一、研究背景與動機

不論是食品製造業、機械製造業、晶片製造業、...等等不同的製造業，都會有個共同的重要問題，就是如何提升自家產品的品質且降低生產成本。但是往往在整個繁瑣的製程當中無法有效掌握問題所在，總是在產生生產完成之後才知道良率如何。所以，如何從產品的品質或是良率，反推產品的製程，或是產品的成分含量比率，透過分析的方式來達到產品品質是非常重要的課題。

二、研究目的

此研究目的，最主要是收集紅、白酒相關資訊，透過迴歸分析方式來找出影響品質的中要因素，並透過調整該獨立變數的值來達到高品質的的紅、白酒。初步會採用分類法則的 C5.0 來嘗試塑模與測試，先瞭解兩種酒類的判斷標準會不會一致。最後，再採取迴歸分析方式來進行塑模分析，找出一個較好的模型，來改良紅、白酒的高品質比率。

三、變數介紹

這兩個資料集(winequality-rec.csv & winequality-white.csv)是“Vinho Verde”酒，紅色和白色的變種酒品，透過常期收集所儲存下來的相關資料。更多資料可以參考：http://www.vinhoverde.pt/en/ 或參考文獻[科爾特斯等，2009]。

由於隱私和後勤問題，只有物理化學（輸入）和感覺（輸出）變量是可用的（例如，沒有關於葡萄類型，葡萄酒品牌，葡萄酒售價等的數據）。

有關紅酒的資料筆數 1599;白酒的筆數 4898。共計有 11 個輸入變數(獨立變數或稱預測變數)以及一個輸出變數 quality(相依變數或稱為解釋變數)。其中完全沒有遺漏值。變數說明如表 3-1。

表 3-1 變數說明

序號	變數名稱	變數說明	變數型態	備註說明
1	fixed.acidity	固定的酸度	real	
2	volatile.acidity	揮發性酸	real	
3	citric.acid	檸檬酸	real	
4	residual.sugar	殘糖	real	

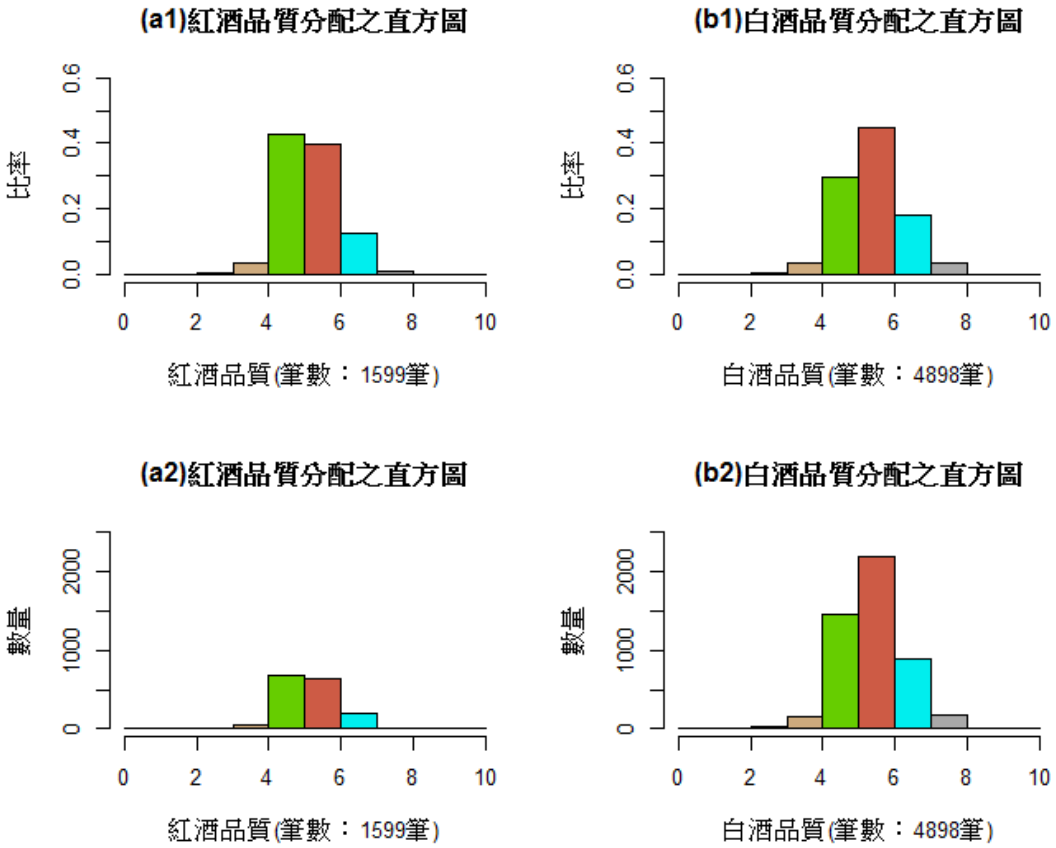
5	chlorides	氯化物	real	
6	free.sulfur.dioxide	游離二氧化硫	integer	
7	total.sulfur.dioxide	二氧化硫總量	integer	
8	density	密度	real	
9	pH	pH 值	real	
10	sulphates	硫酸鹽	real	
11	alcohol	酒精濃度	real	
12	quality	品質	integer	0 到 10 之間的分數

先利用 R 語言將資料讀入查看資料內容如下，分別為 red(紅酒)與 white(白酒)的資料前六筆資料。

```
> red <- read.table(file="winequality-red.csv",header=TRUE,sep=";")
> white <- read.table(file="winequality-white.csv",header=TRUE,sep=";")
> head(red)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1           7.4             0.70        0.00            1.9     0.076
2           7.8             0.88        0.00            2.6     0.098
3           7.8             0.76        0.04            2.3     0.092
4          11.2             0.28        0.56            1.9     0.075
5           7.4             0.70        0.00            1.9     0.076
6           7.4             0.66        0.00            1.8     0.075
  free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality
1                  11                   34  0.9978 3.51      0.56     9.4       5
2                  25                   67  0.9968 3.20      0.68     9.8       5
3                  15                   54  0.9970 3.26      0.65     9.8       5
4                  17                   60  0.9980 3.16      0.58     9.8       6
5                  11                   34  0.9978 3.51      0.56     9.4       5
6                  13                   40  0.9978 3.51      0.56     9.4       5
> head(white)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1           7.0             0.27        0.36           20.7     0.045
2           6.3             0.30        0.34            1.6     0.049
3           8.1             0.28        0.40            6.9     0.050
4           7.2             0.23        0.32            8.5     0.058
5           7.2             0.23        0.32            8.5     0.058
6           8.1             0.28        0.40            6.9     0.050
  free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality
1                  45                  170  1.0010 3.00      0.45     8.8       6
2                  14                  132  0.9940 3.30      0.49     9.5       6
```

```
3          30          97 0.9951 3.26 0.44 10.1 6
4          47         186 0.9956 3.19 0.40  9.9 6
5          47         186 0.9956 3.19 0.40  9.9 6
6          30          97 0.9951 3.26 0.44 10.1 6
>
```

先從紅、白酒的製造品質的分佈情況來瞭解製造品質情況，從下圖可以很明顯，以比率而言(a1)的紅酒與(b1)白酒的大多數百分比都是落於[5,7]的區間內。反而，從實際數量的(a2)與(b2)比較看不出兩者共通性，因為數量大距太大。



為了先瞭解獨立變數對相依變數 quality 的重要性，初步採用 C5.0 的分類法則演算法試著塑模與預測。並且，為了能夠瞭解紅酒與白酒的品質是否會有相同的因素，所以會將此兩個資料集分別各自執行一次，再來評估結果，以做為後續採用迴歸分析之參考與比較。

```
#【C5.0 演算法{C50}】----------
install.packages("C50",repos="http://ftp.yzu.edu.tw/CRAN/")
library("C50")

#【(1)C5.0 training phase】----------
modeling.C50 <- function(data) {
   set.seed(300)
```

```
    s <- sample(1:nrow(data),round(nrow(data)*0.5))
    data.train <- data[s,]
    data.test <- data[-s,]
    rule <- C5.0(factor(data.train$quality)~.,data=data.train)

    #【(2)C5.0 predict(test) phase】----------
    pred <- predict(rule, data.test)
    tb <- table("實際品質"=data.test$quality,"預測品質"=pred)
    return(list(tb=tb,summary=summary(rule)))
}

r <- modeling.C50(data=red)
r$summary
r$tb

w <- modeling.C50(data=white)
w$summary
w$tb
```

表 3-2 紅、白酒的因素使用率

紅酒的因素使用率	白酒的因素使用率
Attribute usage:	Attribute usage:
100.00% alcohol	100.00% alcohol
98.75% total.sulfur.dioxide	84.81% volatile.acidity
88.50% sulphates	79.42% fixed.acidity
73.00% volatile.acidity	79.26% free.sulfur.dioxide
47.63% residual.sugar	65.33% citric.acid
46.50% chlorides	62.92% pH
36.88% free.sulfur.dioxide	60.96% chlorides
28.88% pH	58.35% sulphates
28.75% citric.acid	57.49% residual.sugar
28.50% fixed.acidity	47.49% total.sulfur.dioxide
5.13% density	37.73% density

表 3-3 紅、白酒的預測情況

紅酒的正確情況						
預測品質						
實際品質	3	4	5	6	7	8
3	0	1	2	1	0	0
4	0	0	14	8	1	0

白酒的正確情況							
預測品質							
實際品質	3	4	5	6	7	8	9
3	0	0	5	3	0	0	0
4	1	24	35	26	7	1	0

5	0	9	220	117	8	2	
6	0	11	79	168	35	8	
7	0	0	9	42	51	2	
8	0	0	0	5	5	1	

5	5	25	428	221	32	3	0
6	3	16	233	651	164	20	0
7	0	4	35	187	199	15	0
8	0	2	5	35	39	23	0
9	0	0	1	1	0	0	0

從表 3-2 可以很明顯看出，紅、白酒對於品質的分類方式(變數)大有所不同，也就是在後續的分析當中，理當要採用分割處理，不該將資料集合併處理。

四、初步探索解釋變數

基於所搜集到的資料集如表 1，其中的『quality』品質是唯一可以當成反應變數 Y。其中的 11 個(fixed.acidity、volatile.acidity、citric.acid、residual.sugar、chlorides、free.sulfur.dioxide、total.sulfur.dioxide、density、pH、sulphates 以及 alcohol)皆可當成解釋變數 X。

表 4-1 變數說明

序號	變數名稱	變數說明	變數型態	備註說明
1	fixed.acidity	固定的酸度	real	
2	volatile.acidity	揮發性酸	real	
3	citric.acid	檸檬酸	real	
4	residual.sugar	殘糖	real	
5	chlorides	氯化物	real	
6	free.sulfur.dioxide	游離二氧化硫	integer	
7	total.sulfur.dioxide	二氧化硫總量	integer	
8	density	密度	real	
9	pH	pH 值	real	
10	sulphates	硫酸鹽	real	
11	alcohol	酒精濃度	real	
12	quality	品質	integer	0 到 10 之間的分數

4-1 建立線性模型與探討解釋變數

本節是先利用$Y = \beta_0 + \beta_1 \times X$的模型來產生迴歸模型，再透過迴歸係數$\beta_1$來初步探討解釋變數 X 是否會影響反應變數 Y。

【R code】產出 11 個解釋變數的迴歸模型

```
#【1.產生迴歸方程式】----
red.wine <- read.table(file="winequality-red.csv",header=TRUE,sep=";")
X <- c("fixed.acidity(固定的酸度)","volatile.acidity(揮發性酸)",
       "citric.acid(檸檬酸)","residual.sugar(殘糖)",
       "chlorides(氯化物)","free.sulfur.dioxide(游離二氧化硫)",
       "total.sulfur.dioxide(二氧化硫總量)","density(密度)",
       "pH(pH 值)","sulphates(硫酸鹽)","alcohol(酒精濃度)")

mod <- character()
for (i in 1:11) {
  f <- formula(paste0("quality~",names(red.wine)[i]))
  lm.model <- lm(f, data=red.wine)
```

```
    mod <- c(mod,paste0("Y=",coef(lm.model)[[1]],
                                ifelse(coef(lm.model)[[2]]>0,"+",""),
                                coef(lm.model)[[2]],"X"))
}
names(mod) <- X
mod
```

【執行結果】

fixed.acidity(固定的酸度)

"Y=5.15732186892476+0.0575386437438816X"

volatile.acidity(揮發性酸)

"Y=6.56574550647175-1.76143778011267X"

citric.acid(檸檬酸)

"Y=5.38172490062976+0.938452038802968X"

residual.sugar(殘糖)

"Y=5.61605450900269+0.00786511808072905X"

chlorides(氯化物)

"Y=5.82948465954584-2.21184171639912X"

free.sulfur.dioxide(游離二氧化硫)

"Y=5.69810722357722-0.00391086710132548X"

total.sulfur.dioxide(二氧化硫總量)

"Y=5.84717920042681-0.00454415145598421X"

density(密度)

"Y=80.2385380208004-74.8460136014855X"

pH(pH 值)

"Y=6.63592282675891-0.301983125787391X"

sulphates(硫酸鹽)

"Y=4.8477495343891+1.19771232303137X"

alcohol(酒精濃度)

"Y=1.87497488699707+0.360841765335039X"

【說明】

從以上 11 個解釋變數的迴歸模型來初步探討，『residual.sugar』(殘糖)與『total.sulfur.dioxide』(二氧化硫總量)兩者的迴歸係數β_1的估計量都趨近於 0，表示這兩個解釋變數對 Y 並沒有任何的貢獻。不過，必須等後續的 t-test 才能確定是否將這兩個 X 去除掉。

4-2 透過繪圖的視覺化來探索解釋變數 X

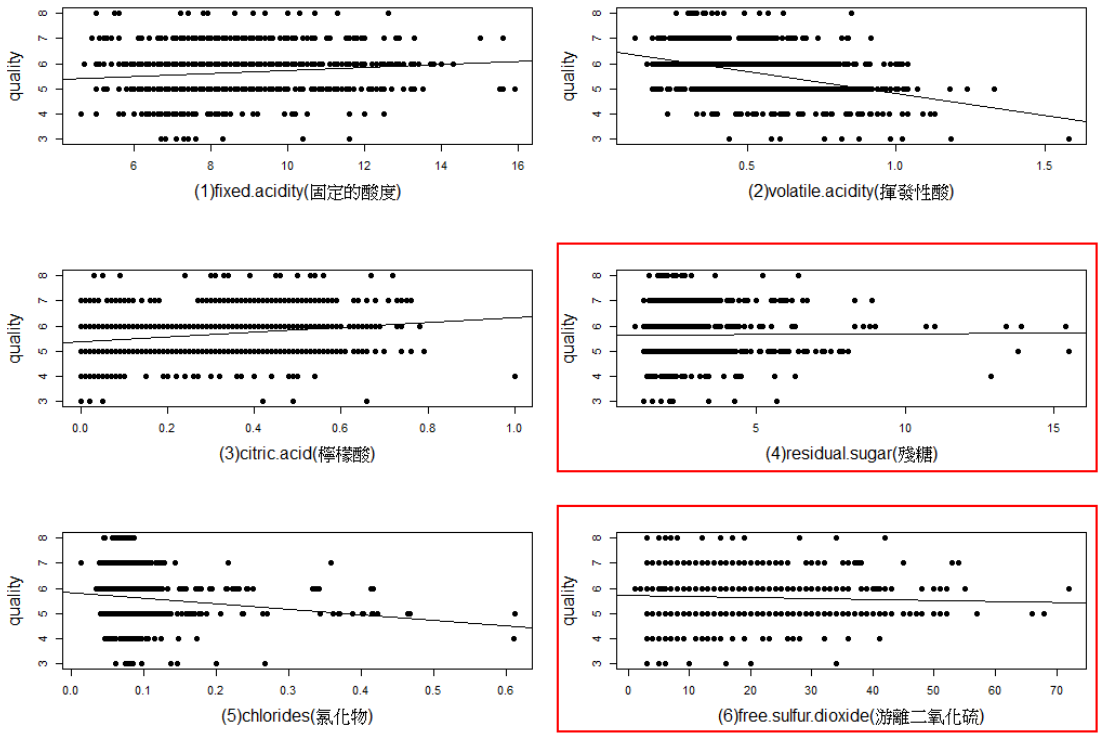
本節主要的目的是透過繪出 X-Y 散佈圖與迴歸線的關係，再與前一章的迴歸係數做一比較與呼應。

