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.

```r
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.

```r
# Descriptive statistics of the whole dataset
desc
```

```
##   variable                           label
## 1    prate     participation rate, percent
## 2    mrate            401k plan match rate
## 3  totpart          total 401k participants
## 4   totelg    total eligible for 401k plan
## 5      age                 age of 401k plan
## 6   totemp  total number of firm employees
## 7     sole = 1 if 401k is firm's sole plan
## 8  ltotemp                   log of totemp
```

```r
str(data)
```

```
## 'data.frame':    1534 obs. of  8 variables:
##  $ prate  : num  26.1 100 97.6 100 82.5 ...
##  $ mrate  : num  0.21 1.42 0.91 0.42 0.53 ...
##  $ totpart: num  1653 262 166 257 591 ...
##  $ totelg : num  6322 262 170 257 716 ...
##  $ age    : int  8 6 10 7 28 7 31 13 21 10 ...
##  $ totemp : num  8709 315 275 500 933 ...
##  $ sole   : int  0 1 1 0 1 1 1 0 1 1 ...
##  $ ltotemp: num  9.07 5.75 5.62 6.21 6.84 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
##  - attr(*, "formats")= chr  "%7.0g" "%7.0g" "%7.0g" "%7.0g" ...
##  - attr(*, "types")= int  254 254 254 254 251 254 251 254
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "participation rate, percent" "401k plan match rate" "total 401k parti
##  - attr(*, "version")= int 10
```

```
summary(data)
```

```
##      prate            mrate           totpart           totelg
##  Min.   :  3.00   Min.   :0.0100   Min.   :   50.0   Min.   :   51.0
##  1st Qu.: 78.10   1st Qu.:0.3000   1st Qu.:  156.2   1st Qu.:  176.0
##  Median : 95.70   Median :0.4600   Median :  276.0   Median :  330.0
##  Mean   : 87.56   Mean   :0.7315   Mean   : 1354.2   Mean   : 1628.5
##  3rd Qu.:100.00   3rd Qu.:0.8300   3rd Qu.:  749.5   3rd Qu.:  890.5
##  Max.   :200.00   Max.   :4.9100   Max.   :58811.0   Max.   :70429.0
##       age            totemp            sole           ltotemp
##  Min.   : 4.00   Min.   :    58   Min.   :0.0000   Min.   : 4.060
##  1st Qu.: 7.00   1st Qu.:   261   1st Qu.:0.0000   1st Qu.: 5.565
##  Median : 9.00   Median :   588   Median :0.0000   Median : 6.377
##  Mean   :13.18   Mean   :  3568   Mean   :0.4876   Mean   : 6.686
##  3rd Qu.:18.00   3rd Qu.:  1804   3rd Qu.:1.0000   3rd Qu.: 7.498
##  Max.   :51.00   Max.   :144387   Max.   :1.0000   Max.   :11.880
```

```
# Descriptive statistics of prate
summary(data$prate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   78.10   95.70   87.56  100.00  200.00
```

```
round(stat.desc(data$prate, desc = TRUE, basic = TRUE, norm = TRUE), 2)
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##      1534.00         0.00         0.00         3.00       200.00
##        range          sum       median         mean      SE.mean
##       197.00    134314.70        95.70        87.56         0.44
## CI.mean.0.95          var      std.dev     coef.var     skewness
##         0.87       300.95        17.35         0.20        -0.95
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##        -7.56         4.36        17.44         0.78         0.00
```

```
round(quantile(data$prate, probs = c(1, 5, 10, 25, 50, 75, 90, 95, 99,
                                      100)/100), 1)
```

```
##    1%    5%   10%   25%   50%   75%   90%   95%   99%  100%
##  31.8  54.0  62.8  78.1  95.7 100.0 100.0 100.0 100.0 200.0
```

