Problem Set #4 Econometrics IVI Lizal, Fall 2005

Sample Selection Bias

This problem concerns predicting productivity of new workers in a large American manufacturing firm. There are five variables: ***y*** - an observed standardized physical productivity measure for the worker after the training period, ***sex*** - a dummy for the workers' sex (males are coded 1), ***dex*** - a score on a physical dexterity exam administered before the worker was hired, ***lex*** - the number of years of education of the worker, and ***quit*** - dummy whether the person quit within the first six months (quitters are coded 1).

1) Estimate the model

$$y_i = \alpha_0 + \alpha_1 sex_i + \alpha_2 dex_i + \alpha_3 lex_i + \alpha_4 lex_i^2 + u_i .$$

a) Test hypotheses:

i) H_0: $\alpha_3 = \alpha_4 = 0$

ii) H_0: $\alpha_4 = 0$

and interpret the results in *economic* terms.

b) Given the results of part a) draw a diagram illustrating the dependence of "expected productivity" on education. Set dexterity at its mean and sex=0. Interpret the picture. How would it change for the men? Suppose you thought the *shape* of education effect was different for men and women; reestimate your model respecified model. Does this improve things?

c) Use the delta-method to construct the a confidence interval for lex^*= level of education maximizing expected productivity.

d) Now consider the possibility that the variability of productivity depends on the *sexdexlex* variables. Propose a model of this type, estimate it, test the hypothesis that variability is independent of these variables, and reestimate part a) in accordance with your results. Be explicit about what you have done here, please.

2) Now consider a similar model for quits

$$logit(Prob(quit_i=1)) = \beta_0 + \beta_1 sex_i + \beta_2 dex_i + \beta_3 lex_i + \beta_4 lex_i^2 + \varepsilon_i$$

where *quit*=1 if the worker quit within the first 6 months after employment, and is 0 otherwise.

a) Estimate this model by *logit*, interpret the estimated parameters, in particular the estimated education effect, draw picture as in part 1b) above of the probability of quitting as a function of years of education.

b) Evaluate the *logit* specification by computing the Pregibon diagnostics suggested in class and interpret.

3)Finally we wish to reconsider the *sexdexlex* productivity model of Part 1) exploring the consequences of "sample selectivity". Suppose instead of observing the entire sample of 683 individuals, we instead observe productivity only for those who didn't quit. To facilitate the analysis of this subsample you may want to use the sort command.

a) Use the Heckman two-step procedure to estimate the productivity equation of Part 1), using only the non-quitters. Note, that you will need to save the inverse Mills ratio from the probit part (stored as @Mills in TSP after the probit part).

b) Compare and contrast the results from Part 1) with your previous results using the full sample, and the results from naively applying OLS to the restricted sample of non-quitters. In particular, discuss how the inferences drawn above earlier are altered by the sample selection bias of non-quitters.