CHAPTER 5

Statistical Aspects of Regression

Statistics is a field of study based on mathematics and probability theory. However, since this book assumes you have no knowledge of these topics, a complete understanding of statistical issues in the regression model will have to await further study.[1] What we will do instead in this chapter is to:

- discuss what statistical methods in the regression model are designed to do;
- show how to carry out a regression analysis using these statistical methods and interpret the results obtained;
- provide some graphical intuition in order to gain a little insight into where statistical results come from and why these results are interpreted in the manner that they are.

We begin by stressing a distinction which arose in Chapter 4 between the regression coefficients, α and β, and the ordinary least squares (OLS) estimates of the regression coefficients, $\hat{\alpha}$ and $\hat{\beta}$. Remember that we began Chapter 4 with a regression model of the form:

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

for $i = 1, \ldots, N$ observations. As noted previously, α and β measure the relationship between Y and X. We pointed out that we do not know what this relationship is, i.e., we do not know what precisely α and β are. We derived OLS estimates which we then labeled $\hat{\alpha}$ and $\hat{\beta}$. We emphasized that α and β are the unknown true coefficients

while $\hat{\alpha}$ and $\hat{\beta}$ are merely estimates (and almost certainly not precisely the same as α and β).

These considerations lead us to ask whether we can gauge how accurate these estimates are. Fortunately we can, using statistical techniques. In particular, these techniques enable us to provide *confidence intervals* for and carry out *hypothesis tests* on our regression coefficients.

To provide some jargon, we say that OLS provides *point estimates* for β (e.g. $\hat{\beta} = 0.000842$ is the point estimate of β in the regression of deforestation on population density in Chapter 4). You can think of a point estimate as your best guess at what β is. Confidence intervals provide *interval estimates*, allowing us to make statements that reflect the uncertainty we may have about the true value of β, for example, that "we are confident that β is greater than 0.0006 and less than 0.0010". We can obtain different confidence intervals corresponding to different levels of confidence. For instance, in the case of a 95% confidence interval we can say that "we are 95% confident that β lies in the interval"; in the case of a 90% confidence interval we can say that "we are 90% confident that β lies in the interval"; and so on. The degree of confidence we have in a chosen interval (e.g. 95%) is referred to as the *confidence level*.

The other major activity of the empirical researcher is *hypothesis testing*. An example of a hypothesis that a researcher may want to test is $\beta = 0$. If the latter hypothesis is true, then this means that the explanatory variable has no explanatory power. Hypothesis testing procedures allow us to carry out such tests.

Both confidence intervals and hypothesis testing procedures are explained further in the rest of this chapter. For expository purposes, we focus on β, since it is usually more important than α in economic problems. However, all the procedures we discuss for β apply equally to α.

Which Factors Affect the Accuracy of the Estimate $\hat{\beta}$?

We have artificially simulated four data sets for X and Y from regression models with $\alpha = 0$ and $\beta = 1$. XY-plots for these four different data sets are presented in Figures 5.1, 5.2, 5.3 and 5.4.

All of these data sets have the same true coefficient values of $\alpha = 0$ and $\beta = 1$ and we hope to obtain $\hat{\alpha}$ and $\hat{\beta}$ values that are roughly equal to 0 and 1, respectively, when we run a regression using any of these four data sets. However, if you imagine trying to fit a straight line (as does OLS) through these XY-plots, you would not expect all four of these lines to be equally accurate.

How confident would you feel about the accuracy of the straight line that you have just fitted? It is intuitively straightforward to see that the line fitted for

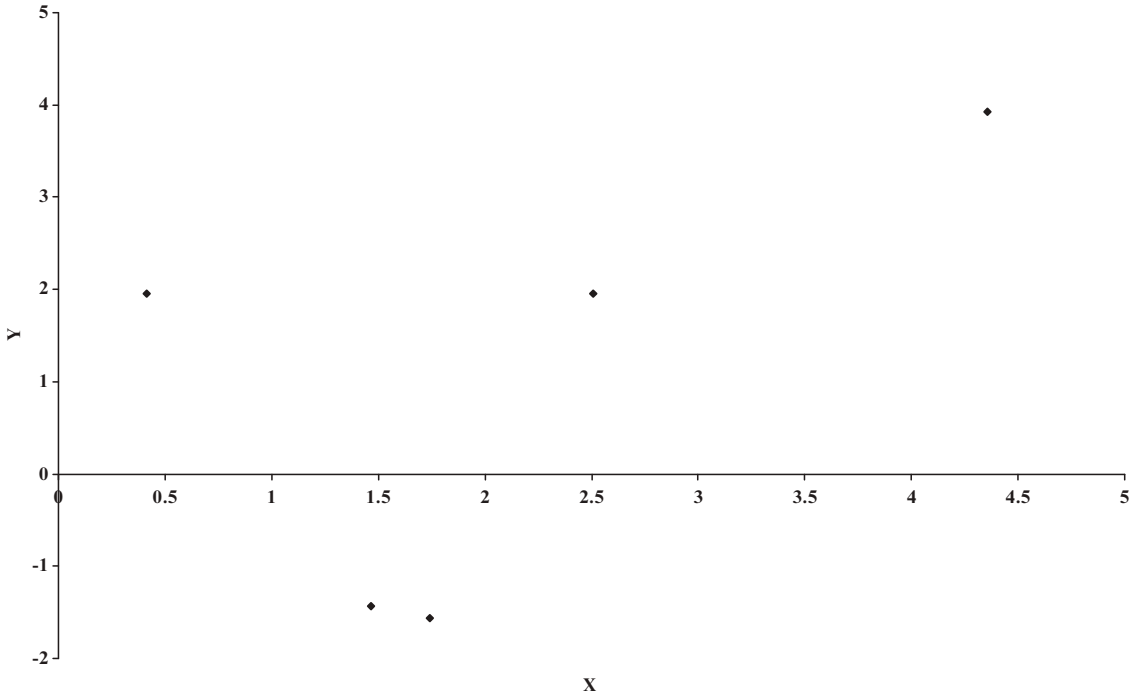

Figure 5.1 *XY*-plot with a very small sample size.

