Econometrics I

Lecture 5: Extended Example: The Wage Equation

Chris Conlon

Fall 2025

Mincerian Regression

- Recall the Mincerian regression (wage equation):

$$\ln wage_i = \beta_0 + \beta_{ed} Education_i + \beta_{exp} Experience_i + \beta_{Fem} Female_i + \cdots + \varepsilon_i$$

- Let's revisit estimating this with the Cornwell and Rupert (NLSY) data.

Process the data

```r
suppressMessages(library(tidyverse))
suppressMessages(library(fixest))
suppressMessages(library(marginaleffects))

# first, the Cornwell and Rupert regression
data <- read.csv('./cornwell-rupert.csv') %>%
  mutate(ED_LEVEL=cut(ED,c(0,8,11,12,15,16,17),
         labels = c("NOHS", "SOMEHS", "HS", "SOMECOL","COL","POST"),
         right=TRUE))

# check that we did it correctly
table(data$ED,data2$ED_LEVEL)
```

Interpreting $\widehat{\beta}$

```r
reg_1 <- feols(LWAGE ~ ED + EXP + I(EXP^2) + WKS + OCC + SOUTH + SMSA
  + MS + UNION + FEM, data = data)

# dropping the constant
reg_2 <- feols(LWAGE ~ -1 + i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                 SOUTH + SMSA + MS + UNION + FEM, data = data2)

# not dropping the constant -- which category is omitted?
reg_3 <- feols(LWAGE ~ 1+ i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                 SOUTH + SMSA + MS + UNION + FEM, data = data2)

# change the omitted category -- how do coefficients change?
reg_4 <- feols(LWAGE ~ 1+ i(ED_LEVEL,ref="COL") + EXP + I(EXP^2) +
                 WKS + OCC + SOUTH + SMSA + MS + UNION + FEM, data = data2)

etable(list(reg_1,reg_2,reg_3,reg_4), export='./table_1.png')
```

Notice I've used the `i(·)` command to make categorical variables into dummies and `I(·)` to do polynomial terms.

Dependent Variable:	LWAGE			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	5.245***		5.655***	6.161***
	(0.0717)		(0.0634)	(0.0597)
ED	0.0565***			
	(0.0026)			
EXP	0.0404***	0.0410***	0.0410***	0.0410***
	(0.0022)	(0.0022)	(0.0022)	(0.0022)
I(I(EXP$\hat{2}$))	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	(4.78×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})
WKS	0.0045***	0.0046***	0.0046***	0.0046***
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
OCC	-0.1405***	-0.1386***	-0.1386***	-0.1386***
	(0.0147)	(0.0151)	(0.0151)	(0.0151)
SOUTH	-0.0721***	-0.0762***	-0.0762***	-0.0762***
	(0.0125)	(0.0126)	(0.0126)	(0.0126)
SMSA	0.1390***	0.1436***	0.1436***	0.1436***
	(0.0121)	(0.0121)	(0.0121)	(0.0121)
MS	0.0674***	0.0692***	0.0692***	0.0692***
	(0.0206)	(0.0207)	(0.0207)	(0.0207)
UNION	0.0901***	0.0940***	0.0940***	0.0940***
	(0.0129)	(0.0130)	(0.0130)	(0.0130)
FEM	-0.3892***	-0.3819***	-0.3819***	-0.3819***
	(0.0252)	(0.0253)	(0.0253)	(0.0253)
ED_LEVEL = NOHS		5.655***		-0.5066***
		(0.0634)		(0.0284)
ED_LEVEL = SOMEHS		5.795***	0.1400***	-0.3666***
		(0.0624)	(0.0249)	(0.0236)
ED_LEVEL = HS		5.903***	0.2482***	-0.2584***
		(0.0609)	(0.0229)	(0.0194)
ED_LEVEL = SOMECOL		5.991***	0.3364***	-0.1702***
		(0.0610)	(0.0268)	(0.0206)
ED_LEVEL = COL		6.161***	0.5066***	
		(0.0597)	(0.0284)	
ED_LEVEL = POST		6.188***	0.5337***	0.0271
		(0.0589)	(0.0295)	(0.0213)
Fit statistics				
Observations	4,165	4,165	4,165	4,165
R^2	0.41826	0.41724	0.41738	0.41738

