Economics 250a
Problem Set #2

This problem set uses data from the 2012 Annual Social and Economic Survey (ASES) supplement to the CPS (formerly known as the "March CPS"). In the course web page directory for problem set 2 you will find a ***stata*** data set, along with a ***sas*** program read2x.sas (and associated log and listing files) and a second program "extract2x.sas" that takes the data extract created by read2x and creates the smaller working data set "cps2012.dta" for this problem set.

Background

cps2012.dta contains data for n=130,964 men and women in the 2012 AESE. The sample includes people age 18-70, and has annual data for 2011.

You will also find the codebook for the "raw" CPS, called march2012-codebook.pdf, and a more complete document called march2012-documentation.pdf. The data and these codebooks are taken directly from the Census website (http://www.bls.census.gov/cps_ftp.html#cpsmarch). A snapshot of the content of this web site is also included in the materials in the course site.

The raw CPS data file is a hierarchical data set with 3 levels of records: household (1 per physical address), family (1 per family; often there is 1 family per house but not always) and person (1 per person). The program read2x.sas reads the raw data set and creates a file with 1 record per adult. It reads a household record and pulls off some "household-level" data. Then it loops over families in the household, and within each family loops over people. A record is "spit out" for each person over age 16. Note the column locations for each variable that is input from each of the three types of records. Using these you can go back to the codebook and find the way the "raw" variable is coded.

For example, the variable "classly" is read from column 291 of the person record. Here is an extract from the codebook of the entry for the variable "LJCW" in column 291 of the person record. Officially this is called "Item 46e". ("46e" is a legacy from many decades of CPS questionnaires – this used to be question 46e in the paper survey).

D LJCW 1 291 (0:7)
Item 46e - Class of worker
U WORKYN = 1
V 0 .Not in universe
V 1 .Private
V 2 .Federal
V 3 .State
V 4 .Local
V 5 .Self employed incorporated, yes
V 6 .Self employed incorporated, no
V .or farm
V 7 .Without pay

note the first line is “U WORKYN = 1”. This is shorthand for the “universe” for this question is people with variable WORKYN=1 (which is in col. 251, and asks if the respondent did any work at a job or business last year). The “V” rows are the possible values (valid codes). The ASES questionnaire has a lot of questions about hours, and various sources of income last year. It also asks about the characteristics of the main (or longest) job held last year, including industry, occupation, and “class” – which is what this variable is measuring.

The ASES also has data for the ¼ of the sample who are in the so-called ‘outgoing rotation group’ in the CPS survey – people in this group are asked about their main **current** job (in March 2012), including hourly pay if they are paid by the hour, or weekly pay and weekly hours, if not.

In this problem set you will get a chance to experiment with a few very simple labor supply models for men and women. Our key variables will be:

annhrs = total annual hours = weeksly * hrswkly (these last 2 are read directly from the CPS person record

wagesal = total earnings from wage and salary jobs, read from col. 364-370. (We will use this instead of earnings including self-employment and farm income).

age, gender (represented by the dummy variable “female”) and education (“yrsed”, constructed from the variable “higrad” in cols. 25-26),

famearn = total wage-salary earnings of everyone in the family (including this person), read from cols 48-54 of the family record

wkpay = weekly average pay (in March 2012) for people in OGR
wage_ogr = hourly pay or weekly pay / hours per week for people in OGR

NOTE: the sample contains any adult male or female. To get male “husbands” of dual-headed families, restrict attention to (famkind=1 and relhead = (1,3)). To get female “wives” of dual-headed familes, restrict attention to (famkind=1 and relhead =(1,4)).

famkind: 1=husband/wife families 2=male head only 3=female head only
relhead: 1=reference person with family in house, 2=lone reference person
3=husband of reference person 4=wife of reference person

<u>Questions</u>

1. For men and women separately, construct and graph the age profiles of:
    a) the dummy “workly” that =1 if annhrs>0
    b) hours per week last year for those who work (use “hrswkly”, setting the variable to missing for nonworkers)
    c) average annual hours for those who work
    d) average annual earnings for those who work

You should obtain profiles that look like the ones shown in Lecture 2, though those data are for 2008. Try the same exercise, looking ONLY at husbands and wives.

