Lecture 6 EEP 118

Spring 2025

Sofia Villas-Boas

Lecture Plan- Lecture 6

Pset 1 - see instructions to submit on pdf of pset in Bcourses and gradescope

Recap Lecture 5

Multiple Regression Chapter 3

1. **Motivation**
2. **Interpretation in the population model**
3. **OLS estimation with dataset**
   1. **Mechanics**
   2. **Interpret beta_hat**
   3. **Dummy variables**
4. **Adding/Omitting variables → Omitted Variable Bias (OVB)**

Study chapters 3.1-3.3.

Daily Assignment 6 posted - ungraded

Take away Lecture 5 :
Statistical Properties of Estimator betahat

1. beta_hat are random variables
2. beta_hat are unbiased (E(Beta0hat)=Beta0, E(beta1hat)=beta1 if

Simple Linear Regression (SLR) Assumptions

SLR1, Y linear in parameters
SLR2, { (xi,yi), i=1,…n} random sample in the population
SLR3 variation in x in sample
SLR4 E(u|x)=0

3. Repeating the same random sampling of N=630 observations gives different estimates, but if you were to average them up, you would find a biased estimator for the population parameter, because $E(\hat{\beta}) = \beta$

4. Increasing sample size increases the precision of the estimate, because given SLR5

$$\mathrm{var}(\hat{\beta}_1) = \frac{\mathrm{var}(u)}{\mathrm{SST}_x}$$

$$var(\widehat{\beta_1}) = \frac{\boldsymbol{\sigma_u^2}}{SST_x}$$

Since we do not know $\boldsymbol{\sigma_u^2}$ we estimate it by $\boldsymbol{s^2}$ $= \frac{SSR}{(n-2)} = \frac{\sum_i u\_hat_i^2}{(n-2)}$

Then the estimated variance of beta_hat is

$$\widehat{var(\widehat{\beta_1})} = \frac{\widehat{\sigma_u^2}}{SST_x} = \frac{SSR}{(n-2)\,SST_x}$$

Increasing n sample size increases the precision of the estimate, because $var(\widehat{\beta_1})$

$$\widehat{var(\widehat{\beta_1})} = \frac{\widehat{\sigma_u^2}}{SST_x} = \frac{SSR}{(n-2)\, SST_x}$$

Practical take-away

$\widehat{var(\widehat{\beta_1})}$ is the measure of the variation we can expect across the different estimators $(\widehat{\beta}\ )$of β

Many samples, then many $\widehat{\beta}$, all distributed around the true (unknown) value of β with standard error $\widehat{se(\widehat{\beta}\ )}$

HOW TO REDUCE $\widehat{se(\widehat{\beta}\ )}$?

- Large n
- Large variation in x
- Small variation in u

Lecture Plan- Lecture 6

Pset 1 - instructions to submit on handout in Bcourses/ Gradescope

Multiple Regression Chapter 3

1. **Motivation**
2. **Interpretation in the population model**
3. **OLS estimation with dataset**
   1. **Mechanics**
   2. **Interpret beta_hat**
   3. **Dummy variables**
4. **Adding/Omitting variables → Omitted Variable Bias (OVB)**

Study chapters 3.1-3.3.

Daily Assignment 6 posted

Multiple regression

1. Motivation

- We specify the linear model

$$\boldsymbol{ln(wage) = \beta_0 + \beta_1 education + \beta_2 gender + \beta_3 experience + \cdots u}$$

- Suppose we are interested in the role of experience on wages, so in the parameter ________
- Other variables are included even if we are not interested in their parameters
- We include these other variables to take them out of the error term , out of u (because if these variables were not included and if they are correlated with experience they would create bias, as you will see later)

Population Model

- In the general population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where u is the disturbance term

- For one observation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

- Key assumption: $E[u|x_1\ x_2\ \ldots\ x_k] = 0$ then

E[y|x]= $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k$.

2. Interpretation

Let the population regression model be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
β_1 measures the effect on y of a change in x_1 by one unit, holding all other factors fixed (x_2 and u).

β_1 measures the effect on E[y] of a change in x_1 by one unit, holding x_2 fixed and assuming E[u|x]=0.

β_1 is <u>true unknown value</u> from the population regression

Lets estimate it, then $\widehat{\beta_1}$ is an <u>estimator</u>, (formula) to compute an <u>estimate</u> (a value) with a sample

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2$$

$\widehat{\beta_1}$ measures the effect on the predicted $\hat{y}$ of a change in x_1 by one unit, holding x_2 fixed.

