Experimental Design and Analysis 4

110024516 邱繼賢

Problem 1. (3.1)

建構模型

$$\text{weight}_{ij} \ = \ \eta \ + \ \alpha_i \ + \ \tau_j \ + \ \epsilon_{ij} \quad , \quad \epsilon_{ij} \ \sim \ N(0,\sigma^2) \quad , \quad i=1,...,6 \quad , \quad j=1,2$$

where α_i is the block (lake) effect, and τ_j is the treatment (scale) effect

Our data matrix :

```
library(dplyr)
library(knitr)
rock = data.frame(lake = as.factor(c(1,2,3,4,5,6,1,2,3,4,5,6)),
                  scale = as.factor(c(1,1,1,1,1,1,2,2,2,2,2,2)),
                  weight = c(8,14,16,19,18,12,11,16,20,18,20,15))
kable(rock, col.names = c("lack(i)","scale(j)","weight(ij)"))
```

lack(i)	scale(j)	weight(ij)
1	1	8
2	1	14
3	1	16
4	1	19
5	1	18
6	1	12
1	2	11
2	2	16
3	2	20
4	2	18
5	2	20
6	2	15

(a)

Compute the within block difference $d_i = \text{weight}_{i1} - \text{weight}_{i2} = \{-3, -2, -4, 1, -2, -3\}$, and do the paired t test

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases}$$

test statistic

$$t_{\text{paired}} = \frac{\overline{d}}{s_d/\sqrt{N}} = \frac{-2.166667}{1.722401/\sqrt{6}} = -3.081297$$

(b)

p-value $= 2 \times P\left(t_{N-1} > |-3.081297|\right) = 0.02742918 < 0.05$

$\Rightarrow$ Reject H_0, so the two scales are significant different

We can check the answer by the ANOVA table

```
fit1 = lm(weight ~ lake + scale, rock)
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## lake       5 135.417 27.0833 18.2584 0.003154 **
## scale      1  14.083 14.0833  9.4944 0.027429 *
## Residuals  5   7.417  1.4833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\Rightarrow$ p-value (0.027429) are the same

Problem 2. (3.9)

(1) Paired comparison design

因為某些 latent variables 像是年齡、性別、體重…等身體狀況的不同皆可能影響血壓，所以在尋找受試者時盡量將 latent variables 相似的兩位受試者分為一組 (ex：雙胞胎)，即為一個 block，然後將一劑藥物 A 和一劑安慰劑隨機分配給同一組的兩位受試者 (可採取雙盲實驗)，然後紀錄兩小時內的血壓變化並計算同一組 block 內的差距，這樣即可以消除不同 block 所帶來的血壓變異，也就是所謂的 Paired t test (如 **3.1**)

(2) Unpaired design

但其實上述的方法非常的不切實際，要同時找到那麼多組 latent variables 相似的受試者非常不容易，unpaired 的方法就是直接將所有受試者「隨機」的平均分配成兩組，一組施打藥物 A 一組施打安慰劑 (同樣採取雙盲實驗)，然後記錄這兩組受試者兩小時內的血壓變化，分別計算平均數 ($\overline{y}_1$，$\overline{y}_2$) 和變異數 (s_1^2，s_2^2)，然後進行 Unpaired t test，這樣的方法雖然沒辦法消除受試者之間的變異，但由於我們分組時是採用「隨機」的方式，可以一定程度的減少其所帶來的影響，另外可以考慮降低所有受試者之間的差異 (ex：同一間療養院的長者、同一間學校的學生)，也可以一定程度提升實驗的精準度。

$$t_{\text{unpaired}} \;=\; \frac{\overline{y}_2 - \overline{y}_1}{\sqrt{s_2^2/N + s_1^2/N}} \;\sim\; t_{2N-2}$$

Problem 3. (3.13)

因為變數 *power, log(speed)* 皆大致為 three evenly spaced levels，定義各自的 linear and quadratic contrast

$$\text{power}_l \;=\; \begin{cases} -1/\sqrt{2}\ , \ \text{power} \;=\; 40 \\ \quad\ 0\ , \ \text{power} \;=\; 50 \\ \ 1/\sqrt{2}\ , \ \text{power} \;=\; 60 \end{cases} ,\quad \text{power}_q \;=\; \begin{cases} \ 1/\sqrt{6}\ , \ \text{power} \;=\; 40 \text{ or } 60 \\ -2/\sqrt{6}\ , \ \text{power} \;=\; 50 \end{cases}$$

$$\text{log-speed}_l \;=\; \begin{cases} -1/\sqrt{2}\ , \ \text{speed} \;=\; 6.42 \\ \quad\ 0\ , \ \text{speed} \;=\; 13 \\ \ 1/\sqrt{2}\ , \ \text{speed} \;=\; 27 \end{cases} ,\quad \text{log-speed}_q \;=\; \begin{cases} \ 1/\sqrt{6}\ , \ \text{speed} \;=\; 6.42 \text{ or } 27 \\ -2/\sqrt{6}\ , \ \text{speed} \;=\; 13 \end{cases}$$

建構模型

$$\text{strength} \;\sim\; \text{power}_l \;+\; \text{power}_q \;+\; \text{log-speed}_l \;+\; \text{log-speed}_q \;+\; \text{power}_l : \text{log-speed}_l$$

因為資料為 single replicate，沒有足夠的 degree of freedom 分配給所有的交互作用項，故模型只選擇放入 $\text{power}_l : \text{log-speed}_l$ 兩個 linear contrast 之間的交互作用，其餘有包含 quadratic contrast 之間的交互作用項則不放入模型

