Linear Model Assignment 6

110024516 統研碩一邱繼賢

Problem 1.
匯入資料 salary 並計算 $percentage\ increase\ =\ 100 \times (Y84 - Y83)/Y83$ 存成變數 PI(percentage increase)，建構模型

$$g_1\ :\ PI\ \sim\ Y84\ +\ Y83\ +\ share\ +\ rev\ +\ inc\ +\ age$$

```
library(dplyr)
library(latex2exp)
salary = read.table("salary.txt", skip = 1)
colnames(salary) = c("y84", "y83", "share", "rev", "inc", "age")
salary = salary %>% mutate(PI = 100*(y84-y83)/y83)
g1 = lm(PI ~ share + rev + inc + age, data = salary)
summary(g1)
```

```
##
## Call:
## lm(formula = PI ~ share + rev + inc + age, data = salary)
##
## 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
## share       -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
```

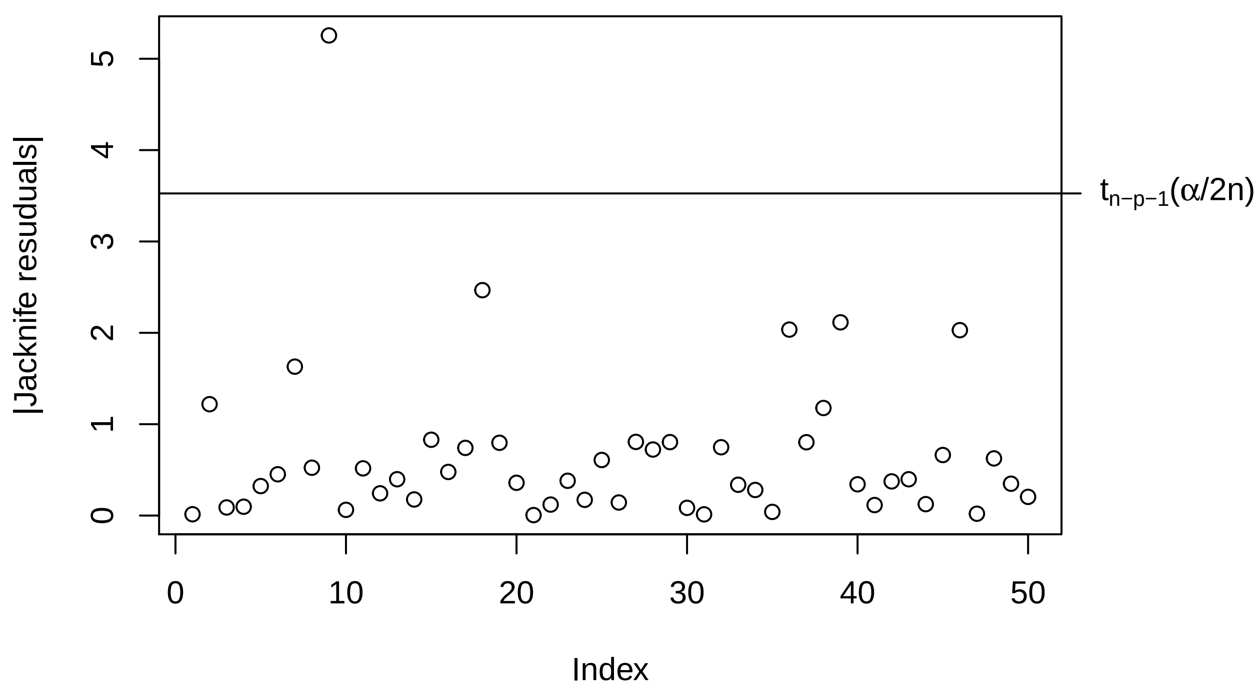
計算 jacknife residuals：

$$t_i\ =\ \frac{\hat{\varepsilon}_i}{\sqrt{1-h_i}\ \hat{\sigma}_{(i)}}\ =\ r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$$

然後用 jacknife residuals 做 multiple test of outlier $\Rightarrow$ conclude an outlier if $|t_i|\ >\ t_{n-p-1}(\alpha/2n)$

```
rjack1 = rstudent(g1)
par(mar = c(5,4,4,5), mfrow = c(1,1))
plot(abs(rjack1), ylab = "|Jacknife resuduals|")
t_value = qt(1-0.05/(2*50), 50-1-5)
```

```r
abline(h = t_value)
axis(4, at = t_value, labels = TeX("t_{n-p-1}($\\alpha$/2n)"), las = 2)
identify(1:50, abs(rjack1), row.names(salary))
```




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

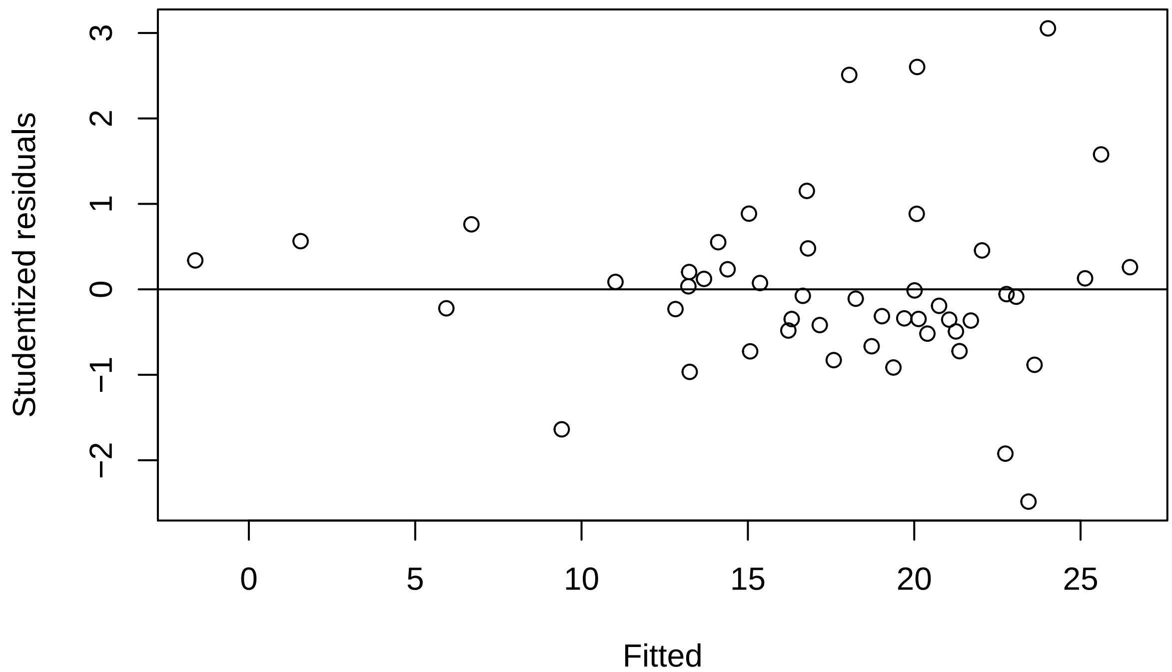
如上圖所示，判定第九個觀測值為 outlier

將 outlier 移除後重新建構模型，然後計算 studentized residuals

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{1-h_i}\,\hat{\sigma}}$$

用來繪製 residual plot

```r
g1.2 = lm(PI ~ share + rev + inc + age, subset = (abs(rjack1)<t_value), data = salary)
rstud1 = rstandard(g1.2)
par(mfrow = c(1,1))
plot(g1.2$fit,rstud1,
     xlab = "Fitted", ylab = "Studentized residuals")
abline(h = 0)
```

