Predicting Hosuehold Income of NYC Residents From Housing Survey Data

Colin Yip
cdyip

Due Wed, March 16, at 8:00PM

Contents

Introduction

The cost of housing in New York City is among the most expensive in the United States. Therefore, we speculate that housing data will be a useful tool to predict the income of New York City residents. In this paper, we predict the household incomes of survey participants from the respondent's age, the number of maintenance deficiencies over a three year period, and the year the respondent moved to NYC.

Exploratory Data Analysis

Data

The New York City Housing and Vacancy Survey gathered data from a random sample of 299 New York City households. We analyzed the relationship between three explanatory variables (age, maintenance deficiences, and years in NYC) and the response variable (household income). The variables are summarized below.

Age: the age of the respondent (in years)

Maintenance Deficiencies: number of maintenance deficiencies of the residence, between 2002 and 2005

Year Moved To NYC: the year the respondent moved to NYC

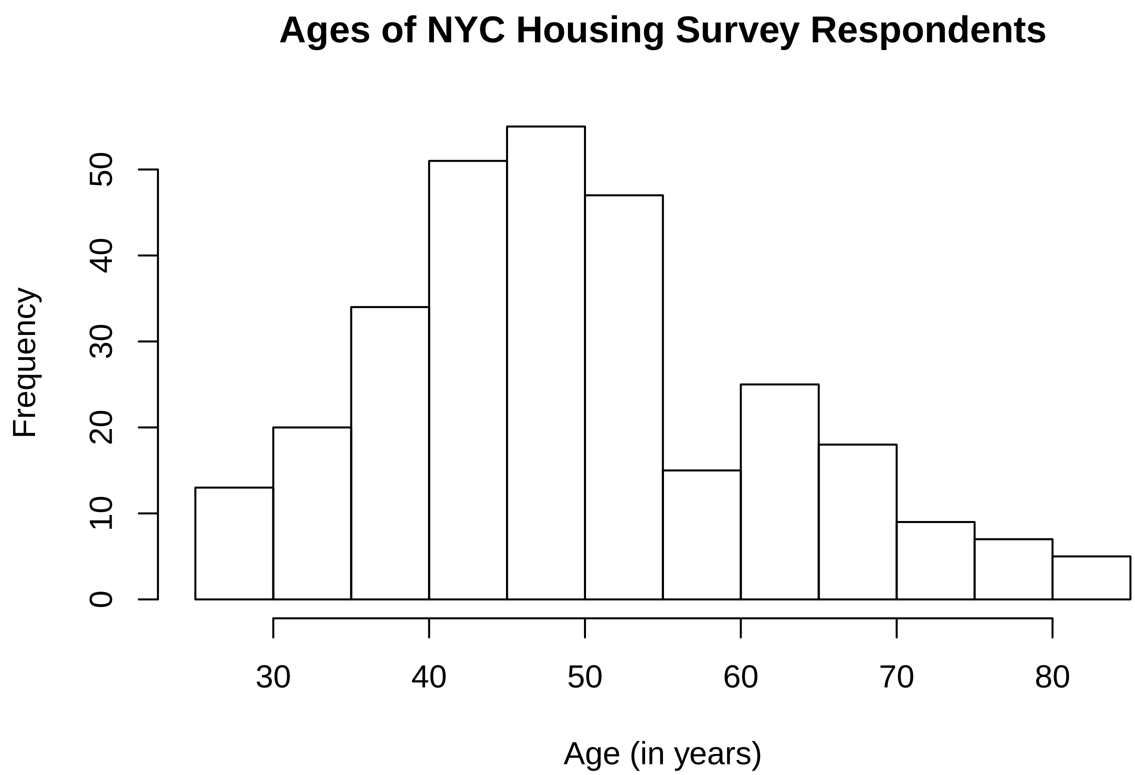
Income: Total household income (in $)