```r
composite = read.table("Composite.txt", header = T)
composite = composite %>%
    mutate(power.l = c(-1,-1,-1,0,0,0,1,1,1)/sqrt(2),
           power.q = c(1,1,1,-2,-2,-2,1,1,1)/sqrt(6),
           log_speed.l = c(-1,0,1,-1,0,1,-1,0,1)/sqrt(2),
           log_speed.q = c(1,-2,1,1,-2,1,1,-2,1)/sqrt(6))
fit3.1 = lm(strength ~ power.l+power.q+log_speed.l+log_speed.q+power.l:log_speed.l, composite)
summary(fit3.1)
```

```
##
## Call:
## lm(formula = strength ~ power.l + power.q + log_speed.l + log_speed.q +
##     power.l:log_speed.l, data = composite)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  0.50722  0.04556 -0.55278 -1.34111  0.56222  0.77889  0.83389 -0.60778
##        9
## -0.22611
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          31.0322     0.4038  76.853 4.86e-06 ***
## power.l               8.6361     0.6994  12.348  0.00114 **
## power.q              -0.3810     0.6994  -0.545  0.62377
## log_speed.l          -1.0465     0.6994  -1.496  0.23146
## log_speed.q          -3.9001     0.6994  -5.577  0.01138 *
## power.l:log_speed.l   2.4700     1.2114   2.039  0.13417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.211 on 3 degrees of freedom
## Multiple R-squared:  0.9845, Adjusted R-squared:  0.9586
## F-statistic: 38.05 on 5 and 3 DF,  p-value: 0.006475
```

- power_q 對模型貢獻不顯著，故刪除該變數
- log-speed_l 雖然對模型貢獻不顯著，但是 log-speed_q 結果為顯著，故兩項皆保留
- $\text{power}_l : \text{log-speed}_l$ 對模型貢獻不顯著，故刪除該變數

重新配飾模型

$$\text{strength} \sim \text{power}_l + \text{log-speed}_l + \text{log-speed}_q$$

```r
fit3.2 = update(fit3.1, .~.-power.q-power.l:log_speed.l)
summary(fit3.2)
```

```
##
## Call:
## lm(formula = strength ~ power.l + log_speed.l + log_speed.q,
##     data = composite)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9
##  1.5867 -0.1100 -1.9433 -1.0300  0.8733  1.0900 -0.5567 -0.7633  0.8533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   31.032      0.493  62.942 1.92e-08 ***
## power.l        8.636      0.854  10.113 0.000162 ***
## log_speed.l   -1.046      0.854  -1.225 0.274961
## log_speed.q   -3.900      0.854  -4.567 0.006018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 5 degrees of freedom
## Multiple R-squared:  0.9614, Adjusted R-squared:  0.9383
## F-statistic: 41.55 on 3 and 5 DF,  p-value: 0.0005869
```

$\Rightarrow$ 變數 power_l 和 log-speed_q 皆呈現顯著

模型估計係數如下

$$\widehat{\text{strength}} = 31.032 + 8.636\ \text{power}_l - 1.046\ \text{log-speed}_l - 3.9\ \text{log-speed}_q$$

觀察其 ANOVA table

