Regression Analysis of Red Wine Quality

Erik Spor - Tuesday 3 P.M.

Michael Rhine - Wednesday 8 A.M.

1. **Introduction**

In this paper, we study the quality of red wine from the “Wine Quality Data Set” from the UC Irvine Machine Learning Repository. We examine all possible predictors for wine quality from the data set (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) and analyze which are statistically significant in determining the quality score. From this, we create a linear regression equation to model the relationship between the predictors and red wine quality. Of the seven predictors that best predict wine quality, we look at how they impact wine quality.

1. **Questions of Interest**
2. Which variables in the “Wine Quality Data Set” best predict red wine quality?
3. How do the predictors affect the quality?
4. What is the best linear regression model for red wine quality?
5. **Regression Method**

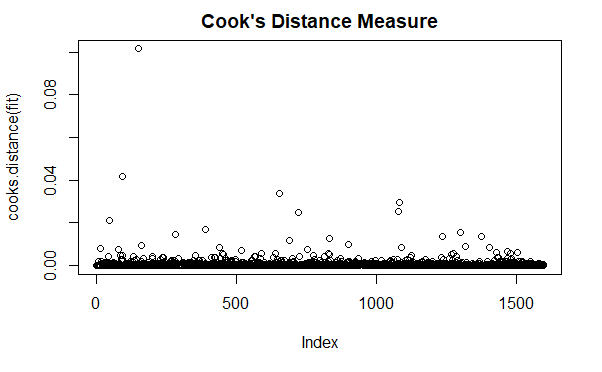
The predictors for red wine quality are found through the use of stepwise regression with Akaike’s Information Criterion to choose predictors, which can be done with the step() function in R. To supplement the results of step(), best subsets regression is used and the results compared.

To examine how the predictors affect quality, scatterplot and correlation matrices are used (R functions pair() and cor() respectively). These depict whether each predictor has a positive or negative correlation with red wine quality.

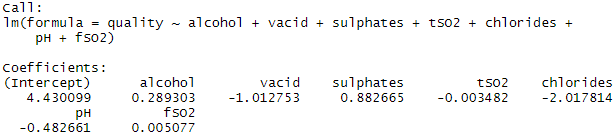
To obtain a more accurate final model, possible interaction terms and higher order terms are checked for and added if significant. We then perform residual analysis to ensure LINE requirements. We examine the Res. vs Fit model to ensure linearity and non-constant variance, and create a qqnorm plot to check for the normal distribution of the residuals.

1. **Regression Analysis, Results and Interpretation**

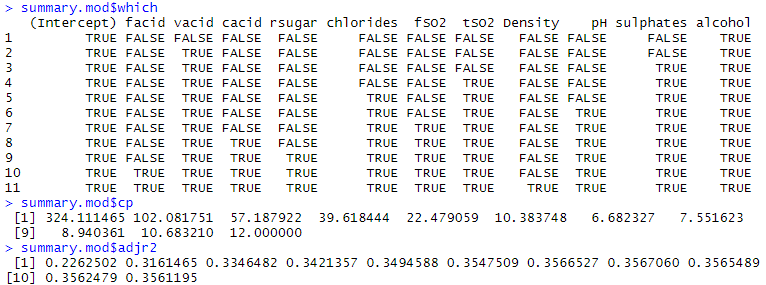
Question 1

We first import our variables and check for any influential points using Cook’s Distance Measure. This graph indicates there are no values greater than 0.5, but we also run a which() function to ensure there are no values greater than 0.5. We can conclude from this that there are no influential data points.

Stepwise regression is then conducted to find the best predictors of red wine quality:



The result is compared to that of best subsets regression with criteria of Mallow’s Cp and adjusted R-squared values.



Through examination of the computation of Mallow's Cp and adjusted R-squared values, we can conclude that the best of the 7-predictors models best predicts wine quality. The highest adjusted R-squared value occurs in the eighth position while the eighth Mallow’s Cp value is close to p, the number of parameters. While it may not be closest to its number of parameters - that award would go to the model containing 11 parameters - and thus having the least bias, we still decide on using the 7-predictor model as our other tests conclude it as best.

