W271 - Homework 2: OLS Inference

Lei Yang, Ron Cordell, Subhashini Raghunathan

Feb 11, 2016

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

```
# load packages
library(car)
library(lmtest)
library(sandwich)
# set work dir, clear workspace, load data, show description
setwd("~/Desktop/W271Data")
rm(list=ls())
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
```

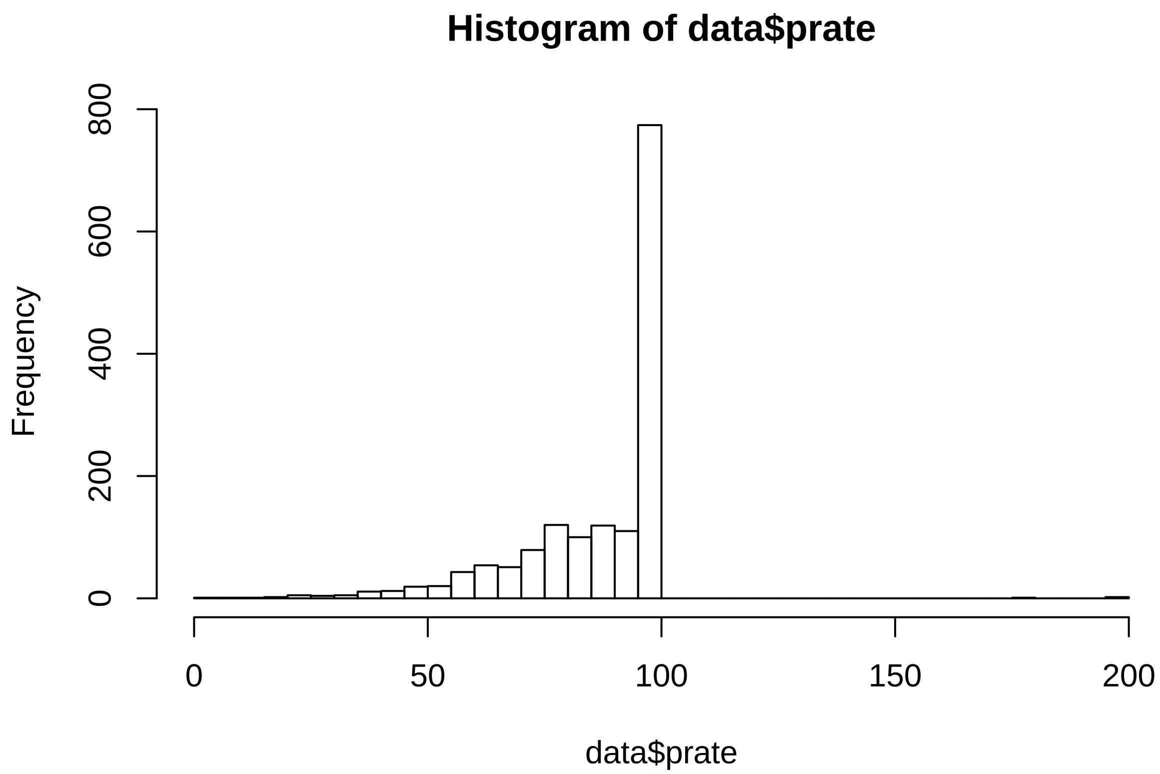
1. Your dependent variable will be *prate*, representing the fraction of a company's employees participating in its 401k plan. Because this variable is bounded between 0 and 1, a linear model without any transformations may not be the most ideal way to analyze the data, but we can still learn a lot from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

We first show the summary of *prate*, then plot a histogram:

```
# show prate summary
summary(data$prate)
```

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

```
# histogram
hist(data$prate, breaks = 30)
```



from the summary, we can see there are **3** records with invalid *prate* values (> 100):

```
# show records with invalid prate value
data[data$prate>100, ]
```

```
##      prate mrate totpart totelg age totemp sole  ltotemp
## 106  200.0  1.07     801    801   7   8546    0 9.053219
## 933  177.2  0.17     404    523   6    645    0 6.469250
## 1263 200.0  0.19     514    514   9    756    0 6.628041
```

```
# update data dataframe
data <- data[data$prate<=100, ]
```

and the distribution of *prate* is not Normal, with big negative skew, because most employees are participating in most companies. Finally, We will get rid of the 3 records with $> 100\%$ values.

