University of California, Santa Barbara

Final Project of

Regression Analysis

PSTAT 126: Regression Analysis

Regression Analysis of Red Wine

Celeste Herrera

Niveditha Lakshminarayanan

June 7, 2020

1. Introduction

The project will concentrate on what factors contribute to the overall quality of red wine, using the “quality” attribute from the red_wine dataset provided by the UC Irvine Machine Learning Repository. We are trying to determine whether our response variable, the “quality” of red wine, measured in this dataset by integers on a scale of 0-10, can be predicted by any of the following predictors provided: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Additionally, we want to find out which of the following has the greatest impact on our output of acknowledging how the quality of wine is.

2. Question of Interest

Can the quality of wine be predicted by the following attributes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol?

Between all the pairs of variables regarding red wine quality, which correlations are large and positive, which are large and negative, and which are small?

What is the predicted quality of red wine with average (given the data) status, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol levels? What is the appropriate 95% confidence interval for the mean amounts?

3. Regression Method

We are trying to pertain our model to comply with the four LINE conditions before we are actually able to answer our question of interest.

4. Regression Analysis, Results and Interpretation

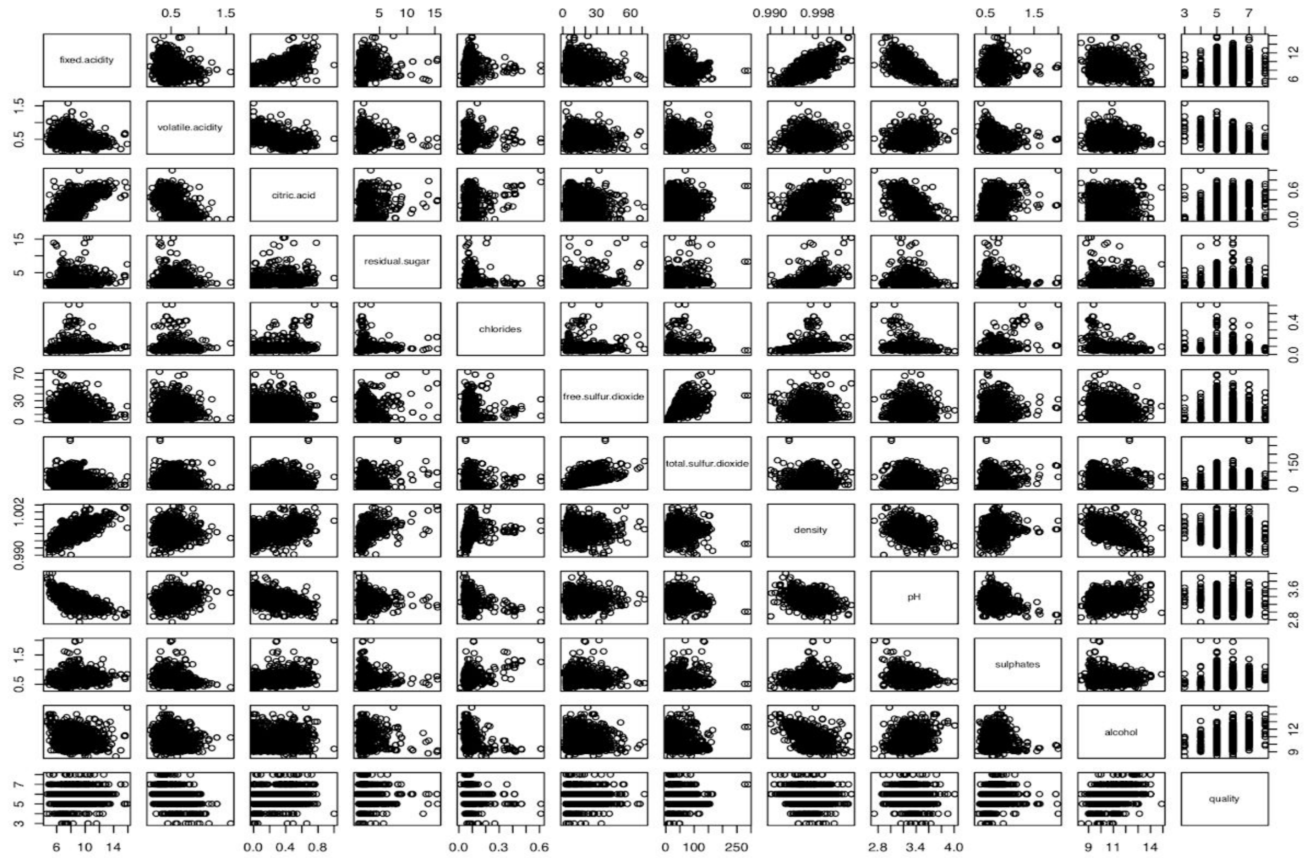
We create our model by first introducing our independent and dependent variables.