From this, an initial linear regression equation is built:

x1 = alcohol

x2 = volatile acid

x3 = sulphates

x4 = pH

x5 = chlorides

x6 = total sulfur dioxide

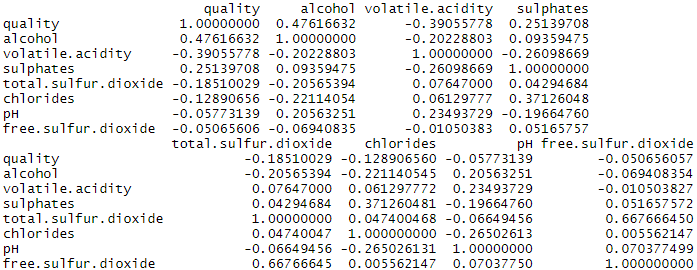
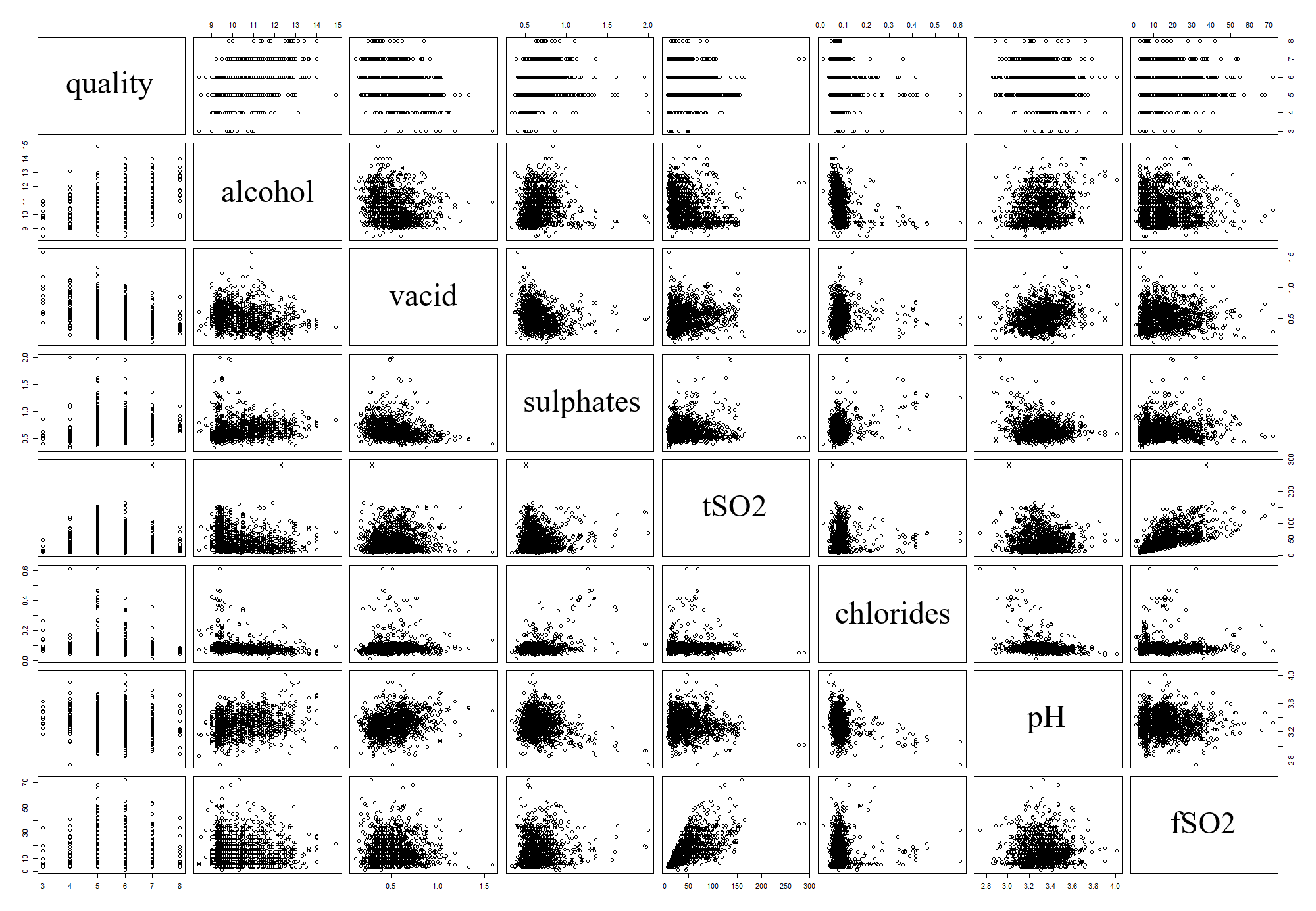
x7 = free sulfur dioxide

y = quality

y =

Question 2

Scatterplot and correlation matrices are made to display the relationships between predictors and the quality of the red wine.



