Linear Regression

Consider the wine data. The data come from a study of Pinot Noir wine quality. The dataset contains 38 observations and 7 variables: Quality, Clarity, Aroma, Body, Flavor, Oakiness, and Region. The goal is to develop a model that relates the quality of Pinot Noir with its features. The model can potentially be used to predict the quality of the wine.

a) **Figure** represent the Scatterplot matrix for wine data. There is a strong positive correlation between response variable `Quality` and predictor variables `Flavor` and `Aroma`, moderate correlation between `Quality` and `Body`. Moreover, there is a strong positive correlation between predictor variables `Aroma` and `Flavor` and `Body` and `Flavor`.

```
library(car)
library(lmtest)
library(ggplot2)

wine<-read.table("wine.txt",header = TRUE)
#View(wine)
wine$Region<-as.factor(wine$Region)
str(wine)
```

```
'data.frame':   38 obs. of  7 variables:
 $ Clarity : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Aroma   : num  3.3 4.4 3.9 3.9 5.6 4.6 4.8 5.3 4.3 4.3 ...
 $ Body    : num  2.8 4.9 5.3 2.6 5.1 4.7 4.8 4.5 4.3 3.9 ...
 $ Flavor  : num  3.1 3.5 4.8 3.1 5.5 5 4.8 4.3 3.9 4.7 ...
 $ Oakiness: num  4.1 3.9 4.7 3.6 5.1 4.1 3.3 5.2 2.9 3.9 ...
 $ Quality : num  9.8 12.6 11.9 11.1 13.3 12.8 12.8 12 13.6 13.9 ...
 $ Region  : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 3 1 ...
```

```
cor(wine[1:6])
```

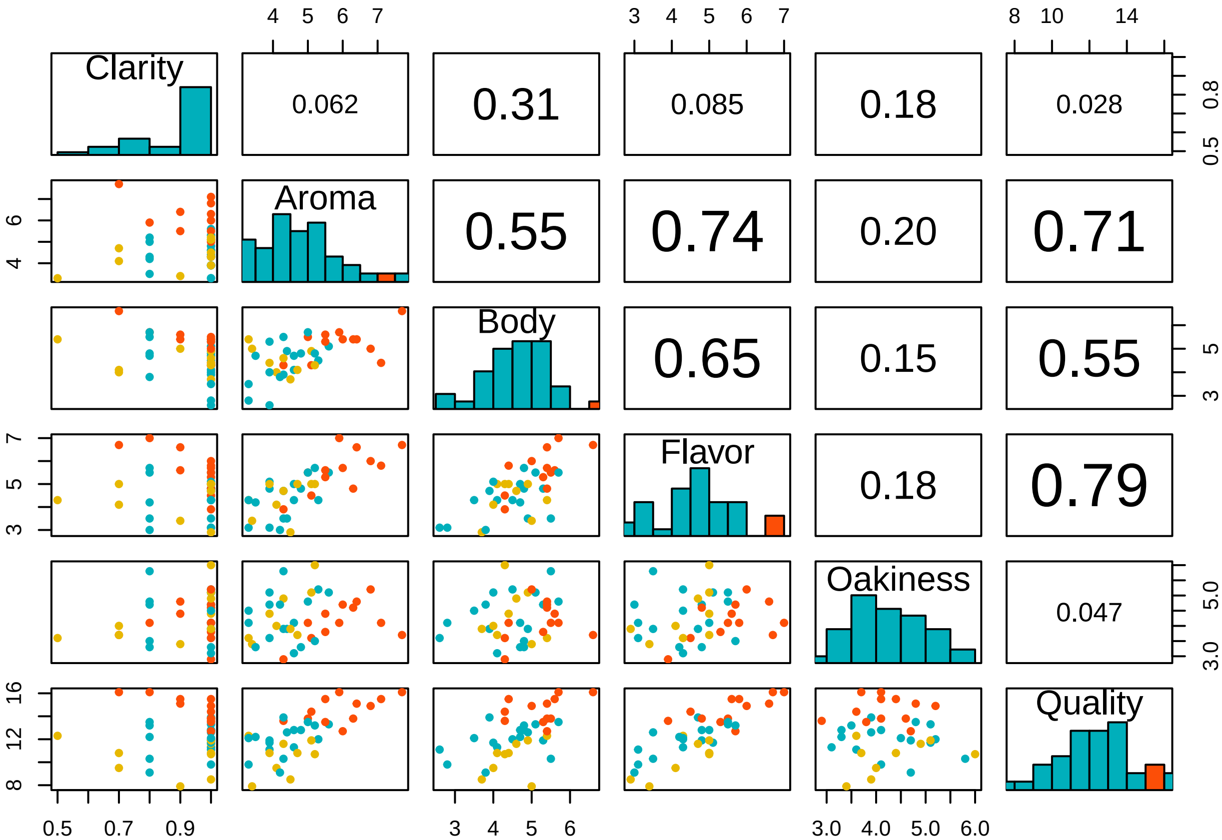
```
             Clarity     Aroma       Body      Flavor    Oakiness     Quality
Clarity   1.00000000 0.0619021 -0.3083783 -0.08515993  0.18321471  0.02844131
Aroma     0.06190210 1.0000000  0.5489102  0.73656121  0.20164445  0.70732432
Body     -0.30837826 0.5489102  1.0000000  0.64665917  0.15210591  0.54870219
Flavor   -0.08515993 0.7365612  0.6466592  1.00000000  0.17976051  0.79004713
Oakiness  0.18321471 0.2016444  0.1521059  0.17976051  1.00000000 -0.04704047
Quality   0.02844131 0.7073243  0.5487022  0.79004713 -0.04704047  1.00000000
```

