Chapter 2: Introduction to quantitative methodology

Aims of the chapter

This chapter aims to familiarise you with some quantitative methodology, mainly regression analysis, to the point where you are able to consume (understand and critically assess) empirical academic papers in development economics. It is meant as a reference chapter that you can return to when reading the remaining chapters of this subject guide.

Learning outcomes

By the end of this chapter, and having completed the Essential reading and Activities, you should be able to:

- recognise when the topics introduced in this chapter are applied in subsequent chapters and reading.

When you have completed the course you should be able to:

- read and critically assess empirical research employing the aspects of quantitative methodology covered in this chapter.

One purpose of this chapter is to familiarise you with quantitative techniques used in the papers you will read throughout the course. Therefore, we urge you to read this chapter thoroughly now (even if you have taken another Statistics course) and review this chapter when reading these papers.

Essential reading

Ray, Debraj *Development Economics*. (Chichester: Princeton University Press, 1998). Appendix 2, pp.777–804.

Further reading

Levitt, Steven D. 'Using electoral cycles in police hiring to estimate the effect of police on crime', *The American Economic Review* 1997, 87(3), pp.270–90.

Introduction

The second chapter of this course is about quantitative methodology, mainly regression analysis. The techniques reviewed here are used extensively, but by no means exclusively, in research in development economics and development policy analysis. Scholars use regression analysis in their attempts to answer questions such as: What are the drivers of economic growth? Is openness to trade related to growth? Can a particular aid programme (a microfinance programme, a programme providing books to school children, budget support for developing country governments) be considered successful? As you can see from these questions, the aspects of quantitative methodology covered in this chapter can be applied to every topic in the remainder of the subject guide: economic growth, international trade, aid, credit, etc.

There is a reason why we start with methodology and not with any of these topics specific to international development. The knowledge about methodology that you will need is similar for all these topics, and we wanted to provide you with a chapter where it is all neatly put together. In fact, the basics of all the methodology you will need to know are in this chapter. In the subsequent chapters, you will see these basics being applied to different areas again and again. This is to say that while the subject matter of this chapter is very important for understanding the rest of the course, you do not need to have fully grasped it all by the end of this chapter. You will practise with it again and again throughout the whole course. You can revisit relevant sections of this chapter every time you are asked to read an empirical paper and will probably understand the material better every time you do so.

A disadvantage of providing you with a methodology reference chapter is that you may feel a bit overwhelmed reading it for the first time. Again, please remember that reading empirical papers and understanding their methodology is a skill that will be built up with time and practice. And as with many skills (like playing the piano, or public speaking), once you learn them, they stick with you and they can be applied in other contexts. If you learn the piano playing Mozart, you will probably do pretty okay playing Bach. Like piano-playing skills, the skills you learn reading an empirical paper about trade can be employed reading empirical work on child labour, the environment or even topics that are completely unrelated to development.

A final thing to remember is that this course aims to teach you how to be a good consumer of empirical research. After working through the subject guide, we hope that you can read an empirical paper, understand most of the quantitative methodology being employed and make an informed argument about how convincing you find the empirical strategy. We do not ask you to be a producer of empirical research. You are not expected to be able to do regression analysis, use statistical programmes or do any calculations whatsoever. You will also never be asked to reproduce numerical results of an empirical paper. In terms of our earlier example, an examination question will never be: 'By what percentage does the productivity of firms go up, if trade barriers are lowered by 10 per cent, according to paper X?' but it may be similar to: 'Why would we expect lower trade barriers to increase firm productivity? How does paper X test this expectation? Do you feel that the authors of paper X have successfully avoided [potential problems with this analysis]?'.

Sampling

When carrying out research, we often want to draw conclusions about a large number of entities: a **population**. We may want to know the average income of **all people in a country**, the productivity of **all firms in a country**, the effect of having new schoolbooks on the attendance of **all school children in a region**. We can go and collect data on the whole population, but this is likely to be time consuming and expensive. When done correctly, sampling can provide a much cheaper alternative to a complete census at a small accuracy cost. **Sampling** is the process by which we select a subset of observations from all possible observations in a population.

When taking a sample, the first step is to define the **target population**. This consists of all the observations about which the research aims to draw conclusions. The **sampled population** consists of all observations

that have some chance of ending up in the sample. Ideally, the target population and the sampled population should be the same. An example: you want to investigate the average number of rooms in London houses. You take a sample from all houses in a London landline phone register. In this case the target population is all houses in London. The sampled population, however, is all houses in London with a landline. Houses without a landline have no chance of ending up in the sample: they are in the target population, but not in the sampled population.

Once the target population is determined and we have a suitable sample frame, we take a sample. Various sampling techniques exist, but here we will focus on a **simple random sample**. In a simple random sample, every member of the population has an equal, known probability of ending up in the sample. You can think of this as throwing all population members into a hat and randomly picking a number of them. Proper random samples of a sufficient size can deliver strikingly accurate predictions about the total population.

If it is not the case that each possible observation has an equal, known probability of ending up in the sample, and/or if the target population is different from the sampled population, the sample is **biased**. It is important to think about the impact that bias can have on the conclusions of the research. When observations that are more or less likely to end up in the sample are systematically different from each other on a characteristic relevant for the outcome, this impact is especially serious. Return to our earlier example: you want to draw conclusions about the average number of rooms in all houses in London, but take a sample of all houses in London with landlines. Houses with landlines may be significantly different from houses without landlines. Maybe older inhabitants are more likely to have a landline, whereas younger ones only have a mobile phone. Age of the inhabitants of a house may be relevant for the number of rooms in the house, as older people may be able to afford bigger houses. Then, the average number of rooms in your sample may be higher than that in the total population of all London houses.

There are many types of sampling biases of which consumers of research have to be aware. We will review a number of them. First, **self-selection bias**. This occurs when subjects themselves have some say over whether or not they are included in the sample. An example: when mailing out surveys, receivers can choose whether or not to send them back. People interested enough to send the survey back may be systematically different from non-respondents. Secondly, **strategic bias** occurs when subjects feel they can manipulate policy choices by participating in the research, impacting the composition of the sample. An example: imagine a microfinance organisation doing a survey of poverty in a village where the organisation will set up a programme. If villagers feel that they can increase their chances of obtaining credit by participating in the survey, the eventual sample may be very different compared to a situation where villagers believe participating will not benefit them (for example when a government statistics bureau is doing a regular poverty survey). Finally, **interviewer bias**. This occurs when the subject responds to characteristics of the interviewer in such a way that the composition of the sample is influenced. When the subject of research is of a sensitive nature this may be especially serious. Consider doing a survey on condom use for example. Whether the interviewers are male or female is likely to impact the willingness of men and women to participate, biasing the sample.

Normal distribution and the central limit theorem

A **probability distribution** is a graphical or mathematical representation that tells us what the relative probabilities are of any of the possible outcomes of a process. For example, we can plot all possible outcomes of a process on the horizontal axis of a graph, and the percentage of the time that these outcomes occur on the vertical axis. Consider the following example: we observe the number of absent children in a 5-child village school, for 45 days. This may have the following result:

No. of children absent	No. of days this is observed	Relative frequency	Probability
0	1	1/45	0.022
1	5	5/45	0.111
2	12	12/45	0.267
3	19	19/45	0.422
4	6	6/45	0.133
5	2	2/45	0.044

Table 2.1

The probability distribution can be represented by the following graph:

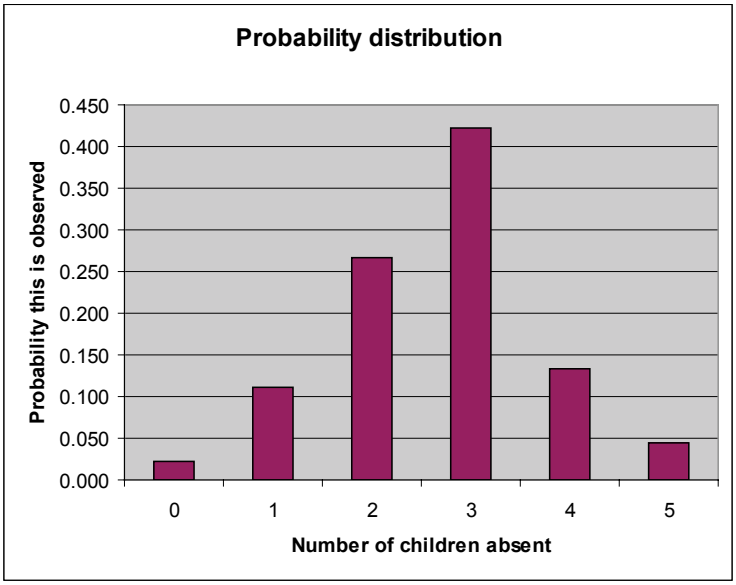

Figure 2.1

The **normal distribution** is a special kind of probability distribution. It is bell-shaped and symmetric, as in the graph below.