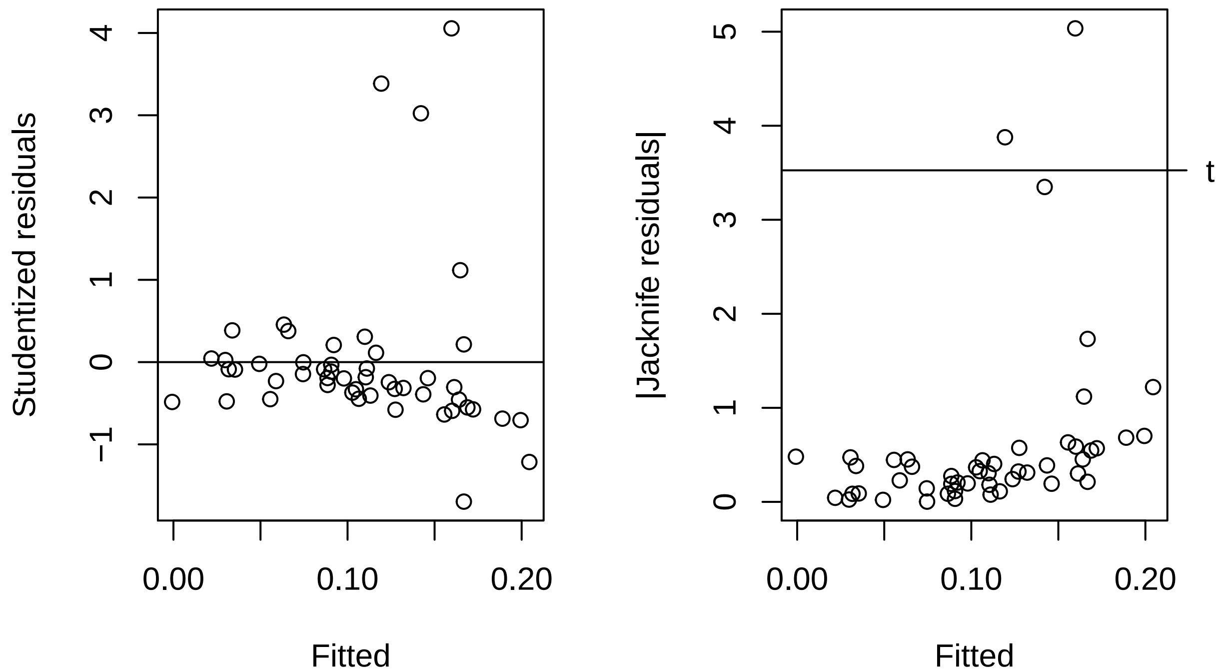


可以觀察出 studentized residuals 有隨著 fitted value 變大而變異增加，將 response variable 變換後重新建構模型

$$g_{1.2}\ :\ \frac{1}{PI+1} \sim Y84\ +\ Y83\ +\ share\ +\ rev\ +\ inc\ +\ age$$

一樣計算新模型的 studentized residual 並繪圖

```
g1.2 = lm(1/(PI+1) ~ share + rev + inc + age, data = salary)
rstud1.2 = rstandard(g1.2)
rjack1.2 = rstudent(g1.2)
par(mfrow = c(1,2))
plot(g1.2$fit, rstud1.2, xlab = "Fitted", ylab = "Studentized residuals")
abline(h = 0)
plot(g1.2$fit, abs(rjack1.2), xlab = "Fitted", ylab = "|Jacknife residuals|")
abline(h = t_value)
axis(4, at = t_value, labels = "t", las = 2)
```



可以發現變換後新模型下大致呈現 constant variance 的現象，只有少數幾個點較為遠離 0 值，但在檢查了他們的 jacknife residual 後可以將那些觀測值視為 outlier。

Problem 2.
建構模型

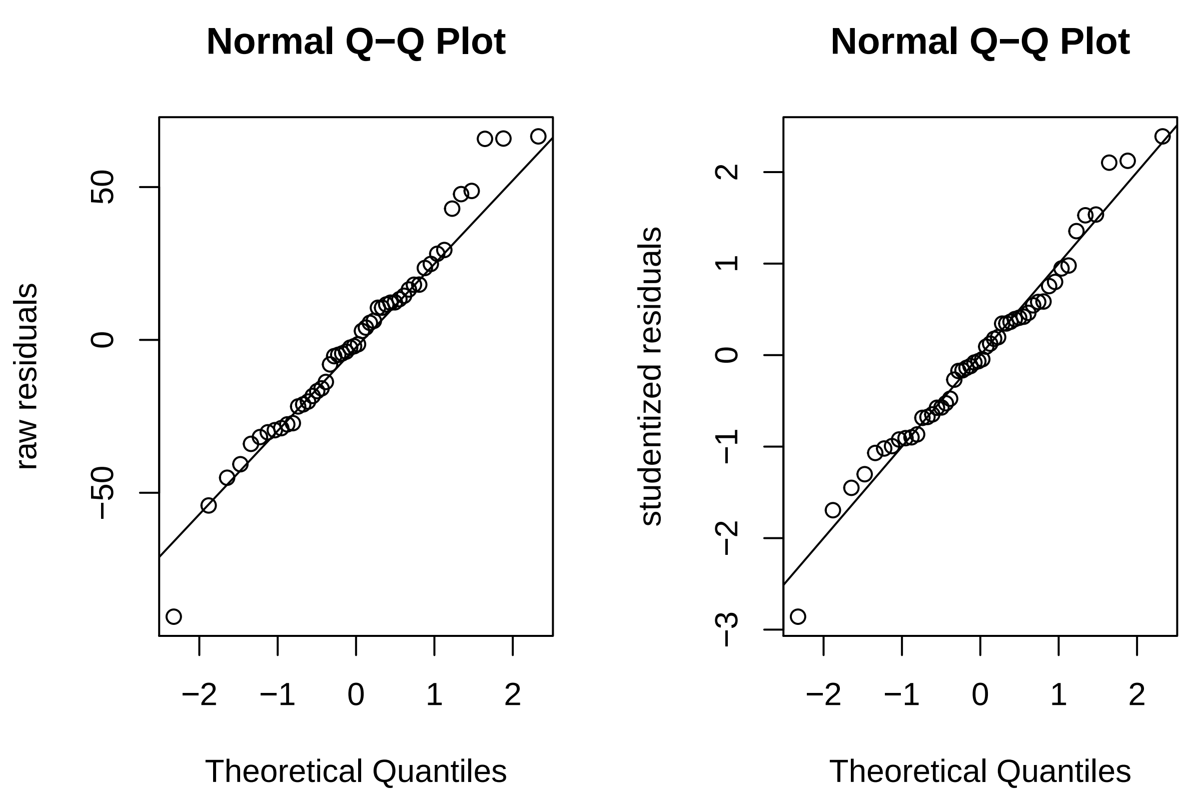
$$g_{2.1}\ :\ total\ \sim\ expend\ +\ salary\ +\ ratio\ +\ takers$$

```
sat = read.table("sat.txt", skip = 1)
colnames(sat) = c("state", "expend", "ratio", "salary", "takers", "verbal", "math", "total")
g2.1 = lm(total ~ expend + salary + ratio + takers, data = sat)
summary(g2.1)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 1045.971536  52.869760  19.7839283 7.857530e-24
## expend         4.462594  10.546528   0.4231339 6.742130e-01
## salary         1.637917   2.387248   0.6861110 4.961632e-01
## ratio         -3.624232   3.215418  -1.1271418 2.656570e-01
## takers        -2.904481   0.231260 -12.5593745 2.606559e-16
```