```r
anova(fit3.2)
```

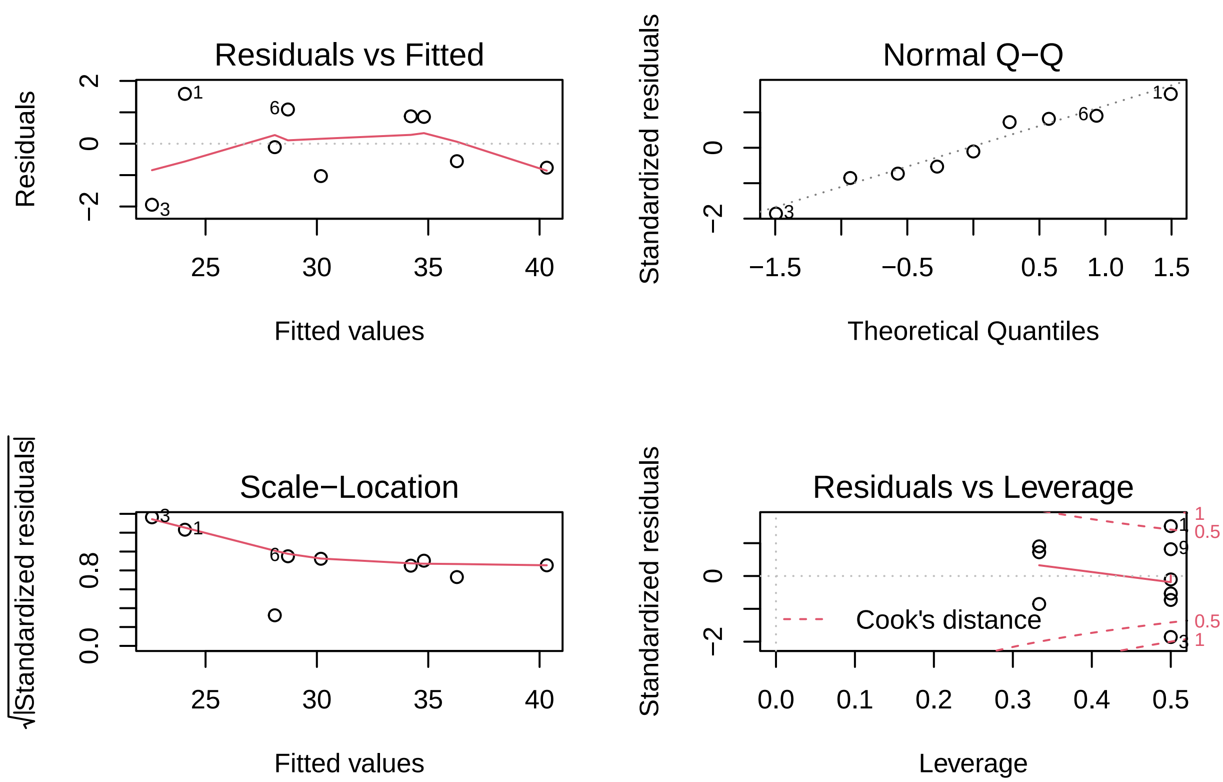
```
## Analysis of Variance Table
##
## Response: strength
```

```
##              Df  Sum Sq Mean Sq  F value   Pr(>F)
## power.l       1 223.748 223.748 102.2746 0.000162 ***
## log_speed.l   1   3.286   3.286   1.5018 0.274961
## log_speed.q   1  45.633  45.633  20.8587 0.006018 **
## Residuals     5  10.939   2.188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

⇒ 結果和上面做回歸分析所得結果一致 (look at the p-value)

