Linear Model Assignment 6

邱繼賢、廖偉傑、鄭雅珊

Problem 1.

首先資料讀取以及新增變數 **per**=100*(Y84-Y83)/Y83.

```
salay.data = read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/salary.txt",
                        header = T, fileEncoding = "UTF-8-BOM") %>% mutate(per=100*(Y84-Y83)/Y83)
```

接著配適模型

$$per = \beta_0 + \beta_1 SHARES + \beta_2 REV + \beta_3 INC + \beta_4 AGE + \varepsilon$$

```
fit.1 <- lm(per~SHARES+REV+INC+AGE, salay.data)
summary(fit.1)
```

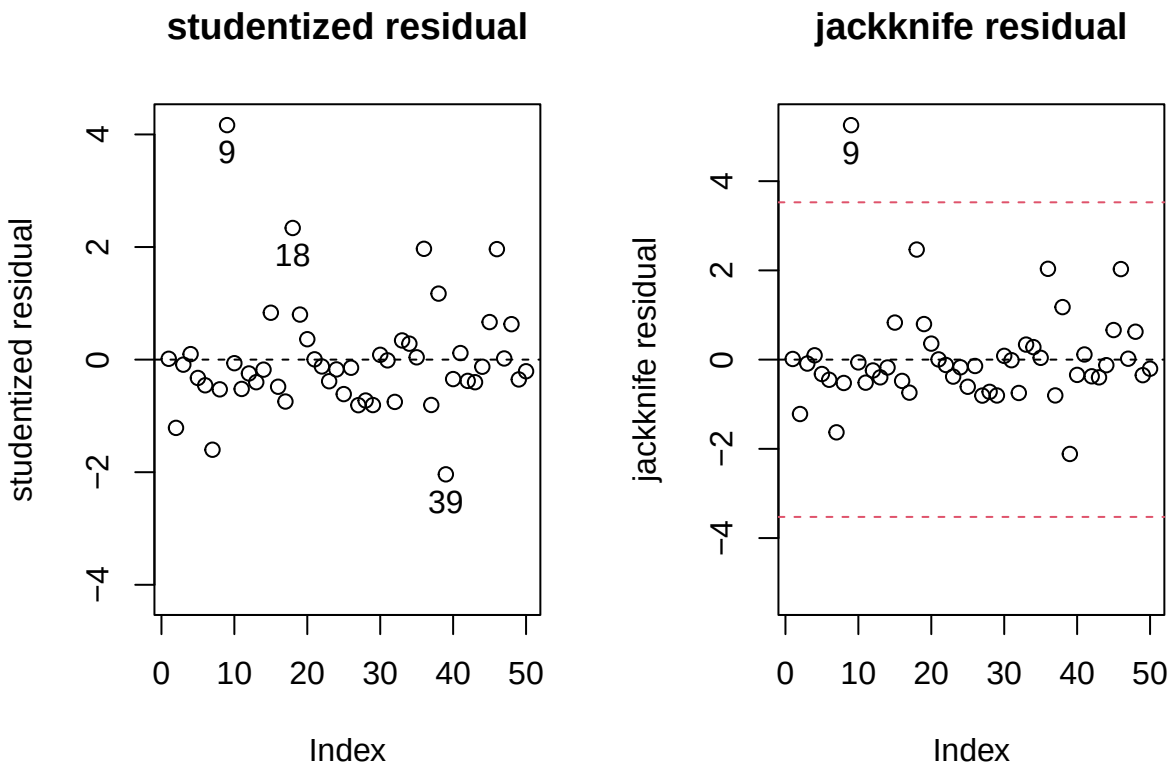
```
##
## Call:
## lm(formula = per ~ SHARES + REV + INC + AGE, data = salay.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.133 -12.519  -4.066   2.846 109.322
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.509e+01  3.571e+01   1.543    0.130
## SHARES      -3.857e-06  3.717e-06  -1.038    0.305
## REV         -7.237e-04  7.695e-04  -0.940    0.352
## INC          9.744e-03  1.655e-02   0.589    0.559
## AGE         -5.713e-01  6.232e-01  -0.917    0.364
##
## Residual standard error: 26.81 on 45 degrees of freedom
## Multiple R-squared:  0.05754,    Adjusted R-squared:  -0.02623
## F-statistic: 0.6869 on 4 and 45 DF,  p-value: 0.6048
```

根據上表的配適結果可發現模型並沒有很好的表現。

接下來檢查在此模型中是否有 outlier 出現的問題，可透過觀察 studentized residual 和 jackknife residual ，其中 jackknife residual 可使用 t 檢定來檢驗是否存在 outlier，因牽涉到多重檢定，透過 Bonferroni correct 調整顯著水準，校正後的顯著水準為 $\alpha^* = 0.05/50$，在這裡 t 檢定所使用到的自由度為 $n-1-(p+1) = 44$。

```
par(mfrow=c(1,2))

stu.1 <- rstandard(fit.1)

jack.1 <- rstudent(fit.1)
critical.v <- qt(.05/(50*2), 44, lower.tail = F)
```

```
plot(stu.1, ylab = 'studentized residual', main = 'studentized residual',ylim=c(-4.2,4.2))
abline(h = 0, lty = 2)
text((1:50)[abs(stu.1) > 2], stu.1[abs(stu.1) > 2], pos=1, labels = (1:50)[abs(stu.1) > 2])

plot(jack.1, ylab = 'jackknife residual', main = 'jackknife residual',ylim=c(-5.3,5.3))
abline(h = c(0, -critical.v, critical.v), col = c(1, 2, 2), lty = 2)
out.idx <- abs(jack.1) > critical.v
text((1:50)[out.idx], jack.1[out.idx], pos=1, labels = (1:50)[out.idx])
```



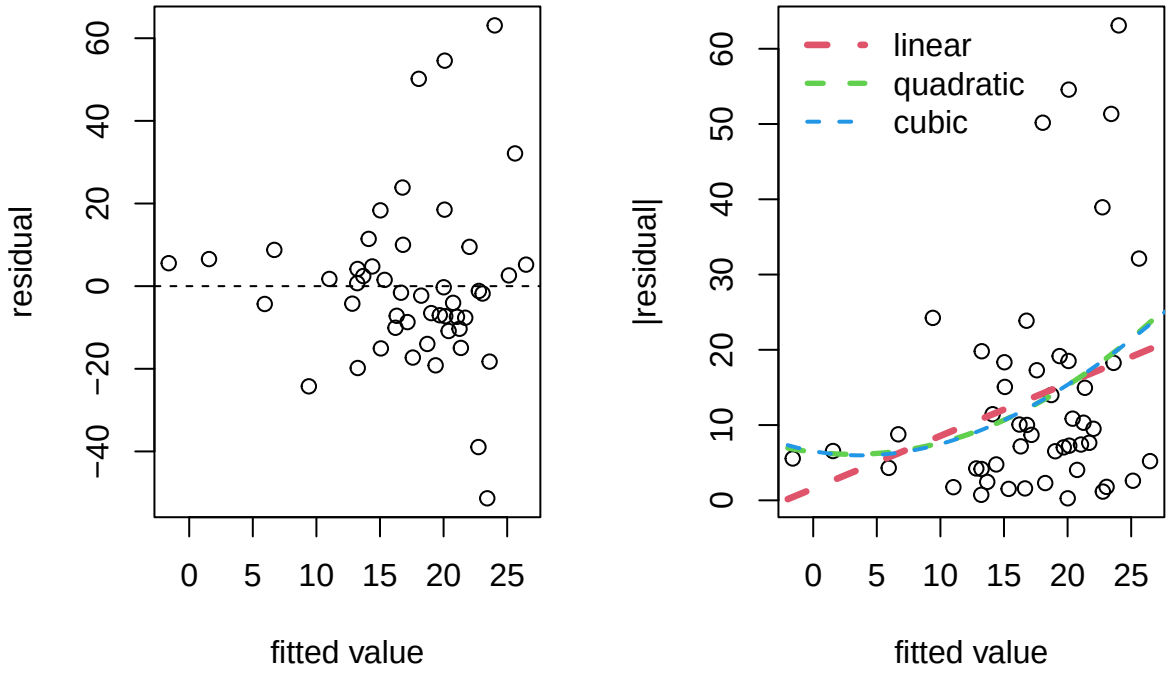
根據上面兩張圖，可發現第 9 資料有明顯較大的 studentized residual ，且其 jackknife residual 也超過臨界值 $t_{44}(1-\alpha^*/2)$，因此可推論第 9 資料為 outlier ，因此在後面的分析終將排除第 9 資料。

下一步檢查同質變異數假設是否成立

這邊可以由 raw residual 和 fitted value 的趨勢檢查，並且可以從 |residual| 和 fitted value 所配適的多項式曲線以觀測 $\sigma_x \propto \mu_x^k$ 的狀況，進而觀察在假設不成立的狀況下可能可以用到的轉換函數。

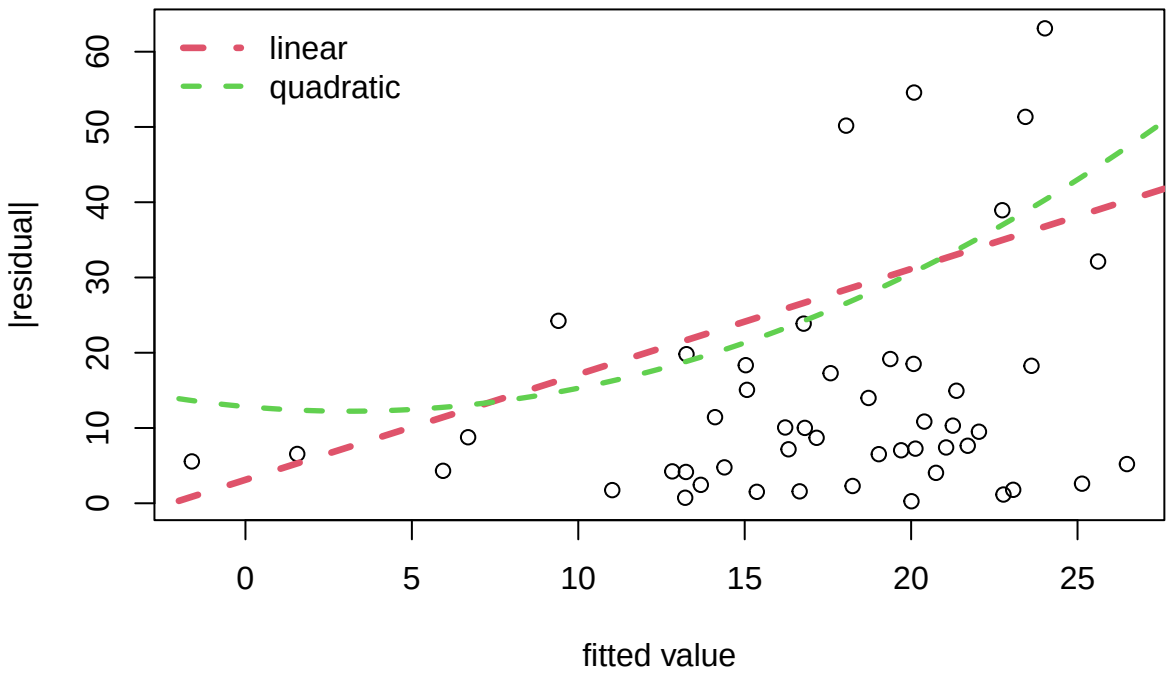
```
fit.11 <- lm(per~SHARES+REV+INC+AGE, salay.data, subset = c(-9))
fit.11.data <- data.frame("fitted" = fit.11$fit, "residual" = fit.11$res)

par(mfrow=c(1,2))

plot(fit.11.data$fitted, fit.11.data$residual, xlab = 'fitted value', ylab = 'residual')
abline(h = 0, lty = 2)

plot(fit.11.data$fitted, abs(fit.11.data$residual), xlab = 'fitted value', ylab = '|residual|')
for (k in 1:3) {
```

```
  rfit1 <- lm(abs(residual) ~ poly(fitted,k), fit.11.data)
  curv.fn<-function(x) predict(rfit1, newdata = data.frame("fitted"=x))
  curve(curv.fn(x), xlim = c(-2,30), n = 1000, col = k+1, lty = 2, lwd=4-k/1.5, add = T)
}
legend('topleft',legend = c('linear','quadratic','cubic'),
       lty = 2, col = c(2,3,4),lwd = 4-(1:3)/1.5, bty="n")
```



