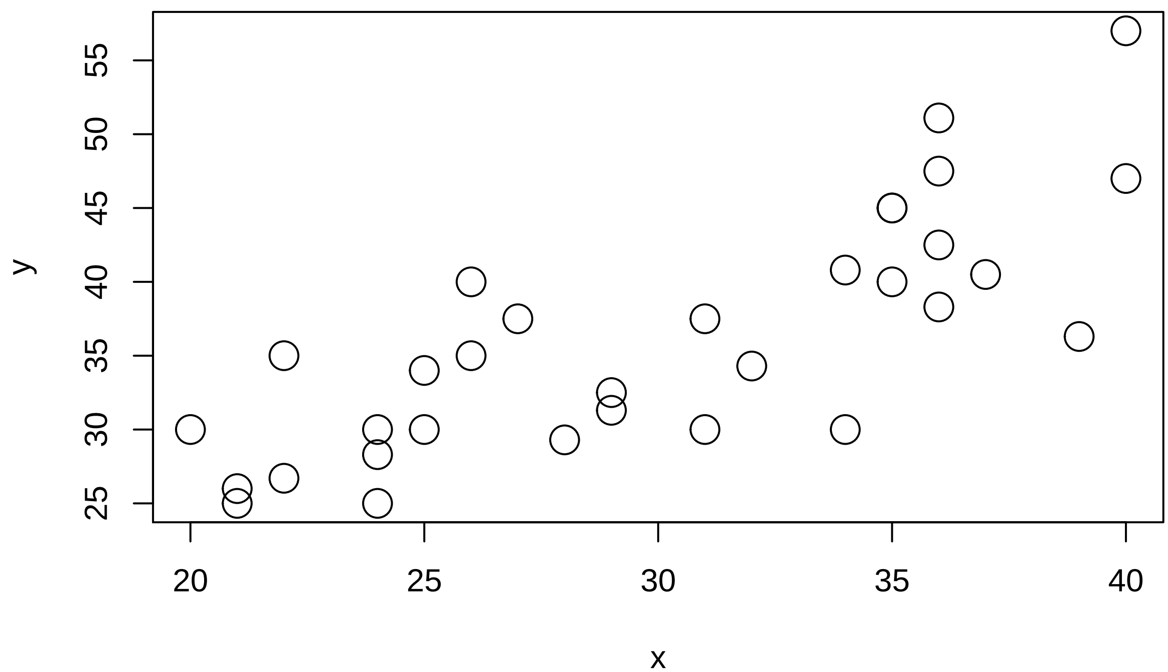
Linear Model Assignment 5

110024516 統研碩一邱繼賢

Problem 1.
a.

```
travel = read.table("travel.txt", skip = 1)
names(travel) = c("obs", "n", "x", "y")
plot(travel$x, travel$y, xlab = "x", ylab = "y", cex = 2)
```



(1) x 和 y 存在著正相關的現象。

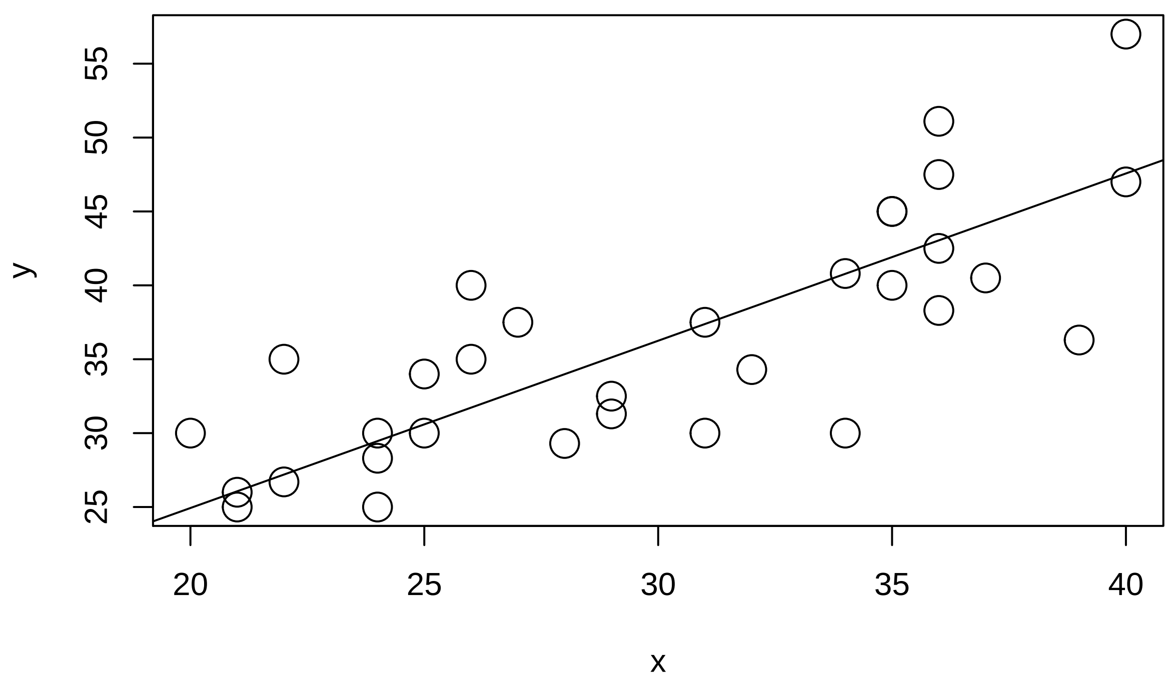
(2) x 對 y 普遍存在著低估的現象，即對兩地點移動所需時間估計普遍小於實際測量後的平均值。