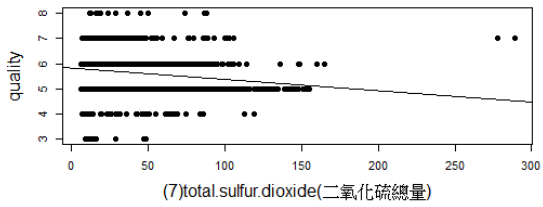
【R code】繪出 11 個解釋變數的散佈圖與迴歸線

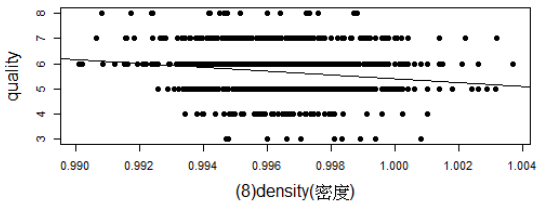
```
#【2.畫 scatter plot 和迴歸線】----
```

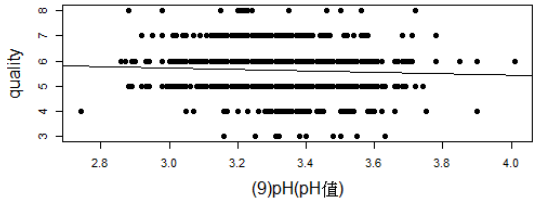
```
for (i in 1:11) {
   if (i==1 || i==7) {
      dev.new()
      par(mfrow=c(3,2))
   }
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   plot(red.wine[,i],red.wine$quality,cex.lab=1.5,
         xlab=paste0("(",i,")",X[i]),ylab="quality",pch=19)
   abline(lm.model)
}
par(mfrow=c(1,1))
```

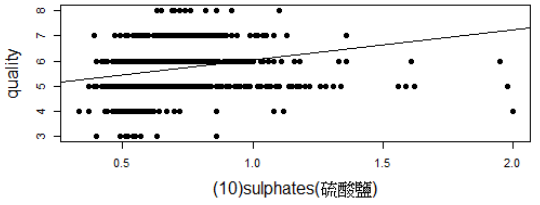
【執行結果】

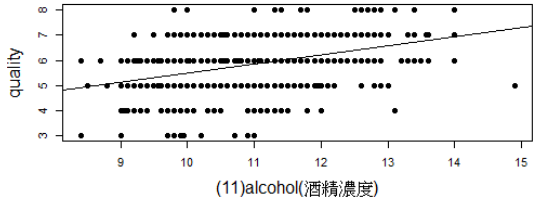












【說明】

因為β_1所代表的迴歸線的斜率，所以從 11 張圖中可以很清楚地看出『residual.sugar』(殘糖)與『total.sulfur.dioxide』(二氧化硫總量)兩者的迴歸線趨於水平線。不過，這也只是與前一章的迴歸係數β_1互相應證。至於是否確定這兩個解釋變數應該被摒除，還要再透過下一章的 t-test 來決定。

由於β_0所代表的是截距，也就是當 X=0 時，Y 的值。以這個個案來看，因為 Y 代表酒品的品質，分數越高代表品質越好。所以，當某一個解釋變數 X 消失時(也就是$\beta_1 = 0$)，β_0代表該酒品的品質分數。因此，當β_0越大且β_1的斜率又呈現負數時，代表該因素是造成品質下降的因素之一。

4-3 針對β_0與β_1的 t-test

【R code】針對β_0與β_1的 t-test

```
#【3.針對 beta0、beta1 做 t 檢定】----
beta0.p.value <- numeric(0)
beta1.p.value <- numeric(0)
for (i in 1:11) {
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   s <- summary(lm.model)
   beta0.p.value <- c(beta0.p.value, coef(s)[1,4])
   beta1.p.value <- c(beta1.p.value, coef(s)[2,4])
}
```

```
names(beta0.p.value) <- X
names(beta1.p.value) <- X
names(beta0.p.value)[beta0.p.value<=0.05]
names(beta1.p.value)[beta1.p.value<=0.05]
```

【執行結果】

```
> names(beta0.p.value)[beta0.p.value<=0.05]
 [1] "fixed.acidity(固定的酸度)"
 [2] "volatile.acidity(揮發性酸)"
 [3] "citric.acid(檸檬酸)"
 [4] "residual.sugar(殘糖)"
 [5] "chlorides(氯化物)"
 [6] "free.sulfur.dioxide(游離二氧化硫)"
 [7] "total.sulfur.dioxide(二氧化硫總量)"
 [8] "density(密度)"
 [9] "pH(pH 值)"
[10] "sulphates(硫酸鹽)"
[11] "alcohol(酒精濃度)"
> names(beta1.p.value)[beta1.p.value<=0.05]
 [1] "fixed.acidity(固定的酸度)"
 [2] "volatile.acidity(揮發性酸)"
 [3] "citric.acid(檸檬酸)"
 [4] "chlorides(氯化物)"
 [5] "free.sulfur.dioxide(游離二氧化硫)"
 [6] "total.sulfur.dioxide(二氧化硫總量)"
 [7] "density(密度)"
 [8] "pH(pH 值)"
 [9] "sulphates(硫酸鹽)"
[10] "alcohol(酒精濃度)"
```

【說明】

從以上的結果可以明確地看出，在以下的檢定中，所有的解釋變數都顯著，拒絕$\beta_0 = 0$；換言之，這 11 個解釋變數的β_0皆不為 0。

$$\begin{cases} H_{0:}: \beta_0 = 0 \\ H_a: \beta_0 \neq 0 \end{cases}$$

不過，針對以下的檢定中，卻只有 10 個解釋變數是顯著，獨缺一個解釋變數『residual.sugar』(殘糖)。換言之，我們無法拒絕它的$\beta_0 = 0$。與前兩章所提到的這兩個『residual.sugar』(殘糖)與『total.sulfur.dioxide』(二氧化硫總量)解釋變數有可能不影響 Y，命中了『residual.sugar』(殘糖)。

$$\begin{cases} H_{0:}: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$$

表 4-2 : 11 個解釋變數的β_0與β_1檢定結果

序號	變數名稱	變數說明	β_0 test	β_1 test
1	fixed.acidity	固定的酸度		
2	volatile.acidity	揮發性酸		
3	citric.acid	檸檬酸		
4	residual.sugar	殘糖		X
5	chlorides	氯化物		
6	free.sulfur.dioxide	游離二氧化硫		
7	total.sulfur.dioxide	二氧化硫總量		
8	density	密度		
9	pH	pH 值		
10	sulphates	硫酸鹽		
11	alcohol	酒精濃度		

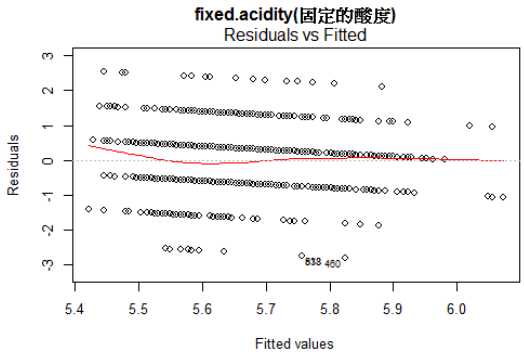
4-4 模型的殘差繪圖診斷

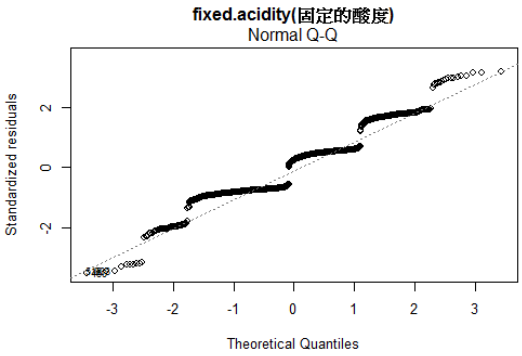
一個好的迴歸模型，殘差應該服從 Normal Distribution，所以可以透過 Q-Q plot 來察看是否標準殘差是否在對角線上。以及殘差與配適值之間的關係，殘差應該隨機分配於水平線殘差=0 的上下。

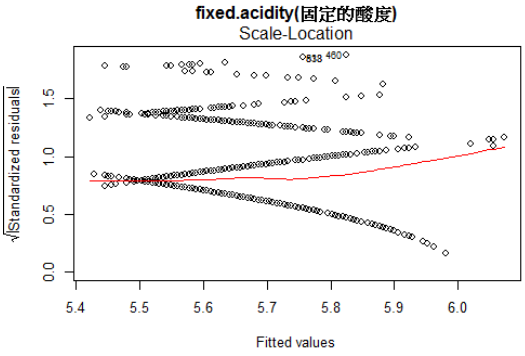
【R code】

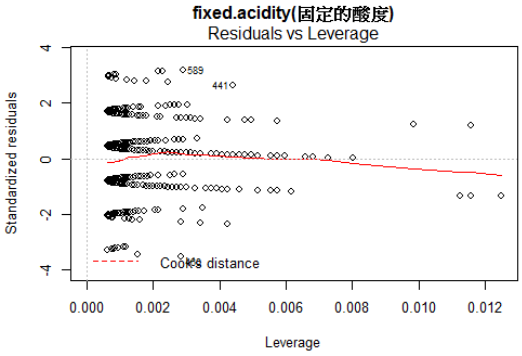
```
#【4.繪出四張圖，模型診斷】----
for (i in 1:11) {
   dev.new()
   par(mfrow=c(2,2))
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   plot(lm.model,main=X[i])
}
par(mfrow=c(1,1))
```

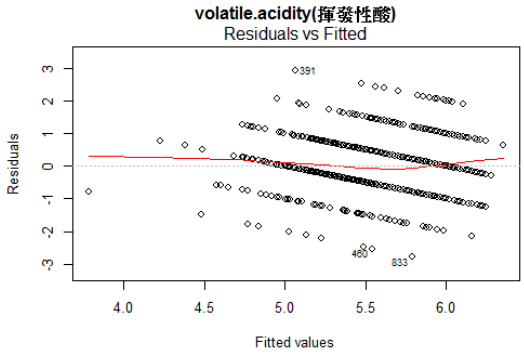
【執行結果】

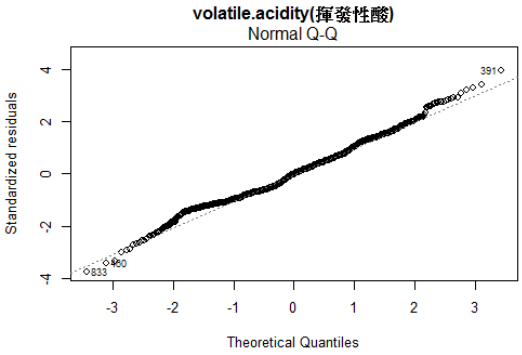

fixed.acidity(固定的酸度)
Residuals vs Fitted
Residuals
Fitted values

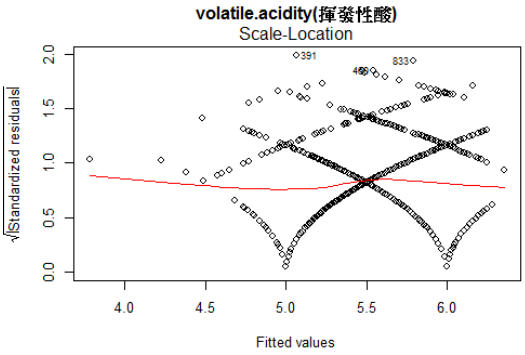

fixed.acidity(固定的酸度)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

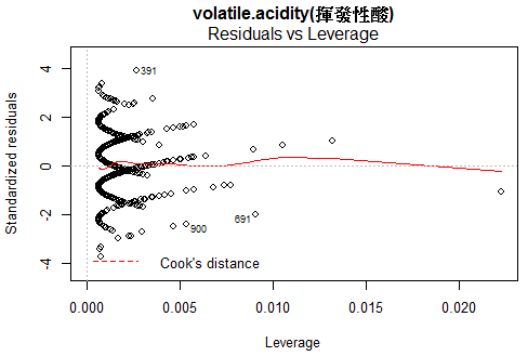

fixed.acidity(固定的酸度)
Scale-Location
√|Standardized residuals|
Fitted values

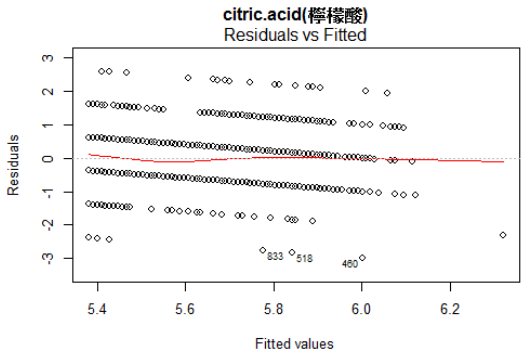

fixed.acidity(固定的酸度)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

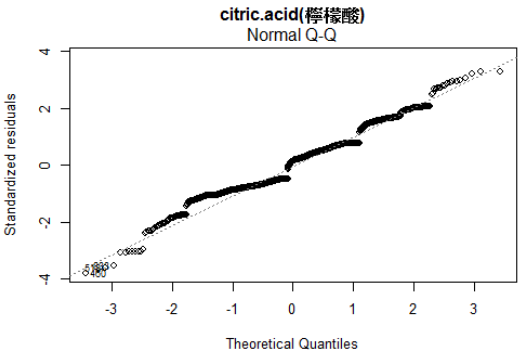

volatile.acidity(挥发性酸)
Residuals vs Fitted
Residuals
Fitted values

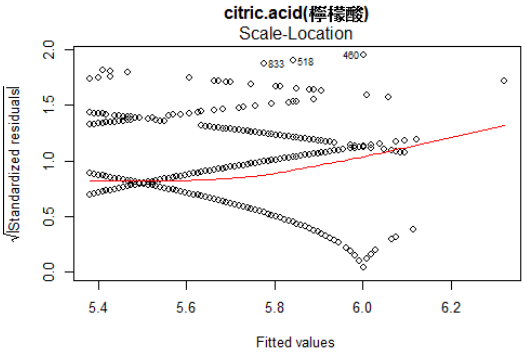

volatile.acidity(挥发性酸)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

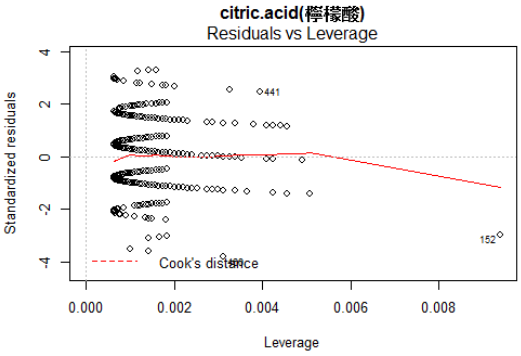

volatile.acidity(挥发性酸)
Scale-Location
√|Standardized residuals|
Fitted values

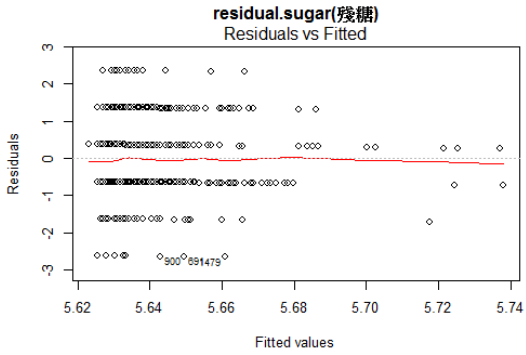

volatile.acidity(挥发性酸)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

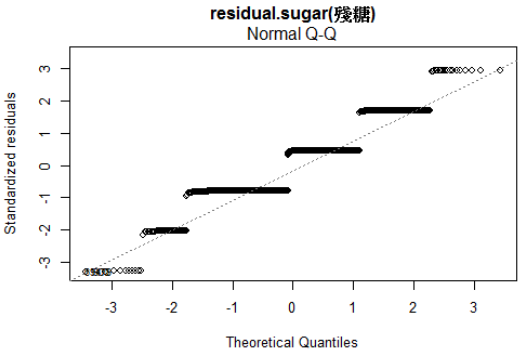

citric.acid(檸檬酸)
Residuals vs Fitted
Residuals
Fitted values
833
518
460

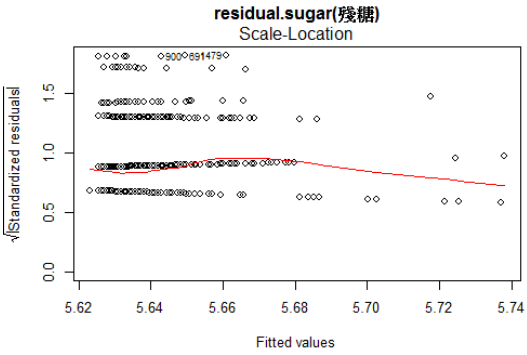

citric.acid(檸檬酸)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

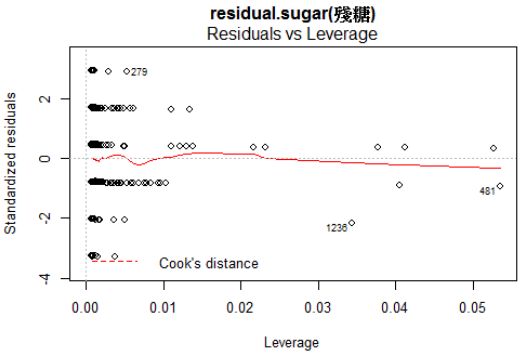

citric.acid(檸檬酸)
Scale-Location
√|Standardized residuals|
Fitted values
833
518
460

