HW 6-Linear Model

ID : 111024517　　　Name : 鄭家豪

due on 12/15

Problem 1

Read data and fit linear model

這裡對於原資料檔的讀取進行一些調整，好方便讀取，並計算題目所需的 response variable(100 ×(Y84-Y83)/Y83)，令其名稱為”**increase**”:

```
data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/salary.txt",
                   header = T,fill = T)
colnames(data) <- c(names(data)[2:7]," ")
data <- data[,-7]
increase <- 100*(data$Y84-data$Y83)/data$Y83
data <- cbind(increase,data)
```

先配適一個 linear model: $y = X\beta + \epsilon$,   where $\epsilon \sim N(0, \sigma^2 I)$

```
fit <- lm(increase~.-Y84-Y83,data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = increase ~ . - Y84 - Y83, data = 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
```

```r
shapiro.test(fit$res)
```

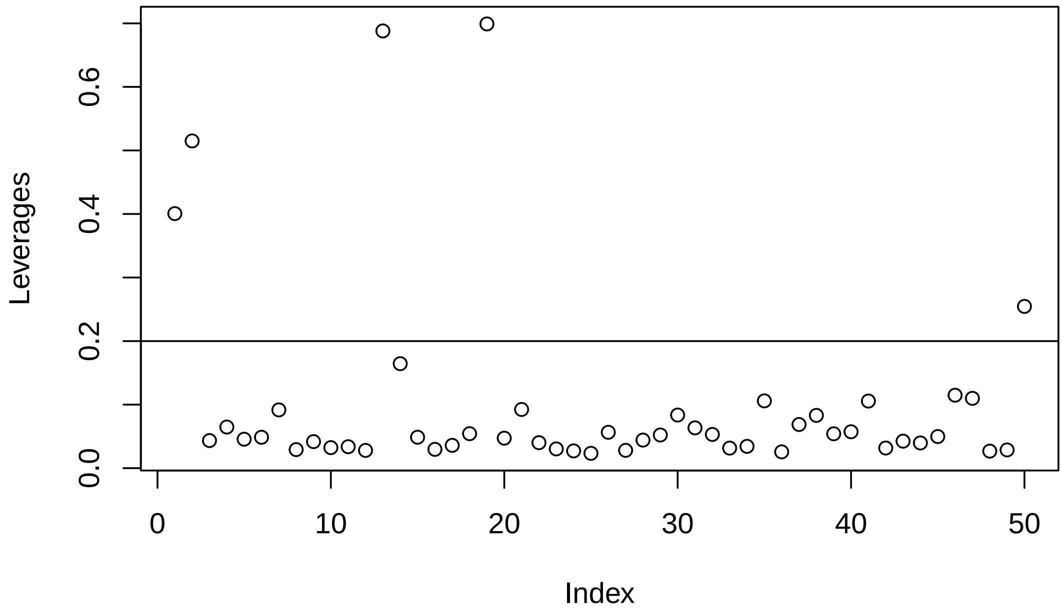
```
##
##  Shapiro-Wilk normality test
##
## data:  fit$res
## W = 0.823, p-value = 3.101e-06
```

這裡得出，不符合 normality 假設，我們做 Diagonstics 找出問題所在。

Diagonstics(Leverage)

根據”rule of thumb” 來找出哪些資料具有 Leverage 大的性質:

```r
x <- model.matrix(fit)
lev <- hat(x)
plot(lev,ylab = "Leverages")
abline(h=2*5/50)
```




```
which(lev>2*5/50)
```

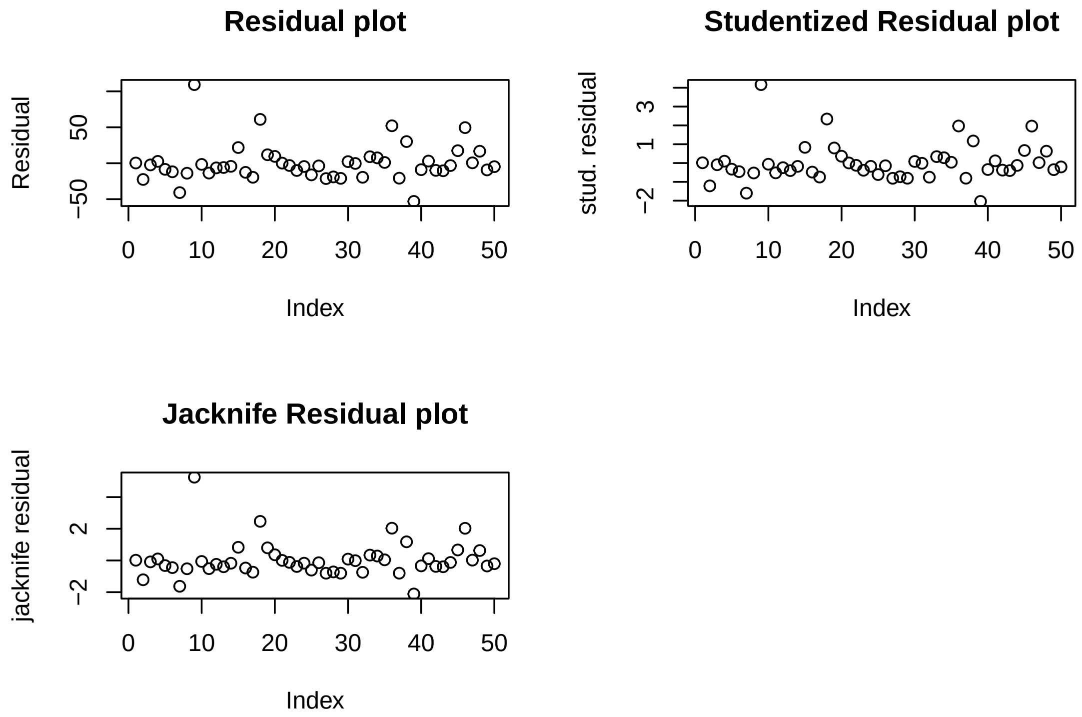
```
## [1]  1  2 13 19 50
```

這裡，我們得知第 1、2、13、19 以及 50 筆資料具有大的 leverage。

接著我們來診斷 outlier

Diagonstics(outlier)

