Applied Regression and Time Series Analysis
Homework 2: OLS Inference

Paul Laskowski and Jeffrey Yau

January 1, 2016

Instructions

The weekly assignment serves two purposes: (1) Extend the materials taught in the asynchronized materials; some new concepts or techniques are introduced in the weekly assignment. (2) Ensure that you have learned the concepts, techniques, theories, statistical models covered in a specific week. Below are some guidelines:

- **Submission: Submit 2 files. Missing one of the two files will result in a 50% reduction in grade.**

    1. A report (in pdf format) detailing your answers and all the steps to arrive at your answers
    2. A well-documented R-script, jupyter notebook, or Rmd file detailing all of the codes used to arrive at your answers.

- Late submission will not receive any credit.
- Answers need to be typed. Latex, which you will likely have to use in R markdown, is great for typesetting documents with mathematical symbols.
- All the steps used to arrive at your final answers need to be shown clearly. These steps are as important as the final answer.
- The final answer of each question needs to be very easy identified; the use of bold fonts, highlights, or circling will help.
- This is a group project. Form a group with 3 or 4 people.
- Although this is a group project, we encourage you to attempt all of the exercises before discussing with your teammates. Do not use the "division-of-labor" approach. Each of the students in a group is expected to make sufficient contribution to the lab. If any of your teammate does not make sufficient contribution, please contact your instructor.
- DO NOT copy and paste or even leverage on the solutions we gave to the students in previous semesters. Violation will be reported to the Director of the MIDS program and the Office that oversees UC Berkeley Academic Integrity. In any case, the lab has various subtle changes that make those answers not directly applicable.

Data

In the United States, a $401K$ is a type of retirement savings plan that is tied to a worker's place of employment. Employees that put money into a $401K$ enjoy certain tax benefits. Moreover, many employers have a policy of promoting 401K use, by matching some percentage of an employee's contributions. If an employer matches at, say, 50%, for every dollar that an employee puts into a $401k$, the employer will put in another 50 cents.

The file **401k_w271.RData** contains data on 401k contributions that were filed with the IRS on form 5500. It was collected by Professor L. E. Papke and may have been further modified by the instructors to test your proficiency.

Exercises

Complete the following exercises, following the best practices outlined in class. Place your answers in a written report (pdf, word, or jupyter notebook format) along with relevant R statements and output.

Load the **401k_w271.RData** dataset and look at the value of the function *desc*() to see what variables are included.

1. Your dependent variable will be prate, representing the fraction of a company's employees participating in its 401k plan. Because this variable is bounded between 0 and 1, a linear model without any transformations may not be the most ideal way to analyze the data, but we can still learn a lot from it. Examine the prate variable and comment on the shape of its distribution.

2. Your independent variable will be mrate, the rate at which a company matches employee 401k contributions. Examine this variable and comment on the shape of its distribution.

3. Generate a scatterplot of prate against mrate. Then estimate the linear regression of prate on mrate. What slope coefficient did you get?

4. Is the assumption of zero-conditional mean realistic? Explain your evidence. What are the implications for your OLS coefficients?

5. Is the assumption of homoskedasticity realistic? Provide at least two pieces of evidence to support your conclusion. What are the implications for your OLS analysis?

6. Is the assumption of normal errors realistic? Provide at least two pieces of evidence to support your conclusion. What are the implications for your OLS analysis?

7. Based on the above considerations, what is the standard error of your slope coefficient?

8. Is the effect you find statistically significant, and is it practically significant?