# Estimating New York City Household Income

Jonathan Lai jklai

Due Wed, October 26, at 11:59PM

## Contents

```
library("knitr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("interactions")
library("leaps")
```

## Introduction

New York City (NYC) is one of the biggest city in the world. Apart from its bustling lights, abundant skyscrapers, and being a cultural and financial center, NYC is also infamous for its expensive housing and rental prices. Living in NYC is expensive. So The New York City Housing and Vacancy Survey has surveyed the New York City population every 3 years to understand housing conditions and the households living in New York City. More specifically, we will focus on predicting the income of New York City households by using characteristics of New York residences that may potentially have an effect on household income.

# Exploratory Data Analysis

## Background Information

In this dataset `nyc`, there are 299 samples collected from households in New York City. There are 4 variables, 3 potential predicting explanatory variables and 1 response variable `Income`.

Variable Descriptions:

1. `Income` refers to the total household income (in dollars) of a NYC resident
2. `Age` refers to the New York City household respondents' age (in years)
3. `MaintenanceDef` refers to the number of maintenance deficiencies of the residence between 2002 and 2005
4. `NYCMove` refers to the year the New York City household respondent moved to NYC

To get an idea of what kind of data is in the dataset `nyc`, below are the first and last few samples in the dataset:

```
head(nyc)
```

```
## # A tibble: 6 x 4
##   Income   Age MaintenanceDef NYCMove
##    <dbl> <dbl>          <dbl>   <dbl>
## 1   8400    77              1    1981
## 2  17510    53              2    1986
## 3  19200    33              4    1992
## 4  42717    55              1    1969
## 5   5000    58              2    1989
## 6  30000    29              4    1994
```

```
tail(nyc)
```

```
## # A tibble: 6 x 4
##   Income   Age MaintenanceDef NYCMove
##    <dbl> <dbl>          <dbl>   <dbl>
## 1   5000    58              2    1989
## 2  36000    46              5    1994
## 3  17510    53              2    1986
## 4  12000    32              4    1995
## 5  14000    31              2    1984
## 6  18000    45              4    2004
```

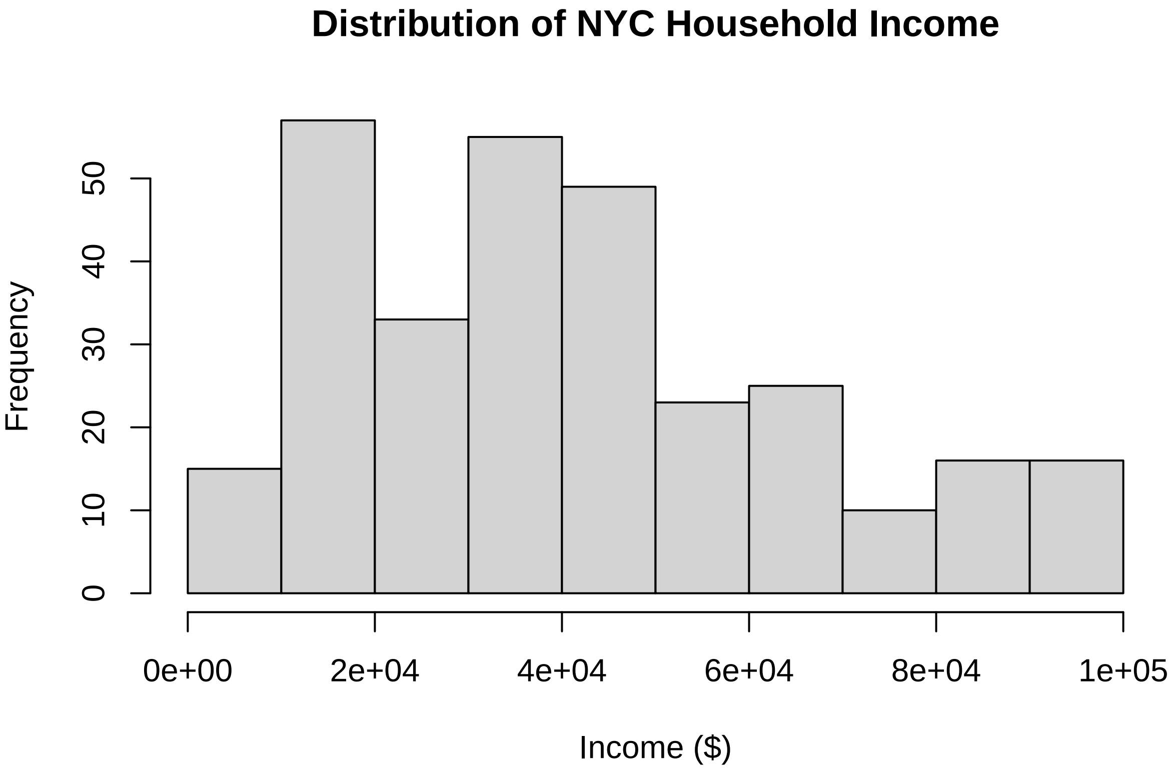
The first data point describes a 77 year old NYC resident who had 1 maintenance deficiency in their residence between 2002 and 2005 and who moved in 1981 that has a household income of $8400.

The last data point describes a 45 year old NYC resident who had 4 maintenance deficiency in their residence between 2002 and 2005 and who moved in 2004 that has a household income of $18000.

### Univariate Analysis

Below are the distributions of all the variables.

```
hist(nyc$Income,
     main = "Distribution of NYC Household Income",
     xlab = "Income ($)")
```

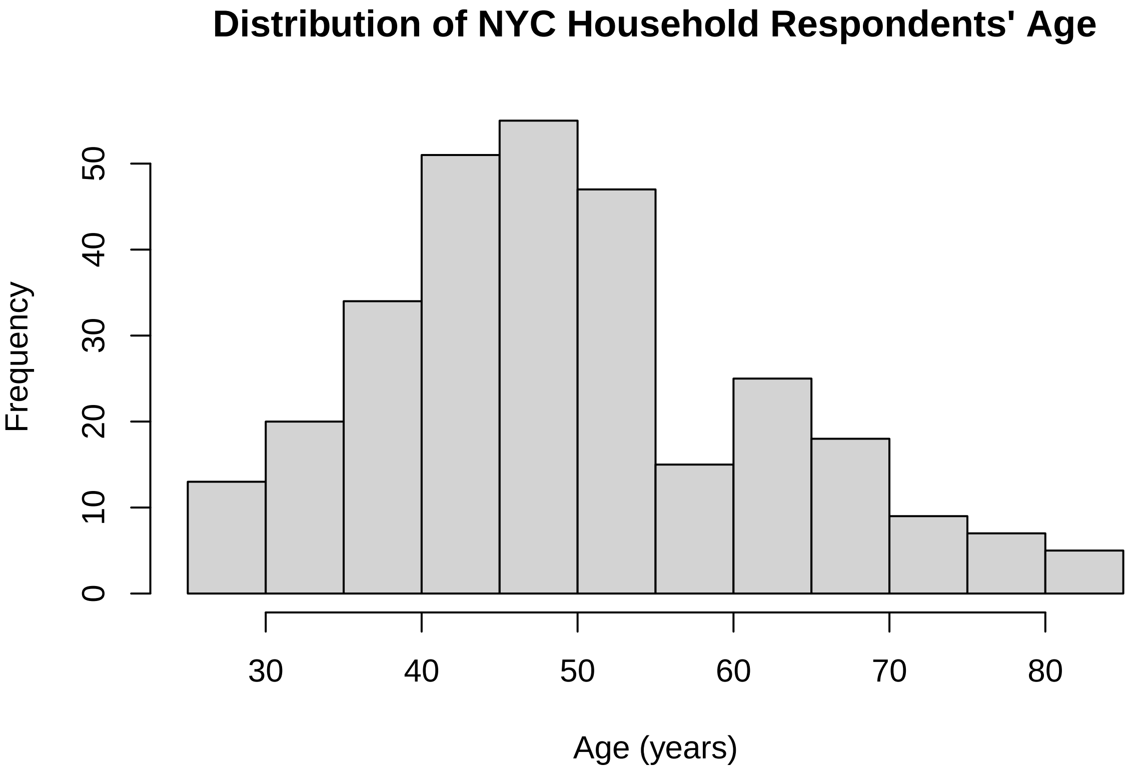


```
summary(nyc$Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1440   21000   39000   42266   57800   98000
```

The distribution of the household income from the New York City respondents appears to be slightly skewed to the right, bimodal, with no outliers. The mean household income is $42266 but the median is smaller at $39000. The range household income is $96560 and the interquartile range is $36800.

```
hist(nyc$Age,
     main = "Distribution of NYC Household Respondents' Age",
     xlab = "Age (years)")
```

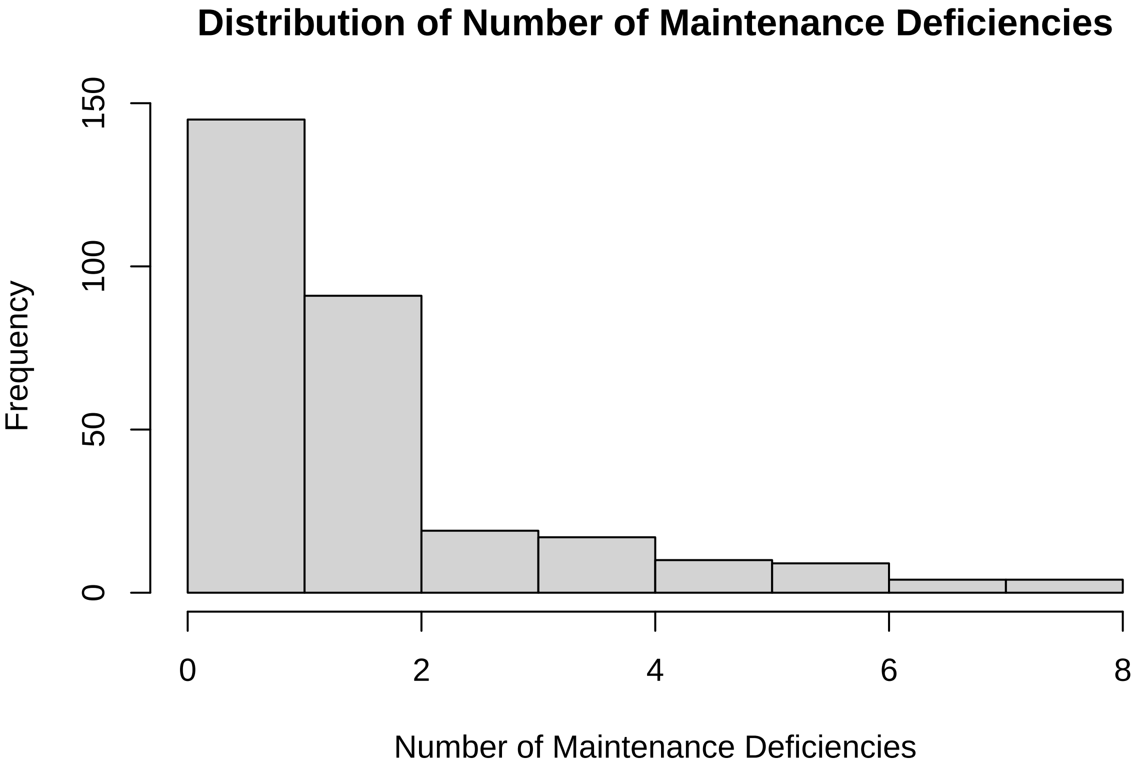


```
summary(nyc$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   26.00   42.00   49.00   50.03   58.00   85.00
```

The distribution of the age of the New York City respondents appears to be very slightly skewed to the right, unimodal, with no outliers. The mean age of the NYC respondents is 50.03 years old and the median is close at 49 years old. The range age of the NYC respondents is 59 years old and the interquartile range is 16 years old.

```
hist(nyc$MaintenanceDef,
     main = "Distribution of Number of Maintenance Deficiencies",
     xlab = "Number of Maintenance Deficiencies")
```

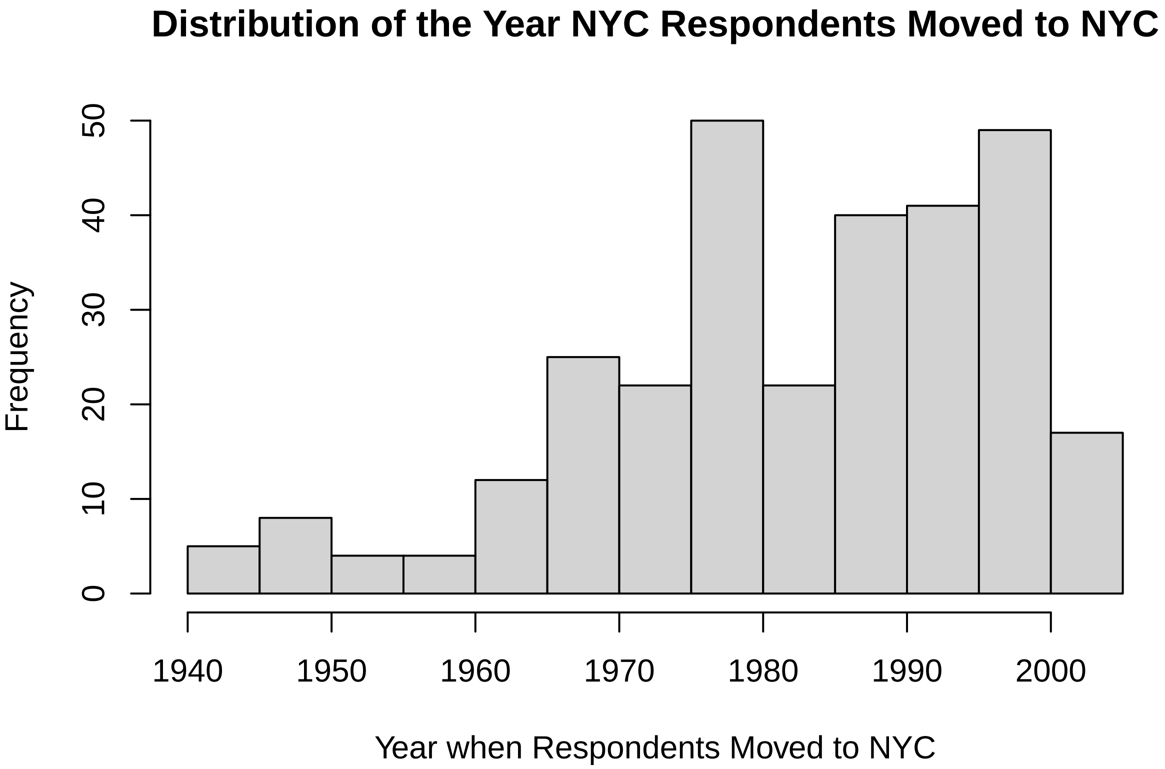


```
summary(nyc$MaintenanceDef)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    2.00    1.98    2.00    8.00
```