From these, it is evident that alcohol level, followed by sulphates, has the strongest positive correlation with red wine quality. Volatile acidity has the strongest negative correlation with quality, followed by total sulfur dioxide and level of chlorides. pH and free sulfur dioxide both have a relatively weak negative correlation with quality.

Based on these correlations, we can predict an ideal quality score by respectively maximizing and minimizing the positive-trended and negative-trended predictor variables. We generate a predicted score of 9.383, with a 95% confidence interval of (8.948, 9.817). This score would be outstanding; the highest quality score in our dataset is only 8.

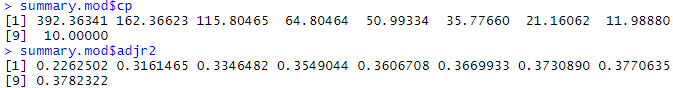
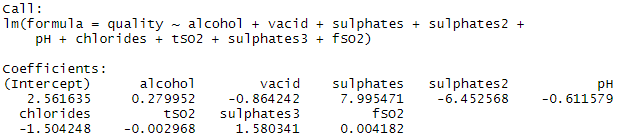
It should be noted however, that this model is possibly underspecified. We must note the lack of account of possible interactions and higher order terms, which could yield biased predicted values. Because of this, it is necessary to determine a correctly specified regression model, so that we may correctly estimate our population slopes and yield unbiased predictions.

Question 3

From the scatterplot matrix and correlation matrix there is one multicollinearity problem: between total sulfur dioxide and free sulfur dioxide (which is to be expected as the free sulfur dioxide is a subset of the total sulfur dioxide). We ran a general linear F-test to see if the inclusion of their interaction was necessary. Our null hypothesis was our initial, reduced model (without the interaction between free SO2 and total SO2) and our alternative hypothesis was our full model (with the interaction) with α = 0.05. The resultant F-statistic is 0.655, along with a p-value of 0.4184, which is greater than 0.05. We fail to reject our null hypothesis and do not add the interaction to our model.

We then tested for significance of higher order terms and checked to see if they would improve the model. We used our initial, reduced model (without higher order terms) as the null hypothesis and the full model (with all higher orders) as the alternative hypothesis with α = 0.05. From this, we found the second and third order terms for sulphates (p < 2e-16 and p = 1.68e-08 respectively) to be statistically significant.

Stepwise and best subsets regression are conducted once more.



These results show that the model including the significant higher order terms is better due to less bias and a higher adjusted R-squared than the initial model.

From the stepwise regression, an updated linear regression equation is built.

