Stats Final Project

Connor Stevens and Sunny Lee

April 2021

1 Introduction

For our MATH2200 Final Project, we took a look at a housing data set from the late 2000s. This housing set had over 2,000 houses and 26 different variables that we could choose from. Ultimately we decided test variables and whether or not they affected the sales price of houses in the area.

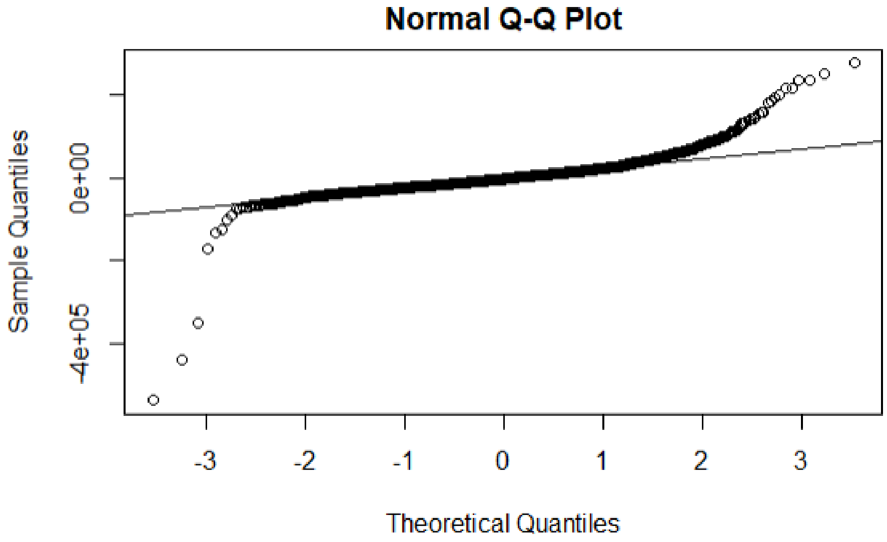
2 Data Description

In a four year span from 2006 to 2010, 2425 single-family households sold in a Midwestern town. As those houses sold, records were kept on different kinds of values for the house, (such as sale price, the condition it was in, the size of the lot, etc.). With these variables, we are able to run a multiple linear regression model and non-parametric tests on our given data. This will allow us to see what variables had the most effect on sale price and whether the information given to us is enough. As we do these tests and find these statistics, we must remember that in 2008 there was a big housing market crash here in the United States. Since our time period for this data spans over that time, some of our data finding may not be useful to use towards today's housing market.

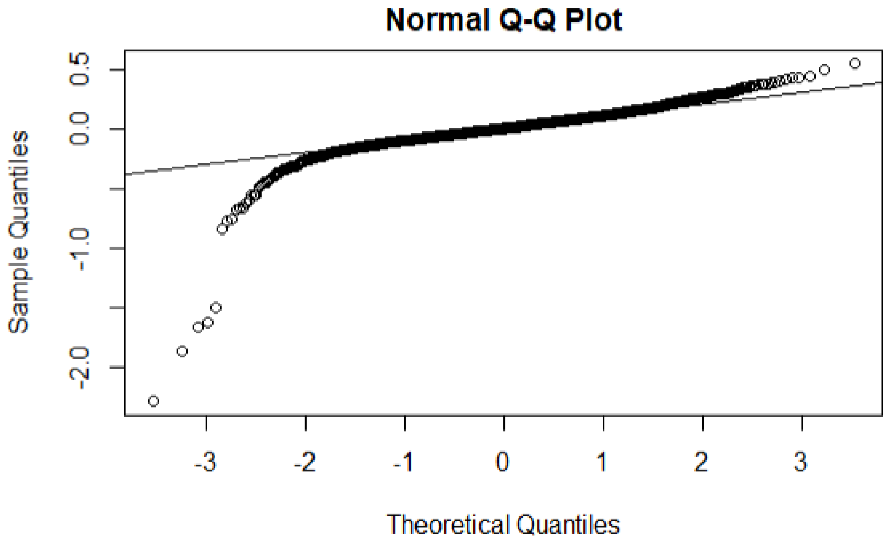
Before running any tests or fitting models to our data, we first wanted to get an idea of what we were dealing with. So we ran a summary on our housing data and found some interesting finds. First, we can see that the cheapest house sold still sold for \$12,789 while the most expensive house sold for \$755,000. We can also see that the average year houses were sold was 2008, right around the housing market crash.

