36-401 DA Exam 2

Sarah Li (sarahli)

November 22, 2022

Introduction

Every three years, the New York City government will conduct an extensive Housing and Vacancy Survey on the current housing conditions of NYC. Our research group has been tasked by a member of the watchdog group to analyze the relationship between household incomes and several other demographic and housing quality measurements. **(1)** In particular, they are interested in investigating two questions: (1) whether the average household income is different for the Caucasian population and Hispanic population, and (2) whether the relationship between age and household income is different depending on whether or not leakage has occurred in a resident's apartment. Each of these hypotheses will be conducted controlling for its appropriate respective predictor variables.

The sample we will be working with is from the Manhattan borough, and the raw data consists of 3373 observations with 11 columns. The response variable for our modeling is Income, and the predictors are Age, HeatBreaks (number of heating equipment breakdowns since 2002), MaintenanceDef (number of maintenance deficiencies between 2002 and 2005), Gender (Male/Female), Ethnic (Caucasian, African-American, Hispanic, Asian, Other), Health (ordinal self-assessment of health status), MiceRats, CracksHoles, BrokenPlaster, and WaterLeakage.

Exploratory Data Analysis

(2) We first conduct exploratory data analyses on the ratio variables of the data set. Looking at these variables, we notice a number of truncated values for Income around ~10 million dollars and a few negative values, a total of 81 abnormal observations. After removing them, we decide to work with the data set of 3292 rows and 11 columns. From

the distributions in Fig.1, Income, Age, HeatBreaks, and MaintenanceDef are all skewed right. There is an expected right skew in breaks/deficiencies, as more apartments should deal with fewer problems than not. The right skew in Income follows the typical trend of earnings in the U.S, with the lower and middle class being more saturated than the elitist class. The centers, or medians, for the variables respectively are 47000$ in total house income, 47 years of age, 1 heating equipment breakdown since 2002, and 2 maintenance deficiencies between 2002-2005. Their inclusive (min, max) range values are (600, 1773211) dollars, (19, 90) years, (1, 5) Heating breakdowns, (0, 8) Maintenance Deficiencies.

Figure 1: Histograms of Ratio Variables

(2 cont.) We then conduct EDA on the categorical variables. From the proportions plot in Fig.2, we note that there are more observed male than female tenants; a major proportion of residents are Caucasian followed by Hispanic, African-American, Asian, then a small

proportion of Other ethnicity; there is a higher proportion of no mice/rats reported in the last 90 days, no cracks/holes in interior walls, no broken plaster on ceilings/walls, and no water leaks compared to yes; and most self-reported health scores are between 1-3 compared to 4-6, indicating that most residents have a reported lower health statuses. These comparisons are not surprising.

