# Experimental Design and Analysis Homework 2

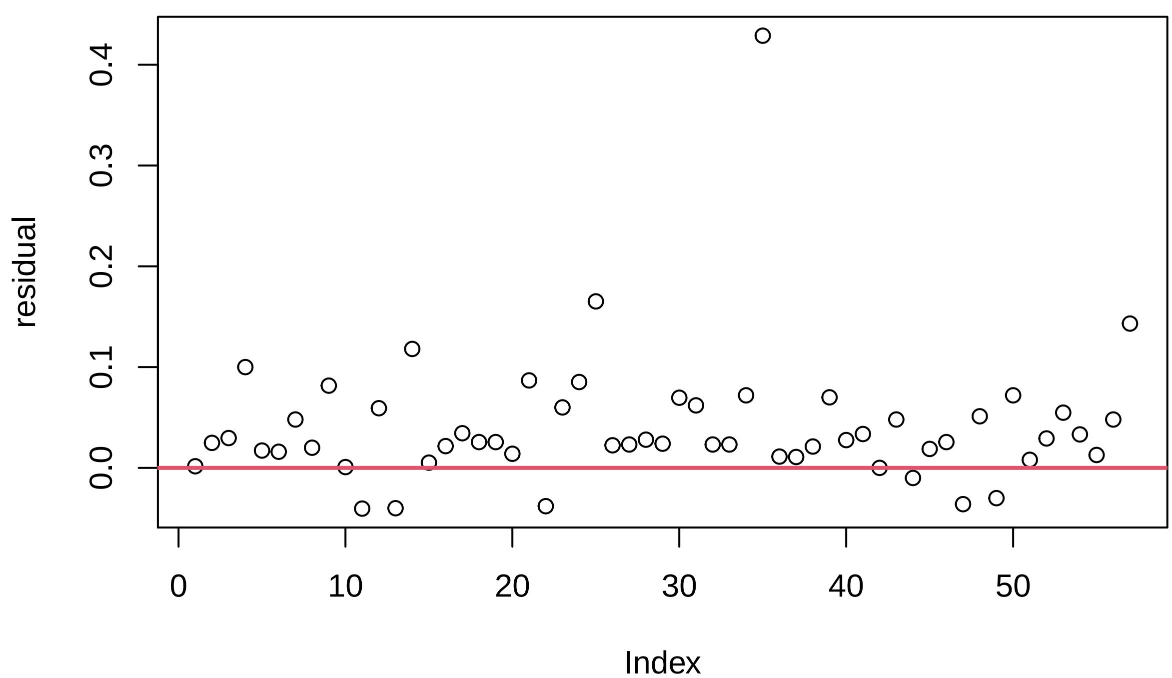
110024516 邱繼賢

## Problem 1.

### (a)

繪製 $y_i - 0.44x_i$ 對 *index* 的 residual plot

```
library(dplyr)
rainfall = read.table("rainfall.txt", header = T)
rainfall = rainfall %>% mutate(res = y-0.44*x, fit = 0.44*x)
plot(rainfall$res, ylab = "residual") ; abline(h = 0, lwd = 2, col = 2)
```



從此 residual plot 中看出，大部分的 residual($y_i - 0.44x_i$) 都大於零，可以推論出無截距項的模型 $y = 0.44x$ 將某種正值的規律給當作隨機加進了 residual 之中，導致 residual 的 mean 並不等於零。

### (b)

首先建構無截距項的模型 *model1*

$$y_i = \beta_1 x_i + \epsilon_i$$

```
fitb.1 = lm(y ~ x -1 , data = rainfall)
summary(fitb.1)
```

```
##
## Call:
## lm(formula = y ~ x - 1, data = rainfall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11930  0.00308  0.01773  0.03547  0.38669
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x 0.455425   0.004484   101.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07106 on 56 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9945
## F-statistic: 1.032e+04 on 1 and 56 DF,  p-value: < 2.2e-16
```

