STAT 425 Case Study Executive Summary

Authors: Wenxiao Yang (wenxiao5), Giang Le (gianghl2), Yuan Chang (changy4)

Introduction

In this case study, we focus on the red wine data set. There are ten independent variables and one dependent variable. The ten independent variables are: fixed acidity (It is the most acids involved with wine or fixed or nonvolatile), volatile acidity (It is the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste), citric acid (it is an organic compound), residual sugar (it is the natural grape sugars left over in a wine after the alcoholic fermentation is complete), chlorides (it is a compound of chlorine), free sulfur dioxide (it is the sulfur dioxide that not reacted), total sulfur dioxide (it is the total amount of sulfur dioxide), density (it is mass divided by volume), pH (it is a measurement of hydroge), sulphates (a salt or ester of sulfuric acid). The one dependent variable is alcohol level.

The goal of our case study is to fit a model that best describes the relationship between the alcohol (which is the dependent variable) and physiochemical information (which are independent variables).

Regression Analysis

The goal of our regression analysis is to fit a linear model to predict the response variable **alcohol** from the remaining ten predictors. We began this analysis by reviewing the Correlation Matrix between the ten predictors and ran a full model containing all ten physiochemical information predictors. We observed that there is some correlation between **fixed.acidity** and **citric.acid** (0.67), **fixed.acidity** and **density** (0.67), **fixed.acidity** and **pH** (-0.68), **total.sulfur.dioxide** and **free.sulfur.dioxide** (0.67).

The Full Model Regression Result shows that nine predictor variables are statistically significant and one predictor (**free.sulfur.dioxide**) is not. The R^2 value of this model is 0.668, which could be interpreted that approximately 66.8 percent of the variance in the response variable is explained by the model.

Because collinearity was detected by our review of the correlation matrix and **free.sulfur.dioxide** is not a significant predictor in the full model, we considered removing **free.sulfur.dioxide** first, and potentially we could also remove other predictors such as **fixed.acidity** and **citric.acid** from the model.

Removing **free.sulfur.dioxide** from the predictors produced a new model with the same R^2. All predictors are statistically significant and the p-value of **total.sulfur.dioxide** also decreased significantly, suggesting that this is an improved model compared to the first regression model. See Reduced Model for the full summary.

We confirmed this by running a partial F-test to compare two models. Our null hypothesis is that the reduced model is adequate (or the partial slope of **free.sulfur.dioxide** is 0) and the alternative hypothesis is that the full model is required (the partial slope of **free.sulfur.dioxide** is not 0). The p-value of this partial F-test is greater than 0.05, so we fail to reject the null hypothesis. Therefore, we have evidence that the reduced model is preferred to the full model and we can drop **free.sulfur.dioxide**.

We confirmed that no collinearity is present in the current model by calculating the condition number (~5.26 < 30).

Diagnostics and Remedial Measures

We ran several diagnostic checks to verify outliers, high leverage points, and highly influential points, as well as our model assumptions.

To detect unusual observations, we combined graphical and numerical tools for the diagnostics. We tried to find observations with leverages more than 2p/n where p is the number of predictors plus 1 and n is the number of observations. We found a number of high leverage points. We applied the outliers test using studentized residuals and found no outliers. Finally, we checked for highly influential points using the Cook's distance and because the maximum Cook's distance is 0.07, we can conclude that there is no highly influential points. Although there are no high influential points based on the rule-of-thumb of the Cook's distance, there are two observations where the Cook's distance are much larger than that in other observations. See Detecting Unusual Observation Plots to review the plots that checked for high leverage and highly influential points.

By plotting the residuals vs. fitted values of the current model and running the studentized Breusch-Pagan test, we found that the assumption of constant error variance is violated. We also checked for normality by plotting a normal Q-Q plot, a histogram of residuals, and running a Kolmogorov-Smirnov test. Even though the histogram of residuals does not show a skewed distribution of residuals, the Kolmogorov-Smirnov test result suggested that the assumption of normality of errors does not hold. See Checking Model Assumptions to view these results in more detail.

