Introduction

In this chapter we consider the many reasons for learning how to do regression. Regression, in its simplest form, is a way of summarising the relationship between two continuous variables. It does this by working out the equation of the straight line that best fits the data on the two variables in question. The simplest and most widely used method for computing this straight line is to find the equation that minimises the squared deviations of each observed point from this line. Hence, the nickname, “Ordinary Least Squares” (OLS) regression. This is the main method considered in this book, though other approaches are considered, particularly in the final chapter.

Why Study Regression?

To those new to the topic, regression analysis can seem difficult and elusive. Possible problems seem unlimited, diagnosis ambiguous and often tortuous, and model development can, in the end, seem more of an art than a science. Yet, regression persists as one of (if not *the*) most popular methods of analysing relationships between continuous variables. Whilst popularity is in itself a good reason to study the topic—journal articles, government-funded reports and academic texts increasingly assume that the reader has familiarity with the technique—the goal of this course is to do more than equip the reader to understand other people’s work. My hope is that by the end of the book you will be able to *do* regression. There are good reasons why regression is so popular and these reasons are worth considering for a moment since they may prove decisive in choosing the appropriate technique for your own research.

Regression is elegant

This claim may start to have a rather hollow ring to it once you have immersed yourself in the messiness of regression diagnostics. But, at heart, regression is an elegant, even beautiful, technique. It uses an astoundingly simple algorithm (minimum sum of squared deviations) to do a very complex job – namely, draw a hypothetical ‘line of best fit’ between a single dependent variable and any number of explanatory variables. Without breaking sweat, it can plot a surface in multi-dimensional space and derive the confidence intervals in every dimension under consideration. And OLS estimates have the wonderful quality of being BLUE – they are the Best Linear Unbiased Estimates. That is, it can be demonstrated mathematically that it is impossible to find better unbiased linear estimates provided certain basic assumptions are met. One only realises, I think, the power and simplicity of regression when one contemplates the alternatives. Other options are so complex as to be beyond the scope of a text such as this, and as a result students new to regression are deprived of this comparison. You will have to take my word for it when I say that, provided the underlying assumptions are met to a reasonable degree, it is highly unlikely that you will find an alternative technique that is simpler, faster and easier to interpret.

Regression controls for other determinants

A great deal of research, particularly where human subjects are involved, cannot be done under the controlled conditions of a scientific experiment. If I want to know the relationship between educational achievement and IQ, it is rather unlikely that I will be given the opportunity to remove several thousand children from the loving care of their parents, place them in an isolated control room for twenty years, and observe how their educational career pans out! From a research perspective, though, it would be an appealing thing to do, since it would allow me to compare "like with like". In order to isolate the effect of IQ on educational performance I need to strip out the effect of all other possible determinants of educational performance. Given the option, the simplest way to do this is choose subjects that vary only by IQ. That is, I will select children of a similar age, gender, race etc. and expose each to exactly the same living and social conditions for a prolonged period. I will then be able to conclude that any difference in educational performance is either random (i.e. without "systematic cause") or driven by IQ.

If it were not for regression analysis, social researchers would spend their idle moments wishfully pondering such experiments. (So perhaps we should, in the very least, thank regression analysis for saving thousands of children from clutches of desperate social scientists…). For what regression does, at least in principle, is give you an estimate of the effect of an explanatory variable (such as IQ) on the variable you are trying to explain (such as educational performance) while holding constant the effects of all the other variables you have included in the regression (age, gender, social class etc.).

At the risk of whipping you up into a state of frenzied excitement, I should add that regression will allow you to do even more than that. It will actually, as a matter of course, provide you with estimates of the effect of each of the other explanatory variables as well, in each case holding constant the effects of the remaining explanatory variables. In that sense, it is actually superior to the controlled experiment approach since it allows you to simultaneously consider the effects of multiple causes (whereas in the pure experiment approach, you can only consider the impact of one effect since your sample has been selected in such a way as to exclude all other effects). Unfortunately, regression can only hold constant variables included in the model—omitted factors have an unknown and potentially powerful effect. This is one of the main challenges for regression modellers: how to plausibly account for relevant drivers. Nevertheless, regression is a huge improvement on bivariate methods (such as correlation coefficients) which make no attempt at controlling for other effects.

Regression quantifies relationships

Implicit in the above discussion is the notion that regression 'measures' the relationship between variables. This is worth highlighting. For if you have just completed a course in basic inference, you will be aware that some tests of a relationship between variables do not give you an estimate of the strength of that relationship; only whether or not the relationship is likely to exist were we to have the privilege of examining data on the population as a whole. What we would really like to know, however, is not just whether educational performance is affected by IQ (that is almost certain) but to what extent? How *sensitive* is the dependent variable to a particular determinant? If we could quantify this effect we could distinguish between minor causes and major ones. We could also use the estimates to predict the effect of a change in policy or circumstances. We could even estimate the value of the dependent variable for someone outside our sample, so long as we have data on each of the explanatory variables for them.

Regression is an established technology

After you have completed this course and become more familiar than you would have liked with all the possible violations of the assumptions that under-gird regression, you will wonder whether regression actually tells you anything meaningful at all about the relationship between variables. Bear in mind, however, during that moment of personal crisis, the reason why we are able to consider these violations is that tried and tested techniques have been developed to test for and remedy them. These tests and solutions are not perfect, but remember that equally severe violations are possible in other techniques, it is just that often no adequate (or accessible) diagnosis or remedy are available. Regression is so widely used, and so well established, that just about every aspect of it has been explored at length. It is a tried and tested technology.

