Data Exercise 1
Econ 322: Econometrics
Fall 2019
Yutian Yang

Upload this data exercise to Sakai by Tuesday, Oct 15th at 2:50 p.m.

Your document should include responses to the question. The answers themselves should be attached to the Sakai assignment as *Yourname_p1*.pdf file (export to .pdf from whatever software you use). Make sure your name is at the top. You may work together in small groups of 2–3 if you would like, but the write-up must be your own; you may not turn in identical write-ups. Please write the names of your collaborators on your homework as well. Also, include your .R script as a separate upload, with yournamen_p1.Rscript as the filename.

Question: "What affects your test score?"

- Download the test score data of 17 thousand Chinese junior high school students and the codebook from Sakai. The codebook gives you information about the variables in the data set. It is important to read it before interpreting your results.
- First, set your working directory to the folder where your data is located (Session > Set Working Directory > Choose Directory).
- Read the data into RStudio:
- Answer the following questions based on regressions and summarized statistics:

1. Given that the survey was done in 2013, generate a new variable named "age" based on the birth year of the student;

2. Run an OLS regression with the studying time, cognitive ability, age and the square of age as explanatory variables to estimate their influence on math score; based on the regression outcome, describe whether **studying time** and **cognitive ability** influences student's math test score; if they do, how they influence; also, specify the **Null Hypothesis** and **Alternative Hypothesis** of each estimated coefficients;

3. Generate the following dummy variables:
(1) **grade9**: the variable has value 1 if the student if in grade 9, otherwise has value 0;
(2) **high_income**: the variable has value 1 if the income status is higher than 3, otherwise has value 0;
(3) **mother_high_edu**: the variable has value 1 if the education level of the student's mother is equal or above 7;
(4) **father_high_edu:** the variable has value 1 if the education level of the student's father is equal or above 7;

4. What proportion of the students are in Grade 9; what proportion of students are female; what proportion of students are older than 13; what proportion of students are from a high-income family?

5. Include the gender information and the dummy variables you generated in Q3 into the OLS regression of Q2, use **robust standard errors** for the regression, show the outcome; based on the regression outcome, describe whether **studying time**, **cognitive ability, gender, from a high income family, having highly educated mother and father** influences student's math test score; if they do, how they influence;

6. Run two OLS regressions with the studying time, cognitive ability, age and the square of age as explanatory variables to estimate their influence on English score for female and male students**, separately**; based on the regression outcome, describe whether **studying time** and **cognitive ability** influences female and male student's English test score differently;

7. Include the dummy variables you generated in Q3 and the number of siblings into the OLS regression of Q6, use **robust standard errors** for the regression, show the outcome and compare the difference between influences of on female and male.

8. Find the correlation between **cognitive ability and math score, cognitive ability and English score.**

9. Are the means of math score of the females and males significantly different? Show your answer based on hypothesis testing. Specify the Null Hypothesis and the Alternative Hypothesis before you present the test outcome.

10. Are the means of English score of the females and males significantly different? Show your answer based on hypothesis testing. Also find the standard error of the difference between the means of the two groups. Specify the Null Hypothesis and the Alternative Hypothesis before you present the test outcome.

Note: For all the answers based on regressions, create a regression table with:

- the mean of the estimated coefficients
- the standard errors of coefficients,
- Indicate the *p*-value (*** indicates $p < 0.01$, ** indicates $p < 0.05$, and * indicates $p < 0.10$).
- the 95% confidence intervals,
- when talking about the significant level of an estimated coefficient, always say it is **significant at a certain level (1%, 5% or 10%)** or not**;**
- Do not just copy the R output and paste it into the table. Instead, create your own professional-looking table that reports the R output in a way that is easy to read and understand.
- The R tutorial has instructions on how to do these things. You should also Google around for R help!

Your table should be of "publication quality" with appropriate titles, labels, and table notes.

Programming Resources

- R Videos and Workshop Materials libguides.rutgers.edu/data/data_R
- Tutorials for Learning R r-bloggers.com/how-to-learn-r-2
- Introduction to R Seminar stats.idre.ucla.edu/r/seminars/intro

Additional Notes

- It is good practice to save your code with the file extension .R.
- Your code should have comments at the top. At a minimum, include your name and that the code is for Econ 322, data assignment 1.