651 Empirical Economics: Assignment 2

Per Hjertstrand* Andrew Proctor†

September 2017

Instructions

The assignment gives a maximum of 5 points. Answers should be provided in English. The key to a good score is to be clear and precise in your answers. For theory questions, be sure to show your work.

Answers to the written exercises should either be typed or, if handwritten, (legibly) scanned. Pictures of handwritten exercises taken from a camera (or smartphone) will not be accepted. Answers to the Stata-based exercises should be submitted in the form of a well-commented Stata **do-file and log file**. Your do-files should contain your code, output, and discussion of results (do not provide discussion in a seperate document. Please save your log files in the *text* format.

*Per.Hjertstrand@ifn.se
†Andrew.Proctor@phdstudent.hhs.se

Part A: Theory Exercises

Answer the solve the following problems in Wooldridge (questions are from the 6^{th} edition. Please let Andrew know if you need access to the questions).

- **Chapter 8:** Problems 1, 2 & 7.i-ii
- **Chapter 10:** Problems 1.i-ii & 5
- **Chapter 13:** Problem 6

And the following (modified from Stock and Watson's *Introduction to Econometrics*):

You have access to a large panel dataset (n=10,000) for workers over the years 2000-2016 and are interested in exploring the effect of education on earnings. The variables in your data set include earnings, age, gender, education, and union status.

a Give some examples of unobserved person-specific variables that might be correlated with both education and earnings.

b Give some examples of unobserved time-specific variables (affecting everyone) that might be correlated with both education and earnings.

c How would you control for the person-specific and time-specific variables above in a panel data regression?

d Can you estimate the effect of someone's gender on earnings in the regression you suggest above?

e Give some examples of variables that might be correlated with education and earnings that vary both over time and across individuals.

f Explain why the residuals might suffer from the two main types of violations of the standard error distribution assumptions (as discussed in the panel data seminar).

g Finally, imagine you want to also control for an aggressive element of earnings by including the lag of earnings (that is, earnings in the previous year). Will a fixed effects regression give reliable estimates for the following regressors: age, education, union status, and the lag of earnings. Why or why not? (Hint: check the fixed effects regression models assumptions.)

Part B: Stata Exercises

For this exercise, you will be exploring determinants of economic growth. Some features of an economy that are commonly hypothesized to affect growth are: openness to trade, the state of human capital and infrastructure, the country's natural resource endowment and its legal, political, and social institutions.

To try to explore these determinants, you will be using data compiled in the "World Development Indicators" dataset, from the World Bank. The data can be downloaded in 2 ways: To download the CSV files, go to: https://data.worldbank.org/data-catalog/world-development-indicators. Select the "WDI (CSV)-ZIP" file (you will need to extract it). If instead you would like to work directly within Stata, you can instead install and use the command **wbopendata**.

You will be working with the following indicators from the World Development Indicators dataset:[1]

- NY.GDP.PCAP.KD
- IC.LGL.DURS
- SG.GEN.PARL.ZS
- TM.TAX.MRCH.WM.AR.ZS
- NY.GDP.TOTL.RT.ZS
- SE.SEC.CMPT.LO.ZS
- SH.H2O.SAFE.ZS
- SH.ANM.ALLW.ZS
- SE.SEC.ENRR

Using this dataset:

1 **Data Preparation:**

(a) Import the dataset, keeping only if the indicator code matches those above. If using **wbopendata** to import the data, import the data in the 'wide' format.

(b) In addition to data for countries, the World Indicators dataset also includes information for a number of suprenational areas. To remove this, enter in the following code:

```
**** Code to drop supranational regions ****
local regionstodrop ARB CSS CEB EAR EAS EAP TEA EMU ///
ECS ECA TEC EUU FCS HPC HIC IBD IBT IDB IDX ///
```

[1] To interpret the indicators, you should consult the "Metadata Glossary".

```
IDA LTE LCN LAC TLA LDC LMY LIC LMC MEA MNA ///
TMN MIC NAC OED OSS PSS PST PRE SST SAS TSA ///
SSF SSA TSS UMC WLD INX

foreach region in `regionstodrop' {
drop if countrycode=="`region'"
}
**** End of code to drop supranational regions ****
```

(c) Replace the indicator codes values listed above with descriptive one-word names to be used as variables once the data is suitably reshaped.

(d) **Reshape** the dataset so it is in the standard format for Stata.

- You may want to drop the country name and indicator name first
- You may also need to correct the year variable you create when reshaping it into a column (long).

(e) **Drop** observations where the year is less than 1970.

(f) Encode the string "country code" variable so that it can be used as the group identifier in the analysis. Then, use **xtset** to provide Stata with the panel structure of your analysis.

