14.32 Pset 2 Solutions

due Thursday, October 12, 2023

1. Question 1

   a) No. For there to be omitted variable bias in this case, the omitted variable X_2 must be correlated with the other regressor X_1. This estimator will not suffer from committed bias.

   b) No. The death rates calculated are equal to the coefficient from the OLS regression of the form

   $$Death_i = \beta_1 6\_Hours + \beta_2 7\_Hours + \beta_3 8\_Hours + \beta_4 9\_Hours + e_i,$$

   where β_1 is the death rate of people who sleep 6 hours, β_2 is the death rate of people who sleep 7 hours, etc. However, we have no information on whether or not this was a controlled experiment, or whether the exogeneity condition was satisfied. We can easily think of a scenario in which there exists a third missing variable that affects the death rate and is correlated with the other regressor (amount of sleep per night). One such example is age: younger people tend to sleep less every night for multiple reasons- partying, or many hours of work, or attending to small children through the night; while retirees may have a luxury of uninterrupted night sleep. We also expect to see higher death rate among older population. This suggests that the regression above would suffer from OVB, with coefficients β_1 an β_2 likely to be biased downward, while β_4 is likely biased up.

   c) i) To test the hypothesis $H_0 : \beta_1 = \beta_2$, we can transform the regression by regressing Y_i on $(X_{1i}+X_{2i})$ and $(X_{1i}-X_{2i})$. We suggest running the following regression model:

   $$Y_i = \beta_0 + b_1(X_{1i} + X_{2i}) + b_2(X_{1I} - X_{2i}) + e_i \ \textbf{(1)}.$$

   There is one-to-one transformation of the regressors and one-to-one transformation of coefficients, where $\beta_1 = b_1 + b_2$ **(2)** and $\beta_2 = b_1 - b_2$ **(3)**.
   From equation (3), we see that $b_2 = b_1 - \beta_2$, and substituting in equation (2) we have that

   $$\begin{aligned} b_2 &= \beta_1 - b_2 - \beta_2 \\ 2b_2 &= \beta_1 - \beta_2 \\ b_2 &= \frac{\beta_1 - \beta_2}{2} \end{aligned}$$

   Thus, running the transformed regression (1) and testing hypothesis $H_0 : b_2 = 0$ will be equivalent to testing if $\beta_1 = \beta_2$.

ii) Using a similar reasoning to part i), we can set $\beta_2 = \frac{b_2 - b_1}{2}$ and $\beta_1 = b_1 + b_2$. Then, we can rewrite the regression model to be

$$Y_i = \beta_0 + (b_1 + b_2) \cdot X_{1i} + \frac{b_2 - b_1}{2} \cdot X_{2i} + e_i$$
$$Y_i = \beta_0 + b_1 X_{1i} + b_2 X_{1i} + \frac{1}{2} X_{2i} b_2 - \frac{1}{2} X_{2i} b_1 + e_i$$
$$Y_i = \beta_0 + b_1 (X_{1i} - \frac{1}{2} X_{2i}) + b_2 (X_{1i} + \frac{1}{2} X_{2i}) + e_i \quad \textbf{(4)}$$

Running the transformed regression of Y_i on $(X_{1i} - \frac{1}{2}X_{2i})$ and $(X_{1i} + \frac{1}{2}X_{2i})$ and testing the $H_0 : b_1 = 0$ will be equivalent to testing if $\beta_1 + 2\beta_2 = 0$.

iii) Using the same method as parts i) and ii), we have

$$\beta_1 = \frac{b_2 + 1 - b_1}{2}$$

and

$$\beta_2 = \frac{1 - b_1 - b_2}{2}$$

Rearranging, we have

$$b_1 = 1 - \beta_1 - \beta_2$$

and

$$b_2 = \beta_1 - \beta_2$$

Our regression can be rewritten as

$$2Y_i = 2\beta_0 + (b_2 + 1 - b_1) X_{1i} + (1 - b_1 - b_2) X_{2i} + e_i$$
$$2Y_i = 2\beta_0 + b_1 (X_{1i} + X_{2i}) + b_2 (X_{1i} - X_{2i}) + X_{1i} + X_{2i} + e_i$$

Let us define $Y_i^* = 2Y_i - X_{1i} - X_{2i}$, our regression then becomes

$$Y_i^* = 2\beta_0 + b_1 (X_{1i} + X_{2i}) + b_2 (X_{1i} - X_{2i}) + e_i$$

Running the transformed regression and testing the b_1 coefficient will be equivalent to testing if $\beta_1 + \beta_2 = 1$.