The distribution of the number of maintenance deficiencies of the respondents' residence in New York City appears to be skewed to the right, unimodal, with no outliers. The mean number of maintenance deficiency is 1.98 and the median is slightly larger at 2 maintenance deficiencies. The range number of maintenance deficiency is 8 and the interquartile range is 1.

```
hist(nyc$NYCMove,
     main = "Distribution of the Year NYC Respondents Moved to NYC",
     xlab = "Year when Respondents Moved to NYC")
```



```
summary(nyc$NYCMove)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1942    1973    1985    1983    1995    2004 
```

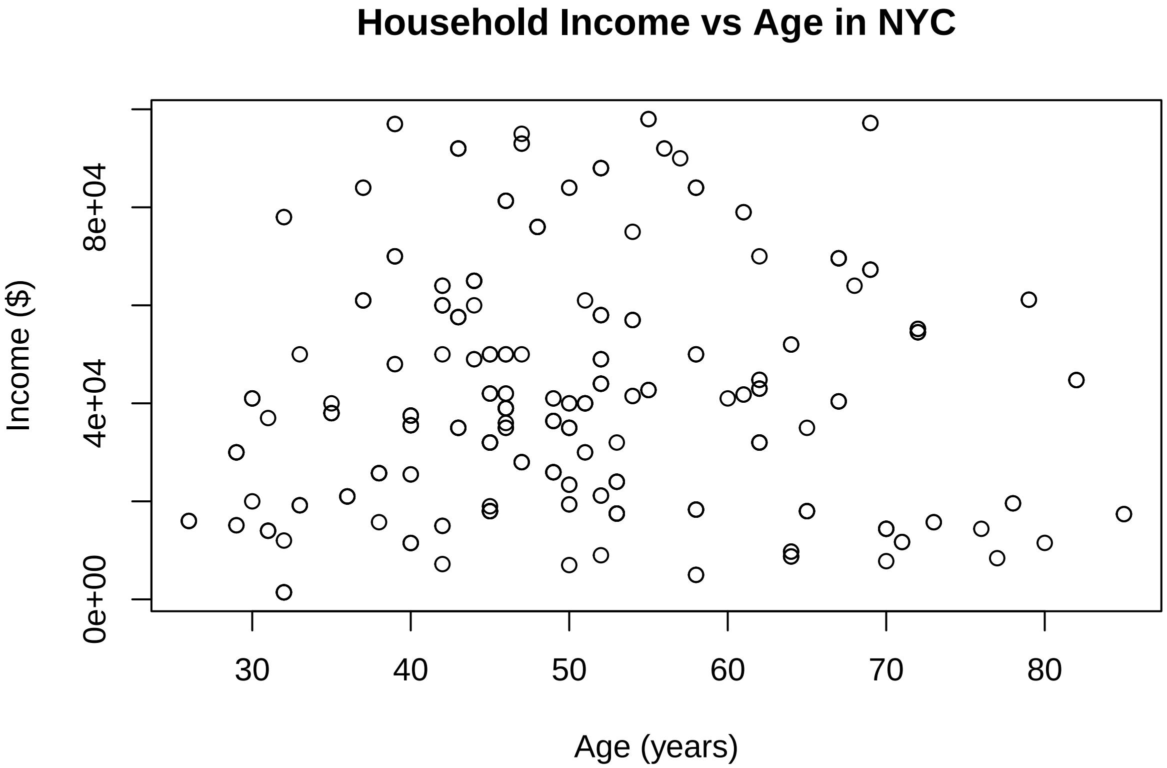
The distribution of the year the New York City respondents moved to NYC appears to be skewed to the left, bimodal, with no outliers. The mean year when the NYC respondents moved to NYC is 1983 and the median is 1985. The range year when the NYC respondents moved to the city is 62 and the interquartile range is 22.

## Bivariate Analysis

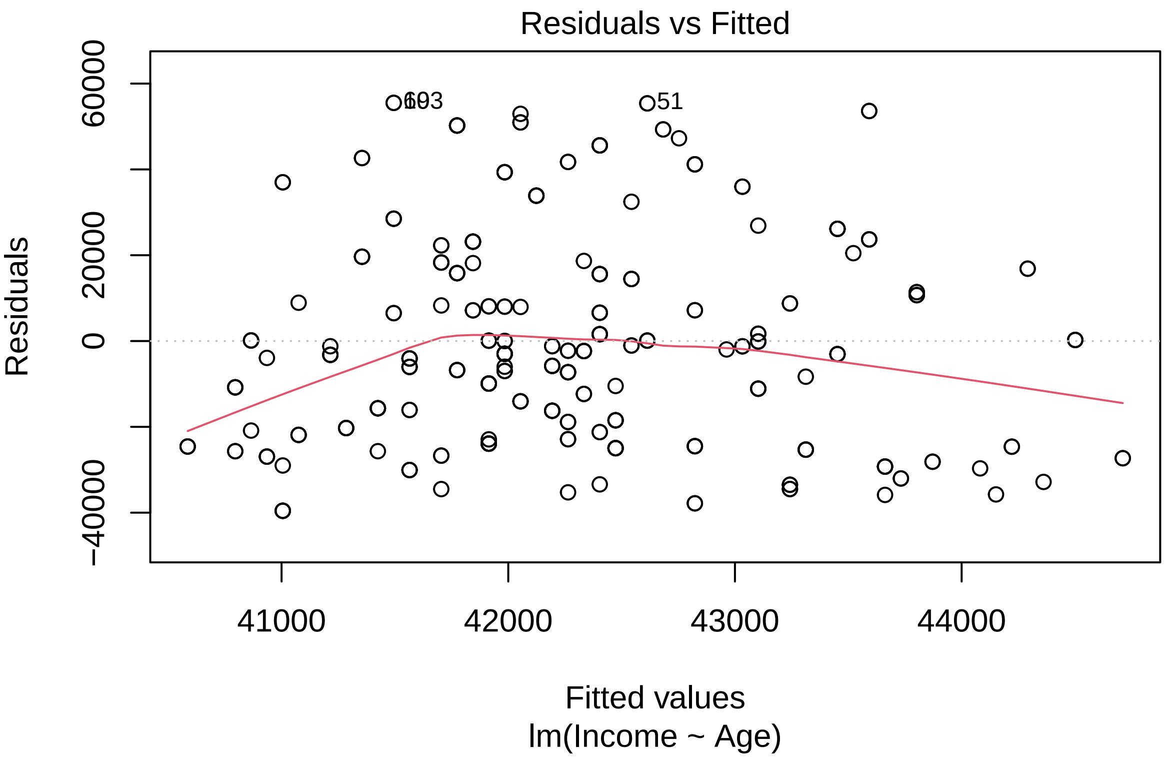
After analyzing the distribution of all the variables, we have to investigate which explanatory variables are most suitable to predict the `Income` (response variable) for creating a linear regression model.

### Variable Age

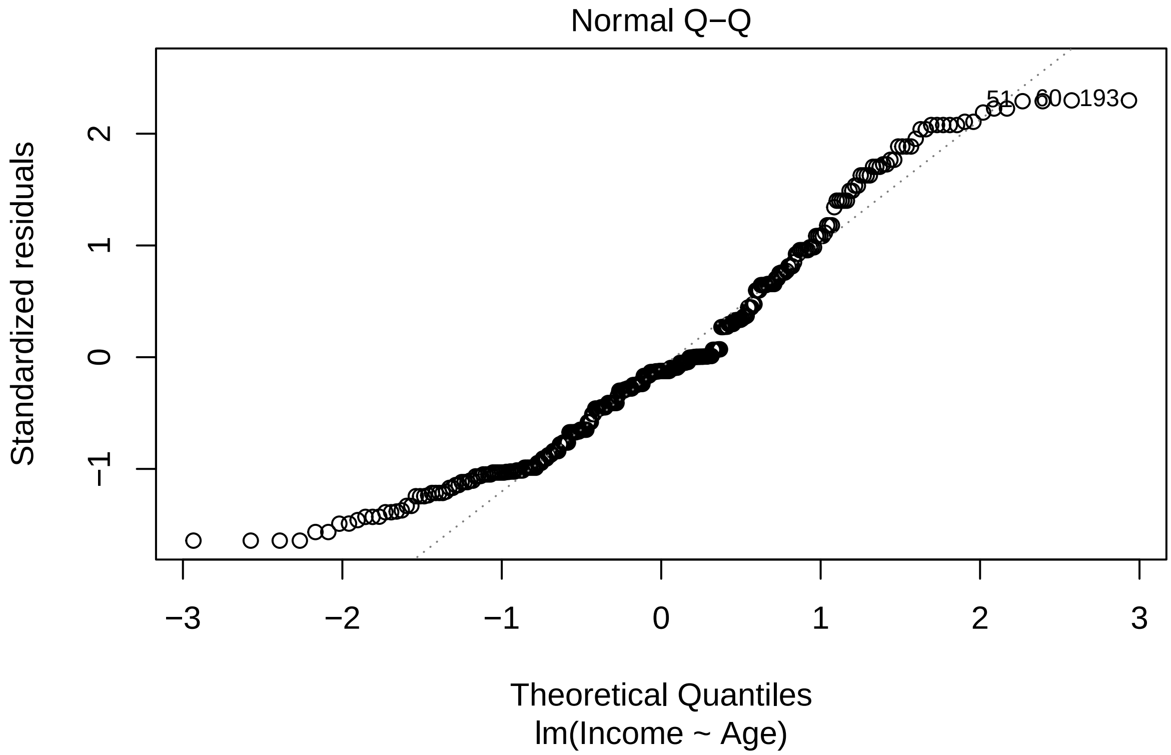
```
plot(Income ~ Age, data = nyc,
     main = "Household Income vs Age in NYC",
     xlab = "Age (years)",
     ylab = "Income ($)")
```



```
agelm1 <- lm(Income~Age, data = nyc)
plot(agelm1, which = 1)
```



```
plot(agelm1, which = 2)
```



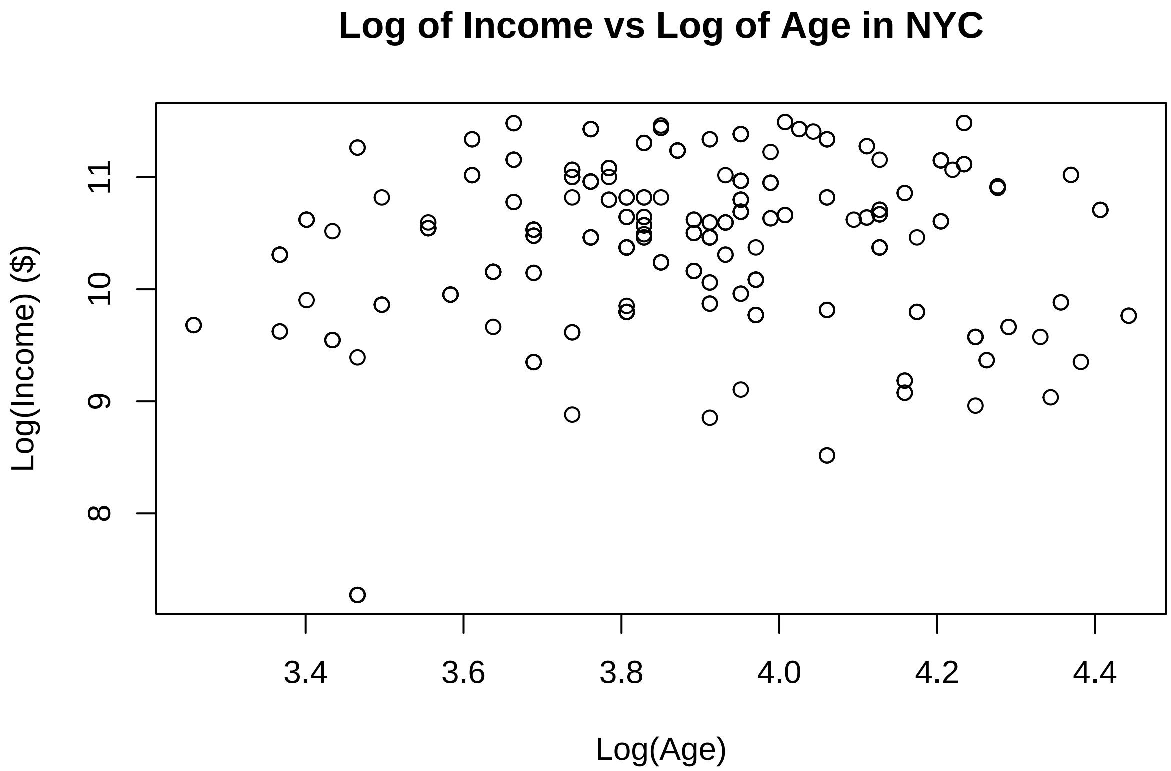
```
summary(agelm1)
```

```
##
## Call:
```

```
## lm(formula = Income ~ Age, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39565 -20285  -2984  15826  55505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38767.78    5816.70   6.665 1.29e-10 ***
## Age            69.92     112.84   0.620    0.536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24230 on 297 degrees of freedom
## Multiple R-squared:  0.001291,   Adjusted R-squared:  -0.002072
## F-statistic: 0.3839 on 1 and 297 DF,  p-value: 0.536
```

Since the scatterplot between response variable `Income` and explanatory variable `Age` appears to have a very weak correlation (or none at all), a residual mean that doesn't equal to 0, and a residual distribution that is not normally distributed, we can transform the variables involved in this scatterplot. The coefficient of determination is very low at 0.13%, meaning only 0.13% of the variability in the household income can be explained by the age of the respondents in NYC.

```r
nyc$logIncome <- log(nyc$Income)
nyc$logAge <- log(nyc$Age)
plot(logIncome ~ logAge, data = nyc,
     main = "Log of Income vs Log of Age in NYC",
     xlab = "Log(Age)",
     ylab = "Log(Income) ($)")
```