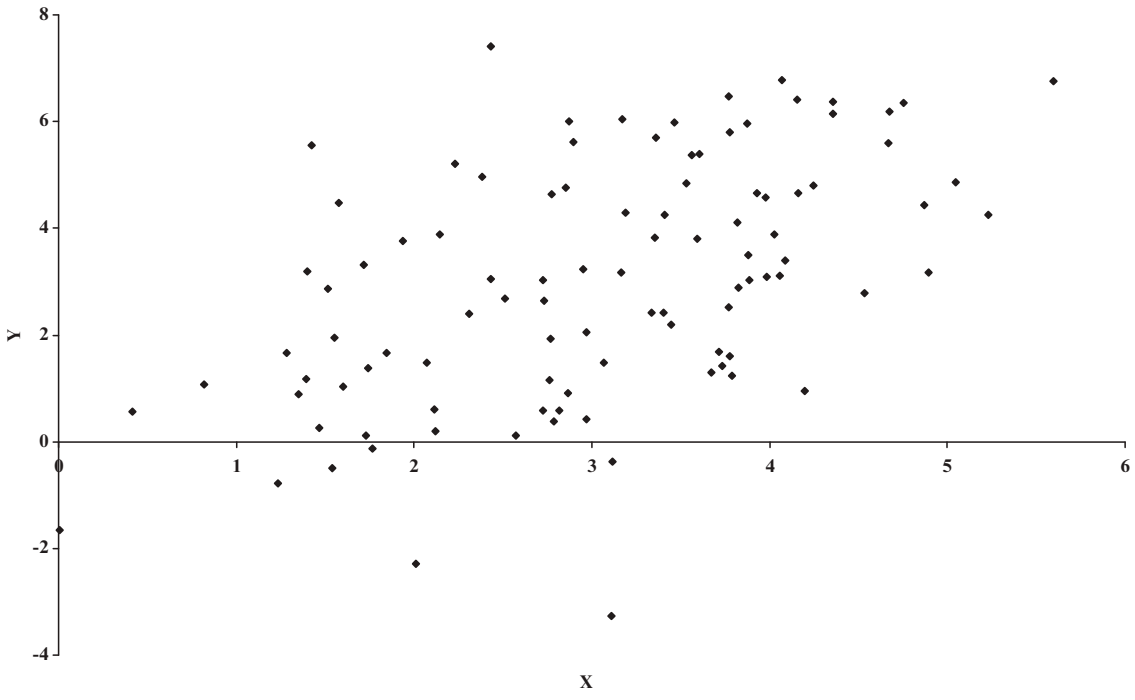

Figure 5.2 *XY*-plot with a large sample size and a large error variance.

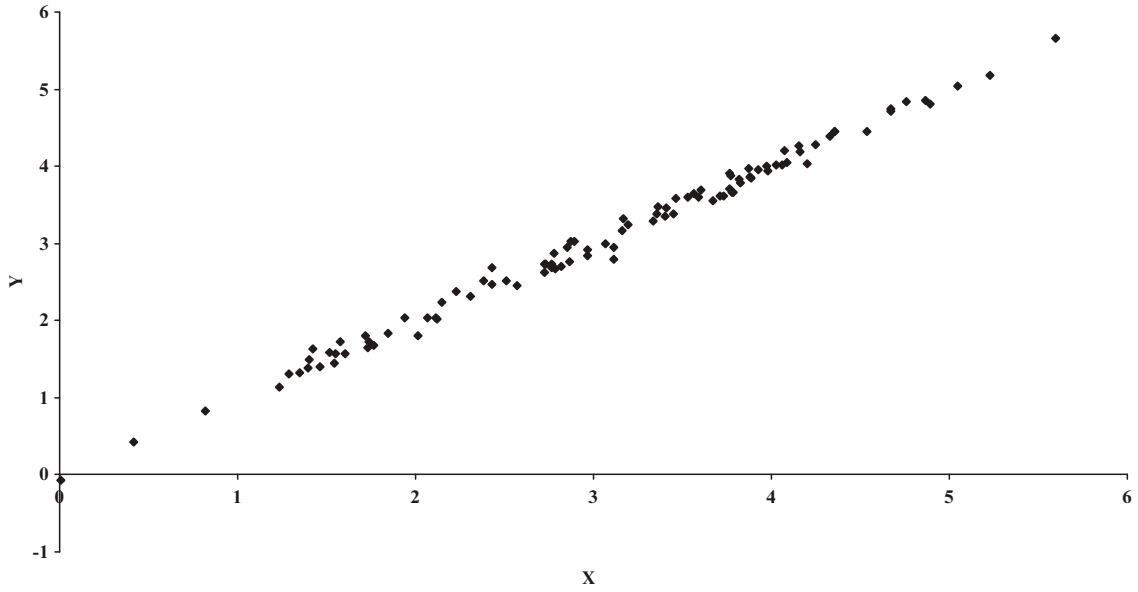

Figure 5.3 *XY*-plot with a large sample size and a small error variance.