Because our response variable contains only positive values, Box-Cox transformation is a possible remedy to violations of assumptions. The lambda value found by the Box-Cox method is -1.53 (see Box-Cox Lambda), so we transformed the response variable by applying the following function to the response variable.

$$g(\mathbf{y}) = \frac{y_i^{\lambda} - 1}{\lambda}$$

and reran the regression model as model_bx. All predictors are found to be statistically significant and the R^2 value is 0.649. See Box-Cox Model for the result summary.

We checked the unusual observations and assumptions again by using a combination of graphical and numerical tools like before and found that we have high leverage observations, no outliers, and no highly influential observations. By the studentized Breusch-Pagan test, we found that homoscedasticity still does not hold and by the Kolmogorov-Smirnov test we found that normality also still does not hold.

We checked for non-linearity by plotting nine added-variable plots, one for each predictor variable. All variables look fine except for residual.sugar because here the points do not look like they scatter around the fitted line, so we decided to transform it to log(residual.sugar). (See Added-Variable Plot for Residual.Sugar) for more detail. After the transformation, the added-variable plot for residual-sugar looks much better.

Based on the non-linearity diagnostic, we took the logs of both the residual.sugar variable and the response variable and re-ran the multivariate linear regression. This modelN has a higher R^2 value compared to previous models (0.71) but the volatile.acidity variable's p value increased. See Log-Transformed Model Result for more detail.

We decided to remove volatile.acidity even though it is still a significant predictor at alpha 0.05, but it is not at lower significance levels. ModelN2 is the model without volatile.acidity and the R^2 value is 0.70. We re-ran the diagnostic checks and found some high leverage points, one outlier, and no highly influential points. By the studentized Breusch-Pagan test, we found that homoscedasticity still does not hold and by the Kolmogorov-Smirnov test we found that normality also still does not hold.

Finally, we tried to apply Generalized Least Squares regression, keeping the log transformations applied earlier to see if we could fix the violations of assumptions. All predictor variables are statistically significant at the lowest significance level reported in R and the R^2 value increased to 0.76, or approximately 76% of variance in the response variable can be explained by the model. This is an improvement compared to our previous models. See GLS Result for more detail.

We re-ran the diagnostic checks again and we found some high leverage points, no outliers, and no highly influential points. By the studentized Breusch-Pagan test, we found that homoscedasticity still does not hold and by the Kolmogorov-Smirnov test we found that normality also still does not hold. Despite being unable to find a model that satisfies these assumptions, by measure of fit, generalized least square regression gave us a better model relative to other methods. It is possible that more advanced techniques need to be used to handle violations of model assumptions.

Conclusion

In conclusion, we found a nicely fitted model for the red wine data set. By eliminating two independent variables, doing diagnostic checks, and then applying the Generalized Least Square Regression, we successfully built our regression model. There are eight independent variables in the regression model and they are: fixed acidity, citric acid, residual sugar (transformed to log), chlorides, total sulfur dioxide,density, pH, and sulphates. In this Generalized From the model, we can conclude that these eight independent variables are statistically significant in predicting the alcohol level in a multivariate linear model. By checking the diagnostics, we found no outliers, no highly influential points, and some high leverage points which are likely to be harmless as they are neither outliers nor highly influential points. Despite the persisting violations of assumptions, by measure of fit, generalized least square regression gave us a better model relative to other methods.

