Homework 2

Rohan Thakur, Charles Kekeh and Megan Jasek

February 7, 2016

```
# Load the dataframe
load("401k_w271.RData")
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
```

Question 1

```
summary(data$prate)
```

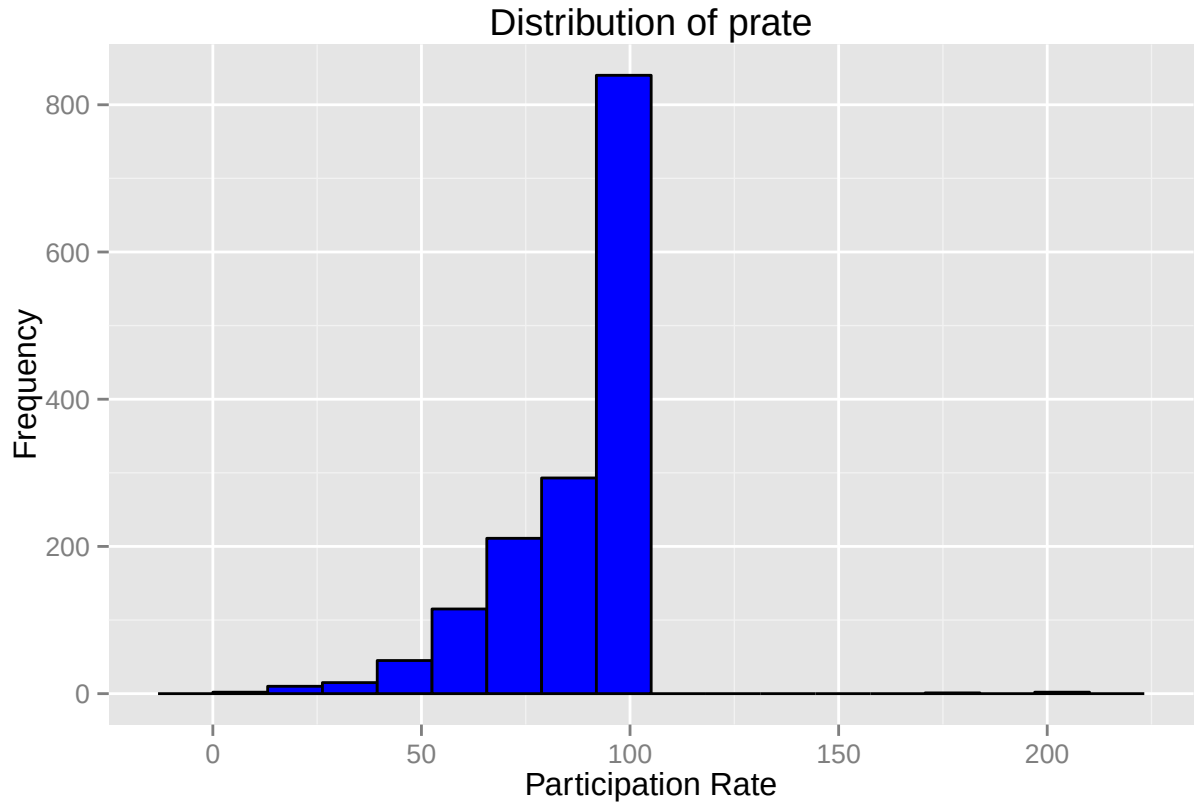
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   78.10   95.70   87.56  100.00  200.00
```

```
print(quantile(data$prate, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%     99%
##  31.763  53.995  62.760  78.100  95.700 100.000 100.000 100.000 100.000
##    100%
## 200.000
```

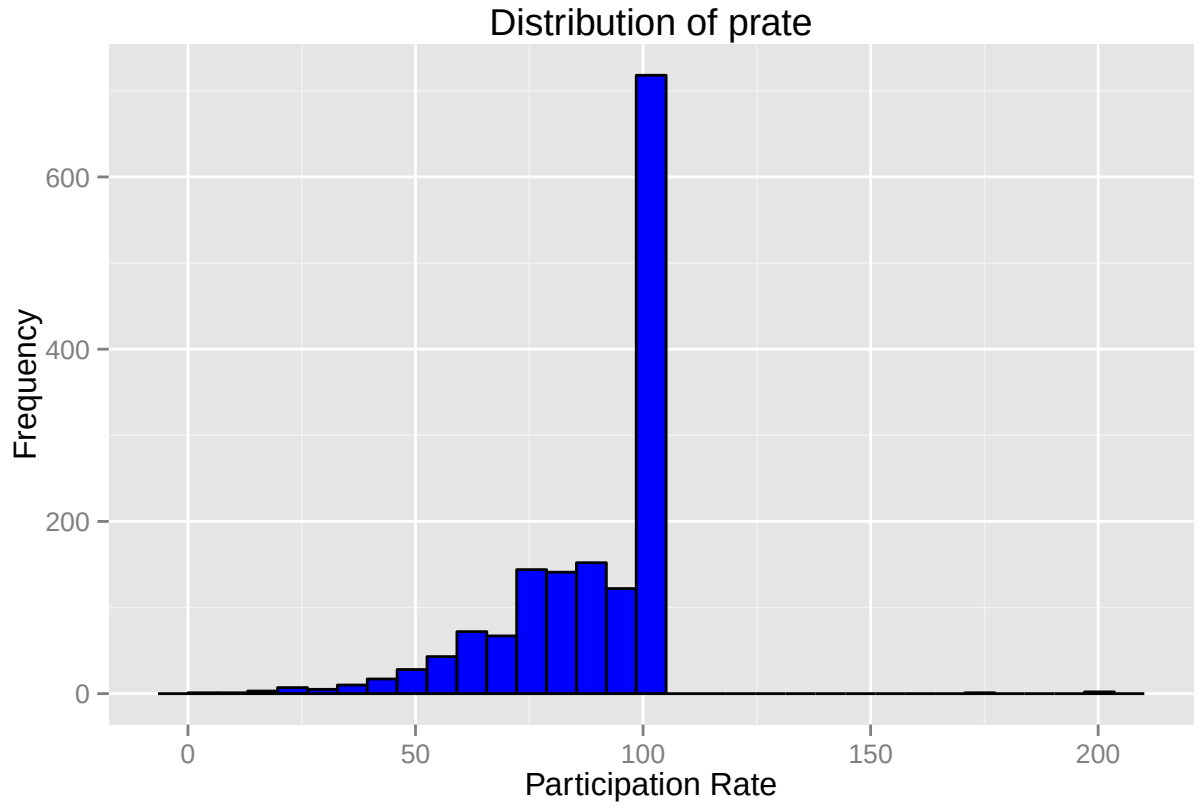
```
# Plot the histogram of prate at 15 bins
prate.hist <- ggplot(data, aes(prate)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black",
        binwidth = (range(data$prate)[2] - range(data$prate)[1])/15) +
    labs(title = "Distribution of prate", x = "Participation Rate",
        y = "Frequency")