2. Suppose you were asked the question: how much of the difference in average hours per year of men versus women is due to the fact that more women don't work at all, versus differences in the hours of work conditional on working.
Hint: If $y \geq 0$ is a random variable, and $D=1[y>0]$, then $E[y]=P(D=1)*E[y|y>0]$.
Now consider how to decompose the difference in unconditional means for y_1 and y_2. You can simplify your answer by focusing on men and women in some age range.

3. In "read2x/extract2x" I constructed the variable "wage" = wagesal/annhrs. Obviously, it has some correlation with annual hours that will confound a labor supply study. I also constructed a "trimmed" wage: twage = 4 if wage<4; twage=400 if wage>400. This is called "Winsorizing". You will want to use twage (or some variant) since otherwise there are some pretty crazy outliers. Look at the distribution of "wage" and compare that to the distribution of twage (the trimmed variant). If you want, try making your own trimmed wage.

a) using "logwage" = log(twage), run a simple "Mincer" wage model, separately for men and women. For this you will use the "experience" variable exp = age-yrsed-5. (I reset exp to 0 if negative). The basic Mincer model is

log (wage) = a+ b*years of education + cubic in experience

(i) compare the "return to education" (the coefficient b) for men and women

(ii) plot the experience profiles for men and women. You should find that for women, wages rise much less with experience.

b) now run an augmented Mincer model that all the above terms plus a dummy for people with 16 or more years of education (coded as the variable "collplus") You should find that even controlling for years of education there is a jump at college completion. Compare the "jumps" for men and women. OPTIONAL: You can also look at the effect of the variables "maplus" which is a dummy for people with a masters or more, and "phd" which is a dummy for having phd or professional degree.

c) Let's focus on the wage effects of the race/ethnicity variables black, asian and Hispanic. (look at the programs to see how these are defined)

(i) consider a basic Mincer model with these 3 dummies. Compare the earnings disadvantages for black and Hispanic workers relative to the omitted group (non-Hispanic white) for men and women.

(ii) now add dummies for immigrant ("imm") and second-generation ("gen2") status. How do these change the Hispanic and Asian dummies?

(iii) Now add dummies for city size (cbsa1-cbsa6, created from the variable cbsa_size, which was read from col. 55 of the household record). What happens? Why?

(iv) Finally, add dummies for "class" of worker: fedwkr, statewkr, localwkr created from "classly"

Do these effect the race/ethnicity variables?

4. In this part we will fit some very simple labor supply models. **For simplicity we will focus on people with annual hrs > 0 and looking ONLY at husbands and wives.** The models will use the functional form:

(*) Annual Hours = a + b*Log(wage) + c*Non-labor income + other controls

We will use "other family earnings" = famearn – wagesal as our measure of non-labor income. Take a look at the distributions of other family earnings for husbands vs. wives.

a) for men and women separately, look at the correlations between annual hours, log wages, other family earnings, age and education.

b) for men and women separately, fit a simple model of the form (*) with no other controls. What are the implied estimates for *mpe* and *σ,* the compensated elasticity of labor supply?

c) Now add years of education and a cubic in experience to the labor supply model as additional controls. What happens to the estimates for *mpe* and *σ*? Do you think these should be added or not?

d) Now we'll try to construct an IV for wages, using wage_ogr as the instrument for wage. To do this we have to focus on people with a valid response for wage_ogr (those in the OGR supplement who have a job in March 2012).

(i) re-fit your labor supply models using ONLY data for people with wage_ogr>0.
(ii) now fit an IV model using log(wage_ogr) as the instrument. What happens to the estimates for *mpe* and *σ.* Can you defend this instrument?