對此模型做 diagnostic

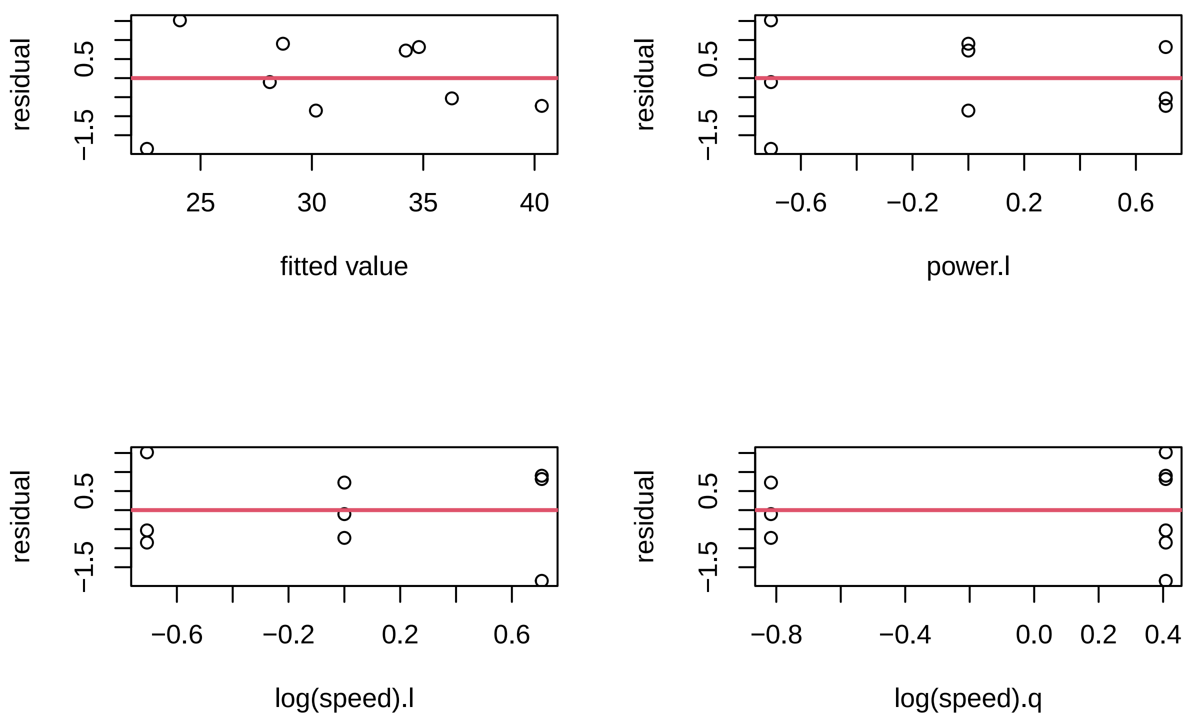
```
par(mfrow = c(2,2))
plot(fit3.2)
```



⇒ 並沒有出現明顯的 outlier，QQ plot 也顯示 residual 接近 normal distribution

以 studentized residual 對 fitted value 以及各變數繪製 residual plots

```r
par(mfrow = c(2,2))
res = rstandard(fit3.2)
plot(fit3.2$fitted.values,res, xlab="fitted value",ylab="residual") ; abline(h=0, col=2, lwd=2)
plot(composite$power.l, res, xlab="power.l",ylab="residual") ; abline(h=0, col=2, lwd=2)
plot(composite$log_speed.l, res, xlab="log(speed).l",ylab="residual") ; abline(h=0, col=2, lwd=2)
plot(composite$log_speed.q, res, xlab="log(speed).q",ylab="residual") ; abline(h=0, col=2, lwd=2)
```



⇒ 大致呈現正常，沒有出現明顯的 non-constant variance 或是 mean curvature

接下來將變數 *power* 和 *log(speed)* 視為 3 levels qualitative variables，建構模型

$$\text{strength}_{ij} = \eta + \text{power}_i + \text{log-speed}_j + \epsilon_{ij} \quad , \quad i,j = 1,2,3 \quad , \quad \epsilon_{ij} \sim N(0,\sigma^2)$$