2. Question 2

a) We can interpret the coefficient on *Income* as follows: For every \$1000 increase in per capita income, Obama wins 0.68 of a percentage point more of the vote. In a real world sense, this is fairly large, as the range of per capita income between states can be in the tens of thousands of dollars. In a statistical sense, this is also large because the standard error is low (less than a third of the coefficient), indicating that this coefficient is statistically significant.

b) The regression error is likely heteroskedastic because voting habits will fluctuate depending on the income variability in the state. For example, Gelman 2007 finds that in poor states, rich people are likely to vote Republican, while in richer states with higher incomes, even wealthy people tend to vote Democrat. **(Other acceptable answers permitted)**

c) Using the given coefficient of 0.68, we know that the difference in percent voting for Obama between Massachusetts and Indiana will be

$$\begin{aligned} & 0.68 \cdot (51.7 - 35.0) \\ & = 0.68 \cdot 16.7 \\ & = 11.36 \end{aligned}$$

d) We can calculate the confidence set of β by adding and subtracting $1.96 \times$ the standard error, giving us

$$[0.68 - (1.96 \cdot 0.18), 0.68 + (1.96 \cdot 0.18)]$$

$$[0.3272, 1.0328]$$

Using the upper and lower bounds of this confidence set as β, we find that the confidence set for the predicted difference in part c) will be

$$[(16.7 \cdot 0.3272), (16.7 \cdot 1.0328)]$$

$$[5.4642, 17.2478]$$

e) The coefficient changed because the first regression had substantial omitted variable bias. Specifically, the coefficient on $Income$ in the first regression was positively biased because of a missing variable that was positively correlated with both $Income$ and $PctObama$, which is $PctCollege$.

f) Yes, the coefficient on $Income$ could still suffer from OVB. Specifically, assuming that older people tend to earn more ($PctOld$ and $Income$ are positively correlated), and older people are less likely to vote for Obama ($PctOld$ and $PctObama$ are negatively correlated), then the coefficient on $Income$ would be negatively biased, and thus would increase if $PctOld$ were included as a regressor.

3. Question 3

a) [See Stata file for results]. The coefficient on height is positive and quite large. It can be interpreted as follows: Every inch in height increases annual wages by 707 dollars.

b) This could lead to OVB in the regression from part a) because, if all variation in earnings is being attributed to height, rather than being attributed to schooling AND height. If height and schooling are positively correlated as Case and Paxson suggest, and schooling is positively correlated with earnings, then the estimated coefficient is positively biased, meaning it is too large.

c) i) There is a large change in the coefficient, from 0.708 to 0.453. This is consistent with the cognitive ability explanation, as we assumed the first predicted coefficient would be too large.

ii) Because the 4 variables are indicator variables, we omit one of the indicators, and the effect of this variable is captured in the constant term (when all other indicators are 0).

iii) The joint test shows that the probability of all coefficients being 0 is close to 0 with an f-statistic of 1100.65, indicating that amount of schooling almost surely has an effect on earnings.

iv) These coefficients measure the difference in earnings between two people of the same height but with different schooling. As predicted, the less schooling one has, the less they earn. Thus, the indicator for the least amount of schooling (LT_HS), has the most negative coefficient, while the constant term (capturing the effect of the $College$ variable), is positive.

d) [See stata file for results]. The coefficient on the residuals is equivalent to the coefficient on height in part c. This is because the residuals in the regression of $Height$ on LT_HS, HS and $Some\_Col$ contain all variability in $Height$ that is *not* explained by differences in schooling. Thus, when we regress $Earnings$ on these residuals, it is equivalent to regressing $Earnings$ on $Height$ with the schooling variables added as controls.