- $R^2 = 99.46\%$ 非常高，但是此為無截距項的模型，$R^2$ 數值並沒有意義
- $\hat{\sigma} = 0.07106$
- $\beta_1$ 對此模型有顯著貢獻，以下再進一步檢定其是否 $= 0.44$

檢定：

$$\begin{cases} H_0 : \beta_1 = 0.44 \\ H_1 : \beta_1 \neq 0.44 \end{cases}$$

```r
fitb.1_test = lm(y ~ offset(0.44*x)-1, data = rainfall)
anova(fitb.1_test, fitb.1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ offset(0.44 * x) - 1
## Model 2: y ~ x - 1
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     57 0.34249
## 2     56 0.28274  1  0.059751 11.834 0.001105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\Rightarrow$ pvalue $= 0.001105 < 0.05$，$\beta_1$ 顯著不等於 0.44，這與我們根據理論得出的比例係數 0.44 有所不合，再加上 **(a)** 中我們所做出的結論，可以推斷出無截距項的模型並不適合此筆資料，其中有一些規律被我們給忽略了。

再來建構有截距項的模型 *model2*

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

```r
fitb.2 = lm(y ~ x, data = rainfall)
summary(fitb.2)
```

```
##
## Call:
## lm(formula = y ~ x, data = rainfall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09314 -0.02529 -0.01205  0.01689  0.38304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.035787   0.012210   2.931  0.00491 **
## x           0.443652   0.005817  76.264  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.06668 on 55 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9905
## F-statistic:  5816 on 1 and 55 DF,  p-value: < 2.2e-16
```

- $R^2 = 99.06\%$ 表現得非常好
- $\hat{\sigma} = 0.06668$ 小於前一個模型的 $\hat{\sigma}$
- $\beta_0, \beta_1$ 皆呈現顯著不為零

一樣進行檢定：

$$\begin{cases} H_0 : \beta_1 = 0.44 \\ H_1 : \beta_1 \neq 0.44 \end{cases}$$

```r
fitb.2_test = lm(y ~ offset(0.44*x), data = rainfall)
anova(fitb.2_test, fitb.2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ offset(0.44 * x)
## Model 2: y ~ x
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     56 0.24630
## 2     55 0.24455  1 0.0017526 0.3942 0.5327
```

$\Rightarrow$ pvalue $= 0.5327 > 0.05$，故可以推斷 $\beta_1$ 和 0.44 並沒有顯著差異，符合我們使用理論所推導出的比例係數 0.44

### (c)

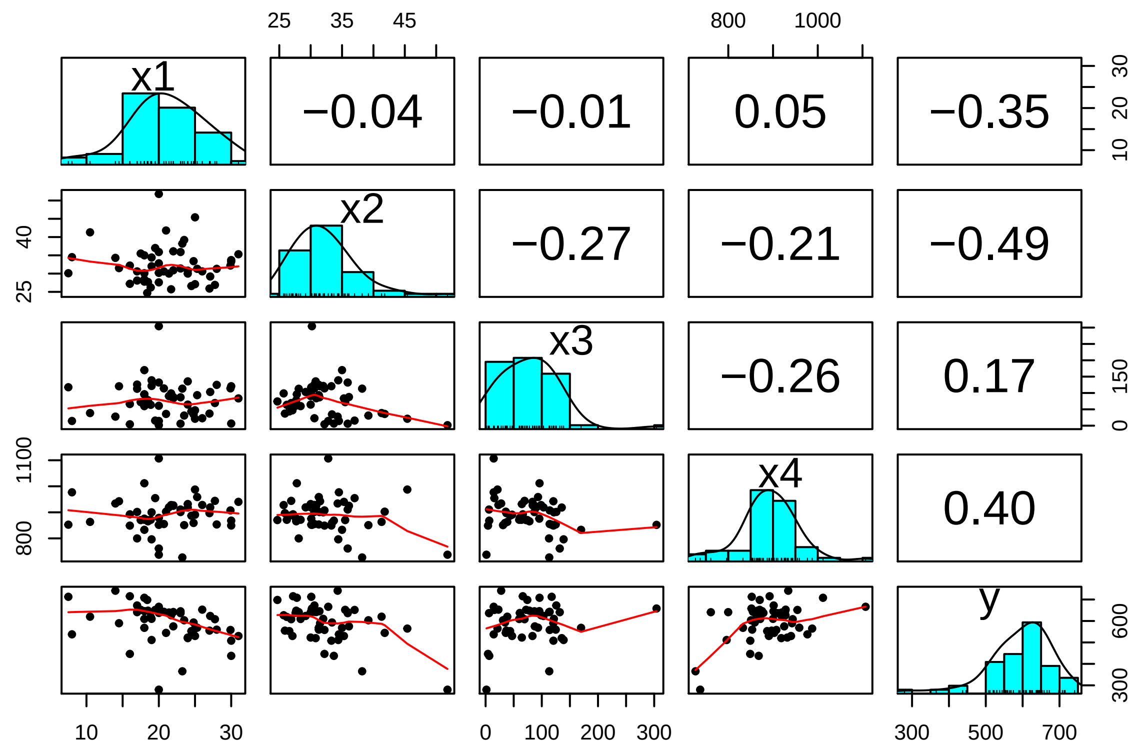
根據 **(a),(b)** 所做出的結論，我會選擇使用加入了截距項的模型 *model2*

$$\hat{y}_i = 0.035787 + 0.443652x_i$$

**Problem 2.**

**(a)**