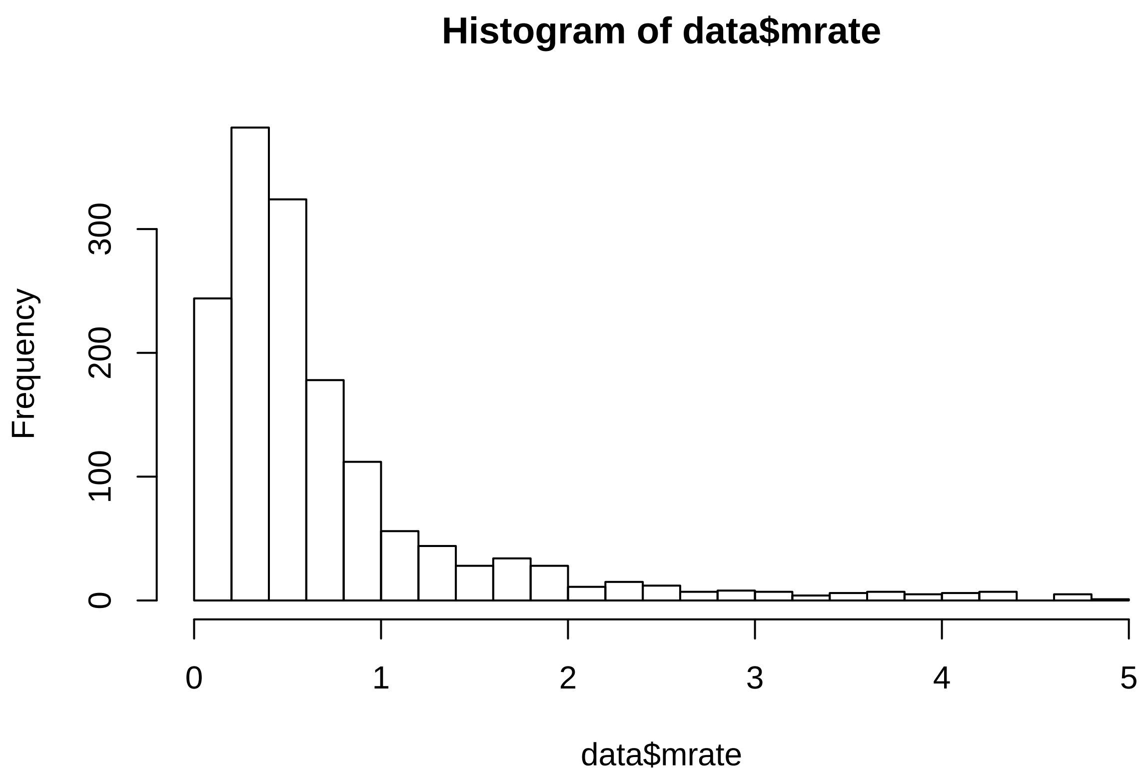
2. Your independent variable will be *mrate*, the rate at which a company matches employee 401k contributions. Examine this variable and comment on the shape of its distribution.

From the histogram of *mrate*:

```
# show summary and histogram of mrate variable
summary(data$mrate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.010   0.300   0.460   0.732   0.830   4.910
```

```r
hist(data$mrate, breaks=20)
```



the distribution of *mrate* is not Normal either, with big positive skew. In addition, all records have valid percentage value.

