Econometrics, Chapter 1 Outline

In this chapter we will discuss curve fitting and introduce the regression model. Justification for the least squares procedure is provided.

1 Curve fitting

- Types of data

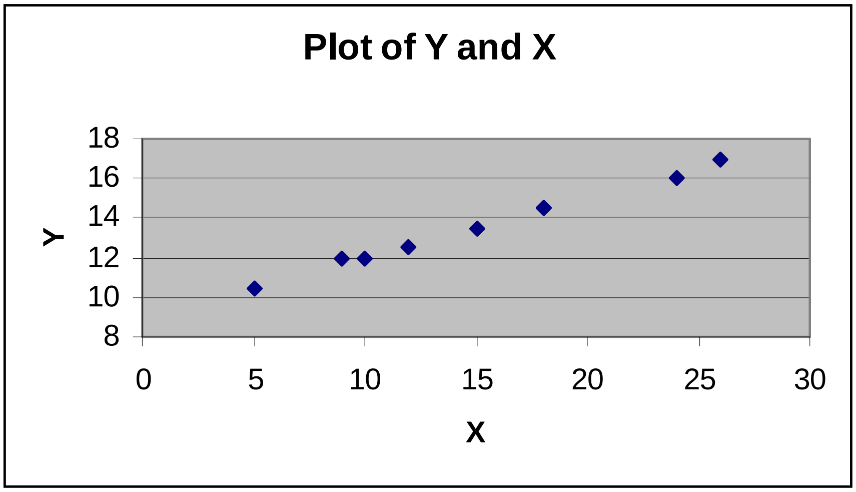
1. time series – describes the movement of a variable over time
2. cross-section – looks at the individual characteristics of a firm or individual at a point in time
3. pooled data – combination of time series and pooled data

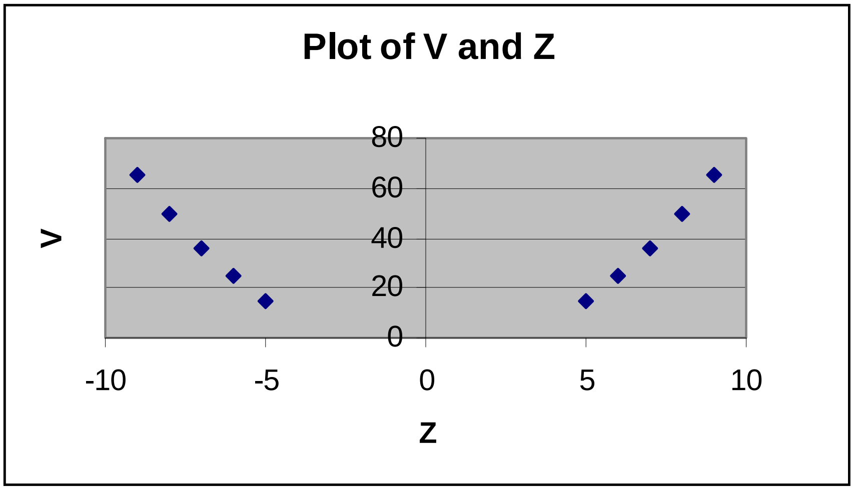
Suppose we hypothesize that there is a relationship between two variables, X and Y. Economic theory would tell us what we should expect this relationship to be, but we will use econometrics to estimate this relationship. In order to estimate this relationship, we need data. If we were to have every possible observation on a variable, we would then have the population of that variable. However, most of the time we will have a sample of the available data. Thus, we will need to perform statistical tests to see if the estimates we obtain are significant (we will get to that later in the course). For now, suppose we have two variables, X and Y. I have created some numbers for X and Y and placed them in the table below; the scatterplot shows the graphical relationship between X and Y. I have also created two more variables, V and Z, and created a table and a scatterplot for them as well.