3. OLS estimation with data set

Mechanics:

Find $\widehat{\beta_0}\ \widehat{\beta_1}\ \ \widehat{\beta_2}\ \ldots\ \widehat{\beta_k}$ such that they minimize

$$\sum \hat{u}_i^2 = \sum_i (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \widehat{\beta_k} x_{ik})^2$$

No easy formula…

Use Lecture6.R

Lets look at a data set now: Data source: Current Population Survey 2006. Sample of 526 households (Lecture6.dta)

N=526

In R type

```
my_data <- read_dta("Lecture6.dta")
```

Data description

wage *average hourly earnings (in $)*
educ *years of education*
exper *years potential experience*
female 1=female, 0=male
nonwhite =1 if nonwhite
services =1 if in services industry
profocc =1 if in professional occupation
married =1 if respondent is married

Called a Dummy Variable, 0 or 1

Click on my_data on right or type/ run in R file the line **head(my_data)**

Lecture6.R × | my_data ×

Filter

	wage	educ	exper	tenure	nonwhite	female	married	numdep	sms
1	3.75	2	39	13	0	0	1	0	
2	2.92	3	51	30	1	0	0	0	
3	3.51	4	39	15	0	0	1	5	
4	3.00	4	48	0	1	0	1	0	
5	3.00	4	36	0	0	0	1	1	
6	5.20	6	47	13	1	0	1	0	
7	3.95	6	49	6	0	0	1	6	
8	2.90	6	49	7	0	0	1	0	
9	3.76	6	6	0	0	0	0	4	
10	6.25	7	39	21	1	0	1	0	
11	4.95	7	25	17	0	0	1	5	
12	6.00	8	44	28	0	0	1	0	
13	10.00	8	13	0	1	0	0	0	

describe(my_data,skew=FALSE)

Summary statistics of the variables, ignore skewness etc

$$s_x = \sqrt{\frac{(x-\bar{x})^2}{N-1}}$$

vars		n	mean ($\bar{x}$)	sd	min	max	range	se
wage	1	526	5.90	**3.69**	0.53	24.98	24.45	0.16
educ	2	526	12.56	2.77	0.00	18.00	18.00	0.12
exper	3	526	17.02	13.57	1.00	51.00	50.00	0.59
nonwhite	5	526	0.10	0.30	0.00	1.00	1.00	0.01
female	6	526	**0.48**	0.50	0.00	1.00	1.00	0.02
married	7	526	0.61	0.49	0.00	1.00	1.00	0.02
services	17	526	0.10	0.30	0.00	1.00	1.00	0.01
profocc	19	526	0.37	0.48	0.00	1.00	1.00	0.02

female=0/1 is a Dummy Variable,

Sample average=

$\overline{female}$ *is a* %, **47.908% here**

Generate log of wage and summary stats of this new variable added as a new column to the matrix dataframe my_data

```
my_data$lwage<-log(my_data$wage)
describe(my_data$lwage, skew=FALSE)
```

	vars	n	mean	sd	min	max	range	se
X1	1	526	1.62	0.53	-0.63	3.22	3.85	0.02

reg1<-lm(lwage ~ educ+exper+female,my_data)

summary(reg1)

```
lm(formula = lwage ~ educ + exper + female, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.89584 -0.26362 -0.03871  0.26765  1.28241 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.480836   0.105016   4.579 5.86e-06 ***
educ         0.091290   0.007123  12.816  < 2e-16 ***
exper        0.009414   0.001449   6.496 1.93e-10 ***
female      -0.343597   0.037667  -9.122  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4289 on 522 degrees of freedom
Multiple R-squared:  0.3526,	Adjusted R-squared:  0.3488 
F-statistic: 94.75 on 3 and 522 DF,  p-value: < 2.2e-16
```

$$\ln(\widehat{wage}) = 0.48 + .091\ \text{educ} + .009\ \text{exp} - .3436\ \text{female}$$

$$(.11) \quad (.007) \qquad (.0015) \qquad (.037)$$

$n = 526$, $R^2 = .353$

NICE WAY TO PRESENT RESULTS

```
reg1<-lm(lwage ~ educ+exper+female,my_data)
summary(reg1)
```

$$\ln(\widehat{wage}) = 0.48 + .091\ \text{educ} + .009\ \text{exp} - .3436\ \text{female}$$
$$\qquad\quad (.11) \quad (.007) \qquad\quad (.0015) \qquad\quad (.037)$$

