Lecture 2
Costas Meghir

- We return to the classical linear regression model to learn formally how best to *estimate* the unknown parameters. The model is

$$Y_i = a + bX_i + u_i$$

- where *a* and *b are the coefficients to be estimated*

Assumptions of the Classical Linear Regression Model

- **Assumption 1:** $E(u_i \mid X) = 0$ The expected value of the error term has mean zero *given any value of the explanatory variable*. Thus observing a high or a low value of *X* does not imply a high or a low value of *u*.

  X **u**nderline: ***X*** and **u** are **uncorrelated.**

- This implies changes in X are not associated with changes in u in any particular direction - Hence the associated changes in *Y* can be attributed to the impact of *X*.
- This assumption allows us to interpret the estimated coefficients as reflecting causal impacts of *X* on *Y*.
- Note that we condition on the *whole* set of data for *X in the sample not* on just one X_i.

- **Assumption 2**: HOMOSKEDASTICITY (Ancient Greek for Equal variance)

$$Var(u_i \mid X) \equiv E(u_i - E(u_i \mid X) \mid X)^2 = E(u_i^2 \mid X) = s^2$$

where s^2 is a positive and finite constant that ***<u>does not depend on X</u>***

- This assumption is not of central importance, at least as far as the interpretation of our estimates as causal is concerned.
- The assumption will be important when considering hypothesis testing
- This assumption can easily be relaxed. We keep it initially because it makes derivations simpler

- **Assumption 3**: The error terms are uncorrelated with each other.

$$\operatorname{cov}(u_i, u_j \mid X) = 0 \quad \forall i, j, \quad i \neq j$$

- When the observations are drawn sequentially over time (time series data) we say that there is *no serial correlation* or *no autocorrelation*.
- When the observations are cross sectional (survey data) we say that we have *no spatial correlation.*
- This assumption will be discussed and relaxed later in the course.

- **Assumption 4**: The variance of *X* must be non-zero.

$$Var(X_i) > 0$$

- This is a crucial requirement. It states the obvious: To identify an impact of *X* on *Y* it must be that we observe situations with different values of *X*. In the absence of such variability there is no information about the impact of *X* on *Y*.
- **Assumption 5:** The number of observations *N* is larger than the number of parameters to be estimated.

Fitting a regression model to the Data

- Consider having a sample of N observations drawn randomly from a population. The object of the exercise is to *estimate* the unknown coefficients a and b from this data.
- To fit a model to the data we need a method that satisfies some basic criteria. The method is referred to as an **estimator**. The numbers produced by the method are referred to as **estimates**; i.e. we need our estimates to have some desirable properties.
- We will focus on two properties for our estimator:
  - Unbiasedness
  - Efficiency [We will leave this for the next lecture]

Unbiasedness

- We want our estimator to be unbiased.
- To understand the concept first note that there actually exist *true* values of the coefficients which of course we do not know. These reflect the true underlying relationship between *Y* and *X*. We want to use a technique to estimate these true coefficients. Our results will only be *approximations* to reality.
- An unbiased estimator is such that *the average of the estimates, across an infinite set of different samples of the same sizeN, is equal to the true value*.
- Mathematically this means that

$$E(\hat{a}) = a \qquad and \qquad E(\hat{b}) = b$$

where the ^ denotes an estimated quantity.

An Example

	$\hat{b}$	$\hat{a}$
Sample 1	1.5841877	1.2185099
Sample 2	2.5563998	.82502003
Sample 3	1.3256603	1.3752522
Sample 4	2.1068873	.92163564
Sample 5	2.1198698	1.0566855
Sample 6	1.8185249	1.048275
Sample 7	1.6573014	.91407965
Sample 8	2.9571939	.78850225
Sample 9	2.2935987	.65818798
Sample 10	2.3455551	1.0852489
Average across samples	**2.0765179**	**.9891397**
Average across **500** samples	**2.0049863**	**.98993739**

Each sample has 14 observations in all cases (*N*=14*)*

True Model: $Y_i = 1 + 2X_i + u_i$ Thus *a=1* and *b=2*

Ordinary Least Squares (OLS)

