PSTAT126 Group Project Step 3

Hanya Ansari, Carina Yuen, Daren Aguilera

2023-12-3

Introduction:

Wine Quality Based on Physicochemical Tests from UCI Machine Learning Repository https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/code?datasetId=4458&searchQuery=R

No Missing Attribute Values: 0

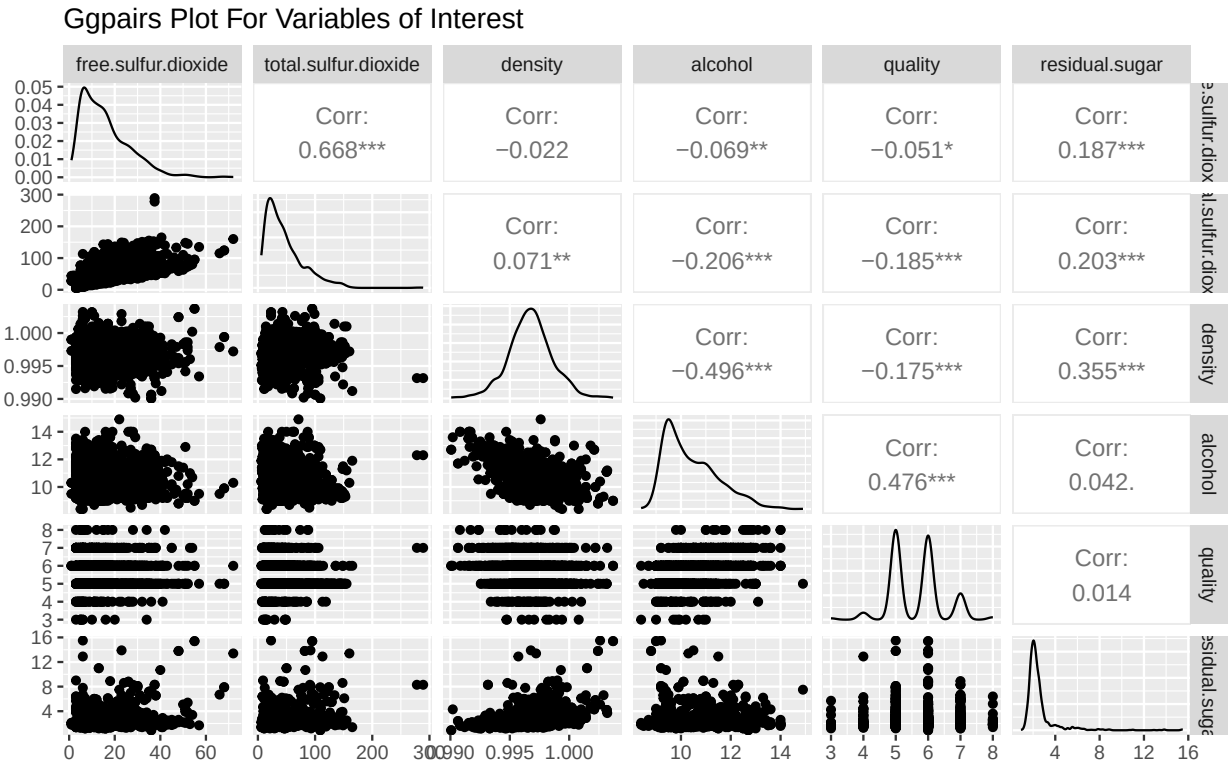
Number of Instances: red wine: 1599

Number of Variables: 12 total, 11 continuous, 1 discrete (fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol and 1 integer output variable: quality score between 0 and 10)

Plots on Explanatory and Response Variables

```
knitr::opts_chunk$set(echo = FALSE,
                      message = F,
                      warning  = F,
                      fig.width = 4,
                      fig.height = 3,
                      fig.align = 'center',
                      fig.pos = 'H')
```

ggpairs plot



Interpretation

The upper triangle of the code output represents the 15 correlation coefficients (6 choose 2 number of pairs of wine parameters). The diagonal displays the graph of the distribution.