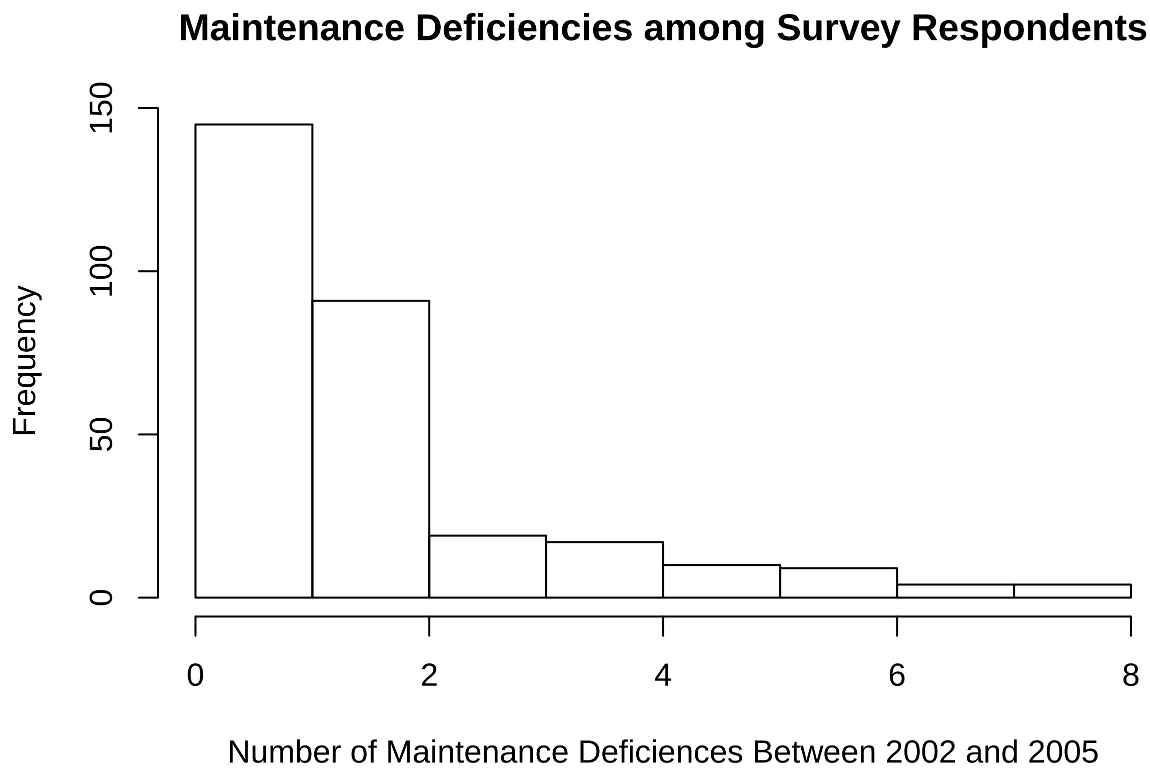
The first ten lines of data are displayed below.