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

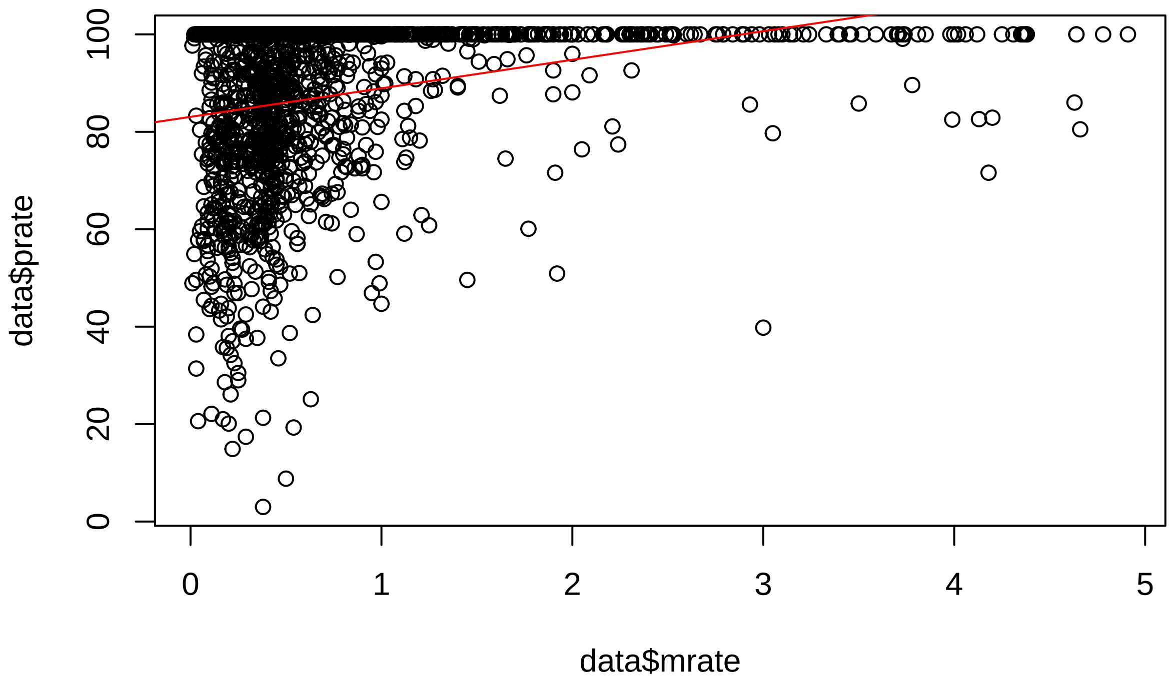
We first generate the scatterplot, then estimate the linear regression for two highly skewed variables:

```r
# generate scatterplot of prate against mrate
plot(data$mrate, data$prate)
# build a linear regression model of prate on mrate
m3 <- lm(prate ~ mrate, data = data)
# evaluate model coefficients
summary(m3)
```

```
##
## Call:
## lm(formula = prate ~ mrate, 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)  83.0618     0.5641  147.24   <2e-16 ***
## mrate         5.8623     0.5275   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
## 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
# overlay our linear model on the scatterplot
abline(m3, col = "red")
```

