CHAPTER 4

An introduction to simple regression

Regression is the most important tool applied economists use to understand the relationship among two or more variables. It is particularly useful for the common case where there are many variables (e.g. unemployment and interest rates, the money supply, exchange rates, inflation, etc.) and the interactions between them are complex.

To give an example, in the summer of 1998 a great deal of attention in the UK media focussed on the proper level at which interest rates should be set. In particular, the manufacturing sector complained that interest rates were too high. They argued that high interest rates encouraged foreigners to invest their money in the UK which, in turn, caused the pound to appreciate. A higher pound made it difficult for UK firms to export their products, resulting in falling sales, increased layoffs and rising unemployment.

But this is only part of the story. Still others believed that interest rates were too low, and argued that higher interest rates were necessary to choke off inflationary pressures due to a relationship between inflation and interest rates. Thus, an important economic question (i.e. interest rate determination) was at stake, and a large number of variables – interest rates, exchange rates, inflation, manufacturing output, exports, unemployment – must be considered in arriving at an answer to the problem. All these variables (and more) shaped the discussion of what the relevant interest rate should be.

As a second example, consider the problem of trying to explain the price of houses. The price of a house depends on many characteristics (e.g. number of bedrooms, number of bathrooms, location of house, size of lot, etc.). As in the above example,

many variables must be included in a model seeking to explain why some houses are more expensive than others.

These two examples are not unusual. Most problems in economics are of a similar level of complexity. Unfortunately, the basic tool you have encountered so far – simple correlation analysis – cannot handle such complexity. For these more complex cases – that is, those involving more than two variables – regression is the tool to use.

Regression as a best fitting line

As a way of understanding regression, let us begin with just two variables (Y and X). We refer to this case as simple regression. Multiple regression, involving many variables, will be discussed in Chapter 6. Beginning with simple regression makes sense since graphical intuition can be developed in a straightforward manner and the relationship between regression and correlation can be illustrated quite easily.

Let us return to the XY-plots used previously (e.g. Figure 2.3 which plots population density against deforestation or Figure 3.1 which plots lot size against house price). We have discussed in Chapters 2 and 3 how an examination of these XY-plots can reveal a great deal about the relationship between X and Y. In particular, a straight line drawn through the points on the XY-plot provides a convenient summary of the relationship between X and Y. In regression analysis, we formally analyze this relationship.

To start with, we assume that a linear relationship exists between Y and X. As an example, you might consider Y to be the house price variable and X to be the lot size variable from data set HPRICE.XLS. Remember that this data set contained the sales price of 546 houses in Windsor, Canada along with several characteristics for each house. It is sensible to assume that the size of the lot affects the price at which a house sells.

We can express the linear relationship between Y and X mathematically as:[1]

$$Y = \alpha + \beta X,$$

where α is the intercept of the line and β is the slope. This equation is referred to as the **regression line**. If in actuality we knew what α and β were, then we would know what the relationship between Y and X was. In practice, of course, we do not have this information. Furthermore, even if our **regression model**, which posits a linear relationship between Y and X, were true, in the real world we would never find that our data points lie precisely on a straight line. Factors such as measurement error mean that individual data points might lie close to but not exactly on a straight line.

For instance, suppose the price of a house (Y) depends on the lot size (X) in the following manner: $Y = 34{,}000 + 7X$ (i.e. $\alpha = 34{,}000$ and $\beta = 7$). If X were 5,000 square feet, this model says the price of the house should be $Y = 34{,}000 + 7 \times 5{,}000 = \$69{,}000$. But, of course, not every house with a lot size of 5,000 square feet will

have a sales price of precisely $69,000. No doubt in this case, the regression model is missing some important variables (e.g. number of bedrooms) that may affect the price of a house. Furthermore, the price of some houses might be higher than they should be (e.g. if they were bought by irrationally exuberant buyers). Alternatively, some houses may sell for less than their true worth (e.g. if the sellers have to relocate to a different city and must sell their houses quickly). For all these reasons, even if $Y = 34{,}000 + 7X$ is an accurate description of a straight line relationship between Y and X, it will not be the case that every data point lies exactly on the line.

Our house price example illustrates a truth about regression modeling: **the linear regression model will always be only an approximation of the true relationship**. The truth may differ in many ways from the approximation implicit in the linear regression model. In economics, the most probable source of error is due to missing variables, usually because we cannot observe them. In our previous example, house prices reflect many variables for which we can easily collect data (e.g. number of bedrooms, number of bathrooms, etc.). But they will also depend on many other factors for which it is difficult if not impossible to collect data (e.g. the number of loud parties held by neighbors, the degree to which the owners have kept the property well-maintained, the quality of the interior decoration of the house, etc.). The omission of these variables from the regression model will mean that the model makes an error.

We call all such errors e. The regression model can now be written as:

$$Y = \alpha + \beta X + e\,.$$

In the regression model, Y is referred to as the **dependent** variable, X the **explanatory** variable, and α and β, **coefficients**. It is common to implicitly assume that the explanatory variable "causes" Y, and the coefficient β measures the influence of X on Y. In light of the comments made in the previous chapter about how correlation does not necessarily imply causality, you may want to question the assumption that the explanatory variable causes the dependent variable. There are three responses that can be made to this statement.

