NYC Housing: Predicting Income

Oskar Eisgruber
oeisgrub

Due Wed, March 4, at 8:00PM

Contents


```
library("knitr")
library("cmu202")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
```

```
########################
### Loading the data
#######################
# First, download the .rda data file from canvas.
# Then, upload it to shimmer as you would for a lab or HW.
# Finally, to load the data, use the reader::read_rds() function on the .rda file in quotes, and assign
# for example:
nyc <- readr::read_rds("nyc.rda")
```

```
nyc <- readr::read_rds("nyc.rda")
```

Introduction

If you can make it in New York City, you can make it anywhere! New York City is refereed to by some as the world's capital. People of every nationality and ethnicity crowd into the city's awesome concrete structures in order to find work in the epicenter of the land of opportunity. As a result of the paramount levels of demand for a slice of real estate in New York City, the average rent is one of the highest in the world. Furthermore, New York City is an old city with its history dating back to before America was even a country. As a result, some of the buildings are old, and in need of repair. However, the prices are still high, even in the buildings that are in need of repairs, which results in people with high incomes living in unassuming apartments. The goal of this project is to analyze the relationship between income and some other variables that relate to housing in New York City. Your research groupis approached by a consumer advo-cate watchdog group that is trying to determine the relationship between the household income and several demographic and housing quality measurements. They believe

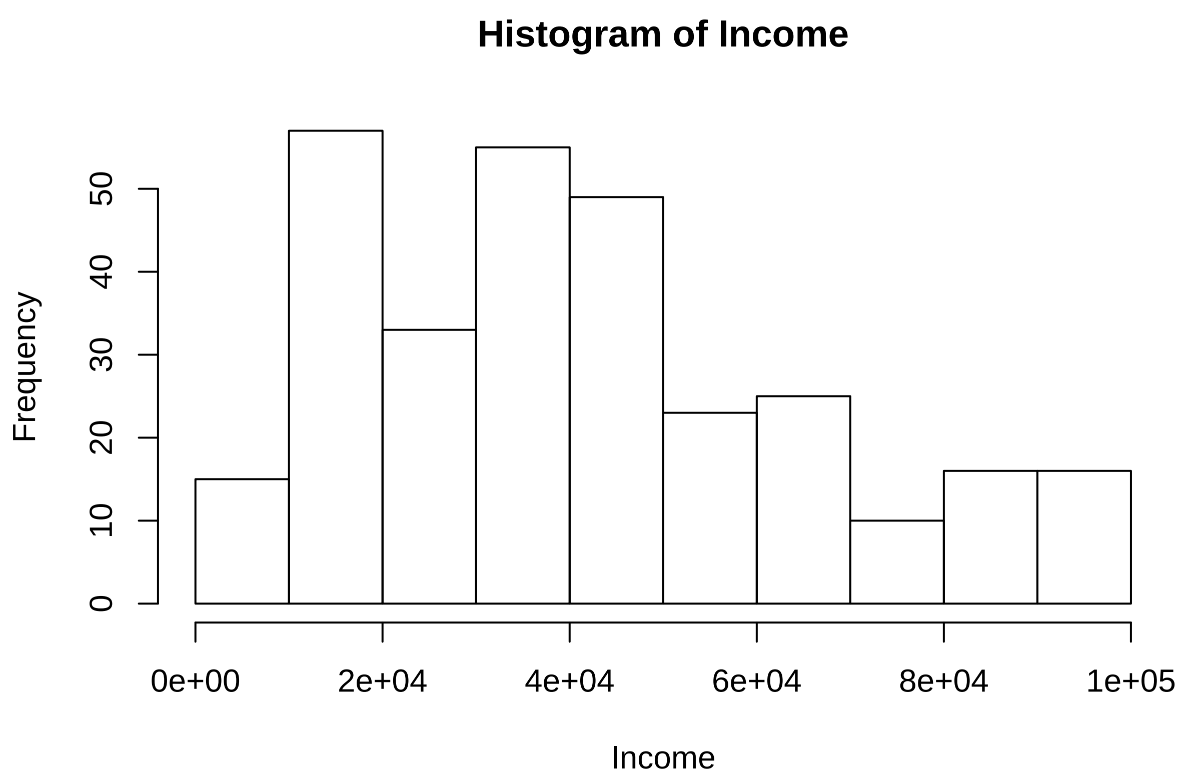
that there is a multivariatelinear regression normal error model underlying the relationship of income with several predictors. - Research Problem: "How the Other Half Lives" Gordon Weinberg

Exploratory Data Analysis

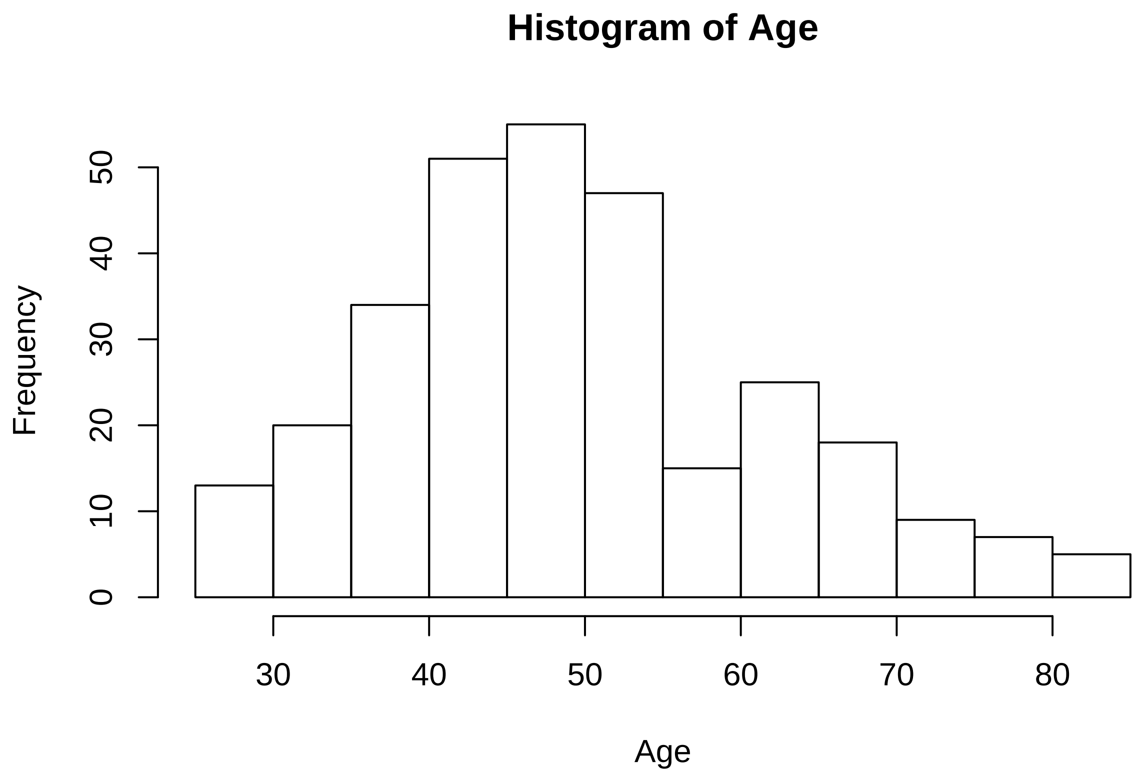
Data Description: **"The New York City Housing and Vacancy Survey is done every three years in an attempt to accurately understand the current housing conditions in the New York City. The survey is well-designedand has an admirably high response rate." - Research Problem: "How the Other Half Lives" Gordon Weinberg**

The definitions of the variables are listed below: Income: total household income (in$) Age: respondent's age (in years) MaintenanceDef: number of maintenance deficiencies between 2002 and 2005 NYCMove: the year the respondent moved to New York City.