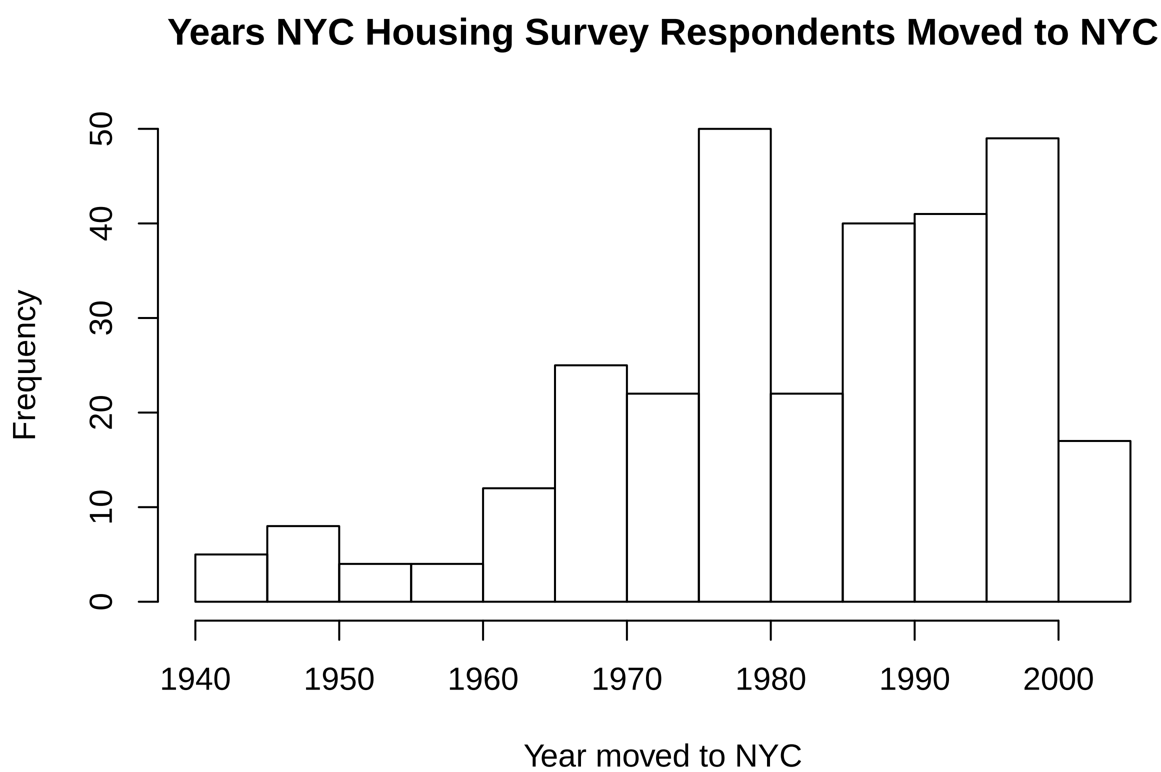
```
## # A tibble: 10 x 4
##    Income   Age MaintenanceDef NYCMove
##     <dbl> <dbl>          <dbl>   <dbl>
##  1   8400    77              1    1981
##  2  17510    53              2    1986
##  3  19200    33              4    1992
##  4  42717    55              1    1969
##  5   5000    58              2    1989
##  6  30000    29              4    1994
##  7  18000    45              4    2004
##  8  14400    70              1    1942
##  9  92000    43              1    1989
## 10  35000    50              1    1995
```