First, note that we talk about the regression **model**. A model specifies how different variables interact. For instance, models of land use posit that population pressures cause rural farmers to expand their lands by cutting down forests, thus causing deforestation. Such models have the causality "built-in" and the purpose of a regression involving Y = deforestation and X = population density is to measure the magnitude of the effect of population pressures only (i.e. the causality assumption may be reasonable and we do not mind assuming it). Secondly, we can treat the regression purely as a technique for generalizing correlation and interpret the numbers that the regression model produces purely as reflecting the association between variables. (In other words, we can drop the causality assumption if we wish.) Thirdly, we can acknowledge that the implicit assumption of causality can be a problem and develop new methods. This issue will be discussed briefly in the last chapter of this book.[2]

In light of the error, e, and the fact that we do not know what α and β are, the first problem in regression analysis is how we can figure approximately, or **estimate**, what α and β are. It is standard practice to refer to the estimates of α and β as $\hat{\alpha}$ and $\hat{\beta}$ (i.e. $\hat{\alpha}$ and $\hat{\beta}$ are actual numbers that the computer calculates, for instance, $\hat{\alpha} =$ 34,136 and $\hat{\beta} = 6.599$, which are estimates of the unknown true values $\alpha = 34{,}000$ and $\beta = 7$). In practice, the way we find estimates is by drawing a line through the points on an *XY*-plot which fits best. Hence, we must define what we mean by "best fitting line".

Before we do this, it is useful to make a distinction between **errors** and **residuals**. The error is defined as the distance between a particular data point and the true regression line. Mathematically, we can rearrange the regression model to write $e_i = Y_i - \alpha - \beta X_i$. This is the **error** for the ith observation. However, if we replace α and β by their estimates $\hat{\alpha}$ and $\hat{\beta}$, we get a straight line which is generally a little different from the true regression line. The deviations from this estimated regression line are called **residuals**. We will use the notation "u" when we refer to residuals. That is, the residuals are given by $u_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$. If you find the distinction between errors and residuals confusing, you can probably ignore it in the rest of this book and assume errors and residuals are the same thing. However, if you plan on further study of econometrics, this distinction becomes crucial.

If we return to some basic geometry, note that we can draw one (and only one) straight line connecting any two distinct points. Thus, in the case of two points, there is no doubt about what the best fitting line through an *XY*-plot is. However, typically we have many points – for instance, our deforestation/population density example has 70 different countries and the *XY*-plots 70 points – and there is ambiguity about what is the "best fitting line". Figure 4.1 plots three data points (A, B and C) on an *XY* graph. Clearly, there is no straight line that passes through all three points. The line I have drawn does not pass through any of them; each point, in other words, is

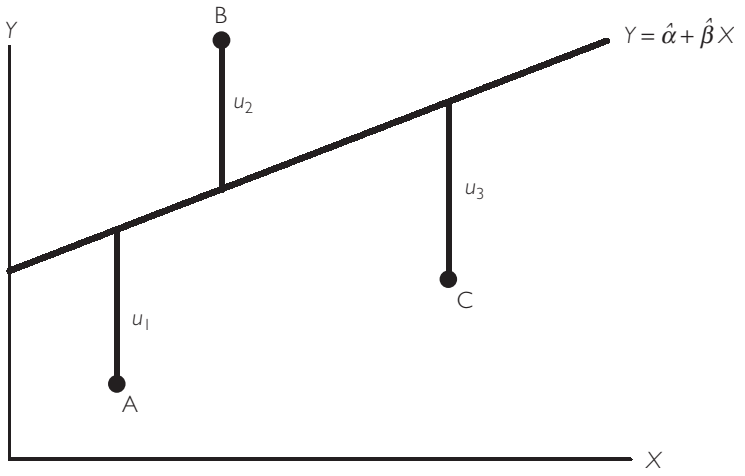

Fig. 4.1 Best fitting line for three data points.

a little bit off the line. To put it another way: the line drawn implies residuals that are labeled u_1, u_2 and u_3. The residuals are the vertical difference between a data point and the line. A good fitting line will have small residuals.

The usual way of measuring the size of the residuals is by means of the sum of squared residuals (**SSR**), which is given by:

$$\text{SSR} = \sum_{i=1}^{N} u_i^2,$$

for $i = 1, \ldots, N$ data points. We want to find the best fitting line which minimizes the sum of squared residuals. For this reason, estimates found in this way are called **least squares** estimates (or ordinary least squares – **OLS** – to distinguish them from more complicated estimators which we will not discuss until the last chapter of this book).

In practice, software packages such as Excel can automatically find values for $\hat{\alpha}$ and $\hat{\beta}$ which will minimize the sum of squared residuals. The exact formulae for $\hat{\alpha}$ and $\hat{\beta}$ can be derived using simple calculus, but we will not derive them here (see Appendix 4.1 for more details).

Example: The regression of deforestation on population density

Consider again the data set FOREST.XLS, which contains data on population density and deforestation for 70 tropical countries. It makes sense to assume that population density influences deforestation rather than the other way around. Thus we choose deforestation as the dependent variable (i.e. Y = deforestation) and population as the explanatory variable (i.e. X = population density). Using Excel (Tools/Data Analysis/Regression) we obtain $\hat{\alpha} = 0.60$ and $\hat{\beta} = 0.000842$. To provide some more jargon, note that when we estimate a regression model it is common to say that "we run a regression of Y on X".