- Quality = quality (response variable)
- Fixed acidity = fixed_acid
- volatile acidity = volatile_acid
- citric acid = citric_acid
- residual sugar = residual_sugar
- Chlorides = chloride
- free sulfur dioxide = free_sulfur
- total sulfur dioxide = total_sulfur
- Density =density
- pH = ph
- Sulphates = sulphate
- Alcohol = alcohol

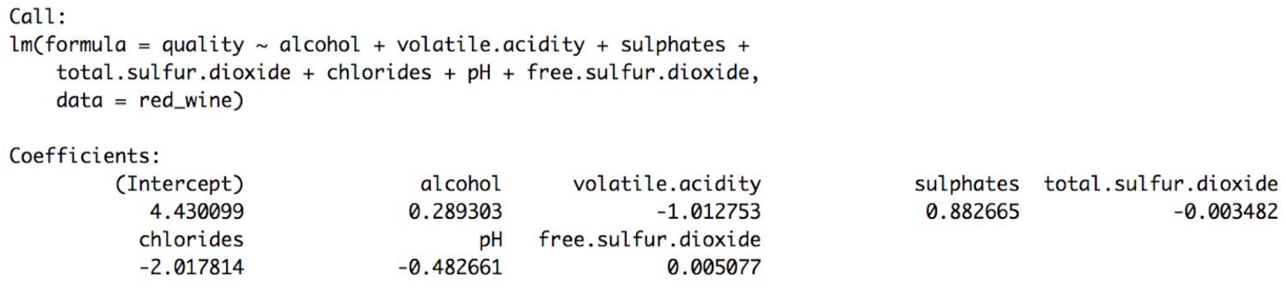
Quality is the variable we are making the attempt to predict with the predictor variables stated above.

As for future reference, we are considering every independent variable provided by the dataset, and are not omitting values either.

We begin our analysis by plotting each potential predictor (using the predictors stated previously) against the response variable (quality) using the pairs() function in R. The results yield the following set of plots:



After performing the pairs() function we move on to perform the stepwise regression using the step() function, to indicate which of the variables are the best for predicting the Quality values.

