Section 3. Simple Regression - Omitted Variable Bias

1. Sampling Properties of OLS estimators
2. What's a Covariance?
3. More sampling properties of OLS estimators
4. Standard Errors of OLS estimators
5. Confidence intervals
6. CLT demonstration
7. Omitted Variables and Omitted Variable Bias (prelude to Section 4)

1. Sampling Properties of OLS estimators

Review:

- Unbiased: $E(b_0) = \boldsymbol{\beta_0}$, $E(b_1) = \boldsymbol{\beta_1}$
- Consistent: $\text{plim}(b_0) = \boldsymbol{\beta_0}$, $\text{plim}(b_1) = \boldsymbol{\beta_1}$

New:

- Asymptotically Normal by the CLT

Derivation

$$b_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$$

$$= \frac{\sum(X_i-\bar{X})Y_i - \cancel{\sum(X_i-\bar{X})\bar{Y}}^{\,0}}{\sum(X_i-\bar{X})X_i - \cancel{\sum(X_i-\bar{X})\bar{X}}^{\,0}}$$

$$= \frac{1}{\sum(X_i-\bar{X})X_i}\sum(X_i-\bar{X})[\beta_0+\beta_1X_i+u_i]$$

$$= \frac{1}{\underbrace{\sum(X_i-\bar{X})X_i}}\left[\cancel{\sum(X_i-\bar{X})}^{\,0}\beta_0+\beta_1\underbrace{\sum(X_i-\bar{X})X_i}+\sum(X_i-\bar{X})u_i\right]$$

$$= 0 + \beta_1 + \frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})X_i}$$

$$E(b_1|X) = E\left(\beta_1+\frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})X_i}\Big|X\right) = \beta_1+\frac{1}{\sum(X_i-\bar{X})X_i}\left[\sum E(X_i-\bar{X})u_i|X\right]$$

$$\left[\sum(X_i-\bar{X})E(u_i|X)\right]$$

$$= \beta_1 + 0 = \beta_1 \quad \text{FOR ALL POSSIBLE } X$$

$$\Rightarrow E(b_1) = \beta_1$$

LEMMA:

#1) $\sum_{i=1}^{N}(X_i-\bar{X}) = \sum X_i - \sum\frac{\sum X_i}{N}$

$= \sum X_i - \cancel{N}\cdot\frac{\sum X_i}{\cancel{N}} = 0$

#2) $\sum_{i=1}^{N}(X_i-\bar{X})C$, C is a constant

$= (X_1-\bar{X})C + (X_2-\bar{X})C + \ldots (X_N-\bar{X})C$

$= 0$

#3) $E(Z|X) = A$ FOR ALL X

$\Rightarrow E(Z) = A$

Derivation

$$\boxed{b_1 = \beta_1 + \frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X})^2}}$$

$$\operatorname{plim} b_1 = \operatorname{plim}(\beta_1) + \operatorname{plim}\left(\frac{\sum (X_i - \bar{X}) u_i / N}{\sum (X_i - \bar{X})^2 / N}\right)$$

$$= \beta_1 + \frac{Cov(X,u) \times 1}{V(X) \times 1} = 0$$

$$Cov(X,u) = E(X - E(X))(u - E(u))$$

$$= E[E(X - E(X))\underbrace{(u - E(u)) \mid X}_{0}] = 0$$

L·L·N

$$E(Z) = \mu_Z \Rightarrow \operatorname{plim}(\bar{Z}) = \mu_Z$$

$$\Rightarrow \operatorname{plim} b_1 = \beta_1 + 0 = \beta_1$$

$$E\left[\overbrace{\frac{\sum (X_i - \bar{X}) u_i}{(N-2)}}^{\downarrow} \frac{(N-2)}{N}\right] = \frac{N-2}{N}$$