Note also that it is actually very easy to calculate these numbers in most statistical software packages. Appropriately, we will turn instead to the more important issue: how we interpret these numbers.

Example: Cost of production in the electric utility industry

The file ELECTRIC.XLS contains data on the costs of production (measured in millions of dollars) for 123 electric utility companies in the US in 1970. Interest centers on understanding the factors which affect costs. Hence, Y = cost of production is the dependent variable. The costs incurred by an electric utility company can potentially depend on many factors. One of the most important

of these is undoubtedly the output (measured as thousands of kilowatt hours of electricity produced) of the company. We would expect companies that are producing more electricity will also be incurring higher costs (e.g. because they have to buy more fuel to generate the electricity). Hence, X = output is a plausible explanatory variable. If we run the regression of costs on output, we obtain $\hat{\alpha} = 2.19$ and $\hat{\beta} = 0.005$.

Example: The effect of advertising on sales

The file ADVERT.XLS contains data on annual sales and advertising expenditures (both measured in millions of dollars) for 84 companies in the US. A company executive might be interested in trying to quantify the effect of advertising on sales. This suggests running a regression with dependent variable Y = sales and explanatory variable X = advertising expenditures. Doing so, we obtain the value $\hat{\alpha} = 502.02$ and $\hat{\beta} = 0.218$, which is indicative of a positive relationship between advertising and sales.

Interpreting OLS estimates

In the previous example of the relationship between deforestation and population density, we obtained OLS estimates for the intercept and slope of the regression line. The question now arises: how should we interpret these estimates? The intercept in the regression model, α, usually has little economic interpretation so we will not discuss it here. However, β is typically quite important. This coefficient is the slope of the best fitting straight line through the XY-plot. In the deforestation/population density example, $\hat{\beta}$ was positive. Remembering the discussion on how to interpret correlations in the previous chapter, we note that since $\hat{\beta}$ is positive X and Y are positively correlated. However, we can go further in interpreting $\hat{\beta}$ if we differentiate the regression model and obtain:

$$\frac{dY}{dX} = \beta .$$

Even if you do not know calculus, the verbal intuition of the previous expression is not hard to provide. Derivatives measure how much Y changes when X is changed by a small (marginal) amount. Hence, β can be interpreted as the **marginal effect** of X on Y and is a measure of how much X influences Y. To be more precise, we can interpret β as a measure of how much Y tends to change when X is changed by one unit.[3] The definition of "unit" in the previous sentence depends on the particular

data set being studied and is best illustrated through examples. Before doing this, it should be stressed that regressions measure tendencies in the data (note the use of the word "tends" in the explanation of β above). It is not necessarily the case that every observation (e.g. country or house) fits the general pattern established by the other observations. In Chapter 2 we called such unusual observations outliers and argued that, in some cases, examining outliers could be quite informative. In the case of regression, outliers are those with residuals that stand out as being unusually large. Hence, examining the residuals from a regression is a common practice. (In Excel you can examine the residuals by clicking on the box labeled "Residuals" in the regression menu.)

Example: The regression of deforestation on population density (continued from page 53)

In the deforestation/population density example we obtained $\hat{\beta} = 0.000842$. This is a measure of how much deforestation tends to change when population density changes by a small amount. Since population density is measured in terms of the number of people per 1,000 hectares and deforestation as the percentage forest loss per year, this figure implies that if we add one more person per 1,000 hectares (i.e. a change of one unit in the explanatory variable) deforestation will tend to increase by 0.000842%.

Alternatively, we could present this information as follows. The population density varies quite a bit across countries: from below 100 people to over 2,500 people per 1,000 hectares. Hence it is not surprising that a change of one person per hectare will have little effect on deforestation. We could multiply everything by 100 and say that "increasing population density by 100 people per thousand hectares will tend to increase deforestation by 0.0842%". Even the latter number may seem insignificant, but note that an increase of annual deforestation rates by 0.0842% per year will result in a country losing an extra 5% of its forest over 50 years. In the long run and over a large area – the spatial and time scales in which environmental economists are accustomed to thinking – this degree of forest loss can be substantial.

Example: Cost of production in the electric utility industry (continued from page 54)

In the regression of company costs on output, we obtained $\hat{\beta} = 0.005$. Remember that β units is the effect on the dependent variable of a one unit change in the explanatory variable. Since output is measured in thousands of kWh, a one

unit change in the explanatory variable is one thousand kWh. Since costs are measured in millions of dollars, β units is β million dollars. Combining these facts we can say that "increasing output by one thousand kWh tends to increase costs by \$5,000 (i.e. $0.005 \times 1{,}000{,}000 = 5{,}000$)".

Of course, we could also express this in terms of a decrease of one unit. That is, we could say, "decreasing output tends to decrease costs by \$5,000".

Example: The effect of advertising on sales (continued from page 54)

Both advertising and sales are measured in millions of dollars and we found $\hat{\beta} = 0.218$. Following the same line of reasoning above, we can say that a one million dollar increase in advertising tends to be associated with a \$218,000 increase in sales (i.e. $1{,}000{,}000 \times 0.218 = 218{,}000$). This result would seem to indicate that spending on advertising is rather counterproductive since an extra \$1,000,000 spent on advertising would only translate into an extra \$218,000 in sales.