NICE WAY TO PRESENT RESULTS

n = 526 , $R^2 = 0.353$

- $\widehat{\beta_1}$ is the marginal effect of education on predicted $ln(\widehat{wage})$ holding experience and gender constant

- Holding experience and gender fixed, a one year increase in education leads to a **0.091** increase in **predicted ln(wages)** which is a **9.1%** increase in **predicted wage (in levels)**

Regression of ln(wage) on education (with increasing number of other controls)

Regression of log(wage) on education

```
reg3<-lm(lwage ~ educ,my_data)
```

$$\widehat{\log(wage)} = 0.584 + .0827\,educ \qquad R^2 = .1858$$
$$\quad (.097) \quad (.0076) \qquad n = 526$$

Regression of ln(wage) on education (with increasing number of other controls)

predicted

Estimated coefficient

Regression of log(wage) on education

```
reg3<-lm(lwage ~ educ,my_data)
```

$$\widehat{\log(wage)} = 0.584 + .0827\,educ \qquad R^2 = .1858$$

$$(.097) \qquad (.0076) \qquad n = 526$$

(standard error)

Regression of ln(wage) on education (with increasing number of other controls)

predicted

Estimated coefficient

Regression of log(wage) on education

reg3<-lm(lwage ~ educ,my_data)

$$\widehat{\log(wage)} = 0.584 + .0827\,educ \qquad R^2 = .1858$$

$$(.097) \qquad (.0076) \qquad n = 526$$

(standard error)

R squared = 18.58%

Regression of ln(wage) on education (with increasing number of other controls)

Regression of log(wage) on education (with increasing number of other controls)

reg3<-lm(lwage ~ educ,my_data)

$\widehat{\log(wage)} = 0.584 + .0827\,educ$ $\quad R^2 = .1858$

$\qquad (.097) \quad (.0076)$ $\quad n = 526$

R squared = 18.58%

reg2<-lm(lwage ~ educ+exper,my_data)

$\widehat{\log(wage)} = 0.217 + .098\,educ + .0103\,exp$ $\quad R^2 = .249$

$\qquad (.11) \quad (.008) \quad (.0016)$ $\quad n = 526$

reg1<-lm(lwage ~ educ+exper+female,my_data)

$\widehat{\log(wage)} = 0.48 + .091\,educ + .009\,exp - .3436\,female$ $\quad R^2 = .353$

$\qquad (.11) \quad (.007) \quad (.0015) \quad (.037)$ $\quad n = 526$

reg0<-lm(lwage ~ educ+exper+female+services,my_data)

$\widehat{\log(wage)} = 0.51 + .0899\,educ + .0096\,exp - .329\,female - .228\,services$ $\quad R^2 = .369$

$\qquad (.104) \quad (.007) \quad (.0014) \quad (.037) \quad (.062)$ $\quad n = 526$

R squared improves with more controls

Later in class, ignore for now significance

Sign (Significance) Size

Lets go equation by equation

Regression of log(wage) on education (with increasing number of other controls)

reg lwage educ

$\widehat{\log(\text{wage})} = 0.584 + .0827\ \text{educ}$ $\quad R^2 = .1858$

$\qquad (.097) \qquad (.0076)$ $\quad n = 526$

Δeduc=1 => Δ lnwagehat=0.0827 $SIZE\ is$ Δw/w=0.0827, positive (SIGN) , (significant) a 8.27 % wage_hat increase holding everything else constant

Next equation

reg lwage educ exper

$\widehat{\log(\text{wage})} = 0.217 + .098\ \text{educ} + .0103\ \text{exp}$ $\quad R^2 = .249$

$\qquad (.11) \qquad (.008) \qquad (.0016)$ $\quad n = 526$

Δexper=1 => Δ lnwagehat=0.0103 $SIZE\ is$ Δw/w=0.0103, positive (SIGN) , (significant) a 1.03 % wagehat increase holding all else, namely education) constant

0.098 is the partial effect of one year of education on predicted ln wages, holding all else constant, that is, Δexper=0

Answering the question on why betahat educ changes with and without controlling for experience

reg lwage educ exper

$$\widehat{\log(wage)} = 0.217 + .098\,educ + .0103\,exp \qquad R^2 = .249$$
$$(.11) \qquad (.008) \qquad (.0016) \qquad n = 526$$