$Cov(X, u)$

$$E\left[\frac{\sum (X_i - \bar{X})^2}{(N-1)} \frac{(N-1)}{N}\right]$$

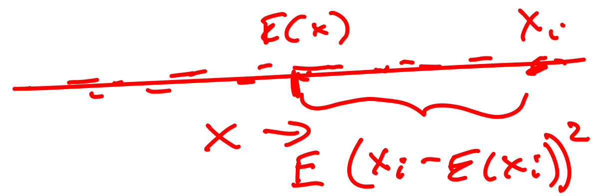
$$= V(X)\left(\frac{N-1}{N}\right)$$

$$\operatorname{plim}\left(\frac{N-1}{N}\right) = 1$$

$$\operatorname{plim}\left(\frac{N-2}{N}\right) = 1$$

2. What's a Covariance?

VARIANCE: $E(X_i - E(X))^2$

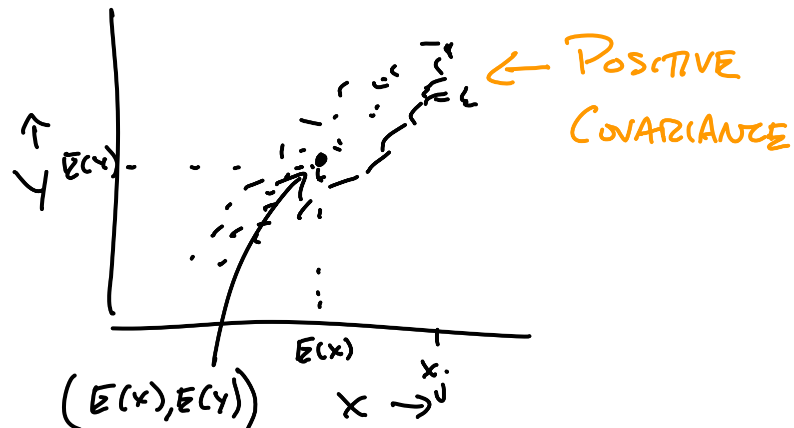


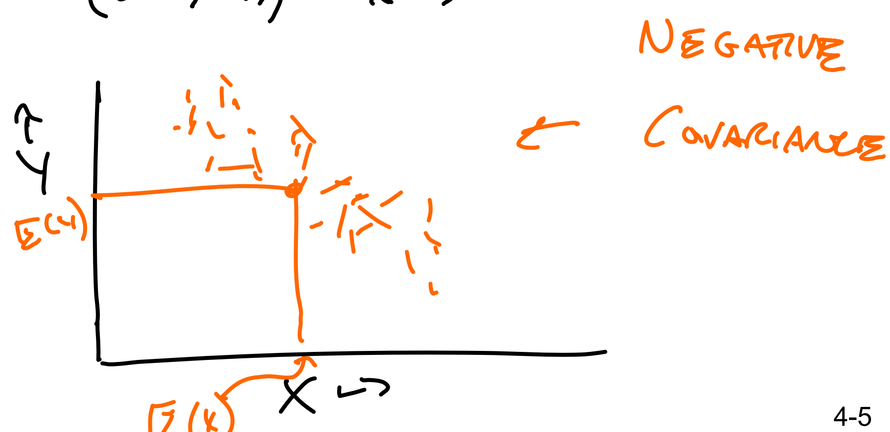
Covariance: $E([X_i - E(X)][Y_i - E(Y)])$

SAMPLE COVARIANCE OF X AND Y

(under random sampling) $\dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{(N-2)}$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$





3. Sampling Properties of OLS estimators (cont.)

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$b_1 \overset{a}{\sim} N(\beta_1, \sigma^2_{\hat{\beta}_1})$

$\hat{V}(b_1) = \dfrac{\sum(x_i-\bar{x})^2 e_i^2/n}{\left[\dfrac{\sum(x_i-\bar{x})^2}{n-2}\right]^2}\left(\dfrac{1}{n}\right)$

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i-\mu_X)u_i]}{[\text{var}(X_i)]^2}. = V(b_1) \qquad (4.14)$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

b_0

We like precision so we like ↑ BIG n

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \quad \text{where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right)X_i. \qquad (4.15)$$

Why asymptotically N(.)?

$$b_1 = \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2} = \frac{1}{\sum(X_i-\bar{X})^2}\sum(X_i-\bar{X})[\beta_0+\beta_1X_i+u_i] = 0+\beta_1+\frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})^2}$$

WHAT'S THE SAMPLING DISTRIBUTION OF b_1?