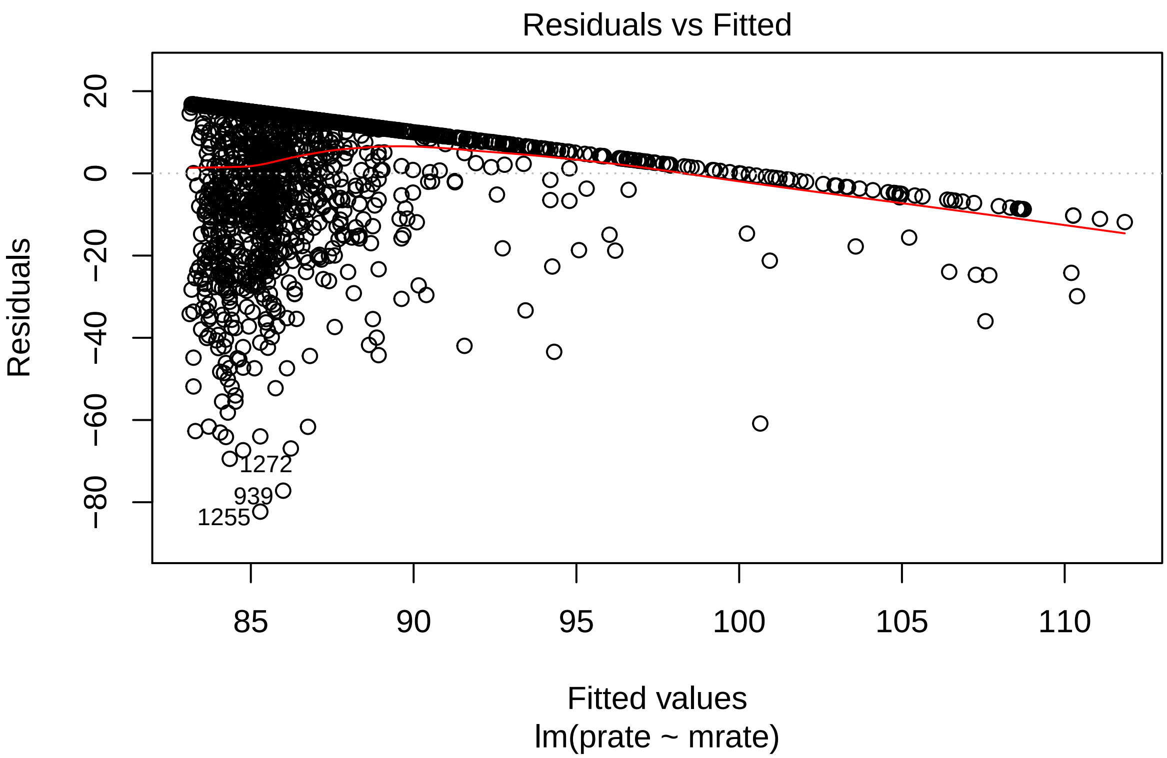


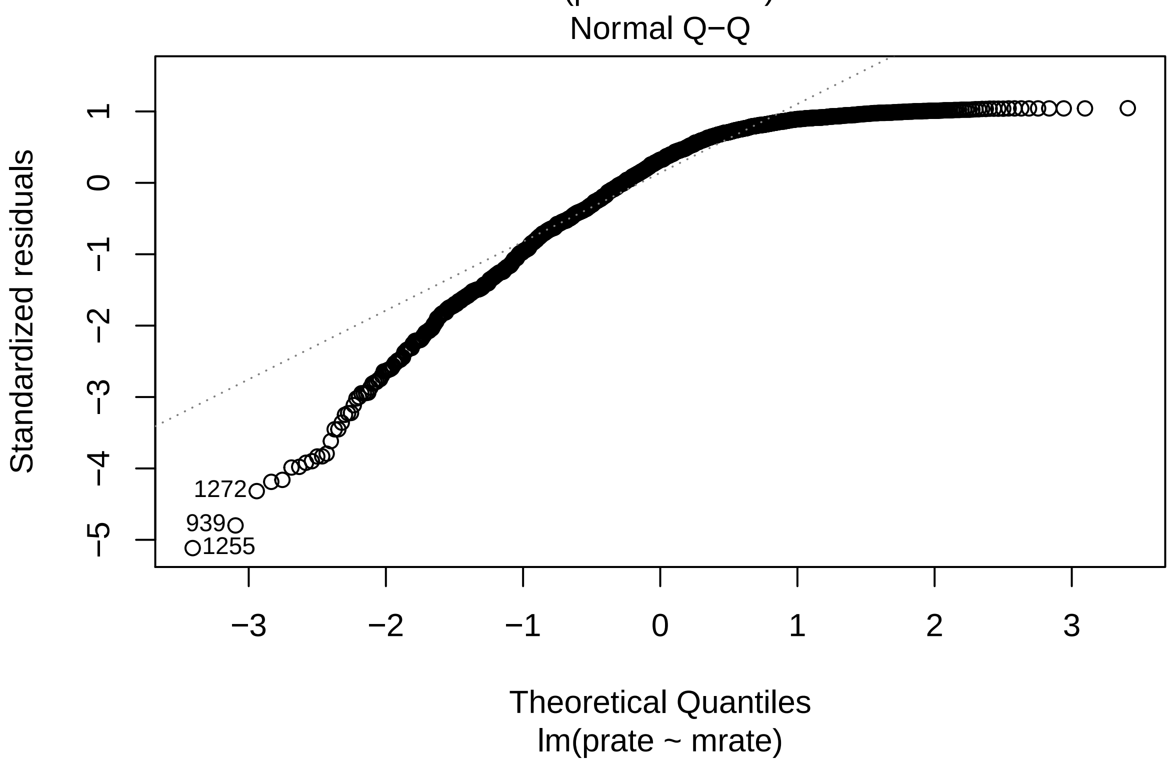
the slope coefficient of *mrate* is `5.8623`.