要找出在此模型下的 outlier，這裡我們呈現”raw residual”、“studentized residual”、“jacknife residual”:



可以發現，這三張圖呈現的 pattern 都很相似，且可以發現在第 1~ 第 10 筆觀察值之間，會存在明顯的 outlier。我們使用 multiple testing(H_0 ：no outlier in the n observations v.s. H_1 ：at least one outlier) 來鑑別是否存在 outlier(reject H0 if $|t_i| > t_{n-p-1}(\alpha/2n)$):

```
unique(which(abs(rstudent(fit)) > qt(1-0.05/(2*50),df=50-5-1)))
```

```
## [1] 9
```

這裡結果顯示出:

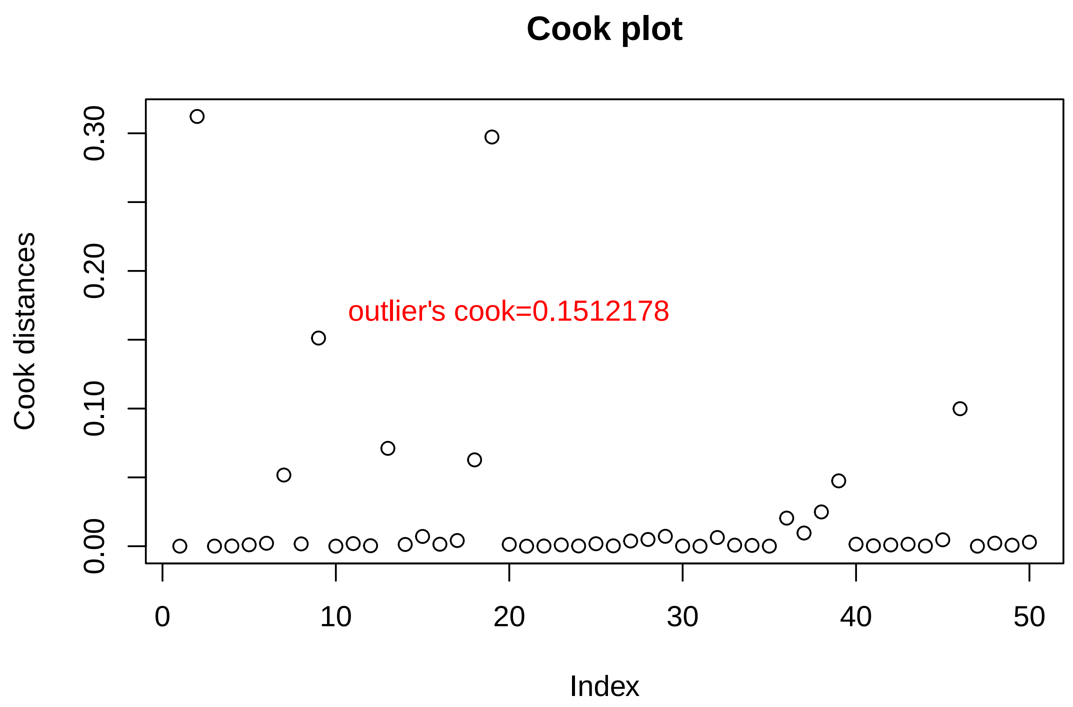
在 $\alpha = 0.05$ 之下，這組資料存在離群值，其中第 9 筆觀察值是絕對值數值上最明顯的 outlier。

接著我們檢查是否有 influential observation。

Diagonstics(influential)

因為 Cook's statistics/distances(scale and unit free) 是 residual 和 leverage 的線性組合，我們使用其來檢驗哪些觀察值是 influential observation:

```
cook <- cooks.distance(fit)
plot(cook,ylab="Cook distances",main = "Cook plot")
text(x=20, y=cook[9]+0.02,
     labels = c("outlier's cook=0.1512178"),
     col="red")
```



這裡很明顯看出，沒有任何一個 Cook's statistics 是比 1 還要大，即使是 outlier 也一樣，我們來觀察 outlier 的 residual 值和 leverage 各是多少:

```
fit$residuals[9]
```

```
##        9
## 109.3221
```

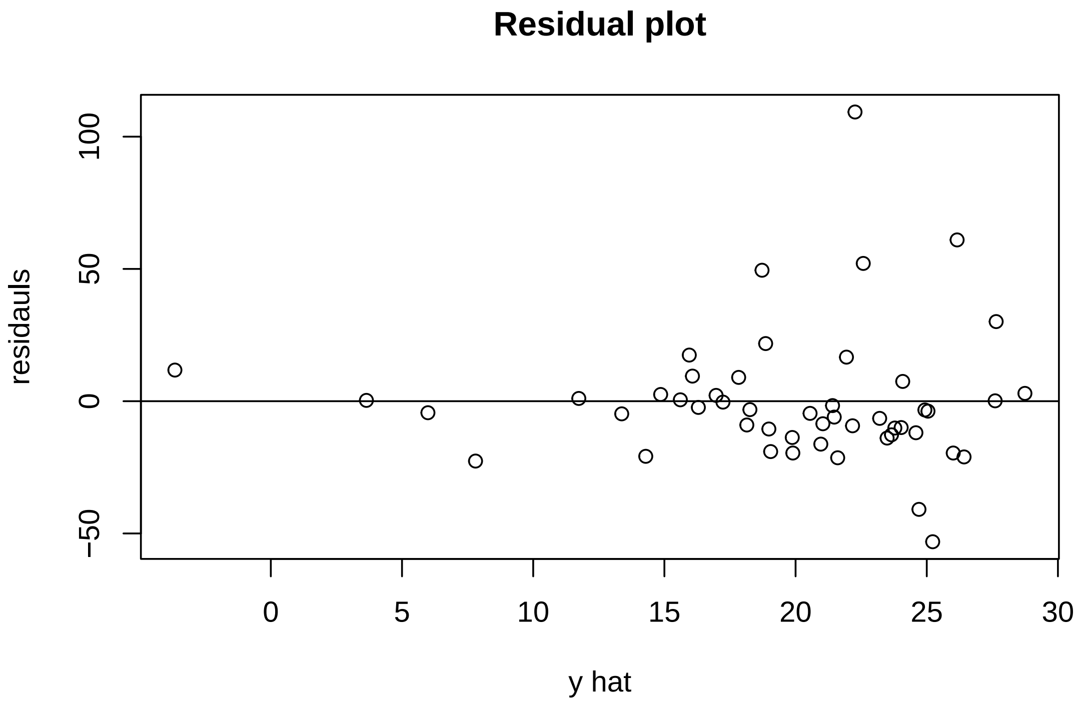
```
lev[9]
```

```
## [1] 0.0417519
```