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y, use = "complete.obs"))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste(prefix, txt, sep = "")
  if (missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex =  cex.cor * (1 + r) / 2)
}

panel.hist <- function(x, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
```

```
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks
  nB <- length(breaks)
  y <- h$counts
  y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y,  ...)
}

#Scatter plot matrix for wine data
my_cols <- c("#00AFBB", "#E7B800", "#FC4E07")
pairs(wine[,1:6],col = my_cols[wine$Region],pch=20,upper.panel = panel.cor,diag.panel  = panel.hist,oma=c(2,2,2,
```



b) From **Figure1** we can see that histogram for variable `Quality` is slightly left skewed. Therefore to explore whether transformation is necessary for variable `Quality` we examine residual plots for multiple linear regression model for `Quality` vs all other predictor variables.

```
full.model<-lm(Quality~.,data=wine)
```

```
#Evaluating model assumptions
shapiro.test(full.model$residuals)
```

```

    Shapiro-Wilk normality test

data:  full.model$residuals
W = 0.98569, p-value = 0.8993
```

```
bptest(full.model)
```

```

    studentized Breusch-Pagan test
```

```
data:  full.model
BP = 8.2839, df = 7, p-value = 0.3082
```

```
durbinWatsonTest(full.model)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.2151442      1.540071   0.092
 Alternative hypothesis: rho != 0
```

```
par(mfrow=c(1,4))
qqnorm(full.model$residuals, xlab = "Expected value", ylab = "Residual", main = "Normal Probability Plot",pch =
qqline(full.model$residuals)
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)

plot(x = full.model$fitted.values, y = full.model$residuals, abline(0,0), xlab = "Fitted Value",ylab = "Residual
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)

plot(resid(full.model),type = "l", main = "Time series plot",cex.lab=0.8,xaxt="n",yaxt="n",cex.main=0.8)
abline(h=0)
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)