Regression can be used to predict and simulate

One of the reasons regression is so widely used is that it can be employed to make predictions and to simulate the outcome of changes to particular determinants. Suppose, for example, one builds a regression model of crime and find that the rate of unemployment is a key driver. Having estimated the relationship between crime and unemployment you could then simulate what would happen to crime rates if the government were to intervene to bring down unemployment. Or suppose you build a regression model that estimates the relationship between sea levels and the concentration of CO_2 in the atmosphere. You could then forecast what sea levels would be in 100 years time if carbon dioxide emissions continue to rise at their current rate. This facility makes regression a very practical tool for social, economic, financial, and scientific prediction and simulation.

Regression is fast

Because of the simplicity of the algorithm used to compute a regression equation, it is lightning-fast in computational terms when compared with, say, maximum likelihood techniques. This is because there is what mathematicians call an ‘analytical solution’ to the regression equations. More complex techniques have no such solution and rely on computationally intensive ‘numerical methods’ (brute force searches over millions of iterations) to find the optimal line of best fit. Even with the burgeoning power and speed of computer technology, when you have a large data set and many thousands of observations, regression can save you a great deal of time, particularly when you have hundreds if not thousands of models to experiment with.

Regression is an excellent starting point

Because of its speed and simplicity, even if it is not the technique of choice, regression is a great place to start. In fact, it has become the benchmark. So when statisticians or researchers develop or apply a more sophisticated technique, they will often also present the results of regression analysis. It is often surprising, however, just how small the improvement the more complex technique affords over the ordinary least squares (OLS) regression. I have despaired myself, having spent months researching and applying some elusive specialist method, only to find that it provides negligible improvement (and sometimes not even that!) over my first regression estimate… Also, more advanced methods often operate along similar lines to regression with similar or related intuition, so its worth getting your head round basic regression even if your ultimate goal is to master a more sophisticated technique (indeed, many advanced texts assume that you understand regression).

It is also worth noting that the violations of the regression assumptions, which we shall spend so much time considering, are themselves a very useful tool in the hands of the intelligent researcher. For they provide a lead on where to go next in the investigative process. If I can pinpoint which assumption is violated by my data, I can search the appropriate literature for solutions to that problem. So in a sense, the regression assumptions and associated violations provide a structure to one's research strategy. They may also allow you to critique and revise established findings in the literature, providing you with the breakthrough you need to further your career.

Overview of Contents and Style

The aim of this text is to introduce students to he finer points of regression analysis, and to do so using the widely used software package, SPSS. Whilst there are a range of texts that cover regression analysis in SPSS, few currently explore the topic in any great depth. This may be because SPSS has limited built-in diagnostic test procedures for regression analysis compared with other packages such as Stata.

The lack of pre-packaged diagnostic tests can be overcome by making use of the powerful Matrix syntax facility in SPSS. Whilst these syntax routines seem more cumbersome than running canned procedures in other packages, for the student learning regression analysis for the first time, I believe they have a distinct advantage over their automated equivalents in that they encourage students to become familiar with the rationale of tests and solutions. As such, they make a far stronger connection between what is presented in econometrics texts and the actual interpretation of the output from diagnostic tests and solutions. Where the syntax is opaque and unintelligible to the reader, little is lost since it is usually only the first few lines of a routine that need to be amended for it to be applied to a particular problem. Moreover, a great benefit of such syntax is that they allow much more flexibility than canned procedures which are often misapplied without the user knowing (because a different formula should be used for small sample sizes, for example).

For keener students, familiarity with matrix routines will open up the possibilities of developing their own syntax. This is surely a powerful advantage since anyone serious about becoming a social or business statistician is likely to need to learn how to write their own syntax routines. At some point, you are likely to find that the specific test or correction method needed for a particular problem is not available in any current software program (or available in another program with which you are not familiar or would prove costly to purchase). Also, programming skills are fairly transferable across software packages, as the general structure and method of a syntax routine is often quite similar.

Structure

We begin by looking at covariance, correlation and regression in chapter 1, followed by prediction and analysis of variance in chapter 2. These topics set the scene for later topics, such as multicolinearity, which is explored later. Non-linear relationships and their implications are introduced in chapter 3, followed by F-tests and structural break tests in chapter 4 which, in applied research, are often closely related (if there is a structural break, can a non-linear transformation be found to heal the fracture, or must the sample be split?). The issue of omitted and irrelevant variables is introduced in chapter 5, followed by heteroscedasticity and multicolinearity in the subsequent two chapters. Chapter 8 introduces logistic regression, the first step in understanding how to model categorical dependent variables.

Exercise:

You probably won't be able to answer these questions in full at this stage – they are intended to get you thinking about what you have read so far and some of issues that motivate the topics that lie ahead:

1. Describe in your own words what you think the advantages of regression analysis.
2. What do you think the limitations of regression might be?
3. What is the intuition behind regression estimation?
4. What does OLS stand for? What does it mean?
5. What do we mean when we say OLS estimates are BLUE?
6. Find out what statisticians mean when they say:
    a. an estimate is "unbiased"
    b. an estimate is "efficient"
7. What does Achen (1982 p.9) mean when he says that, "Nowhere is the gap between practice and its justification so large as in social use of statistical methods?" How does he justify this claim? To what extent do these arguments apply to your own discipline today?

Further Reading:

Section 1 "Introduction" of Achen, C. H. (1982) *Interpreting and Using Regression*, A Sage University Paper, Quantitative Applications in the Social Sciences No. 29, Sage Publications: London, pages 7-12.