

Economics 326
Methods of Empirical Research in Economics
Lecture 16: Large sample results: Consistency

Vadim Marmer
University of British Columbia

March 29, 2011

Why we need the large sample theory

- ▶ We have shown that the OLS estimator $\hat{\beta}$ has some desirable properties:
 - ▶ $\hat{\beta}$ is unbiased if the errors are strongly exogenous:
 $E(U|X) = 0$.
 - ▶ If in addition the errors are homoskedastic then
 $\widehat{Var}(\hat{\beta}) = s^2 / \sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator of the conditional variance of the OLS estimator $\hat{\beta}$.
 - ▶ If in addition the errors are normally distributed (given X) then
 $T = (\hat{\beta} - \beta) / \sqrt{\widehat{Var}(\hat{\beta})}$ has a t distribution which can be used for hypotheses testing.

Why we need the large sample theory

- ▶ If the errors are only weakly exogenous:

$$E(X_i U_i) = 0,$$

the OLS estimator is in general biased.

- ▶ If the errors are heteroskedastic:

$$E(U_i^2 | X_i) = h(X_i),$$

the "usual" variance formula is invalid; we also do not have an unbiased estimator for the variance in this case.

- ▶ If the errors are not normally distributed conditional on X then T - and F -statistics do not have t and F distributions under the null hypothesis.
- ▶ The asymptotic or large sample theory allows us to derive approximate properties and distributions of estimators and test statistics by assuming that the sample size n is very large.

Convergence in probability and LLN

- ▶ Let θ_n be a sequence of random variables indexed by the sample size n . We say that θ_n converges in probability if

$$\lim_{n \rightarrow \infty} P(|\theta_n - \theta| \geq \varepsilon) = 0 \text{ for all } \varepsilon > 0.$$

- ▶ We denote this as $\theta_n \rightarrow_p \theta$ or $p \lim \theta_n = \theta$.
- ▶ An example of convergence in probability is a Law of Large Numbers (LLN):

Let X_1, X_2, \dots, X_n be a random sample such that $E(X_i) = \mu$ for all $i = 1, \dots, n$, and define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then, under certain conditions,

$$\bar{X}_n \rightarrow_p \mu.$$

- ▶ Let X_1, \dots, X_n be a sample of independent identically distributed (iid) random variables. Let $EX_i = \mu$. If $Var(X_i) = \sigma^2 < \infty$ then

$$\bar{X}_n \rightarrow_p \mu.$$

- ▶ In fact when the data are iid, the LLN holds if

$$E|X_i| < \infty,$$

but we prove the result under a stronger assumption that $Var(X_i) < \infty$.

Markov's inequality

- ▶ **Markov's inequality.** Let W be a random variable. For $\varepsilon > 0$ and $r > 0$,

$$P(|W| \geq \varepsilon) \leq \frac{E|W|^r}{\varepsilon^r}.$$

- ▶ With $r = 2$, we have **Chebyshev's inequality**. Suppose that $EX = \mu$. Take $W \equiv X - \mu$ and apply Markov's inequality with $r = 2$. For $\varepsilon > 0$,

$$\begin{aligned} P(|X - \mu| \geq \varepsilon) &\leq \frac{E|X - \mu|^2}{\varepsilon^2} \\ &= \frac{\text{Var}(X)}{\varepsilon^2}. \end{aligned}$$

- ▶ Probability of observing an outlier (a large deviation of X from its mean μ) can be bounded by the variance.

Proof of the LLN

Fix $\varepsilon > 0$ and apply Markov's inequality with $r = 2$:

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \varepsilon) &= P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right) \\ &= P\left(\left|\frac{1}{n} \sum_{i=1}^n (X_i - \mu)\right| \geq \varepsilon\right) \\ &\leq \frac{E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)\right)^2}{\varepsilon^2} \\ &= \frac{1}{n^2 \varepsilon^2} \left(\sum_{i=1}^n E(X_i - \mu)^2 + \sum_{i=1}^n \sum_{j \neq i} E(X_i - \mu)(X_j - \mu) \right) \\ &= \frac{1}{n^2 \varepsilon^2} \left(\sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j \neq i} \text{Cov}(X_i, X_j) \right) \\ &= \frac{n\sigma^2}{n^2 \varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty \text{ for all } \varepsilon > 0. \end{aligned}$$

Averaging and variance reduction

- ▶ Let X_1, \dots, X_n be a sample and suppose that

$$\begin{aligned}E(X_i) &= \mu \text{ for all } i = 1, \dots, n, \\ \text{Var}(X_i) &= \sigma^2 \text{ for all } i = 1, \dots, n, \\ \text{Cov}(X_i, X_j) &= 0 \text{ for all } j \neq i.\end{aligned}$$