plot(full.model,which = 5,caption = "",main="Residuals vs Leverage",cex.lab=0.8,xaxt="n",yaxt="n",cex.main=0.8)
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)
```

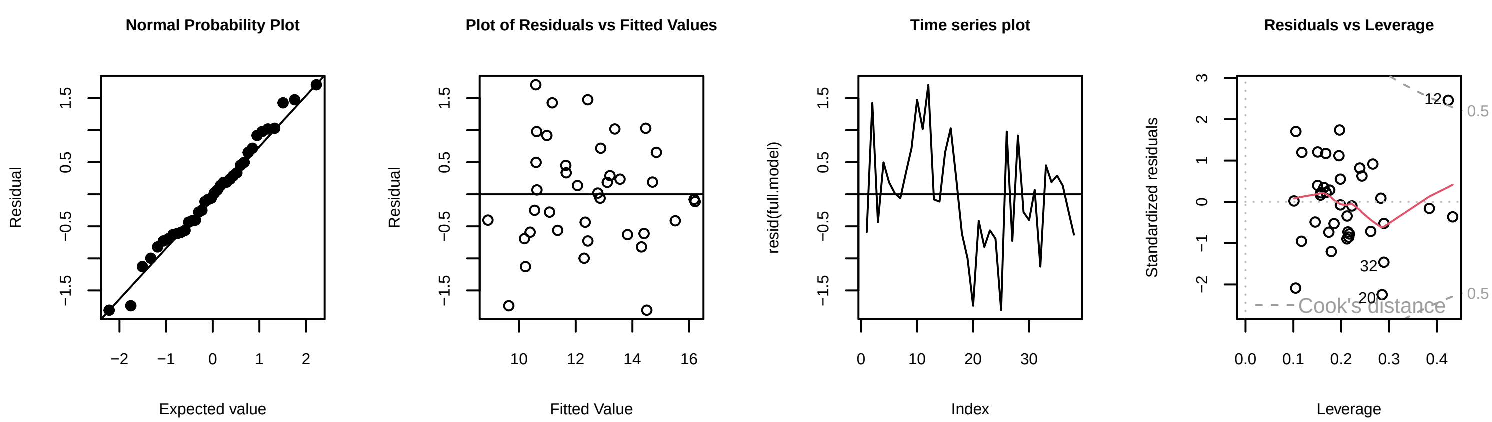


Figure 1: *Accesing model assumptions for Quality vs all other predictors*

Figure1 represent residual plots for the multiple linear regression model for `Quality` vs all other predictor variables.

- From the qqplot we notice that points follow a straight line and the shapirowilks test coincides with the normal QQ plot with pvalue 0.8993> 0.05 implying normality holds.
- We do not see discernible curve pattern to the residuals vs. fitted plot and Breush-Pagan test with pvalue= $0.3082 > 0.05$ indicating constant variance assumption holds.
- Error terms does shows pattern implying independence of Error Terms.
- Pvalue for Durbin-Watson test=$0.108 > 0.05$ imply that autocorrelation does not present in the model.
- From cooks distance we can see there is one influential observation and that is 12 th observation.

Model assumption holds for this model and therefore `Quality` is appropriate as a response variable and transformation is not necessary.

c) For each predictor, simple linear regression model was fitted to predict the response `Quality` and **Table 1** present the P-values for model significance. In the simple linear regression testing for model significance is $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. All models are significant, except models `Quality` vs `Clarity` and `Quality` vs `Oakiness`.

Predctor variable	Clarity	Aroma	Body	Flavor	Oakiness	Region
P-value	0.865	6.87×10^{-7}	0.000361	3.68e-09	0.7791	6.587×10^{-8}

Table 1: Summary of results for LDA and QDA

```
Call:
lm(formula = Quality ~ Clarity, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5257 -1.3227  0.0947  1.2773  3.7681 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.0034     2.5610   4.687 3.89e-05 ***
Clarity       0.4692     2.7486   0.171    0.865    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.073 on 36 degrees of freedom
Multiple R-squared:  0.0008089, Adjusted R-squared:  -0.02695 
F-statistic: 0.02914 on 1 and 36 DF,  p-value: 0.8654


Call:
lm(formula = Quality ~ Aroma, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4726 -0.8574 -0.0091  0.8346  2.2563 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.9583     1.1050   5.392 4.51e-06 ***
Aroma         1.3365     0.2226   6.004 6.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.466 on 36 degrees of freedom
Multiple R-squared:  0.5003,	Adjusted R-squared:  0.4864 
F-statistic: 36.04 on 1 and 36 DF,  p-value: 6.871e-07


Call:
lm(formula = Quality ~ Body, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9669 -0.8386  0.0620  1.2204  3.4502 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.0580     1.6441   3.685 0.000748 ***
Body          1.3618     0.3458   3.938 0.000361 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.734 on 36 degrees of freedom
```

```
Multiple R-squared:  0.3011,    Adjusted R-squared:  0.2817
F-statistic: 15.51 on 1 and 36 DF,  p-value: 0.0003612