We noticed that free sulfur dioxide and total sulfur dioxide have a relatively high correlation value of 0.668. This makes sense because the amount of Total Sulfur Dioxide encompasses the Free Sulfur Dioxide, so as we expect that as the Free Sulfur Dioxide increases, so does the total. Another interesting relationship we noticed was that alcohol content and density had a relatively large (in magnitude) correlation value of -0.496. This can be seen visually in the scatter plot of density and alcohol.

There is not any obvious simple transformations that would improve the fit drastically. We did notice that the density and residual sugar scatter plot had a pattern that roughly looked like an exponential, so we will include the natural log transformation and evaluate if it made a difference.

Interaction Variables

```
# modified model with interaction term between free sulfur dioxide and total sulfur dioxide
fit2_qual_mod<- lm(quality_data~x_f + poly(x_v, 2, raw = T) + x_c+ x_r + poly(x_ch, 2, raw=T) +
                     poly(x_fs, 2, raw=T)+ poly(x_ts, 2, raw=T) + x_d + poly(x_p,2,raw=T) +
                     poly(x_s,2,raw=T) + poly(x_a, 2, raw=T)+ x_fs:x_ts + x_d:x_a)
# poly second order model

# modified model with interaction term and third order poly
fit3_qual_int <- lm(quality_data~x_f+poly(x_v, 3, raw = T) + x_c + x_r +  poly(x_ch, 3, raw=T) +
```

```
                        poly(x_fs, 3, raw=T) + poly(x_ts, 3, raw=T) + x_d +poly(x_p, 3, raw=T)+
                        poly(x_s,3,raw=T)+poly(x_a, 3, raw=T)+x_fs:x_ts +x_d:x_a)
```

Our two models shown above both contain interaction terms between free sulfur dioxide and total sulfure dioxide, as well as density and alcohol. The interactions terms above are not necessary to the final fitting of our model, although we see an increase in adjusted R^2 for the model which we included these interactions. These aren't neccessary since the final model we chose does not incorporate these interactions, although are retained in the report for their interesting result and possible exploration further on.

Computational and Statistical Model Choices

For feature engineering, we used backwards elimination on the full model in our project step 2.

```
## Analysis of Variance Table
##
## Model 1: quality_data ~ x_f + poly(x_v, 3, raw = T) + x_c + x_r + poly(x_ch,
##     3, raw = T) + poly(x_fs, 3, raw = T) + poly(x_ts, 3, raw = T) +
##     x_d + poly(x_p, 3, raw = T) + poly(x_s, 3, raw = T) + poly(x_a,
##     3, raw = T)
## Model 2: quality_data ~ x_f + poly(x_v, 3, raw = T) + x_c + x_r + poly(x_ch,
##     3, raw = T) + poly(x_fs, 3, raw = T) + poly(x_ts, 3, raw = T) +
##     poly(x_p, 3, raw = T) + poly(x_s, 3, raw = T) + poly(x_a,
##     3, raw = T)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1573 624.49
## 2   1574 625.42 -1  -0.92632 2.3333 0.1268
```