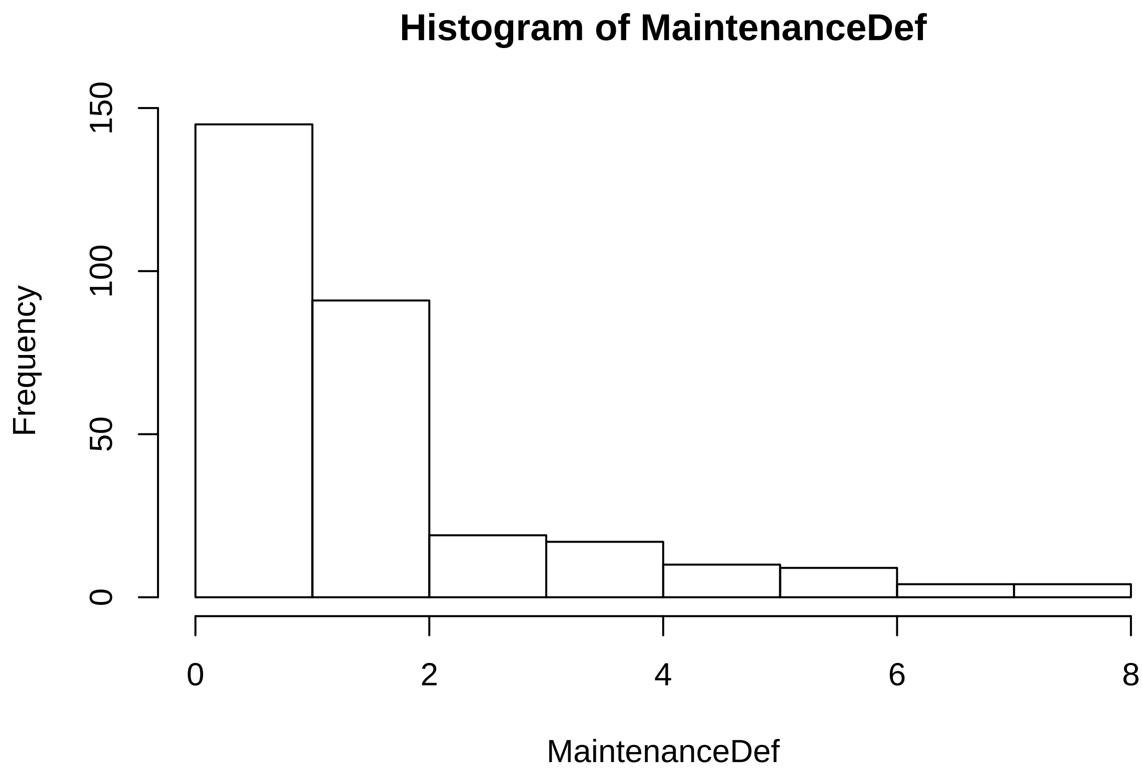
```r
hist(nyc$Income,
  main = "Histogram of Income",
  xlab = "Income")
```



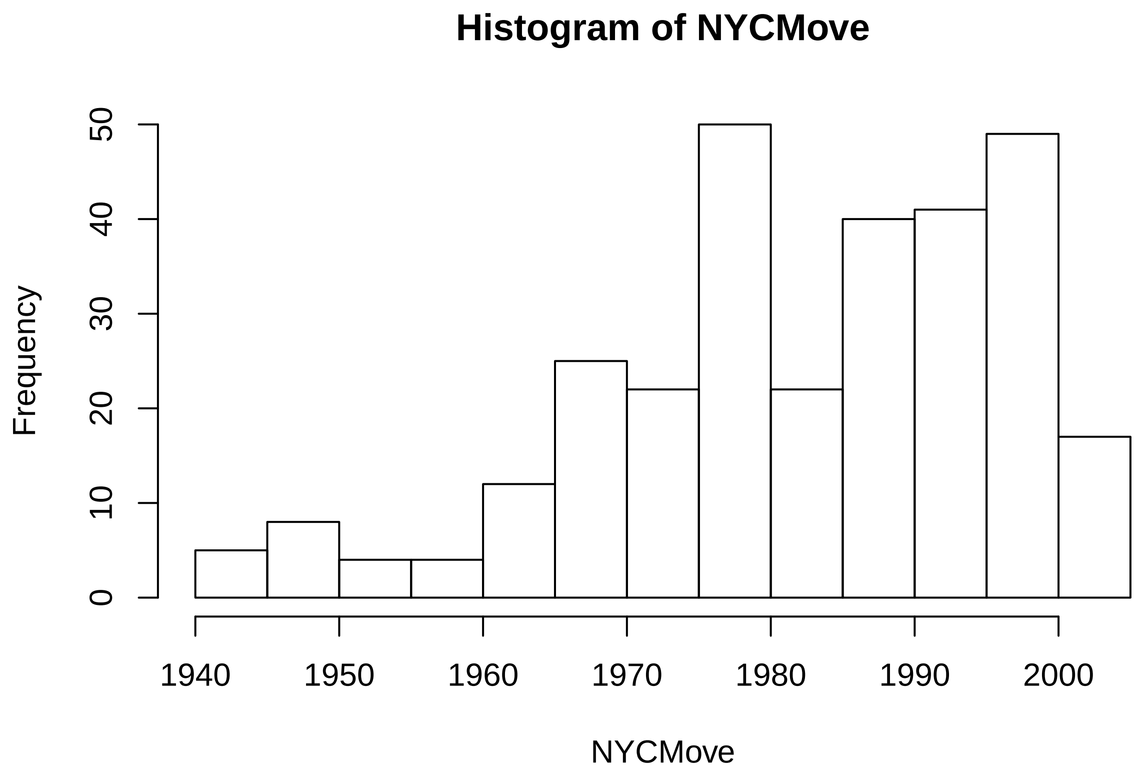

```r
hist(nyc$Age,
  main = "Histogram of Age",
  xlab = "Age")
```




```r
hist(nyc$MaintenanceDef,
  main = "Histogram of MaintenanceDef",
  xlab = "MaintenanceDef")
```




```r
hist(nyc$NYCMove,
  main = "Histogram of NYCMove",
  xlab = "NYCMove")
```



