Step-3

Stella Ramirez, Liuqian Bao, Andrew Cheng

2023-11-21

Introduction

Our data source is from: Anna Montoya, DataCanary. (2016).House Prices - Advanced Regression Techniques. Kaggle.Obtained from https://kaggle.com/competitions/house-prices-advanced-regression-techniques (https://kaggle.com/competitions/house-prices^advanced-regression-techniques)

The population that we are inferring our results on are all residential houses in Ames, Iowa.

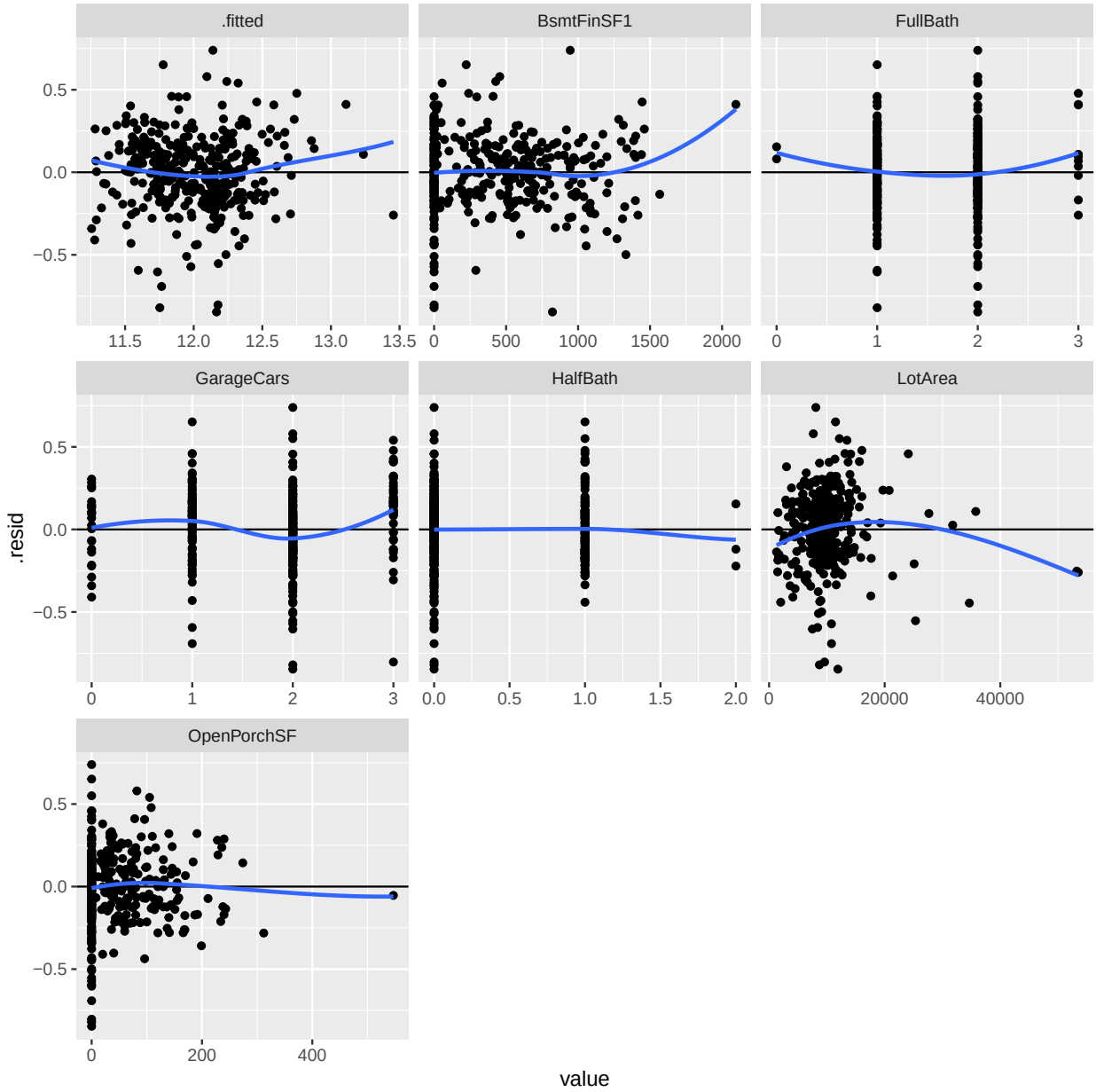
In this step, we did computational model building and statistical model building. First, we selected two models that are interesting to us manually and used ANOVA test and R^2 values to cross validate and determine the better model. Then, we used the step() function to automatically select an optimal with AIC and F-tests. After selecting a single best model, we fitted the test data and made prediction intervals using the optimal model.