Call:
lm(formula = Quality ~ Flavor, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.38583 -0.72226 -0.00756  0.62006  2.52822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9414     0.9911   4.986 1.57e-05 ***
Flavor        1.5719     0.2033   7.732 3.68e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.271 on 36 degrees of freedom
Multiple R-squared:  0.6242,    Adjusted R-squared:  0.6137
F-statistic: 59.79 on 1 and 36 DF,  p-value: 3.683e-09


Call:
lm(formula = Quality ~ Oakiness, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6483 -1.3886 -0.0527  1.2907  3.6429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.9916     1.9918   6.522  1.4e-07 ***
Oakiness     -0.1304     0.4614  -0.283    0.779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.071 on 36 degrees of freedom
Multiple R-squared:  0.002213,  Adjusted R-squared:  -0.0255
F-statistic: 0.07984 on 1 and 36 DF,  p-value: 0.7791


Call:
lm(formula = Quality ~ Region, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8765 -0.8532  0.2395  0.9167  1.9235

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.9765     0.3180  37.662  < 2e-16 ***
Region2      -1.5320     0.5405  -2.834  0.00757 **
Region3       2.6069     0.4944   5.273 7.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.311 on 35 degrees of freedom
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.5891
F-statistic: 27.52 on 2 and 35 DF,  p-value: 6.587e-08
```

d) Multiple regression model to predict the response using all of the predictors.

$$Quality = 7.81437 + 0.01705 Clarity + 0.08901 Aroma + 0.07967 Body - 0.34644 Oakiness - 1.51285 Region2 + 0.97259 Region3$$

- When testing the significance of jth predictor: i.e $H_0 : \beta_j = 0$ vs $H_0 : \beta_j \neq 0$, we can reject the null hypothesis for predictors `Flavor` and `Region` as there p values 6.25×10^{-5} and 2.92×10^{-4} respectively < 0.05. Thus we can conclude that each predictor `Flavor` and `Region` is associated with response `Quality` after adjusting for the other predictors.
- When testing for model significance: i.e $H_0 : \beta_1 = \cdots = \beta_p = 0$ vs H_1 : atleast one $\beta_j \neq 0$ we reject the null hypothesis and conclude that model is significant as p value $= 3.295 \times 10^{-10}$.
- Adjusted R^2 is 0.7997 indicates that approximately 80% proportion of total variation explained by the regression.

```
full.model<-lm(Quality~.,data=wine)
summary(full.model)
```

```
Call:
lm(formula = Quality ~ ., data = wine)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.80824 -0.58413 -0.02081  0.48627  1.70909 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.81437    1.96944   3.968 0.000417 ***
Clarity      0.01705    1.45627   0.012 0.990736    
Aroma        0.08901    0.25250   0.353 0.726908    
Body         0.07967    0.26772   0.298 0.768062    
Flavor       1.11723    0.24026   4.650 6.25e-05 ***
Oakiness    -0.34644    0.23301  -1.487 0.147503    
Region2     -1.51285    0.39227  -3.857 0.000565 ***
Region3      0.97259    0.51017   1.906 0.066218 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9154 on 30 degrees of freedom
Multiple R-squared:  0.8376,    Adjusted R-squared:  0.7997 
F-statistic:  22.1 on 7 and 30 DF,  p-value: 3.295e-10
```

```
region.model<-lm(Quality~.-Region,data=wine)

anova(full.model,region.model)
```

```
Analysis of Variance Table

Model 1: Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + Region
Model 2: Quality ~ (Clarity + Aroma + Body + Flavor + Oakiness + Region) - 
    Region
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     30 25.140                                  
2     32 43.248 -2   -18.108 10.804 0.0002924 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e) Build a reasonably good multiple regression model for these data.

- First, performance of all possible models were compared using $R^2_{adjusted}, MSE_p, BIC$ and Mallows' C_p. According to the plots of $R^2_{adjusted}, MSE_p, BIC$ and Mallows' C_p vs number of variables, only 3 variables are enough to explain the model as after 3 variables there is no much variation added to the model.

