

**Fordham University**

**Gabelli School of Business**

**ISGB 799V-001**

**R Statistical Programming**

**Housing Project**

**Final Report**

**Instructor:**

Professor John Campbell

**Group 6:**

Chenhui (Cedric) Li

Xinyue (Katrina) Chen

Yonglin Liu

Bao Ngoc (Jane) Dinh

04/27/2020

1. **Introduction**

The dataset provides us with data on sold houses. For each house, we have the sale price and some characteristics of the houses. Some of the data are structural characteristics of the house (GarageCars, YearBuilt, HalfBath….). Some of the data are quality-related, often judged on a scale of score (Poor-Excellent). We hope to use the data to analyze factors that can affect the Sale Price of a house and create a statistical model with high accuracy to predict Sale Price.

1. **Pre-analysis**

2.1 Summary Statistics

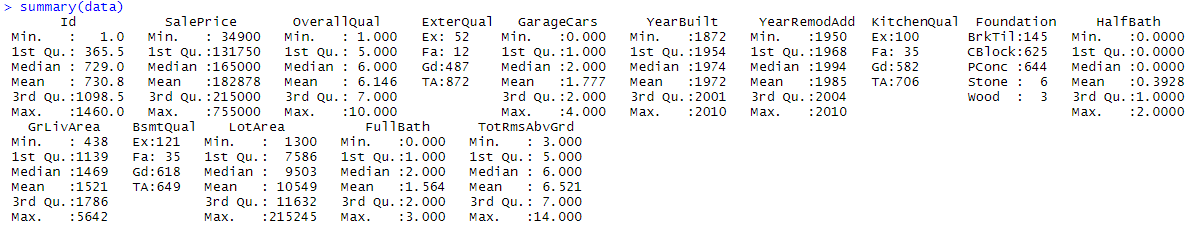


Table 2.1: Summary for All Unprocessed Variable

We aim to select variables that need preprocessing through conducting the summary for all variables. From figure 2.1, there exists a big difference between minimum and maximum value for variables like: GrLivArea, Lot area and SalePrice. Therefore, we did visualizations on these variables to observe the distribution.

2.2 Initial Plots

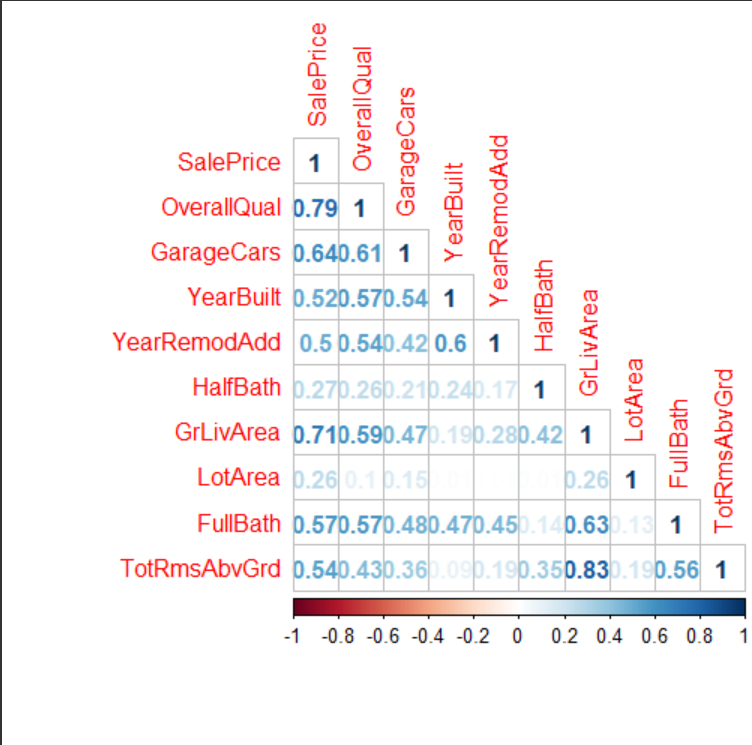


Figure 2.2.1: Heatmap of Correlation Matrix for Numerical Variables

Figure 2.2.1 illustrated which variables had a good correlation with one another. We were mostly interested in variables that correlated well with SalePrice, the variable we were trying to predict. We considered any variable with a correlation of higher than 0.5 a good correlation. Variables that correlate well with SalePrice were: GarageCars, YearBuilt, GrLivArea, FullBath, OverallQual and TotRmsAbvGrd. OverallQual and GrLivArea have the strongest positive correlation with SalePrice and make them likely to be good predictors for SalePrice. GrLivArea also has a strongly correlated with TotRmsAbvGrd and with FullBath, implying a multicollinearity problem.