- ▶ Consider the mean of the average:

$$\begin{aligned}E(\bar{X}_n) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu.\end{aligned}$$

Averaging and variance reduction

- ▶ Consider the variance of the average:

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j \neq i}^n \text{Cov}(X_i, X_j) \right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \sigma^2 + \sum_{i=1}^n \sum_{j \neq i}^n 0 \right) \\ &= \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

- ▶ The variance of the average approaches zero as $n \rightarrow \infty$ if the

Convergence in probability: properties

- ▶ **Slutsky's Lemma.** Suppose that $\theta_n \rightarrow_p \theta$, and let g be a function continuous at θ . Then,

$$g(\theta_n) \rightarrow_p g(\theta).$$

- ▶ If $\theta_n \rightarrow_p \theta$, then $\theta_n^2 \rightarrow_p \theta^2$.
- ▶ If $\theta_n \rightarrow_p \theta$ and $\theta \neq 0$, then $1/\theta_n \rightarrow_p 1/\theta$.
- ▶ Suppose that $\theta_n \rightarrow_p \theta$ and $\lambda_n \rightarrow_p \lambda$. Then,
 - ▶ $\theta_n + \lambda_n \rightarrow_p \theta + \lambda$.
 - ▶ $\theta_n \lambda_n \rightarrow_p \theta \lambda$.
 - ▶ $\theta_n / \lambda_n \rightarrow_p \theta / \lambda$ provided that $\lambda \neq 0$.

Consistency

- ▶ Let $\hat{\beta}_n$ be an estimator of β based on a sample of size n .
- ▶ We say that $\hat{\beta}_n$ is a consistent estimator of β if as $n \rightarrow \infty$,

$$\hat{\beta}_n \rightarrow_p \beta.$$

- ▶ Consistency means that the probability of the event that the distance between $\hat{\beta}_n$ and β exceeds $\varepsilon > 0$ can be made arbitrary small by increasing the sample size.

Consistency of OLS

► Suppose that:

1. The data $\{(Y_i, X_i) : i = 1, \dots, n\}$ are iid.
2. $Y_i = \beta_0 + \beta_1 X_i + U_i$, where $E(U_i) = 0$.
3. $E(X_i U_i) = 0$.
4. $0 < \text{Var}(X_i) < \infty$.

- Let $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$ be the OLS estimators of β_0 and β_1 respectively based on a sample of size n . Under Assumptions 1-4,

$$\hat{\beta}_{0,n} \rightarrow_p \beta_0,$$

$$\hat{\beta}_{1,n} \rightarrow_p \beta_1.$$

- The key identifying assumption is Assumption 3:
 $\text{Cov}(X_i, U_i) = 0$.

Proof of consistency

- Write

$$\begin{aligned}\hat{\beta}_{1,n} &= \frac{\sum_{i=1}^n (X_i - \bar{X}_n) Y_i}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X}_n) U_i}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ &= \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}.\end{aligned}$$

- We will show that

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i &\rightarrow_p 0, \\ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 &\rightarrow_p \text{Var}(X_i),\end{aligned}$$

- Since $\text{Var}(X_i) \neq 0$,

$$\hat{\beta}_{1,n} = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \rightarrow_p \beta_1 + \frac{0}{\text{Var}(X_i)} = \beta_1.$$

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i \rightarrow_p 0$$

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i = \frac{1}{n} \sum_{i=1}^n X_i U_i - \bar{X}_n \left(\frac{1}{n} \sum_{i=1}^n U_i \right).$$

By the LLN,

$$\frac{1}{n} \sum_{i=1}^n X_i U_i \rightarrow_p E(X_i U_i) = 0,$$

$$\bar{X}_n \rightarrow_p E(X_i),$$

$$\frac{1}{n} \sum_{i=1}^n U_i \rightarrow_p E(U_i) = 0.$$

Hence,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i &= \frac{1}{n} \sum_{i=1}^n X_i U_i - \bar{X}_n \left(\frac{1}{n} \sum_{i=1}^n U_i \right) \rightarrow_p 0 - E(X_i) \cdot 0 \\ &= 0. \end{aligned}$$

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \rightarrow_p \text{Var}(X_i)$$

► First,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2\bar{X}_n X_i + \bar{X}_n^2) \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2\bar{X}_n \frac{1}{n} \sum_{i=1}^n X_i + \bar{X}_n^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2\bar{X}_n \bar{X}_n + \bar{X}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2. \end{aligned}$$

► By the LLN, $\frac{1}{n} \sum_{i=1}^n X_i^2 \rightarrow_p E(X_i^2)$ and $\bar{X}_n \rightarrow_p EX_i$.

► By Slutsky's Lemma, $\bar{X}_n^2 \rightarrow_p (EX_i)^2$.