- Next, stepwise selection was carried out and variables `Flavor`, `Oakiness` and `Region` identified as the most important predictors. Although variable `Oakiness` was selected using stepwise method, it is not significant in the model `Quality` vs `Flavor`, `Oakiness` and `Region`. Moreover it does not add much of the variation to the model as $R^2_{adjusted}$ for model with `Oakiness` and without it is 0.8164 and 0.8087 respectively. Therefore variables `Flavor` and `Region` used for the final model.
- Then look for the pairwise interactions between `Flavor` and `Region` and it is not significant as pvalue=0.3378 $>$ 0.05.
- **Figure** represent residual plots for the multiple linear regression model for `Quality` vs vs `Flavor` and `Region`. Model assumptions holds for this model. From the qqplot we notice that points follow a straight line and the shapirowilks test coincides with the normal QQ plot with pvalue 0.9577 $>$ 0.05 implying normality holds.We do not see discernible curve pattern to the residuals vs. fitted plot and Breush-Pagan test with pvalue= 0.3817 $>$ 0.05 indicating constant variance assumption holds.Error terms does not shows pattern implying independence of Error Terms. Pvalue for Durbin-Watson test=0.148 $>$ 0.05 imply that autocorrelation does not present in the model.

```
#perfoarming stepwise regression
step.lm <- step(full.model,direction = "both",trace=FALSE)
summary(step.lm)
```

```

Call:
lm(formula = Quality ~ Flavor + Oakiness + Region, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.81290 -0.59794  0.03423  0.42452  1.71484 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.1208     1.0164   7.990 3.23e-09 ***
Flavor        1.1920     0.1772   6.727 1.15e-07 ***
Oakiness     -0.3183     0.2039  -1.561 0.128060    
Region2      -1.5155     0.3614  -4.193 0.000194 ***
Region3       1.0935     0.4009   2.728 0.010130 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8763 on 33 degrees of freedom
Multiple R-squared:  0.8363,    Adjusted R-squared:  0.8164 
F-statistic: 42.14 on 4 and 33 DF,  p-value: 1.595e-12
```