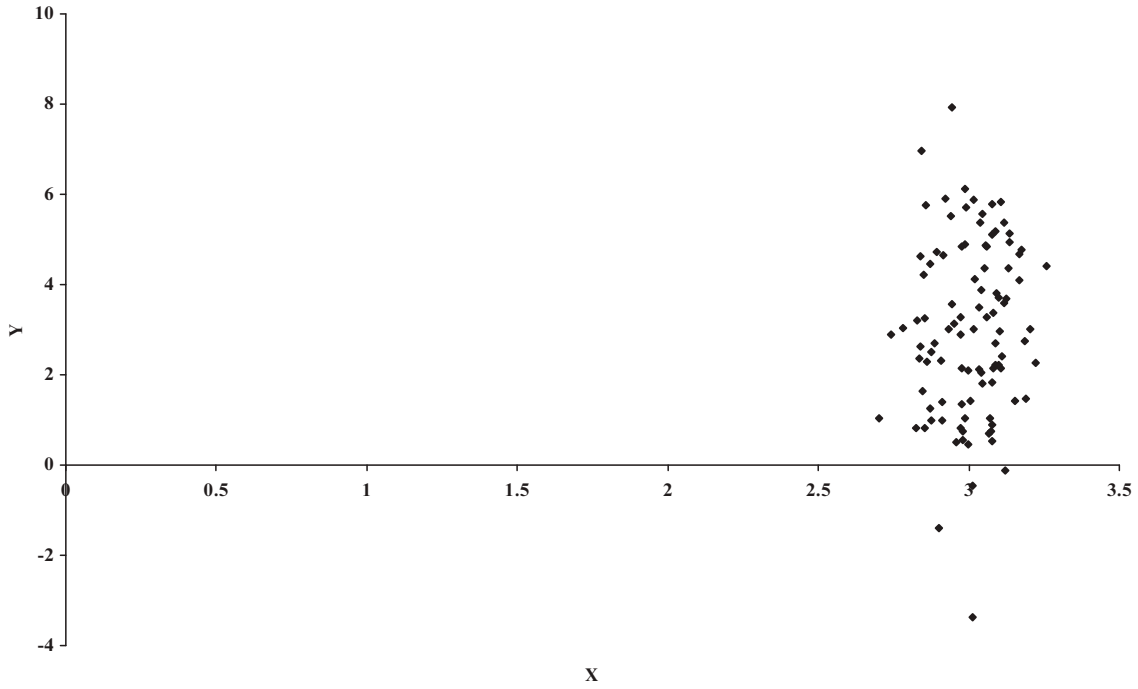

Figure 5.4 *XY*-plot with a limited range of *X* values.

Figure 5.3 would be the most accurate. That is, the straight-line relationship between X and Y "leaps out" in Figure 5.3. Even if you were to draw a best-fitting line by hand through this XY-plot you would find that the intercept (α) was very close to zero and the slope (β) close to 1. In contrast, you would probably be much less confident about the accuracy of a best-fitting straight line that you drew for Figures 5.1, 5.2 and 5.4.

These figures illustrate three main factors that affect the accuracy of OLS estimates and the uncertainty that surrounds our knowledge of what the true value of β really is:

- Having more data points improves the accuracy of estimation. This can be seen by comparing Figure 5.1 ($N = 5$) with Figure 5.3 ($N = 100$).
- Having smaller errors improves accuracy of estimation. Equivalently, if the SSR is small or the variance of the errors is small, the accuracy of the estimation will be improved. This can be seen by comparing Figure 5.2 (large variance of errors) with Figure 5.3 (small variance of errors).[2]
- Having a larger spread of values (i.e. a larger variance) of the explanatory variable (X) improves accuracy of estimation. This can be seen by comparing Figure 5.3 (values of the explanatory variable spread all the way from 0 to 6) to Figure 5.4 (values of the explanatory variable are clustered around 3).

The influence of these three factors is intuitively reasonable. With regards to the first two factors, it is plausible that having either more data or smaller errors should increase the accuracy of estimation. The third factor is perhaps less intuitive, but a simple example should help you to understand it.

Suppose you are interested in investigating the influence of education levels (X is the number of years of schooling) on the income people receive (Y). To understand the nature of this relationship, you will want to go out and collect data on all types of people (e.g. people with no qualifications, people with secondary school education, people with some post-secondary vocational training, people with a university degree, people with a PhD and so on). In other words, you will want to survey a broad spectrum of the population in order to capture as many of these different education levels as possible. In statistical jargon, this means that you will want X to have a high variance. If you do not follow this strategy – for example, were you to survey only those people possessing a PhD – you would get a very unreliable picture of the effect of education on income. If you only surveyed PhDs, you would not know whether the relationship between education and income was positive. For instance, without collecting data on people who quit school at age 16 you would not know for sure that they are making less income than the PhDs.

In summary, having a large spread of values (i.e. a larger variance) for the explanatory variable, X, is a desirable property in an analysis, whereas having a large spread of values (i.e. a larger variance) for the error, e, is not.

Calculating a Confidence Interval for β

The above three factors are reflected in a commonly used interval estimate for β: the confidence interval. This interval reflects the uncertainty surrounding the accuracy of the estimate $\hat{\beta}$. If the confidence interval is small, it indicates accuracy. Conversely, a large confidence interval indicates great uncertainty over the true value of β. In many cases, researchers choose to present the confidence interval in addition to the OLS point estimate.

The mathematical formula for the confidence interval for β is:[3]

$$[\hat{\beta} - t_b s_b, \hat{\beta} + t_b s_b].$$

An equivalent way of expressing the equation above is to say that there is a high level of confidence that the true value of β obeys the following inequality:

$$\hat{\beta} - t_b s_b \leq \beta \leq \hat{\beta} + t_b s_b.$$

The equations above use three numbers that must be calculated, $\hat{\beta}$, t_b and s_b. The first of these, $\hat{\beta}$, is the OLS estimate which we have already discussed in detail; the latter two you may not have seen before. The confidence interval will be calculated automatically by your statistical software package or spreadsheet. Thus, you can calculate confidence intervals without knowing either the above formula or the precise definitions of t_b and s_b. At the most basic level, you can just think of $\hat{\beta}$, t_b and s_b as three numbers calculated by the computer. However, it is worthwhile to have at least some intuition about where the confidence interval comes from as this will aid in your understanding of results. Below, we discuss each of the three numbers required to calculate a confidence interval, relating them to the issues raised in the material above on the factors affecting the accuracy of estimation of $\hat{\beta}$.

First, $\hat{\beta}$ is always included in the confidence interval (in fact, it will be right in the middle of it).

Secondly, s_b is the standard deviation of $\hat{\beta}$. Somewhat confusingly, s_b is often referred to as the *standard error* as opposed to the standard deviation. In Chapter 2, we introduced the standard deviation as a measure of dispersion (i.e. spread or variability) of a variable. For instance, Figure 2.2 plots a histogram for the variable GDP per capita using the cross-country data set GDPPC.XLS. In Chapter 2, we argued that the standard deviation of GDP per capita was a measure of how much GDP per capita varied across countries. We can treat $\hat{\beta}$ as a variable in the same way as GDP per capita is a variable. In other words, we can calculate its standard deviation and use it as a measure of our uncertainty about the accuracy of the estimate. Large values of s_b imply a large degree of uncertainty – $\hat{\beta}$ may be a very inaccurate estimate of β. In contrast, small values of s_b imply a small degree of uncertainty – $\hat{\beta}$ will be an accurate estimate of β.