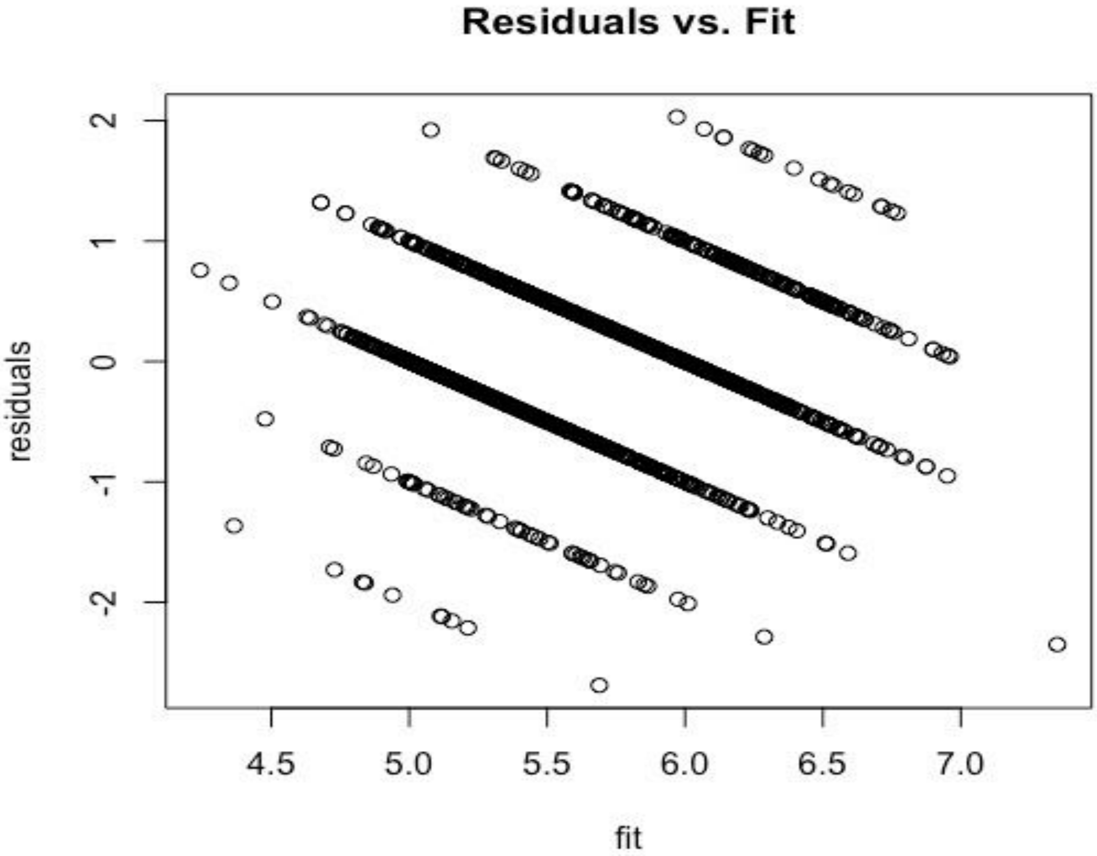


```
Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,
    data = red_wine)

Coefficients:
         (Intercept)               alcohol      volatile.acidity             sulphates  total.sulfur.dioxide
            4.430099              0.289303             -1.012753              0.882665             -0.003482
           chlorides                    pH   free.sulfur.dioxide
           -2.017814             -0.482661              0.005077
```