3 Analysis - Linear Regression

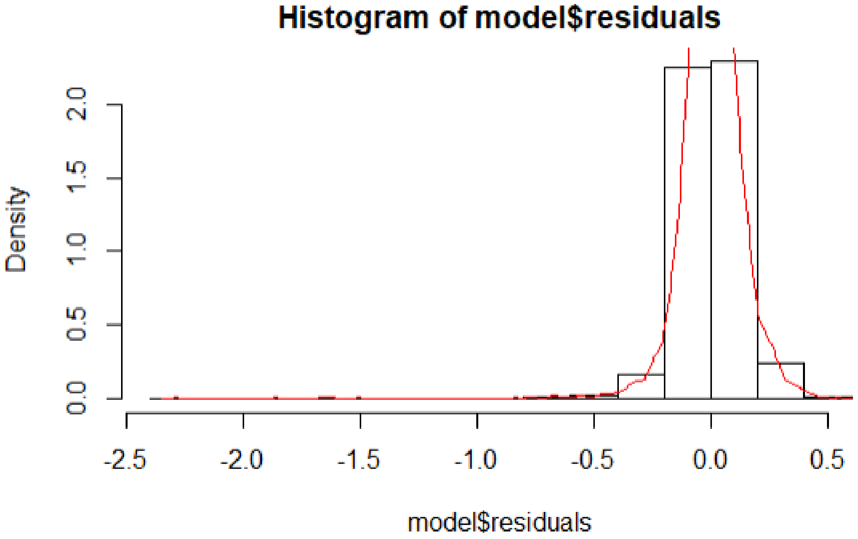
Our first scenario we wanted to model was what were the significant predictors for sale price of a house. To find out, we used multiple linear regression model. First, we fit a linear model to sale price vs every other predictor. We then examined the p-values for each variable to determine whether or not it was significant. We also took note of the adjusted r-squared value for this model, which was 0.8458. After putting all our significant predictors in a linear model, we got an adjusted r-squared value of 0.8189. This is not a big drop-off for the r-squared value considering the fact we got rid of 11 predictors. Then, we wanted to check for outliers to make sure we had the best transformation of the data we could. So, we plotted the residuals on a Quartile-Quartile Norm plot to see if there were any outliers.



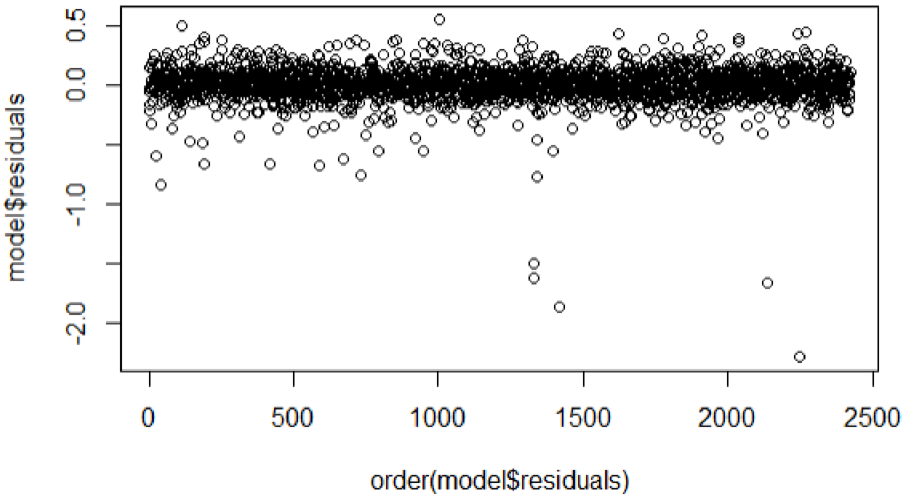
Sure enough, we saw a lot of points fall below and above our normal line. This indicated we should try to transform our response variable. After square rooting sales price and taking the logarithm of sales price, we decided the outliers were most reduced when taking the log of sales price. Doing this bumped up our adjusted r-squared value to 0.8676.



After making sure our predictors were significant, we started testing our assumptions for linear regression. We plotted a histogram for the residuals to make sure they were normally distributed. From the histogram, it looked fairly normal, except for a few outliers.