- Holding experience constant one more year of education increases log wagehat by 0.098, or increases wagehat by 9.8%
- Before not controlling for experience beta_hat=0.0827, so the marginal effect was smaller was 8.27% and not 9.8%

Dummy variable among the x's, among the regressors

Example female, coded as 0/1

$\ln(\widehat{wage})$ =0.48 + .091 educ + .009 exp **- .3436 female**

(.11) (.007) (.0015) (.037)

Female: $\ln\ (\widehat{wage_f})$ =0.48 + .091 educ + .009 exp - .3436

Male: $\ln\ (\widehat{wage_m})$ =0.48 + .091 educ + .009 exp

Then $\ln\ (\widehat{wage_f) - \ln}\ (\widehat{wage_m})$ =-0.3436

What does this mean?

Dummy variable among the x's, among the regressors

Example female, coded as 0/1

$\ln(\widehat{wage})$ =0.48 + .091 educ + .009 exp - .3436 female

(.11) (.007) (.0015) (.037)

Female: $\ln\ (\widehat{wage_f})$ =0.48 + .091 educ + .009 exp - .3436

Male: $\ln\ (\widehat{wage_m})$ =0.48 + .091 educ + .009 exp

Then $\ln\ (wage_f) \widehat{- \ln}\ (\widehat{wage_m})$ =**-0.3436**

Difference in ln is a percent difference

What does this mean?

The difference in predicted wage between men and women holding education and experience constant is that women earn significantly less than men, namely 34.36 percent less !!

(<u>Sign</u> = negative; <u>Significant</u>, yes-later in class;
<u>Size</u>: The difference in predicted wage between women and men is that women significantly earn less 34.36 % than men.

4. Adding/ Omitting Variables

Baseline : regress lwage on a constant and educ exper female

$$\widehat{\log(wage)} = \underset{(.11)}{0.48} + \underset{(.007)}{.091}\, educ + \underset{(.0015)}{.009}\, exp - \underset{(.037)}{.3436}\, female \qquad R^2 = .353 \quad n = 526$$

<u>Adding</u> an irrelevant variable:

$$\widehat{\log(wage)} = \underset{(.11)}{0.48} + \underset{(.007)}{.09}\, educ + \underset{(.001)}{.009}\, exp - \underset{(.038)}{.343}\, female - \underset{(.062)}{.009}\, \textbf{nonwhite} \qquad R^2 = .353 \quad n = 526$$

From top to bottom regression, no change in parameters, no change in R squared and no change in parameters, when added nonwhite – **we added an irrelevant variable**

4. Adding/ Omitting Variables

Baseline : regress lwage on a constant and educ exper female

$$\widehat{\log(wage)} = \underset{(.11)}{0.48} + \underset{(.007)}{.091}\,educ + \underset{(.0015)}{\mathbf{.009}}\,\mathbf{exp} - \underset{(.037)}{.3436}\,female \qquad R^2 = .353,\; n = 526$$

Omitting an important variable not correlated with the female indicator if we only care about interpreting the gender gap in wages independent variables:

$$\widehat{\log(wage)} = \underset{(.09)}{0.83} + \underset{(.007)}{.077}\,educ - \underset{(.04)}{.361}\,female \qquad R^2 = .30,\; n = 526$$

From top to bottom regression, we omit experience, drop in R squared when omitted experience, and no major change in parameters – we omitted a relevant variable experience that is likely not very correlated with the other independent variables educ and female

Adding/ Omitting an important variable correlated with the other x's

Omitted variable bias

$\widehat{\log(wage)} =$	0.68	$+ .069\ educ$	$+ .008\ exp$	$- .315\ female$	$+ .236\ profocc$	$R^2 = .385$
	$(.11)$	$(.008)$	$(.001)$	$(.04)$	$(.045)$	$n = 526$
$\widehat{\log(wage)} =$	1.55	$+$	$+ .004 exp$	$- .319\ female$	$+ .432\ profocc$	$R^2 = .298$
	$(.04)$		$(.0014)$	$(.04)$	$(.04)$	$n = 526$

- From top to bottom regression, we omit education, drop in R squared when omitted educ,
- and change in parameters – we omitted a relevant variable experience that is correlated with the other independent variables exper and profocc

