NYC 36-202 Project 1

Sean Jin
seanj

Introduction

New York City is the largest city in the United States as it is considered to be one fo the cultural and financial centers of the world. It is also known for having an exorbitant cost of living with rents for a one bedroom apartment in the thousands of dollars. Because these buildings may be in poor condition, there is a high demand that almost guarantees tenants, regardless of the cost. Having a large income may not guarantee good living conditions with respect to housing quality. It is important to find the impact of such conditions among several variables that can lead potentially lead to such different living conditions.

Exploratory Data Analysis

Data

For the data about NYC, we analyze a random sample of 299 respondents and 4 variables. Due to our interest in, household income and 3 independent variables: age, maintenance deficiencies, and year move to NYC.

Variable	Description
Income	Total House income (in $)
Age	Respondent's Age (in years)
Maintentance Deficiencies	Number of Maintenance deficencies between 2002 and 2005
Year	The Year the Respondent moved to NYC

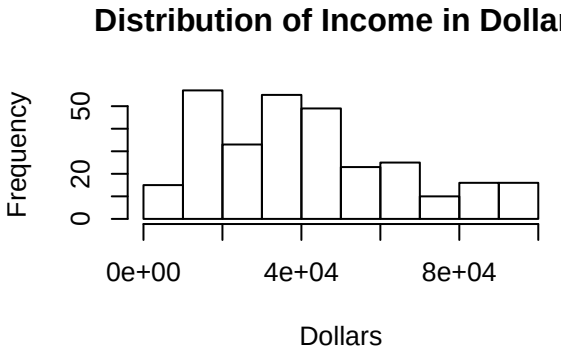
Let's examine the first few lines of data:

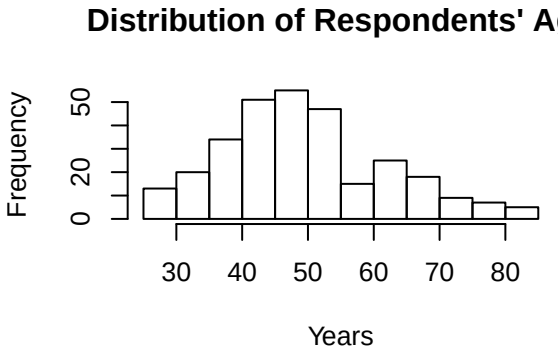
Income	Age	MaintenanceDef	NYCMove
8400	77	1	1981
17510	53	2	1986
19200	33	4	1992
42717	55	1	1969
5000	58	2	1989
30000	29	4	1994

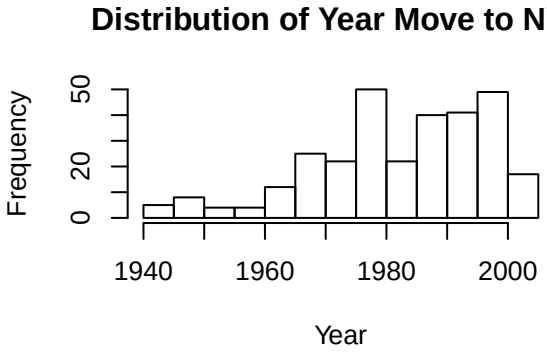
Univariate Exploration:

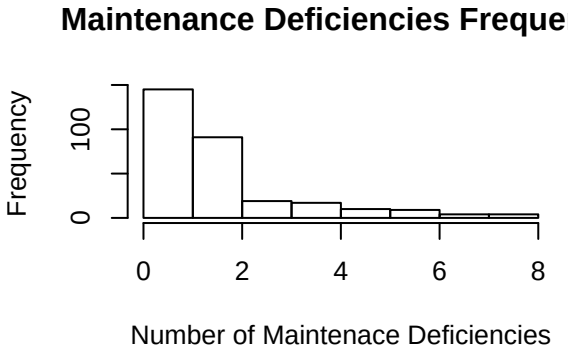
	Income	Age	MaintenanceDef	NYCMove
Min.	1440	26.00	0.00	1942
1st Quartile	21000	42.00	1.00	1973
Median	42266	50.03	2.00	1983
Mean	42266	50.03	1.98	1983

