14.32: Econometrics

Problem Set 2

due Thursday, October 12, 2023 at noon

1. Answer the following theoretical questions.

   (a) You have data (Y_i, X_{1i}, X_{2i}) that satisfies the exogeneity conditions for a regression of Y on X_1 and X_2. You are interested in causal effect of X_1 on Y. Suppose X_1 and X_2 are uncorrelated. You estimate the effect of interest by regressing Y on X_1. Does this estimator suffer from omitted bias?

   (b) A recent study found that the death rate for people who sleep 6 to 7 hours per night is lower than the death rate for people who sleep 8 or more hours. The data used is a random survey of Americans aged 30 to 102. Each survey respondent was tracked for 4 years. The death rate for people sleeping 7 hours was calculated as the ratio of the number of deaths over the span of the study among people sleeping 7 hours to the total number of survey respondents who slept 7 hours. This calculation was then repeated for people sleeping 6 hours and so on. Based on this summary would you recommend that Americans who sleep 9 hours per night consider reducing their sleep to 6 or 7 hours if they want to prolong their lives?

   (c) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$. Make transformations to the regression model so that you can use a single t-test on one coefficient to test the following hypotheses. In particular, your regressors should be linear transformations of the data in the original model, constructed in such a way that your new specifications can recover the original model exactly.

      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)

2. Policy debates and personalities get much attention in presidential elections. We examine whether economic conditions also matter. What was the effect of state-level economic conditions on the election of Barack Obama in 2008? We study this question using cross-sectional data on 50 states for 2008.

First, consider regression (1), for which the dependent variable is the percentage of the state vote going to Obama, *Income* is per capita income, in thousands of 2008 dollars. Heteroskedasticity-robust standard errors are in parentheses.

$$\widehat{PctObama} = \underset{(7.4)}{21.6} + \underset{(0.18)}{0.68} \times Income; \qquad (1)$$

$$R^2 = 0.200$$

(a) Interpret the coefficient on *Income* in regression (1). Is it large in real world sense? Is it statistically large?

(b) In your judgment, is the regression error in regression (1) heteroskedastic or homoskedastic? Briefly explain.

(c) The per capita income in Massachusetts is 51.7 (i.e. $ 51,700 in 2008 dollars) and in Indiana is 35.0, for a difference of 16.7. Use regression (1) to predict the difference in percent voting for Obama between Massachusetts and Indiana.

(d) Compute a 95% confidence interval for the predicted difference in part (c).

Now you run a regression (2):

$$\widehat{PctObama} = \underset{(7.7)}{31.3} - \underset{(0.29)}{0.27} \times Income + \underset{(0.26)}{1.03} \times PctCollege - \underset{(0.08)}{0.11} \times PctWhite; \qquad (2)$$

$$R^2 = 0.476$$

where *PctCollege* is percent of the state population with a college degree or higher, *PctWhite* is percent of the state population that is white.

(e) The coefficient on *Income* changes substantially from regression (1) to (2). Explain why the coefficient changed and provide an explanation of the sign (direction) of the change.

(f) Suppose older voters preferred McCain because he was older than Obama. Do you think the coefficient on Income suffers from omitted variable bias because the fraction of older voters in the state has been omitted? If so, how would you expect the coefficient on Income in (2) to change if the fraction of older voters in the state were included as a regressor? Explain.

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.