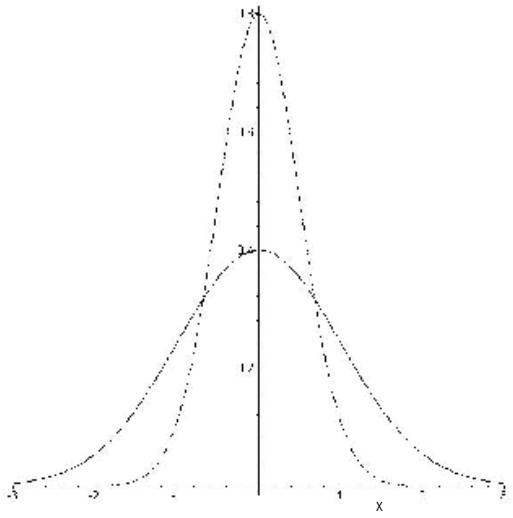

Figure 2.2

As it is symmetric, the **mean** of the normal distribution lies exactly 'in the middle of the distribution'. In the graph above, the mean is 0. Here we will call the mean of this normal distribution μ ('mu'). The bulk of observations is concentrated around the mean. This means that extreme outcomes (such as 3 in the graph above) are very unlikely, whereas outcomes close to the mean are more likely.

The **standard deviation** of a normal distribution is a number that represents how widely observations are spread around the mean. A large standard deviation means a wide 'spread', a small standard deviation means a small 'spread'. In the graph below, the dashed curve has a smaller standard deviation than the solid curve.

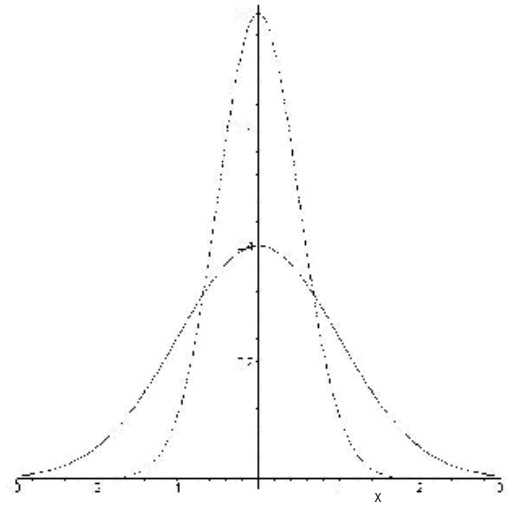

Figure 2.3

Regardless of how wide or narrow they are, all normal distributions have the following properties in common:

- 68 per cent of the probability density of a normal distribution is located within 1 standard deviation around its mean. In Figure 2.4 below: 68 per cent of observations lie between the dashed lines at ±1.
- 95 per cent of the probability density of a normal distribution is located within 2 standard deviations around its mean. In Figure 2.5 below: 95 per cent of observations lie between the dashed lines at ±2.

For example: suppose we know that the height of UK men is normally distributed, with a mean of 175cm and a standard deviation of 5cm. Using this information, we know that 68 per cent of UK men are between 170 (175–5) cm and 180 (175+5) cm tall. We also know that 95 per cent of UK men are between 165 (175–2*5) cm and 185 (175+2*5) cm tall.

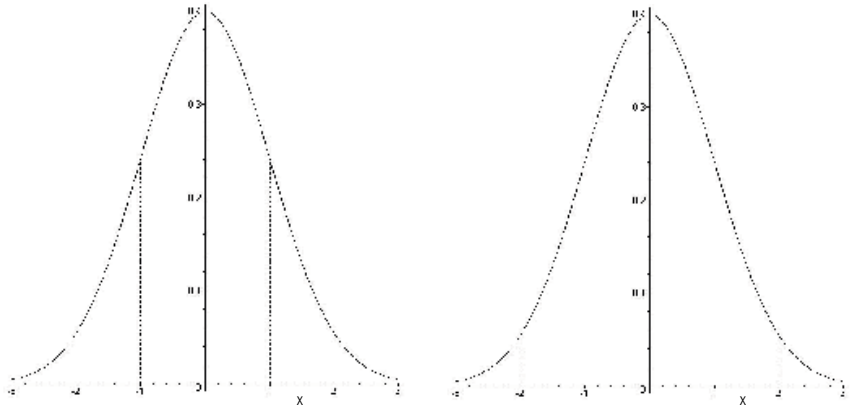

Figure 2.4 and Figure 2.5

Imagine that you draw a large number of random samples from the same target population. You calculate the sample mean of each of these samples. You then draw a probability distribution of the sample mean: what shape would it have? The **central limit theorem** states that as you draw an increasing number of successively random samples from a population, the distribution of sample means will approach a normal distribution. As the number of samples tends to infinity, then the sampling distribution converges to $N(\mu, \sigma^2/n)$. **The distribution of the population itself is irrelevant for this statement.**

Activity

Experiment with the central limit theorem on this website:

- http://onlinestatbook.com/stat_sim/sampling_dist/

Instructions:

1. When you arrive on this website, press 'Begin'. You see four graphs. The first graph shows the distribution of the population from which a sample is taken (the parent population).
2. You can choose between normal, uniform, skewed and customised distributions. Choose the uniform distribution first. A black block will appear because each outcome is equally likely to appear in a (continuous) uniform distribution.

3. The second graph is entitled 'Sample Data'. Press the 'Animated' button and you will see how a sample is drawn from the distribution above. Press 'Animated' a few times. Each time a new blue column is added to the third graph representing the distribution of the sample means. It is the mean of each sample that you obtained by pressing 'Animated'.
4. You can choose to draw 5, 1,000, or 10,000 samples at a time. Accordingly, 5, 1,000, or 10,000 blue sample means are added to the distribution of sample means. Press these buttons a few times. Soon you will see a normal distribution appear. Note that this happens even though the target population is *not* normally distributed.
5. On the right-hand side of the 'Distribution of Means' graph you can choose the sample size of the samples you draw. When you choose 'N=2' a sample mean distribution with a large *standard error* ('standard error' is the technical name for the standard deviation of the sample mean) will appear. If you choose a higher value, e.g. 'N=25' a more narrow distribution will appear.
6. Now repeat the exercise for a skewed distribution and you will see that again the distribution of means will become normal. Experiment with the simulation as much as you like!

Hypothesis testing

As explained in the introduction, theories or observations of the real world can give rise to certain predictions. To test these predictions, we formulate them as **hypotheses**. For each prediction, we formulate two hypotheses: the **null hypothesis (H_0)** and the **alternative hypothesis (H_1)**. Together, these hypotheses should cover all possible outcomes. Usually, H_0 concludes that 'nothing happened', whereas H_1 concludes that 'something happened'. For example:

H_0	H_1
A treatment has had no effect on an outcome variable	A treatment has had an effect on an outcome variable
Group A is not different from group B	Group A is different from group B
Variable *X* is unrelated to variable *Y*	Variable *X* is related to variable *Y*

Table 2.2

H_0 is the hypothesis to be tested – effectively the 'working hypothesis'. We observe something in the real world and then calculate how likely it would be for us to observe what we did, **assuming that H_0 is correct**. If this is sufficiently unlikely, we **reject H_0** in favour of the alternative hypothesis.

Consider the following example: we may think that UK men aged 18–25 are taller on average than all UK men. We know that UK men are 175cm tall on average. We formulate the null and alternative hypotheses:

a. H_0: UK men aged 18-25 are not taller on average than all UK men.

b. H_1: UK men aged 18-25 are taller on average than all UK men.

We take a random sample of all UK men aged 18–25. The men in our sample are 179cm tall on average. Now we ask ourselves: what is the probability that we would get a sample with a mean of 179cm, when the *true* population mean height of men aged 18–25 is 175cm (assuming that H_0 is true)? We calculate (do not worry how for now) that this probability is 4 per cent.

So we now know that getting this sample with an average of 179cm is unlikely if H_0 were true. But is it 'sufficiently unlikely' to reject H_0? The researcher **chooses** the probability level at which to reject H_0. This is called the significance level. Common significance levels are: 10 per cent, 5 per cent and 1 per cent. When the probability that H_0 is true given our observations is below 10 per cent, 5 per cent or 1 per cent respectively, H_0 is rejected at the respective significance level. In our example, H_0 is rejected when we choose a significance level of 10 per cent or 5 per cent, but we fail to reject H_0 when we choose a significance level of 1 per cent. (Note the phrase 'fail to reject' – we *never* 'accept H_0' merely we state that we have insufficient evidence to reject it.)

When deciding whether or not to reject the null hypothesis we can make two types of error:

- A **Type I error** occurs if we reject H_0 even though it is in fact correct.
- A **Type II error** occurs if we fail to reject H_0 even though it is in fact wrong.

The researcher controls the probability of making a Type I error by setting the significance level. If we choose a decision rule to reject H_0 when the probability that it is correct is below 5 per cent, there is a maximum 5 per cent chance that we reject H_0 when it is in fact correct. (Basically this would occur if we obtained an unlucky or 'unrepresentative' sample.) If we set our significance level at 1 per cent, the probability of making a Type I error gets smaller. As may be evident from this example: the smaller the chance of a Type I error, the larger the chance of a Type II error – hence a trade-off between the two error types exists (for a given sample size). Because we do not know the 'truth' we cannot calculate the probability of a Type II error.