Univariate Exploratory Data Analysis

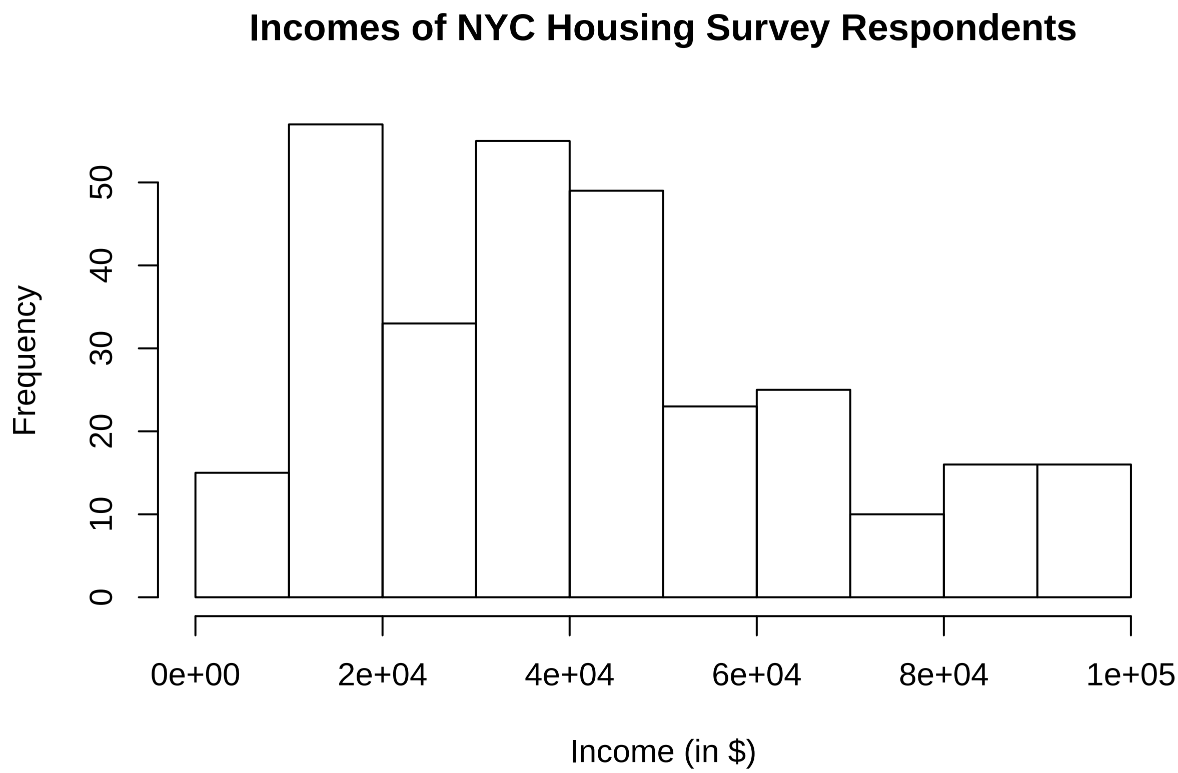
We first analyze the distribution of each variable individually using histograms.




Maintenance Deficiencies among Survey Respondents
Frequency
0
50
100
150
0
2
4
6
8
Number of Maintenance Deficiences Between 2002 and 2005



Years NYC Housing Survey Respondents Moved to NYC
Frequency
0
10
20
30
40
50
1940
1950
1960
1970
1980
1990
2000
Year moved to NYC



The summaries of the distributions of these varaibles are below.

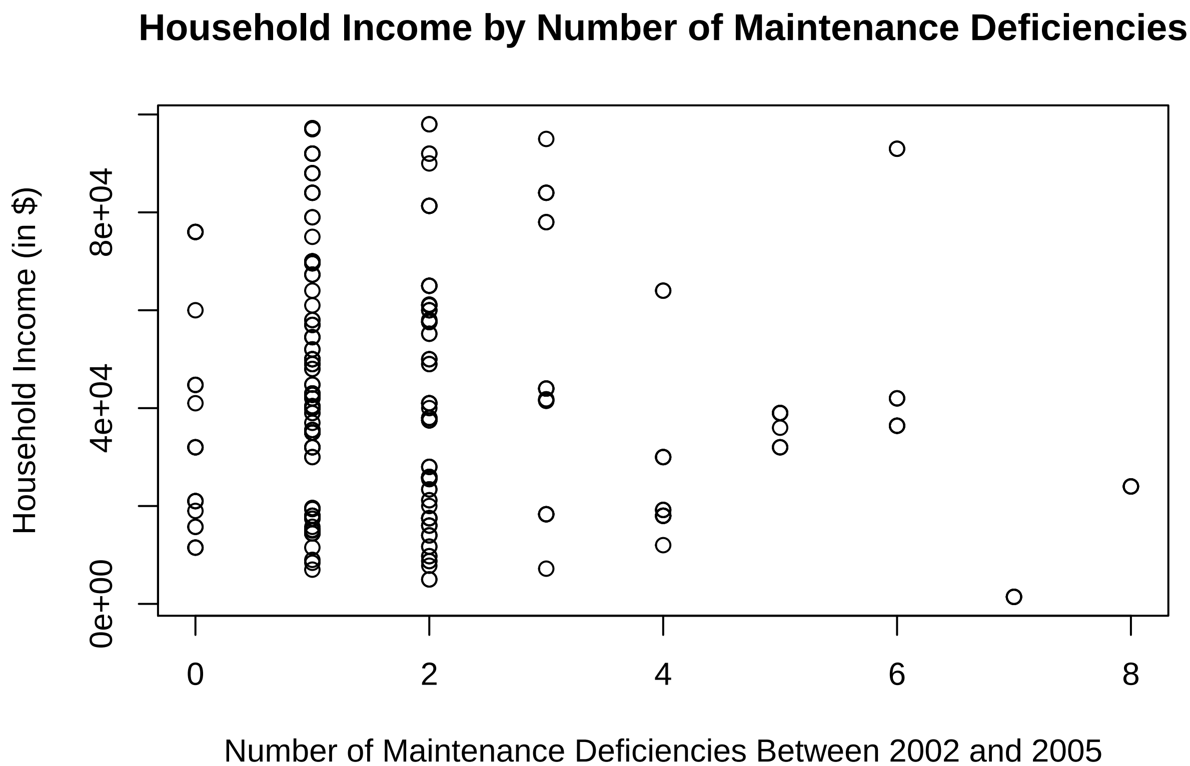
```
##      Income           Age         MaintenanceDef     NYCMove
##  Min.   : 1440   Min.   :26.00   Min.   :0.00   Min.   :1942
##  1st Qu.:21000   1st Qu.:42.00   1st Qu.:1.00   1st Qu.:1973
##  Median :39000   Median :49.00   Median :2.00   Median :1985
##  Mean   :42266   Mean   :50.03   Mean   :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
```