Does this mean that the company executive running this regression should decide to reduce advertising expenditures? Possibly, but not necessarily. The reason for this uncertainty relates to the issue of causality and the question of how correlation or regression results can be interpreted (see Chapter 3 or earlier in this chapter). That is, if the regression truly is a causal one (i.e. it is the case that advertising has a direct influence on sales), then we can interpret the \$218,000 figure as indicative of what the effect of a change in advertising will be. However, if it is not causal, then it is risky to use the regression result to provide strategic advice to a company. Indeed, it is possible that larger companies tend to have egomaniacs as bosses and egomaniacs enjoy seeing their companies advertised. If this (possibly implausible) story is true then we would expect to see larger companies advertising more – exactly what our regression has found. Such an interpretation would imply that it is possible that advertising is not directly influencing sales. The apparent positive relationship between advertising and sales from the regression analysis may be due solely to the behavior of the bosses of large companies.

Deciding whether it is reasonable to assume that a regression model captures a causal relationship in which one variable directly influences another is very difficult, and it is hard to offer any general rules on the subject. Perhaps the best advice is to draw on common sense and economic theory to guide you in interpretation.

Exercise 4.1

The Excel data set FOREST.XLS contains data on Y = deforestation, X = population density, W = change in cropland and Z = change in pasture land.

(a) Run a regression of Y on X and interpret the results.
(b) Run a regression of Y on W and one of Y on Z and interpret the results.
(c) Create a new variable, V, by dividing X by 100. What are the units in terms of which V is measured?
(d) Run a regression of Y on V. Compare your results to those for (a). How do you interpret your coefficient estimate of β? How does $\hat{\alpha}$ differ between (a) and (d)?
(e) Experiment with scaling dependent and explanatory variables (i.e. by dividing them by a constant) and see what effect this has on your coefficient estimates.

Fitted values and R^2: measuring the fit of a regression model

In the preceding discussion we learned how to calculate and interpret regression coefficients, $\hat{\alpha}$ and $\hat{\beta}$. Furthermore, we explained that regression finds the "best fitting" line in the sense that it minimizes the SSR. However, it is possible that the "best" fit is not a very good fit at all. Appropriately, it is desirable to have some measure of fit (or a measure of how good the best fitting line is). The most common measure of fit is referred to as the R^2. It relates closely to the correlation between Y and X. In fact, for the simple regression model, it is the correlation squared. This provides the formal statistical link between regression and correlation. However, the previous discussion should make the informal links between correlation and regression clear. Both are interested in quantifying the degree of association between different variables and both can be interpreted in terms of fitting lines through XY-plots.

To derive and explain R^2, we will begin with some background material. We start by clarifying the notion of a **fitted value**. Remember that regression fits a straight line through an XY-plot, but does not pass precisely through each point on the plot (i.e. an error is made). In the case of our deforestation/population density example, this meant that individual countries did not lie on the regression line. The fitted value for observation i is the value that lies on the regression line corresponding to the X_i value for that particular observation (e.g. house, country). In other words, if you draw a straight vertical line through a particular point in the XY-plot, the intersection of this vertical line and the regression line is the fitted value corresponding to the point you chose.

Alternatively, we can think of the idea of a fitted value in terms of the formula for the regression model:

$$Y_i = \alpha + \beta X_i + e_i .$$

Remember that adding *i* subscripts (e.g. Y_i) indicates that we are referring to a particular observation (e.g. the *i*th country or the *i*th house). If we ignore the error, we can say that the model's prediction of Y_i should be equal to $\alpha + \beta X_i$. If we replace α and β by the OLS estimates $\hat{\alpha}$ and $\hat{\beta}$, we obtain a so-called "fitted" or "predicted" value for Y_i:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i .$$

Note that we are using the value of the explanatory variable and the OLS estimates to predict the dependent variable. By looking at actual (Y_i) versus fitted ($\hat{Y}_i$) values we can gain a rough impression of the "goodness of fit" of the regression model. Many software packages allow you to print out the actual and fitted values for each observation. An examination of these values not only gives you a rough measure of how well the regression model fits, they allow you to examine individual observations to determine which ones are close to the regression line and which are not. Since the regression line captures general patterns or tendencies in your data set, you can see which observations conform to the general pattern and which do not.

Exercise 4.2

Using the data in FOREST.XLS (see Exercise 4.1), run a regression of Y on X using Excel with the box clicked on labeled "Line Fit Plot" in the regression menu. Graphically and numerically compare the actual to the fitted values (i.e. look at the columns labeled "Residual Output" and the accompanying display chart).

We have defined the residual made in fitting our best fitting line previously. Another way to express this residual is in terms of the difference between the actual and fitted values of Y. That is:

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

Software packages such as Excel can also plot or list the residuals from a regression model. These can be examined in turn to give a rough impression of the goodness of fit of the regression model. We emphasize that unusually big residuals are outliers and sometimes these outliers are of interest.