Interpreting $\widehat{\beta}$

```r
reg_1 <- feols(LWAGE ~ ED + EXP + I(EXP^2) + WKS + OCC + SOUTH + SMSA
  + MS + UNION + FEM, data = data)

# dropping the constant
reg_2 <- feols(LWAGE ~ -1 + i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                    SOUTH + SMSA + MS + UNION + FEM, data = data2)

# not dropping the constant -- which category is omitted?
reg_3 <- feols(LWAGE ~ 1+ i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                    SOUTH + SMSA + MS + UNION + FEM, data = data2)

# change the omitted category -- how do coefficients change?
reg_4 <- feols(LWAGE ~ 1+ i(ED_LEVEL,ref="COL") + EXP + I(EXP^2) +
                    WKS + OCC + SOUTH + SMSA + MS + UNION + FEM, data = data2)

etable(list(reg_1,reg_2,reg_3,reg_4), export='./table_1.png')
```

Note on interpreting effects with $\log(y_i) \approx 1 + \beta$:

- $\exp(-.3892) = .6826$
- $\exp(.05654) = 1.057$

Dependent Variable:	LWAGE			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	5.245***		5.655***	6.161***
	(0.0717)		(0.0634)	(0.0597)
ED	0.0565***			
	(0.0026)			
EXP	0.0404***	0.0410***	0.0410***	0.0410***
	(0.0022)	(0.0022)	(0.0022)	(0.0022)
I(I(EXP$\hat{2}$))	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	(4.78×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})
WKS	0.0045***	0.0046***	0.0046***	0.0046***
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
OCC	-0.1405***	-0.1386***	-0.1386***	-0.1386***
	(0.0147)	(0.0151)	(0.0151)	(0.0151)
SOUTH	-0.0721***	-0.0762***	-0.0762***	-0.0762***
	(0.0125)	(0.0126)	(0.0126)	(0.0126)
SMSA	0.1390***	0.1436***	0.1436***	0.1436***
	(0.0121)	(0.0121)	(0.0121)	(0.0121)
MS	0.0674***	0.0692***	0.0692***	0.0692***
	(0.0206)	(0.0207)	(0.0207)	(0.0207)
UNION	0.0901***	0.0940***	0.0940***	0.0940***
	(0.0129)	(0.0130)	(0.0130)	(0.0130)
FEM	-0.3892***	-0.3819***	-0.3819***	-0.3819***
	(0.0252)	(0.0253)	(0.0253)	(0.0253)
ED_LEVEL = NOHS		5.655***		-0.5066***
		(0.0634)		(0.0284)
ED_LEVEL = SOMEHS		5.795***	0.1400***	-0.3666***
		(0.0624)	(0.0249)	(0.0236)
ED_LEVEL = HS		5.903***	0.2482***	-0.2584***
		(0.0609)	(0.0229)	(0.0194)
ED_LEVEL = SOMECOL		5.991***	0.3364***	-0.1702***
		(0.0610)	(0.0268)	(0.0206)
ED_LEVEL = COL		6.161***	0.5066***	
		(0.0597)	(0.0284)	
ED_LEVEL = POST		6.188***	0.5337***	0.0271
		(0.0589)	(0.0295)	(0.0213)
Fit statistics				
Observations	4,165	4,165	4,165	4,165
R^2	0.41826	0.41724	0.41738	0.41738

Interpreting $\widehat{\beta}$

```r
reg_1 <- feols(LWAGE ~ ED + EXP + I(EXP^2) + WKS + OCC + SOUTH + SMSA
  + MS + UNION + FEM, data = data)

# dropping the constant
reg_2 <- feols(LWAGE ~ -1 + i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                 SOUTH + SMSA + MS + UNION + FEM, data = data2)

# not dropping the constant -- which category is omitted?
reg_3 <- feols(LWAGE ~ 1+ i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                 SOUTH + SMSA + MS + UNION + FEM, data = data2)

# change the omitted category -- how do coefficients change?
reg_4 <- feols(LWAGE ~ 1+ i(ED_LEVEL,ref="COL") + EXP + I(EXP^2) +
                 WKS + OCC + SOUTH + SMSA + MS + UNION + FEM, data = data2)

etable(list(reg_1,reg_2,reg_3,reg_4), export='./table_1.png')
```

- In Regression #1, what does the coeffcient β_{ED} mean?
- What about in Regression #2? What is the interpretation of $\beta_{somecol}$?
- How about in Regression #3? and #4?