In other chapters, we have put mathematical formulae in appendices. However, to properly draw out the connections between the formula for the confidence interval and the graphical intuition provided in Figures 5.1–5.4, a small amount of mathematics is required. We present (but do not derive) the following formula for the standard deviation of $\hat{\beta}$:

$$s_b = \sqrt{\frac{SSR}{(N-2)\sum (X_i - \bar{X})^2}}.$$

This expression, which measures the variability or uncertainty in $\hat{\beta}$, reflects all of the issues raised in the context of our discussion of Figures 5.1, 5.2, 5.3 and 5.4.

Looking at the formula for the confidence interval, we can see that the larger s_b is, the wider the confidence interval is. If we combine this consideration with a careful analysis of the components of the formula for s_b, we can say that s_b and, hence, the width of the confidence interval:

- vary directly with SSR (i.e. more variable errors or residuals imply less accurate estimation);
- vary inversely with N (i.e. more data points imply more accurate estimation);
- vary inversely with $\sum (X_i - \bar{X})^2$ (i.e. more variability in X implies more accurate estimation).[4]

We stress that these three factors (i.e. N, SSR and the standard deviation of X), which affect the width of the confidence interval, are the same as those discussed above as affecting the accuracy of the OLS estimate $\hat{\beta}$.

The third number in the formula for the confidence interval is t_b. It is hard to provide much intuition about this number without some knowledge of statistics.[5] Some informal intuition for what it means, however, can be obtained from Example 5.1.

Example 5.1: Election polls

You may have encountered point estimates and something akin to a confidence interval in political polls, which are regularly taken in the weeks and months before an election. These are usually carried out by staffers telephoning a few hundred potential voters and asking them which party they intend to support on election day. Suppose Party A is running in the election. The newspaper reports that 43% of those surveyed will support Party A. This is the newspaper's point estimate of what voters will do on election day. Of course, in reality the actual result on election day will rarely, if ever, be exactly that indicated by the

pre-election poll. This discrepancy illustrates a point we stressed earlier in this chapter in the context of the regression model: a point estimate (e.g. $\hat{\beta}$) will rarely, if ever, be identical to the true value (e.g. β).

Newspapers typically recognize that their surveys will not be precisely accurate and often add statements to their coverage such as: "This result is accurate to within +/−2 percentage points." Although they do not explicitly say it, they are getting this result from a confidence interval (usually a 95% confidence interval).[6] An equivalent statement would be: "We are 95% confident that Party A will receive between 41% and 45% of the vote on election day."

This example provides some additional intuition about what confidence intervals are. If you understand this example, you can also see that different confidence levels imply different confidence intervals. As a trivial example, consider the 100% confidence level. We can be certain that Party A is going to receive between 0% and 100% of the vote on election day. A 100% confidence interval for Party A's percentage of the vote would thus be [0,100].

Now consider the other extreme: how confident can we be that Party A is going to receive almost precisely 43% of the vote? Probably not very confident for, as noted, in reality we rarely find that opinion polls and election day results match identically. For this reason, a confidence interval right around 43% (e.g. [42.9,43.1]) will have a very low confidence level (perhaps 10%).

Note that, the more confident you wish to be about your interval, the wider it becomes. For instance, 99% confidence intervals will always be wider than 95% confidence intervals. The number t_b controls the confidence level. If the level of confidence is high (e.g. 99%) t_b will be large, while if the level of confidence is low (e.g. 50%) t_b will be small.

To return to the general statistical theory of regression, we should stress (without explanation beyond that given in Example 5.1) the following:

- t_b decreases with N (i.e. the more data points you have the smaller the confidence interval will be).
- t_b increases with the level of confidence you choose.

Researchers usually present 95% confidence intervals, although other intervals are possible (e.g. 99% or 90% confidence intervals are sometimes presented). A useful (but formally incorrect) intuition for 95% confidence intervals is conveyed by the following statement: "There is a 95% probability that the true value of β lies in the 95% confidence interval." A correct (but somewhat awkward) interpretation of this statement is: "If you repeatedly used (in different data sets) the above formula for calculating confi-

dence intervals, 95% of the confidence intervals constructed would contain the true value for β." Similar statements can be made for 99% or 90% confidence intervals, simply by replacing "95%" with the desired confidence level.

The preceding material is intended to provide some intuition and motivation for the statistical theory underlying confidence intervals. Even if you do not fully understand this material, confidence intervals can be calculated quite easily in most spreadsheets and statistical software packages.

Example 5.2: Confidence intervals for the data sets in Figures 5.1–5.4

Figures 5.1 through 5.4 contained four different data sets, all of which were artificially generated on the computer with $\alpha = 0$ and $\beta = 1$. Thus, unlike with real data, in this case we know the true values of α and β. Remember that the data set used in Figure 5.3 has some very desirable properties, i.e. large sample size, spread-out values for the explanatory variables, and small errors. These properties are missing to varying degrees in the other three data sets. Table 5.1 contains OLS point estimates, $\hat{\beta}$, and 90%, 95% and 99% confidence intervals for these four data sets.

Table 5.1 OLS point estimates and confidence intervals for four data sets.

Data Set	$\hat{\beta}$	90% confidence interval	95% confidence interval	99% confidence interval
Figure 5.1	0.91	[–0.92, 2.75]	[–1.57, 3.39]	[–3.64, 5.47]
Figure 5.2	1.04	[0.75, 1.32]	[0.70, 1.38]	[0.59, 1.49]
Figure 5.3	1.00	[0.99, 1.01]	[0.99, 1.02]	[0.98, 1.03]
Figure 5.4	1.52	[–1.33, 4.36]	[–1.88, 4.91]	[–2.98, 6.02]

The following points are worth emphasizing:

- Reading across any row, we can see that as the confidence level gets higher the confidence interval gets wider. The widest interval is the 99% confidence interval for the data set in Figure 5.4. In this case, if you want to be 99% confident, you have to say β could be anywhere between –2.98 and 6.02.
- The data set in Figure 5.3 – the one with the most desirable properties of all the data sets – yields an OLS estimate of 1.00 which is equal to the true value to two decimal places (more precisely, $\hat{\beta} = 1.002577$ for this data set).

