Predicting NYC Household Income

Liam Gersten
lgersten

Wed, 3/24/2021

Contents

Introduction **1**

Modeling **8**

Prediction **17**

Discussion **17**

Works Cited **18**

Introduction

Rent in New York City can be exorbitantly expensive. Despite the abundance of jobs and opportunities in what some may consider the culturally distinct hub of world commerce, the island of Manhattan in particular stands out as unaffordable for most. Even small flats tend to cost more than an estimated $1,000 per square foot (Warren). Such a trend warrants exploration of other factors and their relationships with cost. These include but aren't limited to resident age, income, maintenance deficiencies, or year of arrival. More specifically, we'll be using the reputable New York City Housing and Vacancy Survey to predict resident income from different combinations of the aforementioned variables.

Exploratory Data Analysis

Data

The data obtained via the survey contains 299 entries and 4 variables, all of which can be used as quantitative predictors/estimators during analysis. Since we are primarily interested in predicting resident income, we will be exploring and modeling the relationship between total household income (in $) and four variables:

Age: respondent's age (in years)

MaintenanceDef: number of maintenance deficiencies of the resident from 2002 to 2005

NYCMove: the year the respondent moved to New York City

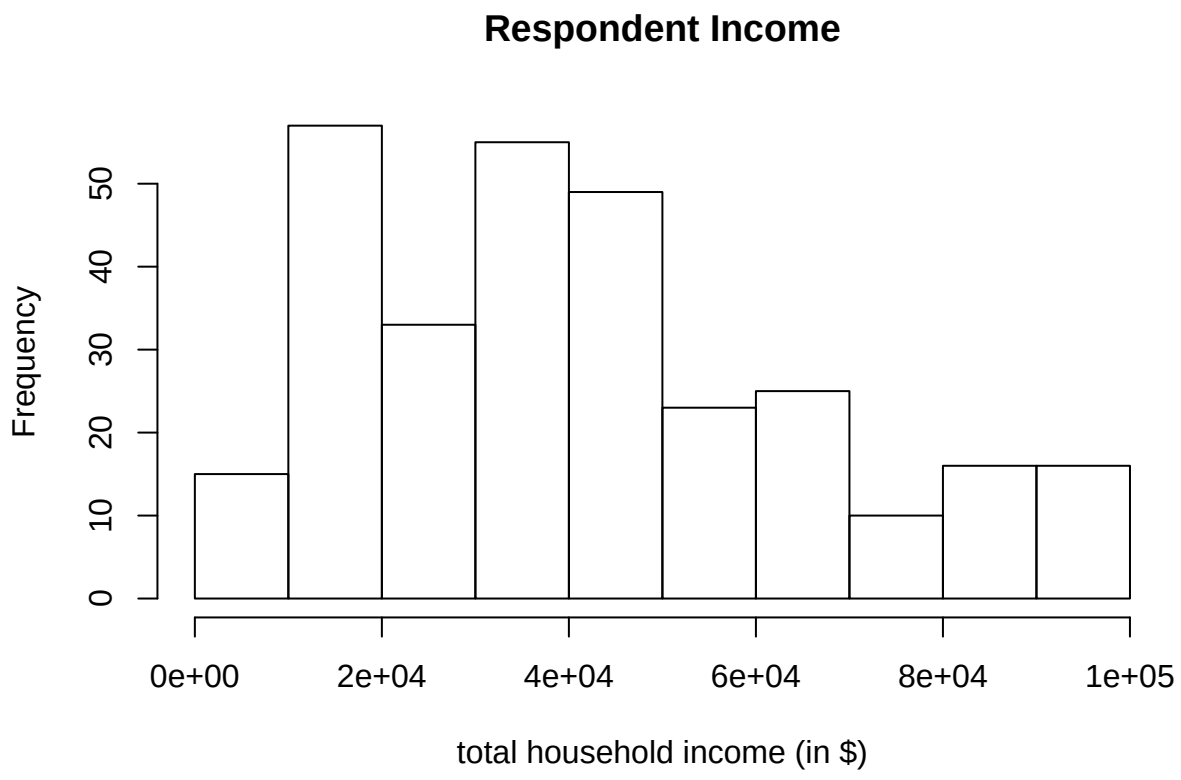
We are primarily interested in presenting one or more linear/multiple regression models to help understand the relationship between income and said variables.

The header or first few lines of the data are as follows:

```
## # 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
```

Univariate Exploratory Data Analysis

To start, we will visualize and explore each variable individually using histograms and numerical summaries.