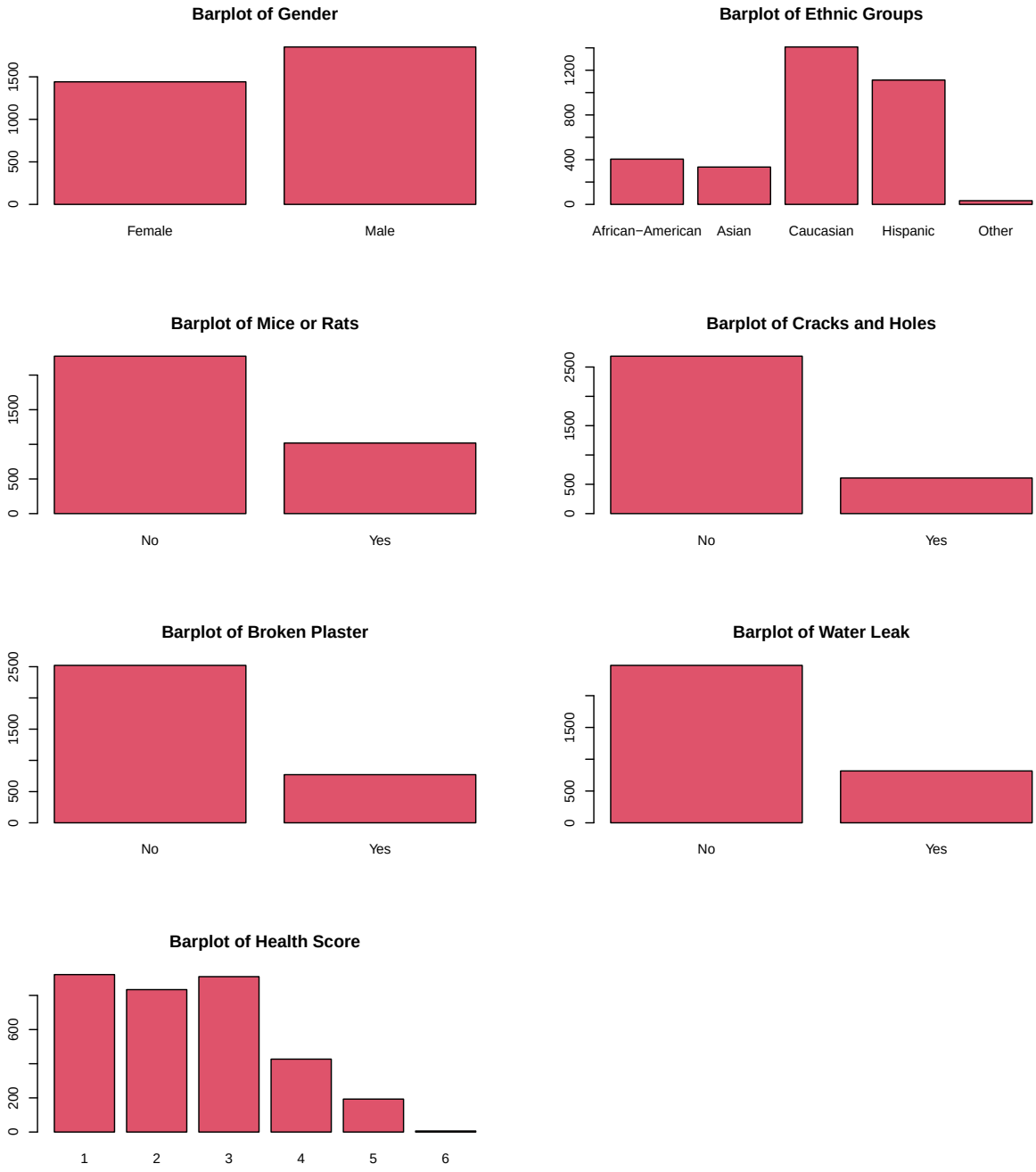

Figure 2: Barplots of Categorical Variables

(3) We then proceed to do multivariate EDA. Since our continuous variables all skewed right, we attempt to log transform them into a more normal shape and compare the linearity of the scatter plots against the response variable Income. In the original pairs plot (results omitted) we observed that Age, HeatBreaks, and Maintenancedef plotted with Income resulted in very heteroskedastic shapes. The points between Income~Age were mostly cluttered around the bottom of the plot, with a few scattered above; the points between Income~HeatBreaks were skewed right, following the trend of the HeatBreaks distribution; Income~MaintenanceDef was also skewed right, following the trend the MaintenanceDef histogram from above. Since linearity is a major assumption of multilinear models, we transform Income and plot the pairs outcome again in the Fig.3. This time, we see that the scatter plots are much more evenly distributed with LogIncome, though still not quite linear. Fortunately, there doesn't seem to be signs of extreme multicollinearity among the predictors variables with one another. Subsequently, after attempting to transform more of the predictors (results omitted), we notice the plots don't change and linearity doesn't improve significantly, so we proceed with log transforming only Income.

