Lecture 2: OLS

Linear Regression

$$y = x'\beta + \varepsilon$$

Three assumptions must be satisfied

(1) $\varepsilon_i \sim iid(0, \sigma^2) \Rightarrow E(\varepsilon) = 0, \; Var(\varepsilon) = \sigma^2 \cdot I_{N \times N}$

(2) $E(\varepsilon_i | X_{ik}) = 0$ for all k

(3) X has full column rank K

$$\begin{array}{c|ccc} i \backslash k & 1 & 2 & 3 \\ \hline 1 & 1 & 1 & 0 \\ 2 & 0 & 1 & 0 \\ 3 & 1 & 0 & 0 \end{array}$$

$\hat{\beta}_{OLS} = (X'X)^{-1}X'y \rightarrow$ From minimizing $\varepsilon'\varepsilon$ (the sum of squared residuals)

$$\min_{\beta} E(\hat{\varepsilon}'\hat{\varepsilon}) = E\left[(y - x'\beta)'(y - x'\beta)\right] = E\left[(y - x'\beta)^2\right]$$

$$\Rightarrow E\left[2(\hat{y} - \hat{x}'\hat{\beta})(-\hat{x})\right] = 0$$

$$E\left[(-2)\hat{x}(\hat{y} - \hat{x}'\hat{\beta})\right] = 0$$

$$\cancel{E(-2)}\left[E(\hat{x}\hat{y}) - E(\hat{x}\hat{x}')E(\hat{\beta})\right] = 0$$

$$E(\hat{x}\hat{y}) = E(\hat{x}\hat{x}')\hat{\beta}$$

$$E(\hat{x}\hat{x}')^{-1}E(\hat{x}\hat{y}) = \hat{\beta}$$

$$(x'x)^{-1}(x'Y) = \hat{\beta}$$

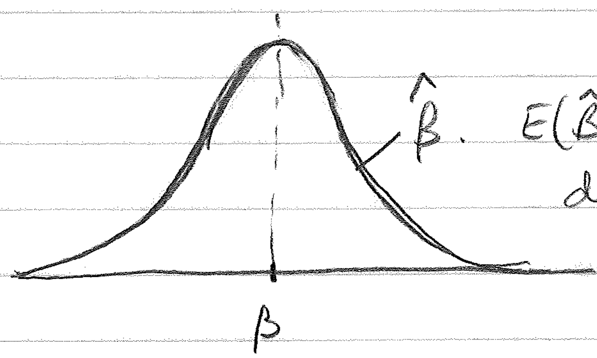
$E(\hat{\beta}_{OLS}) = \beta \Rightarrow$ OLS is unbiased

$$\widehat{Var}(\hat{\beta}_{OLS}) = \hat{\sigma}^2 (X'X)^{-1} \text{ where } \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-K} = \frac{SSR}{n-K}$$

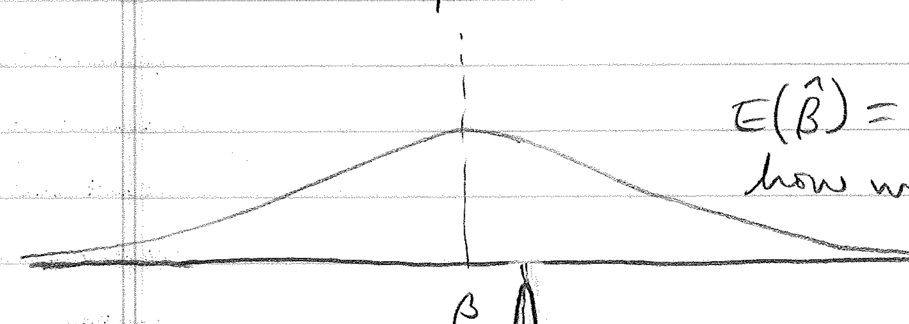
$\Rightarrow$ OLS is <u>efficient</u>

$\hat{\beta}_{OLS}$ is BLUE.