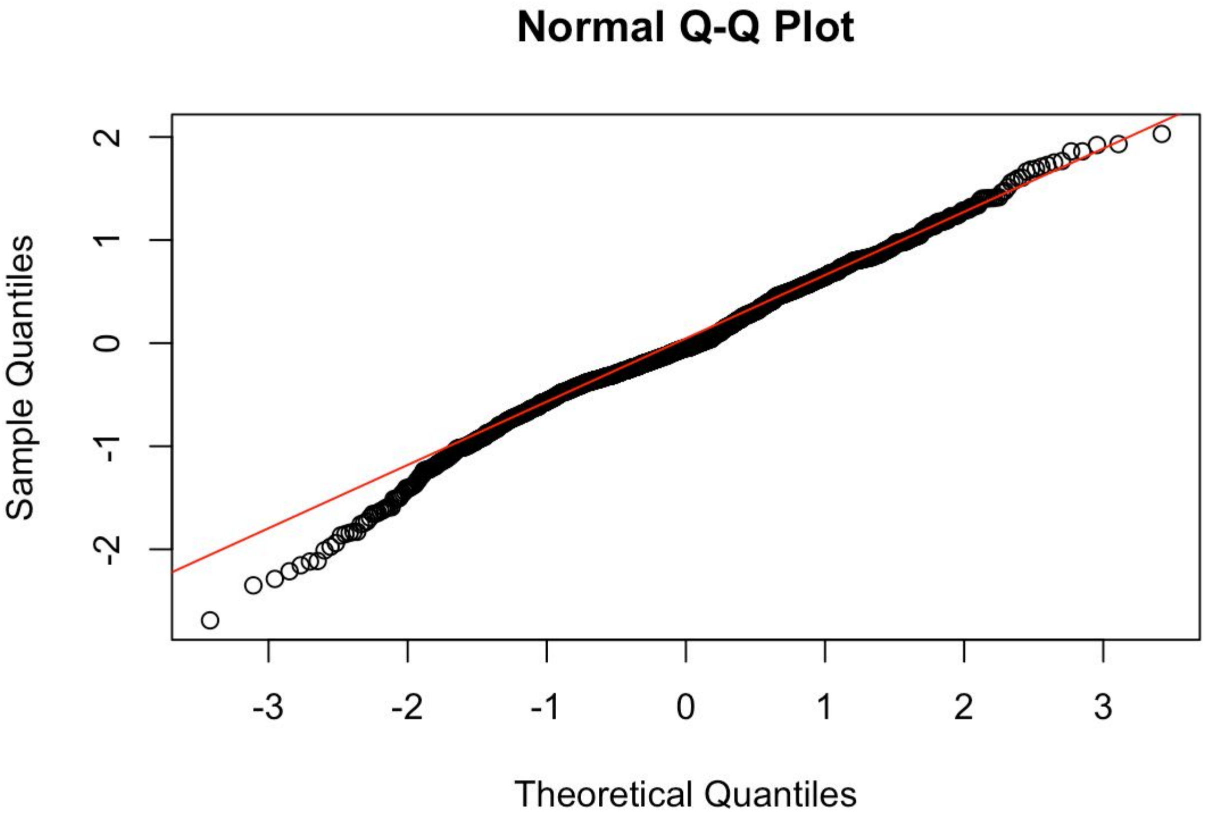
From the results above, it is clear that the step() function suggests that the simplest and best fitting regression model can be obtained by using alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and free sulfur dioxide as predictors for quality.

We now run the leaps() procedure to reinforce that the model found with the steps() function is the simplest and best fitting regression model to predict for red wine quality, since the leaps() procedure in R performs a thorough search through all of the predictor variables to find the best subsets to predict the value of the response variable in linear regression. We specifically used the leaps procedure setting the “method” component to “adjr2” to use the adjusted R squared values as a criteria since the adjusted R squared value does not necessarily increase as more predictors are added, helping us identify which predictors should be included and excluded. Running the leaps() procedure with specifically the selected predictor values from the step() function yields very little difference in adjusted R squared values than running the leaps() procedure on all of our predictor variables, demonstrating that having more variables in the model will create a negligible effect on our results, thereby confirming supporting our decision to use the simplified model with predictors alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and free sulfur dioxide acting on quality. **often several equally good models**

