Final Project Template

Sarah Worthington, Bruce Decker, Austin Rogers

Abstract

< We used the wine data from UC Irvine to see what affects wine's alcohol(%vol). We wondered if residual sugar affects alcohol differently for red versus white wine, and if quality of the wine impact alcohol(%vol), controlling for all other variables. We then used transformations and model selection to refine our model. We also checked linearity assumptions using diagnostic plots We conclude that quality is a significant predictor of alcohol. There was also a significant interaction between sugar and wine type. We hope our results help wine experts in assessing a wine's alcohol percent of total volume. >

1. Problem and Motivation

< The wine industry is a 300 billion dollar industry world wide. Billions of people drink wine therefore having a model to help assess the alcohol percent of total volume is important. We focused on basis questions people typically ask, such as type of wine, quality, and sugar content. So our project is useful for consumers interested in knowing more about their wine alcohol(%vol). This data set was collected from various wine samples originating in northern Portugal, and is found on the UC Irvine machine learning repository. The data set was created to help researchers model the effects of psychochemical and sensory variables on wine. Interestingly no variables such as grape type, wine price, or wine label are included for privacy reasons.However, the variables provided will still be sufficient to develop an effective model for assessing alcohol(%vol), and examining other variables of interest.>

2. Questions of Interest

< 1. Does the residual sugar content affect alcohol different for red versus white wines? 2. What is the association between quality of the wine and alcohol(%vol.) after controlling for all other variables >

3. Data

< The data set was a sampling of many different properties of a variety of Portuguese wines. Our response variable was alcohol(%vol.). The possible predictors were fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, chlorides, quality, and type. Total sulfur dioxide and chlorides were dropped from the final model during model selection.We were most interested in the predictors quality, residual sugar and type because they were consumer friendly, recognizable, and informative. > Data:

```
Wine_red <- read.table("winequality-red.csv", sep = ";", header= TRUE)
Wine_red$type <- "red"
Wine_White <- read.table("winequality-white.csv", sep = ";", header= TRUE)
Wine_White$type <- "white"
Wine <- rbind(Wine_red,Wine_White)
Wine$quality <- as.numeric(Wine$quality)
```

4. Regression Methods

< Once we have our model to answer question one we assigned wine a dummy variable with red wine as the default group. Then we created a scatterplot of the effects of residual sugar on alcohol(%vol), color coded by wine type, to see if there was a visual difference. Afterwards, based on reasonable evidence in the graph that residual sugar content differs by type of wine, we then tested the interaction between type and residual sugar content by fitting a non-parallel multiple regression model holding all other variables in the model

constant. For our second question we did a regular multiple linear regression test to test the significance of quality on alcohol(%vol) after controlling for other variables. We then broke quality into high(5-9) and low categories(1-4) to assess the difference between high and low quality wines.After which, we will create a box plot of the groups to see if there are any obvious differences. Then we will run a regression with quality dummy coded into high and low groups. Last we reassessed our model to see if it was still a good fit.>

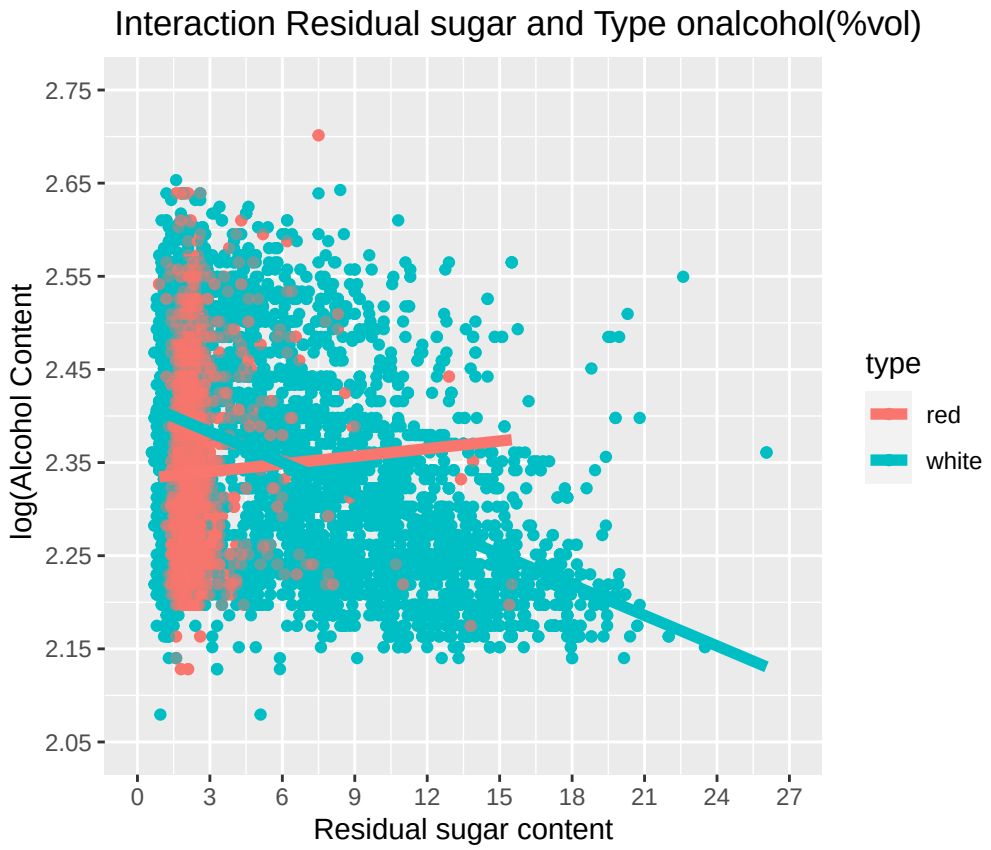
5. Regression Analysis, Results, and Interpretation