	Income	Age	MaintenanceDef	NYCMove
3rd Quartile	57800	58.00	2.00	1995
Max.	98000	85.00	8.00	2004
SD.	24201	12.44	1.62	14.1





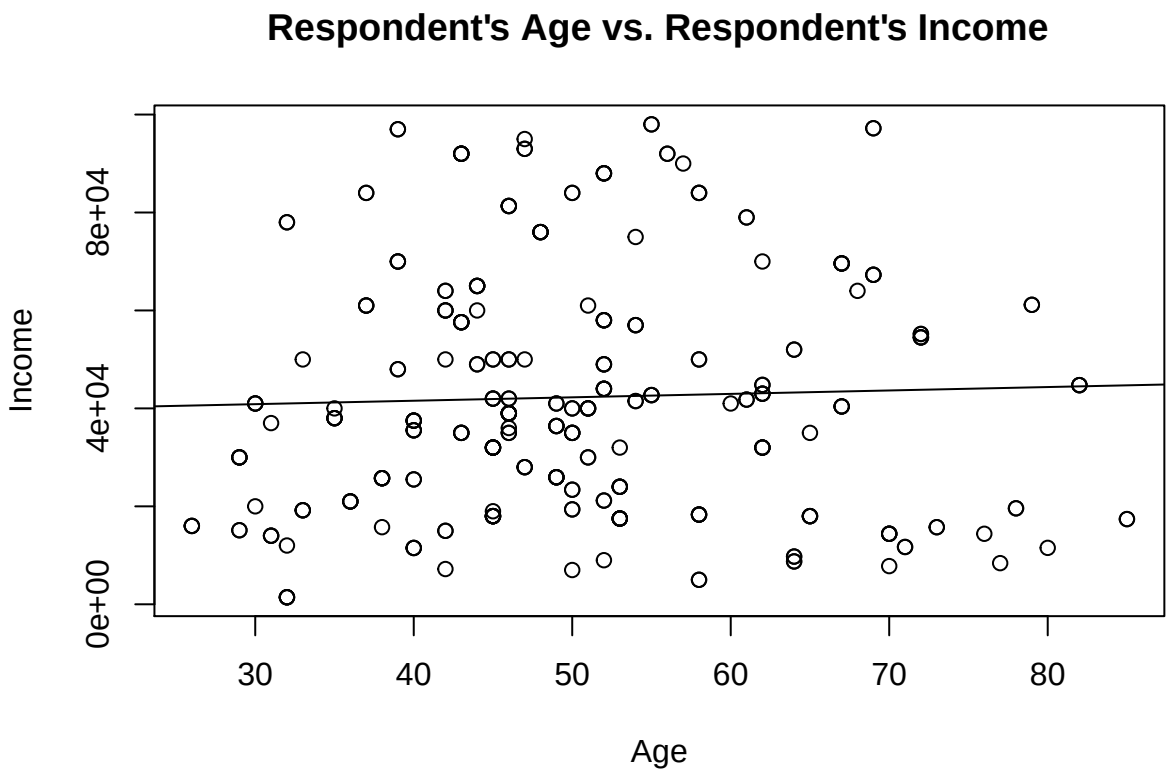


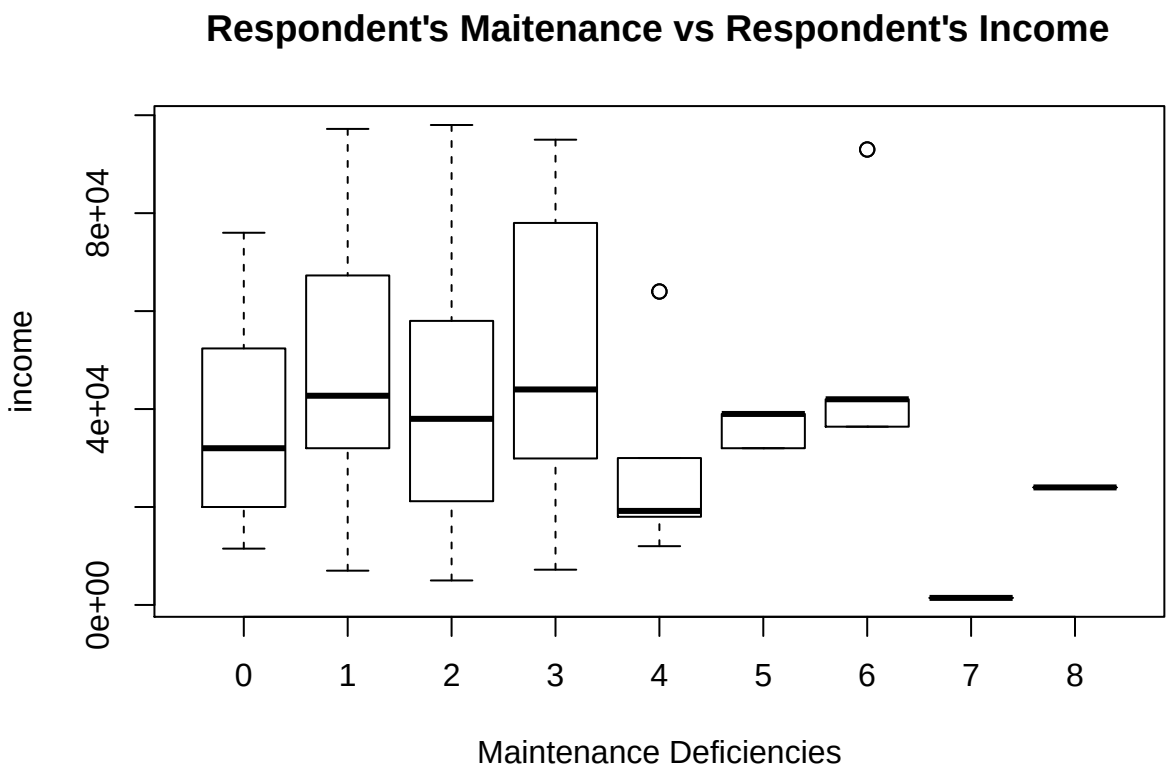


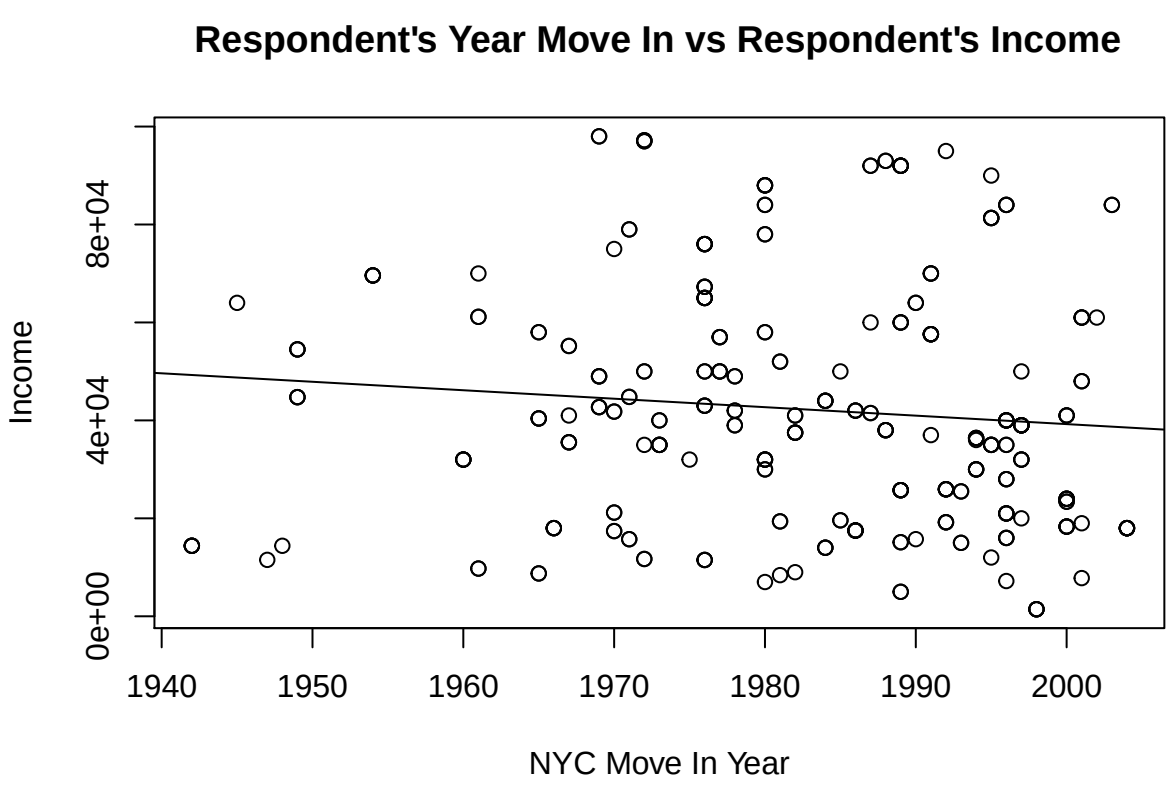
After looking at both the graphs and the summary statistics of our variables, we make the following observations: the distribution of respondent's income is slightly right skewed, with the median and mean being slightly off, but close to each other. The distribution is centered around 39000 and the standard deviation is about 24201 dollars. The distribution in regards of age could be unimodal, except for the dip at the ages from 50-60, but the median and mean are very similar. There seems to be an increase in regard to the number of people who move to NYC up to the most recent years, the exception being from 1975 to 1980, and the median and mean seem to be very similar (1985 and 1983). The number of Maintenace Deficiences range from 0 to 8 with an average of 1.98 and a median of 2.00. The most number of maintenance deficiencies happen to be 1, next being 2.