plot(prate.hist)
```



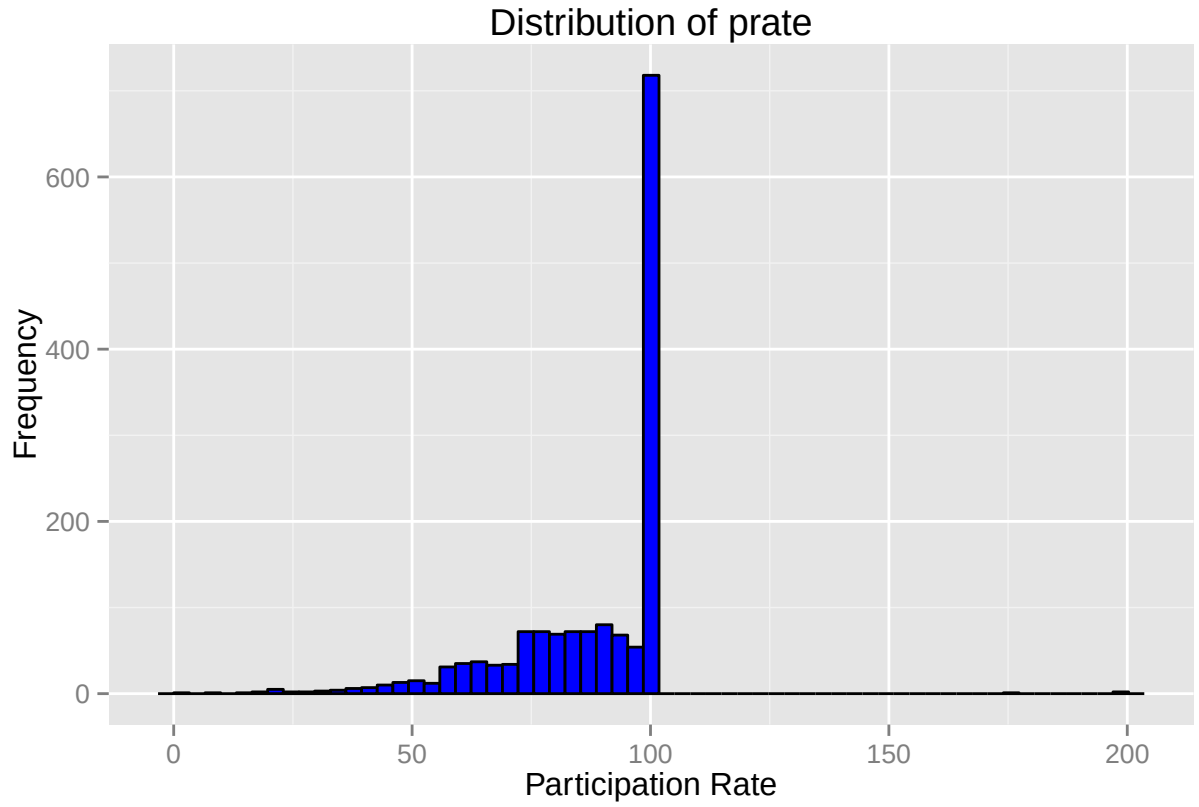

```
# Plot the histogram of prate at 30 bins
prate.hist <- ggplot(data, aes(prate)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black",
        binwidth = (range(data$prate)[2] - range(data$prate)[1])/30) +
    labs(title = "Distribution of prate", x = "Participation Rate",
        y = "Frequency")

plot(prate.hist)
```




```
# Plot the histogram of prate at 60 bins
prate.hist <- ggplot(data, aes(prate)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black",
        binwidth = (range(data$prate)[2] - range(data$prate)[1])/60) +
    labs(title = "Distribution of prate", x = "Participation Rate",
        y = "Frequency")

plot(prate.hist)
```



The variable has higher frequency at higher values of participation rates with a particularly large spike at 100% participation, indicating that most companies have all employees participating in the 401k. There also seem to be erroneous values greater than 100% which we will code as NA.