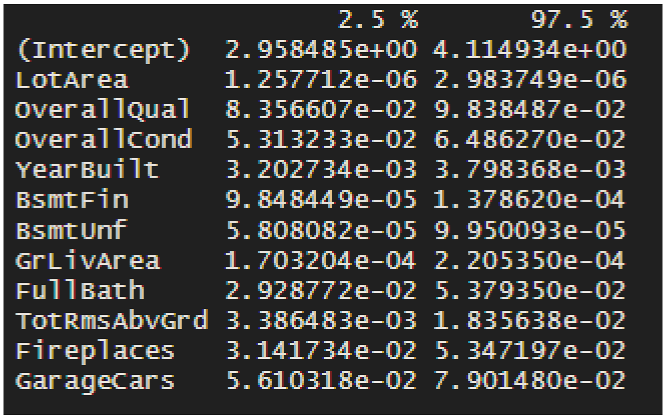


So we could now go check and see if the residuals were independent. To do so, we plotted the residuals against the order of residuals. From the plot we can see that the seems to be no clear pattern among the residuals.



Thus, we can conclude that the residuals were independent. Finally, we needed to make sure there was no multi-collinearity going on within our predictors. After running the VIF command in R, we could see that there was no multi-collinearity occurred between any of our variables. Our hypothesis test is that our H_0 is no coefficients are 0. Alternatively, our alternative hypothesis, H_a is that at least one coefficient is 0. Thus, we could finally start our analysis on our linear model.

Looking at our linear model, we can see that every single predictor increases the log cost of a house. This would make sense, considering we would expect the price to increase when the house can hold more or is more recently built. The predictors that change the log price the most, as in they have the highest coefficient, are overall quality, garage cars, and overall condition. Overall quality and condition make sense to affect price of a house, but it did catch our eye that how many cars the garage fit changed the log price more than overall condition. The predictor that changed the log price the least happened to be the area of the lot. Now, we needed to check our hypothesis test. As we could see from our linear model summary, we could see each predictor had a p-value of below 0.05, thus making it significant. To double check our work, we used the confint() command in r to run a 95% confidence interval on all coefficient values.

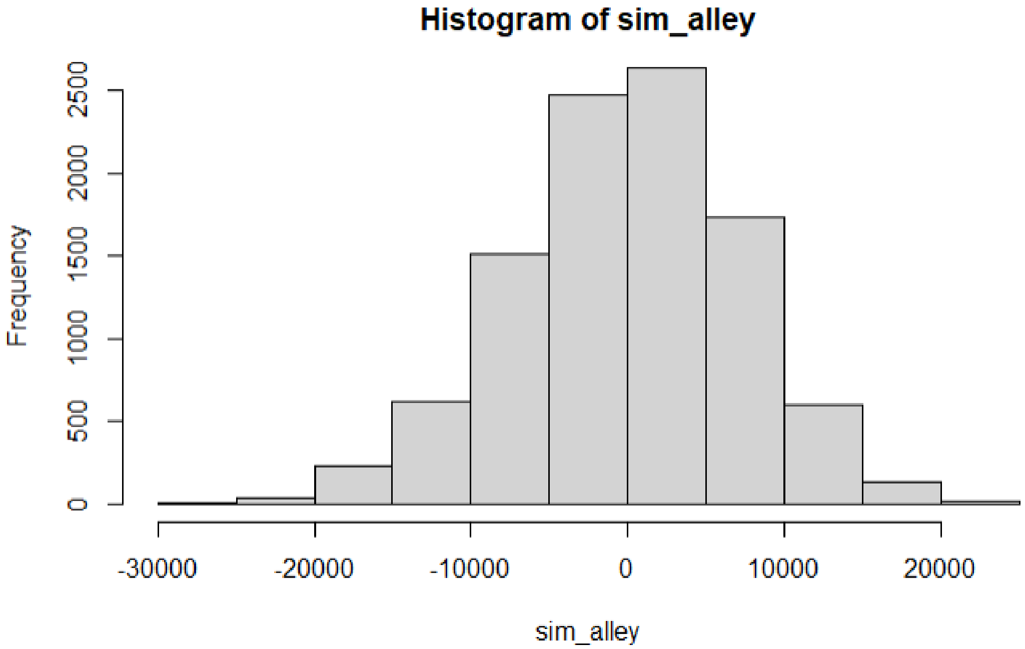


```
                     2.5 %       97.5 %
(Intercept)  2.958485e+00 4.114934e+00
LotArea      1.257712e-06 2.983749e-06
OverallQual  8.356607e-02 9.838487e-02
OverallCond  5.313233e-02 6.486270e-02
YearBuilt    3.202734e-03 3.798368e-03
BsmtFin      9.848449e-05 1.378620e-04
BsmtUnf      5.808082e-05 9.950093e-05
GrLivArea    1.703204e-04 2.205350e-04
FullBath     2.928772e-02 5.379350e-02
TotRmsAbvGrd 3.386483e-03 1.835638e-02
Fireplaces   3.141734e-02 5.347197e-02
GarageCars   5.610318e-02 7.901480e-02
```

