A}$ has 3 columns, but $\mathbf{B}$ only has 2 rows. On the other hand, the product $\mathbf{BA}$ exists: $\mathbf{B}$ has 2 columns and this corresponds to the number of rows in $\mathbf{A}$. The resulting matrix, $\mathbf{C}$, is of order 2×3 and contains the following elements.

- c_{11} is the inner-product of the 1st row in $\mathbf{B}$ and the 1st column in $\mathbf{A}$:

$$c_{11} = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -1 \cdot -1 + 1 \cdot 1 = 2$$

- c_{12} is the inner-product of the 1st row in $\mathbf{B}$ and the 2nd column in $\mathbf{A}$:

$$c_{12} = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = -1 \cdot 0 + 1 \cdot 2 = 2$$

- c_{13} is the inner-product of the 1st row in $\mathbf{B}$ and the 3rd column in $\mathbf{A}$:

$$c_{13} = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = -1 \cdot 1 + 1 \cdot 3 = 2$$

- c_{21} is the inner-product of the 2nd row in $\mathbf{B}$ and the 1st column in $\mathbf{A}$:

$$c_{21} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 1 \cdot -1 - 1 \cdot 1 = -2$$

- c_{22} is the inner-product of the 2nd row in $\mathbf{B}$ and the 2nd column in $\mathbf{A}$:

$$c_{22} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = 1 \cdot 0 - 1 \cdot 2 = -2$$

- c_{23} is the inner-product of the 2nd row in $\mathbf{B}$ and the 3rd column in $\mathbf{A}$:

$$c_{23} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = 1 \cdot 1 - 1 \cdot 3 = -2$$

Thus,

$$\mathbf{C} = \mathbf{BA} = \begin{pmatrix} 2 & 2 & 2 \\ -2 & -2 & -2 \end{pmatrix}$$

What the example illustrates is an important difference with scalar algebra. In scalar algebra, the order in which the scalars are multiplied is irrelevant: $pq = qp$. This is not true for matrices: in general, $\mathbf{AB} \neq \mathbf{BA}$. Indeed, as the example shows, it is even possible that one of these matrix products exists whereas the other does not.

Another important property that we frequently rely on in regression analysis concerns the transpose of a matrix product:

> Assuming conformable matrices, $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$. In plain English, the transpose of a product is equal to the product of the transposes *in reversed order*.

For some matrices, multiplication of the matrix with itself yields the same matrix again: $\mathbf{AA} = \mathbf{A}$. In this case, we say that the matrix $\mathbf{A}$ is an idempotent matrix.

B.2.4 The Inverse

Let us briefly return to scalar algebra. Here, we can define the inverse or reciprocal of a scalar k as another scalar k^{-1} such that $kk^{-1} = 1$. We can multiply other scalars by the inverse, for example, $k^{-1}m$. This means that we divide m by k.

We can do something analogous in matrix algebra. For some matrix $\mathbf{A}$, we can define the inverse $\mathbf{A}^{-1}$ such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

We can also use the inverse in multiplications such as $\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{B}\mathbf{A}^{-1}$. However, this cannot be interpreted as a division of the elements of $\mathbf{B}$ by $\mathbf{A}$. Instead, this is another matrix product.

In scalar algebra, the inverse cannot always be computed. Specifically, if $k = 0$ then k^{-1} is undefined. This is also true with inverses of matrices. The inverse requires that: (1) the matrix $\mathbf{A}$ is square; and (2) that this matrix has a non-zero determinant. The determinant is a single number that is associated with a matrix and provides important information about the matrix. If the determinant is 0, for example, the matrix is not full rank. In this case, the matrix is said to be *singular*. When a matrix can be inverted, we say that it is *regular*.

Computation of the inverse can be quite complicated and is not something we have to worry about much. What is useful is awareness of some of the properties of the inverse. These include:

1. $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$.

2. $\left(\mathbf{A}^{\top}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{\top}$.

3. $(k\mathbf{A})^{-1} = k^{-1}\mathbf{A}^{-1}$, where k is a scalar.

4. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices of order $m \times m$. Then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Inverses come in particularly handy when we try to solve systems of equations.

B.3 Representing Equations Through Matrices

B.3.1 A Single Linear Equation

The regression function for a single unit is a single linear equation. The generic linear equation takes the form of

$$y = a_1x_1 + a_2x_2 + \cdots + a_Kx_K$$

It is quite easy to represent this equation in matrix form using the inner-product. Define $\mathbf{a}^{\top} = (a_1\ a_2 \cdots a_K)$ and $\mathbf{x}^{\top} = (x_1\ x_2 \cdots x_K)$. then the linear equation is given by

$$y \quad = \quad \mathbf{a}^{\top}\mathbf{x} = \mathbf{x}^{\top}\mathbf{a}$$

B.3.2 A System of Linear Equations

Across units, the regression function is a system of n equations. The generic system of equations takes the form of

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1K}x_K \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2K}x_K \\ \vdots &\quad \vdots \\ y_M &= a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{MK}x_K \end{aligned}$$

We can represent this by defining the vector $\mathbf{y}^\top = (y_1\ y_2 \cdots y_M)$, the vector $\mathbf{x}^\top = (x_1\ x_2 \cdots x_K)$, and the matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MK} \end{pmatrix}$$

The system of equations now is equal to

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

B.3.3 A Single Quadratic Equation

In statistics, for example in least squares estimation, we sometimes encounter single equations that are quadratic in the variables. An example is the equation

$$y = a_{11}x_1^2 + (a_{12} + a_{21})x_1x_2 + a_{22}x_2^2$$

Such an equation can be represented economically using matrices:

$$y = \mathbf{x}^\top\mathbf{A}\mathbf{x}$$

Here, $\mathbf{x}^\top = (x_1\ x_2)$ and

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

It is easy to see that this produces the desired equation: first create the matrix product $\mathbf{s}^\top = \mathbf{x}^\top\mathbf{A}$ and then create the matrix product $\mathbf{s}^\top\mathbf{x}$. The notation encompasses any

quadratic equation, regardless of the number of predictors. All one needs to do is to increase the length of $\mathbf{x}$ and the dimensionality of $\mathbf{A}$ as is appropriate.

B.4 Solving Linear Equations

B.4.1 Regular Systems

Consider a system $\mathbf{y} = \mathbf{A}\mathbf{x}$ of M equations in M unknowns. The vector $\mathbf{y}$ and matrix $\mathbf{A}$ are known. The elements of $\mathbf{x}$ are unknown. We assume $\mathbf{A}$ to be regular, which means that it has a non-zero determinant and can be inverted. In this case, the system of equations can be solved quite easily using $\mathbf{A}^{-1}$. Specifically, by multiplying $\mathbf{A}^{-1}$ into both sides of the system we obtain

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{y} &= \mathbf{A}^{-1}\mathbf{A}\mathbf{x} \\ &= \mathbf{I}\mathbf{x} \\ &= \mathbf{x} \end{aligned}$$

In principle, we can solve any system of equations in this manner, no matter how many equations it encompasses.

As an example, consider the following system of 3 equations in 3 unknowns:

$$\begin{aligned} 0 &= -x_1 + x_2 + x_3 \\ 4 &= 2x_1 + x_2 - x_3 \\ 7 &= x_1 + 3x_2 + 2x_3 \end{aligned}$$

We define $\mathbf{y}^\top = (0\ 4\ 7)$, $\mathbf{x}^\top = (x_1\ x_2\ x_3)$, and

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 1 \\ 2 & 1 & -1 \\ 1 & 3 & 2 \end{pmatrix}$$

The determinant of $\mathbf{A}$ is -5, which means that it can be inverted. The inverse is[2]

$$\mathbf{A}^{-1} = \begin{pmatrix} -\frac{5}{5} & -\frac{1}{5} & -\frac{2}{5} \\ \frac{5}{5} & \frac{3}{5} & -\frac{1}{5} \\ -\frac{5}{5} & -\frac{4}{5} & \frac{3}{5} \end{pmatrix}$$

[2]We can obtain this, for example, by running the `solve` command in R.

The solution to the system of equation is now

$$\begin{aligned}\mathbf{x} &= \mathbf{A}^{-1}\mathbf{y} \\ &= \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\end{aligned}$$

Thus, $x_1 = 2$, $x_2 = 1$, and $x_3 = 1$. It is easily verified that these are the correct solutions for the system.

B.4.2 Irregular Systems

Let us revisit the previous example, but now we change the last equation:

$$\begin{aligned}0 &= -x_1 + x_2 + x_3 \\ 4 &= 2x_1 + x_2 - x_3 \\ 7 &= -x_1 + 4x_2 + 2x_3\end{aligned}$$

Thus,

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 1 \\ 2 & 1 & -1 \\ -1 & 4 & 2 \end{pmatrix}$$

When we now try the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, it does not work. The determinant of $\mathbf{A}$ is 0 and the inverse is not defined. We say that $\mathbf{A}$ is singular.

What is the problem? If we look at the last row of $\mathbf{A}$, we see that it is a perfect linear function of the first two rows. Specifically, if we multiply the first row by 3 and then add the second row, we can recover the last row perfectly: $a_{31} = 3a_{11} + a_{12}$; $a_{32} = 3a_{12} + a_{22}$; and $a_{33} = 3a_{13} + a_{23}$. The third row of $\mathbf{A}$ is thus *linearly dependent* on the other two rows and, as such, it does not contain any new information. Consequently, we have a system that effectively only has two equations. With three unknowns, this cannot produce a unique solution for $\mathbf{x}$.