Dependent Variable:	LWAGE			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	5.245***		5.655***	6.161***
	(0.0717)		(0.0634)	(0.0597)
ED	0.0565***			
	(0.0026)			
EXP	0.0404***	0.0410***	0.0410***	0.0410***
	(0.0022)	(0.0022)	(0.0022)	(0.0022)
I(I(EXP^2))	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	(4.78×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})
WKS	0.0045***	0.0046***	0.0046***	0.0046***
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
OCC	-0.1405***	-0.1386***	-0.1386***	-0.1386***
	(0.0147)	(0.0151)	(0.0151)	(0.0151)
SOUTH	-0.0721***	-0.0762***	-0.0762***	-0.0762***
	(0.0125)	(0.0126)	(0.0126)	(0.0126)
SMSA	0.1390***	0.1436***	0.1436***	0.1436***
	(0.0121)	(0.0121)	(0.0121)	(0.0121)
MS	0.0674***	0.0692***	0.0692***	0.0692***
	(0.0206)	(0.0207)	(0.0207)	(0.0207)
UNION	0.0901***	0.0940***	0.0940***	0.0940***
	(0.0129)	(0.0130)	(0.0130)	(0.0130)
FEM	-0.3892***	-0.3819***	-0.3819***	-0.3819***
	(0.0252)	(0.0253)	(0.0253)	(0.0253)
ED_LEVEL = NOHS		5.655***		-0.5066***
		(0.0634)		(0.0284)
ED_LEVEL = SOMEHS		5.795***	0.1400***	-0.3666***
		(0.0624)	(0.0249)	(0.0236)
ED_LEVEL = HS		5.903***	0.2482***	-0.2584***
		(0.0609)	(0.0229)	(0.0194)
ED_LEVEL = SOMECOL		5.991***	0.3364***	-0.1702***
		(0.0610)	(0.0268)	(0.0206)
ED_LEVEL = COL		6.161***	0.5066***	
		(0.0597)	(0.0284)	
ED_LEVEL = POST		6.188***	0.5337***	0.0271
		(0.0589)	(0.0295)	(0.0213)
Fit statistics				
Observations	4,165	4,165	4,165	4,165
R^2	0.41826	0.41724	0.41738	0.41738

Formulating Linear Hypotheses

```
> print(linearHypothesis(reg_5, c("FEM = MALE"),
+                        vcoev = vcovHC(reg_5, type = "HC1")))

Linear hypothesis test:
FEM - MALE = 0

Model 1: restricted model
Model 2: LWAGE ~ -1 + ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA + MS +
    UNION + FEM + MALE

  Res.Df Df  Chisq Pr(>Chisq)
1   4155
2   4154  1 238.93  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> print(linearHypothesis(reg_6, c("FEM = 0"),
+                        vcov = vcovHC(reg_6, type = "HC1")))

Linear hypothesis test:
FEM = 0

Model 1: restricted model
Model 2: LWAGE ~ 1 + ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA + MS +
    UNION + FEM

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1   4155
2   4154  1 263.33  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can test the same hypothesis with an F-test two different ways:

- $H_0 : \beta_M = \beta_F$ if we omit the constant.
- $H_0 : \beta_F = 0$ if we include the constant.

Correlation in F-Tests

```r
# Number of observations
n <- 200
# Generate random data
# Random integer values between 18 and 65
AGE <- sample(22:65, n, replace = TRUE)
# same as age-22 but one less for some observations
EXP <- AGE - 22 - rbinom(n=n,size=1,prob=0.4)
EXP[EXP < 0] <- 0 # replace negative values with 0
LWAGE = 2.5 + .02*AGE + .03*EXP + .5*rnorm(n)
# create data frame
df <- data.frame(LWAGE,AGE,EXP)

# estimate OLS
reg <- feols(LWAGE ~ AGE + EXP, data = df)
summary(reg)
linearHypothesis(reg,c("AGE+EXP=0"),
        vcov = vcovHC(reg, type = "HC1"))
```

```
> reg <- feols(LWAGE ~ AGE + EXP, data = df) # estimate OLS
> summary(reg)
OLS estimation, Dep. Var.: LWAGE
Observations: 200
Standard-errors: IID
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 2.082345   1.630139 1.277404  0.20296
AGE         0.040453   0.072772 0.555890  0.57892
EXP         0.005372   0.072809 0.073781  0.94126
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.495771   Adj. R2: 0.585937
> linearHypothesis(reg,c("AGE+EXP=0"),
+                  vcov = vcovHC(reg, type = "HC1"))

Linear hypothesis test:
AGE  + EXP = 0

Model 1: restricted model
Model 2: LWAGE ~ AGE + EXP

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1    198
2    197  1 317.52  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notice that t-stats are not significant but F-test is huge, why?