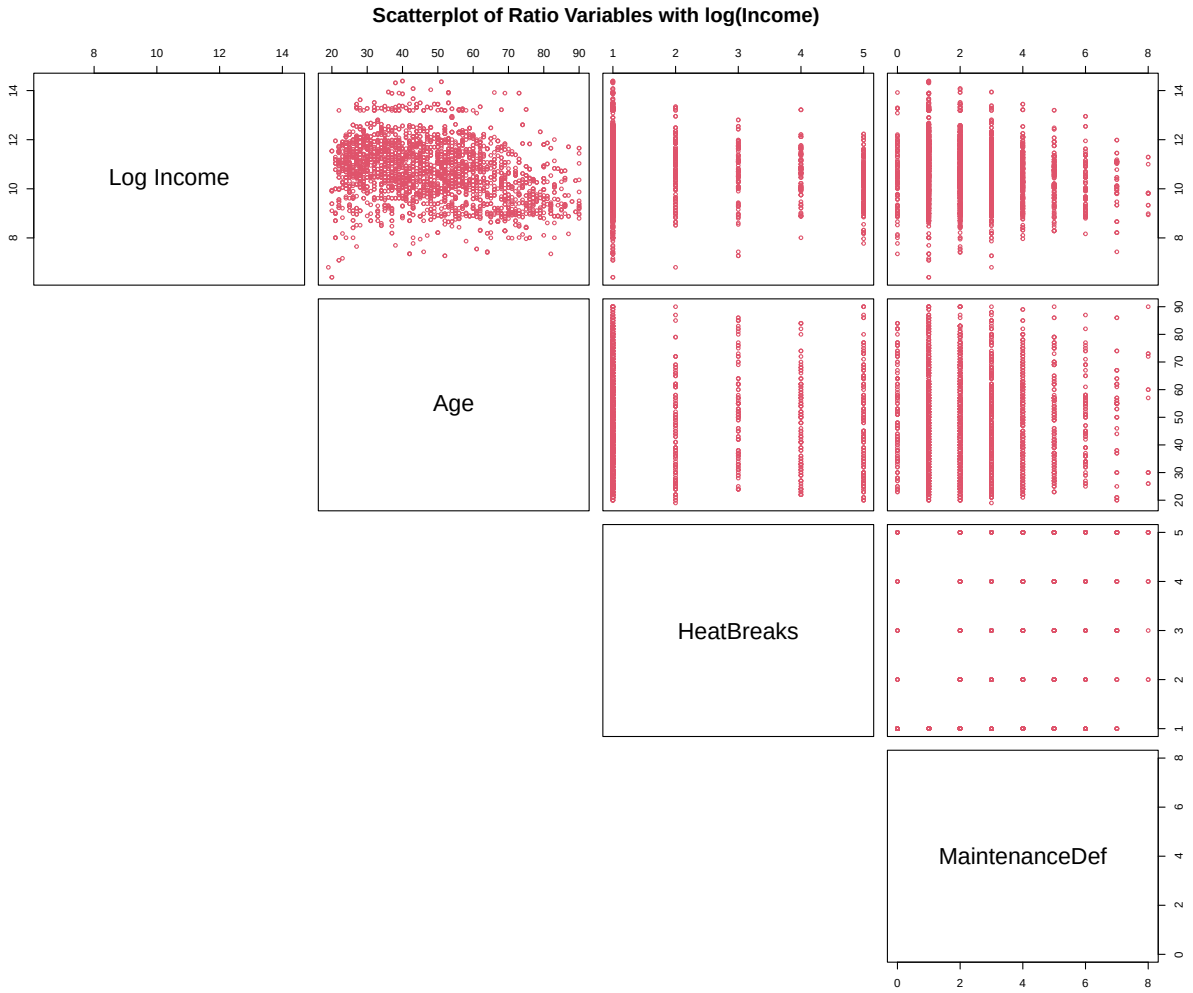

Figure 3: Pairs Plot of Ratio Variables with Log Income

(3 cont.) We follow the same procedure plotting Income with the categorical variables and observe that every single boxplot range and interquartile range were grouped on the lower ends, with a significant amount of outliers with large ranges on the positive ends. After log transforming Income and plotting the side-by-side boxplots with the categorical variables in Fig.4, we confirm that the distributions are much more evenly distributed across Incomes. We point out that, given the data set, females have overall slightly higher Log Incomes than males; Caucasians and Asians have slightly higher Log Incomes than African-Americans and Hispanics; Residents with no rodents, cracks/holes, broken plaster, and water leaks have slightly higher Log Incomes than those with these disturbances. Interestingly, those with reported lower Health Statuses have higher Log Incomes than those with lower ratings. Observing much better distributions, we proceed with Log Income as our response variable for modeling.