B.4.3 The Rank of a Matrix

The situation of linear dependence that we encountered in the previous example means that the matrix $\mathbf{A}$ is not full rank. By rank we mean the following.

> The rank of a matrix is the number of linearly independent (LIN) vectors that constitute the matrix.

In the example, we could make the following partition:

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \mathbf{a}_3^\top \end{pmatrix}$$

where $\mathbf{a}_1^\top = (-1\ 1\ 1)$, $\mathbf{a}_2^\top = (2\ 1\ -1)$, and $\mathbf{a}_3^\top = (-1\ 4\ 2)$. In order to achieve full rank, all three vectors should be LIN, which means that none should be a perfect linear function of the others. That is not true here, however, because $\mathbf{a}_3^\top = 3\mathbf{a}_1^\top + \mathbf{a}_2^\top$. Thus, there are only two LIN vectors and the rank is 2. This has implications for the invertibility of the matrix:

> If a matrix is not full rank, then, it is singular. This means that it has a determinant of 0 and cannot be inverted.

It is useful to connect the concept of rank with that of order. This is quite easy to do in a square matrix.

> A square matrix is full rank if and only if its rank is equal to the number of rows = columns in the matrix.

In rectangular matrices, the number of rows may deviate from the number of columns and this has implications for the definition of full rank.

> A $r \times c$ matrix is full rank if the rank is equal to $min(r, c)$.

For example, the maximum achievable rank in a 4×2 matrix is 2—the minimum of the pairing of the number of rows and columns. If the number of LIN vectors is equal to 2, then we would say that the matrix is full rank.

B.5 Matrix Differentiation

We conclude the discussion of matrix algebra by revisiting the topic of derivatives. Any equation or system of equation(s) represented through matrices can be differentiated. The principles for doing this are no different than in ordinary algebra. The big difference is that we collect all of the derivatives in matrices.

B.5.1 Differentiating a Scalar with Respect to a Vector

Earlier, we discussed the function $y = a_1x_1 + a_2x_2 + a_3x_3$ and represented this as $y = \mathbf{a}^\top\mathbf{x} = \mathbf{x}^\top\mathbf{a}$. We are now interested in taking the partial derivatives of y with

respect to all of the x variables. From Appendix A, we know that this produces the following results: $\partial y/\partial x_1 = a_1$, $\partial y/\partial x_2 = a_2$, and $\partial y/\partial x_3 = a_3$. The question is how to arrange these derivatives in matrix form. Here we follow the following convention.

> A derivative of a scalar with respect to a column vector produces a column vector; a derivative of a scalar with respect to a row vector produces a row vector.

Accordingly,

$$\begin{aligned}\frac{\partial y}{\partial \mathbf{x}} &= \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \frac{\partial y}{\partial x_3} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \mathbf{a} \\ \frac{\partial y}{\partial \mathbf{x}^\top} &= \begin{pmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \frac{\partial y}{\partial x_3} \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} = \mathbf{a}^\top\end{aligned}$$

B.5.2 Differentiating a Vector with Respect to a Vector

Differentiation of a vector with respect to a vector becomes important in the context of systems of equations. Consider, for example, the following system of two linear equations:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{aligned}$$

Earlier, we saw that we can represent this system as $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{y}^\top = (y_1\ y_2)$, $\mathbf{x}^\top = (x_1\ x_2\ x_3)$, and

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$$

For this system of equations, we can identify six partial derivatives: (1) $\partial y_1/\partial x_1 = a_{11}$; (2) $\partial y_2/\partial x_2 = a_{12}$; (3) $\partial y_1/\partial x_3 = a_{13}$; (4) $\partial y_2/\partial x_1 = a_{21}$; (5) $\partial y_2/\partial x_2 = a_{22}$; and (6) $\partial y_2/\partial x_3 = a_{23}$. The question is again how to arrange these in matrix form. If we want to place all of the derivatives of y_1 in a single row and do the same with the derivatives for y_2, then we get the following result:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial y_1}{\partial \mathbf{x}^\top} \\ \frac{\partial y_2}{\partial \mathbf{x}^\top} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \mathbf{A}$$

We see that the partial derivative of $\mathbf{y}$ can be partitioned into two partial derivatives, one for y_1 and the other for y_2. We want each of these partitions to produce a row

vector and, as per the convention outlined earlier, this means that we differentiate with respect to a row vector, in this case $\mathbf{x}^\top$. If we want to place the first derivatives in two column vectors, then the same logic produces the following result.

$$\frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial \mathbf{x}} & \frac{\partial y_2}{\partial \mathbf{x}} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix} = \mathbf{A}^\top$$

B.5.3 Differentiation of Quadratic Functions

Consider the quadratic equation $y = \mathbf{x}^\top \mathbf{A}\mathbf{x}$, where $\mathbf{x}$ is of order $K \times 1$. To find the partial derivatives of y with respect to the elements of $\mathbf{x}$ we borrow the idea of differentiation by parts from calculus. Define the $K \times 1$ vector $\mathbf{u} = \mathbf{A}\mathbf{x}$ and the $1 \times K$ vector $\mathbf{v} = \mathbf{x}^\top \mathbf{A}$. Then y can be expressed in two different ways: (1) $y = \mathbf{x}^\top \mathbf{u}$ and (2) $y = \mathbf{v}\mathbf{x}$. Both expressions fall in the category of differentiating a scalar with respect to a vector. Based on the earlier results,

$$\begin{aligned} \frac{\partial \mathbf{x}^\top \mathbf{u}}{\partial \mathbf{x}} &= \mathbf{u} \\ \frac{\partial \mathbf{v}\mathbf{x}}{\partial \mathbf{x}} &= \mathbf{v}^\top \end{aligned}$$

The theory of differentiation by parts now states that

$$\begin{aligned} \partial y/\partial \mathbf{x} &= \mathbf{u} + \mathbf{v}^\top \\ &= \mathbf{A}\mathbf{x} + \left(\mathbf{x}^\top \mathbf{A}\right)^\top \\ &= \mathbf{A}\mathbf{x} + \mathbf{A}^\top \mathbf{x} \\ &= \left(\mathbf{A} + \mathbf{A}^\top\right)\mathbf{x} \end{aligned}$$

A special case arises when $\mathbf{A}$ is symmetric. In this case, $\mathbf{A}^\top = \mathbf{A}$ and $\partial y/\partial \mathbf{x} = 2\mathbf{A}\mathbf{x}$.

Appendix C

Regression Proofs

C.1 Simple Regression

C.1.1 R-Squared and Correlation

In simple regression analysis, the coefficient of determinant is equal to the square of the correlation coefficient. From Equation 1.5, we know that

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

The denominator is equal to $(n-1)s_Y^2$, where s_Y is the standard deviation of Y. The regression residuals are given by $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. From Equation 3.3, we know that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, so that $e_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$. From Equation 3.4, we know that $\hat{\beta}_1 = s_{XY}/s_X^2$. Thus,

$$e_i = (y_i - \bar{y}) - \frac{s_{XY}}{s_X^2}(x_i - \bar{x})$$

and

$$\begin{aligned}
\sum_i e_i^2 &= \sum_i (y_i - \bar{y})^2 - 2\frac{s_{XY}}{s_X^2}\sum_i (y_i - \bar{y})(x_i - \bar{x}) + \left(\frac{s_{XY}}{s_X^2}\right)^2 \sum_i (x_i - \bar{x})^2 \\
&= (n-1)s_Y^2 - 2\frac{s_{XY}}{S_X^2}(n-1)s_{XY} + \left(\frac{s_{XY}}{s_X^2}\right)^2 (n-1)s_X^2
\end{aligned}$$

This may be written as $(n-1)(s_X^2 s_Y^2 - s_{XY}^2)/s_X^2$. With this alternative expression for the SSE in simple regression analysis, the coefficient of determination becomes

$$\begin{aligned} R^2 &= 1 - \frac{(n-1)s_X^2 s_Y^2 - s_{XY}^2}{(n-1)s_X^2 s_Y^2} \\ &= 1 - 1 + \frac{s_{XY}^2}{s_X^2 s_Y^2} \\ &= \frac{s_{XY}^2}{s_X^2 s_Y^2} \end{aligned}$$

We know that the Pearson product moment correlation coefficient is $r = s_{XY}/(s_X s_Y)$, so that it follows that $R^2 = r^2$.