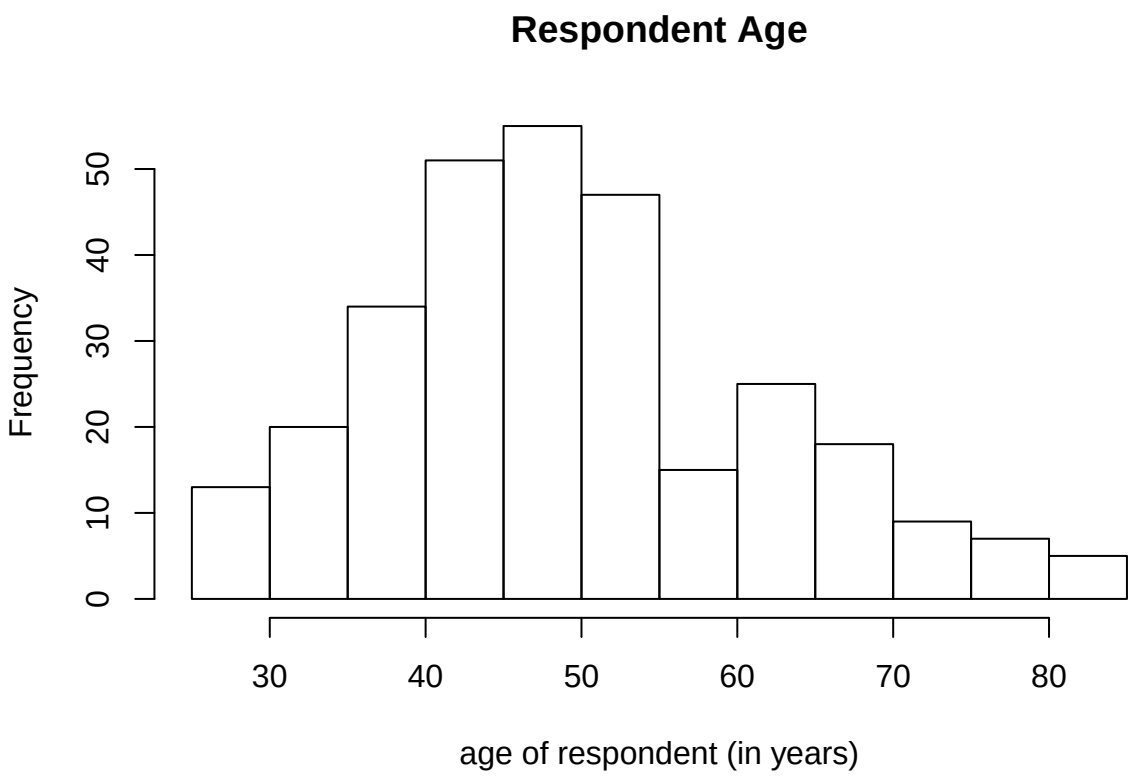


Summary

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

Observations

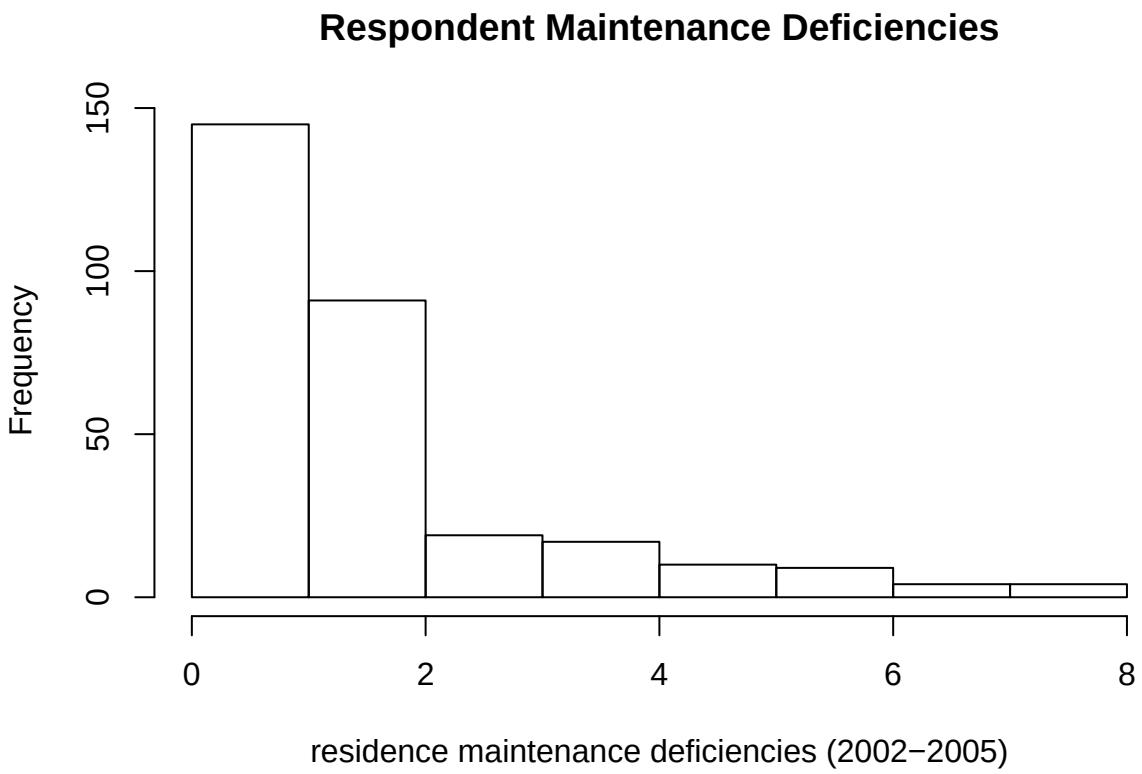
The distribution of Respondent income appears slightly skewed to the right with two peaks in the two lower quartiles which may indicate that the distribution is bimodal. There are possible outliers towards the upper extrema. The household income of most respondents ranges from $20,000 to $40,000. It should be noted that the median household income of respondents is $39,000, which is roughly $30,000 less than the median household income for the United States as a whole (Semega et AL).



Summary

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

Observations The distribution of respondent age appears roughly symmetric and unimodal with a single peak between the ages of 40 and 50. All respondents were adults and none were older than 85.

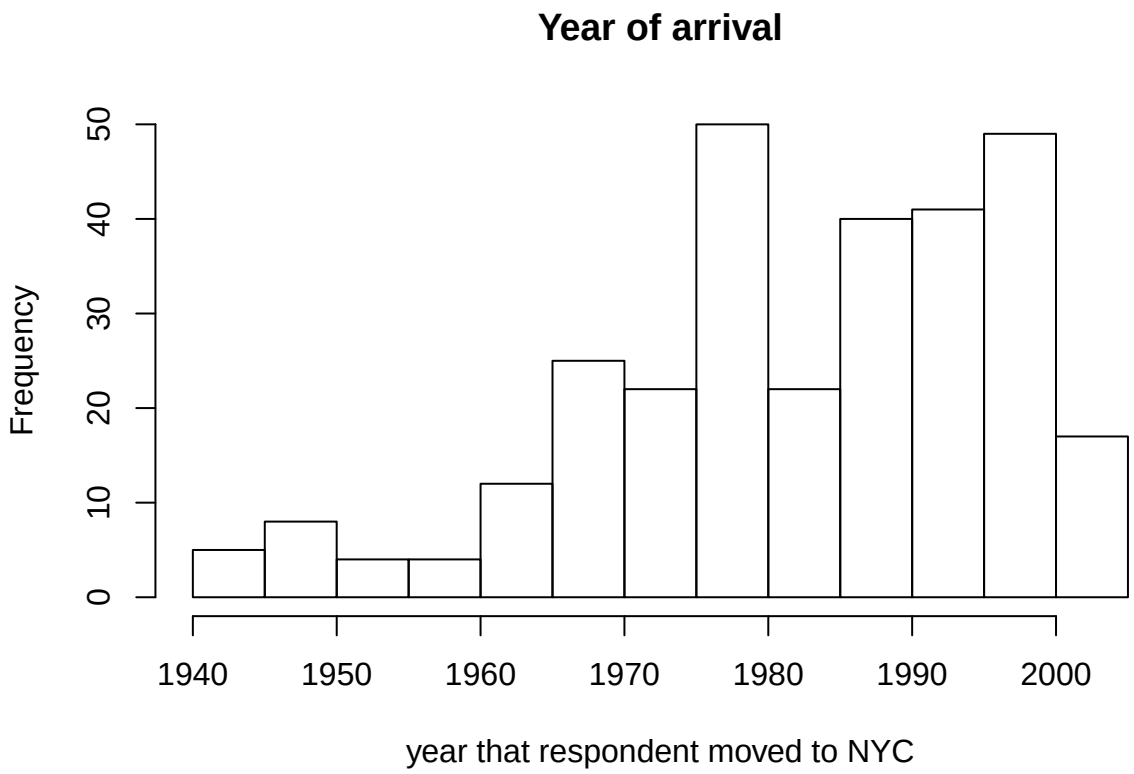


Summary

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

Observations

The distribution of respondent maintenance deficiencies is strongly skewed to the right with severe outliers towards the right extrema. On closer inspection, we can identify a small handful of outliers between 6 and 8 deficiencies.