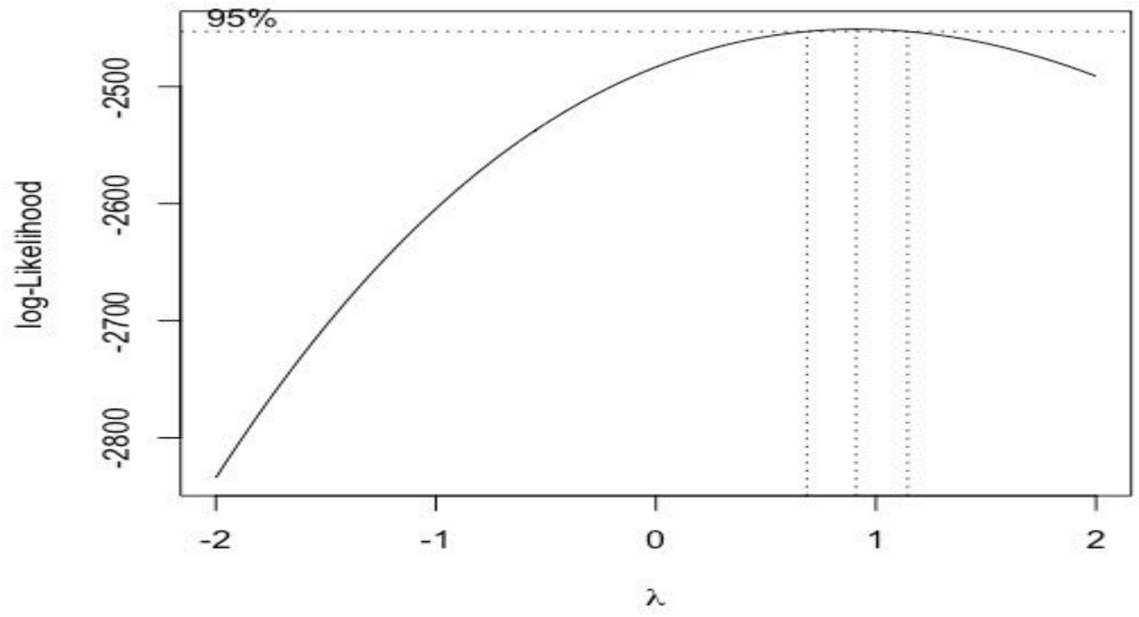
Our next step is to see whether our model abides by the four L.I.N.E conditions. We will check the assumptions by first observing the Residuals vs. Fit plot of the model to indicate whether the linearity and equal variance assumptions are met. From the plot we notice that the data appears to be evenly spread out with no clustering or fanning in any areas, suggesting that the Residuals vs. Fit plot exhibits equal variance. We also notice many linear patterns across the graph, which we posit are due to the fact that the response variable, Quality, is a categorical variable since the dataset measures quality as only integer values from 0-10. Due to the categorical nature of the response variable, the Residuals vs. Fit plot does not have all of the data linearly organized into one single line, but instead in multiple lines. By taking this property into account, we can conclude that the Residuals vs. Fit plot does meet the linearity criterion. So, Residuals vs. Fit plot appears to be well-behaved, suggesting that no transformation is necessary for the data.



We must now check that the normality assumption is met by observing the Normal Q-Q plot of the residual values of our model. We see that the plot below exhibits an approximately linear pattern, with the points slightly skewed towards the beginning of the Q-Q plot. Despite this problem with the Q-Q plot, the majority of the data points exhibit a linear trend around the normal line, so we can conclude that our normality condition is met for this model, and supports our hypothesis that no transformation is needed for the data.



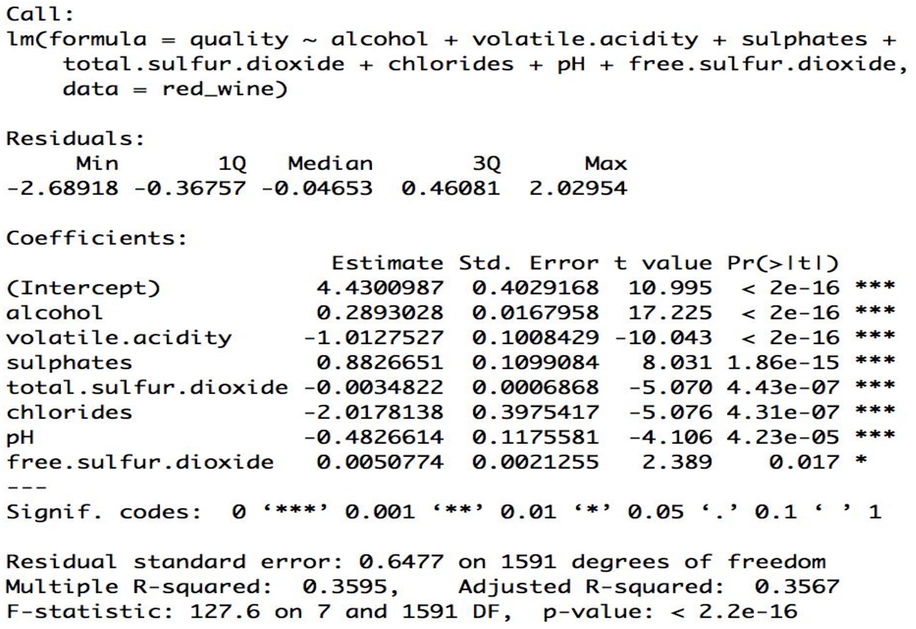
We reinforce our belief that there is no need for a transformation with a BoxCox plot. We can see that the dashed lines demarcate that the 95% confidence interval contains the value λ = 1, and it is known that if 1 is in the 95% CI of the BoxCox plot, we can choose to do no transformation.