a.
將 raw residuals 和 studentized residuals 對 normal distribution 做 Q-Q Plot，如下圖所示，可發現各點均大致落在一直線上，故可以推斷此筆數據符合 normality assumption。

```
rstud2 = rstandard(g2.1)
par(mar = c(5,4,4,2), mfrow = c(1,2))
qqnorm(g2.1$res, ylab = "raw residuals")
qqline(g2.1$res)
qqnorm(rstud2, ylab = "studentized residuals")
abline(0,1)
```



b.
計算各觀測值的 leverage：$h_i = H_{ii}$，並與 $2p/n = 2 \times 5/50 = 0.2$ 進行比較，若較大則視為 large leverage point

```
x = model.matrix(g2.1)
lev = hat(x)
par(mar = c(5,4,4,4), mfrow = c(1,1))
plot(lev, ylab = "Leverages", main = "index plot of leverages")
abline(h = 2*5/50)
axis(4, at = 2*5/50, labels = "2p/n", las = 2)
identify(1:50, lev, sat$state)
```

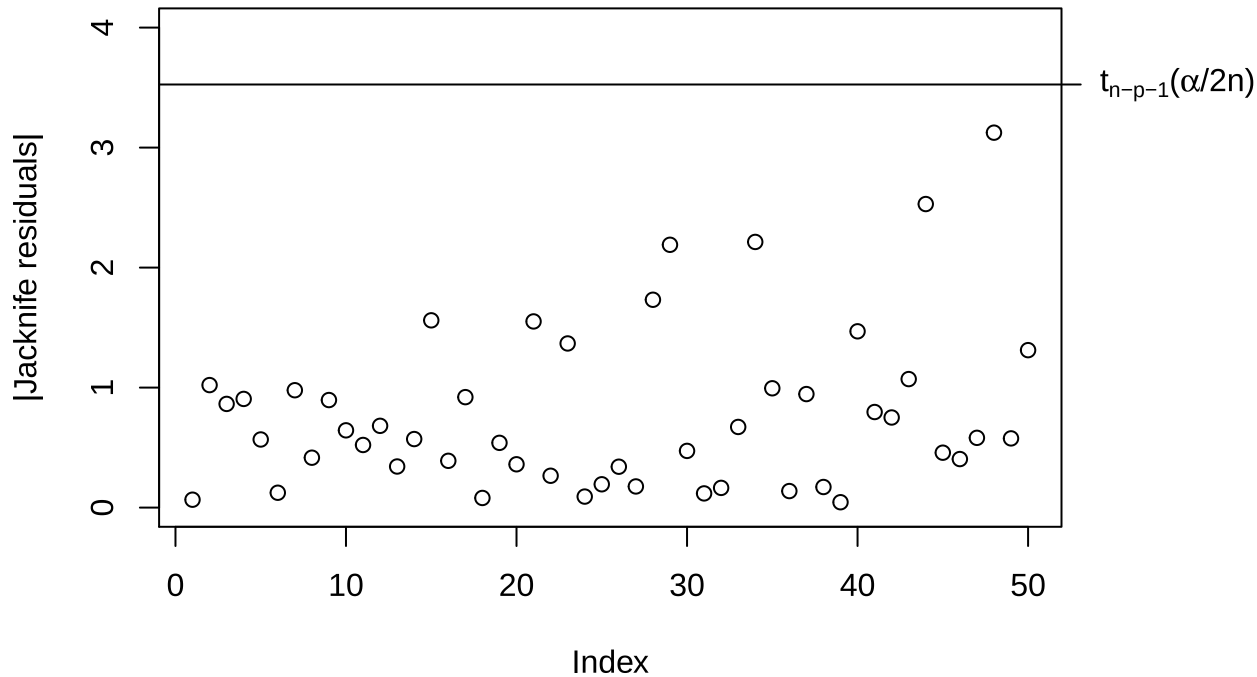

```
## integer(0)
```

由上圖可知，共有四個點超過 $2p/n$，分別是 California, Connecticut, New Jersey, Utah 四州，故推斷它們為 large leverage points。

c.
計算 Jacknife residuals t_i 和 critical value $t_{n-p-1}(\alpha/2n)$ 並做比較，若 $|t_i| > t_{n-p-1}(\alpha/2n)$，則視為 outlier

```
rjack2 = rstudent(g2.1)
par(mar = c(5,4,4,5), mfrow = c(1,1))
plot(abs(rjack2), ylab = "|Jacknife residuals|", ylim = c(0,4))
t_value2 = qt(1-0.05/(2*50), 50-5-1)
abline(h = t_value2)
```

```
axis(4, at = t_value2, labels = TeX("t_{n-p-1}($\\alpha$/2n)"), las = 2)
identify(1:50, abs(rjack2), sat$state)
```