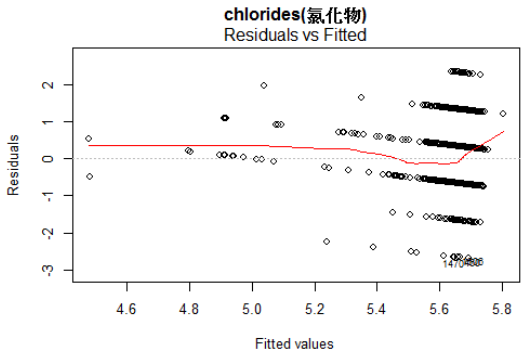

citric.acid(檸檬酸)
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
441
152

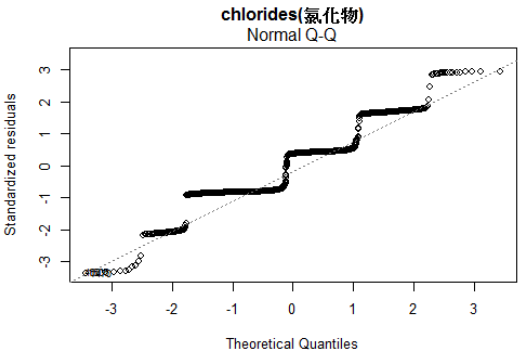

residual.sugar(殘糖)
Residuals vs Fitted
Residuals
Fitted values
900
691479

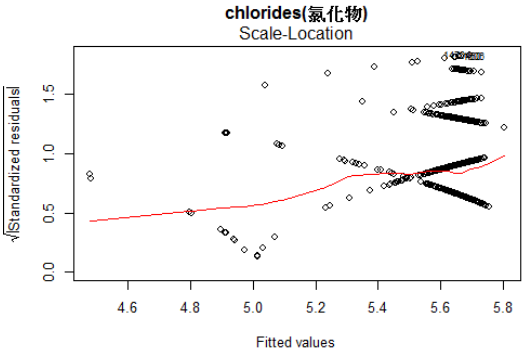

residual.sugar(殘糖)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

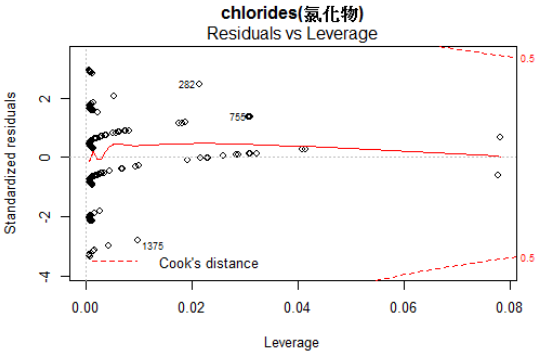

residual.sugar(殘糖)
Scale-Location
√|Standardized residuals|
Fitted values
900
691479

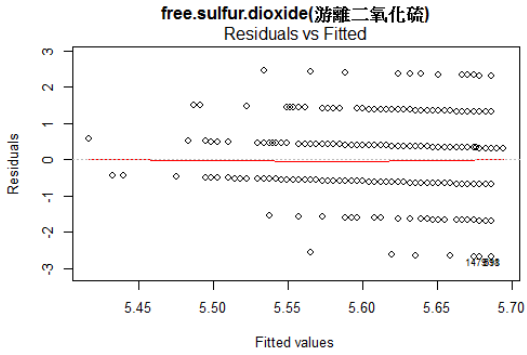

residual.sugar(殘糖)
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
279
481
1236

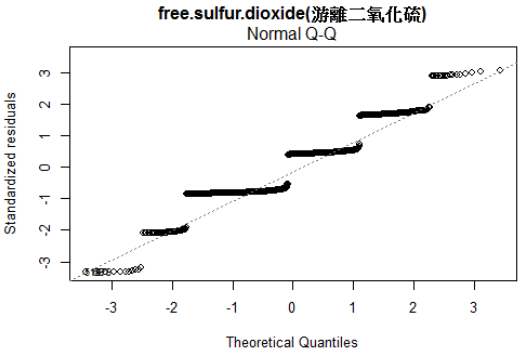

chlorides(氯化物)
Residuals vs Fitted
Residuals
Fitted values

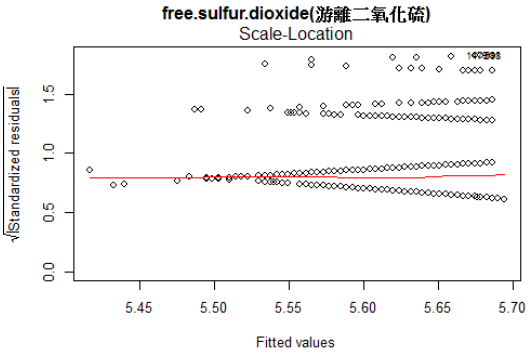

chlorides(氯化物)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

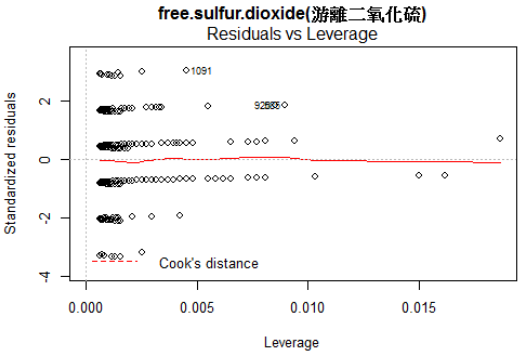

chlorides(氯化物)
Scale-Location
√|Standardized residuals|
Fitted values

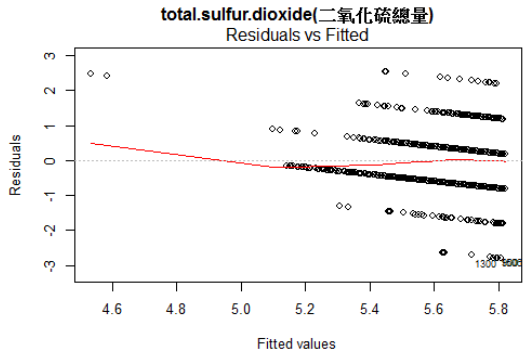

chlorides(氯化物)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

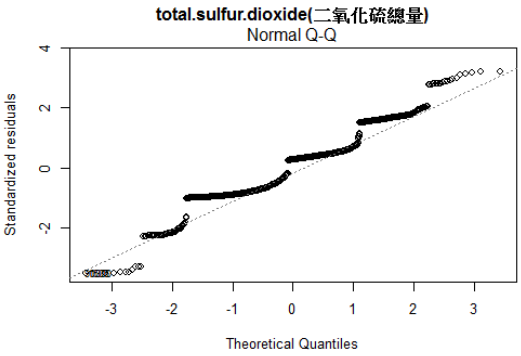

free.sulfur.dioxide(游離二氧化硫)
Residuals vs Fitted
Residuals
Fitted values

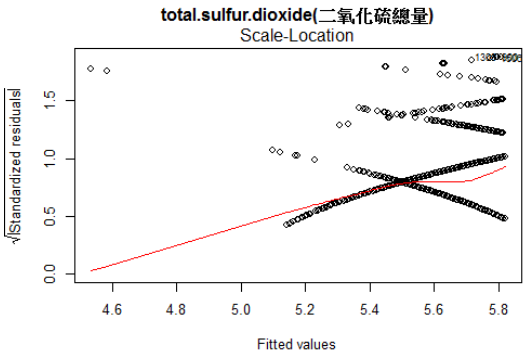

free.sulfur.dioxide(游離二氧化硫)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

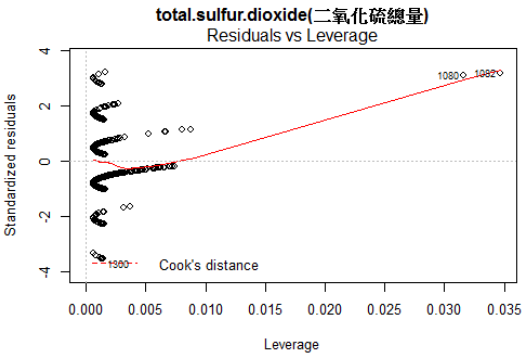

free.sulfur.dioxide(游離二氧化硫)
Scale-Location
√|Standardized residuals|
Fitted values

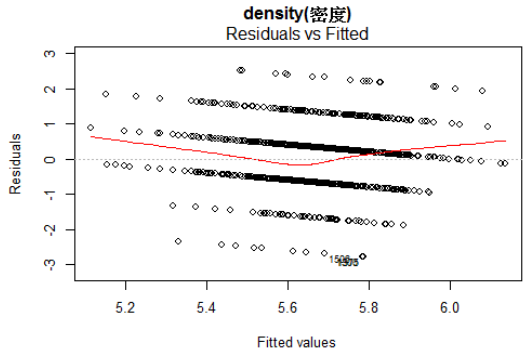

free.sulfur.dioxide(游離二氧化硫)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

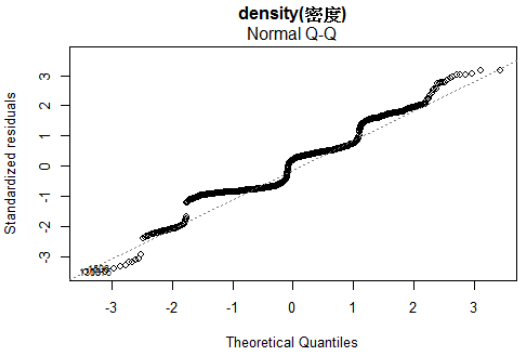

total.sulfur.dioxide(二氧化硫總量)
Residuals vs Fitted
Residuals
Fitted values

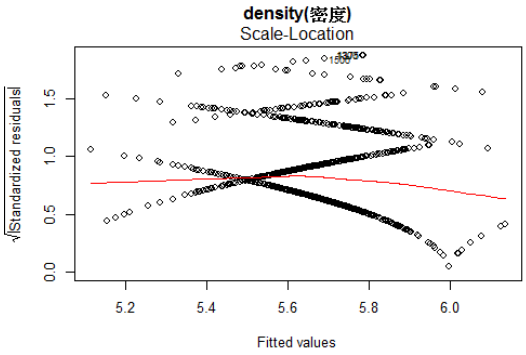

total.sulfur.dioxide(二氧化硫總量)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

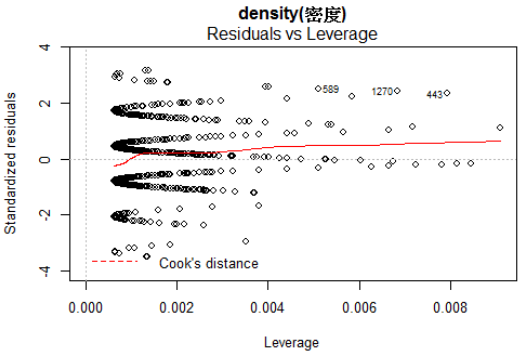

total.sulfur.dioxide(二氧化硫總量)
Scale-Location
√|Standardized residuals|
Fitted values

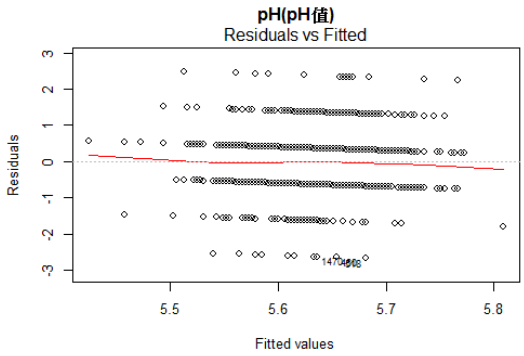

total.sulfur.dioxide(二氧化硫總量)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