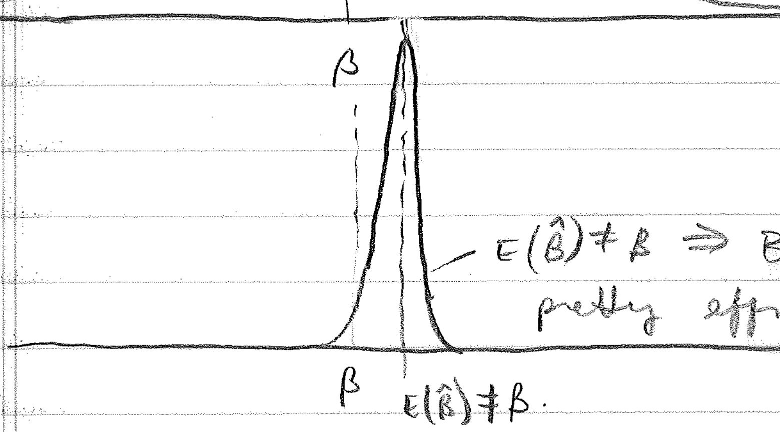
→ Side note: What's all this discussion about unbiased and efficient?



$E(\hat{\beta}) = \beta$ but there's a distribution of $\hat{\beta}$ values.



$E(\hat{\beta}) = \beta$ but look at how wide the distribution is! <u>Inefficient!</u>



$E(\hat{\beta}) \neq \beta \Rightarrow$ Biased. But it's also pretty efficient.

When the 3 assumptions are satisfied, OLS is the best (most efficient) <u>linear</u> unbiased estimator.

Identification Problems with X ⇒ Bias?

1. Omitted Variables

True model:

$$y_i = \alpha + \gamma s_i + \theta A_i + \varepsilon_i$$

↳ ability, unobserved, and correlated with s_i and y_i.

$$E(\varepsilon_i \mid s_i, A_i) = 0$$

$$\Rightarrow E(\hat{\gamma}) = \gamma \quad \text{Unbiased.}$$

Estimated model:

$$y_i = \tilde{\alpha} + \tilde{\gamma} s_i + \tilde{\varepsilon}_i$$

$E(\tilde{\varepsilon}_i \mid s_i) \neq 0$ bias due to omitted A_i.

$$\Rightarrow E(\hat{\tilde{\gamma}}) \neq \gamma \quad \text{Biased.}$$

But how biased are we? What <u>direction</u> is the bias?

$$E(\hat{\tilde{\gamma}}) = \gamma + \theta \cdot \frac{Cov(s, A)}{Var(s)}$$

where does this formula come from?

$$\hat{\tilde{\gamma}} = (s's)^{-1}(s'Y)$$

now let's substitute the true y_i

$$= (s's)^{-1}\left[s'(s\gamma + A\theta + \varepsilon)\right]$$

$$= (s's)^{-1}\left[(s's)\gamma + (s'A)\theta + s'\varepsilon\right]$$

$$= \cancel{(s's)^{-1}(s's)}\gamma + (s's)^{-1}(s'A)\theta + s'\varepsilon$$

$$= \gamma + (s's)^{-1}(s'A)\theta + s'\varepsilon$$

$$E(\hat{\tilde{\gamma}}) = \gamma + E(s's)^{-1}E(s'A)\theta + \underbrace{E(s'\varepsilon)}_{=0}$$

$$E(\hat{\tilde{\gamma}}) = \gamma + \theta\,\frac{Cov(s,A)}{Var(s)}$$

So if θ is expected to be positive (ability increases income controlling for s) and if $Cov(s,A) > 0$ (ability increases schooling), then bias is positive and $\tilde{\gamma}$ overstates the true γ.

```
clear
use restricted92

reg lnw computer exp

reg lnw computer ed exp

reg ed computer exp

*calculate the bias*
display 1.601328*.0790976

*does this get us back to the biased coeff estimate on computer?*
display .1266612+.1789369
```

2. Bad Controls

If we're worried about OVB, why not just throw everything we can into a regression?

a.) $\hat{Var}(\hat{\beta}_{OLS}) = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-K}(X'X)^{-1}$

The more K regressors you have, the smaller the denominator. If including another regressor does not reduce $\hat{\varepsilon}'\hat{\varepsilon}$, then all it does is increase the variance.