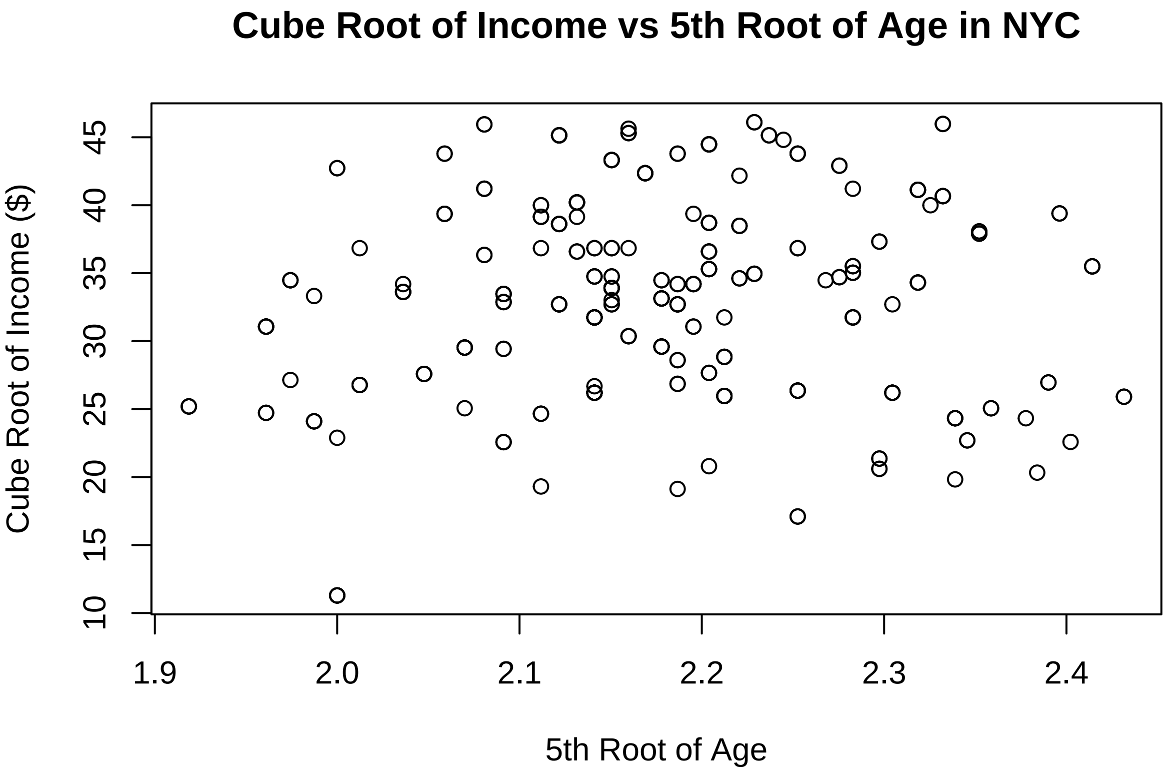


```r
agelm2 <- lm(logIncome ~ logAge, data = nyc)
summary(agelm2)
```

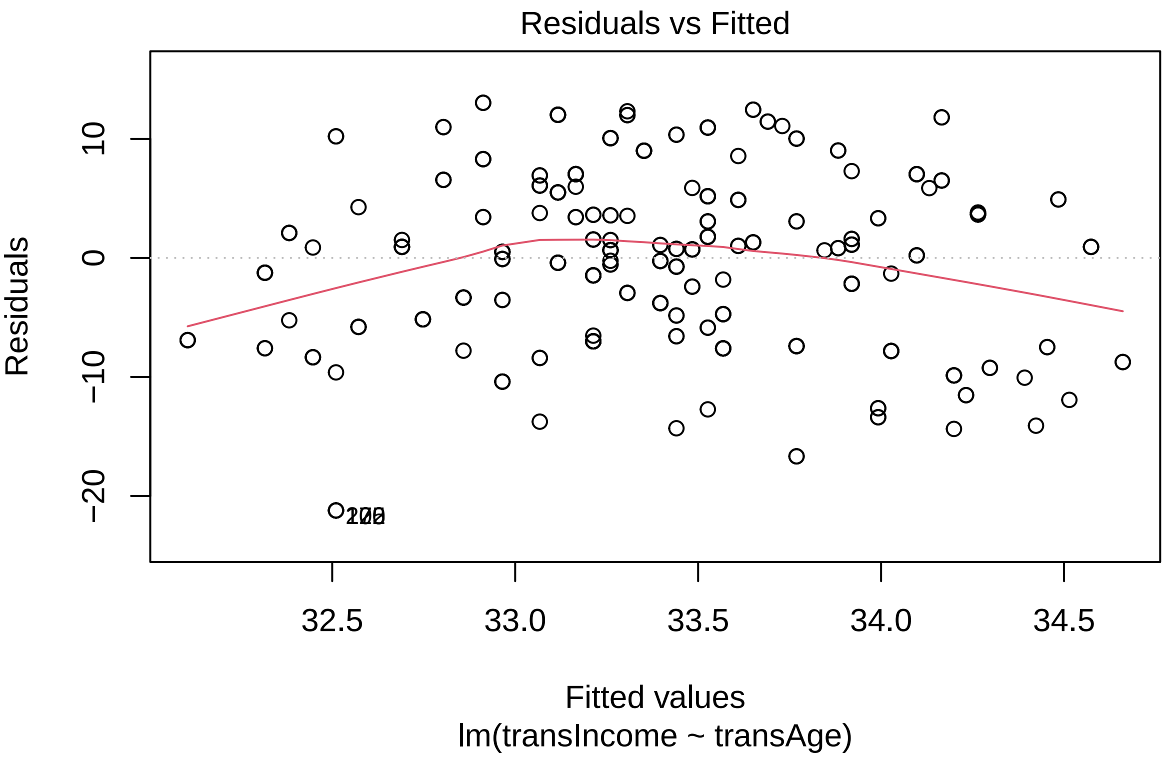
```
##
## Call:
## lm(formula = logIncome ~ logAge, data = nyc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0533 -0.4069  0.1423  0.5042  1.1005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.3383     0.6597   14.16   <2e-16 ***
## logAge        0.2849     0.1696    1.68    0.094 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7279 on 297 degrees of freedom
## Multiple R-squared:  0.009413,   Adjusted R-squared:  0.006078
## F-statistic: 2.822 on 1 and 297 DF,  p-value: 0.09402
```

Since the individual distribution of `Income` is skewed to the right, we used the logarithmic function to transform `Income` to be normally distributed. And also since the individual distribution of `Age` is skewed to the right, we transformed the variable using the logarithmic function too. However, despite producing a higher coefficient of determination value at 0.94%, using logarithmic transformation violates the normality error assumption based on the deviation in the QQ plot. A better transformation can be applied.

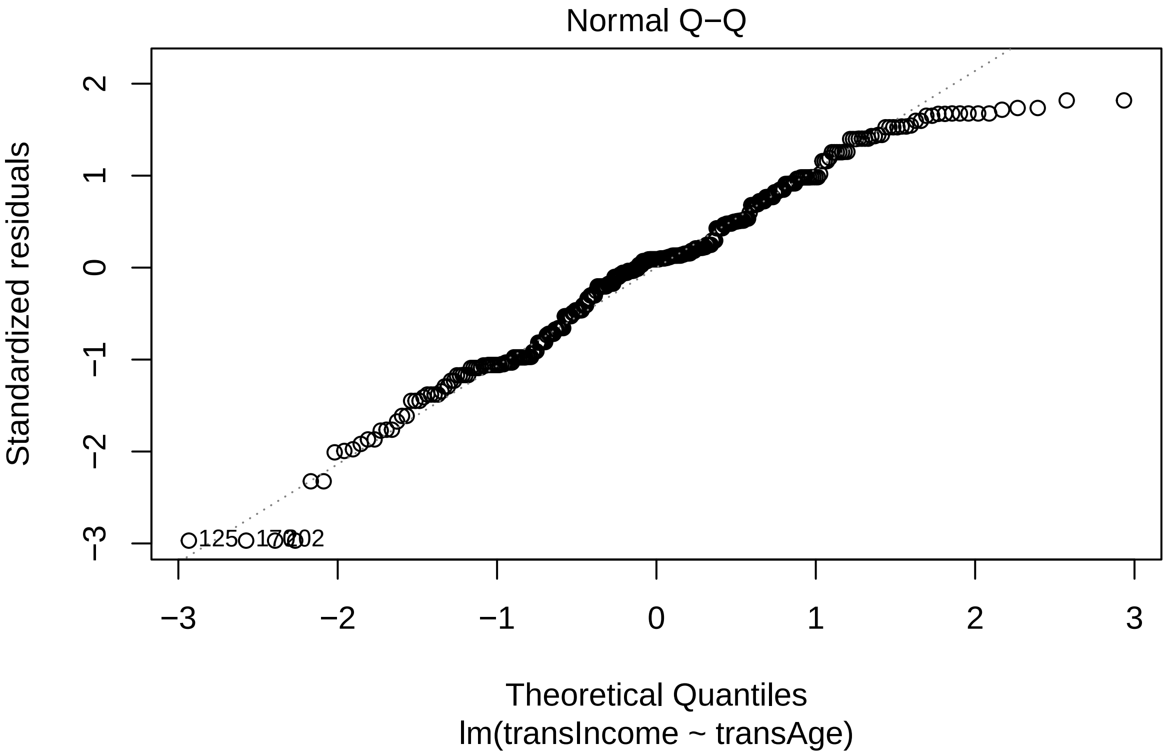
```
nyc$transIncome <- (nyc$Income)^(1/3)
nyc$transAge <- (nyc$Age)^(1/5)
plot(transIncome ~ transAge, data = nyc,
     main = "Cube Root of Income vs 5th Root of Age in NYC",
     xlab = "5th Root of Age",
     ylab = "Cube Root of Income ($)")
```



```
agelm3 <- lm(transIncome ~ transAge, data = nyc)
plot(agelm3, which = 1)
```



```
plot(agelm3, which = 2)
```



```
summary(agelm3)
```

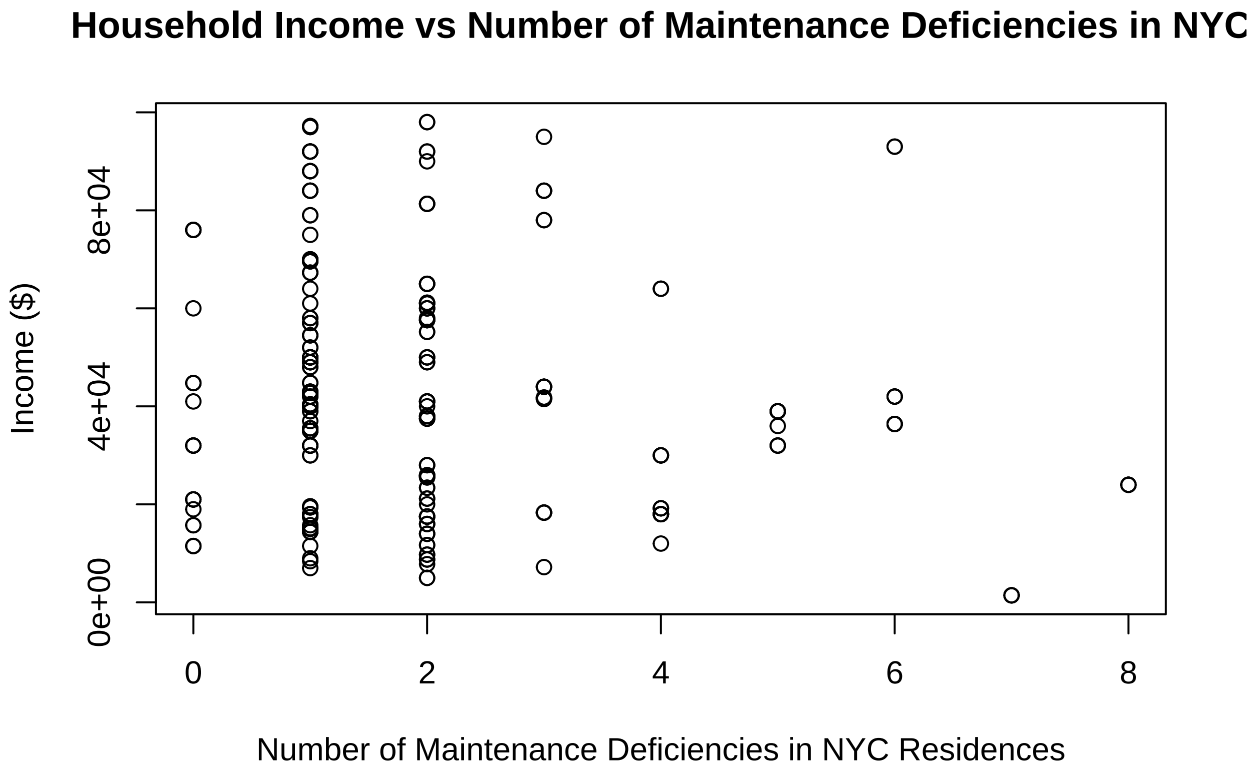
```
##
## Call:
```

```
## lm(formula = transIncome ~ transAge, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2179  -5.1586   0.6516   5.1823  13.0345
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.545      8.401   2.684  0.00769 **
## transAge       4.983      3.855   1.292  0.19722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.191 on 297 degrees of freedom
## Multiple R-squared:  0.005593,   Adjusted R-squared:  0.002244
## F-statistic:  1.67 on 1 and 297 DF,  p-value: 0.1972
```