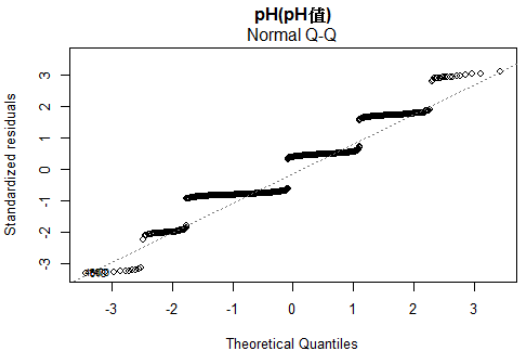

density(密度)
Residuals vs Fitted
Residuals
Fitted values

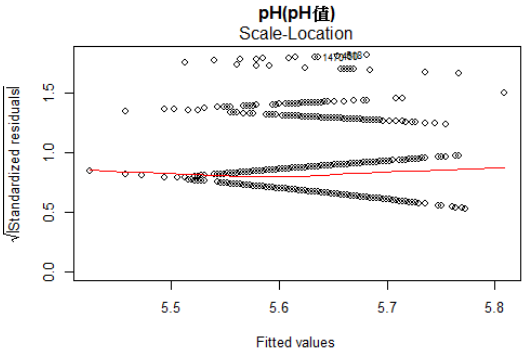

density(密度)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

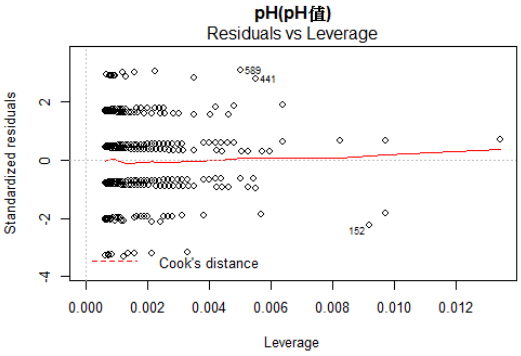

density(密度)
Scale-Location
√|Standardized residuals|
Fitted values

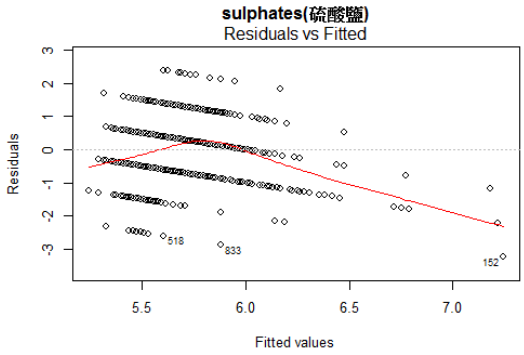

density(密度)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

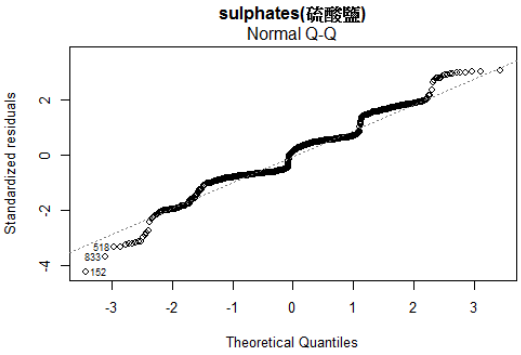

pH(pH值)
Residuals vs Fitted
Residuals
Fitted values

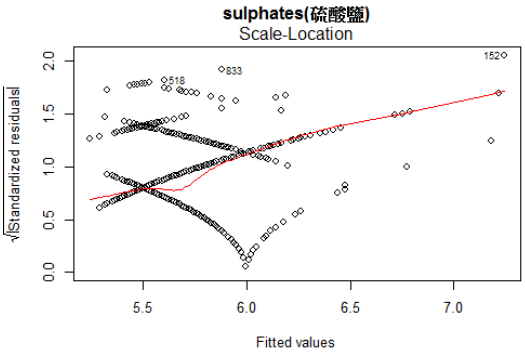

pH(pH值)
Normal Q-Q
Standardized residuals
Theoretical Quantiles

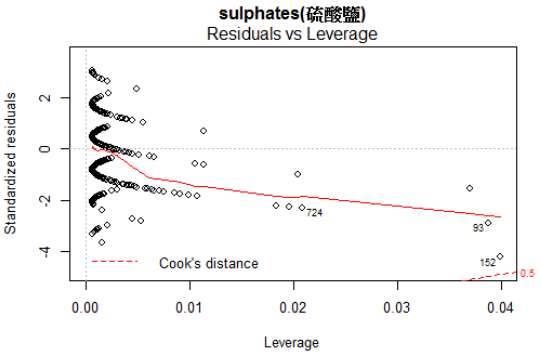

pH(pH值)
Scale-Location
√|Standardized residuals|
Fitted values

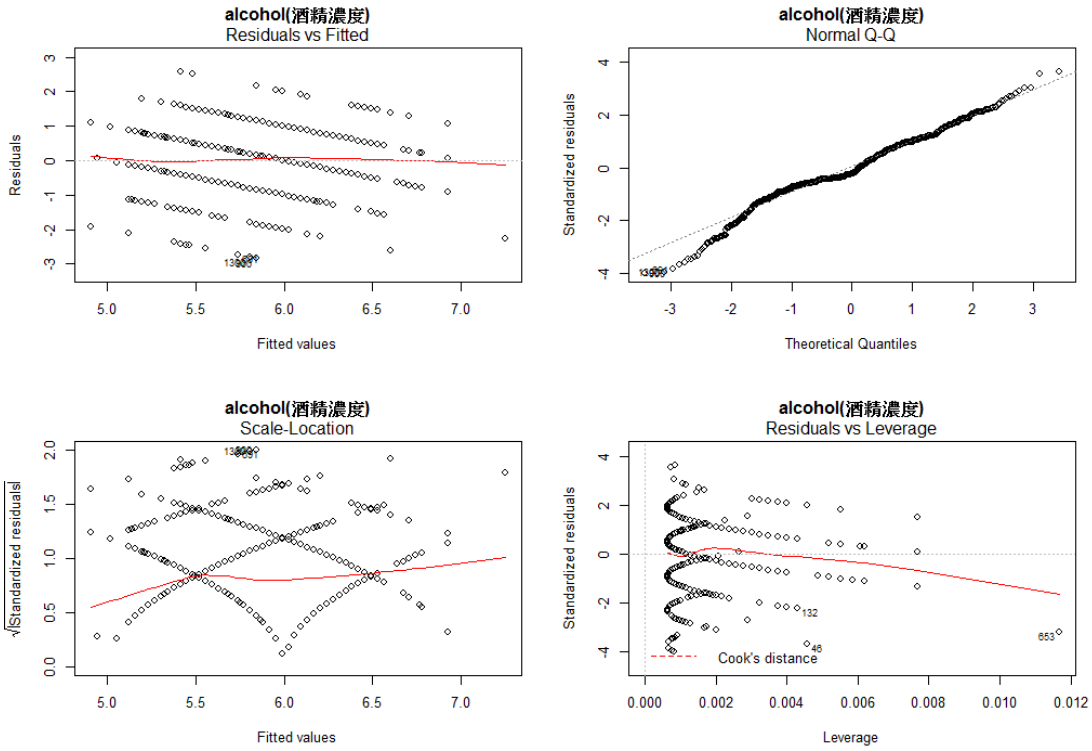

pH(pH值)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


sulphates(硫酸鹽)
Residuals vs Fitted
Residuals
Fitted values


sulphates(硫酸鹽)
Normal Q-Q
Standardized residuals
Theoretical Quantiles


sulphates(硫酸鹽)
Scale-Location
√|Standardized residuals|
Fitted values


sulphates(硫酸鹽)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



【說明】

從"殘差"(Residuals)和"配適值"Fitted values"來看，必須要儘量符合隨機分配。Fixed.acidity(固定的酸度)、volatile.acidity(揮發性酸)、citric.acid(檸檬酸)、residual.sugar(殘糖)、free.sulfur.dioxide(游離二氧化硫)、density(密度)、pH(pH 值)以及 alcohol(酒精濃度)皆符合隨機分配 Residuals=0 的水平線上下分配。

從『Q-Q Plot』來看，volatile.acidity(揮發性酸)的"殘差"最符合 Normal Distribution。Alcohol(酒精濃度)次之，倘若去除離群值，也會很符合 Normal Distribution。Fixed.acidity(固定的酸度)、citric.acid(檸檬酸)、density(密度)、total.sulfur.dioxide(二氧化硫總量) 以及 sulphates(硫酸鹽)更次之。

表 4-3 : 11 個解釋變數的繪圖整理

序號	解釋變數	變數說明	β_0 test	β_1 test	Q-Q Plot	殘差與配適值
1	fixed.acidity	固定的酸度				◎
2	volatile.acidity	揮發性酸			◎	◎
3	citric.acid	檸檬酸				◎
4	residual.sugar	殘糖		X		◎
5	chlorides	氯化物				
6	free.sulfur.dioxide	游離二氧化硫				◎
7	total.sulfur.dioxide	二氧化硫總量				
8	density	密度				◎
9	pH	pH 值				◎

10	sulphates	硫酸鹽				
11	alcohol	酒精濃度			◎	◎

4-5 預測

本節透過以下一組 11 個新值對應這 11 個解釋變數，套入此迴歸模型來看看配適出來的新值為何?

{ 6.9, 0.680, 0.12, 1.75, 0.030, 6, 13, 0.9912, 3.48, 0.42, 12.5 }

【R code】透過此迴歸模型來 fit 新值

```
#【5.predict 新值】----
fit <- numeric(0)
new.X <- c(6.9, 0.680, 0.12, 1.75, 0.030, 6, 13, 0.9912, 3.48, 0.42, 12.5)
for (i in 1:11){
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   new.dataframe <- data.frame(new.X[i])
   names(new.dataframe) <- names(red.wine)[i]
   p <- predict.lm(lm.model, new.dataframe, interval="prediction")
   fit <- c(fit, p[1,"fit"])
}
names(fit) <- X[1:11]
fit
```

【執行結果】

```
         fixed.acidity(固定的酸度)
                          5.554339
       volatile.acidity(揮發性酸)
                          5.367968
            citric.acid(檸檬酸)
                          5.494339
          residual.sugar(殘糖)
                          5.629818
             chlorides(氯化物)
                          5.763129
 free.sulfur.dioxide(游離二氧化硫)
                          5.674642
total.sulfur.dioxide(二氧化硫總量)
                          5.788105
                density(密度)
                          6.051169
                          pH(pH 值)
```

```
                                  5.585022
                sulphates(硫酸鹽)
                                  5.350789
                alcohol(酒精濃度)
                                  6.385497
```

4-6 整體性的分析，針對 11 個解釋變數的 ANOVA 變異數分析

F 檢定和 t 檢定的等價性($F = t^2$)，此章將驗證是否與前面的 t-test 是否相同的結果。以下將針對以下假設來檢定。

$$\begin{cases} H_{0:}: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$$

【R code】

```
#【6.ANOVA 變異數分析】----
anova <- data.frame()
for (i in 1:11) {
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   anova <- rbind(anova,data.frame(anova(lm.model)[1,"F
value"],anova(lm.model)[1,"Pr(>F)"]))
}
rownames(anova) <- X[1:11]
colnames(anova) <- c("F.value","p.value")
anova
```

【執行結果】

```
                                    F.value      p.value
fixed.acidity(固定的酸度)          24.9600375 6.495635e-07
volatile.acidity(揮發性酸)        287.4444497 2.051715e-59
citric.acid(檸檬酸)                86.2577262 4.991295e-20
residual.sugar(殘糖)                0.3011837 5.832180e-01
chlorides(氯化物)                  26.9856084 2.313383e-07
free.sulfur.dioxide(游離二氧化硫)   4.1085023 4.283398e-02
total.sulfur.dioxide(二氧化硫總量) 56.6578176 8.621703e-14
density(密度)                      50.4052231 1.874957e-12
pH(pH 值)                           5.3404622 2.096278e-02
sulphates(硫酸鹽)                 107.7404330 1.802088e-24
alcohol(酒精濃度)                 468.2670106 2.831477e-91
```

【說明】

從以上的結果，residual.sugar(殘糖)再次被淘汰，就如同前面的 t-test 的結果相同。於是我們再將結果加入表 4。

表 4-4 : 11 個解釋變數 ANOVA 的 F 檢定

序號	解釋變數	變數說明	β_0 test	β_1 test	Q-Q Plot	殘差與配適值	F test
1	fixed.acidity	固定的酸度				◎	
2	volatile.acidity	揮發性酸			◎	◎	◎
3	citric.acid	檸檬酸				◎	
4	residual.sugar	殘糖		X		◎	X
5	chlorides	氯化物					
6	free.sulfur.dioxide	游離二氧化硫				◎	
7	total.sulfur.dioxide	二氧化硫總量					
8	density	密度				◎	
9	pH	pH 值				◎	
10	sulphates	硫酸鹽					◎
11	alcohol	酒精濃度			◎	◎	◎

4-7 R-Squared 的分析

R-Squared 對於模型的貢獻在於線性迴歸的解釋能力，所以針對 11 個解釋變數做一個比較。

【R code】

```
#【7.R-Squared & Adjusted R-Squared 分析】----
R.square <- data.frame()
for (i in 1:11) {
  f <- formula(paste0("quality~",names(red.wine)[i]))
  lm.model <- lm(f, data=red.wine)
  R.square <- R.square <- rbind(R.square,
                                data.frame(summary(lm.model)$r.squared,
                                           summary(lm.model)$adj.r.squared))
}
rownames(R.square) <- X[1:11]
colnames(R.square) <- c("R.squared","R.Adj.squared")
R.square
```

【執行結果】

```
                                    R.squared R.Adj.squared
fixed.acidity(固定的酸度)          0.0153888116  0.0147722736
volatile.acidity(揮發性酸)         0.1525353797  0.1520047193
citric.acid(檸檬酸)                0.0512445152  0.0506504291
residual.sugar(殘糖)               0.0001885579 -0.0004374981
chlorides(氯化物)                  0.0166169012  0.0160011322
free.sulfur.dioxide(游離二氧化硫)  0.0025660361  0.0019414688
```

total.sulfur.dioxide(二氧化硫總量) 0.0342621170 0.0336573969
density(密度) 0.0305967362 0.0299897211
~~pH(pH 值) 0.0033329135 0.0027088264~~
sulphates(硫酸鹽) 0.0632004914 0.0626138918
alcohol(酒精濃度) 0.2267343681 0.2262501692

【說明】

根據以上 11 個解釋變數的 R-squared & Adjusted R-squared 的表現來看，可以將表現太差的，也就是對整體解釋能力太低(未達到 1%)的三個解釋變數去除，包括 residual.sugar(殘糖)、free.sulfur.dioxide(游離二氧化硫)以及 pH(pH 值)。相對地，如果取 0.05 以上，volatile.acidity(揮發性酸)、sulphates(硫酸鹽)與 alcohol(酒精濃度)在此處的表現算是最好的三個，所以將其加入表 5。

表 4-5 : 11 個解釋變數的 R-squared

序號	解釋變數	變數說明	β_0 test	β_1 test	Q-Q Plot	殘差與配適值	F test	R squared
1	fixed.acidity	固定的酸度				◎		
2	volatile.acidity	揮發性酸			◎	◎	◎	◎
3	citric.acid	檸檬酸				◎		
4	residual.sugar	殘糖		X		◎	X	X
5	chlorides	氯化物						
6	free.sulfur.dioxide	游離二氧化硫				◎		X
7	total.sulfur.dioxide	二氧化硫總量						
8	density	密度				◎		
9	pH	pH 值				◎		X
10	sulphates	硫酸鹽					◎	◎
11	alcohol	酒精濃度			◎	◎	◎	◎

4-8 初步變數分析結論

針對前面所做的資料分析匯整表 6，從表 6 中可以非常清楚看出 "residual.sugar"(殘糖) 這個解釋變數幾乎是可以被淘汰不納入迴歸分析的考慮因素之一。反之，表現最好的包括以下三個：

- ✓ volatile.acidity (揮發性酸)
- ✓ alcohol (酒精濃度)
- ✓ sulphates (硫酸鹽)

表 4-6 : 11 個解釋變數的檢定總整理

序號	解釋變數	變數說明	β_0 test	β_1 test	Q-Q Plot	殘差與配適值	F test	R squared
1	fixed.acidity	固定的酸度				◎		

2	volatile.acidity	揮發性酸			◎	◎	◎	◎
3	citric.acid	檸檬酸				◎		
4	**residual.sugar**	**殘糖**		**X**		**◎**	**X**	**X**
5	chlorides	氯化物						
6	free.sulfur.dioxide	游離二氧化硫				◎		X
7	total.sulfur.dioxide	二氧化硫總量						
8	density	密度				◎		
9	pH	pH 值				◎		X
10	sulphates	硫酸鹽					◎	◎
11	alcohol	酒精濃度			◎	◎	◎	◎

五、模型診斷與矯正之測量

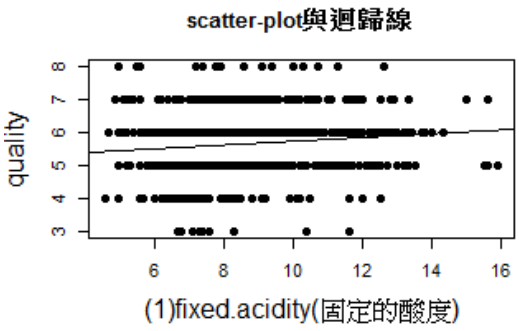
本章會先從殘差(Residuals)的圖型分佈診斷、無母數圖型分析(Levene)和 Brown-Forsythe、Breusch-Pagan 檢定，最後再透過解釋變數的轉換和反應變數的轉換(Box-Cox)來調整模型。

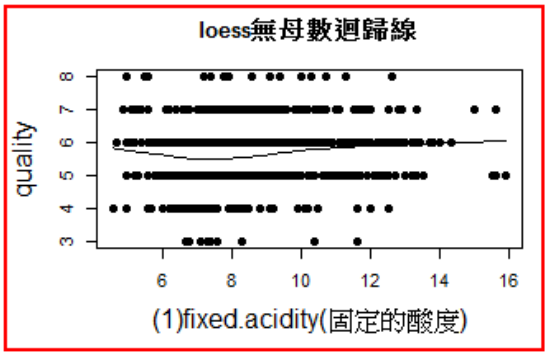
5-1 迴歸線與無母數分析圖

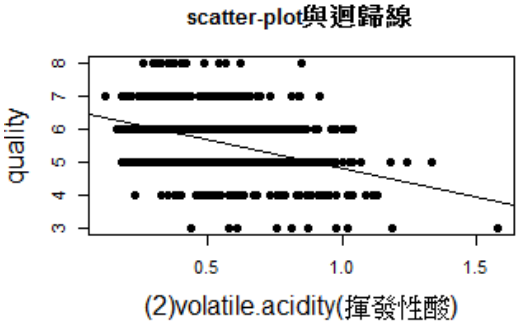
【R code】

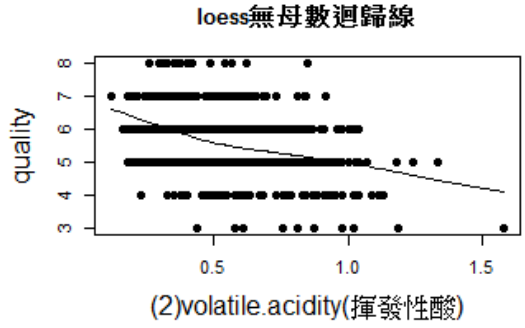
```
#【5-1.繪製散佈圖與迴歸線&無母數(lowess or loess)分析】----
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
mod <- character()
cnt <- 0
for (i in choice.X) {
   if (cnt%%3==0) {
      dev.new()
      par(mfrow=c(3,2))
   }
   cnt <- cnt + 1
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   plot(red.wine[,i],main="scatter-plot 與迴歸線",
         red.wine$quality,cex.lab=1.5,
         xlab=paste0("(",i,")",X[i]),ylab="quality",pch=19)
   abline(lm.model)
   plot(red.wine[,i],main="loess 無母數迴歸線",
         red.wine$quality,cex.lab=1.5,
         xlab=paste0("(",i,")",X[i]),ylab="quality",pch=19)
   lines(lowess(red.wine[,i],red.wine$quality))
}
par(mfrow=c(1,1))
```