Cross Validation

Comparing the models for partitioned data (on fixed acidity, alcohol, and volatile acidity) and the original dataset. The error values for the trained dataset are similar to that of the original dataset. For example, for fixed acidity, the trained MSE was 0.3926479 and the original data set's MSE is 0.3805819. This shows by fitting the model we were able to measure predictive accuracy within our dataset.

```
## [1] 0.3912048
```

```
mse(model=fit_train, data=data_partition$test)
```

```
## [1] 0.3790097
```

```
mse(model=fit_train2, data=data_partition$test)
```

```
## [1] 0.3401661
```

```
mse(model=fit_train3, data=data_partition$test)
```

```
## [1] 0.01498261
```

```r
mse(model=fit_normal, data=data_partition$test)
```

```
## [1] 0.3613544
```

```r
mse(model=fit_normal2, data=data_partition$test)
```

```
## [1] 0.3272631
```

```r
mse(model=fit_normal3, data=data_partition$test)
```

```
## [1] 0.01460971
```

```
## [1] "BIC"
```

```r
BIC(fit_normal)
```

```
## [1] 3127.203
```

```r
BIC(fit_normal2)
```

```
## [1] 2956.867
```

```r
BIC(fit_normal3)
```

```
## [1] -1877.124
```

```r
summary(fit_normal3)
```

```
##
## Call:
## lm(formula = volatile.acidity ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47100 -0.08658 -0.01539  0.06470  0.78039
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.047e+01  4.257e+00  -4.808 1.67e-06 ***
## fixed.acidity         1.125e-02  5.242e-03   2.147    0.032 *
## citric.acid          -6.276e-01  2.527e-02 -24.834  < 2e-16 ***
## residual.sugar       -2.109e-03  3.035e-03  -0.695    0.487
## chlorides             7.871e-01  8.301e-02   9.482  < 2e-16 ***
## free.sulfur.dioxide  -2.629e-03  4.347e-04  -6.047 1.83e-09 ***
## total.sulfur.dioxide  1.082e-03  1.458e-04   7.420 1.90e-13 ***
## density               2.105e+01  4.344e+00   4.845 1.39e-06 ***
## pH                    2.180e-02  3.880e-02   0.562    0.574
## sulphates            -1.476e-01  2.329e-02  -6.337 3.04e-10 ***
```