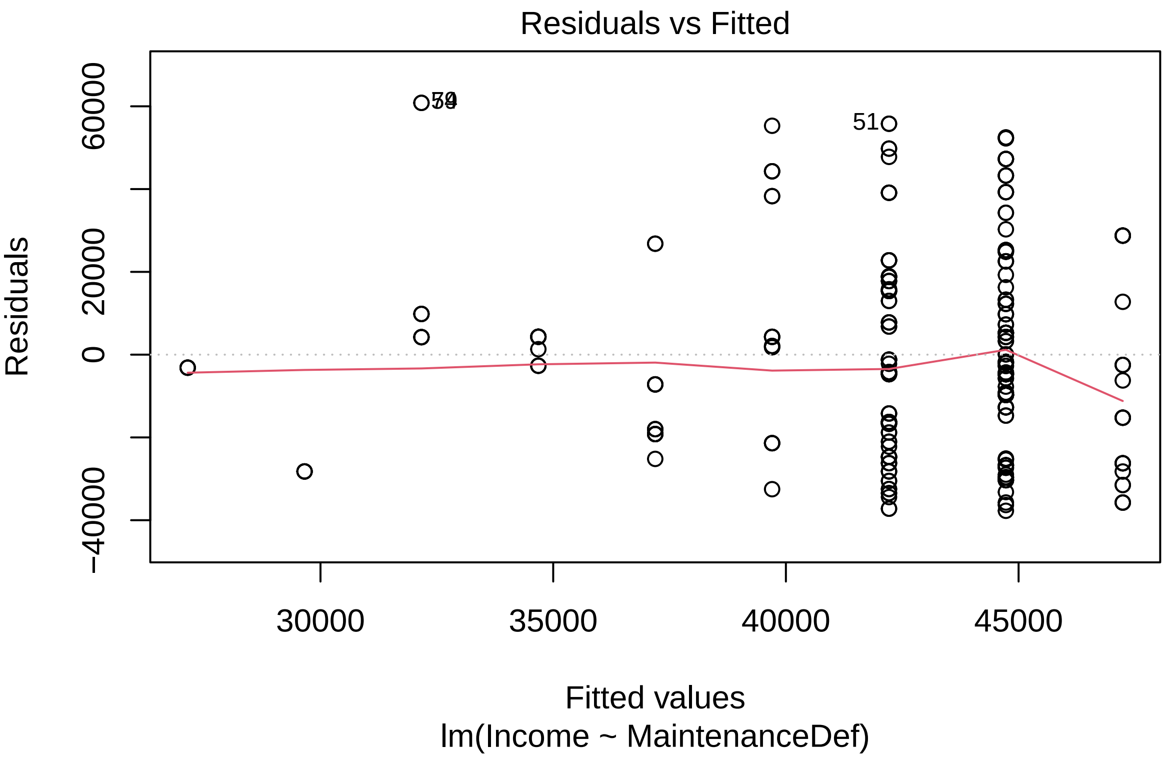
After transforming both the variables `Income` and `Age` with root functions, the model yields a larger coefficient of determination at 0.56% than the linear regression model without any transformation. This means that 0.56% of the variability in the cube root of household income can be explained by the 5th root of age of NYC respondents. This linear regression model also meets the error assumptions of having a constant sigma, a residual mean of roughly 0, a patternless residual plot that establishes independence, and data points that are close to the QQ plot line which establishes normality.

**Variable MaintenanceDef**

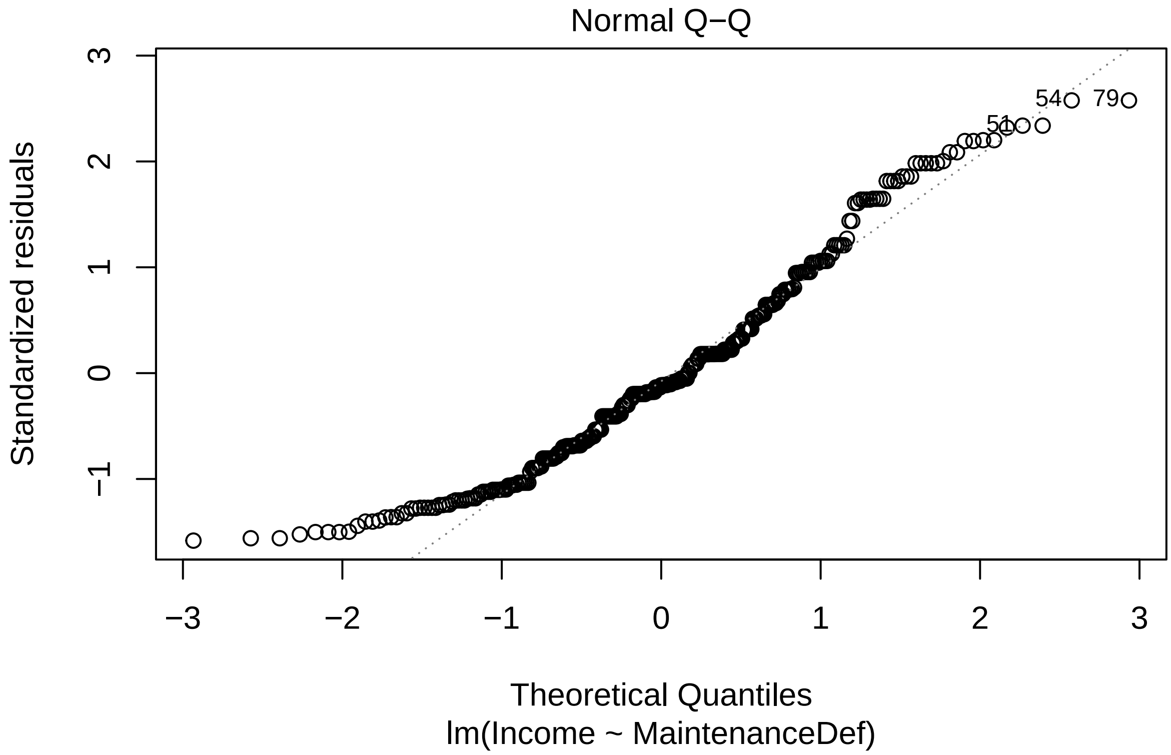
```
plot(Income ~ MaintenanceDef, data = nyc,
     main = "Household Income vs Number of Maintenance Deficiencies in NYC",
     xlab = "Number of Maintenance Deficiencies in NYC Residences",
     ylab = "Income ($)")
```



```
MainDeflm1 <- lm(Income ~ MaintenanceDef, data = nyc)
plot(MainDeflm1, which = 1)
```



```
plot(MainDeflm1, which = 2)
```



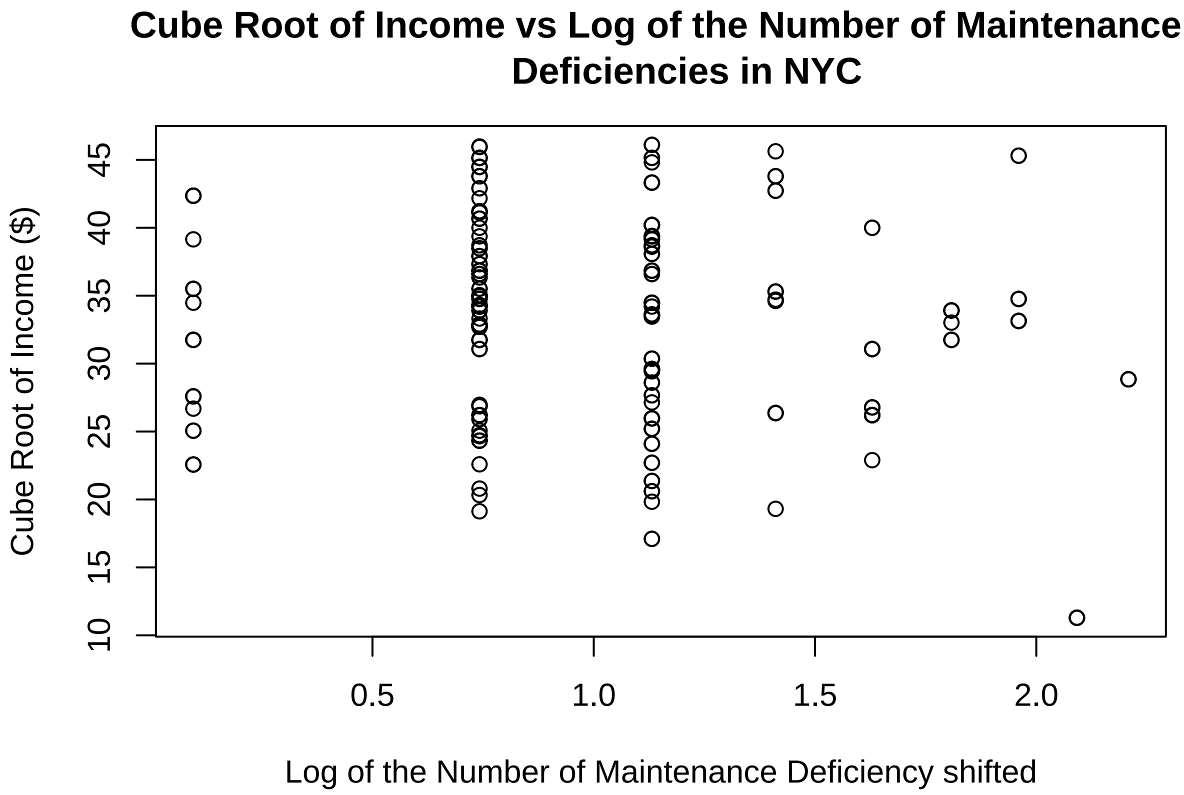
```
summary(MainDeflm1)
```

```
##
## Call:
```

```
## lm(formula = Income ~ MaintenanceDef, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37727 -19004  -2727  15385  60831
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     47238.6     2184.7  21.622  < 2e-16 ***
## MaintenanceDef  -2511.6      854.6  -2.939  0.00355 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23900 on 297 degrees of freedom
## Multiple R-squared:  0.02826,    Adjusted R-squared:  0.02499
## F-statistic: 8.637 on 1 and 297 DF,  p-value: 0.003553
```

Since the scatterplot between response variable `Income` and explanatory variable number of maintenance deficiencies of respondents' residences between 2002-2005 appears to have a non linear regression relationship, a residual distribution that is not normally distributed, we can transform the variables involved in this scatterplot. The coefficient of determination is 2.83%, meaning the 2.83% of variability in the household income can be explained by the number of maintenance deficiencies in NYC respondents' residences between years 2002 and 2005.

```
MainDefShift <- nyc$MaintenanceDef + 1.1
nyc$logMainDef <- log(MainDefShift)
plot(transIncome ~ logMainDef, data = nyc,
     main = "Cube Root of Income vs Log of the Number of Maintenance
     Deficiencies in NYC",
     xlab = "Log of the Number of Maintenance Deficiency shifted",
     ylab = "Cube Root of Income ($)")
```

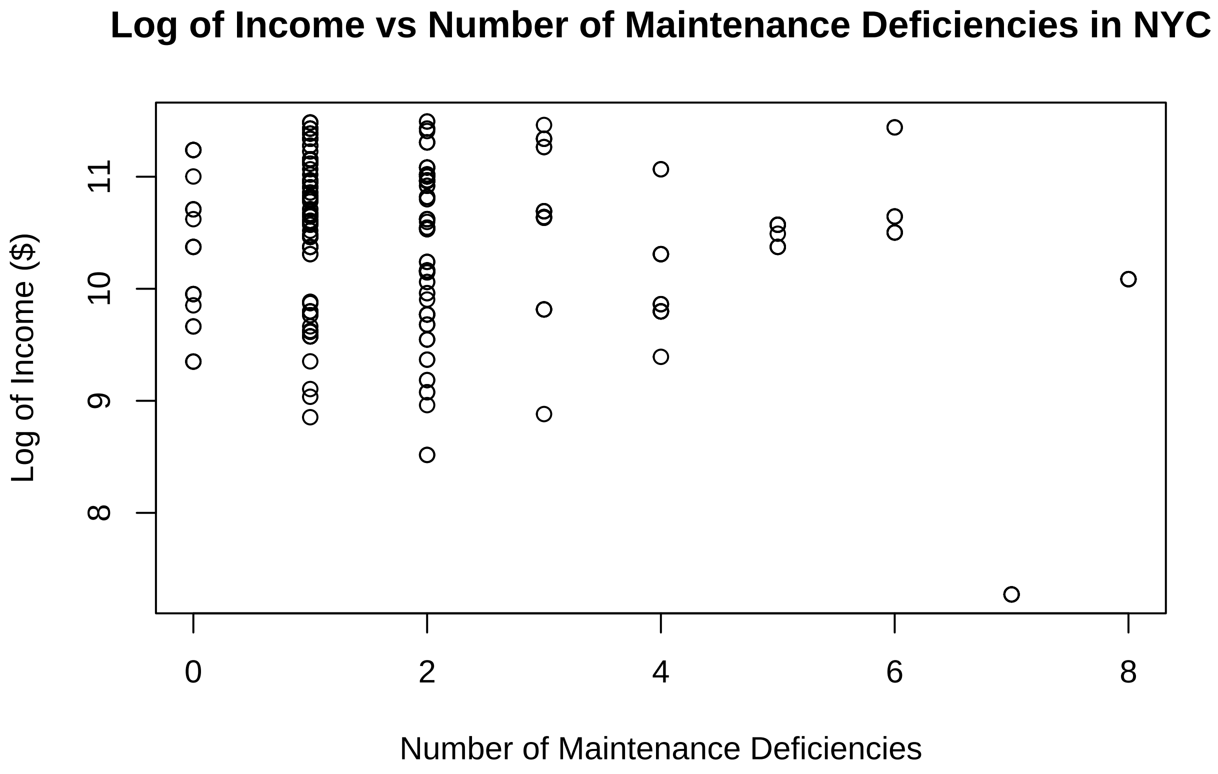


```r
MainDeflm2 <- lm(transIncome ~ logMainDef, data = nyc)
summary(MainDeflm2)
```

```
##
## Call:
## lm(formula = transIncome ~ logMainDef, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4021  -5.2466   0.3817   5.5296  14.2834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.9118     0.9856  36.436  < 2e-16 ***
## logMainDef   -2.4941     0.8853  -2.817  0.00517 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.117 on 297 degrees of freedom
## Multiple R-squared:  0.02603,    Adjusted R-squared:  0.02275
## F-statistic: 7.936 on 1 and 297 DF,  p-value: 0.005171
```