Model Selection and Improvements:

```
#final model
Wine.best.lm <- lm(log(alcohol) ~ fixed.acidity +  volatile.acidity + residual.sugar + free.sulfur.diox
```

< Our model: $\log(Alcohol)_i = \beta_0 + (\beta_1 \text{quality}_i) + (\beta_2 \text{fixed.acidity}_i) + (\beta_3 \text{volatile.acidity}_i) + (\beta_4 \text{residual sugar}_i) + (\beta_5 \text{type}_i) + (\beta_6 \text{free.sulfur.dioxide}_i) + (\beta_7 \text{density}_i) + (\beta_8 \text{ph}_i) + (\beta_9 \text{sufates}_i)$ As you can see in the appendix, after reading in the and checking the model assumption plots, we performed model selection using BIC method to eliminate unnecessary predictors, included total.sulfur.dioxide and chlorides. Next we eliminated some outliers and perform transformations.Using Avplots we checked the predictors, and as they all appeared fairly linear we did not transform them. Then we used the box cox method, which suggested a log transform the response variable alcohol. These changes greatly improved our diagnostic plots and model, which was terrible at first. The we tested the model assumptions including linearity, constant variance, and normality assumptions. Independence, was not tested, and may be of some concern. Constant variance was assessed with a scale location plot, and a residuals versus fitted plot.The plots for constant variance were reasonable after our aforementioned adjustments.Although there was one suspect point, overall the variance appeared fairly constant. Normality assessed with a qq plot was pretty deviant, even after our model adjustments.However, the sample size is massive,so this assumption is reasonably met based on the central limit theorem. Linearity was assessed with a residuals versus fitted plot and seemed reasonably met, despite a deviant few points at either end. After all of our transformations the needed assumptions appear met. Question 1:Checking for interaction between type and residual sugar:

```
#plotting the interaction of residual sugar and type onalcohol(%vol)
Wine1 <- Wine[-c(3263,4381,5501,3126,3017,3253, 6345),]
Winelm <- lm(log(alcohol) ~ residual.sugar*type, data = Wine[-c(3263,4381,5501,3126,3017,3253, 6345),])
Wine1$fitted <- Winelm$fitted.values
ggplot(data = Wine1, mapping = aes(x = residual.sugar, y =log(alcohol),color = type)) + geom_point(aes(
geom_point(data = subset(Wine, type == 'red'),alpha=.4) + geom_line(aes(x = residual.sugar, y =fitted,
  labs(title = " Interaction Residual sugar and Type onalcohol(%vol)",
           x = "Residual sugar content",
           y = " log(Alcohol Content") +
           scale_x_continuous(limits = c(0,27),
           breaks = seq(0,27,by = 3)) +
           scale_y_continuous(limits = c(2.05,2.75),
           breaks = seq(2.05,2.75,by = .1)) +
           theme(aspect.ratio = 1)
```




```
logmodel.int <- lm(log(alcohol) ~ fixed.acidity +  volatile.acidity + residual.sugar*type + free.sulfur
summary(logmodel.int)
```

