Econ 665

Assignment #2:

Due Date: November 1, 2017

General Instructions

Please answer the following sections:. You should submit your answers in a well-written text, using Stata output as well as the do file that you used to get the answers. In your report, you can use another document for your analysis but make sure it answers the questions in each section below. While your answers should be submitted individually and **written independently**, you are allowed to work in cooperation with your classmates. It is expected, however, that you give proper credit to your classmates who facilitated your understanding.

Question 1: Women Wages, race and unions (%35)

Please use the attached file (nlsw88) for this assignment on our mycourse page. It is a data set that we used in class from the National Longitudinal Survey of employed Women in 1988 . You should use the sub-sample for which hourly wage, race and an indicator of union status are available., i.e.

```
Use nslw88
Keep if !missing(wage+race+union)

Generate lwage=log(wage)
```

Question 1.a): test the hypothesis that race explains much of the variation in lwage.

Question 1.b). Recall the model stating that the lwage depends on race, union and job tenure. Consider this model to search for evidence of statistical discrimination (that is of differential effects across types of individuals). Test a model that includes interactions of the factors race and tenure. Please give an explanation of each of the coefficients from the regression

Question 1.c) Using the model used in the previous question, test the hypothesis that union workers earn premium wages.

Question 1.d) Using the commands seen in class, check the model you estimated in b) and c) . Check for some of the violations of the basic assumptions of OLS estimation.

Question 2: Distance to College and years of Completed Education (35%)

Please find attached the data set collegedistance.dta (on the mycourse page) These data are taken from the *HighSchool and Beyond* survey conducted by the US Department of Education in 1980, with a follow-up in 1986. The survey included students from approximately 1100 high schools.

The data used here were supplied by Professor Cecilia Rouse of Princeton University and were used in her paper "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment," *Journal of Business and Economic Statistics*, April 1995, Vol. 12, No. 2, pp 217-224.

Series in Data Set

Name	Description
Ed	Years of Education Completed (See below)
female	1 = Female/0 = Male
Black	1 = Black/0 = Not-Black
Hispanic	1 = Hispanic/0 = Not-Hispanic
Bytest	Base Year Composite Test Score. (These are achievement tests given to high school seniors in the sample)
dadcoll	1 = Father is a College Graduate/ 0 = Father is not a College Graduate
momcoll	1 = Mother is a College Graduate/ 0 = Mother is not a College Graduate
incomehi	1 = Family Income > $25,000 per year/ 0 = Income ≤ $25,000 per year.
ownhome	1= Family Owns Home / 0 = Family Does not Own Home
urban	1 = School in Urban Area / = School not in Urban Area
cue80	County Unemployment rate in 1980
stwmfg80	State Hourly Wage in Manufacturing in 1980
dist	Distance from 4yr College in 10's of miles, i.e. 1 represents 10 miles, 2 represents 20 miles, etc..
tuition	Avg. State 4yr College Tuition in $1000's

ed: Years of Education: Rouse computed years of education by assigning 12 years to all members of the senior class. Each additional year of secondary education counted as a one year. Students with vocational degrees were assigned 13 years, AA degrees were assigned 14 years, BA degrees were assigned 16 years, those with some graduate education were assigned 17 years, and those with a graduate degree were assigned 18 years.

Please answer the following questions

Question 2a) Run a regression of years of completed education (*ed*) on distance to the nearest college (*dist*). What is the intercept? What is the estimated slope? What does it say? Can you answer the question: how does the average value of years of completed schooling change when colleges are built close to where students go to high school?

Question 2b) Run a regression of *ed* on *dist*, but this time include additional regressors to control for characteristics of the student, the student's family and the local labor market. In particular include as additional regressors :*bytest, female, black, Hispanic, incomehi, ownhome, dadcoll, cue80* and *stwmfg80* . Now, what is the estimated effect of *dist* on *ed*?