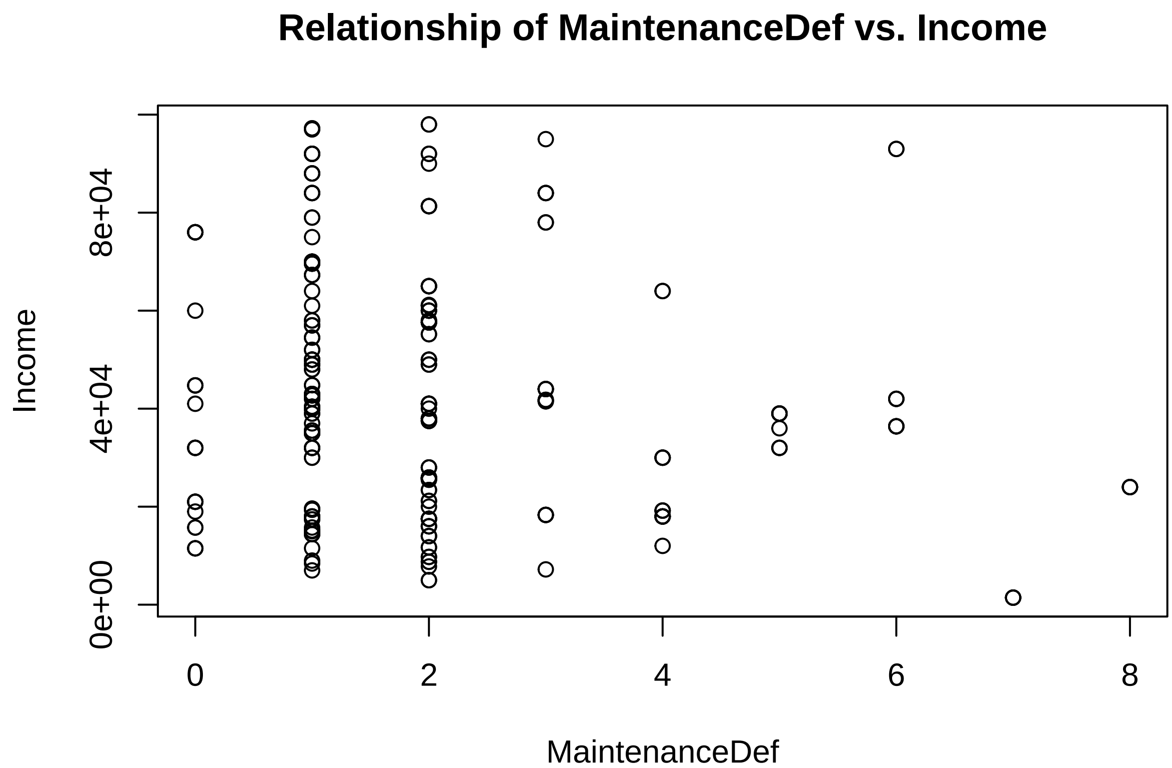

```r
summary(nyc)
```

```
##      Income           Age        MaintenanceDef     NYCMove
##  Min.   : 1440   Min.   :26.00   Min.   :0.00   Min.   :1942
##  1st Qu.:21000   1st Qu.:42.00   1st Qu.:1.00   1st Qu.:1973
##  Median :39000   Median :49.00   Median :2.00   Median :1985
##  Mean   :42266   Mean   :50.03   Mean   :1.98   Mean   :1983
##  3rd Qu.:57800   3rd Qu.:58.00   3rd Qu.:2.00   3rd Qu.:1995
##  Max.   :98000   Max.   :85.00   Max.   :8.00   Max.   :2004
```

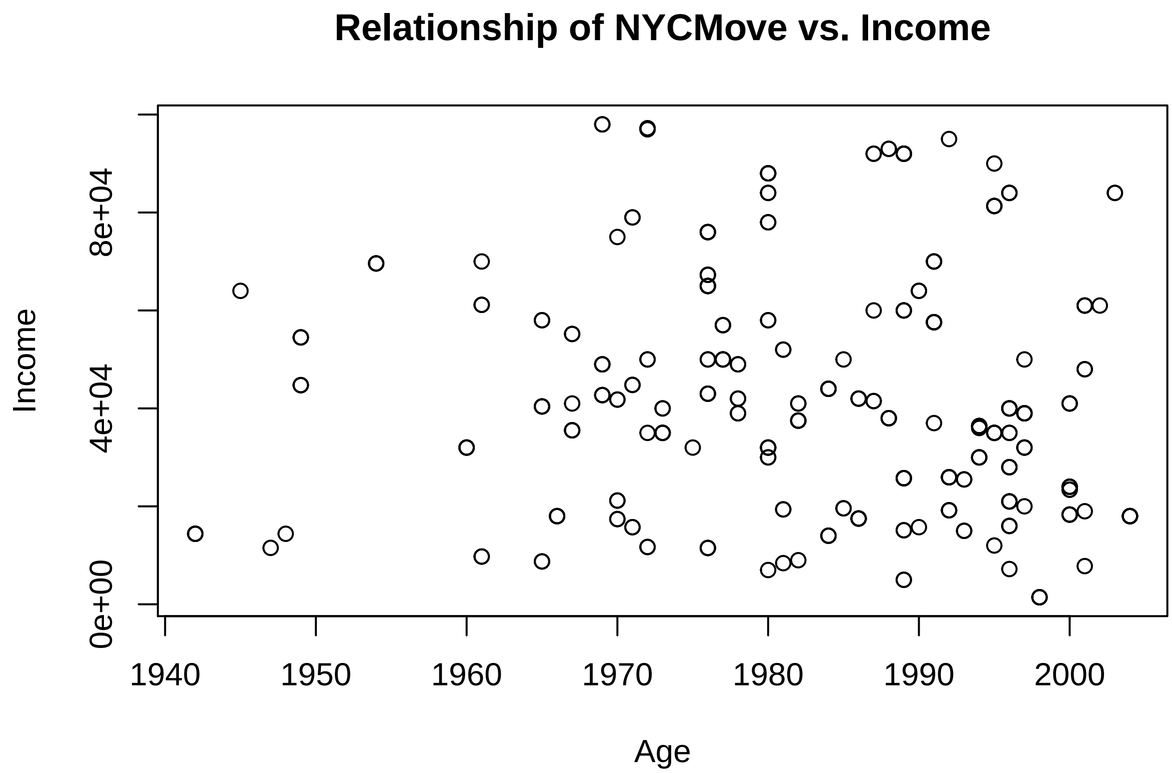
The histogram of Income is skewed right, unimodal, with possible outliers between 8x10^4 and 1x10^5. The histogram of Age is weakly skewed right, it resmebles a bell shape, but not symmetric enough to be classified as so, unimodal, with no outliers. The histogram of MaintenanceDef is skeweded right, unimodal, and does not have any outliers. The histogram of NYCMove is skewed left, unimodal, with a possible outlier between 1945-1950.

```r
plot(Income~Age,
  data = nyc,
  main = "Relationship of Age vs. Income",
  xlab = "Age",
  ylab = "Income")
```


```
plot(Income~MaintenanceDef,
  data = nyc,
  main = "Relationship of MaintenanceDef vs. Income",
  xlab = "MaintenanceDef",
  ylab = "Income")
```