(g) For GDP per capita, create a new variable equal to the log of GDP per capita.

(h) One of the measures of human capital quality we want to use in this analysis is the average lower secondary school completion rate for individuals in the labor force. Our data, however, reports only this completion rate for each year. To proxy the completion rate in the overall labor force, instead create a variable that is the average completion rate over the timespan (t-15,t-3), where t is the year of the current observation. Assuming lower secondary school ends at 15, this should roughly capture the completion rate for individuals aged 18-30 in the dataset.

i. To create this variable, I suggest you first create a new variable equal to the secondary school completion rate, if this is non-missing, otherwise 0.

ii. Then create two new variables, called nonmissingtotal and completionsum, each initially equal to zero.

iii. Then, store the sum of the 3^{rd} though 15^{th} lags of the completion rate in completionsum, and the count of nonmissing values for the 3^{rd} though 15^{th} lags of the completion rate. You can either

do a long way (using **replace**) or shortly using a **forval** loop over the lag values 3 through 15. (Note: You will definitely want to use the **lag** operator to complete this part).

iv. Finally, create a variable that is the secondary school completion average, which is equal to your completionsum / nonmissingtotal.

2 **Exploratory Analysis:** The outcome variable for this analysis is log of GDP per capita. The explanatory variables are the other indicators I have asked you to keep in the datset (but substituting the derived average of secondary school completion for the raw rate). For the dependent variable and all the explanatory variables:[2]

(a) Use the **mdesc** command to explore how many missing values there are in the dataset.

(b) Provide summary statistics (mean, standard deviation, min and max).

3 **Regression Analysis:**

(a) Regress the outcome variable on the explanatory variables (no need to comment about it yet).

(b) Then repeat this regression (still using **regress**), but include country and year fixed effects.[3] How do your results for the explanatory variables change?

(c) Test if your results from 2 are heteroskedastic. What does this indicate about the model?

(d) Now repeat your regression from (3b), but specifying it in a manner that uses heteroskedastic-robust standard errors. How do your results change?

(e) Repeat the regression in (3b) again, but now in a manner that allows there to be correlation in the errors for each country. How do the results change?

(f) Now specify (3b) as a fixed effects regression. Remember that you can include the group fixed effects by using the option **fe**. Perform this regression first making the assumption of homoscedastic and uncorrelated standard errors.

(g) Test for autocorrelation using either the Wooldridge (2002) test or the Inoue and Solo (2006) test. Suggest some reasons why the errors might be autocorrelated.

[2] You do not need to repeat the results you get in commands, just the Stata output will suffice.

[3] Instead of indicator variables for the fixed effects yourself, just use the **i.** prefix to the (encoded) country and year variables to tell Stata to create indicators from them

(h) Now run the fixed effects regression in a manner that allows for correlation in the residuals for each country. Interpret your results for this regression (be specific) and comment on how your results changed between this regression and (3f).

(i) Discuss whether you think we should trust the results of this regression. Specifically, focus on whether you think the fixed effects regression model Assumption FE.4 holds. List at least 4 concerns that you might have with this regression.

(j) Can you think of any reasons why the effect of some of these variables might be attenuated when using a fixed effects regression?

4 For three person groups only:

(a) You worry about omitted variable bias stemming from the fact that there is persistence in GDP and growth, so that the GDP in the previous year has a strong independent affect on the GDP in the next. Trying to address this, rerun your regression in (3h) but now including the lag of per capita GPD.

(b) State why including the lagged dependent variable in (4a) likely violates one of the fixed effects regression error assumptions.

(c) Try modifying the regression (3h) by using at least two of the following alternative explanatory variables either instead of some of those in the original analysis, or in addition to them:[4]

- Domestic credit to the private sector (indicator code: FS.AST.PRVT.GD.ZS)
- Tax rate (indicator code: GC.TAX.TOTL.GD.ZS)
- Sanitation rate (indicator code: SH.STA.ACSN)
- Tuburculosis rate (indicator code: SH.TBS.INCD)
- Female mortality rate (indicator code: SP.DYN.AMRT.FE)
- Tertiary enrollment rate (indicator code: SE.TER.ENRR)
- Cellphones usage rate (indicator code: IT.CEL.SETS)

(d) Interpret how your results change. Discuss how the new variables may address omitted variable bias, but also discuss give examples of some omitted variables that may be correlated with your new explanatory variables.

(e) How does the number of observations per group change with your new regression specification? How is that likely to affect your hypothesis testing in your analysis?

Good luck!

[4] Be sure to consult the Metadata Glossary for the precise definitions. You may omit some of the original explanatory variables if you wish.