Assignment 2: Who Wants to be a Billionaire (in a non-linear world)?

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			

Due Date and Weight

- **Submit via the LMS by 3pm on Friday, 19 October.**
- Group registration for this assignment closes at **5pm on Monday 15 October**.
- No late assignments will be accepted.
- This assignment is worth 10% of your final mark in ECOM20001.
- There are 45 marks in total.

What You Must Submit via the LMS

- **Assignment answers**, no more than 12 pages with 12 point font. 5 points out of 45 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.

Additional Instructions

- You may submit this assignment on your own, or in groups of two, to which you have been assigned in your tutorials by your tutor. Groups of people attending different tutorials are not allowed.

- You must complete the assignment in **no more than 12 A4 pages** with 12-point Arial, Times New Roman, Helvetica, Cambria or Calibri font, at least 1.15 points interline space, and with at least 2.54 cm margins on all sides of the page. 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.

    - Note: Questions 4 and 5 require you to construct tables of results; submitting RStudio output alone for these questions will result in deducted marks.

- **Figures** may also 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 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. Remember to always clearly label the x-axis and the y-axis of your figures, and to add legends if there is more than one line in the figure.

- 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 transfer to the exam at the following link:

    - https://students.unimelb.edu.au/admin/special/

Who Wants to be a Billionaire (in a non-linear world)?

In this assignment, we continue to explore which countries have more billionaires. But we now use longitudinal data (i.e., with several years of observation for each country) and we examine the relationships with population size, natural resources rents, GDP per capita and the rule of law.

Getting Started

Please create an Assignment2 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following files into the Assignment2 folder:

- Billionaires_clean2.csv

This dataset contains the following 7 variables:[1]

- country --- Name of the country
- year --- Year of observation (2005 to 2013)
- numbil0 --- Number of billionaires in country, 0 if none in the Forbes list
- natrent --- Total natural resources rents in $US
- pop --- Population size
- gdppc --- GDP per capita of the country, in thousands of current $US
- roflaw --- Rule of law index (0 to 1)

There is complete data for 9 years (2005 to 2013) across 63 different countries. The data is unbalanced in that some countries have data for all 9 years (like Australia), while others have data for some years but not all (like Saudi Arabia, which has data available for only 2 years). In total there are 444 country-year pairs.

[1] The reference for the dataset is Treisman, D. (2016). Russia's Billionaires. *American Economic Review*, 106(5), 236-41. The data are from selected countries from 2005 to 2013.

Assignment Questions

1. Report and discuss summary statistics only for numbil0, natrent, pop, gdppc, roflaw. What does a typical country look like? Do any variables require rescaling? If so, rescale as you deem to be appropriate, and work with the rescaled data for the remainder of the assignment.

   Your discussion should be no more than **3 sentences** long.

   (3 marks)

2. Provide the following three scatter plots using the ggplot() command in R which also plots a line of best fit with the scatter plot (option "formula = y ~ poly(x,1)"):

   • numbil0 on the vertical axis, natrent on the horizontal axis

   • numbil0 on the vertical axis, pop on the horizontal axis

   • natrent on the vertical axis, pop on the horizontal axis

   Briefly interpret the pattern in each scatter plot. Based on your figures, what would be the direction of omitted bias with an OLS regression coefficient in a simple linear regression with numbil0 as the dependent variable and natrent as the independent variable?

   In total, your discussion of the scatter plots and omitted variable bias should be no more than **4 sentences long**.

   (4 marks)

3. Provide the following scatter plot using the ggplot() command in R which also plots a line of best fit with the scatter plot (option "formula = y ~ poly(x,1)"):

   • numbil0 on the vertical axis, gdppc on the horizontal axis

   • numbil0 on the vertical axis, roflaw on the horizontal axis

   • gdppc on the vertical axis, roflaw on the horizontal axis

   Briefly interpret the pattern in the scatter plot. Based on your figures, what would be the direction of omitted bias with an OLS regression coefficient in a simple linear regression with numbil0 as the dependent variable and gdppc as the independent variable?

   In total, your discussion of the scatter plots and omitted variable bias should be no more than **4 sentences long**.

   (4 marks)

4. Estimate the following 5 multiple linear regression models where numbil0 is the dependent variable in each regression. The list of independent variables in each respective regression is:

   • natrent

   • natrent, pop

   • natrent, pop, gdppc

   • natrent, pop, gdppc, roflaw

   • natrent, pop, gdppc, roflaw, plus an appropriate set of year dummy variables to control for year of the sample

   You need to construct the year dummy variables in your code; they are not provided. In my solution code, I label the dummies d2005, d2006, d2007, d2008, d2008, d2009, d2010, d2011, d2012, d2013. You may want to do the same.