C.1.2 Variance of the Predicted Values

We are interested in obtaining a formula for the sampling variance of the predicted values that emerge from a simple regression analysis. These predictions are given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Thus, we are interested in $V[\hat{y}_i]$. By definition,

$$V[\hat{y}_i] = E\left[(\hat{y}_i - E[\hat{y}_i])^2\right]$$

Expanding the expectation of the predicted values, we have $E[\hat{y}_i] = E[\hat{\beta}_0 + \hat{\beta}_1 x_i] = E[\hat{\beta}_0] + E[\hat{\beta}_1]x_i = \beta_0 + \beta_1 x_i$. Here we took advantage of the Gauss-Markov theorem and its implication that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Thus,

$$\begin{aligned} V[\hat{y}_i] &= E\left\{\left[(\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\beta_0 + \beta_1 x_i)\right]^2\right\} \\ &= E\left\{\left[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i\right]^2\right\} \\ &= \underbrace{E\left[(\hat{\beta}_0 - \beta_0)^2\right]}_{Var[\hat{\beta}_0]} + 2x_i \underbrace{E\left[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\right]}_{Cov[\hat{\beta}_0, \hat{\beta}_1]} + x_i^2 \underbrace{E\left[(\hat{\beta}_1 - \beta_1)^2\right]}_{Var[\hat{\beta}_1]} \end{aligned}$$

We know that $Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)$, $Var[\hat{\beta}_1] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$, and $Cov[\hat{\beta}_0, \hat{\beta}_1] = -\bar{x} Var[\hat{\beta}_1]$. Substitution of these results yields

$$\begin{aligned} V[\hat{y}_i] &= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \left(\frac{\sum_i (x_i - \bar{x})^2}{n} + \bar{x}^2 + x_i^2 - 2\bar{x} x_i \right) \\ &= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \left(\frac{\sum_i (x_i - \bar{x})^2}{n} + (x_i - \bar{x})^2 \right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right) \end{aligned}$$

C.2 Multiple Regression

C.2.1 Residuals

By definition, the vector of residuals is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, the residuals may also be written as

$$\mathbf{e} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

It is then easy to see that $\mathbf{e}$ and $\hat{\mathbf{y}}$ are uncorrelated:

$$\begin{aligned} E[\hat{\mathbf{y}}^\top \mathbf{e}] &= E\left[(\mathbf{H}\mathbf{y})^\top (\mathbf{I} - \mathbf{H})\mathbf{y} \right] \\ &= E\left[\mathbf{y}^\top \mathbf{H}^\top \mathbf{I}\mathbf{y} - \mathbf{y}^\top \underbrace{\mathbf{H}^\top \mathbf{H}}_{\mathbf{H}^\top} \mathbf{y} \right] \\ &= E\left[\mathbf{y}^\top \mathbf{H}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{H}^\top \mathbf{y} \right] \\ &= \mathbf{0} \end{aligned}$$

The second line takes advantage of both the symmetry of the hat matrix ($\mathbf{H} = \mathbf{H}^\top$) and its idempotency ($\mathbf{H}^\top \mathbf{H} = \mathbf{H}^\top$).

In the expression we derived, the residuals are a function of the observed dependent variable. We can also make them a function of the errors of the population regression

model. Since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we can write the residuals as

$$\begin{aligned}
\mathbf{e} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \underbrace{\mathbf{H}}_{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}\mathbf{X}\boldsymbol{\beta} - \mathbf{H}\boldsymbol{\varepsilon} \\
&= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{H}\boldsymbol{\varepsilon} \\
&= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\boldsymbol{\varepsilon} \\
&= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}
\end{aligned}$$

The residuals are thus a function of the errors. Under the assumption that $E[\boldsymbol{\varepsilon}] = \mathbf{0}$, it can then be easily shown that $E[\mathbf{e}] = \mathbf{0}$. That is, the residuals have a mean of zero. It can also be demonstrated that $\mathbf{V}[\mathbf{e}] = \sigma^2(\mathbf{I}-\mathbf{H})$. Thus, the variance-covariance matrix of the residuals differs from the variance-covariance matrix of the errors, which under the usual assumptions is $\mathbf{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$. The differences manifest themselves both on the diagonal and off-diagonal. Since the diagonal hat values may be different from each other, the variances of the individual residuals may differ even if $\boldsymbol{\varepsilon}$ is homoskedastic. Since the off-diagonal hat values may not be 0, the covariances between the residuals may be non-zero even if $\boldsymbol{\varepsilon}$ does not suffer from autocorrelation.[1]

C.2.2 OLS

The least squares criterion is given by $S = \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}$. Substituting $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and expanding yields

$$\begin{aligned}
S &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{y}^\top - \boldsymbol{\beta}^\top\mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

The last line follows from the fact that $\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}$ are identical scalars and may hence be added together.

To minimize S, we need to differentiate it with respect to the elements in the vector $\boldsymbol{\beta}$. We can reformulate S in the following manner:

$$S = s_1 + s_2 + s_3,$$

[1] Specifically, $0 < h_{ii} < 1$ and $-.5 < h_{ij} < .5$.

where

$$\begin{aligned} s_1 &= \mathbf{y}^\top \mathbf{y} \\ s_2 &= -2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} \\ s_3 &= \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \end{aligned}$$

Setting $\mathbf{X}^\top \mathbf{y} = \mathbf{a}$ and $\mathbf{X}^\top \mathbf{X} = \mathbf{C}$, this can be simplified further:

$$\begin{aligned} s_1 &= \mathbf{y}^\top \mathbf{y} \\ s_2 &= -2\boldsymbol{\beta}^\top \mathbf{a} \\ s_3 &= \boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} \end{aligned}$$

Here it is important to note that $\mathbf{C}$ is a symmetric matrix.

We can now differentiate S with respect to the vector $\boldsymbol{\beta}$. Because the derivative of a sum is equal to the sum of the derivatives, this amounts to

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = \frac{\partial s_1}{\partial \boldsymbol{\beta}} + \frac{\partial s_2}{\partial \boldsymbol{\beta}} + \frac{\partial s_3}{\partial \boldsymbol{\beta}}$$

Since s_1 is not a function of $\boldsymbol{\beta}$, it follows that

$$\frac{\partial s_1}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

where $\mathbf{0}$ is a $(K+1) \times 1$ vector of 0s. The derivative of s_2 follows from the results on scalar-vector differentiation, presented in Appendix B.5:

$$\begin{aligned} \frac{\partial s_2}{\partial \boldsymbol{\beta}} &= -2\mathbf{a} \\ &= -2\mathbf{X}^\top \mathbf{y} \end{aligned}$$

Finally, the derivative of s_3 follows from the results about differentiating quadratic functions (Appendix B.5):

$$\begin{aligned} \frac{\partial s_3}{\partial \boldsymbol{\beta}} &= 2\mathbf{C}\boldsymbol{\beta} \\ &= 2\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \end{aligned}$$

Thus, the derivative of S is given by

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{\beta}} &= \mathbf{0} - 2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\ &= 2\left(\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top \mathbf{y}\right)\end{aligned}$$

If we set this to $\mathbf{0}$, we can derive the OLS estimator as is shown in Chapter 5.1.

C.2.3 Gauss-Markov Theorem

The Gauss-Markov theorem states that the OLS/ML estimator of the regression coefficients is the best linear unbiased estimator if Assumptions 4.2-4.3 are met. To prove this result, we proceed as follows.

OLS Is a Linear Estimator It is easy to demonstrate that the OLS estimator is a linear estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{A}}\mathbf{y} \\ &= \mathbf{A}\mathbf{y}\end{aligned}$$

Together, $\hat{\boldsymbol{\beta}}$, $\mathbf{A}$, and $\mathbf{y}$ form a linear system of equations and, hence, it follows that the OLS estimator is a linear function of the data contained in $\mathbf{y}$.

OLS Is Unbiased To demonstrate that OLS is unbiased, we begin by re-writing the estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{y} \\ &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon}\end{aligned}$$

We now take the expectation of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1} E[\mathbf{X}^\top\boldsymbol{\varepsilon}] \\ &= \boldsymbol{\beta} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \underbrace{\mathbf{0}}_{\text{Assumption 4.3}} \\ &= \boldsymbol{\beta} \end{aligned}$$

Under Assumption 4.3, then, the estimator is a linear unbiased estimator.

OLS Is Best We can demonstrate that OLS is the best linear unbiased estimator by contrasting it with the generic linear estimator

$$\tilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y},$$

where $\mathbf{C}$ is a $(K+1) \times n$ matrix such that $\mathbf{C} = \mathbf{B} + \mathbf{A}$, $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, and $\mathbf{B}$ is conformable generic matrix. It can be demonstrated that

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= [\mathbf{B} + \mathbf{A}]\mathbf{y} \\ &= [\mathbf{B} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top][\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] \\ &= \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\varepsilon} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} \\ &= \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{I}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\varepsilon} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} \\ &= [\mathbf{B}\mathbf{X} + \mathbf{I}]\boldsymbol{\beta} + [\mathbf{B} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\boldsymbol{\varepsilon} \end{aligned}$$

Viewed in this light, $\hat{\boldsymbol{\beta}}$ is a special case of $\tilde{\boldsymbol{\beta}}$ that arises when $\mathbf{B} = \mathbf{0}$.