- The data set in Figure 5.3 yields confidence intervals which are much narrower than those for Figures 5.1, 5.2 and 5.4. This makes sense since we would expect the OLS estimate using the data set in Figure 5.3 to be more accurate than the other data sets.
- The data sets in Figures 5.1, 5.2 and 5.4 yield a variety of results. Figure 5.2 contains a data set of the sort usually found in a well-designed empirical project (rarely does one get a data set as good as Figure 5.3). This data set has mostly desirable properties, but the errors are moderately large, reflecting the measurement error and imperfections in the underlying economic theory which so often occur in practice. For this representative data set, $\hat{\beta} = 1.04$ which is not too far off the true value of $\beta = 1$. With respect to this data set, we can make statements of the form: "the value of β lies in the interval [0.70,1.38] with a 95% confidence level" or "we are 99% confident that β lies between 0.59 and 1.49".

Exercise 5.1

The data sets used to calculate Figures 5.1, 5.2, 5.3 and 5.4 are in FIG51.XLS, FIG52.XLS, FIG53.XLS and FIG54.XLS.

(a) Calculate the OLS estimates $\hat{\alpha}$ and $\hat{\beta}$ for these four data sets. How close are they to 0 and 1 (the values we used to artificially simulate the data)?
(b) Calculate confidence intervals for α for the four data sets. Examine how the width of the confidence interval relates to N and the variability of the errors.
(c) Calculate 99% and 90% confidence intervals for the data sets. How do these differ from the 95% confidence intervals in part (b)?

Example 5.3a: The regression of deforestation on population density

Let us go back to our deforestation (Y) and population density (X) data set (FOREST.XLS). We saw in Chapter 4 that $\hat{\beta} = 0.000842$. In other words, the marginal effect of population density on deforestation was 0.000842. A 95% confidence interval for this effect is [0.00061,0.001075], indicating (with a great deal of certainty) that the marginal effect of population on deforestation is greater than 0.00061 and less than 0.001075.

Example 5.4a: The regression of lot size on house price

In Chapter 4, we investigated the effect of lot size (X) on the sale price (Y) of a house, using data on 546 houses sold in Windsor, Canada (see data set HPRICE.XLS). Running a regression of Y on X we obtain the following estimated relationship:

$$Y = 34\,136 + 6.59X$$

or, equivalently, $\hat{\alpha} = 34\,136$ and $\hat{\beta} = 6.59$. We can say that the OLS estimate of the marginal effect of X on Y is 6.59. This means that our best guess would be that increasing lot size by an extra square foot is associated with a \$6.59 increase in house price.

The 95% confidence interval for β is [5.72,7.47]. This means that, although the effect of lot size on house price is estimated at \$6.59, we are not certain that this figure is exactly correct. However, we are extremely confident (i.e. 95% confident) that the effect of lot size on house price is at least \$5.72 and at most \$7.47. This interval would be enough for a potential buyer or seller to have a good idea of the value of lot size.

Exercise 5.2

The file ADVERT.XLS contains data on annual sales (Y) and advertising expenditure (X) (both measured in millions of dollars) for 84 companies in the USA.

(a) Run a regression of Y on X and obtain 95% confidence intervals for α and β.

(b) Write a sentence explaining verbally what the 95% confidence interval for β means in terms of the possible range of values that the effect of the explanatory variable on the dependent variable may take.

Exercise 5.3

The file ELECTRIC.XLS contains data on the costs of production (Y, measured in millions of dollars) and output (X, measured in thousands of kilowatt hours) for 123 electricity utility companies in the USA. Repeat Exercise 5.2 for this data set.

Testing whether $\beta = 0$

Hypothesis testing is another exercise commonly carried out by the empirical economist. As with confidence intervals, we will not go into the statistical theory that underlies hypothesis testing. Instead, we focus on the practical details of how to carry out hypothesis tests and interpret the results. Classical hypothesis testing involves specifying a hypothesis to test. This is referred to as the *null hypothesis*, and is labeled as H_0. It is compared to an *alternative hypothesis*, labeled H_1. A common hypothesis test is whether $\beta = 0$. Formally, we say that this is a test of H_0: $\beta = 0$ against H_1: $\beta \neq 0$.

Note that, if $\beta = 0$ then X does not appear in the regression model; that is, the explanatory variable fails to provide any explanatory power whatsoever for the dependent variable. If you think of the kinds of questions of interest to economists (e.g. "Does education increase an individual's earning potential?", "Will a certain advertising strategy increase sales?", "Will a new government training scheme lower unemployment?") you will see that many are of the form "Does the explanatory variable have an effect on the dependent variable?" or "Does $\beta = 0$ in the regression of Y on X?". The purpose of the hypothesis test of $\beta = 0$ is to answer this question.

The first point worth stressing is that hypothesis testing and confidence intervals are closely related. In fact, one way of testing whether $\beta = 0$ is to look at the confidence interval for β and see whether it contains zero. If it does not then we can, to introduce some statistical jargon, "reject the hypothesis that $\beta = 0$" or conclude "X has significant explanatory power for Y" or "β is significantly different from zero" or "β is statistically significant". If the confidence interval does include zero then we change the word "reject" to "accept" and "has significant explanatory power" to "does not have significant explanatory power" and so on. This confidence interval approach to hypothesis testing is exactly equivalent to the formal approach to hypothesis testing discussed below.

Just as confidence intervals came with various levels of confidence (e.g. 95% is the usual choice), hypothesis tests come with various *levels of significance*. If you use the confidence interval approach to hypothesis testing, then the level of significance is 100% minus the confidence level. That is, if a 95% confidence interval does not include zero, then you may say "I reject the hypothesis that $\beta = 0$ at the 5% level of significance" (i.e. 100% − 95% = 5%). If you had used a 90% confidence interval (and found it did not contain zero) then you would say: "I reject the hypothesis that $\beta = 0$ at the 10% level of significance."

The alternative way of carrying out hypothesis testing is to calculate a *test statistic*. In the case of testing whether $\beta = 0$, the test statistic is known as a t-statistic (or t-ratio or t-stat). It is calculated as:

$$t = \frac{\hat{\beta}}{s_b}.$$

Large values[7] of t indicate that $\beta \neq 0$, while small values indicate that $\beta = 0$. Mathematical intuition for the preceding sentence is given as: if $\hat{\beta}$ is large relative to its standard deviation, s_b, then we can conclude that β is significantly different from zero. The question arises as to what we mean by "large" and "small". In a formal statistical sense, the test statistic is large or small relative to a "critical value" taken from statistical tables of the "Student-t distribution". A discussion of how to do this is given in Appendix 5.1. Fortunately, we do not have to trouble ourselves with statistical tables since most common computer software packages print out something called a *P-value* automatically. The P-value provides a direct measure of whether t is "large" or "small". A useful (but formally incorrect) intuition would be to interpret the P-value as measuring the probability that $\beta = 0$. If the P-value is small, $\beta = 0$ is unlikely to be true. Accordingly, the following strategy is commonly used.