```
# Creating a clean version of the variable
data$prate.clean = data$prate
data$prate.clean[data$prate.clean > 100] = NA
summary(data$prate.clean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   78.05   95.70   87.35  100.00  100.00       3
```

Question 2

```
summary(data$mrate)
```

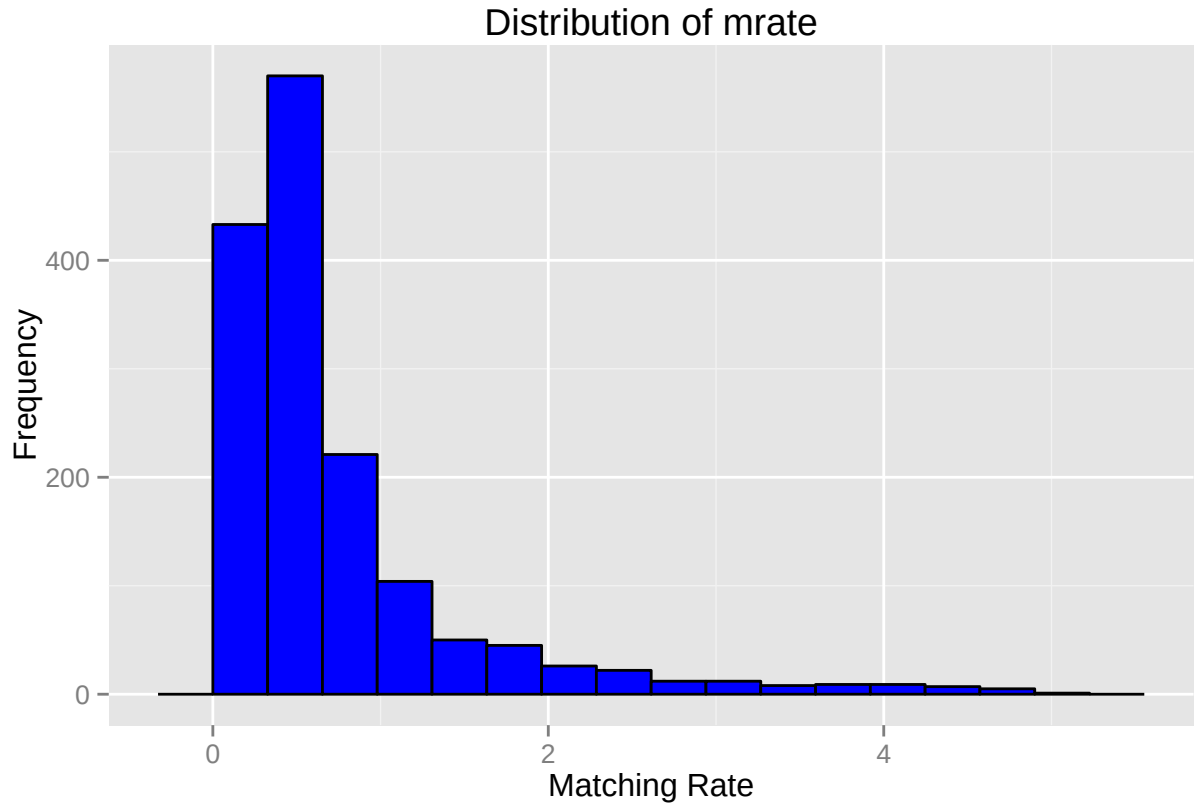
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.3000  0.4600  0.7315  0.8300  4.9100
```

```
print(quantile(data$mrate, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%     99%    100%
## 0.0300 0.1100 0.1600 0.3000 0.4600 0.8300 1.6570 2.3635 4.1267 4.9100
```

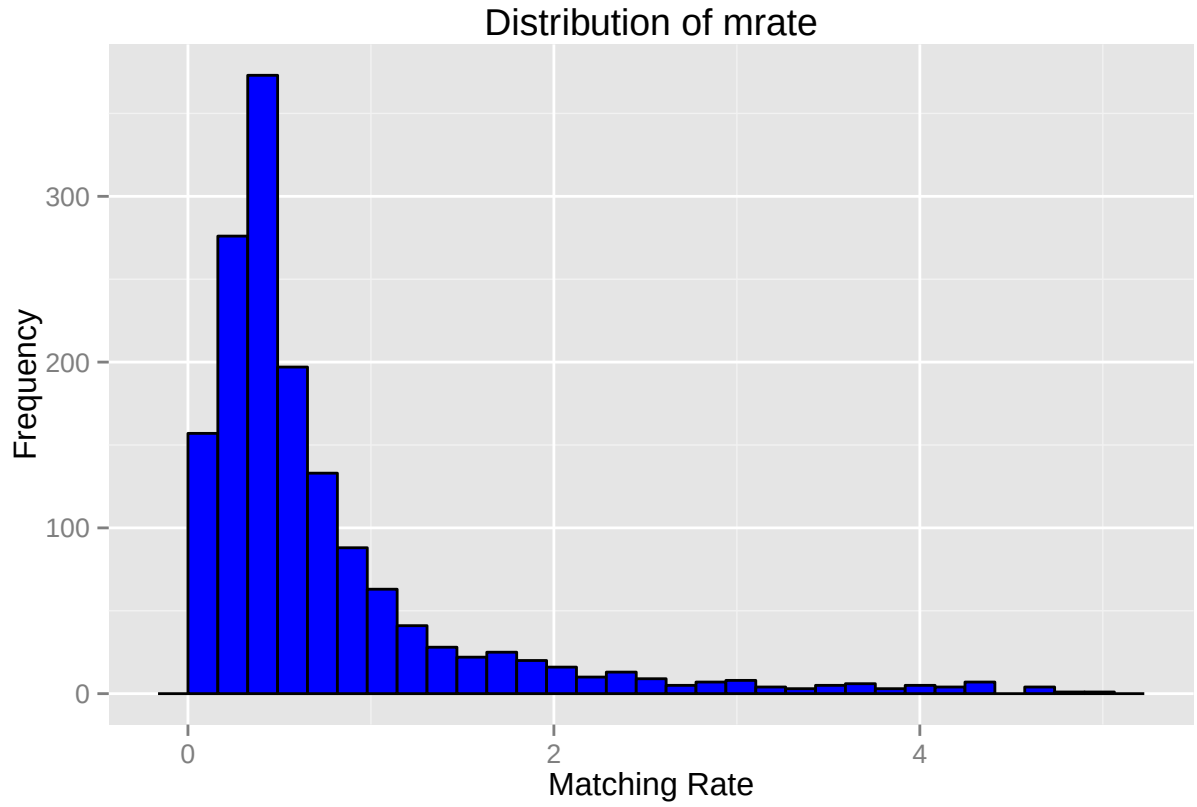
```r
# Plot the histogram of mrate at 15 bins
mrate.hist <- ggplot(data, aes(mrate)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black",
        binwidth = (range(data$mrate)[2] - range(data$mrate)[1])/15) +
    labs(title = "Distribution of mrate", x = "Matching Rate",
        y = "Frequency")

plot(mrate.hist)
```