So how do we calculate the probability that H_0 is true given our observations? We have an observed sample mean ($\bar{x}$) and hypothetical population mean (μ_0). Because of the central limit theorem, we know that the sample mean is (approximately) normally distributed. We also have a good approximation of the standard error of the sampling distribution ($s/\sqrt{n}$, which approximates $\sigma/\sqrt{n}$, where σ is the population standard deviation). We calculate by how many standard errors the observed sample mean and the hypothetical population mean are apart. This number is called the ***t*-statistic**.

For example, our calculated sample mean and the mean under H_0 are two standard errors apart: the *t*-statistic is 2. Written down as a formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The *t*-distribution is very similar to the standard normal distribution N(0,1), but for small samples it has slightly fatter tails, as we had to estimate the standard error of the sampling distribution and therefore introduced a bit more uncertainty. However, these distributions are very similar when we take a sample of reasonable size. (As a rule of thumb, we may use the standard normal distribution as an approximation for the *t*-distribution for 'large' sample sizes, for example $n \geq 50$.)

An overview of critical values of when the *t*-statistic converges to the standard normal (i.e. when the sample size is (very) large):

- If the observed sample mean and hypothetical population mean are further apart than 1.96 standard errors we can reject H_0 at the 5 per cent level.

- If the observed sample mean and hypothetical population mean are further apart than 1.64 standard errors we can reject H_0 at the 10 per cent level.
- If the observed sample mean and hypothetical population mean are further apart than 2.58 standard errors we can reject H_0 at the 1 per cent level.

When sample sizes are a bit smaller, the tails of the *t*-distribution are a bit fatter and the critical values are a bit larger. Critical values for different 'degrees of freedom' are included in the back of most statistics textbooks. For the most part, using a rule of thumb value that you will reject the null hypothesis for any *t*-statistic that is 2 or larger in absolute value (for statistical significance at the 5 per cent level) is a pretty common and relatively benign practice.

The **p-value** reports the probability mass in the sampling distribution that lies outside of your estimate. For example, a p-value of .035 means that 3.5 per cent of the probability mass of the sampling distribution under the null lies outside your sample estimate (technically this definition is for a one-sided test – but you do not need to worry about one-sided and two-sided tests). Thus if you want to test the null at 5 per cent you will look for a p-value of .05 or lower. From the example above, our p-value of .035 means we can reject the null hypothesis at 5 per cent, but not at 1 per cent.

Basics of regression

In this section, we are going to bring together the material we have covered so far to learn the basics of regression. **Regression analysis** in essence does two things: it establishes a conditional correlation between two variables and it determines how likely it is that this correlation is due to chance.

A **correlation** is a systematic relationship between two variables. For example: aid spending is related to poverty reduction, or: trade openness is related to growth. Note that a correlation may be positive or negative: **more** aid spending is related to **less** poverty (a negative correlation) and *more* trade openness is related to **more** growth (positive correlation). Note that a correlation between X and Y does not necessarily mean that X **causes** Y. Correlation does not necessarily imply causation. You will see why in the next section.

Establishing that two variables are correlated in a sample might tell us something, but is of limited policy relevance by itself. Between most variables, there will be some correlation due to pure chance. If you let 100 people roll a die three times, you will probably be able to find people whose rolls seem to be correlated. However, as rolling a die is random, we can comfortably say that this correlation is due to chance.

We need to apply what we learned about hypothesis testing here: if the probability of observing a particular relationship between X and Y, when they are in fact unrelated, is below a certain significance level (say there is less than a 5 per cent chance that we would observe a particular relationship in a sample if in fact there was none in reality in the population), we conclude that there is a **significant** relation between X and Y. For example: we observe a positive correlation between education and wage: the more education the higher the wage. From the central limit theorem we know there is less than a 3 per cent chance of observing this relationship if education and wage were unrelated. If our significance level was set at 5 per cent, we can now say that the positive relationship between education and wage is statistically significant, at the 5 per cent level. Note that the relationship is not significant at the 1 per cent level.

How does establishing a correlation work more technically? Although we do not go into the full technical details, we will illustrate this graphically. Imagine that you have two variables that you think are related: X and Y. We can plot these on a graph.

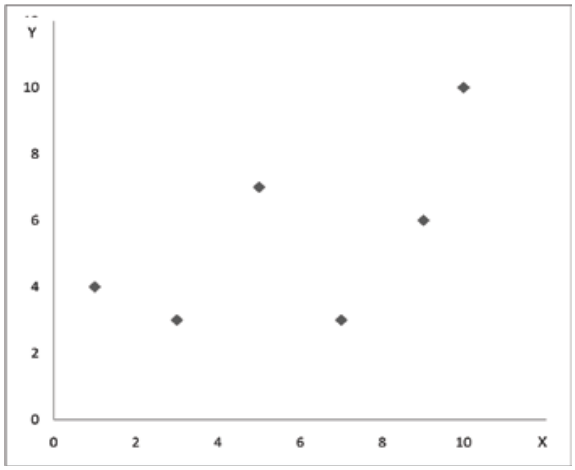

Figure 2.6

Now we think about a single line that will best depict the relationship between X and Y, given these points (see below). In this case, the line is upward-sloping suggesting a positive relationship. A negative relationship can be depicted as a downward-sloping line.

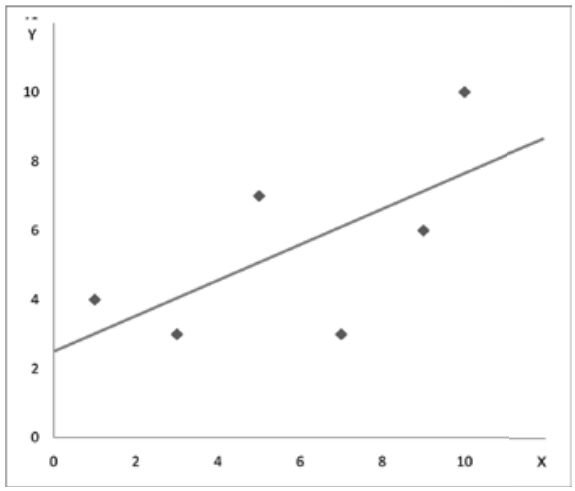

Figure 2.7

How can we determine which is the 'best' line? Statisticians have defined the 'best' line as the line that minimises the sum of squared errors. The error is the vertical distance of each individual point to the regression line. Statistical software can single out the one line for which the sum of these squared distances is smallest. Since the sum of squared errors is central in this type of regression, it is also termed '**Ordinary Least Squares**' **(OLS)**.

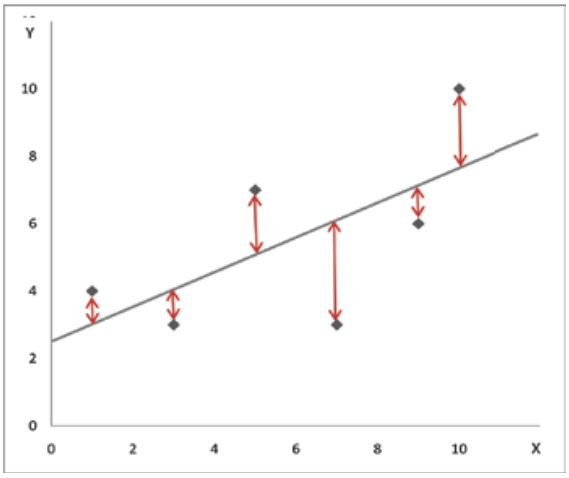

Figure 2.8

The graph illustrates how a regression line is derived.

You may remember that lines are usually described using a slope, *a*, and an intercept, *b*:

$y = ax + b$

Analogously, the regression equation is:

$y = \alpha + \beta x + \varepsilon$

This simply says that each point in our graph can be described by the regression line (with an intercept, α, and a slope, β) plus the error, ε (the vertical deviation of a given point from the regression line). We call *y* the **dependent variable** or outcome variable, and *x* the independent variable or **explanatory variable**. β is a **regression coefficient** (or **slope coefficient**). Note that a positive β indicates a positive relationship between *x* and *y*, whereas a negative β indicates a negative relationship. No *linear* relationship at all would mean $\beta = 0$. In this equation, there is one explanatory variable. In practice, we will mostly see regressions with more than one explanatory variable, resulting in a multiple linear regression equation:

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots. + \beta_k x_k + \varepsilon$

Activity

On this website www.calpoly.edu/~srein/StatDemo/All.html, you will see a regression graph. Add, delete or move points in the graph and see how the regression line changes. In particular, try the following:

- When the regression model opens, you see an example of a negative correlation. Add points to turn it into a positive correlation.
- Click on 'show residuals' to see how large the error for each point is.
- Clear all points by clicking on 'clear all points'. Add two points that display a clear positive relationship. Now add one more point that does not confirm this relationship (having a large error; being very far from the line). Observe the impact of adding this one point on the position of the regression line.
- Clear all points. Add twenty points that display a clear positive relationship. Again add one more point that does not confirm this relationship. How does the impact on the position of the regression compare to the previously observed impact?