Exercise 4.3

(a) Using the data in FOREST.XLS (see Exercise 4.1) run a regression of Y on X using Excel with the boxes labeled "Residuals" and "Residual Plots" in the regression menu clicked on. How would you interpret the residuals? Are there any outliers?

(b) Repeat question (a) for the other variables, W and Z in this data set.

To illustrate the kind of information with which residual analysis can provide us, take a look at your computer output from Exercise 4.3 (a). In the Residual Output, observation 39 has a fitted value of 2.93 and a residual of −1.63. By adding these two figures together (or by looking at the original data), you can see that the actual deforestation rate for this country is 1.3. What do all these numbers imply? Note that the regression model is predicting a much higher value (2.93) for deforestation than actually occurred (1.3) in this country. This means that this country may be doing much better at protecting its forests than the regression model implies, and, consequently, is making better efforts at forest conservation than are other countries. This kind of information may be important to policymakers in other countries, particularly as this outlier country may provide useful lessons that can be applied to them.

The ideas of a residual and a fitted value are important in developing an informal understanding of how well a regression model fits. However, we still lack a formal numerical measure of fit. At this stage, we can now derive and motivate such a measure: R^2.

Recall that variance is the measure of dispersion or variability of the data. Here we define a closely related concept, the total sum of squares or TSS:

$$\text{TSS} = \sum (Y_i - \overline{Y})^2,$$

Note that the formula for the variance of Y is $TSS/N - 1$ (see Chapter 2). Loosely speaking, the $N - 1$ term will cancel out in our final formula for R^2 and, hence, we ignore it. So think of TSS as being a measure of the variability of Y. The regression model seeks to explain the variability in Y through the explanatory variable X. It can be shown that the total variability in Y can be broken into two parts as:

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

where RSS is the regression sum of squares, a measure of the explanation provided by the regression model.[4] RSS is given by:

$$\text{RSS} = \sum (\hat{Y}_i - \overline{Y})^2.$$

Remembering that SSR is the sum of squared residuals and that a good fitting regression model will make the SSR very small, we can combine the equations above to yield a measure of fit:

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}}$$

or, equivalently,

$$R^2 = \frac{\text{RSS}}{\text{TSS}}.$$

Intuitively, the R^2 measures the proportion of the total variance of Y that can be explained by X. Note that TSS, RSS and SSR are all sums of squared numbers and, hence, are all non-negative. This implies TSS ≥ RSS and TSS ≥ SSR. Using these facts, it can be seen that $0 \leq R^2 \leq 1$.

Further intuition about this measure of fit can be obtained by noting that small values of SSR indicate that the regression model is fitting well. A regression line which fits all the data points perfectly in the XY-plot will have no errors and hence SSR = 0 and $R^2 = 1$. Looking at the formula above, you can see that values of R^2 near 1 imply a good fit and that $R^2 = 1$ implies a perfect fit. In sum, high values of R^2 imply a good fit and low values a bad fit.

An alternative source of intuition is provided by the RSS. RSS measures how much of the variation in Y the explanatory variables explain. If RSS is near TSS, then the explanatory variables account for almost all of the variability and the fit will be a good one. Looking at the previous formula you can see that the R^2 is near one in this case.

Example: Cost of production in the electric utility industry (continued from page 56)

In the regression of Y = cost of production on X = output for the 123 electric utility companies, $R^2 = 0.92$. This is a number that is quite high and close to one, indicating that the fit of the regression line is quite good. Put another way, 92% of the variation in costs across companies can be explained by the variation in output. Note that if you simply calculate the correlation between output and cost you obtain $r_{XY} = 0.96$. This correlation squared is exactly equal to R^2 (i.e. $0.96^2 = 0.92$).

This example highlights the close relationship between correlation and regression. Notice that the R^2 from the regression of Y on X is exactly equal to the square of the correlation between Y and X. Regression is really just an extension of correlation. Yet, regression also provides you with an explicit expression for the marginal effect (β), which is often important for policy analysis.

Example: The effect of advertising on sales (continued from p. 56)

The R^2 from the regression of sales on advertising expenditures using data set ADVERT.XLS is 0.09. This relatively small number indicates that variations in advertising expenditures across companies account for only a small proportion of the variation in sales. This finding is probably reasonable, in that you would expect factors other than advertising (e.g. product quality, pricing, etc.) to play a very important role in explaining the sales of a company.

Exercise 4.4

(a) Using the data in FOREST.XLS (see Exercise 4.1), run a regression of Y on X using Excel. What is the R^2?
(b) Calculate the correlation between Y and X.
(c) Discuss the relationship between your answers in (a) and (b).
(d) Redo (a) for various regressions involving the variables W, X, Y and Z in the data set. Comment on the fit of each of these regressions.

Nonlinearity in regression

So far, we have used the linear regression model and fit a straight line through XY-plots. However, this may not always be appropriate. Consider the XY-plot in Figure 4.2. It looks like the relationship between Y and X is not linear. If we were to fit a straight line through the data, it might give a misleading representation of the relationship between Y and X. In fact, we have artificially generated this data by assuming the relationship between Y and X is of the form:

$$Y_i = 6X_i^2 ,$$

such that the true relationship is quadratic. A cursory glance at the XY-plots can often indicate whether fitting a straight line is appropriate or not.