```
## alcohol               2.867e-02  5.490e-03   5.223 2.00e-07 ***
## quality              -4.432e-02  4.953e-03  -8.948  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1311 on 1587 degrees of freedom
## Multiple R-squared:  0.468,  Adjusted R-squared:  0.4643
## F-statistic: 126.9 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Model Selection

We partitioned the data into 70% training and 30% test sets and looked at stepwise regression with backwards elimination and found there was not a huge difference in the model quality, so we decided to use criterion based methods. Each of the criterion based methods balance model complexity and model fit. We choose the BIC method to prioritize selection consistency across our trained model. From this we deduced that the fit_normal3 has the lowest BIC value of -1877.124, so we choose this model to best fit the dataset.

Interpretation of Beta Coefficients

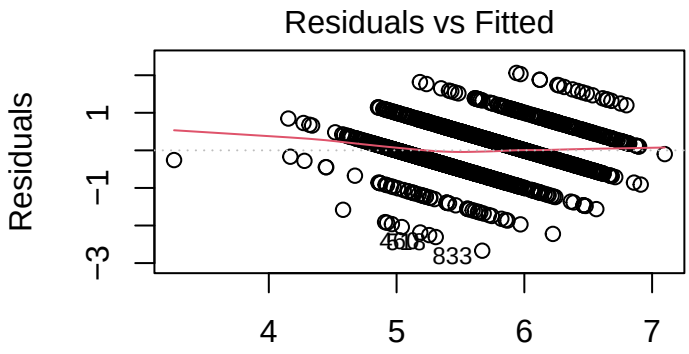
Looking at the coefficients of our chosen model fit_normal3, the intercept was about -20.470, fixed acidity: 0.011, citric acid: -0.628, residual sugar: -0.002, chlorides: 0.787, free.sulfur.dioxide: -0.003, total.sulfur.dioxide:0.001, density: 21.046, pH: 0.022, sulphates: -0.148, alcohol: 0.029, quality: -0.044.

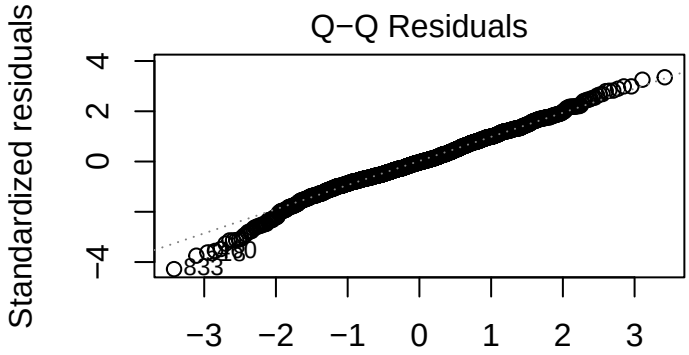
The criteria with coefficients with low p values (for alpha=0.05) were fixed.acidity, citric.acid, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulphates, alcohol and quality.

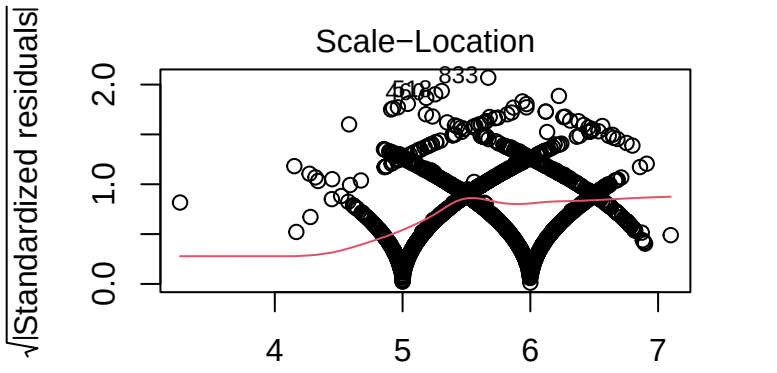