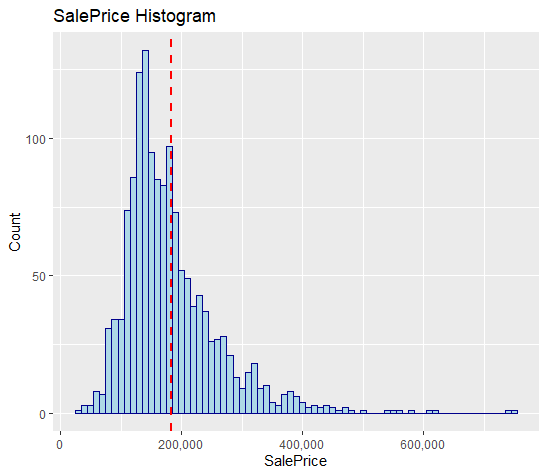


Figure 2.2.2. Histogram of Sales Price

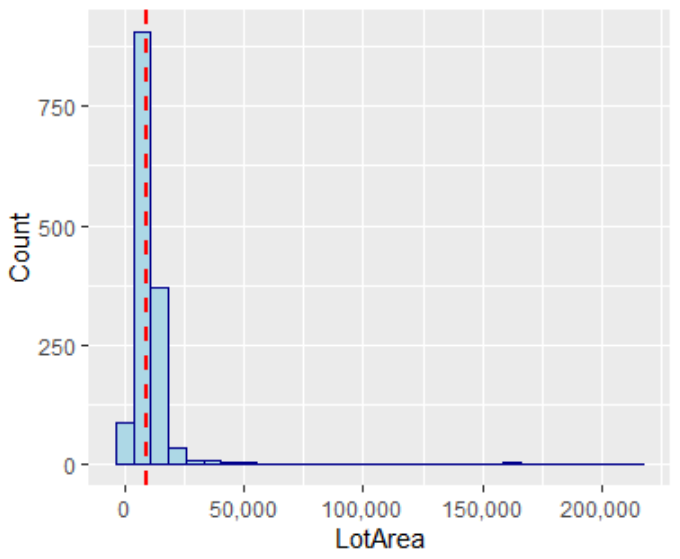
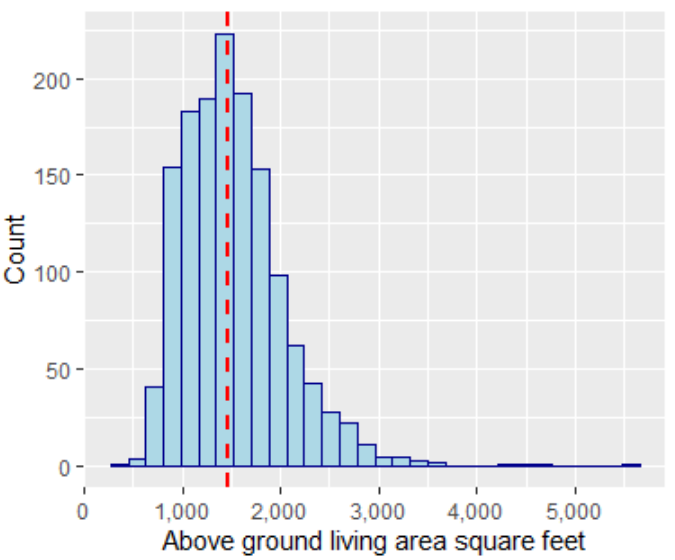


Figure 2.2.3. Histogram of Above ground living area Figure 2.2.4. Histogram of Lot Area

Figures 2.2.2 to 2.2.4 illustrate the frequency distributions of sales price, above ground living area, and lot area. These three features are highly right-skewed as we can see a small number of outliers with values much higher than the third quartiles. The extreme values are usually difficult to predict accurately by the model and may adversely affect the performance of the linear regression model.

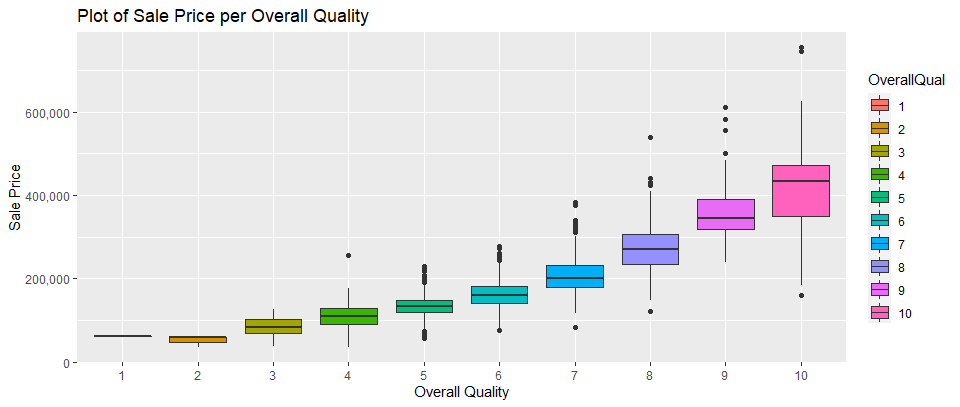


Figure 2.2.5. Plot of Sale Price per Overall Quality

Overall, as the house quality gets better, the sale price gets higher. However, as quality improves, the variance in sale price in the score category varies. In overall quality score 3-8, there isn’t any skew. ‘2’ and ‘10’ for overall quality have left skew- data points varies are further apart below the median. In ‘9’ there was a right skew - data points are further apart above the median.

2.3 Feature Development

We have transformed some of the predictor variables to make them useful to our models. According to the summary statistics, the overall quality variable, by default, was recognized as a character. In fact, the overall quality is a continuous variable ranging from one to ten. Therefore, we changed the data type to numeric, allowing us to explore the correlation with other numeric variables. In addition, we calculated the years since the original construction date and since the remodel date for each house by using the YearBuilt and YearRemodAdd variables. Furthermore, in order to reduce the influence caused by outliers, we drop outliers below 1st and above 99th percentile for variables SalePrice, GrLivArea, Lot Area in the improved linear regression model.