So, from observing the Residuals vs. Fit plot, the Q-Q plot, and the BoxCox plot, it is clear that the data already follows the L.I.N.E conditions without need for transformation.

Now we can begin to answer our updated question of interest:

Can the quality of wine be predicted by alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and free sulfur dioxide?

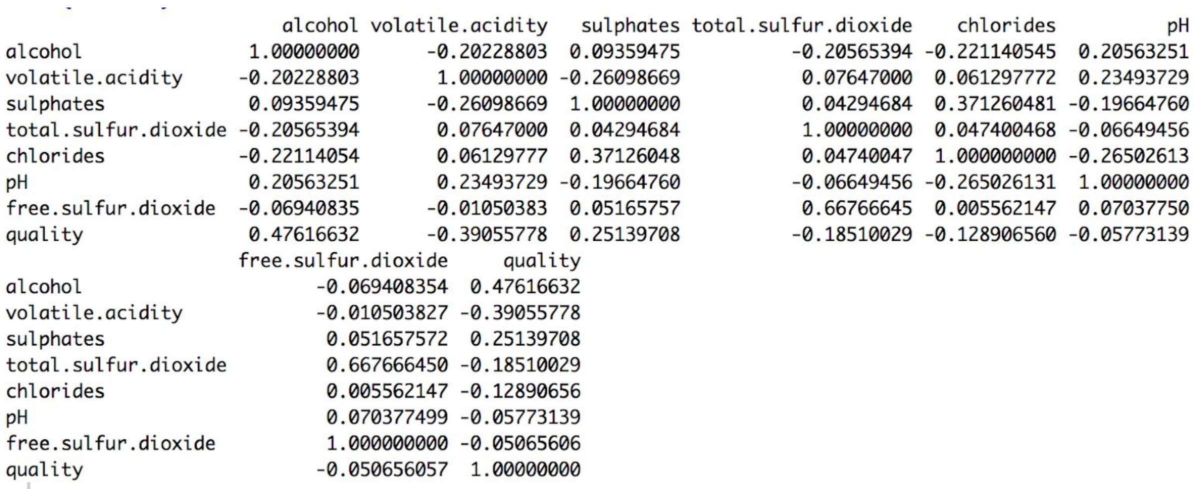


```
Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,
    data = red_wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

We notice from the output of the summary() function of our updated model that the coefficient of determination is 0.3595, implying that 35.95% of all variation of quality is explained by taking the variation of the alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and free sulfur dioxide variables into account. While this number appears to be quite small, summary(mod1), where mod1 is the regression model with all of our original predictors, still supports that this is not due to a poor choice of predictors for our narrowed model as the adjusted R-squared for the original model is smaller than the adjusted R-squared of our final model. Since adjusted R-squared does not always necessarily increase when more predictors are added like the R-squared coefficient of determination tends to, the fact that the new model has a larger adjusted R-squared once again supports that our new model is the most efficient linear regression model to use.

Furthermore, the F-statistic of the associated given model we used is relatively large for there being 7 and 1591 degrees of freedom. It additionally has a very small associated p-value, indicating that the model is well adjusted and confirming our original assumption that was originally predicted.

Now that we have proven that our updated model is appropriate, we can answer the next questions. To determine the relationship between all of the variables, we can look at the correlation matrix from the cor() function.



