ECON-122 Introduction to Econometrics

Agnieszka Postepska

July 13-14, 2011

Algebraic vs. statistical properties of an estimator

- algebraic properties hold by construction and they hold in any sample that we draw - pure math
- now we return to the population model - our estimators for β_0 and β_1 are random variables -their values (estimates) depend on the sample we draw
- if the estimator is a random variable we can look at the properties of its distribution over different random samples from the population

Statistical properties - introduction

- in order to establish properties of an estimator, we need to have *repeated samples* from the population - knowledge of the way in which observations are generated is crucial because the estimator will only have certain properties if the observations are generated in the desired way - so in some estimating situations and NOT in all
- this is why econometricians developed so many different estimators so that researchers can draw inference in so many different estimating scenarios - think of textbooks as catalogs of all these different estimators
- now we'll learn how such a catalog is structured

Introduction to the four assumptions

- As we already learned some structure of the basic linear regression model and we will get to know it very well we will build the road map of how to use these catalogs around this model
- in many standard estimation situation, OLS is the optimal estimator
- the model consists of assumptions concerning the way in which data is generated
- by changing these assumptions we are creating new estimating situations, in many of which OLS is no longer the desirable estimator
- it is crucial to know and understand these assumptions and their relevance for practice as many of econometric problems (and all of the problems we will study in this class) can be characterized as situations in which one (or more) of these assumptions is violated in a particular way

Cook book procedure for using the catalog:

- model the estimation situation in a general linear regression model
- pinpoint the ways in which the assumptions of the basic regression model are violated
- look up a textbook whether OLS estimator retains its properties under observed violation
- if YES proceed with OLS, if NOT find out what the alternative is
- if not certain whether assumptions are met - use one of the test also provided in textbooks

It is vital to use the appropriate estimator - there is no reason why a statistical package will not return you the estimates if the assumptions are violated -the problem is that these numbers are meaningless if the assumptions based on which properties of an estimator are derived, are violated.

The four assumptions

- A1: Linearity in parameters $y_i = \beta_0 + \beta_1 x_i + u_i$
- A2: Random Sampling - everybody from the underlying population has the same chance of being sampled; observations are iid
- A3: Sample variation in the explanatory variable
- A4: Zero conditional mean: $E(u|x) = 0$
- A5: Homoskedasticity: $Var(u|x) = \sigma^2$ - variance of the error term is constant given any value of the independent variable

What are those statistical properties of the estimators that we worry about so much and why?

- **unbiasedness**: the mean of the sampling distribution of the parameter is equal to the true value, i.e., $E(\hat{\beta}) = \beta$
- **efficiency**: smallest variance of the sample distribution of the parameter

Very often in econometrics we can find many unbiased estimators - in this setting we choose the one with the smallest variance.

Sample distribution of the estimators

- suppose you are taking 2000 repeated samples
- from every sample you compute the estimate
- as the samples differ, the estimates are different too
- the manner in which these estimates are distributed is called *sample distribution* of the estimated coefficient
- it is simply the probability density function of the estimates
- get the 2000 estimates and construct a histogram - use this histogram to approximate the frequencies of different estimates
- understanding the concept of sampling distribution is crucial in understanding econometrics - estimators are said to be "good" if their sampling distribution have "good" properties: unbiasedness and small variance are the two properties of the sampling distribution

Unbiasedness of an estimator in detail

- Def: the mean of the sampling distribution of the parameter is equal to the true value, i.e., $E(\hat{\beta}) = \beta$
- in other words: an estimator is said to be unbiased if in a situation in which we could take repeated sample an infinite number of times, we would get the correct estimate (true value of the parameter) "on average"
- in practice we only have one sample and that is why we cannot test whether it is the case or not
- we will see below that unbiasedness of OLS is derived using the assumptions we have seen before
- if these assumption do not hold, we cannot be certain that our estimator is unbiased
- information on how data was generated is crucial in constructing the sampling distribution
- therefore, using the "right" estimator is crucial for meaningful results!

Unbiasedness of OLS

Unbiasedness of OLS estimator:

$$E(\hat{\beta}_0) = \beta_0 \qquad \text{and} \qquad E(\hat{\beta}_1) = \beta_1$$

Now some math to prove it...