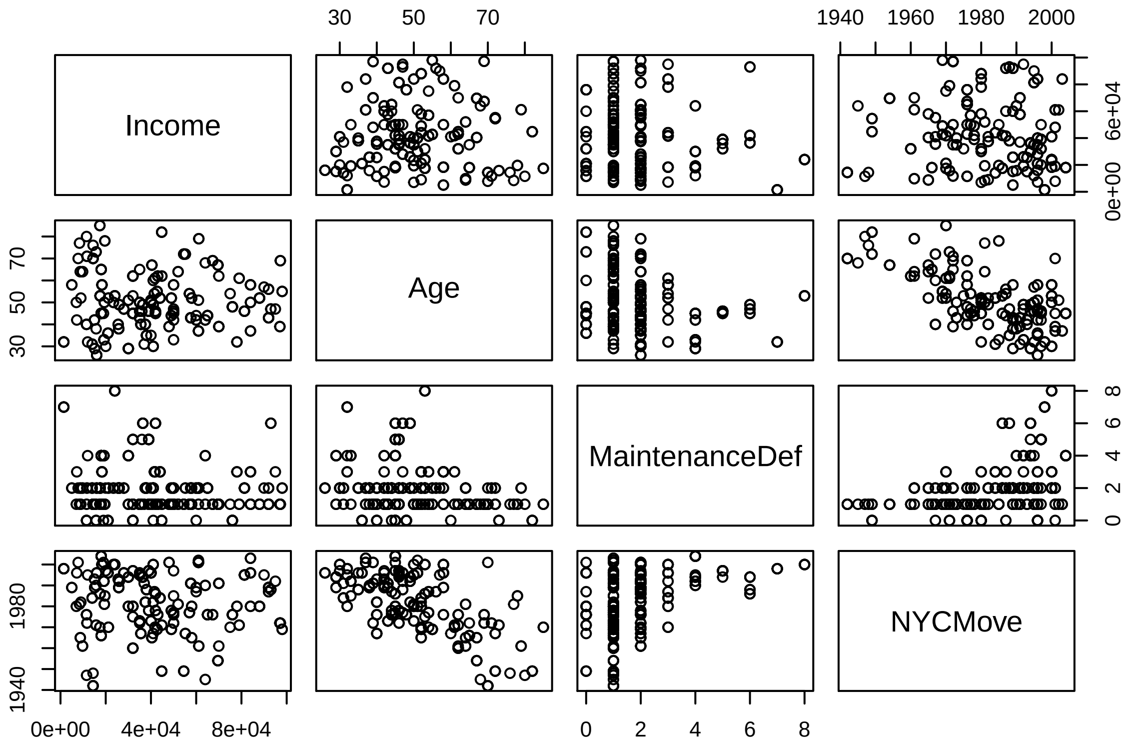

```
plot(Income~NYCMove,
  data = nyc,
  main = "Relationship of NYCMove vs. Income",
  xlab = "Age",
  ylab = "Income")
```




```r
cor(nyc)
```

```
##                     Income         Age MaintenanceDef    NYCMove
## Income          1.00000000  0.03593162     -0.1681017 -0.1009987
## Age             0.03593162  1.00000000     -0.2486687 -0.6365920
## MaintenanceDef -0.16810175 -0.24866870      1.0000000  0.4563387
## NYCMove        -0.10099865 -0.63659204      0.4563387  1.0000000
```

```r
pairs(nyc)
```



There is a weak positive linear relationship between Age and Income. There is a weak negative linear relationship between MaintenanceDef and Income. There is a weak negative linear relationship between NYCMove and Income. There is a negative linear relationship between MaintenanceDef and Age. There is a a strong negative linear relationship between NYCMove and Age. The strength of this relationship is concerning for multicollinearity. There is a strong positive parabolic relationshiup between NYCMove and MintenanceDef. The strength of this relationship is concerning for multicollinearity.

Modeling

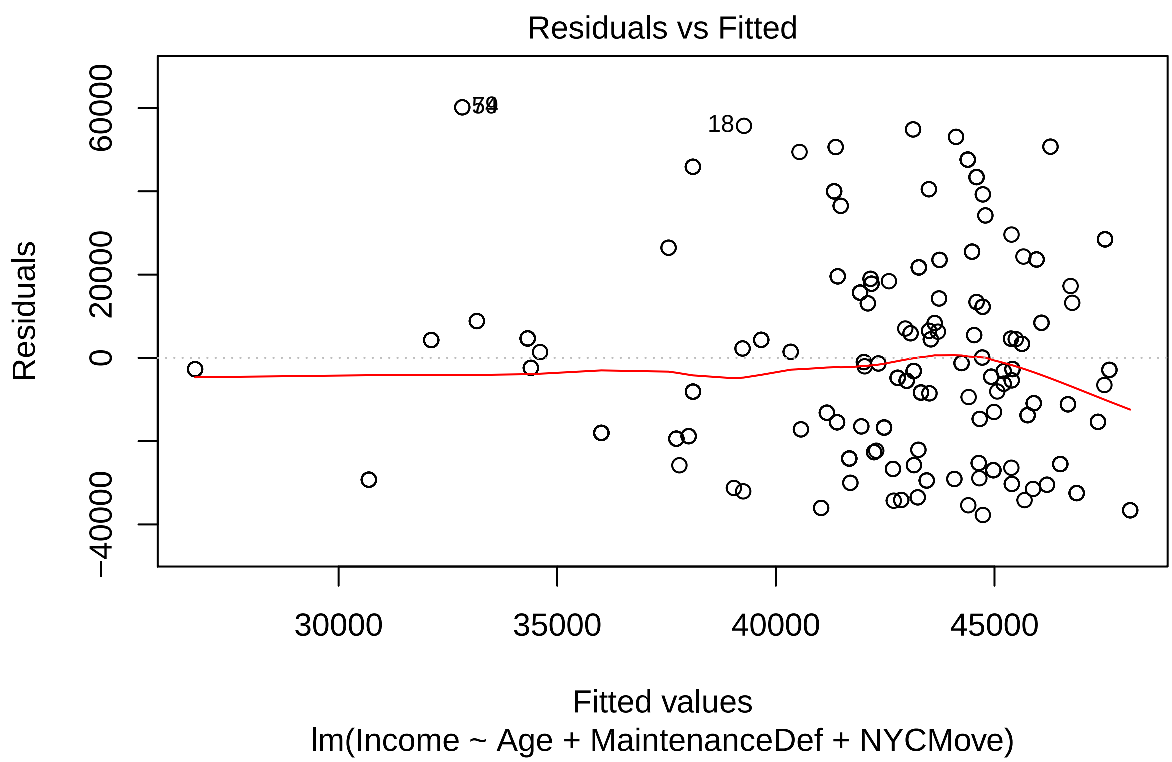
```
nyc.mod <- lm(Income ~ Age + MaintenanceDef + NYCMove,
              data = nyc)
summary(nyc.mod)
```

```
##
## Call:
## lm(formula = Income ~ Age + MaintenanceDef + NYCMove, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37734 -18010  -2878  14971  60171
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     237408.41  278939.01   0.851   0.3954
```