從右圖可以發現左側至中間部分的 σ_x 比右側部分來得小，有 σ_x 隨著 μ_x 增加而增加的趨勢，因此觀察 |residual| vs fitted value 的圖和配適的多項式曲線，可發現的確有一次或二次的趨勢。接著畫上 $2\sigma_x$ ，觀察其包住資料的方式是否合適。(大約要有 95% 資料會落在 $2\sigma_x$ 下)

```
plot(fit.11.data$fitted, abs(fit.11.data$residual), xlab = 'fitted value', ylab = '|residual|')
for (k in 1:2) {
  rfit1 <- lm(abs(residual) ~ poly(fitted,k), fit.11.data)
  curv.fn<-function(x) 2*predict(rfit1, newdata = data.frame("fitted"=x))
  curve(curv.fn(x), xlim = c(-2,30), n = 1000, col = k+1, lty = 2, lwd=4-k/1.5, add = T)
}
legend('topleft', legend = c('linear','quadratic'),
       lty = 2, col = c(2,3),lwd = 4-(1:2)/1.5, bty="n")
```

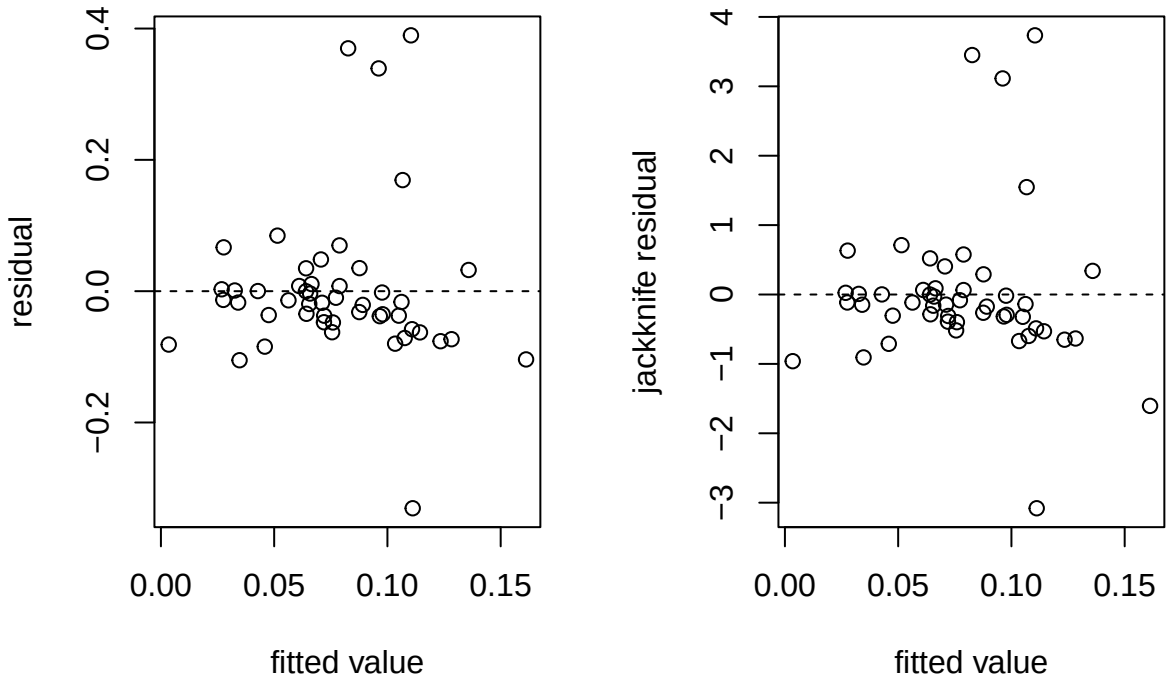


可以發現二次曲線（綠）涵蓋數據的程度較為合適，因此結論出 $\sigma_x \propto \mu_x^2$，因此在這裡選擇需要對 response 做倒數轉換，但因 **per** 中有觀測值為 0 的資料，在做轉換前將資料向右平移 2，因此令轉換為

$$g(y) = 1/(y+2)$$

接著將轉換後的反應變數重新配適模型並同樣觀察 residual plot 和 jackknife residual。

```
fit.12 <- lm(I(1/(per+2))~SHARES+REV+INC+AGE, salay.data, subset = c(-9))

par(mfrow=c(1,2))
plot(fit.12$fitted, fit.12$res, xlab = 'fitted value', ylab = 'residual')
abline(h = 0, lty = 2)

plot(fit.12$fitted, rstudent(fit.12), xlab = 'fitted value', ylab = 'jackknife residual')
abline(h = 0, lty = 2)
```



觀察上面二圖的整體趨勢，可發現 variance 並未像轉換前呈現喇叭狀，反而呈現較一致的 pattern ，因此透過倒數轉換確實有改善 non-constant variance 問題。

最後觀察轉換後模型的配適結果

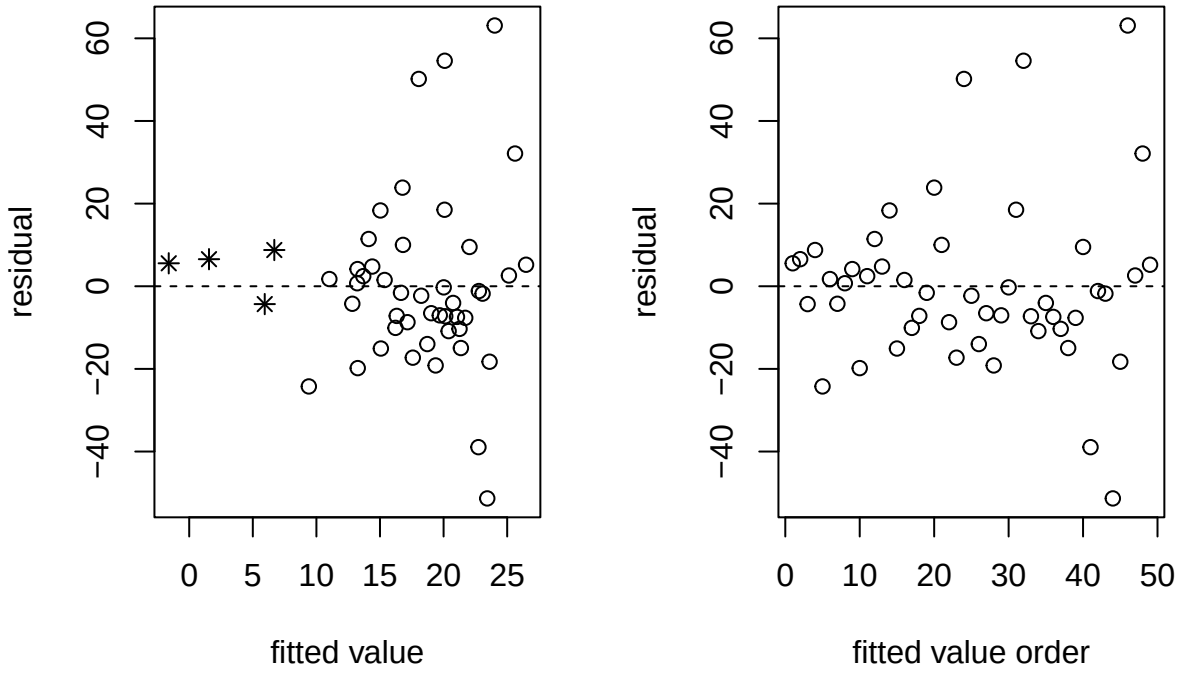
```
summary(fit.12)
```

```
##
## Call:
## lm(formula = I(1/(per + 2)) ~ SHARES + REV + INC + AGE, data = salay.data,
##     subset = c(-9))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33059 -0.04765 -0.01736  0.00788  0.38961
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.647e-01  1.619e-01  -1.018    0.314
## SHARES      -3.305e-09  1.686e-08  -0.196    0.845
## REV         -1.287e-06  3.498e-06  -0.368    0.715
## INC          5.209e-05  7.567e-05   0.688    0.495
## AGE          4.232e-03  2.824e-03   1.499    0.141
##
## Residual standard error: 0.1215 on 44 degrees of freedom
## Multiple R-squared:  0.07157,    Adjusted R-squared:  -0.01284
## F-statistic: 0.8479 on 4 and 44 DF,  p-value: 0.5026
```

配適結果依然顯示模型並沒有特別好的表現。

註：這題在觀察需注意另一種可能，之所以會覺得有 non-constant variance，主要是由於左邊四筆 residual 都較接近 0 (四筆對應到下圖中的 * 符號)，因此用 4 筆觀測值來判斷其 variance 大小可能極不準確，故也有可能總結為“無證據支持 non-constant variance”，為了支持這一說法，將 residual 依據 fitted values 的大小排序，再繪製成下方右圖，其讓左圖中每個相鄰點的 x 座標距離相等，若是有明顯的 non-constant variance 的跡象，則右圖也應會看到其現象，但右圖卻呈現類似於 constant variance 的感覺，這也可以用來支持 non-constant variance 的現象不顯著。

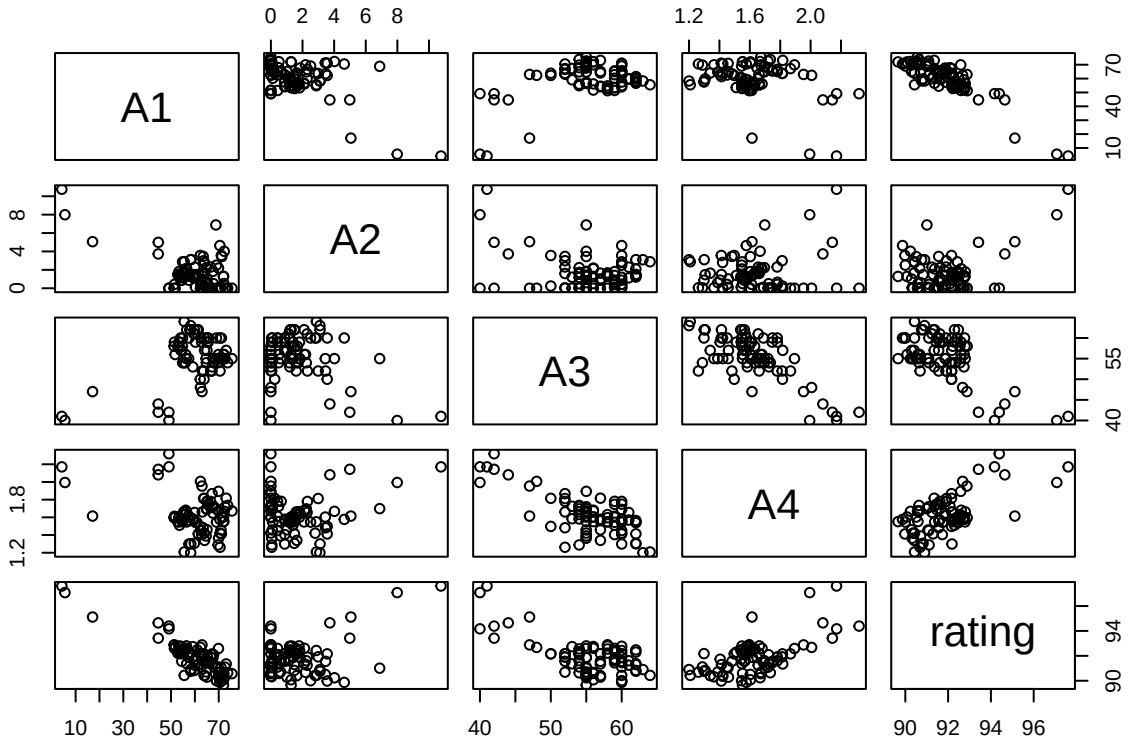
```
par(mfrow=c(1,2))
plot(fit.11$fit, fit.11$res, pch = ifelse(fit.11$fit < 8, 8, 1),
     xlab = 'fitted value', ylab = 'residual')
abline(h = 0, lty = 2)
plot(fit.11$res[order(fit.11$fit)], xlab = 'fitted value order',ylab = 'residual')
abline(h = 0, lty = 2)
```



Problem 2.

