ECOM20001: Econometrics 1

Assignment 2

Student Information

To receive an assignment grade, you must fill out the information in this table and include this page as your assignment cover page.

Name	Student ID Number	Tutor	Tutorial Day & Time	Tutorial Location
Sally Probability	422552	Richard Hayes	Tue 10:15am	The Spot 4452
Markus Statistics	653223	Sahiba Narang	Wed 4:15pm	The Spot 3054

Due Date and Weight

- **Submit via the LMS by 8am on Monday, 6 May 2019.**
- No late assignments will be accepted.
- This assignment is worth 5% of your final mark in ECOM20001.
- There are 50 marks in total.

What You Must Submit via the LMS

- **Assignment answers**, no more than 10 A4 pages with 12 point font. 5 points out of 25 will be deducted if you answers exceed 10 A4 pages.
- The **R code** that generates your results. Specifically, copy-and-paste your R code in an Appendix at the end of your assignment document (e.g., in the .docx file) so that it can be viewed and tested by markers. The R code Appendix does not count toward your 10 page answer limit. You may alter and shrink the R code font to less than 12 point font so that it is easier to read. 2 points out of 50 will be deducted if you do not include your R code.

Additional Instructions

- You may submit this assignment in groups of one or two. Students in a group of two are allowed to be in different tutorials.
- You must complete the assignment in no more than 10 A4 pages with 12 point Arial, Times New Roman, Helvetica, Cambria or Calibri font. The assignment cover page does not count as one of the 10 A4 pages.
- To save time, you may cut and paste RStudio output directly into your answers in reporting empirical results. You are also free to create your own better-formatted tables based on your RStudio output, which is of course good practice in learning how to present empirical results.
- Figures may also may be copied and pasted directly into your assignment answers. They may be scaled down in size to meet the 10 page limit, but please ensure that your figures are readable. If they are not, marks will be deducted.
- Marks will be deducted if interpretations of results are incorrect, imprecise, unclear, or not well-scaled. Similarly, marks will be deducted if figures or tables are incorrect, unclear, not properly labeled, not well-scaled, or missing legends.
- This R code in the Appendix at the end of your assignment (as discussed on the previous page) must be clearly commented and easy for the subject tutors to follow. If the code is not well commented and easy to follow, marks will be deducted.
- Students with a genuine reason for not being able to submit the assignment on time can apply for special consideration to have the assignment mark transferred to the exam at the following link:
    - https://students.unimelb.edu.au/admin/special/

Getting Started

Please create an Assignment1 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following data file into the Assignment1 folder:

- as2_billions.csv

This dataset contains the following 11 variables:

- country: country name
- year: year
- numbil: number of billionaires in the country
- pop: population of the country
- gdppc: Gross Domestic Product (GDP) per capita
- gattwto08: number of years in the General Agreement on Tariffs and Trade (GATT) or World Trade Organisation (WTO) as of 2008
- fullprivproc: privatisation proceeds in the economy (billions of dollars)
- topint08: top marginal income tax rate (%) in 2008
- rintr: real interest rate in (%) in 2008
- roflaw: rule of law index in 2008
- nrrents: natural resource rents as a % of GDP in 2008

In total, the dataset contains this information for 1,094 (country, year) observations, with the sample spanning 9 years (2005-2013) across 123 unique countries.

About the Assignment

In this assignment, we explore macroeconomic variables determines the number of billionaires, and hence extreme wealth, in a country. The reference for this dataset is Treisman, D. (2016): “Russia’s Billionaires,” *American Economic Review: Papers and Proceedings*, 106(5), 236-241.

Questions

1. **(4 marks)** Compute summary statistics (mean, standard deviation, min, max) for all variables in the dataset. Describe in words a typical observation based on the sample means. All determine which variables should be rescaled and choose them using an appropriate scaling factor. You answer should be no more than 3 sentences.

2. **(3 marks)** Sort the data in terms of numbil so that the first observation is the one with the most billionaires in the sample.[1] Report, in a table, the list of country and year for the top 20 observations. In no more than 3 sentences total, answer the following questions:

   - Which countries can be found in the top 20 observations in terms of numbil?
   - Does a particular country stand out in terms of the number of times it is in the top 20, and the number of billionaires it has relative to the next country in terms of billionaire counts?

3. **(3 marks)** Construct the following three scatter plots, where in each case the first variable (before the vs.) goes on the vertical axis and the second variable (after the vs.) goes on the horizontal axis. Ensure correct axes and graph titles.