觀察其 ANOVA table

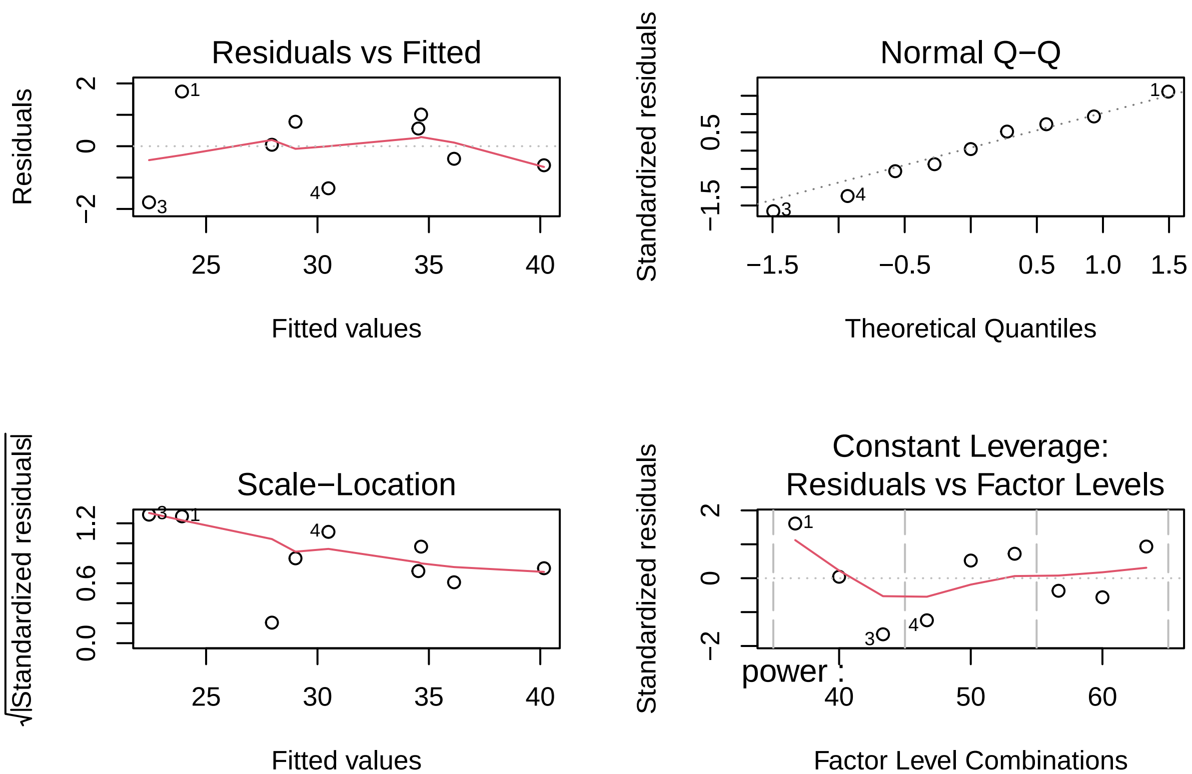
```r
composite$power = as.factor(composite$power)
composite$log_speed = as.factor(log(composite$speed))
fit3.3 = lm(strength ~ power + log_speed, composite)
anova(fit3.3)
```

```
## Analysis of Variance Table
##
## Response: strength
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## power      2 224.184 112.092 42.6893 0.002003 **
## log_speed  2  48.919  24.459  9.3151 0.031242 *
## Residuals  4  10.503   2.626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

⇒ 變數 *power* 和 *log(speed)* 皆呈現顯著，與前一個模型結果一致

一樣對此模型做 diagnostic

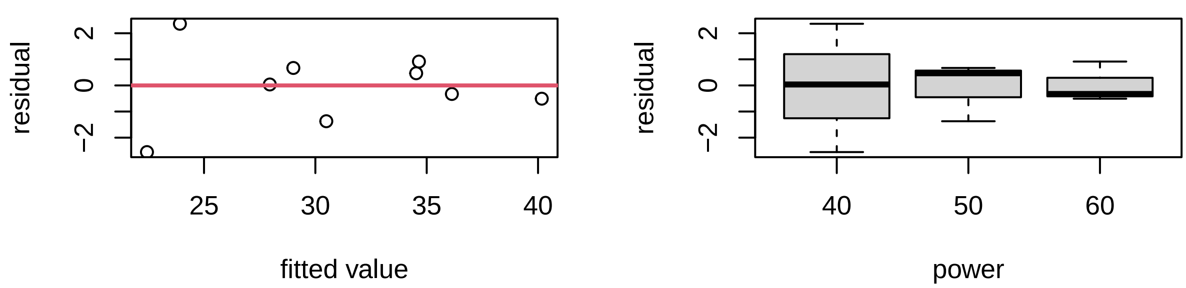
```r
par(mfrow = c(2,2))
plot(fit3.3)
```

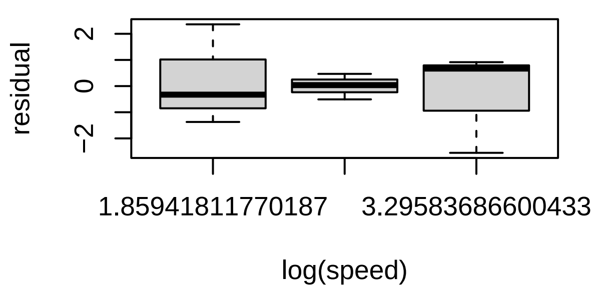


⇒ 並沒有出現明顯的 outlier，QQ plot 也顯示 residual 接近 normal distribution

以 studentized residuals 對 fitted values 以及兩個 treatments 繪製 residual plots

