Problem set 3

July 19th, 2011
due July 21st, 2011

Question 1

A study relating college GPA to time spent in various activities has a data set which includes, for each student observed, the average number of hours spent each week in four activities: studying (st), sleeping (sl), working (w), leisure (l). Any activity is put into one of four categories, so that for each student the sum of hours in the four activities must be 168. It is proposed that the following model should be used:

$$GPA_i = \beta_0 + \beta_1 st_i + \beta_2 sl_i + \beta_3 w_i + \beta_4 l_i + \epsilon_i$$

- Can this model be used to obtain parameter estimates? Consider this in the context of the five assumptions.
- How could the model be reformulated so that the parameter estimates have a useful interpretation?

Question 2

Use STATA and the data set *bwght.dta* to help answer this question. The data set contains 1388 observations on 14 variables. The variables of interest are *bwght* (birth weight in ounces, bw_i), *cigs* (average maternal cigarettes consumption per day during pregnancy, c_i) and $faminc$ (family income in thousands of dollars, fi_i).

$$bw_i = \beta_0 + \beta_1 c_i + \beta_2 sl_i + \beta_3 fi_i + \epsilon_i$$

However, suppose we do not have the information on family income so the following model is estimated instead:

$$bw_i = \beta_0 + \beta_1 c_i + \beta_2 sl_i + u_i$$

- What is the expression for the bias in the estimator $\hat{\beta}_1$ obtained using this model? What sign do you expect it to take? Explain.
- Using Stata, run regressions for both the *true* and the *observed* models, and report the coefficients estimates. Does this support your intuition? (You might want to run an extra regression here.)

- Are the results from these regressions sufficient to prove that existence and direction of bias in the *observed* model?
- Would you expect to see further change in the coefficients with the inclusion of an additional variable providing a measure of the mother's health.

Provide the do-file and output log from your work using STATA.

Question 3

Use STATA and the data set WAGE2.dta to help answer this question. The data set contains 935 observations on 17 variables. The variables of interests are *wage* (average monthly earnings, w_i), *educ* (years of education, s_i), *exper* (years of work experience, e_i) and *tenure* (years with current employer, t_i). Consider the following three models, assuming that the third specification is true model:

$$w_i = \beta_0 + \beta_1 s_i + \epsilon_i$$
$$w_i = \alpha_0 + \alpha_1 s_i + \alpha_2 e_i + u_i$$
$$w_i = \gamma_0 + \gamma_1 s_i + \gamma_2 e_i + \gamma_3 t_i + v_i$$

- Using STATA, estimate each model using OLS. Report the coefficients, standard errors and R^2 for each specification.
- Consider the return to schooling coefficients: how do the estimates change between the different specification? Provide an explanation for this.
- Now consider the return to experience coefficients: how do the estimates vary between the second and third specifications. Was this change you were expecting? Explain.

Provide the do-file and output log from work using STATA.

Question 4

Suppose you have the following model:

$$y = \beta_0 + \beta_1 x + u$$

where all the assumptions hold and the sample size is n. Your friend suggests to etsimate $\beta's$ with:

$$\tilde{\beta} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{y_i - y_1}{x_i - x_1}$$

Your friend also tells you that this estimator is unbiased. In what sense is it a good suggestion and in what sense is it a bad one? Show how you would check whether it's unbiased (you don't have to do the whole calculations).

Question 5

If the errors in the classical regression model are NOT normally distributed, although the OLS estimator is no longer BLUE, it is still unbiased. True, false or uncertain?