Due to the skewness to the right of the individual distribution of `Income`, we used the cube root transformation again. And we used the logarithmic function for transformation on the variable number of maintenance deficiencies to establish normality. It has a higher coefficient of determination at 2.60%, meaning 2.60% of the variability in the household income can be explained by the log of the number of maintenance deficiencies in the NYC respondents' household between 2002 and 2005. The residual mean is roughly 0, normality can be established based on the QQ plot, independence can also be established because the residual plot is somewhat patternless, and it has a constant spread. However, a better model can be implemented.

```
plot(logIncome ~ MaintenanceDef, data = nyc,
     main = "Log of Income vs Number of Maintenance Deficiencies in NYC",
     xlab = "Number of Maintenance Deficiencies",
     ylab = "Log of Income ($)")
```

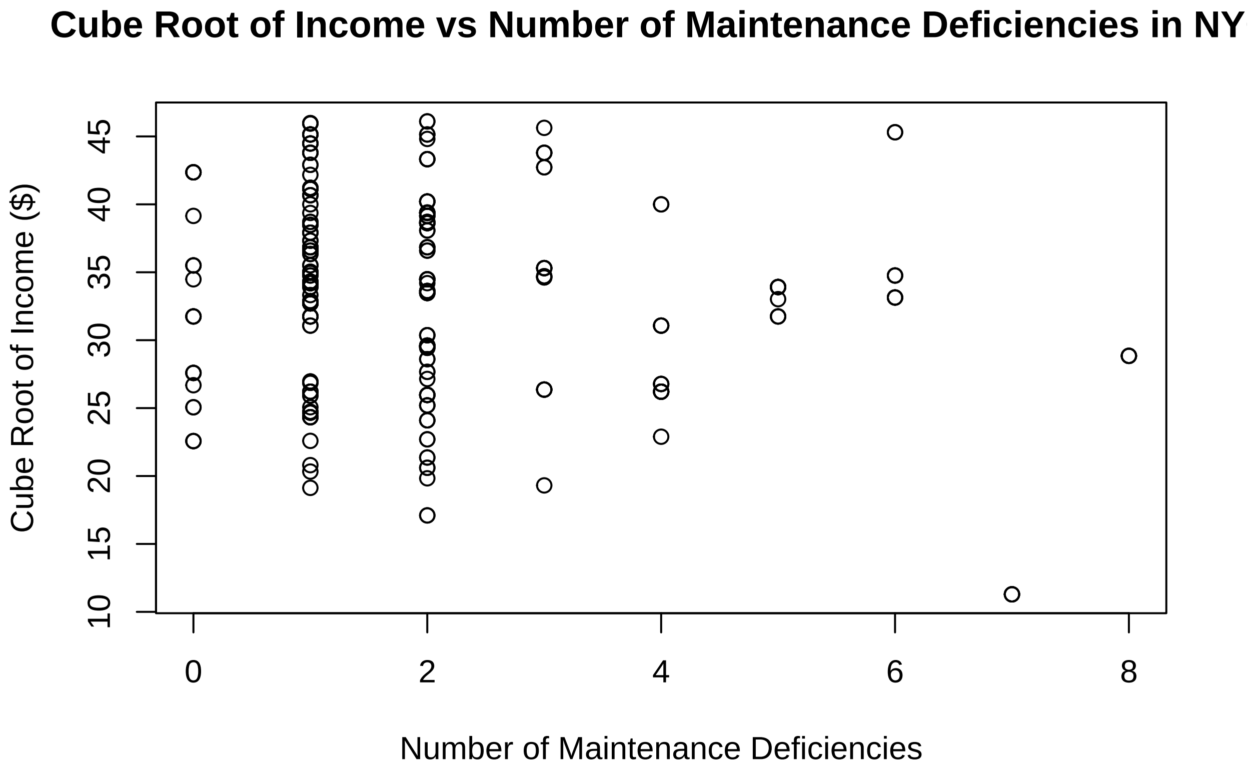


```
MainDeflm3 <- lm(logIncome ~ MaintenanceDef, data = nyc)
summary(MainDeflm3)
```

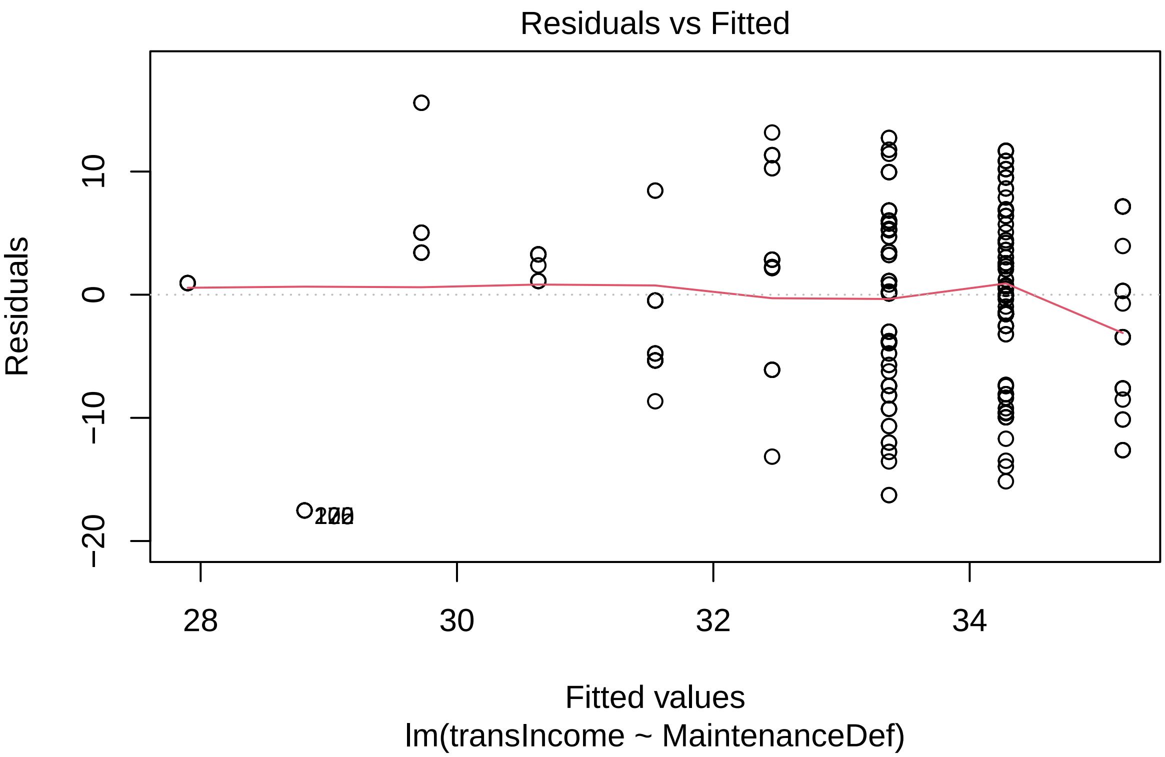
```
##
## Call:
## lm(formula = logIncome ~ MaintenanceDef, data = nyc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6267 -0.4042  0.1033  0.5192  1.4327
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.65927    0.06489 164.261  < 2e-16 ***
## MaintenanceDef -0.10860    0.02538  -4.278 2.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7098 on 297 degrees of freedom
## Multiple R-squared:  0.05805,    Adjusted R-squared:  0.05487
## F-statistic:  18.3 on 1 and 297 DF,  p-value: 2.545e-05
```

Due to the skewness to the right of the individual distribution of `Income`, we used the logarithmic transformation for `Income`. It has a higher coefficient of determination at 5.81%, meaning 5.81% of the variability in the log of household income can be explained by the number of maintenance deficiencies in the NYC respondents' household between 2002 and 2005. Despite having a higher coefficient of determination, normality cannot be established as the data points deviate from the QQ plot line too much.

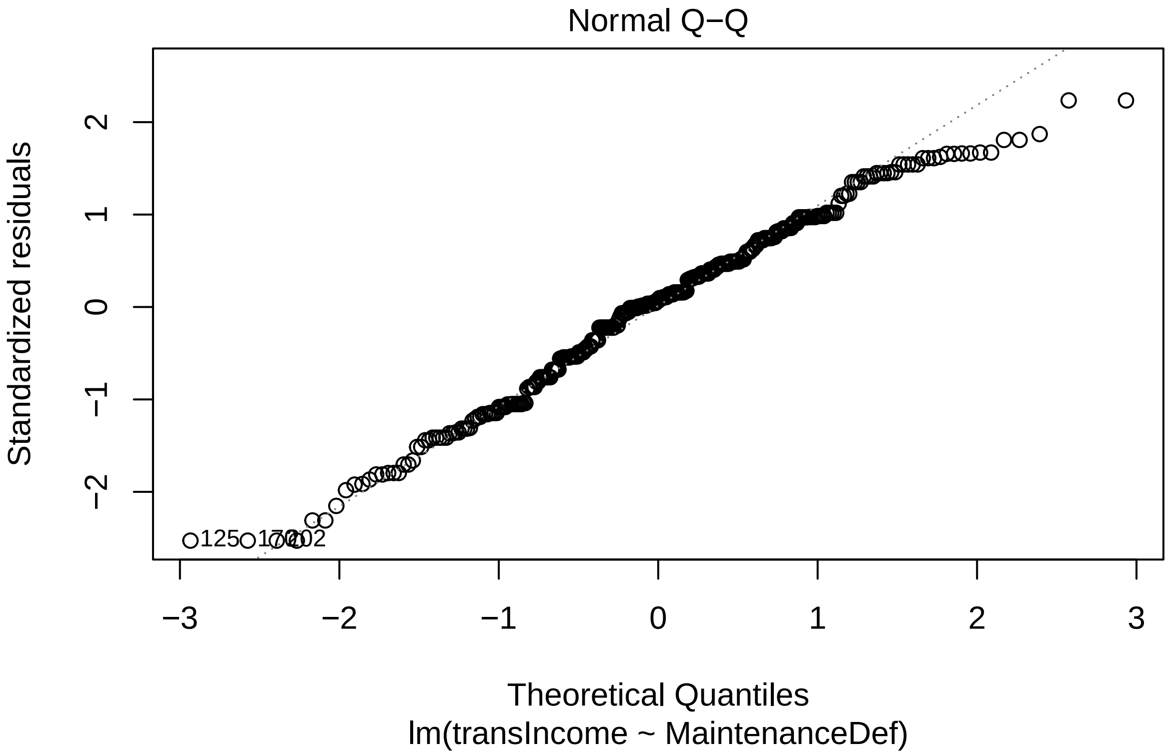
```
plot(transIncome ~ MaintenanceDef, data = nyc,
     main = "Cube Root of Income vs Number of Maintenance Deficiencies in NYC",
     xlab = "Number of Maintenance Deficiencies",
     ylab = "Cube Root of Income ($)")
```



```
MainDeflm4 <- lm(transIncome ~ MaintenanceDef, data = nyc)
plot(MainDeflm4, which = 1)
```



```
plot(MainDeflm4, which = 2)
```



```
summary(MainDeflm4)
```

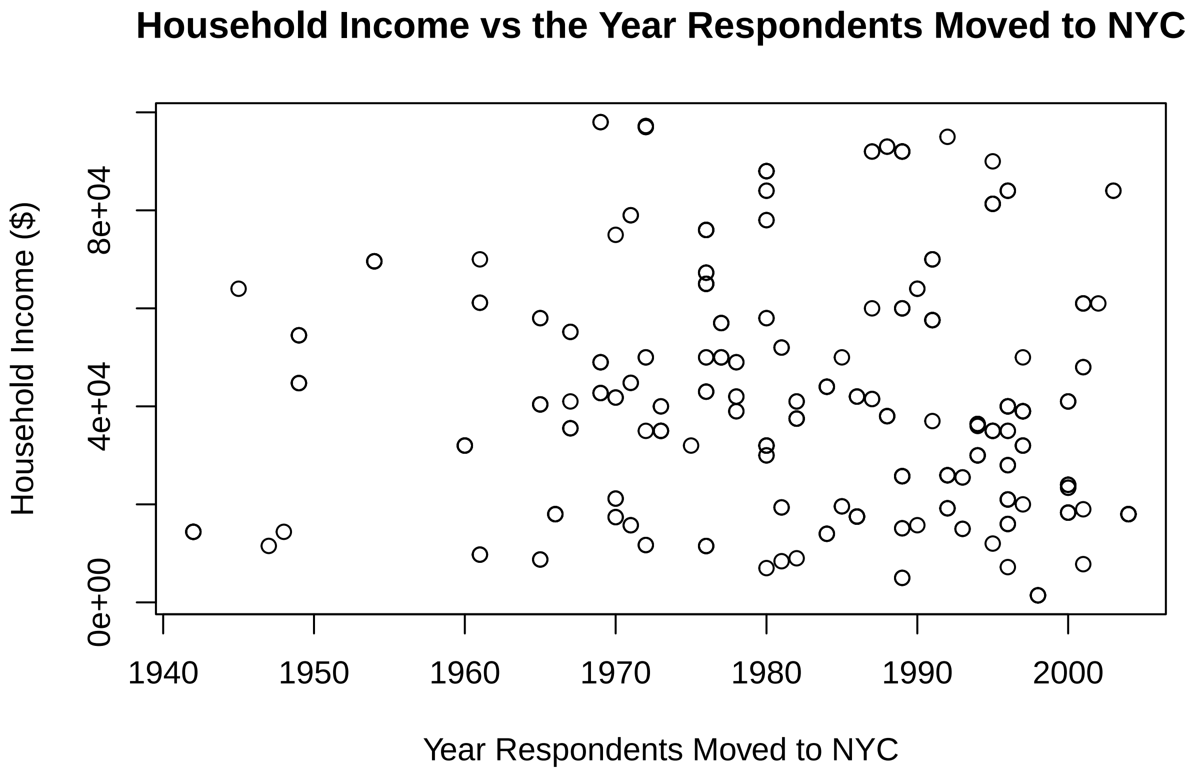
```
##
## Call:
```

```
## lm(formula = transIncome ~ MaintenanceDef, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5184  -5.0544   0.4777   5.2489  15.5837
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.1946     0.6452  54.546  < 2e-16 ***
## MaintenanceDef  -0.9120     0.2524  -3.613 0.000355 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.058 on 297 degrees of freedom
## Multiple R-squared:  0.04211,    Adjusted R-squared:  0.03888
## F-statistic: 13.06 on 1 and 297 DF,  p-value: 0.000355
```