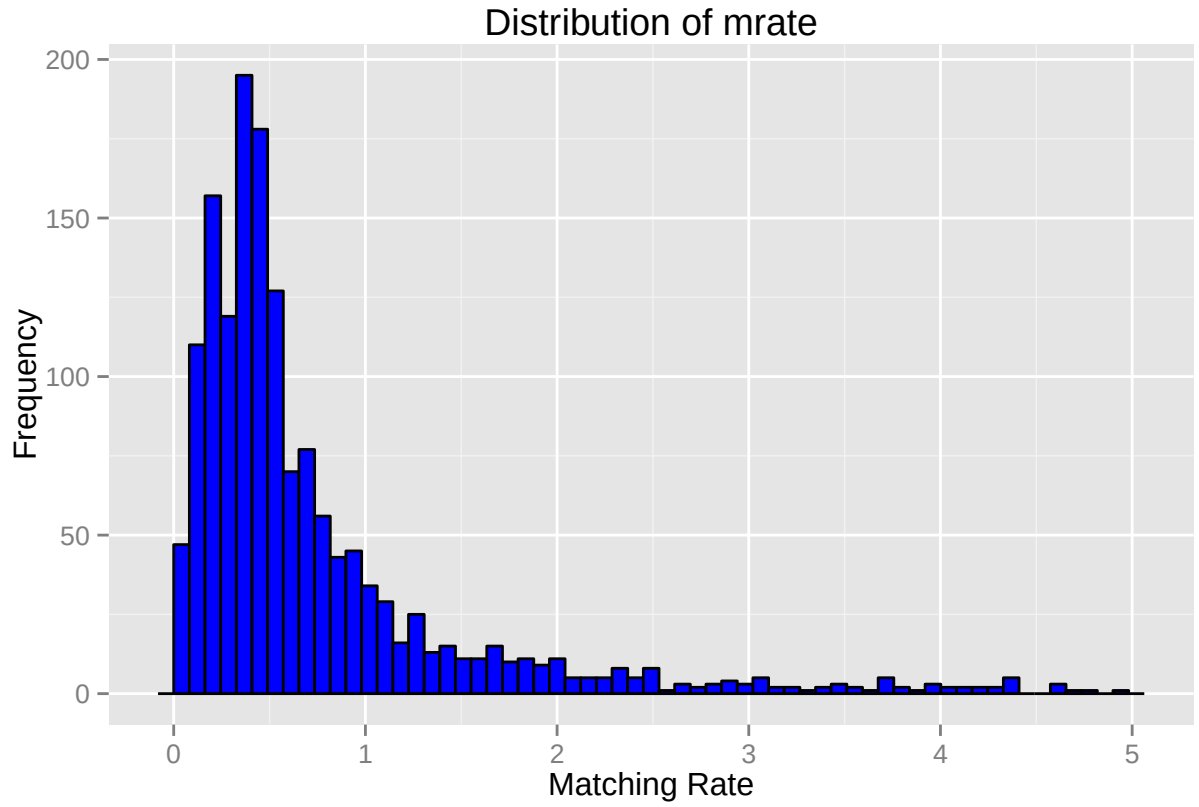

```r
# Plot the histogram of mrate at 30 bins
mrate.hist <- ggplot(data, aes(mrate)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black",
        binwidth = (range(data$mrate)[2] - range(data$mrate)[1])/30) +
    labs(title = "Distribution of mrate", x = "Matching Rate",
        y = "Frequency")

plot(mrate.hist)
```




```
# Plot the histogram of mrate at 60 bins
mrate.hist <- ggplot(data, aes(mrate)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black",
        binwidth = (range(data$mrate)[2] - range(data$mrate)[1])/60) +
    labs(title = "Distribution of mrate", x = "Matching Rate",
        y = "Frequency")

plot(mrate.hist)
```



mrate is heavily positively skewed, with most companies matching between 30% and 83%. Though the variable has a mean of 73%, the median here - 46% - is a better measure of central tendency due to some large outliers.