首先對資料繪製 scatter plot 觀察變數間的關係：

```
octane=read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/octane.txt")
pairs(octane)
```



圖中顯示 A1~A4 彼此的 scatter plots 上，有一些點跟其他大部分的點相距較遠，這些點可能會造成 large leverage，但由 A1~A4 對 rating 的 scatter plots 上, 這些相距較遠的點, 並沒有造成 non-linear 的影響。

接著配適模型

$$\Omega : rating = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta_4 A_4 + \epsilon$$

```
g=lm(rating~.,data=octane)
summary(g)
```

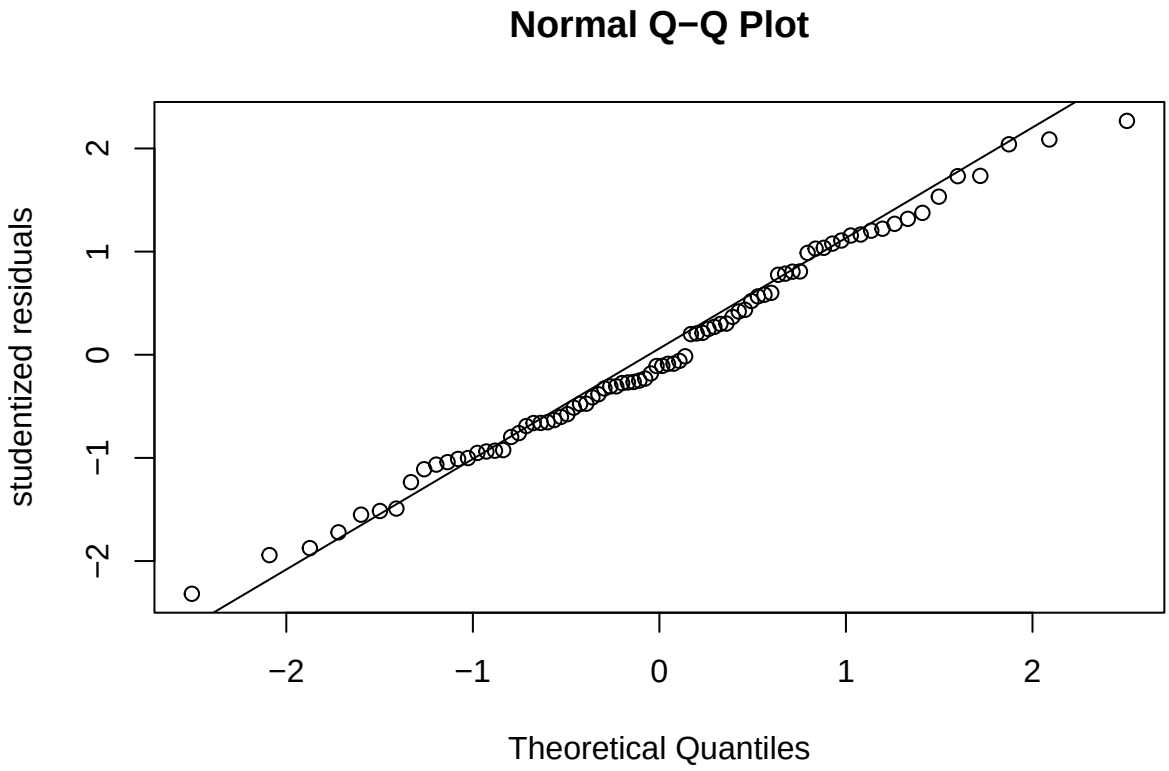
```
##
## Call:
## lm(formula = rating ~ ., data = octane)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00612 -0.28588 -0.04679  0.32159  0.98069
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 95.853150   1.224877  78.255  < 2e-16 ***
## A1          -0.092821   0.005235 -17.729  < 2e-16 ***
## A2          -0.126798   0.032157  -3.943 0.000176 ***
## A3          -0.025381   0.013971  -1.817 0.073160 .
## A4           1.967603   0.324573   6.062 4.65e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4415 on 77 degrees of freedom
```

```
## Multiple R-squared:  0.9056, Adjusted R-squared:  0.9007
## F-statistic: 184.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

從 summary report 中可發現 R^2 高達 0.9056，顯示該模型對 rating 有很高的解釋能力，但解釋變數 A3 的係數並不顯著。

1. 檢查常態假設：

```
rstud <- rstandard(g)
qqnorm(rstud, ylab = "studentized residuals")
qqline(rstud)
```



由於所有 studentized residuals 均落在斜直線附近，進一步做 Shapiro-Wilk normality test

$$H_0 : \text{the studentized residuals are normally distributed vs. } H_1 : \text{not } H_0$$

```
shapiro.test(rstud)
```

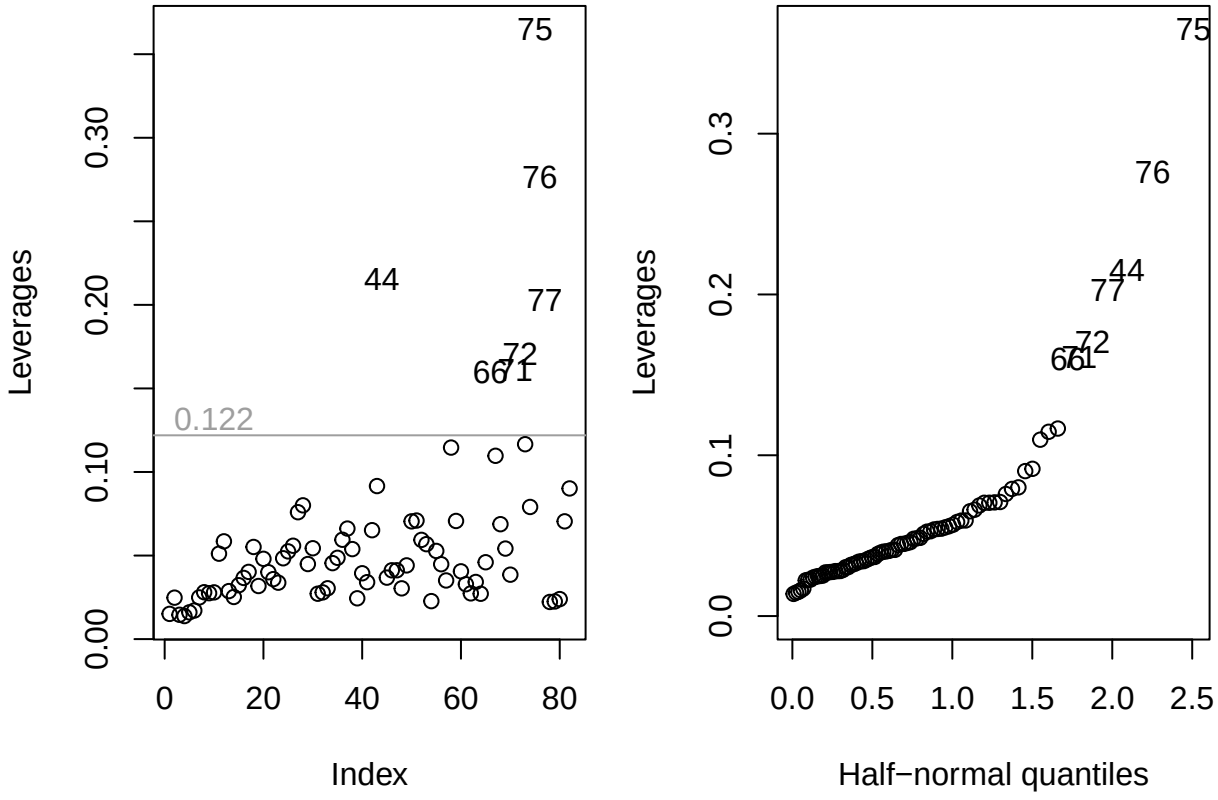
```
##
##  Shapiro-Wilk normality test
##
## data:  rstud
## W = 0.98887, p-value = 0.7073
```

由於 p-value=0.7073>0.05，在顯著水準 $\alpha = 0.05$ 下，沒有足夠證據拒絕 H_0，因此常態假設成立。

2. 檢查是否有 large leverage point ("rule of thumb": $h_i > 2p/n$)：

繪製 leverage vs. index plot 及 half-normal plot 進行觀察：