The multiple R-squared value is 0.468 and the adjusted R-squared value is 0.4643. Interpreting this, the multiple R squared value measures the amount of variation in the response variable that can be explained by the predictor variables, and the adjusted R squared represents the values that have been adjusted for the number of predictors in the model. A high R^2 is not neccessarily a guarantee that the model will accurately describe the population because as the number of predictors increase, the R^2 value naturally increases, but this may not be indicative of the model improving.

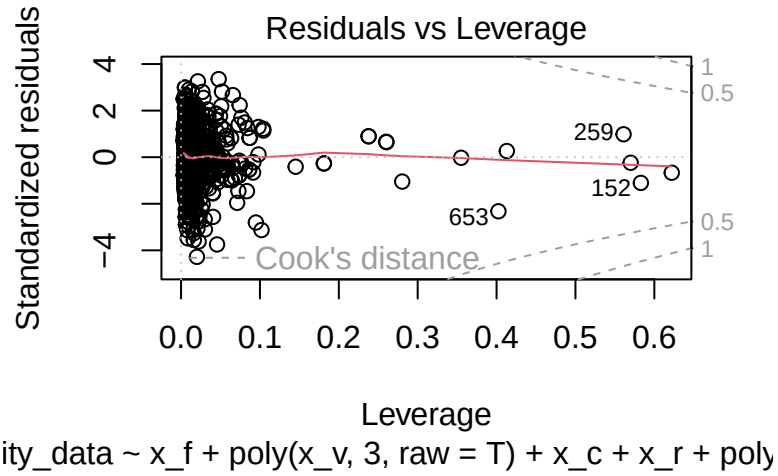
Analysis of the residuals and influence plots

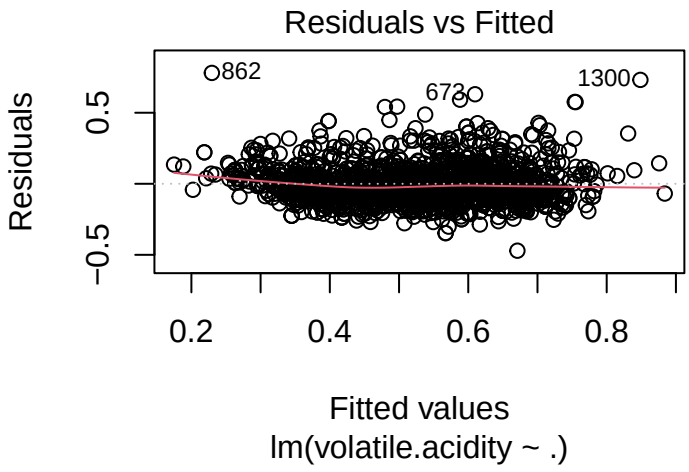
Comparing the full model and the BIC, we can see the Scale-Location plot improves drastically in regards to the violating the homoscedasticity assumption. There is no clear pattern in the residuals vs fitted plots. The normal QQ plot indicates that the data does follow a relatively normal distribution. However, upon looking at the Scale-Location, it violates the assumption of constant variance as there a distinct trend/shape on the plot.

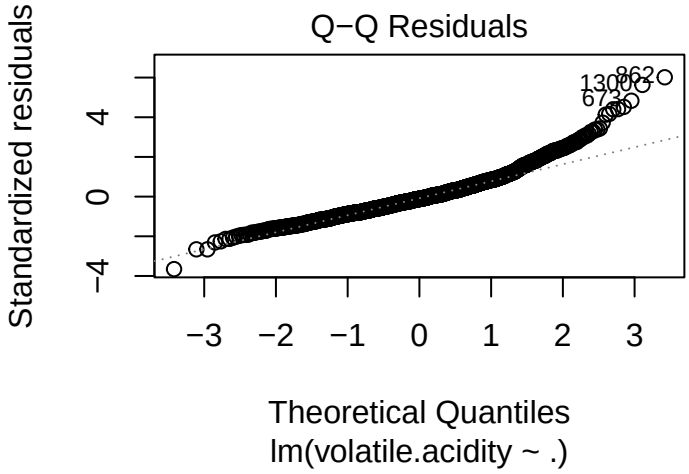

Residuals vs Fitted
Residuals
1
−1
−3
833
4
5
6
7
Fitted values
ity_data ~ x_f + poly(x_v, 3, raw = T) + x_c + x_r + poly

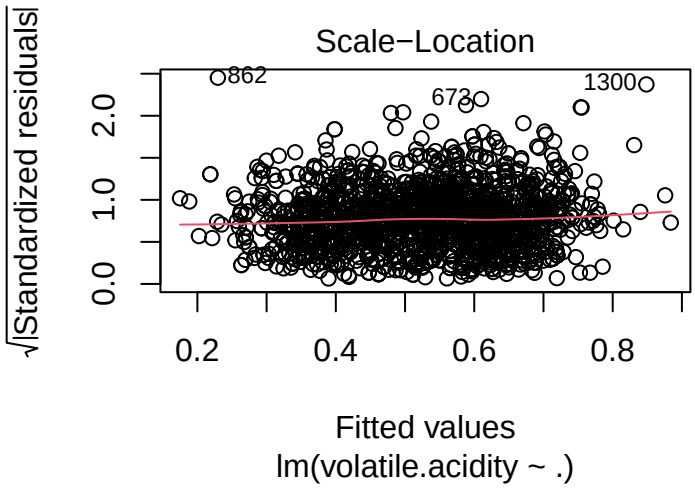

Q−Q Residuals
Standardized residuals
4
2
0
−4
833
−3
−2
−1
0
1
2
3
Theoretical Quantiles
ity_data ~ x_f + poly(x_v, 3, raw = T) + x_c + x_r + poly

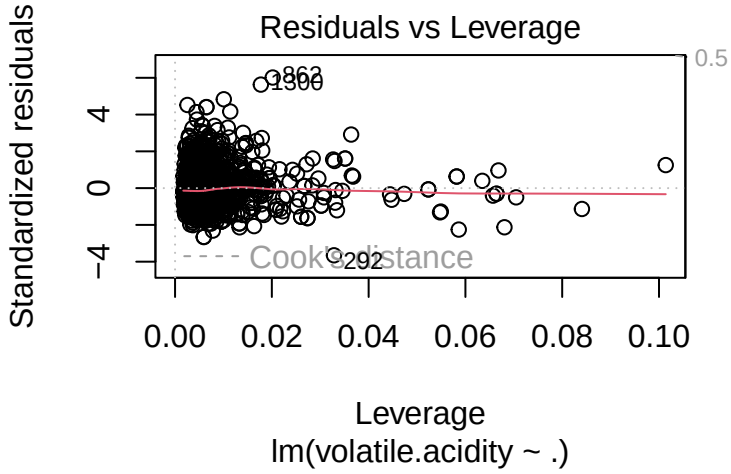

Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
4
5
6
7
Fitted values
ity_data ~ x_f + poly(x_v, 3, raw = T) + x_c + x_r + poly


Residuals vs Leverage
Standardized residuals
−4
0
2
4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
0.5
0.5
1
259
152
653
Cook's distance
Leverage
ity_data ~ x_f + poly(x_v, 3, raw = T) + x_c + x_r + poly


Residuals vs Fitted
862
673
1300
Residuals
0.5
−0.5
0.2
0.4
0.6
0.8
Fitted values
lm(volatile.acidity ~ .)


Q−Q Residuals
862
1300
673
Standardized residuals
4
0
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(volatile.acidity ~ .)





Confidence Intervals and Prediction Intervals

