**STAC32 PROJECT**

**Shu Tu 999339096**

**Zhengshuo Xu 998300351**

**Introduction**

A number of different wines of the Pinot Noir type were rated by experts. A number of different qualities were assessed, "Clarity", "Aroma", "Body", "Flavor" and "Oakiness", as well as an overall “Quality". In addition, the region from which the wine came was recorded (labeled 1, 2 or 3).

The data can be found in [http://www.utsc.utoronto.ca/%7ebutler/c32/wine.txt.](http://www.utsc.utoronto.ca/~butler/c32/wine.txt.)

We choose to use this data set because we are personally curious about what really affects the quality of the wine.

In this project, the main goal is to study the relationships between the variable "Quality” and the other six variables "Clarity", "Aroma", "Body", "Flavor", "Oakiness" and "region". A variety of tests and methods will be performed to analyze the data. Our assumptions are "Clarity", "Aroma", "Oakiness" ,"Body" and "Flavor" are all positively correlated to "Quality" while "Region" is not correlated to "Quality".

**Literature review**

1. “Clarity is an essential quality required by consumers” (Neggers, 2006)
2. Marais(1983) demonstrates that flavor and aroma have the most significant effect on wine quality.
3. Pucktte(2012) mentions that “body is a snapshot of the overall impression of a wine”.

**Methods**

First, we check the normality of each qualities and the overall "Quality" except "Region".

> par(mfrow=c(2,3))

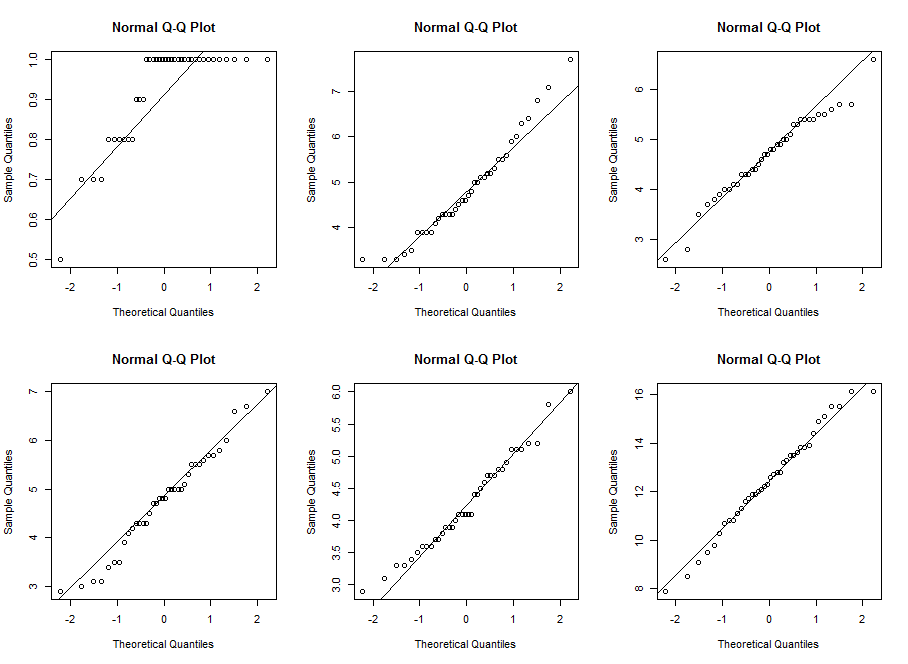
> for (i in 1:6)

+ {

+ qqnorm(wine[,i])

+ qqline(wine[,i])

+ }



From the plots we see that the overall "Quality" are the closest to normal.

> wine=read.table("wine.txt",header=T)

> head(wine)