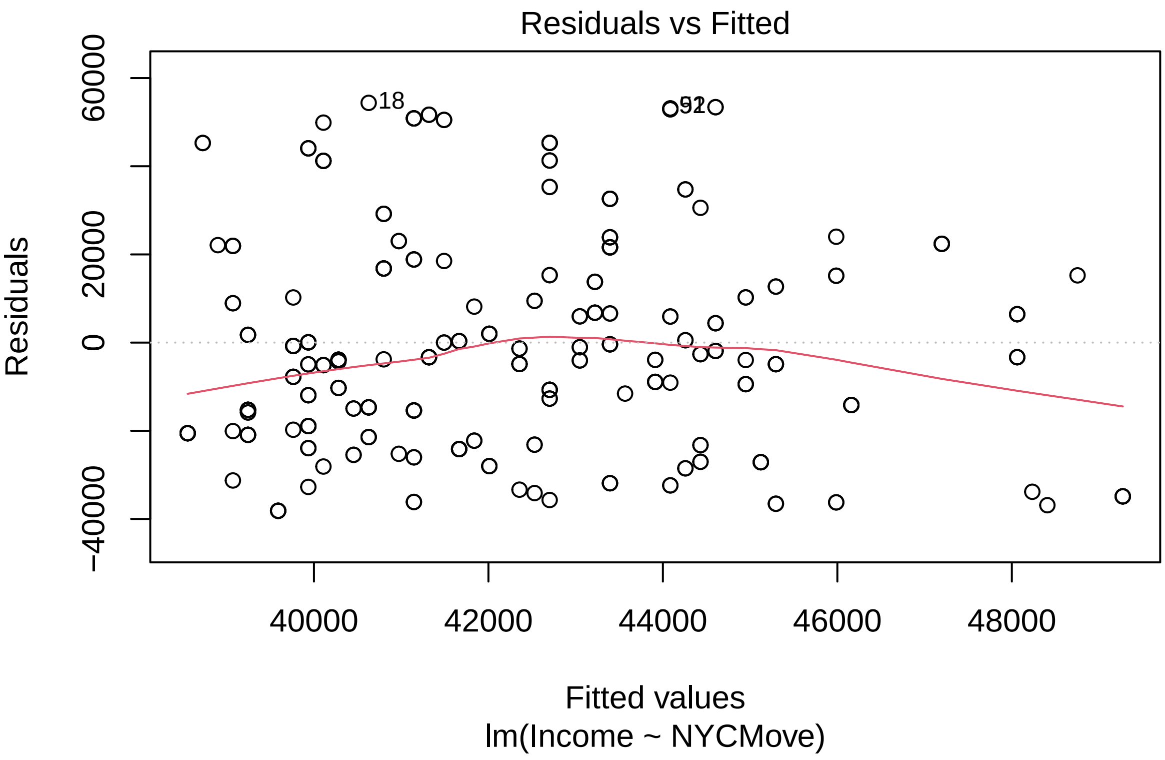
After transforming only variable `Income` with a root function, the model yields a larger coefficient of determination at 4.21% than the orginal linear regression model without any transformation. This means that 4.21% of the variability in the cube root of household income can be explained by the number of maintenance deficiencies in NYC respondents' households between 2002 and 2005. This linear regression model also meets the error assumptions of having a constant sigma, a residual mean of roughly 0, a patternless residual plot that establishes independence, and data points that are close to the QQ plot line that establishes normality. While exploring the bivariate relationship between `Income` and the number of maintenance deficiency in NYC respondents' households between 2002 and 2005, the root function wasn't used on the number of maintenance deficiency because the logarithmic transformation creates a better coefficient of determination value overall.

### Variable NYCMove

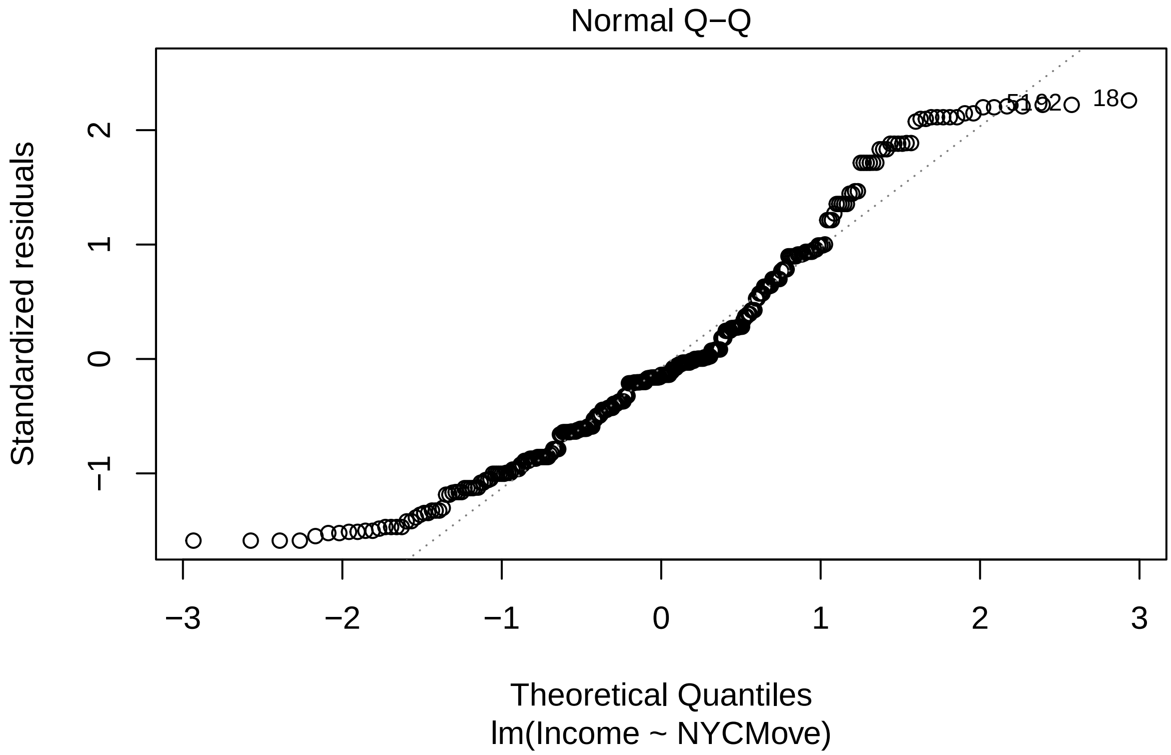
```
plot(Income ~ NYCMove, data = nyc,
     main = "Household Income vs the Year Respondents Moved to NYC",
     xlab = "Year Respondents Moved to NYC",
     ylab = "Household Income ($)")
```



```
NYCMovelm1 <- lm(Income ~ NYCMove, data = nyc)
plot(NYCMovelm1, which = 1)
```



```r
plot(NYCMovelm1, which = 2)
```



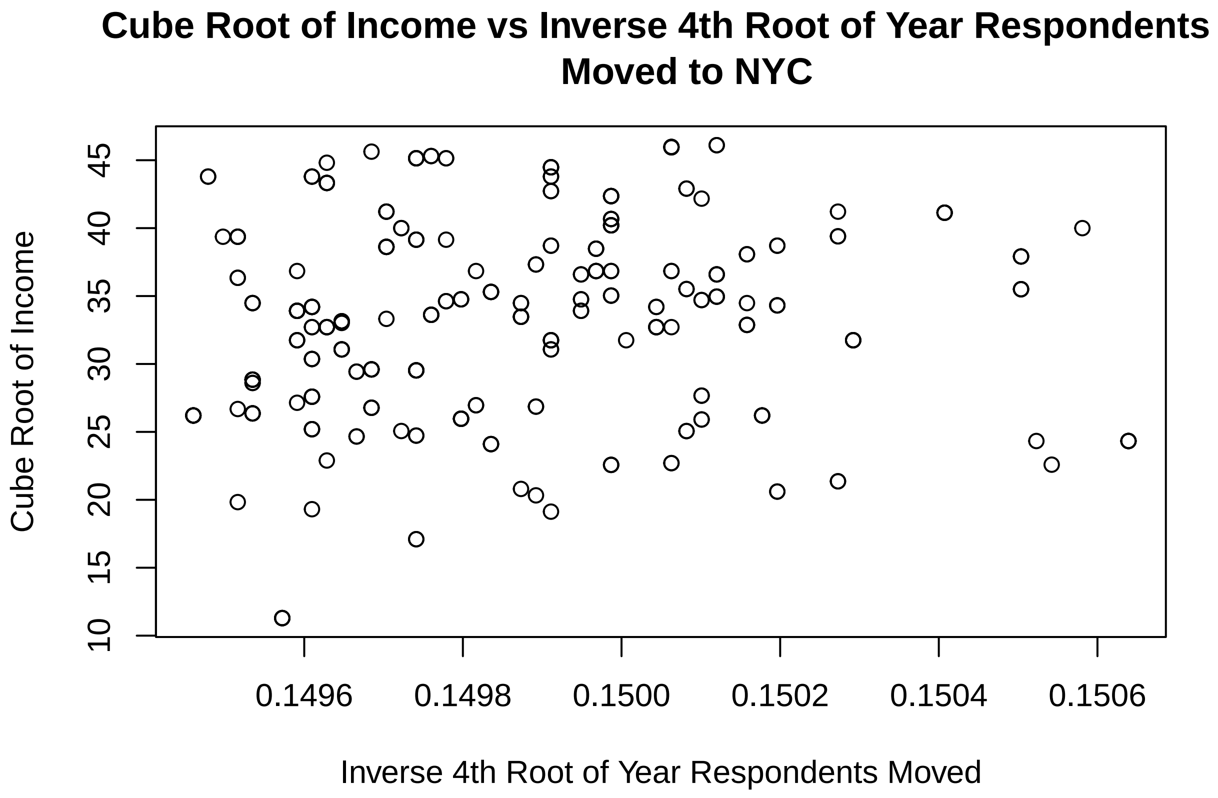
```r
summary(NYCMovelm1)
```

```
##
## Call:
```

```
## lm(formula = Income ~ NYCMove, data = nyc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38150 -18936  -3319  15272  54373
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 385030.64  195923.67   1.965   0.0503 .
## NYCMove       -172.89      98.82  -1.750   0.0812 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24120 on 297 degrees of freedom
## Multiple R-squared:  0.0102, Adjusted R-squared:  0.006868
## F-statistic: 3.061 on 1 and 297 DF,  p-value: 0.08123
```

Since the scatterplot between response variable `Income` and explanatory variable year respondents moved to NYC appears to have a very weak correlation (or none at all), a residual mean that doesn't equal to 0, and a residual distribution that is not normally distributed, we can transform the variables involved in this scatterplot. The coefficient of determination is low at 1.02%, meaning only 1.02% of the variability in the household income can be explained by the year respondents moved to NYC.

```
nyc$transNYCMove <- (nyc$NYCMove)^(-1/4)
plot(transIncome ~ transNYCMove, data = nyc,
     main = "Cube Root of Income vs Inverse 4th Root of Year Respondents
     Moved to NYC",
     xlab = "Inverse 4th Root of Year Respondents Moved",
     ylab = "Cube Root of Income")
```

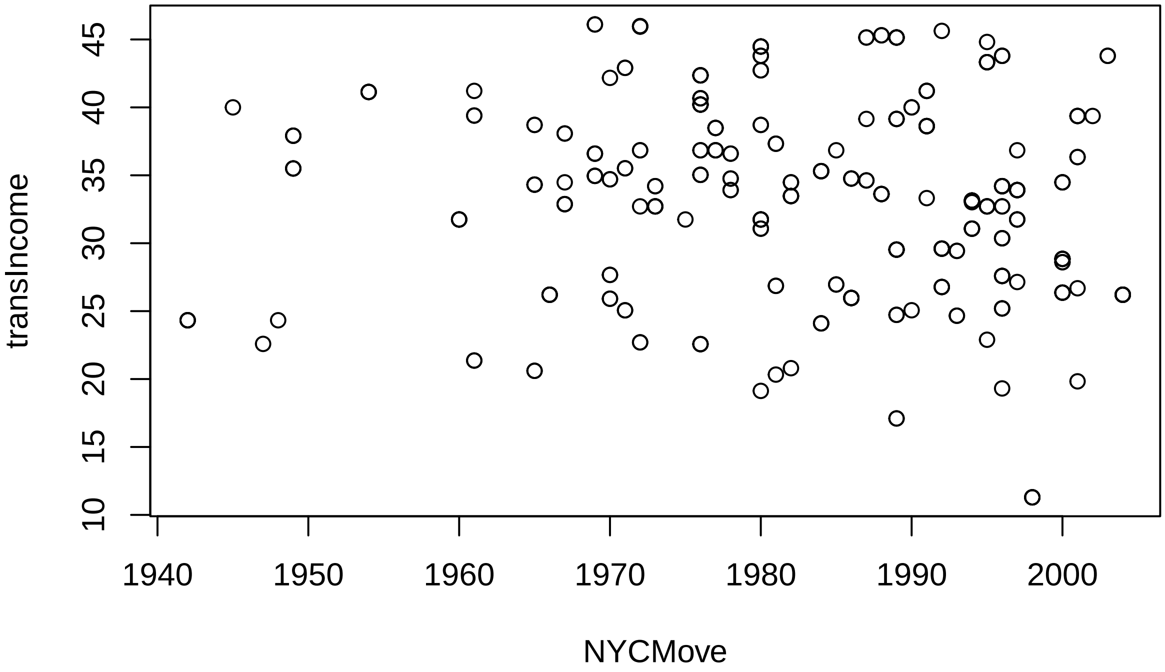


```
NYCMovelm2 <- lm(transIncome ~ transNYCMove, data = nyc)
summary(NYCMovelm2)
```