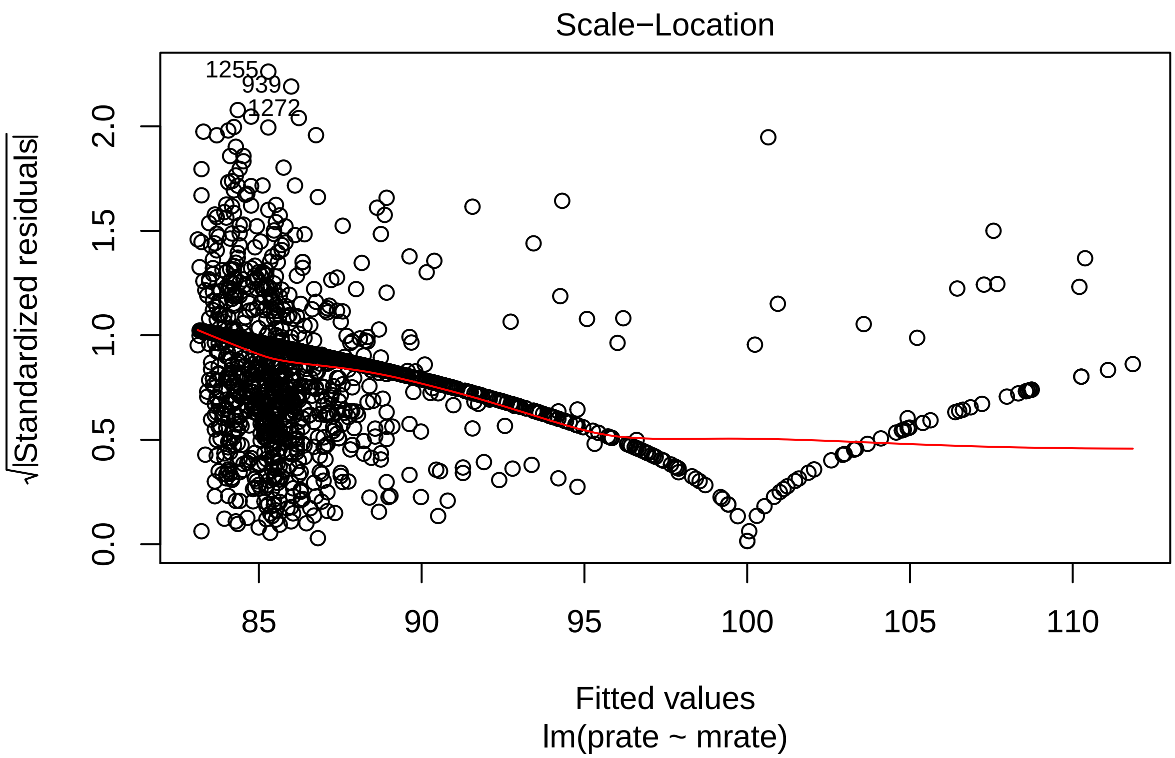
4. Is the assumption of zero-conditional mean realistic? Explain your evidence. What are the implications for your OLS coefficients?

