Introduction to Econometrics
Economics 4811: Spring 2018

Problem Set 2

Due: *Beginning of class, February 14th, 2018*

You may collaborate with other students. Please hand in your own work. At the top of your answer sheet, please identify the students with whom you have consulted to complete this problem set.

The data for this problem set are available on the class Canvas page. Click on the link to the data file and save the data on your local disk.

Question 1

The data **ps2q1.dta** includes information on wages, attained education (*educ*), parent's education (*motheduc* and *fatheduc*), a measure of cognitive ability (*abil*) and several other variables for 1,230 working men in 1991.

(i) Use gen to create $abil^2$ and *ln(motheduc)* and *ln(fatheduc)*. Estimate the following regression:

$$educ_i = \alpha_0 + \alpha_1 \ln(motheduc_i) + \alpha_2 \ln(fatheduc_i) + \alpha_3 abil_i + \alpha_4 abil^2{}_i + e_i$$

Interpret $\hat{\alpha}_1$ *and* $\hat{\alpha}_2$.

(ii) What is the predicted effect of an additional unit of ability on attained education?

(iii) Using the estimates of $\widehat{\alpha_3}$ and $\widehat{\alpha_4}$, use calculus to find the value of *abil* where *educ* is minimized (take the derivative of the estimated regression in (i) relative to *abil* holding parent's education constant; take also the second derivative so you know whether you are calculating a minimum or a maximum.)

(iv) Given your answer in (iii), explain, in a way that your non-Econ, non-mathematical friend would understand, what the result in part (iii) means in terms of the relationship between *educ* and *abil*?

(v) Now estimate the regression:

$$educ_i = \delta_0 + \delta_1 motheduc_i + \delta_2 fatheduc_i + \delta_3 abil_i + u_i$$

Interpret $\hat{\delta}_1, \hat{\delta}_2$ *and* R^2.

(vi) Generate the variable *pareduc=motheduc+fatheduc*. If you were to estimate the regression:

$$educ_i = \beta_0 + \beta_1 pareduc_i + \beta_2 abil_i + \varepsilon_i$$

What assumption would you be making about the effect of father's education and mother's education on education of their children?

(vii) Now consider the regression

$$educ_i = \phi_0 + \phi_1 pareduc_i + \phi_2 fatheduc + \phi_3 abil_i + v_i$$

How is ϕ_1 related to δ_1? Give an intuitive or mathematical explanation for your answer. Check your intuition is correct by estimating this regression and comparing the coefficients.

(viii) Consider the following regressions:

$$educ_i = \gamma_0 + \gamma_1 pareduc_i + \gamma_2 abil_i + r_i$$

$$educ_i = \varphi_0 + \varphi_1 pareduc_i + \omega_i$$

a. Under what condition(s) will your estimate of γ_1 be the same as your estimate of φ_1?
b. Derive a general relationship that links γ_1 with φ_1.
c. Show and estimate the subsidiary regression that will establish whether that condition holds. Show the estimated coefficients.
d. Use estimates from your subsidiary regression to confirm that your general relationship is correct.
e. Can you conclude ability is an omitted variable in $educ_i = \varphi_0 + \varphi_1 pareduc_i + \omega_i$? How was its exclusion biasing the effects of parent's education on education?

(ix) Consider the following regressions:

$$educ_i = \delta_0 + \delta_1 motheduc_i + \delta_2 fatheduc_i + \delta_3 abil_i + u_i$$

$$educ_i = \lambda_0 + \lambda_1 motheduc_i + \lambda_2 fatheduc_i + w_i$$

a. Derive a general relationship that links δ_1 with λ_1.
b. Show and estimate the subsidiary regression(s) that will establish whether that condition holds. Show the estimated coefficients.
c. Use estimates from your subsidiary regression to confirm that your general relationship is correct.

Question 2

Use the data in **ps2q2.dta** to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The unit of analysis is at the school level.

(i) Estimate the relationship between *math4* (as the dependent variable) and *pctsgle* (as the independent variable). What is your estimate of the intercept and the slope? Interpret these estimates. Does your result make economic sense?

(ii) What is the variance of *math4*? What is the total sum of squares from the regression in (i)? How are they related? What is the residual sum of squares from the regression in (i)?

(iii) Compute R^2 using the information in (ii). Confirm your calculation is correct by cross-checking with the STATA output. Interpret R^2.

(iv) A colleague argues that an analysis with the model

$$\ln(math4_i) = \beta_0 + \beta_1 pctsgle_i + u_i$$

is better than the model in part (ii), because the R^2 from a regression of this new model is bigger than the R^2 from the one from part (i). Is the statement about which R^2 is higher true? Do you agree with the economist's argument about why his model is better?

(v) What is the interpretation of $\hat{\beta}_1$ from the model in part (iv)?

(vi) We know that regressions like

$$\ln(math4_i) = \beta_0 + \beta_1 pctsgle_i + u_i$$

are unlikely to be internally valid because of omitted variable bias.

a. Provide an example of an important factor excluded from this model which would cause the estimate of β_1 to get smaller. Explain why that factor would make β_1 get smaller.

b. Provide an example of an important factor excluded from this model which would cause the estimate of β_1 to get larger. Explain why that factor would make β_1 get larger.

(vii) Even if we had all the important other factors in the model, what is another fundamental reason we would be skeptical that this type of model could provide internal validity for the causal relationship between single-parent households and student's math performance?

Question 3

State whether each of the following statements is true, false or uncertain, and provide a brief justification either with a counterexample, argument, or algebra. I need to see some work - No credit is granted for simply saying "true", "false", or "uncertain."

Consider the following two regressions based on a sample of 6000 individuals:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + u_i \qquad (1)$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i \qquad (2)$$

The OLS estimate of α_1 is negative and larger than the OLS estimate of β_1 (i.e. $\hat{\beta}_1 < \hat{\alpha}_1 < 0$)

(i) The covariance between X_1 and Y is positive.

(ii) The R^2 in regression (1) must be smaller (<) than the R^2 in regression (2).

(iii) The SST in regression (2) must be larger (>) than the SST in regression (1).

(iv) If X_1 and X_2 have a negative relationship, then the OLS estimate of β_2 is positive.