Summary

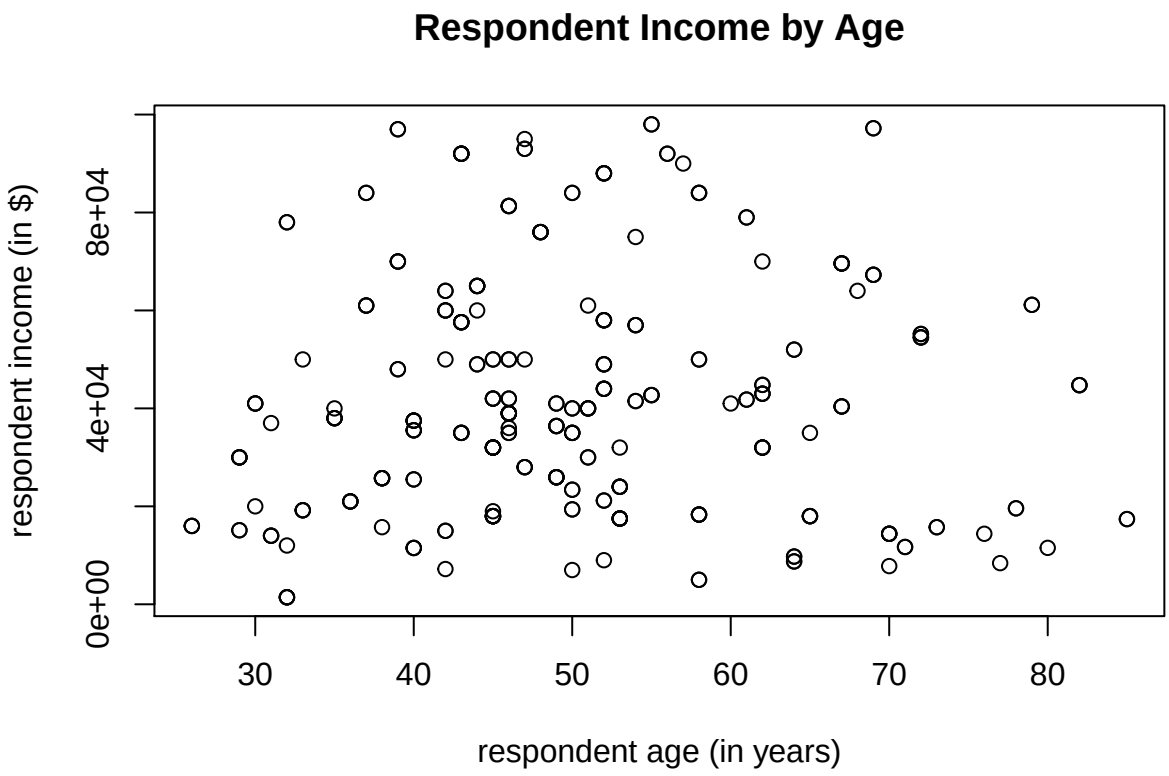
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1942    1973    1985    1983    1995    2004
```

Observations

The distribution of arrival years is skewed to the left and possibly bimodal with twin peaks in the upper two quartiles and possible outliers towards the lower extrema.

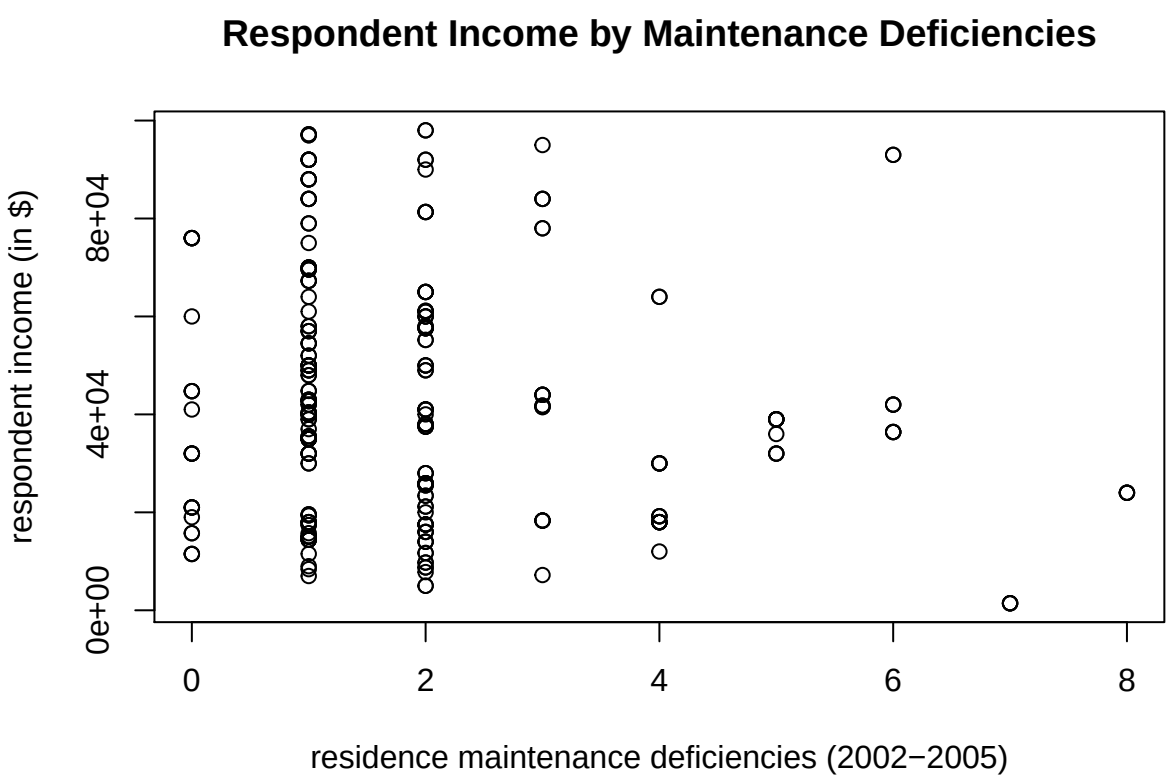
Bivariate Exploratory Data Analysis

Next, we will see and comment on three scatterplots, each representing some relationship between Income and one of its possible predictors



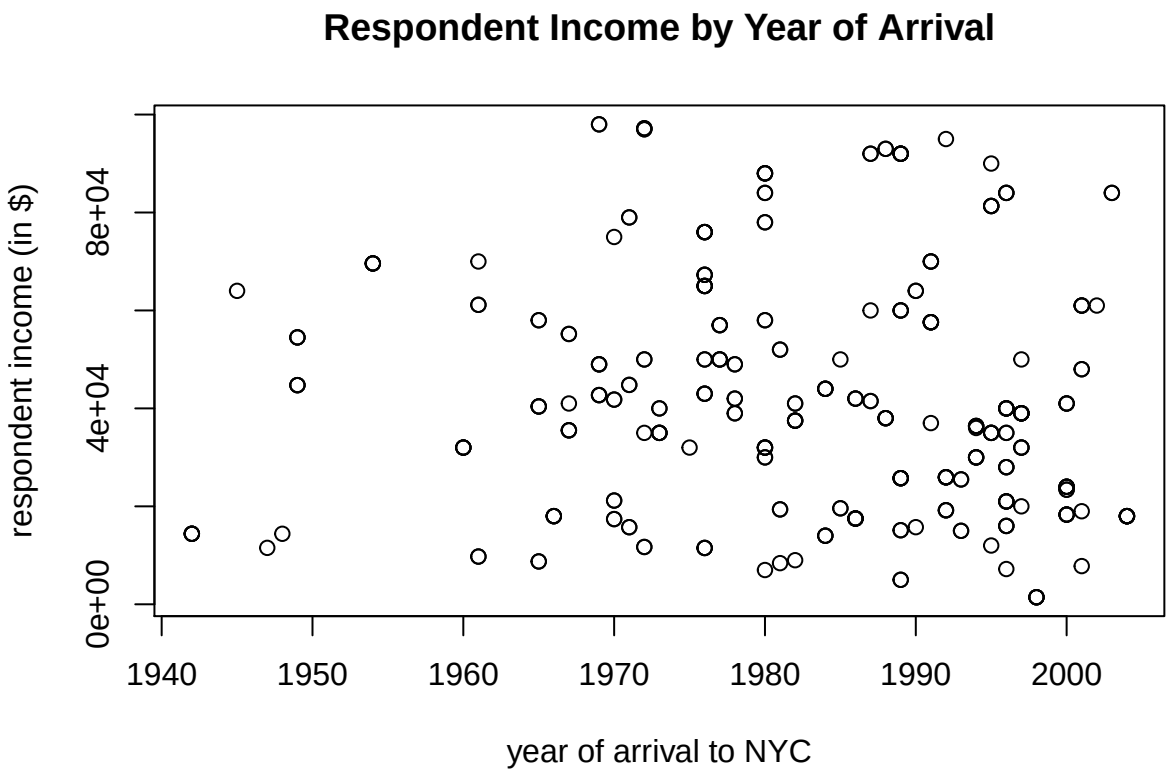
Observations

There appears to be a positive but weak linear association between respondent income and age. In general, a respondent's income increases as their age increases.



Observations

