14.32: Econometrics

Problem Set 2

due Monday, October 6, 2025 at 1pm

1. Answer the following theoretical questions.

   (a) You have data (Y_i, X_{1i}, X_{2i}) generated from the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$, where e_i satisfies the usual exogeneity condition. You are interested in estimating the causal effect of X_1 on Y. Suppose X_1 and X_2 are positively correlated, and $\beta_2 > 0$. If you regress Y only on X_1, will your estimate of β_1 be biased? In what direction?

   (b) A researcher runs a simple regression of income (Y) on years of education (X_1) and finds a strong positive effect. Another researcher argues that this estimate is likely upward biased because it omits cognitive ability (X_2). Explain why omitted variable bias may be present, what is the most likely sign of the bias and how it affects the interpretation of the education coefficient.

   (c) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$. Transform the regression so that you can use a t-statistic to test

      i $H_0 : \beta_1 = \beta_2$

      ii $H_0 : \beta_1 + 2\beta_2 = 0$

      iii $H_0 : \beta_1 + \beta_2 = 1$ (you may redefine the dependant variable)

   (d) True or False: In a multiple regression model, the coefficient $\hat{\beta}_1$ on regressor X_1 always equals the coefficient from a simple regression of Y on X_1. Briefly justify your answer.

2. Policy makers often debate whether increasing public school funding leads to better student outcomes. This question is particularly important in the context of persistent achievement gaps across states.

   To study this question, a researcher collects data from 50 U.S. states in 2019. The dependent variable is *AvgScore*, the average score on a standardized 8th grade math test (out of 500 points). One key independent variable is *Spend*, per-pupil public school expenditure (in thousands of dollars). Heteroskedasticity-robust standard errors are reported in parentheses.

$$\widehat{AvgScore} = \underset{(15.2)}{392.5} + \underset{(1.25)}{2.90} \times Spend; \quad R^2 = 0.180 \tag{1}$$

(a) Interpret the coefficient on *Spend* in Regression (1). Is it large in a real-world sense? Is it statistically significant?

(b) Suppose Mississippi spends \$8,000 per pupil and Massachusetts spends \$16,000. Predict the difference in average test scores between the two states using Regression (1).

(c) Compute a 95% confidence interval for the coefficient on *Spend*. What does this interval imply about the precision of the estimate?

(d) Do you think the regression error is likely to be homoskedastic or heteroskedastic in this context? Explain briefly.

The researcher suspects that other state-level characteristics may also affect student performance. In particular, states differ in income levels and demographics. She runs a new regression:

$$\widehat{AvgScore} = \underset{(14.8)}{361.2} + \underset{(1.90)}{0.85} \times Spend + \underset{(0.21)}{1.75} \times PctCollege - \underset{(0.12)}{0.62} \times PctPoverty; \quad R^2 = 0.510 \tag{2}$$

Where:

- *PctCollege* is the percent of adults with a college degree or higher,
- *PctPoverty* is the percent of the population below the poverty line.

(e) The coefficient on *Spend* decreased substantially from Regression (1) to Regression (2). Why might this have happened? Explain both the direction and magnitude of the change.

(f) The researcher is concerned about omitted variable bias from not including *ClassSize* (average class size in public schools). Would including this variable likely increase or decrease the estimated effect of *Spend*? Justify your reasoning.

(g) Based on Regression (2), interpret the coefficient on *PctCollege*. What does this suggest about the relationship between adult education levels and student performance?

(h) What is the adjusted R^2 trying to capture in this context? Would you expect it to be higher or lower than the reported R^2?

3. Being tall may pay off (in the literal sense). There is much indirect evidence that being tall gives a person an advantage in terms of higher salary, especially in sales and management. The file Earnings and Height.dta contains data on earnings (annual labor earn- ing of an individual in 2012 in US dollars), height (in inches without shoes), gender and educational attainment for a sample of 17,870 US workers, which is taken to be a subset of data from the US National Health Interview Survey for 1994.

   a. Run an OLS regression of *Earnings* on *Height*. Discuss the interpretation, the sign and the size of the coefficient.

   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.

   b. 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?

   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).

c. Run a regression of *Earnings* on *Height*, including LT_HS, HS, and $Some\_Col$ as control variables.

   i. Compare the estimated coefficient on *Height* in two regressions. 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?)

d. Run an OLS regression of *Height*, on LT_HS, HS, and $Some\_Col$, get residuals from this regression. Regress *Earnings* on the residuals you just obtained, compare the results with the ones you obtained in c. Discuss.