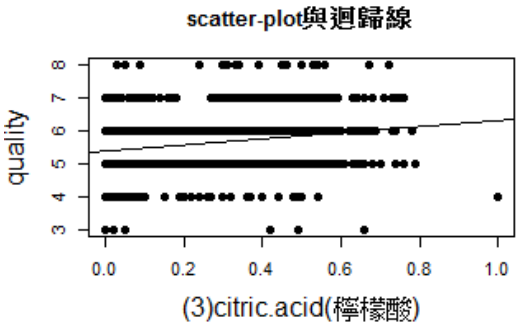
【執行結果】

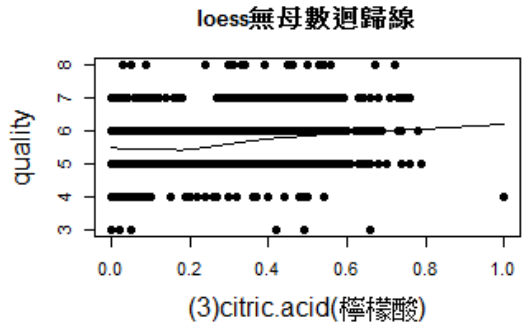

scatter-plot與迴歸線
quality
3 4 5 6 7 8
6 8 10 12 14 16
(1)fixed.acidity(固定的酸度)

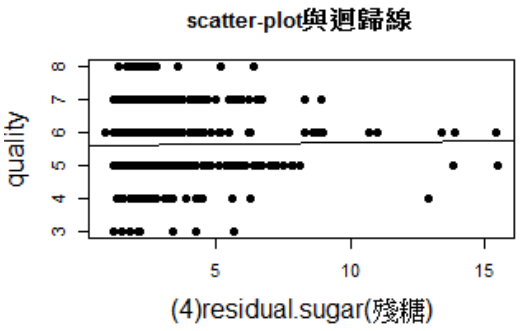

loess無母數迴歸線
quality
3 4 5 6 7 8
6 8 10 12 14 16
(1)fixed.acidity(固定的酸度)

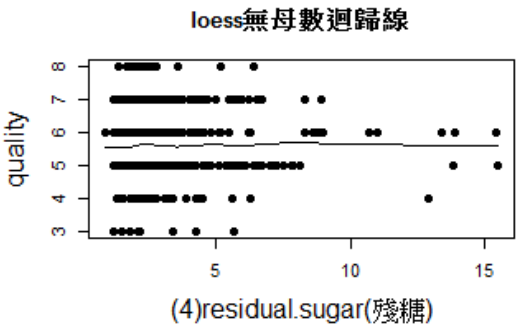

scatter-plot與迴歸線
quality
3 4 5 6 7 8
0.5 1.0 1.5
(2)volatile.acidity(揮發性酸)

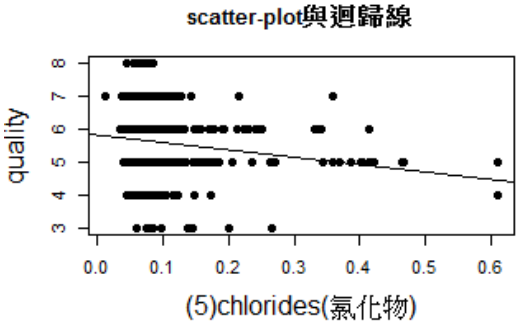

loess無母數迴歸線
quality
3 4 5 6 7 8
0.5 1.0 1.5
(2)volatile.acidity(揮發性酸)

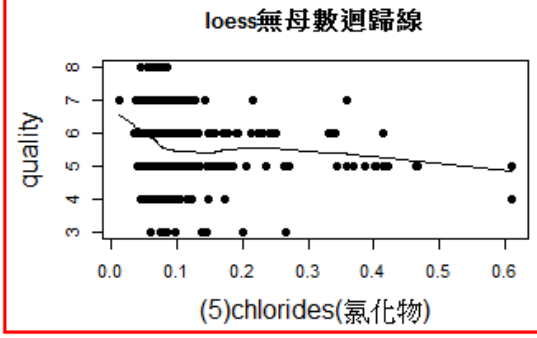

scatter-plot與迴歸線
quality
3 4 5 6 7 8
0.0 0.2 0.4 0.6 0.8 1.0
(3)citric.acid(檸檬酸)

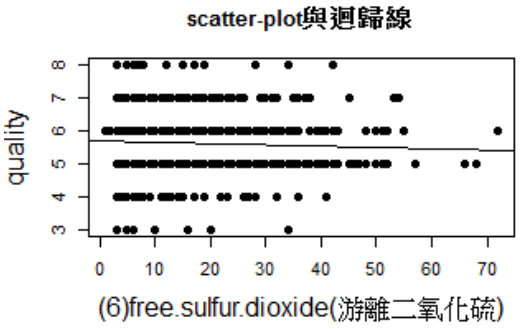

loess無母數迴歸線
quality
3 4 5 6 7 8
0.0 0.2 0.4 0.6 0.8 1.0
(3)citric.acid(檸檬酸)

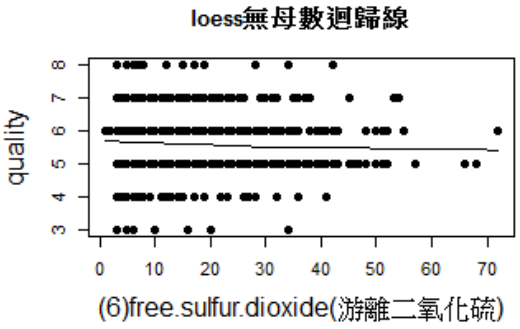

scatter-plot與迴歸線
quality
(4)residual.sugar(殘糖)

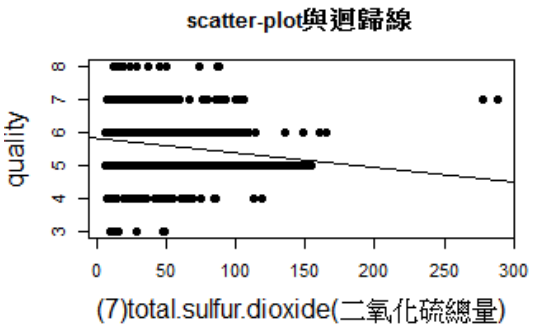

loess無母數迴歸線
quality
(4)residual.sugar(殘糖)

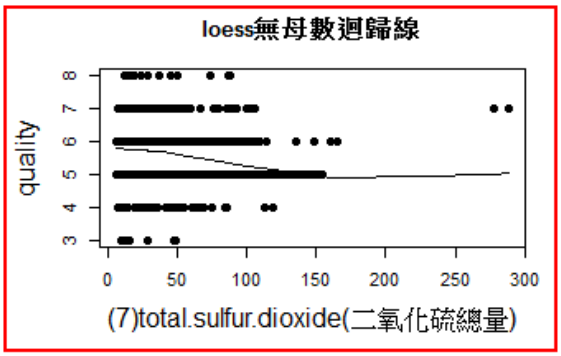

scatter-plot與迴歸線
quality
(5)chlorides(氯化物)

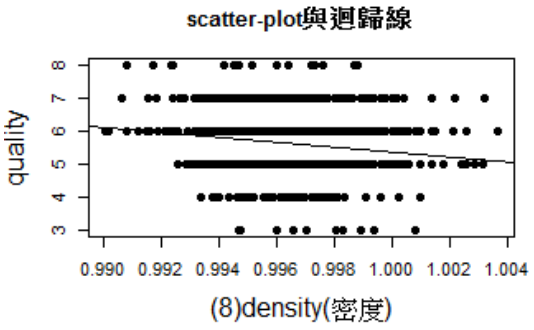

loess無母數迴歸線
quality
(5)chlorides(氯化物)

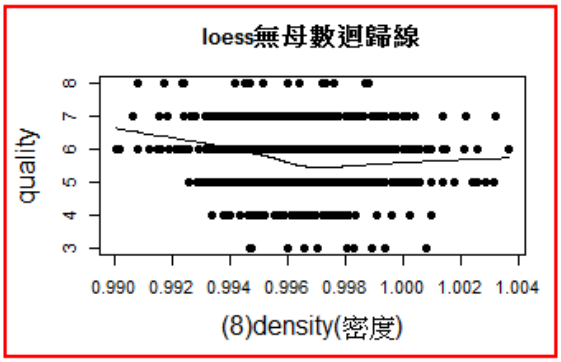

scatter-plot與迴歸線
quality
(6)free.sulfur.dioxide(游離二氧化硫)

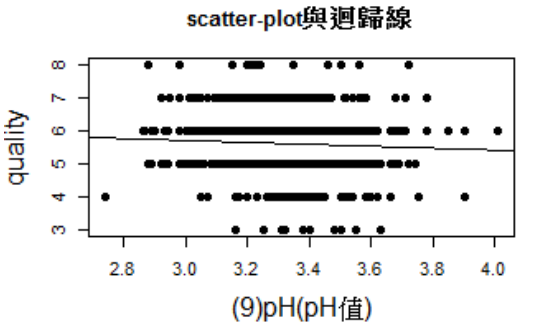

loess無母數迴歸線
quality
(6)free.sulfur.dioxide(游離二氧化硫)

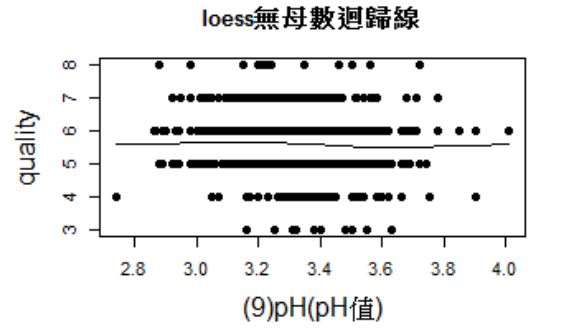

scatter-plot與迴歸線
quality
(7)total.sulfur.dioxide(二氧化硫總量)

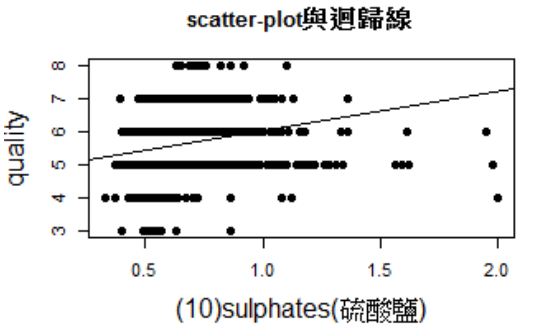

loess無母數迴歸線
quality
(7)total.sulfur.dioxide(二氧化硫總量)

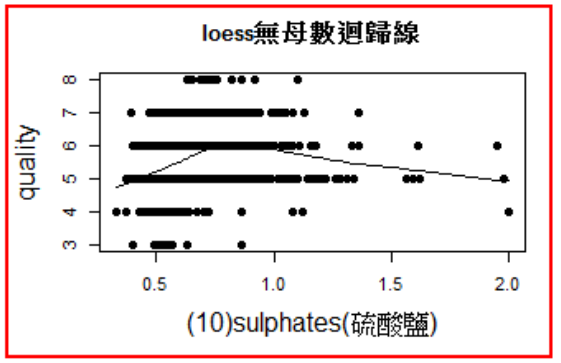

scatter-plot與迴歸線
quality
(8)density(密度)

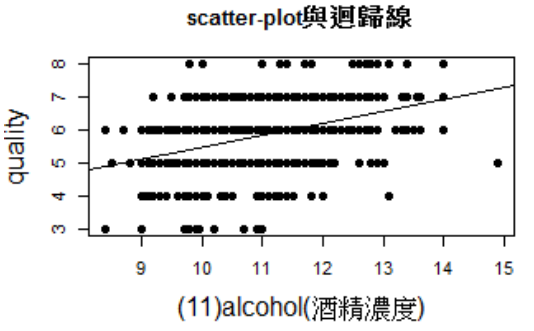

loess無母數迴歸線
quality
(8)density(密度)

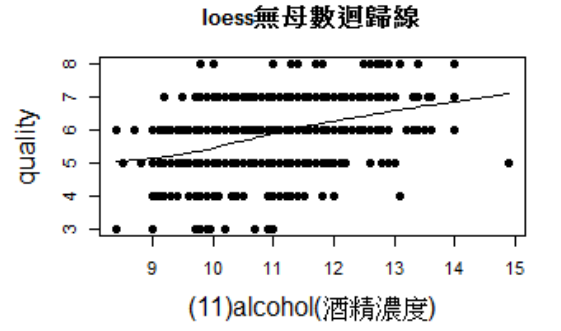

scatter-plot與迴歸線
quality
(9)pH(pH值)


loess無母數迴歸線
quality
(9)pH(pH值)


scatter-plot與迴歸線
quality
(10)sulphates(硫酸鹽)


loess無母數迴歸線
quality
(10)sulphates(硫酸鹽)


scatter-plot與迴歸線
quality
(11)alcohol(酒精濃度)

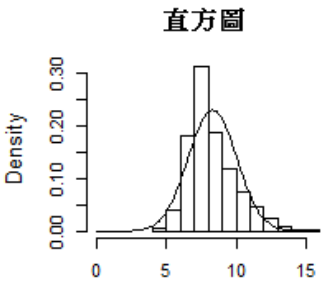

loess無母數迴歸線
quality
(11)alcohol(酒精濃度)

5-2 繪製 X 的直方圖(常態分配) & 盒鬚圖(box-plot)

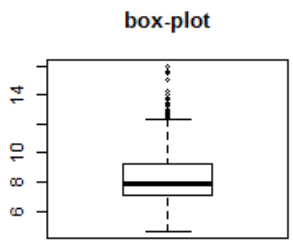
【R code】

```
#【5-2.繪製直方圖(常態分配) & 盒鬚圖(box-plot)分析】----
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
mod <- character()
cnt <- 0
for (i in choice.X) {
   if (cnt%%3==0) {
      dev.new()
      par(mfrow=c(3,3))
   }
   cnt <- cnt + 1
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   hist(red.wine[,i],main="直方圖",freq=FALSE,cex.lab=1.2,
           xlim=c(min(0,red.wine[,i]),max(red.wine[,i])),
           xlab=paste0("(",i,")",X[i])
           )
   curve(dnorm(x,mean=mean(red.wine[,i]),sd=sd(red.wine[,i])),
            add=TRUE)
   boxplot(red.wine[,i],cex.lab=1.2, outline=TRUE,
                main="box-plot", xlab=paste0("(",i,")",X[i]))
   boxplot(red.wine[,i],cex.lab=1.2, outline=FALSE,
                main="box-plot(不顯示 outlier)", xlab=paste0("(",i,")",X[i]))
}
par(mfrow=c(1,1))
```

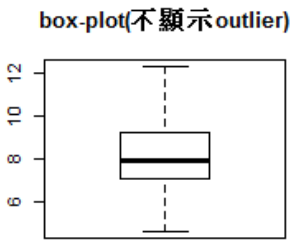
【執行結果】



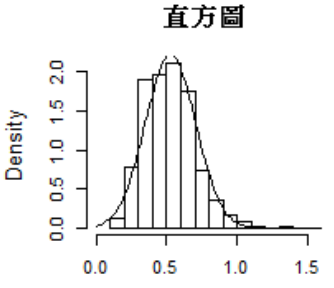
(1)fixed.acidity(固定的酸度)



