Financial Econometrics
Shanghai Jiao Tong University
Summer School – 2025
Empirical Exercise 5B
With Solutions

Fatih Kansoy

Reference: Stock, J.H. and Watson, M.W., Introduction to Econometrics (4th Edition)

Data: `Earnings_and_Height.csv`

Empirical Exercise 7.2. In the empirical exercises on earnings and height in Chapters 4 and 5, you estimated a relatively large and statistically significant effect of a worker's height on his or her earnings. One explanation for this result is omitted variable bias: Height is correlated with an omitted factor that affects earnings. For example, Case and Paxson (2008) suggest that cognitive ability (or intelligence) is the omitted factor. The mechanism they describe is straightforward: Poor nutrition and other harmful environmental factors in utero and in early childhood have, on average, deleterious effects on both cognitive and physical development. Cognitive ability affects earnings later in life and thus is an omitted variable in the regression.

(a) Suppose that the mechanism described above is correct. Explain how this leads to omitted variable bias in the OLS regression of *Earnings* on *Height*. Does the bias lead the estimated slope to be too large or too small? [*Hint:* Review Equation (6.1).]

If the mechanism described above is correct, the estimated effect of height on earnings should disappear if a variable measuring cognitive ability is included in the regression. Unfortunately, there isn't a direct measure of cognitive ability in the data set, but the data set does include years of education for each individual. Because students with higher cognitive ability are more likely to attend school longer, years of education might serve as a control variable for cognitive ability; in this case, including education in the regression will eliminate, or at least attenuate, the omitted variable bias problem.

Use the years of education variable (*educ*) to construct four indicator variables for whether a worker has less than a high school diploma ($LT\_HS = 1$ if $educ < 12$, 0 otherwise), a high school diploma ($HS = 1$ if $educ = 12$, 0 otherwise), some college ($Some\_Col = 1$ if $12 < educ < 16$, 0 otherwise), or a bachelor's degree or higher ($College = 1$ if $educ \geq 16$, 0 otherwise).

(b) Focusing first on women only, run a regression of (1) *Earnings* on *Height* and (2) *Earnings* on *Height*, including *LT_HS*, *HS*, and *Some_Col* as control variables.

(i) Compare the estimated coefficient on *Height* in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.

(ii) The regression omits the control variable *College*. Why?

(iii) Test the joint null hypothesis that the coefficients on the education variables are equal to 0.

(iv) Discuss the values of the estimated coefficients on *LT_HS*, *HS*, and *Some_Col*. (Each of the estimated coefficients is negative, and the coefficient on *LT_HS* is more negative than the coefficient on *HS*, which in turn is more negative than the coefficient on *Some_Col*. Why? What do the coefficients measure?)

(c) Repeat (b), using data for men.

Solution:

Key Concepts

Omitted Variable Bias Formula (Equation 6.1):

$$\text{Bias}(\hat{\beta}_1) = \beta_2 \cdot \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

where X_2 is the omitted variable, β_2 is its true coefficient, and X_1 is the included regressor.

Joint Hypothesis Test (F-test):

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0$$

Dummy Variable Trap: With J categories, include only $J - 1$ dummies. The omitted category is the reference group.

Data Overview

Variable	Description	Mean	SD	Min	Max
earnings	Annual earnings ($)	46,875	26,923	4,726	84,055
height	Height (inches)	66.96	3.97	48	84
educ	Years of education	13.54	2.64	0	19

n = 17,870. Women: 9,974 (55.8%). Men: 7,896 (44.2%).

Education Categories:

Variable	Definition	% of Sample
LT_HS	Less than high school (educ < 12)	10.1%
HS	High school diploma (educ = 12)	36.4%
Some_Col	Some college (12 < educ < 16)	24.9%
College	Bachelor's or higher (educ ≥ 16)	28.7%

(a) Omitted Variable Bias Theory

The Case-Paxson Mechanism:

- Poor early-life conditions → lower cognitive ability AND shorter height
- Cognitive ability → higher earnings
- Height is positively correlated with cognitive ability (both affected by early nutrition)

Applying the OVB Formula:
Let X_1 = Height, X_2 = Cognitive Ability (omitted), Y = Earnings.

$$\text{Bias}(\hat{\beta}_{\text{Height}}) = \underbrace{\beta_{\text{CogAbility}}}_{>0} \times \underbrace{\frac{\text{Cov}(\text{Height}, \text{CogAbility})}{\text{Var}(\text{Height})}}_{>0} > 0$$

Answer: The bias is **positive**.
The simple regression of Earnings on Height **overestimates** the true causal effect of height. Part of what appears to be a "height effect" is actually the effect of cognitive ability, which is correlated with height.
If we could control for cognitive ability, the height coefficient would **decrease** (become smaller in magnitude or possibly become insignificant).

(b) Women Only Analysis

Regression Results

	Reg (1) Height Only	**Reg (2)** + Education
Height	511.22^{***}	135.14
	(98.90)	(92.55)
LT_HS		$-31{,}857.81^{***}$
		(963.77)
HS		$-20{,}417.89^{***}$
		(626.18)
Some_Col		$-12{,}649.07^{***}$
		(685.30)
Constant	$12{,}650.86^{**}$	$50{,}749.52^{***}$
	(6,383.74)	(6,013.68)
R^2	0.0027	0.1382
n	9,974	9,974

Standard errors in parentheses. $^{**}p < 0.05$, $^{***}p < 0.01$

(b)(i) Does the Height Coefficient Change?