There seems to be a very weak, positive linear association between maintenance deficiencies and respondent's income. As the number of maintenance deficiencies increases, respondent income seems to somewhat increase.

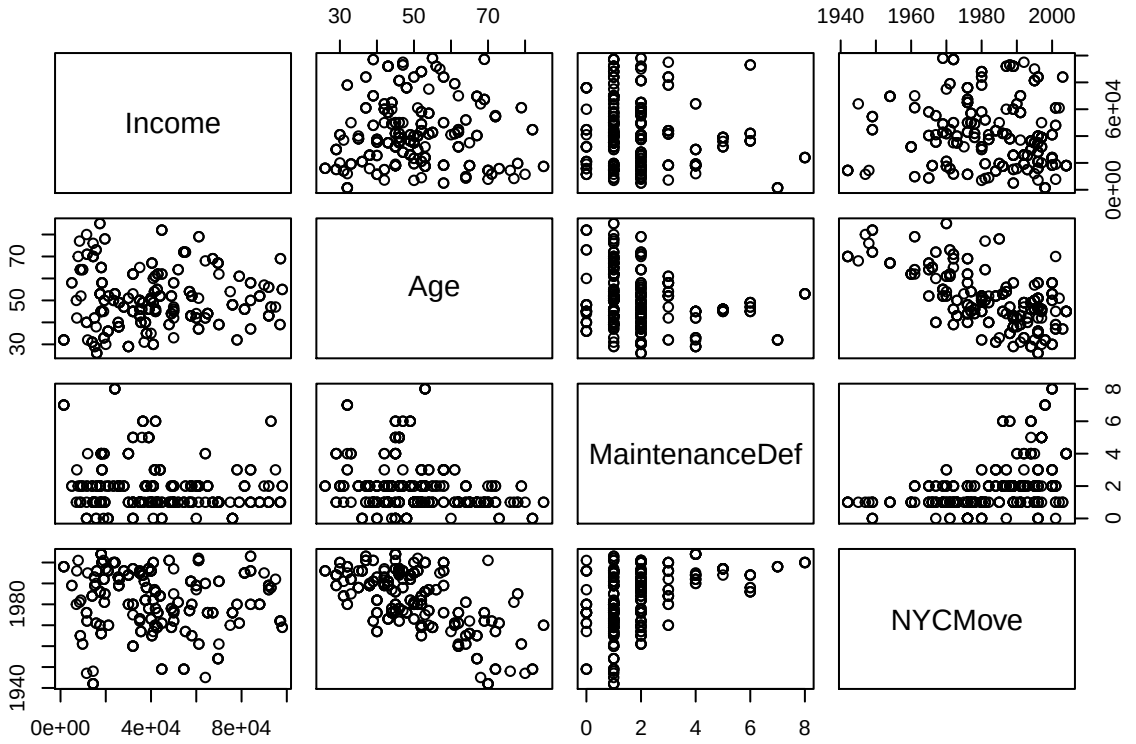


Observations There appears to be a weak, negative linear relationship between year of arrival to NYC and respondent income. Generally speaking, newer arrivals tend to have a slightly lower household income.

Modeling

Notice that three of the four histograms produced showed some skewness, some stronger than others. As such, we may need to try one or more transformations on variables so as to validate the assumptions needed for a multiple linear regression model. Furthermore, all predictors showed some relationship with Income, so we aim to include all of them in our model. First, we must check for signs multicollinearity.

Relationships Between Quantitative Variables