2.4 Initial hypothesis

With the correlation plot and pre-analysis of the data, we believe that not all variables will be important or significant in predicting sale price. We believe that variables that correlate strongly could be the important variables, such as OverallQual and GrLivArea, to predict the sale price. We will be able to tell which one is important and which one isn’t through p-value and other indicators. As we encounter variables that aren’t as helpful, we can either find ways to improve the variable or completely remove them. We also expect that some of our variables will have multicollinearity issues. For example, we have a few variables related to quality of parts of the house. These variables might be highly correlated, which can be detected with Variance Inflation Factors.

1. **Modeling**

Before creating the regression models, we split the data into a training set of 70% and 30% for the testing dataset, split on sale price. This way of partition is a common split as 30% of the 1423 observations ensure there is a reasonable amount of data for testing the models.

3.1 Linear Regression Model

We first generated a linear regression model using all the variables, except for the sale ID and the built and remodel dates as replaced by Age and YearRemodAdd. The sale ID is meaningless and will not add any predictor power to the model.

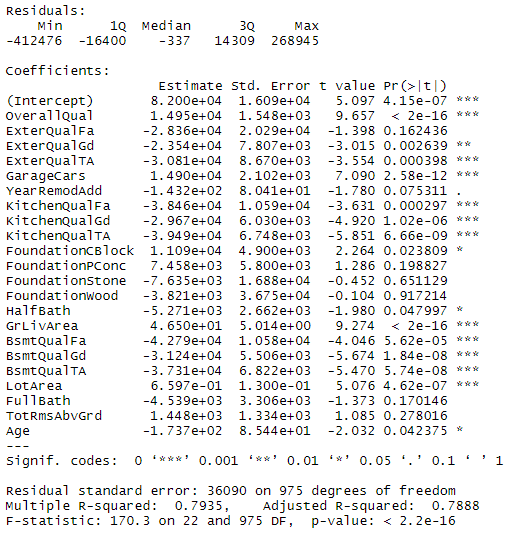


Figure 3.1.1 Model 1 Coefficient output

The adjusted R-squared indicates 78.9% of the variation in sale price can be explained by the variables in the linear regression model. The value is reasonably high as it is not far away from 1, implying the model can explain a fair amount. The F-statistic is far larger than 1, meaning the model fits the data way better than random guessing by predicting the mean of sale price for all the observations.

The intercept value suggests that when all predictors are 0, our predicted value for sale price would be $82,000. The estimate of overall quality suggests that for every one unit increase in the overall quality, the sale price of the house increases by $14,950, holding all other variables constant.

As we can see in the coefficients table, most of the variables have p-values below 0.05, indicating the coefficient estimates of these variables, such as overall quality and age, are statistically significant. But variables such as Foundation, FullBath, and TotRmsAbvGrd have a p-value larger than 0.05, we tried dropping these three variables to see if adjusted R-squared improved. In fact, R-squared for both training and testing data were about the same.

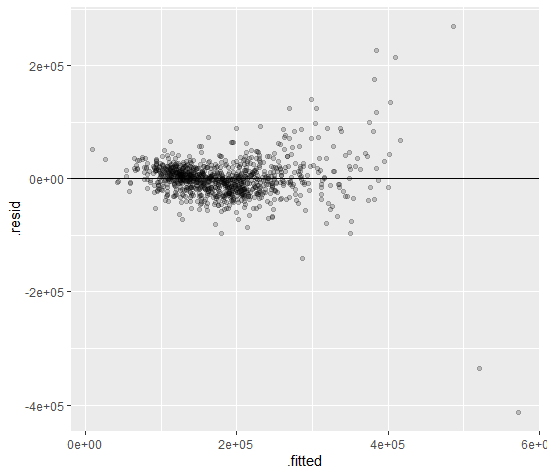
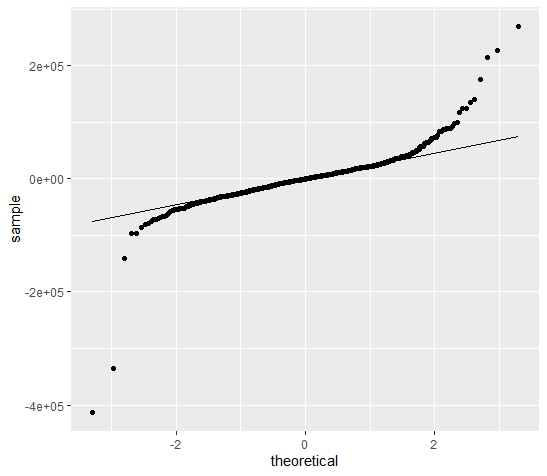


Figure 3.1.2 QQ-plot for Model 1 Figure 3.1.3 Residual Plot for Model 1

To check the four assumptions of the linear regression model, we plotted a QQ plot and residual plot. The above QQ plot showed some non-normality of the model residuals at the bottom left and top right corners of the chart as those dots deviated away from the straight line. The residual plot shows that variances of the residuals are not the same for different values of x, especially for the x values ranging beyond 300,000. The linearity assumption is met because a linear pattern can be observed from the residual plot.