```
gas = read.table("Gasoline.txt", header = T)
library(psych)
pairs.panels(gas, ellipses = F)
```




- x1 (汽油稅) 和 y (汽油消耗量) 呈現負相關，相關係數 = −0.35
- x2 (人均收入) 和 y (汽油消耗量) 呈現負相關，相關係數 = −0.49
- x3 (鋪設高速公路長) 和 y (汽油消耗量) 呈現些微正相關，相關係數 = 0.17
- x4 (持牌司機人數) 和 y (汽油消耗量) 呈現正相關，相關係數 = 0.4
- x1,x2,x3,x4 四個變數之間的相關係數數值並不高，推測在做模型時不會產生嚴重的共線性狀況

**(b)**

建構模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

```
fit_14 = lm(y ~ x1+x2+x3+x4, data = gas)
summary(fit_14)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = gas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -151.942  -30.757    2.443   41.201  115.262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 439.9743   171.8339   2.560 0.013801 *
## x1           -6.2927     1.7155  -3.668 0.000632 ***
## x2           -6.1718     1.8726  -3.296 0.001895 **
## x3            0.2766     0.1843   1.501 0.140249
## x4            0.5210     0.1499   3.476 0.001122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.3 on 46 degrees of freedom
## Multiple R-squared:  0.5034, Adjusted R-squared:  0.4602
## F-statistic: 11.66 on 4 and 46 DF,  p-value: 1.28e-06
```

- 變數 x1,x2,x4 對 response y 有顯著影響