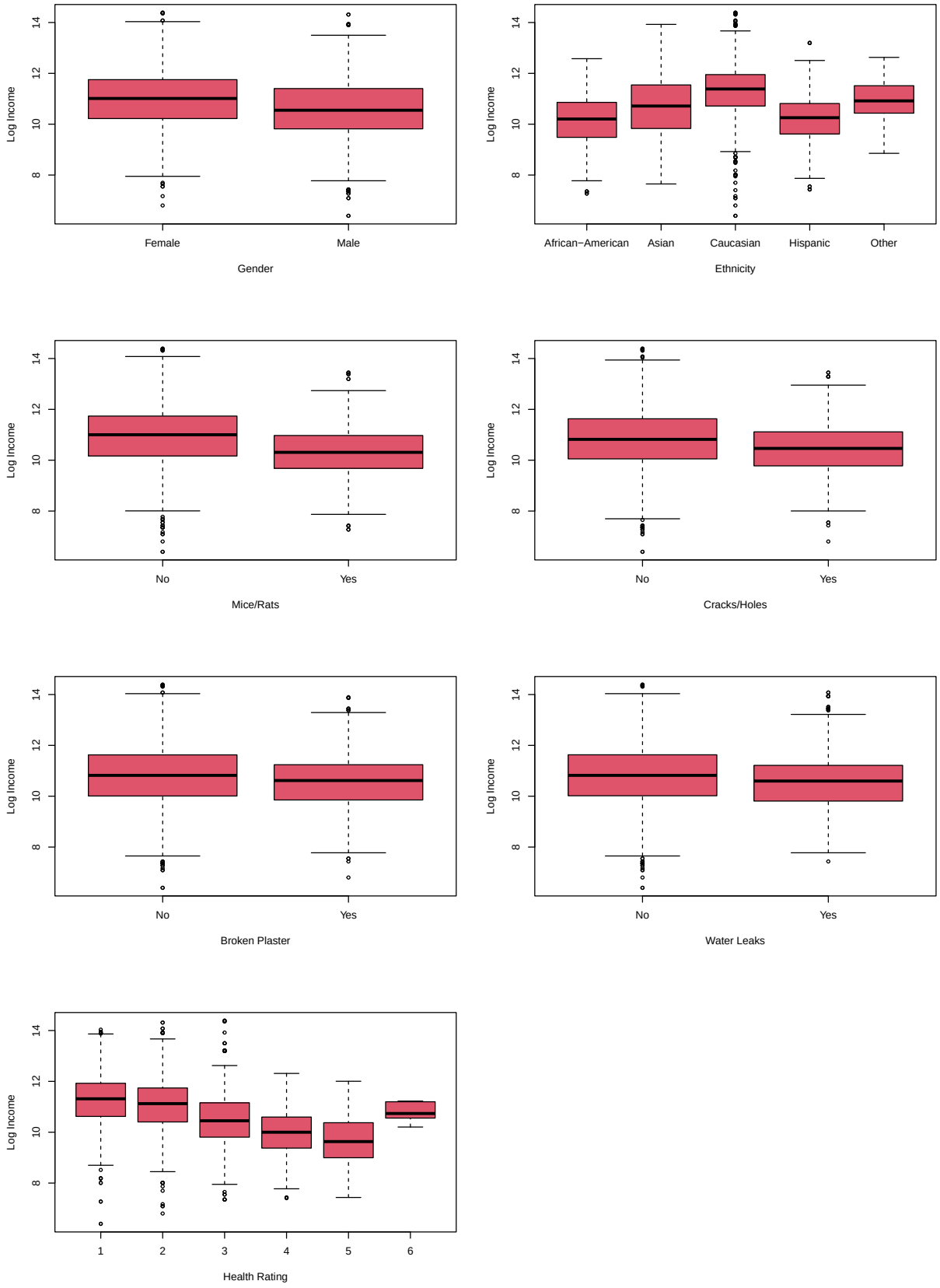

Figure 4: Boxplots of Factor Variables with Log Income

Initial Modeling and Diagnostics

Modeling

(4) We propose an initial model that includes all the predictor variables with the transformed response variable LogIncome:

$$\begin{aligned} LogIncome = \beta_0 + \beta_1 * Age + \beta_2 * HeatBreaks + \beta_3 * MaintenanceDef + \beta_4 * \mathbb{1}_{Gender=M} + \\ \beta_5 * \mathbb{1}_{Ethnic=AfricanAmerican} + \beta_6 * \mathbb{1}_{Ethnic=Asian} + \beta_7 * \mathbb{1}_{Ethnic=Hispanic} + \\ \beta_8 * \mathbb{1}_{Ethnic=Other} + \beta_9 * \mathbb{1}_{Health=2} + \beta_{10} * \mathbb{1}_{Health=3} + \\ \beta_{11} * \mathbb{1}_{Health=4} + \beta_{12} * \mathbb{1}_{Health=5} + \beta_{13} * \mathbb{1}_{Health=6} + \\ \beta_{14} * \mathbb{1}_{MiceRats=Yes} + \beta_{15} * \mathbb{1}_{CracksHoles=Yes} + \beta_{16} * \mathbb{1}_{BrokenPlaster=Yes} + \\ \beta_{17} * \mathbb{1}_{WaterLeakage=Yes} + \epsilon \end{aligned}$$

We treat Income, Age, HeatBreaks, and MaintenanceDef as the ratio or continuous variables. Gender, MiceRats, CracksHoles, BrokenPlaster, and WaterLeakage are all categorical/factor variables with 2 binary levels, while Ethnic is a categorical/factor variable with 5 levels. Health is treated as a ordinal categorical variable with 6 levels, since a higher the score indicates better health. As shown previously, log transforming Income significantly improved the initial model's linearity assumptions in both EDA and BDA. **(5)** We set the baseline for Ethnic to be Caucasian instead of the default. We also acknowledge the research group was interested in the relationship between age and household income depending on water leakages but have decided to omit the interaction between Age and WaterLeakage. These were decided due to statistical reasons below.

Diagnostics

(6)(9) For a multiple linear model to be considered a good fit of the observed data, it must follow a set of assumptions:

1. *The observations must be random, independent and identically distributed.* Indeed, we have no reason to believe that separate observations of residents and their apartments are samples dependently, since the data set is a subset of government official data.

2. *Linearity*. After transforming our response variable Income, we have concluded that the scatterplots of each predictor against LogIncome is evenly distributed but not so noticeably linear. Rather, many of the plots were more uniform than linear in a positive/negative direction. The side-by-side boxplots had many outliers but were roughly symmetric about the median otherwise after transformation. Overall, we are hesitant to confidently say that this assumption was not violated.

3. *The errors/residuals are uncorrelated and follow a normal distribution with mean 0 and* σ^2. Looking at our residuals vs. predictors plot in Fig.5, we observe that the errors are roughly centered around 0 overall. However, we also notice unusual heteroskedastic patterns, such as slight fanning toward the left direction for MaintenenceDef and HeatBreaks, Ethnic, and Health. The residuals of Age have a gap and a few outliers in the positive direction. Looking at the residuals vs. fitted values, we can see that the points moderately follow a homoskedastic scatter about 0. There is a slight divide in the upper right, but it doesn't seem too concerning. We assess the normality of the errors by looking at the q-q plot; there is slight deviation from the line on the lower left but nothing else too abnormal. Overall, there is no strong evidence indicating that the errors are uncorrelated, but we will still keep in mind the abnormal outward fanning.

4. *Constant Variance.* As noted previously, there were slight heteroskedastic patterns (open fanning toward the left) for MaintenenceDef and HeatBreaks, Ethnic, and Health. This goes against the constant variance assumption of the residuals for each predictor that is independent of another predictor. We also assessed the multicollinearity of the predictors using the variation inflation factor test vif() and concluded that the only predictor with a slightly higher vif is MaintenanceDef (3.62). We will consider the multicollinearity of MaintenanceDef with the other predictors in our conclusion. **(8)** Finally, to add on to the plots of residuals, we perform Cook's Distance on the outliers to see if there are any influential points that might have leveraged or increased the data variance.

Looking at the Fig.6, we can see that observations 2970, 2004, and 1344 stand out among the calculations, but we rest assured since the highest value in these sorted quantiles is 3.1892e-16, which is way below the F distribution's median with (18, 3274) degrees of freedom. None of the data points in our multiple linear model are considered influential and leveraging.