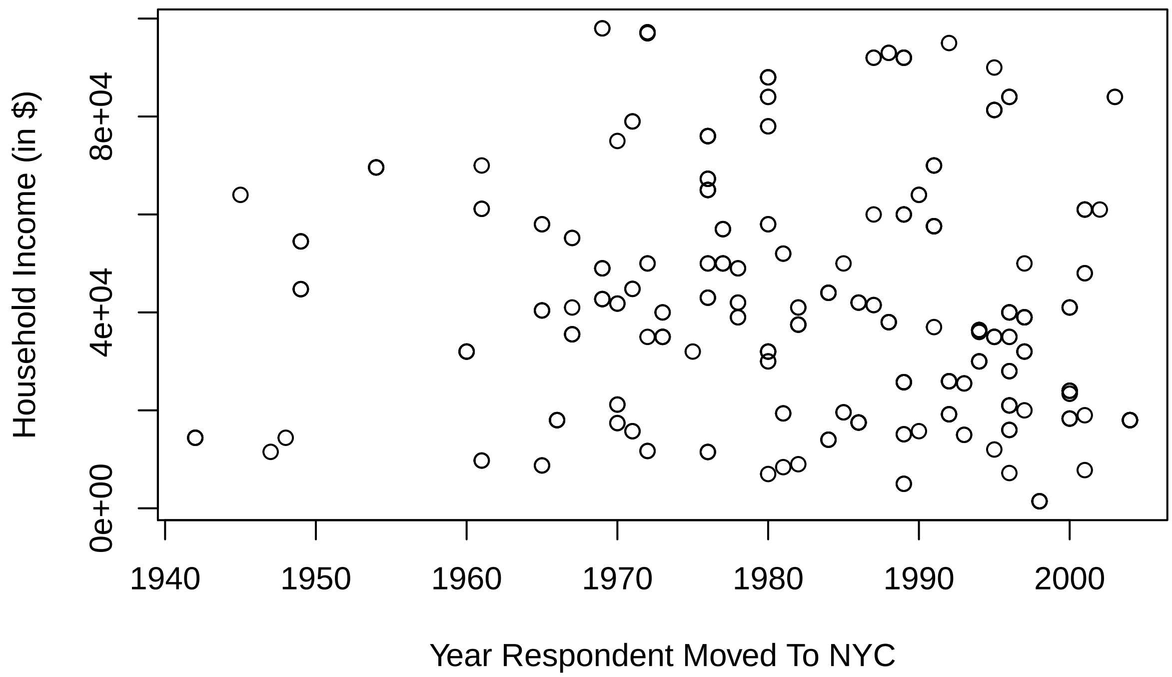
From the histograms and summaries, we observe that the distrbution of Age appears either unimodal or bimodal; we need more data to verify which it is. The average age of survey respondents is about 50 years old. The distribution of Maintenance Defiencies appears unimodal and strongly skewed right, with a median of 2 maintenance deficiencies. The distribution of Year Moved To NYC is either unimodal or bimodal; once again, we need more data to verify which it truly is. It is skewed left, with most respondents moving to NYC between 1960 and 2004. The distribution of Income is either unimodal or bimodal, and roughly symmetric. The average household income is $42266, with a range of $1440 to $98000