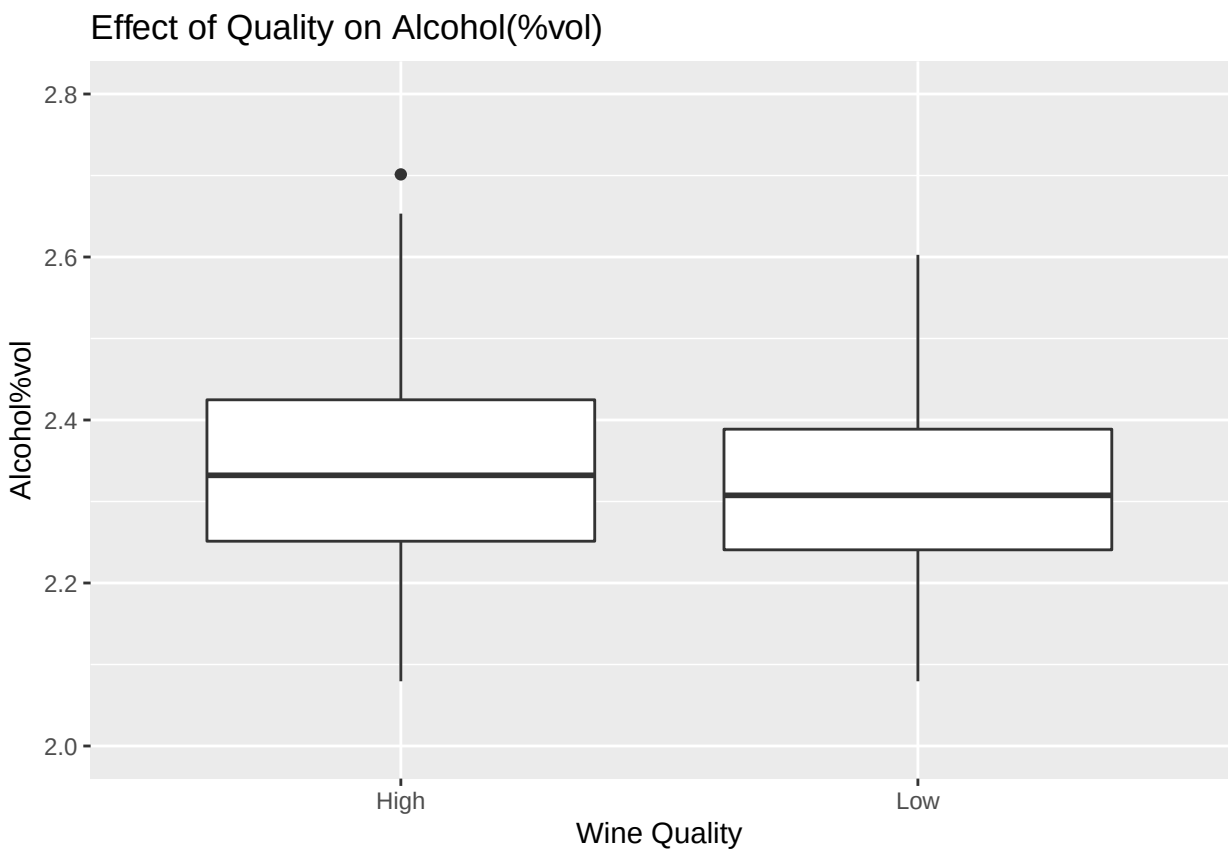
```
##
## Call:
## lm(formula = log(alcohol) ~ fixed.acidity + volatile.acidity +
##     residual.sugar * type + free.sulfur.dioxide + density + pH +
##     sulphates + quality + type, data = Wine[-c(3263, 4381, 5501,
##     3126, 3017, 3253, 6345), ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.214346 -0.025518 -0.001446  0.023259  0.242328
##
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)               6.471e+01  4.238e-01  152.680  < 2e-16 ***
## fixed.acidity             5.399e-02  6.721e-04   80.334  < 2e-16 ***
## volatile.acidity          3.801e-02  4.443e-03    8.555  < 2e-16 ***
## residual.sugar            2.806e-02  7.688e-04   36.500  < 2e-16 ***
## typewhite                -9.443e-02  3.192e-03  -29.585  < 2e-16 ***
## free.sulfur.dioxide      -2.051e-04  3.654e-05   -5.612 2.08e-08 ***
## density                  -6.405e+01  4.322e-01 -148.199  < 2e-16 ***
## pH                        2.513e-01  4.326e-03   58.098  < 2e-16 ***
## sulphates                 9.510e-02  4.177e-03   22.768  < 2e-16 ***
## quality                   6.668e-03  7.093e-04    9.400  < 2e-16 ***
## residual.sugar:typewhite -6.805e-03  7.608e-04   -8.944  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04209 on 6479 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.8569
## F-statistic:  3887 on 10 and 6479 DF,  p-value: < 2.2e-16
```

```
## there appears to be a strong statistical difference between wine type and residual sugar
```