   Present your regression results in a table that has 6 columns.

   • The first column contains the independent variable names

   • Columns 2 through 6 contain the regression results for each of the 5 multiple linear regression models listed above. Specifically, each regression coefficient estimate should be reported with its standard error below it in parentheses ( ).

   • Report **heteroskedasticity-robust standard errors** in this question, and for all other regressions for the remainder of the assignment. (NB: The R code making use of the "stargazer" package that was provided to you with the solutions to Assignment 1 was designed to report homoscedastic standard errors only)

   • To avoid the dummy variable trap, make **2005** your base year in any relevant regressions.

   • Put ** and * markers on the coefficient estimates to indicate statistical significance of a test of the 2-sided null that the coefficient equals 0 at the 1% and 5% levels, respectively.

   • At the bottom of each column the regression Adjusted R-Squared and number of observations used in the regression should be reported. For all other regressions that appear in the assignment, continue to report the Adjusted R-Squared and number of observations used.

   • An example table at the end of this document provides an example of the table structure you must follow. Be sure to include a clear table title and footnote.

In **no more than 10 sentences**, discuss the main results from the table.

   • Discuss the main results from the table, focusing on the coefficient estimate on natrent. Discuss its statistical significance and discuss the magnitude of the coefficient estimates

that are statistically significant. What happens as you add pop and then roflaw to the regressors?

• What makes more sense for interpreting the results: describing a one-unit change in roflaw or a 0.1-unit change in roflaw?

• As part of your discussion, provide an example where omitted variable bias in the regression impacted the coefficient estimate on gdppc and the intuition behind the change in the gdppc coefficient once the variable was controlled for in the regression.

• Interpret the coefficient estimate on the 2013 year-dummy.

(7 marks)

5. Run the following three regressions:

   • Regression 1

      • dependent variable: numbil0

      • independent variables: log(pop), natrent, gdppc, roflaw

   • Regression 2

      • dependent variable: log(numbil0)

      • independent variables: pop, natrent, gdppc, roflaw

   • Regression 3

      • dependent variable: log(numbil0)

      • independent variables: log(pop), natrent, gdppc, roflaw

   Present your results in a table like you constructed for question 4. An example table at the end of this document provides an example of the table structure you must follow. Be sure to include a clear table title and footnote.

   In **4 sentences max**, interpret the main results, focusing on the coefficient estimate on either pop or log(pop) and its statistical significance and magnitude.

   (4 marks)

6. Construct a set of interactions between log(pop) and each of the year dummy variables you created. So, for example, in the solution code I create a new variable called lnpop_d2006 which is the product of log(pop) and d2006: lnpop_d2006= log(pop) x d2006. Create this interaction, as well as similar interactions for all other year dummies in the data. Using your set of interactions, run the following regression:

- dependent variable: log(numbil0)
- independent variables: log(pop), lnpop_d2006, lnpop_d2007, lnpop_d2008, lnpop_d2009, lnpop_d2010, lnpop_d2011, lnpop_d2012, lnpop_d2013, natrent, gdppc, roflaw, d2006, d2007, d2008, d2008, d2009, d2010, d2011, d2012, d2013

To present your findings, you may simply copy-and-paste your R-output for the regression directly into your answers document. Continue to make 2005 the base year in your regressions to avoid the dummy variable trap.

- Interpret the main results from the regression, focusing on the statistical significance and magnitude of the coefficient on log(pop), and the statistical significance of the coefficients of regressors involving log(pop) and year dummy variables interactions. (3 sentences max)
- Interpret the magnitude and statistical significance of the impact of pop on numbil0 in 2013. (4 sentences max)
- Also carefully explain two ways in which this regression avoids the dummy variable trap. (2 sentences max)

In total, your interpretation of the regression results and dummy variable trap should be no more than **9 sentences long**.

(8 marks)

7. Using your regression from question 6, test the joint null hypothesis tests based on the regression coefficients for the following regressors, reporting the F-statistic, degrees of freedom, and p-value for the test:

   lnpop_d2006=lnpop_d2007=lnpop_d2008=lnpop_d2009=lnpop_d2010=lnpop_d2011=lnpop_d2012=lnpop_d2013=0

   Interpret your test results in **no more than 2 sentences**.