x1 = alcohol

x2 = volatile acid

x3 = sulphates

x4 = pH

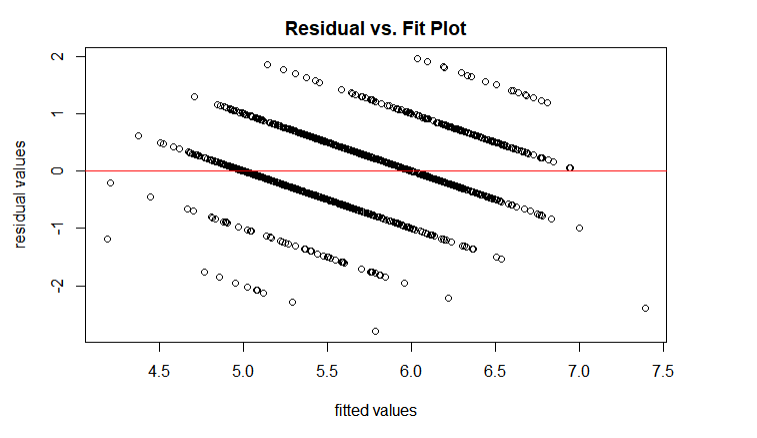
x5 = chlorides

x6 = total sulfur dioxide

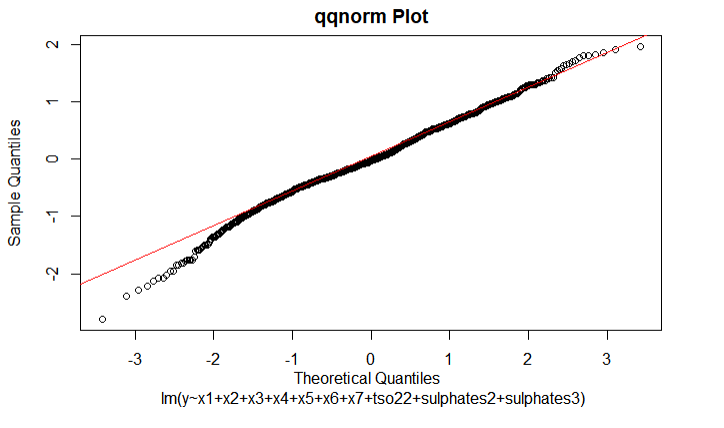
x7 = free sulfur dioxide

y = quality

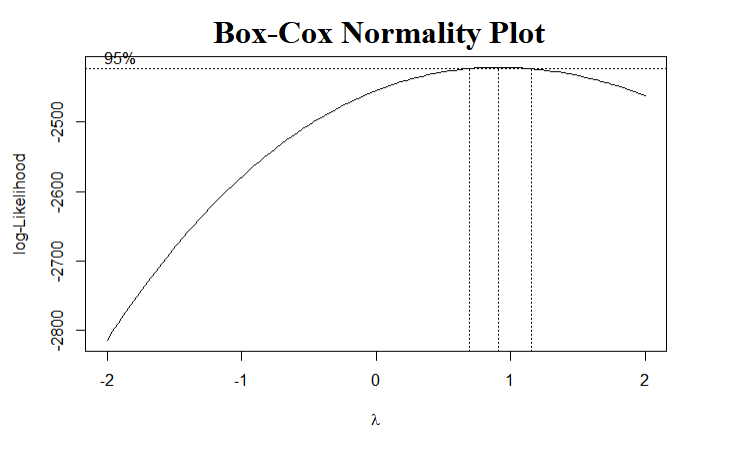
We are then able to perform residual analysis to ensure LINE requirements. If any requirement is not met, we must perform a transformation of the data. A Residual vs. Fit did not suggest non-linearity and the variance may be considered constant. There is a noticeable outlier, but our prior use of Cook’s Distance Measure has already determined that no points are influential.



A qqnorm plot did indicate normality, but there did appear to be a left-skew with the data.



To ensure normality, a box cox transformation was computed and allowed us to conclude that no transformations on the response were necessary, as 1 fell inside the 95% confidence interval for lambda.



Our LINE conditions are met. We can conclude that our final, correctly specified model is:

Returning to the end of question 2, we can now predict a less biased, more accurate score with the respectively maximized and minimized positive-trended and negative-trended predictors. The new predicted result is 8.042 with a 95% confidence interval of (7.282, 8.802).

1. **Conclusion**

We are confident that alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH, and free sulfur dioxide are statistically significant predictors for our model and can all be used to predict an accurate quality score of red wine with minimal bias. Additionally, the second and third order terms for sulphates are statistically significant in creating a robust, more accurate model.

Of these predictors, the volatile acidity, total sulfur dioxide, and chloride levels have the strongest negative relationship with red wine quality, while the alcohol and sulphate levels have the strongest positive relationship. Meanwhile, the levels of pH and free sulfur dioxide have little impact on the quality. Due to the low impact of these two predictors, it is possible the model could be improved by removing them, but they remain due to our failure to reject them in our analysis. As the data is only related to the Portuguese "Vinho Verde" wine, these results do not pertain to all red wine, but they should not be entirely discounted either.

1. **Appendix**

**Question 1**

# Importing and renaming variables

rwine <- read.csv("winequality-red.csv")

quality = rwine$quality

facid = rwine$fixed.acidity

vacid = rwine$volatile.acidity

cacid = rwine$citric.acid

rsugar = rwine$residual.sugar

chlorides = rwine$chlorides

fSO2 = rwine$free.sulfur.dioxide

tSO2 = rwine$total.sulfur.dioxide

Density = rwine$density

pH = rwine$pH

sulphates = rwine$sulphates

alcohol = rwine$alcohol

#Find (& remove) influential data points via Cook's Distance Measure

fit=lm(quality ~ facid + vacid + cacid + rsugar + chlorides +

fSO2 + tSO2 + Density + pH + sulphates + alcohol)

plot(cooks.distance(fit), main = "Cook's Distance Measure")

which(cooks.distance(fit)>.5)

# Stepwise regression of winequality-red using AIC

mod0 = lm(quality ~ 1)

modup = lm(quality ~ facid + vacid + cacid + rsugar + chlorides +

fSO2 + tSO2 + Density + pH + sulphates + alcohol)

step(mod0, scope = list(lower = mod0, upper = modup))

# Compare subsets with Mallow's Cp and adjusted R-squared

mod = regsubsets(cbind(facid, vacid, cacid, rsugar, chlorides, fSO2,

tSO2, Density, pH, sulphates, alcohol), quality, nvmax = 11)

summary.mod <- summary(mod)

summary.mod$which

summary.mod$cp

summary.mod$adjr2

#initial fitted model

x1=alcohol

x2=vacid

x3=sulphates

x4=tSO2

x5=chlorides

x6=pH

x7=fSO2

y=quality

fit=lm(y~x1+x2+x3+x4+x5+x6+x7)