By looking at the pairs plots, we found that: 1. The explanatory variables FullBath and GarageCars are moderately correlated(Corr: 0.457***). 2. The explanatory variables ExterQual(exterior quality) and BsmtFinSF1(basement finished area) are somewhat correlated (judged by their box plots).

1: For our first computational model, we chose to fit a model with all of our quantitative variables.

```
##
## Call:
## lm(formula = log(SalePrice) ~ LotArea + GarageCars + BsmtFinSF1 +
##     FullBath + OpenPorchSF + HalfBath, data = h2_partition$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84617 -0.13342  0.00401  0.14426  0.73890
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.101e+01  4.085e-02 269.481  < 2e-16 ***
## LotArea     1.374e-05  2.373e-06   5.787 1.62e-08 ***
## GarageCars  2.136e-01  1.895e-02  11.272  < 2e-16 ***
## BsmtFinSF1  2.092e-04  2.966e-05   7.053 9.75e-12 ***
## FullBath    1.977e-01  2.490e-02   7.941 2.92e-14 ***
## OpenPorchSF 7.831e-04  1.945e-04   4.025 7.01e-05 ***
## HalfBath    1.532e-01  2.618e-02   5.851 1.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2284 on 342 degrees of freedom
## Multiple R-squared:  0.6978, Adjusted R-squared:  0.6925
## F-statistic: 131.6 on 6 and 342 DF,  p-value: < 2.2e-16
```

In the following Graphs, we plotted the residuals for each variable to assess the model fit. We performed this twice, concluding that it is necessary to take the log of the response in order to correct the constant variance assumption. We have omitted the graphs of the un-transformed model, opting to display those of the model more useful to us.



Normality Check for Computational Model 1:
Then, we performed a normality check, to make sure the model was adequate.

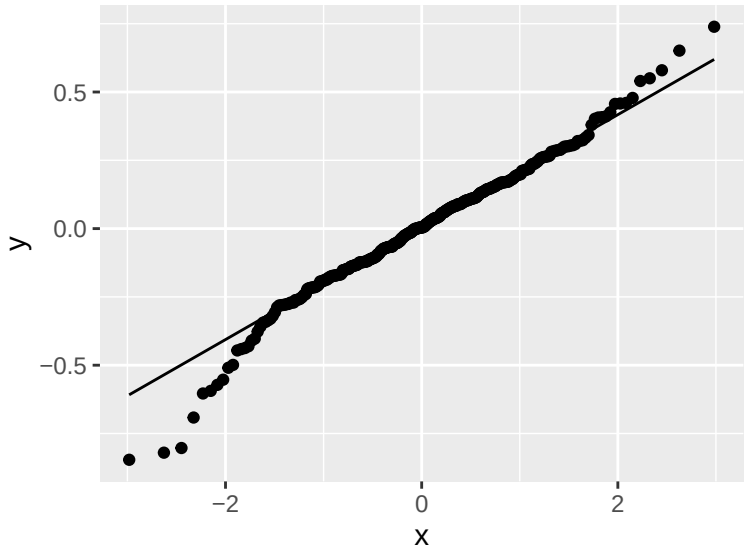


Figure 1: Normality Check for Computational Model 1

2: For our second computational model, we used the variable GarageCars because it had high correlation, as seen in our beginning plots, and KitchenQual becuase we wanted to test a highly correlated categorical variable, in addition to quantitative.

```
##
## Call:
## lm(formula = log(SalePrice) ~ GarageCars + KitchenQual, data = h2_partition$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80995 -0.15561  0.00358  0.13887  0.92052
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.04967    0.08199 146.957  < 2e-16 ***
## GarageCars     0.24193    0.02220  10.899  < 2e-16 ***
## KitchenQualFa -0.82102    0.11004  -7.461 7.07e-13 ***
## KitchenQualGd -0.35048    0.06239  -5.617 4.00e-08 ***
## KitchenQualTA -0.59185    0.06591  -8.980  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2645 on 344 degrees of freedom
## Multiple R-squared:  0.5923, Adjusted R-squared:  0.5876
## F-statistic:   125 on 4 and 344 DF,  p-value: < 2.2e-16
```

Again, we transformed SalePrice with a log function to make the variance more constant. Then we performed the following normality check on this model as well.