Clarity Aroma Body Flavor Oakiness Quality Region

1 1 3.3 2.8 3.1 4.1 9.8 1

2 1 4.4 4.9 3.5 3.9 12.6 1

3 1 3.9 5.3 4.8 4.7 11.9 1

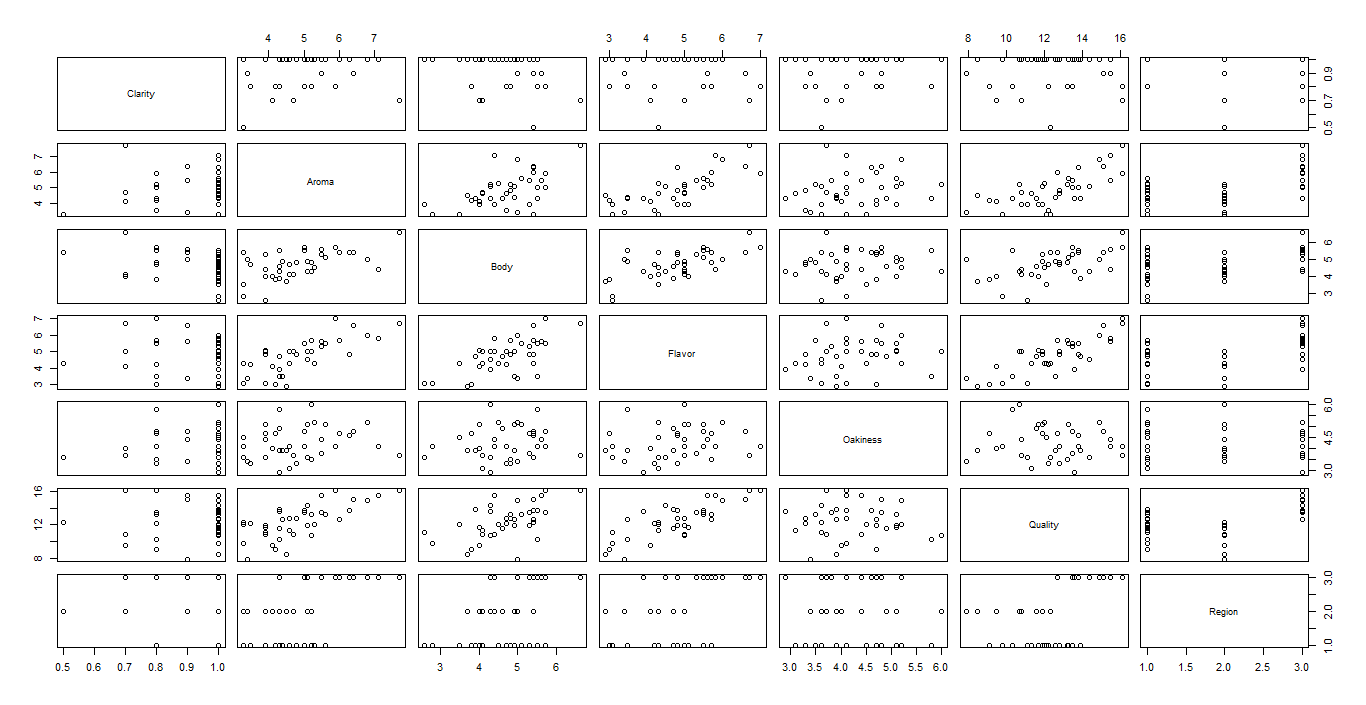
4 1 3.9 2.6 3.1 3.6 11.1 1

5 1 5.6 5.1 5.5 5.1 13.3 1

6 1 4.6 4.7 5.0 4.1 12.8 1

> attach(wine)

> pairs(wine)



From the pairs plot, we can briefly see that the overall "Quality " and "Aroma" are positively related(Quality up as Aroma increase); "Flavor" is also positively related to "Quality "; And there is a weak positive relationship between " Quality" and "Body". "Aroma", "Body" and "Flavor" are positively related to each other.

We want to study the relations among these variables with respect to "Region" respectively.

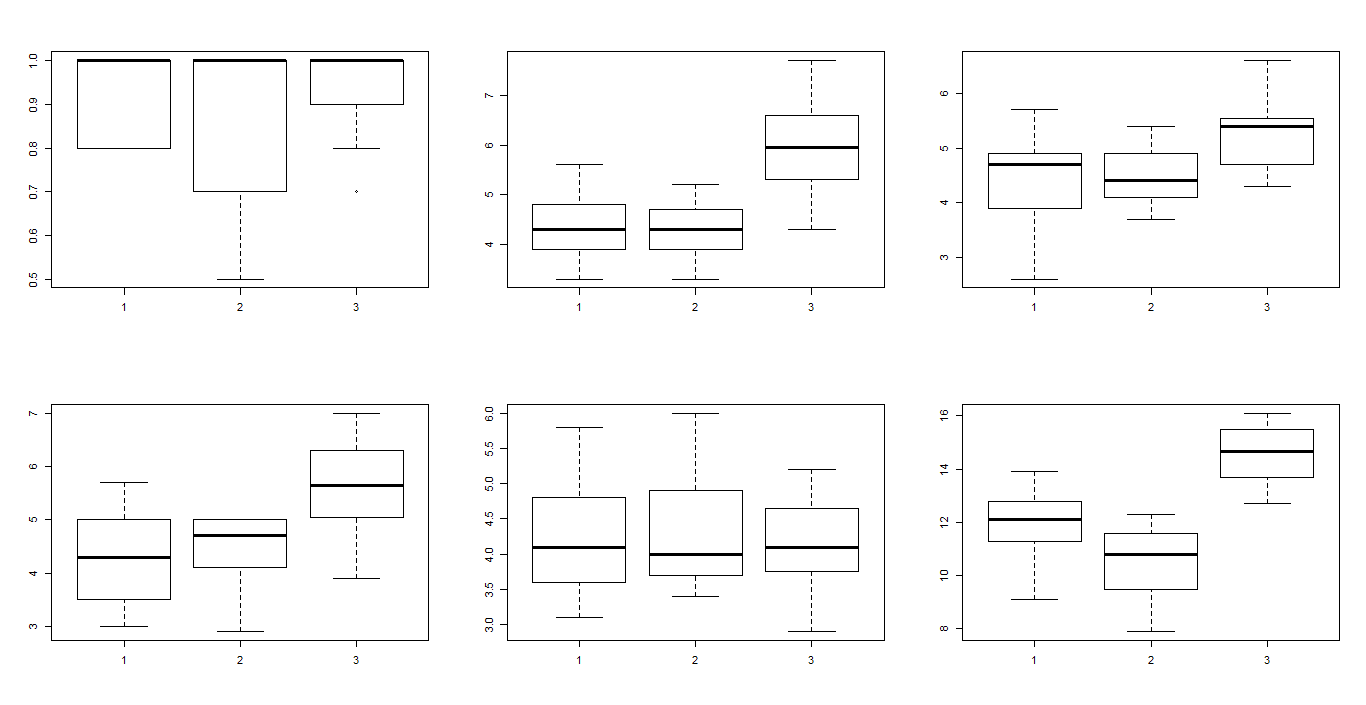
> par(mfrow=c(2,3))

> for (i in 1:6)

+ {

+ boxplot(wine[,i]~Region)

+ }



From these boxplots with respect to "Region", we can see that:

1) For "Clarity":

the medians of Clarity are the same for three regions(equal to 1), and the spread of region1 is almost the same as that of region3; the spread of region2 is way larger than the other two regions. And we see that there is an outlier in the region3, whose clarity is 0.7.

> index\_3=c(1:n)[Region==3]

> min(Clarity[index\_3])

[1] 0.7

2) For "Aroma":

the medians of Aroma for region1 and region2 are the same(at around 4.2); region3 has the largest median(at 6), which is much bigger than that of the first two regions, and the largest spread as well.