- If the P-value is less than 5% (usually written as 0.05 by the computer) then t is "large" and we conclude that $\beta \neq 0$.
- If the P-value is greater than 5% then t is "small" and we conclude that $\beta = 0$.

The preceding test used the 5% level of significance. However, if we were to replace the figure 5% in the above expressions with 1% (i.e. we reject $\beta = 0$ if the P-value is less than 1%) our hypothesis test would be carried out at the 1% level of significance.

As an aside, it is worth noting that we are focusing on the test of $\beta = 0$ partly because it is an important one, but also because it is the test that is usually printed out by computer packages. You can use it without fully understanding the underlying statistics. However, in order to test other hypotheses (e.g. H_0: $\beta = 1$ or hypotheses involving many coefficients, in the multiple regression case in Chapter 6) you would need more statistical knowledge than is covered here (see Appendix 5.1 for more details). The general structure of a hypothesis test is always of the form outlined above:

1. Specify the hypothesis being tested.
2. Calculate a test statistic.
3. Compare the test statistic to a critical value.

The first of these three steps is typically easy, but the second and third can be much harder. In particular, to obtain the test statistic for more complicated hypothesis tests will typically require some extra calculations beyond merely running the regression. Obtaining the critical value will involve the use of statistical tables. Hence, if you wish to do more complicated hypothesis tests you will have to resort to a basic statistics or econometrics textbook.

As a practical summary, note that regression techniques provide the following information about β:

- $\hat{\beta}$, the OLS point estimate, or best guess, of what β is;
- the 95% confidence interval, which gives an interval in which we are 95% confident that β lies;

- the standard deviation (or standard error) of $\hat{\beta}$, s_b, which is a measure of how accurate $\hat{\beta}$ is;
- the test statistic, t, for testing $\beta = 0$;
- the P-value for testing $\beta = 0$.

These five components, ($\hat{\beta}$, the confidence interval, s_b, t and the P-value) are usually printed out in a row by statistical software packages. In practice, the most important are $\hat{\beta}$, the confidence interval and the P-value. You can usually interpret your empirical findings without explicit reference to t and s_b. Example 5.3b serves to illustrate how regression results are presented and can be interpreted.

Example 5.3b: The regression of deforestation on population density

If we regress deforestation (Y) on population density (X), the output in Table 5.2 is produced.

Table 5.2 Regression of deforestation on population density.

	Coefficient	Standard error	t-stat	P-value	Lower 95%	Upper 95%
Intercept	0.599965	0.112318	5.341646	1.15E-06	0.375837	0.824093
X variable	0.000842	0.000117	7.227937	5.5E-10	0.00061	0.001075

The row labeled "Intercept" contains results for α and the row labeled "X variable" contains results for β. We focus discussion on this latter row. The column labeled "Coefficient" presents the OLS estimate and, as we have seen before, $\hat{\beta} = 0.000842$, indicating that increasing the population density by one person per hectare is associated with an increase in deforestation rates of 0.000842%. The columns labeled "Lower 95%" and "Upper 95%" give the lower and upper bounds of the 95% confidence interval. For this data set, as discussed previously, the 95% confidence interval for β is [0.00061,0.001075]. Thus, we are 95% confident that the marginal effect of population density on deforestation is between 0.00061% and 0.001075%.

The columns labeled "Standard error" and "t-stat" indicate that $s_b = 0.000117$ and $t = 7.227937$. These numbers are not essential to carrying out a hypothesis test of $\beta = 0$ when the P-value is given. For most purposes, we can ignore these two columns.[8]

The hypothesis test of $\beta = 0$ can be done in two equivalent ways. First, we can use the 95% confidence interval for β of [0.00061,0.001075]. Since this interval

does not contain 0, we can reject the hypothesis that $\beta = 0$ at the 5% level of significance. In other words, there is strong evidence for the hypothesis that $\beta \neq 0$ and that population density has significant power in explaining deforestation. Second, we can look at the P-value which is 5.5×10^{-10} and much less than 0.05.[9] This means that we can reject the hypothesis that population density has no effect on deforestation at the 5% level of significance. In other words, we have strong evidence that population density does indeed affect deforestation rates.

Exercise 5.4

Using Table 5.2 (or running a regression yourself using data set FOREST.XLS), test the hypothesis that $\alpha = 0$.

Exercise 5.5

In addition to deforestation rate (Y), data set FOREST.XLS also contains data on W, the percentage increase in cropland (labeled "Crop ch"), and Z, the percentage change in pasture land (labeled "Pasture Ch").

(a) Run a regression of Y on W and interpret your results. Can you reject the hypothesis that expansion of cropland has an effect on deforestation rates?
(b) Run a regression of Y on Z and interpret your results. Can you reject the hypothesis that expansion of pastureland has an effect on deforestation rates?

Exercise 5.6

Use data sets FIG51.XLS, FIG52.XLS, FIG53.XLS and FIG54.XLS.

(a) Test whether $\beta = 0$ using the confidence interval approach for each of the four data sets.
(b) Test whether $\beta = 0$ using the P-value approach for each of the four data sets. Use the 5% level of significance.
(c) Redo parts (a) and (b) for α.
(d) Redo parts (a), (b) and (c) using the 1% level of significance.
(e) Are your results sensible in light of the discussion in this chapter of the factors affecting the accuracy of OLS estimates?

Example 5.4b: The regression of lot size on house price

Previously, we found a 95% confidence interval in the regression of Y, house price, on X, lot size, to be [5.27,7.47]. Since this interval does not contain zero, we can reject the hypothesis that $\beta = 0$ at the 5% level of significance. Lot size does indeed seem to have a statistically significant effect on house prices.

Alternatively, we see that the P-value is 6.77×10^{-42}, which is much less than 0.05. As before, we can reject the hypothesis that $\beta = 0$ at the 5% level of significance. Note also that, since, 6.77×10^{-42} is less than 0.01 we can also reject the hypothesis that $\beta = 0$ at the 1% level of significance. This is strong evidence indeed that lot size affects house price.