$v_i = (X_i-\bar{X})u_i \quad E(v_i) = 0 \Leftarrow \quad E(v_i|X) = E((X_i-\bar{X})u_i|X) = (X_i-\bar{X})E(u_i|X) = 0$

v_i iid? ✓

$V(v_i) = V[(X_i-\bar{X})u_i]$ is finite. ✓

$\sum v_i/n = \bar{v} \overset{A}{\sim} N(0, V(\bar{v}))$

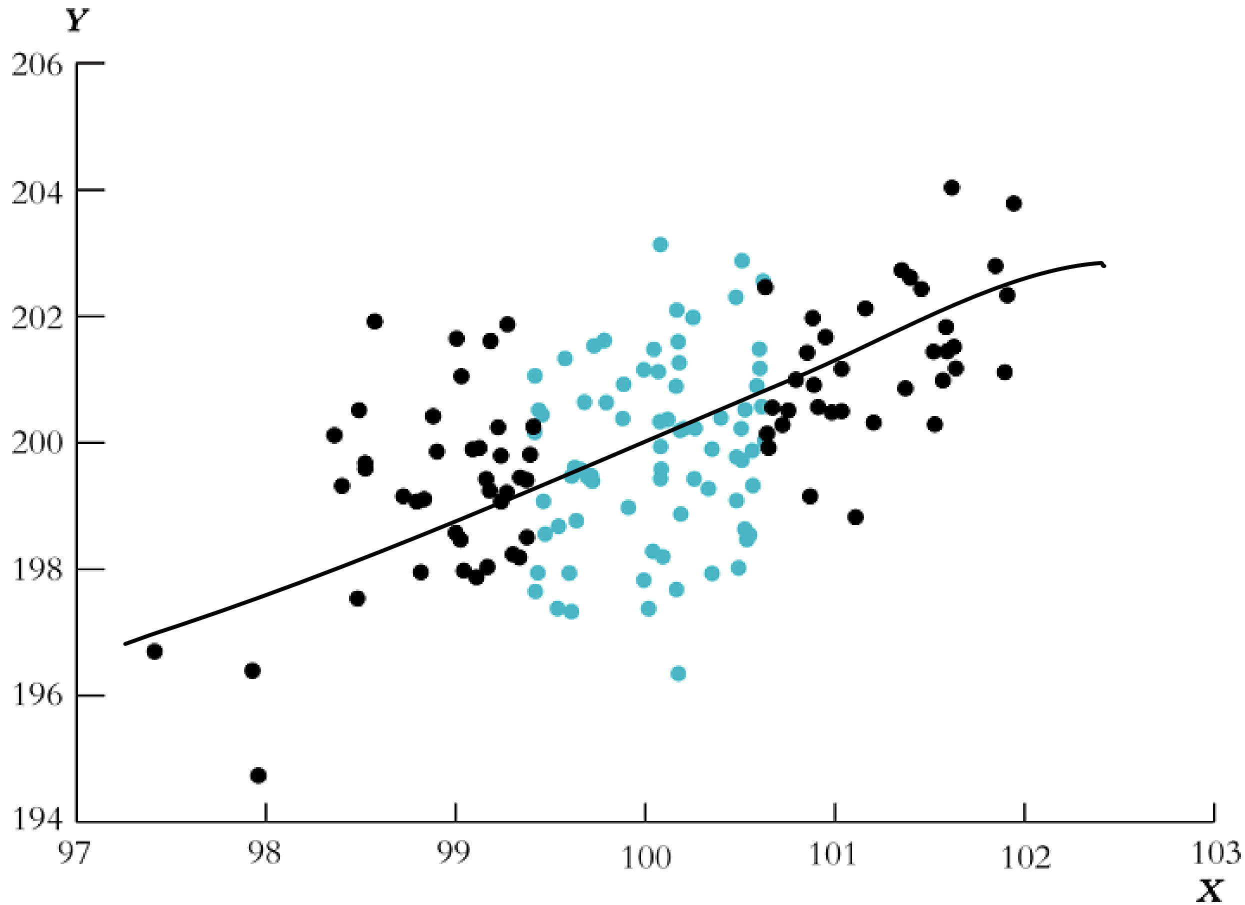
$$b_1 = \beta_1 + \frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})^2}$$