- Clear all points and temporarily remove the regression line by clicking on 'no regression line'. Attempt to add twenty-five points that are completely randomly distributed over the graph. Bring back the regression line by clicking on 'show regression line'. Did you succeed in creating no correlation between x and y whatsoever?
- Clear all points and temporarily remove the regression line. Add twenty points that have a U-shaped relationship: high levels of y for both very low and very high levels of x and low levels of y for medium levels of x. Perhaps this represents the level of health care an individual needs at various ages: high levels of health care as a more vulnerable child or senior and low levels of health care as a resilient adult. Bring back the regression line. How well do you feel that the regression line reflects the relationship created?

At the bottom of the graph, you can see the regression equation that describes the regression line you can create by adding, deleting and moving points.

The values of and when the regression model opens are α=70.08 and β=-0.472 respectively. Try the following, clicking 'clear points' each time you start a new item:

- Create a line with a large negative value of β.
- Create a line with a small negative value of β.
- Create a line with a positive value of β.
- Attempt to create a line with $\beta = 0$.
- Attempt to create a line with $\alpha = 0$.

Returning to the one explanatory variable model, we reformulate our question: how likely is it that we observe a given relationship between *X* and *Y* in our sample when in reality, in the whole population, *X* and *Y* are unrelated? Alternatively, if we estimate a slope coefficient to be 5.46, we want to know, 'how likely is it that we could derive a slope estimate of 5.46 from a sample when the true underlying relationship (true slope coefficient) in the population was 0?' Or, more simply: How likely is it that we observe a given $\hat{\beta}$ (where $\hat{\beta}$ ('beta hat')is our point estimate of the true β), whereas in reality β=0?

To distinguish the relationship between *X* and *Y* observed in a sample from the 'true', unknown relationship, we call the observed relationship $\hat{\beta}$. We now need to apply what we learned in previous sections. From hypothesis testing:

- We set H_0: β=0. There is no linear relationship between *X* and *Y*.
- H_1: $\beta \neq 0$. There is a linear relationship between *X* and *Y*.

From the central limit theorem we can analytically derive the shape of the sampling distribution of $\hat{\beta}$ and derive a *t*-statistic. We use the *t*-statistic to determine the level of statistical significance of our estimate.

If we observe a $\hat{\beta}$ that is sufficiently unlikely given H_0, we reject H_0 and say the slope coefficient is statistically significantly different from 0. Note that we are using the term 'significantly' very specifically here! The researcher can determine what is 'sufficiently unlikely' his/herself by choosing a significance level.

The sample distribution of $\hat{\beta}$ under H_0: $\beta = 0$ can be visualised as in the graph below.

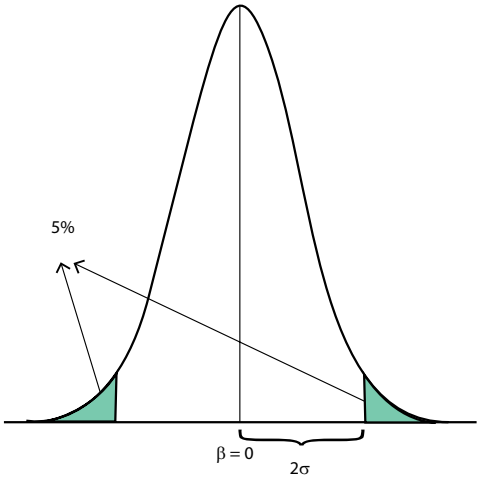

Figure 2.9

We omit technical details, but for $\hat{\beta}$ to be significant at the 5 per cent level the criteria listed below hold. If one of these is true, this automatically means they are all true, so in practice it is only necessary to check one:

- The chance of observing $\hat{\beta}$ when the 'true' $\beta=0$ is smaller than 5 per cent. This 5 per cent is represented by the blue shaded area in the graph. **The p-value is smaller than 0.05**.
- The observed $\hat{\beta}$ is at least two times the standard error away from 0.
- The distance between $\hat{\beta}$ and 0 (which is simply $\hat{\beta}$) divided by the sample standard error is also called the *t*-statistic. **The *t*-statistic of $\hat{\beta}$ is bigger than 2 or smaller than –2**. Note that the actual critical value associated with the 5 per cent level may be slightly smaller or slightly larger than 2, depending on the sample size. But we use 2 as a rule of thumb.

Now, we have all the information we need to read a regression table. When reading a regression table, pay attention to these features:

- The dependent variable '*y*' is usually mentioned at the top or bottom of the table.
- All explanatory variables included in the regression (the different '*x*'s), are listed in the rows of the table.
- The resulting $\hat{\beta}$'s, or coefficients, are in the corresponding rows.
- In parentheses under these coefficients, we are provided with information on the significance of the coefficient. This can be provided in 3 forms, analogous to the three criteria that hold when $\hat{\beta}$ is significant.
- A p-value, indicating the chance that $\hat{\beta}$ is observed when $\beta=0$. A p-value lower than 5 per cent indicates significance at the 5 per cent level.
- A standard error. A coefficient that is larger than twice the standard error is significant at the 5 per cent level.
- A *t*-statistic. A *t*-statistic larger than 2 indicates significance at 5 per cent level.
- Often, the author indicates significance using asterisks (*).

An example of a regression table is given below. The authors attempt to investigate the impact of independent media on public food distribution and calamity relief expenditure.

Figure removed due to copyright restrictions

Figure 2.10

From: Besley, Timothy and Robert Burgess 'The political economy of government responsiveness. Theory and evidence from India'. The Quarterly Journal of Economics *117(4), 2002, pp.1415–451. © Used with kind permission.*

Endogeneity

In the previous section we have said correlation does not imply causation. Although there may be a (significant) correlation between A and B, this does not necessarily mean that A has some causal effect on B. A number of endogeneity problems may inhibit us from drawing such a conclusion. We will discuss **omitted variable bias** and **reversed causality**. Before going into this, this section will give an introduction to endogeneity.

Imagine we want to investigate the effect of a 'treatment' on an outcome. Ideally, we would like **all observations to be identical** and to differ only in whether or not they get the treatment (as in panel A in the figure below). This way, we can have confidence that the observed effect on the outcome variable is due to the treatment, and not due to anything else. For example: We want to examine the effect of taking vitamins on health. We can administer vitamins to a group of genetically identical mice in a laboratory and observe any changes in their health.

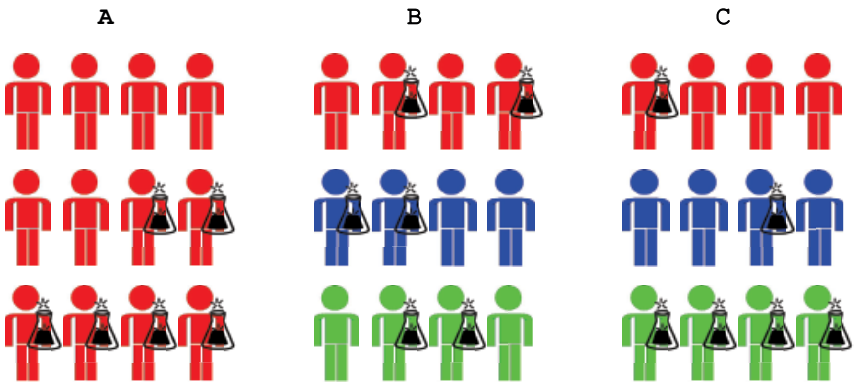

Figure 2.11

Please see the PDF on the VLE for a colour version of this Figure.

However, ideal experiments like A will rarely occur in the social sciences. Most analyses must be done on **non-experimental data**. Observations are likely to be different across *many* characteristics, not only whether or not they get the 'treatment'. We want to make sure that any effect we observe on the outcome variable is solely due to the treatment, not due to these other characteristics.

One way to ensure this would be for the **treatment to be distributed at random**, which makes it likely that the observations receiving the treatment will not be systematically different from the observations not receiving the treatment (as in panel B). To go back to our example: we can bring in a group of participants, randomly assign half of them to take vitamins, give the rest a placebo and observe any health effects.

Conducting an experiment as in B is not always possible. We may have to rely on **observed data** on treatment and outcome. It is possible that individuals who received the treatment are systematically different with respect to some characteristics, from those who did not receive the treatment, as in panel C. We cannot be completely sure that any observed effect on the outcome is due to the treatment, rather than these characteristics. It may also be difficult to observe what 'came first': the outcome or the treatment.

Go back to our example once more: we observe the health status of a number of people and ask them about their vitamin use. People taking vitamins are on average healthier than people who don't. However, people choosing to take vitamins may be different with respect to a number of characteristics: they may have a healthier lifestyle in general, be better informed about health issues, etc. These characteristics may influence both their health status and their decision to take vitamins. We cannot conclude that taking vitamins improves health necessarily.