Marginal Effects

Our regression specification contained both linear terms and quadratic terms for experience

$$\ln wage_i = \beta_0 + \beta_{exp} Experience_i + \beta_{exp^2} Experience_i^2 + \beta X_i + \varepsilon_i$$

We can compute the marginal effect of an additional year of experience as:

$$\frac{\partial \ln wage_i}{\partial Experience_i} = \beta_{exp} + 2\beta_{exp^2} Experience_i \equiv g(\boldsymbol{\beta}).$$

- Note: We cannot simply interpret β_{exp} or β_{exp^2} on their own.
- To compute standard errors on marginal effects, we can't just look at the t-stats.
- This is also an issue in nonlinear models like Probit and Logit (later).

Delta Method II

- Suppose we have an asymptotic distribution for an estimator:

$$\sqrt{n}\left(\boldsymbol{b}-\boldsymbol{\beta}\right) \Rightarrow_d \mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}\right).$$

- Then the asymptotic distribution of a function of the estimator is

$$\sqrt{n}\left(g\left(\boldsymbol{b}\right)-g\left(\boldsymbol{\beta}\right)\right) \Rightarrow_d \mathcal{N}\left(\mathbf{0},\left(\nabla g\left(\boldsymbol{\beta}\right)\right)'\boldsymbol{\Sigma}\nabla g\left(\boldsymbol{\beta}\right)\right),$$

  where $\nabla g\left(\boldsymbol{\beta}\right)$ is the gradient of $g\left(\boldsymbol{\beta}\right)$:

$$\nabla g\left(\boldsymbol{\beta}\right) = \left(\ \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_2}, \cdots, \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_K}\ \right)^T.$$

- Note that we can estimate $\nabla g\left(\boldsymbol{\beta}\right)$ with $\nabla g\left(\boldsymbol{b}\right)$.
- Notice how the covariance between the coefficients $\boldsymbol{\Sigma}$ matters!
- What is the expression for the standard error of the marginal effect of experience?

Delta Method / Marginal Effects in R

```
>
> library(marginaleffects)
> avg_slopes(reg_2, variables = "EXP", vcov = "HC1")

 Estimate Std. Error    z Pr(>|z|)     S  2.5 % 97.5 %
   0.0134   0.000591 22.8   <0.001 378.7 0.0123 0.0146

Term: EXP
Type: response
Comparison: dY/dX
```

Here we get the answer correct

What is the exercise we perform in each case?

```
> # don't use the I() to construct interactions
> reg_wrong <- feols(LWAGE ~ -1 + i(ED_LEVEL) + EXP + EXP2 + WKS + OCC +
+                       SOUTH + SMSA + MS + UNION + FEM,
+                     data = data %>% mutate(EXP2 = EXP^2))
> avg_slopes(reg_wrong, variables = "EXP", vcov = "HC1")

 Estimate Std. Error    z Pr(>|z|)     S  2.5 % 97.5 %
    0.041     0.0022 18.6   <0.001 254.8 0.0367 0.0453

Term: EXP
Type: response
Comparison: dY/dX
```

Here we get the answer wrong

Bootstrap/ Simulated Asymptotic Distribution

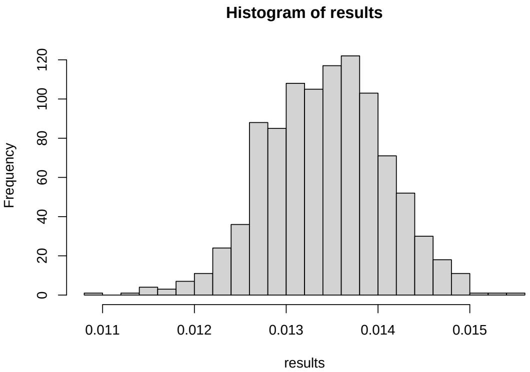
- Given the the asymptotic distribution of a parameter estimate