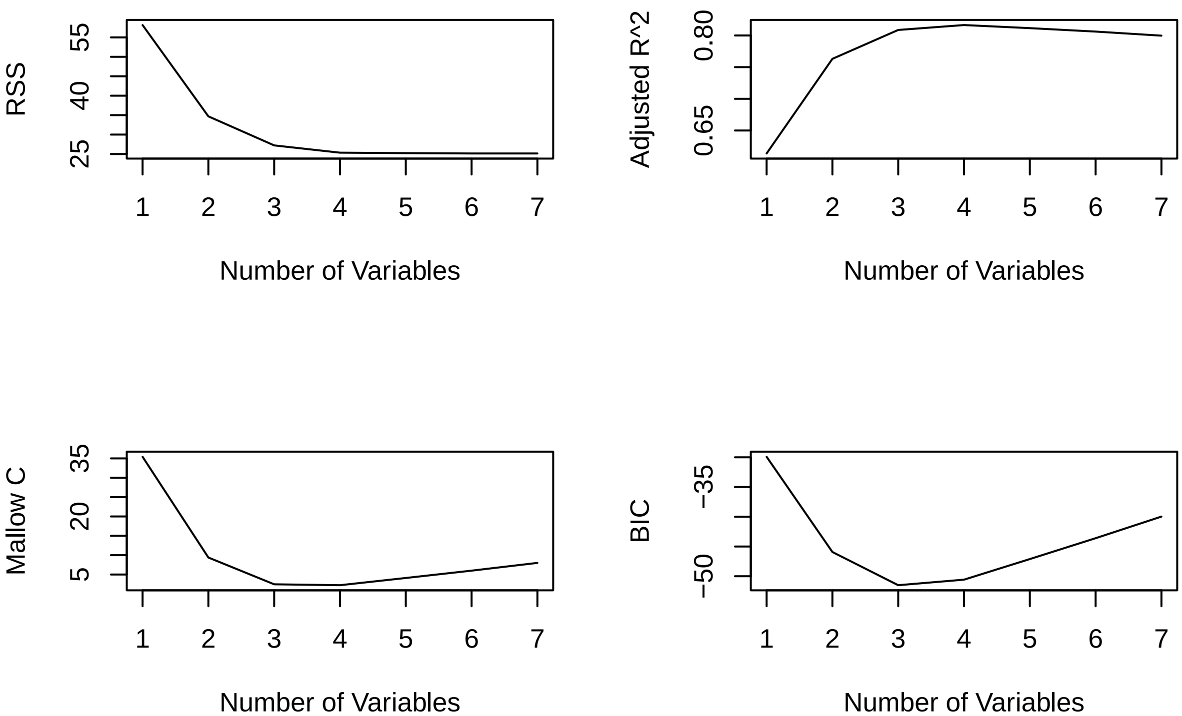
```
library(leaps)
regfit = regsubsets(Quality~.,data=wine) #full model
regsumm = summary(regfit)
regsumm
```

```

Subset selection object
Call: regsubsets.formula(Quality ~ ., data = wine)
7 Variables  (and intercept)
         Forced in Forced out
Clarity      FALSE      FALSE
Aroma        FALSE      FALSE
Body         FALSE      FALSE
Flavor       FALSE      FALSE
Oakiness     FALSE      FALSE
Region2      FALSE      FALSE
Region3      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         Clarity Aroma Body Flavor Oakiness Region2 Region3
1  ( 1 ) " "     " "   " "  "*"    " "      " "     " "    
2  ( 1 ) " "     " "   " "  "*"    " "      "*"     " "    
```

```
3  ( 1 ) " "     " "    " "  "*"     " "       "*"      "*"
4  ( 1 ) " "     " "    " "  "*"     "*"       "*"      "*"
5  ( 1 ) " "     "*"    " "  "*"     "*"       "*"      "*"
6  ( 1 ) " "     "*"    "*"  "*"     "*"       "*"      "*"
7  ( 1 ) "*"     "*"    "*"  "*"     "*"       "*"      "*"
```

```r
#Plot of Variable selection criteria with all variables.
# need to compare the performance of the different models for choosing the best number of variables for reduce
par(mfrow = c(2,2))
plot(regsumm$rss, xlab = "Number of Variables", ylab = "RSS", type ="l")
plot(regsumm$adjr2, xlab = "Number of Variables", ylab = "Adjusted R^2", type ="l")
plot(regsumm$cp, xlab = "Number of Variables", ylab = "Mallow C", type ="l")
plot(regsumm$bic, xlab = "Number of Variables", ylab = "BIC", type ="l")
```




```r
par(mfrow = c(1,1))
```

```r
#evaluating interactions
m1<-lm(Quality~Flavor+Region,data = wine)
m2<-lm(Quality~Flavor+Region+Flavor*Region,data = wine)

anova(m1,m2)
```

```
Analysis of Variance Table

Model 1: Quality ~ Flavor + Region
Model 2: Quality ~ Flavor + Region + Flavor * Region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     34 27.213
2     32 25.429  2    1.7845 1.1229 0.3378
```

```r
#final model
Reduce.model<-lm(Quality~Flavor+Region,data = wine)
summary(Reduce.model)
```

```

Call:
lm(formula = Quality ~ Flavor + Region, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.97630 -0.58844  0.02184  0.51572  1.94232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0943     0.7912   8.967 1.76e-10 ***
Flavor        1.1155     0.1738   6.417 2.49e-07 ***
Region2      -1.5335     0.3688  -4.158 0.000205 ***
Region3       1.2234     0.4003   3.056 0.004346 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8946 on 34 degrees of freedom
Multiple R-squared:  0.8242,    Adjusted R-squared:  0.8087
F-statistic: 53.13 on 3 and 34 DF,  p-value: 6.358e-13
```

```r
#Evaluating model assumptions
shapiro.test(Reduce.model$residuals)
```

```
    Shapiro-Wilk normality test

data:  Reduce.model$residuals
W = 0.98843, p-value = 0.9577
```

```r
bptest(Reduce.model)
```

```
    studentized Breusch-Pagan test

data:  Reduce.model
BP = 3.0648, df = 3, p-value = 0.3817
```

```r
durbinWatsonTest(Reduce.model)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.1866772      1.603486   0.156
 Alternative hypothesis: rho != 0
```

```r
#model assumptions
par(mfrow=c(1,4))
qqnorm(Reduce.model$residuals, xlab = "Expected value", ylab = "Residual", main = "Normal Probability Plot",pch
qqline(Reduce.model$residuals)
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)

plot(x = Reduce.model$fitted.values, y = Reduce.model$residuals, abline(0,0), xlab = "Fitted Value",ylab = "Resi
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)

plot(resid(Reduce.model),type = "l", main = "Time series plot",cex.lab=0.8,xaxt="n",yaxt="n",cex.main=0.8)
abline(h=0)
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)

plot(Reduce.model,which = 5,caption = "",main="Residuals vs Leverage",cex.lab=0.8,xaxt="n",yaxt="n",cex.main=0.8
axis(2,cex.axis=0.8)
axis(1,cex.axis=0.8)
```