```
lev <- lm.influence(g)$hat
n <- dim(octane)[1]; p <- sum(lev)
par(mfrow=c(1,2),mar=c(5,4,1,1))
plot(lev, ylab="Leverages", type = "n"); abline(h=2*p/n,col=8)
text(10,2*p/n+0.01,round(2*p/n,3),col=8)
points(seq(n)[lev<2*p/n], lev[lev<2*p/n])
text(seq(n)[lev>2*p/n], lev[lev>2*p/n], seq(n)[lev>2*p/n])
"halfnorm" <-
function (x, nlab = 2, labs = as.character(1:length(x)), ylab = "Sorted Data",...)
{
  x <- abs(x); labord <- order(x); x <- sort(x)
  i <- order(x); n <- length(x); ui <- qnorm((n + 1:n)/(2 * n + 1))
  plot(ui, x[i], xlab = "Half-normal quantiles", ylab = ylab, ylim=c(0,max(x))
       , type = "n", ...)
  if(nlab < n)
    points(ui[1:(n - nlab)], x[i][1:(n - nlab)])
  text(ui[(n - nlab + 1):n], x[i][(n - nlab + 1):n], labs[labord][(n - nlab + 1):n])
}
halfnorm(lev, nlab = 7, labs=seq(n), ylab="Leverages")
```



從圖中可見第 44, 66, 71, 72, 75, 76, 77 七筆資料 $h_i > 2p/n = 0.122$ $(n = 82, p = 5)$，視為 large leverage points。另外列出 $h_i, i = 44, 66, 71, 72, 75, 76, 77$ 於下表：

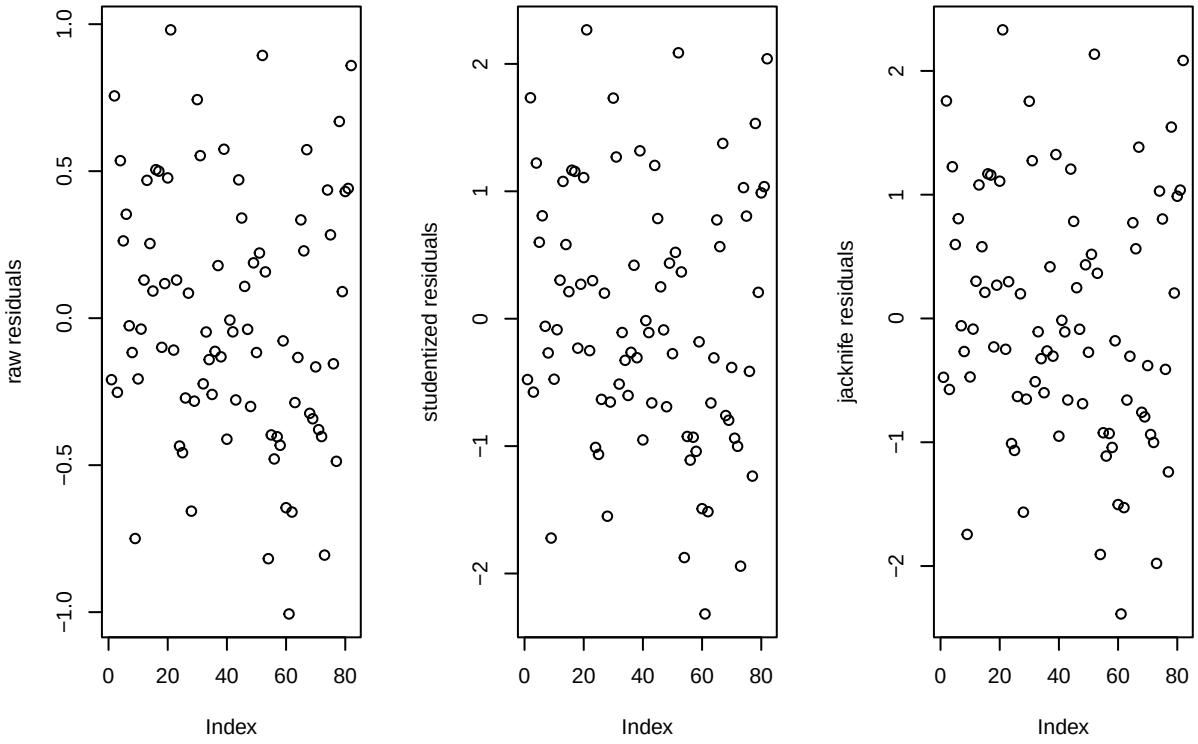
```
library(knitr)
matrix(round(lev[lev>2*p/n],3), ncol = sum(lev>2*p/n)) %>%
  `colnames<-`(c(which(lev>2*p/n))) %>% `rownames<-`('$h_i$') %>% kable()
```

	44	66	71	72	75	76	77
h_i	0.216	0.16	0.161	0.171	0.365	0.276	0.203

3. 檢查是否有 outlier：

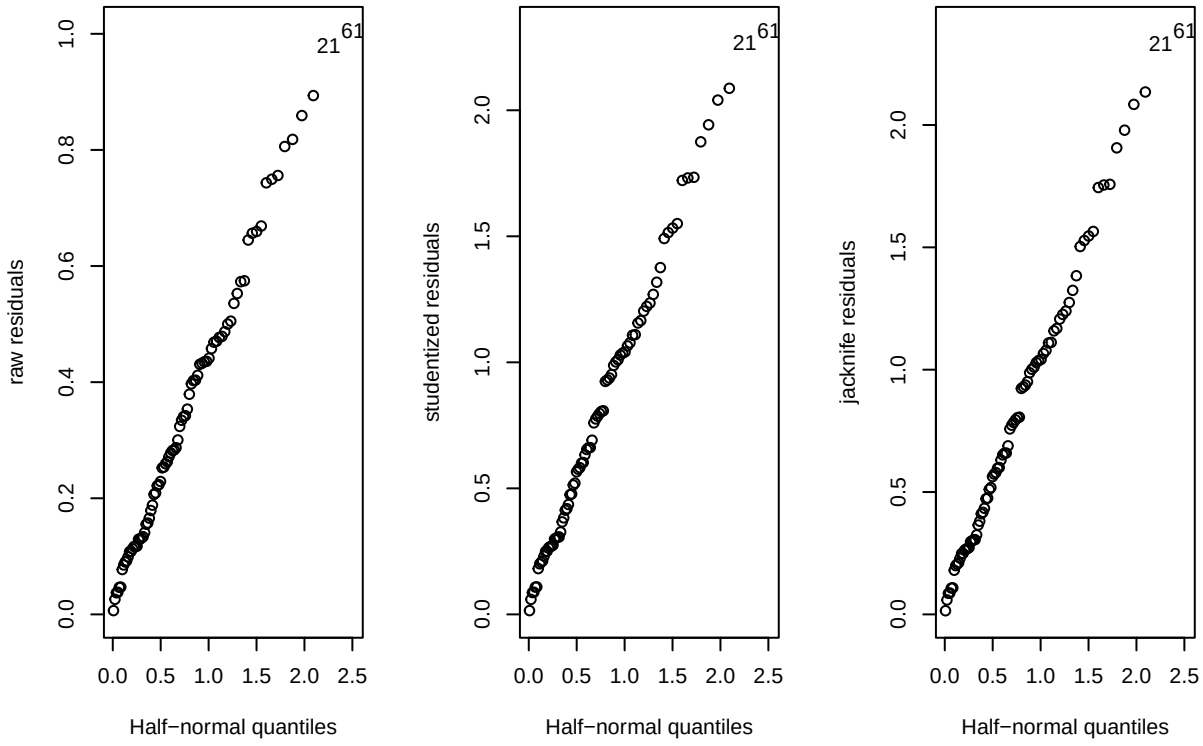
首先繪製 82 筆觀測值的 raw residual, studentized residual 及 jacknife residuals 圖如下：

```
par(mfrow=c(1,3))
rstud <- rstandard(g); rjack <- rstudent(g)
plot(g$res,ylab="raw residuals"); plot(rstud,ylab="studentized residuals");
plot(rjack,ylab="jacknife residuals")
```



可以發現三張圖的差異不大，進一步繪製 half-normal plot 檢查：

```
par(mfrow=c(1,3))
halfnorm(g$res, labs=seq(n), ylab="raw residuals")
halfnorm(rstud, labs=seq(n), ylab="studentized residuals")
halfnorm(rjack, labs=seq(n), ylab="jacknife residuals")
```



可發現 raw residual, studentized residual 及 jacknife residuals 最大的兩點均落在第 21 及 61 兩筆觀測值。

接著我們使用 jacknife residuals (t_i) 來檢查是否存在 outlier。

```
rjack[abs(rjack)==max(abs(rjack))]
```

```
##        61
## -2.387102
```

```
qt(1-0.05/(2*n), n-p-1)
```

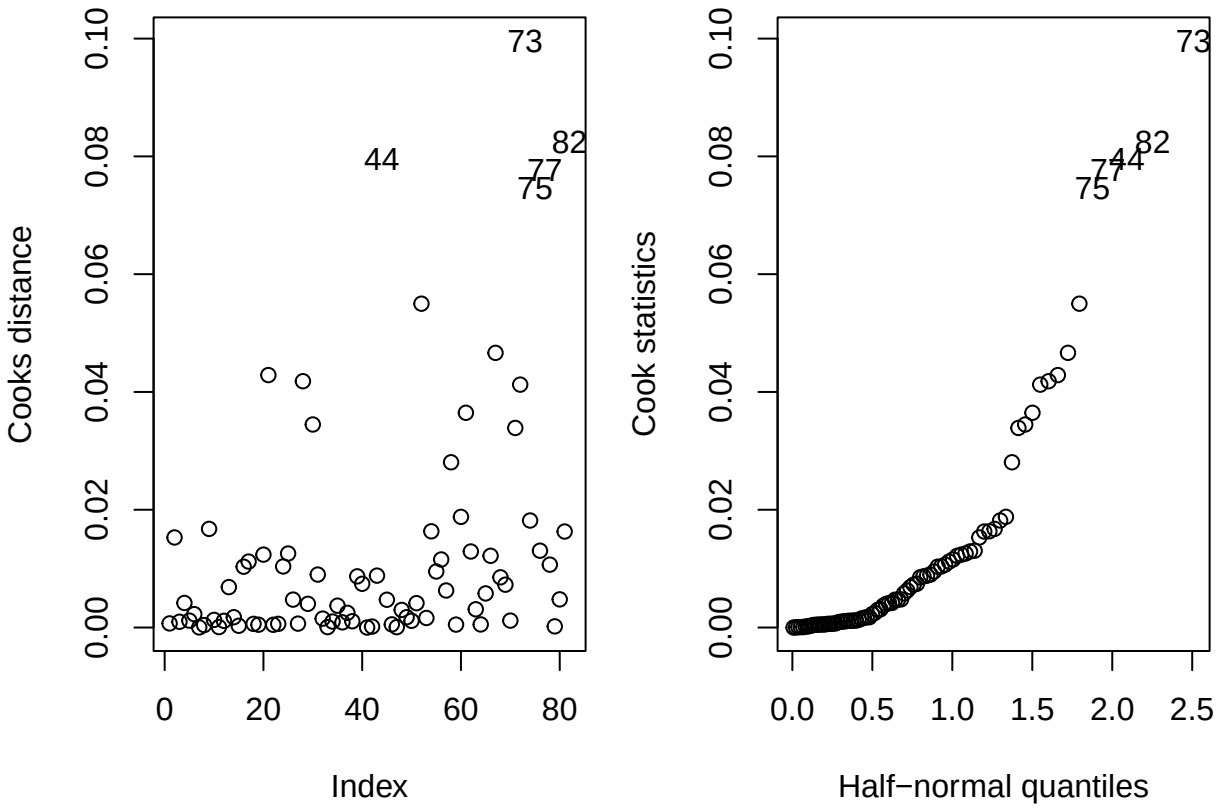
```
## [1] 3.576536
```

其中第 61 筆資料的 jacknife residual 值最大，進一步做 Bonferroni test，在顯著水準 $\alpha = 0.05$ 下，$|t_{61}| = 2.387 < t_{n-p-1}(\alpha/(2n)) = t_{76}(3.04 \times 10^{-4}) = 3.577$，因此判定該組資料沒有 outlier 存在。

4. 檢查是否有 influential point：

繪製 82 筆資料的 Cooks distance (D_i) vs. index plot 及 half-normal plot 如下圖：

```
cook <- cooks.distance(g)
par(mfrow=c(1,2),mar=c(5,4,1,1))
plot(cook, ylab="Cooks distance", type = "n")
points(seq(n)[!(cook %in% tail(sort(cook),5))], cook[!(cook %in% tail(sort(cook),5))])
text(seq(n)[cook %in% tail(sort(cook),5)], cook[cook %in% tail(sort(cook),5)],
     seq(n)[cook %in% tail(sort(cook),5)])
halfnorm(cook, nlab = 5, labs=seq(n), ylab="Cook statistics")
```



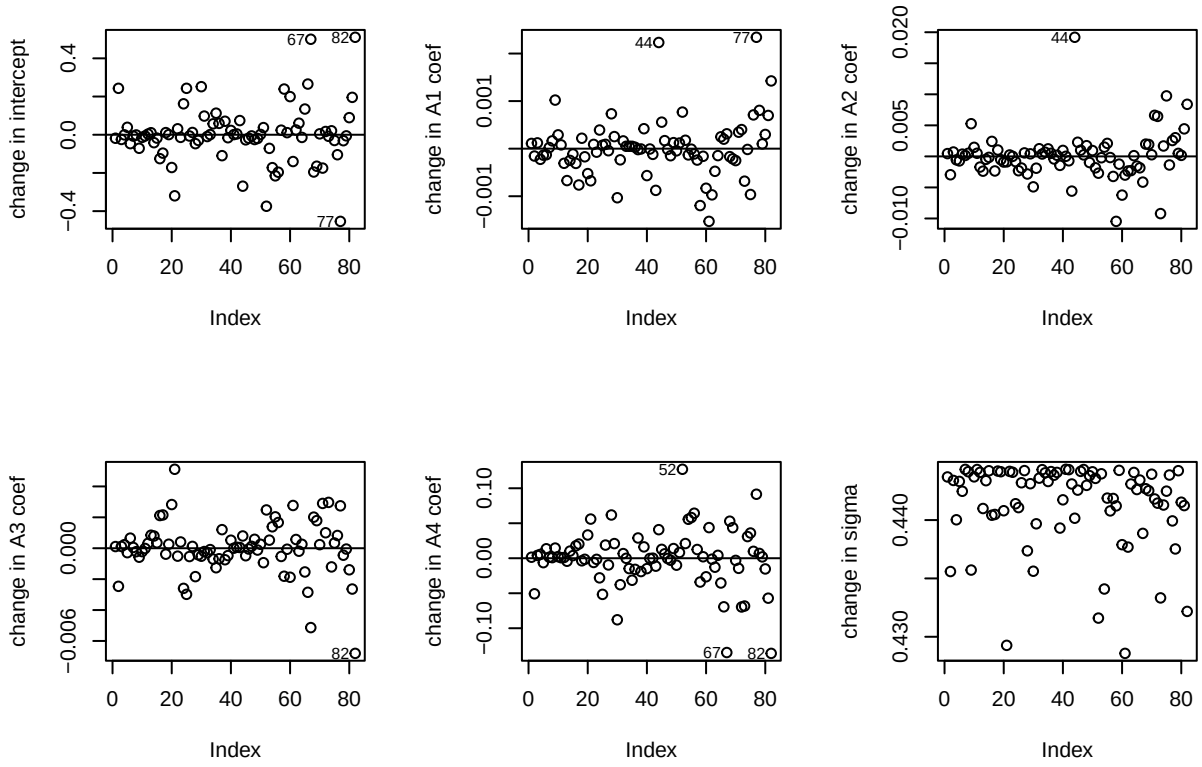
其中 D_i 最大的前五筆資料為第 44, 73, 75, 77, 82 個觀測值，但從圖中可知所有觀測值的 D_i 皆小於 1，所以並無真正明顯的 influential point。列出 $D_i, i = 44, 73, 75, 77, 82$ 於下表：