Exercise 5.7

We have used the file ADVERT.XLS before. Remember that it contains data on the sales and advertising expenditure of 84 companies. Set up and run a regression using this data and discuss your results verbally as you would in a report. Include a discussion of the marginal effect of advertising on sales and a discussion of whether this marginal effect is statistically significant.

Hypothesis Testing Involving R^2: The F-Statistic

Most computer packages which include regression also print out results for the test of the hypothesis H_0: $R^2 = 0$. The definition and interpretation of R^2 was given in Chapter 4. Recall that R^2 is a measure of how well the regression line fits the data or, equivalently, of the proportion of the variability in Y that can be explained by X. If $R^2 = 0$ then X does not have any explanatory power for Y. The test of the hypothesis $R^2 = 0$ can therefore be interpreted as a test of whether the regression explains anything at all. For the case of simple regression, this test is equivalent to a test of $\beta = 0$.

In Chapter 6, we discuss the case of multiple regression (where there are many explanatory variables), in which case this test will be different. To preview our discussion of the next chapter, note that the test of $R^2 = 0$ will be used as a test of whether all of the explanatory variables jointly have any explanatory power for the dependent variable. In contrast, the t-statistic test of $\beta = 0$ is used to investigate whether a single explanatory variable has explanatory power.

The strategy and intuition involved in testing $R^2 = 0$ proceed along the same lines as above. That is, the computer software calculates a test statistic which you must then

compare to a critical value. Alternatively, a P-value can be calculated which directly gives a measure of the plausibility of the null hypothesis $R^2 = 0$ against the alternative hypothesis, $R^2 \neq 0$. Most statistical software packages automatically calculate the P-value and, if so, you don't need to know the precise form of the test statistic or how to use statistical tables to obtain a critical value. For completeness, though, we present the test statistic, the *F*-statistic,[10] which is calculated as:

$$F = \frac{(N-2)R^2}{(1-R^2)}.$$

As before, large values of the test statistic indicate $R^2 \neq 0$ while small values indicate $R^2 = 0$. As for the test of $\beta = 0$, we use the P-value to decide what is "large" and what is "small" (i.e. whether R^2 is significantly different from zero or not). The test is performed according to the following strategy:

- If the P-value for the *F*-statistic is less than 5% (i.e. 0.05), we conclude $R^2 \neq 0$.
- If the P-value for the *F*-statistic is greater than 5% (i.e. 0.05), we conclude $R^2 = 0$.

This strategy provides a statistical test with a 5% level of significance. To carry out a test at the 1% level of significance, merely replace 5% (0.05) by 1% (0.01) in the preceding sentences. Other levels of significance (e.g. 10%) can be calculated in an analogous manner.

Example 5.3c: The regression of deforestation on population density

In the case of the deforestation and population density data set, $F = 52.24308$. Is this "large"?

If you said "yes", you are right, since the P-value for the *F*-statistic is 5.5×10^{-10}, which is less than 0.05. We can conclude in light of this finding that population density does have explanatory power for *Y*. Formally, we can say that "R^2 is significantly different from zero at the 5% level" or that "*X* has statistically significant explanatory power for *Y*" or that "the regression is significant". Note that the P-value for the *F*-statistic is equal to the P-value in the test of $\beta = 0$, stressing the equivalence of these two tests in the case of simple regression.

Exercise 5.8

Use data sets FIG51.XLS, FIG52.XLS, FIG53.XLS and FIG54.XLS. Test whether $R^2 = 0$ for each of the four data sets. Compare your results with those of Exercise 5.6.

Example 5.5: Cost of production in the electricity utility industry

We used the file ELECTRIC.XLS in Chapter 4. Recall that it contains data on the costs of production (Y) and output (X) in 123 electricity utility companies. If we run the regression of Y on X, we obtain the results shown in Table 5.3. Furthermore, $R^2 = 0.916218$. The P-value for testing $R^2 = 0$ is 5.36E-67.

Table 5.3 Regression of costs of production on output.

	Coefficient	Standard error	t-stat	P-value	Lower 95%	Upper 95%
Intercept	2.186583	1.879484	1.163395	0.246958	−1.534354	5.90752
X variable	0.004789	0.000132	36.37623	5.36E-67	0.004528	0.005049

It is worthwhile, by way of summary of the material in Chapters 4 and 5, to illustrate how the results presented in Table 5.3 might be written up in a formal report. A typical report would include presentation of the statistical material in a table followed by a verbal summary discussing the economic intuition behind the analysis and the statistical findings in the light of this intuition. The report might go as follows:

The above table presents results from an OLS regression using the electricity industry data set. Since we are interested in investigating how different output choices by firms influence their costs of production, we select the cost of production as our dependent variable and output as the explanatory variable. The table reveals that the estimated coefficient on output is 0.004789 and suggests that electricity utility firms with higher levels of output tend to have higher costs of production. In particular, increasing output by 1000 kWh tends to increase costs by $4789.

It can be observed that the marginal effect of output on costs is strongly statistically significant, since the P-value is very small (much smaller, say, than 1%). An examination of the 95% confidence interval shows that we can be quite confident that increasing output by 1000 kWh is associated with an increase in costs of at least $4528 and at most $5049. An examination of R^2 reinforces the view that output provides a large part of the explanation for why costs vary across utilities. In particular, 92% of the variability in costs of production across firms can be explained by different output levels. The P-value for the F-statistic is much smaller than 1%, indicating significance of R^2 at the 1% level.

Chapter Summary

1. The accuracy of OLS estimates depends on the number of data points, the variability of the explanatory variable and the variability of the errors.
2. The confidence interval provides an interval estimate of β (i.e. an interval in which you can be confident that β lies). It is calculated in most computer software packages.
3. The width of the confidence interval depends on the same factors as affect the accuracy of OLS estimates. In addition, the width of the confidence interval depends on the confidence level (i.e. the degree of confidence you want to have in your interval estimate).
4. A hypothesis test of whether $\beta = 0$ can be used to find out whether the explanatory variable belongs in the regression. The P-value, which is calculated automatically in most spreadsheet or statistical computer packages, is a measure of how plausible the hypothesis is.
5. If the P-value for the hypothesis test of whether $\beta = 0$ is less than 0.05 then you can reject the hypothesis at the 5% level of significance. Hence, you can conclude that X does belong in the regression.
6. If the P-value for the hypothesis test of whether $\beta = 0$ is greater than 0.05 then you cannot reject the hypothesis at the 5% level of significance. Hence, you cannot conclude that X belongs in the regression.
7. A hypothesis test of whether $R^2 = 0$ can be used to investigate whether the regression helps explain the dependent variable. A P-value for this test is calculated automatically in most spreadsheet and statistical computer packages and can be used in a similar manner to that outlined in points 5 and 6.