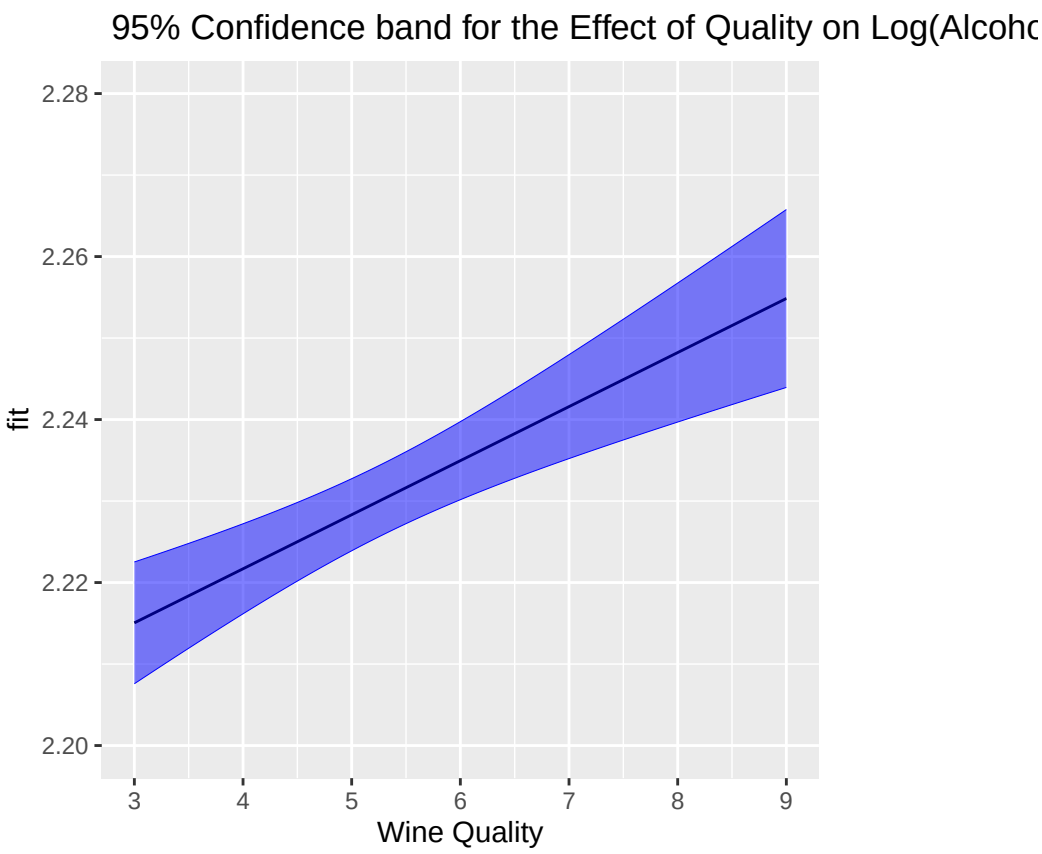
< The null hypothesis,is that the interaction between sugar and wine type is equal to 0 controlling for all other variables. The alternative hypothesis is that the interaction between sugar and wine type is not equal to 0 controlling also else equal. Our results show that the interaction of wine type and sugar on alcohol is significant after controlling for all the other variables.The effect on log(alcohol) is negative and significant, and the estimated coefficient of -.006805 has a p value of 2e-16, which is extremely statistically significant. Thus, we conclude there is a strong reason to use a non-parallel model and that the the type of wine does make a difference in residual sugar content's effect on alcohol(%vol) even after controlling for all other variables. Our results indicate that the type of wine moderates the effects of residual sugar on alcohol(%vol).
> Question 2:Checking for an association between quality of wine and alcohol(%vol)

```
#overall impact of quality
summary(Wine.best.lm)
#Using dummy variables to investigate further:
Wine$qualityHL[Wine$quality<5] <- 0
Wine$qualityHL[Wine$quality>4] <- 1
Wine$qualityHL <- as.character(Wine$qualityHL)
Wine$qualityHL[Wine$qualityHL == "0"] <-"Low"
Wine$qualityHL[Wine$qualityHL == "1"] <-"High"
#boxplot of Quality Dummy Variables
ggplot(Wine, aes(x=qualityHL, y=log(alcohol))) +
    geom_boxplot() + labs(title = "Effect of Quality on Alcohol(%vol)",
       x = "Wine Quality",
       y = "Alcohol%vol") + scale_y_continuous(limits = c(2 ,2.8),
        breaks = seq(2 ,2.8,by = .2))
```




```
HighLowLog <- lm(log(alcohol) ~ fixed.acidity +  volatile.acidity + residual.sugar + free.sulfur.dioxid
summary(HighLowLog)
#Prepare the data for 95% confidence bands
EffectPlotModel <- lm(log(alcohol) ~ quality + fixed.acidity +
volatile.acidity + residual.sugar + free.sulfur.dioxide + density + pH +
sulphates + type, data = Wine[-c(3263,4381,5501,3126,3017,3253, 6345),])
wineVec <- seq(min(Wine$quality), max(Wine$quality), length.out = 6497)
NewData <- data.frame(quality = wineVec, fixed.acidity =
rep(median(Wine$fixed.acidity), 6497), volatile.acidity =
rep(median(Wine$volatile.acidity), 6497), residual.sugar =
rep(median(Wine$residual.sugar), 6497), free.sulfur.dioxide =
rep(median(Wine$free.sulfur.dioxide), 6497), density = rep(median(Wine$density),
6497), pH = rep(median(Wine$pH), 6497), sulphates = rep(median(Wine$sulphates),
6497), type = rep(median(Wine$type), 6497))
WineFittedValues <- predict(EffectPlotModel, newdata = NewData, se.fit =
TRUE)
schef2 <- sqrt(7*qf(0.95, 7, nrow(Wine) - 8))
EffPlotDF <- data.frame(quality = wineVec, fit = WineFittedValues$fit ,
CBlwr = WineFittedValues$fit - schef2*WineFittedValues$se.fit, CBupr =
WineFittedValues$fit + schef2*WineFittedValues$se.fit)
#plot
ggplot(data = EffPlotDF)+
geom_line(mapping = aes(x = quality, y = fit), color = 'black')+
geom_line(mapping = aes(x = quality, y = CBupr), color = 'blue', size =
.01)+
geom_line(mapping = aes(x = quality, y = CBlwr), color = 'blue', size =
.01)+
```

```
geom_ribbon(aes(ymin = CBlwr, ymax = CBupr, x = quality), fill = 'blue',
alpha = .5)+
  labs(title = " 95% Confidence band for the Effect of Quality on Log(Alcohol) of Wine",
       x = "Wine Quality",
       y = "fit") + scale_y_continuous(limits = c(2.20,2.28),
        breaks = seq(2.20,2.28,by = .02)) +
        scale_x_continuous(limits = c(3,9),
        breaks = seq(3,9,by = 1)) +
        theme(aspect.ratio = 1)
```