b. 以每組地點間的 travelers 數量為權重 ($w_i \propto n_i$)，建構回歸模型如下：

$$S^{-1}Y = S^{-1}X\beta + S^{-1}\epsilon \text{ , where } S = diag(\frac{1}{\sqrt{w_1}}, ..., \frac{1}{\sqrt{w_n}}) \text{ , then } \Sigma = SS^T$$

```r
w = travel$n
g = lm(y ~ x, data = travel, weights = w)
summary(g)
```

```
##
## Call:
## lm(formula = y ~ x, data = travel, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -20.278  -7.661  -0.680   4.543  33.219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2932     4.5903   0.500    0.621
## x             1.1319     0.1475   7.676 1.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.01 on 30 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.6514
## F-statistic: 58.93 on 1 and 30 DF,  p-value: 1.458e-08
```



c. In order to check model g for lack of fit. Construct saturated model ga and do the anova test to compare the two models as below :

$$\begin{cases} H_0 : \text{g model fitted better} \\ H_1 : \text{ga model fitted better} \end{cases} \iff \begin{cases} H_0 : \text{g model is not lack of fit} \\ H_1 : \text{g model is lack of fit} \end{cases}$$

```r
ga = lm(y ~ factor(x), data = travel, weights = w)
anova(g, ga)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ factor(x)
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     30 3006.20
## 2     15  945.47 15    2060.7 2.1796 0.07132 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value $= 0.07132 \ > \ 0.05 \ \Rightarrow$ fail to reject H_0

$\therefore$ We do not detect lack of fit for model g.

Problem 2.

a. Take the number of fathers in each category as weight($w_i \ \propto \ n_i$), then contruct the Weighted Least Square as below :

$$\text{model } g1 \ : \ S^{-1}Y \ = \ S^{-1}X\beta \ + \ S^{-1}\epsilon \ , \text{ where } S^{-1} \ = \ diag(\sqrt{w_1}, ..., \sqrt{w_n})$$

```
height = read.table("height.txt", skip = 2)
father_h = height[,1]
son_h = height[,2]
w = height[,3]
g1 = lm(son_h ~ father_h, weights = w)
summary(g1)
```

```
##
## Call:
## lm(formula = son_h ~ father_h, weights = w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39024 -0.77499  0.04766  1.15672  1.67501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.5820     2.2486   14.49 4.87e-08 ***
## father_h      0.5297     0.0332   15.96 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.147 on 10 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9584
## F-statistic: 254.6 on 1 and 10 DF,  p-value: 1.926e-08
```

b. Construct model g2 : $height \ of \ son \ = \ height \ of \ father \ + \ error \ , \ with \ w_i \ \propto \ n_i$

and then do the anova test for comparing g1 and g2 models :

$$\begin{cases} H_0 : \text{g2 model fits better} \\ H_1 : \text{g1 model fits better} \end{cases}$$

```
g2 = lm(son_h ~ offset(father_h)-1, weights = w)
anova(g2, g1)
```

```
## Analysis of Variance Table
##
## Model 1: son_h ~ offset(father_h) - 1
## Model 2: son_h ~ father_h
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     12 384.54
## 2     10  13.17  2    371.37 141.03 4.706e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\because p-value \ < \ 0.05 \ \Rightarrow$ reject H_0

$\therefore$ g1 model is a better model for fitting. g1 model is not approriate to be simplified to g2.

Problem 3.

