ECON4004 Econometrics 2

Lab 3

Question 1 (based on Wooldridge, Exercise C17.8)

The dataset **JTRAIN2.dta** contains data on a job training experiment for a group of men. Men could enter the program starting in January 1976 through about mid-1977. The program ended in December 1977. The idea is to test whether participation in the job training program had an effect on unemployment probabilities and earnings in 1978.

(i) The variable *train* is the job training indicator. How many men in the sample participated in the job training program? What was the highest number of months a man actually participated in the program?

(ii) Run a linear regression of train on several demographic and pretraining variables: u*nem74* (denoting being unemployed in 1974), *unem75* (denoting being unemployed in 1975), *age*, *educ* (denoting years of education), *black*, *hisp* (denoting being Hispanic, which is an ethnic categorization and not a racial one; e.g. there are both black and white Hispanics), and *married*. Are these variables jointly significant at the 5% level?

(iii) Estimate a probit version of the linear model in part (ii). Compute the likelihood ratio test for joint significance of all variables. What do you conclude?

(iv) Based on your answers to parts (ii) and (iii), does it appear that participation in job training can be treated as exogenous for explaining 1978 unemployment status? Explain.

(v) Run a simple regression of unem78 on train and report the results in equation form. What is the estimated effect of participating in the job training program on the probability of being unemployed in 1978? Is it statistically significant?

(vi) Run a probit of unem78 on train. Does it make sense to compare the probit coefficient on train with the coefficient obtained from the linear model in part (v)?

(vii) Find the fitted probabilities from parts (v) and (vi). Explain why they are identical. Which approach would you use to measure the effect and statistical significance of the job training program?

(viii) Add all the variables from part (ii) as additional controls to the models from parts (v) and (vi). Are the fitted probabilities now identical? What is the correlation between them?

(ix) Using the model from part (viii), estimate the average partial effect of *train* on the 1978 unemployment probability. How does the estimate compare with the OLS estimate from part (viii)?

Note: The average partial affect (APE), also known as average marginal effect (AME) first computes the marginal effect of a given regressor for each observation, while taking the values of all other regressors as given. Then it averages the marginal effect over all observations. Hence, for a continuous variable cvar, the APE (AME) is equal to

$$APE_{cvar} = \frac{1}{n}\sum_{i=1}^{N} \varphi\,[(\hat{\beta}_0 + cvar_i\hat{\beta}_{cvar} + sum\ of\ other\ regressors\ multiplied\ by\ their\ coefficients)\hat{\beta}_{cvar}],$$

where φ denotes the normal density function (not the cumulative distribution function).

It is important to note that all regressors, including the continuous variable whose APE we calculate are evaluated at their observed values for each observation.

On the other hand, for a binary variable bvar, the APE(AME) is equal to

$$APE_{bvar} = \frac{1}{n}\sum[\Phi(\hat{\beta}_0 + cvar_i\hat{\beta}_{bvar} + sum\ of\ other\ regresors\ multiplied\ by\ their\ coefficients) - \Phi(\hat{\beta}_0 + sum\ of\ other\ regressors\ multiplied\ by\ their\ coefficients)]$$

In this case, all regressors other than the binary variable whose APE we calculate are evaluated at their observed values for each observation. The binary variable, on the other hand, first takes the value 1 for all observations, and then the value 0 for all observations, so that we can calculate the difference in the two cumulative distribution functions Φ.

(x) Estimate the average partial effects of the remaining regressors in (ix) on the 1978 unemployment probability. How does the estimate compare with the OLS estimate from part (viii)?

Question 2 ((based on Wooldridge, Exercise C17.14) The data set **happiness.dta** contains independently pooled cross sections for the even years from 1994 through 2006, obtained from the General Social Survey. The dependent variable for this problem is a measure of "happiness," *vhappy*, which is a binary variable equal to one if the person reports being "very happy" (as opposed to just "pretty happy" or "not too happy").

(i) Estimate a probit probability model relating *vhappy* to *occattend* (denoting attendance of religious services between several times a year and 2-3 times per month) and *regattend* (denoting attendance of religious services more often that 2-3 times per month) and include a full set of year dummies. Find the average partial effects for *occattend* and *regattend*. How do these compare with those from estimating a linear probability model?

(ii) Define a variable, *highinc*, equal to one if family income is above $25,000. Include *highinc*, *unem10* (denoting having been unemployed in the last 10 years), *educ* (denoting years of education), and *teens* (denoting the number of household members 13 thru 17 years old) to the probit estimation in part (ii). Is the APE of *regattend* affected much? What about its statistical significance?

(iii) Discuss the APEs and statistical significance of the four new variables in part (ii). Do the estimates make sense?

(iv) Controlling for the factors in part (ii), do there appear to be differences in happiness by gender or race? Justify your answer.