3) For "Body":

the medians of Body among 3 regions do not differ much, but the spread of region1 is the largest.

4) For "Flavor":

region3 has the biggest median of Flavor and its spread is pretty much the same as region1's.

5) For "Oakiness":

three regions have pretty much the same median and spread for Oakiness.

6) For "Quality":

region3 has the largest median and the smallest spread for Quality while region2 has the smallest and the largest spread.

> aggregate(Quality~Region,wine,mean)

Region Quality

1 1 11.97647

2 2 10.44444

3 3 14.58333

Using ANOVA to test the means among these three regions.

> region=factor(Region)

> wine.aov=aov=aov(Quality~region,data=wine)

> summary(wine.aov)

Df Sum Sq Mean Sq F value Pr(>F)

region 2 94.62 47.31 27.52 6.59e-08 \*\*\*

Residuals 35 60.17 1.72

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

The p-value is very small, and we conclude that not all 3 means are the same, at least one is different from others. So we use Tukey's test to compare three region's mean differences in one test.

> TukeyHSD(wine.aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Quality ~ region, data = wine)

$region

diff lwr upr p adj

2-1 -1.532026 -2.854777 -0.2092757 0.0201945

3-1 2.606863 1.397042 3.8166831 0.0000206

3-2 4.138889 2.723960 5.5538182 0.0000001

The result has shown that "Quality" for region3 is significantly higher than those for the other two regions.

Overall, region3 is the best region over almost all the qualities of wine and region2 is relatively worse than the other two regions.

We then start to study the correlation among the qualities:

> cor(wine)

Clarity Aroma Body Flavor Oakiness Quality Region

Clarity 1.00000000 0.0619021 -0.3083783 -0.08515993 0.18321471 0.02844131 -0.02032013

Aroma 0.06190210 1.0000000 0.5489102 0.73656121 0.20164445 0.70732432 0.62010056

Body -0.30837826 0.5489102 1.0000000 0.64665917 0.15210591 0.54870219 0.43544030

Flavor -0.08515993 0.7365612 0.6466592 1.00000000 0.17976051 0.79004713 0.50591814

Oakiness 0.18321471 0.2016444 0.1521059 0.17976051 1.00000000 -0.-0.05956862

Quality 0.02844131 0.7073243 0.5487022 0.79004713 -0.04704047 1.00000000 0.50704891

Region -0.02032013 0.6201006 0.4354403 0.50591814 -0.05956862 0.50704891 1.00000000

From the correlation function in R, we can see that "Aroma" and "Flavor" have really strong positive relationship with "Quality".

Next, we use two sample t-test to check the relationship between the qualities of wine against the overall "Quality". First, we test if "Oakiness" affects the overall "Quality". we split "Oakiness" into 2 groups: "hioak" group with those Oakiness greater than the mean Oakiness, and "lowoak" group with those Oakiness less than the mean Oakiness, then we calculate the means of "Quality" corresponding to these two groups, and see that they are almost the same. Lastly, we run the t-test, and it also gives me the result that there is no difference in means of "Quality" between "hioak" and "lowoak"(since p-value=0.9326 >0.05, we accept the null hypothesis), hence we conclude that "Oakiness" does not associate with "Quality".

> mean(Oakiness)

[1] 4.255263

> table(Oakiness>=mean(Oakiness))

FALSE TRUE

21 17

> hioak=c(1:n)[Oakiness>=mean(Oakiness)]

> lowoak=c(1:n)[Oakiness<mean(Oakiness)]

> oakg=cut(Oakiness,c(0,mean(Oakiness),n),labels=c("lowoak","hiwoak"))

> oakiness=factor(Oakiness)

> levels(oakg)

[1] "lowoak" "hiwoak"

> oakg

[1] lowoak lowoak hiwoak lowoak hiwoak lowoak lowoak hiwoak lowoak lowoak lowoak lowoak lowoak lowoak

[15] lowoak hiwoak hiwoak lowoak lowoak lowoak hiwoak lowoak lowoak lowoak hiwoak hiwoak hiwoak hiwoak

[29] hiwoak lowoak hiwoak hiwoak hiwoak hiwoak hiwoak lowoak hiwoak lowoak

Levels: lowoak hiwoak

> aggregate(Quality~oakg,wine,mean)

oakg Quality

1 lowoak 12.46190

2 hiwoak 12.40588

> t.test(Quality~oakg,data=wine)

Welch Two Sample t-test

data: Quality by oakg

t = 0.0851, df = 35.934, p-value = 0.9326

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.278688 1.390732

sample estimates:

mean in group lowoak mean in group hiwoak

12.46190 12.40588

We also tested the rest of the qualities of wine using the same method:

> hiclar=c(1:n)[Clarity>=mean(Clarity)]

> lowclar=c(1:n)[Clarity<mean(Clarity)]

> clarg=cut(Clarity,c(0,mean(Clarity),n),labels=c("lowclar","hiclar"))

> Clarity=factor(Clarity)

> t.test(Quality~clarg,data=wine)

Welch Two Sample t-test

data: Quality by clarg

t = -0.0111, df = 16.298, p-value = 0.9913

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.777269 1.758808

sample estimates:

mean in group lowclar mean in group hiclar

12.43077 12.44000

> table(Aroma>=mean(Aroma))

FALSE TRUE

21 17

> hiaro=c(1:n)[Aroma>=mean(Aroma)]

> lowaro=c(1:n)[Aroma<mean(Aroma)]

> arog=cut(Aroma,c(0,mean(Aroma),n),labels=c("lowaro","hiaro"))

> Aroma=factor(Aroma)

> t.test(Quality~arog,data=wine)

Welch Two Sample t-test

data: Quality by arog

t = -5.0842, df = 35.026, p-value = 1.242e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-3.660104 -1.571268

sample estimates:

mean in group lowaro mean in group hiaro

11.26667 13.88235

> table(Body>=mean(Body))