Y	X
13.5	15
12	9
14.5	18
12.5	12
10.5	5
17	26
14.5	18
12	10
16	24

V	Z
49.6	8
24.4	6
64.9	-9
24.4	-6
14.5	5
36.1	7
36.1	-7
64.9	9
49.6	-8
14.5	-5

When we attempt to fit "curves" to the data, what we are actually going to do is attempt to fit a straight line to the data. This is where the term "linear" comes into play in our analysis. In some cases, attempting to fit a straight line through the data may not be the best option. Looking at X and Y above, it seems that a straight line might yield a good approximation of the relationship between X and Y, but it does not look like a straight line (we are only going to try to fit one line to ALL of the data points) will yield a very good approximation of the relationship between V and Z. You should be aware that


Plot of Y and X
18
16
14
12
10
8
Y
0
5
10
15
20
25
30
X


Plot of V and Z
80
60
40
20
0
V
-10
-5
0
5
10
Z

sometimes a linear approach is NOT the best approach to approximating the relationship between two variables – however, we will focus on linear regression techniques.

1.1 What line to fit?

We can attempt to fit an infinite number of lines through the X and Y data. We could connect the lowest point (5,11) and the highest point (26,17). We could try to draw a line in by hand that looks like it will fit and then try to measure the slope and intercept by hand. However, the method that we will use is the **least squares method**. We use the least squares method to find the line of best fit. The line of best fit is defined as the line which minimizes the sum of the squared (vertical) deviations of the points of the graph from the points of the straight line that we choose. So what we do is draw a line through the data, measure how far the data point is vertically from the line (that is the deviation), and square that value (that is the squared deviation). We do this for each data point and then sum the squared values. This gives us our "sum of squared deviations". What we would then do is draw a different line through the data and find the sum of the squared deviations of that line – if the sum of squared deviations of the second line was lower than the sum of the squared deviations of the first line, then the second line would be a better fit than the first line. Of course, we could draw a third line and repeat the process and see if it has a lower sum of squared deviations than the second line. Hopefully you get the idea.

1.2 Why use least squares?

We could use other methods of trying to find the line of best fit. Two alternative criteria that the book proposes is using the sum of the deviation values themselves (NOT squaring them) and using the sum of absolute value of the deviations. When using the sum of the deviations we would want to try to get the sum as close to zero as possible. One reason that we do NOT want to use the sum of the deviations without squaring them has to do with the following example. Suppose we have two data points. The X value of both data points is 10. The Y value of the first data point is 17 and the Y value of the second data point is 7. Clearly, the line that best approximates this relationship is a vertical line at $X = 5$. However, ANY line that passes through the point (5,12) will have a sum of deviations equal to zero, since one data point will be 5 units above the line and the other will be 5 units below the line. Thus it may be possible to find a line that has a sum of deviations equal to zero that does not give a good approximation of the data. As has also been suggested, we could try to minimize the sum of the absolute value of the deviations, as this would eliminate the problem of having two data points cancel each other out (since we are summing only positive values). There are two reasons we do not use absolute value. The first is that using the sum of the absolute value of the deviations puts less weight on a data point that is very, very far away from the line than the least squares method does while it puts more weight than least squares on data points that are fairly close to the line. The second reason that we use the least squares method rather than the absolute value method is mathematical. For those of you that have had calculus, whenever you see "minimize" or "maximize" you should think derivative. If you recall what the absolute value function looks like, it has a kink in it (it is not smooth), which makes it nondifferentiable, which makes it a mess to work with. The least squares method also has the added bonus in that it permits statistical testing of the estimates of the slope and intercept that we will obtain.

2 Derivation of least squares

Note: Before reading this section you should be familiar with the use of summation operators. See the notes on the appendix to chapter 1 for a review.

2.1 Proposing a model

After we have looked at our data, we need to propose a model. As I have said, we will work with LINEAR models in this class. Most of our models will look like:

$Y = \alpha + \beta X$

CAUTION: You should take note of one thing that the authors of the book do. They will use x (lowercase) and X (capital) to mean two DIFFERENT things. Also, y (lowercase) and Y (capital) mean DIFFERENT things. The lowercase letters mean that the variables are measured in "deviations form". I will explain what that means when we get there. The capital letters just refer to the actual values for X and Y. I bring this up because I was working one of the problems at the end of one of the chapters and I forgot that they did this and I spent about 2 hours on the problem before I realized that the reason I couldn't get the right answer was because I was using capital X's when the book had used lowercase X's

Notice that when we write down our model, we put Y on the left-hand side and X on the right-hand side. We have, in effect, decided that X has an influence on Y. We call X (or, more generally, the right-hand side variable) the independent variable because we are assuming it is NOT influenced by anything. We call Y the dependent variable because we are assuming that X helps determine Y.

