W271-2 – Spring 2016 – **HW 2**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 10, 2016

Contents

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.

```
load("401k_w271.Rdata")
```

Question 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.

$12,345,678.900$

Question 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.

Question 3

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

Question 4

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

Question 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?

Question 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?

Question 7

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

Question 8

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