From the expression above, it is clear that $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ if Assumption 4.3 holds, $\mathbf{B}\mathbf{X} = \mathbf{0}$ and if $E[\mathbf{B}\boldsymbol{\varepsilon}] = \mathbf{0}$. We assume that both conditions hold true so that $\tilde{\boldsymbol{\beta}}$ is a linear unbiased estimator. Making the appropriate substitutions,

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + [\mathbf{B} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\boldsymbol{\varepsilon}$$

We now derive the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$. By definition, $\mathbf{V}[\tilde{\boldsymbol{\beta}}] = E[(\tilde{\boldsymbol{\beta}} - E[\tilde{\boldsymbol{\beta}}])(\tilde{\boldsymbol{\beta}} - E[\tilde{\boldsymbol{\beta}}])^\top]$. Under the assumption that $\tilde{\boldsymbol{\beta}}$ is unbiased, this

reduces to

$$\begin{aligned}\mathbf{V}[\tilde{\boldsymbol{\beta}}] &= E[(\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}})^\top] \\ &= E\left[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top\right] - \boldsymbol{\beta}\boldsymbol{\beta}^\top\end{aligned}$$

Substituting the expression for $\tilde{\boldsymbol{\beta}}$, we can show that

$$\begin{aligned}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \boldsymbol{\beta}\boldsymbol{\varepsilon}^\top\mathbf{B}^\top + \boldsymbol{\beta}\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \\ &\quad \mathbf{B}\boldsymbol{\varepsilon}\boldsymbol{\beta}^\top + \mathbf{B}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{B}^\top + \mathbf{B}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \\ &\quad \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\beta}^\top + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{B}^\top + \\ &\quad \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\end{aligned}$$

We now take the expectation of this expression, where we keep four things in mind: (1) $\mathbf{BX} = \mathbf{0}$; (2) $E[\mathbf{B}\boldsymbol{\varepsilon}] = \mathbf{0}$; (3) $E[\mathbf{X}^\top\boldsymbol{\varepsilon}] = \mathbf{0}$; and (4) $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\mathbf{I}$ (Assumption 4.2).

$$\begin{aligned}E[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top] &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \boldsymbol{\beta}\underbrace{E[(\mathbf{B}\boldsymbol{\varepsilon})^\top]}_{\mathbf{0}^\top} + \boldsymbol{\beta}\underbrace{E[\boldsymbol{\varepsilon}^\top\mathbf{X}]}_{\mathbf{0}^\top}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \\ &\quad \underbrace{E[\mathbf{B}\boldsymbol{\varepsilon}]}_{\mathbf{0}}\boldsymbol{\beta}^\top + \sigma^2\mathbf{B}\mathbf{B}^\top + \sigma^2\underbrace{\mathbf{BX}}_{\mathbf{0}}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \\ &\quad \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\underbrace{E[\mathbf{X}^\top\boldsymbol{\varepsilon}]}_{\mathbf{0}^\top}\boldsymbol{\beta}^\top + \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\underbrace{(\mathbf{BX})^\top}_{\mathbf{0}^\top} + \\ &\quad \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \sigma^2\left[\mathbf{B}\mathbf{B}^\top + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]\end{aligned}$$

Substituting this result into the expression for the variance of $\tilde{\boldsymbol{\beta}}$, we get

$$\mathbf{V}[\tilde{\boldsymbol{\beta}}] = \sigma^2\left[\mathbf{B}\mathbf{B}^\top + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]$$

where $\mathbf{B}\mathbf{B}^\top$ is positive definite.[2]

The only thing that is now left to do is to compare this variance to the variance of the OLS estimator. As we shall demonstrate in the next section of the Appendix,

$$\mathbf{V}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

Consequently,

$$\mathbf{V}[\tilde{\boldsymbol{\beta}}] - \mathbf{V}[\hat{\boldsymbol{\beta}}] = \sigma^2 \mathbf{B}\mathbf{B}^\top$$

Due to the fact that $\sigma^2 > 0$ and $\mathbf{B}\mathbf{B}^\top$ is positive definite, it follows that the expression above is greater than $\mathbf{0}$. This means that any other linear unbiased estimator has a greater variance than the OLS estimator and, as such, is less efficient. Or in other words, OLS is the best linear unbiased estimator.

C.2.4 Bias in the MLE of the Regression Variance

The MLE of the regression variance is given by $\hat{\sigma}^2 = \mathbf{e}^\top \mathbf{e}/n$. To compute the bias in this estimator, we start by showing the expectation of $\mathbf{e}^\top \mathbf{e}$. We have already seen that $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon} - \mathbf{H}\boldsymbol{\varepsilon}$. Consequently, the sum of squared residuals may be written as

$$\begin{aligned}
\mathbf{e}^\top \mathbf{e} &= (\boldsymbol{\varepsilon}^\top - \boldsymbol{\varepsilon}^\top \mathbf{H})(\boldsymbol{\varepsilon} - \mathbf{H}\boldsymbol{\varepsilon}) \\
&= \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top \mathbf{H}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top \mathbf{H}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top \mathbf{H}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top \mathbf{H}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^\top \mathbf{M}\boldsymbol{\varepsilon} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_i \varepsilon_j m_{ij}
\end{aligned}$$

[2] One can think of positive definiteness as the matrix equivalent of saying that something is strictly positive. More precisely, a matrix $\mathbf{Q}$ is positive definite if for all non-zero vectors $\mathbf{x}$, the product $\mathbf{x}^\top \mathbf{Q}\mathbf{x}$ returns a positive value.

where $m_{ij} = 1 - h_{ij}$. The expectation of the sum of squared residuals is

$$E[\mathbf{e}^\top \mathbf{e}] = E\left[\sum_i \sum_j \varepsilon_i \varepsilon_j m_{ij}\right] = \sigma^2 \sum_i m_{ii}$$

This result follows from the fact that $E[\varepsilon_i \varepsilon_j] = \sigma^2$ if $i = j$ and 0 otherwise (courtesy of Assumption 4.2). We know that $\sum_i m_{ii} = \sum_i 1 - \sum_i h_{ii} = n - (K+1)$. Consequently,

$$E[\mathbf{e}^\top \mathbf{e}] = \sigma^2(n - K - 1)$$

This means that

$$E[\hat{\sigma}^2] = \frac{n-K-1}{n}\sigma^2$$

Now let

$$s^2 = \frac{n}{n-K-1}\hat{\sigma}^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n-K-1}$$

It is easily demonstrated that this estimator is unbiased:

$$E[s^2] = \frac{n}{n-K-1}E[\hat{\sigma}^2] = \frac{n}{n-K-1}\frac{n-K-1}{n}\sigma^2 = \sigma^2$$

C.2.5 Standard Errors of the Regression Coefficients

By definition, $\mathbf{V}[\hat{\boldsymbol{\beta}}] = E[(\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}])^\top]$. If we can assume that the OLS/ML estimator is unbiased, then

$$\mathbf{V}[\hat{\boldsymbol{\beta}}] = E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right] = E\left[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top\right] - \boldsymbol{\beta}\boldsymbol{\beta}^\top$$

We know that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$. Consequently,

$$\begin{aligned}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top = {} & \boldsymbol{\beta}\boldsymbol{\beta}^\top + \boldsymbol{\beta}\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\beta}^\top + \\ & \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\end{aligned}$$

Taking expectations, we have

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top\right] &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \boldsymbol{\beta}E\left[\boldsymbol{\varepsilon}^\top\mathbf{X}\right]\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}E\left[\mathbf{X}^\top\boldsymbol{\varepsilon}\right]\boldsymbol{\beta}^\top + \\ &\quad \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \end{aligned}$$

Under Assumption 4.3, the second and third terms vanish. Under Assumption 4.2, $E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] = \sigma^2\mathbf{I}$. Thus,

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top\right] &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left[\sigma^2\mathbf{I}\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \end{aligned}$$

Substitution yields

$$\begin{aligned} \mathbf{V}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}\boldsymbol{\beta}^\top + \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \end{aligned}$$

C.2.6 Standard Errors of the Predicted Values

For the predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}$, we can derive the standard errors in the following way. By definition, $V[\hat{y}_i] = E\left[(\hat{y}_i - E[\hat{y}_i])^2\right]$. We can easily demonstrate that $E[\hat{y}_i] = E[\hat{\beta}_0] + E[\hat{\beta}_1]x_{i1} + \cdots + E[\hat{\beta}_K]x_{iK} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$, assuming that the OLS estimators are unbiased. Consequently,

$$\begin{aligned} \hat{y}_i - E[\hat{y}_i] &= \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}\right) - \\ &\quad (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}) \\ &= (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_{i1} + \cdots + (\hat{\beta}_K - \beta_K)x_{iK} \end{aligned}$$

We know that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \cdots - \hat{\beta}_K\bar{x}_K$. We also can show that $\bar{y} = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_K\bar{x}_K + \bar{\varepsilon}$ or $\beta_0 = \bar{y} - \beta_1\bar{x}_1 - \cdots - \beta_K\bar{x}_K - \bar{\varepsilon}$. Thus,

$$\begin{aligned}\hat{\beta}_0 - \beta_0 &= \left(\bar{y} - \hat{\beta}_1\bar{x}_1 - \cdots - \hat{\beta}_K\bar{x}_K\right) - \\ &\quad (\bar{y} - \beta_1\bar{x}_1 - \cdots - \beta_K\bar{x}_K - \bar{\varepsilon}) \\ &= -(\hat{\beta}_1 - \beta_1)\bar{x}_1 - \cdots - (\hat{\beta}_K - \beta_K)\bar{x}_K + \bar{\varepsilon}\end{aligned}$$

Substitution yields

$$\begin{aligned}\hat{y}_i - E[\hat{y}_i] &= -(\hat{\beta}_1 - \beta_1)\bar{x}_1 - \cdots - (\hat{\beta}_K - \beta_K)\bar{x}_K + \bar{\varepsilon} + \\ &\quad (\hat{\beta}_1 - \beta_1)x_{i1} + \cdots + (\hat{\beta}_K - \beta_K)x_{iK} \\ &= (\hat{\beta}_1 - \beta_1)(x_{i1} - \bar{x}_1) + \cdots + (\hat{\beta}_K - \beta_K)(x_{iK} - \bar{x}_K) + \bar{\varepsilon}\end{aligned}$$