These examples illustrate the difference between an exogenous and an endogenous variable:

- Variation in an **exogenous variable** – is determined outside the model.
- Variation in an **endogenous variable** – is determined within the model.

When we allocate vitamins at random, variation in vitamin use is not determined by some process relevant to health. The variation can be considered exogenous. However, when we observe vitamin use, variation is a function of processes within the system (people's choice to take vitamins is related to all the other choices they make, their preferences), that may be relevant to health. This variation can be considered endogenous.

One endogeneity problem is **omitted variable bias**. This section will explain what omitted variable bias is and how it can impact conclusions drawn from a regression. An **omitted variable** is correlated with both an explanatory variable and the outcome variable in a regression, but is itself not included in the regression. If such an omitted variable exists, the relationship between an explanatory variable and the outcome variable observed in a regression, may be different from the 'true' relationship between these variables. This is called omitted variable bias.

An example

We run a regression with the number of babies born in a particular area as a dependent variable and the number of storks in that area as an explanatory variable. The regression shows that the number of storks is significantly correlated to the number of babies born. We may conclude that babies are delivered by storks. Schematically:

Observed relationship

Regression equation: #babies = $\alpha + \beta$ #storks + ε

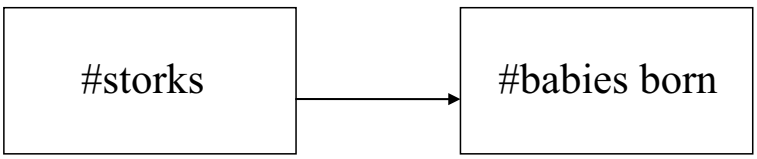

Figure 2.12

You may already suspect that the observed relationship is different from the actual relationship. The omitted variable in this case may be whether or not the area is in the countryside. 'Countryside' is correlated to the number of babies born, as people may prefer raising their children in a greener, safer environment. We can also expect storks to prefer to live in the countryside. Hence, the omitted variable causes variation in both the dependent and independent variables: areas in the countryside have both a higher number of storks and a higher number of children born. There is no actual causal relationship between the number of children born and the number of storks. Schematically:

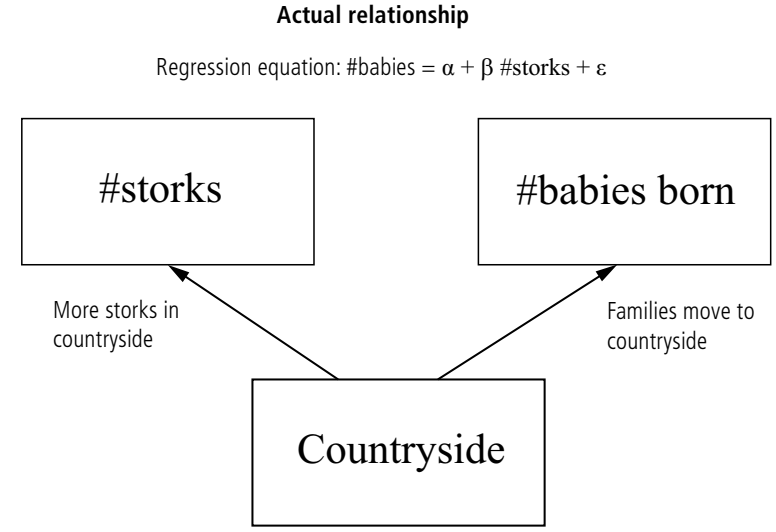

Figure 2.13

The previous example is an example of **upward bias**: because of the omitted variable, the relationship between the modelled explanatory variable and the outcome variable appears more strongly positive than the 'true' relationship. An omitted variable can also result in **downward bias**: because of the omitted variable, the relationship between the explanatory variable and the outcome variable appears more strongly negative than it is in reality. Schematically:

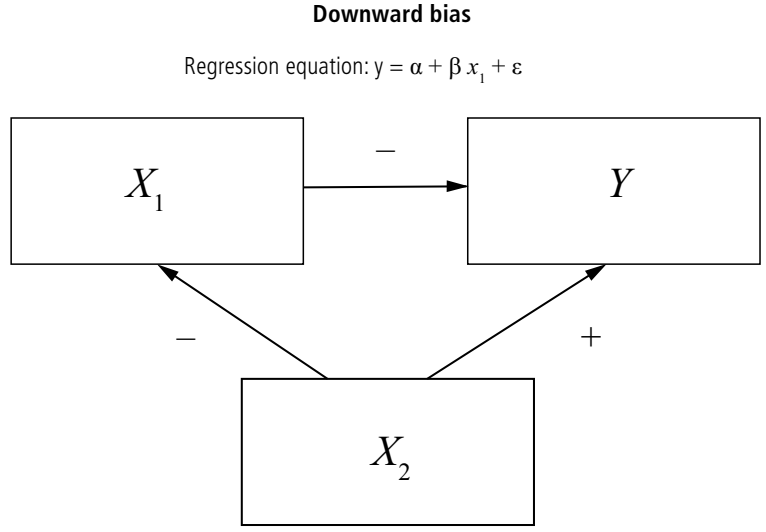

Figure 2.14

An example

We want to investigate whether government financial support to schools increases pupil performance. We find a negative relationship between the amount of money a school received from the government and the average performance of its pupils. We may conclude that the more money a school receives from the government, the worse its students do. However, an omitted variable may be school wealth. The government may target financially needy schools, so wealthier schools are less likely to

receive government support. Pupils in wealthier schools can be expected to perform better, as the school can afford better teachers, materials etc. In this case, the omitted variable makes the impact of government spending on pupil performance appear more negative than it is. In reality, government spending may still improve student performance, but the effect of the omitted variable may be stronger, and the regression coefficient may have a negative sign.

Reversed causality is another problem associated with endogeneity. Reversed causality occurs when variation in the outcome variable causes variation in the explanatory variable, rather than the other way round. Reversed causality can make an explanatory variable appear a stronger cause of the outcome variable than it is in reality. For example, we may observe a positive correlation between the number of policemen and the crime rate. This may mean that policemen cause people to commit crimes. It may also mean that in cities with a higher crime rate, more policemen get hired (which sounds more likely).

Activity

Consider the following explanatory and outcome variables. For each of these cases, answer the following questions:

- What is a possible omitted variable?
- Which mechanisms connect this omitted variable to both the explanatory and outcome variable?
- What is the direction of the bias (upward or downward)?

Explanatory variable	Outcome variable
A microfinance programme	Poverty
Trade openness	GDP growth
Taking a statistics course	Knowing that correlation does not equal causation
Educational attainment	Earnings
Distance of individuals' houses to the town centre	Expenses made for travel (£)
Farm size	Farm productivity

Table 2.3

Control variables, dummy variables and interaction terms

This section will cover different types of variables that are commonly used in regression analysis, and their interpretation.

Including a **control variable** in a regression is one way to attempt to deal with omitted variable bias. Take the following example: we are interested in the question of whether a lower student-teacher ratio (str) improves students' performance. To assess this, we regress average test scores of students in a district on the average student teacher ratio:

$$\text{test score} = \alpha + \beta_1 \text{str} + \varepsilon$$

The regression shows a negative relationship between the average student-teacher ratio in a district and the average test scores. Students in districts with lower student-teacher ratios achieved higher test scores on average.

We may suspect that this result is biased downward, as we can think of several potential omitted variables:

- Perhaps district wealth is both related to higher test scores (students in richer districts are in a more conducive environment to learn) and a lower student-teacher ratio (schools in richer districts can afford to hire more teachers).
- Perhaps the value that parents put on education is correlated with test scores (parents valuing education more require their children to work harder in school) and with student-teacher ratios (parents putting high value on education push the school to hire more teachers).

To address the first problem, we include district wealth in our regression: we **control for** district wealth.

$$\text{test score} = \alpha + \beta_1 \text{str} + \beta_2 \text{wealth} + \varepsilon$$

You can think about this as artificially holding wealth across districts constant; we compare student-teacher ratios and test scores of districts that artificially have the same wealth. If our suspicion that district wealth is an omitted variable was correct, we would see that $\widehat{\beta}_1$ increases (it becomes 'less negative') compared to our regression without a control variable. We expect that omitting district wealth makes the relationship between student-teacher ratios and student test scores seem more negative than it is in reality and controlling for district wealth should bring $\widehat{\beta}_1$ closer to its 'true' value.

Variables that you include in a regression to exclude a source of omitted variable bias, while not being primarily interested in its causal effect on the outcome variable, are sometimes called control variables. The distinction between explanatory variable and control variable, however, is very loose and is not always made. The name 'control variable' has to do with the *function* of the variable in relation to the purpose of the regression analysis, not because there is anything intrinsically different about them compared to other variables.

Including control variables cannot solve all omitted variable problems. Consider our second potential omitted variable: parents' valuation of education. This is very hard to observe, let alone to measure and to include in a regression. Some sources of omitted variable bias can be dealt with by including control variables, others cannot.