   - numbil vs. gdppc
   - numbil vs. gattwto08
   - gdppc vs. gattwto08

4. **(6 marks)** Suppose you ran a single linear regression where the dependent variable numbil and the independent variable is gdppc. Further suppose that gattwto08 is an omitted variable. Based on the three scatter plots from question 3, carefully explain what the direction of the bias would be for the slope coefficient on gdppc in the single linear regression of numbil on gdppc. You answer should have no more than 5 sentences.

5. **(4 marks)** Using the as.numeric() command in R, construct dummy variables for all years in the dataset, and denote these dummy variables as d2005, d2006, d2007, d2008, d2009, d2010, d2011, d2012, d2013. Suppose you tried to run a regression of numbil on a constant and d2005, d2006, d2007, d2008, d2009, d2010, d2011, d2012, d2013. In no more than 3 sentences describe the problem that you would run into.

[1] To do this question in R, you use the order() and head() commands. Alternatively, you can simply sort the data in an Excel spreadsheet and present the results based on this.

6. **(8 marks)** Run the following 8 regressions (labelled Reg (1) to Reg (8)), where in each case the dependent variable is numbil, the regression includes a constant, and each bullet point below lists the other independent variables to be included:

   - Reg (1): pop
   - Reg (2): pop, gdppc
   - Reg (3): pop, gdppc, gattwto08
   - Reg (4): pop, gdppc, gattwto08, fullprivproc
   - Reg (5): pop, gdppc, gattwto08, fullprivproc, topint08
   - Reg (6): pop, gdppc, gattwto08, fullprivproc, topint08, rintr
   - Reg (7): pop, gdppc, gattwto08, fullprivproc, topint08, rintr, roflaw, nrrents
   - Reg (8): pop, gdppc, gattwto08, fullprivproc, topint08, rintr, roflaw, nrrents, d2006, d2007, d2008, d2009, d2010, d2011, d2012, d2013

7. **(10 marks)** Based on the regression results table from question 6, answer the following questions:

   - Compare the results in Reg (2) and Reg (3). Does the change in the coefficient on gdppc correspond to the patterns you documented in question 4 above? Please answer in no more than 2 sentences.
   - Compare the results across Reg (2) to Reg (8) in the table. Across which columns does the coefficient on gdppc start to “settle down,” that is not change drastically across consecutive columns? Please answer in no more than 3 sentences.
   - Focusing on the results in Reg (8). Which variables have a statistically significant regression coefficient (from 0) at the 5% level? Ignoring the constant, provide 2 seperate interpretations for each of these significant coefficients: (1) for one interpretation consider a change in the independent variable by 1 unit; and (2) for the other consider a 1 standard deviation change in the independent variable. You should be able to provide all of these interpretations for all of the statistically significant regressors in no more than 10 sentences.

8. **(2 marks)** Compare the R-squared and adjusted R-squared for Reg (5) and Reg (8). Explain how it is possible for Reg (8) to have a higher R-Squared, but smaller adjusted R-Squared than Reg (5). Please answer in no more than 3 sentences.

9. **(2 mark)** Would you reject any of the regressions in the table from question 6 at the 5% level of significance? Please answer in no more than 2 sentences.[2]

10. **(6 marks)** Based on the Reg (8) regression specification, conduct the following joint hypotheses tests, where "B_var" denotes the regression coefficient for the variable var, and != means "not equals." For each test, report the relevant test statistic, degrees of freedom, and p-value. Briefly interpret the results of each test in no more than 2 sentences per test.

    1. H0: B_roflaw=0 AND B_nrrents=0 vs H1: B_roflaw!=0 OR B_nrrents!=0
    2. H0: B_topint08=B_rintr vs. H1: B_topint08!=B_rintr

HINT: for this question, you will need to use the **linearHypothesis()** command in R. See tutorial 9 on the LMS site and tute9.R specifically for detailed explanations and examples of how to use it.

11. **(2 marks)** R-code: we will review and mark your R code according to the following scheme:

    - 2/2 if R code is correct and organised and commented like the solution code for the assignment.
    - 1/2 if R code is correct, but hard to follow or not well commented.
    - 0/2 if R code is incorrect and/or a complete mess, or not submitted.

[2] You may conduct these tests based on the regression F-statistics produced by the 'stargazer' command in R. Recall from Tutorial 8 (see tute8.R specifically) that these F-statistics do not account for heteroskedasticity. You may alternatively update the regression F-statistics in the 'stargazer' table by computing them 'by hand' for each regression using the linearHypothesis() command in R (as we do in tute8.R), which correctly accounts for heteroskedasticity.