$$\boldsymbol{b} \sim_d \mathcal{N}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right),$$

  we have an estimated density function $\hat{f}$. Let $\hat{f}$ be the multivariate normal density with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma}$.
- We can simulate the asymptotic distribution of $g\left(\boldsymbol{b}\right)$ by
  - Simulating draws $\boldsymbol{b}_m$ for $m = 1, 2, \ldots M$ from $\hat{f}$
  - Computing $g\left(\boldsymbol{b}_m\right)$ for each draw
  - Then $\left(g\left(\boldsymbol{b}_1\right), g\left(\boldsymbol{b}_2\right), \ldots, g\left(\boldsymbol{b}_M\right)\right)$ will be a simulated asymptotic distribution for $g\left(\boldsymbol{b}\right)$
- This can be useful when you have code to compute $g\left(\cdot\right)$, but computing the derivative $g\left(\cdot\right)$ would be difficult. For example, when $g\left(\cdot\right)$ represents an complex behavioral (or equilibrium) model.
- Also if you are too lazy to load `marginaleffects`

Boostrap Marginal Effects in R

```r
# quick bootstrap comparison
n <- dim(data)[1]
results <- rep(0,1000)
for (i in 1:1000){
  my_weights <-  rmultinom(1,n,rep(1/n,n))/n
  my_reg <- feols(LWAGE ~ -1 + i(ED_LEVEL) + EXP + I(EXP^2) + WKS + OCC +
                  SOUTH + SMSA + MS + UNION + FEM,
                  data = data,
                  # multiplier bootstrap
                  weights=my_weights)
  results[i]<- my_reg$coefficients[7] +
    sum(2 * my_reg$coefficients[8] * my_weights *data$EXP)
}
print(mean(results))
# [1] 0.01342801
print(quantile(results,c(0.025,.975)))
#2.5%      97.5%
#   0.01219617 0.01463365
hist(results,20)
```



How does the distribution of results compare to delta method? Can we do better?

Heterogeneous Effects

- When we have a model of the form
$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$
we're implicitly saying that the effect of X_1 is the same for all individuals.
- Often we would like to relax this, allowing different groups to have different slopes with respect to X_1.
- This is easy as long as the group membership is observed in the data. We simply interact the regressor with dummy variables:
$$Y_i = \beta_0 + \beta_{0F} D_{Fi} + \beta_1 X_{1i} + \beta_2 X_{1i} D_{Fi} + \varepsilon_i$$
where D_{Fi} is a dummy variable for whether individual i is female. Note that we have allowed for the intercepts and slopes to vary by sex here.
- Run this on your own and experiment with `i`$(\cdot)$ operator and : and $*$.

Mincerian Regression: Measurement Error

- What happens if one of the variables of interest is measured with error?
- Let's say the the recorded education might be one year more or less than the person's actual education.
- Note: this may already be happening in the data, but let's make it happen more.

```
noise <- sample(-1:1,dim(data)[1],replace=T)

reg_8 <- feols(LWAGE ~ ED_NOISY + EXP + I(EXP^2) + WKS + OCC + SOUTH + SMSA
+ MS + FEM + UNION, data = data %>% mutate(ED_NOISY = + noise))
summary(reg_8)

etable(list(reg_1,reg_8), export='./table_noise.png')
```

Dependent Variable:	LWAGE	
Model:	(1)	(2)
Variables		
Constant	5.245***	6.154***
	(0.0717)	(0.0613)
ED	0.0565***	
	(0.0026)	
EXP	0.0404***	0.0381***
	(0.0022)	(0.0023)
I(I(EXP$\hat{2}$))	-0.0007***	-0.0007***
	(4.78×10^{-5})	(5.04×10^{-5})
WKS	0.0045***	0.0035***
	(0.0011)	(0.0011)
OCC	-0.1405***	-0.3169***
	(0.0147)	(0.0129)
SOUTH	-0.0721***	-0.1088***
	(0.0125)	(0.0131)
SMSA	0.1390***	0.1652***
	(0.0121)	(0.0127)
MS	0.0674***	0.0833***
	(0.0206)	(0.0217)
UNION	0.0901***	0.0637***
	(0.0129)	(0.0135)
FEM	-0.3892***	-0.4154***
	(0.0252)	(0.0265)
ED_NOISY		-0.0104
		(0.0071)
Fit statistics		
Observations	4,165	4,165
R^2	0.41826	0.35299
Adjusted R^2	0.41686	0.35143

