Math 342 Analysis of the Correlations Between House Features and Price in King County

Jonathan Lee Jeffrey Smith Phelix Tang

leej207@wwu.edu smith292@wwu.edu tangp3@wwu.edu

1 Introduction

Over the last couple decades, Seattle has experienced massive population growth. This growth has led to greater demand in housing, which has skyrocketed prices of homes in all of King County. In this analysis, we will be using multilinear regression to test if it is possible to determine which home attributes are significant in predicting prices of homes in the King County area between May 2014-May 2015. Our team has a vested interest in this topic as we are all from the King County area and wish to know if predicting prices of homes is possible given certain variables. Obtaining this information could tell us which home characteristics lead to more costly homes and vice-versa.

The data in use was retrieved from house sales in King County, which includes Seattle from May 2014 to May 2015. No method of data collection was given. There are 21614 samples in the dataset, and each sample includes 21 observations. After reviewing the dataset, we have agreed to filter some of the information for us to accurately interpret the data. The response variable is the price of the house. The predictors are number of bedrooms, number of bathrooms, square-foot of living space, square-foot of the lot, number of floors, waterfront, condition, grade, square-foot above land, square-foot below land, year built, and year renovated. Categorical observations are waterfront, and year renovated. During our exploratory data analysis of the dataset using RStudio, we can assume that there will be some magnitude of multicollinearity between certain observations. Such possible observations with suspected correlation include number of bedrooms, bathrooms, and size of the house, as well as year built and condition.

After our initial interpretation of the data, we are relatively confident that the price of homes can be predicted. We believe price is related to these variables as they describe quantity, quality, and age of the property.

2 Regression analysis

The first step of the regression analysis was to determine which columns (or predictors) should be used as feature vectors. Our initial dataset had 21 columns. We intuitively removed seven columns from the dataset. These columns include: "id," "date," "lat," "long," "zip code," "sqft living15," "sqft lot15." These columns were removed immediately due to several reasons such as unneeded information or lack of description. From our interpretation of the data, several of the columns seemed to be significant – for example, "view" – however the data didn't include a description on how "view" was graded. Due to the lack of quantitative information, "view" was removed from the dataset. After running several tests on the predictors, such as running correlation matrices, testing

for normality and constant variance, and using the T-test, it was decided to include only 11 columns in the final model. These columns were chosen based on features that passed certain requirements for significance in predicting price.

In our exploratory data analysis, we plotted price vs several columns, and through that process it was discovered that there were numerous data points that were extreme outliers. For example, when price vs number of bedrooms was plotted, there was one observation with 33 bedrooms when there should have been just three. Through this process, we were able to process and clean our dataset to remove errors in certain observations that could have skewed the R^2, variance inflation factors, and but not limited to p-values.

The initial starting feature vectors, when doing regression analysis, showed that normality was more of a polynomial curve rather than a linear curve. In attempts to correct the Normal Q-Q plot, we squared a couple vectors individually and found that the best predictor to square was square foot living. Although this didn't fully straighten out the Normal Q-Q plot, it significantly brought the two tails down close to the From there we checked for multicollinearity, and did the VIF test. The only real noticeable thing was the correlation between sqft and sqft2 which was to be expected. After all those changes we decided to use the osslr function to see which subset of our vectors would give us the best results, we found that the model we chose had the highest adjusted R^2 value, and that it overall was the best model out of all the subsets. Our R^2 value is .6778, Which is to be expected. A house price is based solely on whatever the buyer is willing to pay. Plus we didn't have columns like which neighborhood / location the house was in which also plays another huge factor in determining price.