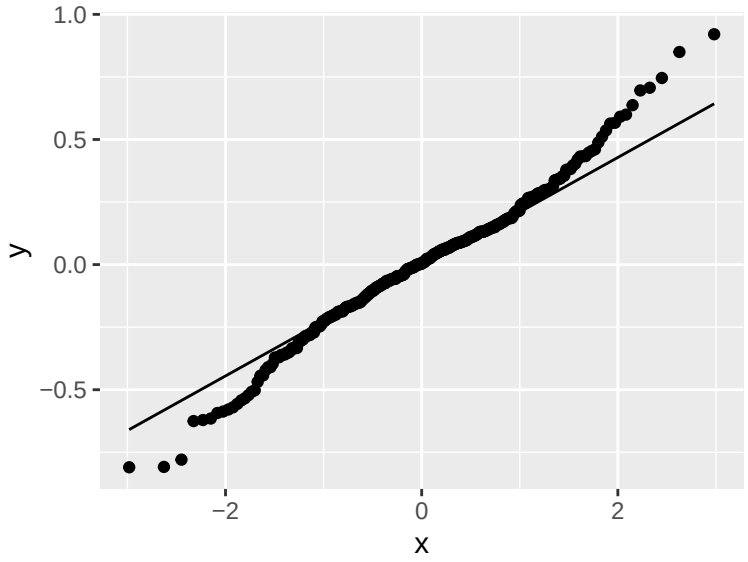


Figure 2: Normality Check for Computational Model 2

Normality Check for Computational Model 2

Cross Analyzing the Two Computational Models

Anova Function to Cross Validate

```
## Analysis of Variance Table
##
## Model 1: log(SalePrice) ~ LotArea + GarageCars + BsmtFinSF1 + FullBath +
##     OpenPorchSF + HalfBath
## Model 2: log(SalePrice) ~ GarageCars + KitchenQual
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    342 17.835
## 2    344 24.059 -2   -6.2242 59.677 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test results in very small p-value, which allows us to conclude that model 1, the model with all of our numerical variables, are significantly better than the other model.

Adjusted R^2 to Cross Validate

The adjusted R^2 of model 1 is 0.6935, which means that the model explains 69.35% of the variation in the response is explained by model 1. On the other hand, the adjusted R^2 of model 2 is 0.5876, which tells us that model 1 explains more variation in the response. Judged by this, model 1 is the better model.

Statistical Model

Next, we used backward selection to assess which variables would be best to use in our model. To do so, we begin with a full model and subtract the predictors one at a time.

We chose to omit the code output of each of these processes, as it is extremeley lengthy. The result of backward selection was a model including LotArea, GarageCars, BsmtFinSF1, FullBath, ExterQual, OpenPorch, HalfBath, and KitchenQual.

After doing so, we decided to also try forward selection to assess which variables would be best to use in our model. To do so, we begin with an empty model and add the predictors one at a time.

The result of forward selection was a model including ExterQual, LotArea, GarageCars, BsmtFinSF1, HalfBath, FullBath, KitchenQual, and OpenPorchSF.

In order to perform these two types of selection, keeping the response as the log of SalePrice, we used the AIC as the criteria. The model returned will have the predictors with the lowest AIC's. Both of these methods of selection returned the same model with 8 predictors, so we will use that as our statistical model.

The following displays the summary of our fitted statistical model. As you can see, the variables chosen were ExterQual, LotArea, GarageCars, BsmtFinSF1, HalfBath, FullBath, KitchenQual, and OpenPorchSF.