Squaring the left-hand side yields

$$\begin{aligned}(\hat{y}_i - E[\hat{y}_i])^2 &= \left[(\hat{\beta}_1 - \beta_1)(x_{i1} - \bar{x}_1) + \cdots + (\hat{\beta}_K - \beta_K)(x_{iK} - \bar{x}_K) + \bar{\varepsilon}\right]^2 \\ &= \sum_k (\hat{\beta}_k - \beta_k)^2(x_{ik} - \bar{x}_k)^2 + \\ &\quad 2\sum_{j<k}(\hat{\beta}_j - \beta_j)(\hat{\beta}_k - \beta_k)(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) + \\ &\quad \sum_k (\hat{\beta}_k - \beta_k)(x_{ik} - \bar{x}_k)\bar{\varepsilon} + \\ &\quad \bar{\varepsilon}^2\end{aligned}$$

We conclude the derivation by taking expectations over both sides. Here, we keep in mind that (1) $E[(\hat{\beta}_k - \beta_k)^2] = Var[\hat{\beta}_k]$; (2) $E[(\hat{\beta}_j - \beta_j)(\hat{\beta}_k - \beta_k)] = Cov[\hat{\beta}_j, \hat{\beta}_k]$; (3) $E[\bar{\varepsilon}] = 0$; and $E[\bar{\varepsilon}^2] = \sigma^2/n$. It then follows that

$$\begin{aligned}E\left[(\hat{y}_i - E[\hat{y}_i])^2\right] &= \sum_k (x_{ik} - \bar{x}_k)^2 Var[\hat{\beta}_k] + \\ &\quad 2\sum_{j<k}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)Cov[\hat{\beta}_j, \hat{\beta}_k] + \frac{\sigma^2}{n}\end{aligned}$$

C.2.7 ANOVA

By definition the sample variation in Y is given by $\sum_i (y_i - \bar{y})^2$. In the sample, $y_i = \hat{y}_i + e_i$. Hence,

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i [(\hat{y}_i + e_i) - \bar{y}]^2 \\ &= \sum_i (\hat{y}_i \bar{y})^2 + \sum_i e_i^2 + 2\sum_i (\hat{y}_i - \bar{y}) e_i\end{aligned}$$

The last term vanishes. It can be written as $\sum_i \hat{y}_i e_i - \bar{y}\sum_i e_i$. In Chapter 4, we saw that the predicted values and the residuals are orthogonal so that $\sum_i \hat{y}_i e_i = 0$. We also know that the residuals sum to 0, so that $\bar{y}\sum_i e_i = 0$. Consequently,

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2$$

We can also present this result in matrix form. To do so, we begin by considering the SST. Expansion of the scalar result yields:

$$\sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2 = \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2$$

In matrix form, $\sum_i y_i^2 = \mathbf{y}^\top \mathbf{y}$. To obtain $\sum_i y_i$, we generate n dimensional vector $\mathbf{j}^\top = (1\ 1 \cdots 1)$. We can then write $\sum_i y_i = \mathbf{j}^\top \mathbf{y}$. Of course, we have the square of $\sum_i y_i$. In matrix form, this can be written as $(\sum_i y_i)^2 = (\mathbf{j}^\top \mathbf{y})^\top \mathbf{j}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{J} \mathbf{y}$, where $\mathbf{J} = \mathbf{j}\mathbf{j}^\top$ is a $n \times n$ matrix consisting entirely of 1s. Thus, we have the following result:

$$SST = \mathbf{y}^\top \mathbf{y} - \frac{1}{n}\mathbf{y}^\top \mathbf{J} \mathbf{y} = \mathbf{y}^\top \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{y}$$

We now turn to the SSE. We can write $\sum_i e_i^2$ as $\mathbf{e}^\top \mathbf{e}$, which can be expanded

as

$$\begin{aligned}
\mathbf{e}^\top\mathbf{e} &= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{y}^\top\mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{y} + \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{y}^\top\mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{y} + \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{X}\underbrace{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}}_{\hat{\boldsymbol{\beta}}} \\
&= \mathbf{y}^\top\mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{y} + \hat{\boldsymbol{\beta}}^\top\mathbf{I}\mathbf{X}^\top\mathbf{y} \\
&= \mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{y}
\end{aligned}$$

This can also be written a bit differently. Setting $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, the last term becomes $\mathbf{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{y}^\top\mathbf{H}\mathbf{y}$, so that

$$SSE = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{H}\mathbf{y} = \mathbf{y}^\top\left(\mathbf{I}-\mathbf{H}\right)\mathbf{y}$$

We obtain the SSR from the equation $SST = SSR + SSE$ or $SSR = SST - SSE$:

$$\begin{aligned}
SSR &= \mathbf{y}^\top\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{y} - \mathbf{y}^\top\left(\mathbf{I}-\mathbf{H}\right)\mathbf{y} \\
&= \mathbf{y}^\top\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{y}
\end{aligned}$$

This is equivalent to $SSR = \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{y} - (1/n)\mathbf{y}^\top\mathbf{J}\mathbf{y}$.

C.2.8 Shortcomings of t-Tests When Testing Joint Hypotheses

Consider the following regression model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$. We formulate the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. Kmenta (1997) shows that

$$F = \frac{t_1^2 + t_2^2 + 2t_1 t_2 r_{XZ}}{2(1-r_{XZ}^2)},$$

where t_1 is the test statistic associated with $\hat{\beta}_1$, t_2 is the test statistic associated with $\hat{\beta}_2$, and r_{XZ} is the correlation between the two predictors. If $r_{XZ} = 0$, then $F = .5(t_1^2 + t_2^2)$. Since this is simply the arithmetic mean of the two squared t-

test statistics, we would not expect differences in the statistical conclusions that we derive from the t-tests and the F-test. However, if $r_{XZ} \to 1$, i.e., there is near-perfect multicollinearity, then F can be large even if the two t-test statistics are small. In this case, individual t-tests might suggest that the parameters are zero in the population, while the F-test might suggest the opposite.

C.2.9 Expected Mean Squares

In this appendix, we derive the expectations of MSR and MSE. To reduce complexity, we do this for the simple regression model. Following Appendix C.2.4, we know that

$$E[MSE] \quad = \quad E[s^2] = \sigma^2$$

In the simple regression model, $MSR = SSR/1 = SSR$. Taking the expectation, we get

$$\begin{aligned}
E[MSR] &= E[SSR] \\
&= E\left[\sum_i (\hat{y}_i - \bar{y})^2\right] \\
&= E\left[\sum_i (\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} - \bar{y})^2\right] \\
&= E\left[\sum_i (\underbrace{\bar{y} - \hat{\beta}_1 \bar{x}}_{\hat{\beta}_0} + \hat{\beta}_1 x_i - \bar{y})^2\right] \\
&= E\left[\hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2\right] \\
&= \underbrace{E[\hat{\beta}_1^2]}_{\text{Everything else is fixed}} \sum_i (x_i - \bar{x})^2
\end{aligned}$$

This can be expanded further. Noting that $\text{Var}(\hat{\beta}_1) = E[\hat{\beta}_1^2] - (E[\hat{\beta}_1])^2$, it follows that $E[MSR] = [\text{Var}(\hat{\beta}_1) + (E[\hat{\beta}_1])^2] \sum_i (x_i \bar{x})^2$. Since $\hat{\beta}_1$ is unbiased,

this may also be written as $E[MSR] = \text{Var}(\hat{\beta}_1) \sum_i (x_i - \bar{x})^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$. From Chapter 3, we know that $\text{Var}(\hat{\beta}_1) = \sigma^2 / \sum_i (x_i - \bar{x})^2$. Substitution yields

$$E[MSR] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$$

Under the null hypothesis $\beta_1 = 0$, this reduces to $E[MSR] = \sigma^2$.

C.2.10 Derivation of the F-test Statistic

Consider the multiple regression model $y_i \sim \mathcal{N}(\mu_i, \sigma)$ with $\mu_i = \beta_0 + \beta_k x_{ik}$. Under the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$, the model reduces to $y_i \sim \mathcal{N}(\mu, \sigma)$, where $\mu = \beta_0$. Under the null hypothesis, then, we can consider the observations to be n independent draws from the identical normal distribution with mean μ and variance σ^2.

This result is important because it allows us to invoke Cochran's theorem (Cochran, 1934). This theorem states that, for a series of normally and independently distributed (n.i.d.) variables with mean μ and variance σ^2, the SST may be decomposed into Q sums of squares, SS_q, each with degrees of freedom ν_q. Furthermore, the terms SS_q/σ^2 are independent chi-squared variates, as long as $\sum_{q=1}^{Q} \nu_q = n - 1$.