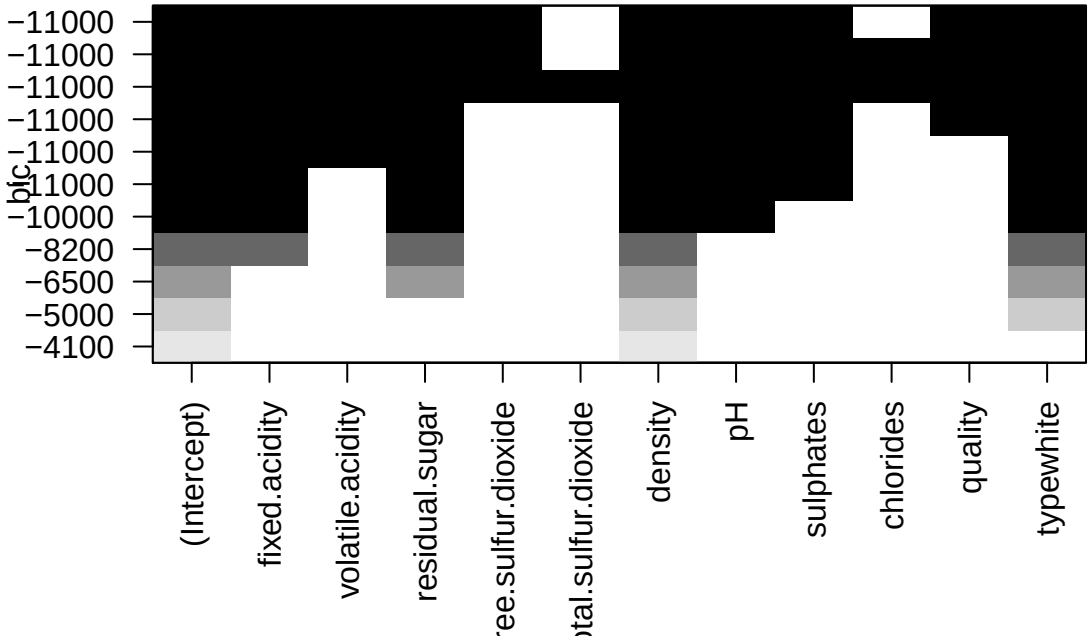


< First we tested if quality was significant overall. Our null hypothesis for that test was the coefficient for quality is equal to 0 all else equal and the alternative hypothesis was the coefficient for quality is not equal to 0 controlling for all other variables. Then we ran a second set of tests where the null hypothesis was that all of the coefficients for the quality dummy variables, high and low, are equal to 0 all else equal. The alternative hypothesis at least one coefficient for the quality dummy variables, high or low, is not equal to 0 all else equal. Based on our first test in general there appears to be a strong (statistically significant) positive relationship between quality of wine and alcohol(%vol).We then set the high quality group of wines(5-9) to be our dummy variable, and graphed a box plot of the two groups. The graph shows some differences between high and low groups, so it is worth further investigation. After running a mutiple linear regression it appears that all else equal High quality wines have an estimated -1.097 effect on the average log (alcohol)(%vol) when compared to low quality wines. Also all else equal changing quality of wine from high to low changes the average alcohol percent volume by $100(e^{-1.097} - 1)$ =-66.6128804. Higher quality wines appear to have more alcohol(%vol). While overall quality is important it may not be the best indicator of how much alcohol is in the wine, but rather having a high or low quality generally may be more important. Our results, show there is an effect of quality overall, but the main difference is really between high an low groupings of levels rather than each different level. We also created a 95% confidence interval band graph for the effects of quality on alcohol controlling for all other variables in the model. This allows us to see graphically the positive relationship

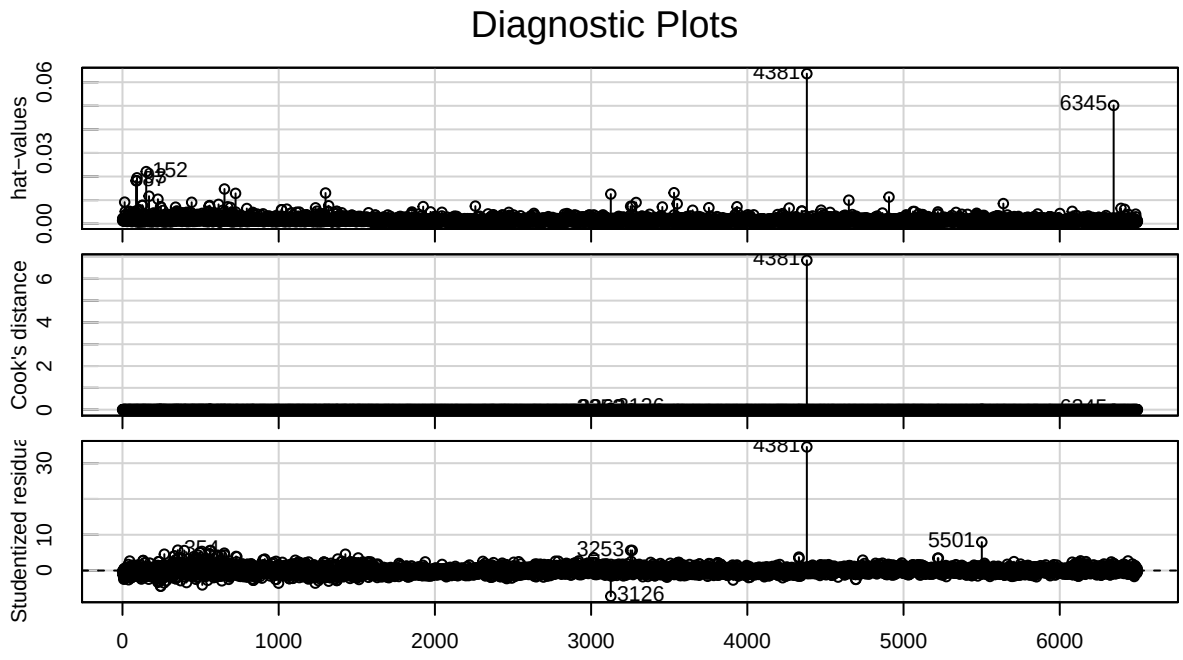
between quality and alcohol.> # 6. Conclusions < Our results showed several large and significant effects for our questions of interest. For question one, different types of wine were shown to have different residual sugar content which in turn affects their alcohol(%vol). This difference is also very statistically significant For question two we found that in fact there were large differences between lower quality wines and higher quality wines, and quality overall was a significant predictor of a wine's alcohol(%vol). These questions have provided reliable analysis that can be used to help consumers understand the alcohol(%vol) of the wine they are consuming.When considering our statistical analysis it is important to consider that we removed a few outliers Although we have large sample size potentially, by filtering out the outliers, we may be ignoring important differences in the data. There also does appear to be some multicollinearity within the model, particularly between density and residual.sugar and density and fixed acidity. However, the multicollinearity is not super large, and so we will not be using weighted least squares regression. Given our data set it also is important to consider unequal group sizes specifically within type of win, there are far more white than red wines. More data should be collected and future analysis should focus on assessing more red wines so comparisons are more robust. Our results are also only generalizable to Portuguese wines, and so more data should be collected with other types of wine to see if these patterns hold up.Last as alcohol was not a completely random variable we should be aware of possible violations of independence. In addition, because our response variable alcohol is reported as a percent it is finite instead of continuous. To account for this alcohol was log transformed. In conclusion, despite its limitations, our analysis of Portugues wine is relevant to helping consumers understand more about how quality, type, and residual sugar all impact alcohol percent volume.