$\hookrightarrow$ Why might an X_K have no effect on $\hat{\varepsilon}'\hat{\varepsilon}$? $\frac{Cov(\tilde{X}_k, y)}{Var(\tilde{X}_k)} = \beta_k \approx 0$.

b) Suppose we want

$$\ln y_i = \alpha + \beta_1 college_i + \ldots + \varepsilon_i$$

What is the effect of college on wages?

What happens if we also control for occupational category ($wc_i = 1$ if white collar, $= 0$ if blue collar).

Now we're regressing,

$$\ln y_i = \tilde{\alpha} + \tilde{\beta}_1 \text{college}_i + \tilde{\beta}_2 wc_i + \ldots + \tilde{\varepsilon}_i$$

What's the interpretation? First, assume college is randomly assigned, but people can select into white collar or blue collar jobs.

$$E[Y_i | C_i = 1] - E[Y_i | C_i = 0] = E[Y_{1i} - Y_{0i}]$$
$$E[wc_i | C_i = 1] - E[wc_i | c_i = 0] = E[wc_{1i} - wc_{0i}]$$

Now, let's include a control for wc_i with $wc_i = 1$.

$$E[Y_i | wc_i = 1, c_i = 1] - E[Y_i | wc_i = 1, c_i = 0] =$$

$$E[Y_{1i} | wc_{1i} = 1] - E[Y_{0i} | wc_{0i} = 1] =$$

$$\underbrace{E[Y_{1i} - Y_{0i} | wc_{1i} = 1]}_{\text{causal effect}} + \underbrace{\{E[Y_{0i} | wc_{1i} = 1] - E[Y_{0i} | w_{0i} = 1]\}}_{\text{selection bias}}$$

$$\underbrace{E[Y_{1i} | wc_{1i} = 1] - E[Y_{0i} | wc_{1i} = 1]}_{} + E[Y_{0i} | wc_{1i} = 1] - E[Y_{0i} | w_{0i} = 1]$$
$$E[Y_{1i} - Y_{0i} | wc_{1i} = 1] + \{E[Y_{0i} | wc_{1i} = 1] - E[Y_{0i} | w_{0i} = 1]\}.$$

causal effect: Difference in wages for those who work a white collar job because they have a college degree

selection bias: college changes the composition of workers.

because college affects who becomes a white collar worker, someone who is a white collar worker but did not go to college might just be very talented (and therefore earn more). And someone who is white collar <u>only</u> because he is college educated might not be so talented (and therefore earn less).

→ WC_i is a bad control because it can be <u>caused</u> by $college_i$

```
clear
use restricted92

reg lnw ed exp exp2

sort occ
areg lnw ed exp exp2, absorb(occ)
```


9

3. Bad Functional Form, Misspecification

How do you choose between specifications?

(1) $\ln w = \alpha + \beta_1\, educ + \beta_2\, exp + \beta_3\, exp^2 + \varepsilon_i$

(2) $\ln w = \alpha + \beta_1\, educ + \beta_2\, educ^2 + \beta_3\, exp + \beta_4\, exp^2 + \beta_5\, exp \times educ + \varepsilon_i$

(3) $\ln w = \alpha + \beta_1 \ln educ + \beta_2 \ln exp + \beta_3 [\ln educ \times \ln exp] + \varepsilon_i$

It's an art more than a science.

Theory — a. You need to have some theory guiding you in choosing what to include as regressors.

Play & Justify — b. Run regressions and see what you get. Does a different specification make a difference? Do the results make intuitive sense (with respect to some theory)?

Test — c. Test against alternative specifications.

J - Test

$H_0: \; y = X\beta + \varepsilon_0$

$H_1: \; y = Z\delta + \varepsilon_1$

Regress $y = (1-\lambda)X\beta + \lambda Z\delta + \varepsilon_2$

If H_0 is true, then $\hat{\lambda} = 0$

In practice:

i. reg y on Z. Obtain $\hat{\delta}_{OLS}$. Predict $\hat{y}_1 = Z\hat{\delta}_{OLS}$

ii. reg y on X and $\hat{y}_1$.

$$y = (1-\lambda)X\beta + \lambda \hat{y}_1 + \varepsilon_3.$$

Test $\hat{\lambda} = 0$ using a t-test.

