Economics 202 Professor Zabel
Econometrics Spring 2016

Homework 3: Due Thursday, February 11 @ 1:30 pm

Full Credit will not be Given Unless You Show Your Work

Make sure to include any regression results in a table using outreg.

Part I – Book Problem: Chapter 3: 4, C4 in Wooldridge. For problem C4, the dataset ATTEND.RAW can be downloaded from Trunk. There are 680 observations on the following variables

1. attend	classes attended out of 32
2. termgpa	GPA for term
3. priGPA	cumulative GPA prior to term
4. ACT	ACT score
5. final	final exam score
6. atndrte	percent classes attended
7. hwrte	percent homework turned in
8. frosh	=1 if freshman, 0 otherwise
9. soph	=1 if sophomore, 0 otherwise
10. skipped	number of classes skipped
11. stndfnl	(final - mean)/sd

To input the data into Stata use the command (given that you downloaded attend.raw to your c: drive):

```
infile “variable names” using c:\attend.raw
```

Part II – Data Problem

You will be accessing data from the March Supplement to the 2015 Current Population Survey. This is a random sample of the civilian non-institutional population of the United States living in housing units and members of the Armed Forces living in civilian housing units on a military base or in a household not on a military base. The Annual Social and Economic (ASEC) Supplement (in march) of the CPS provides the usual monthly labor force data, but in addition, provides supplemental data on work experience, income, noncash benefits, and migration.

To download the dataset

1. go to: http://www.census.gov/programs-surveys/cps.html
2. Click on "Data"
3. Click on "View all Data"
4. Under Microdata click on DataWeb FTP
5. Scroll down to CPS March Supplement and then download
    a. 2015 Tech Documentation and
    b. 2015 Data File

Note that the data file is zipped and you will need to unzip it. I use 7-zip which you can download for free.

Variable	Definition
hid	Unique id
htype	Household type (1-9) – see user guide for definitions
relationship	Relationship to reference person
age	Age in years
marital_status	Categories of marital status – see user guide for definitions
sex	Gender indicator – male or female
educ	Categories for educational attainment – see user guide for definitions
race	Race indicator – see user guide for definitions
wsal	Annual wage/salary earnings in $1000s of dollars
metro_status	1 if in metropolitan area, 2 otherwise
nper	Number of persons in the family
nkids	Number of children (age<=17) in the family
region	Census region

Download the do file cps_2015 (from Trunk) and run. Then make the following changes to the data

1. Limit the dataset to observations where
   a. 25 ≤ age ≤ 64
   b. $10,000 ≤ wsal ≤ $1,000,000
   c. Metro status is identified
2. Put wsal is $1,000's

Problems to answer with the CPS data:

1. Describe your data; include the period of analysis, the geographic area from which the sample is taken, the number of observations, the observation level of the data.

2. A. What is the percentage of women in the dataset?
   B. What is the median number of years of education in the sample?
   C. What are the 25th, median, and 75th percentiles of the wage/salary earnings distribution (use sum command with "detail" option)?

To determine the median years of education, you will need to look in the user guide, first click on – Person Record" under "Data Dictionary" on the left-hand-side (starts on page 8-15), then look for the variable called "A_HGA".

3. Generate a graph of the distribution of wage/salary earnings (wsal) and give the graph a title. Use the stata command "hist" with the option "title("name") where "name" is your title. Include this graph in your homework. Do the same for the log of wsal.

Generating Binary Variables (that is variables with two values; 0 and 1): Suppose I have a variable, X, that is an indicator of ten categories (so it takes on the values 1-10). If I want to create a binary variable, Y, that is 1 for values of:

X = 5	gen Y = X==5	
X = 1 to 5:	gen Y = X<=5	
X between 4 and 7:	gen Y = X>=4 & X<=7	& is "and"
X = 1 or 3	gen Y = X==1 | X==3	| is "or"

Generate binary variables for female, residing in a metro area (metro), married, separated, divorced, white only (white), having a high school degree but not a Bachelor's degree (hs), having at least a Bachelor's degree (col) and the four census regions (northeast, south, midwest, and west).

You will need to look in the codebook to determine which values of the base variables to use to generate these binary variables.

4. Include a table of summary statistics for wsal, age, hs, col, white, metro, married, separated, divorced, female, nkids, nper, northeast, south, midwest, and west using the "tabstat" command in Stata. Call this Table 1. Make sure to give that table a title.

5. Run a regression of ln(wsal) on age, hs, col, white, metro, married, separated, divorced, female, nkids, nper, south, midwest, and west. (call this Model 1).

A. Include the regression results in a table. Make sure to display at least 4 decimal places. Call this Table 2
B. Interpret the estimated coefficients for col, nper, and south. Why is the sign of the coefficient estimate for nper negative?
C. Determine the percent difference in expected wages/salary earnings for a family with one child and one with 4 children, all else equal. Is this difference economically significant?

6. Instead of using nkids as a regressor, consider the following specification that generates binary variables for nkids = 1, nkids = 2, nkids = 3, and nkids >= 4 (the left out group is nkids = 0).

A. Why might this specification make more sense than using nkids?
B. Estimate Model 1, replacing nkids with the above 4 binary variables. Call this Model 2. Include the regression results in Table 2. Interpret the coefficient estimate for nkidsge4. Based on the results, explain why Model 2 is preferred to Model 1.
C. Based on the results for Model 2, propose a different specification, Model 3, for the impact of the number of children on wsal that is potentially "better" than Model 2. Include the results for this model in Table 2. Explain how you would determine if Model 2 or Model 3 is preferred (that is, describe a procedure you would us to choose between the two models but you DON'T necessarily have to carry out this procedure).

Part III – Article: Read through Section II (up to page 161) of “Are Restaurants Really Supersizing America?” by Michael L. Anderson and David A. Matsa, *American Economic Journal: Applied Economics*, 3(1): 152—88.

1. Why is current evidence that points to fast food restaurants as a culprit in the increasing obesity rates in the U.S. not causal evidence of this relationship?

2. How do the authors plan to show the causal relationship?

3. Why might regulating what fast food restaurants can serve or restricting where they can locate not necessarily affect obesity rates?

4. Based on the implications of the model in Section I, explain why the question “Do fast food restaurants increase obesity?” can only be answered empirically.

5. Explain the data that comes from the Behavioral Risk Factor Surveillance System (BRFSS). What are the sources of the data? What is the unit of observation and what is the sample size? What years do the data cover? What geographic area are the data drawn from?

Part IV – Paper Topic

ON A SEPARATE PAGE and handed in separately:

List up to three topics, questions you would like to answer, hypothesis you would like to test for your paper