APPENDIX

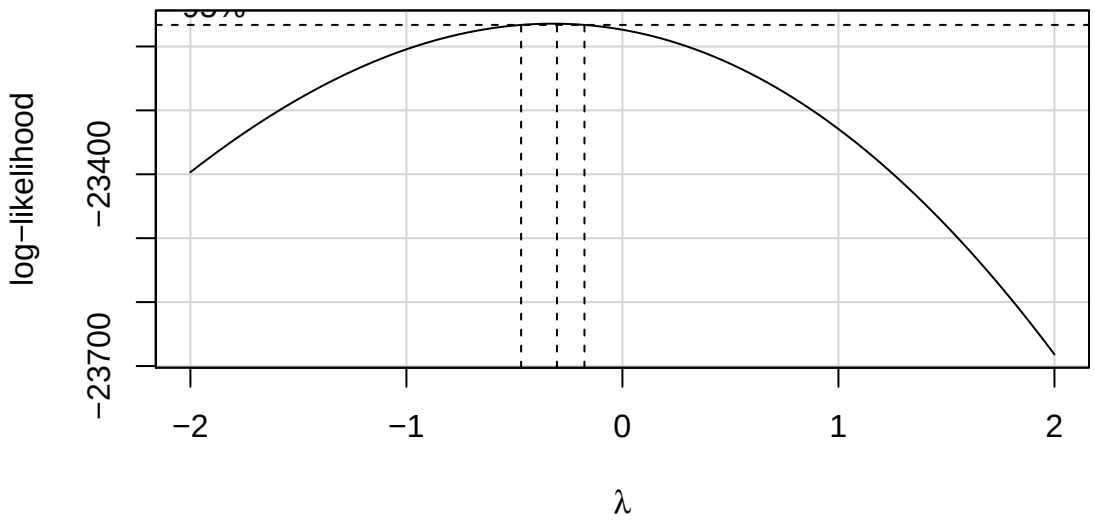
```
#nonadjusted model
Lmmodel <- lm(alcohol ~ fixed.acidity +  volatile.acidity + residual.sugar + free.sulfur.dioxide + dens
# model selection
Wine.bs <- regsubsets(alcohol ~ fixed.acidity +  volatile.acidity + residual.sugar + free.sulfur.dioxid
plot(Wine.bs,method = 'bic')
```



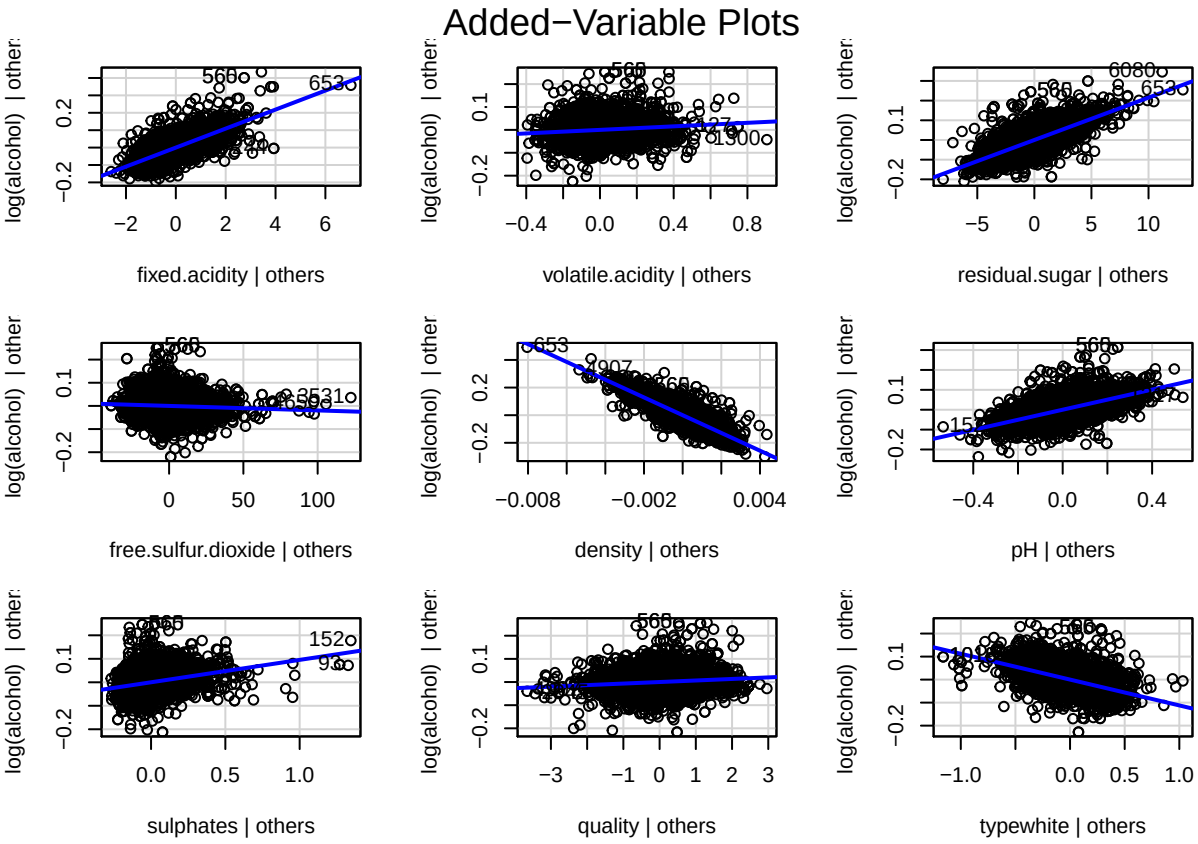

```
# Dropped total.sulfur.dioxide and chlorides from the model.
# Assessing leverage and outliers
influenceIndexPlot(Lmmodel, vars = c('hat', 'Cook','Studentized'), id = list(n = 5))
```