Appendix

Correlation Matrix

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                  1.00            -0.26        0.67           0.11
## volatile.acidity              -0.26             1.00       -0.55           0.00
## citric.acid                    0.67            -0.55        1.00           0.14
## residual.sugar                 0.11             0.00        0.14           1.00
## chlorides                      0.09             0.06        0.20           0.06
## free.sulfur.dioxide           -0.15            -0.01       -0.06           0.19
## total.sulfur.dioxide          -0.11             0.08        0.04           0.20
## density                        0.67             0.02        0.36           0.36
## pH                            -0.68             0.23       -0.54          -0.09
## sulphates                      0.18            -0.26        0.31           0.01
##                       chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity              0.09               -0.15                -0.11    0.67
## volatile.acidity           0.06               -0.01                 0.08    0.02
## citric.acid                0.20               -0.06                 0.04    0.36
## residual.sugar             0.06                0.19                 0.20    0.36
## chlorides                  1.00                0.01                 0.05    0.20
## free.sulfur.dioxide        0.01                1.00                 0.67   -0.02
## total.sulfur.dioxide       0.05                0.67                 1.00    0.07
## density                    0.20               -0.02                 0.07    1.00
## pH                        -0.27                0.07                -0.07   -0.34
## sulphates                  0.37                0.05                 0.04    0.15
##                             pH sulphates
## fixed.acidity            -0.68      0.18
## volatile.acidity          0.23     -0.26
## citric.acid              -0.54      0.31
## residual.sugar           -0.09      0.01
## chlorides                -0.27      0.37
## free.sulfur.dioxide       0.07      0.05
## total.sulfur.dioxide     -0.07      0.04
## density                  -0.34      0.15
## pH                        1.00     -0.20
## sulphates                -0.20      1.00
```

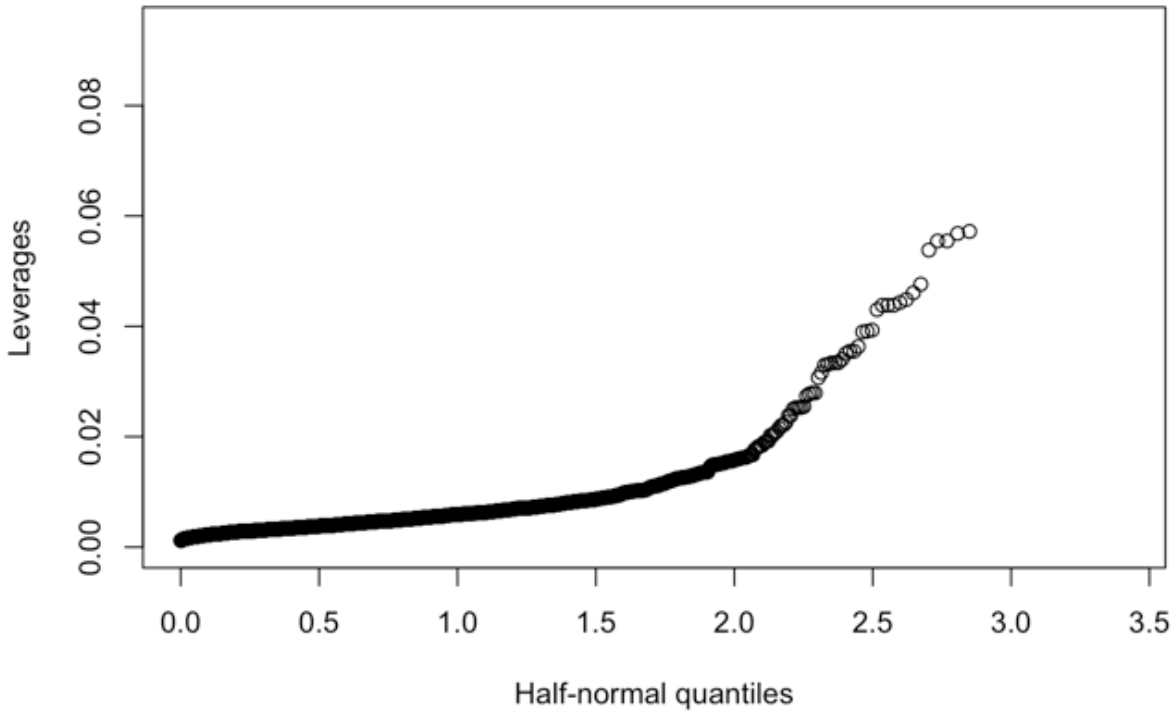
Full Regression Model Summary

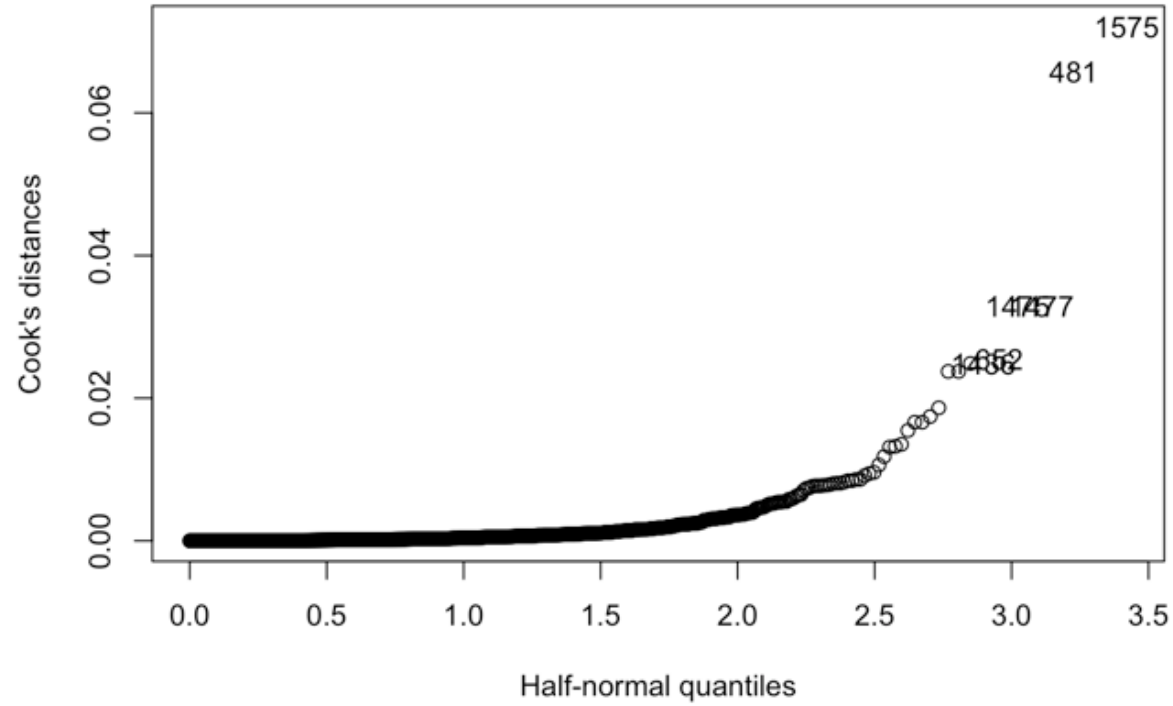
```
