

# EXTRA CREDIT

## Problem Set 2

ECON 306 - Introduction to Econometrics

Spring 2020

Delina E Agnosteva

**Due Date: May 3 at 11:59PM on Canvas.**

**INSTRUCTIONS:** Solve the following questions to the best of your ability. Come see me and ask me if you do not know how to solve any of these questions, even before the due date. I will work with you if you are having trouble solving these.

*This is an optional, extra-credit assignment. To receive full credit for this assignment, the problem set needs to be submitted to **Canvas in a single PDF document** containing your 1) Stata log file, 2) any figures (scatterplots, histograms, etc.), and 3) any written explanations and answers. All of these components need to be attached together in that order. Late submissions will NOT be accepted. DO NOT email! No assignments will be accepted via email.*

First of all, for this problem set, you will have to submit the Stata log file. Stata can record your session into a file called a log file but does not start a log automatically; you must tell Stata to record your session. By default, the resulting log file contains what you type and what Stata produces in response, recorded in a format called Stata Markup and Control Language (SMCL). The file can be printed or converted to plain text for incorporation into documents you create with your word processor. You can find more information here: <https://www.stata.com/manuals13/u15.pdf>.

So, in the beginning of your Stata .do file write the following command: *log using session* (or a different file name). Then, at the very end of your .do file, include *translate session.smcl session.pdf*. This would translate SMCL files to plain text, which is a format more useful for inclusion into a word processing document. Or, even better, you can translate your Stata SMCL log files directly into PDF files and then use Adobe Acrobat to merge PDF files together. You will *need to* turn in this log file to receive full credit for this assignment.

I would strongly suggest compiling the log file in Stata after you have completed all of your code and can run it smoothly without any errors. In that way, your log file would not contain any lines of code that does not produce any results or any duplicate results. Please do your best to include comments in your code (using the \* sign in your Stata .do file) and to make the solutions to the different problems as clearly marked as possible. Otherwise, the graders might have to penalize you, if they cannot follow your work. And then I will have to re-grade your work and the whole process becomes highly inefficient.

## Part I - Smocking.xlsx Data.

Believe it or not, workers used to be able to smoke inside office buildings. Smoking bans were introduced in several areas during the 1990s. Supporters of these bans argues that in addition to eliminating the externality of secondhand smoke, they would encourage smokers to quit by reducing their opportunities to smoke. In this assignment, you will estimate the effect of workplace smoking bans on smoking, using data on a sample of 10,000 U.S. indoor workers from 1991 and 1993.

Smoking is a cross-sectional data set with observations on 10,000 indoor workers, which is a subset of a 18,090-observation data set collected as part of the National Health Interview Survey in 1991 and then again (with different respondents) in 1993. The data set contains information on whether individuals were, or were not, subject to a workplace smoking ban, whether or not the individuals smoked and other individual characteristics. These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Matthew Farrelly and Edward Montgomery “Do Workplace Smoking Bans Reduce Smoking?” American Economic Review, September 1999, Vol. 89, No. 4, 728-747.

Use the data to complete the following:

**Problem 1.** Estimate the probability of smoking for (i) all workers, (ii) workers affected by workplace smoking bans, and (iii) workers not affected by workplace smoking bans.

**Problem 2.** What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

**Problem 3.** Estimate a linear probability model with *smoker* as the dependent variable and the following regressors: *smkban*, *female*, *age*, *age*<sup>2</sup>, *hsdrop*, *hsgrad*, *colsome*, *colgrad*, *black*, *hispanic*. Compare the estimated effect of a smoking ban from this regression with your answer from Problem 2. Suggest an explanation, based on the substance of this regression, for the change in the estimated effect of a smoking ban between Problem 2 and Problem 3.

**Problem 4.** Test the hypothesis that the coefficient on *smkban* is 0 in the population version of the regression in Problem 3 against the alternative that it is nonzero, at the 5% significance level.

**Problem 5.** Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in Problem 3. Does the probability of smoking increase or decrease with the level of education?

