Econometrics A (Econ 210)

Problem Set 5

Mohsen Mirtaher - Fall 2015

Due: Nov 12, 2015; TA Session

1. In this question, we will explore how omitting a relevant variable in the regression makes the OLS estimates biased. Suppose the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, but you only observe x_1 and x_2.

   Set $\beta_0 = 1$, $\beta_1 = 1.5$, $\beta_2 = 2$ and $\beta_3 = 1$

   For all the exercises below, simulate:

   - $\varepsilon \sim N(0, 1)$
   - $x_1 \sim N(2, 1.5)$

   (a) Let $n = 100$.

      i. Draw (independently!) $x_2 \sim N(1, 1.5)$ and $x_3 \sim N(1, 1.5)$. Estimate OLS on a constant, x_1 and x_2. Report the estimated values and t values of the coefficients. Are the estimations biased?

      ii. Now draw x_2 and x_3 from a joint normal with mean $[1, 1]$ and variance-covariance $= \begin{pmatrix} 1.5 & 1 \\ 1 & 1.5 \end{pmatrix}$. What happens to the OLS estimates of β_1 and β_2? Why?

      iii. Examine whether changing β_3 increases or reduces the bias?

(b) Now let $n = 1000$ and repeat (i) and (ii). What is the effect of increasing n?

(c) Now repeat (b) with $k = 500$ repetitions, so that you get 500 vectors of $\hat{\beta}^{OLS}$ estimates for each part (i) and (ii). Plot histograms of $\hat{\beta}_2^{OLS}$ corresponding with each part (i) and (ii).

2. A researcher is going to estimate the effect of education on income. She categorizes the education into three categories: High school dropouts, High school graduates, and College graduates. She creates three binary variables corresponding to each educational level as follows: $D_i^{HD} = 1$ for the high school dropouts and zero otherwise, $D_i^{HG} = 1$ for individuals whose highest education level is high school diploma and zero otherwise, and $D_i^C = 1$ for individuals who have college degree and zero otherwise. Y_i is the income of individual i. The sample mean of income for each education category is $\bar{Y}_{HD}$, $\bar{Y}_{HG}$, and $\bar{Y}_C$, respectively. The researcher is thinking about running three models as the following

$$Y_i = \alpha_0 D_i^{HD} + \alpha_1 D_i^{HG} + \alpha_2 D_i^C + \epsilon_i \tag{1}$$

$$Y_i = \beta_0 + \beta_1 D_i^{HG} + \beta_2 D_i^C + \nu_i \tag{2}$$

$$Y_i = \gamma_0 + \gamma_0 D_i^{HD} + \gamma_1 D_i^{HG} + \gamma_2 D_i^C + \mu_i \tag{3}$$

(a) Show that OLS procedure cannot be applied to estimate model (3).

(b) Show that in model (1) $\hat{\alpha}_0^{OLS} = \bar{Y}_{HD}$, $\hat{\alpha}_1^{OLS} = \bar{Y}_{HG}$, $\hat{\alpha}_2^{OLS} = \bar{Y}_C$.

(c) Show that in model (2) $\hat{\beta}_0^{OLS} = \bar{Y}_{HD}$, $\hat{\beta}_1^{OLS} = \bar{Y}_{HG} - \bar{Y}_{HD}$, $\hat{\beta}_2^{OLS} = \bar{Y}_C - \bar{Y}_{HD}$.

(d) Discuss the difference in the interpretation of the coefficients between model (1) and model (2).

3. The file *ceosal2.csv* or *ceosal2.dta* contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.

The **.dta* file has variable labels. For the **.csv* format, you can find a description here:

http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal2.des

(a) Estimate a model relating annual salary to firm sales and market value. Transform the variables so that the coefficients can be interpreted as elasticities. Give a quantitative interpretation of the results.

(b) Add *profits* to the equation above. Now try with the log of profits instead. What happens to the coefficients? Does it make more sense to use the level or the log of profits?

(c) Calculate the correlation of *profits* with the other exogenous variables. What does this imply for the OLS estimates?

(d) Add the variable *ceoten* (CEO tenure) and its square to the model in (b). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?

4. Use the data set cps08.dta. This is CPS data on average hourly earnings. You should report the coefficients and standard errors in one table, to make the comparisons easier.

(a) Regress average hourly earnings on age and dummy variables for gender and a college degree. Report coefficients and standard errors.

(b) Generate an interaction term between female and bachelor, and run a regression including this variable. Report coefficients and standard errors. How would you interpret the coefficient on the interaction term?

(c) Test the null hypothesis that there is no interaction between gender and college degree at the 5% significance level.

(d) Generate an interaction term between bachelor and age. Run a regression including this term in addition to those in part (b), so that your model looks like

$$ahe = \beta_0 + \beta_1 bachelor + \beta_2 female + \beta_3 bachelor \times female + \beta_4 age + \beta_5 age \times bachelor + U$$

Report coefficients and standard errors. How would you interpret the coefficient on this term?

(e) Test the hypothesis that $\beta_4 = \beta_5$ at the 5% significance level, where your alternative hypothesis is that $\beta_5 > \beta_4$. What do you conclude?