```
matrix(round(cook[cook %in% tail(sort(cook),5)],3), ncol = 5) %>%
  `colnames<-`(c(which(cook %in% tail(sort(cook),5)))) %>% `rownames<-`('$D_i$') %>%
  kable()
```

	44	73	75	77	82
D_i	0.08	0.1	0.075	0.078	0.083

繪製 change in the estimated coefficients vs. index plot 觀察移除單筆資料對參數估計的變動影響如下圖：

```
ginf <- lm.influence(g)
par(mfrow=c(2,3))
for(i in 1:5){
plot(ginf$coef[,i],ylab=c("change in intercept","change in A1 coef",
                          "change in A2 coef","change in A3 coef",
                          "change in A4 coef","change in sigma")[i])
  abline(h=0); cl <- c(0.4,0.002,0.019,0.006,0.1)
text(seq(n)[abs(ginf$coef[,i])>cl[i]]-5, ginf$coef[,i][abs(ginf$coef[,i])>cl[i]],
     seq(n)[abs(ginf$coef[,i])>cl[i]],cex=0.8)
}
plot(ginf$sig,ylab="change in sigma")
```



統整 2-4 點的觀察於下表：

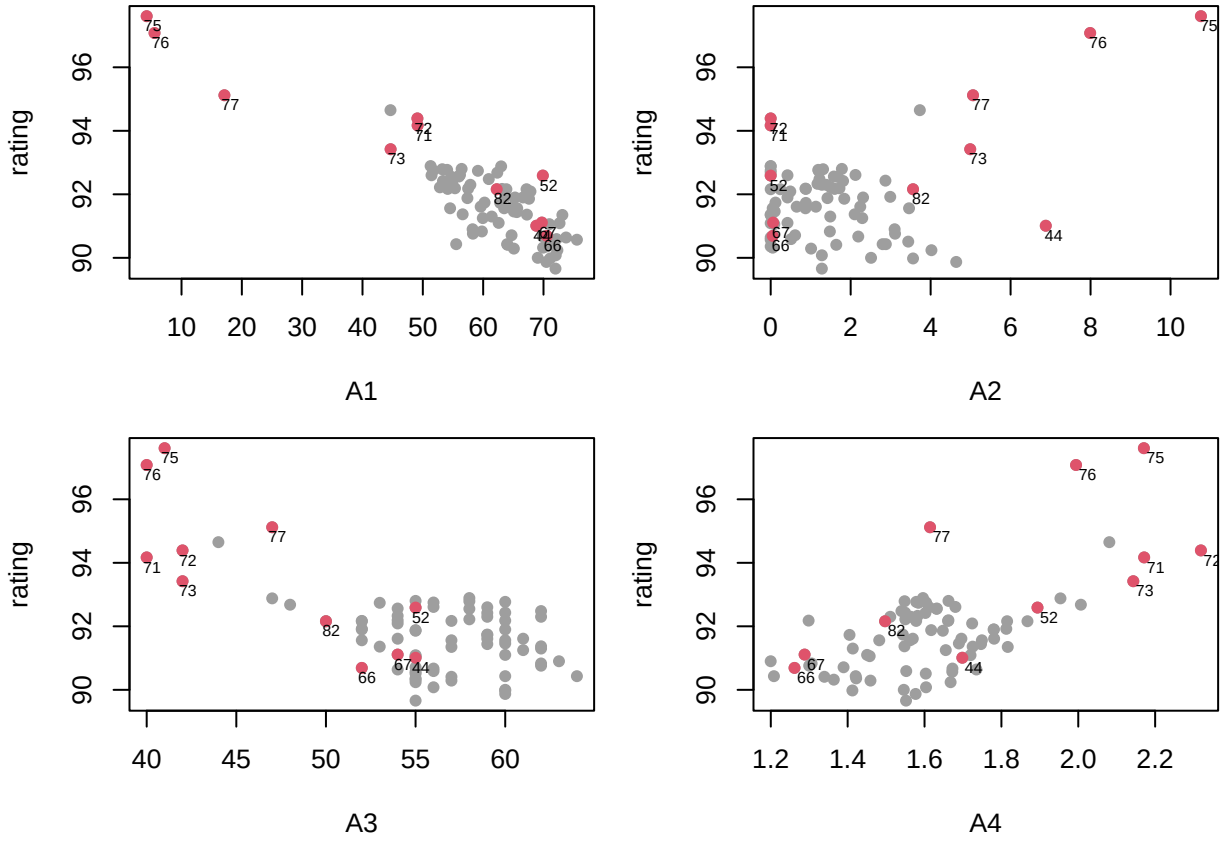
```
M <- matrix(rep("",11*8), ncol = 11)
M[1,c(1,3,5,6,8,9,10)] <- "v"; M[3,c(1,7,8,10,11)] <- "v"; M[4,c(4,10,11)] <- "v"
M[5,c(1,10)] <- "v"; M[6,1] <- "v"; M[7,11] <- "v"; M[8,c(2,4,11)] <- "v"
M %>%
  `colnames<-`(c(44,52,66,67,71,72,73,75,76,77,82)) %>%
  `rownames<-`(c('Large leverage','Outlier','Cook stat.','Change in intercept',
                 'Change in A1 coef','Change in A2 coef','Change in A3 coef',
                 'Change in A4 coef')) %>% kable()
```

	44	52	66	67	71	72	73	75	76	77	82
Large leverage	v		v		v	v		v	v	v	
Outlier											
Cook stat.	v						v	v		v	v
Change in intercept				v						v	v
Change in A1 coef	v									v	
Change in A2 coef	v										
Change in A3 coef											v
Change in A4 coef		v		v							v

- 雖然第 44, 66, 71, 72, 75, 76, 77 筆資料有較大的 leverage，但真正使參數估計變化較多的是第 44 筆資料，在 A2 的係數估計，移除第 44 筆資料明顯較移除其他筆資料的改變更大。
- 針對 intercept, A1, A3, A4 及 σ 的估計變化，上表也列出變化較大的資料點，不過從圖形可看出，和移除其他資料點的改變差異不會太大。

最後以散布圖呈現所有資料的解釋變數值和 rating 間的關係，以及標示上表列出的資料點在圖中的位置：

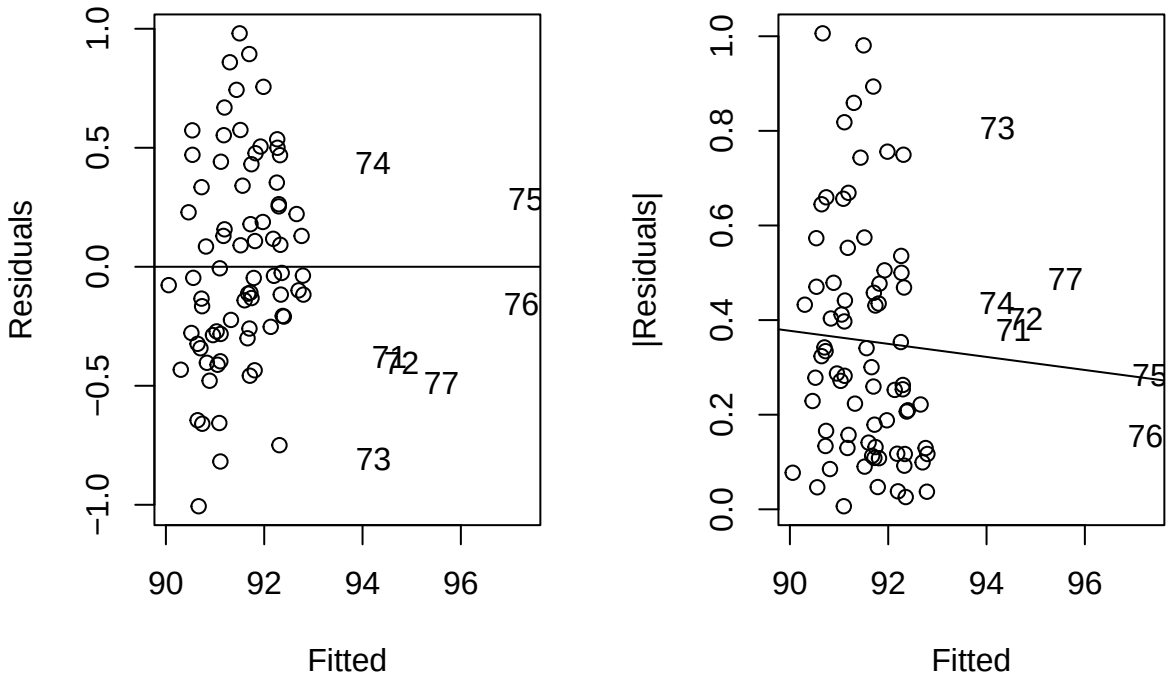
```
par(mfrow=c(2,2),mar=c(4,4,1,1));id1 <- c(44,52,66,67,71,72,73,75,76,77,82)
for(i in 1:4){
  plot(octane[,i],octane$rating,xlab=names(octane)[i],ylab=names(octane)[5],pch=16,col=8)
  points(octane[id1,i], octane$rating[id1],col=2,pch=16)
  text(octane[id1,i]+c(1,0.2,0.3,0.03)[i], octane$rating[id1]-0.3, id1,cex=0.6)
}
```



5. 檢查是否存在 non-constant variance：

首先繪製 residuals vs. fitted value plot：

```
par(mfrow=c(1,2))
plot(g$fit, g$res, xlab="Fitted", ylab="Residuals", type = "n"); abline(h=0);
points(g$fit[g$fit<94], g$res[g$fit<94])
text(g$fit[g$fit>94], g$res[g$fit>94], seq(n)[which(g$fit>94)])

plot(g$fit,abs(g$res),xlab="Fitted",ylab="|Residuals|", type = "n")
points(g$fit[g$fit<94], abs(g$res)[g$fit<94])
text(g$fit[g$fit>94], abs(g$res)[g$fit>94], seq(n)[which(g$fit>94)])
abline(summary(lm(abs(g$res) ~ g$fit)))
```



從上面兩張圖發現，residual 或 residual 取絕對值的變化看似隨著 fitted value 增加略為減少，其中 fitted value 偏大 ($\widehat{\text{rating}} > 94$) 的 7 個點當中，有五個點 71, 72, 75, 76, 77 是 large leverage。

進一步做 score Test for non-constant error variance：

$$H_0 : \text{constant error variance vs. } H_1 : \text{the error variance changes}$$