```
par(mfrow = c(2,2))
res = rstudent(fit3.3)
plot(fit3.3$fitted.values, res, xlab="fitted value",ylab="residual") ; abline(h = 0, col = 2, lwd = 2)
plot(composite$power, res, xlab="power",ylab="residual")
plot(composite$log_speed, res, xlab="log(speed)",ylab="residual")
```





⇒ 看起來有一點 non-constant variance 的現象，但我們每個 treatment 的 3 個 levels 下只有三次實驗值，可以再增加實驗 replication 的次數後，若依舊呈現此現象，則可以考慮對反應變數做 tranformation 或是用 weighted leasted square 的方式來估計回歸係數。

Problem 4. (3.28)

Show the data matrix as below

```
resistor = data.frame(plate = as.factor(c(1,1,1,2,2,2,3,3,3,4,4,4)),
                      shape = as.factor(c("A","C","D","A","B","D","A","B","C","B","C","D")),
                      noise = c(1.11,0.95,0.82,1.7,1.22,0.97,1.6,1.11,1.52,1.22,1.54,1.18))
kable(resistor, col.names = c("plate","shape", "noise"))
```

plate	shape	noise
1	A	1.11
1	C	0.95
1	D	0.82
2	A	1.70
2	B	1.22
2	D	0.97
3	A	1.60
3	B	1.11
3	C	1.52
4	B	1.22
4	C	1.54
4	D	1.18

(a)

Because each pair of treatments appear in the same number ($\lambda = 2$) of blocks, it is a Balanced Incomplete Block Design (BIBD).

We have $t = 4$ treatments (shape), $b = 4$ blocks (plate) of size $k = 3$, each treatment replicated $r = 3$ times.

(b)

建構模型

$$\text{noise}_{ij} = \eta + \alpha_i + \tau_j + \epsilon_{ij} \ , \ \epsilon_{ij} \sim N(0, \sigma^2)$$

where α_i is the block (plate) effect with $i = 1, 2, 3, 4$, and τ_j is the treatment (shape) effect with $j = A, B, C, D$

觀察模型的 ANOVA table

```
out = aov(noise ~ plate + shape, resistor)
summary(out)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## plate        3 0.3474 0.11579   8.455 0.0211 *
## shape        3 0.4651 0.15502  11.319 0.0115 *
## Residuals    5 0.0685 0.01369
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\Rightarrow$ shape 的 p-value 呈現顯著，代表不同的形狀對噪音的影響有著顯著差距