Under H_0 the n.i.d. assumption with identical mean and variance is met. In the ANOVA, we decompose SST into two components: (1) SSR with K degrees of freedom and (2) SSE with $n - K - 1$ degrees of freedom. Noting that the degrees of freedom add to $n - 1$, we now know that

$$\begin{aligned} \frac{SSR}{\sigma^2} &\sim \chi^2_K \\ \frac{SSE}{\sigma^2} &\sim \chi^2_{n-K-1} \end{aligned}$$

We now turn to another result from mathematical statistics, namely that the ratio of two independent chi-squared variates, each divided by the degrees

of freedom, follows an F-distribution. Hence,

$$\frac{\frac{\chi^2_K}{K}}{\frac{\chi^2_{n-K-1}}{n-K-1}} = \frac{\frac{SSR}{K\sigma^2}}{\frac{SSE}{(n-K-1)\sigma^2}} = \frac{MSR}{MSE} \sim \mathcal{F}[K, n-K-1]$$

C.2.11 Variance of the Fitted Values

By definition, $\mathsf{Var}(\hat{y}_i) = E\left[\hat{y}_i - E[\hat{y}_i])^2\right]$. We know that $E[\hat{y}_i] = E[y_i]$. Substitution yields, $\mathsf{Var}(\hat{y}_i) = E\left[(\hat{y}_i - E[y_i])^2\right]$. We know that $\hat{y}_i = \hat{\beta}_0 + \sum_k \hat{\beta}_k x_{ik}$ and $E[y_i] = \beta_0 + \sum_k \beta_k x_{ik}$. Hence,

$$\mathsf{Var}(\hat{y}_i) = E\left[\left(\hat{\beta}_0 - \beta_0 + \sum_k (\hat{\beta}_k - \beta_k) x_{ik}\right)^2\right]$$

We also know that $\hat{\beta}_0 = \bar{y} - \sum_k \hat{\beta}_k \bar{x}_k$. Further, $\beta_0 = y_i - \sum_k \beta_k x_{ik} - \varepsilon$. We manipulate the latter equation by summing β_0 n times and then dividing by n:

$$\begin{aligned}
\frac{1}{n}\sum_i \beta_0 &= \beta_0 \\
&= \frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_k \sum_i \beta_k x_{ik} - \frac{1}{n}\sum_i \varepsilon_i \\
&= \bar{y} - \sum_k \beta_k \bar{x}_k - \bar{\varepsilon}
\end{aligned}$$

Consequently,

$$\begin{aligned}
\hat{\beta}_0 - \beta_0 &= \left(\bar{y} - \sum_k \hat{\beta}_k \bar{x}_k\right) - \left(\bar{y} - \sum_k \beta_k \bar{x}_k - \bar{\varepsilon}\right) \\
&= -\sum_k (\hat{\beta}_k - \beta_k)\bar{x}_k + \bar{\varepsilon}
\end{aligned}$$

Substitution into the formula for the variance of the fitted values now yields

$$\mathsf{Var}(\hat{y}_i) = E\left[\left((\hat{\beta}_k - \beta_k)(x_{ik} - \bar{x}_k) + \bar{\varepsilon}\right)^2\right]$$

We now expand and distribute the expected value operator over the constituent terms:

$$\begin{aligned}\mathsf{Var}(\hat{y}_i) &= \sum_k E\left[(\hat{\beta}_k - \beta_k)^2\right](x_{ik} - \bar{x})^2 + \\ &\quad 2\sum_{j<k} E\left[(\hat{\beta}_j - \beta_j)(\hat{\beta}_k - \beta_k)\right](x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \\ &\quad 2\sum_k E\left[\hat{\beta}_k - \beta_k\right] E\left[\bar{\varepsilon}(x_{ik} - \bar{x})\right] + \\ &\quad E\left[\bar{\varepsilon^2}\right]\end{aligned}$$

The terms in the third line disappear because we assume $\hat{\beta}_k$ to be unbiased, so that $E[\hat{\beta}_k - \beta_k] = E[\hat{\beta}_k] - \beta_k = \beta_k - \beta_k = 0$. In the first line, we recognize the terms $E\left[\left(\hat{\beta}_k - \beta_k\right)^2\right]$ as variances $\mathsf{Var}(\hat{\beta}_k)$. In the second line, the terms $E\left[\left(\hat{\beta}_j - \beta_j\right)\left(\hat{\beta}_k - \beta_k\right)\right]$ are covariances of the type $\mathsf{Cov}(\hat{\beta}_j, \hat{\beta}_k)$. Finally, $E[\bar{\varepsilon}^2]$ is equal to $\mathsf{Var}(\bar{\varepsilon})$. Being the variance of a sample mean, this is equal to σ^2/n. Substituting these results, we get

$$\begin{aligned}\mathsf{Var}(\hat{y}_i) &= \sum_k \mathsf{Var}(\hat{\beta}_k)(x_{ik} - \bar{x})^2 + \\ &\quad 2\sum_{j<k} \mathsf{Cov}(\hat{\beta}_j, \hat{\beta}_k)(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) + \\ &\quad \frac{\sigma^2}{n}\end{aligned}$$

C.2.12 Adjusted R^2

By definition,

$$\bar{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)} = 1 - \frac{SSE}{SST}\frac{n-1}{n-K-1}$$

From the definition of the coefficient of determination, we know that $\frac{SSE}{SST} = 1 - R^2$. Substitution yields

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-K-1}$$

C.2.13 R^2 and the F-Statistic

We know that

$$F = \frac{MSR}{MSE} = \frac{SSR}{SSE} \frac{n-K-1}{K}$$

We also know that

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

From the last result, we now derive the following expressions:

$$\begin{aligned} SSR &= R^2 \cdot SST \\ SSE &= \left(1 - R^2\right) \cdot SST \end{aligned}$$

Substitution in the formula for F yields

$$F = \frac{R^2 \cdot SST}{(1 - R^2) \cdot SST} \frac{n-K-1}{K} = \frac{R^2}{1 - R^2} \frac{n-K-1}{K}$$

C.3 Model Fit and Comparison

C.3.1 Kullback-Leibler Information

Kullback-Leibler Information is a criterion that measures how far a model strays from the truth. Let $f(x)$ denote the true model. Further, let $g(x|\theta)$ denote the model that we estimate, which postulates that the observed data depend on the parameter(s) θ. The Kullback-Leibler criterion is now defined as (see Burnham

and Anderson, 2004, pp. 266-267):

$$\begin{aligned} I[f,g] &= \int f(x) \ln\left(\frac{f(x)}{g(x|\theta)}\right) dx \\ &= \int f(x) \ln\left[f(x)\right] dx - \underbrace{\int f(x) \ln\left[g(x|\theta)\right] dx}_{\text{Definition of the mean}} \\ &= C - E_f[\ln(g(x|\theta))] \end{aligned}$$

Here, C may be treated as a constant, since the true model is assumed to always be the same. Further, $E_f[\ln(g(x|\theta))]$ is the expectation for the estimated model. The best model is the one that minimizes I.

As an example consider the normal distribution. Let the true distribution be normal with a mean of μ_T and a variance of σ_T^2. The estimated model also stipulates a normal distribution but with a mean of μ and a variance of σ^2. After some mathematics, we can show that

$$I(f,g) = \frac{1}{2}\left[\ln\left(\frac{\sigma^2}{\sigma_T^2}\right) - 1 + \frac{\sigma_T^2 + (\mu_t - \mu)^2}{\sigma^2}\right]$$

Now imagine $\mu_T = 0$ and $\sigma_T^2 = 1$. Further, let g_1 denote the normal distribution with mean 0 and variance 2, whereas g_2 is the normal distribution with mean 1 and variance 1. It is now easy to show that $I(f,g_1) = 0.097$ and $I(f,g_2) = 0.5$. Clearly, g_1 is closer to the truth than g_2.

C.3.2 The Akaike Information Criterion

The Kullback-Leibler criterion is a theoretical result. The statistician Akaike showed how this result can be combined with maximum likelihood estimates obtained from empirical data. Specifically, he showed that (see Burnham and Anderson, 2004, p. 268):

$$\ell - K = C - \hat{E}_{\hat{\theta}}[I(f,\hat{g})]$$

Here, ℓ is the log-likelihood (see Chapter 3), K is the number of estimated parameters in g, and $C - \hat{E}_{\hat{\theta}}[I(f,\hat{g})]$ is the expected value of the Kullback-

Leibler Information at the maximum likelihood estimates. Note that K is the bias we incur when we use the log-likelihood as the estimator of the expected Kullback-Leibler information. Subtracting K from the log-likelihood produces an unbiased estimator.

The Akaike Information Criterion (AIC) is equal to minus two times the expected value of the Kullback-Leibler Information. The multiplication by -2 is completely arbitrary. Akaike did it because minus two times the log-likelihood, which is also known as the deviance, was used widely in statistical inference.

One of the major benefits of using the AIC is that we do not need to know the true model. Imagine, we have two competing models with probability densities g and h, respectively. We can now write

$$\begin{aligned} AIC_1 &= -2\ell_1 + 2K_1 = -2C + 2\hat{E}_{\hat{\theta}_1}[I(f,\hat{g})] \\ AIC_2 &= -2\ell_2 + 2K_2 = -2C + 2\hat{E}_{\hat{\theta}_2}[I(f,\hat{h})] \end{aligned}$$

Each of the AIC values still contains the component C, which pertains to the true model, which is generally unknown. However, when we subtract the minimum AIC then this term disappears. It is easy to show this. For argument's sake, let g yield the better—i.e., smaller—value of AIC. We can now subtract AIC_1 from each of the terms above:

$$\begin{aligned} \Delta_1 &= AIC_1 - AIC_1 \\ &= 0 \\ \Delta_2 &= AIC_2 - AIC_1 \\ &= \left\{-2C + 2\hat{E}_{\hat{\theta}_2}[I(f,\hat{h})]\right\} - \left\{-2C + 2\hat{E}_{\hat{\theta}_1}[I(f,\hat{g})]\right\} \\ &= 2\left\{\hat{E}_{\hat{\theta}_2}[I(f,\hat{h})] - \hat{E}_{\hat{\theta}_1}[I(f,\hat{g})]\right\} \end{aligned}$$

We observe that the terms involving C drop out, which means that we do not need to know what the true model is.