**Problem 6.** Repeat Problem 3 - Problem 5 using a probit model.

**Problem 7.** Repeat Problem 3 - Problem 5 using a logit model.

**Problem 8.**

a) Mr. A is a white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that Mr. A is not subject to a workplace smoking ban, calculate the probability that Mr. A smokes. Carry out the calculation again, assuming that he is subject to a workplace smoking ban. What is the effect of the smoking ban on the probability of smoking?

b) Repeat a) for Ms. B, a female, black, 40-year-old college graduate.

c) Repeat a) and b) using the linear probability model.

## Part II - Income\_Democracy.xlsx Data.

Do citizens demand more democracy and political freedom as their incomes grow? That is, is democracy a normal good? Use the dataset *Income\_Democracy*, which contains a panel data set for 195 countries for the years 1960, 1965, ... 2000. The data were supplied by Professor Daron Acemoglu and are a subset of the data used in his paper with Simon Johnson, James Robinson, and Pierre Yared, "Income and Democracy" *American Economic Review*, 2008, 98:3: 808-842. The dataset contains an index of political freedom/democracy for each country in each year, together with data on each country's income and various demographic controls. The income and demographic controls are lagged five years relative to the democracy index to allow time for democracy to adjust to changes in these variables.

Use the data to complete the following:

**Problem 1.** Is the data set a balanced panel? Explain.

**Problem 2.** The index of political freedom/democracy is label *Dem\_ind*.

a) What are the minimum and maximum values of *Dem\_ind* in the data set? What are the mean and standard deviation of *Dem\_ind* in the data set? What are the 10th, 25th, 50th, 75th, and 90th percentiles of its distribution?

- b) What is the value of *Dem\_ind* for the United States in 2000? Averaged over all years in the data set?
- c) What is the value of for Libya in 2000? Averaged over all years in the data set?
- d) List five countries with an average value of *Dem\_ind* greater than 0.95; less than 0.10; and between 0.3 and 0.7.

**Problem 3.** The logarithm of per capita income is labeled *Log\_GDP**PC*. Regress *Dem\_ind* on *Log\_GDP**PC*. Use standard errors that are clustered by country.

- a) How large is the estimated coefficient on *Log\_GDP**PC*? Is the coefficient statistically significant?
- b) If per capita income in a country increases by 20%, by how much is *Dem\_ind* predicted to increase? What is a 95% confidence interval for the prediction? Is the predicted increase in *Dem\_ind* large or small? Explain.
- c) Why is it important to use clustered standard errors for the regression? Do the results change if you do not use clustered standard errors?

**Problem 4.**

- a) Suggest a variable that varies across countries but plausibly varies little – or not at all – over time and that could cause omitted variable bias in the regression in Problem 3.
- b) Estimate the regression in Problem 3, allowing for country fixed effects. How do your answers to Problem 3 a) and Problem 3 b) change?
- c) Exclude the data for Azerbaijan, and rerun the regression. Do the results change? Why or why not?
- d) Suggest a variable that varies over time but plausibly varies little – or not at all – across countries and that could cause omitted variable bias in the regression in Problem 3.
- e) Estimate the regression in Problem 3, allowing for time and country fixed effects. How do your answers to Problem 3 a) and Problem 3 b) change?
- f) There are additional demographic controls in the data set. Should these variables be included in the regression? If so, how do the results change when they are included?

**Problem 5.** Based on your analysis, what conclusions do you draw about the effects of income on democracy?

## Part III - Income\_Democracy.xlsx Data.

This problem is intended to help you hone your coding skills even further. There are times we need to do some repetitive tasks in the process of data preparation, analysis or presentation. For example, we might have to compute a set of variables in the same manner or to rename, relabel, create a series of variables, or repetitively re-code the values of a number of variables. Given that the whole point of programming is to automate a particular process, we wish to minimize the number of times we have to “manually” do the same steps. To that end, researchers often use *loops*. And once you complete this exercise, you would have too.