```
library(car); ncvTest(g)
```

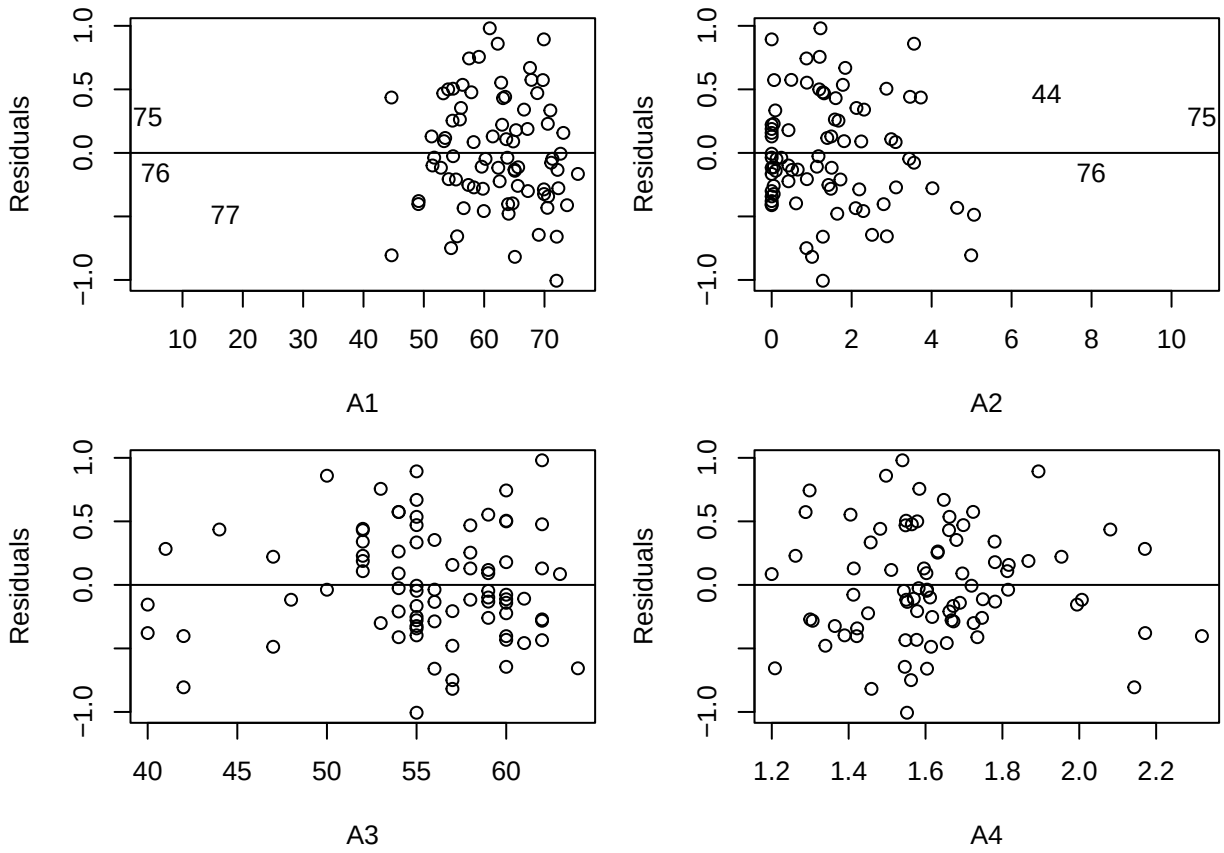
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.5271729, Df = 1, p = 0.4678
```

由於 p-value= 0.4678>0.05，在顯著水準 $\alpha = 0.05$ 下，沒有足夠證據拒絕 H_0，因此判定模型不違反 constant variance 假設，此處不對資料進行任何轉換。

另外也繪製 residuals vs. A1-A4 plot 進行觀察：

```
par(mfrow=c(2,2),mar=c(4,4,1,1))
for(j in 1:4){
  if(j==1){
      plot(octane[,j], g$res,xlab=names(octane)[j], ylab="Residuals", type = "n");
    points(octane[octane[,j]>40,j], g$res[octane[,j]>40])
    text(octane[octane[,j]<40,j], g$res[octane[,j]<40], seq(n)[which(octane[,j]<40)])
  }else if(j==2){
      plot(octane[,j], g$res,xlab=names(octane)[j], ylab="Residuals", type = "n");
    points(octane[octane[,j]<6,j], g$res[octane[,j]<6])
    text(octane[octane[,j]>6,j], g$res[octane[,j]>6], seq(n)[which(octane[,j]>6)])
```

```
  }else{
    plot(octane[,j], g$res,xlab=names(octane)[j], ylab="Residuals")
  }
  abline(h=0)
}
```



(1) Residual 隨著 A1 增加有變大的趨勢，但若排除 75, 76, 77 三點，其資料點看起來就沒有特別的 pattern

(2) Residual 隨著 A2 增加而變小，但若排除 44, 75, 76 三點，其資料點看起來就沒有特別的 pattern

(3) Residual vs. A3 及 Residual vs. A4 並無呈現 non-constant variance pattern

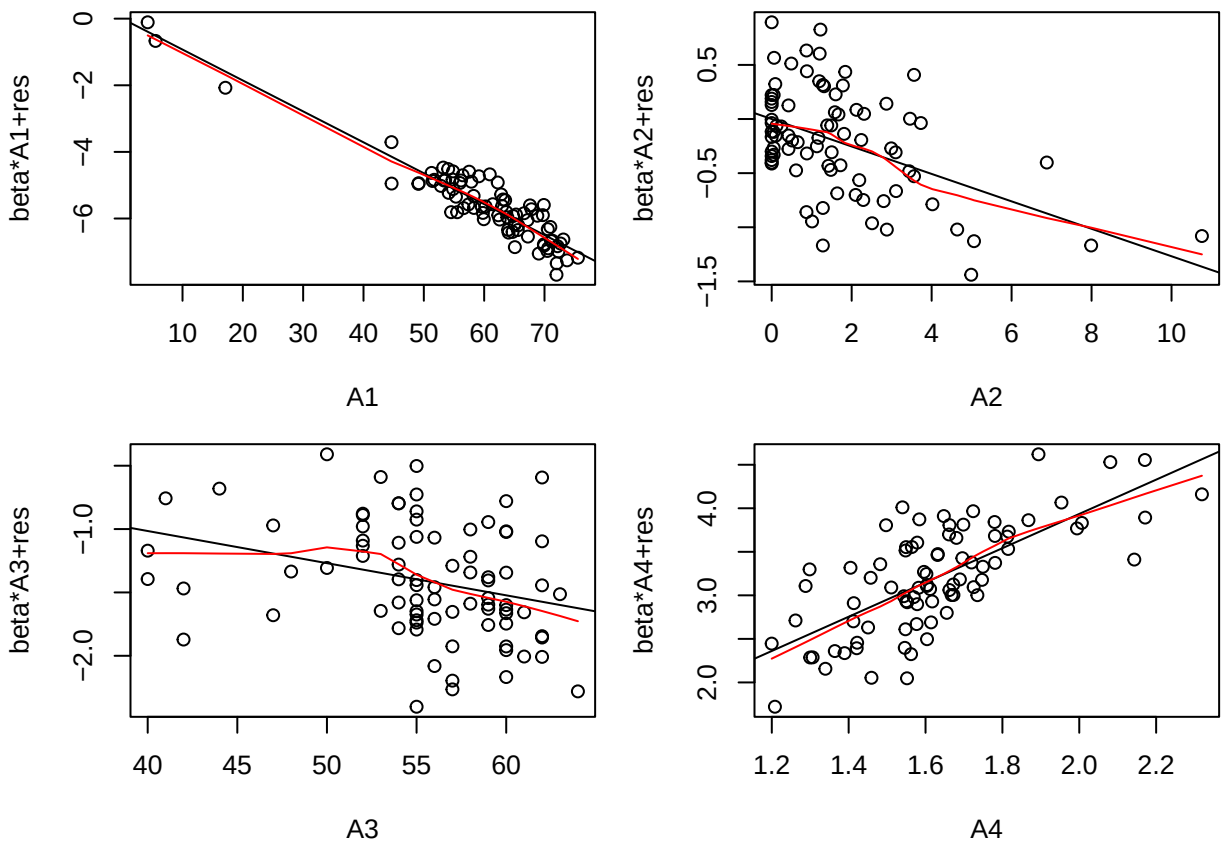
在 (1)(2) 點的觀察，造成 non-constant variance 的資料點皆為 large leverage points，而 large leverage 的 residual variance 較小，因此由這些點造成 non-constant variance 的情況並無法推定 non-constant variance 為真，仍需要 A1<40 和 A2>6 更多資料來佐證此一假設。

6. 檢查是否有 mean curvature pattern：

以下繪製 partial residuals plot 進行觀察：

```
prplot <- function(g,i)
{
# Partial residuals plot for predictor i
  xl<-attributes(g$terms)$term.labels[i]; yl<-paste("beta*",xl,"+res",sep="")
  x<-model.matrix(g)[,i+1]
  plot(x,g$coeff[i+1]*x+g$res,xlab=xl,ylab=yl); abline(0,g$coeff[i+1])
  invisible()
}
par(mfrow=c(2,2),mar=c(4,4,1,1))
```

```
for(i in 1:4){
  prplot(g,i)
  lines(lowess(x = octane[,i],y = g$residuals+g$coeff[i+1]*octane[,i], f = 0.8),
col = "red")
}
```



partial residuals plot 顯示 A3 變數對 residual 有 quadratic 的關係存在，因此考慮加入二次項解釋變數到模型當中 (A2 看起來也有些微 quadratic pattern 但配適二次項係數並不顯著)，以下配適模型

$$\Omega_2 : rating = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta_4 A_4 + \beta_5 A_3^2 + \epsilon$$

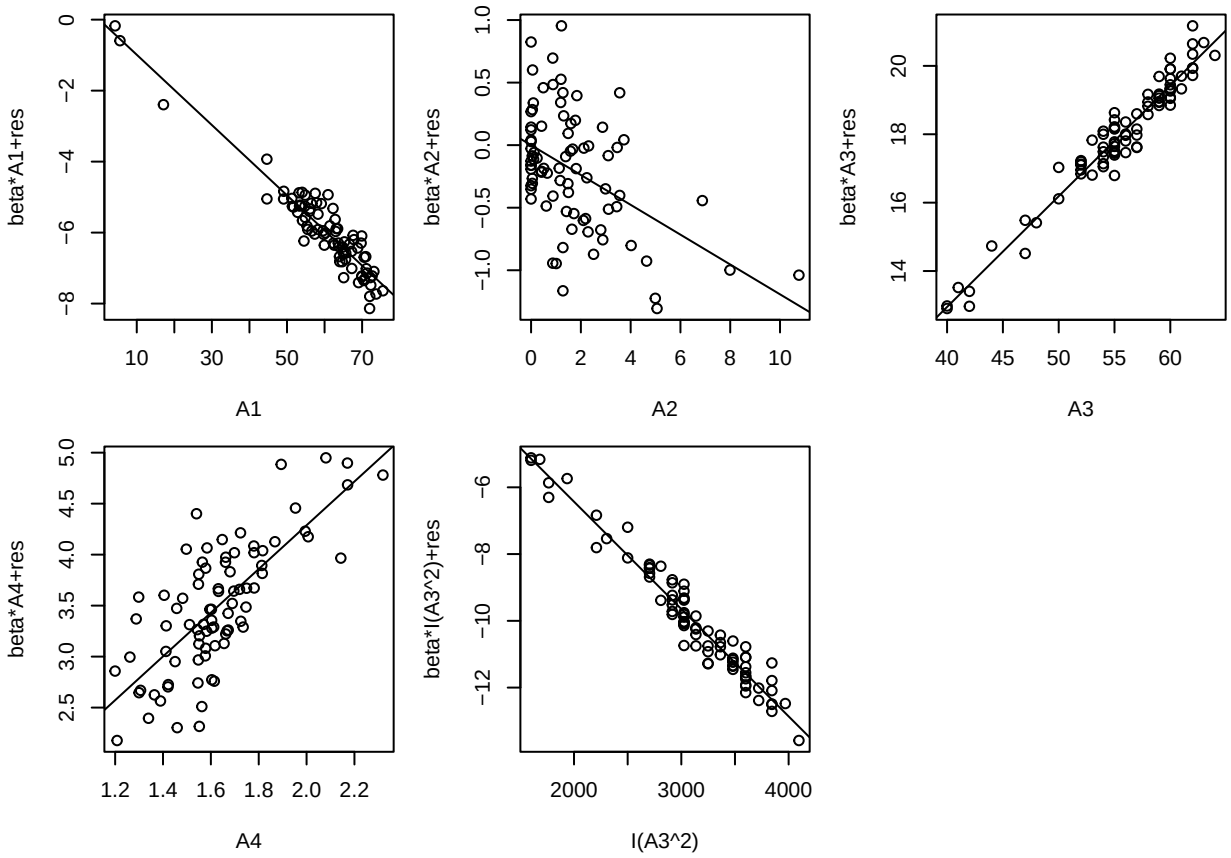
```
g2 <- lm(rating~.+I(A3^2), data = octane)
summary(g2)
```

```
##
## Call:
## lm(formula = rating ~ . + I(A3^2), data = octane)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01167 -0.31776 -0.04766  0.28219  1.09923
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.546383   4.734309  18.281  < 2e-16 ***
## A1          -0.099002   0.005966 -16.595  < 2e-16 ***
## A2          -0.119257   0.031741  -3.757 0.000335 ***
```

```
## A3           0.323742   0.172336   1.879 0.064140 .
## A4           2.144055   0.329804   6.501 7.53e-09 ***
## I(A3^2)     -0.003216   0.001582  -2.032 0.045622 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4328 on 76 degrees of freedom
## Multiple R-squared:  0.9105, Adjusted R-squared:  0.9046
## F-statistic: 154.6 on 5 and 76 DF,  p-value: < 2.2e-16
```