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

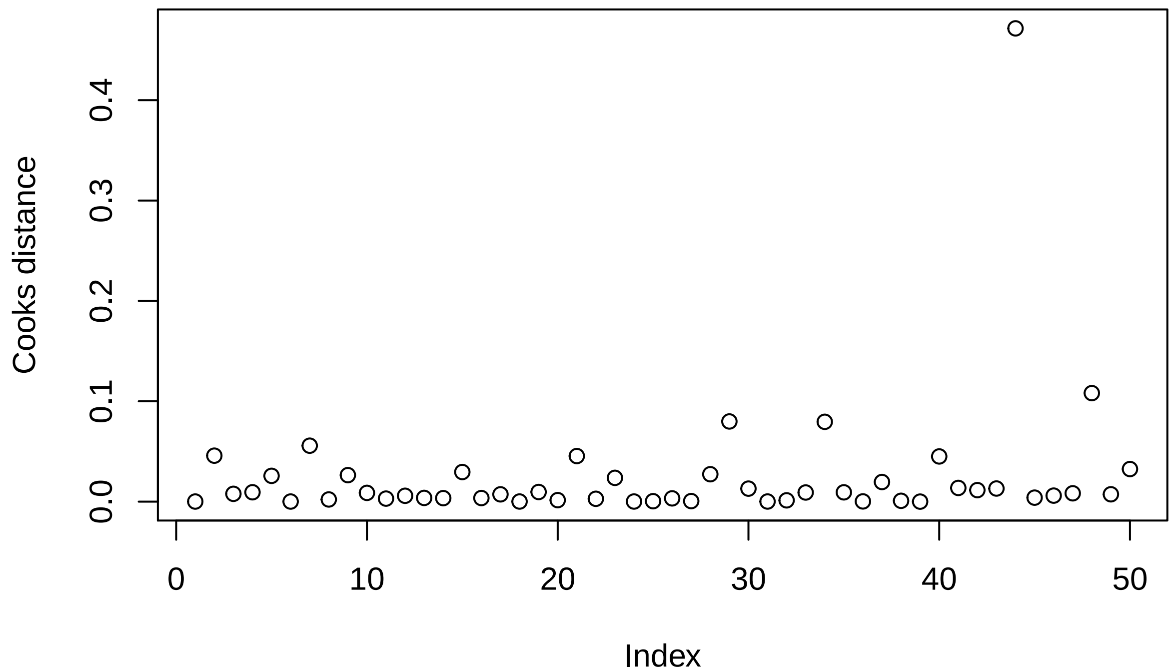
如上圖所示，所有觀測值的 |Jacknife residual| 皆小於 critical value，故推斷此筆數據沒有 outlier。
但事實上，Bonferroni critical value $t_{n-p-1}(\alpha/2n)$ 是一種非常保守的多重檢定 critical value ，而上圖之中，West Virginia, Utah 兩州的 Jacknife residuals 也都已經相當接近該數值了，所以它們也有機會被判定為 outlier。

d.

計算各觀測值的 Cook's statistics/distances

$$D_i = (\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X)(\hat{\beta} - \hat{\beta}_{(i)})/(p\ \hat{\sigma}^2) = (1/p)\ r_i^2\ (h_i/(1-h_i))$$

```
cook = cooks.distance(g2.1)
plot(cook, ylab = "Cooks distance")
identify(1:50, cook, sat$state)
```




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

由上圖可發現，有一點的 Cook's statistic 數值特別高，就是 Utah，而且從 b. 和 c. 兩題可知，因為 Utah 的 leverage 和 |Jacknife residual| 數值都偏高，所以 Cook's statistic 也理所當然地較大，故我們可以推斷 Utah 就是 influential point。
然後將 Utah 此筆資料移除後，重新建構新模型，並計算 $(\hat{\beta} - \hat{\beta}_{(i)})/\hat{\beta}$ 比較新舊模型的係數變化：

```
g2.d = lm(total ~ expend + salary + ratio + takers,
          subset = (cook<0.3), data = sat)
summary(g2.d)$coef
```

```
##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 1093.8459730 53.4225501  20.47536050 4.042501e-24
## expend        -0.9427394 10.1921677  -0.09249645 9.267235e-01
## salary         3.0964294  2.3282862   1.32991787 1.903986e-01
## ratio         -7.6391442  3.4279050  -2.22851688 3.100332e-02
## takers        -2.9308044  0.2187703 -13.39671685 3.945690e-17
```

```
(summary(g2.1)$coef[2:5,1]-summary(g2.d)$coef[2:5,1])/summary(g2.1)$coef[2:5,1]
```

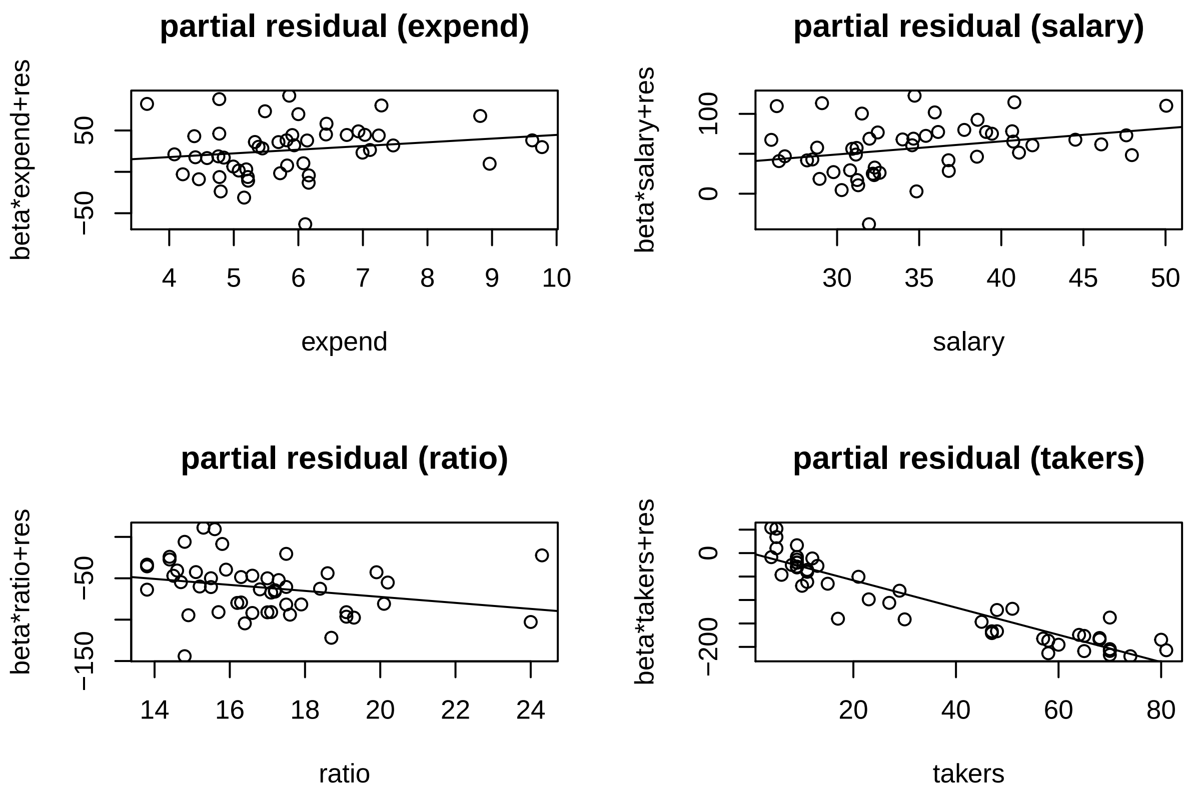
```
##       expend       salary        ratio       takers
##  1.211253663 -0.890467644 -1.107796537 -0.009063192
```

和原模型的係數比較發現，除了 $takers$ 的係數沒有太大的變化之外，其餘三變數的係數都有滿明顯的變化，特別是 $expend$ 的係數，從原本的 4.4626 變成 -0.9427，代表僅僅移除 Utah 一點後，$expend$ 對 $total$ 從正相關變成了負相關，可見此點的存在與否對於整個回歸模型有很明顯的影響。

e.