Adding/ Omitting an important variable correlated with the other x's

Omitted variable bias

$$\widehat{\log(wage)} = 0.68 + .069\,educ + .008\,exp - .315\,female + .236\,profocc \qquad R^2 = .385$$
$$\qquad (.11) \qquad (.008) \qquad (.001) \qquad (.04) \qquad (.045) \qquad n = 526$$

$$\widehat{\log(wage)} = 1.55 + \quad + .004exp - .319\,female + .432\,profocc \qquad R^2 = .298$$
$$\qquad (.04) \qquad \qquad (.0014) \qquad (.04) \qquad (.04) \qquad n = 526$$

. correlate lwage educ exp female profocc nonwhite (obs=526)

	lwage	educ	exper	female	profocc	nonwhite
lwage	1.0000					
educ	0.4311	1.0000				
exper	0.1114	-0.2995	1.0000			
female	-0.3737	-0.0850	-0.0416	1.0000		
profocc	0.4451	0.4968	-0.0056	-0.1774	1.0000	
nonwhite	-0.0389	-0.0847	0.0144	-0.0109	-0.0886	1.0000

- From top to bottom regression, we omit education, drop in R squared when **omitted educ**,
- and change in parameters – we omitted a relevant variable experience that is correlated with the other independent variables exper and profocc

Corr(exp,edu)<0

Adding/ Omitting an important variable correlated with the other x's

Omitted variable bias

$$\widehat{\log(wage)} = 0.68 + .069\,educ + .008\,exp - .315\,female + .236\,profocc \qquad R^2 = .385$$
$$(.11) \quad (.008) \quad (.001) \quad (.04) \quad (.045) \qquad n = 526$$

$$\widehat{\log(wage)} = 1.55 + \quad + .004\,exp - .319\,female + .432\,profocc \qquad R^2 = .298$$
$$(.04) \quad \quad (.0014) \quad (.04) \quad (.04) \qquad n = 526$$

```
. correlate lwage educ exp female profocc nonwhite (obs=526)
             |    lwage     educ    exper   female  profocc nonwhite
       lwage |   1.0000
        educ |   0.4311   1.0000
       exper |   0.1114  -0.2995   1.0000
      female |  -0.3737  -0.0850  -0.0416   1.0000
     profocc |   0.4451   0.4968  -0.0056  -0.1774   1.0000
    nonwhite |  -0.0389  -0.0847   0.0144  -0.0109  -0.0886   1.0000
```

Corr(profocc,edu)>0

- From top to bottom regression, we omit education, drop in R squared when omitted educ,
- and change in parameters – we omitted a relevant variable experience that is correlated with the other independent variables exper and profocc

Corr(exp,edu)<0

Omitted Variable Bias

- General Issue- why do we add variables to a regression?
  - To improve the estimation (R squared) and
  - to control for that added variable.

- When controlling for an additional variable do we affect the estimated parameters?

Omitted Variable Bias

4.b. Omitted Variable bias in our example as illustration

- Suppose the model

$$lnwage = \beta_0 + \beta_2 educ + \beta_1\, profocc + \cdots + u,$$

where profocc=0 or 1

- And the **Underspecified** model

$$lnwage = \widetilde{\beta_0} + \widetilde{\beta_1} profocc + \cdots + \tilde{u}$$

Omitted Variable Bias (OVB)

4.b. Omitted Variable bias in our example as illustration

$$\widehat{lnwage} = 0.68 + .069\ \text{educ} + 0.008\ \text{exp} - .315\ \text{female} + .236\ \text{profocc}$$

$$\widehat{lnwage} = 1.55 + \quad + .004\text{exp} \quad - .319\ \text{female} + .432\ \text{profocc}$$

0.236 means that controlling for education (and gender and experience) prof occ respondents earn 23.6% wages above the others.

If we neglect to include education, comparing workers with prof occ to those not (proff occ=0), the difference between the two groups is **0.432**

0.432 (includes profocc and educ effect) because those with profocc are also more educated.

WE WILL SHOW THAT Since corr(profocc,educ)=0.4968 >0 and beta hat of educ also is positive, then ignoring education will bias the beta hat of profocc upwards! We over estimate the effect of profocc on wages if we omit education…. LETs GO !

OVB

- True model y $= \beta_0 + \beta_1 profoc + \beta_2 educ + u$ (A)
- Underspecified model y $= \widetilde{\beta_0} + \widetilde{\beta_1} profoc + \tilde{u}$

Hoes does β_1 relate to $\widetilde{\beta_1}$?

