BUFN 745 Problem Set 1

Spring 2023

Professor Alex He

Due Monday, 2/20/2023 at 11:59pm EST

Please submit an electronic version on Elms

1 Short Questions [15 Points, 5 points each]

For True/False questions please briefly explain your reasoning.

1. Suppose we have a panel data of of 100 firms over 20 years at a monthly frequency. We are interested in how firms' bond ratings affect their investment. Give an example of an omitted variable that can be fixed by Fixed Effects or First Differences. Give another example of an omitted variable that cannot be fixed by Fixed Effects or First Differences.

2. Suppose you are interested in the effect of being included in the S&P 500 index on stock prices. Give at least two reasons why OLS might not be the best linear unbiased estimator.

3. Is the following statement true or false?
   "If fixed effects model and random effects model give very different estimates, then the estimate from the random effects model is consistent."
   Please explain the reasoning for your answer.

2 Housing Prices [20 Points]

In this exercise you will work with a dataset containing housing prices from a large sample of neighborhoods. **Use Python and Google Colab to answer the following questions.** You may write your text responses either in the document or in the Colab (Jupyter) notebook. Please use the template Python file (Problem Set 1 template.ipynb) in the shared Google Drive folder or on Canvas to load the data and guide you through the questions.

1. (4 points) Run a regression with the following model, and report the R^2 of the regression.

$$price_i = \beta_0 + \beta_1 crime_i + \beta_2 nox_i + \beta_3 dist_i + \beta_4 radial_i + \beta_5 proptax_i + \epsilon_i$$

2. (4 points) Run a Breusch-Pagan test for heteroskedasticity, and report the result of this test. Is there heteroskedasticity present? If there is, fix this problem by running a second regression using White standard errors. Explain what has (and hasn't) changed.

3. (4 points) Run an F-test to check if *proptax* and *radial* jointly have no effect. What do you conclude? [Hint: If you found there was heteroskedasticity in part 2, think about which regression you should use for this test.]

4. (4 points) Run a Ramsey RESET test for functional form specification, and report the result of this test. Is the linear model misspecified?

5. (4 points) Run a regression using a log variant of the original model, as shown below. [Hint: You will need to create a new variable that is the natural log of price.] Make sure to use White standard errors if you found there was heteroskedasticity. Compare the goodness-of-fit statistics for this model and the original model. Which one is a better fit?

$$\log(price_i) = \beta_0 + \beta_1 crime_i + \beta_2 nox_i + \beta_3 dist_i + \beta_4 radial_i + \beta_5 proptax_i + \epsilon_i$$

3 Value of Patents [30 Points]

In this exercise we study the value of patents for firms. Suppose we have a panel dataset for a sample of publicly listed firms. For each firm we have the number of patents granted to that firm in each year and the stock price and market value by the end of each year.

1. (3 points) Consider an OLS model:

$$y_{it} = \beta_0 + \beta_1 p_{it} + \epsilon_{it}$$

where y_{it} is the market value, and p_{it} is the number of patents for firm i in year t. Suppose α_i is unobserved firm-specific characteristic, like firm culture and managerial talent. What assumption do you need for the OLS estimator to be unbiased? Do you think the assumption is satisfied here and why?

2. (3 points) If the assumption from 1 is satisfied, is the OLS estimator BLUE? What procedure can you do to get an estimator that is BLUE?

3. (4 points) If the assumption from 1 is not satisfied, what are the two ways to get unbiased and consistent estimates? Use one or two sentences to describe each approach.

Use Python and Google Colab to answer the following questions.

4. (2 point) How many firms have at least one patent? How many firms have no patent?

5. (2 points) Calculate market capitalization by multiplying the stock price by shares outstanding. What is the average market capitalization?

6. (3 points) Create a variable "log_cap" that is the log market capitalization. Run an OLS regression of *log* market capitalization on the number of patents. What is the coefficient?

7. (5 points) Run a regression of *log* market capitalization on the number of patents with **firm and year fixed effects**. What is the coefficient?

8. (5 points) Run a **first difference** regression of *log* market capitalization on the number of patents. What is the coefficient?

9. (3 points) Using the average firm's market capitalization, what is the value of a patent using the three methods (OLS, FE, and FD)? (for example, if the average market capitalization is 1 million, and coefficient is 0.01, then one patent increases the market capitalization by 1 percent $\approx \log(1 + 0.01) - \log(1)$, which is 0.01×1 million = 10,000)

4 Predicting Sovereign Default [30 Points]

In this exercise will use data from World Bank from 1991 to 2017 to predict default of sovereign debt. Please use Python and Google Colab to answer the following questions.

1. (5 points) Split the data into a training sample (before 2010) and a test sample (2010 and afterwards). Fit an OLS model to the training sample and predict the default in the test sample. What is the problem with the predicted value of default?

2. (6 points) Estimate a logit model for the training sample and use bootstrap to get the standard errors of coefficients.

3. (4 points) What is the marginal effect for Total International Reserves ("irtd") based on the logit model?

4. (5 points) Estimate a probit model for the training sample and use bootstrap to get the standard errors of coefficients.

5. (4 points) What is the marginal effect for Total International Reserves ("irtd") based on the probit model?

6. (6 points) A common way to assess the predictive efficiency of a binary model is the ROC curve. It plots the true positive rate ($= \frac{\text{True Positive}}{\text{True Positive+False Negative}}$) against the false positive rate ($\frac{\text{False Positive}}{\text{False Positive+True Negative}}$). If a test is the diagonal line from bottom left corner to upper right corner, then the test is completely useless. The bigger the area is between the ROC curve and the diagonal line, the better the test (Read more about ROC curve in https://en.wikipedia.org/wiki/Receiver_operating_characteristic). Plot the ROC curve for logit and probit using the test sample and compare the accuracy of the two predictions. [Hint: from the *sklearn.metrics* library, use *roc_curve* function to plot the curve, and use *roc_auc_score* function to calculate the area under curve]

7. (extra points: 5 points) Use machine learning methods (such as PCA) to select variables from the list of all variables and then run a logit regression and plot the ROC curve. Does this improve the prediction?