```
##
## Call:
## lm(formula = log(SalePrice) ~ ExterQual + LotArea + GarageCars +
##     BsmtFinSF1 + HalfBath + FullBath + KitchenQual + OpenPorchSF,
##     data = h2_partition$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70395 -0.09629  0.01682  0.11618  0.70776
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.176e+01  8.946e-02 131.465  < 2e-16 ***
## ExterQualFa   -6.986e-01  1.457e-01  -4.796 2.44e-06 ***
## ExterQualGd   -1.551e-01  7.448e-02  -2.083 0.038046 *
## ExterQualTA   -3.406e-01  7.961e-02  -4.279 2.45e-05 ***
## LotArea        1.511e-05  2.082e-06   7.258 2.75e-12 ***
## GarageCars     1.251e-01  1.832e-02   6.829 4.01e-11 ***
## BsmtFinSF1     1.810e-04  2.566e-05   7.054 9.95e-12 ***
## HalfBath       1.264e-01  2.269e-02   5.569 5.25e-08 ***
## FullBath       1.352e-01  2.244e-02   6.025 4.45e-09 ***
## KitchenQualFa -3.373e-01  9.213e-02  -3.661 0.000291 ***
## KitchenQualGd -1.565e-01  5.447e-02  -2.873 0.004318 **
## KitchenQualTA -2.445e-01  5.965e-02  -4.099 5.20e-05 ***
## OpenPorchSF    4.086e-04  1.708e-04   2.393 0.017280 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1948 on 336 degrees of freedom
## Multiple R-squared:  0.7839, Adjusted R-squared:  0.7762
## F-statistic: 101.6 on 12 and 336 DF,  p-value: < 2.2e-16
```

Looking at our R^2 value for this model shows that 78% of the variation in the response is explained by this model.

The following shows the residual plots we investigated to ensure that taking the log of the response helps with the model assumptions.

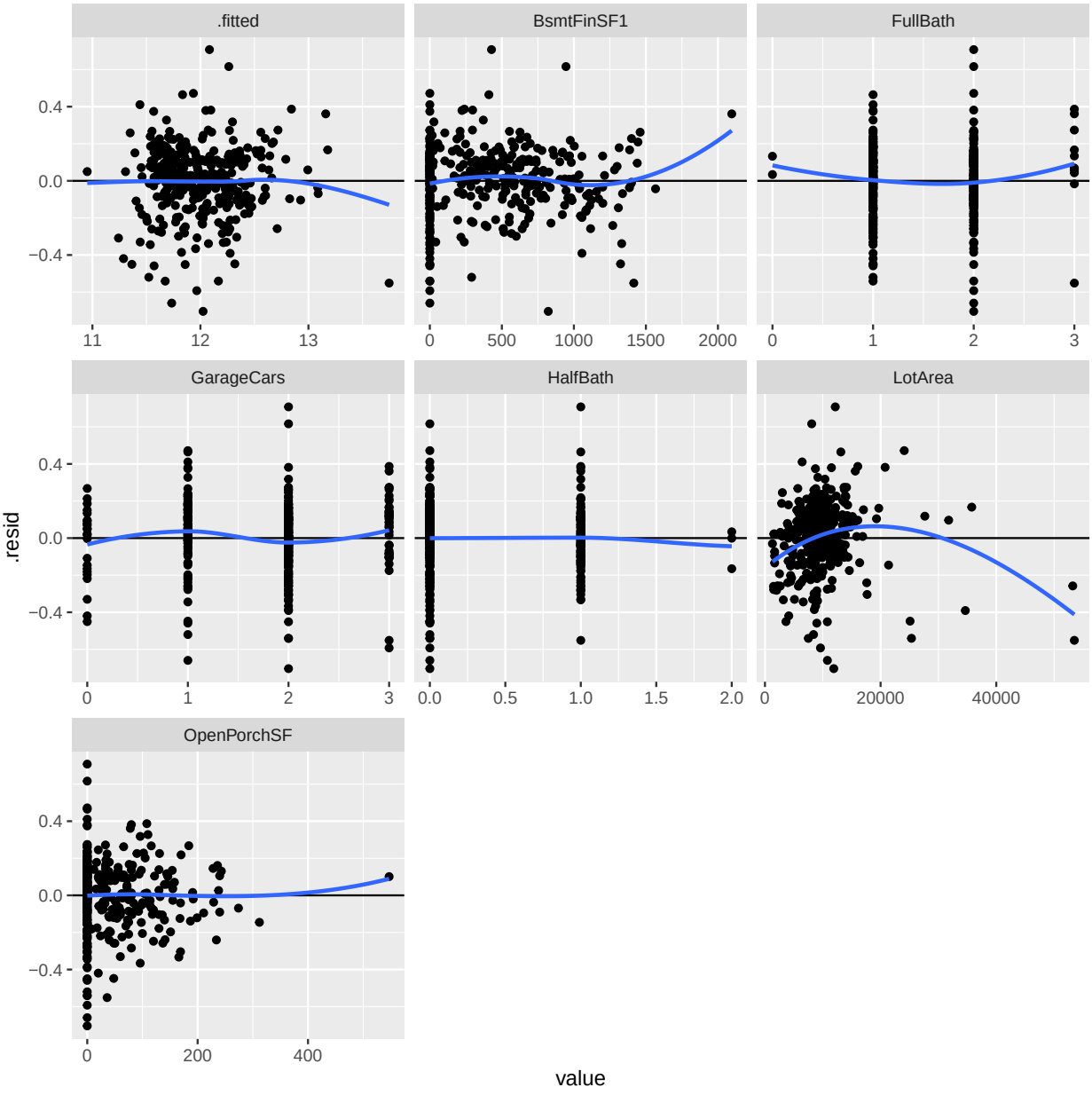


Figure 3: Residual Graphs for the Statistical Model

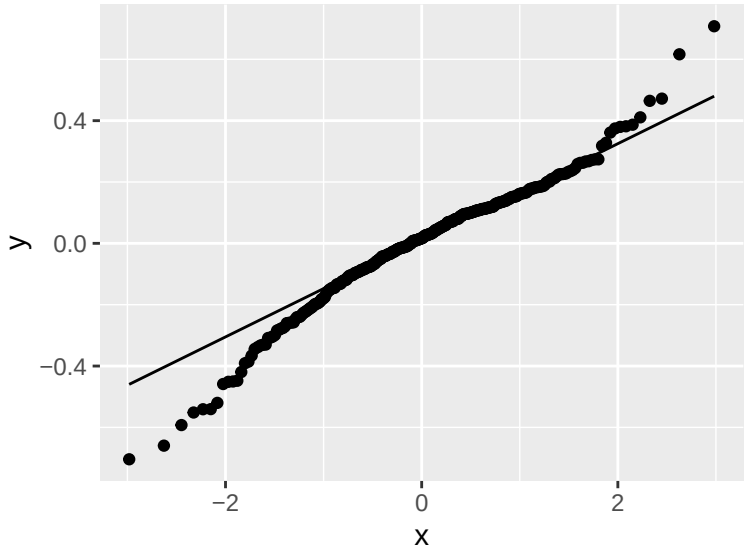


Figure 4: Normality Check for Statistical Model

Reporting on Test Data

In the following section, we fit our test data with the linear model based on the predictors chosen from our selection process.