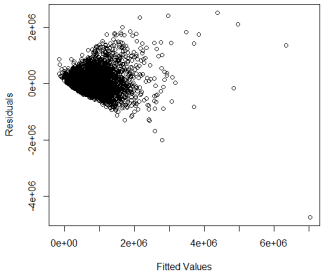

Figure 1: Constant Variance

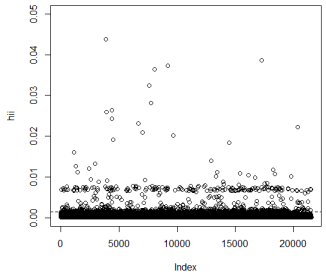

Figure 2: Influential Points

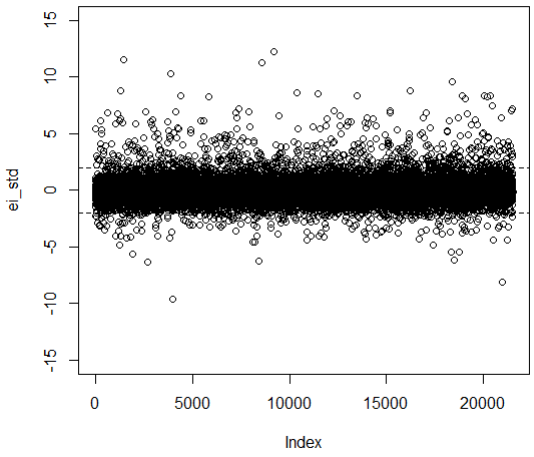

Figure 3: Outliers

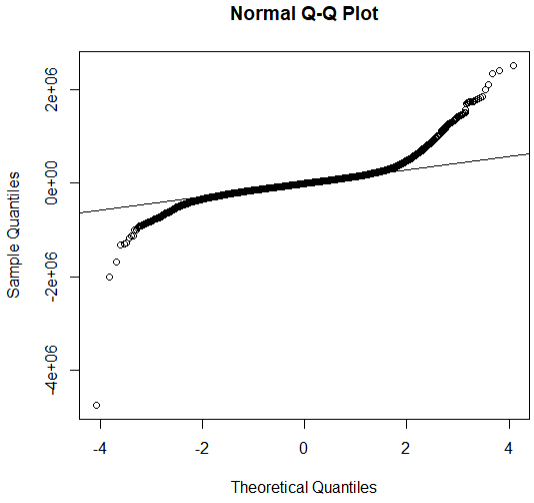

Figure 4: QQNorm

3 Model Diagnostics

Our model has around 22,000 data points, so it's expected to not exactly fit normality and linearity. 99% of the standard deviation of our residuals fit between the 2 and -2 bounds. We've tried our best to get our model to fit the QQplot, but obviously the upper and lower bounds of the house prices won't fit a linear model exactly. Most of our outliers tend to be houses with smaller square footage, that happen to be in a very good location, or by the water which would make sense. I think one of our biggest issues in terms of getting our R^2 value up, would be if we had a column with neighborhoods. A house in a very wealthy neighborhood would have it's price raised up accordingly. Another issue with our data is that since we believe the data was collected via web-scraping, there's bound to be several mis-inputted data points, we've managed to catch a few of them, but with a dataset this large, it's practically impossible to go in and catch every single error.

4 Discussion

The model was used to predict the cost of a 2600 square foot house in King County in 2015 with 4 bedrooms, 2.5 bathrooms, and was built in 2004. This house is also not on a waterfront, has a lot size of 4250 square feet, has not been renovated, and has a condition rating of 4 out of 5, and a grade rating of 8 out of 12. The actual cost of this house in 2015 was $542,000, and the predicted value was $528,548.9. After calculating the prediction interval for this house, we are 95 confident that a house with these parameters costs between $119,708.7 and $937,389.2. The calculated prediction interval covers a shockingly wide range of prices, however, the distinction between the selling price and the value of a house must be drawn. Houses do not always sell at value. This discrepancy can make predicting the selling price of a house difficult from the parameters about size and condition, as the overall market was not included in the model. Location was also not included in this model. Proximity to schools, transportation centers, and other social hubs would be an important factor in predicting the price of a house and should be a parameter investigated for future models.

As stated above, the adjusted R^2 squared value showed the regression model accounted for 67.78% of the variation in the data. This model is not overwhelmingly reliable in terms of predicting the selling prices of houses. These parameters would be more adept at predicting the value of houses.