$$= \beta_1 + \frac{\sum v_i}{\sum(X_i-\bar{X})^2} = \beta_1 + \frac{\sum v_i/n}{\sum(X_i-\bar{X})^2/n}$$

$$\Rightarrow b_1 \overset{A}{\sim} N(\beta_1, V(b_1))$$

FIGURE 4.5 The Variance of $\hat{\beta}_1$ and the Variance of X

The colored dots represent a set of X_i's with a small variance. The black dots represent a set of X_i's with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.



$Q \sim N(\mu_Q, V(Q))$

4. Standard Errors

$Z = \frac{Q - \mu_Q}{\sqrt{V(Q)}} \sim N(0,1)$

- CLT: b_1 approx $\sim N(\boldsymbol{\beta_1}, \mathbf{V(b_1)})$

  b_0 approx $\sim N(\boldsymbol{\beta_0}, \mathbf{V(b_0)})$

$t = \frac{Q - \mu_Q}{\sqrt{\hat{V}(Q)}} \overset{A}{\sim} N(0,1)$

- **"Standard Errors" are consistent estimators of standard deviations of b_o and b_1 . (S&W p. 133, 151,180).**
- **So "t-statistics" have a standard normal distribution.**

std error $(b_1) = \sqrt{\hat{V}(b_1)}$

std. error $(b_0) = \sqrt{\hat{V}(b_0)}$

$t = \frac{b_1 - \beta_1}{\sqrt{\hat{V}(b_1)}} = \frac{b_1 - \beta_1}{\text{std.error}(b_1)} \overset{A}{\sim} N(0,1)$

5. Confidence Intervals

Confidence Intervals for β_1

A 95% two-sided confidence interval for β_1 is an interval that contains the true value of β_1 with a 95% probability, that is, it contains the true value of β_1 in 95% of all possible randomly drawn samples. Equivalently, it is also the set of values of β_1 that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, it is constructed as

$$95\% \text{ confidence interval for } \beta_1 = (\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)). \quad (4.27)$$

..and the same for $\boldsymbol{\beta_0}$

Confidence Intervals

$b_1 \overset{A}{\sim} N(\beta_1, V(b_1))$ by CLT

Confidence intervals for β_0, β_1

$$P\left(-1.96 \le \left(\frac{b_1-\beta_1}{se(b_1)}\right) \le 1.96\right) \approx 0.95$$

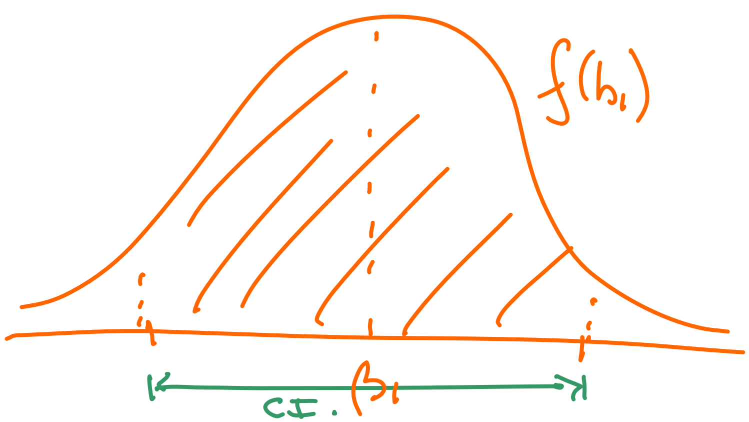
$$P\left[-1.96\, se(b_1) \le b_1-\beta_1 \le 1.96\, s.e.(b_1)\right] \approx 0.95$$