```
##                      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
```

Observations

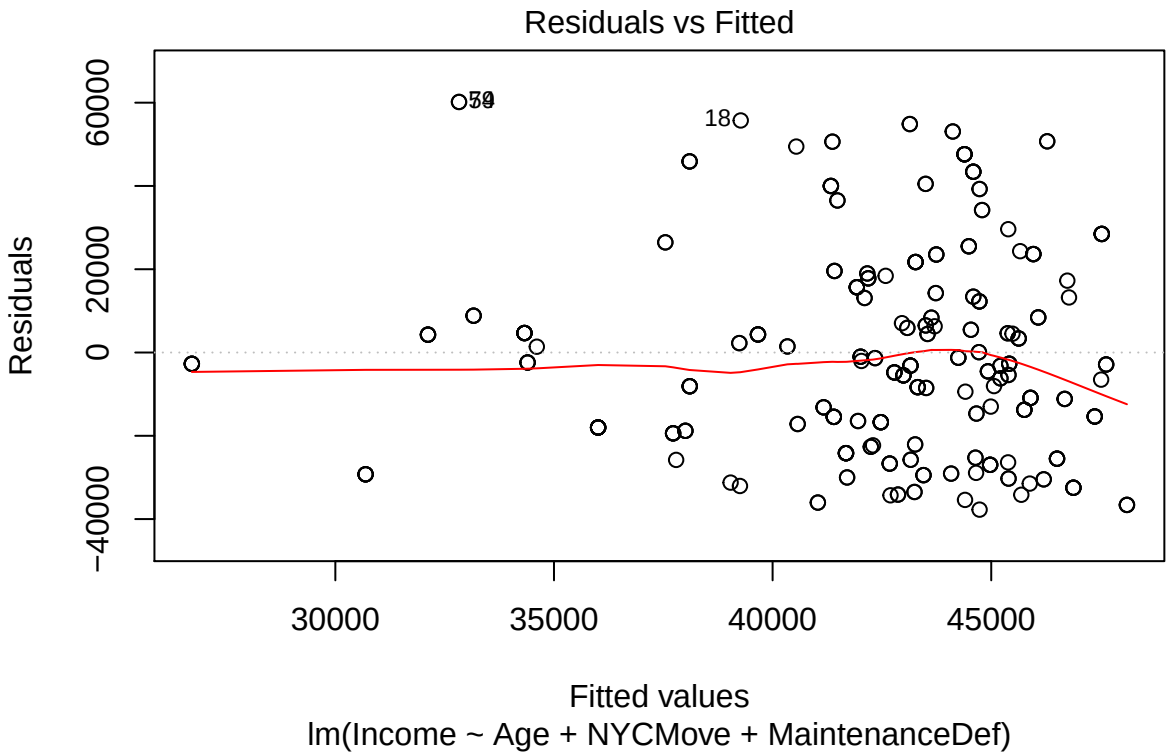
We notice that there seems to be a mildly strong, negative linear association between Age and NYCMove, with a correlation coefficient of -0.63659204. This may indicate possible multicollinearity, so we will test vifs. To do this, we produce a multilinear regression model with all three predictors before testing for vifs.

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

Since no variables produce a vif greater than 2.5, we can proceed without worry of strong multicollinearity.

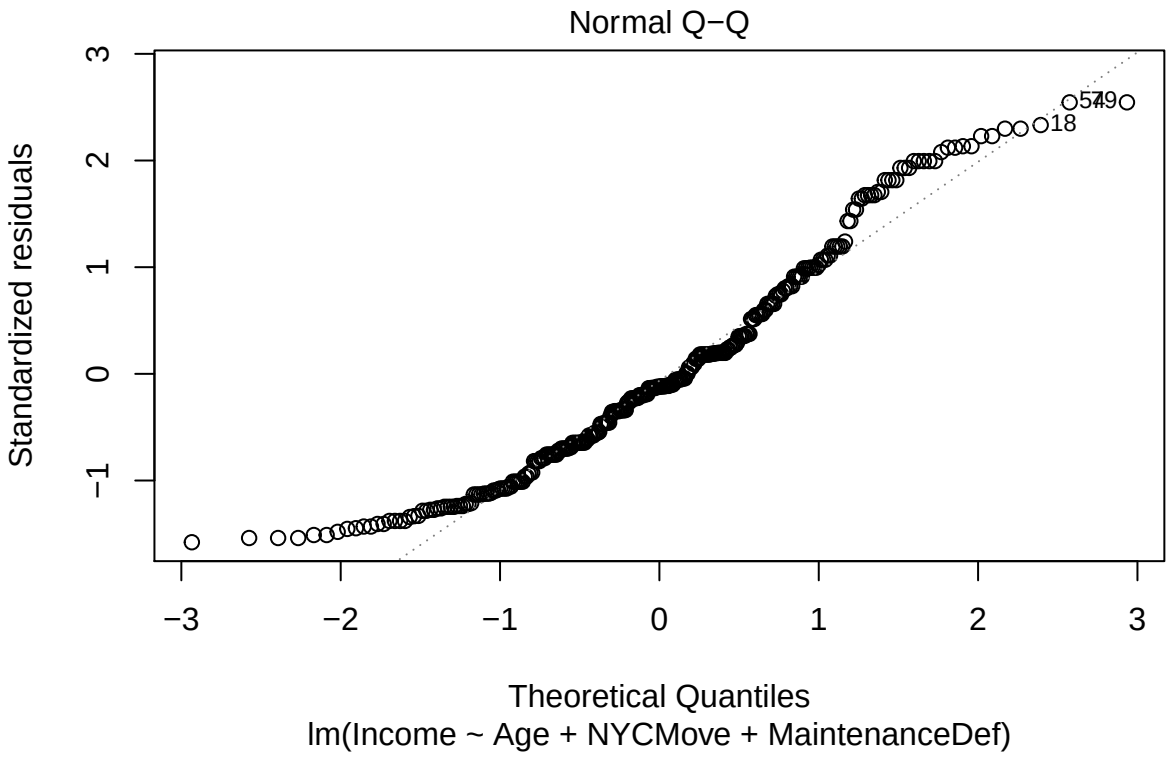
Diagnostics

Before attempting to validate the error assumptions of a multiple regression model via residual diagnostics, we must first use scatterplots to validate linearity assumptions for each predictor. We know from our bivariate exploratory data analysis that the scatterplots of all three