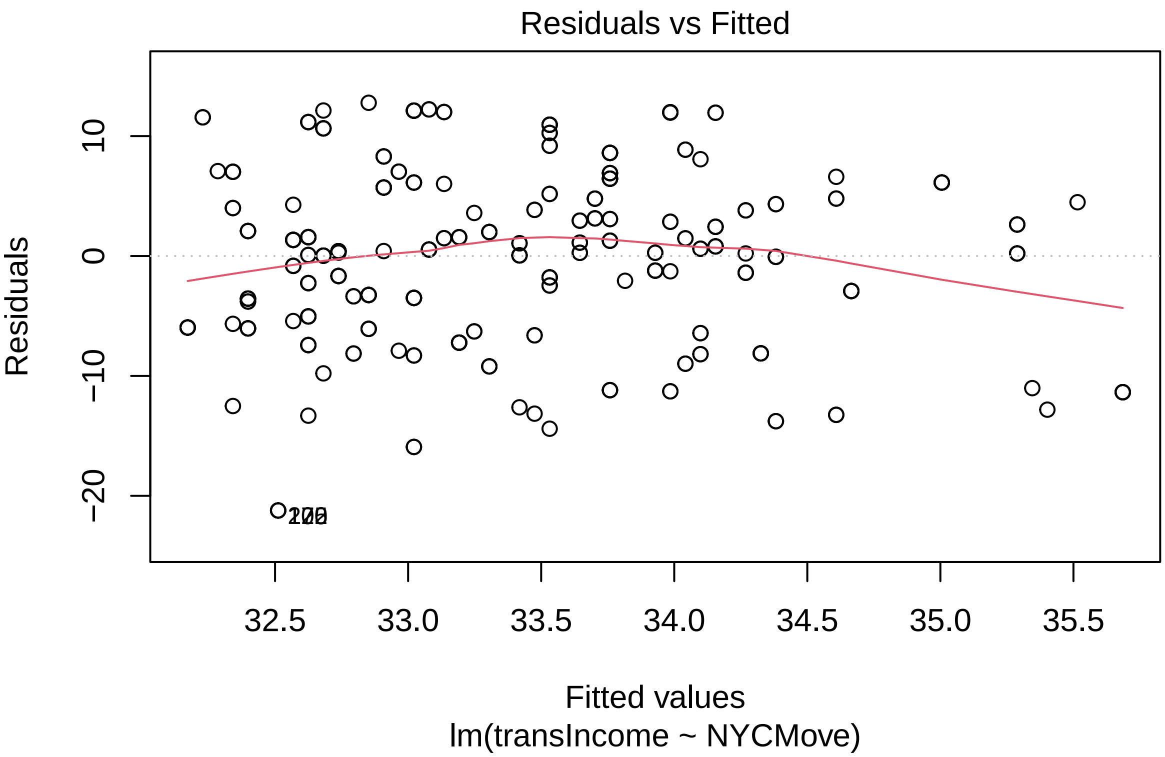
```
##
## Call:
## lm(formula = transIncome ~ transNYCMove, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2322  -5.0461   0.3949   4.9918  12.7717
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -409.8      232.1  -1.765   0.0785 .
## transNYCMove   2957.3     1548.9   1.909   0.0572 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.167 on 297 degrees of freedom
## Multiple R-squared:  0.01212,    Adjusted R-squared:  0.008798
## F-statistic: 3.645 on 1 and 297 DF,  p-value: 0.0572
```

Since the individual distribution of `Income` is skewed to the right, we used the root function to transform `Income` to be normally distributed. And also since the individual distribution of the year respondents moved to NYC is skewed to the left, we transformed the variable by using a inverse root function. The coefficient of determination value is 1.21%, meaning 1.21% of the variability in the cube root of `Income` can be explained by the inverse 4th root of the year when the respondents moved to NYC.

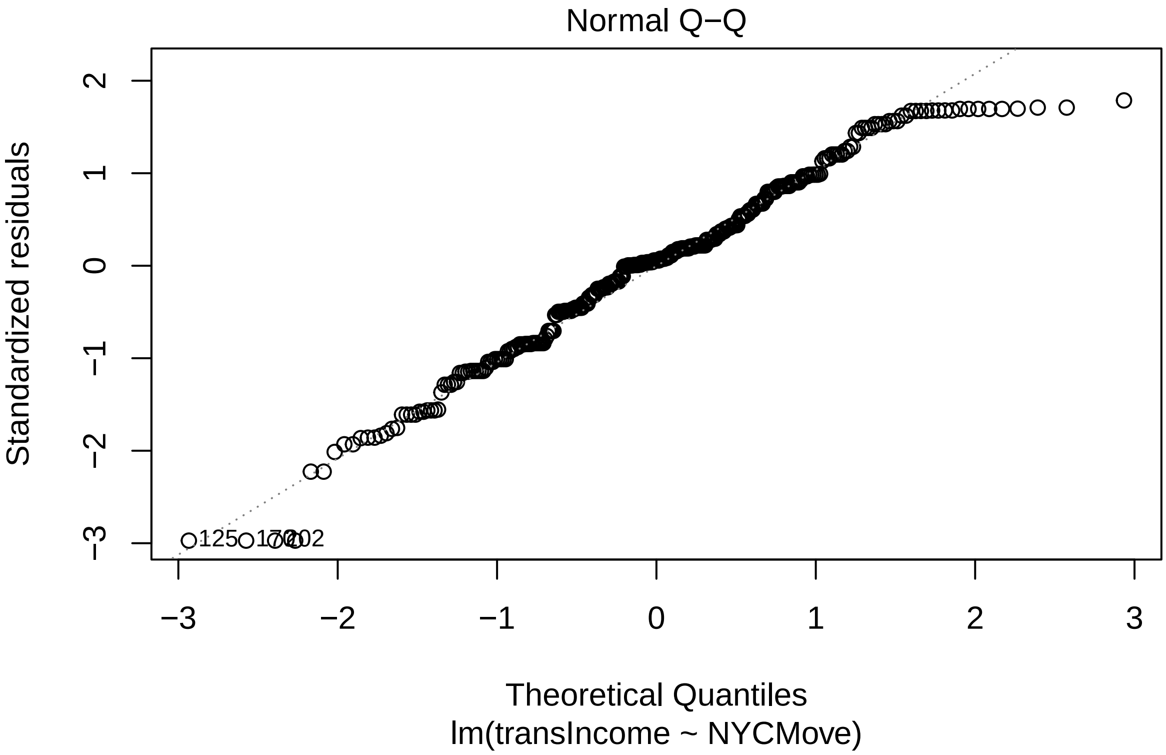
```
plot(transIncome ~ NYCMove, data = nyc)
```



```
NYCMovelm3 <- lm(transIncome ~ NYCMove, data = nyc)
plot(NYCMovelm3, which = 1)
```



```
plot(NYCMovelm3, which = 2)
```



```
summary(NYCMovelm3)
```

```
##
## Call:
```

```
## lm(formula = transIncome ~ NYCMove, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2193  -5.0359   0.4027   4.9830  12.7772
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.74065   58.21523   2.503   0.0128 *
## NYCMove      -0.05667    0.02936  -1.930   0.0546 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.166 on 297 degrees of freedom
## Multiple R-squared:  0.01239,    Adjusted R-squared:  0.009061
## F-statistic: 3.725 on 1 and 297 DF,  p-value: 0.05456
```

After transforming only variable `Income` with a root function, the model yields a similar coefficient of determination at 1.24% than the linear regression model with the cube root of `Income` and inverse 4th root of the year the respondents moved to NYC. This means that 1.24% of the variability in the cube root of household income can be explained by the year the NYC respondents moved to NYC. So we should choose the untransformed version of the variable the year respondents moved to NYC to keep the model simple. This linear regression model also meets most of the error assumptions of having a constant sigma, a patternless residual plot that establishes independence, and data points that are close to the QQ plot line that establishes normality. If we were to use a pure root or logarithmic function to transform the variable the year the respondents moved to NYC, the distribution of the variable the year the respondents moved to NYC would be even more skewed. If we were to use an exponential function after shifting the variable the year respondents moved to NYC (to reduce the numeric value size of the year respondents moved to NYC), the distribution of that variable will become skewed to the right. So keeping the variable `NYCMove`, the year respondents moved to NYC, unchanged will be the best.

# Modeling

To determine which explanatory variables to use in the final linear regression model for predicting household income, the below correlation matrix shows the correlation values.

```
nyc.ordered <- subset(nyc, select = c(Income, logIncome, transIncome, Age, logAge, transAge, Maintenanc
round(cor(nyc.ordered), digits = 2)
```

```
##                Income logIncome transIncome   Age logAge transAge
## Income           1.00      0.90        0.96  0.04   0.08     0.07
## logIncome        0.90      1.00        0.98  0.05   0.10     0.09
## transIncome      0.96      0.98        1.00  0.04   0.08     0.07
## Age              0.04      0.05        0.04  1.00   0.99     0.99
## logAge           0.08      0.10        0.08  0.99   1.00     1.00
## transAge         0.07      0.09        0.07  0.99   1.00     1.00
## MaintenanceDef  -0.17     -0.24       -0.21 -0.25  -0.24    -0.24
## logMainDef      -0.13     -0.19       -0.16 -0.25  -0.24    -0.25
## NYCMove         -0.10     -0.12       -0.11 -0.64  -0.61    -0.61
## transNYCMove     0.10      0.12        0.11  0.64   0.61     0.61
##                MaintenanceDef logMainDef NYCMove transNYCMove
## Income                  -0.17      -0.13   -0.10         0.10
## logIncome               -0.24      -0.19   -0.12         0.12
## transIncome             -0.21      -0.16   -0.11         0.11
## Age                     -0.25      -0.25   -0.64         0.64
## logAge                  -0.24      -0.24   -0.61         0.61
## transAge                -0.24      -0.25   -0.61         0.61
## MaintenanceDef           1.00       0.95    0.46        -0.46
## logMainDef               0.95       1.00    0.46        -0.45
## NYCMove                  0.46       0.46    1.00        -1.00
## transNYCMove            -0.46      -0.45   -1.00         1.00
```

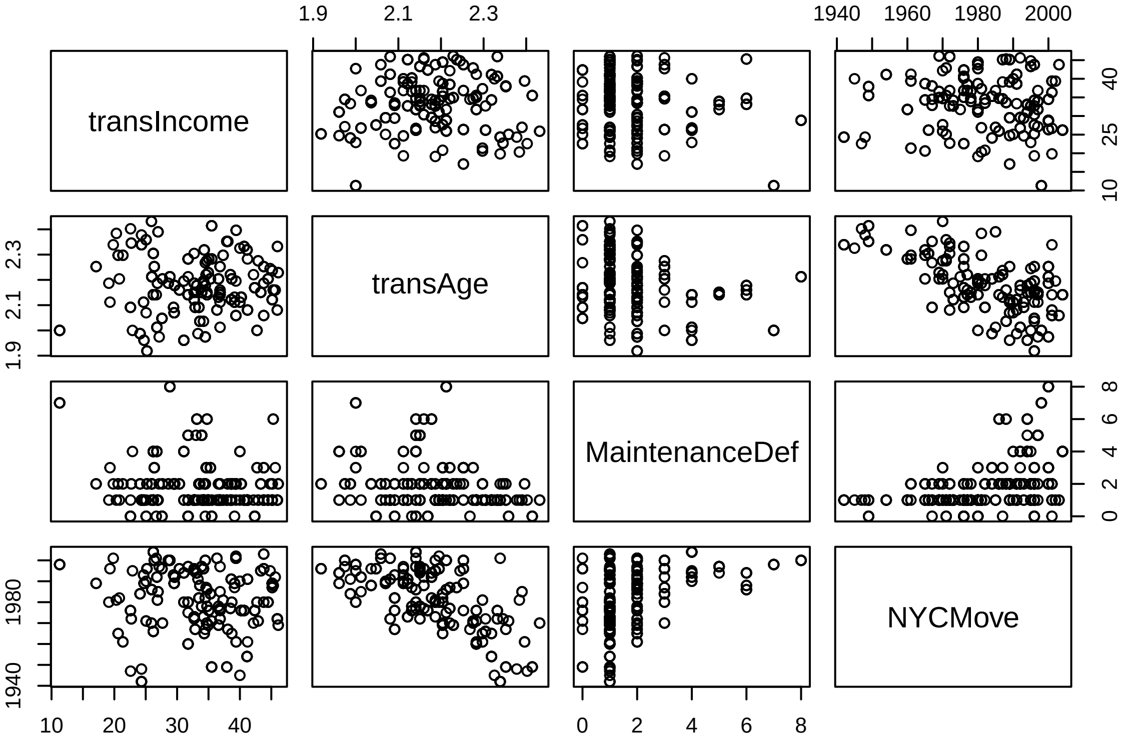
Looking across the 3 household income variables, not transforming the `Income` variable yields to lowest correlation overall. But although using the logarithmic transformation for the `Income` variable may yield the highest correlation coefficients, many of the models that used the log of `Income` led to violations of the normality error assumption. So we should use the variable labled `transIncome`, which is the cube root of the household income, as the response variable for the final model.

Across the response variable the cube root of household income of NYC respondents, the log of `Age` (labled `logAge`) has the highest correlation among all the `Age` explanatory variables, but it violates the normality assumption. So we will use the transformed version of variable `Age` which is the 5th root; the original number of Maintenance Deficiency (labled `MaintenanceDef`) has the highest correlation among all the `Maintenance Deficiency` explanatory variables; the original year the respondents moved to NYC (labled `NYCMove`) has the highest correlation among the 2 explanatory variables that describe the year respondents moved to NYC.

So the chosen variables are the cube root of `Income`, the 5th root of `Age`, the original number of `Maintenance Deficiencies` in NYC residences between the years 2002 and 2005, and the original year the respondents moved to NYC.

Below is the pairs plot of the selected variables to illustrate the relationship between all 4 variables concisely.

```
nyc.transformed <- subset(nyc, select = c(transIncome, transAge, MaintenanceDef, NYCMove))
pairs(nyc.transformed)
```



## Dangerous Multicollinearity

Based on the pairs plot, the correlation between the 3 explanatory variables doesn't seem to be strong. But to make sure that the data with the selected variables doesn't have dangerous multicollinearity, the vif value must be checked.

```
nyclm <- lm(transIncome ~ transAge + MaintenanceDef + NYCMove, data = nyc)
car::vif(nyclm)
```

```
##      transAge MaintenanceDef        NYCMove
##      1.610464       1.267024       1.916164
```

The above vif values do not exceed the 2.5 value, meaning there is not a sign for dangerous multicollinearity and we should not be concerned between the relationships between the 3 explanatory variables.