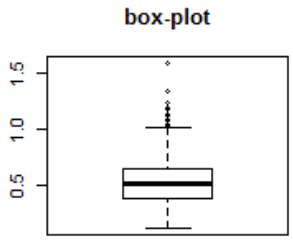
(1)fixed.acidity(固定的酸度)



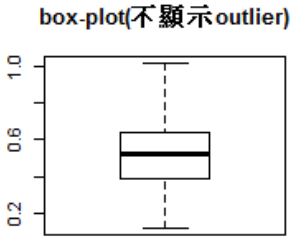
(1)fixed.acidity(固定的酸度)



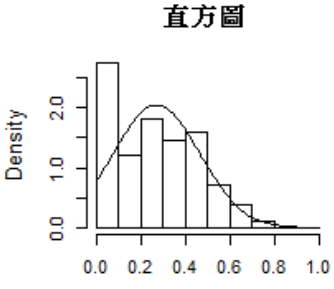
(2)volatile.acidity(揮發性酸)



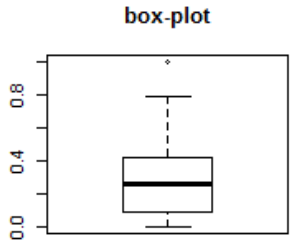
(2)volatile.acidity(揮發性酸)



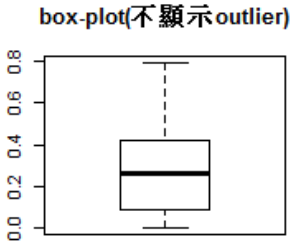
(2)volatile.acidity(揮發性酸)



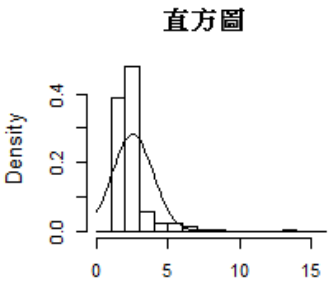
(3)citric.acid(檸檬酸)



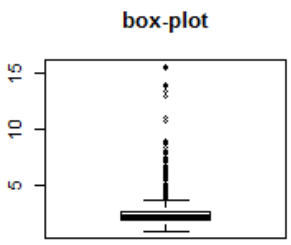
(3)citric.acid(檸檬酸)



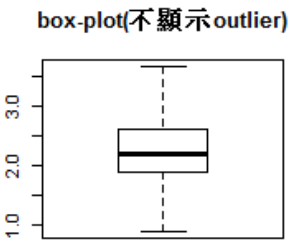
(3)citric.acid(檸檬酸)



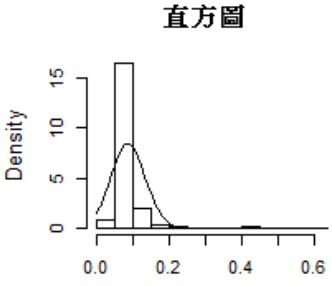
(4)residual.sugar(殘糖)



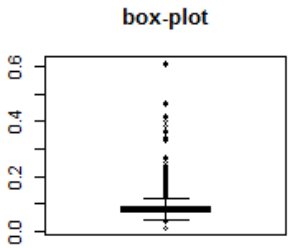
(4)residual.sugar(殘糖)



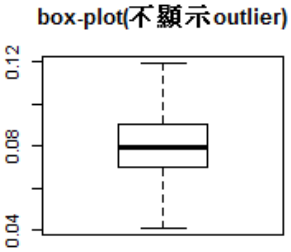
(4)residual.sugar(殘糖)



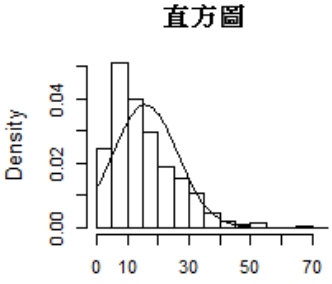
(5)chlorides(氯化物)



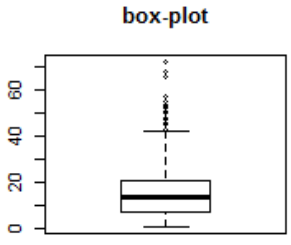
(5)chlorides(氯化物)



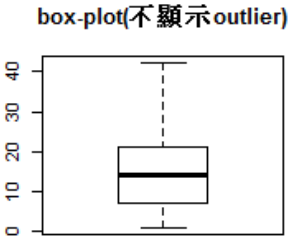
(5)chlorides(氯化物)



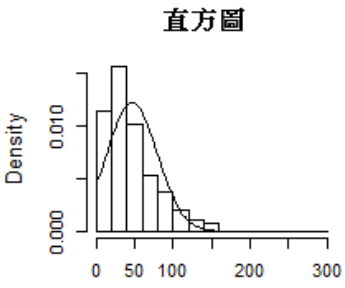
(6)free.sulfur.dioxide(游離二氧化硫



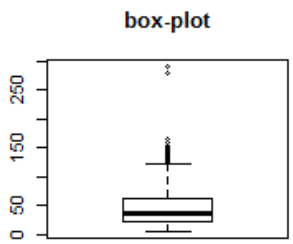
(6)free.sulfur.dioxide(游離二氧化硫



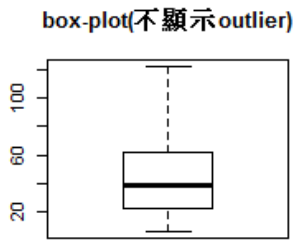
(6)free.sulfur.dioxide(游離二氧化硫



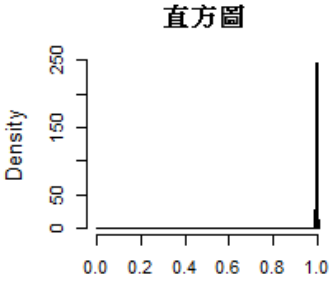
(7)total.sulfur.dioxide(二氧化硫總量



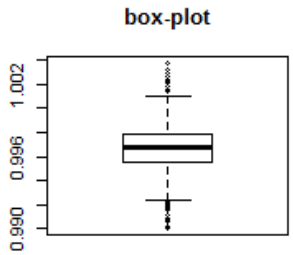
(7)total.sulfur.dioxide(二氧化硫總量



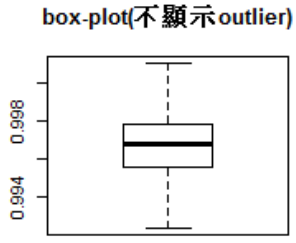
(7)total.sulfur.dioxide(二氧化硫總量



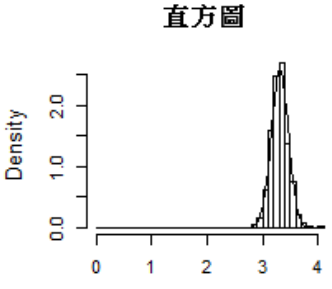
(8)density(密度)



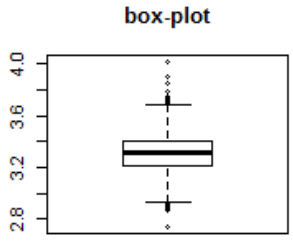
(8)density(密度)



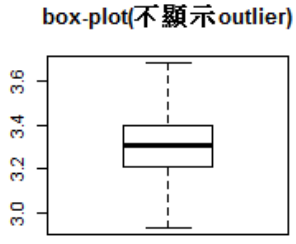
(8)density(密度)



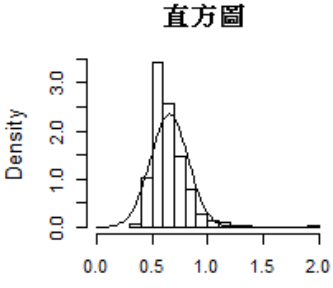
(9)pH(pH值)



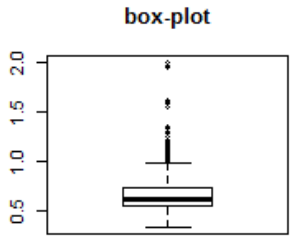
(9)pH(pH值)



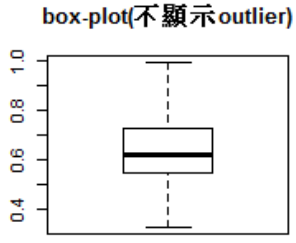
(9)pH(pH值)



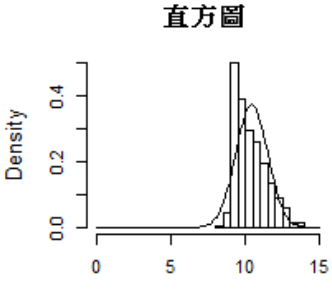
(10)sulphates(硫酸鹽)



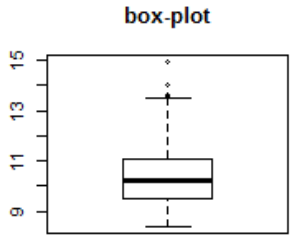
(10)sulphates(硫酸鹽)



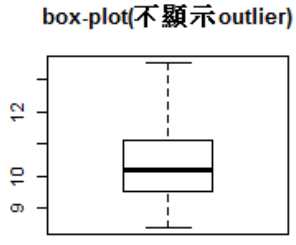
(10)sulphates(硫酸鹽)



(11)alcohol(酒精濃度)



(11)alcohol(酒精濃度)



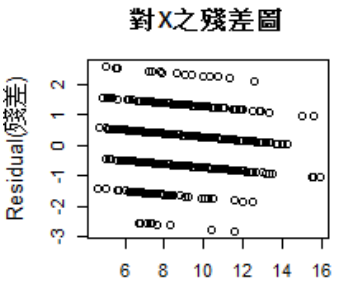
(11)alcohol(酒精濃度)

5-3 殘差對 X 的散佈圖，以及殘差的常態分配圖(boxplot & Q-Q plot)

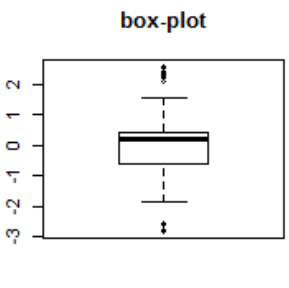
【R code】

```
#【5-3.殘差分析】----
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
mod <- character()
cnt <- 0
for (i in choice.X) {
   if (cnt%%3==0) {
      dev.new()
      par(mfrow=c(3,3))
   }
   cnt <- cnt + 1
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   plot(red.wine[,i],resid(lm.model),cex.lab=1.2,
         xlab=paste0("(",i,")",X[i]),ylab="Residual(殘差)",
         main="對 X 之殘差圖")
   boxplot(resid(lm.model),cex.lab=1.2, outline=TRUE,
            main="box-plot", xlab=paste0("(",i,")",X[i]))
   qqnorm(resid(lm.model))
   qqline(resid(lm.model),col="red")
}
par(mfrow=c(1,1))
```

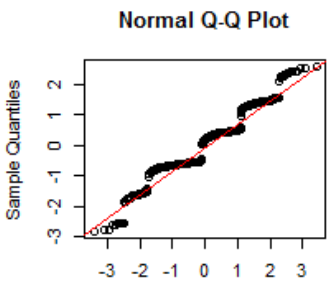
【執行結果】

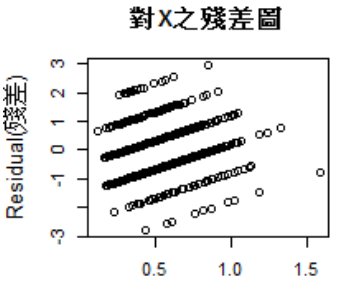


(1)fixed.acidity(固定的酸度)

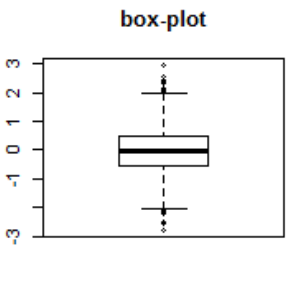


(1)fixed.acidity(固定的酸度)

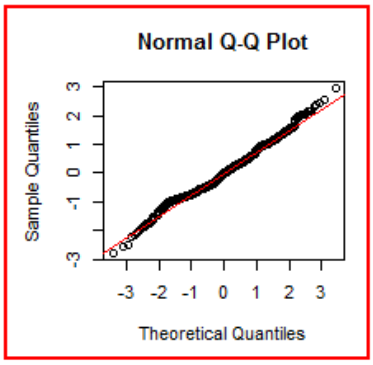


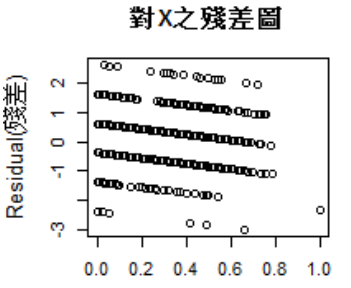


(2)volatile.acidity(揮發性酸)

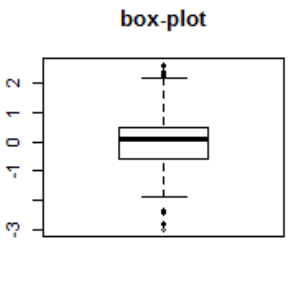


(2)volatile.acidity(揮發性酸)

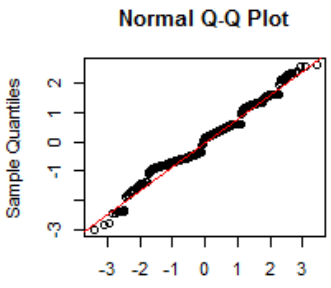


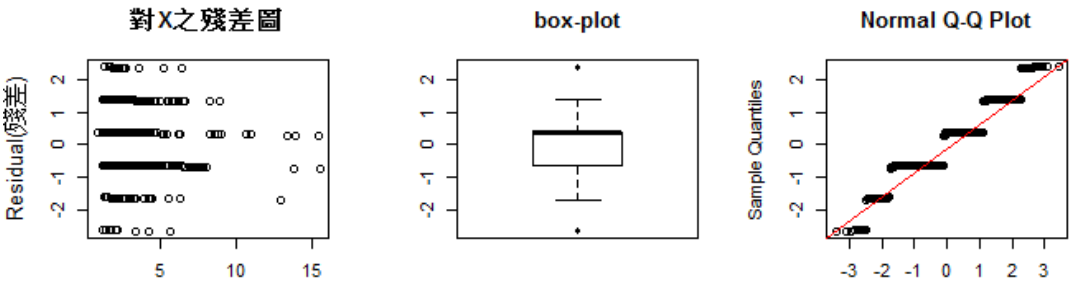


(3)citric.acid(檸檬酸)



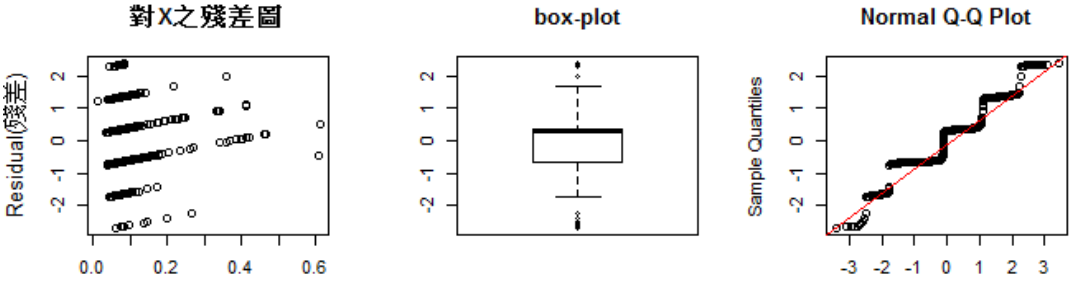
(3)citric.acid(檸檬酸)




對X之殘差圖
Residual(殘差)
5
10
15
(4)residual.sugar(殘糖)

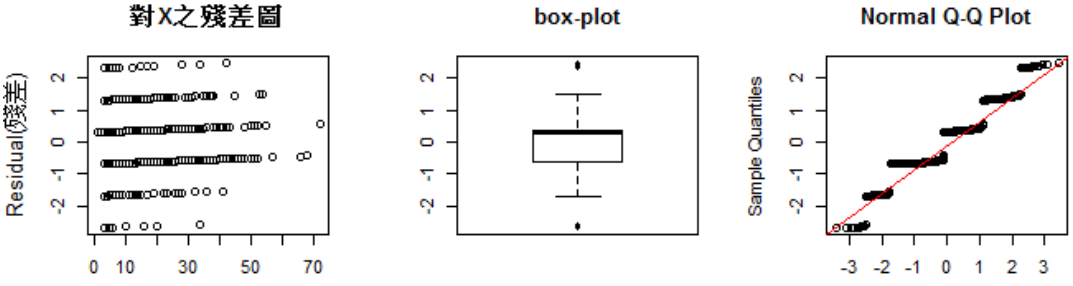
box-plot
(4)residual.sugar(殘糖)