We will do an arithmetic transformation and multiply the values of mrate by 100 in order to keep it consistent in format with the prate variable.

```
# Creating a transformed version of the variable
data$mrate.clean = data$mrate * 100
summary(data$mrate.clean)
```

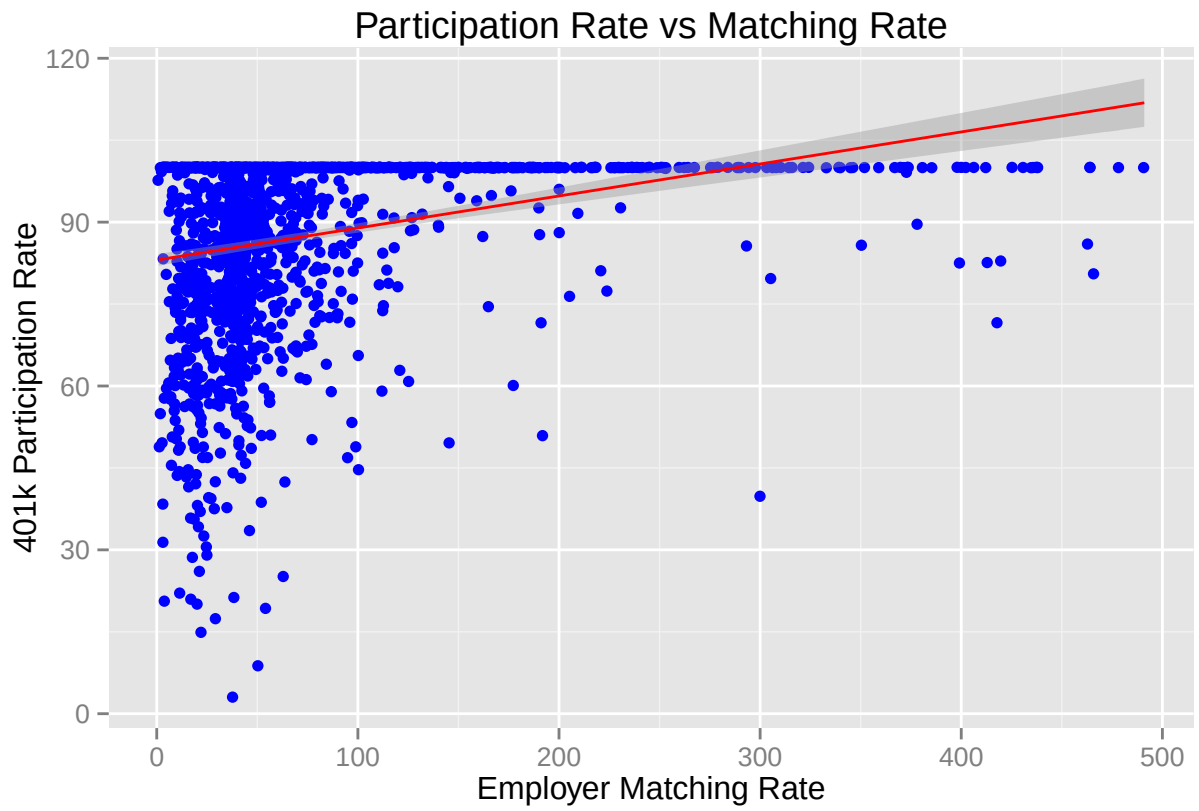
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   30.00   46.00   73.15   83.00  491.00
```

Question 3

```
# Create a scatterplot of prate vs mrate
scatter.prate.mrate <- ggplot(data, aes(mrate.clean,
    prate.clean)) + geom_point(colour = "Blue", position = "jitter") +
    geom_smooth(method = "lm", colour = "Red") + labs(y = "401k Participation Rate",
    x = "Employer Matching Rate", title = "Participation Rate vs Matching Rate")
plot(scatter.prate.mrate)
```

```
## Warning: Removed 3 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```




```r
# Running linear regression
model = lm(prate.clean ~ mrate.clean, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = prate.clean ~ mrate.clean, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.289  -8.200   5.186  12.723  16.821
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.061788   0.564121  147.24   <2e-16 ***
## mrate.clean  0.058623   0.005275   11.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.09 on 1529 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.07475,    Adjusted R-squared:  0.07414
## F-statistic: 123.5 on 1 and 1529 DF,  p-value: < 2.2e-16
```

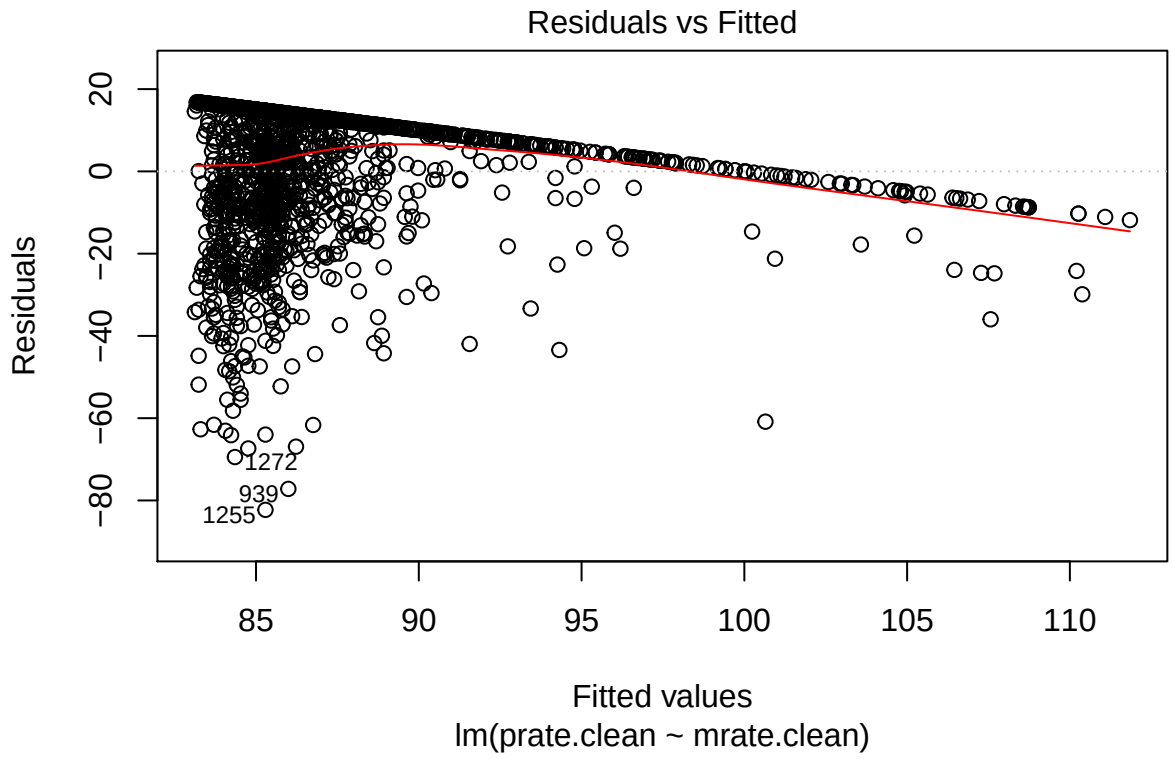
```r
# Print the coefficient
print(model$coefficients[2])
```