predictors vs income appear reasonably linear. Some outliers can be observed in both "respondent income by maintenance deficiencies" and "respondent income by year of arrival." Since the linearity assumption for a multiple regression model has been reasonably validated for each predictor, we will now proceed with validating error assumptions using residual diagnostics for the full model.



Observations

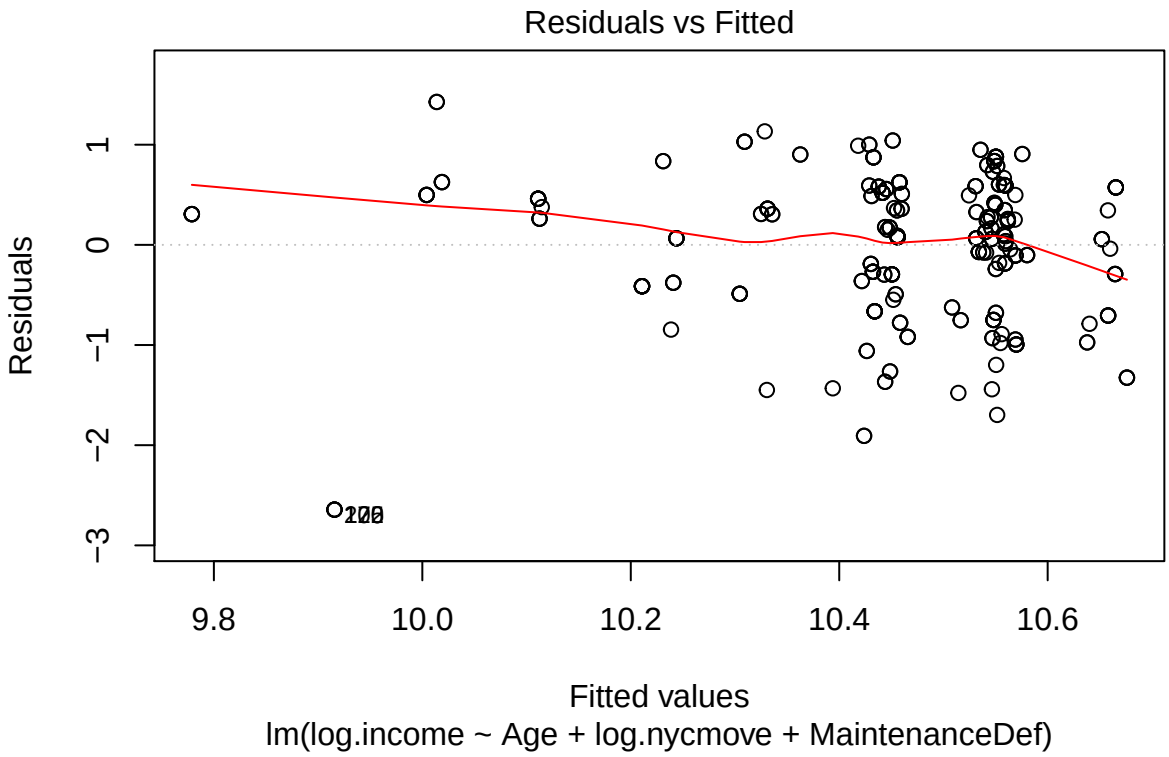
Since residuals are shown to be scattered above and below the 0-line without pattern, centered around 0, and constantly spread above and below the 0-line, we can reasonably assume that the errors are independent, have mean 0, and have constant standard deviations.

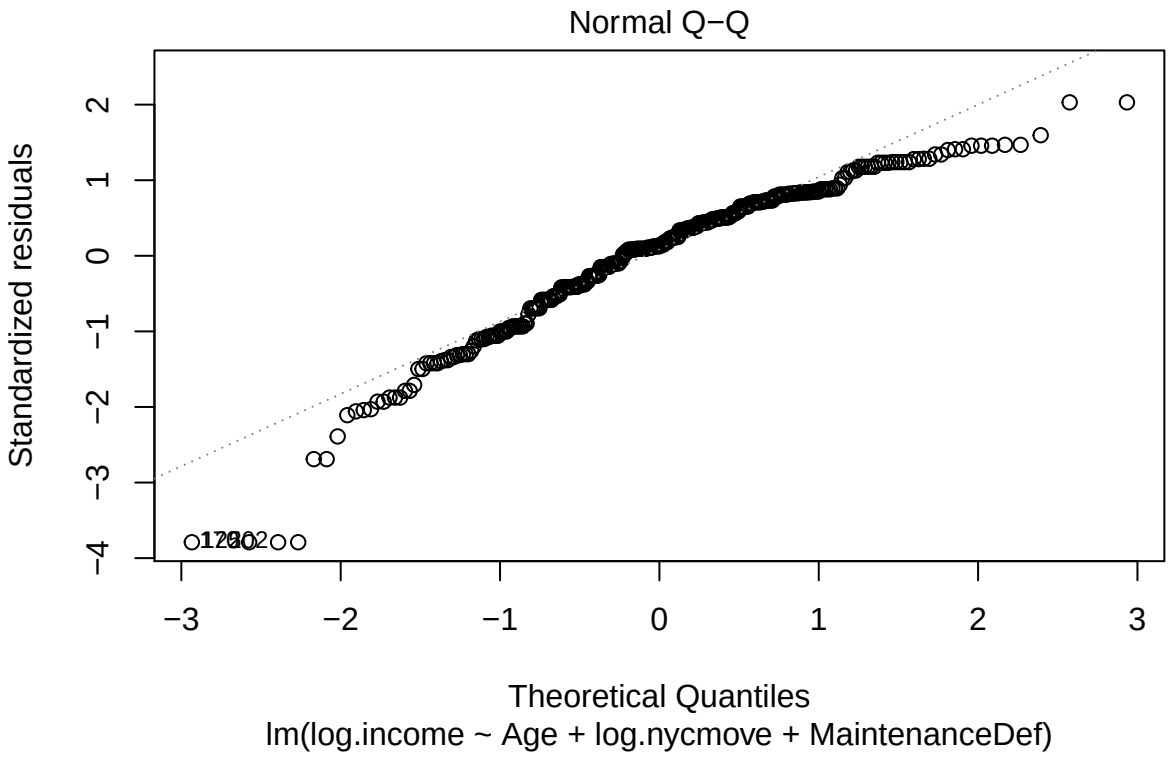


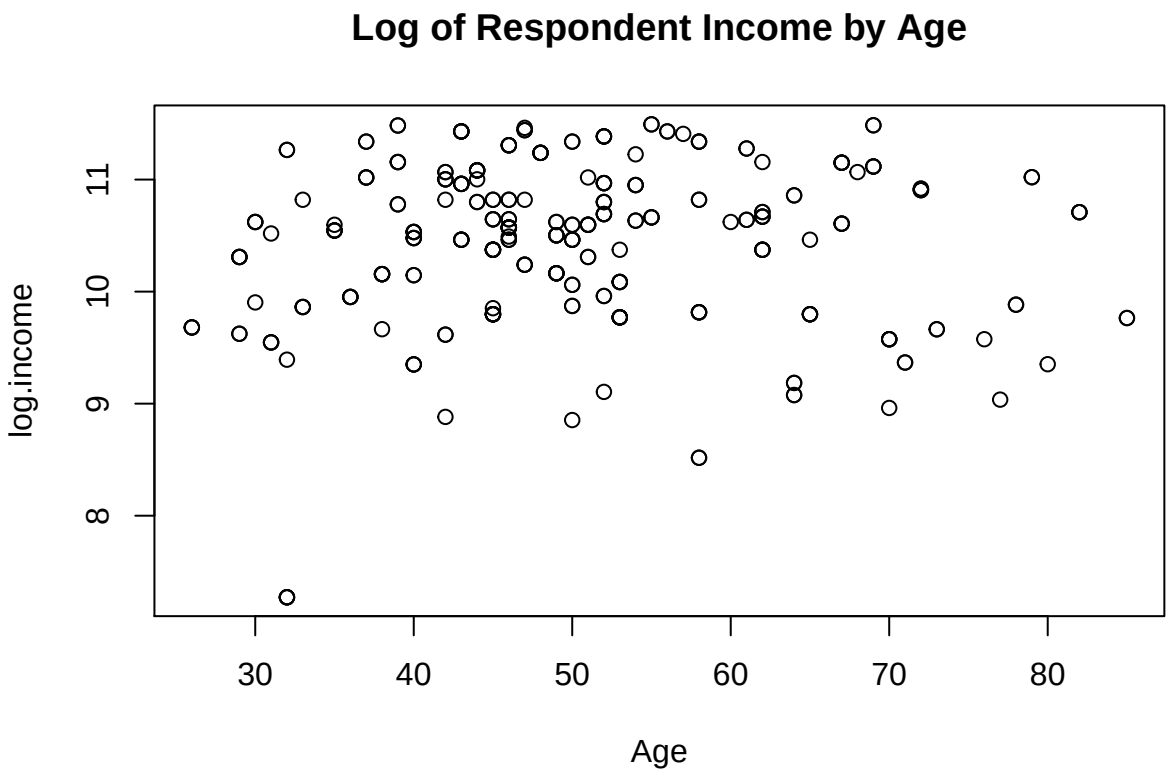
Observations

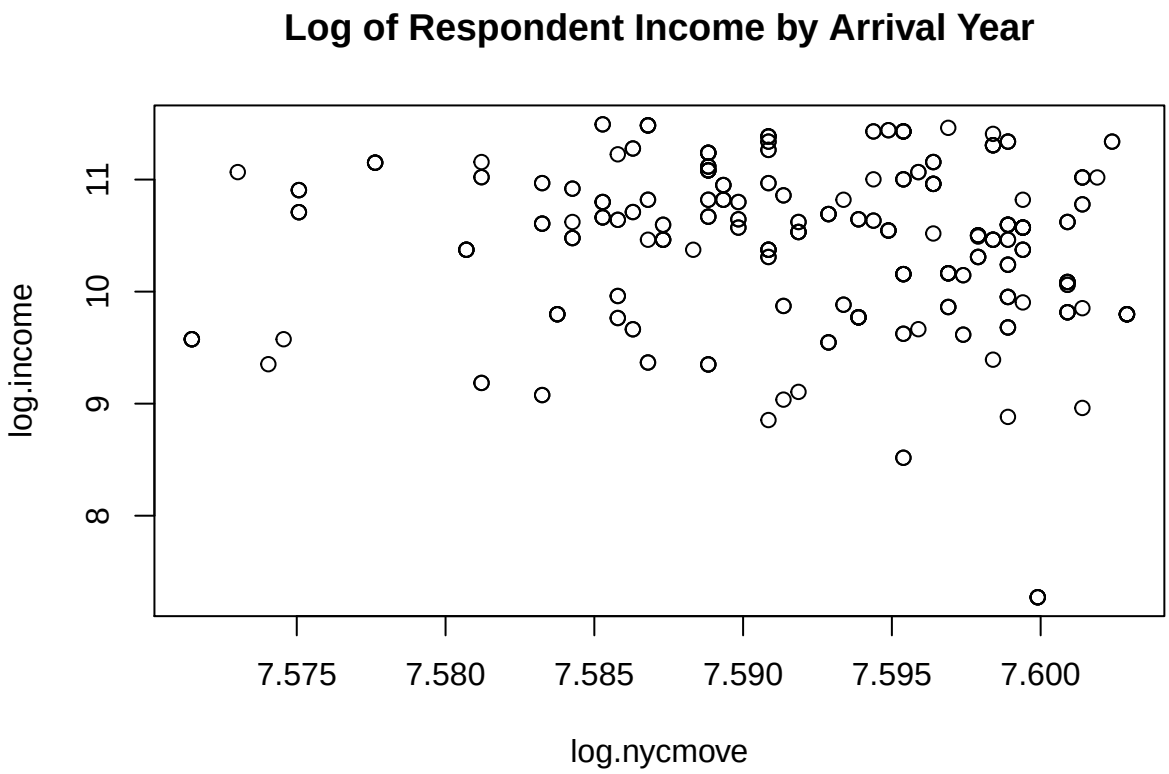
The normal Q-Q plot show systematic deviation from its line. As such, log transformations will be applied to certain variables until we can validate that the errors of the model are normally distributed.