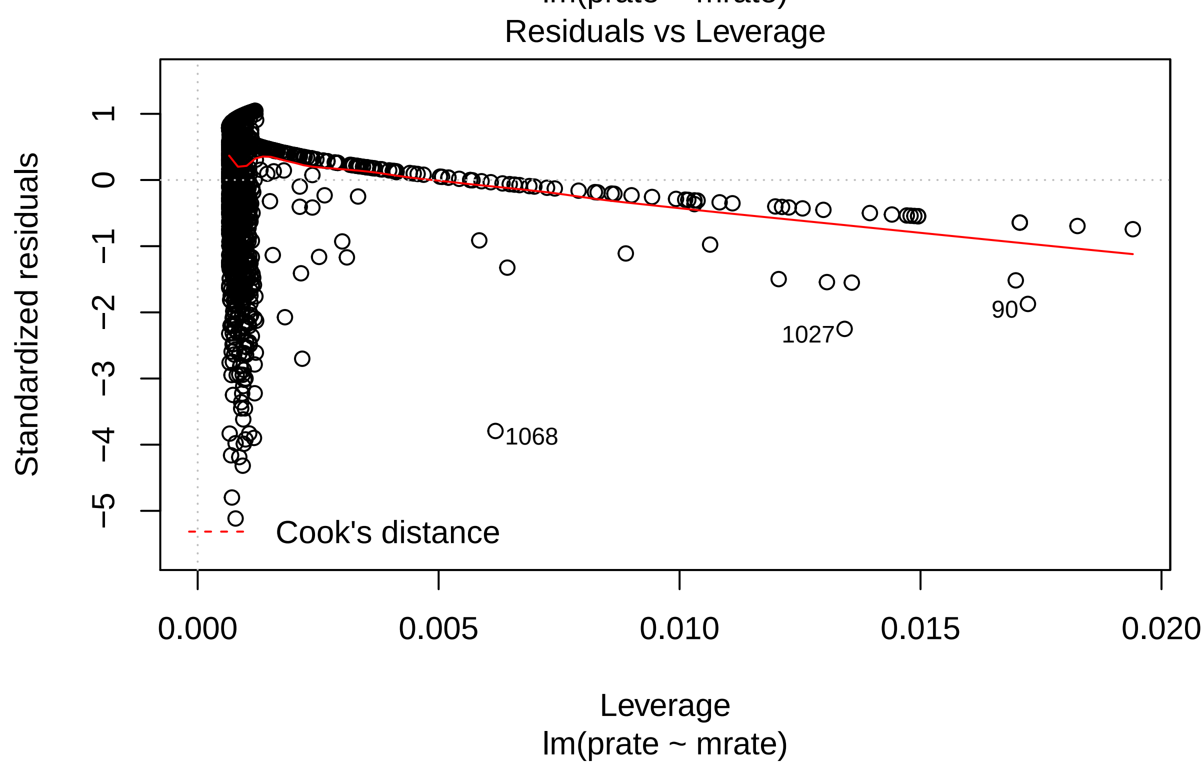
To evaluate the "zero-conditional mean" assumption $MLR4'$, we want to check the *Residuals vs. Fitted Value* plot. And from below we can see, the residuals first increases above zero, then decreases to below zero, as the prediction increases. Also note that those negative errors are mostly associated with the fitted values that are larger than 100%, which is not really valid.

```r
# model diagnostic
plot(m3)
```


Residuals vs Fitted
Residuals
20
0
−20
−40
−60
−80
1272
939
1255
85
90
95
100
105
110
Fitted values
lm(prate ~ mrate)


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 ~ mrate)





The intercept `83.0618` indicates that even without any corporate matching, on average their still will be `83.0618%` employees participate 401k. And the slope `5.8623` implicates that with every 1% increase of corporate matching, the participation rate will go up `5.8623%`.

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?

From both the *Residuals vs Fitted* and *mrate vs prate* plots, we can see the error variance is bigger toward left and reduces toward right, which can be attributed to less data in the region of large matching rate. Thus we seem to have violation in homoskedasticity. In addtion, we can perform Breusch-Pagan test to check the null hypothesis for homoskedasticity:

```
bp <- bptest(m3)
bp
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m3
## BP = 27.921, df = 1, p-value = 1.264e-07
```

The p-value `0.0000` indicates we can reject the null, and in favor of heteraskedasticity. In order to accommodate the effect, we use robust standard errors instead:

```
# robust standard error
re = coeftest(m3, vcov = vcovHC)
re
```