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


對X之殘差圖
Residual(殘差)
0.0
0.2
0.4
0.6
(5)chlorides(氯化物)

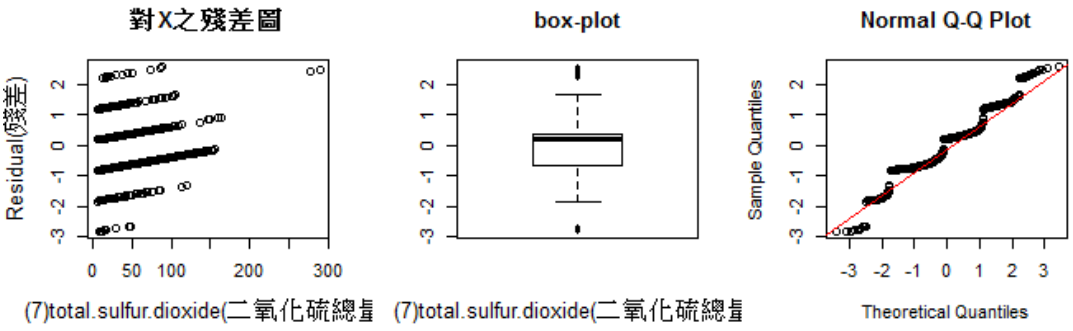
box-plot
(5)chlorides(氯化物)

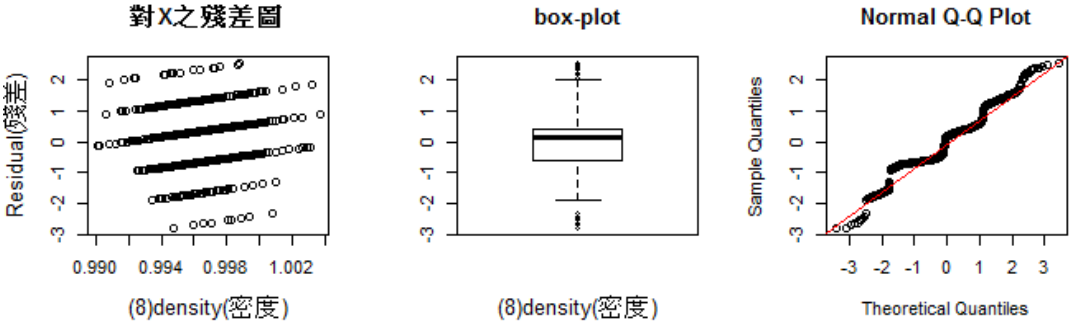
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

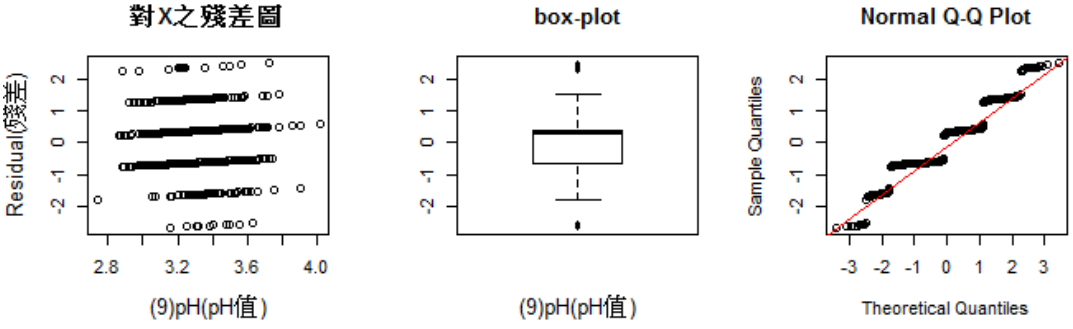

對X之殘差圖
Residual(殘差)
0
10
30
50
70
(6)free.sulfur.dioxide(游離二氧化硫

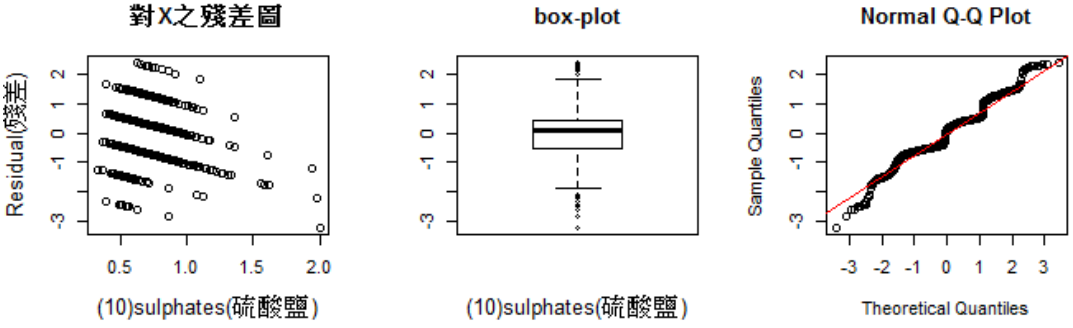
box-plot
(6)free.sulfur.dioxide(游離二氧化硫

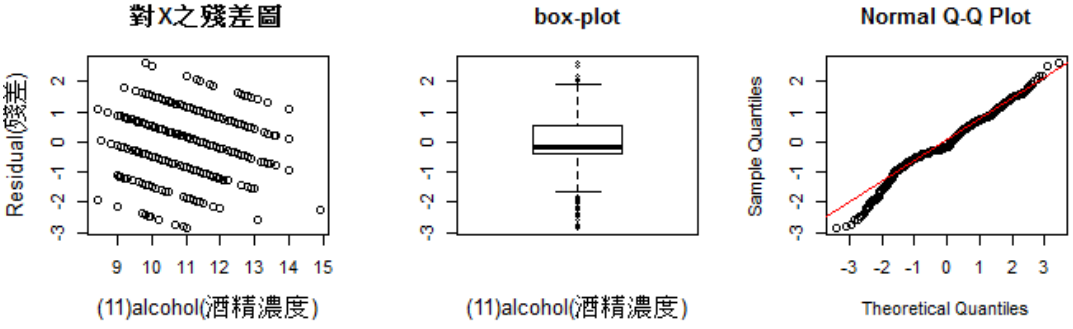
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


對X之殘差圖
Residual(殘差)
(7)total.sulfur.dioxide(二氧化硫總量
box-plot
(7)total.sulfur.dioxide(二氧化硫總量
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


對X之殘差圖
Residual(殘差)
(8)density(密度)
box-plot
(8)density(密度)
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


對X之殘差圖
Residual(殘差)
(9)pH(pH值)
box-plot
(9)pH(pH值)
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


對X之殘差圖
Residual(殘差)
(10)sulphates(硫酸鹽)
box-plot
(10)sulphates(硫酸鹽)
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


對X之殘差圖
Residual(殘差)
(11)alcohol(酒精濃度)
box-plot
(11)alcohol(酒精濃度)
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

5-4 Brown-Forsythe 與 Breusch-Pagan 檢定

【R code】

```
#【5-4-1.針對 residual 進行  Brown-Forsythe Test】----
install.packages("lawstat",repos = "http://cran.csie.ntu.edu.tw/")
library(lawstat)
# Brown-Forsythe Test  是採用  location="median"
# Levene Test  是採用  location="mean"
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
result <- data.frame()
p.value <- numeric(0)
for (i in choice.X) {
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   bf <- levene.test(resid(lm.model), red.wine[,i], location="median",
correction.method="zero.correction")
   p.value = c(p.value,bf$p.value)
}
result <- data.frame(X.variable=X,bf.p.value=p.value);
result[order(result$bf.p.value),]

#【5-4-2.針對 residual 進行  Breusch-Pagan Test】----
install.packages("lmtest",repos = "http://cran.csie.ntu.edu.tw/")
library(lmtest)
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
p.value <- numeric(0)
for (i in choice.X) {
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   bp <- bptest(lm.model)
   p.value = c(p.value,bp$p.value)
}
result <- cbind(result,bp.p.value=p.value);
result[order(result$bp.p.value),c(1,3)]

#【4-2.挑選 Brown-Forsythe Test 與 Breusch-Pagan Test 皆顯著】
result[result$bf.p.value < 0.05 & result$bp.p.value < 0.05 ,]
```

【執行結果】

	X.variable	bf.p.value

11 alcohol(酒精濃度) 1.251119e-08
10 sulphates(硫酸鹽) 3.045298e-02
7 total.sulfur.dioxide(二氧化硫總量) 8.915088e-02
3 citric.acid(檸檬酸) 1.703422e-01
2 volatile.acidity(揮發性酸) 3.402836e-01
5 chlorides(氯化物) 3.938061e-01
9 pH(pH 值) 5.882744e-01
4 residual.sugar(殘糖) 7.624561e-01
6 free.sulfur.dioxide(游離二氧化硫) 8.659762e-01
1 fixed.acidity(固定的酸度) 9.742309e-01
8 density(密度) 9.973763e-01

X.variable bp.p.value
10 sulphates(硫酸鹽) 7.927499e-19
11 alcohol(酒精濃度) 6.454021e-08
7 total.sulfur.dioxide(二氧化硫總量) 2.515996e-06
6 free.sulfur.dioxide(游離二氧化硫) 2.628626e-04
3 citric.acid(檸檬酸) 9.713534e-03
4 residual.sugar(殘糖) 9.049318e-02
5 chlorides(氯化物) 1.400066e-01
8 density(密度) 1.558287e-01
2 volatile.acidity(揮發性酸) 3.436290e-01
1 fixed.acidity(固定的酸度) 3.526455e-01
9 pH(pH 值) 6.342166e-01

BF & BP 共同顯著項目

X.variable bf.p.value bp.p.value
10 sulphates(硫酸鹽) 3.045298e-02 7.927499e-19
11 alcohol(酒精濃度) 1.251119e-08 6.454021e-08

表 5-1 : Brown-Forsythe Test 與 Breusch-Pagan Test 的結果比較

序號	X.variable	bf.p.value	bp.p.value
1	fixed.acidity(固定的酸度)	9.742309e-01	3.526455e-01
2	volatile.acidity(揮發性酸)	3.402836e-01	3.436290e-01
3	citric.acid(檸檬酸)	1.703422e-01	**9.713534e-03***
4	residual.sugar(殘糖)	7.624561e-01	9.049318e-02
5	chlorides(氯化物)	3.938061e-01	1.400066e-01
6	free.sulfur.dioxide(游離二氧化硫)	8.659762e-01	**2.628626e-04***
7	total.sulfur.dioxide(二氧化硫總量)	8.915088e-02	**2.515996e-06***
8	density(密度)	9.973763e-01	1.558287e-01

9	pH(pH 值)	5.882744e-01	6.342166e-01
10	sulphates(硫酸鹽)	**3.045298e-02***	**7.927499e-19***
11	alcohol(酒精濃度)	**1.251119e-08***	**6.454021e-08***

【說明】

以上針對 11 個解釋變數進行 Brown-Forsythe 檢定，只有(10) sulphates(硫酸鹽) 和(11)alcohol(酒精濃度)兩個變數顯著，也就代表其殘差的變異數不是常數，其他變數的殘差變異數都是常數。至於 Breusch-Pagan 檢定，結果(3)citric.acid(檸檬酸)、(6)free.sulfur.dioxide(游離二氧化硫)、(7)total.sulfur.dioxide(二氧化硫總量)、(10)sulphates(硫酸鹽)以及(11)alcohol(酒精濃度)顯著，代表這幾個變數的誤差項不是常數變異數。兩個檢定皆顯著的只有(10) sulphates(硫酸鹽) 和(11)alcohol(酒精濃度)兩個變數。

5-5 配適不佳的 F 檢定(*F* Test for Lack of Fit)

本資料集符合在一個或多個 X 水準下重複多個觀測值的情況下，所以以下使用『*F* Test for Lack of Fit』來進行 11 個解釋變數的配適檢定。

【R code】

```
#【5-5.Proc reg (Lack of Fit) 看配適的迴歸線】----
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
p.value <- numeric(0)
for (i in choice.X) {
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   SSE <- anova(lm.model)[2,2]
   SSE.df <- anova(lm.model)[2,1]
   #以下區分不同的 X 水準
   lm.model <- lm(red.wine$quality~factor(red.wine[,i]))
   SSPE <- anova(lm.model)[2,2]
   SSPE.df <- anova(lm.model)[2,1]
   SSLF <- SSE - SSPE
   SSLF.df <- SSE.df-SSPE.df
   F.stat <- (SSLF/SSLF.df)/(SSPE/SSPE.df)
   p.val <- 1.0 - pf(F.stat, (SSE.df-SSPE.df), SSPE.df)
   p.value = c(p.value,p.val)
}
result <- data.frame(X 變數=X,p.val=p.value)
result[order(result$p.val),]
```

【執行結果】

	X 變數	p.val
1	fixed.acidity(固定的酸度)	1.838951e-05

```
2                volatile.acidity(揮發性酸) 2.445089e-03
3                     citric.acid(檸檬酸) 8.846394e-09
4                  residual.sugar(殘糖) 1.056613e-07
5                       chlorides(氯化物) 3.024990e-05
6   free.sulfur.dioxide(游離二氧化硫) 6.806567e-01
7  total.sulfur.dioxide(二氧化硫總量) 2.380280e-02
8                         density(密度) 2.865407e-07
9                              pH(pH 值) 4.710541e-03
10                  sulphates(硫酸鹽) 0.000000e+00
11                  alcohol(酒精濃度) 1.330848e-04
```

【說明】

依據以上配適不佳之 F 檢定結果，除了『free.sulfur.dioxide(游離二氧化硫)』不顯著以外，其他 X 變數全都顯著。也就是說，在 11 個解釋變數中，只有『(6)free.sulfur.dioxide(游離二氧化硫)』的迴歸函數是線性函數，其他皆不是。

5-6 解釋變數 X 的轉換

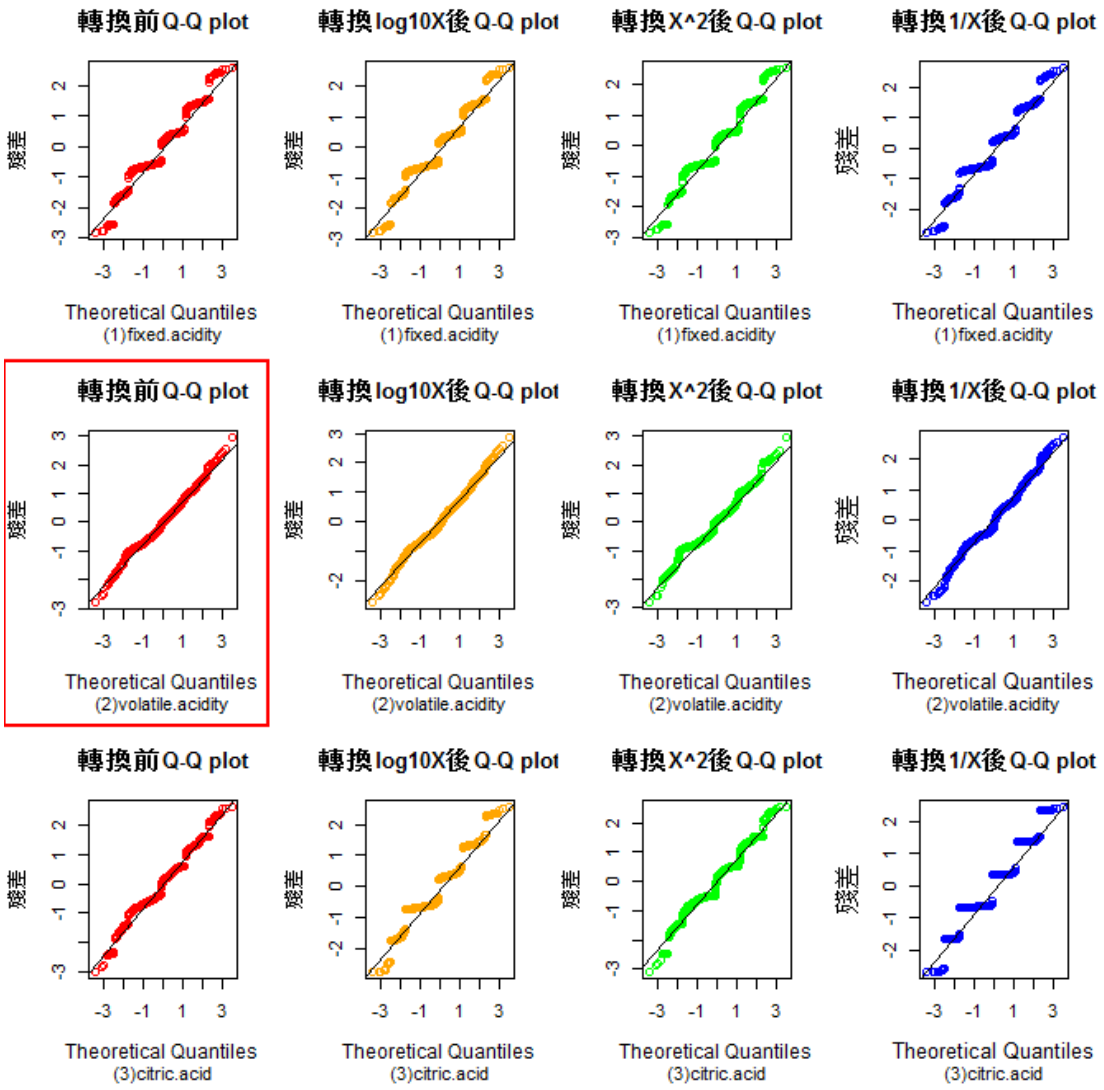
由於本資料集的分佈較為特殊，所以以下同時採用三種模式來進行 11 個 X 變的轉換，再透過轉換後的 Q-Q plot 來進行觀測。