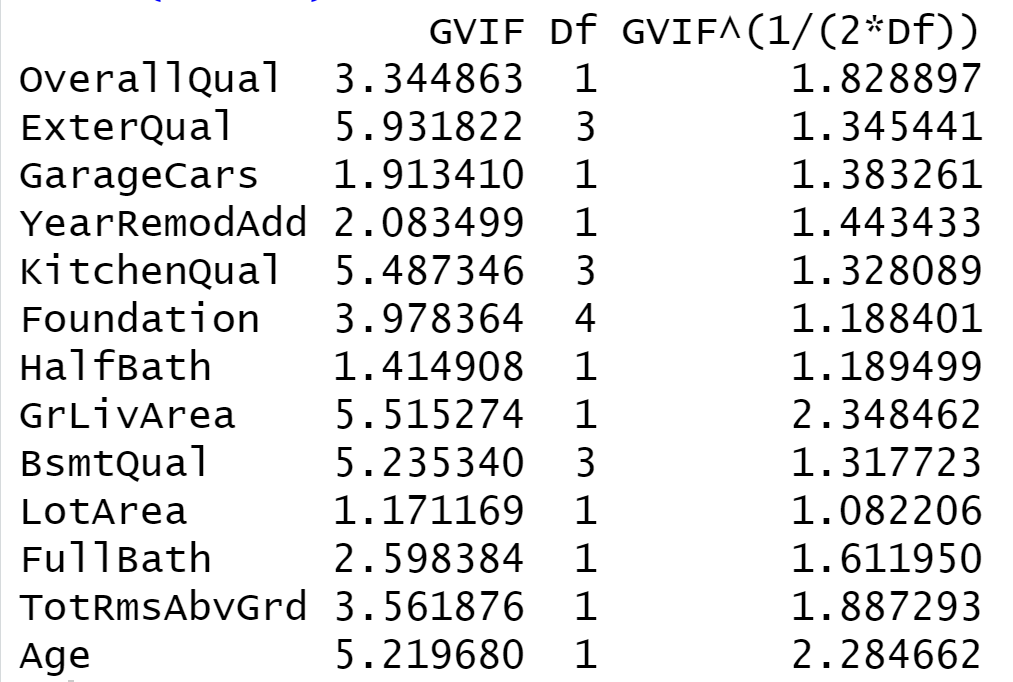


Figure 3.1.4 Variance Inflation Factor for Model 1

We used Variance Inflation Factor (VIF) to check for the multicollinearity issue: no variables have a strong collinearity but some variables with VIF higher than five signal the problem. We tried to adjust the multicollinearity by removing those variables from the model individually and altogether. The partial F test shows that the full model is significantly better than the nested models with some variables removed. The p values less than alpha 0.05 indicates that the complex model significantly improved on the reduced models.