Bivariate Exploration

Now that we have a good understanding of the distribution of the individual variables in this data, we can start to find the relationship between other variables. We are mostly interested in how variables interact with respondent's income, because if we understand the relationship between income and the other variables, we can find a way to find what may result in a person's income based on age, maintenance deficiencies, and move in year in order to improve conditions in NYC as a whole.


Respondent's Age vs. Respondent's Income
Income
0e+00
4e+04
8e+04
30
40
50
60
70
80
Age

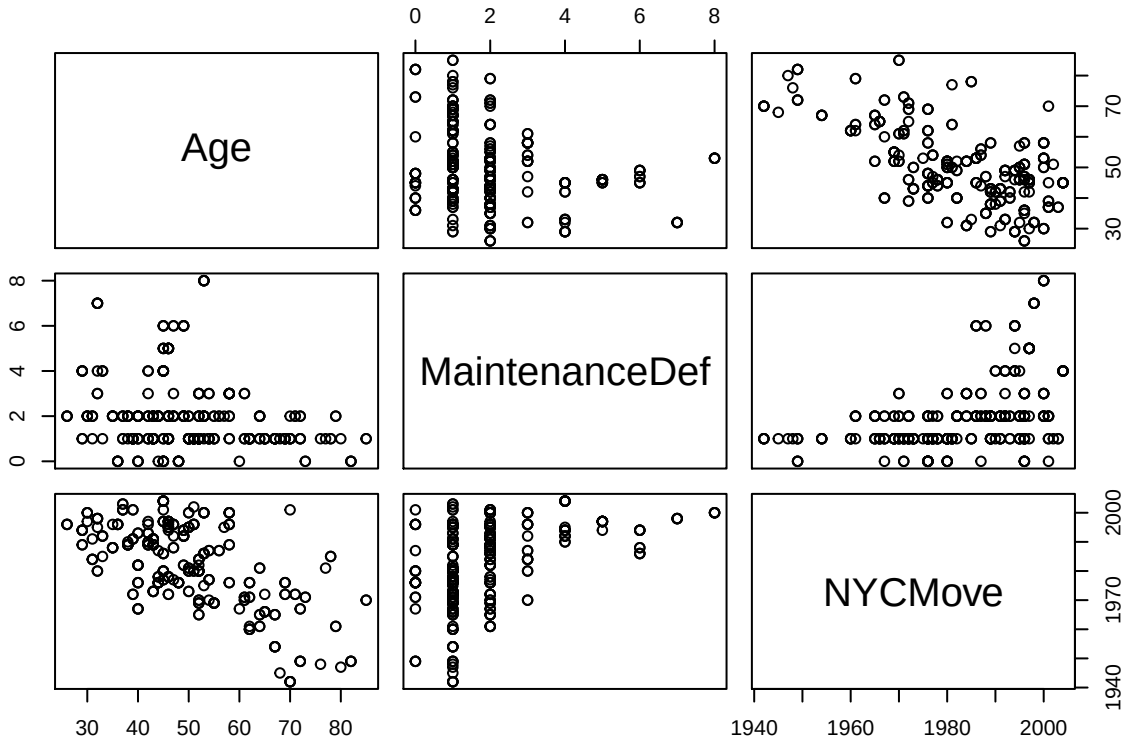

Respondent's Maitenance vs Respondent's Income
income
0e+00
4e+04
8e+04
0
1
2
3
4
5
6
7
8
Maintenance Deficiencies



Through analysis of our graphs, we find that the income relating seem to have a distribution that is not linear. Even the distribution among the number of maintenance deficiencies differ regardless of income as even the distribution within income and maintenance deficiencies have no set trend. Even the regression line for the year that the respondnent's moved in and their age displays that they have little correlation with income based on the magnitude of the regression line.

Modeling

After exploring and visualizing the relationships among our variables, we now turn to building a linear regression model to predict patient satisfaction. We start by looking at the histogram of our response variable. Some of these graphs seem to be similar, so there is a good indication of multicolinearity, especially the relationship between maintenance deficiencies and year that the respondent has moved in as well and the relationship between maintenance deficiencies and the age of the respondent.