iii. Now reg y on X. Obtain $\hat{\beta}_{OLS}$. Predict $\hat{y}_2 = X\hat{\beta}_{OLS}$

iv. reg y on Z and $\hat{y}_2$. Test $\hat{\lambda} = 0$.

```
clear
use restricted92

reg lnw ed exp exp2

xi: reg lnw i.educat exp exp2

gen lned=ln(ed)
gen lnexp=ln(exp)
gen lnedXlnexp=lned*lnexp

reg lnw lned lnexp lnedXlnexp
*throw this last one out; doesn't make sense*


*Now test specifications 1 and 2.  Which is better?*
xi: reg lnw i.educat exp exp2
predict lnw_hat1

reg lnw ed exp exp2 lnw_hat1

***
reg lnw ed exp exp2
predict lnw_hat2

xi: reg lnw i.educat exp exp2 lnw_hat2


************************
xi: reg lnw ed i.educat exp exp2
```

4. Measurement Error

Case 1: y_i^* measured with error

$y_i^* = X_i'\beta + \varepsilon_i$ but observe $y_i = y_i^* + u_i$

$y_i = X_i'\beta + (\varepsilon_i + u_i)$

if $E(u_i \cdot X_i) = 0$, $\hat{\beta}_{OLS}$ unbiased but error variance increases (larger SEs).

if $E(u_i \cdot X_i) \neq 0$, $\hat{\beta}_{OLS}$ biased.

Case 2: X measured with error.

$y_i = \gamma \cdot s_i^* + \varepsilon_i$, s_i^* = true schooling

observe $s_i = s_i^* + u_i$, $u_i \sim iid(0, \sigma_u^2)$

$E(u_i \cdot s_i) = 0 \Rightarrow$ They're independent!

$y_i = \gamma s_i + (\varepsilon_i - \gamma u_i) = \gamma s_i + \tilde{\varepsilon}_i$

$\Rightarrow$ $\boxed{\hat{\gamma}_{OLS} \text{ biased down} \Rightarrow \text{attenuation bias}}$

why?

$$Cov(\tilde{\varepsilon}_i, s_i) = Cov(-\gamma u_i, u_i) = -\gamma \sigma_u^2 < 0$$

There's lower correlation between observed schooling and earnings due to misreporting in schooling (some variation not due to true variation in treatment).

$$E(\hat{\gamma}_{OLS}) = \gamma + \frac{-\gamma \sigma_u^2}{Var(s_i)} = \gamma - \gamma\left(\frac{\sigma_u^2}{\sigma_s^2}\right) = \gamma\left(1 - \underbrace{\frac{\sigma_u^2}{\sigma_s^2}}\right)$$

$$\lambda = \frac{\sigma_u^2}{\sigma_s^2} = \frac{\text{Noise}}{\text{Total Variance}}$$

$$\lambda = \frac{\sigma_u^2}{\sigma_s^2 + \sigma_u^2} = \frac{\text{Noise}}{\text{Signal} + \text{Noise}}$$

If $\lambda = 0.1 \Rightarrow 10\%$ attenuation bias in bivariate regression.

Now add X_i's to the regression:

$$y_i = \gamma \cdot s_i + X_i'\beta + (\varepsilon_i - \gamma u_i)$$

$$E(\hat{\gamma}_{OLS}) = \gamma\left(1 - \frac{\lambda}{\underbrace{1 - R^2_{s,x}}}\right)$$

$R^2_{s,x}$ = R-squared from regression of s_i on X_i's.

$R^2_{s,x} \uparrow \Rightarrow$ attenuation bias $\uparrow$ for fixed λ.

↳ If X_i's correlated with $s_i^* \Rightarrow$ soak up the signal in s_i.

↳ If (u_i, X_i) independent, then X_i soaks up no noise variance.

```
clear
use restricted92

reg lnw ed exp exp2

set seed 1000
gen ed_error1=invnorm(uniform())

gen edmis1=ed+ed_error

reg lnw edmis exp exp2

gen ed_error2=2*invnorm(uniform())
gen edmis2=ed+ed_error2

reg lnw edmis2 exp exp2

reg edmis2 female mar femmar

reg lnw edmis2 exp exp2 female mar femmar
```