Question 2c) Is the effect of *dist* on *ed* in the regression in b) substantively different from the regression in a). Based on this, does the regression in a) seem to suffer from important omitted variable bias?

Question 2d) The value of the coefficient on *dadcoll* is positive. What does this coefficient measure?

Question 2 e) Explain why *cue80* and *swfmg80* appear in the regression. Are the signs of the estimated coefficients what one would expect? Interpret the magnitudes of the coefficients. (hint: if not sure, use the margins command with the options dydx (varlist) and eyex(varlist) . See help margins)

Question 2 f) Bob is a black male. Hish high school was 20 miles from the nearest college. His base-year composite score (*bytest*) was 58. His family income in 1980 was $26000 and his family owned a home. His mother attended college, but his father did not. The unemployment rate in his county was 7.5% and the state average manufacturing hourly wage was $9.75. Predict Bob's years of completed schooling using the regression in 2b). Is he predicted to have a college degree?

Question 3. Village Program Placement and Female Participation in Microcredit program. (30%)

Recall the data set in problem 1. It is a data set from the Bangladesh Household Survey that was administered in 1998. The information was collected at the household and community levels. It contains information on 1,129 households. If you recall, some of these households have members who are participants in a micro-credit project while some households do not have participants. The variable "dfmd" is the indicator variable if the household have a female member who is a micro-credit participant.

We want to use this data set to start thinking of what one can do when one wants to measure the impact of a program on its participants (potential and actual participants). The program in this case is microcredit and we want to assess whether it leads to an increase in expenditures.

We first assume that the placement of the program was done randomly across the villages: that is we will assume that some villages received the treatment (i.e. the microcredit program) randomly and that other villages were then used as control villages (also chosen randomly). We will use the information that a woman has participated in a microcredit program to create a “treatment” variable at the village level.

To create indicators of a program in the village, use the data set hh_98.

```
use hh_98
```

Now let’s create a unique village indicator variable for the data set:

```
gen vill=thanaid*10+villid
```

Now, use the egen command to create the indicator variable.

```
egen progvillm=max(dmmfd),by(vill)
egen progvillf=max(dfmfd),by(vill)
```

Now that this is done, please answer the following questions

Question 3.a) Compute the average treatment effect of village program placement. Test the hypothesis that the households in villages with a female program in microcredit placement have an effect on the log of total per capita expenditure (*lexptot*).

First, use the difference in means command:

```
ttest lexptot, by(progvillf)
```

Then, use the regression command

```
reg lexptot progvillf
```

Discuss the results.

Question 3.b) The preceding regression estimates the overall impact of the village programs on the per capita expenditure of households. Yet, things may be different from the impact on the expenditure after holding factors constant, that is specifying the model adjusted for covariates that affect the outcomes of interest.

Now, regress the same outcome against the village program indicator variable, plus other factors that may influence the expenditure. Note that because the data come from a weighted survey, one will use the survey weights with the option [pw=”var”].

Please run the following regression:

```
reg lexptot progvillf sexhead agehead educhead lnland vaccess pcirr rice wheat milk
oil egg (pw=weight]
```

What happens to the result? Please comment.

Question 3.c) Even though microcredit program placement assignment is random across villages, the participation decision may not be. For instance, only those households with fewer than 50 decimals of lands can participate in the microcredit program.

This might mean we need to look at a different variable: perhaps we should look at the program participation for females rather than the village variable we created above. So, look at the hypothesis using the dfmfd variable:

```
ttest lexptot, by(dfmfd)
```

```
reg lexptot dfmd
```

Is there a different between the participants and the non-participants now?

Question 3 d) Similar to 3.b), please control for other household and village-level covariates in the female participation equation:

```
reg lexptot dfmfd sexhead agehead educhead lnland vaccess pcirr rice wheat milk
oil egg [pw=weight]
```

How are the results now? What can one say about the female participation in microcredit?

Question 3 e) What can one say about the impact of program placement? What about the impact of program participation? Are the impacts different? Could one capture both effects? Run a regression and discuss.