   (3 marks)

8. Run the following regression

   - dependent variable: log(numbil0)
   - independent variables: log(gdppc), roflaw, log(gdppc)_roflaw, log(pop), natrent, where log(gdppc)_roflaw =(log(gdppc)) x roflaw.

   To present your findings, you may simply copy-and-paste your R-output for the regression directly into your answers document.

Interpret the statistical significance and signs (but not the magnitudes) of the regression coefficients on log(gdppc), roflaw, log(gdppc)_roflaw in the regression.

Together what do they suggest about the relationship between the rule of law and the number of billionaires as GDP per capita increases?

In total, your answer should **not exceed 4 sentences**.

(3 marks)

9. Using your regression results in question 8, compute the elasticity of numbil0 with respect to gdppc for roflaw =0.1, 0.2, …, 1.

   • For each of these 10 partial effects, test whether the elasticity is statistically different from 0. Explain why and how you proceed.

   • Report your results in a table that has 10 rows (for roflaw =0.1, 0.2, …, 1) and 4 columns: roflaw value, elasticity for a given roflaw value, the F-statistic for the test of the null that the elasticity for a given roflaw value equals 0, and the corresponding p-value for the test. Explain how you carry out these tests.

   • Discuss your findings and briefly provide a potential economic explanation.

   In total, your answer should **not exceed 4 sentences.**

   (4 marks)

10. Your R Code will be graded as part of the assignment.

    - 5/5 if the code is as clear as the code from the tutorials
    - 3/5 if anything in the code is unclear or uncommented
    - 0/5 if the code is an incomprehensible mess
    - If the code does not run and/or the results of the assignment cannot be replicated by your tutor, you could fail the assignment!

    (5 marks)

Table Structure for Question 4
FAKE DATA EXAMPLE; FOR ILLUSTRATIONS PURPOSES ONLY

Billionaires and country characteristics

	(1)	(2)	(3)	(4)	(5)
natrent	0.025	-0.012	-0.014	0.018	0.014
	(0.016)	(0.018)	(0.017)	(0.013)	(0.013)
pop		0.069**	0.091**	0.103**	0.104**
		(0.016)	(0.017)	(0.016)	(0.016)
gdppc			0.119**	-0.035*	-0.040**
			(0.032)	(0.015)	(0.015)
roflaw				23.923**	25.005**
				(2.916)	(2.915)
d2006					1.988
					(2.308)
d2007					1.916
					(1.876)
d2008					2.618
					(2.284)
d2009					1.101
					(2.241)
d2010					2.697
					(2.101)
d2011					7.590**
					(2.565)
d2012					3.507
					(2.123)
d2013					6.691**
					(2.305)
Constant	8.763**	6.713**	2.538*	-10.317**	-13.991**
	(0.778)	(0.661)	(1.050)	(1.847)	(2.357)
N	321	321	321	321	321
adj. R^2	0.003	0.086	0.146	0.274	0.299
F	2.456	10.982	11.296	23.332	9.986

Dependent variable is the number of billionaires in the country. natrent is natural resources rents in billions of dollars. pop is total population of a country in millions. gdppc is GDP per capita in thousands of dollars. roflaw is an index between 0 and 1 on the level of the rule of law in a country. dXXXX is a dummy variable for the year XXXX. Heteroskedasticity robust standard errors in parentheses.
Statistical significance from two-sided tests of the null of no effect marked as * for 5% and ** for 1%

Table Structure for Question 5
FAKE DATA EXAMPLE; FOR ILLUSTRATIONS PURPOSES ONLY

Billionaires and population size

	(1) numbil0	(2) log(numbil0)	(3) log(numbil0)
lnpop	1.610**		0.234**
	(0.307)		(0.046)
pop		0.008**	
		(0.002)	
natrent	0.040*	0.009**	0.009**
	(0.017)	(0.003)	(0.003)
gdppc	-0.024	-0.006	-0.000
	(0.021)	(0.006)	(0.006)
roflaw	18.772**	2.838**	2.180**
	(2.811)	(0.601)	(0.606)
Constant	-6.916**	1.926**	1.306**
	(2.492)	(0.373)	(0.404)
N	321	321	321
adj. R^2	0.178	0.113	0.131
F	16.544	11.158	14.762

numbil0 is the number of billionaires in the country. natrent is natural resources rents in billions of dollars. pop is total population of a country in millions. gdppc is GDP per capita in thousands of dollars. roflaw is an index between 0 and 1 on the level of the rule of law in a country. Heteroskedasticity robust standard errors in parentheses.
Statistical significance from two-sided tests of the null of no effect marked as * for 5% and ** for 1%