## Constructing the Final Linear Regression Model for Estimating Household Income

```
best.nyc.subset <- regsubsets(transIncome ~. , data = nyc.transformed, nvmax = 4)
summary(best.nyc.subset)
```

```
## Subset selection object
## Call: regsubsets.formula(transIncome ~ ., data = nyc.transformed, nvmax = 4)
## 3 Variables  (and intercept)
##                Forced in Forced out
## transAge           FALSE      FALSE
## MaintenanceDef     FALSE      FALSE
## NYCMove            FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          transAge MaintenanceDef NYCMove
## 1  ( 1 ) " "      "*"            " "
## 2  ( 1 ) "*"      "*"            " "
## 3  ( 1 ) "*"      "*"            "*"
```

```
nyclm3 <- lm(transIncome ~ transAge + MaintenanceDef + NYCMove, data = nyc)
summary(nyclm3)
```

```
##
## Call:
## lm(formula = transIncome ~ transAge + MaintenanceDef + NYCMove,
##     data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3911  -4.9985   0.5382   5.2259  15.4690
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    39.678761  85.929440   0.462  0.64459
## transAge        1.517469   4.816191   0.315  0.75293
## MaintenanceDef -0.871787   0.284957  -3.059  0.00242 **
## NYCMove        -0.003968   0.040151  -0.099  0.92135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.079 on 295 degrees of freedom
## Multiple R-squared:  0.04282,    Adjusted R-squared:  0.03309
## F-statistic: 4.399 on 3 and 295 DF,  p-value: 0.004784
```

If we were to use all 3 explanatory variables, the coefficient of determination would be 4.28%, meaning 4.28% of the variability in the cube root of household income can be explained by the 5th root of `Age`, the number of maintenance deficiencies in the NYC residences between 2002 and 2005, and the year the respondents moved to NYC. But since the p-value of the 5th root of `Age` and the year the respondents moved to NYC is greater than any reasonable significance level, we cannot reject the null hypothesis, meaning there is not significant evidence that suggests a relationship between the 5th root of `Age` and the cube root of household income and there is not a relationship between the year respondents moved to NYC and the cube root of household income. So a better linear regression model should be created.

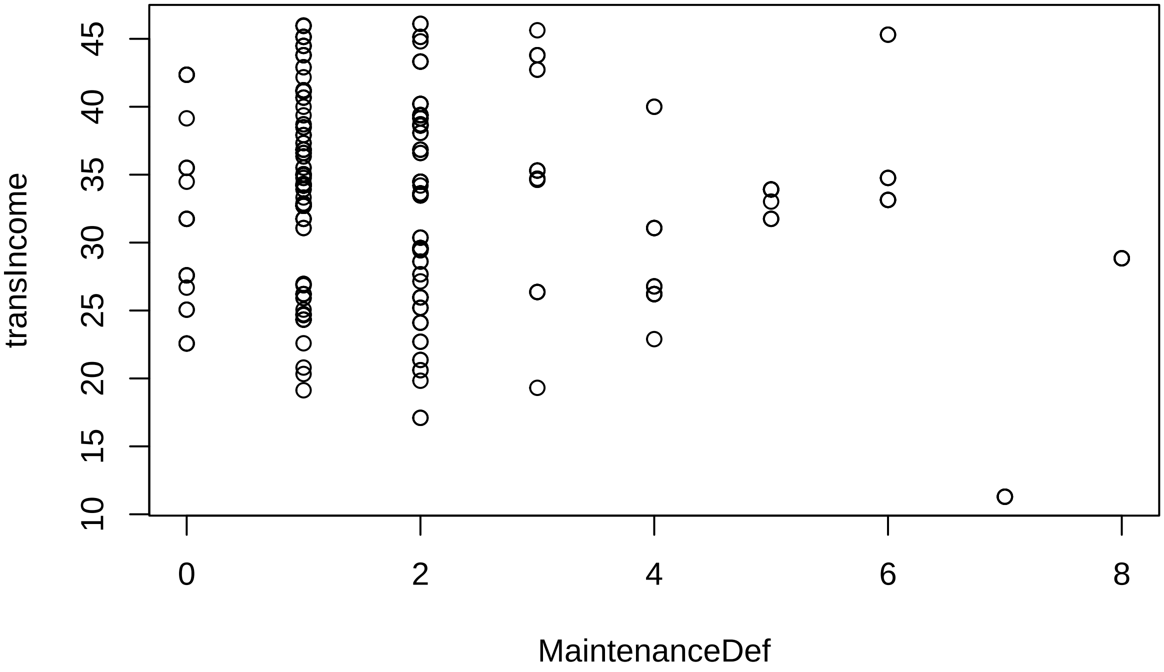
```
nyclm2 <- lm(transIncome ~ transAge + MaintenanceDef, data = nyc)
summary(nyclm2)
```

```
##
## Call:
## lm(formula = transIncome ~ transAge + MaintenanceDef, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3466  -5.0608   0.5522   5.2880  15.4975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     31.2302     8.6438   3.613 0.000356 ***
## transAge         1.7953     3.9036   0.460 0.645911
## MaintenanceDef  -0.8831     0.2604  -3.392 0.000789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

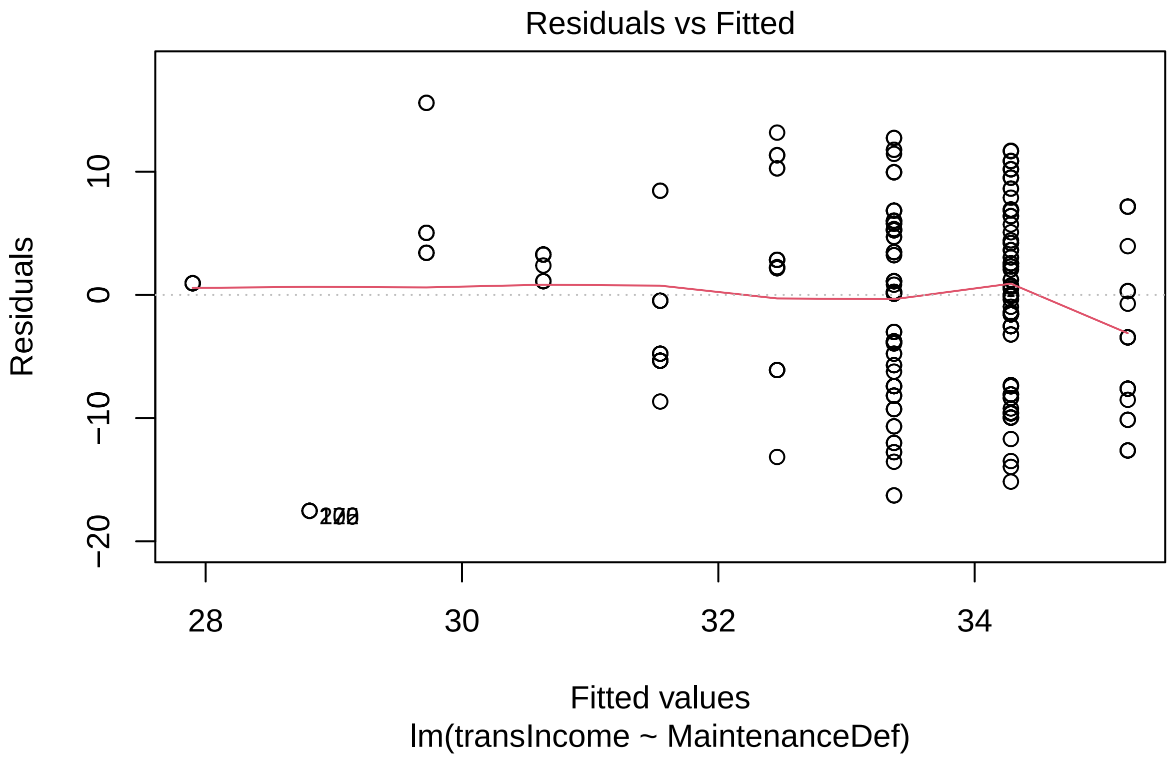
```
##
## Residual standard error: 7.067 on 296 degrees of freedom
## Multiple R-squared:  0.04279,    Adjusted R-squared:  0.03632
## F-statistic: 6.616 on 2 and 296 DF,  p-value: 0.001545
```

If we were to use 2 explanatory variables (the 5th root of `Age` and the number of maintenance deficiencies), the coefficient of determination would be 4.28%, meaning 4.28% of the variability in the cube root of household income can be explained by the 5th root of `Age` and the number of maintenance deficiencies in the NYC residences between 2002 and 2005. But since the p-value of the 5th root of `Age` is greater than any reasonable significance level, we cannot reject the null hypothesis, meaning there is not significance evidence that suggests a relationship between the 5th root of `Age` and the cube root of household income. So a better linear regression model can be used to predict household income of NYC residents.

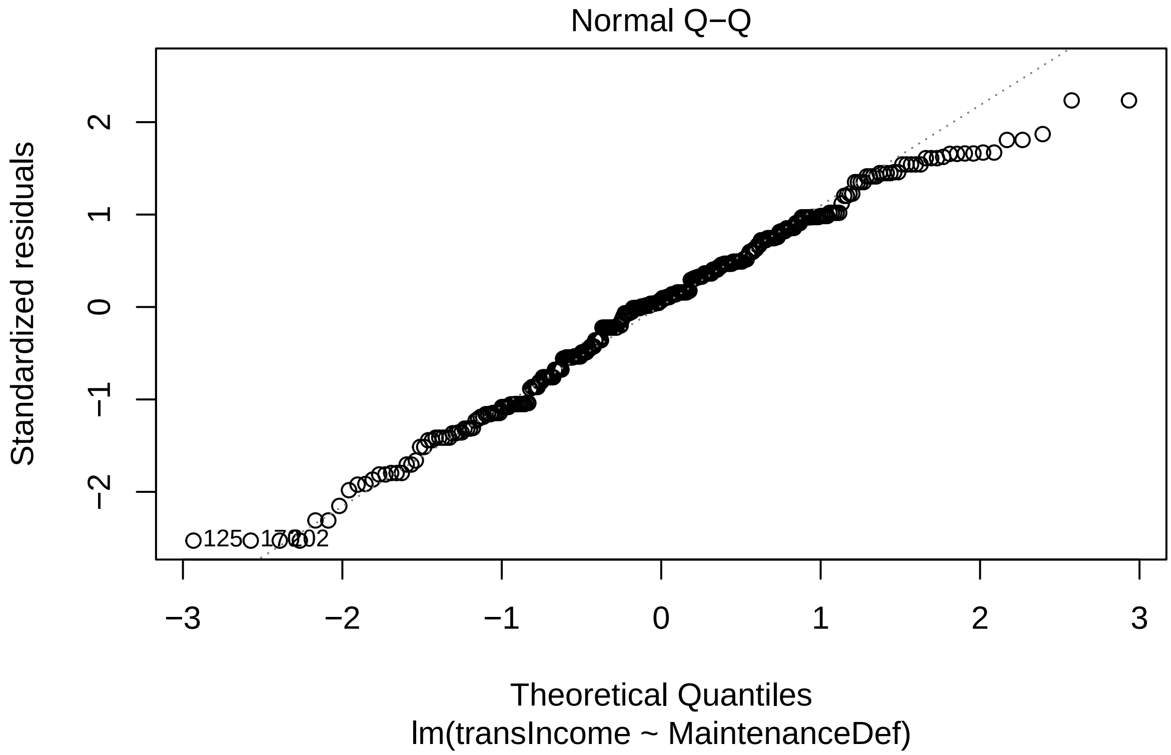
```
plot(transIncome ~ MaintenanceDef, data = nyc)
```



```
nyclm1 <- lm(transIncome ~ MaintenanceDef, data = nyc)
plot(nyclm1, which = 1)
```



```
plot(nyclm1, which = 2)
```



```
summary(nyclm1)
```

```
##
## Call:
```

```
## lm(formula = transIncome ~ MaintenanceDef, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5184  -5.0544   0.4777   5.2489  15.5837
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.1946     0.6452  54.546  < 2e-16 ***
## MaintenanceDef  -0.9120     0.2524  -3.613 0.000355 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.058 on 297 degrees of freedom
## Multiple R-squared:  0.04211,    Adjusted R-squared:  0.03888
## F-statistic: 13.06 on 1 and 297 DF,  p-value: 0.000355
```

If we use only one explanatory variable, the number of maintenance deficiencies in the NYC residences between 2002 and 2005, the coefficient of determination value is not too much smaller than the linear regression models that used more explanatory variables. The coefficient of determination value is similar at 4.21%, meaning 4.21% of the variability in the cube root of household income can be explained by the number of maintenance deficiencies in the NYC residences between 2002 and 2005. The p-value is smaller than a 5% significance level, so we can reject the null hypothesis. This means that there is sufficient evidence that suggests a relationship between the number of maintenance deficiencies and the cube root of household income of NYC residents.

So the final regression model is

$$Income = (\beta_0 + \beta_1(MaintenanceDef))^3 + error$$

. Or the cube root of Income is equal to

$$(\beta_0 + \beta_1(MaintenanceDef)) + error$$

# Prediction

The predicted Income is equal to (35.1946 -0.9120(MaintenanceDef))^3

If a client is interest in predicting income for a household with 3 maintenance deficiencies, whose respondent's age is 53 and who moved to NYC in 1987.

predicted Income = (35.1946 -0.9120*(3))^3 =

```r
(35.1946 -0.9120*(3))^3
```

```
## [1] 34197.11
```

So we predict that a household with 3 maintenance deficiencies in their residence, whose respondent's age is 53 and who moved to NYC in 1987, would have a household income of $34197.11 or have the cube root of household income of $32.46.

# Discussion

While we may have some idea of what a New York City resident's household income is given the number of maintenance deficiencies at their residence, it is important to acknowledge that the this regression model is not reliable because of its low coefficient of determination value. Despite having the highest coefficient of determination among all the other regression models that was tried, only 4.21% of the variability in the log of household income can be explained by the number of maintenance deficiency in the residence, which is extremely low. And even though it meets all the error assumptions of having a residual mean of roughly 0, a constant spread, patternless residual plot, and data points close to the QQ plot line, the model's linear relationship is still very weak.

Further more, a higher-order regression line may be more appropriate for predicting the household income of the NYC residents based on their age, the number of maintenance deficiencies in their residences between 2002 and 2005, and the year they moved to NYC. To improve our prediction of NYC residents' household income, we could, firstly, use more recent data to present data that is more current, and, secondly, use other explanatory variables, like years of education, birth place (categorical variable that explores whether respondents are born in the US or not), monthly rent amount for current residences, etc. So there, certainly, is a way to further improve the current linear regression model to predict NYC residents' household income.