Bivariate Exploratory Data Analysis

We next examine the relationship between the response variable (Income) and the explanatory variables (Age, Maintenance Deficiencies, and Year Moved To NYC). One scatterplot is created for each explanatory variable to compare the relationship of each explanatory variable with Income.


Household Income by Age of Respondent
Household Income (in $)
0e+00
4e+04
8e+04
30
40
50
60
70
80
Age (in years)


Household Income by Number of Maintenance Deficiencies
Household Income (in $)
0e+00
4e+04
8e+04
0
2
4
6
8
Number of Maintenance Deficiencies Between 2002 and 2005



From the scatterplots, we observe that the relationship between Age and Income is linear and positive, but very weak. As age increases, household income tends to increase. The realtionship between Maintenance Deficiencies and Household Income appears linear and negative, and also weak. It appears that as maintenance deficiencies in a household increase, income tends to decrease. Finally, the relationship between Year Moved To NYC and Household Income appears linear, negative and weak. As the year the respondent moved to NYC increases, household income tends to decrease.

Modeling

From the bivariate EDA above, we observed that the relationship between Income and each of the response variables appears linear. Therefore, a multiple linear regression model seems appropriate.

We will not perform any transformations on the data because the distribution of Income is roughly symmetric.

Next, we create a multiple linear regression model with all the explanatory variables included and check for multicolllinearity. Below is the correlation matrix of the data.

```
##                 Income   Age MaintenanceDef NYCMove
## Income            1.00  0.04          -0.17   -0.10
## Age               0.04  1.00          -0.25   -0.64
## MaintenanceDef   -0.17 -0.25           1.00    0.46
## NYCMove          -0.10 -0.64           0.46    1.00
```

From the correlation matrix, we observe that the correlation of -0.64 between Year Moved To NYC and Age raises the concern that there may be dangerous multicollinearity. To formally verify, we check the vif's below.

