ECOM20001: Econometrics 1

Assignment 2

Student Information

You must fill out this table and include it as a front cover page for your assignment. Only students whose name and student ID number are included on the cover page will receive marks. Groups of up to 3 students are allowed.

Name	Student ID Number
Sally Probability	422552
Yin Statistics	653223
Ahmed Hypothesis	883190

Due Date and Weight

- **Submit via LMS by 8am on 11 May 2020**
- 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 LMS

- **Assignment answers**, no more than 8 A4 pages with 12 point font and 1 cm margins (max). 5/50 marks deducted if any of these restrictions are violated.
- 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 marks will be deducted if you do not include your R code.**

Additional Instructions

- You may submit this assignment in groups of up to three students. Students in a group are allowed to be in different tutorials. You are also most welcome to submit your assignment not as part of a group.
- You must complete the assignment in no more than 8 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 8 A4 pages, nor does the R code in the Appendix.
- 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 8 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 Assignment2 folder on your computer, and go to the Canvas site for ECOM 20001 and download the following data file into the Assignment2 folder:

- as2_crime.csv

This dataset contains the following 9 variables:

- state: name of state in the United States
- year: year
- robbery_rate: number of robberies in the state per 100,000 people
- assault_rate: number of assaults in the state per 100,000 people
- burglary_rate: number of burglaries in the state per 100,000 people
- black: percentage of the population that is black/African American
- income: average household income
- age: average individual age
- female: percentage of the population that is female

In total, this dataset contains 11 years (2000-2010 inclusive) of crime and demographic data for each of the 50 American states, yielding 550 (state, year) observations in total.[1]

About the Assignment

In this assignment, we study what determines the level of crime in a given state, as measured in terms of the rates of robberies, assaults, and burglaries. We are interest in particular in the relationship between the share of the population that is black and the level of crime in a state. There is substantial body of research in economics on large black/white disparities in the United States that identifies where they exist (e.g, crime, income, education), why they exist (e.g., discrimination, poverty traps), and what policy can do to mitigate them (affirmative action policies, anti-discrimination laws). In this assignment, we seek to identify whether crime tends to be higher or lower in states whose populations have a larger share of black citizens.

[1] The crime data comes from Doleac, Jennifer L. (2017): “The Effects of DNA Databases on Crime,” *American Economic Journal: Applied Economics*, 9(1), 165-201.

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. Also determine which demographics variables should be rescaled for regression analyses as these will enter as independent variables in regressions below. Rescale them using an appropriate scaling factor and create a new variable called "var_rescale" where "var" is the name of the original variable you rescaled.[2]

2. **(6 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 and for each scatter plot ensure you overlay a single-linear regression line that helps visualise the relationship between the two variables.

   - robbery_rate vs. black
   - robbery_rate vs. income
   - black vs. income

3. **(6 marks)** Suppose you ran a single linear regression where the dependent variable robbery_rate and the independent variable is black. Further suppose that income is an omitted variable in the regression. Based on the signs of the estimated regression line slopes in the three scatter plots from question 2, carefully explain what the direction of the bias would be for the slope coefficient on black in the single linear regression of robbery_rate on black. Also determine whether the coefficient would be too large or too small in magnitude based on the sign we expect for the coefficient on black (based on the historical experience with racial discrepancies in the US) in a regression of robbery_rate on black.

4. **(4 marks)** Using the as.numeric() command in R, construct dummy variables for all years in the dataset, and denote these dummy variables as d2000, d2001, d2002, d2003, d2004, d2005, d2006, d2007, d2008, d2009, d2010. Suppose you tried to run a regression of robbery_rate on a constant and d2000, d2001, d2002, d2003, d2004, d2005, d2006, d2007, d2008, d2009, d2010. What problem would arise if you tried to run this regression? Also run this regression in R. What does R do to avoid the problem? (note: you do not have to report the regression results for your answer, just state what R does to fix the problem).

[2] To take a completely irrelevant example, if you were to rescale year by 1000, then you are to create a new variable in R called year_rescale=year/1000.

5. **(6 marks)** Run the following 5 regressions — labelled Reg (1) to Reg (5) — where in each case the dependent variable is robbery_rate, the regression includes a constant, and each bullet point below lists the other independent variables to be included:

   - Reg (1): black
   - Reg (2): black, income
   - Reg (3): black, income, age
   - Reg (4): black, income, age, female
   - Reg (5): black, income, age, female, d2001, d2002, d2003, d2004, d2005, d2006, d2007, d2008, d2009, d2010

where note for any variable in the list that you have scaled from question 1 corresponds to its rescaled counterpart. Construct your table using the stargazer() command in R. For each regression report heteroskedasticity-robust standard errors[3] for each coefficient estimate, the adjusted R-squared for model fit, and the number of observations used in running the regression.

6. **(10 marks)** Based on the regression results table from question 5, answer the following questions:

   A. Compare the results in Reg (1) and Reg (2). Does the change in the coefficient on black correspond to the patterns you documented in questions 2 and 3 above?

   B. Compare the results across Reg (2) to Reg (5) in the table. Across which columns does the coefficient on black start to “settle down,” that is not change drastically across consecutive columns? In Reg (5) is the regression coefficient on black statistically significant at the 5% level?

   C. Focusing on the year dummy variables coefficient estimates in Reg (5), explain what the base group is and interpret whether robbery_rate is statistically significantly higher or lower in particular year s across the US states using a 5% level of significance. Also interpret the coefficients on year dummies that are statistically significantly different from 0.

   D. Focusing on the results in Reg (5). Ignoring the constant, provide 2 seperate interpretations for how robbery_rate changes with black each of these

[3] Assume heteroskedastic errors throughout the entire assignment in conducting all hypothesis tests.

significant coefficients. In particular: (1) how much does robbery_rate change when black increases by 1 unit?; and (2) how much does robbery_rate change when black increases by 1 standard deviation? Which interpretation is more relevant given the scale of black and its variation in the dataset?

E. Finally, based on the more relevant predicted change in robbery_rate from in question 6D, compare the predicted change in robbery_rate to the sample mean of robbery_rate and comment on whether you think it is large-magnitude change in robbery_rate.

7. **(4 marks)** Construct another regression table, except this time run the following three regressions:

- Reg (1): dependent variable is robbery_rate
- Reg (2): dependent variable is assault_rate
- Reg (3): dependent variable is burglary_rate

Where in each regression the set of independent variables to be used is black, income, age, female, d2001, d2002, d2003, d2004, d2005, d2006, d2007, d2008, d2009, d2010. Construct your table using the stargazer() command in R. For each regression report heteroskedasticity-robust standard errors for each coefficient estimate, the adjusted R-squared for model fit, and the number of observations used in running the regression. Notice that Reg (1) in this table simply repeats Reg (5) from the regression table from question 5.

8. **(8 marks)** Answer the following questions based on the regression output from question 7:

A. Using a 5% significance level, interpret economic magnitude of the coefficients on black along with their statistical significance Reg (2) and (3).

B. Compare the R-Squared's across Regs (1)-(3). Which form of crime is best predicted by the set of regressors considered?

C. For Reg (1)-(3), compute the 95% confidence interval of the respective changes in robbery_rate, assault_rate, and burglary_rate associated with a 1 standard deviation increase in black.

D. Compute the overall regression F-statistic for Reg (1)-(3) estimated in question 7. Report the F-statistic, relevant degrees of freedom for the test, and the p-value of the test, ensuring that your F-statistics and p-values account for heteroskedasticity. Provide an interpretation of the statistical significance of the results, and also state in plain language/words what conclusion you can draw from each test for Reg (1)-(3).

9. **(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.