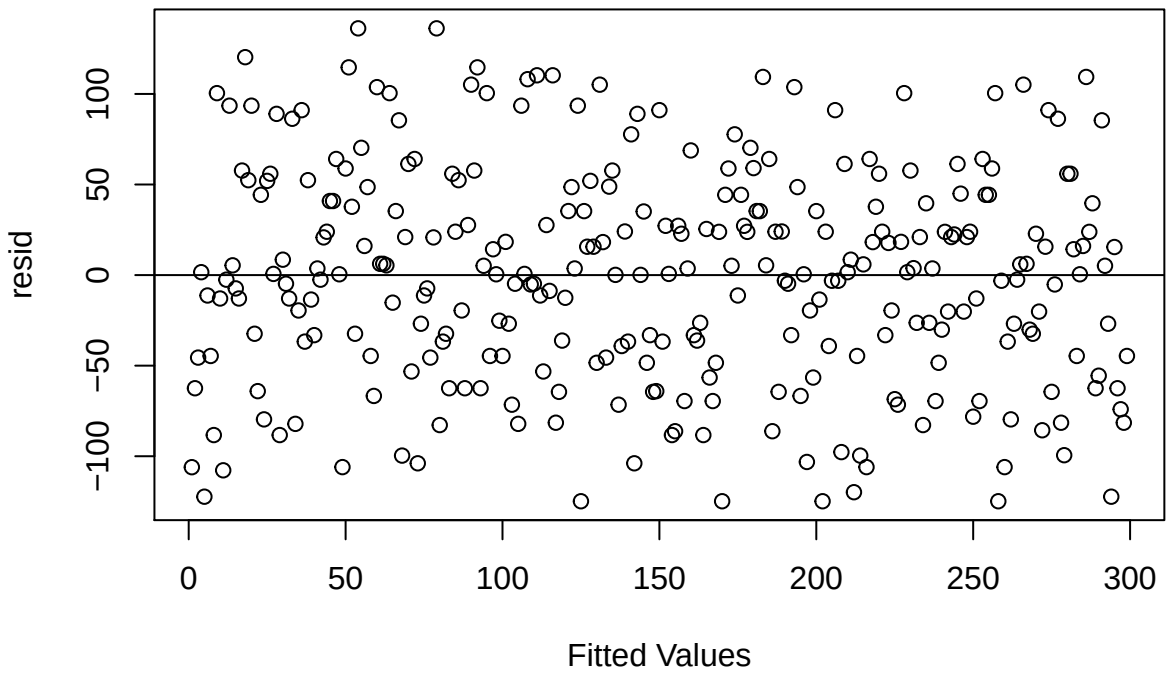


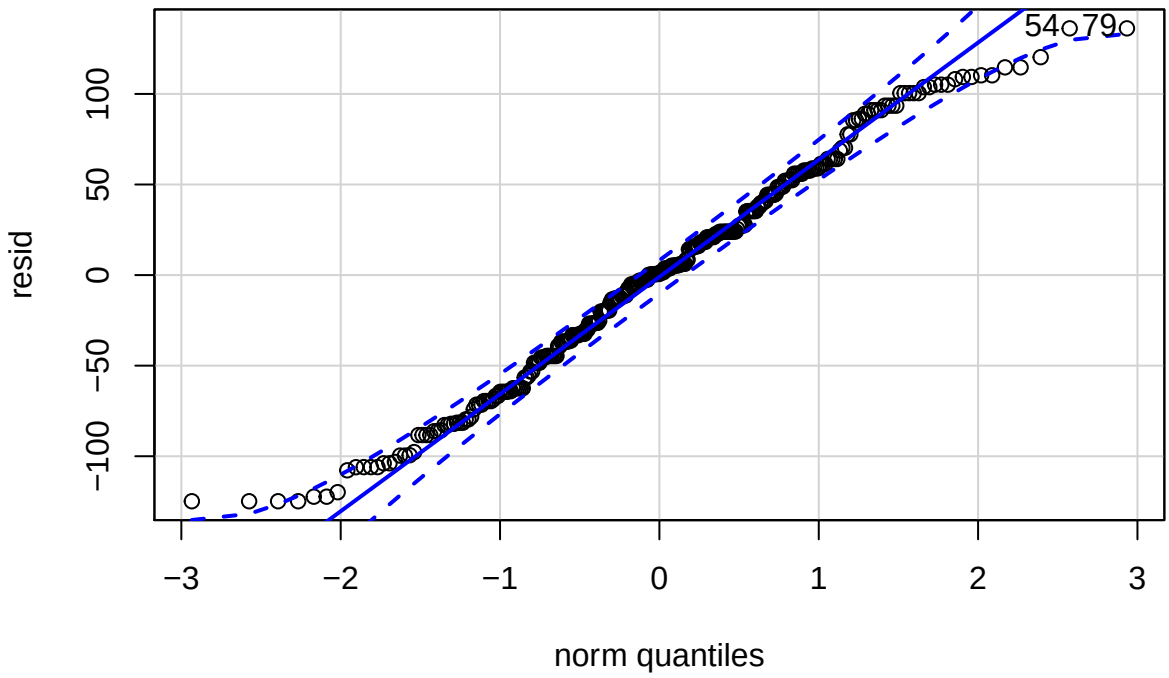
Through the analysis of our graphs, we can determine that age has a negative linear relationship with the year that people move into NYC. The seems to be some positive linear relationship between maintenance deficiencies and the year that the respondent moved in and a negative relationship between the maintenance deficencies and the age of the respondent. There seems to be a strong relationship between age and year that a person has moved into NYC, and the relationship seems to be very similar, so we would want to fomrally check by using variation inflation factors for each of these variables in the full model.

Age	MaintenanceDef	NYCMove
1.6876	1.2677	1.9997

Because all of the variation inflation factors are less than 2.5, there should be no concern about multicollinearity.