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

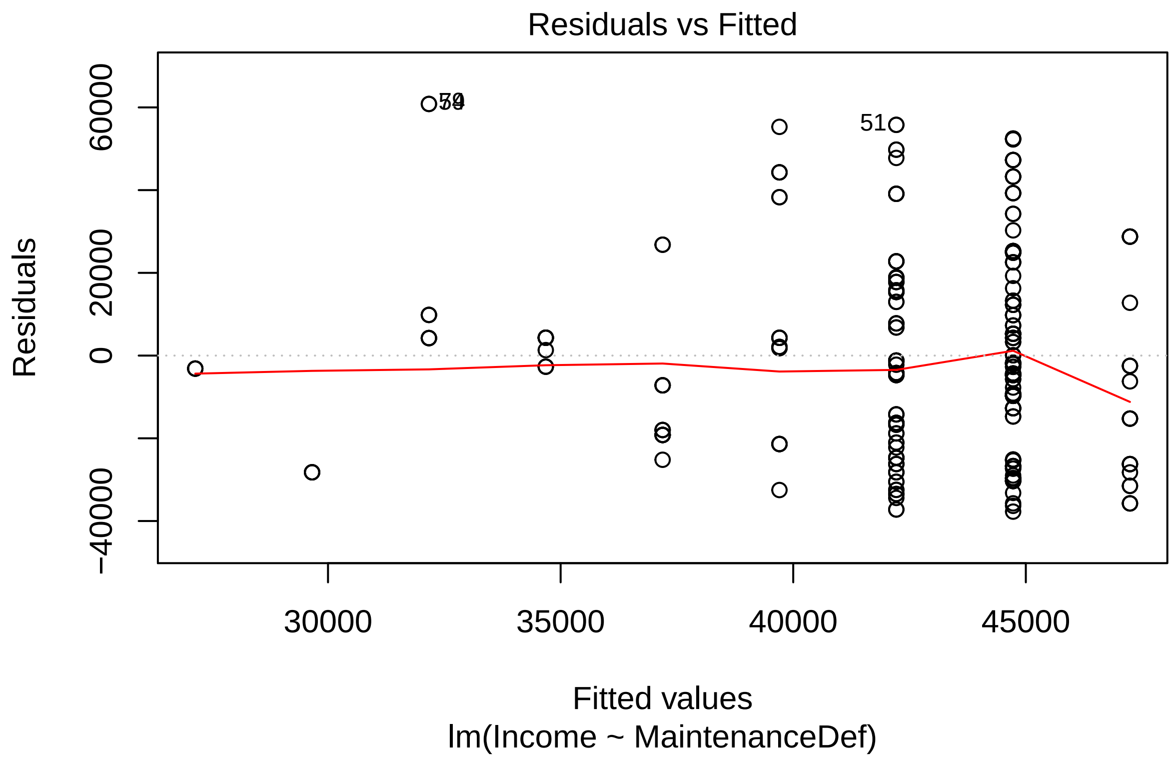
None of the vifs are over 2.5. Therefore, we conclude that there is no dangerous multicollinearity.

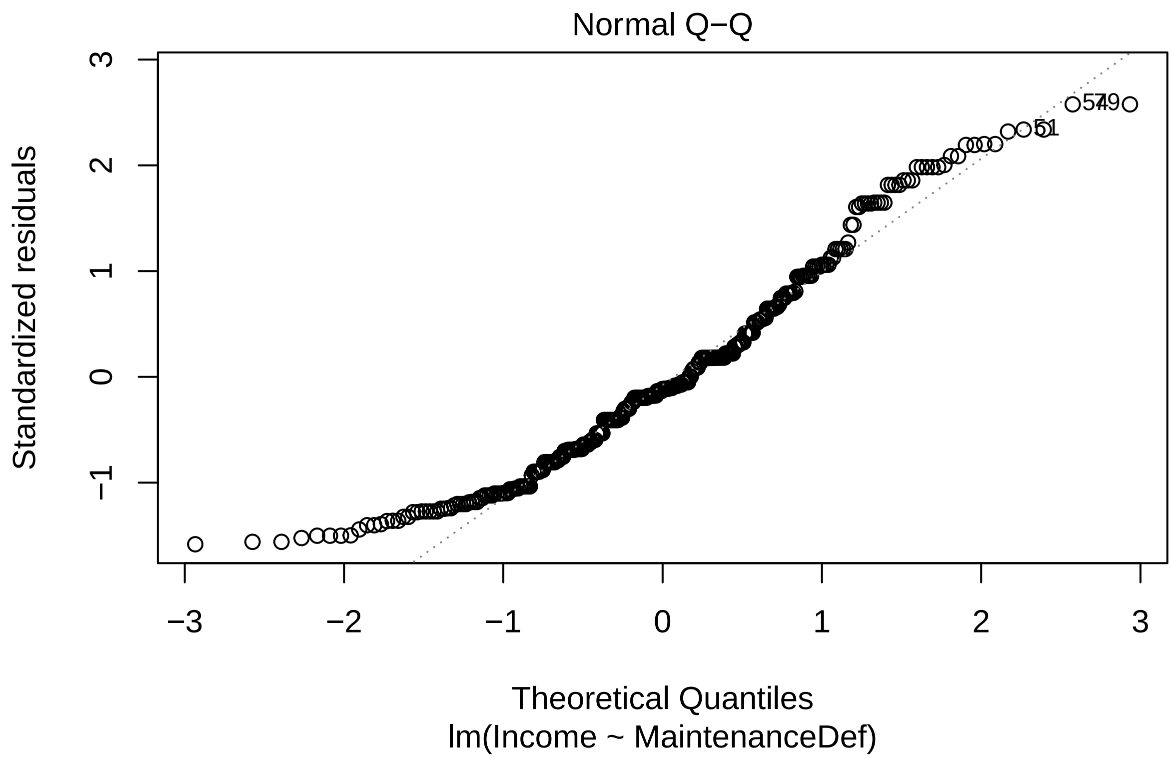
Next, we check the best submodels to determine which variables we should include.