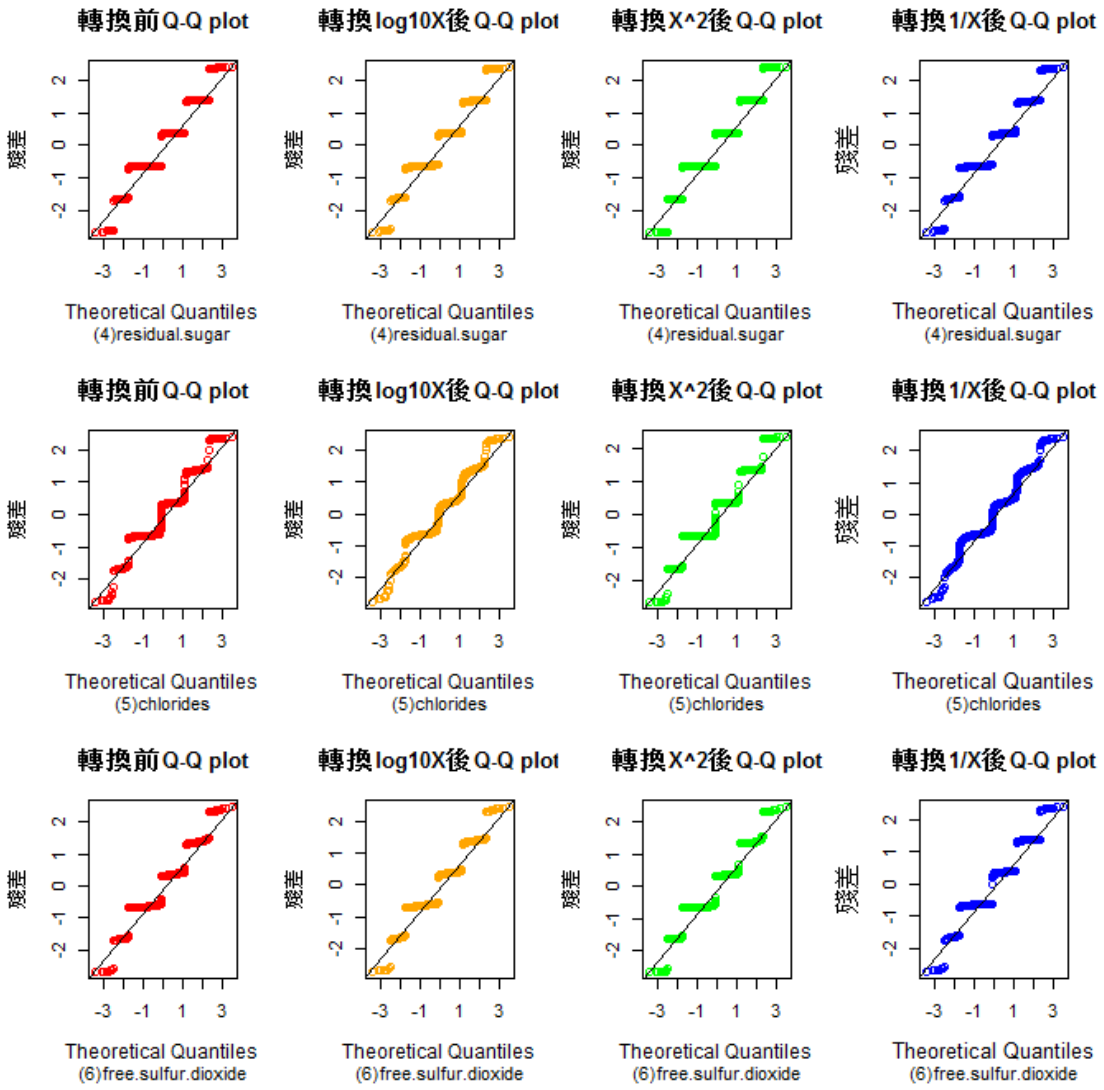
【R code】

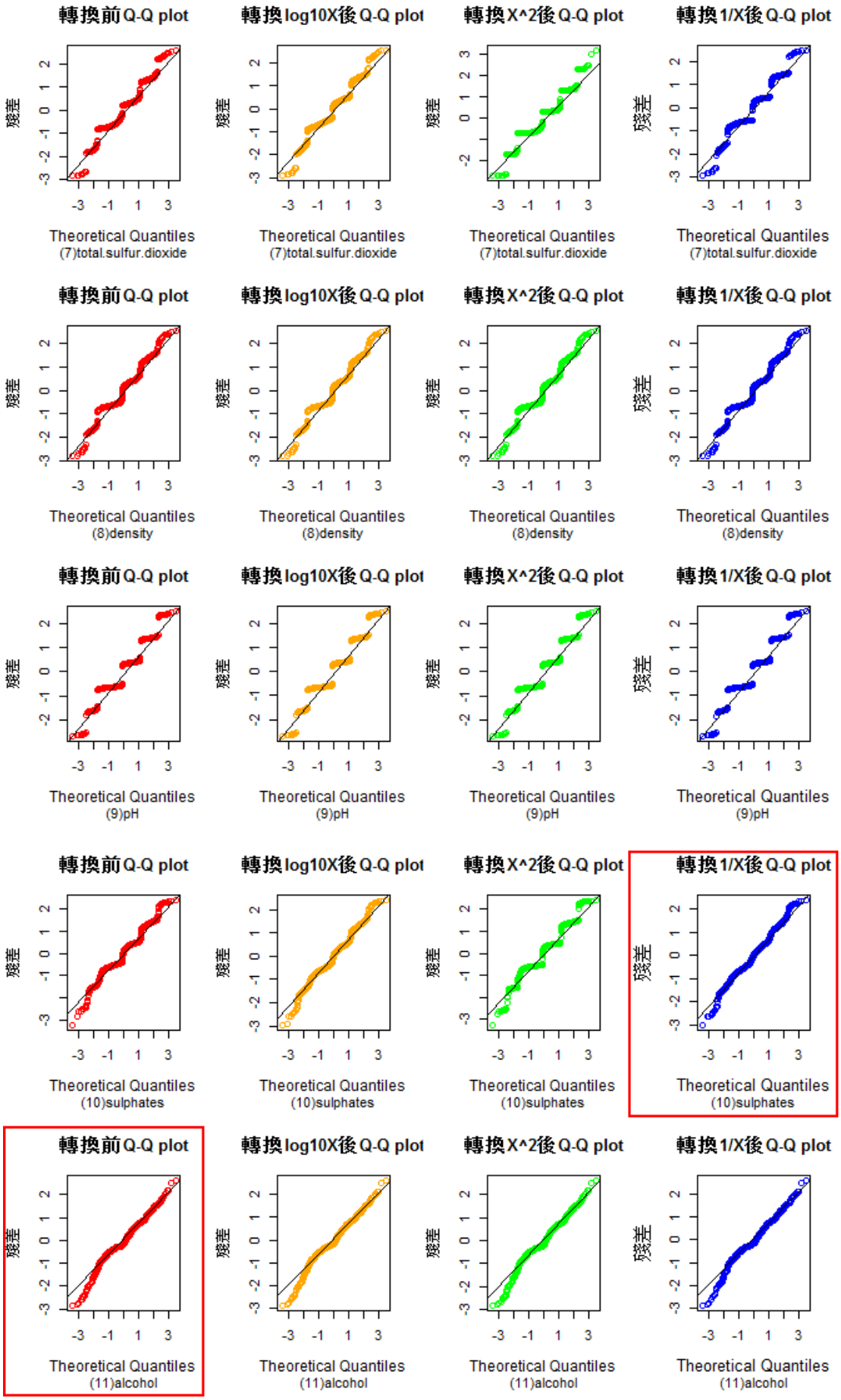
```
#【5-6.資料轉換 X】----
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
mod <- character()
cnt <- 0
for (i in choice.X) {
  if (cnt%%3==0) {
    dev.new()
    par(mfrow=c(3,4))
  }
  cnt <- cnt + 1
  f <- formula(paste0("quality~",names(red.wine)[i]))
  lm.model <- lm(f, data=red.wine)
  qqnorm(resid(lm.model),col="red",cex.lab=1.1,ylab="殘差",
         main="轉換前 Q-Q plot",
         sub=paste0("(",i,")",names(red.wine)[i]))
  qqline(resid(lm.model))

  x.new <- log10(ifelse(red.wine[,i]==0,1,red.wine[,i]))
  lm.model <- lm(red.wine$quality~x.new)
  qqnorm(resid(lm.model),col="orange",cex.lab=1.1,ylab="殘差",
```

```
            main="轉換 log10X 後 Q-Q plot",
            sub=paste0("(",i,")",names(red.wine)[i]))
  qqline(resid(lm.model))

  x.new <- (red.wine[,i])^2
  lm.model <- lm(red.wine$quality~x.new)
  qqnorm(resid(lm.model),col="green",cex.lab=1.1,ylab="殘差",
            main="轉換 X 平方後 Q-Q plot",
            sub=paste0("(",i,")",names(red.wine)[i]))
  qqline(resid(lm.model))

  x.new <- ifelse(red.wine[,i]==0,0,1/red.wine[,i])
  lm.model <- lm(red.wine$quality~x.new)
  qqnorm(resid(lm.model),col="blue",cex.lab=1.2,ylab="殘差",
            main="轉換 1/X 後 Q-Q plot",
            sub=paste0("(",i,")",names(red.wine)[i]))
  qqline(resid(lm.model))
}
par(mfrow=c(1,1))
```

【執行結果】


轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(1)fixed.acidity
轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(2)volatile.acidity
轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(3)citric.acid


轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(4)residual.sugar
轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(5)chlorides
轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(6)free.sulfur.dioxide


轉換前Q-Q plot
轉換log10X後Q-Q plot
轉換X^2後Q-Q plot
轉換1/X後Q-Q plot
殘差
Theoretical Quantiles
(7)total.sulfur.dioxide
(8)density
(9)pH
(10)sulphates
(11)alcohol

【說明】

從以上 11 個解釋變數，利用三種轉換方式(Log10、X2 以及 1/X)所得結果，可以得到較好的結果，僅有(10)sulphates(硫酸鹽)透過 1/x 的轉換是最好的一個。而(2)volatile acidity(揮發性酸)原本就已經很符合常態，(11)alcohol(酒精)也還可以。

5-7 殘差的常態分配檢定(Shapiro-Wilk)

【R code】

```
#【5-7 殘差的常態分配檢定(Shapiro-Wilk)】
choice.X <- c(1:11)
# choice.X <- c(2,10,11)     #可以只選擇某些解釋變數
p.value <- numeric(0)
for (i in choice.X) {
   f <- formula(paste0("quality~",names(red.wine)[i]))
   lm.model <- lm(f, data=red.wine)
   s <- shapiro.test(resid(lm.model))
   p.value = c(p.value,s$p.value)
}
result <- cbind(X[1:11],p.value);result
```

【執行結果】

	p.value
[1,] "fixed.acidity(固定的酸度)"	"1.32503571687972e-27"
[2,] "volatile.acidity(揮發性酸)"	"8.68866795703351e-10"
[3,] "citric.acid(檸檬酸)"	"3.02180649952336e-18"
[4,] "residual.sugar(殘糖)"	"3.9217831204756e-35"
[5,] "chlorides(氯化物)"	"1.24763891653593e-31"
[6,] "free.sulfur.dioxide(游離二氧化硫)"	"1.41771759914014e-32"
[7,] "total.sulfur.dioxide(二氧化硫總量)"	"4.43113672448884e-24"
[8,] "density(密度)"	"4.55664047578531e-24"
[9,] "pH(pH 值)"	"5.66241131707243e-32"
[10,] "sulphates(硫酸鹽)"	"3.31594008875393e-21"
[11,] "alcohol(酒精濃度)"	"8.86454266844593e-16"

【說明】

通常使用 Shapiro-Wilk 來做常態分配檢定，根據官方資料顯示(參考如下)，通常 alpha 會定為 0.1，也就是 p.value < 0.1 才算是顯著。從以上的檢定結果，全數顯著，也代表全部的殘差都不是常態分配。

(http://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html)

5-8 試著利用混合模式進行解釋變數 X 的轉變，配合反應變數 Y 的 Box-Cox 轉換

【R code】

```
install.packages("DescTools")
```

```
library(DescTools)
m <- matrix(0,nrow=11,ncol=4)
Y.new <- BoxCox(red.wine$quality, lambda = BoxCoxLambda(red.wine$quality))
for (c in 1:4) {
   for (r in 1:11) {
      if (c==1) X.new <- red.wine[,r]
      else if (c==2) X.new <- log10(ifelse(red.wine[,r]==0,1,red.wine[,r]))
      else if (c==3) X.new <- (red.wine[,r])^2
      else X.new <- ifelse(red.wine[,r]==0,0,1/red.wine[,r])
      lm.model <- lm(Y.new ~ X.new)
      s <- shapiro.test(resid(lm.model))
      m[r,c]=s$p.value
   }
}
rownames(m)<-names(red.wine)[1:11]
colnames(m)<-c("X","log10X","X^2","1/X")
m
```

【執行結果】

```
                              X       log10X          X^2          1/X
fixed.acidity        1.188601e-27 4.723230e-28 5.271179e-28 3.788489e-29
volatile.acidity     2.995254e-10 2.584447e-10 2.285190e-16 3.008798e-16
citric.acid          2.628627e-18 8.365066e-29 1.751684e-22 2.285618e-34
residual.sugar       3.686254e-35 2.140458e-34 1.084084e-35 6.162761e-34
chlorides            1.116165e-31 1.040096e-26 2.957112e-34 2.273959e-26
free.sulfur.dioxide  1.458037e-32 2.366421e-32 1.251387e-33 1.875937e-33
total.sulfur.dioxide 4.837623e-24 4.434966e-23 8.203097e-30 1.839133e-28
density              4.560237e-24 4.748431e-24 4.378203e-24 4.942915e-24
pH                   4.743017e-32 4.789191e-32 4.454705e-32 4.589167e-32
sulphates            3.191898e-21 2.893229e-14 2.706116e-28 1.768080e-10
alcohol              2.192373e-15 1.147749e-14 2.698173e-16 4.257392e-14
```

【說明】

經過以上同時對不同的 X 以及 Y 變數的不同轉換，以及利用 Shapiro-Wilk 對轉換後的殘差進行常態檢定方式，很不幸地，全數都顯著，也就是經過轉換後，殘差仍然無法呈現常態分配。

參考資料

http://archive.ics.uci.edu/ml/datasets/Wine+Quality