```
##
## Call:
## lm(formula = log(SalePrice) ~ ExterQual + LotArea + GarageCars +
##     BsmtFinSF1 + HalfBath + FullBath + KitchenQual + OpenPorchSF,
##     data = h2_partition$test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62300 -0.09981  0.00000  0.09839  0.59175
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.178e+01  1.009e-01 116.735  < 2e-16 ***
## ExterQualFa   -8.778e-01  2.016e-01  -4.353 2.59e-05 ***
## ExterQualGd   -1.905e-01  8.960e-02  -2.126 0.035242 *
## ExterQualTA   -3.509e-01  9.376e-02  -3.743 0.000266 ***
## LotArea        4.491e-06  8.398e-07   5.348 3.60e-07 ***
## GarageCars     1.080e-01  2.677e-02   4.034 9.04e-05 ***
## BsmtFinSF1     1.950e-04  3.560e-05   5.476 1.99e-07 ***
## HalfBath       1.002e-01  3.104e-02   3.229 0.001550 **
## FullBath       1.725e-01  3.130e-02   5.512 1.69e-07 ***
## KitchenQualFa -2.845e-01  1.116e-01  -2.548 0.011924 *
## KitchenQualGd -7.999e-02  7.493e-02  -1.068 0.287600
## KitchenQualTA -1.889e-01  7.870e-02  -2.400 0.017734 *