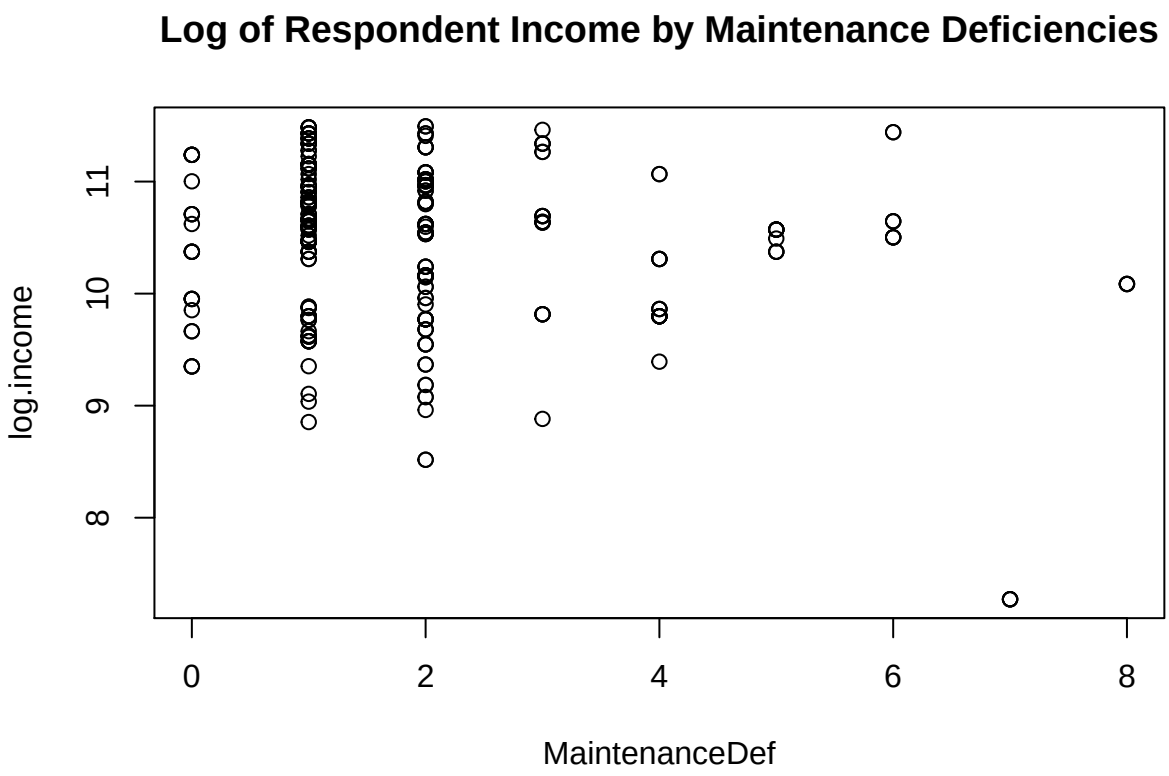
After testing combinations of transformations for each variable, we find that taking the natuaral log of both Income and NYCMove alleviates some the error normality violation as seen on the new Q-Q plot. We also see that all previous error assumptions hold true after inspecting the updated residual diagnostics, and updated scatterplots show a reasonably linear association.


Residuals vs Fitted
Residuals
1
0
−1
−2
−3
9.8
10.0
10.2
10.4
10.6
Fitted values
lm(log.income ~ Age + log.nycmove + MaintenanceDef)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log.income ~ Age + log.nycmove + MaintenanceDef)


Log of Respondent Income by Age
log.income
8
9
10
11
30
40
50
60
70
80
Age


Log of Respondent Income by Arrival Year
log.income
8
9
10
11
7.575
7.580
7.585
7.590
7.595
7.600
log.nycmove



Summary of Chosen Model