For this exercise, we are still going to use the Income\_Democracy.xlsx data from Part II. The Income\_Democracy.xlsx, as you know, are panel data, which means that the variables of

interest vary both over time and over entities. Sometimes, after performing our estimations in a panel data setting (using variation over time and over entities), however, we might want to "zoom in" the specific time periods to investigate for anything more pronounced happening during one or more time periods and thus driving the results. To that end, we might re-estimate our baseline model *per year*.

Your task is to create a loop that estimates the model from Problem 3 of Part II for each individual year in the data. Stata has several different ways of doing loops. You need to figure out which one is the most appropriate to employ and what its specific syntax is.

Then, using Stata command(s) create a Word (.docx) table with the results from these regressions. Your table should have 9 columns (one for each regression model) and include the number of observations, the  $R^2$ , the adjusted  $R^2$ . Include this table in your file with the written explanations and answers to questions and, once you are done, convert the Word file into a PDF file.

## Part IV - CPS2015.xlsx Data.

Each month the Bureau of Labor Statistics in the U.S. Department of Labor conducts the "Current Population Survey" (CPS), which provides data on labor force characteristics of the population, including the level of employment, unemployment, and earnings. Approximately 54,000 randomly selected U.S. households are surveyed each month. The sample is chosen by randomly selecting addresses from a database comprised of addresses from the most recent decennial census augmented with data on new housing units constructed after the last census.

In this exercise, you will investigate the relationship between a worker's age and earnings. Generally, the older workers have more job experience, leading to higher productivity and higher earnings. Use the data to complete the following:

**Problem 1.** Run a regression of average hourly earnings on age, sex, and education. If age increases from 25 to 26, how are earnings expected to change? If age increases from 33 to 34, how are earnings expected to change?

**Problem 2.** Run a regression of the logarithm of average hourly earnings,  $\ln(AHE)$ , on *Age*, *Female*, and *Bachelor*. If *Age* increases from 25 to 36, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**Problem 3.** Run a regression of the logarithm of average hourly earnings,  $\ln(AHE)$ , on  $\ln(Age)$ , *Female*, and *Bachelor*. If *Age* increases from 25 to 36, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**Problem 4.** Run a regression of the logarithm of average hourly earnings,  $\ln(AHE)$ , on *Age*,  $Age^2$ , *Female*, and *Bachelor*. If *Age* increases from 25 to 36, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**Problem 5.** Do you prefer the regression in Problem 3 to Problem 2? Explain.

**Problem 6.** Do you prefer the regression in Problem 4 to Problem 2? Explain.

**Problem 7.** Do you prefer the regression in Problem 4 to Problem 3? Explain.

**Problem 8.** Plot the regression relation between  $Age$  and  $\ln(AHE)$  from Problem 2, Problem 3, and Problem 4 for males with a high school diploma. Describe the similarities and difference between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

**Problem 9.** Run a regression of  $\ln(AHE)$  on  $Age$ ,  $Age^2$ ,  $Female$ , and  $Bachelor$ , and  $Female \times Bachelor$ .

- a) What does the coefficient on the interaction term measure?
- b) Amelia is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of  $\ln(AHE)$ ?
- c) Dalia is a 30-year-old female with a high school diploma. What does the regression predict for her value of  $\ln(AHE)$ ?
- d) What is the predicted difference between Amelia's and Dalia's earnings?
- e) Adrian is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of  $\ln(AHE)$ ?
- f) Daniel is a 30-year-old male with a high school diploma. What does the regression predict for his value of  $\ln(AHE)$ ?
- g) What is the predicted difference between Adrian's and Daniel's earnings?

**Problem 10.** Is the effect of  $Age$  on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

**Problem 11.** Is the effect of  $Age$  on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

**Problem 12.** Using Stata command(s) create a Word (.docx) table with the results from Problem 1 through Problem 4 and from Problem 9 through Problem 11. Your table should have 7 columns (one for each regression model) and include the number of observations, the  $R^2$ , and report heteroskedasticity-robust standard errors. Include this table in your file with the written explanations and answers to questions and, once you are done, convert the Word file into a PDF file.