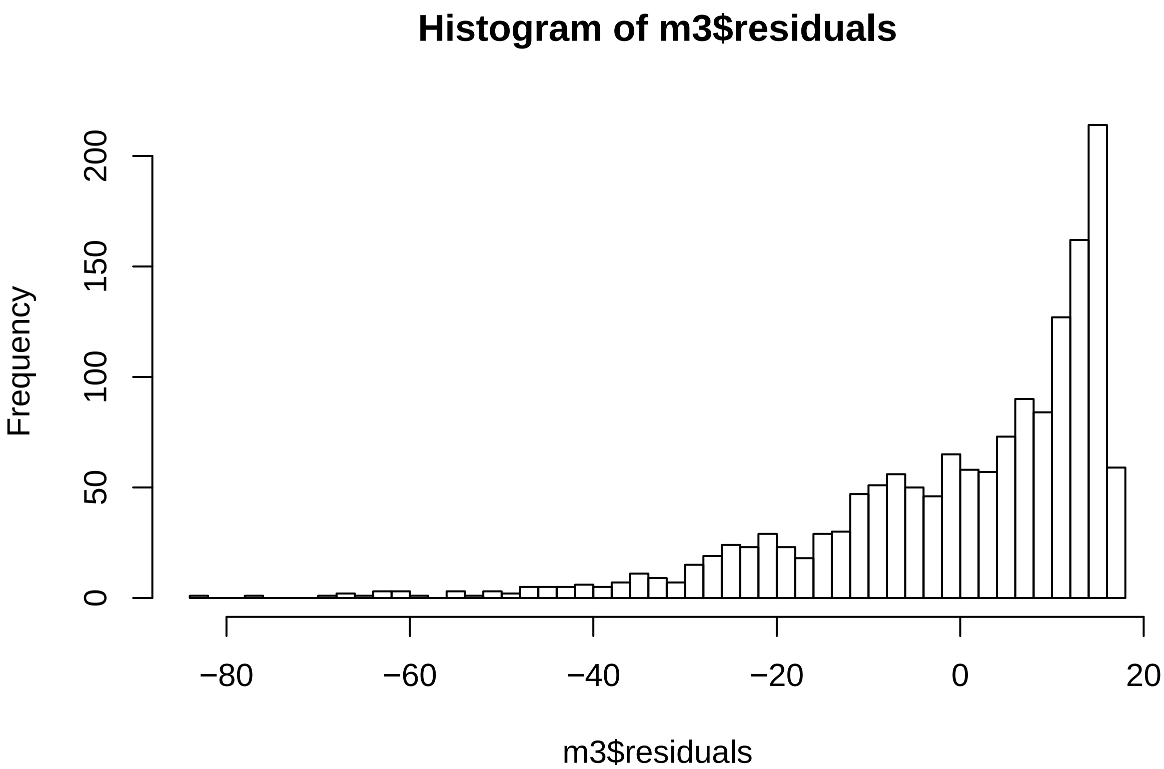
```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 83.06179    0.61310 135.479 < 2.2e-16 ***
## mrate        5.86227    0.47015  12.469 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

notice that the estimate is identical with standard error assumption, but the standard error of the intercept is bigger, to address homoskedasticity.

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?

From the below histogram of residuals, we can see it has negative skew, and is not normal.

```
hist(m3$residuals, breaks = 50)
```



In addition, from the above *QQ plot for standardized residuals*, we can observe the negative skew as well. But because we have a large sample size `1531`, we can get normality of our sampling distributions.

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

From the robust error calculation above, the standard error of slope coefficent is `0.4702`. Noticed that this standard error is even smaller than the one obtained above (0.5275) without robust error.

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

To test overall model significance, we use the wald test, which generalizes the usual F-test of overall significance, but allows for a heteroskedasticity-robust covariance matrix.

```r
# run Wald test
wt <- waldtest(m3, vcov = vcovHC)
wt
```

```
## Wald test
##
## Model 1: prate ~ mrate
## Model 2: prate ~ 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
```

With a $p-value$ of `0.0000` from the Wald test, the model is overall statistically significant. However, the correlation coefficient between *prate* and *mrate* is only `0.2734`, which is < 0.3 and only implies small effect. In addition, given that the standard deviation of *prate* is `16.7247`, the slope of *mrate* (`5.8623`) is only `35.05%` of one standard deviation. Thus the treatment from corporate matching rate on the participation rate is practically insignificant.