## OpenPorchSF    4.452e-04  2.449e-04   1.818 0.071169 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1723 on 138 degrees of freedom
## Multiple R-squared:  0.8135, Adjusted R-squared:  0.7973
## F-statistic: 50.16 on 12 and 138 DF,  p-value: < 2.2e-16
```

Just in case, we performed a normality check to make sure the assumption holds within our test data.

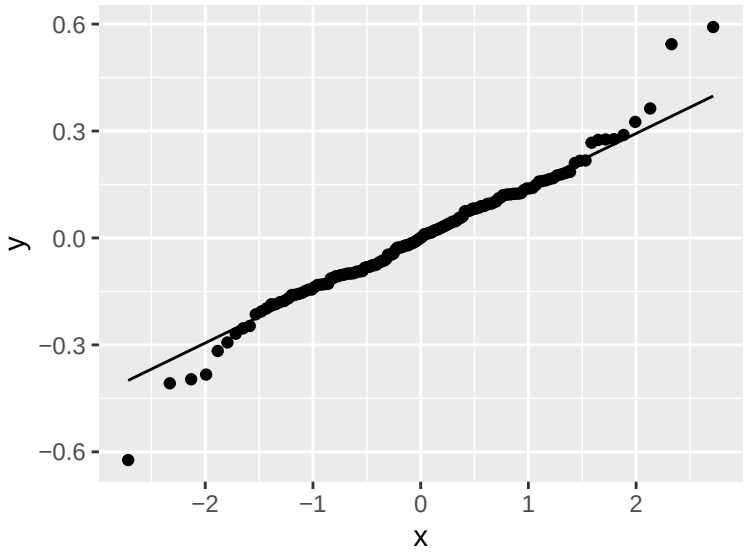


Figure 5: Normality Check for the Test Data Model

Looking for Influence points

Plot of the Rows and Their Residuals

The following shows the row values plotted with their respective residuals.

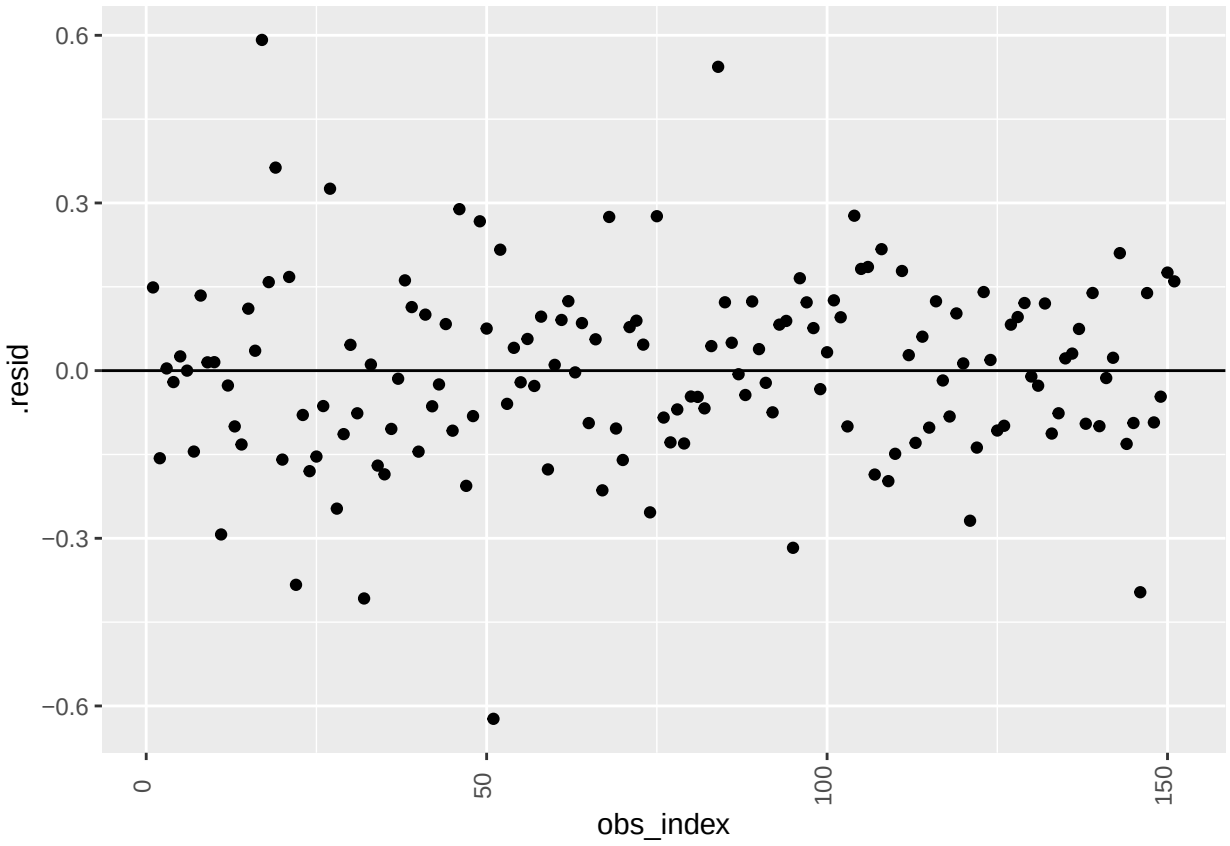
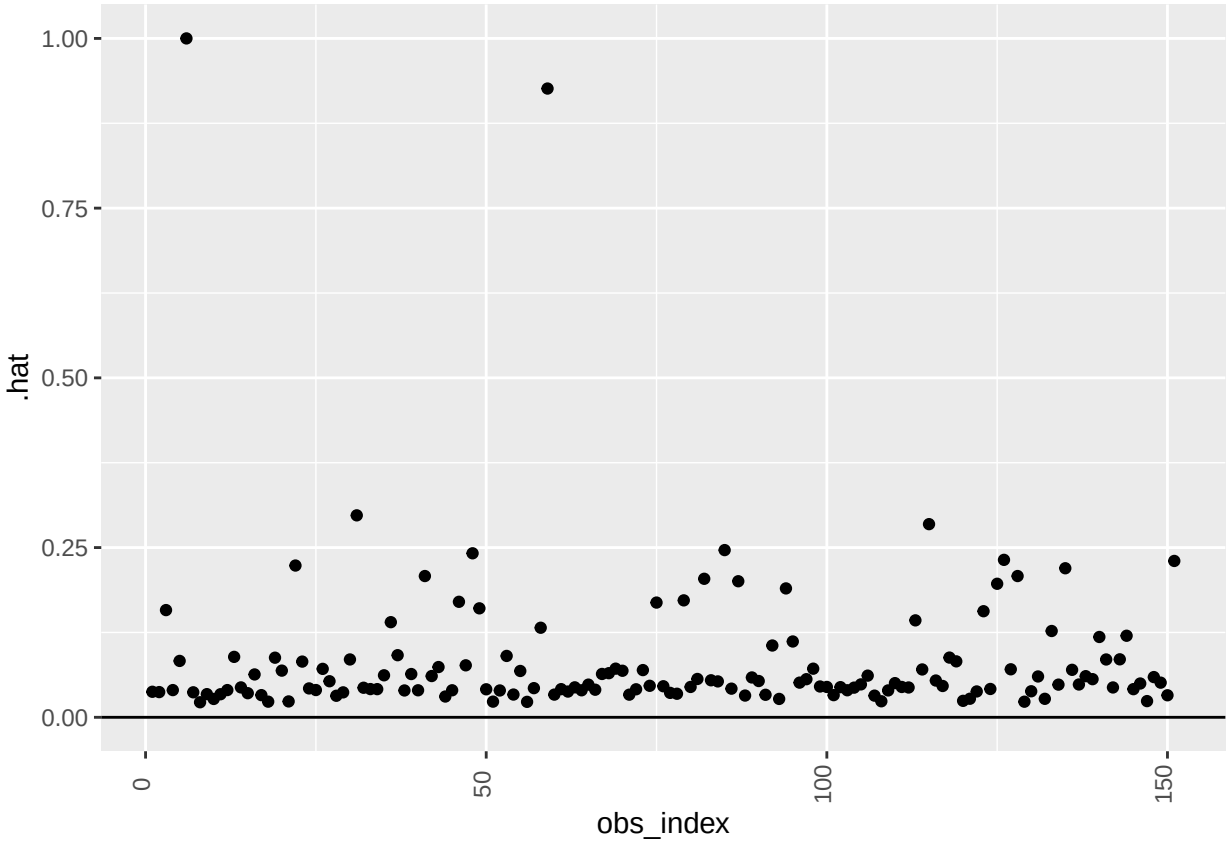


Figure 6: Rows with Residuals

Plot of the Rows and Their .hat Values

This graph shows the row values with their .hat values. We use these graphs to investigate influence points.



Next, we wanted to pull out the highest values for each .cooksd, .resid, and .hat, as we can potentially use these to indicate influential points.