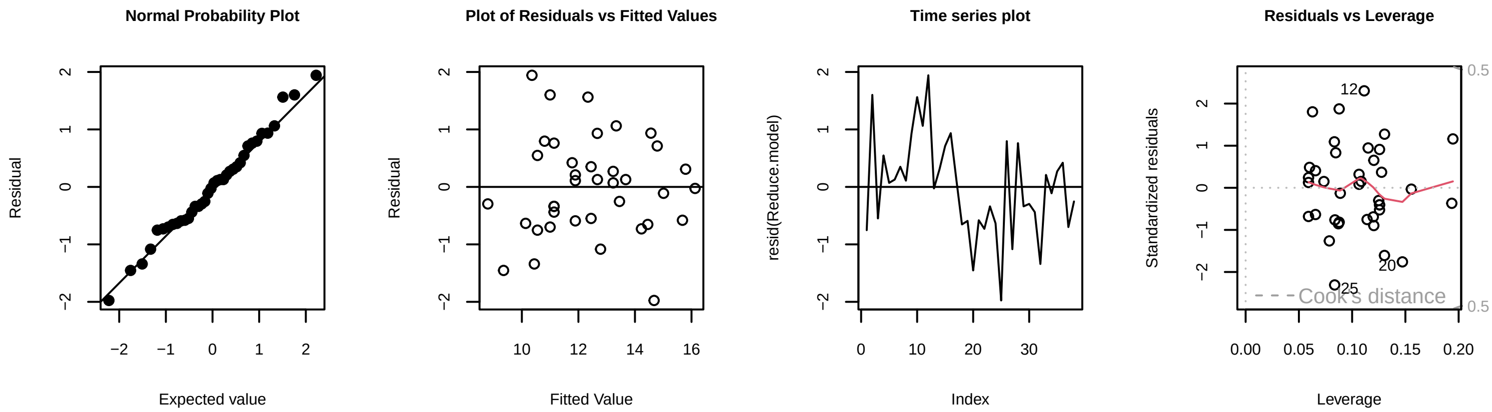


Figure 2: *Accesing model assumptions for Quality vs Flavor and region*

f) final model:

$$Quality = 7.0943 + 1.1155Flavor - 1.5335Region2 + 1.2234Region3$$

Adjusted R-squared: 0.8087 and p-value: 6.358e-13 $< 0.05 \Rightarrow$ model is significant. Moreover p values for testing $H_0 : \beta_j = 0$ vs $H_0 : \beta_j \neq 0$ are $2.49 \times 10^{-7}, 2.46 \times 10^{-6}$ for `Flavor` and `Quality` respectively $\Rightarrow$ that each predictor is significant .

```
anova(Reduce.model,lm(Quality~Flavor,data=wine))
```

```
Analysis of Variance Table

Model 1: Quality ~ Flavor + Region
Model 2: Quality ~ Flavor
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     34 27.213
2     36 58.173 -2    -30.96 19.341 2.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

g) `Quality` of a wine from `Region 1` with `Flavor` equal to its mean value (4.7684) is 12.4137.

- 95% Prediction interval : (10.53775,14.28967). Thus we can be 95% confident that this new observation will fall within (10.53775,14.28967)
- 95% Confidence interval : (11.95152,12.8756). Thus We can be 95% confident that the average `Quality` of a wine from `Region 1` with `Flavor` equal to its mean value is between 11.95152 and 12.8756.

```
newdat<-data.frame(Flavor=mean(wine$Flavor),Region="1")
predict(Reduce.model,newdat)
```

```
       1
12.41371
```

```
predict(Reduce.model, newdata = newdat, interval = "prediction",level=0.95)
```

```
       fit      lwr      upr
1 12.41371 10.53775 14.28967
```

```
predict(Reduce.model, newdata = newdat, interval = "confidence",level=0.95)
```

```
       fit      lwr     upr
1 12.41371 11.95152 12.8759
```