FALSE TRUE

17 21

> hibody=c(1:n)[Body>=mean(Body)]

> lowbody=c(1:n)[Body<mean(Body)]

> bodyg=cut(Body,c(0,mean(Body),n),labels=c("lowbody","hibody"))

> Body=factor(Body)

> t.test(Quality~bodyg,data=wine)

Welch Two Sample t-test

data: Quality by bodyg

t = -2.57, df = 34.248, p-value = 0.01469

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.8638688 -0.3350108

sample estimates:

mean in group lowbody mean in group hibody

11.55294 13.15238

> table(Flavor>=mean(Flavor))

FALSE TRUE

17 21

> hifla=c(1:n)[Flavor>=mean(Flavor)]

> lowfla=c(1:n)[Flavor<mean(Flavor)]

> flag=cut(Flavor,c(0,mean(Flavor),n),labels=c("lowfla","hifla"))

> Flavor=factor(Flavor)

> t.test(Quality~flag,data=wine)

Welch Two Sample t-test

data: Quality by flag

t = -3.4647, df = 32.704, p-value = 0.001504

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-3.248654 -0.844343

sample estimates:

mean in group lowfla mean in group hifla

11.30588 13.35238

> table(Region>=mean(Region))

FALSE TRUE

17 21

> hireg=c(1:n)[Region>=mean(Region)]

> lowreg=c(1:n)[Region<mean(Region)]

> regg=cut(Region,c(0,mean(Region),n),labels=c("lowreg","hireg"))

> Region=factor(Region)

> t.test(Quality~regg,data=wine)

Welch Two Sample t-test

data: Quality by regg

t = -1.3347, df = 31.897, p-value = 0.1914

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.1045737 0.4384673

sample estimates:

mean in group lowreg mean in group hireg

11.97647 12.80952

From the results we see that "Aroma", "Body", "Flavor", all have p-values smaller than 0.05, i.e., they are all more or less related to the overall "Quality".

In order to further study the association between the overall "Quality" with the other qualities of wine , we build a regression function with "Quality" as the response variable and the other qualities of wine as the predictor variables.

> wine.lm=lm(Quality~Clarity+Aroma+Body+Flavor+Oakiness+Region,data=wine)

> summary(wine.lm)

Call:

lm(formula = Quality ~ Clarity + Aroma + Body + Flavor + Oakiness +

Region, data = wine)

Residuals:

Min 1Q Median 3Q Max

-2.83614 -0.57561 -0.06547 0.66181 1.70485

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.98433 2.26965 1.755 0.089056 .

Clarity 2.34751 1.76362 1.331 0.192872

Aroma 0.49731 0.30536 1.629 0.113516

Body 0.27841 0.34091 0.817 0.420357

Flavor 1.16987 0.30958 3.779 0.000673 \*\*\*

Oakiness -0.69229 0.28483 -2.431 0.021058 \*

Region -0.03381 0.29568 -0.114 0.909694

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 1.181 on 31 degrees of freedom

Multiple R-squared: 0.7207, Adjusted R-squared: 0.6667

F-statistic: 13.33 on 6 and 31 DF, p-value: 2.037e-07

From the summary of the initial regression function, we see that the p-value of "Flavor" equals to 0.000673, which is very small, hence we conclude that "Flavor" is really significant and has a strong explanatory power of "Quality" (i.e., the level of "Flavor" has strong effect on the "Quality" rating). "Oakiness" also has a p-value less than 0.05, which indicates the significance too. Then we use the "backward selection" to remove all the non-significant predictor variables.

> step(wine.lm, direction="backward")

Start: AIC=18.9

Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region

Df Sum of Sq RSS AIC

- Region 1 0.0182 43.248 16.916

- Body 1 0.9300 44.160 17.709

<none> 43.230 18.900

- Clarity 1 2.4707 45.701 19.012

- Aroma 1 3.6988 46.929 20.020

- Oakiness 1 8.2380 51.468 23.528

- Flavor 1 19.9132 63.143 31.297

Step: AIC=16.92

Quality ~ Clarity + Aroma + Body + Flavor + Oakiness

Df Sum of Sq RSS AIC

- Body 1 0.9118 44.160 15.709

<none> 43.248 16.916

- Clarity 1 2.4577 45.706 17.016

- Aroma 1 4.2397 47.488 18.470

- Oakiness 1 8.5978 51.846 21.806

- Flavor 1 19.8986 63.147 29.299

Step: AIC=15.71

Quality ~ Clarity + Aroma + Flavor + Oakiness

Df Sum of Sq RSS AIC

- Clarity 1 1.6936 45.853 15.139

<none> 44.160 15.709

- Aroma 1 5.3545 49.514 18.058

- Oakiness 1 8.0807 52.241 20.094

- Flavor 1 27.3280 71.488 32.014

Step: AIC=15.14

Quality ~ Aroma + Flavor + Oakiness

Df Sum of Sq RSS AIC

<none> 45.853 15.139

- Aroma 1 6.6026 52.456 18.251

- Oakiness 1 6.9989 52.852 18.537

- Flavor 1 25.6888 71.542 30.043

Call:

lm(formula = Quality ~ Aroma + Flavor + Oakiness, data = wine)

Coefficients:

(Intercept) Aroma Flavor Oakiness

6.4672 0.5801 1.1997 -0.6023

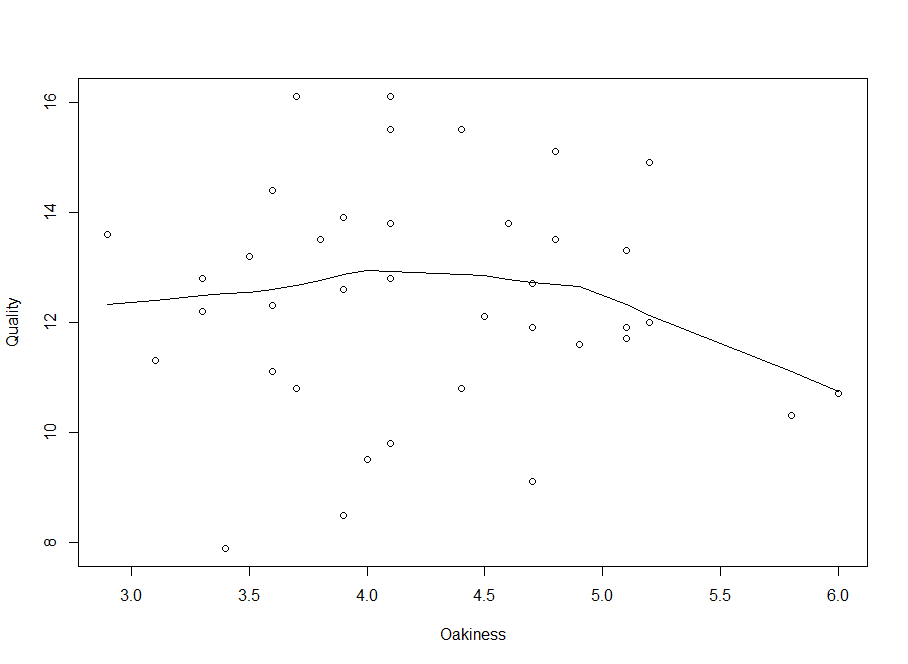
The "backward selection" gave me wine.lm2:

Quality=β0+β1Aroma+β2Flavor+β3Oakiness+ε

shows that only "Aroma", "Flavor" and "Oakiness" are important in identifying the overall quality of the wine; From the coefficients we can tell that "Aroma" and "Flavor" are positively related to "Quality" while "Oakiness" is negatively related to "Quality". It is odd that we cannot find any relationship between "Oakiness" and "Quality" in the pairs plot , the correlation function(-0.047 close to zero) and the two sample t-test has also shown that there is no relationship between "Oakiness" and "Quality". We can also see it directly from the plot of "Oakiness" against "Quality".

> plot(Quality~Oakiness)

> lines(lowess(Quality~Oakiness))



The observations randomly scattered.

So, we plot the residuals and predicted values of this regression model to check its appropriateness.

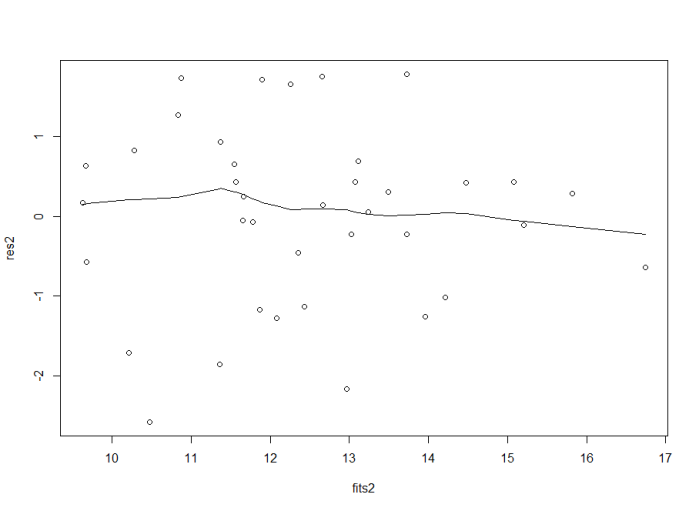
> wine.lm2=lm(Quality~Aroma+Flavor+Oakiness)

> fits2=fitted(wine.lm2)

> res2=resid(wine.lm2)

> plot(fits2,res2)

> lines(lowess(fits2,res2))



Residual plot basically random, lowess curve basically flat. Good.

> summary(wine.lm2)

Call:

lm(formula = Quality ~ Aroma + Flavor + Oakiness)

Residuals:

Min 1Q Median 3Q Max

-2.5707 -0.6256 0.1521 0.6467 1.7741

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.4672 1.3328 4.852 2.67e-05 \*\*\*

Aroma 0.5801 0.2622 2.213 0.033740 \*

Flavor 1.1997 0.2749 4.364 0.000113 \*\*\*

Oakiness -0.6023 0.2644 -2.278 0.029127 \*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 1.161 on 34 degrees of freedom

Multiple R-squared: 0.7038, Adjusted R-squared: 0.6776

F-statistic: 26.92 on 3 and 34 DF, p-value: 4.203e-09

The R-squared has not dropped much (from 72% in wine.lm to 70.4% in wine.lm2), which suggests that taking out those variables has not done much damage.

we then test which model is better:

H0: two models equally good versus HA: wine.lm is better

> anova(wine.lm2,wine.lm)

Analysis of Variance Table

Model 1: Quality ~ Aroma + Flavor + Oakiness

Model 2: Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region

Res.Df RSS Df Sum of Sq F Pr(>F)

1 34 45.853

2 31 43.230 3 2.6236 0.6271 0.6029

The p-value= 0.6029, which is not small(greater than 0.05), hence we cannot reject the null.

If we delete "Oakiness" from the regression model:

> wine.lm3=lm(Quality~Aroma+Flavor)

> summary(wine.lm3)

Call:

lm(formula = Quality ~ Aroma + Flavor)

Residuals:

Min 1Q Median 3Q Max

-2.19048 -0.60300 -0.03203 0.66039 2.46287

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.3462 1.0091 4.307 0.000127 \*\*\*

Aroma 0.5180 0.2759 1.877 0.068849 .