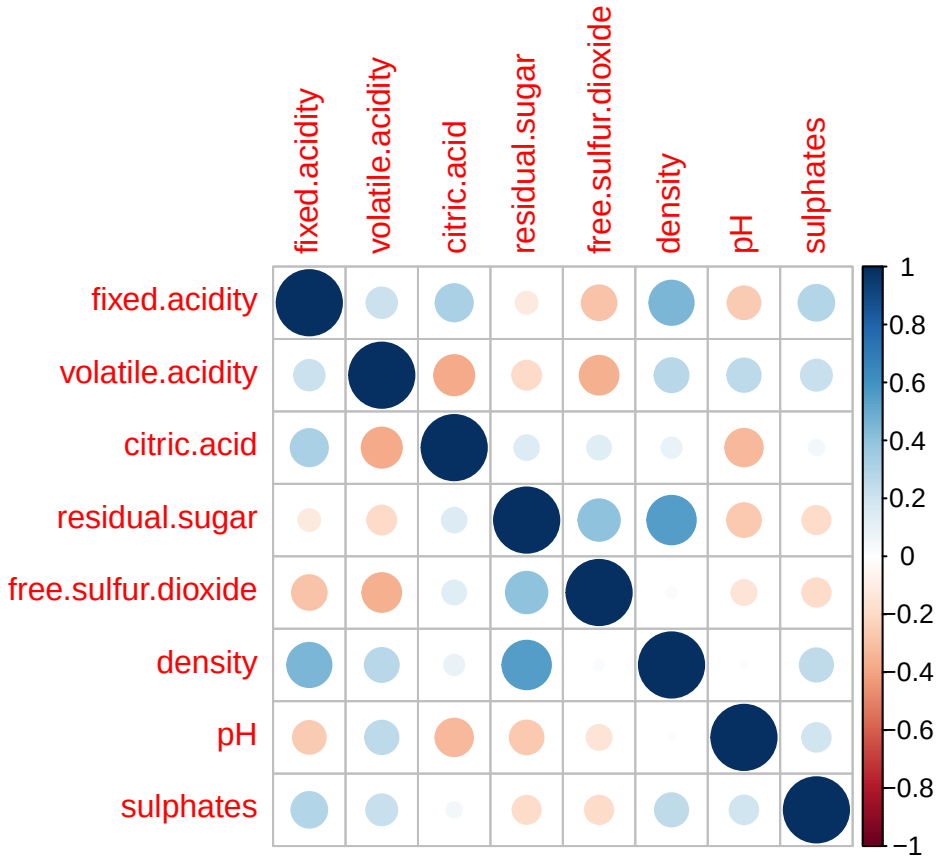

```
Lmmodel1 <- lm(alcohol ~ fixed.acidity +  volatile.acidity + residual.sugar + free.sulfur.dioxide + dens
#transforming the Response
boxCox(Lmmodel1)
```




```
# transforming the predictors
avPlots(Wine.best.lm)
```

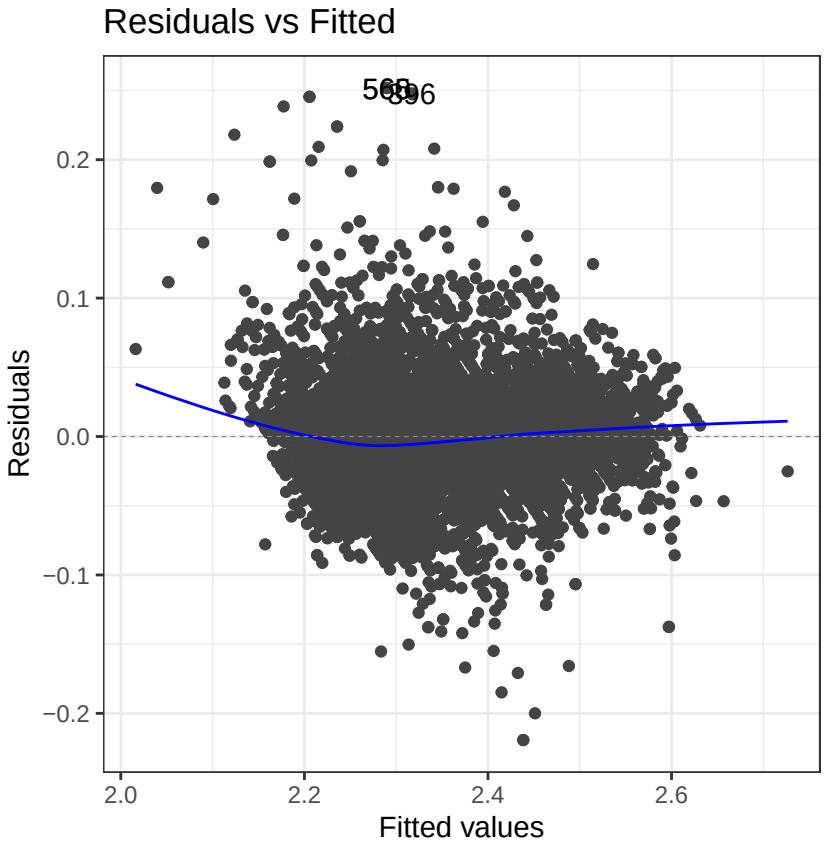



```
#checking mutli-collinearity
corrplot(cor(Wine[, c(1:4,6,8:10)]))
```