從 summary report 中可發現 R^2 提升至 0.9105，雖然解釋變數 A3 的係數仍不顯著，但二次項的係數顯著。以下再繪製 partial residuals plot：

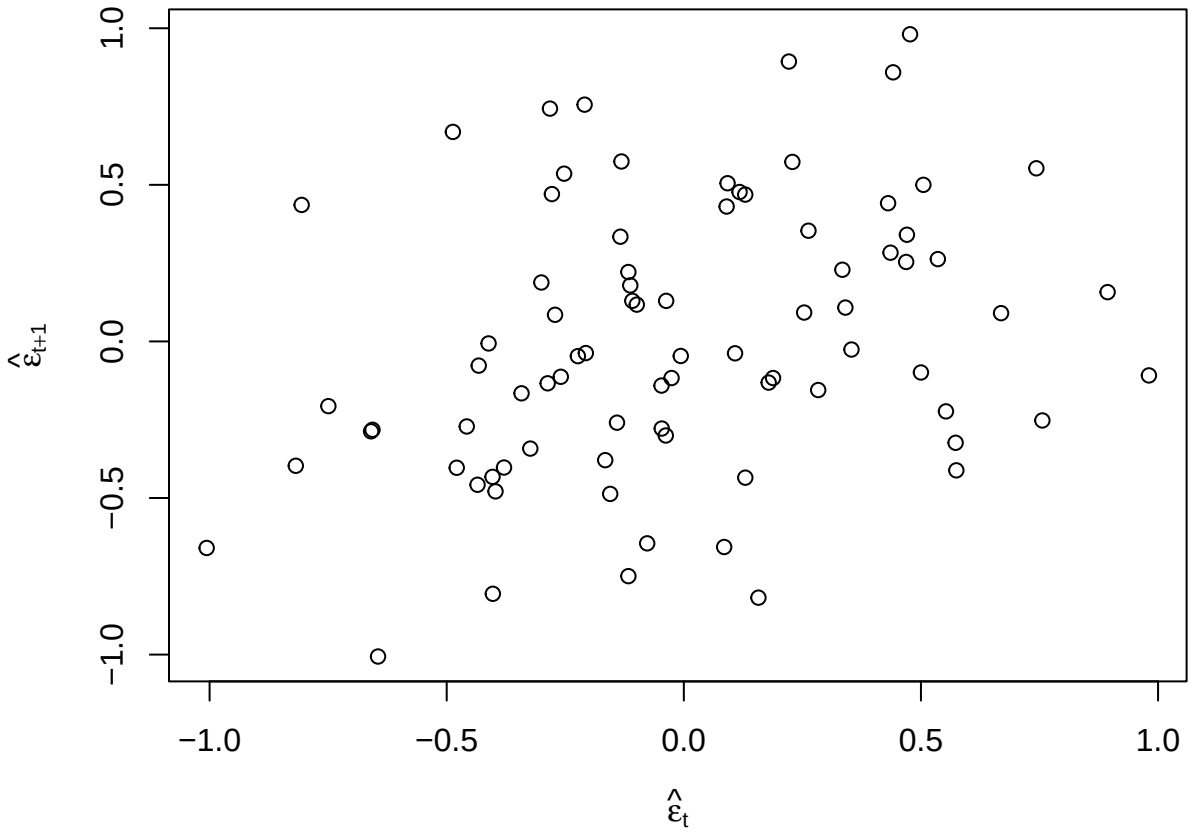
```
par(mfrow=c(2,3),mar=c(4,4,1,1))
for(i in 1:5) prplot(g2,i)
```



從圖中發現並無更明顯的 mean structure，因此最終使用模型 Ω_2 來解釋該組資料。

7. (FYI) 假設每一筆數據在 data set 中的前後順序正好反映 time order，檢查是否有 correlated error 的狀況。以下繪製 $\hat{\epsilon}_t$ against $\hat{\epsilon}_{t+1}$ plot：

```
par(mar=c(4,5,1,1))
plot(g$res[-n],g$res[-1],xlab=expression(hat(epsilon)[t]),
     ylab=expression(hat(epsilon)[t+1]))
```



圖形顯示似乎存在 correlated error 的狀況，進一步做 Durbin-Watson test：

```
library(lmtest)
dwtest(g, data=octane)
```

```
##
##  Durbin-Watson test
##
## data:  g
## DW = 1.3423, p-value = 0.0003063
## alternative hypothesis: true autocorrelation is greater than 0
```

由於 p-value$= 3.063 \times 10^{-4}$，在顯著水準 $\alpha = 0.05$ 下，有足夠證據拒絕 H_0，顯示有 correlated error 的狀況。可考慮使用 GLS 進行分析，此處假設 correlated errors 服從 AR(1) structure：

```
library(nlme)
g3 <- gls(rating~.,data=octane, correlation=corAR1(form=~1))
#form=~1: using the order of the observations in the data as a covariate, and no groups
summary(g3)
```

```
## Generalized least squares fit by REML
##   Model: rating ~ .
##   Data: octane
##        AIC     BIC    logLik
##   122.0083 138.415 -54.00417
##
## Correlation Structure: AR(1)
##  Formula: ~1
```

```
##  Parameter estimate(s):
##      Phi
## 0.420975
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 95.45862 1.4312417  66.69637  0.0000
## A1          -0.08461 0.0069060 -12.25177  0.0000
## A2          -0.10619 0.0377895  -2.81007  0.0063
## A3          -0.02844 0.0179745  -1.58206  0.1177
## A4           1.99287 0.3932281   5.06797  0.0000
##
##  Correlation:
##    (Intr) A1     A2     A3
## A1 -0.065
## A2 -0.192  0.524
## A3 -0.847 -0.360 -0.031
## A4 -0.851  0.004  0.039  0.576
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.38917392 -0.66797405 -0.06196949  0.62459688  2.17263985
##
## Residual standard error: 0.4604631
## Degrees of freedom: 82 total; 77 residual
```

```
intervals(g3)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                    lower        est.        upper
## (Intercept) 92.60865067 95.45861693 98.308583189
## A1          -0.09836225 -0.08461066 -0.070859067
## A2          -0.18143986 -0.10619134 -0.030942820
## A3          -0.06422843 -0.02843670  0.007355035
## A4           1.20985207  1.99286915  2.775886233
##
##  Correlation structure:
##         lower     est.     upper
## Phi 0.1548294 0.420975 0.6301519
##
##  Residual standard error:
##     lower      est.     upper
## 0.3740666 0.4604631 0.5668143
```

Problem 3.

匯入資料：

```
data3 = read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/vehicle.txt",
                   skip = 1)
colnames(data3) = c("ACC","WHP","SP","G")
```

(a)

建構 linear model :

$$ACC = \beta_0 + \beta_1 WHP + \beta_2 SP + \beta_3 G + \epsilon$$

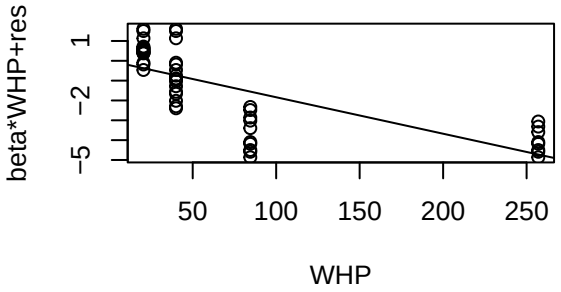
```
fit3.1 = lm(ACC ~ ., data3)
summary(fit3.1)
```

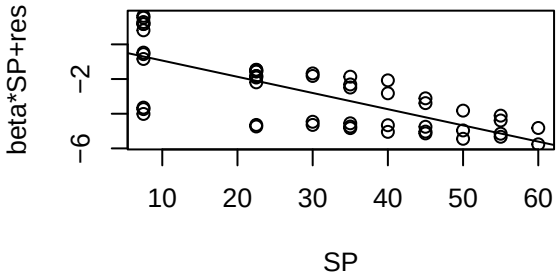
```
##
## Call:
## lm(formula = ACC ~ ., data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3124 -0.9003  0.2486  0.9489  2.3477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.19949    0.60087  11.982 9.57e-16 ***
## WHP         -0.01838    0.00269  -6.833 1.62e-08 ***
## SP          -0.09347    0.01307  -7.149 5.45e-09 ***
## G           -0.15548    0.09040  -1.720   0.0922 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.47 on 46 degrees of freedom
## Multiple R-squared:  0.624,  Adjusted R-squared:  0.5995
## F-statistic: 25.45 on 3 and 46 DF,  p-value: 7.451e-10
```

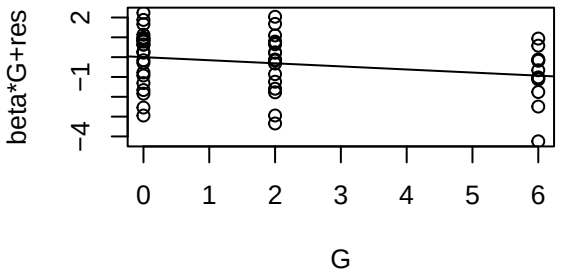
- 變數 G 呈現不顯著
- R^2 只有 62.4%

確認三個變數的 partial residual plots

```
par(mfrow = c(2,2))
prplot(fit3.1,1) ; prplot(fit3.1,2) ; prplot(fit3.1,3)
```







- 變數 *WHP* 看起來有一個開口向上的 mean curvature，可考慮在模型中加入二次項
- 變數 *SP* 有 non-constant variance 的現象，隨著 *SP* 變大，variance 逐漸變小，而且看起來資料點以 fitted line 為界有分群的現象

(b)

在模型中加入變數 *WHP* 的二次項：

```
fit3.2 = update(fit3.1, .~.+I(WHP^2))
summary(fit3.2)
```