Some variables we may want to include in a regression are dichotomous: they can take only one of two values:

- a person's gender (male/female)
- whether a person has a college degree (yes/no)
- whether a person has been to jail (yes/no)

This information could be captured using a dummy variable. A **dummy variable** is a variable that can take on only one of two values: coded either zero or one.

We may, for example, be interested in investigating the effect of education on 'wage'. We suspect that whether an individual is male or female also has an impact on wage. We have a database with information on individuals, including their level of education, gender and the wage they earn. We construct a dummy variable ('male') equalling 1 if a person is male and 0 if a person is not. Ignoring omitted variable problems for now, we estimate the following regression:

wage = $\alpha + \beta_1$ years of schooling + β_2 male + ε

We can interpret β_2 as the effect of being male on wage. β_1 can be interpreted as the effect of schooling, controlling for gender. Graphically, including the dummy variable allows for two regression lines representing the relationship between schooling and wage, one for men and one for women, with different intercepts (but the same slope, β_1). In the graph below:

- α captures the wage of a woman without schooling.
- $\alpha + \beta_2$ captures the wage of a man without schooling.

For both genders, the effect of schooling on wage is captured by β_1. The wage of a man with one year of schooling, is reflected by $\alpha + \beta_2 + \beta_1 * 1$. The wage of a woman with one year of schooling is reflected by $\alpha + \beta_1 * 1$.

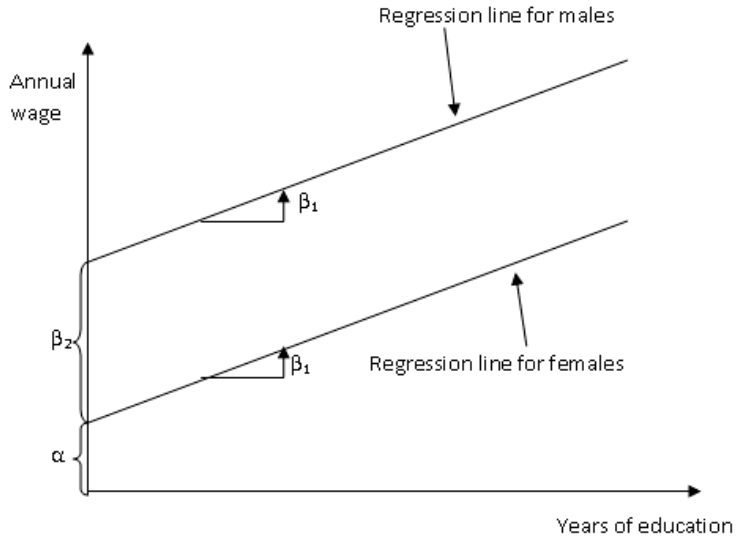

Figure 2.15

Note that we did not include dummy variables for being female **and** male **plus** an intercept term in our regression. Doing so would violate one of the basic assumptions of OLS regression. It would become impossible to estimate the regression equation. All you have to remember about this is that a researcher will omit one category of the dummy variable from the equation if an intercept term is included.

In our example, we assumed that the effect of education is the same for men and women. We did not consider the possibility that the effect of education on wages might be different for men and women. Say we expect that one year of education will steeply increase a man's wage, whilst it has a weaker effect on a woman's wage.

An **interaction term** allows us to incorporate this idea into a regression. An interaction term consists of two explanatory variables multiplied together. A regression with interaction term can take this form:

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) + \varepsilon$

In this equation, β_1 captures the effect of x_1, β_2 captures the effect of x_2 and β_3 captures the *additional* effect of the combination of x_1 *and* x_2.

For our example, the regression equation would be:

wage = $\alpha + \beta_1$ years of schooling + β_2 male + β_3 (years of schooling * male) + ε

where β_3 captures the additional effect of being male **and** having a certain amount of schooling on wage. Graphically, an interaction term allows for different **slopes** (as well as intercepts) of the regression line for different categories of observations. Remember that a dummy variable allows for different **intercepts** only.

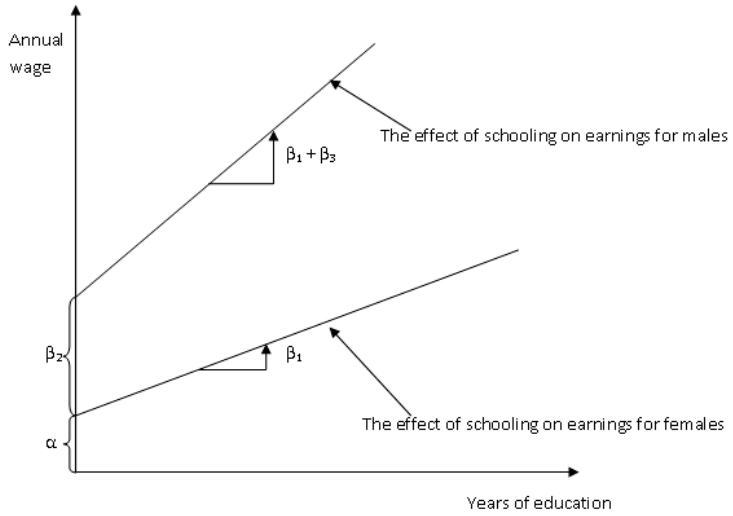

Figure 2.16

In the graph above:

- α captures the wage of a woman without schooling.
- The effect of schooling on wage for women is captured by β_1.
- As before, the wage of a woman with one year of schooling, is reflected by $\alpha + \beta_1 * 1$.
- $\alpha + \beta_2$ captures the wage of a man without schooling.
- The additional effect on wage of the combination of having schooling and being male is captured by β_3.
- The wage of a man with one year of schooling, is reflected by $\alpha + \beta_2 + \beta_1 * 1 + \beta_3 * 1$.

Cross-section, time series and panel data

This section will consider three kinds of data that could be used to run a regression: cross-section, time series and panel data. Different types of data may be more or less suitable to answer different research questions.

Cross-section data contains information on different cross-section units at the same point in time (or an average over time). Cross-section units may be countries, people or any unit of observation arrayed across space. The cross-section dataset only contains information on different values of the cross-section units, and no information on variation **within** these cross-section units over time. Thus, cross-section data may contain characteristics (wealth for example) of different countries, provinces, individuals, households, etc. at a single point in time or an average over a single time period. It may give information about the differences in wealth between countries or households, but no information on how the wealth of individual countries or households developed over time.

An example

Imagine that we want to investigate whether a lower national tax rate attracts more investment from abroad. Cross-section data we could use to try to answer this question could look as follows:

Tax rates and FDI inflows in 1998		
Country	Tax rate (%)	FDI inflow (million$)
Argentina	32	461
China	21	3650
France	42	996
Nigeria	20	210
United States	28	1680
Zambia	17	30

Table 2.4

Time series data contains information on a single unit of analysis at different points in time. It contains variation over time only. Thus, time series data may contain information on how the wealth of an individual country or household developed over time. We could use only one time series, using information about past wealth in a country to predict future development of its wealth. Alternatively, we could investigate the relationship between multiple time series, for example research the relationship between changes in trade policy and changes in wealth in one country.

Returning to the earlier example, investigating the potential relationship between taxes and foreign investment, time series data could look like this:

Tax rate and FDI inflow in China		
Year	Tax rate (%)	FDI inflow (million$)
1960	23	501
1970	35	713
1980	41	278
1990	33	1044
2000	21	3650
2010	26	4120

Table 2.5

Panel data contains information on different cross-sectional units at different points in time. In other words, it combines cross-section and time series data. Panel data contains both variation *between* units of observation and **within** units of observation over time. Thus, panel data may contain information on how the wealth of an individual country or household developed over time, and information on differences in wealth between various countries or households.

Panel data that could be used to do our example research into the relationship between taxes and FDI could look like this:

Tax rates and FDI inflows			
Country	Year	Tax rate (%)	FDI inflow (million$)
China	1990	33	1044
China	2000	21	3650
China	2010	26	4120
Nigeria	1990	13	199
Nigeria	2000	20	210
Nigeria	2010	26	861
Zambia	1990	11	59
Zambia	2000	17	30
Zambia	2010	13	150

Table 2.6

We have identified three kinds of data: cross-section, time series and panel data. None of these types of data is intrinsically 'better' than others. However, when we have a particular research question in mind, a specific type of data may be better suited to answer it. Here is a (non-exhaustive) list of examples:

Cross-section data can be useful when the research question concerns differences between observations, for example in the following cases:

- When one or more of the variables we wish to research changes very little over time.
    - Do countries with a legal system of Anglo-Saxon origin provide better protection for investors than countries with a legal system of a different origin?
    - In a country with low household mobility: do households that live in an area with TV reception have a different attitude towards the ruling government than households in areas without such reception?
- Even when we do expect the variables of interest to change over time, data may only be available at one point in time.

A research question that concerns only one cross-section unit (there is only one world price of oil, exchange rate between the euro and the dollar, Dow Jones index) will necessarily use **time series data**, for example:

- How does the world price of grain respond to changes in the production of bio diesel?
- Given past US interest rates and other variables, what is the most likely US interest rate one month from now?