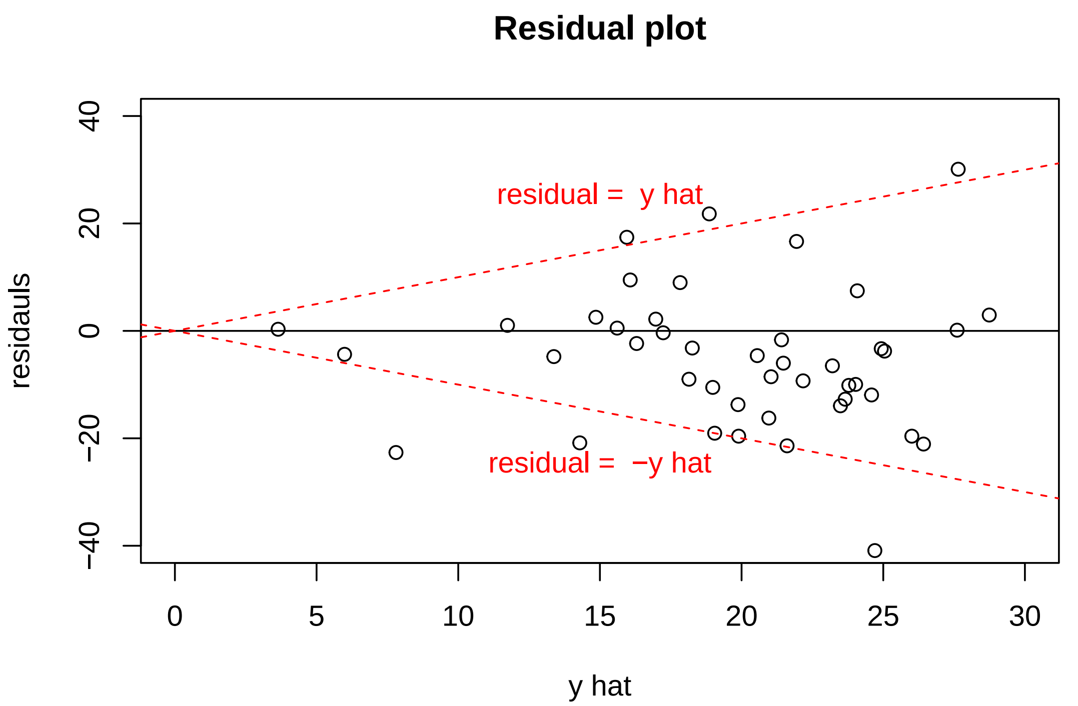
雖然其 residual 值很大，但 leverage 很小，所以 cook's distance of outlier 自然就不會很大。
因為這對於 fitting model 影響不大，這裡我不考慮將其 outlier 給移除掉。

我們來觀察 Residual plot，來看整體的 pattern。

Residual plot



有些點對於觀察 residaul to $\hat{y}$ 來說很礙事，我們將其圖聚焦於比較多點集聚的地方，放大觀察:



雖然有些許點是在紅線外，不過這裡很明顯觀察出，大部份的點是遵循 "$|y| = x$" 的形式變化的，代表為 non-constant variance。為了使之 constant variance ，根據 LNp.7-11 ，$var(y_i) \propto [E(y_i)]^2$，因此進行 $y_i \rightarrow \log(y_i)$ 的轉換。

```r
min(data$increase)
```

```
## [1] -27.90698
```

在進行轉換前，先檢查 y_i 的值是否都是大於 0，由於最小值 = -27.90698，因此需要做平移使得 log 轉換成立。但在做平移前，我們先檢查要平移的量值 (27.91) 所佔 Range of response 的比例:

```r
27.91/(max(data$increase) - min(data$increase))
```

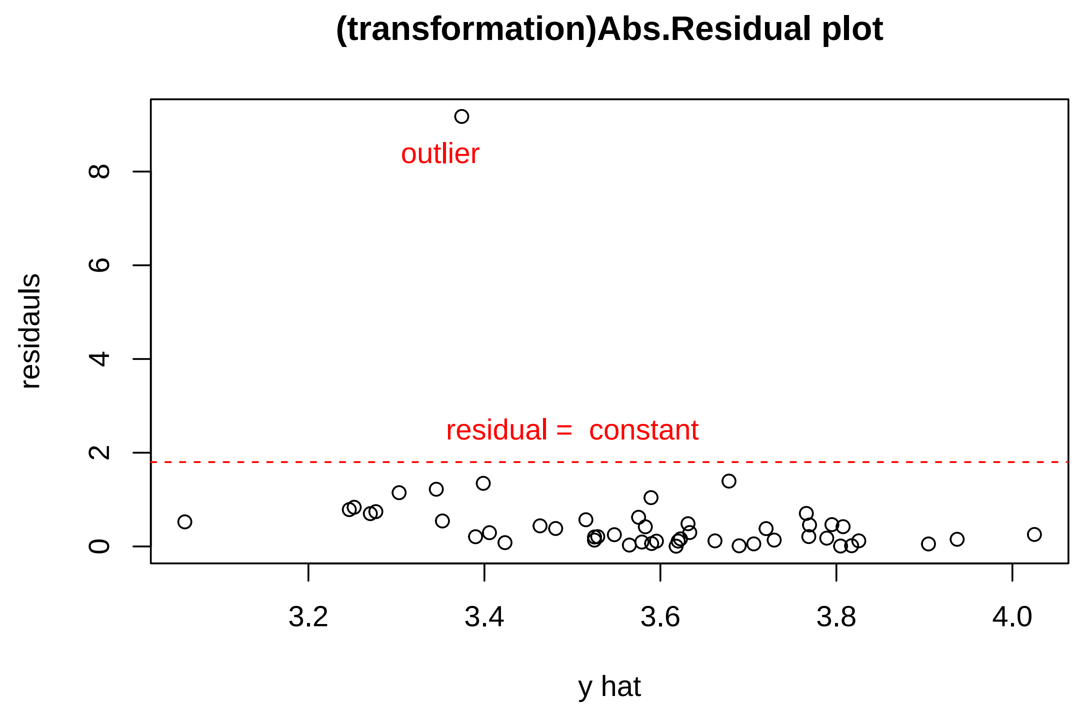
```
## [1] 0.1749902
```

得出平移後，所占的比例約為 17.5% ，以比例來看不算很大的平移。由於不清楚資料背景的意義，假設其平移對 Response 不會造成太大的影響，建立以下模型:

$$\log(y + 27.91) = X\beta + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2 I)$$

```r
fit_transformation <- lm(log(increase+27.91) ~ . -Y84 -Y83,data = data)
summary(fit_transformation)
```

```
##
## Call:
## lm(formula = log(increase + 27.91) ~ . - Y84 - Y83, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1757 -0.1316  0.1133  0.4533  1.3941
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.000e+00  1.967e+00   1.017    0.315
## SHARES       6.297e-09  2.047e-07   0.031    0.976
## REV         -2.291e-05  4.239e-05  -0.540    0.592
## INC          4.744e-04  9.115e-04   0.521    0.605
## AGE          2.802e-02  3.433e-02   0.816    0.419
##
## Residual standard error: 1.477 on 45 degrees of freedom
## Multiple R-squared:  0.02005,    Adjusted R-squared:  -0.06706
## F-statistic: 0.2301 on 4 and 45 DF,  p-value: 0.92
```



進行 log 轉換之後，雖然其模型解釋能力仍然沒有改善，以及在 Absolutely Residual plot 上有一個點特別突兀，但整體來看比較有 constant variance 的感覺，這達到我們的目的。

Problem 2

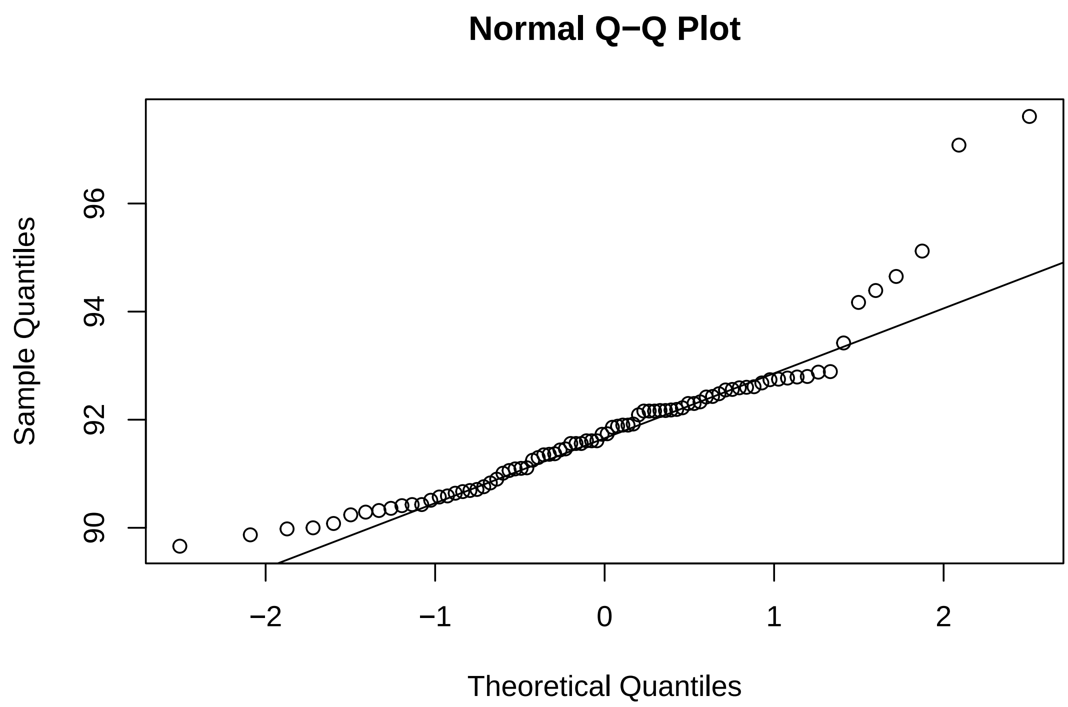
Read data and check normality firstly

```
data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/octane.txt",header
head(data)
```

```
##      A1   A2 A3      A4 rating
## 1 55.33 1.72 54 1.66219  92.19
## 2 59.13 1.20 53 1.58399  92.74
## 3 57.39 1.42 55 1.61731  91.88
## 4 56.43 1.78 55 1.66228  92.80
## 5 55.98 1.58 54 1.63195  92.56
## 6 56.16 2.12 56 1.68034  92.61
```

這組資料的 response variables 為量化連續型資料。

接著檢查 Q-Q plot:



雖然後面與前面的部分有點偏離直線，但整體而言還算是在一條線上，故推測符合 normality 的假設。我們來做 diagonstics 來驗證是否有不正常的狀況:

Diagonstics(Leverage)

先配適一個 linear model: $y_{\text{rating}} = X\beta + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2 I)$

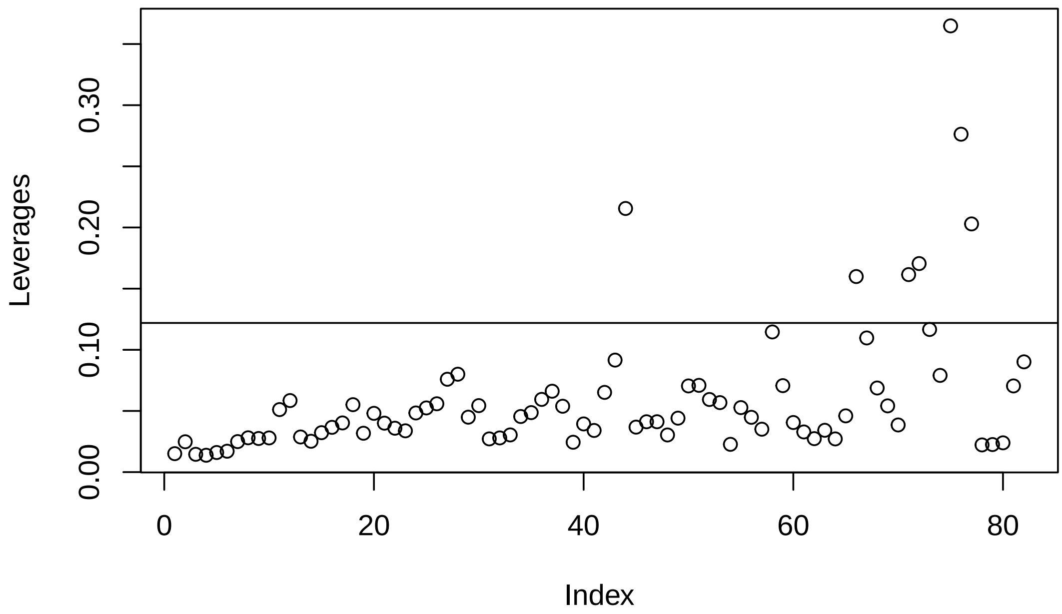
```r
fit <- lm(rating ~ . ,data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = rating ~ ., data = data)
##
## 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
```