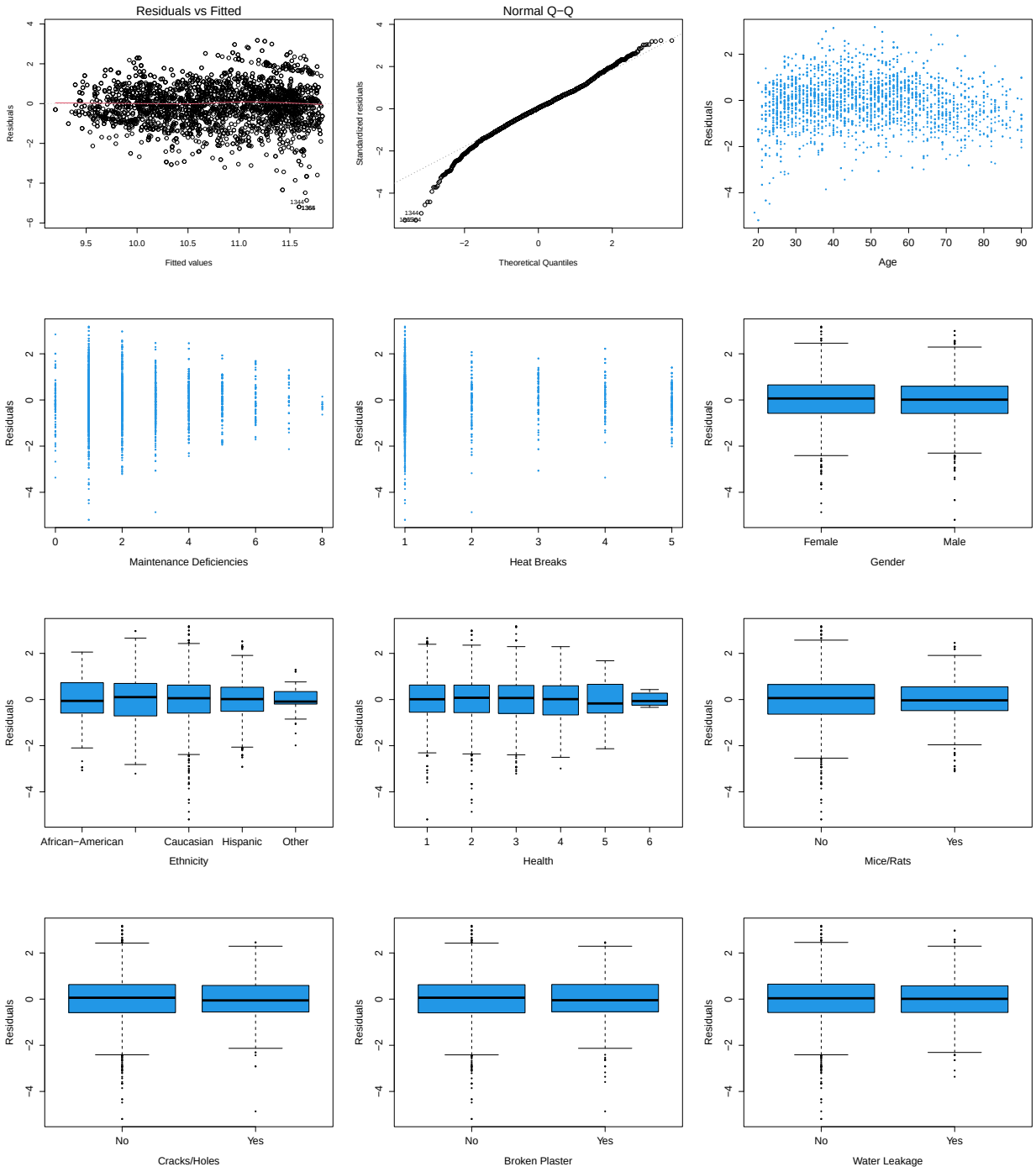

Figure 5: Residuals vs. Fitted for Predictors/Response and Normal Q-Q Plot

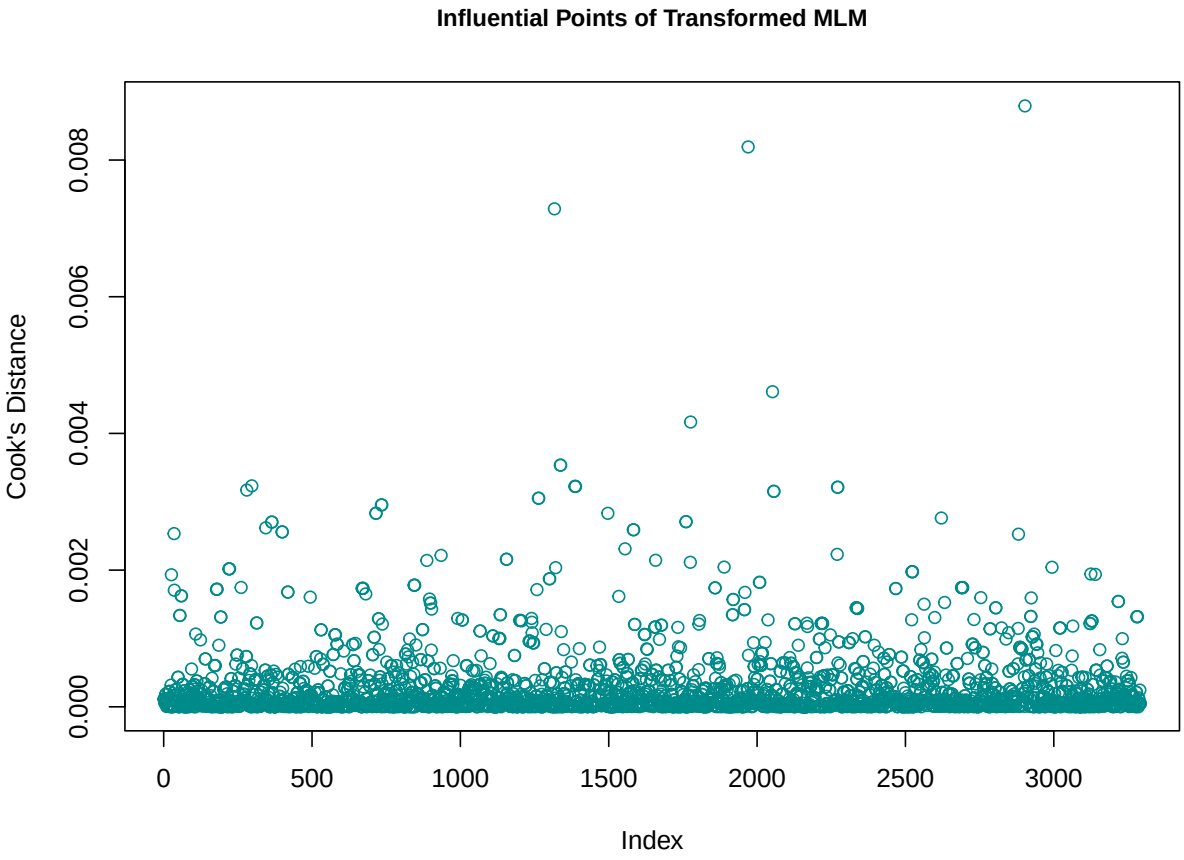

Figure 6: Cook's Distance on the LM

(6 cont.) From the discussions about MLR assumptions, the possible improvements of our diagnostics that we were able to successfully address was transforming Income, which improved the scatterplots of the predictors with Income. **(9 cont.)** However, we could not entirely address the linearity assumption, as the scatterplots still did not display obvious or moderately strong linear relationships, even with further transformations. We also could not address the constant variance assumption, with suspicion of there being multicollinearity of MaintenanceDef with other predictors. **(7)** To see if there were any other significant interactions, we visually inferred from the bivariate analysis conducted beforehand and tried a few models using WaterLeakage as a suspect. We also tried running some interactions with MaintenaceDef given the vif() output; however, we were unable to find any significant interactive relationships using anova.