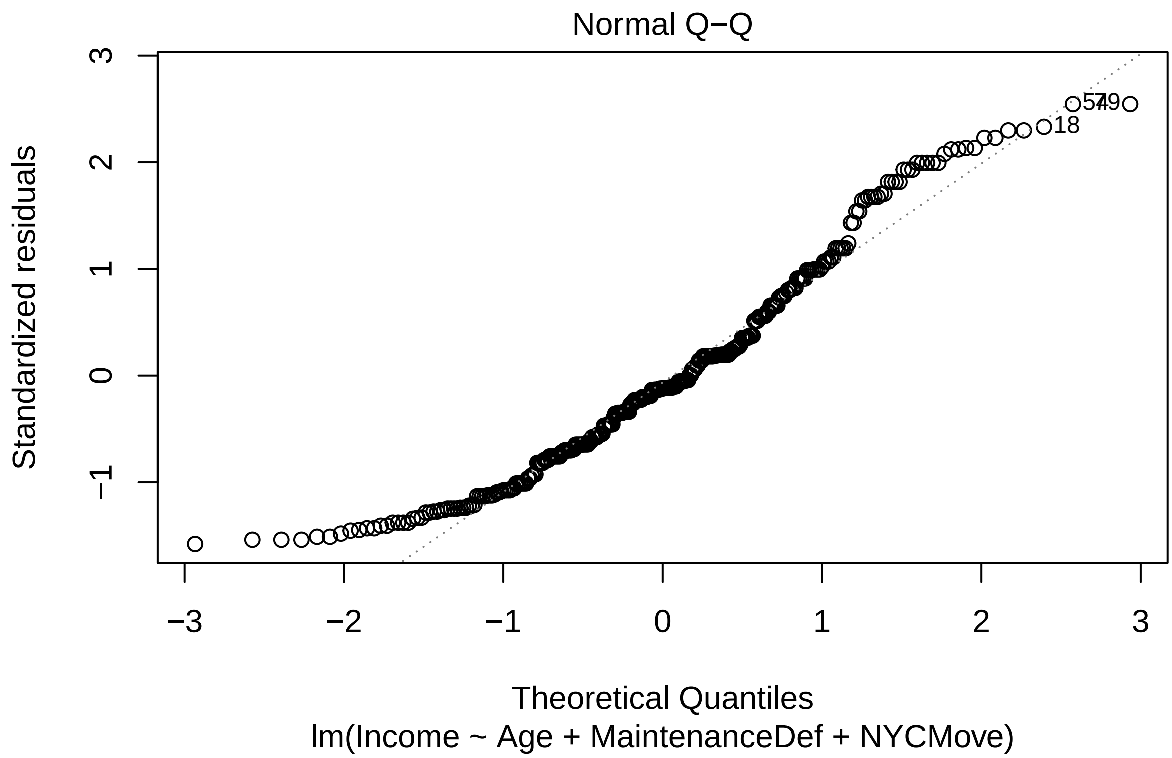
```
## Age                  -71.98     144.97  -0.496   0.6199
## MaintenanceDef   -2273.22     964.72  -2.356   0.0191 *
## NYCMove              -94.34     138.82  -0.680   0.4973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23960 on 295 degrees of freedom
## Multiple R-squared:  0.02981,    Adjusted R-squared:  0.01995
## F-statistic: 3.022 on 3 and 295 DF,  p-value: 0.03005
```

The p-value of the F-Test is below .05, so the null hypothesis is rejected, there is a relationhsip between Income and at least one of the explanatory variables listed. The p-value for every explanatory variable is above .05, except for MaintenanceDef. As a reult the null hypothesis is rejected on the t-test for Age and NYCMove meaning that there is not enough evidence to conclude that there is a linear relationhsip between them and Income. There is enough evidence to conclude that there is a relationship between MaintenanceDef and Income. The R^2 for the linear model is 0.02981 which means that only 2.981 percent of the model can be explained by a linear model. This is low, however, is not unexpected.

```r
plot(nyc.mod,
      which = 1)
```




```r
plot(nyc.mod,
      which = 2)
```



The residual plot is converning becasue at the right end of the graph the residuals are not centered around 0 which calls into question the mean = 0 assumption, and they dont show approximately constant spread above and below the zero line which calls into question the constant standard deviation assumption.

The QQ plot is concerning becasue at either end of the plot, the data varies considerably far from the Nomral Q-Q line. This calls into question the Normality Assumption. However we should not be too worried about the Normality assumption not being met in this case becasue the saample size is large enough.

```
car::vif(nyc.mod)
```

```
##            Age MaintenanceDef        NYCMove
##       1.687649       1.267728       1.999724
```

None of the vif values are above 2.5, so none are concerning for multicollinearity.

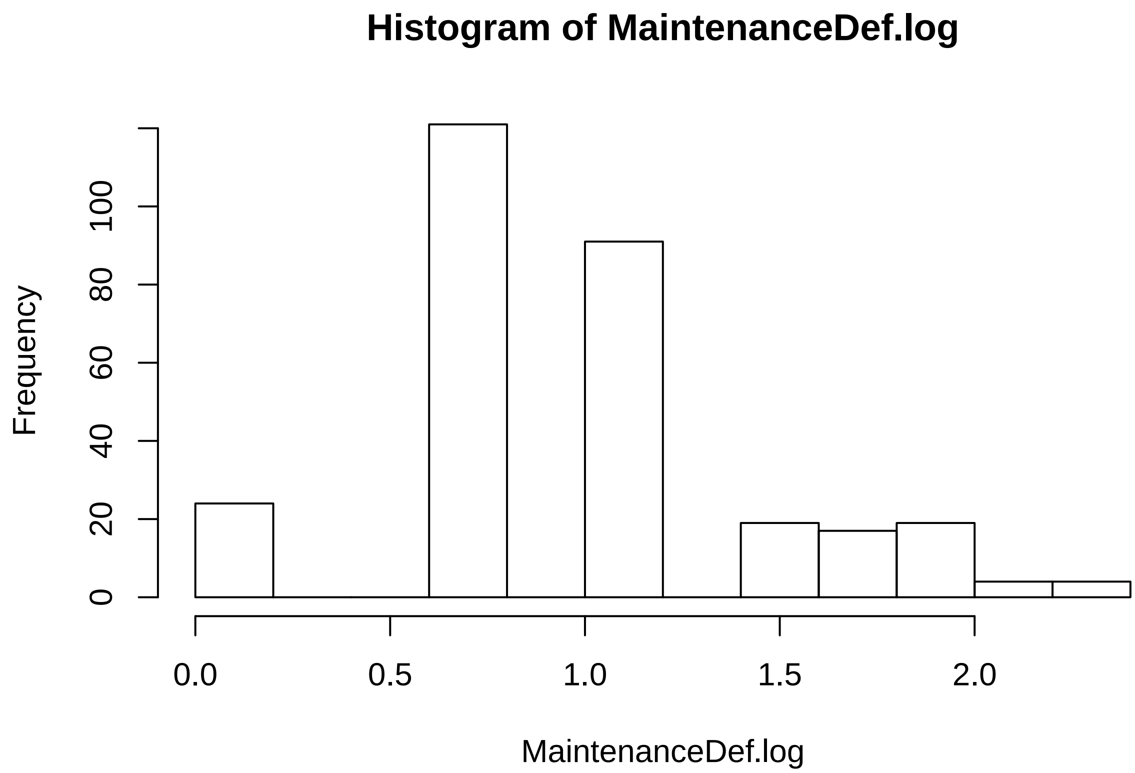
```
min(nyc$MaintenanceDef)
```

```
## [1] 0
```

```
y_shifted <- nyc$MaintenanceDef + 1.1
```