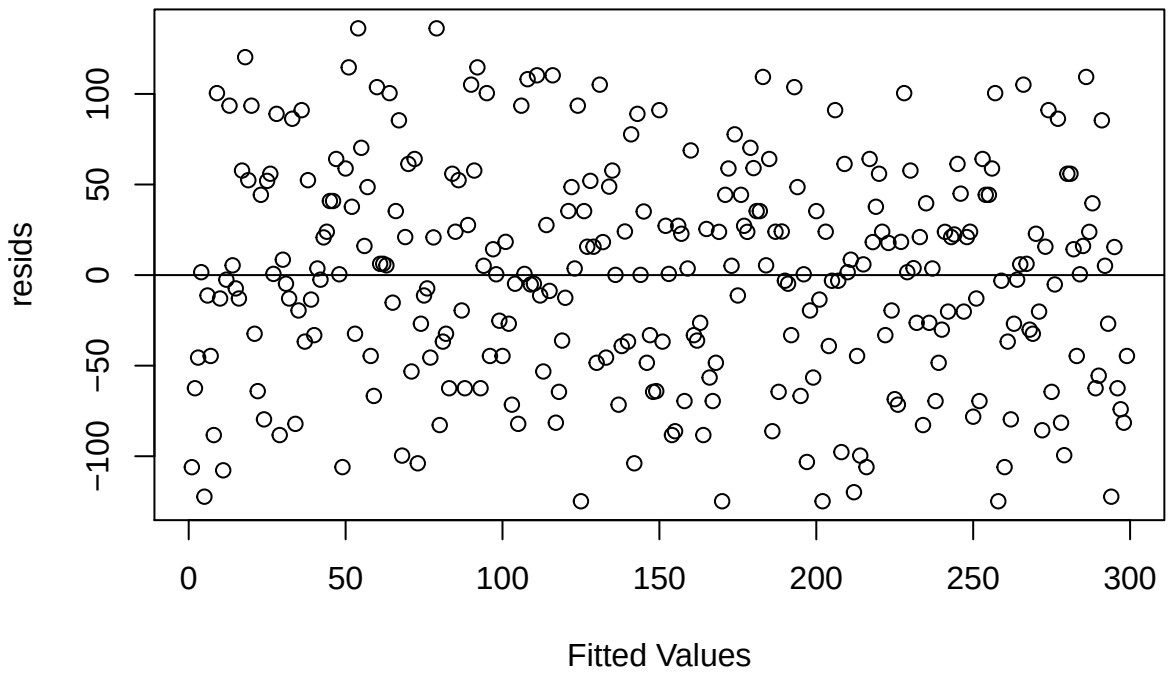
I used the full model and it seems that the zero assumptions are reasonably justified because the residuals have roughly even spread above and below the 0 line and do not have any other behavior. The normaltiy assumption has not been met as not all values fall in the qqplot confidence interval. Some of the values regarding normality do not fit in the range, provided below:

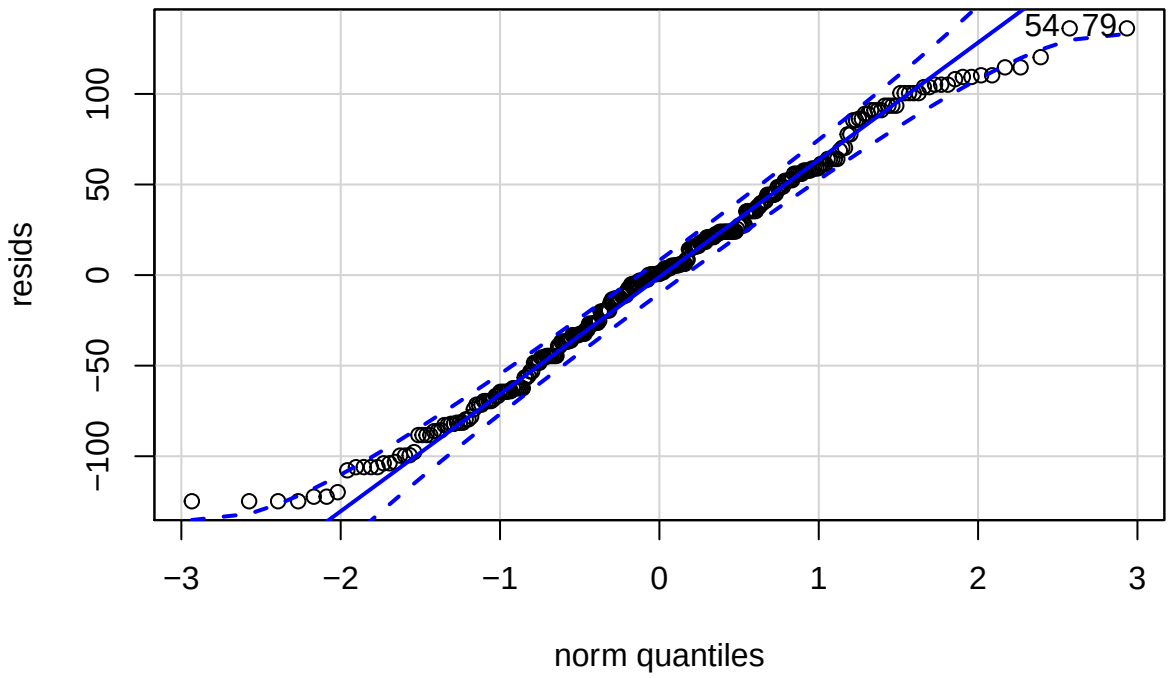

resid
100
50
0
−50
−100
0
50
100
150
200
250
300
Fitted Values




```
## [1] 54 79
```

Below is the new model in regards to taking the square root of the income:


resids
100
50
0
−50
−100
0
50
100
150
200
250
300
Fitted Values




```
## [1] 54 79
```

Note: all of the values in regards to vif have remain the same. The explanatory variables have not changed at all, but the response variable has changed to be the square root of the original. We want to do this as it was not normalized in regards to the respone variable unless we were to take the square root of the original response variable. Zero mean, equal spread, and indepedence have already been met before with the original model, but with the new model that takes the square root of the Income (response variable), zero mean, equal spread, independence, and normalization have all been met. Therefore, we can use this new model in order to predit any values. Here's a summary for these values:

```
##
## Call:
## lm(formula = Income^0.5 ~ Age + MaintenanceDef + NYCMove, data = nyc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.852  -44.651    0.651   42.603  136.212
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    685.5447   699.8086   0.980  0.32808
## Age             -0.2070     0.3637  -0.569  0.56966
## MaintenanceDef  -6.7052     2.4203  -2.770  0.00595 **
## NYCMove         -0.2348     0.3483  -0.674  0.50069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 60.11 on 295 degrees of freedom
## Multiple R-squared:  0.03864,    Adjusted R-squared:  0.02886
## F-statistic: 3.952 on 3 and 295 DF,  p-value: 0.008712
```

Therefore the equation for our final model is:

$\sqrt{Income} = 685.5447 - 0.2070(Age) - 6.7052(MaintenanceDef) - 0.2348(NYCMove)$

Where NYCMove is the year that people have moved into NYC, and MaintenanceDef is the number of Maintenance Deficiencies that they had to encounter. We can note that as a person becomes a year older, the income would decrease by 0.2070 square root dollars; that a person encounters 1 more maintenance deficiency, the income would decrease by 6.7052 square dollars; that a person moves in 1 year later, the income would decrease by 0.2348 square root dollars.