After doing so we can conclude that we are 95% confident that the values fall between the two values you see below in the table. Thus, we can confidently conclude that none of our coefficients are equal to 0.

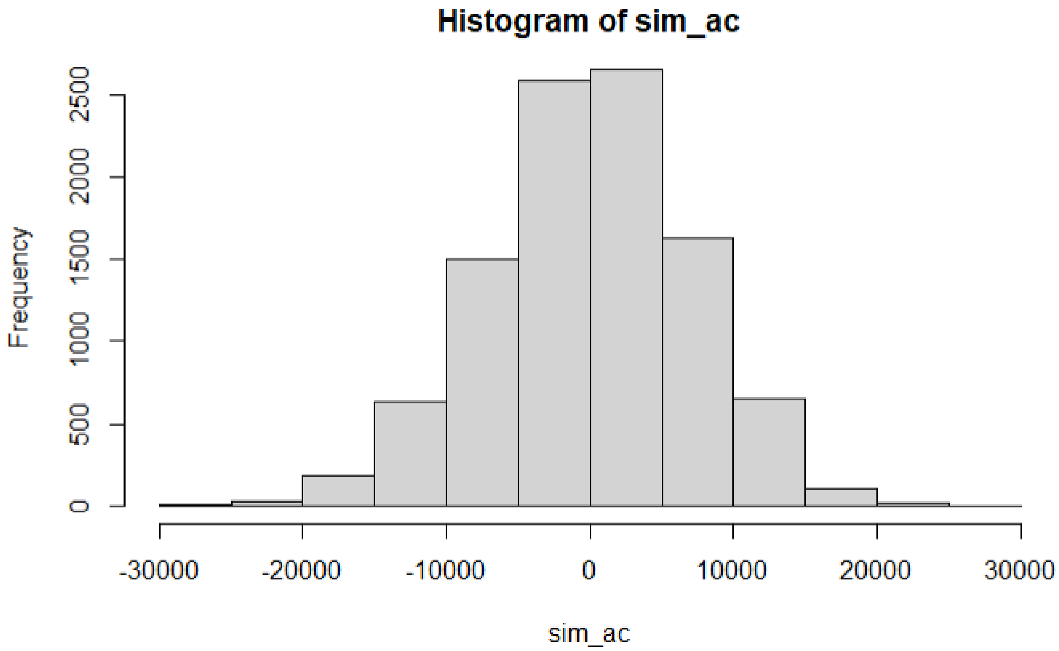
4 Analysis - Non-parametric Tests

For our second scenario, we wanted to check whether alley access, air conditioning and whether a house had a half bath or not had an effect on the sale price of a house. For the alley access, we first split the data and subtract the mean of the sales prices of the houses with alley access and those without alley access which we observe as 50940.21. We then randomly sample without replacement 137, the number of houses with alley access, and the rest, 2288 houses, and subtract the two new means. Doing this simulation 10,000 times, we obtain the following histogram:



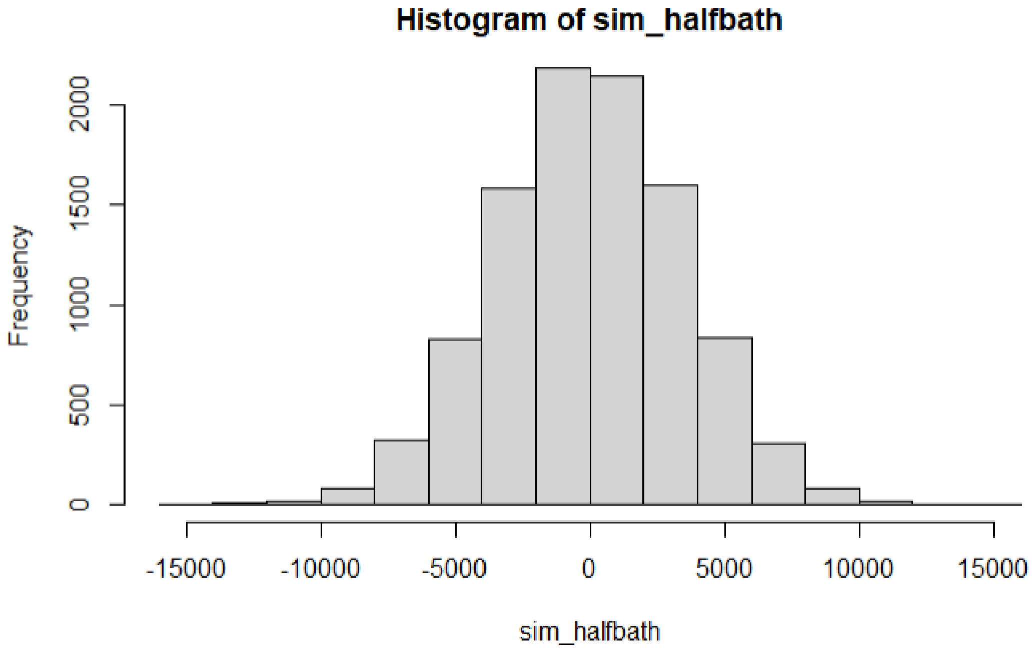
Comparing the simulated means against the observed mean, we observe exactly zero differences in simulated means which are greater than or equal to the observed mean. Thus, our simulated p-value is 0 and since our p-value is zero, we can reject the null hypothesis: $H_o : \mu_{no} = \mu_{yes}$ and accept the alternate hypothesis $H_a : \mu_{no} > \mu_{yes}$. Thus since $\mu_{no} > \mu_{yes}$, we conclude that the sales prices of houses without alley access is higher than those with alley access.

We run the same test except with air conditioning instead of alley access. This time, we observe a mean difference of 90827.8. We also have a different amount of houses with and without air conditioning, and thus in our simulation, we randomly sample 2282 sales prices and 143 sales prices to simulate the differences in our means. Again, running the simulation 10,000 times, we obtain the following histogram:



Since none of the simulated difference in means were greater than our observed mean, 90827.8, we obtain a p-value of zero. Since our simulated p-value is zero, we reject the null hypothesis $H_0 : \mu_{yes} = \mu_{no}$ and accept the alternative hypothesis $H_a : \mu_{yes} > \mu_{no}$. Thus, we conclude that houses with air conditioning on average have a greater sales price than those without.

Finally, we check whether or not houses with one or more half bathrooms on average have a higher sales price. We observe a difference in means of 51322.54 and randomly sample 1385 and 1040 to obtain our simulated means. We run this simulation 10,000 times, and obtain the following histogram:



Exactly no difference in simulated means were above 51322.54, we obtain a p-value of zero and thus we reject the null hypothesis $H_0 : \mu_{yes} = \mu_{no}$ and accept the alternative hypothesis $H_a : \mu_{yes} > \mu_{no}$. Since we reject the null hypothesis, we conclude the average sales prices of houses with a half bathroom is greater than those without.

5 Conclusion

While our linear model has a fairly high adjusted r-squared (0.8676) there are two caveats to it. Firstly, our model does not predict actual sales price for houses. Instead, our model predicts the log of sales price. This is because, to reduce the amount of outliers, we needed to transform our data. Plus, even with the transformation, we could see there were still some low outliers. Secondly, our using R, we found the best model that we could use. Thus, only eleven predictors are present in our model. Overall, our linear model does a fairly good job at predicting the logarithm of sales prices for the houses in that community.

As for our non-parametric tests, no transformations needed to be made because there were no assumptions we needed to abide by. However, we did make a new variable for if a house has a half bath or not. For our non-parametric tests, we were testing to see if three variables that were not included in our linear model still had an effect on the outcome of sales price. We did this by using random sampling on alley access, air conditioning, and our new half bathroom variable. Through these samplings we were able to get very low p-values. Since our p-values were so small, we can reject our null hypothesis that $\mu_{yes} = \mu_{no}$ for all of the three variables. This would lead us to conclude that no alley access, having air conditioning and having a half bathroom generally increases the sales price for houses.