對 $model\ g_{2.1}$ 的各變數繪製 partial residual plots：

```
prplot <- function(g,i)
{
  library(latex2exp)
# Partial residuals plot for predictor i
  xl<-attributes(g$terms)$term.labels[i]
  yl<-paste("beta*",xl,"+res",sep="")
  m = paste("partial residual (",xl,")", sep = "")
  x<-model.matrix(g)[,i+1]
  plot(x,g$coeff[i+1]*x+g$res,xlab=xl,ylab=yl,
       main = m)
  abline(0,g$coeff[i+1])
  invisible()
}
par(mfrow = c(2,2))
prplot(g2.1,1)
prplot(g2.1,2)
prplot(g2.1,3)
prplot(g2.1,4)
```



可觀察到變數 $takers$ 將資料明顯的區分為兩群，將資料區分為 $takers < 40$ 和 $takers > 40$ 分別建構回歸模

型並比較：

```r
g2.2 = lm(total ~ expend + salary + ratio + takers, subset = (takers < 40), data = sat)
summary(g2.2)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 993.717753 84.5009900 11.7598356 5.855759e-11
## expend        7.758139 16.4328738  0.4721109 6.414969e-01
## salary        1.029256  3.3057718  0.3113511 7.584649e-01
## ratio         1.425143  4.6110901  0.3090686 7.601774e-01
## takers       -5.524231  0.8706121 -6.3452260 2.194034e-06
```

```r
g2.3 = lm(total ~ expend + salary + ratio + takers, subset = (takers > 40), data = sat)
summary(g2.3)$coef
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 801.4329379 105.6773368  7.5837730 5.200614e-07
## expend       11.1443785  10.8359315  1.0284652 3.173539e-01
## salary       -0.6354384   2.7190282 -0.2337006 8.178547e-01
## ratio         3.9147437   4.8627271  0.8050511 4.312964e-01
## takers       -0.3003496   0.8868563 -0.3386677 7.387786e-01
```

```r
(summary(g2.2)$coef[2:5,1]-summary(g2.3)$coef[2:5,1])/summary(g2.2)$coef[2:5,1]
```

```
##     expend     salary      ratio     takers
## -0.4364758  1.6173766 -1.7469128  0.9456305
```

四變數的係數都有滿明顯的變化，特別是 *salary* 的係數正的變成負的，由此可知以變數 *takers* 大小所區分的兩群資料，在 preditors 和 response 的關係上有著結構性的差別。

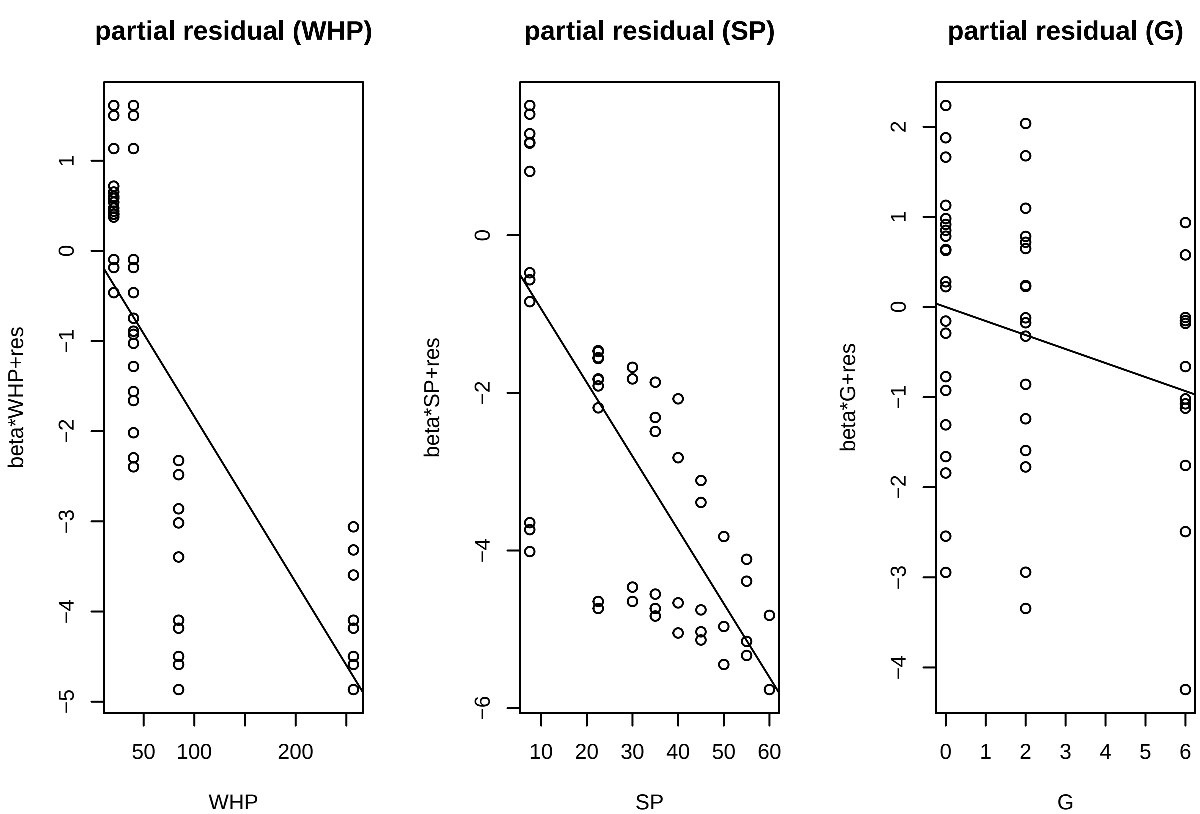
Problem 3.

a. 建構模型

$$g_3 \ : \ ACC \ \sim \ WHP \ + \ SP \ + \ G$$

然後對三個變數分別做 partial residual plots :

```
vehicle = read.table("vehicle.txt", skip = 1)
colnames(vehicle) = c("ACC","WHP","SP","G")
g3 = lm(ACC ~ WHP+SP+G, data = vehicle)
par(mfrow = c(1,3))
prplot(g3,1)
prplot(g3,2)
prplot(g3,3)
```