- The Main method we will focus on is OLS, also referred to as Least squares.
- This method chooses the line so that sum of squared residuals (squared vertical distances of the data points from the fitted line) are **minimised**
- We will show that this method yields an estimator that has very desirable properties. In particular the estimator is **unbiased** and **efficient (**see next lecture**)**
- Mathematically this is a very well defined problem:

$$\min_{a,b}\{S = \frac{1}{N}\sum_{i=1}^{N} u_i^2\} = \min_{a,b} \frac{1}{N}\sum_{i=1}^{N}(Y_i - a - bX_i)^2$$

First Order Conditions

$$\frac{\partial S}{\partial a} = -\frac{2}{N}\sum_{i=1}^{N}(Y_i - a - bX_i) = 0$$

$$\frac{\partial S}{\partial b} = -\frac{2}{N}\sum_{i=1}^{N}[(Y_i - a - bX_i)X_i] = 0$$

This is a set of two simultaneous equations for *a* and *b*. The estimator is obtained by solving for *a* and *b* in terms of means and cross products of the data.

The Estimator

- Solving for *a* we get

$$\hat{a} = \overline{Y} - \hat{b}\hat{X}$$

where the *bar* denotes sample average

- Solving for *b* we get that

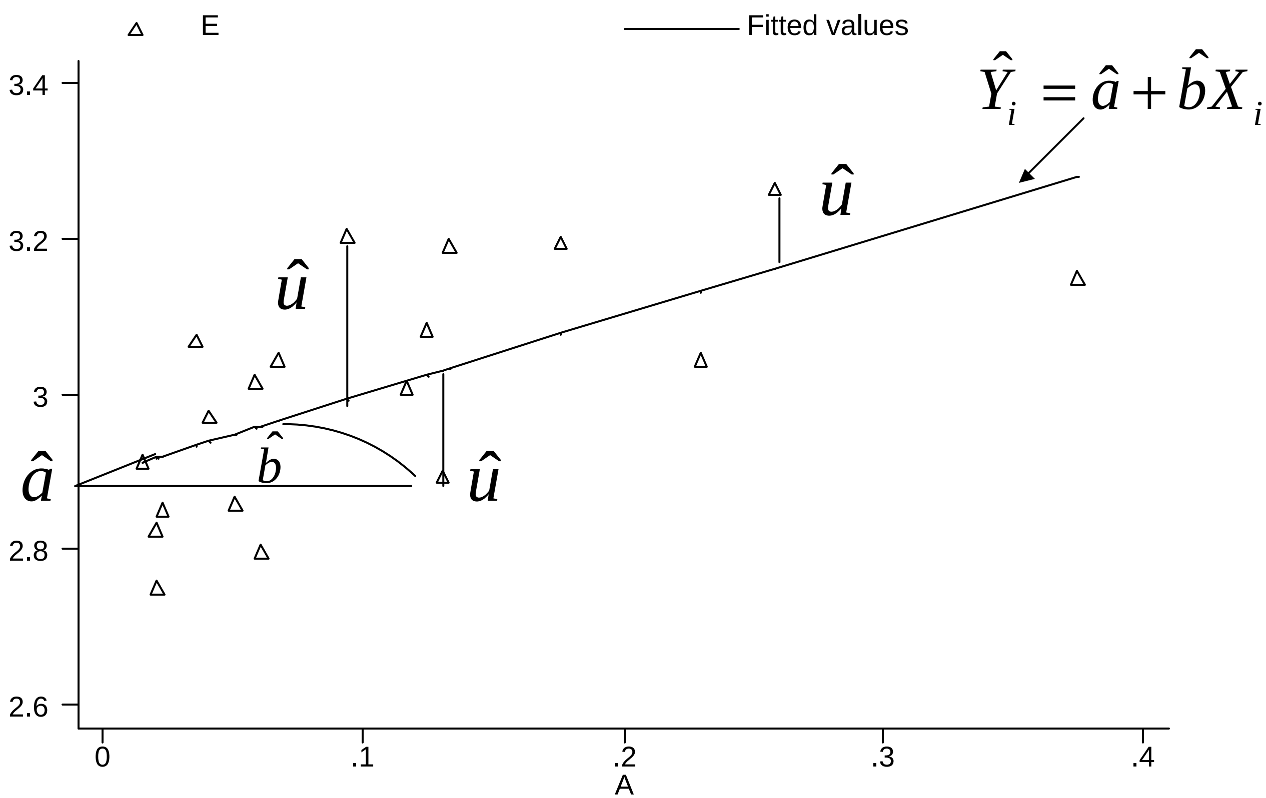
$$\hat{b} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N} (X_i - \overline{X})^2}$$