C.4 Non-Linear Models

C.4.1 Marginal Effect in the Log-Linear Model

The log-linear model is given by

$$\ln y_i = \beta_0 + \sum_k \beta_k \ln x_{ik} + \varepsilon_i$$

Writing this in terms of the dependent variable, we obtain

$$y_i = \exp\left(\beta_0 + \sum_k \beta_k \ln x_{ik} + \varepsilon_i\right)$$

Taking the derivative with respect to x_j yields

$$\frac{\partial y}{\partial x_j} = \frac{\beta_j}{x_j} \exp\left(\beta_0 + \sum_k \beta_k \ln x_{ik} + \varepsilon_i\right) = \beta_j \frac{y}{x_j}$$

Solving for β_j, we obtain

$$\beta_j = \frac{\partial y / y}{\partial x_j / x_j}$$

The numerator is the relative change in the mean, whereas the denominator is the relative change in the predictor of interest. We can turn these relative changes into percentage changes by multiplying both the numerator and the denominator by a factor of 100, which of course does not affect the overall result.

C.4.2 Marginal Effects in Semi-Log Models

The generic log-lin model may be written as $y_i = \exp(\beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i)$. Taking the partial derivative with respect to x_j yields

$$\frac{\partial y}{\partial x_j} = \beta_j \exp\left(\beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i\right) = \beta_j y$$

Solving for β_j, we get

$$\beta_j \quad = \quad \frac{\partial y/y}{\partial x_j}$$

The numerator may be viewed as the relative change in Y, whereas the denominator is the absolute change in x_j.

The generic lin-log model is given by $y_i = \beta_0 + \sum_k \beta_k \ln x_{ik} + \varepsilon_i$. Taking the partial derivative with respect to x_j yields

$$\frac{\partial y}{\partial x_j} \quad = \quad \frac{\beta_j}{x_j}$$

Solving for β_j, we get

$$\beta_j \quad = \quad x_j \frac{\partial y}{\partial x_j} = \frac{\partial y}{\partial x_j / x_j}$$

The numerator is an absolute change in Y, whereas the denominator is a partial change in x_j.

C.5 Interaction Effects

C.5.1 Covariance Between the Interaction and Its Constituent Terms

Consider $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i \times z_i + \varepsilon_i$, where X and Z are covariates. We derive the covariance between the interaction and X. By definition, $\sigma_{xz,x} = E[(xz - E[xz])(x - E[x])]$. Let $x^d = x - E[x]$. Further, we know from mathematical statistics that $x \cdot z = (x^d + E[x])(z^d + E[z]) = x^d z^d + x^d E[z] + z^d E[x] + E[x]E[z]$ and that $E[xz] = E[x^d z^d] + E[x^d]E[z] + E[z^d]E[x] + E[x]E[z] = \sigma_{x,z} + E[x]E[z]$. It then follows that $xz - E[xz] = x^d z^d + x^d E[z] +$

$z^d E[x] - \sigma_{x,z}$. With all of these preparatory steps out of the way, we now get

$$\begin{aligned}
\sigma_{xz,x} &= E[(xz - E[xz])(x - E[x])] \\
&= E[(x^d z^d + x^d E[z] + z^d E[x] - \sigma_{x,z}) x^d] \\
&= E\left[\left(x^d\right)^2 z^d\right] + \left(x^d\right)^2 E[z] + x^d z^d E[x] - x^d \sigma_{x,z} \\
&= E\left[\left(x^d\right)^2 z^d\right] + E\left[\left(x^d\right)^2\right] E[z] + E[x^d z^d] E[x] - E[x^d] \sigma_{x,z} \\
&= E\left[\left(x^d\right)^2 z^d\right] + \sigma_x^2 \mu_z + \sigma_{x,z} \mu_x
\end{aligned}$$

The last equation comes about by realizing that $E[(x^d)^2] = \sigma_x^2$, $E[x^d z^d] = \sigma_{x,z}$, and $E[x^d] = 0$. Under multivariate normality, $E\left[\left(x^d\right)^2 z^d\right] = 0$, since this is the multivariate moment of order 2 and 1, which is by definition 0. The covariance between the interaction and z is derived analogously.

C.5.2 The Effect of Centering

The centered regression model is given by

$$y_i = \beta_0 + \beta_1 x_i^d + \beta_2 z_i^d + \beta_3 x_i^d \times z_i^d + \varepsilon_i$$

Since $x_i^d = x_i - \bar{x}$ and $z_i^d = z_i - \bar{z}$, we may also write the model as

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_2 (z_i - \bar{z}) + \beta_3 (x_i - \bar{x}) \times (z_i - \bar{z}) + \varepsilon_i$$

Upon expansion we get

$$\begin{aligned}
y_i = &(\beta_0 - \beta_1 \bar{x} - \beta_2 \bar{z} + \beta_3 \bar{x} \bar{z}) + \\
&(\beta_1 - \beta_3 \bar{z})\, x_i + \\
&(\beta_2 - \beta_3 \bar{x})\, z_i + \\
&\beta_3 x_i \times z_i + \varepsilon_i
\end{aligned}$$

We see that the effect of the centered interaction term is identical to that of the uncentered term. We also see that the effect of the uncentered version of

x is equal to $\beta_1 - \beta_3\bar{z}$, that the effect of the uncentered version of z is equal to $\beta_2 - \beta_3\bar{x}$, and that the uncentered intercept is $\beta_0 - \beta_1\bar{x} - \beta_2\bar{z} + \beta_3\bar{x}\bar{z}$.

C.6 Influence and Normality

C.6.1 PRESS Residuals

By definition, the PRESS residuals, p_i, are equal to $y_i - \hat{y}_{i(i)}$, where $\hat{y}_{i(i)}$ is the predicted value of the ith observation when the regression coefficients are computed by omitting that observation. Kmenta (1997) shows that

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{(i)}} = \frac{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i e_i}{1 - h_{ii}},$$

where $\hat{\boldsymbol{\beta}_{(i)}}$ is the OLS estimator of $\boldsymbol{\beta}$ that comes about when the ith observation is deleted. Now

$$\begin{aligned}
p_i &= y_i - x_i^\top\hat{\boldsymbol{\beta}}_{(i)} \\
&= y_i - \mathbf{x}_i^\top\left[\hat{\boldsymbol{\beta}} - \frac{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i e_i}{1 - h_{ii}}\right] \\
&= \underbrace{y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}}_{e_i} + \frac{\underbrace{\mathbf{x}_i^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i}_{h_{ii}}\, e_i}{1 - h_{ii}} \\
&= e_i + \frac{h_{ii}e_i}{1 - h_{ii}} \\
&= \frac{e_i}{1 - h_{ii}}
\end{aligned}$$

Glossary

Akaike Information Criterion The Akaike Information Criterion provides a metric for model evaluation that allows the researcher to select the best model from a set. 132

analysis of covariance A statistical model in which the dependent variable depends on factors and their interactions as well as one or more covariates. The covariates are typically centered about their sample means. 167

analysis of variance (1) The decomposition of the sample variation in the dependent variable (SST) into a part that is due to the regression (SSR) and a part that is due to error (SSE). (2) A statistical model in which the dependent variable depends solely on factors and their interactions. 108

autocorrelation The condition that different error terms are correlated with each other: $Cov[\varepsilon_i, \varepsilon_j] \neq 0$ for $i \neq j$. In time series analysis, this is also referred to as serial correlation. 32

ceteris paribus Holding all else equal. 85

coefficient of determination Also known as the R-squared, the coefficient of determination reveals how much of the variation in the dependent variable is accounted for by the regression. Formulaically, $R^2 = SSR/SST = 1 - SSE/SST$. 13

covariate A continuous predictor. 8

cross-sectional data Cross sections of a population of units. A set of units from the same population is analyzed at a single point in time. 32

data generating process A process that is believed to have generated the data that were collected. Often abbreviated as DGP. 22

dependent variable The variable that is being predicted or explained. Typically denoted as Y. 8

discrete change The change in an outcome that is due to a change of δ units in a predictor, while holding all else constant. 22

dummy variable A variable that takes on the values 1 and 0. Typically used in regression analysis to absorb the effects of discrete predictors. 166

elasticity The percentage change in an outcome that we can expect from a one percent increase in a predictor, while holding all else constant. 160

error term A.k.a. the disturbance or simply the error. A stochastic component that remains unobserved and is added to the model to absorb: (1) omitted predictors; (2) measurement error in the dependent variable; and (3) idiosyncratic variation in the dependent variable. The term should be distinguished from residuals, which are observed and not stochastic. Error terms are specified at the level of a population, whereas residuals pertain to samples. 23

estimate The value that an estimator takes in a particular sample. 38

estimator A rule for computing a parameter based on the sample. 38

evidence ratio The ratio of the likelihoods of two models given the data. 146

exogeneity A variable or parameter is exogenous to the extent that it is determined by forces external to the model. Exogeneity can be strong or weak. Under strong exogneity, $E[\varepsilon|x] = 0$. Under weak exogneity, $E[\varepsilon x] = 0$. 33