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

Out of these three models, the simple linear regression model with Maintenance Deficiencies as the only response variable has the highest adjusted R^2 and is the simplest model. Therefore, we proceed with the simple linear regression model.

Using the simple linear regression model, we examine the residual plot and normal qq plot to check that the error assumptions are reasonable.


Residuals vs Fitted
Residuals
60000
20000
0
−40000
59
74
51
30000
35000
40000
45000
Fitted values
lm(Income ~ MaintenanceDef)



The residuals appear to be scattered randomly, so it is reasonable to assume that the errors are independent. Second, the residuals appear to have mean close to 0, so we can assume that the errors have mean of 0. Third, the residuals appear to have constant spread, so it is reasonable to assume that the errors have constant standard deviation. Fourth, in the qq plot, there is some deviation from the line at both tails. However, most of the points are close to the line, so it is reasonable to assume that the errors are normally distributed. The error assumptions are all satisfied, so we proceed with the current model.

Below is a summary of the model.

```
##
## 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
```

We consider this an appropriate model. In the univariate EDA above, we verified that the relationship between Maintenance Deficiencies and Income appears to be linear. The error assumptions of independence, mean 0, constant standard deviation, and normal distribution are all satisfied. Additionally, this model had the highest adjusted R^2 out of all the linear regression models we examined. It is also the simplest.

From the summary, we observe that as the number of maintenance deficiencies increases, income tends to decrease. This verifies the negative relationship between Maintenance Deficiencies and Income that we observed in univariate EDA.

The p-value of 0.003553 is less than 0.05, so the relationship is statistically significant. Overall, this model is the simplest model and had the highest adjusted R^2 out of all the ones we examined, so we are confident that this is an appropriate model to predict the household income of survey respondents.

Prediction

Now, we will predict the income of a household with a 53 year-old respondent who moved to NYC in 1987 and three maintenance deficiencies.

```r
47238.6 - 2511.6*3
```

```
## [1] 39703.8
```

For a household with a 53 year-old respondent who moved to NYC in 1987 and three maintenance deficiencies, we predict a household income of $39703.8. Since the only response variable in our model is Maintenance Deficiencies, we ignored the repondent's age and year they moved to NYC in our prediction.

Discussion

Overall, we conclude that the number of maintenance deficiencies in a household has a negative linear relationship with income, and that this relationship is statistically significant.

Limitations of the model include that the R^2 of 0.02826 is low, so the model accounts for little of the variation in household income. Additionally, the residuals are not completely normally distributed, so the error assumptions may not be satisfied.

Future studies could explore the effect of other variables (borough of nyc, ethnicity, number of people in the household, etc.) on household income. The current model only accounts for the number of maintenance deficiencies and another predictor may help to more accurately predict household income.