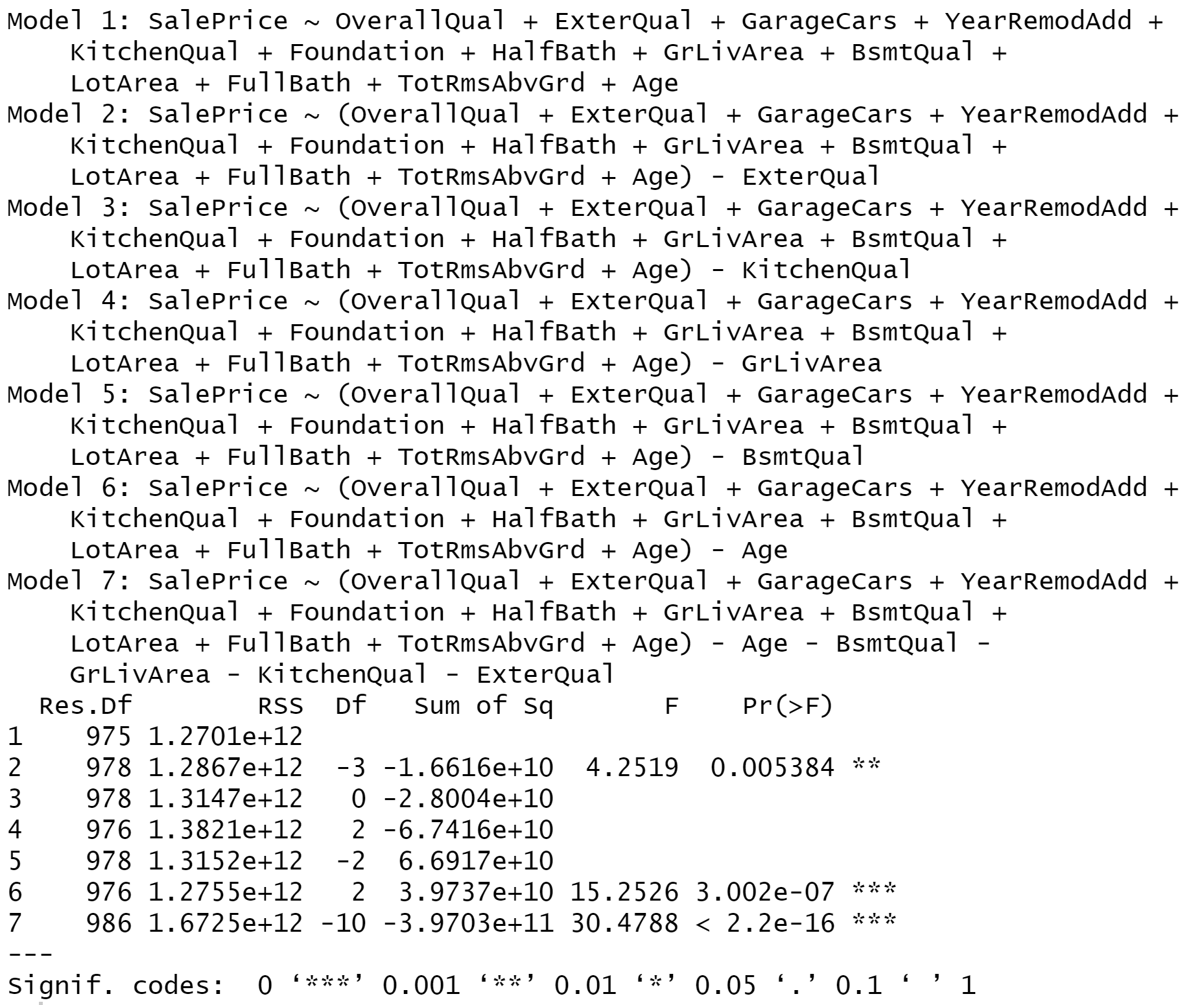


Fig 3.1.5 ANOVA results comparing full and nested models

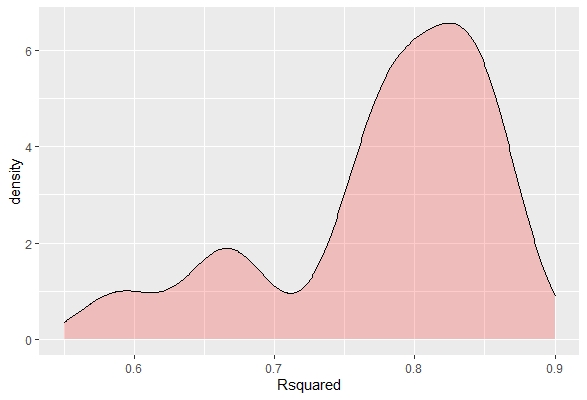
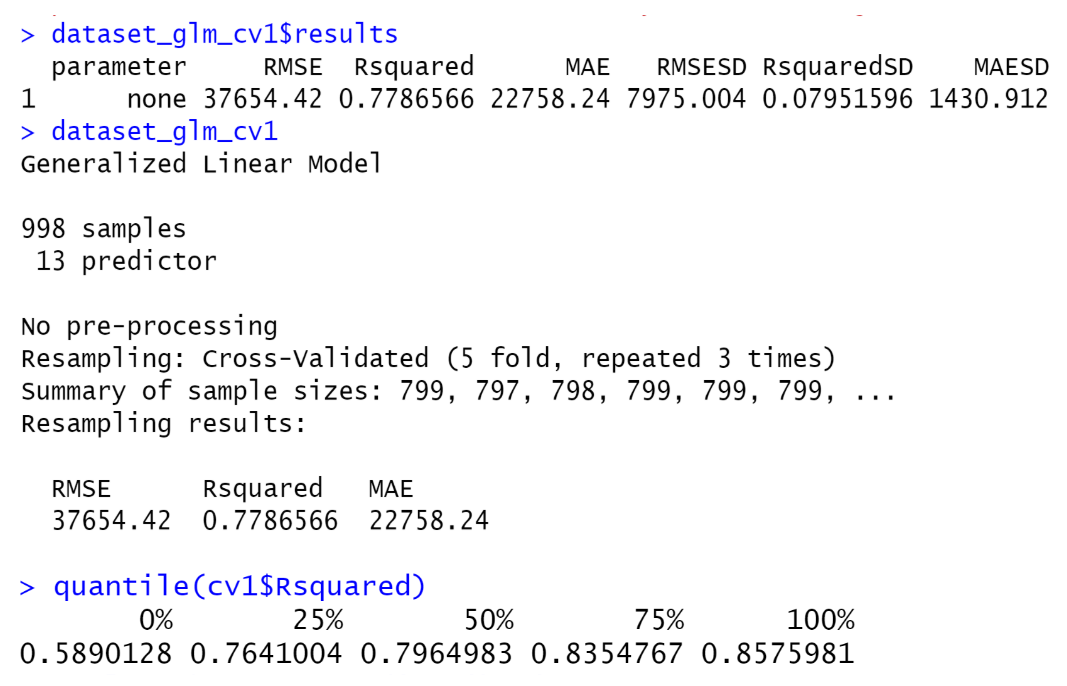


Fig 3.1.6 Model 1 cross validation results

We also used K-fold cross validation specified with five folds repeated three times. We believed setting the number of folds to be five is reasonable because the dataset with about 1400 records is rather small, so splitting into larger fold seems excessive as there will be not many records to train on each time. The validation results show that, in general, we expect the R-squared for model 1 is 77.9%, Root Mean Square Error (RMSE) of around 37654 and Mean Absolute Error (MAE) of 22758. The density plot shows the distribution of R-squared is left-skewed. The R-squared ranged from 58.9% to 85.76%, with a medium of around 79.6%. The R-squared measure on the single test set of 83.15% is in line with the cross validation results since it is between the second and the third quantiles. The higher R-squared on testing data than training suggests that the model does not seem to overfit.

3.2 Linear Regression Model with Improved Dataset

To improve the performance of our linear regression model, we dropped outliers below 1st and above 99th percentile for SalePrice, GrLivArea and LotArea due to their right-skewed distribution. The removal only applied to the training set, and the testing set remained the same. And implemented the linear regression model again with the same variables as the first model.