Flavor 1.1702 0.2905 4.027 0.000288 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 1.229 on 35 degrees of freedom

Multiple R-squared: 0.6586, Adjusted R-squared: 0.639

F-statistic: 33.75 on 2 and 35 DF, p-value: 6.811e-09

> anova(wine.lm3,wine.lm2)

Analysis of Variance Table

Model 1: Quality ~ Aroma + Flavor

Model 2: Quality ~ Aroma + Flavor + Oakiness

Res.Df RSS Df Sum of Sq F Pr(>F)

1 35 52.852

2 34 45.853 1 6.9989 5.1896 0.02913 \*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

The R-squared has dropped dramatically(from 70.4% in wine.lm2 to 65.9% in wine.lm3), and the ANOVA test between two models also shows that wine.lm2 are better than wine.lm3(since its p-value for F test is 0.029, which is smaller than 0.05 and we reject the null hypothesis: two models equally good)

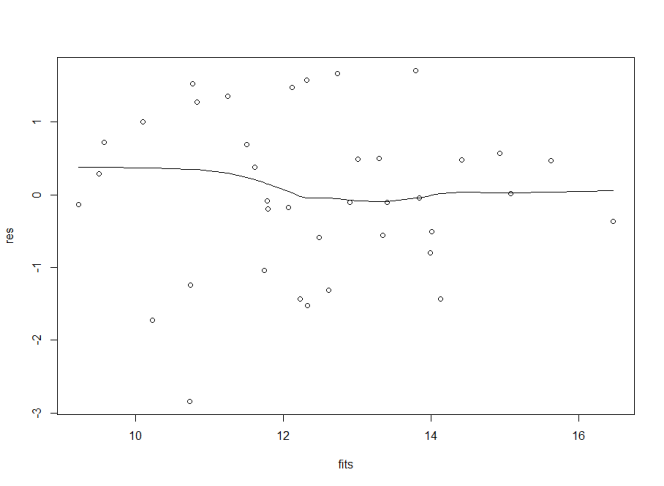
Then we start to check the residuals:

> fits=fitted(wine.lm)

> res=resid(wine.lm)

> plot(fits,res)

> lines(lowess(fits,res))

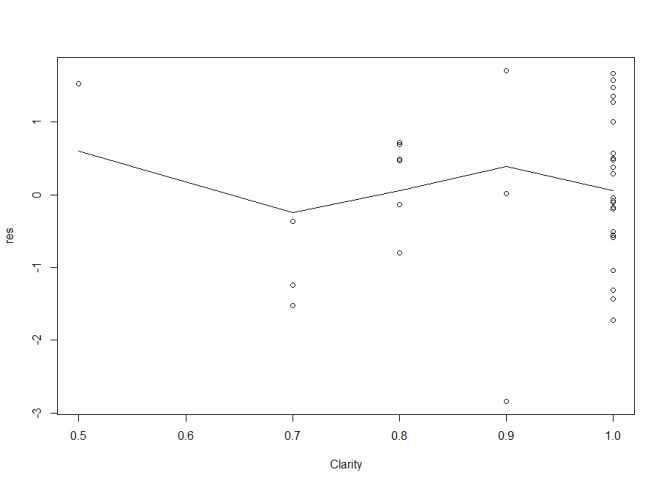


Residual against fitted value plot of the initial regression function seems random, lowess curve basically flat. Not bad.

But residuals against other predictor variables are curves everywhere.

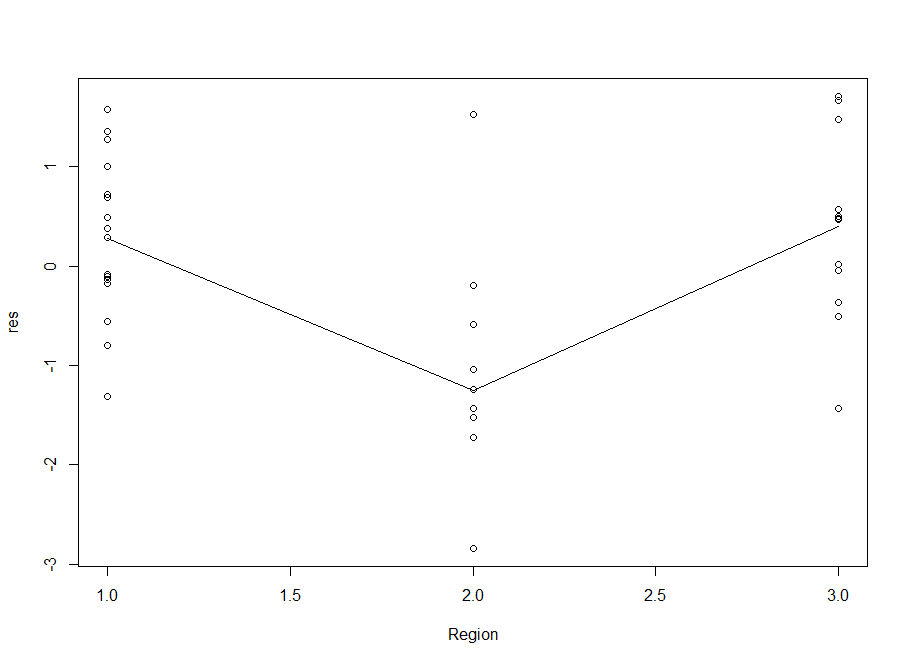
> plot(Clarity,res)

> lines(lowess(Clarity,res))