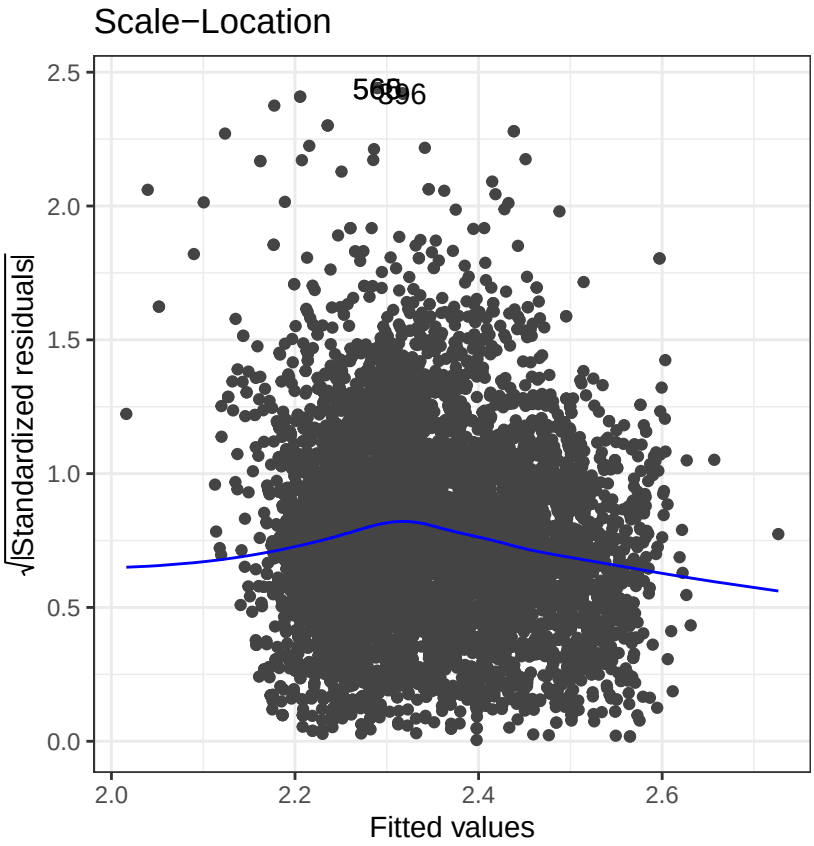



```
#checking model assumptions
#Linearity
autoplot(Wine.best.lm, which = 1, ncol = 1, nrow = 1) + theme_bw() +
        theme(aspect.ratio = 1)
```

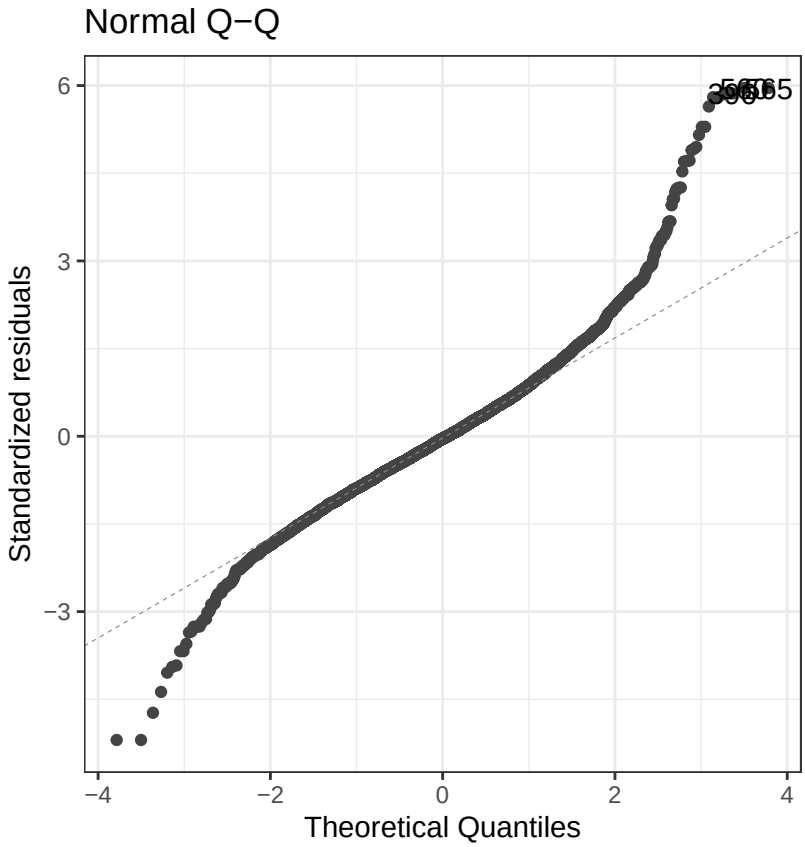
```
## Warning: `arrange_()` is deprecated as of dplyr 0.7.0.
## Please use `arrange()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```




```
#Constant Variance
autoplot(Wine.best.lm, which = 3, ncol = 1, nrow = 1) + theme_bw() +
        theme(aspect.ratio = 1) #looks pretty good
```




```
#normality
autoplot(Wine.best.lm, which = 2, ncol = 1, nrow = 1) + theme_bw() +
       theme(aspect.ratio = 1) #approximately normal
```


Normal Q−Q
Standardized residuals
Theoretical Quantiles
6
3
0
−3
−4
−2
0
2
4