Unbiasdness of $\hat{\beta}_1^{OLS}$

Lets first rewrite the formula for the estimator in the following way:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\beta_0 \sum_i (x_i - \bar{x}) + \beta_1 \sum_i x_i (x_i - \bar{x}) + \sum_i (x_i - \bar{x}) u_i - \bar{y} \sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\beta_1 \sum_i x_i (x_i - \bar{x}) + \sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2} \\
&= \beta_1 + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

since $\sum_i x_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x})^2$
and $\sum_i (x_i - \bar{x}) = \sum_i x_i - \sum_i \bar{x} = n\bar{x} - n\bar{x} = 0$
and $\bar{y} \sum_i (x_i - \bar{x}) = \bar{y} \sum_i x_i - \bar{y} n\bar{x} = \bar{y} n\bar{x} - \bar{y} n\bar{x} = 0$

Now we can proceed to the actual proof of unbiasedness

Proof.

$$\begin{aligned} E(\hat{\beta}_1) &= E[\beta_1 + \frac{\sum_i (x_i - \bar{x})u_i}{\sum_i (x_i - \bar{x})^2} | x_i] \\ &= \beta_1 + \frac{E[\sum_i (x_i - \bar{x})u_i | x_i]}{\sum_i (x_i - \bar{x})^2} \\ &= \beta_1 + \frac{\sum_i (x_i - \bar{x}) E(u_i | x_i)}{\sum_i (x_i - \bar{x})^2} \\ &= \beta_1 \end{aligned}$$

□

Unbiasdness of $\hat{\beta}_0^{OLS}$

Proof.

$$\begin{aligned}
E(\hat{\beta}_0|x_i) &= E(\bar{y} - \hat{\beta}_1\bar{x}|x_i) \\
&= E(\beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x}|x_i) \\
&= E[\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}|x_i] \\
&= E(\beta_0|x_i) + E[(\beta_1 - \hat{\beta}_1)\bar{x}|x_i] + E(\bar{u}|x_i) \\
&= \beta_0 + E[(\beta_1 - \hat{\beta}_1)\bar{x}|x_i] \\
&= \beta_0 \\
\text{as } E(\hat{\beta}_1) &= \beta_1
\end{aligned}$$

□

Which assumptions did we use to prove unbiasedness?

- A1 ?
- A2 ?
- A3 ?
- A4 ?

What happens if any of these assumptions is violated??
Also, remember that unbiasedness is a statistical property - it will hold if we draw a "typical" sample. It is always possible to get an unlucky draw. Therefore, before starting any estimation procedure "stare at" your data...

Efficiency of an estimator

- Def: an estimator is efficient if its sample distribution has the smallest variance
- often a secondary criterion: helps choosing between many unbiased estimators
- the variance of the sampling distribution tells us how far we can expect our estimate to be away from the true value of the parameter
- consider an estimator that always returns value 5.2 - what is the variance of this estimator?
- how desirable this estimator is?
- there usually is some additional constraint, like unbiasedness, that an estimator must satisfy before we can talk about efficiency

Variance of the OLS estimator

Variance of OLS estimator:

$$\begin{aligned} Var(\beta_1) &= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \\ \text{and} & \\ Var(\beta_0) &= \frac{\sigma^2 n^{-1} \sum_i x_i^2}{\sum_i (x_i - \bar{x})^2} \end{aligned}$$

Now some more math. We will only prove the variance of β_1 here.

Proof.

Recall that: $\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}$

And that $Var(A + B) = Var(A) + Var(B) + 2Cov(A, B)$

And that $Var(2A) = 4Var(A)$

Than:

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var(\beta_1) + Var[\frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}] + 2Cov(\beta_1, \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}) \\
&= Var[\frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}] \\
&= \frac{\sum_i (x_i - \bar{x})^2 Var(u_i)}{[\sum_i (x_i - \bar{x})^2]^2} \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

since all covariances are 0 and $Var(u) = \sigma^2$ □

The error variance σ^2

- the formula for the variance of $\hat{\beta}_1$ is useless unless we know σ^2 - highly unlikely...
- therefore, we need to find another estimator for the error variance
- natural candidate - residuals
- once again lets underline the difference between the errors and the residuals:

  errors $\longrightarrow$ unobservable

  residuals $\longrightarrow$ computed from the data after estimates of parameters are obtained

Estimating the error variance σ^2

- unbiased estimator would be $n^{-1}\sum_i u_i^2$
- however, we do not observe u_i
- so lets use the OLS residuals to develop the estimator so use $n^{-1}\sum_i \hat{u}_i^2$
- this estimator turns out to be biased → lets use the variation of this estimator

An unbiased estimator of the error variance is:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i \hat{u}_i^2 \equiv s^2 \text{ (alternative notation)}$$

And therefore, the estimator for the variance of $\hat{\beta}_1$ becomes:

$$Var\hat{(}\beta_1) = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}$$