```
data$prate[data$prate > 100]
```

```
## [1] 200.0 177.2 200.0
```

See Figure 3

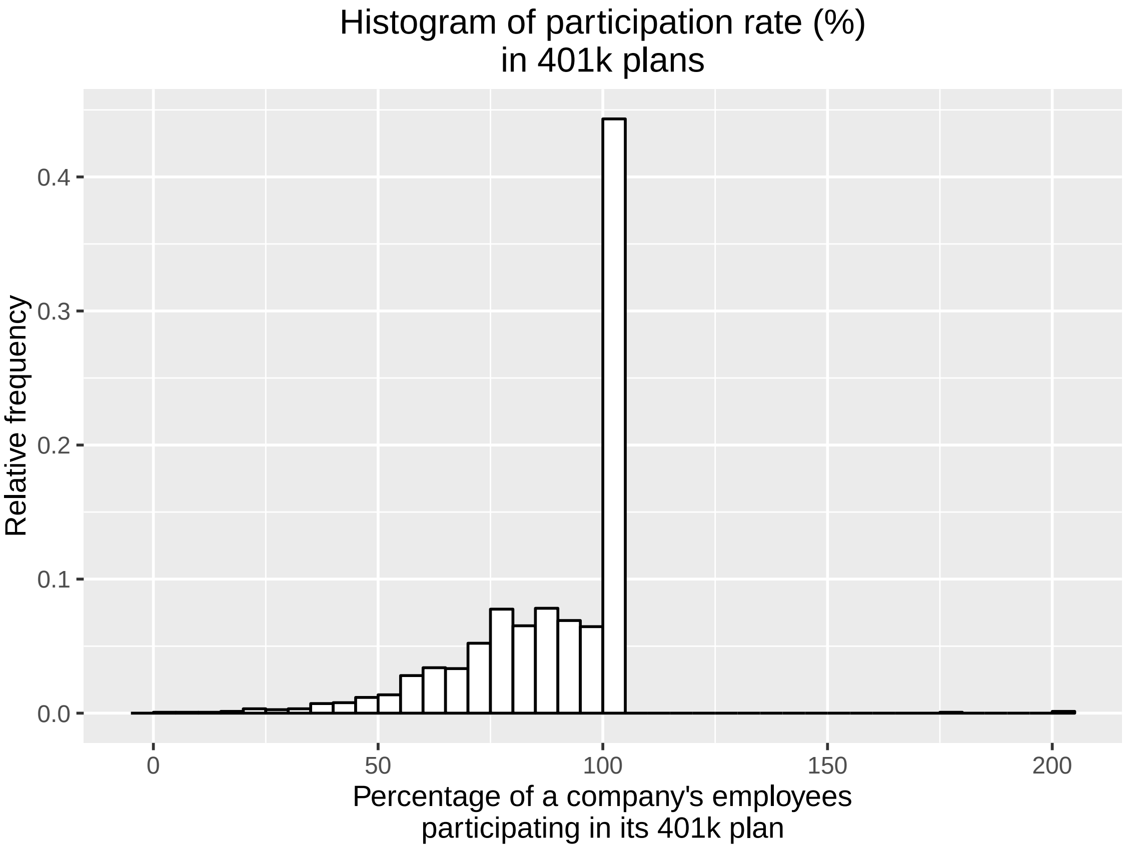

Figure 1: Histogram of participation rate (%) in 401k plans (bin width = 5)

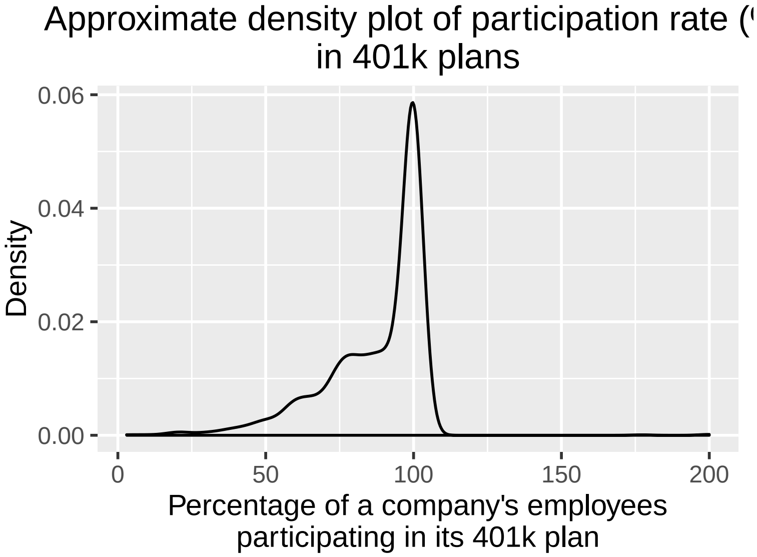

Figure 2: Approximate density plot of participation rate (%) in 401k plans

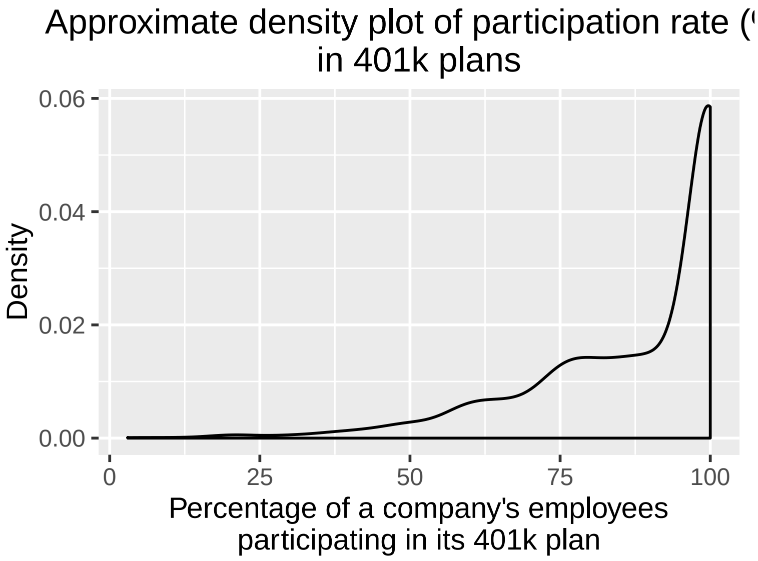

Figure 3: Approximate density plot of participation rate (%) in 401k plans (excluding wrong values, higher than 100%)

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?