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Omitted Variables Bias, Revisited

- Suppose the econometrician only observes regressors $\mathbf{X}$, but the true model is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{z}\gamma + \varepsilon,$$

- The OLS estimator will equal

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\gamma + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

- The last term is mean zero given the strict exogeneity assumption.
- Note that the second term will not be zero if $\mathbf{X}$ and $\mathbf{z}$ are correlated; i.e. if $\mathbf{X}'\mathbf{z} \neq 0$.
- Implication: correlation between omitted variables and the observed regressors makes OLS biased.

Omitted Variables Bias II

- Using the Frisch-Waugh theorem, we can show that

$$E\left[b_{OLS,k}|\boldsymbol{X},\boldsymbol{z}\right] = \beta_k + \gamma\left(\frac{Cov\left(z, x_k|\boldsymbol{X}_{-k}\right)}{Var\left(x_k|\boldsymbol{X}_{-k}\right)}\right)$$

  where $\boldsymbol{X}_{-k}$ refers to all the regressors besides x_k.
- Suppose positive correlation between regressor x_k and omitted variable z.
- Also suppose $\beta_k > 0$ and $\gamma > 0$ so both variables have positive effects.
- Let's compare the average value of the dependent variable for $x_k = 0$ and $x_k = 1$. Two things change between these points:
  - Dependent variable Y increases by β_k because of direct effect of x_k.
  - Value of z should be higher because of the positive correlation between x_k and z. Higher values of z also contribute to a higher dependent variable because $\gamma > 0$.

Omitted Variables Bias in Mincerian Regression

- What sort of variables might the wage equation omit, and how would you expect them to affect the estimated coefficients?

Dependent Variable:	LWAGE			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	5.245***		5.655***	6.161***
	(0.0717)		(0.0634)	(0.0597)
ED	0.0565***			
	(0.0026)			
EXP	0.0404***	0.0410***	0.0410***	0.0410***
	(0.0022)	(0.0022)	(0.0022)	(0.0022)
I(I(EXP2̂))	-0.0007***	-0.0007***	-0.0007***	-0.0007***
	(4.78×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})	(4.8×10^{-5})
WKS	0.0045***	0.0046***	0.0046***	0.0046***
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
OCC	-0.1405***	-0.1386***	-0.1386***	-0.1386***
	(0.0147)	(0.0151)	(0.0151)	(0.0151)
SOUTH	-0.0721***	-0.0762***	-0.0762***	-0.0762***
	(0.0125)	(0.0126)	(0.0126)	(0.0126)
SMSA	0.1390***	0.1436***	0.1436***	0.1436***
	(0.0121)	(0.0121)	(0.0121)	(0.0121)
MS	0.0674***	0.0692***	0.0692***	0.0692***
	(0.0206)	(0.0207)	(0.0207)	(0.0207)
UNION	0.0901***	0.0940***	0.0940***	0.0940***
	(0.0129)	(0.0130)	(0.0130)	(0.0130)
FEM	-0.3892***	-0.3819***	-0.3819***	-0.3819***
	(0.0252)	(0.0253)	(0.0253)	(0.0253)
ED_LEVEL = NOHS		5.655***		-0.5066***
		(0.0634)		(0.0284)
ED_LEVEL = SOMEHS		5.795***	0.1400***	-0.3666***
		(0.0624)	(0.0249)	(0.0236)
ED_LEVEL = HS		5.903***	0.2482***	-0.2584***
		(0.0609)	(0.0229)	(0.0194)
ED_LEVEL = SOMECOL		5.991***	0.3364***	-0.1702***
		(0.0610)	(0.0268)	(0.0206)
ED_LEVEL = COL		6.161***	0.5066***	
		(0.0597)	(0.0284)	
ED_LEVEL = POST		6.188***	0.5337***	0.0271
		(0.0589)	(0.0295)	(0.0213)
Fit statistics				
Observations	4,165	4,165	4,165	4,165
R^2	0.41826	0.41724	0.41738	0.41738