Model Inference and Results

(10.1) We perform a t-test on the research's group first question of interest: Whether the average household income is different for Caucasian and Hispanic populations when other various demographic and housing quality measurements (the other predictors in our model) are the same.

$$H_0 : \beta_7 = 0$$
$$H_A : \beta_7 \neq 0$$

First, we observe from the summary output of our global F-test that the entire model observes a p-value of 2.2e-16, which is statistically significant; however, the multiple R-squared value is only .2946, indicating that the model does not explain the variation in the data that well. Looking at the levels for Ethnic, using Caucasian as the baseline level, we observe that Hispanic has a t-value of -15.139 and a low p-value (2e-16 < 0.05). Thus, we reject H_0 and conclude that there is a difference in the average household Income for Caucasian versus Hispanic residents, controlling for all other variables.

(10.2) Next, we also perform a hypothesis test regarding the research group's second question of interest – that the relationship between age and income is different depending on the occurrence of water leakages in the apartment, for households sharing the same characteristics (predictors). We would add a new β coefficient to our initial model: $\beta_{18} * Age * \mathbb{I}_{WaterLeakge=Yes}$ and perform a test on the following set of hypotheses.

$$H_0 : \beta_{18} = 0$$
$$H_A : \beta_{18} \neq 0$$

Our summary shows that the global F-test results remain about the same. However, looking at the interaction term $Age : WaterLeakeage = Yes$, we observe a t-value of 0.020 and a very high p-value of 0.98424>0.05. Thus, we fail to reject the null hypothesis and cannot conclude that adding the interaction term significant improved the model. There is no evidence suggesting a difference in the slopes of the relationship between Age and LogIncome for residences with and without Waterleakages, controlling for the other predictors.

(11) Though our Global F-test p-value was regarded as statistically significant, there are metrics to obtain a model focusing on high predictive power given our dataset. We approach this alternative model with an exhaustive search using Aikake's Information Criterion (AIC), which measures how well the new model can generalize to unseen data and combat possible over fitting. As a result, the bestglm() decides to keep Age, HeatBreaks, Gender, Ethnic, Health (African America, Asian, and Hispanic), and MiceRats in the new model. The new model omits the level Health (Other), MaintenanceDef, CracksHoles, BrokenPlaster, and WaterLeakage. The predictors that were kept matched every predictor with very low Beta coefficient p-values in the anova summary output.

Conclusion and Discussion

(12) To reiterate, the research team was particularly interested in the household income for Caucasian versus Hispanic populations, as well as whether the relationship between Income and Age differed with the occurrences of water leakages within the apartments. From our hypotheses tests, we obtained evidence for the former but not the latter. The explanation for these findings can be attributed to the fact, generally in the US, Caucasians tend to have higher earnings than the other populations, and there have been no past findings suggesting Age and Income to differ with water leakages. Prior to the tests, we decided to immediately transform Income to LogIncome to better handle linearity. Additionally, to address how well our full model handled the hypotheses and multiple linear assumptions, we refer back to the bivariate analyses and residual plots. In summary, the observations are i.i.d. and contained no outliers; the individual plots of each predictor with LogIncome were not apparently linear but instead potentially uniform; most of

the error plots for the predictors were homoskedastic about 0, some with non constant variance and heteroskedastic/fanning toward the left; the residuals were roughly normal with mean 0. Overall, none of the assumptions were extremely violated especially given the large sample, and, thus, the model fit is usable.

(12 cont.) Though the AIC model chooses the best predictors in the multiple linear regression model to handle the predictive power for LogIncome on testing data, we still abide by the model that included all predictors written in Initial Modeling. Though the p-values of the omitted predictors in the AIC model – MaintenanceDef, CracksHoles, BrokenPlaster, and WaterLeakage – had high p-values in the anova summary output, that does not imply that these predictors are insignificant and should be removed. As demonstrated by the vif() multicollinearity analysis, MaintenanceDef had the potential to be correlated with the other predictors. This can attribute to the higher p-values for the omitted predictors, as well as the heteroskedastic shapes of their residual plots. The potential collinearity may even improve the model. However, we also keep in mind the trade-off of keeping 17 variables in the model with over fitting. Researchers who care more about predictive power unseen data rather than model fit of original data should prefer the simpler AIC model.