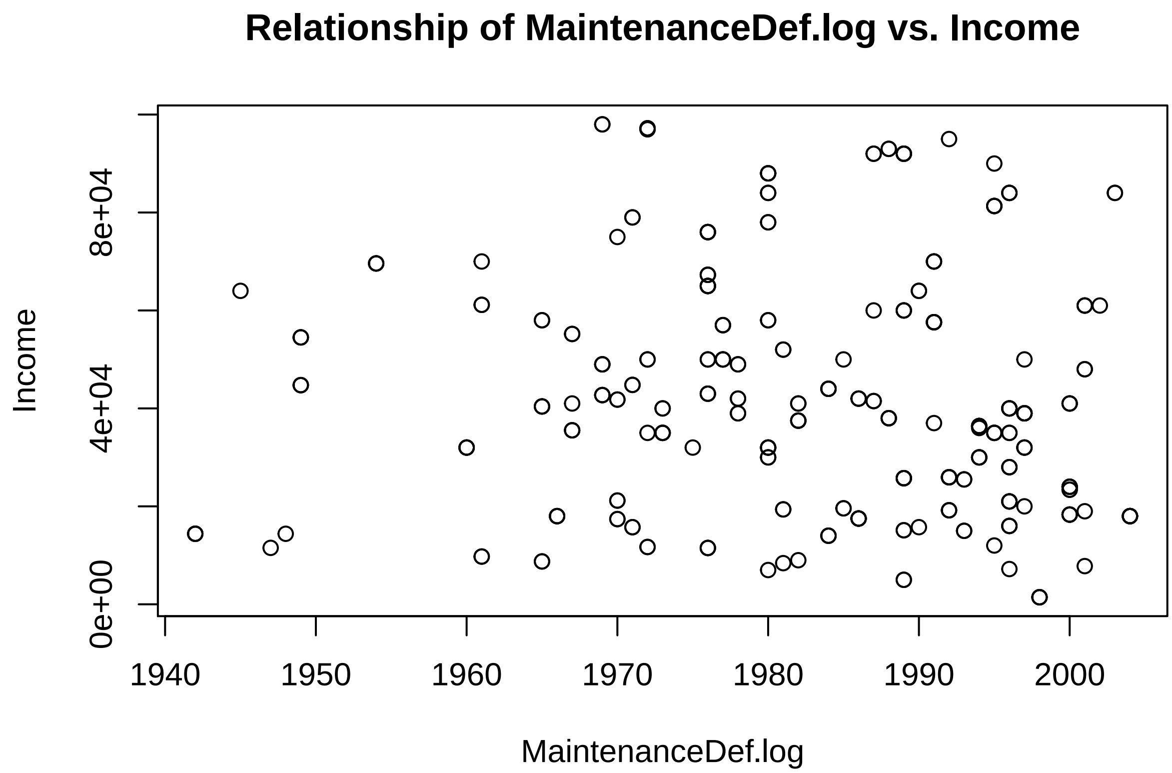
```
nyc$MaintenanceDef.log <- log(y_shifted)
```

```
hist(nyc$MaintenanceDef.log,
  main = "Histogram of MaintenanceDef.log",
  xlab = "MaintenanceDef.log")
```



The histogram of MaintenanceDef.log is unimodal and vaguely symmetric, as in it is not skewed. It is much better than the MaintenanceDef histogram.

```
plot(Income~NYCMove,
  data = nyc,
  main = "Relationship of MaintenanceDef.log vs. Income",
  xlab = "MaintenanceDef.log",
  ylab = "Income")
```



There apprears to be a vaguely linear relationship between MaintenanceDef.log and Income.

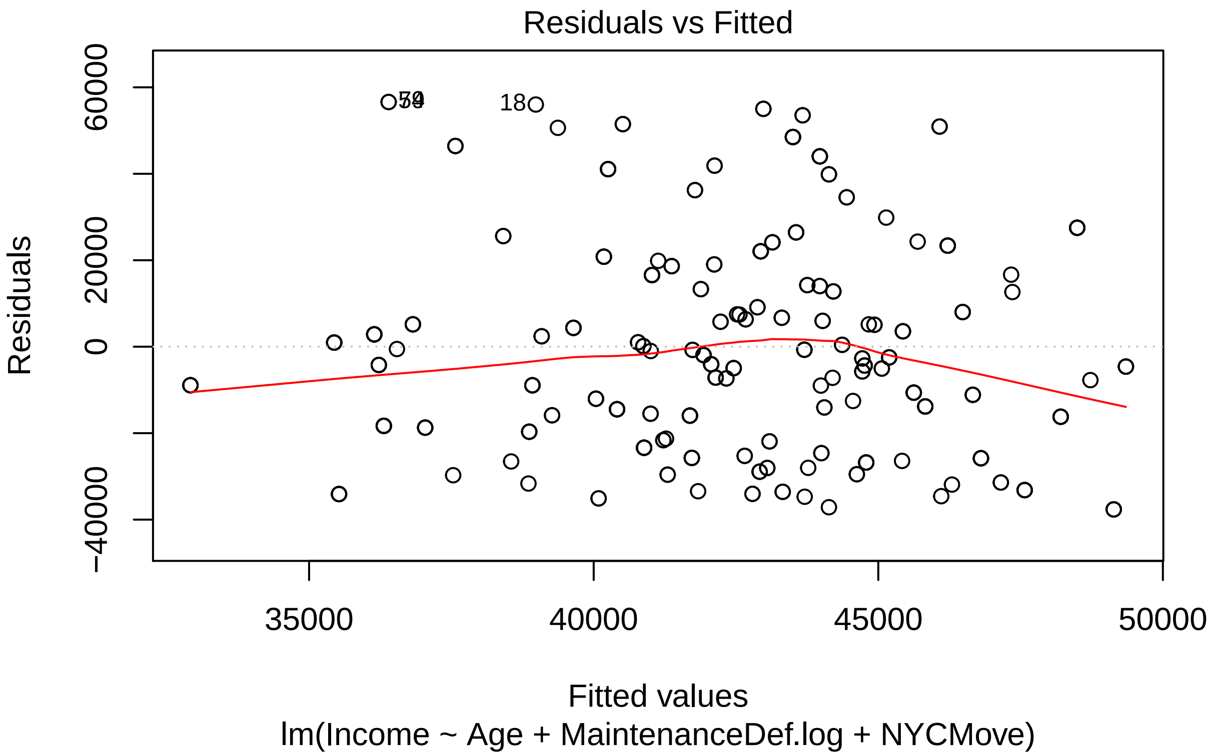
```
nyc.log <- lm(Income ~ Age + MaintenanceDef.log + NYCMove,
              data = nyc)
summary(nyc.log)
```

```
##
## Call:
## lm(formula = Income ~ Age + MaintenanceDef.log + NYCMove, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37637 -18313  -4064  15411  56602
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        315107.45  278872.19   1.130   0.2594
## Age                   -80.21     145.55  -0.551   0.5820
## MaintenanceDef.log  -5676.20    3367.28  -1.686   0.0929 .
## NYCMove              -132.70     139.07  -0.954   0.3408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24070 on 295 degrees of freedom
## Multiple R-squared:  0.02098,    Adjusted R-squared:  0.01103
## F-statistic: 2.108 on 3 and 295 DF,  p-value: 0.09934
```