(1) WHP 的 partial residual plot 似乎存在著 mean curvature，代表只有一次項的 WHP 還不足以解釋此筆資料。

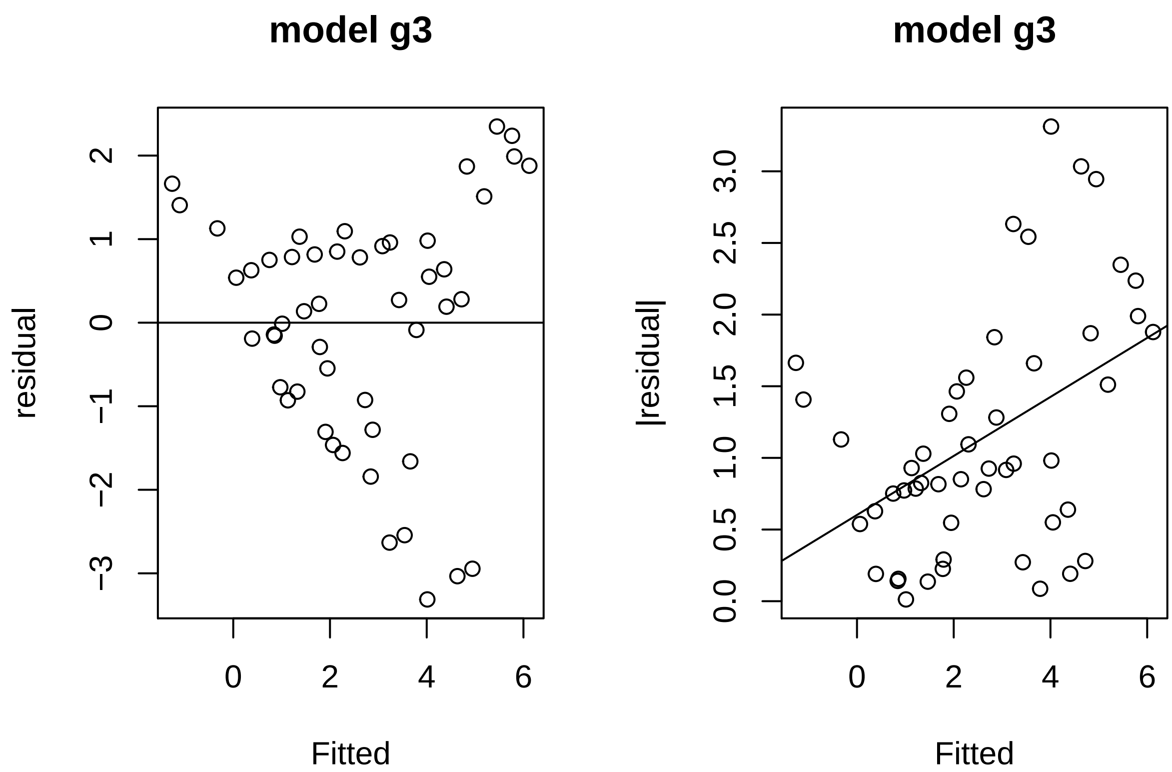
(2) SP 的 partial residual plot 的變異程度隨著 SP 增加而遞減。

b.

繪製 *model* g_3 的 residual plot 和 absolute residual plot :

```
par(mfrow = c(1,2))
plot(g3$fit, g3$res, xlab = "Fitted", ylab = "residual", main = "model g3")
abline(h = 0)
```

```r
plot(g3$fit, abs(g3$res), xlab = "Fitted", ylab = "|residual|", main = "model g3")
abline(lm(abs(g3$res)~g3$fit))
```




```r
summary(lm(abs(g3$res)~g3$fit))$coef
```

```
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 0.6018234 0.18761225 3.207805 0.002383428
## g3$fit      0.2058917 0.05905973 3.486161 0.001057974
```

(1) residual plot 呈現隨著 fitted value 上升而變異增加的趨勢 (推測關係為 $var(y_i) \propto [E(y_i)]^2$)。

(2) absolute residual plot 也有呈現出正相關的趨勢。

(3) 建構 $|residual| \sim fitted\ value$ 回歸模型，其回歸線斜率為正值且 pvalue 結果亦為顯著不等於零。

$\Rightarrow$ 推斷此筆數據具有 non-constant variance，再加上 a. 小題最後所做出的結論，可對模型做出以下改進：

(i) 加入變數 WHP^2

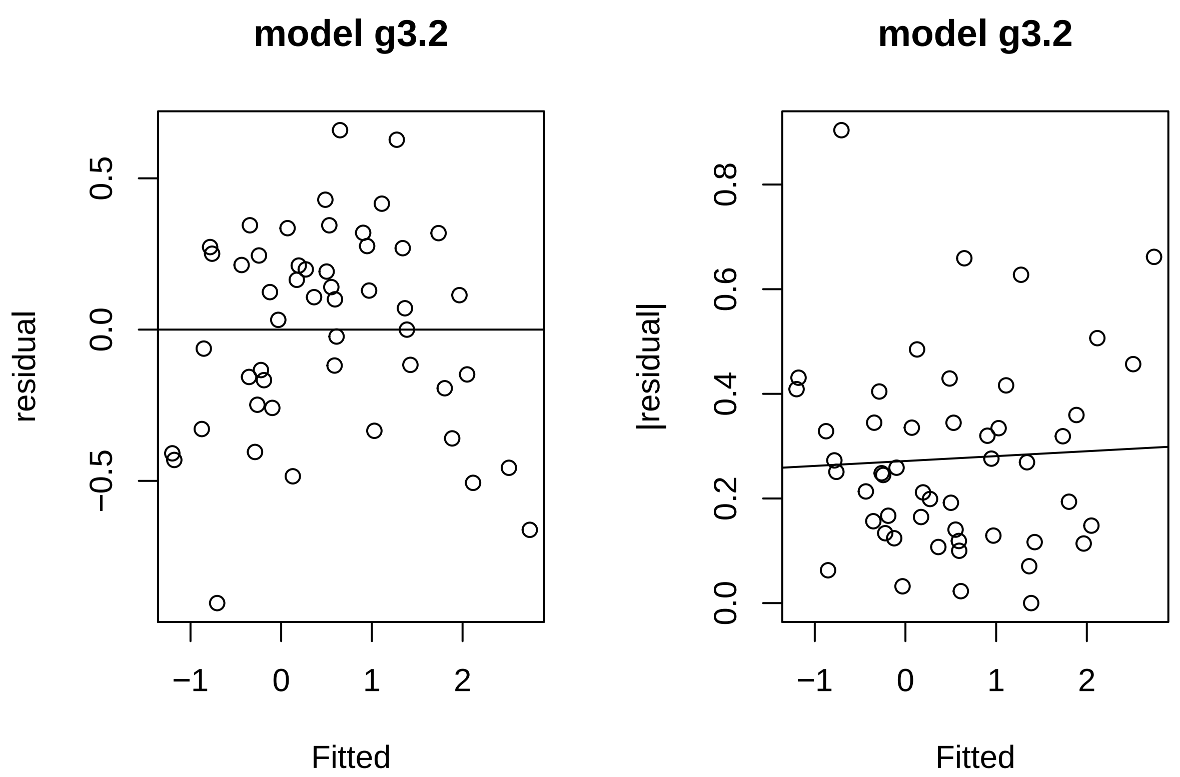
(ii) response variable ACC 改為 $log(ACC)$