What should you do if a quadratic relationship rather than a linear relationship exists? The answer is surprisingly simple: rather than regressing Y on X, regress Y on X^2 instead.

Of course, the relationship revealed by the XY-plot may be found to be neither linear nor quadratic. It may appear that Y is related to $\ln(X)$ or $1/X$ or X^3 or any other transformation of X. However, the same general strategy holds: transform the X variable as appropriate and then run a regression of Y on the transformed variable. You can even transform Y if it seems appropriate.

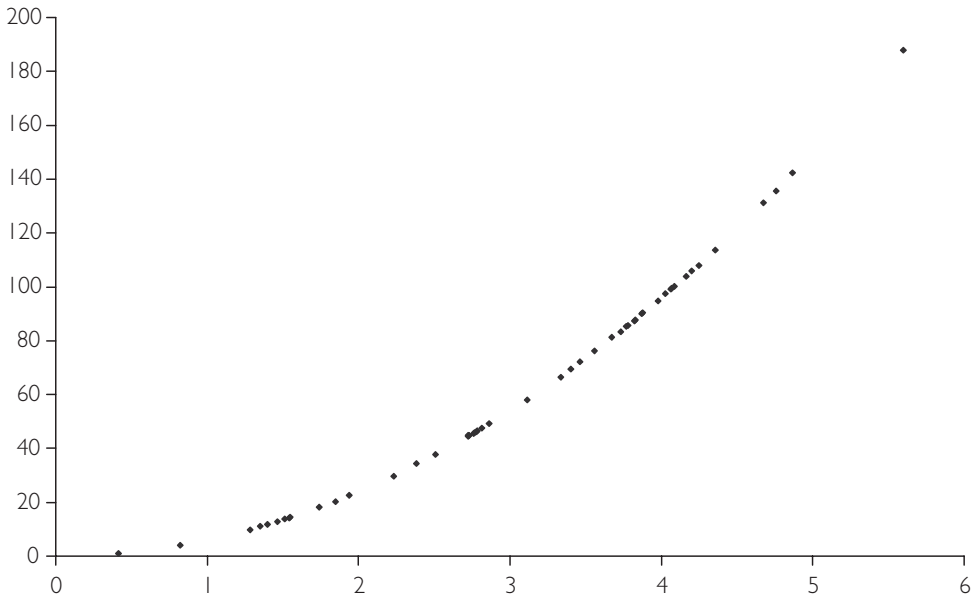

Fig. 4.2 A quadratic relationship between X and Y.

A very common transformation, of both the dependent and explanatory variables, is the logarithmic transformation. Even if you are not familiar with logarithms, they are easy to work with in any spreadsheet or econometric software package, including Excel.[5] Often economists work with natural logarithms, for which the symbol is ln. In this book, we will always use natural logarithms and simply refer to them as "logs" for short. It is common to say that: "we took the log of variable X" or that "we worked with log X". The mathematical notation is $\ln(X)$.[6]

Why is it common to use $\ln(Y)$ as the dependent variable and $\ln(X)$ as the explanatory variable? First, the expressions will often allow us to interpret results quite easily. Second, data transformed in this way often does appear to satisfy the linearity assumption of the regression model.

To fully understand the first point, we need some background in calculus, which is beyond the scope of this book. Fortunately, the intuition can be stated verbally. In the following regression:

$$\ln(Y) = \alpha + \beta \ln(X) + e,$$

β can be interpreted as an **elasticity**. Recall that, in the basic regression without logs, we said that "Y tends to change by β **units** for a one **unit** change in X". In the regression containing both logged dependent and explanatory variables, we can now say that "Y tends to change by β **percent** for a one **percent** change in X". That is, instead of having to worry about units of measurements, regression results using logged variables are always interpreted as elasticities. Logs are convenient for other reasons too. For instance, as discussed in Chapter 2, when we have time series data, the percentage change in a variable is approximately $100 \times [\ln(Y_t) - \ln(Y_{t-1})]$. This transformation will turn out to be useful in later chapters in this book.

The second justification for the log transformation is purely practical: With many data sets, if you take the logs of dependent and explanatory variables and make an *XY*-plot the resulting relationship will look linear. This is illustrated in Figures 4.3 and 4.4. Figure 4.3 is an *XY*-plot of two data series, *Y* and *X*, neither of which has been transformed in any way. Figure 4.4 is an *XY*-plot of ln(*X*) and ln(*Y*). Note that the points in the first figure do not seem to lie along a straight line. Rather the relationship is one of a steep-sloped pattern for small values of *X*, that gradually flattens out as *X* increases. This is a typical pattern for data which should be logged. Figure 4.4 shows that, once the data is logged, the *XY*-plot indicates a linear pattern. An OLS regression will fit a straight line with a high degree of accuracy in Figure 4.4. However, fitting an accurate straight line through Figure 4.3 is a very difficult (and probably not the best) thing to do.

