Economics 202 Professor Zabel
Econometrics Spring 2016

Homework 4: Due Thursday, February 25 @ 1:30 pm
Full Credit will not be Given Unless You Show Your Work

Part I – Book Problem: Chapter 7: C4 in Wooldridge. Include the results from any regressions for this problem in one table. Call it Table I-1.

NEW PART C4.iv In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that the impact on gpa of being an athlete is the same for men and women. What is the expected difference in gpa between women athletes and women non-athletes?

For question C.4, you will need to download the dataset gpa2.txt. There are 4,137 observations in this dataset. The variables in this dataset are in order

1. sat	combined SAT score
2. tothrs	total hours through fall semester
3. colgpa	GPA after fall semester
4. athlete	=1 if athlete, 0 otherwise
5. verbmath	verbal/math SAT score
6. hsize	size graduating class, 100s
7. hsrank	rank in graduating class
8. hsperc	100*(hsrank/hssize)
9. female	=1 if female, 0 otherwise
10. white	=1 if white, 0 otherwise
11. black	=1 if black, 0 otherwise

Read gpa2.txt into Stata using the infile statement.

Part IIA - This is a continuation of the problem from the last homework using the Stata data set **College Distance** that contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986

1. Run a regression of *yrsed* on *dist*. Call this Model 1.
    a. Display the results using the outreg command. Include 3 decimal places and standard errors below the estimates (see pages 2-16 and 2-17 in the class notes) Call this Table 1 – make sure to give the table a title.

   b. What is the percent difference in *yrsed* for someone whose high school is 50 miles from the nearest four-year college compared to someone who's high school is 10 miles from the nearest four-year college? Is this economically significant? (Think about what is the appropriate measure to use).
2. Run the regression that adds the following variables to Model l: *female*, *black* and *Hispanic*. Call this Model 2.
   a. Add the regression results for Model 2 to Table 1 using outreg
   b. Interpret the coefficient estimate for *black*. Is it statistically significant? Is this economically significant?
3. Run the regression that adds the following variables to Model 2: *bytest*, *incomehi*, *ownhome*, *dadcoll*, *momcoll*, *cue80*, and *stwmfg80*. Call this Model 3.
   a. Include a table of summary statistics for all the variables in Model 3. Make sure to include a title.
   b. Add the regression results for Model 3 to Table 1 using outreg
   c. Make the same calculation as in 1b. Is this economically significant?
   d. Is the coefficient for black statistically significant? Is this economically significant? Explain the difference in results from part 2b.
4. Interpret the coefficient estimates for *dadcoll* and *momcoll*. Is the difference in *yrsed* for a person whose parents both went to college as compared to a person where neither parents went to college economically significant? What is a causal mechanism that explains this effect?
5. Explain why *cue80* and *stwgfg80* are included in the regression. Interpret the estimated coefficient estimates for these variables. Are the signs of their estimated coefficients what you would have expected? Explain.
6. Bob is a black, non-Hispanic male. His high school was 20 miles from the nearest college. His base-year composite score (*bytest*) was 58. His family income in 1980 was $26,000, and his family owned a home. His mother attended college but his father did not. The unemployment rate in this county was 7.5%, and the state average manufacturing hourly wage was $9.75. Predict Bob's years of completed schooling using the regression results for Model 3.
7. Add the square of *dist* to the Model 3 and run the regression. Call this Model 4.
   a. Add the regression results for Model 4 to Table 1 using outreg
   b. Make the same calculation as in 1b. Is this economically significant?

Part II – Using IPUMS - The Integrated Public Use Microdata Series (IPUMS-USA) consists of more than fifty high-precision samples of the American population drawn from fifteen federal censuses and from the American Community Surveys of 2000-present. Each record is a person, with all characteristics numerically coded. In most samples persons are organized into households, making it possible to study the characteristics of people in the context of their families or other co-residents. A data extraction system enables users to select only the samples and variables they require. See https://usa.ipums.org/usa/

We will be using the 1% sample from the 2010 American Community Survey (ACS). The ACS provides the same information as the Decennial Census on an annual basis. It is different than the Census in that it is a survey – i.e. a random sample from the U.S. population.

One needs to set up an account (free) to use the IPUMS. Click on the link "Select Data" and then one can select the variables and samples to use. One you have created a data set, you will be notified via e-mail that your sample is ready. Along with the data, there is a code book and a Stata do file that can be used to create a Stata data set.

I have generated a data set for you to use. The original sample includes 3,061,692 observations. The version you will use excludes group quarters (GQ<=2), and is limited to household heads and spouses (RELATE<=2) who are between 25 and 64 years of age 25<=AGE<=64. This leaves a sample of 1,340,362 individuals. The Stata dataset ipums_acs_2010 and the codebook IPUMS_acs_2010_codebook can be downloaded from Trunk.

Generate binary variables for female, married (spouse present or absent), white, black, hispanic, having a high school degree or some college but not a Bachelor's degree (hs_somecol), having at least a Bachelor's degree but not a Masters' degree or higher (ba), having a Masters' degree or higher (maplus), having any health insurance (health_ins), having private health insurance (health_ins_private), employed, Cognitive difficulty (cog_diff), Ambulatory difficulty (amb_diff), Vision or hearing difficulty (sens_diff), veteran. Also put total income (inctot) in $1,000s and generate age^2 (divide this variable by 100) and female·inctot (fem_inc).

1. Create a table of variable definitions. Make sure to number the table (Table 1) and a title.

2. Include the summary statistics for all the generated variables in a table. Make sure to number the table (Table 2) and a title.

3. Run a regression of married on the above set of variables. Include the regression results in a table (using outreg). Include 4 decimal places. Make sure to number the table (Table 3) and a title.

4. Calculate the semi-elasticities for age (evaluated at the mean), health_ins_private, black, cog_diff, and employed.

5. Test that the impacts of total income for men and women are zero against a two-sided alternative.

6. Calculate the semi-elasticities for the difference in the impact on marriage for an individual at the 95^{th} percentile of the income distribution versus a person at the 5^{th} percentile of the income distribution, all else equal. Evaluate at the conditional means for men and women.

Part III – Article: Read sections 1 and 2 of the article "Binge drinking and labor market success: a longitudinal study on young people," by Shao-Hsun Keng and Wallace E. Huffman. Answer the following four questions

A. Imposing taxes on good can be efficiency enhancing if their consumption imposes negative externalities on society. Give five ways in which binge drinking can impose external costs on society.

B. Previous studies find evidence that moderate drinkers have higher wages than non-drinkers. Why might this be the case?

C. What does it mean for binge drinking to be a "rational addiction?" How do the authors allow for the possibility of rational addition in their model?

D. Why might cross-sectional studies produce biased estimates of the impact of binge drinking on earnings? What is the authors' solution to this problem?

Part IV – Paper Topic: On a page handed in separately, update your idea(s) about your paper topic. Include the regression model that you will estimate making clear at what level the data are measured. Also include possible data sources.