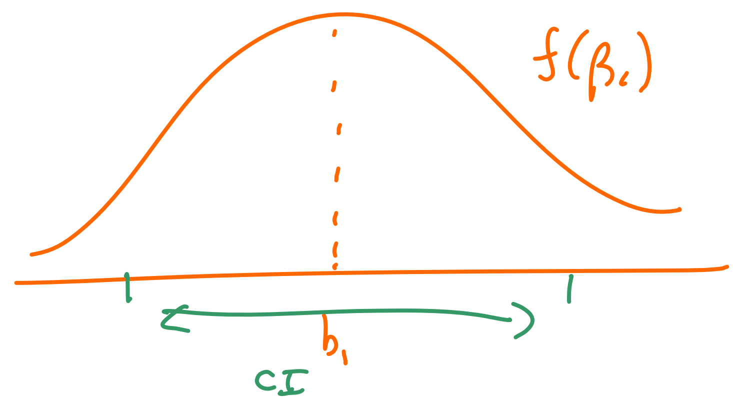
$\Downarrow$ PROBABILITY.

$$P\left[\beta_1 - 1.96\, se(\cdot) \le b_1 \le \beta_1 + 1.96\, se(\cdot)\right] \approx 0.95$$

$\Rightarrow$ STATISTICS

$$P\left[b_1 - 1.96\, se(\cdot) \le \beta_1 \le b_1 + 1.96\, se(\cdot)\right] \approx 0.95$$