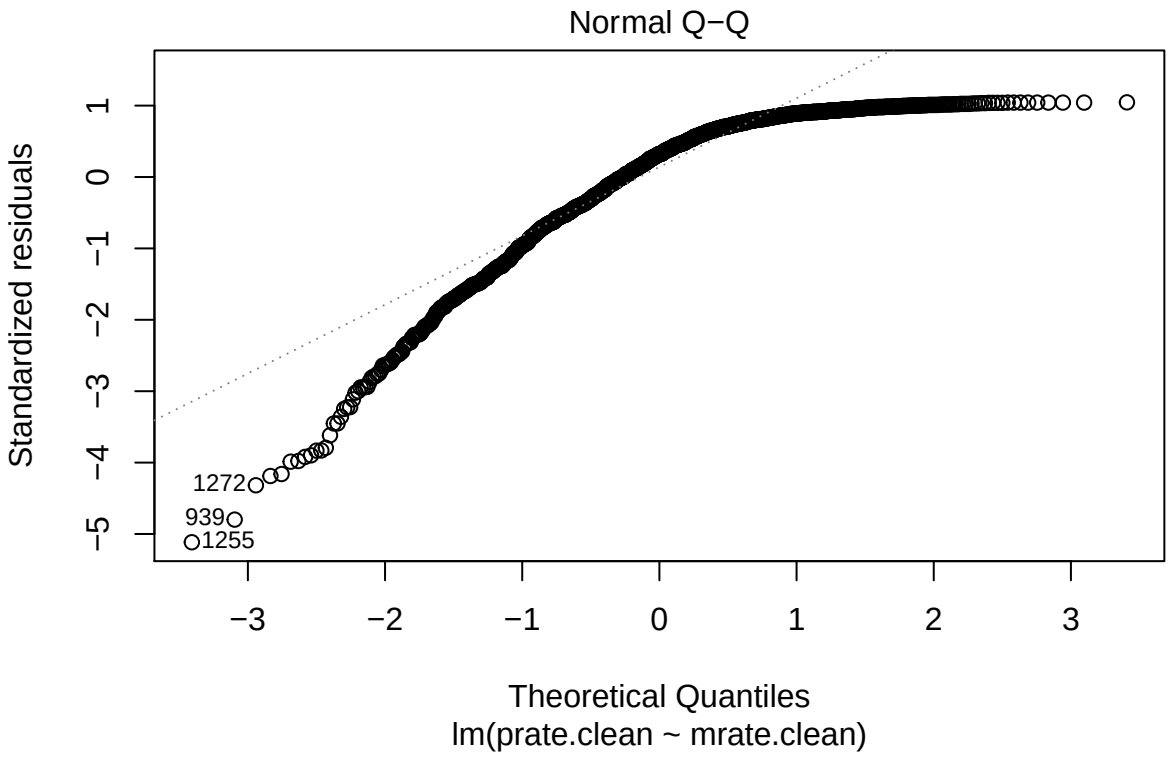
```
## mrate.clean
##  0.05862268
```

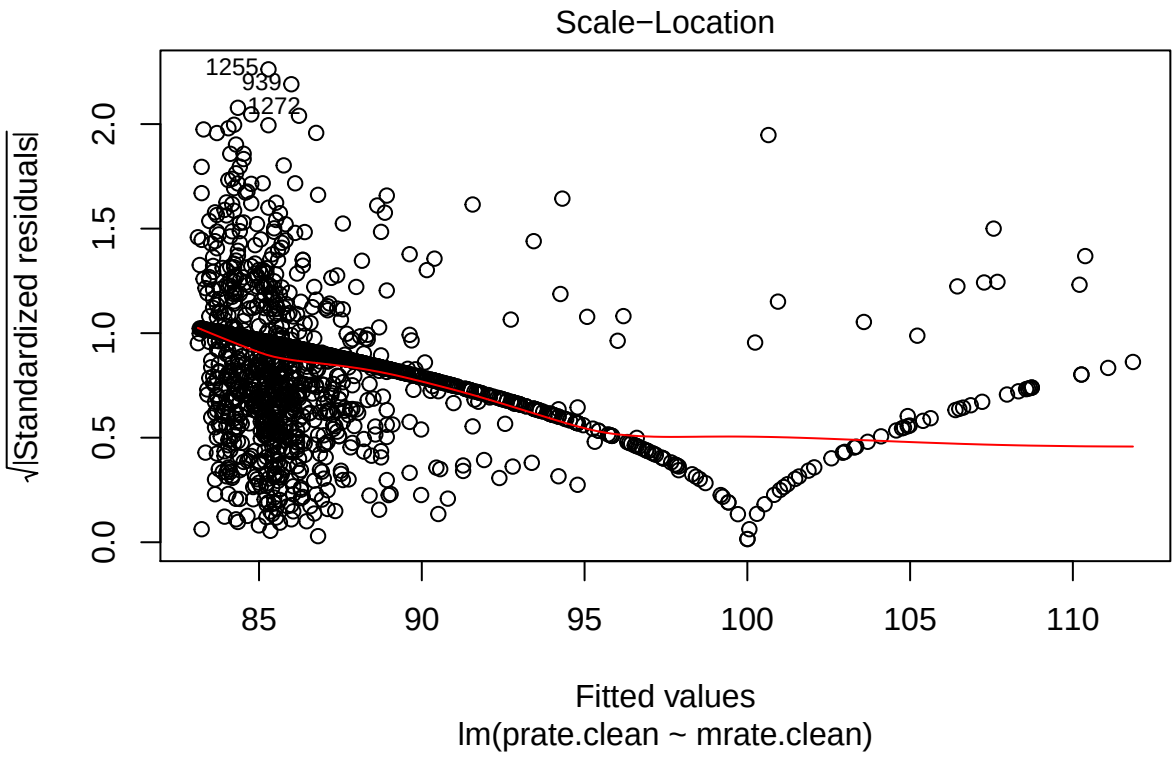
We get a slope coefficient of 0.06. Based on the t statistic, the coefficient is significant at the 0.05 significance level.

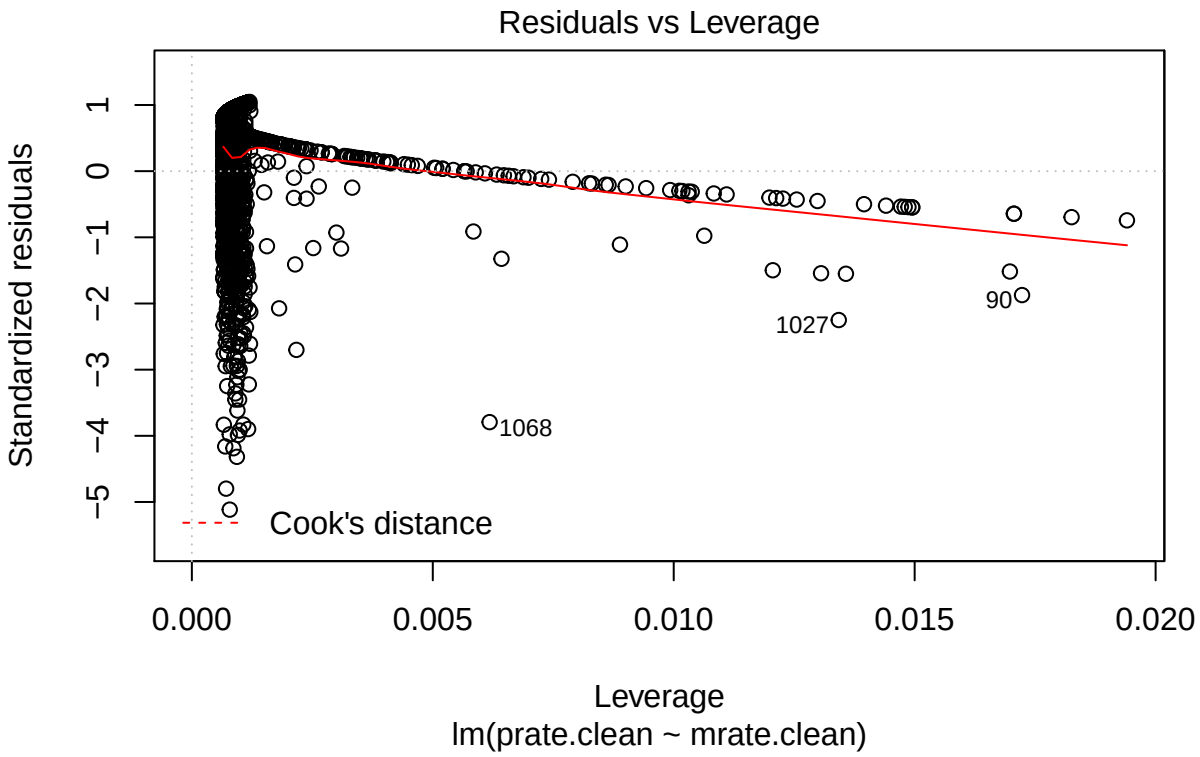
Question 4

```r
# Printing the diagnostic plots
plot(model)
```




Normal Q−Q
Standardized residuals
1
0
−1
−2
−3
−4
−5
1272
939
1255
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(prate.clean ~ mrate.clean)


Scale−Location
1255
939
1272
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
85
90
95
100
105
110
Fitted values
lm(prate.clean ~ mrate.clean)



From residuals vs fitted values graph, we can see that there is a violation of zero-conditional mean, as we see the smoothing curve slope go up and then follow a downward trend. Nonetheless, since we have a large sample size, we do not need this assumption since we can use the assumption of exogeneity. Therefore, this has very little impact on our regression model.