##
## Call:
## lm(formula = alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates, data = redwines[, -c(12)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07175 -0.39267 -0.04056  0.35396  2.44365
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.072e+02  1.308e+01  46.419  < 2e-16 ***
## fixed.acidity         5.324e-01  2.064e-02  25.796  < 2e-16 ***
## volatile.acidity      3.608e-01  1.144e-01   3.154 0.001638 **
## citric.acid           8.306e-01  1.379e-01   6.024 2.11e-09 ***
## residual.sugar        2.844e-01  1.229e-02  23.135  < 2e-16 ***
## chlorides            -1.462e+00  3.956e-01  -3.696 0.000227 ***
## free.sulfur.dioxide  -2.143e-03  2.057e-03  -1.042 0.297517
## total.sulfur.dioxide -2.296e-03  6.881e-04  -3.336 0.000868 ***
## density              -6.174e+02  1.342e+01 -45.998  < 2e-16 ***
## pH                    3.762e+00  1.551e-01  24.263  < 2e-16 ***
## sulphates             1.247e+00  1.037e-01  12.020  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.614 on 1588 degrees of freedom
## Multiple R-squared:  0.6701, Adjusted R-squared:  0.668
## F-statistic: 322.5 on 10 and 1588 DF,  p-value: < 2.2e-16
```

Reduced Model Without Collinearity

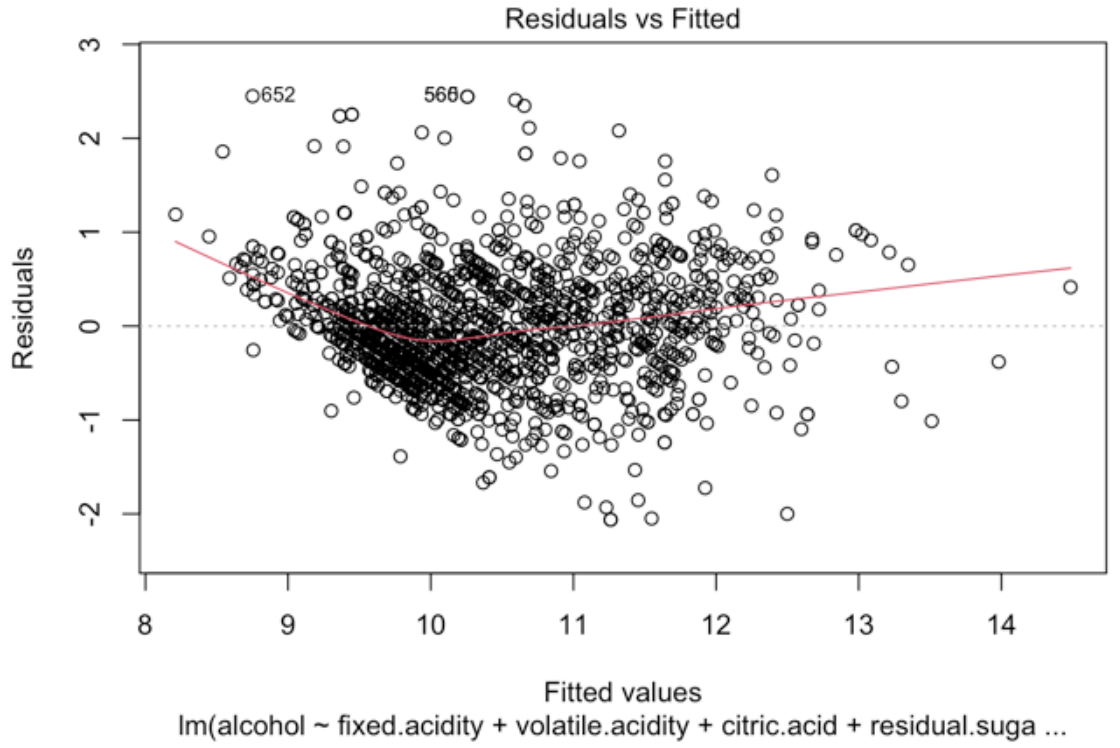
```