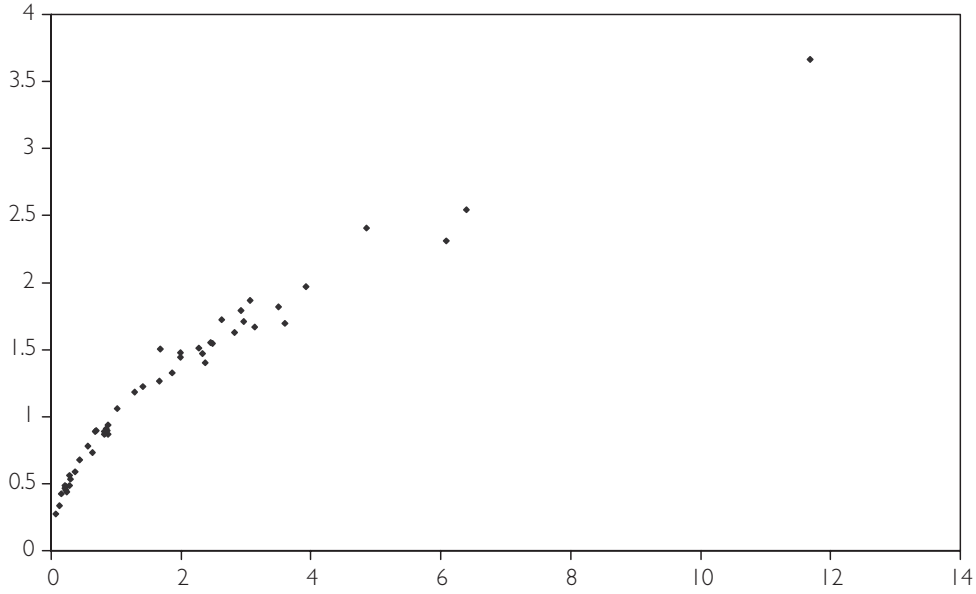

Fig. 4.3 *X* and *Y* need to be logged.

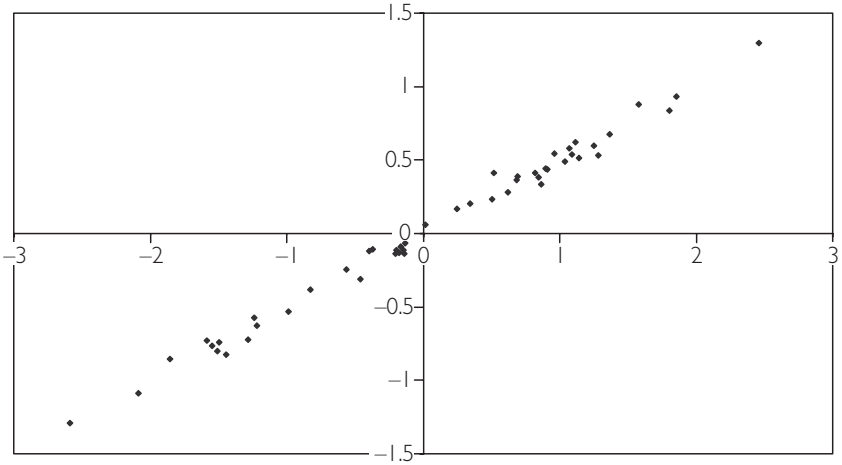

Fig. 4.4 ln(*X*) versus ln(*Y*).

On what basis should you log your data (or for that matter take any other transformation)? There is no simple rule that can be given. Examining *XY*-plots of the data transformed in various ways is often instructive. For instance, begin by looking at a plot of *X* against *Y*. This may look roughly linear. If so, just go ahead and run a regression of *Y* on *X*. If the plot does not look linear, it may exhibit some other pattern that you recognize (e.g. the quadratic form of Figure 4.2 or the logarithmic form of Figure 4.3). If so, create an *XY*-plot of suitable transformed variables (e.g. ln(*Y*) against ln(*X*)) and see if it looks linear. Such a strategy will likely work well in a simple regression containing only one explanatory variable. In Chapter 6, we will move on to cases with several explanatory variables. In these cases, the examination of *XY*-plots may be quite complicated since there are so many possible *XY*-plots that could be constructed.

Exercise 4.5

Using the data in FOREST.XLS examine different *XY*-plots involving the variables *X*, *Y*, *W* and *Z* (see Exercise 4.1 for a definition of these variables). Does there seem to be a nonlinear relationship between any pair of variables? Repeat the exercise using the data in the advertising example (ADVERT.XLS).

Exercise 4.6

Data set EX46.XLS contains two variables, labeled *Y* and *X*.

(a) Make an *XY*-plot of these two variables. Does the relationship between *Y* and *X* appear to be linear?
(b) Calculate the square root of variable *X*. Note the Excel symbol for square root is SQRT.
(c) Make an *XY*-plot of the square root of *X* against *Y*. Does this relationship appear to be linear?

Exercise 4.7

Use the data in the example related to costs of production in the electric utility industry (ELECTRIC.XLS), where *Y* = cost of production and *X* = output.

(a) Run a regression of *Y* on *X*.
(b) Take log transformations of both variables.
(c) Run a regression of ln(*Y*) on ln(*X*) and interpret your results verbally.