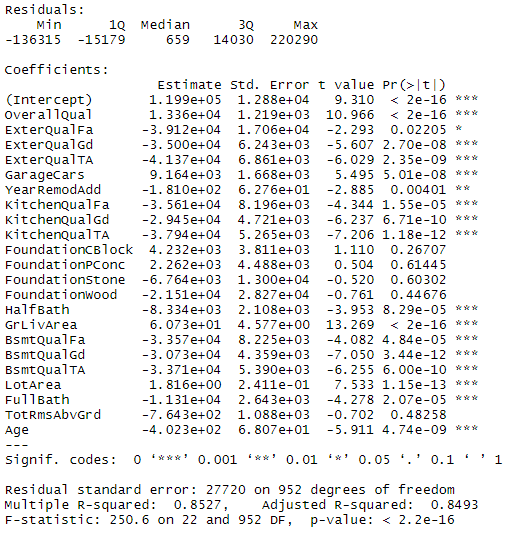


Figure 3.2.1: Model 2 Coefficient Output

The intercept value suggests that when all predictors are 0, our predicted value for sale price would be $119,900. The estimate of overall quality suggests that for every one unit increase in the overall quality, the sale price of the house increases by $13,360, holding all other variables constant.

As we can see in the coefficients table, most of the variables have p-values below 0.05, indicating the coefficient estimates of these variables, such as overall quality and age, are statistically significant.

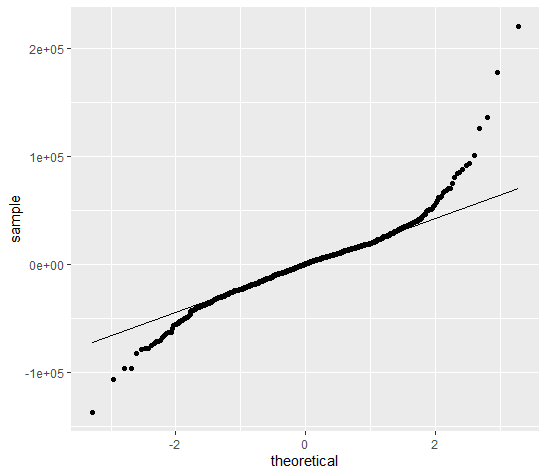
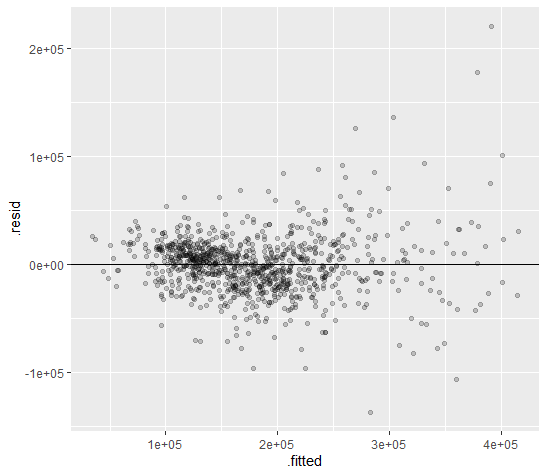
 

Figure 3.2.2: QQ-plot for Model 2 Figure 3.2.3: Residual Plot for Model 2

The problems with the assumptions remained. The QQ plot showed larger non-normality of the residuals. There are residuals at the bottom left and top right corners of the chart as those dots deviated away from the straight line. The residual plot shows that a more constant variances of the residuals as different values of x ranged beyond 300,000. The linearity assumption is met because a linear pattern can be observed from the residual plot.

Besides, we used 5 folds cross validation and 3 repeats to generate linear regression.

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE (Validation)** | **R-square (Validation)** |
| Linear Regression 2 | 32944.57 | 0.835925 |
| Linear Regression 2 (CV) | 28183.38 | 0.844422 |

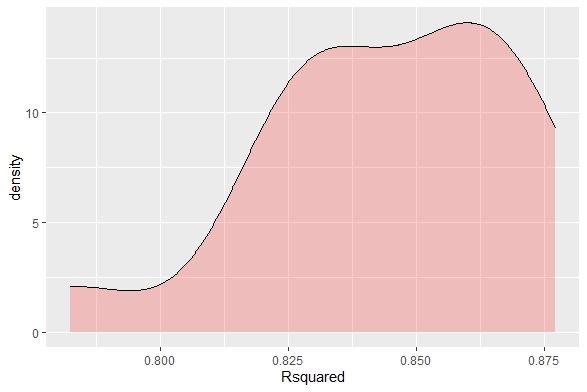


Figure 3.2.4: Model 2 Cross Validation Result

Statistically, we figured the linear regression model with cross validation has better performance based on R-square value and the root-mean square deviation. The reducing value of RMSE infers that our observed data points are closer to the model’s predicted values and indicates a better fit. Linear regression with cross validation has slightly higher average R-square, which indicates 84.4% of the variation in sale price can be explained by the second model. The density plot shows the distribution of R-squared is left-skewed. In general, we can expect the R-square falled in the range between 80% and 87.5%, so the results of model 2 are more consistent. But the R-squared for testing data is lower than that for training, suggesting a slight overfitting issue.

3.3 Ridge Regression

To reduce overfitting and get more consistent results of model 2, we implemented the ridge regression model, which is used to reduce variances by adding bias into regression estimation. Since the result of model 2 showed that the dataset without outliers is relatively better for prediction, we decided to continue our modelings based on the improved dataset. By calling the Glmnet package from R, we set the mode of alpha equal to zero for ridge regression and set the number of folders to five for cross validation. The result of coefficients with “One Standard Error” Lambda is shown below.