Let us specify educ=a+ρ profoc + v (B)

Where ρ has same sign as corr(educ,profoc) in the population

Substituting (B) into (A)

Y $= \beta_0 + \beta_1 profoc + \beta_2$(a+$\rho$ $profoc$+v) $+ u$

$$Y = \underbrace{\beta_0 + \beta_2 a}_{\widetilde{\beta_0}} + \underbrace{(\beta_1 + \beta_2 \rho)}_{\widetilde{\beta_1}}\ profoc + \underbrace{\beta_2 v + . u}_{\tilde{u}}$$

OVB β_1

So, $\widetilde{\beta_1} = \beta_1 + \beta_2 \ \rho$

biased True value

Same sign as correlation

$$\widetilde{\beta_{profocc}} = \overset{+}{\beta_{profocc}} + \overset{+}{\beta_{edu}} \ \overset{+}{\rho} > \ \beta_{profocc} > 0 \quad \Rightarrow \widetilde{\beta_{profocc}} > \beta_{profocc}$$

Corr(profocc,edu)>0

. correlate lwage educ exp female profocc nonwhite (obs=526)

	lwage	educ	exper	female	profocc	nonwhite
lwage	1.0000					
educ	0.4311	1.0000				
exper	0.1114	-0.2995	1.0000			
female	-0.3737	-0.0850	-0.0416	1.0000		
profocc	0.4451	0.4968	-0.0056	-0.1774	1.0000	
nonwhite	-0.0389	-0.0847	0.0144	-0.0109	-0.0886	1.0000

Corr(profocc,edu)>0

Omitting education will bias $\widetilde{\beta_{profocc}}$ upwards

$$\Rightarrow \widetilde{\beta_{profocc}} > \beta_{profocc}$$

Omitted Variable Bias (OVB)

4.b. Omitted Variable bias in our example as illustration

$$\widehat{lnwage} = 0.68 + .069\ educ + 0.008\ exp - .315\ female + .236\ profocc$$

$$\widehat{lnwage} = 1.55 + \qquad + .004 exp \qquad - .319\ female + .432\ profocc$$

0.236 means that controlling for education (and gender and experience) prof occ respondents earn 23.6% wages above the others.

If we neglect to include education, comparing workers with prof occ to those not (proff occ=0), the difference between the two groups is **0.432**

0.432 (includes profocc and educ effect) because those with profocc are also more educated.

WE SHOWED THAT Since corr(profocc,educ)=0.4968 >0 and beta hat of educ also is positive, then ignoring education will bias the beta hat of profocc upwards! We over estimate the effect of profocc on wages if we omit education….

Daily assignment 6:
OVB omitting education will bias experience coefficient?

- True model y $= \beta_0 + \beta_1 exper + \beta_2 educ + u$ (A)
- Underspecified model y $= \widetilde{\beta_0} + \widetilde{\beta_1}\, exper + \tilde{u}$

Hoes does β_1 relate to $\widetilde{\beta_1}$?

Let us specify educ=a+ρ exper + v (B)

Where ρ has same sign as corr(educ,exper) in the population

Substituting (B) into (A)

Y $= \beta_0 + \beta_1 exper + \beta_2$(a+$\rho$ $exper$+v) $+ u$

$$Y= \underbrace{\beta_0 + \beta_2 a}_{\widetilde{\beta_0}} + \underbrace{(\beta_1 + \beta_2 \rho)}_{\widetilde{\beta_1}}\ exper\ \underbrace{\beta_2\, v}_{\tilde{u}} +.\ u$$

OVB β_1

So, $\widetilde{\beta_1} = \beta_1 + \beta_2 \ \rho$

biased

True value

Same sign as correlation

$\widetilde{\beta_{exper}} = \overset{+}{\beta_{exper}} + \overset{+}{\beta_{edu}} \ \overset{-}{\rho} \ < \ \beta_{exper}$ >0 => $\widetilde{\beta_{exper}} < \beta_{exper}$

Corr(exper,edu)<0

Omitting education will bias $\widetilde{\beta_{exper}}$ downwards

=> $\widetilde{\beta_{exper}} < \beta_{exper}$

Coming up Lecture 7

5. Statistical Properties of Estimator beta_hat

Assumptions for Multiple Linear Regression (MLR)

MLR1

MLR2

MLR3

MLR4

Multicollinearity

6. Stat Property MLR5 → var(beta_hat)

7. Goodness of Fit

Study chapters 3.3, 3.4, 3.6

Posted DA 6

Problem set 2 will be posted soon if not already

Ungraded Quiz 2 coming up and is posted for you to check how you are learning the material