```
## [1] 0.4144566
```

```
## [1] 0.3935886
```

```
## [1] 0.02166344
```

```
##
## Call:
## lm(formula = volatile.acidity ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.47100 -0.08658 -0.01539  0.06470  0.78039
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.047e+01  4.257e+00  -4.808 1.67e-06 ***
## fixed.acidity         1.125e-02  5.242e-03   2.147    0.032 *
## citric.acid          -6.276e-01  2.527e-02 -24.834  < 2e-16 ***
## residual.sugar       -2.109e-03  3.035e-03  -0.695    0.487
## chlorides             7.871e-01  8.301e-02   9.482  < 2e-16 ***
## free.sulfur.dioxide  -2.629e-03  4.347e-04  -6.047 1.83e-09 ***
## total.sulfur.dioxide  1.082e-03  1.458e-04   7.420 1.90e-13 ***
## density               2.105e+01  4.344e+00   4.845 1.39e-06 ***
## pH                    2.180e-02  3.880e-02   0.562    0.574
## sulphates            -1.476e-01  2.329e-02  -6.337 3.04e-10 ***
## alcohol               2.867e-02  5.490e-03   5.223 2.00e-07 ***
## quality              -4.432e-02  4.953e-03  -8.948  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1311 on 1587 degrees of freedom
## Multiple R-squared:  0.468,  Adjusted R-squared:  0.4643
## F-statistic: 126.9 on 11 and 1587 DF,  p-value: < 2.2e-16

##         fit       lwr       upr
## 1 0.5246879 0.5173857 0.5319901

##         fit       lwr       upr
## 2 0.7105225 0.4659868 0.9550582
```

Confidence and prediction intervals

With 95% confidence, the mean volatile acidity in the data is estimated to be between 0.5211921 and 0.5363845. Additionally, with 95% confidence, the predicted mean volatile acidity in the data is estimated to be between 0.4439131 and 0.9532625.

Summary

In conclusion, our working dataset is on red wine and its variables as to how they contribute with each other and its impact on overall quality.

From the ggpairs output, free sulfur dioxide and total sulfur dioxide had a relatively high correlation value of 0.668, which makes intuitive sense given the total sulfur dioxide value encompasses the free sulfur dioxide value. Alcohol content and density also had a relatively large (in magnitude) correlation value of -0.486.

In this project step, we used stepwise regression with backward elimination. First, we created our starting model fit3_quality, which is a full polynomial model of the quality data with highest order 3. To perform backward elimination, we dropped the density term (one of the few terms that was not raised to poly(3)). Then we did an analysis comparing the models using anova(fit3_quality, fit_drop_3), and noticed that dropping the term did not change the RSS much, in fact, it increased it from 624.49 to 625.42. We decided to use the BIC method to choose our single best model.

For implementation of interaction variables, our chosen two models contain interaction terms between free sulfur dioxide and total sulfur dioxide, as well as density and alcohol. The interactions terms above are not necessary to the final fitting of our model, although we see an increase in adjusted R2.

The BIC method was applied to the original dataset that was partitioned into 70% training and 30% test sets; Cross validation on fit_normal3 revealed predictive accuracy within our dataset on variables such as fixed acidity, alcohol, and volatile acidity.

The coefficients of our chosen model fit_normal3 revealed that the variables fixed_acidity, citric.acid, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulfates, alcohol, and quality had low p values (for alpha = 0.05). In addition to this, our analysis also indicated that an increase in predictors and R^2 values was not necessarily indicative of model improvement.

The Scale-Location plot revealed an improvement in addressing homoscedasticity issues despite the absence of a clear pattern in the residuals vs fitted plots.

To do our confidence intervals and prediction intervals, we wanted to estimate volatile acidity values as it is a continuous variable. With 95% confidence, the volatile acidity in the data is estimated to be between ~0.52 and 0.54 Additionally, with 95% confidence, the predicted mean volatile acidity value is estimated to be between 0.44 and 0.95.