2.2 Finding the least squares estimates of α and β

Recall that we want to minimize the sum of squared deviations from our line. How do we write the sum of squared deviations mathematically?

$$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

The Y_i is just the actual data value for each Y. The $\hat{Y}_i$ is our predicted value for Y_i which is based on the line we drew. The letter i is the index for our summation notation, and $\sum$ tells us to some up all the squared deviations from 1 to N, where N is the number of observations (data points) we have. Our goal is to minimize this sum of squared deviations. Note that the lowest sum of squared deviations you can ever have is zero since we are adding together numbers that must all either be positive (because we are squaring numbers) or zero.

In order to minimize the sum of squared deviations $\left(\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2\right)$ we first need to plug in for $\hat{Y}_i$. What can we substitute in for $\hat{Y}_i$? Since $\hat{Y}_i$ is our predicted value of Y_i, we know that $\hat{Y}_i$ will be given to us by the equation of our line. So $\hat{Y}_i = \alpha + \beta X_i$, where X_i is the X value that corresponds to the Y value. So we plug in $\alpha + \beta X_i$ for $\hat{Y}_i$. We now have our sum of squared deviations as:

$$\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i)^2$$

Recall that to minimize a function we need to take the derivative and set it equal to zero. But just what are we taking the derivative of? Well, we have two unknowns, α and β, that we are trying to estimate, so we need to take the derivative of our function with respect to α and also with respect to β. So we need two derivatives. Actually, we will take partial derivatives, which are denoted by ∂ rather than total derivatives, which would be the normal dy/dx stuff most people are probably used to. Partial derivatives are easy to take – they just assume that other variables in the equation are constants. For example, if we take the partial derivative of our sum of squared deviations with respect to α, we just treat β as if it were a constant. So:

$$\frac{\partial}{\partial\alpha}\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i)^2 = -2\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i)$$

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i)^2 = -2\sum_{i=1}^{N}X_i(Y_i - \alpha - \beta X_i)$$

Now, set both equations equal to zero, and solve for α and β.

$$-2\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i) = 0$$

$$-2\sum_{i=1}^{N}X_i(Y_i - \alpha - \beta X_i) = 0$$

The book does a pretty decent job of explaining how to solve for α and β on page 17, and I urge you to attempt to solve for α and β on your own (it will give you practice using the summation rules in the appendix and hint: solve for β first) and then use the book if you get stuck. I, however, will just skip to the answers:

$$\alpha = \frac{\sum_{i=1}^{N}Y_i}{N} - \beta\frac{\sum_{i=1}^{N}X_i}{N}$$

$$\beta = \frac{N \sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N \sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i \right)^2}$$

Notice that β is written only in terms of X_i and Y_i, so we can calculate β directly from the observed data. As for α, notice that it includes a β in its solution as well as X_i and Y_i. This is fine since we know that β only consists of X_i and Y_i. We could plug in the formula for β into the formula for α so that we would just have X_i and Y_i in the formula for α, but this would lead to a very messy formula for α. We can now obtain our least squares estimates for α and β, and find the estimated model.

After the book gives you the two formulas for α and β, it discusses "deviations form". Deviations form just means that the each observation of the variable has had its mean subtracted from it. So, instead of using X we would use x, where $x_i = X_i - \bar{X}$, where $\bar{X}$ is the mean of X. I personally do not like deviations form as it obscures some things, but that is precisely why the authors use it, to make the math easier. I will try NOT to use deviations form in class since my board written x and X will probably look the same.

You should now be able to compute the least squares estimates for the data that I have created above.