**(c)**

(1) 直觀判斷上會認為隨著汽油稅收 (x1) 增加，相對應的汽油消耗量 (y) 會跟著有所減少 (負相關)，而回歸模型的係數 (-6.2927)，和兩變數的相關係數 (-0.35) 也都呈現為負值，符合直觀。

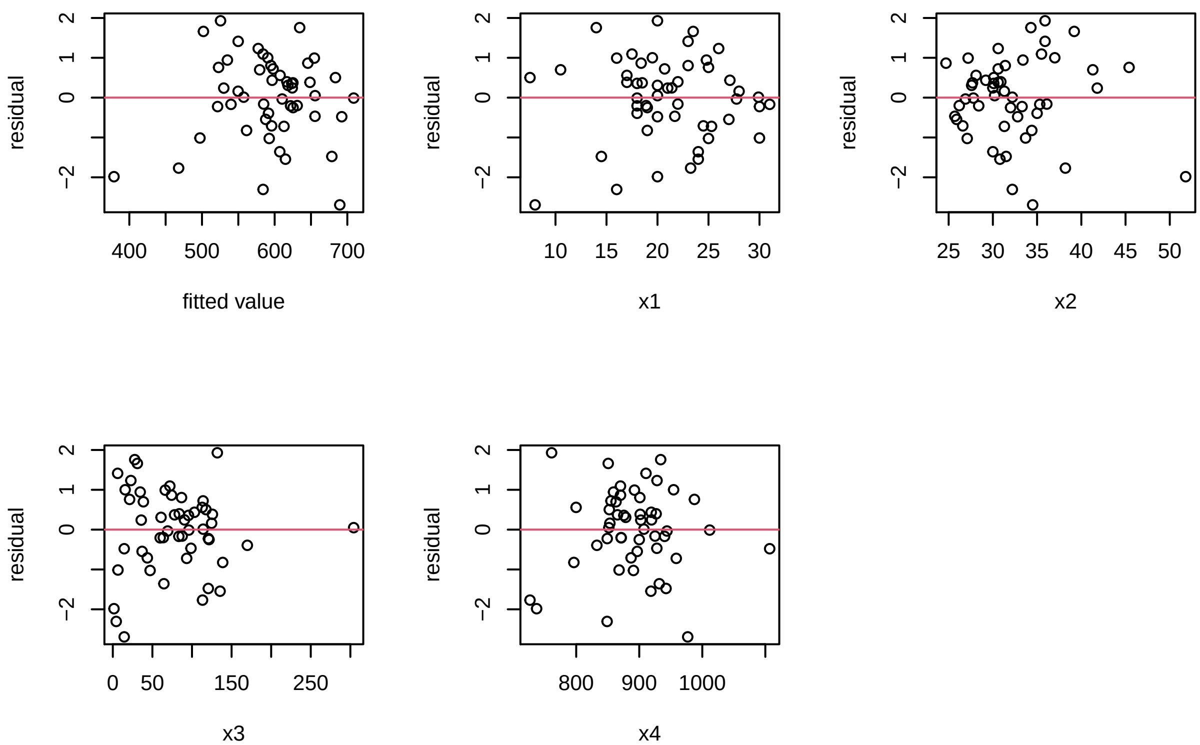
(2) 直觀判斷上會認為隨著人均收入 (x2) 增加，相對應的汽油消耗量 (y) 也會跟著上升 (正相關)，但是實際做回歸模型的係數 (-6.1718)，和兩變數的相關係數 (-0.49) 皆呈現為負值，與直觀上有所衝突，可能的原因是在人均收入較高的地區，交通壅塞，大部分的人都選擇使用大眾運輸工具，反而減少了汽油的消耗量，但實際造成此現象的原因是否如此，還需要更多資訊才能下定論。

(3) 直觀判斷上會認為隨著鋪設的高速公路長 (x3) 或是持牌司機的人數 (x4) 增加，相對應的汽油消耗量 (y) 也會有所增加 (正相關)，而回歸模型的兩係數 (0.2766,0.5210)，以及兩組變數各自的相關係數 (0.17,0.4) 也都呈現為正值，符合直觀。

### (e)

將模型的 studentized residual 對 fitted value, x1, x2, x3, x4 各自繪製 residual plot

```r
par(mfrow = c(2,3))
rstud = rstandard(fit_14)
plot(fit_14$fitted.values, rstud, xlab = "fitted value", ylab = "residual") ; abline(h = 0, col = 2)
plot(gas$x1, rstud, xlab = "x1", ylab = "residual") ; abline(h = 0, col = 2)
plot(gas$x2, rstud, xlab = "x2", ylab = "residual") ; abline(h = 0, col = 2)
plot(gas$x3, rstud, xlab = "x3", ylab = "residual") ; abline(h = 0, col = 2)
plot(gas$x4, rstud, xlab = "x4", ylab = "residual") ; abline(h = 0, col = 2)
```




- 變數 x2 和 x3 所對應的 residual plot 看起來皆有著 non-constant variance，可以考慮使用 weighted least square 來重新建構模型

- 其他變數和 fitted value 所對應的 residual plot 看起來並沒有明顯的 non-constant variance 或是 mean curvature