- Thus the estimator of the slope coefficient can be seen to be the the ratio of the covariance of X and Y to the variance of X
- We also observe from the first expression that the regression line will always pass through the mean of the data
- Define the *fitted values* as

$$\hat{Y}_i = \hat{a} + \hat{b}X_i$$

- These are also referred to as *predicted values*
- *The residual* is defined as

$$\hat{u}_i = Y_i - \hat{Y}_i$$



The Fitted Regression Line

Deriving Properties

- First note that within a sample $\bar{Y} = a + b\bar{X} + \bar{u}$

- Hence

$$Y_i - \bar{Y} = b(X_i - \bar{X}) + (u_i - \bar{u})$$

- Substitute this in the expression for *b* to obtain

$$\hat{b} = \frac{\sum_{i=1}^{N} \left[b(X_i - \bar{X})^2 + (X_i - \bar{X})(u_i - \bar{u}) \right]}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

Properties continued

Hence this leads to

$$\hat{b} = b + \frac{\sum_{i=1}^{N}(X_i - \overline{X})(u_i - \overline{u})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

The second part of this expression is called the sample or estimation error. If the estimator is unbiased then this error will have expected value zero.

Unbiasedness - We will use Assumption 1 only for this proof

$$E(\hat{b} \mid X) = b + E\left[\frac{\sum_{i=1}^{N} (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \mid X \right] =$$

$$b + \left[\frac{\sum_{i=1}^{N} (X_i - \bar{X}) E\{(u_i - \bar{u}) \mid X\}}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right] = b + \left[\frac{\sum_{i=1}^{N} (X_i - \bar{X}) \times 0}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right] =$$

$$b$$

Finally note that since $E(\hat{b} \mid X) = b$ for <u>any X</u> it must be that $E(\hat{b}) = b$

Goodness of Fit

- We measure how well the model fits the data using the R^2.
- This is the ratio of the <u>*explained sum of squares*</u> to the <u>*total sum of squares*</u>
- Define the Total sum of Squares as $TSS = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$
- Define the explained sum of Squares as

$$ESS = \sum_{i=1}^{N}\left[\hat{b}(X_i - \overline{X})\right]^2$$

- Define the residual sum of Squares as

$$RSS = \sum_{i=1}^{N}\left[Y_i - \hat{a} - \hat{b}X_i)\right]^2 = \sum_{i=1}^{N}\hat{u}_i^2$$

- Then we define $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

- The R^2 is a measure of how much of the variance of Y is explained by the regressor X.
- The R^2 computed following an OLS regression is always between 0 and 1.
- A low R^2 is not necessarily an indication that the model is wrong - jus that the included X has low explanatory power.
- The key to whether the results are interpretable as causal impacts is whether the explanatory variable is uncorrelated with the error term.

An Example - The price elasticity of Butter Purchases
Regression of log butter purchases on log price

```
. regr lbp lpbr

     Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  1,    49) =   49.61
      Model |  .317655914     1  .317655914           Prob > F      =  0.0000
   Residual |  .313752725    49  .006403117           R-squared     =  0.5031
-------------+------------------------------           Adj R-squared =  0.4929
      Total |  .631408639    50  .012628173           Root MSE      =  .08002

------------------------------------------------------------------------------
        lbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log price |  -.8421586   .1195669    -7.04   0.000    -1.082437   -.6018798
      _cons |    4.52206   .1600375    28.26   0.000     4.200453    4.843668
```

- --