根據"rule of thumb" 來找出具有 Leverage 大的資料:

```
x = model.matrix(fit)
lev <- hat(x)
plot(lev,ylab = "Leverages")
abline(h=2*5/82)
```



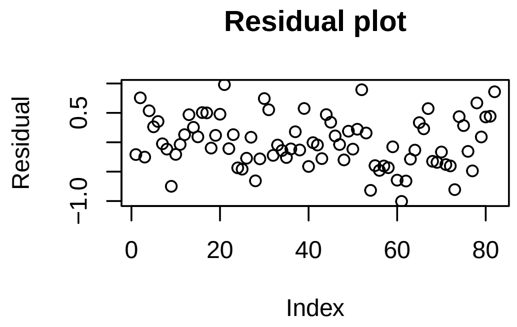

```
which(lev>2*5/82)
```

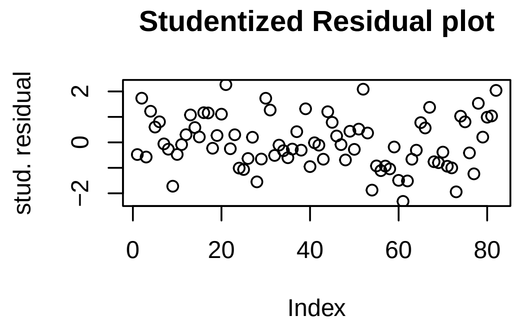
```
## [1] 44 66 71 72 75 76 77
```

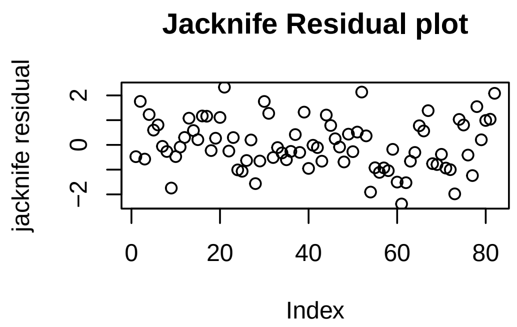
這裡觀察出，第 44, 66,71,72,75~77 筆資料具有 leverage 大的性質。

Diagonstics(outlier)

要找出在此模型下的 outlier，這裡我們呈現"raw residual"、"studentized residual"、"jacknife residual":







乍看之下，感覺沒 outlier，我們用程式來檢驗看看是否存在 outlier(given $\alpha = 0.05$):

$$\mathrm{H_0} : \text{no outlier in the n observations} \quad \text{v.s.} \quad \mathrm{H_1} : \text{at least one outlier}$$

```
unique(which(abs(rstudent(fit)) > qt(1-0.05/(2*82),df=50-5-1)))
```

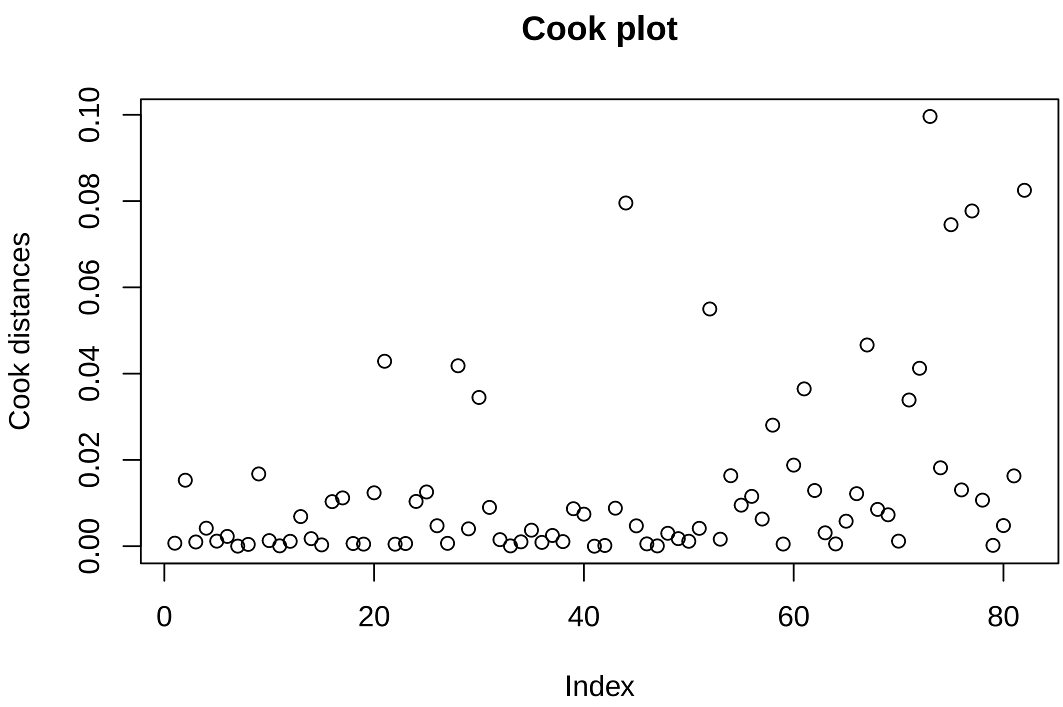
```
## integer(0)
```

其 critical value $= t_{n-p-1}(\alpha/2n) = 3.692514$，由於計算結果取絕對值後沒有超過臨界值，故無法說明這組數據有 outlier。