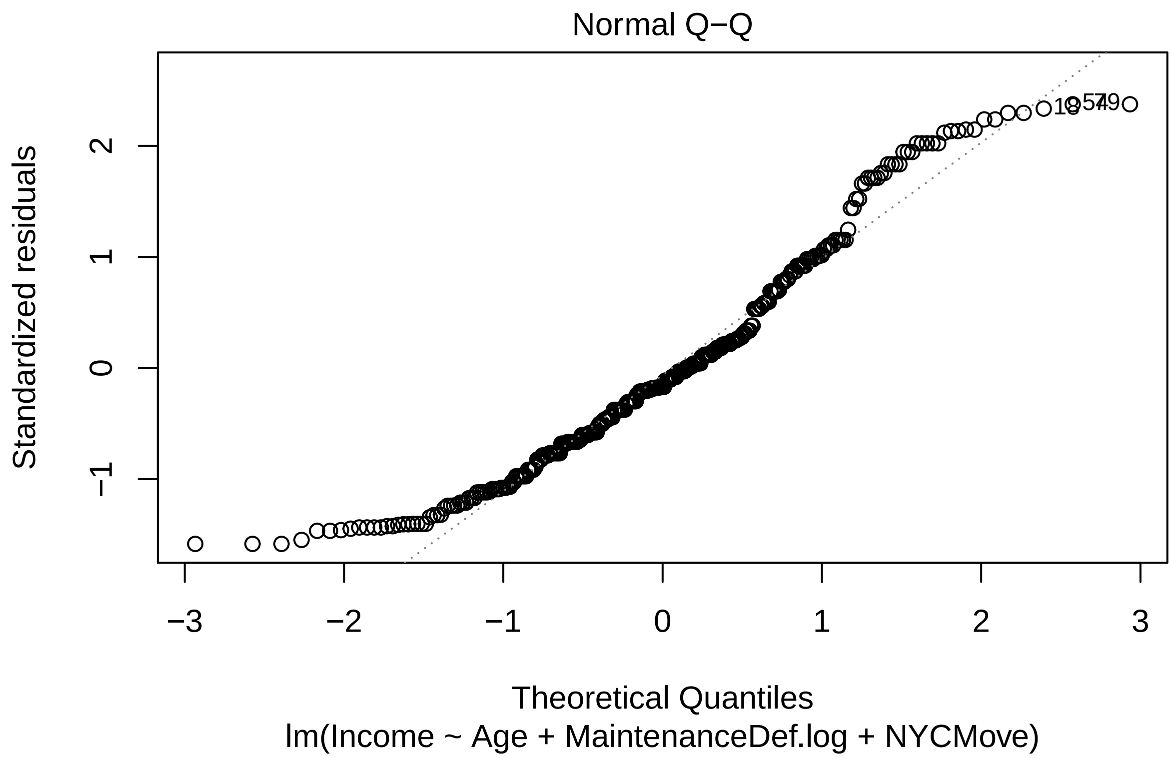
The p-value for the F-Test is 0.09934 whic his greater than .05, but still low enough to be

considered significant. As a result, there is at least one variable that has a significant relationship to Income. As we see with the p-values for the t-tests of the indiviudal variables, only MaintenaceDef.log is significant. The R^2 value is 0.02098, which means that 2.098% of the model can be explained by this model. The R^2 value is lower for this model than the nyc.mod model.

```
plot(nyc.log,
     which = 1)
```




```
plot(nyc.log,
     which = 2)
```



The residual plot looks a lot better than it did with the linear model nyc.mod. The residuals appear to be patternlessly scattered, reasonably centered around 0, and show an approximatley constant spread above and below the zero line. As a result, all of the assumptions seem to be validated. The Q-Q plot also looks better than is did with the linear model myc.mod. The data appears to be closer to the normal Q-Q line. As a result the Normality Assumption appears to also be validated.

```
car::vif(nyc.log)
```

```
##                Age MaintenanceDef.log            NYCMove
##           1.685814           1.264771           1.988725
```

None of the vif values are above 2.5, so none are concerning for multicollinearity.