Panel data can be useful when the research question concerns both differences between entities and over time, for example:

- When we want to investigate the effect of some 'treatment'.
    - Do women that participate in a microfinance scheme experience a greater increase in their sense of empowerment over time than non-participating women?

In the context of panel data, you may come across the term **fixed effects**. To understand this concept, let's go back once more to our example, investigating whether low tax rates attract foreign investment with a panel data set. There are a number of possible omitted variables that might bias the analysis if we run a regression with FDI as a dependent variable and tax rates as an independent variable. Perhaps countries with a low tax rate also have other policies that attract investors. Or perhaps some attribute of Chinese culture both favours low taxes and draws investors. We could introduce numerous control variables to address the omitted variable problems. However, there could still be 'something about China' we do not observe or that we cannot measure, that may make it both prone to have a low interest rate and high levels of FDI.

To (partially) address omitted variable problem(s), we can introduce **country-fixed effects**. We add a dummy variable that equals one if the observation concerns China and zero otherwise. We construct such dummy variables for all countries in the sample, which will make our data look like this:

Tax rates and FDI inflows						
Country	Year	Tax rate (%)	FDI inflow (million$)	Dummy China	Dummy Nigeria	Dummy Zambia
China	1990	33	1044	1	0	0
China	2000	21	3650	1	0	0
China	2010	26	4120	1	0	0
Nigeria	1990	13	199	0	1	0
Nigeria	2000	20	210	0	1	0
Nigeria	2010	26	861	0	1	0
Zambia	1990	11	59	0	0	1
Zambia	2000	17	30	0	0	1
Zambia	2010	18	43	0	0	1

Table 2.7

You can think of the variable 'dummy China' as a control for 'being China'. It controls for both observed and unobserved characteristics of China **that do not change over time**. In our example, fixed effects might control for some attribute of Chinese culture, if we expect this to be stable over the research period. Including fixed effects would not adequately control for GDP differences between research countries: we would expect GDP to change substantially over time.

When estimating a model using panel data and including **fixed effects**, it includes dummy variables for each individual entity, controlling for characteristics of these entities **that do not change over time**. We can thus introduce country-fixed effects, province-fixed effects, household-fixed effects, individual-fixed effects, etc. depending on the data at hand. Another type of fixed effects we can introduce is **time-fixed effects**. When including time-fixed effects, one includes a dummy for each relevant time period (a year dummy, or decade dummy). This controls for characteristics of a period **that do not vary between cross-section units**.

Instrumental variable (IV) regression

Problems of endogeneity make it hard for researchers clearly to identify causal relationships. Using **instrumental variable regression** can potentially solve, or at least mitigate, problems of endogeneity and **identify** causal relationships. The challenge is to find a truly valid instrument. This section will provide an introduction to instrumental variable regression.

Consider a regression equation:

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

We suspect that *x* is an endogenous variable. We can thus not interpret β_1 as the causal effect of *x* on *y*. Attempting to solve this problem, we search for a valid **instrumental variable** or **instrument**. Let us call our instrument *z*. For *z* to be a valid instrument, it has to be exogenous and satisfy two conditions:

- **Relevance condition**: *z* is correlated with *x*.
- **Exclusion restriction**: *z* is not related to *y*, *except* through its correlation with *x*.

Using the instrument *z* we can extract out some exogenous variation in *x*. If we then use this 'extracted' exogenous variation in *x* to estimate the effect of *x* on *y*, we can measure the unbiased causal effect of *x* on *y*.

This is all rather abstract, so let us consider an example (after Levitt 1997, further reading). We want to assess the effect of police presence in a city on crime. Variation in police presence, however, is endogenous. Reversed causality is an especially pressing concern: in cities that have a high crime rate, more police officers get hired. Running a regression that looks like this:

crime rate $= \alpha + \beta$ # police officers $+ \varepsilon$

is unlikely to give us unbiased estimates of the effect of police presence on crime.

Now say that we think that a dummy indicating whether a given year is an election year is a valid instrument for the number of police officers. Would this satisfy our two conditions?

- Is being in an election year correlated to the number of police officers?

Maybe, if we expect that officials aiming to get re-elected would hire more police officers to satisfy voters. It is easy to check this condition: we can run a regression with the number of policy officers as a dependent variable and the election year dummy as an explanatory variable and see if the two are significantly correlated.

- Is being in an election year unrelated to crime except for its effect on crime through the number of police officers?

It may seem reasonable to believe that an election year does not have an independent effect on crime. There is no way to prove this definitively, however, and we have to use our common sense to assess whether the exclusion restriction is satisfied. For example, if we believe that there are more protests in an election year, which may be related to crime (throwing rocks at the police, smashing cars or shop windows), our instrument is not valid. Equally, if we think that an increase in crime will lead citizens to force officials out of office, leading to new elections, our instrument is not valid. In the USA (where the study was done) municipal elections occur

on a fixed schedule, and not too many riots occur surrounding mayoral races. So we are probably okay. Authors using instrumental variables often spend a large part of their paper arguing that their instrument is indeed exogenous and trying to disprove counterarguments like the ones above.

If both conditions are met, we use our instrument to isolate exogenous variation in police presence. In this case, exogenous variation is **change in police presence as a result of being in an election year**. Variation in police presence due to other factors may be endogenous to the model. We would like to filter out the first type of variation. Schematically:

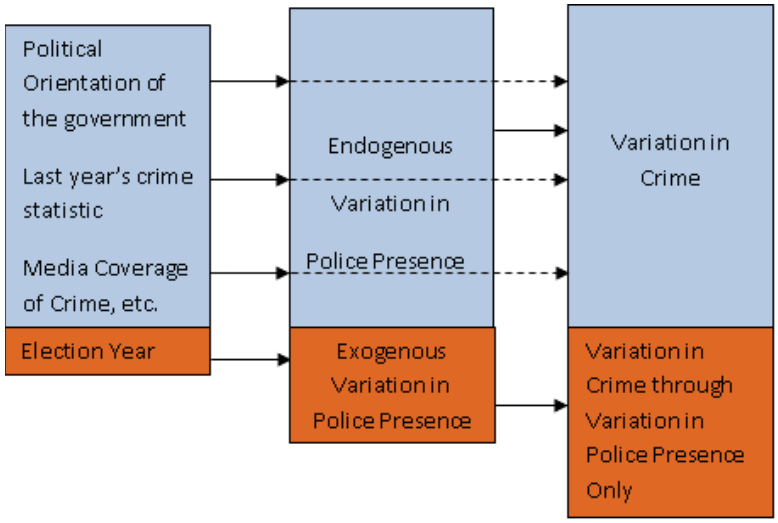

Figure 2.17

Note that an election year does not have to be the only, or even the most important determinant of the number of police officers. It just has to be a source of exogenous variation.

How does this work more technically? Instrumental variable estimation consists of two stages: the 'First Stage Regression' and the 'Second Stage Regression', together known as an IV regression or often a **Two-Stage Least Squares (2SLS)** regression.

In the first stage, we regress the number of police officers (our endogenous variable) on a dummy for election year (our instrument):

$\#\text{ police officers} = \alpha + \beta \text{ election year dummy} + \varepsilon$

If our instrument satisfies the relevance condition, the election year dummy and the number of police officers will be significantly correlated. We now know the size of the effect of being an election year on the number of police officers. We also know which year in our dataset was an election year. We can use this information to predict how many police officers there are in any given year. In this case: α for a non-election year and $\alpha + \beta$ for an election year. Call this predicted number of police officers '$\widehat{\#\text{ of police officers}}$'. Because '$\widehat{\#\text{ of police officers}}$' only contains variation in police presence due to the election year, we have isolated some exogenous variation in police presence.

Now we use this exogenous variation in police presence to estimate the IV regression. We estimate:

$$\text{crime rate} = \alpha + \beta \widehat{\#\text{ of police officers}} + \varepsilon$$

Note that we use the **predicted** number of police officers, not the **actual** number of police officers (which is endogenous). If the exclusion restriction is satisfied, this should allow us to estimate the *causal* effect of police presence on crime.

In terms of the abstract model, the two stages of a two-stage least squares model are:

- Stage 1: $x = \alpha + \beta z + \varepsilon$. We use this regression to construct 'predicted x' or , isolating exogenous variation in x.
- Stage 2: $y = \alpha + \beta \hat{x} + \varepsilon$. Estimating the causal effect of x on y.

To sum up, the instrumental variable approach can **potentially** solve endogeneity issues due to reverse causality and establish a causal relationship between two variables. We must be convinced, however, that the instrument is exogenous, reasonably correlated with the original (endogenous) variable, and that the exclusion restriction holds. Assessing these assumptions is key to assessing the credibility of a paper using instrumental variable estimation. The most important tool for doing so is your common sense!

Difference-in-difference estimation

Difference-in-difference estimation may be employed when we want to investigate the effect of some treatment on an outcome and we have information both over time and across space (or cross-section units) for both the treated and the untreated. In the context of development, this is often the effect of a particular aid programme (providing schools with textbooks, providing microcredit, etc.) on the well-being of the poor (pupil performance, income). Call the cross-sectional units receiving the treatment the **treated group**. We often know the score of the treated group on the outcome variable before and after the treatment.