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.

Then for our Prediction Point estimate:

$\sqrt{Income} = 685.5447 - 0.2070(53) - 6.7052(3) - 0.2348(1987)$

$\sqrt{Income} = 187.9105$

With 95% Confidence we know that our mean for a household with three maintenance deficiencies and whose respondent's age is 53 and who moved to NYC in 1987 would be between 178.9614 and 196.7503 square root dollars. There is also a 95% chance that a random selected respondent whose age is 53, who moved to NYC in 1987, and had to deal with three maintenance deficiencies would have an income between 69.22807 dollars square root dollars and 306.4836 square root dollars.

Discussion

For this analysis, we have learned that income is related to the number of Maintenance Deficiencies, the age of the respondent's, and the year that the respondent has moved into NYC. There were no strong multicollinearity issues between any of our predictors, but we needed to transform our income in order to fit the 4 requirements for our prediction, where we took the square root of the income and set that equal to the sum of the predictor variables. All the predictors affected the Income in some shape or form. I thought initially that the age of the respondent and the year that the respondent moved into NYC would be an issue as they should have a strong linear relationship for the two predictor variables, but because their variation inflation factors were not significant, we did not have to remove these varibles from our original model. We had to do was transform our response variable from our original to have our new model. So, it seems that Income is based on the number of Maintenance Deficiencies, the age of the respondent's, and the year that the respondent has moved into NYC. It's hard to determine income based on those factors as there are studies where the factors of income may be due to their personality, their education, and their life style, which are possible factors for the income of people as a whole. Having an analysis of this regarding income is extremely crucial as the three variables listed may not have the strongest relationship in regards to income.