```
## # A tibble: 3 x 2
## # Groups:   name [3]
##   name     value
##   <chr>    <dbl>
## 1 .cooksd 13.8
## 2 .hat     1.00
## 3 .resid  -0.623
```

After investigation, we found that the data with the highest .hat value is located in row 6 of our test data, and the data with the highest .resid is located in the 51st row of our test data.

Comparing Models With and Without the Unusal Observations

Here is a summary of the fit including the influence points:

```
##
## Call:
## lm(formula = log(SalePrice) ~ ExterQual + LotArea + GarageCars +
##     BsmtFinSF1 + HalfBath + FullBath + KitchenQual + OpenPorchSF,
##     data = h2_partition$test)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62300 -0.09981  0.00000  0.09839  0.59175
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.178e+01  1.009e-01 116.735  < 2e-16 ***
## ExterQualFa   -8.778e-01  2.016e-01  -4.353 2.59e-05 ***
## ExterQualGd   -1.905e-01  8.960e-02  -2.126 0.035242 *
## ExterQualTA   -3.509e-01  9.376e-02  -3.743 0.000266 ***
## LotArea        4.491e-06  8.398e-07   5.348 3.60e-07 ***
## GarageCars     1.080e-01  2.677e-02   4.034 9.04e-05 ***
## BsmtFinSF1     1.950e-04  3.560e-05   5.476 1.99e-07 ***
## HalfBath       1.002e-01  3.104e-02   3.229 0.001550 **
## FullBath       1.725e-01  3.130e-02   5.512 1.69e-07 ***
## KitchenQualFa -2.845e-01  1.116e-01  -2.548 0.011924 *
## KitchenQualGd -7.999e-02  7.493e-02  -1.068 0.287600
## KitchenQualTA -1.889e-01  7.870e-02  -2.400 0.017734 *
## OpenPorchSF    4.452e-04  2.449e-04   1.818 0.071169 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1723 on 138 degrees of freedom
## Multiple R-squared:  0.8135, Adjusted R-squared:  0.7973
## F-statistic: 50.16 on 12 and 138 DF,  p-value: < 2.2e-16
```

And here is the summary of the fit after they have been removed:

```
##
## Call:
## lm(formula = log(SalePrice) ~ ExterQual + LotArea + GarageCars +
##     BsmtFinSF1 + HalfBath + FullBath + KitchenQual + OpenPorchSF,
##     data = h2_testDF_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41392 -0.10223 -0.00065  0.09857  0.58088
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.181e+01  9.646e-02 122.394  < 2e-16 ***
## ExterQualGd   -1.901e-01  8.546e-02  -2.225 0.027731 *
## ExterQualTA   -3.540e-01  8.943e-02  -3.958 0.000121 ***
## LotArea        4.462e-06  8.010e-07   5.570 1.30e-07 ***
## GarageCars     1.051e-01  2.555e-02   4.114 6.67e-05 ***
## BsmtFinSF1     1.915e-04  3.396e-05   5.638 9.45e-08 ***
## HalfBath       9.648e-02  2.962e-02   3.257 0.001418 **
## FullBath       1.680e-01  2.988e-02   5.621 1.02e-07 ***
## KitchenQualFa -2.914e-01  1.065e-01  -2.736 0.007045 **
## KitchenQualGd -8.201e-02  7.147e-02  -1.148 0.253156
## KitchenQualTA -1.861e-01  7.506e-02  -2.479 0.014390 *
## OpenPorchSF    3.989e-04  2.338e-04   1.706 0.090285 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1643 on 137 degrees of freedom
## Multiple R-squared:  0.8182, Adjusted R-squared:  0.8036
## F-statistic: 56.03 on 11 and 137 DF,  p-value: < 2.2e-16
```

After removing the high leverage point at row 6 and the outlier point at row 51, we fit the selected model with 8 variables again. By comparing the summary of the two models, we found that the adjusted R^2 and the F statistics increased slightly, which makes sense because the model has a better fit without the unusual observations. However, the estimated coefficients of the variables and their significance did not change a lot, which means that removing unusual observations does not drastically change the model.

Confidence Interval for Mean Response Next, we computed a confidence interval for the mean response.

```
##        fit     lwr      upr
## 1 11.90099 11.8556 11.94638
```

For a house with average LotArea, GarageCars, BsmtFinSF1(basement finished are), HalfBath, FullBath, OpenPorchSF(open porch area), typical/average exterior and kitchen quality, the log of the sale price is estimated to be between 11.90099 and 11.94638, with 95% confidence.

Prediction Interval At a Value Finally, we computed prediction intervals for a few points. The first is at a randomly chosen row, row 100. The next two are rows 6 and 51, the points of influence.

```
##        fit      lwr      upr
## 1 11.83771 11.48945 12.18596
```

```
##       fit      lwr      upr
## 1 11.3022 10.82035 11.78406
```

```
##        fit      lwr      upr
## 1 11.59978 11.25516 11.94441
```

For a house that is the leverage point in the observation, the log of the sale price is estimated to be between 11.3022 and 11.78406, with 95% confidence.

For a house that is the outlier in the observation, the log of the sale price is estimated to be between 11.59978 and 11.94441, with 95% confidence.

As you can see, the points of influence affect the prediction intervals slightly.

Conclusion For step 3 of our project, we first put aside test data and re-introduced the reader to our data source.

Before we started our analysis of the data through models, we created pairs plots, and from these, we concluded that the data have a moderate correlation.

For each of our models, we applied diagnostic techniques and transformed the response with the log function, in order to obtain a more constant variance. This allows us to view any correlation or relationships between our response and our explanatory variables better, with a more accurate model.

We chose two computational models, one including all of our numerical variables and one that we based off of correlation, including a categorical variable.

We have then created an ANOVA table to cross-validate these models.

We are not using interaction variables, as there is not sufficient evidence showing that variables interact.

Next, we did statistical analysis to determine a good model. Keeping our response the log of SalePrice, we have performed forward and backward selection using AIC as our criteria. Both forward and backward selection returned us the same model that has the same 8 predictors, so we decided to use those to define our model. The variables chosen are ExterQual, LotArea, GarageCars, BsmtFinSF1, HalfBath, FullBath, KitchenQual, and OpenPorchSF.

After choosing our model, we have fit the model and looked at the corresponding R^2 value, which showed that 78% variation in the response was explained by the model we chose.

We then fit our model on our test data and looked for influence points. To do so, we looked at the residual and hat matrix values. Those with high values in each respective category are influence points.

Due to the potential influence of these points on our overall model, we have investigated the fit when removing unwanted data points and without. We concluded that these points do have some impact on the fit of our model, however, they do not drastically change our results.

Finally, we computed a confidence interval for a mean predicted value and several prediction intervals for future predicted values at specific points.

For our confidence interval, it is important to note that for the categorical variables, we found the most commonly occurring values of each. We have used the variable ExterQual as "TA", and the variable KitchenQual as "TA".

In the prediction intervals, we tested each of our leverage points in addition to a general point that followed the trend, in order to see if there was much difference. We found that these points will change the interval, however there is not a huge discrepancy.

The values we obtained from the confidence and prediction intervals were within our expectations in accordance with our model. Overall, we believe that the model we chose is a good fit for our data. We can conclude from our analysis that the most significant predictors are external quality(ExterQual), LotArea, GarageCars, finished basement square footage (BsmtFinSF1), HalfBath, FullBath, kitchen quality (KitchenQual), and open porch square footage (OpenPorchSF).