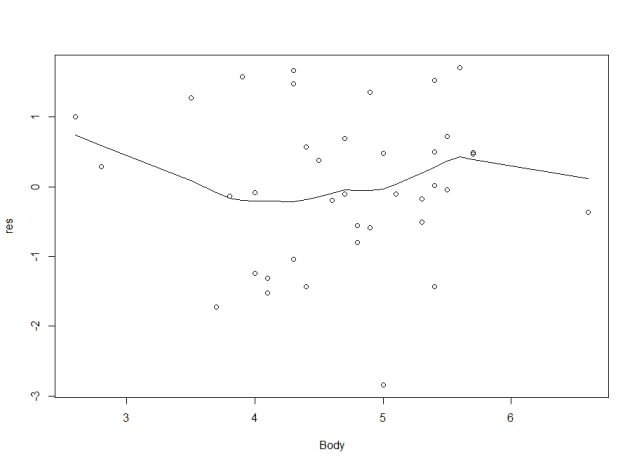
> plot(Region,res)

> lines(lowess(Region,res))



> plot(Body,res)

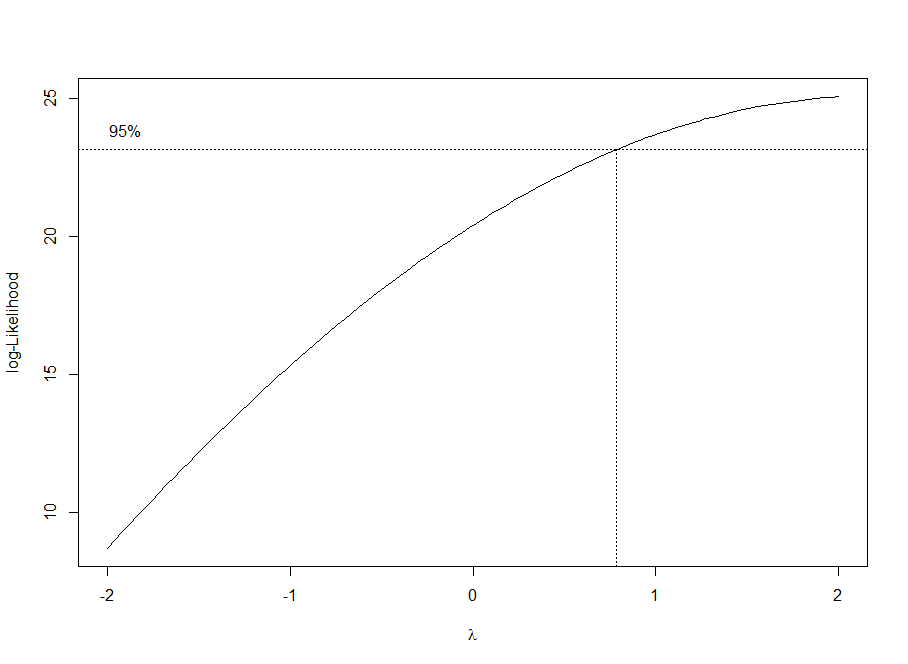
> lines(lowess(Body,res))



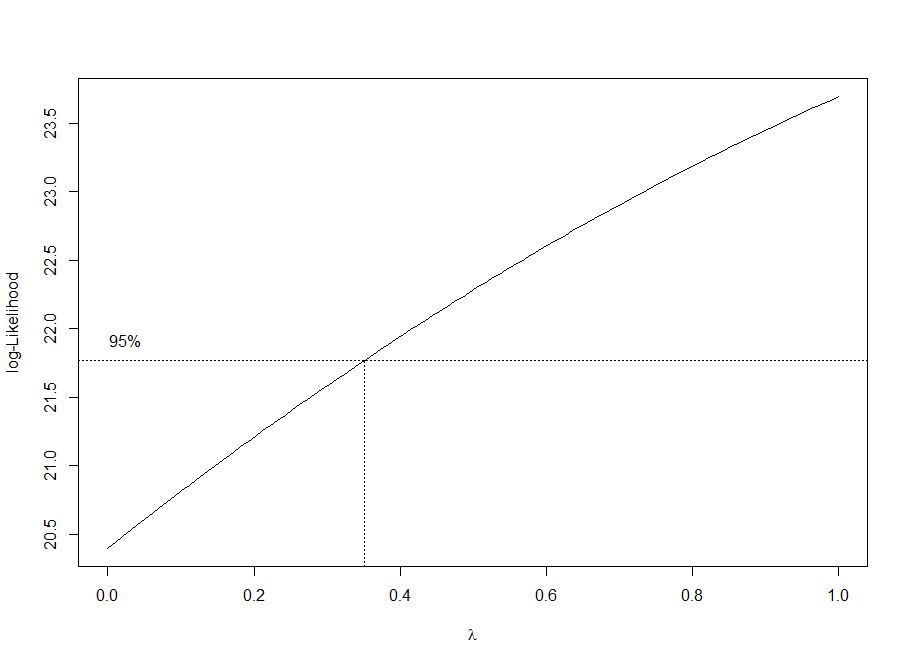
Then we use R's boxcox to produce a plot to find the λ:

> library(MASS)

> boxcox(wine.lm)



> boxcox(wine.lm,lambda=seq(0,1,0.1))



we have that λ=0.35, now we transform the response variable "Quality" into "Quality0.35" and get the new regression function with response variable equals to "Quality0.35", after that we use the "backward selection" again:

> quality=Quality^0.35

> wine.lm4=lm(quality~Clarity+Aroma+Body+Flavor+Oakiness+Region,data=wine)

> step(wine.lm4, direction="backward")