##
## Call:
## lm(formula = alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + total.sulfur.dioxide + density +
##     pH + sulphates, data = redwines[, -c(12)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06145 -0.39706 -0.03917  0.34928  2.44848
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.059e+02  1.302e+01  46.535  < 2e-16 ***
## fixed.acidity         5.300e-01  2.051e-02  25.846  < 2e-16 ***
## volatile.acidity      3.809e-01  1.128e-01   3.377 0.000749 ***
## citric.acid           8.548e-01  1.359e-01   6.289 4.12e-10 ***
## residual.sugar        2.827e-01  1.219e-02  23.198  < 2e-16 ***
## chlorides            -1.487e+00  3.949e-01  -3.766 0.000172 ***
## total.sulfur.dioxide -2.775e-03  5.123e-04  -5.416 7.02e-08 ***
## density              -6.160e+02  1.335e+01 -46.125  < 2e-16 ***
## pH                    3.739e+00  1.534e-01  24.369  < 2e-16 ***
## sulphates             1.242e+00  1.036e-01  11.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.614 on 1589 degrees of freedom
## Multiple R-squared:  0.6699, Adjusted R-squared:  0.668
## F-statistic: 358.2 on 9 and 1589 DF,  p-value: < 2.2e-16
```

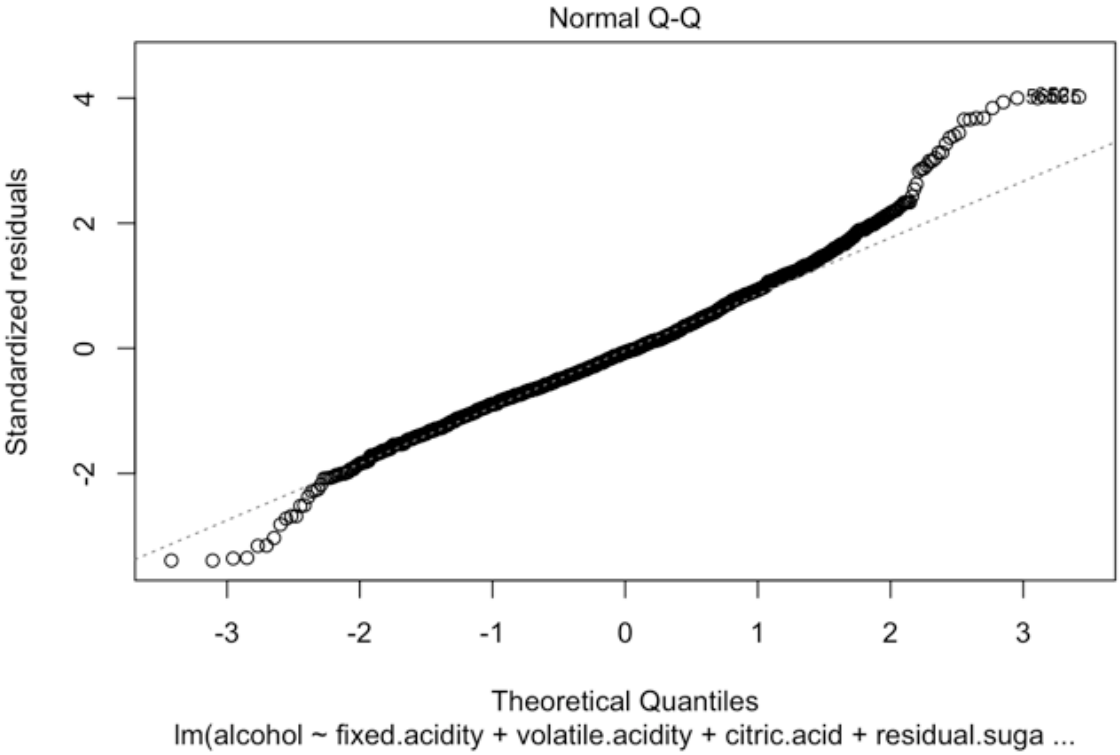
Detecting Unusual Observation Plots



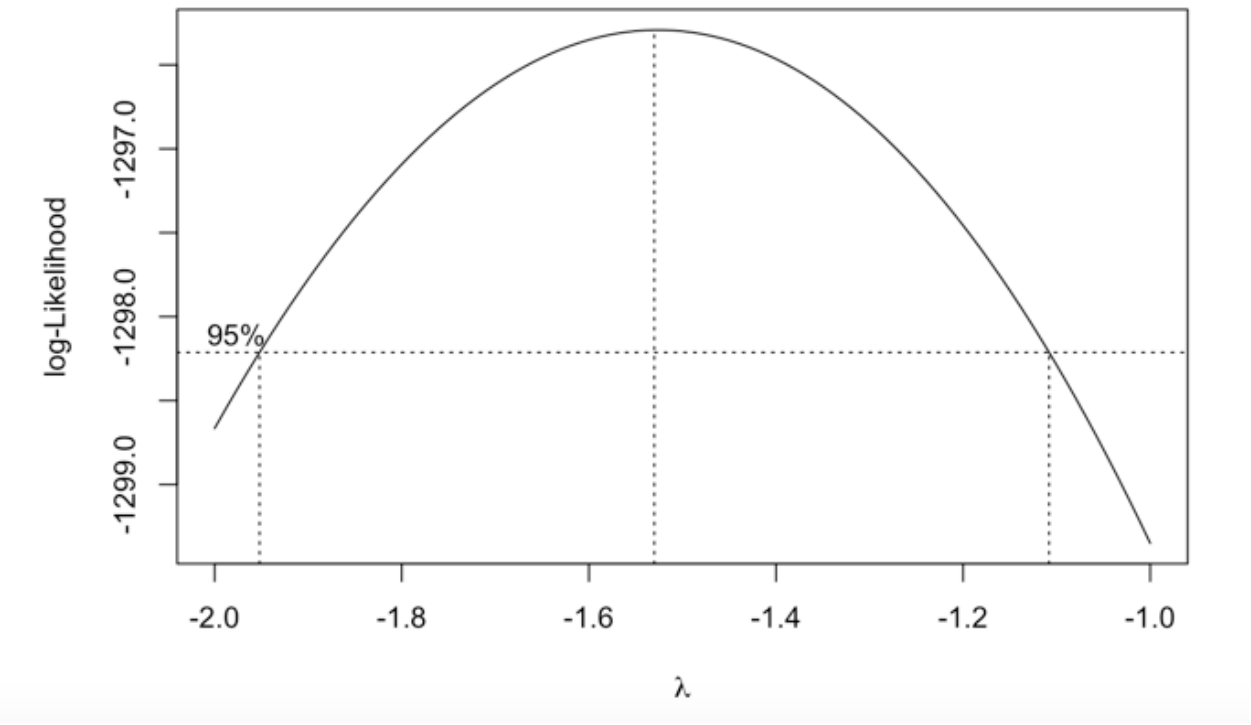


Checking Model Assumptions





Box-Cox Lambda

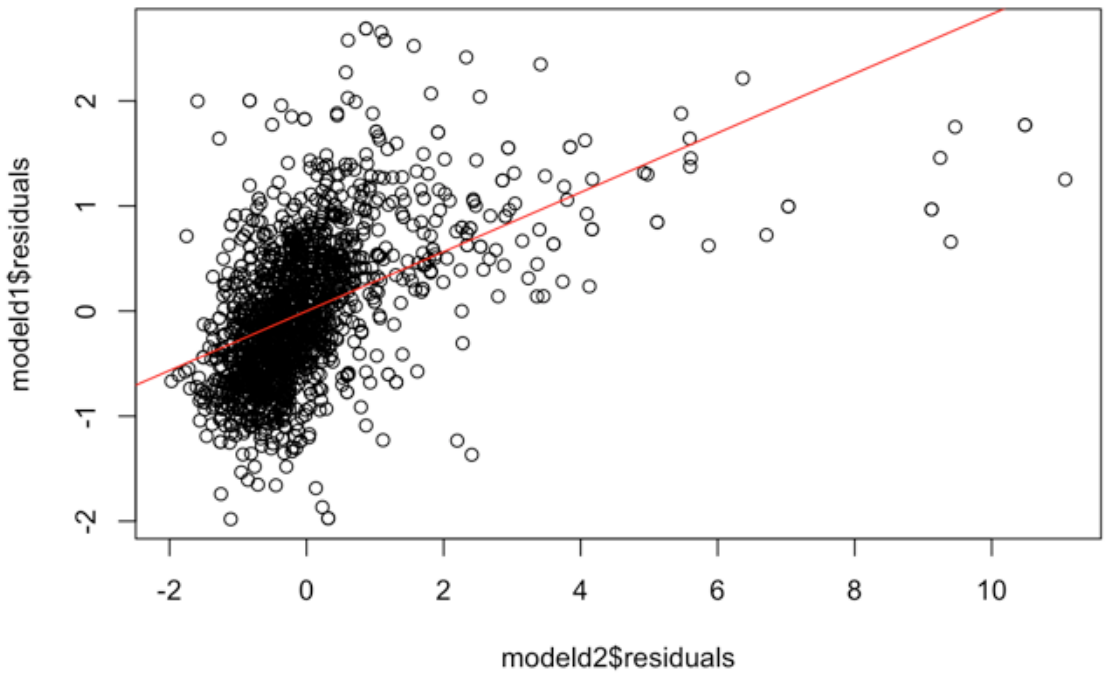


Box-Cox Model