5 Conclusions

The purpose of this model was to predict the price of a house in King County in 2015 from certain parameters about this house. Since none of the assumptions for linear regression were satisfied, we concluded that this model is not reliable for predicting the price of houses. This is evident in the size of the prediction interval for the investigated house, covering a range over $800,000, which was far over the selling price of the house in question. There were no surprising significant variables in the model. We were surprised to see that the number of floors was not significant in predicting the selling price of houses. We reasoned that the area of the living space was a more adept indicator than the number of floors. This model may be more accurate for predicting the value of houses instead of their selling price. A future test could predict the market value of houses from the parameters used above. New variables such as location could also be used. Predicting the selling price of houses would require another set of variables outlining the state of the market. For the model to be accurate as time progresses, a variable for inflation would need to be present.

6 Appendix

```
library('car')
library('olsrr')

# Housing Data #
df <- read.csv("df.csv", TRUE)
df

# Renovation dummy variable #
df$renovated = ifelse(df$yr_renovated==0,0,1)

# Square foot living adjustment #
df$sqft_living2 = df$sqft_living * df$sqft_living
#df$sqft_living3 = df$sqft_living * df$sqft_living * df$sqft_living

# Model in test #
fit = lm(price~date+bedrooms+bathrooms+sqft_living+sqft_lot+floors+waterfront
+condition+grade+sqft_above+sqft_basement+yr_built, data = df)
# Updated model (took out date and floors) #
fit2 = lm(price~bedrooms+bathrooms+sqft_lot+sqft_living
+waterfront+condition+grade+yr_built+sqft_living2+renovated, data = df)

# Anova analysis #
anova(fit)
anova(fit2)

# Correlation matrix of fit 2)
mat = matrix(c(df$price,df$bedrooms,df$bathrooms,
df$sqft_living,df$sqft_living2,df$sqft_lot,
df$floors,df$waterfront,df$condition,
df$grade,df$yr_built), ncol = 11)
cor(mat)
vif(fit2)

# coefficient interpretation #
fit2$coefficients
sum = summary(fit2)
sum
sum$fstatistic

# normality/constant variance tests #
ei = fit2$residuals
qqnorm(ei)
qqline(ei)

plot(fit2$fitted.values, ei, xlab = "Fitted Values", ylab = "Residuals")
hii = hatvalues(fit2)
bm = 2*(15+1)/21517
plot(hii, ylim = c(0,0.05))
abline(h=bm, lty=2)
```

```
ols_step_best_subset(fit2)

new = data.frame(bedrooms=c(4), bathrooms=c(2.5), sqft_living=c(2600),
waterfront=c(0), condition=c(4), grade=c(8), yr_built=(2004),
sqft_living2=c(6760000), renovated=c(0), sqft_lot=c(4250))
predict(fit2, new)
predict(fit2, new, interval = "predict")
```

```
Analysis of Variance Table

Response: price
               Df     Sum Sq    Mean Sq   F value    Pr(>F)
bedrooms        1 2.8851e+14 2.8851e+14  6632.858 < 2.2e-16 ***
bathrooms       1 5.1867e+14 5.1867e+14 11924.299 < 2.2e-16 ***
sqft_living     1 6.6786e+14 6.6786e+14 15354.362 < 2.2e-16 ***
waterfront      1 1.0146e+14 1.0146e+14  2332.664 < 2.2e-16 ***
condition       1 2.1068e+13 2.1068e+13   484.360 < 2.2e-16 ***
grade           1 1.2052e+14 1.2052e+14  2770.887 < 2.2e-16 ***
yr_built        1 1.5937e+14 1.5937e+14  3663.935 < 2.2e-16 ***
sqft_living2    1 8.7522e+13 8.7522e+13  2012.156 < 2.2e-16 ***
renovated       1 9.6819e+11 9.6819e+11    22.259 2.397e-06 ***
sqft_lot        1 3.0374e+12 3.0374e+12    69.831 < 2.2e-16 ***
Residuals   21506 9.3544e+14 4.3497e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Anova

```
 (Intercept)      bedrooms     bathrooms   sqft_living    waterfront     condition
5.910814e+06 -2.272231e+04  6.017954e+04 -4.688731e+01  7.023893e+05  2.480400e+04
       grade      yr_built  sqft_living2     renovated      sqft_lot
1.410639e+05 -3.386585e+03  3.554957e-02  3.602564e+04 -2.927585e-01
```

Figure 6: Coefficents