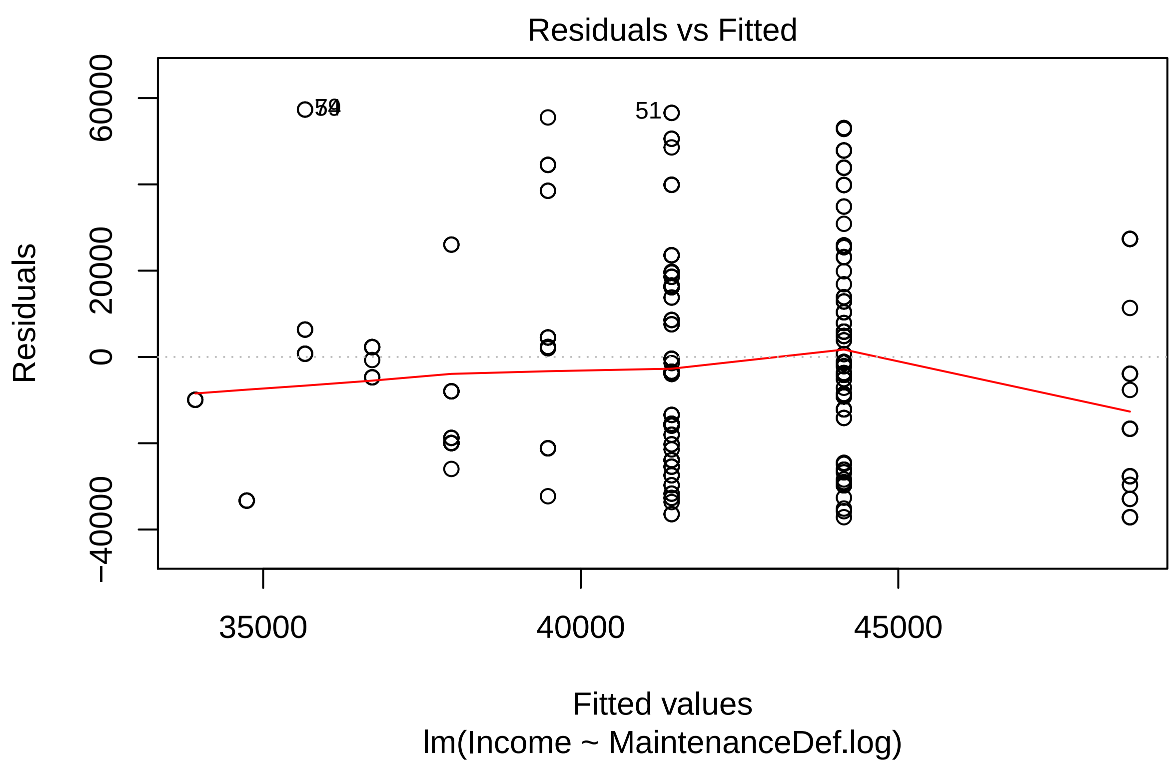
```
nyc.log.less <- lm(Income ~ MaintenanceDef.log,
              data = nyc)
summary(nyc.log.less)
```

```
##
## Call:
## lm(formula = Income ~ MaintenanceDef.log, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37147 -19364  -3744  16169  57341
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

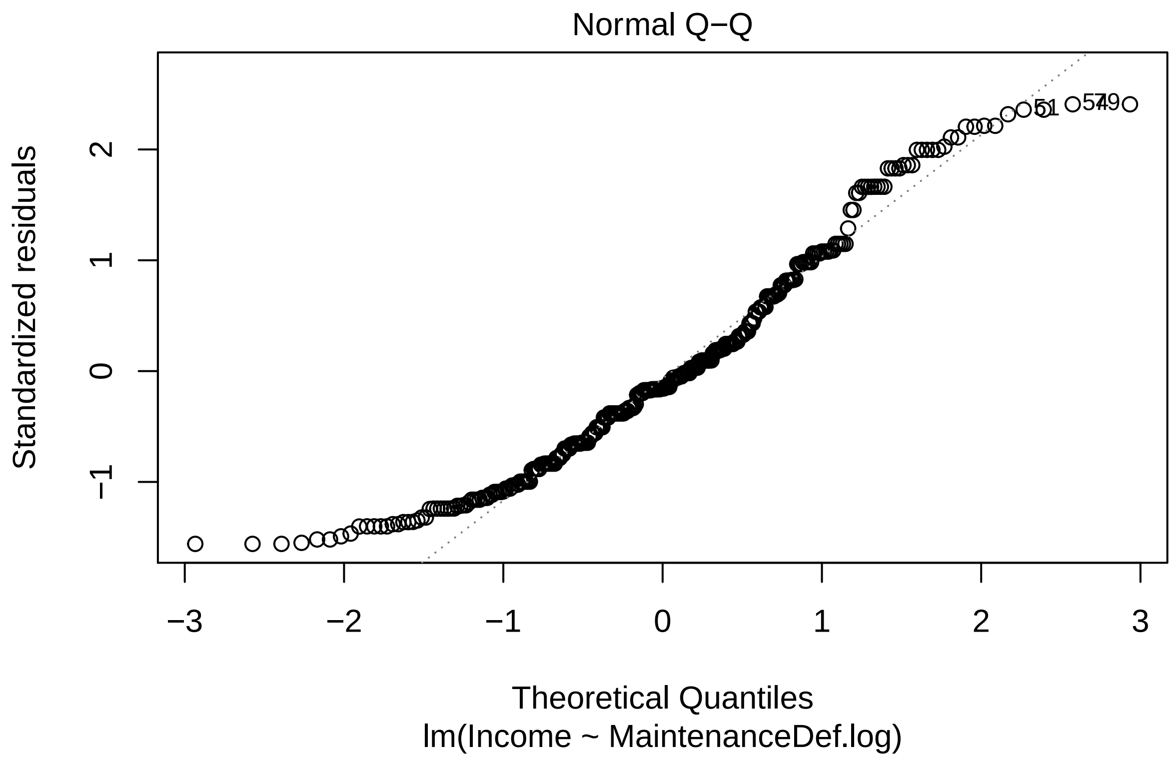
```
## (Intercept)          49311       3327    14.82   <2e-16 ***
## MaintenanceDef.log    -6965       2989    -2.33   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24020 on 297 degrees of freedom
## Multiple R-squared:  0.01796,    Adjusted R-squared:  0.01465
## F-statistic: 5.431 on 1 and 297 DF,  p-value: 0.02045
```

The p-value for the F-Test is 0.02045 which is significant or smaller than .05. As a result, there is at least one variable that has a significant relationship to Income. As we see with the p-values for the t-tests of the indiviudal variables, only MaintenaceDef.log is significant. The R^2 value is 0.01796, which means that 1.786% of the model can be explained by this model. The R^2 value is lower for this model than the nyc.mod and nyc.log model.

```r
plot(nyc.log.less,
      which = 1)
```




```r
plot(nyc.log.less,
      which = 2)
```



At first sight it was look as though the residuals do not look randomly scattered, and that the independence assumption is violated. However we must remember that this is in part due to the nature of the data. Otherwise the other 2 assumptions seem to be validated better than in the linear model but not as well as in the mod.log model. The Q-Q plot also looks better than is did with the linear model myc.mod and similar to how it didin the nyc.log model. The data appears to be closer to the normal Q-Q line. As a result the Normality Assumption appears to be validated.

The models I tried were: - nyc.mod: linear model of nyc with three vbariables: Age, MaintenaceDef, and NYCMove, all untransformed. - nyc.log: linear model of nyc with three variables, Age, NYCMove, and a log transformation of the MaintenanceDef variable. - nyc.log.less: linear model of nyc with only one varibale, the log tranformation of the MaintenanceDef variable. **I ended up choosing the nyc.log model becasue all of the assumptions are validated, and the P-value is near the highest of all of the models I tested. Additionaly, there was no concern for multicolinearity.**

Prediction

We are interested in predicting the income for a household with three maintenance deficiencies and whose respondent's age is 53 and who moved to NYC in 1987.

nyc.log: Income = Age + MaintenanceDef.log + NYCMove Income = beta0(Age) + beta1(MaintenanceDef.log) + beta2(NYCMove) + error

```
-80.21*53 + -5676.20*(log(3) + 1.1) + -132.70*1987 + 315107.45
```

```
## [1] 34701.66
```

The predicted income for a household with three maintenance deficiencies and whose respondent's age is 53 and who moved to NYC in 1987 is 34701.66 using the nyc.log model.

Discussion

The model that fit the data best was found to be nyc.log. It it is a linear model of nyc with three variables: Age, NYCYears, and a log transofrmation of the variable MaintenaceDef. I chose this model becasue all of the assumptions are validated, and the P-value is near the highest of all of the other models I tested. Additionaly, there was no concern for multicolinearity. Although I found the nyc.log model fit my data the best, it is important to note that the R^2 value was very low. ALthough that is expected for this data set, it should be noted. Furthermore, in the model, only the variable MaintenanceDef.log was significant, and its p-value for the t-test was above 0.5. The future directions in which the work can go are to analyze how Income is related to location in New York City. I am aware that location was not a provided variable in this dataset, but if it were included, one could analyze how location relates to Income, mapping out wealthier sub-burroughs and burroughs. Furthermore, if race were a variable, one could analyze how it relates to Income and location to see if there are racial divisons in the city and see where minority populations live compared to majority populations.