(i) 整理 data：

將 data 的第二欄數據 $\times 0.00001 + 0.742$ 伸縮平移到其對應的真實數據值，並以相同的 day 為一組計算其 standard deviation (std)，然後取 standard variance 為其權重 ($w_i \propto \frac{1}{std_i^2}$)，整理後資料呈現如下：

```
library(dplyr)
library(knitr)
crank = read.table("crank.txt", skip = 1)
names(crank) = c("day", "diameter")
crank$diameter = 0.742+0.00001*crank$diameter

crank = crank %>% group_by(day) %>%
    mutate(std = sd(diameter)) %>%
    mutate(weight = 1/std^2) %>%
    ungroup()
kable(crank)
```

day	diameter	std	weight
1	0.74293	2.86e-05	1219512195
1	0.74298	2.86e-05	1219512195
1	0.74290	2.86e-05	1219512195
1	0.74294	2.86e-05	1219512195
1	0.74294	2.86e-05	1219512195
4	0.74293	5.79e-05	298507463
4	0.74300	5.79e-05	298507463
4	0.74288	5.79e-05	298507463
4	0.74285	5.79e-05	298507463
4	0.74289	5.79e-05	298507463
7	0.74289	4.44e-05	507614213
7	0.74290	4.44e-05	507614213
7	0.74292	4.44e-05	507614213
7	0.74295	4.44e-05	507614213
7	0.74300	4.44e-05	507614213
10	0.74293	2.61e-05	1470588235
10	0.74288	2.61e-05	1470588235
10	0.74287	2.61e-05	1470588235
10	0.74287	2.61e-05	1470588235
10	0.74287	2.61e-05	1470588235
13	0.74288	2.12e-05	2222222222
13	0.74286	2.12e-05	2222222222
13	0.74291	2.12e-05	2222222222
13	0.74289	2.12e-05	2222222222
13	0.74286	2.12e-05	2222222222
16	0.74282	7.21e-05	192307692
16	0.74272	7.21e-05	192307692
16	0.74280	7.21e-05	192307692
16	0.74272	7.21e-05	192307692
16	0.74289	7.21e-05	192307692
19	0.74281	6.99e-05	204918033
19	0.74280	6.99e-05	204918033
19	0.74278	6.99e-05	204918033
19	0.74294	6.99e-05	204918033

day	diameter	std	weight
19	0.74290	6.99e-05	204918033
22	0.74290	6.28e-05	253164557
22	0.74292	6.28e-05	253164557
22	0.74282	6.28e-05	253164557
22	0.74277	6.28e-05	253164557
22	0.74289	6.28e-05	253164557

(ii) Test for under control or not

建構模型

$$\begin{cases} g_3 : diameter \;=\; \beta_0 \;+\; \beta_1 \; day \;+\; \epsilon \;,\; with \; weight \;\propto\; \dfrac{1}{std^2} \\ g_4 : diameter \;=\; 0.74275 \;+\; \epsilon \;,\; with \; weight \;\propto\; \dfrac{1}{std^2} \end{cases}$$

判斷 process 是否 under control 的條件即為進行以下檢定：

$$\begin{cases} H_0 : \beta_0 \;=\; 0.74275 \;\; and \;\; \beta_1 \;=\; 0 \\ H_1 : \beta_0 \;\neq\; 0.74275 \;\; or \;\; \beta_1 \;\neq\; 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : g4 \;\; fits \;\; better \\ H_1 : g3 \;\; fits \;\; better \end{cases}$$

```r
g3 = lm(diameter ~ day, weights = weight, data = crank)
g4 = lm(diameter ~ offset(rep(0.74275,40))-1,
        weights = weight, data = crank)
anova(g4,g3)
```

```
## Analysis of Variance Table
##
## Model 1: diameter ~ offset(rep(0.74275, 40)) - 1
## Model 2: diameter ~ day
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     40 715.90
## 2     38  40.35  2    675.54 318.08 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p-value \;<\; 0.05 \;\Rightarrow$ right H_0

$\therefore$ g_3 fits better. The process is out of control.

(iii) Test for lack of fit

建構 saturated model $g_5 : diameter \;\sim\; factor(day) \;,\; with \; weight \;\propto\; \frac{1}{std^2}$

並進行以下檢定

$$\begin{cases} H_0 : g_3 \; fits \; better \\ H_1 : g_5 \; fits \; better \end{cases} \Leftrightarrow \begin{cases} H_0 : g_3 \; is \; not \; lack \; of \; fit \\ H_1 : g_3 \; is \; lack \; of \; fit \end{cases}$$

```r
g5 = lm(diameter ~ factor(day), weights = weight, data = crank)
anova(g3, g5)
```

```
## Analysis of Variance Table
##
## Model 1: diameter ~ day
## Model 2: diameter ~ factor(day)
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     38 40.352
## 2     32 32.000  6     8.352 1.392  0.248
```

$p - value\ =\ 0.248\ >\ 0.05\ \Rightarrow$ fail to reject H_0
$\therefore$ We do not detect lack of fit for model g_3