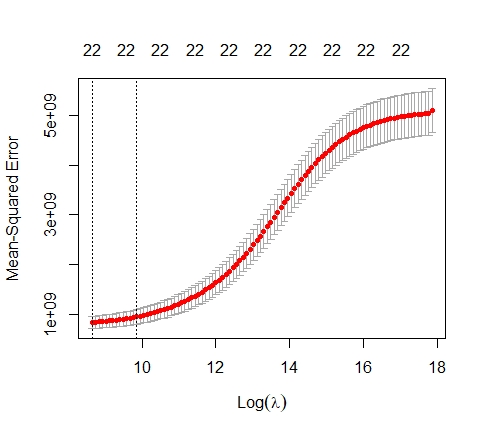
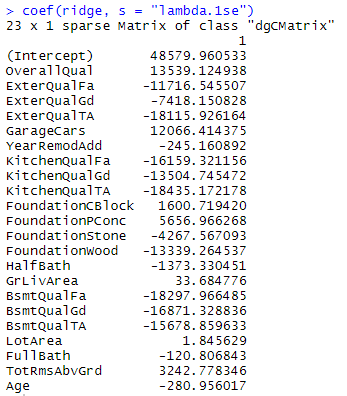
 

Figure 3.3.1: Ridge Regression Output - Lambda (left) and Coefficients (right)

The plot on the left shows the values of lambda, which help identify the values that lead to the lowest average errors across the several held out test sets. The ridge regression model shows the lambda for minimum MSE is around 9, and the largest lambda within one standard deviation from that is slightly below 10.

According to Figure 3.3.1, the intercept value suggests our base value for sale price would be $48,580, which is much lower compared to both linear regression models. KitchenQualTA, the typical/average kitchen quality, has the largest coefficient: in average, if the house has a typical/average kitchen quality, its sales price will drop by $18,435, with all other variables remaining unchanged. The coefficient of LotArea is the smallest, showing that one square feet increase in lot size can only lead to a $1.85 increase in sales price.

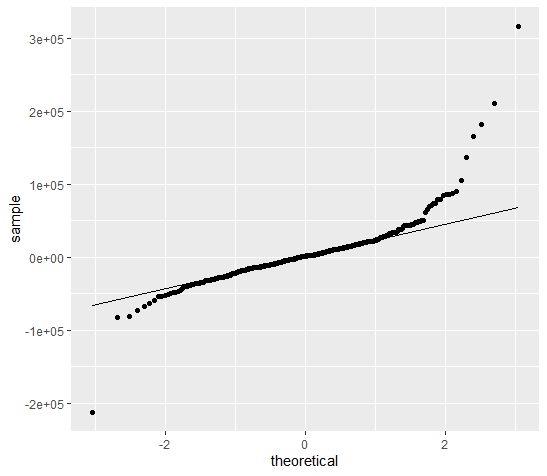
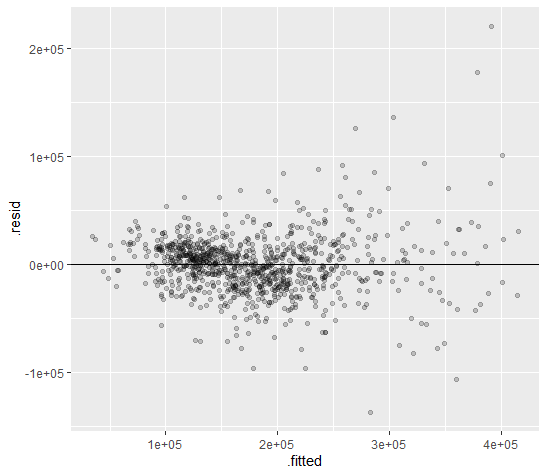
 

Figure 3.3.2: QQ-plot for Ridge Regression Figure 3.3.3: Residual Plot for Ridge Regression

Similar to what we did in linear regression models, we checked the four assumptions of the ridge regression model by plotting the QQ plot and residual plot. The QQ plot on the left showed a non-normality of the model since the tail of the line is sloping upward. Some residuals at the bottom left and top right corners distributed away from the straight line. The residual plot on the right showed that variances of the residuals are not the same for the fitted values over 250,000. The linearity assumption is met in the residual plot because most points are close to the center line and evenly distributed up and down.

3.4 Lasso Regression

In addition to Ridge regression, we implemented Lasso regression. Lasso regression is also a regularization model but works differently compared to ridge regression. We kept calling the glmnet package and setting the number of folders by letting nfold equal to five, but we set the alpha value equal to one instead to change the model to lasso regression. The result in coefficients with “One Standard Error” Lambda is shown below.

The LASSO model shows that the lambda values that lead to the lowest average errors is around 5, and the largest lambda within one standard deviation from that is slightly above 7.

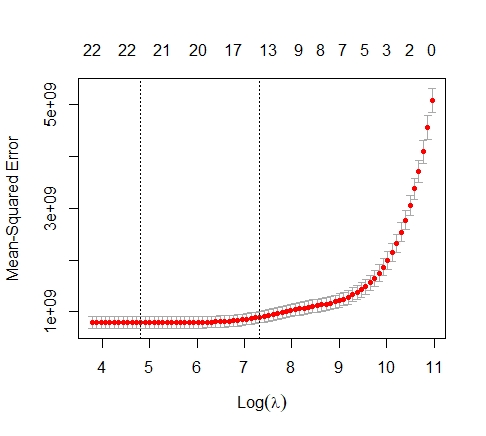
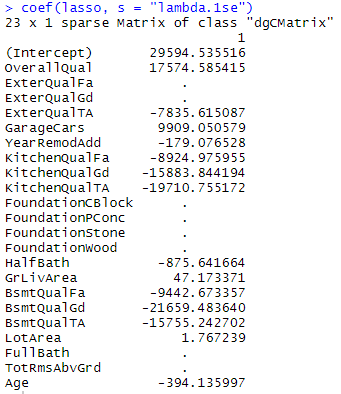
 

Figure 3.4.1: Lasso Regression Output - Lambda (left) and Coefficients (right)

The lasso model simplified our model by shrinking down and eliminating variables to avoid overfitting. We can see that, based on the results in Figure 3.4.1, coefficients of eight variables are zero because they are eliminated during the regularization. The intercept value is $29,595, which is even lower compared to ridge regression. The height of the basement is now having the largest impact on the sales price: the price decreases by $21,659 if the basement height is in good range. The size of the lot still has the smallest impact: one square feet increase in lot size only leads to a $1.77 increase in sales price.