**Question 2**

# Scatterplot matrix

pairs(~ quality + alcohol + vacid + sulphates + tSO2 + chlorides + pH + fSO2)

# Correlation matrix

variables = rwine[,c(12,11,2,10,7,5,9,6),drop=FALSE]

cor(variables)

#Predicted quality score and confidence interval with min/maxed predictors

fit=lm(y~x1+x2+x3+x4+x5+x6+x7)

data.max=data.frame(x1=max(x1),x2=min(x2),x3=max(x3), x4=min(x4), x5=min(x5), x6=min(x6),x7=max(x7))

predict(fit, data.max, interval='predict', level=0.95)

predict(fit, data.max, interval='confidence', level=0.95)

**Question 3**

# General Linear F-Test for I(fSO2\*tSO2)

mod.full=lm(y~x1+x2+x3+x4+x5+x6+x7+I(x4\*x7))

mod.red=lm(y~x1+x2+x3+x4+x5+x6+x7)

anova(mod.red,mod.full)

# Test second order

# alcohol

alcohol2 = alcohol\*\*2

summary(lm(quality ~ alcohol + alcohol2))

# vacid

vacid2 = vacid\*\*2

summary(lm(quality ~ vacid + vacid2))

# sulphates

sulphates2 = sulphates\*\*2

summary(lm(quality ~ sulphates + sulphates2))

# tSO2

tSO22 = tSO2\*\*2

summary(lm(quality ~ tSO2 + tSO22))

# chlorides

chlorides2 = chlorides\*\*2

summary(lm(quality ~ chlorides + chlorides2))

#pH

pH2 = pH\*\*2

summary(lm(quality ~ pH + pH2))

#fSO2

fSO22 = fSO2\*\*2

summary(lm(quality ~ fSO2 + fSO22))

# Put significant second order value into step function

mod0 <- lm(quality ~ 1)

modup <- lm(quality ~ vacid + chlorides + fSO2 + tSO2 + pH +

sulphates + alcohol + chlorides2 + sulphates2)

step(mod0, scope = list(lower = mod0, upper = modup))

# Test third order

sulphates3 = sulphates\*\*3

summary(lm(quality ~ sulphates + sulphates2 + sulphates3))

# Put significant third order value into step function

mod0 <- lm(quality ~ 1)

modup <- lm(quality ~ vacid + chlorides + sulphates3 +

fSO2 + tSO2 + pH + sulphates + alcohol + sulphates2)

step(mod0, scope = list(lower = mod0, upper = modup))

# Test fourth order

sulphates4 = sulphates\*\*4

summary(lm(quality ~ sulphates + sulphates2 + sulphates3 + sulphates4))

# Compare with Mallow's Cp and adjusted R-squared

mod = regsubsets(cbind(vacid, chlorides, fSO2, tSO2, pH, sulphates,

alcohol, sulphates2, sulphates3), quality, nvmax = 11)

summary.mod <- summary(mod)

summary.mod$which

summary.mod$cp

summary.mod$adjr2

#Res. Vs. Fit Plot

fin.mod=lm(y~x1+x2+x3+x4+x5+x6+x7+tso22+sulphates2+sulphates3)

res=resid(fin.mod)

fitt=fitted(fin.mod)

plot(x=fitt,y=res,xlab="fitted values", ylab="residual values",main='Residual vs. Fit Plot')+abline(h=0, col='red')

#qqnorm Plot

qqnorm(res, main='qqnorm Plot',xlab="Theoretical Quantiles \n lm(y~x1+x2+x3+x4+x5+x6+x7+tso22+sulphates2+sulphates3)")

qqline(res, col='red')

#Box-Cox Normality Plot

library(MASS)

boxcox(mod.fin)

#Predicted Quality Score & Confidence Interval with Final Model

fit.fin=lm(y~alcohol+vacid+sulphates+sulphates2+sulphates3+tSO2+chlorides+pH+fSO2)

data.max.int=data.frame(alcohol=max(alcohol), vacid=min(vacid), sulphates=max(sulphates),sulphates2=max(sulphates2),sulphates3=max(sulphates3),tSO2=min(tSO2), chlorides=min(chlorides), pH=min(pH), fSO2=max(fSO2))

predict(fit.fin,data.max.int,interval='predict',level=0.95)

predict(fit.fin,data.max.int,interval='confidence',level=0.95)