	Without Education	**With Education**
$\hat{\beta}_{\text{Height}}$	\$511.22	\$135.14
SE	98.90	92.55
t-statistic	5.17	1.46
p-value	0.000	0.144
Change	−\$376.08 (−73.6%)	

Answer: Yes, there is a **dramatic change**. The height coefficient drops by **73.6%** (from \$511 to \$135 per inch) and becomes **statistically insignificant** ($p = 0.144$).

This is **consistent with the cognitive ability explanation**. When we control for education (a proxy for cognitive ability), most of the apparent height effect disappears. This suggests that the simple regression was capturing the correlation between height and cognitive ability, not a direct causal effect of height.

(b)(ii) Why is College Omitted?

Answer: College is omitted to avoid the **dummy variable trap** (perfect multicollinearity).
Since every observation belongs to exactly one education category:

$$\text{LT\_HS} + \text{HS} + \text{Some\_Col} + \text{College} = 1 \quad \text{for all } i$$

Including all four dummies plus a constant would make the design matrix $\mathbf{X}$ have linearly dependent columns, so $(\mathbf{X}'\mathbf{X})$ would be singular and uninvertible.
By omitting College, it becomes the **reference category**. All education coefficients are interpreted as effects *relative to college graduates.*

(b)(iii) Joint F-Test

Hypothesis:

$$H_0 : \beta_{\text{LT\_HS}} = \beta_{\text{HS}} = \beta_{\text{Some\_Col}} = 0$$

$$H_1 : \text{At least one coefficient} \neq 0$$

Test Result:

$$F(3, 9969) = 522.52, \quad p\text{-value} < 0.0001$$

Answer: Reject H_0 decisively. The education variables are **jointly highly significant**. Education is an important determinant of earnings.

(b)(iv) Interpreting the Education Coefficients

Variable	Coefficient	Interpretation
LT_HS	−$31,858	Earns $31,858 less than College grads
HS	−$20,418	Earns $20,418 less than College grads
Some_Col	−$12,649	Earns $12,649 less than College grads
College	0 (reference)	Reference category

Answer: All coefficients are **negative** because the reference group (College) has the **highest earnings**. Each coefficient measures how much *less* that education group earns compared to college graduates.
The ordering $|\beta_{\text{LT\_HS}}| > |\beta_{\text{HS}}| > |\beta_{\text{Some\_Col}}|$ reflects the **education-earnings gradient**:

- Less than high school: largest earnings penalty ($31,858)
- High school only: moderate penalty ($20,418)
- Some college: smallest penalty ($12,649)

This pattern is consistent with human capital theory: more education leads to higher earnings.

(c) Men Only Analysis

Regression Results

	Reg (1) Height Only	**Reg (2)** + Education
Height	1,306.86***	744.68***
	(100.77)	(94.70)
LT_HS		−31,400.49***
		(978.44)
HS		−20,345.85***
		(692.44)
Some_Col		−12,610.92***
		(758.06)
Constant	−43,130.34***	9,862.74
	(7,068.48)	(6,697.04)
R^2	0.0209	0.1658
n	7,896	7,896

(c)(i) Does the Height Coefficient Change?

	Without Education	**With Education**
$\hat{\beta}_{\text{Height}}$	\$1,306.86	\$744.68
SE	100.77	94.70
t-statistic	12.97	7.86
p-value	0.000	0.000
Change	−\$562.18 (−43.0%)	

Answer: Yes, the height coefficient drops by **43%** (from \$1,307 to \$745 per inch). However, unlike for women, the coefficient **remains statistically significant** ($t = 7.86$, $p < 0.001$).
This is **partially consistent** with the cognitive ability explanation: controlling for education reduces the height coefficient substantially. However, a significant height effect persists even after controlling for education, suggesting either:

- Education is an imperfect proxy for cognitive ability
- There are additional mechanisms linking height to earnings for men (e.g., physical presence, discrimination, occupational sorting)

(c)(iii) Joint F-Test

$$F(3, 7891) = 456.94, \quad p\text{-value} < 0.0001$$

Answer: Reject H_0. Education variables are jointly highly significant for men as well.

(c)(iv) Education Coefficients for Men

Variable	**Men**	**Women**
LT_HS	−\$31,400	−\$31,858
HS	−\$20,346	−\$20,418
Some_Col	−\$12,611	−\$12,649

The education coefficients are very similar for men and women, showing a consistent education-earnings gradient across genders.

Summary: Women vs. Men

	Women		Men	
	No Educ	With Educ	No Educ	With Educ
$\hat{\beta}_{\text{Height}}$	\$511	\$135	\$1,307	\$745
SE	98.90	92.55	100.77	94.70
Significant?	Yes	**No**	Yes	Yes
% Change	-73.6%		-43.0%	
R^2	0.003	0.138	0.021	0.166
F-test (educ)	$F = 522.5$		$F = 456.9$	

Conclusions

1. **OVB Theory Confirmed (Partially):** Controlling for education substantially reduces the height coefficient for both genders, consistent with the cognitive ability hypothesis.

2. **Gender Difference:** For women, the height effect disappears entirely (-73.6%, becomes insignificant). For men, the effect is reduced (-43%) but remains large and significant.

3. **Education as Proxy:** Education is an imperfect proxy for cognitive ability. The persistent height effect for men may reflect:
   - Measurement error in the proxy
   - Other channels (discrimination, physical jobs, leadership perceptions)

4. **Joint Significance:** Education variables are highly significant for both genders ($F > 450$), confirming the strong education-earnings relationship.

5. **Dummy Variable Interpretation:** Coefficients on education dummies are negative because they measure earnings relative to the highest-education reference group (College).