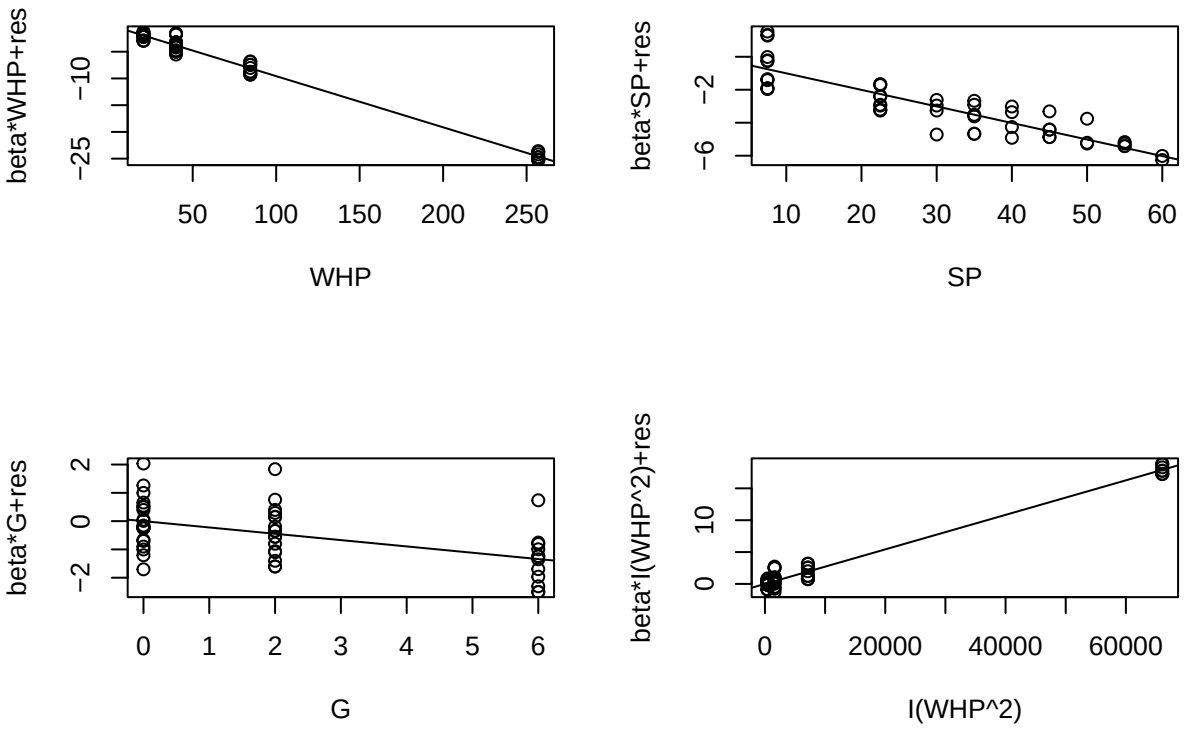
```
##
## Call:
## lm(formula = ACC ~ WHP + SP + G + I(WHP^2), data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70446 -0.64576 -0.05457  0.54006  2.28414
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.011e+01  5.055e-01  19.999  < 2e-16 ***
## WHP         -9.569e-02  9.175e-03 -10.429 1.37e-13 ***
## SP          -1.004e-01  8.188e-03 -12.259 6.08e-16 ***
## G           -2.236e-01  5.689e-02  -3.929  0.00029 ***
## I(WHP^2)     2.710e-04  3.162e-05   8.570 5.18e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9161 on 45 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.8445
## F-statistic: 67.51 on 4 and 45 DF,  p-value: < 2.2e-16
```

- 原本不顯著的變數 G 現在呈現為顯著
- R^2 上升到 85.72%，相較於前一個模型提升非常多

一樣對每個解釋變數繪製 partial residual plots 診斷模型：

```
par(mfrow = c(2,2))
prplot(fit3.2,1)
prplot(fit3.2,2)
prplot(fit3.2,3)
prplot(fit3.2,4)
```



- 變數 WHP 不再有類似二次曲線的 mean curvature
- 變數 SP 的 non-constant variance 現象有所改善，而且資料點不再明顯的被 fitted line 分割成兩群

⇒ 此模型相較於 (a) 中配飾模型來得適合

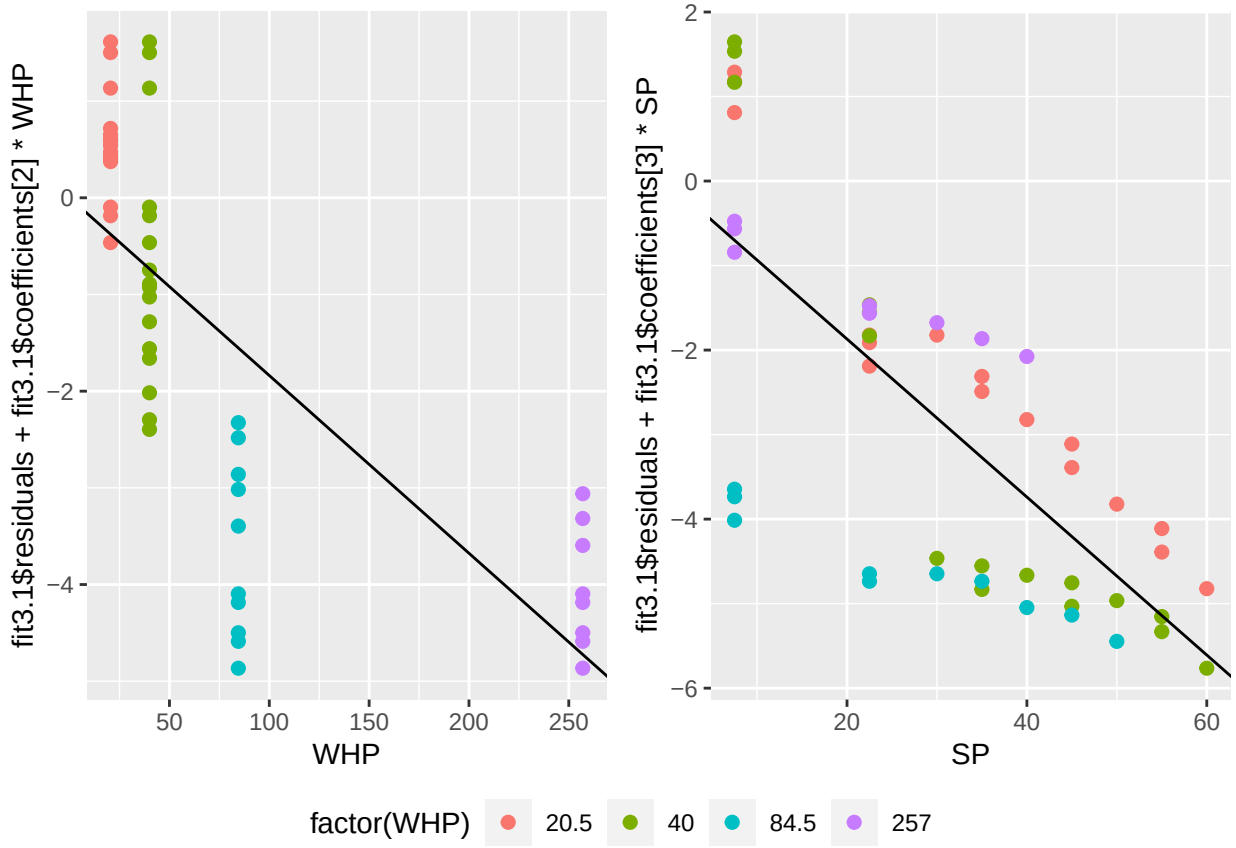
(c)

在 (a) 中變數 SP 的 partial residual plot 有著 non-constant variance 的現象，然而在 (b) 中此現象已被改善，但是我們並沒有對 response variable 採取任何的 transformation 或是利用 weight，這是因為變數 WHP 的二次

效應所造成的誤判：在 (a) 模型 WHP^2 所帶來的變異並沒有被包含在模型的「規律」中，而是被誤植在模型的「隨機」之中，進而造成 SP partial residual plot 有異常的現象；但是在模型 (b) 則是將 WHP^2 所帶來的變異放入模型的「規律」之中，此時 SP partial residual plot 的異常現象就有所改善了。

可以藉由進一步觀察 (a) 中變數 WHP 和 SP partial residual plots 來驗證上述論點：

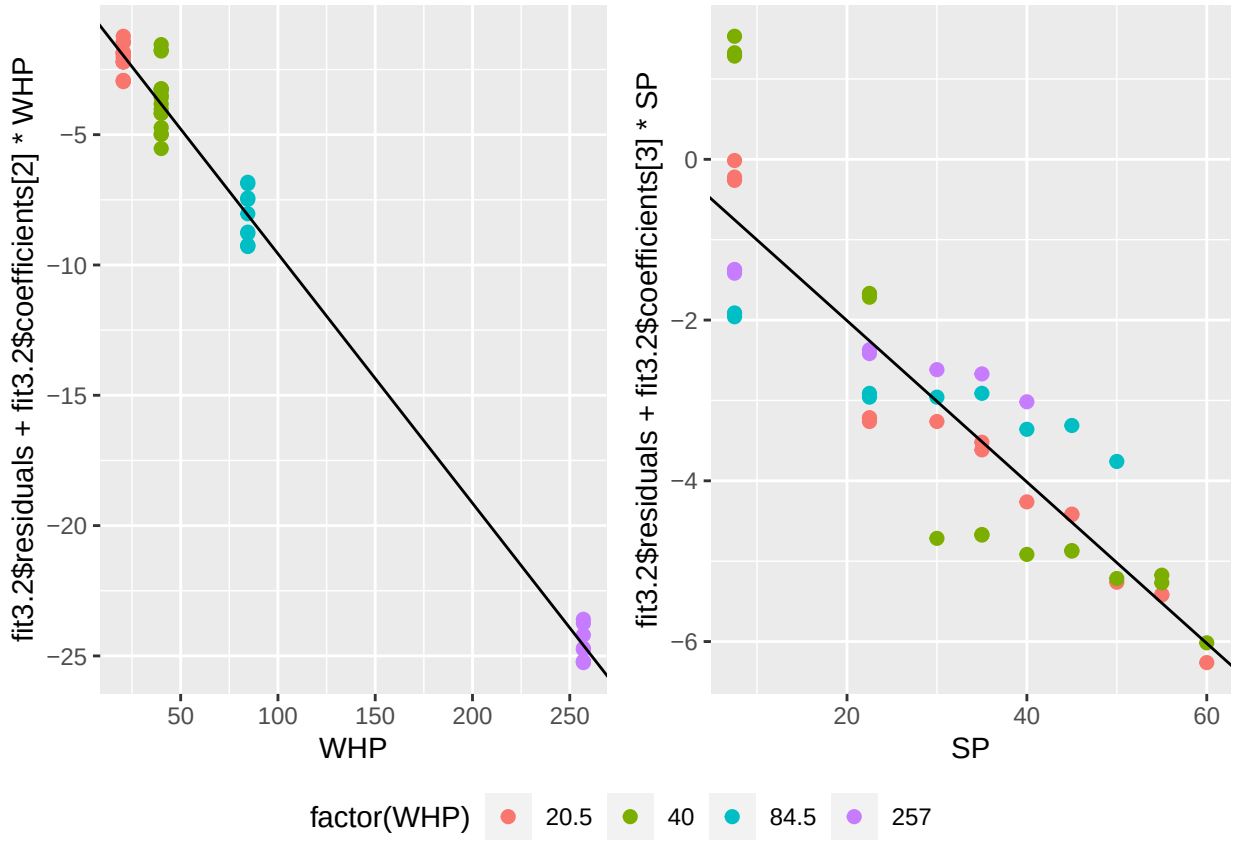
```
library(ggplot2)
library(ggpubr)
p1 = ggplot(data3)+
    geom_point(aes(WHP,fit3.1$residuals+fit3.1$coefficients[2]*WHP,col=factor(WHP)),size=2)+
    geom_abline(slope=fit3.1$coefficients[2],intercept=0)

p2 = ggplot(data3)+
    geom_point(aes(SP,fit3.1$residuals+fit3.1$coefficients[3]*SP,col=factor(WHP)),size=2)+
    geom_abline(slope=fit3.1$coefficients[3],intercept=0)

ggarrange(p1, p2, ncol = 2, common.legend = T, legend = "bottom")
```



可以很清楚的看到 SP partial residual plot 中分割成的上下兩群分別是 $WHP = 257$ or 20.5 以及 $WHP = 40$ or 84.5 兩群，前者的 residual 大多為正，後者的大多為負，而造成此現象就是因為尚未將 WHP^2 考慮進模型的規律之中 (左圖有一開口向上的二次曲線 mean curvature)，這也是使得 SP partial residual plot 有著 non-constant variance 的原因之一。

```
p3 = ggplot(data3)+
    geom_point(aes(WHP,fit3.2$residuals+fit3.2$coefficients[2]*WHP,col=factor(WHP)),size=2)+
    geom_abline(slope=fit3.2$coefficients[2],intercept=0)

p4 = ggplot(data3)+
```

```
    geom_point(aes(SP,fit3.2$residuals+fit3.2$coefficients[3]*SP,col=factor(WHP)),size=2)+
    geom_abline(slope=fit3.2$coefficients[3],intercept=0)

ggarrange(p3, p4, ncol = 2, common.legend = T, legend = "bottom")
```



⇒ 將 WHP^2 放入模型後上述現象得到了明顯的改善。