建構新模型：

$$g_{3.2}\ :\ log(ACC)\ \sim\ WHP\ +\ WHP^2\ +\ SP\ +\ G$$

一樣對其繪製 residual plot 和 absolute residual plot 以及 partial residual plots :

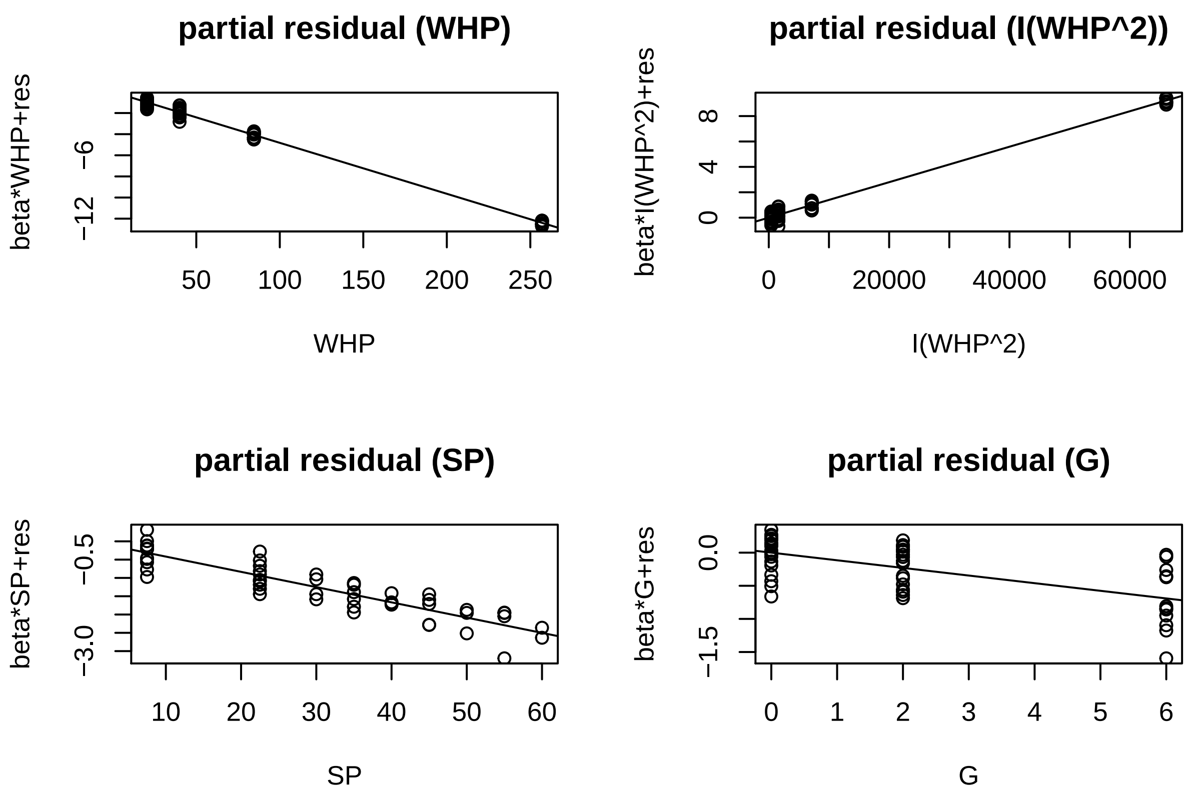
```r
g3.2 = lm(log(ACC) ~ WHP+I(WHP^2)+SP+G, data = vehicle)
par(mfrow = c(1,2))
plot(g3.2$fit, g3.2$res, xlab = "Fitted", ylab = "residual", main = "model g3.2")
abline(h = 0)
plot(g3.2$fit, abs(g3.2$res), xlab = "Fitted", ylab = "|residual|", main = "model g3.2")
abline(lm(abs(g3.2$res)~g3.2$fit))
```




```r
summary(lm(abs(g3.2$res)~g3.2$fit))$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.271580480 0.02967156 9.1528870 4.229992e-12
## g3.2$fit    0.009356901 0.02708719 0.3454364 7.312757e-01
```

```r
par(mfrow = c(2,2))
prplot(g3.2,1)
prplot(g3.2,2)
prplot(g3.2,3)
prplot(g3.2,4)
```



(1) residual plot 不再呈現隨著 fitted value 增加而變異上升的現象。

(2) absolute residual plot 中的回歸線斜率數值非常小，且 pvalue 所呈現的結果也為不顯著不為零。

(3) 各 partial residual plots 也都沒有出現明顯的 mean curvature 以及 unequal variance。

c.

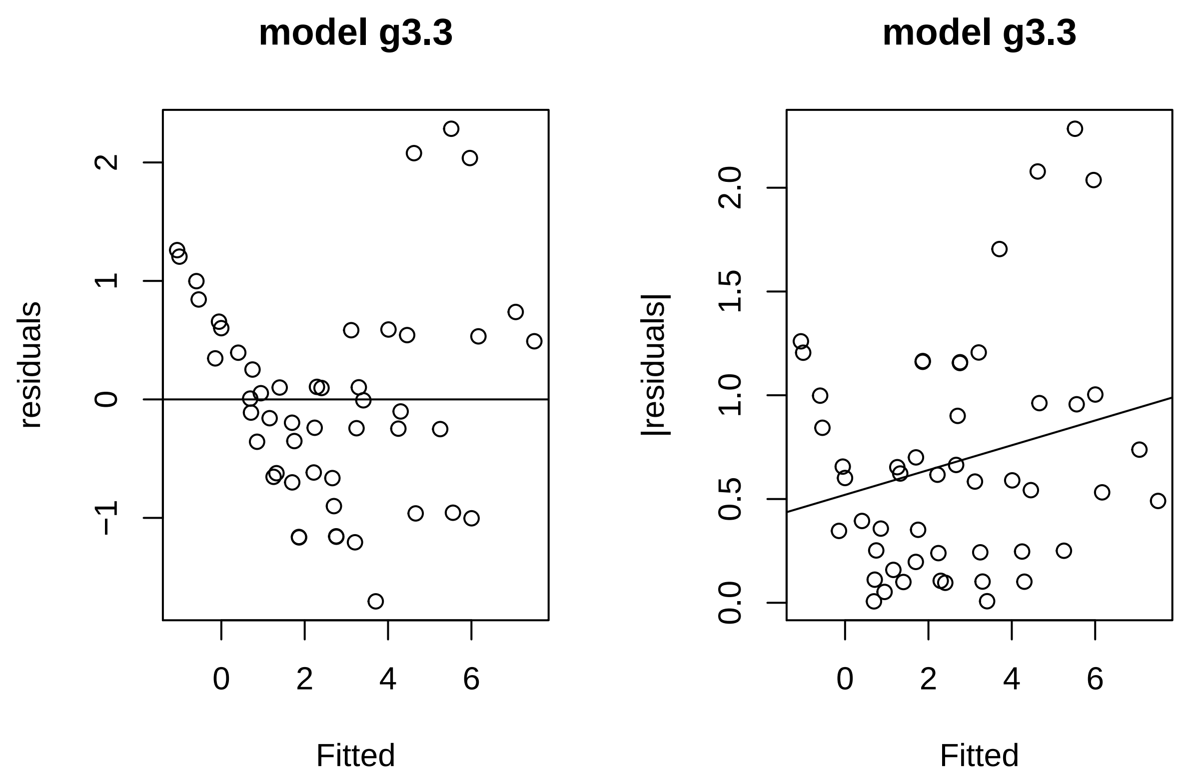
如果不希望 dependent variable 有任何的 weight or transform，那我會建構模型：