```
##
## Call:
## lm(formula = log.income ~ Age + log.nycmove + MaintenanceDef,
##     data = nyc)
##
## Coefficients:
##    (Intercept)             Age     log.nycmove  MaintenanceDef
##      28.375359       -0.001335       -2.325244       -0.106471
```

```
##
## Call:
## lm(formula = log.income ~ Age + log.nycmove + MaintenanceDef,
##     data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64330 -0.39538  0.08962  0.51975  1.42655
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.375359  62.062420   0.457 0.647859
## Age            -0.001335   0.004312  -0.310 0.756989
## log.nycmove    -2.325244   8.160274  -0.285 0.775885
## MaintenanceDef -0.106471   0.028658  -3.715 0.000243 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7121 on 295 degrees of freedom
## Multiple R-squared:  0.0584, Adjusted R-squared:  0.04882
## F-statistic: 6.099 on 3 and 295 DF,  p-value: 0.000488
```

Observations

Although all necessary conditions were satisfied for the use of a multiple regression model including all three possible predictor variables, the model itself is far from a perfect predictor of Income with a Multiple R-squared of only 0.0584. The low value of our R-squared makes some sense as each scatterplot modeling the relationship between Income and its predictors showed a weak to very weak linear association by itself. However, the model is still significant with an F-test p-value of 0.000488. The negative coefficient values for predictor match the directions seen in the updated scatterplots.

Over 30 linear models were tested so as to find the one that both met all assumptions and yielded the highest R-squared/Multiple R-squared. These tests did not involve log transformations of the Age variable, as it was shown to be symmetric during Univariate EDA. In general, models involving log.income yielded higher greater R-squared's than those involving the untransformed Income variable.

Prediction

With our established model, we can now predict the income for a household with three maintenance deficiencies, whose respondent's age is 53 and who moved to NYC in 1987.

Our model is as follows:

$$\text{log.income} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{log.nycmove}) + \beta_3(\text{MaintenanceDef})$$

$$\text{log.income} = 28.375359 - 0.001335(\text{Age}) - 2.325244(\text{log.nycmove}) - 0.106471(\text{MaintenanceDef})$$

In this case:

Age = 53 MaintenanceDef = 3 log.nycmove = log(1987)

$$\text{log.income} = 28.375359 - 0.001335(53) - 2.325244(\log(1987)) - 0.106471(3) = 10.3264$$

```
## [1] 10.3264
```

We predict the log of this income to be 10.3264, which translates to an income of roughly $30,528

This income falls almost $10,000 below the mean income of our data.

Discussion

Through our modeling and analysis of the New York City data set, we can conclude that household income is somewhat related to the age of an occupant, their move-in year, and the number of maintenance deficiencies at that residence from 2002 to 2005. While the model used itself is significant, its predictive power falls short due to a small number of outliers and weak linear relationships between income and its three predictors. Generally speaking, the model shows us that younger residents, more recent move-in dates, and greater maintenance deficiencies are all associated with lower household income in New York.

Works Cited

New York City Housing and Vacancy Survey (NYCHVS). n.d. 22 March 2021. https://www.census.gov/programs-surveys/nychvs.html. Semega, Jessica, et al. “Income and Poverty in the United States: 2019.” 2020. Warren, Katie. “This tiny NYC penthouse costs $1,843 per square foot, but every detail was designed so it ‘functions like one twice its size.’ Take a look inside.” Business Insider 20 February 2020.