接下來用 Tukey method 進行多重比較

```r
library(multcomp)
fitT = glht(out, linfct = mcp(shape = "Tukey"))
summary(fitT)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = noise ~ plate + shape, data = resistor)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## B - A == 0 -0.45500    0.10135  -4.490   0.0234 *
## C - A == 0 -0.15625    0.10135  -1.542   0.4812
## D - A == 0 -0.50375    0.10135  -4.971   0.0156 *
## C - B == 0  0.29875    0.10135   2.948   0.1075
## D - B == 0 -0.04875    0.10135  -0.481   0.9601
## D - C == 0 -0.34750    0.10135  -3.429   0.0649 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

$\Rightarrow$ (A,B) 和 (A,D) 這兩對組合皆有顯著差異

Problem 5. (3.35)

建構模型

$$y_{ijlm} \;=\; \eta \;+\; \text{day}_i \;+\; \text{operator}_j \;+\; \text{machine}_l \;+\; \text{method}_m \;+\; \epsilon_{ijlm} \quad , \quad \epsilon_{ijlm} \;\sim N(0, \sigma^2)$$

where day_i , operator_j , machine_l are block effects with $i, j = 1, ..., 5$, $l = \alpha, \beta, \gamma, \delta, \epsilon$, and method_m is treatment effect with $m = A, B, C, D, E$

觀察模型的 ANOVA table

```
assembly = read.table("assemblymethod.TXT", header = T)
assembly$Day = as.factor(assembly$Day)
assembly$Operator = as.factor(assembly$Operator)
fit5 = lm(Throughput ~ Day+Operator+Machine+Method, assembly)
anova(fit5)
```

```
## Analysis of Variance Table
##
## Response: Throughput
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Day        4  125.2    31.3  1.5343    0.2806
## Operator   4  167.2    41.8  2.0490    0.1800
## Machine    4 3424.8   856.2 41.9706 2.062e-05 ***
## Method     4 2857.6   714.4 35.0196 4.075e-05 ***
## Residuals  8  163.2    20.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\Rightarrow$ treatment effect method_m 呈現顯著，代表使用不同的方法對反應變數 throughput 所造成的影響有顯著差異

再進一步觀察模型中係數的數值

```
summary(fit5)
```

```
##
## Call:
## lm(formula = Throughput ~ Day + Operator + Machine + Method,
##     data = assembly)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -6.4   -1.6    0.2    2.0    3.2
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     102.600      3.725  27.547 3.25e-09 ***
## Day2             -3.200      2.857  -1.120  0.29512
## Day3             -5.200      2.857  -1.820  0.10619
## Day4             -3.000      2.857  -1.050  0.32431
## Day5             -6.600      2.857  -2.310  0.04965 *
## Operator2         4.800      2.857   1.680  0.13140
## Operator3         2.000      2.857   0.700  0.50369
## Operator4        -1.200      2.857  -0.420  0.68548
## Operator5         5.400      2.857   1.890  0.09537 .
## Machinebeta      -7.600      2.857  -2.661  0.02878 *
## Machinedelta     12.200      2.857   4.271  0.00272 **
## Machineepsilon    6.000      2.857   2.100  0.06889 .
## Machinegamma    -21.600      2.857  -7.562 6.54e-05 ***
## MethodB          11.600      2.857   4.061  0.00363 **
## MethodC          -4.000      2.857  -1.400  0.19900
## MethodD          25.400      2.857   8.892 2.03e-05 ***
## MethodE          16.000      2.857   5.601  0.00051 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.517 on 8 degrees of freedom
## Multiple R-squared:  0.9758, Adjusted R-squared:  0.9273
## F-statistic: 20.14 on 16 and 8 DF,  p-value: 9.907e-05
```

可以發現只有 effect method_B , method_C , method_D , method_E 有估計值，代表此模型有著 $\text{method}_A = 0$ 的 baseline constraint，此時其餘四者的估計值代表著使用不同 method 時和使用 method A 時，反應變數平均的差距，可以得到以下關係式

$$\begin{cases} \overline{y}_{11\alpha B} - \overline{y}_{11\alpha A} = 11.6 \\ \overline{y}_{11\alpha C} - \overline{y}_{11\alpha A} = -4 \\ \overline{y}_{11\alpha D} - \overline{y}_{11\alpha A} = 25.4 \\ \overline{y}_{11\alpha E} - \overline{y}_{11\alpha A} = 16 \end{cases} \Rightarrow \overline{y}_{11\alpha D} > \overline{y}_{11\alpha E} > \overline{y}_{11\alpha B} > \overline{y}_{11\alpha A} > \overline{y}_{11\alpha C}$$

所以使用 method D 會明顯優於其他 method