$$g_{3.3} \ : \ ACC \ \sim \ WHP \ + \ WHP^2 \ + \ SP \ + \ G$$

因為 a. 題中 WHP 的 partial residual plot 中有著明顯的 mean curvature，故多加入一項解釋變數 WHP^2 用以解釋該現象。

對此模型繪製 residual plot 和 absolute residual plot：

```
g3.3 = lm(ACC ~ WHP+I(WHP^2)+SP+G, data = vehicle)
par(mfrow = c(1,2))
plot(g3.3$fit, g3.3$res, xlab = "Fitted", ylab = "residuals", main = "model g3.3")
abline(h=0)
plot(g3.3$fit, abs(g3.3$res), xlab = "Fitted", ylab = "|residuals|", main = "model g3.3")
abline(lm(abs(g3.3$res)~g3.3$fit))
```



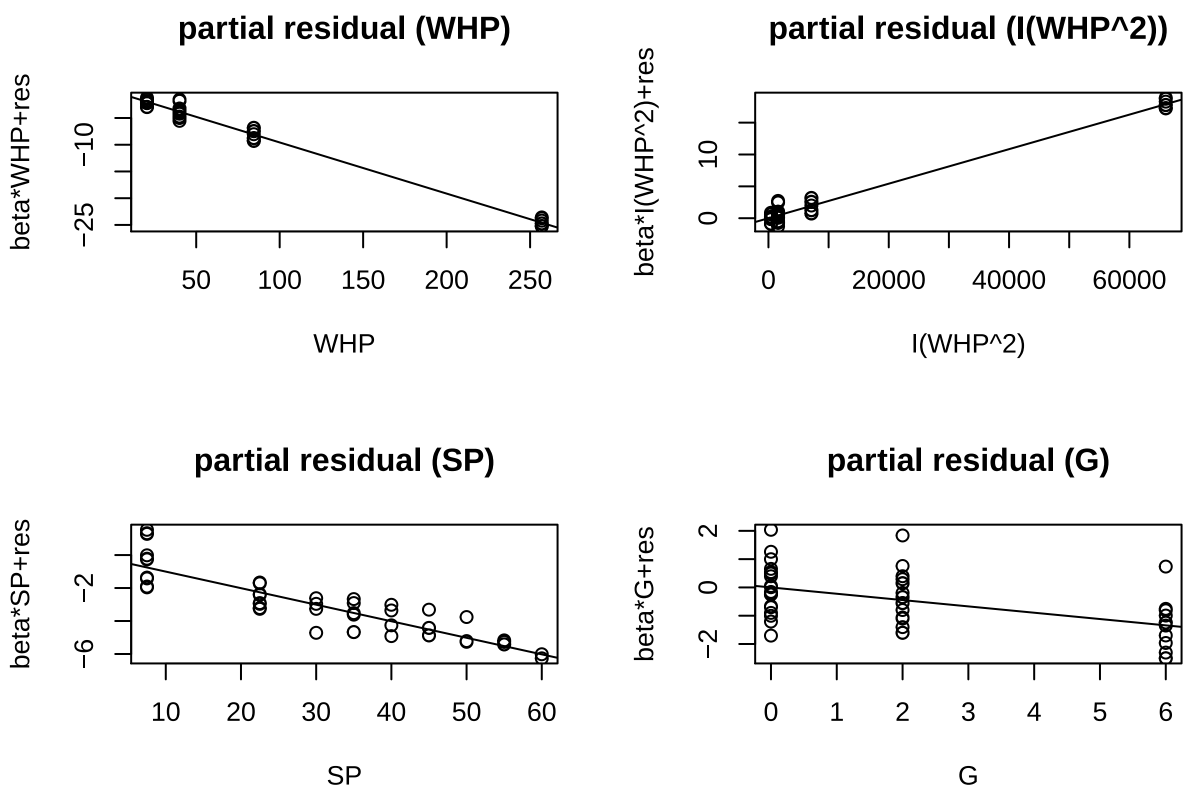

```
summary(lm(abs(g3.3$res)~g3.3$fit))$coef
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.5203516 0.12123593 4.292057 8.542063e-05
## g3.3$fit    0.0596854 0.03602724 1.656674 1.041088e-01
```

residual plot 似乎還是有著 non-constant variance 的現象，表現的不如 *model* $g_{3.2}$ 來得好，但 absolute residual plot 的回歸線斜率的估計和檢定結果已經是足夠靠近零了，故可以推斷此模型下為 constant variance。

再來看此模型下的各 partial residual plots：

```
par(mfrow = c(2,2))
prplot(g3.3,1)
prplot(g3.3,2)
prplot(g3.3,3)
prplot(g3.3,4)
```



一樣在變數 SP 的部分表現不如 $model\ g_{3.2}$，仍保有變異漸減的趨勢，但在變數 WHP 已經沒有 mean curvature 的現象了，這是因為原本在 $model\ g_3$ 中，少考慮了 WHP^2 造成的影響，將其歸入隨機的部分，所以造成了原本模型有著明顯的 non-constant variance 現象。

即使在各 residual plots 上，$model\ g_{3.2}$ 皆表現的比 $model\ g_{3.3}$ 來得好，但是 $model\ g_{3.3}$ 因為沒有對反應變數做函數變換，所以此模型在解釋各係數的意義上會來得更容易於理解，所以若是想研究的內容，係數的數值有著重要意義時 (ex：物理/化學定律)，會更傾向於使用 $model\ g_{3.3}$ 這種模型。