factor A discrete or categorical predictor. 8

hat matrix A $n \times n$ matrix that maps the observed into the fitted values in a regression model. 83

hat value A diagonal element of the hat matrix. 242

homoskedasticity The assumption that the variance of the error term is constant across units: $Var[\varepsilon_i] = \sigma^2 =$ constant. When this assumption is violated, we say that there is heteroskedasticity. 24

influence An attribute of a data point, whereby it has a strong influence on the partial slope coefficient. Influence results from the combination of a data point being a leverage point and an outlier. 239

interaction The product of two variables that allows us to explore moderation in a regression model. 193

intercept The predicted value of the dependent variable when the predictor(s) is/are 0. In simple regression analysis, this is the point where the crosses the y-axis. 5

leverage point An observation with an atypical value on the predictor. 239

linear constraint A linear function involving the parameters of a statistical model. 114

marginal effect The rate of change in an outcome, i.e., the change in an outcome relative to an infinitesimally small change in a predictor, while holding all else equal. If q is the outcome of interest (e.g., the mean), then the marginal effect is given by $\partial q/\partial x$. 21

maximum likelihood An estimation approach whereby the parameter values are chosen in such a manner that they maximize the likelihood of the data. 47

mean squares due to error The average of the squared residuals: $MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-K-1} = s^2$, where K is the number of regression coefficients, excluding the constant. Also known as the mean squared error. 57

method of moments An estimation approach that exploits moment conditions implied by the model in order to obtain an estimator. 44

moderator A variable that influences the relationship between a predictor and the dependent variable. 193

multicollinearity The situation that one or more predictor variables are (near) perfect linear functions of other predictor variables.. 79

nested model A model $\mathcal{M}_j$ is nested inside another model $\mathcal{M}_k$ if it is a subset of the latter: $\mathcal{M}_j \subset \mathcal{M}_k$. This means that we can derive $\mathcal{M}_j$ by constraining one or more parameters in $\mathcal{M}_k$. 133

ordinary least squares An estimation approach whereby the sum of the squared residuals is being minimized with respect to the parameters. 39

outlier An observation with an atypical value on the dependent variable. 239

partial slope A slope coefficient associated with a predictor that is net of the effects of other predictors in the model. 72

population regression function Also known as the conditional expectation function, this gives the conditional expectation of the dependent variable in a linear regression analysis: $\mu_i = E[y_i|\mathbf{x}_i]$. In simple regression analysis, $\mu_i = \beta_0 + \beta_1 x_i$; in multiple regression analysis, $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. 24

population regression model The regression model stated at the level of the population, including the error term. The simple population regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The multiple population regression model is $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$. 24

prediction The value that the dependent variable is expected to take given a set of parameter estimates and a set of assumed values of the predictors. Also known as predicted or fitted value. 8

predictor A variable that is used to predict the dependent variable. Typically denoted as X, this is also known as the regressor or independent variable. 8

regression line In simple regression analysis, a linear equation that generates predictions on the basis of the values taken on by a single predictor. The generic formula is $\hat{y}_i = a + b \cdot x_i$. 5

regression through the origin A regression model in which the constant is dropped. In simple regression analysis, this means that we force the regression line to go through the origin. 11

residual The discrepancy between the observed and predicted values of the dependent variable: $e_i = y_i - \hat{y}_i$. 5

root mean squared error The square root of the MSE or $s = \sqrt{\sum_i e_i^2/(n-K-1)}$. Also known as the residual standard error, this can be used as a measure of model fit. 130

sample regression function A function that is linear in the parameters and that produces predicted values of the dependent variable in the sample based on one or more predictors. In simple regression analysis, this function is equal to $\hat{y}_i = a + b \cdot x_i$. In multiple regression analysis, this is $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$. 8

sample regression model The regression model stated at the level of the sample, including the residuals. The simple sample regression model is $y_i = a + b \cdot x_i + e_i$. The multiple sample regression model is $y_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + e_i$. 8

sample size The number of observations in the sample/data. Typically indicated as n. 9

simple slope The marginal effect of a predictor evaluated at different values of a moderator variable. 201

slope The rate of change in the dependent variable for a change in the predictor variable. 5

specification The process of developing a statistical model, which includes the selection of predictors, the function linking the predictors to the dependent variable, and distributional assumptions about the dependent variable. 23

specification error Any erroneous choice in the model specification. Here, erroneous means that the model postulates a data generating process that departs from the true data generating process. 34

standardized regression coefficient A partial slope coefficient that removes the measurement units of both the dependent variable and the predictor. It is computed as $\hat{\beta}_j^s = \hat{\beta}_j \frac{s_{x_j}}{s_y}$. In simple regression analysis, the standardized regression coefficient is equal to the correlation between the predictor and the dependent variable. 90

sum of squared errors The sum of the squared residuals: $SSE = \sum_{i=1}^{n} e_i^2$. Also known as the residual sum of squares. 12

sum of squares regression The sum of the squares deviations of the predicted values and the mean of the dependent variable: $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$. 108

sum of squares total The variation in the dependent variable, i.e., $(n-1) \cdot s_Y^2$. 13

time series data Data for which the units are successive time points such as days, weeks, months, quarters, or years. 33

variance-covariance matrix of the estimators For the multiple regression model, the VCE of the regression coefficients is given by $\sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$. 103

Bibliography

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Newbury Park, CA: Sage Publications.

Achen, Christopher H. 1990. "What Does 'Explained Variance' Explain?: Reply." *Political Analysis* 2(1):173–184.

Aiken, Leona S. and Stephen G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage Publications.

Baumgartner, F.R., C. Breunig, C. Green-Pedersen, B.D. Jones, P.B. Mortensen, M. Nuytemans and S. Walgrave. 2009. "Punctuated Equilibrium in Comparative Perspective." *American Journal of Political Science* 53(3):603–620.

Belsley, David A., Edwin Kuh and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1):63–82.

Brown, Steven R. and Lawrence E. Melamed. 1990. *Experimental Design and Analysis*. Newbury Park, CA: Sage Publications.

Bryman, A. 1988. *Quantity and Quality in Social Research*. London: Unwin Hyman.

Burnham, Kenneth P. and David R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33(2):261–304.

Card, David and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4):772–793.

Chamberlin, Thomas C. 1965. "The Method of Multiple Working Hypotheses." *Science* 148(3671):754–759.

Chatterjee, Samprit and Ali S. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. New York: Wiley.

Chirot, Daniel and Charles Ragin. 1975. "The Market, Tradition and Peasant Rebellion: The Case of Romania 1907." *American Sociological Review* 40(4):428–444.

Cochran, William G. 1934. "The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance." *Mathematical Proceedings of the Cambridge Philosophical Society* 30(2):178–191.

Davidson, Russell and James G. MacKinnon. 1981. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." *Econometrica* 49(3):781–793.

Engle, R.F., D.F. Hendry and J-F. Richard. 1983. "Exogeneity." *Econometrica* 51(2):277–304.

Fisher, R.A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222:309–368.

Fox, John. 2003. "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software* 8(15):1–9.

Fox, John and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Thousand Oaks, CA: Sage Publications.

Gilchrist, W. 1984. *Statistical Modelling*. Chichester, UK Wiley.

Greene, W.H. 2011. *Econometric Analysis*. 7 ed. Upper Saddle River, NJ: Prentice Hall.

Greenwald, Anthony G., Anthony R. Pratkanis, Michael R. Leippe and Michael H. Baumgardner. 1986. "Under What Conditions Does Theory Obstruct Research Progress?" *Psychological Review* 93(2):216–229.

Grömping, U. 2006. "Relative Importance for Linear Regression in R: The Package relaimpo." *Journal of Statistical Software* 17(1):1–27.

Holm, Sture. 1979. "A Simple Sequentiallly Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6(2):65–70.

Hume, David. 1993. *An Enquiry Concerning Human Understanding*. Indianapolis, IN: Hackett.

Jobson, J.D. 1991. *Applied Multivariate Data Analysis, Vol. I: Regression and Experimental Design*. New York: Springer.

Johnson, Palmer O. and Jerzy Neyman. 1936. "Tests of Certain Linear Hypotheses and Their Applications to Some Educational Problems." *Statistical Research Memoirs* 1:57–93.

Kam, Cindy and Robert J. Franzese. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor, MI: University of Michigan Press.

King, Gary. 1990. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The R-Squared'." *Political Analysis* 2(1):185–200.

Kmenta, J. 1997. *Elements of Econometrics*. 2 ed. Ann Arbor, MI: University of Michigan Press.

Kumbhakar, S.C. and C.A.K. Lovell. 2003. *Stochastic Frontier Analysis*. New York: Cambridge University Press.

Lehmann, Erich L. 1990. "Model Specification: The Views of Fisher and Neyman, and Later Developments." *Statistical Science* 5(2):160–168.

Lewis-Beck, Michael S. and Andrew Skalaban. 1990. "The R-Squared: Some Straight Talk." *Political Analysis* 2(1):153–171.

Lindeman, R.H., P.F. Merenda and R.Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis*. Glenview, IL: Scott and Foresman.

McCullagh, P. and J.A. Nelder. 1983. *Generalized Linear Models*. London: Chapman and Hall.

Wagenmakers, Eric-Jan and Simon Farrell. 2004. "AIC Model Selection Using Akaike Weights." *Psychonomic Bulletin & Review* 11(1):192–196.