Start: AIC=-180.87

quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region

Df Sum of Sq RSS AIC

- Region 1 0.000727 0.22598 -182.75

- Body 1 0.004561 0.22982 -182.11

<none> 0.22526 -180.87

- Aroma 1 0.014525 0.23978 -180.49

- Clarity 1 0.014920 0.24018 -180.43

- Oakiness 1 0.038069 0.26333 -176.93

- Flavor 1 0.105657 0.33091 -168.25

Step: AIC=-182.75

quality ~ Clarity + Aroma + Body + Flavor + Oakiness

Df Sum of Sq RSS AIC

- Body 1 0.004159 0.23014 -184.05

<none> 0.22598 -182.75

- Aroma 1 0.014499 0.24048 -182.38

- Clarity 1 0.014682 0.24067 -182.35

- Oakiness 1 0.037890 0.26387 -178.85

- Flavor 1 0.105093 0.33108 -170.23

Step: AIC=-184.05

quality ~ Clarity + Aroma + Flavor + Oakiness

Df Sum of Sq RSS AIC

- Clarity 1 0.010908 0.24105 -184.29

<none> 0.23014 -184.05

- Aroma 1 0.018822 0.24896 -183.06

- Oakiness 1 0.035565 0.26571 -180.59

- Flavor 1 0.142785 0.37293 -167.71

Step: AIC=-184.29

quality ~ Aroma + Flavor + Oakiness

Df Sum of Sq RSS AIC

<none> 0.24105 -184.29

- Aroma 1 0.024650 0.26570 -182.59

- Oakiness 1 0.029602 0.27065 -181.89

- Flavor 1 0.132631 0.37368 -169.63

Call:

lm(formula = quality ~ Aroma + Flavor + Oakiness, data = wine)

Coefficients:

(Intercept) Aroma Flavor Oakiness

1.99256 0.03545 0.08620 -0.03917

The "backward selection" still leaves us with the same predictor variables as before, i.e., "Aroma", "Flavor" and "Oakiness"

We look up the new regression model:

Quality0.35=β0+β1Aroma+β2Flavor+β3Oakiness+ε

> wine.lm5=lm(quality~Aroma+Flavor+Oakiness,data=wine)

> summary(wine.lm5)

Call:

lm(formula = quality ~ Aroma + Flavor + Oakiness, data = wine)

Residuals:

Min 1Q Median 3Q Max

-0.211542 -0.052475 0.007556 0.048640 0.129878

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.99256 0.09663 20.620 < 2e-16 \*\*\*

Aroma 0.03545 0.01901 1.865 0.070885 .

Flavor 0.08620 0.01993 4.325 0.000126 \*\*\*

Oakiness -0.03917 0.01917 -2.043 0.048828 \*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.0842 on 34 degrees of freedom

Multiple R-squared: 0.6793, Adjusted R-squared: 0.651

F-statistic: 24.01 on 3 and 34 DF, p-value: 1.591e-08

The R-squared for this model is only 67.9%, which is much smaller than that of wine.lm2(70.4%)

All these tests have told us that we should still go with wine.lm2:

i.e., Quality=β0+β1Aroma+β2Flavor+β3Oakiness+ε

Lastly, we attempt to detect if there is any combination-outlier by using "leverages":

> h=hatvalues(wine.lm2)

> cb=cbind(wine,h)

> head(cb)

Clarity Aroma Body Flavor Oakiness Quality Region h

1 1 3.3 2.8 3.1 4.1 9.8 1 0.10112647

2 1 4.4 4.9 3.5 3.9 12.6 1 0.08512861

3 1 3.9 5.3 4.8 4.7 11.9 1 0.08913132

4 1 3.9 2.6 3.1 3.6 11.1 1 0.11529314

5 1 5.6 5.1 5.5 5.1 13.3 1 0.06944827

6 1 4.6 4.7 5.0 4.1 12.8 1 0.03791098

> L=2\*(6+1)/n

> table(cb$h>L)

FALSE

38

> cb[cb$h>=L]

data frame with 0 columns and 38 rows

Zero outlier has been detected.

**Results**

"Aroma" and "Flavor" are strongly positively related to “quality”

"Oakiness" does not associate with "Quality" although it is somehow significant in the model.

“body” is weakly positively related to “quality” but it is significant in the final model.

Our final model is Quality=β0+β1Aroma+β2Flavor+β3Oakiness+ε.

If we use the coefficients the final fitted regression model becomes:

=6.4672+0.5801Aroma+1.1997Flavor-0.6023Oakiness



Therefore 1 unit of increase in "Aroma" will cause an increase of 0.5801 in response variable (overall "Quality") when other variables are held constant;

1 unit of increase in "Flavor" will cause an increase of 1.1997 in response variable (overall "Quality") when other variables are held constant;

1 unit of increase in "Oakiness" will cause a decrease of 0.6023 in response variable (overall "Quality") when other variables are held constant.

**Conclusion and discussion**

Therefore we have seen that out of the six variables, “Aroma” and “Flavor” are positively related to “Quality”, “Body” is weakly related to “Quality” while “Clarity”, “Oakiness” and “Region” are not correlated to flavor.

This has proven that our assumption in the beginning was wrong, and it also surprises that “Clarity” is not correlated to “Quality”, since in literature review, Negger has stated that “Clarity is an essential quality required by consumers”.

What we are confused about is that we see "Oakiness" does not associate with "Quality" from the pairs plot although it is somehow significant in the model and “Body” is weakly positively related to “Quality” but it is significant in the final model. These are contradictions.

**Reference:**

1. [http://www.utsc.utoronto.ca/%7ebutler/c32/wine.txt.](http://www.utsc.utoronto.ca/~butler/c32/wine.txt.)
2. Eggers, N. (2006, June 21). CLARIFICATION AND STABILIZATION TREATMENTS: FINING WINE.

[https://people.ok.ubc.ca/neggers/Chem422A/Clarification%20and%20stablization%20treatments.pdf](https://people.ok.ubc.ca/neggers/Chem422A/Clarification and stablization treatments.pdf)

1. Marais, J. (1983, August). Terpenes in the Aroma of Grapes and Wine: A Review. <http://www.click-on-wine.com/uploads/96974/files/Terpenes-in-the-Aroma.pdf>

# 4. Pucktte, M. (2012 July 23). How Basic Wine Characteristics Help You Find Favorites. http://winefolly.com/review/wine-characteristics/.