► Thus,

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \rightarrow_p E(X_i^2) - (EX_i)^2 = \text{Var}(X_i).$$

Multiple regression

- ▶ Under similar conditions to 1-4, one can establish consistency of OLS for the multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i} + U_i,$$

where $EU_i = 0$.

- ▶ The key assumption is that the errors and regressors are uncorrelated:

$$E(X_{1,i}U_i) = \dots = E(X_{k,i}U_i) = 0.$$

Omitted variables and the inconsistency of OLS

- ▶ Suppose that the true model has two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + U_i,$$
$$E(X_{1,i} U_i) = E(X_{2,i} U_i) = 0.$$

- ▶ Suppose that the econometrician includes only X_1 in the regression when estimating β_1 :

$$\begin{aligned}\tilde{\beta}_{1,n} &= \frac{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) Y_i}{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} \\ &= \frac{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) (\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + U_i)}{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) X_{2,i}}{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} + \frac{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) U_i}{\sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2}.\end{aligned}$$

Omitted variables and the inconsistency of OLS

$$\tilde{\beta}_{1,n} = \beta_1 + \beta_2 \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) X_{2,i}}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} + \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) U_i}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2}.$$

► As before,

$$\begin{aligned} \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) U_i}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} &= \frac{\frac{1}{n} \sum_{i=1}^n X_{1,i} U_i - \bar{X}_{1,n} \bar{U}_n}{\frac{1}{n} \sum_{i=1}^n X_{1,i}^2 - \bar{X}_{1,n}^2} \\ &\xrightarrow{p} \frac{0}{EX_{1,i}^2 - (EX_{1,i})^2} \\ &= \frac{0}{\text{Var}(X_{1,i})} = 0. \end{aligned}$$

Omitted variables and the inconsistency of OLS

$$\tilde{\beta}_{1,n} = \beta_1 + \beta_2 \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) X_{2,i}}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} + \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) U_i}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2}.$$

► However,

$$\begin{aligned} \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) X_{2,i}}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} &= \frac{\frac{1}{n} \sum_{i=1}^n X_{1,i} X_{2,i} - \bar{X}_{1,n} \bar{X}_{2,n}}{\frac{1}{n} \sum_{i=1}^n X_{1,i}^2 - \bar{X}_{1,n}^2} \\ &\xrightarrow{p} \frac{E(X_{1,i} X_{2,i}) - (EX_{1,i})(EX_{2,i})}{EX_{1,i}^2 - (EX_{1,i})^2} \\ &= \frac{\text{Cov}(X_{1,i}, X_{2,i})}{\text{Var}(X_{1,i})}. \end{aligned}$$

Omitted variables and the inconsistency of OLS

- ▶ We have,

$$\begin{aligned}\tilde{\beta}_{1,n} &= \beta_1 + \beta_2 \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) X_{2,i}}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} + \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n}) U_i}{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - \bar{X}_{1,n})^2} \\ &\rightarrow_p \beta_1 + \beta_2 \frac{\text{Cov}(X_{1,i}, X_{2,i})}{\text{Var}(X_{1,i})} + \frac{0}{\text{Var}(X_{1,i})} \\ &= \beta_1 + \beta_2 \frac{\text{Cov}(X_{1,i}, X_{2,i})}{\text{Var}(X_{1,i})}.\end{aligned}$$

- ▶ Thus, $\tilde{\beta}_{1,n}$ is inconsistent unless:
 1. $\beta_2 = 0$ (the model is correctly specified).
 2. $\text{Cov}(X_{1,i}, X_{2,i}) = 0$ (the omitted variable is uncorrelated with the included regressor).

Omitted variables and the inconsistency of OLS

- ▶ In this example, the model contains two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + U_i,$$
$$E(X_{1,i} U_i) = E(X_{2,i} U_i) = 0.$$

- ▶ However, since X_2 is not controlled for, it goes into the error term:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + V_i, \text{ where}$$
$$V_i = \beta_2 X_{2,i} + U_i.$$

- ▶ For consistency of $\tilde{\beta}_{1,n}$ we need $\text{Cov}(X_{1,i}, V_i)$ to be equal to zero, however,

$$\begin{aligned} \text{Cov}(X_{1,i}, V_i) &= \text{Cov}(X_{1,i}, \beta_2 X_{2,i} + U_i) \\ &= \text{Cov}(X_{1,i}, \beta_2 X_{2,i}) + \text{Cov}(X_{1,i}, U_i) \\ &= \beta_2 \text{Cov}(X_{1,i}, X_{2,i}) + 0 \\ &\neq 0, \text{ unless } \beta_2 = 0 \text{ or } \text{Cov}(X_{1,i}, X_{2,i}) = 0. \end{aligned}$$