Chapter summary

1. Simple regression quantifies the effect of an explanatory variable, X, on a dependent variable, Y. Hence, it measures the relationship between two variables.
2. The relationship between Y and X is assumed to take the form, $Y = \alpha + \beta X$, where α is the intercept and β the slope of a straight line. This is called the regression line.
3. The regression line is the best fitting line through an XY graph.
4. No line will ever fit perfectly through all the points in an XY graph. The distance between each point and the line is called a residual.
5. The ordinary least squares (OLS) estimator is the one which minimizes the sum of squared residuals.
6. OLS provides estimates of α and β which are labeled $\hat{\alpha}$ and $\hat{\beta}$.
7. Regression coefficients should be interpreted as marginal effects (i.e. as measures of the effect on Y of a small change in X).
8. R^2 is a measure of how well the regression line fits through the XY graph.
9. OLS estimates and the R^2 are calculated in computer software packages such as Excel.
10. Regression lines do not have to be linear. To carry out nonlinear regression, merely replace Y and/or X in the regression model by a suitable nonlinear transformation (e.g. $\ln(Y)$ or X^2).

Appendix 4.1: Mathematical details

The OLS estimator defines the best fitting line through the points on an XY-plot. Mathematically, we are interested in choosing $\hat{\alpha}$ and $\hat{\beta}$ so as to minimize the sum of squared residuals. The SSR can be written as:

$$\text{SSR} = \sum_{i=1}^{N} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right)^2 .$$

Optional exercise

Take first and second derivatives with respect to $\hat{\alpha}$ and $\hat{\beta}$ of the above expression for SSR. Use these to find values of $\hat{\alpha}$ and $\hat{\beta}$ that minimize SSR. Verify that the solution you have found does indeed minimize (rather than maximize) SSR.

If you have done the previous exercise correctly, you should have obtained the following:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

and

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X},$$

where $\overline{Y}$ and $\overline{X}$ are the means of Y and X (see Chapter 2). These are the OLS estimators for α and β. Note that there are several equivalent ways of writing the formula for $\hat{\beta}$. If you consult other textbooks you will find alternative expressions for the OLS estimator.

These equations can be used to demonstrate the consequences of taking **deviations from means**. By way of explanation, note that we have assumed above that the dependent and explanatory variables, X and Y, are based on the raw data. However, in some cases researchers do not work with just X and Y, but rather with X and Y minus their respective means:

$$y_i = Y_i - \overline{Y}$$

and

$$x_i = X_i - \overline{X}.$$

Consider using OLS to estimate the regression:

$$y = a + bX + e,$$

where we have used the symbols a and b to distinguish them from the coefficients α and β in the regression involving Y and X.

It turns out that the relationship between OLS estimates from the original regression and the one where deviations from means have been taken is a simple one. The OLS estimate of b is always exactly the same as $\hat{\beta}$ and the OLS estimate of a is always zero. In other words, taking deviations from means simplifies the regression model by getting rid of the intercept (i.e. there is no point in including an intercept since its coefficient is always zero). This simplification does not have any effect on the slope coefficient in the regression model. It is unchanged by taking deviations from means and still has the same interpretation as a marginal effect.

It is not too hard to prove the statements in the previous paragraph and, if you are mathematically inclined, you might be interested in doing so. As a hint, note that the means of y and x are zero.

In Chapter 6, we will consider the case where there are several explanatory variables. In this case, if you take deviations from means of the dependent and all of the explanatory variables, you obtain the same result. That is, the intercept disappears from the regression, but all other coefficient estimates are unaffected.

Endnotes

1. Note that, at many places, we will omit multiplication signs for simplicity. For instance, instead of saying $Y = \alpha + \beta \times X$ we will just say $Y = \alpha + \beta X$.
2. Some statistics books draw a dividing line between correlation and regression. They argue that correlation should only be interpreted as a measure of the *association* between two variables, not the *causality*. In contrast, regression should be based on causality in the manner of such statements as: "Economic theory tells us that X causes Y". Of course, this division simplifies the interpretation of empirical results. After all, it is conceptually easier to think of your dependent variable – isolated on one side of the regression equation – as being "caused" by the explanatory variables on the other. However, it can be argued that this division is in actuality an artificial one. As we saw in Chapter 3, there are many cases for which correlation does indeed reflect causality. Furthermore, in future chapters we will encounter some cases in which the regressions are based on causality, some in which they are not, and others about which we are unsure. The general message here is that you need to exercise care when interpreting regression results as reflecting causality. The same holds for correlation results. Common sense and economic theory will help you in your interpretation of either.
3. If you cannot see this construct your own numerical example. That is, choose any values for α, β and X, then use the equation $Y = \alpha + \beta X$ to calculate Y (call this "original Y"). Now increase X by one, leaving α and β unchanged and calculate a new Y. No matter what values you originally chose for α, β and X, you will find new Y minus original Y is precisely β. In other words, β is a measure of the effect on Y of increasing X by one unit.
4. Excel prints out TSS, RSS and SSR in a table labeled ANOVA. The column labeled "SS" contains these three sums of squares. At this stage, you probably do not know what ANOVA means, but we will discuss it briefly in Chapter 7, "Regression with dummy variables".
5. You can calculate the natural logarithm of any number in Excel by using the formula bar. For instance, if you want to calculate the log of the number in cell D4 move to the formula bar and type "= ln(D4)" then press enter.
6. One thing to note about logs is that they are only defined for positive numbers. So if your data contains zeros or negative numbers, you cannot take logs (i.e. the software will display an error message).