```
model_bx=lm((alcohol^(-1.53)-1)/(-1.53) ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlor
ides +  total.sulfur.dioxide + density + pH + sulphates, data=redwines[,-c(12)])
summary(model_bx)
```

```
##
## Call:
## lm(formula = (alcohol^(-1.53) - 1)/(-1.53) ~ fixed.acidity +
##     volatile.acidity + citric.acid + residual.sugar + chlorides +
##     total.sulfur.dioxide + density + pH + sulphates, data = redwines[,
##     -c(12)])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0056555 -0.0010128 -0.0000063  0.0009974  0.0064010
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.076e+00  3.340e-02  62.155  < 2e-16 ***
## fixed.acidity         1.304e-03  5.260e-05  24.785  < 2e-16 ***
## volatile.acidity      9.833e-04  2.893e-04   3.399 0.000692 ***
## citric.acid           2.060e-03  3.487e-04   5.908 4.23e-09 ***
## residual.sugar        6.800e-04  3.126e-05  21.751  < 2e-16 ***
## chlorides            -4.670e-03  1.013e-03  -4.611 4.33e-06 ***
## total.sulfur.dioxide -8.396e-06  1.314e-06  -6.389 2.18e-10 ***
## density              -1.491e+00  3.426e-02 -43.529  < 2e-16 ***
## pH                    9.233e-03  3.936e-04  23.461  < 2e-16 ***
## sulphates             3.181e-03  2.658e-04  11.970  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001575 on 1589 degrees of freedom
## Multiple R-squared:  0.6509, Adjusted R-squared:  0.649
## F-statistic: 329.2 on 9 and 1589 DF,  p-value: < 2.2e-16
```

Added-Variable Plot for Residual.Sugar



Log-Transformed Model