```
                         alcohol volatile.acidity   sulphates total.sulfur.dioxide    chlorides          pH
alcohol               1.00000000      -0.20228803  0.09359475          -0.20565394 -0.221140545  0.20563251
volatile.acidity     -0.20228803       1.00000000 -0.26098669           0.07647000  0.061297772  0.23493729
sulphates             0.09359475      -0.26098669  1.00000000           0.04294684  0.371260481 -0.19664760
total.sulfur.dioxide -0.20565394       0.07647000  0.04294684           1.00000000  0.047400468 -0.06649456
chlorides            -0.22114054       0.06129777  0.37126048           0.04740047  1.000000000 -0.26502613
pH                    0.20563251       0.23493729 -0.19664760          -0.06649456 -0.265026131  1.00000000
free.sulfur.dioxide  -0.06940835      -0.01050383  0.05165757           0.66766645  0.005562147  0.07037750
quality               0.47616632      -0.39055778  0.25139708          -0.18510029 -0.128906560 -0.05773139
                     free.sulfur.dioxide     quality
alcohol                     -0.069408354  0.47616632
volatile.acidity            -0.010503827 -0.39055778
sulphates                    0.051657572  0.25139708
total.sulfur.dioxide         0.667666450 -0.18510029
chlorides                    0.005562147 -0.12890656
pH                           0.070377499 -0.05773139
free.sulfur.dioxide          1.000000000 -0.05065606
quality                     -0.050656057  1.00000000
```

From this, we can see that the largest negative correlation is -0.39055778, between volatile acidity and quality. Although this is not relatively large since this value is closer to 0 than -1, we can say to a certain extent that as volatile acidity value increases, red wine quality decreases. We can also see that the largest positive correlation is 0.47616632, between alcohol and quality. This value is also not overwhelmingly large, but it again demonstrates that, to a certain extent, as the amount of alcohol in red wine increases, the red wine's quality will increase as well.

Finally, we can solve for the prediction of what the quality of red wine would be with average (given the data) status, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol levels? We can find this value along with the value of the 95% confidence interval for the mean amounts using the predict() function. From these functions, we yield that red wine with average values of the independent variables from the new model has a predicted quality of 5.636023. The 95% prediction interval is [4.365106, 6.906939], and 95% confidence interval is [5.60424, 5.667795]. We can see that the prediction interval is accurately wider than the confidence interval since the prediction interval formula utilizes an extra MSE term.

5. Conclusion

Overall, we are assured that the red wine's fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol can be used to predict the classification of the wine measured by the quality. However, following the statement from the UCI Machine Learning Repository Website that “we are not sure if all input variables are relevant”, our various tests found that this overall statement can be condensed,while still precise, to be that the red wine’s alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and free sulfur dioxide have more of an important role as to predicting the overall quality of wine.

Specifically, each of the independent variables have certain effects on red wine quality. Alcohol denotes the percentage of alcohol by volume of the wine type. Volatile acidity represents the amount of steam distillable acids present in wine. Sulphates are responsible for maintaining the wines freshness, and total sulfur dioxide is the portion of Sulfur Dioxide that is free in the wine plus the portion that is bound to other chemicals in the wine. The amount of chlorides depends on the geographic, geologic and climatic conditions of vine culture, and the pH is a way to measure ripeness in relation to acidity. Finally, free sulfur dioxide is used throughout all stages of the winemaking process to prevent oxidation and microbial growth.

Additional types of tests that would have been interesting to administer (but is beyond the content taught in this course) would be tests for multicollinearity, or any types of interdependence between the predictors. From the predictors present, I would guess that there would be a directly proportional relationship between alcohol content and residual sugars, since it would make sense that more sugars would be necessary in the wine as alcohol content increased, so as to make the wine taste less bitter and more sweet. We could have potentially tested for interdependence of predictors to see if there existed any relationships that would better predict wine quality, and could have been able to even eliminate potential redundant predictors to the model.

6. Appendix