Diagonstics(influential)

利用 Cook's statsitics/distances 來找出 influential observations:

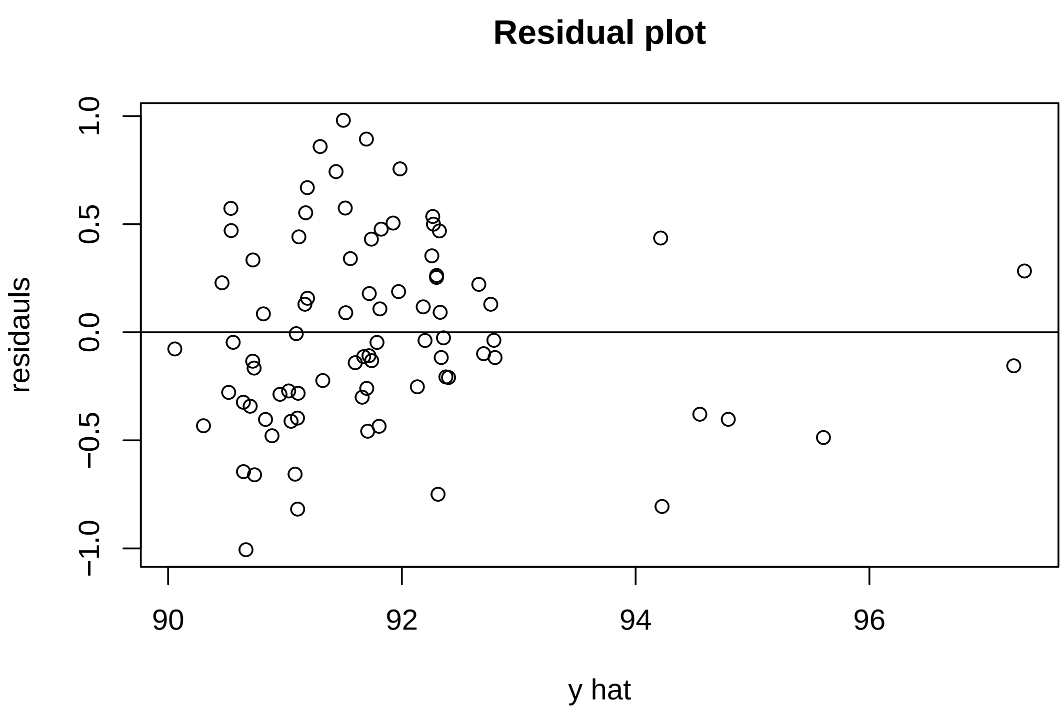
```
cook <- cooks.distance(fit)
plot(cook,ylab="Cook distances",main = "Cook plot")
```



這裡很明顯看出，沒有任何一個 Cook's statistics 是比 1 還要大的，因此沒有 highly influential observation。

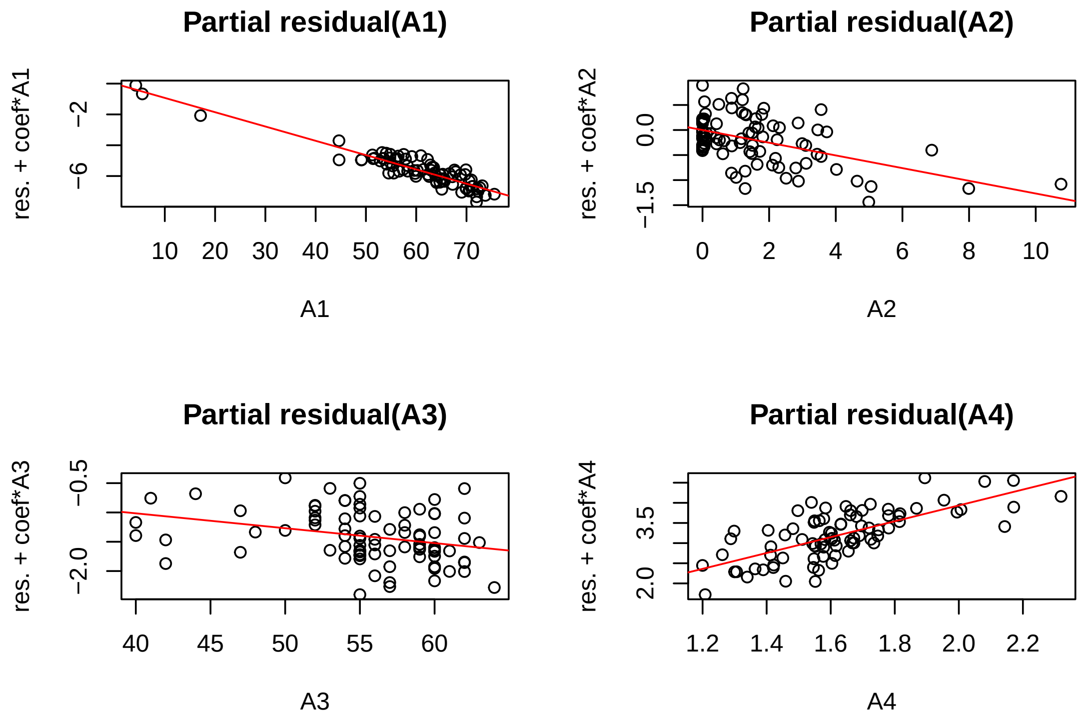
Residual plot

我們來觀察此模型的 residual plot:



就觀察來看，看不太出 non-constant variance 的感覺。

我們來觀察 Partial Residual plot:



從這四張圖來看，感覺沒有很明顯的 mean curvature 的現象。

最後，我們可以用 shapiro.test() 指令來檢驗這模型的常態假設是否顯著:

```
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.98902, p-value = 0.7176
```

其 p-value $= 0.7176 > 0.05$，故不拒絕 H_0: 模型符合 normal。我們這裡並沒有做任何補救措施 (remedy)。

Problem 3

Read data

```
data <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/vehicle.txt",
                   header = T)
```

(a)