```
##
## Call:
## lm(formula = log(alcohol) ~ fixed.acidity + volatile.acidity +
##     citric.acid + log(residual.sugar) + chlorides + total.sulfur.dioxide +
##     density + pH + sulphates, data = redwines[, -c(12)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21937 -0.03359 -0.00364  0.03292  0.24246
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.974e+01  1.146e+00  52.119  < 2e-16 ***
## fixed.acidity         4.882e-02  1.776e-03  27.480  < 2e-16 ***
## volatile.acidity      2.479e-02  9.840e-03   2.519 0.011862 *
## citric.acid           6.561e-02  1.188e-02   5.521 3.93e-08 ***
## log(residual.sugar)   1.238e-01  4.292e-03  28.834  < 2e-16 ***
## chlorides            -1.238e-01  3.448e-02  -3.591 0.000339 ***
## total.sulfur.dioxide -3.102e-04  4.470e-05  -6.939 5.73e-12 ***
## density              -5.930e+01  1.175e+00 -50.477  < 2e-16 ***
## pH                    3.355e-01  1.335e-02  25.128  < 2e-16 ***
## sulphates             1.169e-01  9.030e-03  12.941  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05355 on 1589 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7069
## F-statistic: 429.1 on 9 and 1589 DF,  p-value: < 2.2e-16
```

GLS Model

```
##
## Call:
## lm(formula = log(alcohol) ~ fixed.acidity + citric.acid + log(residual.sugar) +
##     chlorides + total.sulfur.dioxide + density + pH + sulphates,
##     data = redwines, weights = weight)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1487 -0.8459 -0.0803  0.7985  4.9741
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.324e+01  1.040e+00  60.788  < 2e-16 ***
## fixed.acidity         5.170e-02  1.736e-03  29.781  < 2e-16 ***
## citric.acid           4.506e-02  9.327e-03   4.831 1.49e-06 ***
## log(residual.sugar)   1.258e-01  4.370e-03  28.784  < 2e-16 ***
## chlorides            -1.114e-01  2.447e-02  -4.552 5.72e-06 ***
## total.sulfur.dioxide -2.429e-04  4.076e-05  -5.959 3.12e-09 ***
## density              -6.283e+01  1.067e+00 -58.909  < 2e-16 ***
## pH                    3.355e-01  1.196e-02  28.057  < 2e-16 ***
## sulphates             1.201e-01  8.004e-03  15.004  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.288 on 1590 degrees of freedom
## Multiple R-squared:  0.7581, Adjusted R-squared:  0.7569
## F-statistic:   623 on 8 and 1590 DF,  p-value: < 2.2e-16
```