Question 5

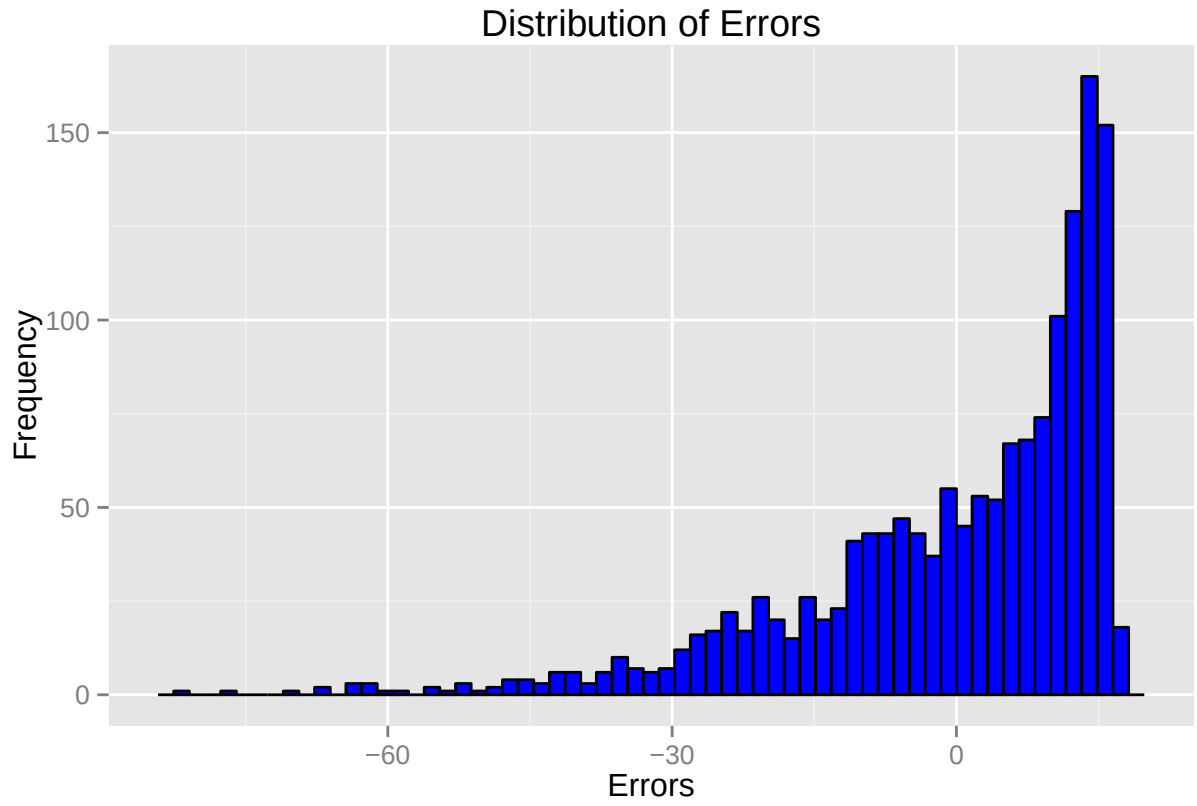
There seems to be a violation of homoskedasticity:
1 - We can see from the residuals vs fitted plot that the variance narrows as we move to higher fitted values.
2 - The same story is told by the scale-location plot where we see that the graph is nowhere close to a horizontal band, which is what we would get if homoskedasticity was met.

We do not look at the Breusch Pagan test since we have a large number of observations, therefore we know almost certainly that we will obtain significance. The implication of heteroskedasticity in the data is that it might cause the standard error of our b_j coefficients to become biased. This may lead to our estimates of those coefficients not being BLUE, or a false negative in the hypothesis test. To correct for this, we will have to use robust standard errors.

Question 6

```
# Plotting Histogram of errors
errors.hist <- ggplot(model, aes(model$residuals)) +
```

```
    theme(legend.position = "none") + geom_histogram(fill = "Blue",
    colour = "Black", binwidth = (range(model$residuals)[2] -
        range(model$residuals)[1])/60) + labs(title = "Distribution of Errors",
    x = "Errors", y = "Frequency")
plot(errors.hist)
```



From inspection of plots, we can see a violation of normality:
1 - When we plot the histogram of errors, we see the negative skew.
2 - The negative skew is also apparent in the Q-Q plot of the standardized residuals.
We do not conduct the Shapiro Wilk test because knowing that we have a very large sample size, we know almost certainly that we will obtain significance.

In terms of implications, despite non-normality from the plots, we can use OLS asymptotics. Since there is a version of the central limit theorem that tells us that the sampling distribution of coefficient estimates approaches normality with large sample sizes, we do not need the normality assumption of our error. Therefore, the finding that our errors do not follow a normal distribution does not have much of an impact on our regression.

Question 7

```
# Based on the violation of homoskedasticity, we
# must run robust standard errors.
coeftest(model, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 83.0617879  0.6130975 135.479 < 2.2e-16 ***
## mrate.clean  0.0586227  0.0047015  12.469 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After running the linear model with robust standard errors, we get 0.0047 as the standard error for the mrate coefficient. This is lower than the 0.0053 without the robust standard errors.

Question 8

```
waldtest(model, vcov = vcovHC)
```

```
## Wald test
##
## Model 1: prate.clean ~ mrate.clean
## Model 2: prate.clean ~ 1
##   Res.Df Df      F    Pr(>F)
## 1   1529
## 2   1530 -1 155.47 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is highly statistically significant. It shows small practical significance, implying that a change of 1% in the matching rate by an employer corresponds to a change of 0.06% in participation in 401k plans by employees. Therefore, it would take a 17% increase in matching rate for a corresponding 1% increase in participation. That interpretation is also somewhat supported by the R^2 statistic of our original linear regression of prate on mrate. In that regression, the R^2 statistic has a value of 0.075, impyining that less than 8% of the variance of the participation rate (prate) is explained by the model.