Appendix 5.1: Using Statistical Tables to Test Whether $\beta = 0$

The P-value is all that you need to know in order to test the hypothesis that $\beta = 0$. Most computer software packages (e.g. Excel, Stata, Gretl or E-views) automatically provide P-values. However, if you do not have such a computer package or are reading a paper which presents the *t*-statistic, not the P-value, then it is useful to know how to carry out hypothesis testing using statistical tables. Virtually any statistics or econometrics textbook will describe the method in detail and will also provide the necessary

statistical table for you to do so. Here we offer only a brief discussion along with a rough rule of thumb which is applicable to the case when the sample size, N, is large.

Remember that hypothesis testing involves the comparison of a test statistic to a number called a "critical value". If the test statistic is larger (in absolute value) than the critical value, the hypothesis is rejected. Here, the test statistic is the t-stat given in the body of the chapter. This must be compared to a critical value taken from the Student-t statistical table. It turns out that this critical value is precisely what we have called t_b in our discussion of confidence intervals. If N is large and you are using the 5% level of significance, then $t_b = 1.96$. This suggests the following rule of thumb:

If the t-statistic is greater than 1.96 in absolute value (i.e. $|t| > 1.96$), then reject the hypothesis that $\beta = 0$ at the 5% level of significance. If the t-statistic is less than 1.96 in absolute value, then accept the hypothesis that $\beta = 0$ at the 5% level of significance.

If the hypothesis that $\beta = 0$ is rejected, then we say that "X is significant" or that "X provides statistically significant explanatory power for Y".

This rule of thumb is likely to be quite accurate if sample size is large. Formally, the critical value equals 1.96 if the sample size is infinity. However, even moderately large sample sizes will yield similar critical values. For instance, if $N = 120$, the critical value is 1.98. If $N = 40$, it is 2.02. Even the quite small sample size of $N = 20$ yields a critical value of 2.09 which is not that different from 1.96. However, you should be careful when using this rule of thumb if N is very small or the t-statistic is very close to 2.00. If you look back at the examples included in the body of this chapter you can see that the strategy outlined here works quite well. For instance, in Example 5.5, the t-statistic for testing whether $\beta = 0$ was 36.4, which is much larger than 1.96. Hence we concluded that output is a statistically significant explanatory variable for cost of production. In this example (and all others), both the P-value and confidence interval approaches lead to the same conclusion as the approximate strategy described in this appendix.

The previous discussion related to the 5% level of significance. The large sample critical value for the 10% level of significance is 1.65. For the 1% level of significance, it is 2.58.

By far the most common hypothesis to test for is H_0: $\beta = 0$. Using the techniques outlined in this appendix we can generalize this hypothesis slightly to that of H_0: $\beta = c$, where c is some number that may not be zero (e.g. $c = 1$). In this case, the test statistic changes slightly, but the critical value is exactly the same as for the test of $\beta = 0$. In particular, the test statistic becomes:

$$t = \frac{\hat{\beta} - c}{s_b}.$$

This will not be produced automatically by a computer package, but it can be calculated quite easily in a spreadsheet or statistical software package. That is, $\hat{\beta}$ and s_b are calculated by the computer and you have to provide c, depending on the hypothesis that you are interested in testing. These three numbers can be combined using the equation above to give you a value for your test statistic. If this value is greater than 1.96 in absolute value, you will conclude that $\beta \neq c$ at the 5% level of significance. The caveats about using this rule of thumb if your sample size is very small apply here.

Endnotes

1. As mentioned previously, a good basic statistics book is the one by Wonnacott and Wonnacott (1990). Introductory econometrics textbooks include those by Hill, Griffiths and Judge (1997) and Koop (2008).
2. If you are having trouble grasping this point, draw a straight line with intercept = 0 and slope = 1 through Figures 5.2 and 5.3 and then look at some of the resulting residuals (constructed as in Figure 4.1). You should see that most of the residuals in Figure 5.2 will be much bigger (in absolute value) than those in Figure 5.3. This will result in a larger SSR (see the formula in Chapter 4) and, since residuals and errors are very similar things, a bigger variance of errors (see the formula for the standard deviation of a variable in the descriptive statistics section of Chapter 2 and remember that the variance is just the standard deviation squared).
3. The notation that "the variable W lies between a and b" or "W is greater than or equal to a and less than or equal to b" is expressed mathematically as "W lies in the interval $[a,b]$". We use this mathematical notation occasionally in this book.
4. Note that, as described in Chapter 2, $\sum (X_i - \bar{X})^2$ is a key component of the standard deviation of X. In particular, large values of this expression are associated with large standard deviations of X.
5. For those with some knowledge of statistics, note that t_b is a value taken from statistical tables for the Student-t distribution. Appendix 5.1 provides some additional discussion about t_b.
6. The choice of a 95% confidence interval is by far the most common one, and whenever a confidence interval is not specified you can assume it is 95%.
7. We mean large in an absolute value sense.
8. In the examples in this book we never use s_b and rarely use t. For future reference, the only places we use t are in the Dickey–Fuller and Engle–Granger tests which are discussed in Chapters 10 and 11, respectively.
9. Note that 5.5E-10 is the way most computer packages write 5.5×10^{-10} which can also be written as 0.00000000055.
10. Formally, the F-statistic is only one in an entire class of test statistics that take their critical values from the so-called "F-distribution". Appendix 12.1 offers some additional discussion of this topic.

References

Hill, C., Griffiths, W. and Judge, G. (1997) *Undergraduate Econometrics*, John Wiley and Sons, Chichester.

Koop, G. (2008) *Introduction to Econometrics*, John Wiley and Sons, Chichester.

Wonnacott, T. and Wonnacott, R. (1990) *Introductory Statistics for Business and Economics*, Fourth edition, John Wiley and Sons, Chichester.