Model:

$$Y_{\text{ACC}} = \beta_0 + \beta_1 X_{\text{WHP}} + \beta_2 X_{\text{SP}} + \beta_3 X_{\text{G}} + \epsilon \ , \ \text{where } \epsilon \sim N(0, \sigma^2 I)$$

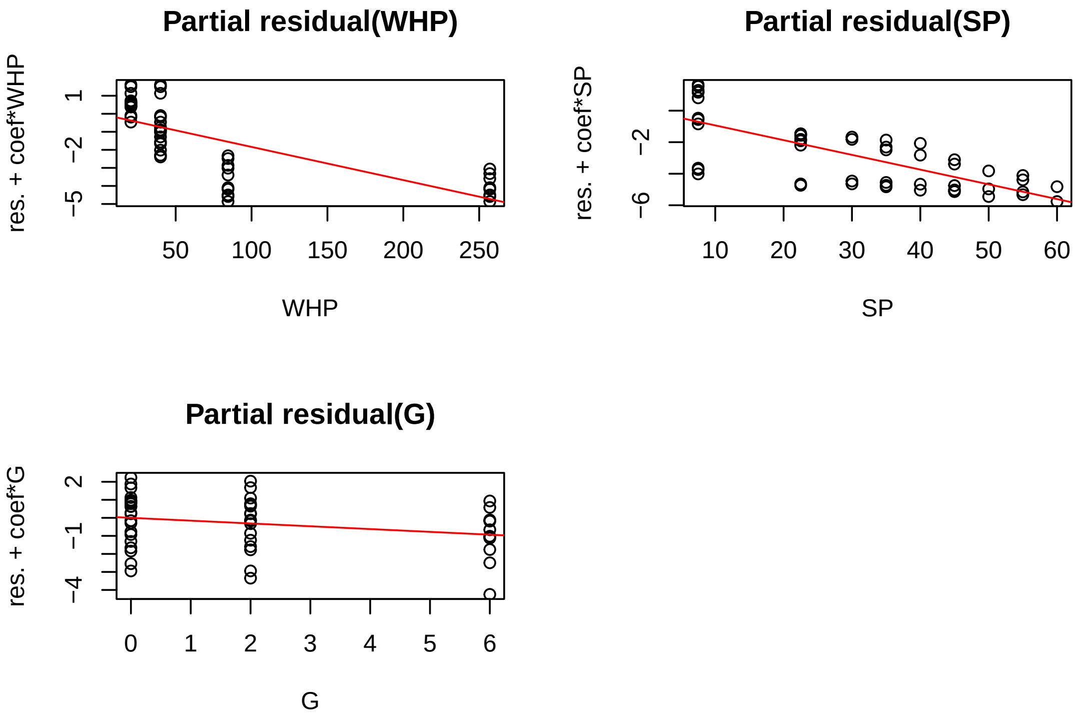
```r
fit <- lm(ACC ~ . ,data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = ACC ~ ., data = data)
##
## 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
```

我們得到模型:

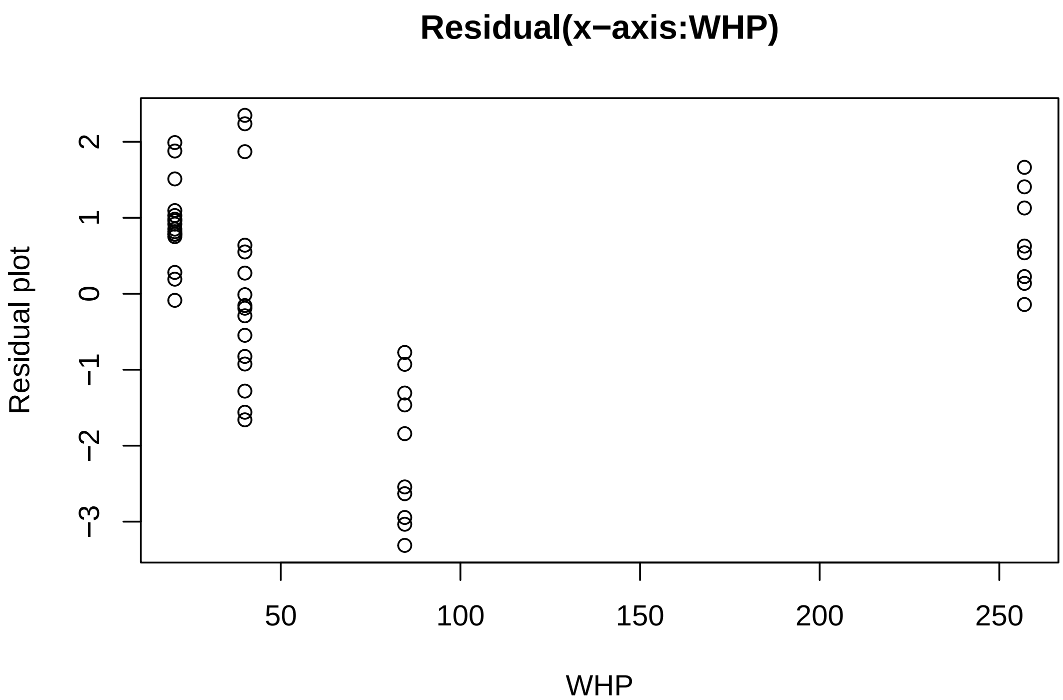
$$\hat{Y}_{\text{ACC}} = 7.19949 - 0.01838 X_{\text{WHP}} - 0.09347 X_{\text{SP}} - 0.15548 X_{\text{G}}$$

Partial residual plot:



(b)

觀察 WHP-residual 之間的關係:



從這裡會發現到，似乎還存在二次項的效應，這裡嘗試增加一個幾變數項: WHP^2

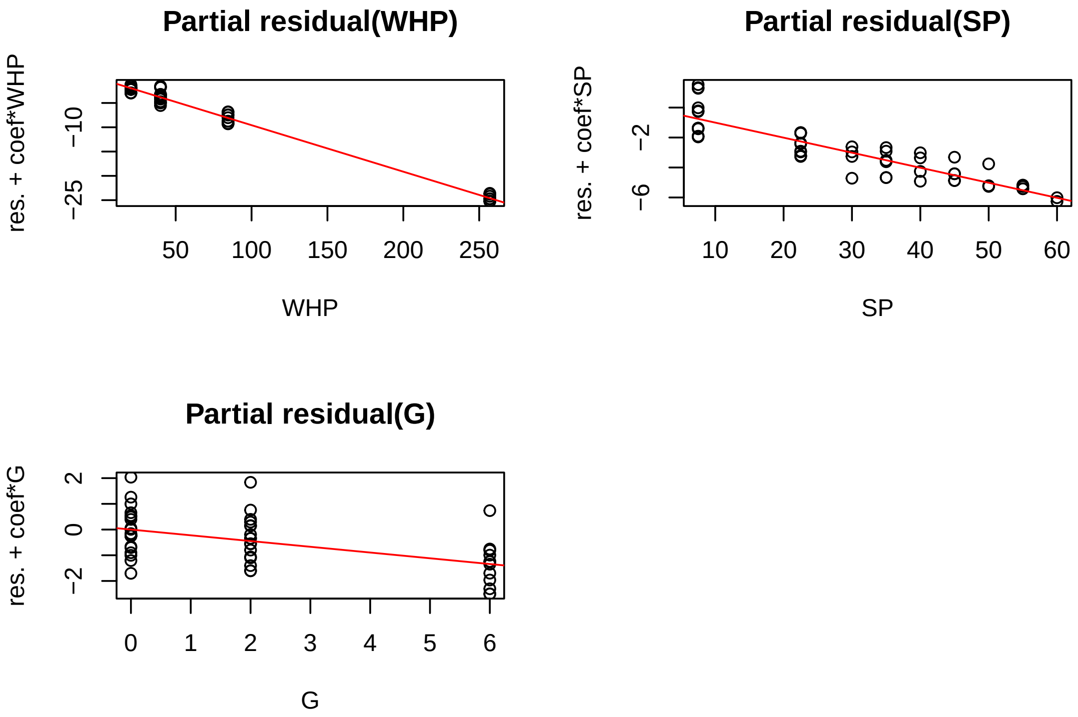
```r
fit_add <- lm(ACC ~ . + I(WHP^2),data = data)
summary(fit_add)
```

```
##
## Call:
## lm(formula = ACC ~ . + I(WHP^2), data = data)
##
## 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
```

得到模型:

$$\hat{Y}_{\text{ACC}} = 10.011 - 0.09569X_{\text{WHP}} - 0.1004X_{\text{SP}} - 0.2236X_{\text{G}} + 2.71 \times 10^{-4} \times \text{WHP}^2$$

接著類似於 (a) ，劃出每個解釋變數對應的 partial residual plot:



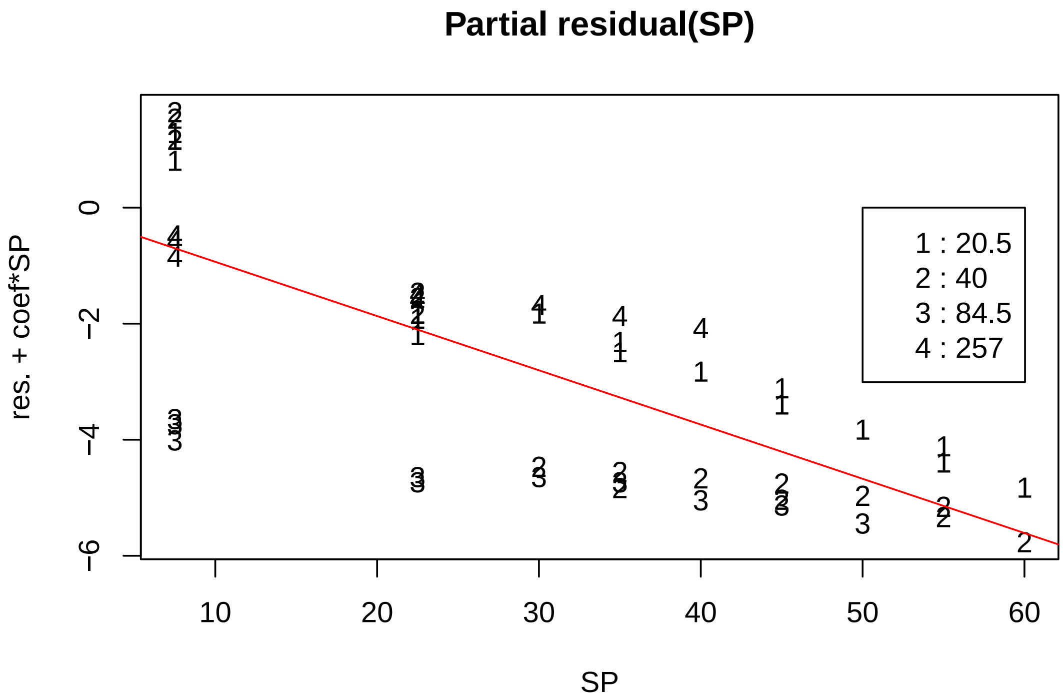
可以發現，加入 WHP^2 項後的新模型，似乎比較合適。

(c)

我們觀察一下 (a) 三個 Partial Residual plot，很明顯地看出 SP 的 Partial Residual，發現 Variance 會隨著 SP 增加而有非遞增的現象，其他兩個 Partial Residual plot 並沒有很明顯的 non-constant variance 狀況。再來加上 (b) 的分析，推測 WHP 可能是解釋這情況的一個關鍵，這裡試著將 WHP 以分群的形式，繪至 SP 的 Partial residual plot 來觀察有甚麼狀況。

```
levels(factor(data$WHP))
```

```
## [1] "20.5" "40"   "84.5" "257"
```



可以發現到，直線上方幾乎都是 WHP = 20.5、257 ; 下方則幾乎是 WHP = 40 、84.5。另外觀察 (b) 中，SP 的 Partial Residual plot，會發現到 non-constant variance 的性質消失，推測原本 error 的部分，包含 WHP^2 的效應，所以在 (b) 加進來 WHP^2 項之後，這個效應就移至模型中規律的部分。另外，考量到 WHP 只有四個值，這裡我們是當連續變數來做分析，這意味著 WHP = 20.5 、257 是相當遠離平均值 (77.38)，故加入 WHP 的二次項會使得，WHP = 最小或最大值時，SP 的 Partial Residual plot 中，會產生極大 residual 的現象將被解決，自然會回到 constant variance 的狀況。