

1 Fitting the Data

- A look at the philosophy and mathematics behind estimation. How to fit the data to the relationship between education and earnings.
- Notice that in the raw data we can only ‘eye-ball’ the difference; conditional means only consider one year against another - there is also the problem of conditioning on other variables (such as age).
- Print outs given in L2_79.log and L2_95.log. Regression read-out for the relationship between education and earnings. To note: descriptive statistics to check that we are using the correct data; estimate using the ‘raw’ figures; estimate using a log conversion of the data. Relationship is log linear to aid the linearization of the model and to give a ‘reading’ of a percentage change.
- Concerned with minimum distance methods of estimation - this case is called Ordinary Least Squares (OLS), or ‘least squares’ for short.
- Estimation of a straight line: for $i = 1 \dots n$ individuals in the cross-section data set

$$\widehat{Y}_i = a + bX_i$$

- Look at the difference between a straight line and the data plot. For each and every point we suppose that there is a deviation between the observation Y and the predicted value of Y from the straight line \widehat{Y} . We can write the deviation of Y from \widehat{Y} as:

$$e_i = Y_i - \widehat{Y}_i$$

- Note that X can vary, there is a corresponding value of Y , but a determined value \widehat{Y} , given the value of a and b .
- Look at the example: L2_ex1.log; L2_ex1.xls. Although only 5 observations on X and Y , nonetheless gives all you need to understand using the larger data sets: L1_79.xls and L1_95.xls.
- As $\sum_{i=1}^n e_i = 0$, work with the sum of squared residuals: $\sum_{i=1}^n e_i^2$.
- How to estimate the parameters of the straight line for the data contained in L2_ex1.xls, or in L1_79.xls and L1_95.xls.
- Divide the problem into 2 parts: consider the intercept (a) then the slope parameter (b).
- Intercept (a): differentiate function ($e_i = Y_i - a$) with respect to a :

$$g(a) = \sum_{i=1}^n e_i^2$$

- Note solution: an important property of OLS estimators - on average, the sum of the error terms is zero. You have an unbiased estimate.

$$\sum_{i=1}^n e_i = 0$$

- Which gives:

$$a = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}$$

- In the absence of a slope parameter, the best point estimator of the dependent variable (Y) is the unconditional sample mean \bar{Y} .
- Slope coefficient (b): differentiate function ($e_i = Y_i - bX_i$) with respect to b :

$$g(b) = \sum_{i=1}^n e_i^2$$

- Solution set at zero gives:

$$\sum_{i=1}^n X_i e_i = 0$$

- Important property required for OLS estimation - correlation of independent stochastic variable (X) and the error term (e) is zero on average.
- Substitute solution into original expression ($e_i = Y_i - bX_i$) to get:

$$b = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

- Gather the 2 conditions together and use to solve $e_i = Y_i - (a + bX_i)$ in terms of X and Y : Use the 2 solutions from the differentiation.
- For a : use $\sum_{i=1}^n e_i = 0$ to give:

$$\sum_i Y_i = na + \sum_i bX_i$$

- Solving gives a as:

$$a = \bar{Y} - b\bar{X}$$

- For b : use $\sum_{i=1}^n X_i e_i = 0$ to give:

$$\sum_i X_i Y_i = \sum_i X_i a + \sum_i X_i^2 b$$

- Use the solution to a to solve for b :

$$b = \frac{(\sum_i X_i Y_i - n \overline{XY})}{(\sum_i X_i^2 - n \overline{X}^2)}$$

- Use to check results on L2_ex1.xls.
- Today's lecture has been concerned with:
 1. Understanding why 'minimum distance' estimation gives you an unbiased estimate of the relationship between 2 variables.
 2. The important results from the optimization are: $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n X_i e_i = 0$ to produce the formulas for the OLS estimates of parameters a and b .
 3. Commit to memory the formulas for a and b and the means by which they are derived.