```
#Set working directory
getwd()
```

```
setwd("/Users/Nivi/Documents/UCSB 2019-20/Spring 2020/PSTAT 126")
getwd()

#Read data into R
red_wine = read.csv(“winequality-red.csv”, header = TRUE)
View(red_wine)

#Scatterplot matrix
pairs(red_wine)

#Save all variables
quality = red_wine$quality
fixed_acid = red_wine$fixed.acidity
volatile_acid = red_wine$volatile.acidity
citric_acid = red_wine$citric.acid
residual_sugar = red_wine$residual.sugar
chloride = red_wine$chloride
free_sulfur = red_wine$free.sulfur.dioxide
total_sulfur = red_wine$total.sulfur.dioxide
density = red_wine$density
ph = red_wine$pH
sulphates = red_wine$sulphates
alcohol = red_wine$alcohol

#Run stepwise regression
mod0 <- lm(quality ~ 1, data = red_wine)
mod1 <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid +
residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol, data =
red_wine)
step(mod0, scope = list(lower=mod0, upper = mod1))
#smallest AIC is -1380.79, so new model is lm(quality ~ alcohol +
volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + pH +
free.sulfur.dioxide, data = red_wine)

#Run leaps procedure with full model
leaps(cbind(fixed_acid, volatile_acid,
citric_acid,residual_sugar,chloride,free_sulfur,total_sulfur,density,p
h,sulphates,alcohol),quality,method = "adjr2")

#Run leaps procedure with reduced model to compare
leaps(cbind(volatile_acid,chloride,free_sulfur,total_sulfur,ph,sulphat
es,alcohol),quality,method = "adjr2")
```

```
#Final model
model = lm(quality ~ alcohol + volatile.acidity + sulphates +
total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide, data =
red_wine)

#Initial Residual vs fitted values
residuals = residuals(model)
fit = fitted(model)
plot(fit, residuals, main = ‘Residuals vs. Fit’)

#Get QQ plot to check for normality
qqnorm(residuals) #skewed
qqline(residuals, col = “red”)

#Run BoxCox to check for whether transformation necessary
library(MASS)
boxcox = boxcox(quality ~ alcohol + volatile.acidity + sulphates +
total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide, data =
red_wine)

#Further analyze model with coefficient of determination, F-statistic, &
p-value
summary(model)


#use cor function
cor_data<-red_wine[,c('alcohol','volatile.acidity','sulphates','total.
sulfur.dioxide','chlorides','pH','free.sulfur.dioxide','quality')]
cor(cor_data)


#prediction interval
model_final =
predict(model_final, newdata = data.frame(alcohol = mean(alcohol),
volatile_acid = mean(volatile_acid),sulphates = mean(sulphates),
total_sulfur = mean(total_sulfur),chloride = mean(chloride), ph =
mean(ph), free_sulfur = mean(free_sulfur)),interval = "prediction",
level = 0.95)
```

```
#confidence interval
predict(model_final, newdata = data.frame(alcohol = mean(alcohol),
volatile_acid = mean(volatile_acid),sulphates = mean(sulphates),
total_sulfur = mean(total_sulfur),chloride = mean(chloride), ph =
mean(ph), free_sulfur = mean(free_sulfur)),interval = "confidence",
level = 0.95)

#regsubsets
library(leaps)
mod_proj =
regsubsets(cbind(fixed_acid,volatile_acid,citric_acid,residual_sugar,c
hloride,free_sulfur,total_sulfur,density,ph,sulphates,alcohol),quality
)
```