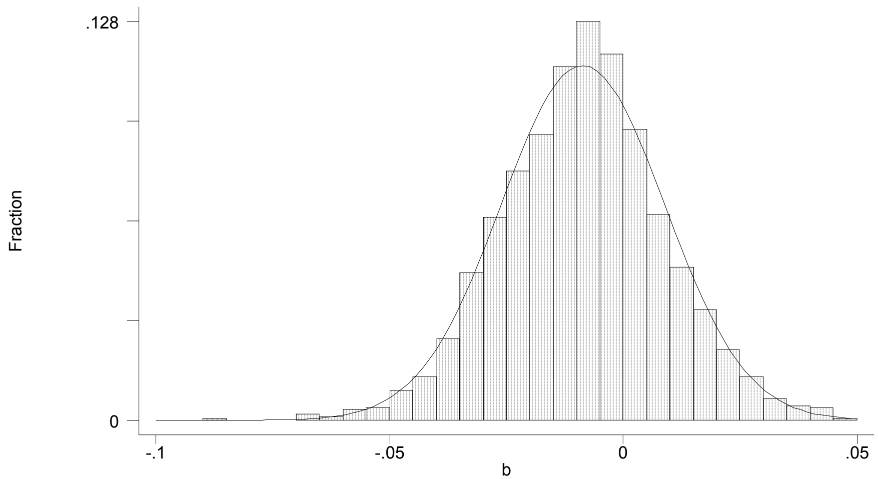


6. CLT demonstrations using Stata

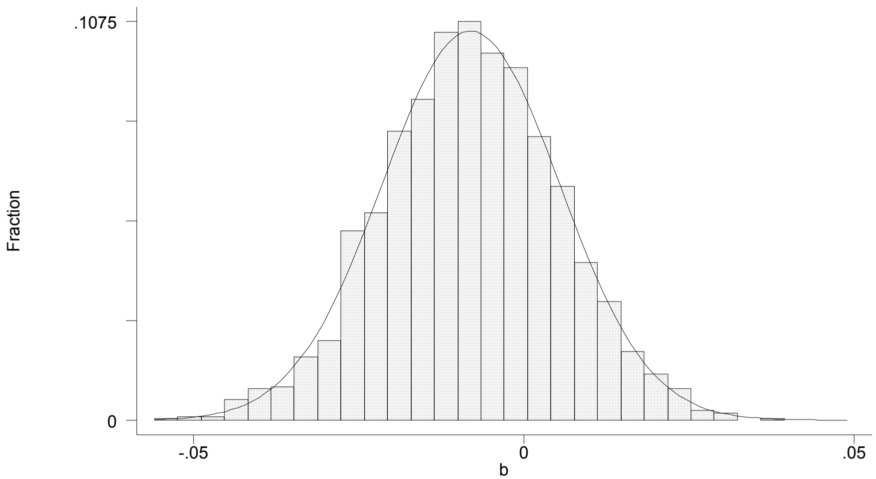
- Stata doesn't mind running a regression
  a few thousand times,
  - which allows us to observe a
  sampling distribution for $\mathbf{b_1}$
  e.g., bootregh00.do

..and the same for $\mathbf{b_0}$

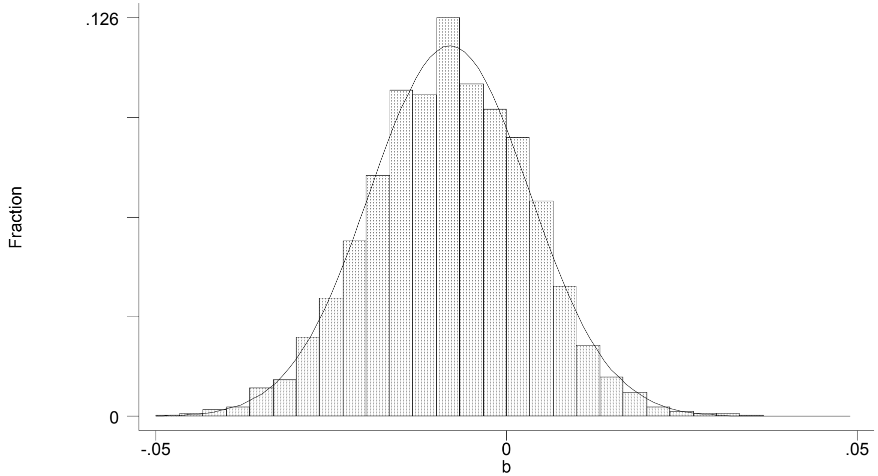
CLT in action: sampling distributions for b_1



n=20, 2000 reps



n=30, 2000 reps



n=40, 2000 reps

Note: In developing countries this slope is about -0.2 children per year of education. Vg

7. Omitted Variables and Omitted Variable Bias

- **What if you left out an important variable?**
- **Many interesting relationships have more than 2 dimensions**
- **Multivariate regression:**

7a. OLS Multivariate regression

The OLS Estimators, Predicted Values, and Residuals in the Multiple Regression Model

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the values of $b_0, b_1, \ldots, b_k$ that minimize the sum of squared prediction mistakes $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki},\ i = 1, \ldots, n, \text{ and} \tag{5.11}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{5.12}$$

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and residual $\hat{u}_i$ are computed from a sample of n observations of $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$. These are estimators of the unknown true population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and error term, u_i.

Look familiar? Same criterion with more variables.

7b. Properties of OLS estimators in Multivariate Regression

- Consistent
- Unbiased
- Approximately N(.) in large samples
- Claim:

7c. Omitted Variable "Bias"

- Short regression

$y = b_0^s + b_1^s x_1 + e^S$ (SR)

- Long regression

$y = b_0^L + b_1^L x_1 + b_2^L x_2 + e^L$ (LR)

- Claim:

$b_1^s = b_1^L + b_2^L b_{21}$,

b_{21} is slope of a regression of x_2 on x_1