For example, we know that pupils on average scored 210 points on a test before they received schoolbooks and 220 points on average after. The **difference** between pupil's scores before and after they received schoolbooks is 10 points. We might say that the schoolbook programme caused the pupils' scores on tests to increase by 10 points. However, we would be wrong in drawing such a causal conclusion: maybe something else changed during the time that the programme was running, causing performance of pupils to go up irrespective of the programme. Maybe the government has increased its budget on education and **all** children in the country (those receiving books **and** those not receiving books through the programme) score better on the test.

To draw credible causal conclusions, we need a **control group**. Ideally, the treatment and control group should have identical characteristics and differ only in whether or not they receive the treatment. The control group is our **counterfactual**: we believe that, **in absence of the treatment**, the outcome variable for the treatment and control group would follow the same trend. Any difference between the trends of the control group and the treatment group can then be attributed to the treatment.

In terms of our example, say that the schoolbook programme has been **randomised**. A group of potential recipient schools was selected, a lottery was held and only half of these schools received books. Pupils in the schools not receiving books form our control group. We also measure the performance of control group pupils before and after the programme: they score on average 220 points before and 228 points after the programme is executed. The **difference** between control group pupils' scores before and after programme execution is 8 points. The **difference** between the **differences** in test scores in the control and treatment group before and after the programme is 2 (10 – 8). (You may understand now why this method is called difference-in-difference estimation.) We conclude that the programme caused pupils' scores to rise by 2 points. For this conclusion to hold, we must credibly argue that the trends in test scores would be the same for the pupils receiving books and those not receiving books, **in absence of the programme**.

More technically, a difference-in-difference regression set-up will look like this:

$$\text{Outcome}_{ti} = \beta_1 + \beta_2\,\text{treatment}_i + \beta_3\,\text{post}_t + \beta_4\,(\text{treatment} * \text{post})_{ti} + \varepsilon_{ti}$$

We observe the outcome variable for an entity (i) in two time periods (t), before and after the treatment could have been received. 'Treatment' is a dummy indicating whether an individual entity has received the treatment and 'post' is a dummy indicating whether the observation is made before or after the treatment period.

Graphically, we could represent the interpretation of the coefficients in the following series of graphs:

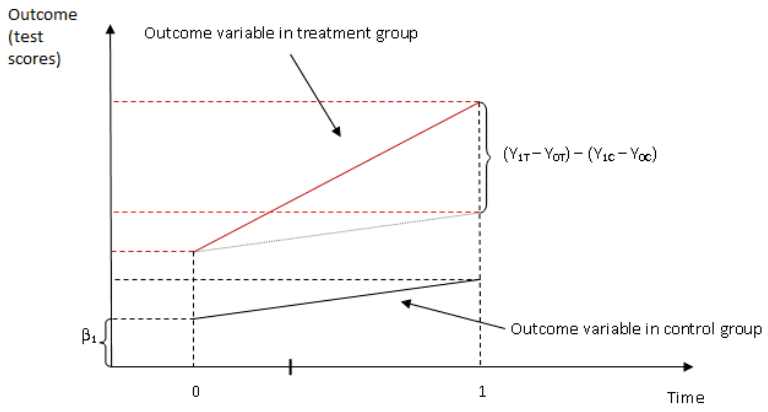

Figure 2.18

Note that the regression equation includes a dummy variable (treated). This allows for two separate regression lines, each with their own intercept. β_1 captures the effect of being in the non-treated group. It is the intercept term for the non-treated group.

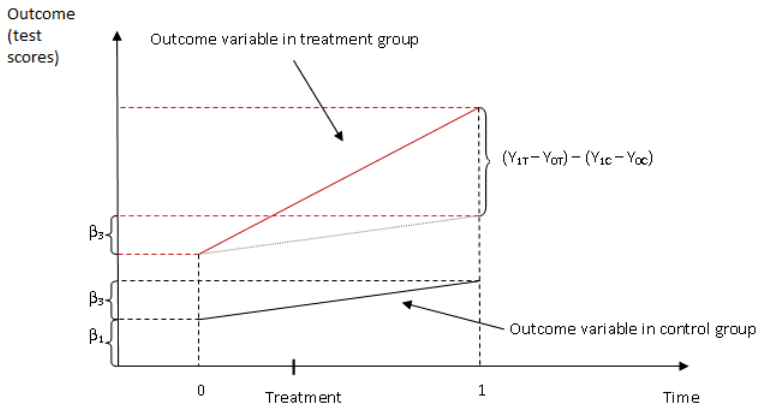

Figure 2.19

β_3 captures the effect of being in the post-treatment period. In absence of the treatment, the treatment and the control group should display the same trend in outcome. This allows us to draw a (hypothetical) line showing how outcome in the treatment group would have developed, in absence of the treatment.

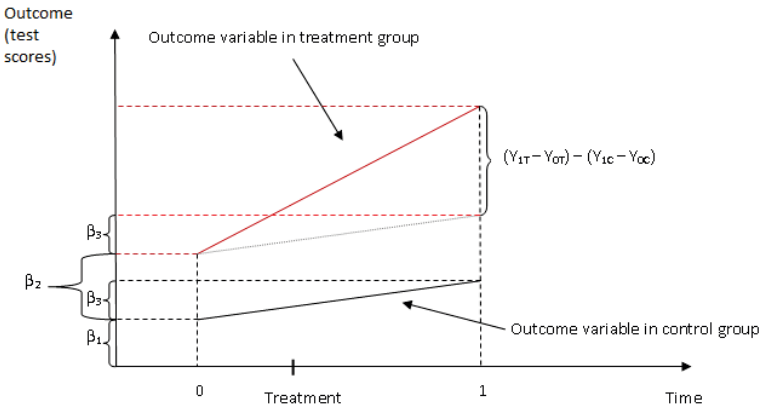

Figure 2.20

β_2 captures the effect of being in the treated group. It is the intercept for the regression line for the treatment group, minus the intercept of the regression line for the control group (β_1).

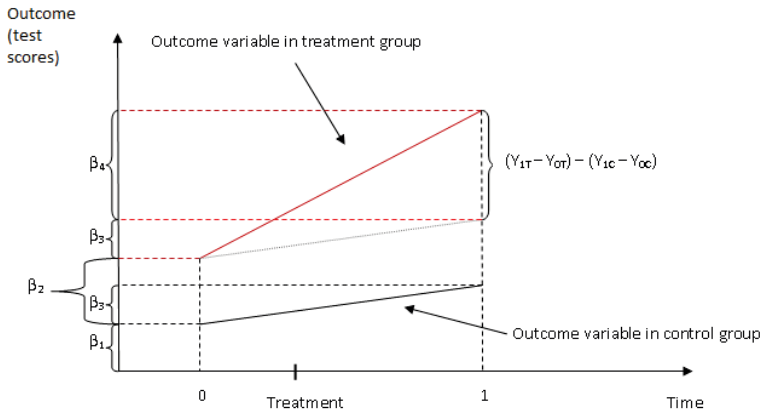

Figure 2.21

Finally, β_4 is our variable of interest. Note that this is the coefficient on an interaction term. It captures the additional effect of **the combination of** being in the treatment group *and* being in the post-treatment period. In other words, it captures the effect of having received the treatment.

Summary

Difference-in-difference estimation is an approach to estimate the causal effect of a policy on a treatment group. Its credibility, however, depends entirely on the validity of the key assumption and chosen control group: In the absence of treatment the difference in outcomes between treatment and control group would not have changed. Is it reasonable to assume that this assumption is true? An important way of assessing the validity of this assumption is to use your common sense. The researcher can also supply evidence that suggests that the key assumption is valid, for example, if the researcher can show a graph of the trend in the outcome before the treatment. If trends look similar, we may be inclined to think the research strategy is valid. Furthermore, a researcher may show that treatment and control group do not differ significantly in characteristics that may be relevant to the research, using certain statistical tests.

Now read

Ray, Debraj *Development Economics*. (Chichester: Princeton University Press, 1998) Appendix 2, pp.777–804.

Ray treats some of the material above in Appendix 2. Generally, he does so in a more technical way. If it helps you to see the same thing explained using a different approach, please use the material in Ray.

Again, if you do not fully understand the material yet, do not worry. You have just laid the basics of all the methodology you will need to learn for the course. It is only natural for you to need more practice to grasp it all. More practice will follow in each chapter from now on.

A reminder of your learning outcomes

Having completed this chapter, and the Essential reading and Activities, you should now be able to:

- recognise when the topics introduced in this chapter are applied in subsequent chapters and reading.

When you have completed the course you should be able to:

- read and critically assess empirical research employing the aspects of quantitative methodology covered in this chapter.

Sample examination questions

There will be no examination questions exclusively testing your knowledge of methodology. However, if you look at the examination questions in the subsequent chapters, you will realise that being able to apply the material in this chapter to the topics in subsequent chapters is necessary to get a good score in the examination.