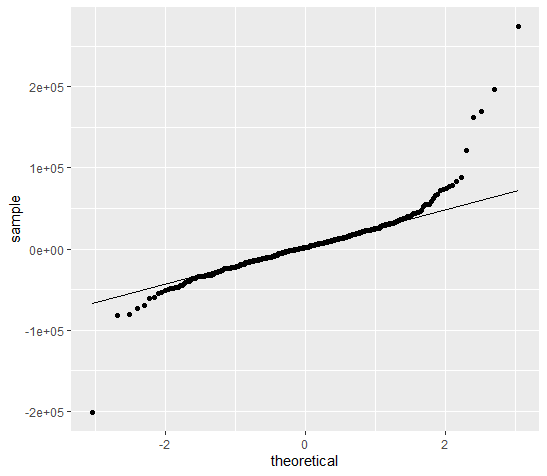
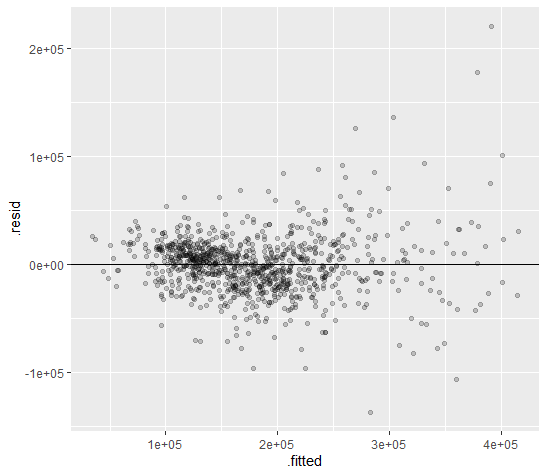
 

Figure 3.4.2: QQ-plot for Lasso Regression Figure 3.4.3: Residual Plot for Lasso Regression

Finally, we checked the four assumptions of the ridge regression model by using the two plots above. It is not surprising that two regularized models are having similar results. Figure 3.4.2 showed a non-normality of the model since all residuals at the bottom left and top right corners distributed away from the straight line. Figure 3.4.3 showed that variances of the residuals are not the same for the fitted values over 250,000, but the linearity assumption is still met in the residual plot because most points are close to and evenly distributed up around the zero line.

3.5 Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **RMSE (train)** | **RMSE (test)** | **R-square (train)** | **R-square (test)** | **Variance of R-square** |
| Linear Regression 1 | 35673.73 | 33389.20 | 0.7935 | 0.8315 | 0.038 |
| Linear Regression 2 | 27391.99 | 32944.57 | 0.8527 | 0.8359 | -0.0168 |
| Ridge Regression | 30183.89 | 37545.60 | 0.8212 | 0.7869 | -0.0343 |
| Lasso Regression | 29564.94 | 35677.49 | 0.8285 | 0.8076 | -0.0209 |

Table 3.5.1: RMSE and R-square Values for Regression Models

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE (Validation)** | **R-square (Validation)** |
| Linear Regression 1 (CV) | 37654.42 | 0.7787 |
| Linear Regression 2 (CV) | 28183.38 | 0.8444 |

Table 3.5.2: RMSE and R-square Values for Cross Validation

Overall, from the table 3.5, we can see the second linear regression model removing outliers is the strongest and the most reliable prediction model, as it has much lower testing RMSE value and highest testing R-square value compared to the other three models. Comparatively, model 2 performed better than model 1 in terms of higher R-squared and lower RMSE for cross validation results, which is reasonable because model 2 predicts well for houses sold at normal prices. And based on our VIF analysis on predictors, we implemented basic linear regression by dropping variables with high VIF values. Since the R-square dropped slightly and the partial models are not statistically more significant based on Anova analysis, we realized simply dropping variables is not the right way to improve the performance. To reduce variances for model 2, we implemented Ridge and Lasso regression to get consistent results. As the above table shows, the training results of the Ridge and Lasso models were worse than model 2. And these regularized models did not reduce the variances because the R-square gaps between training and testing were larger. Therefore, from the cross validation aspect, the model 2 performed the best for predicting sale price for houses without outliers.

1. **Conclusion and Limitation**

4.1 Limitation

We only implemented linear regression models in this project, the most robust model we obtained is limited to the fields of linear regression models. If we apply more complicated models such as random forest, the expected reliable model may differ. In addition, these models are primarily limited by external validations. We apply our model to the same dataset for testing, therefore our robust model may not perform well for housing data from a larger region or at a national scope.

4.2 Conclusion

After we attempted several approaches for modeling, the linear regression model without outliers is our robust model. By interpreting the coefficients of this model, we concluded the kitchen quality, height of the basement, and quality of material on the exterior are prominent variables that affect the sale price of housing. In order to make sure the price of the housing is reasonable, we recommend house-seekers to pay more attention to these variables.