Design of experiment - Homework 07

統碩一 106354003 林健宏

※ Exercise 10.1

Three ANOVA tables are given for the results of a single experiment. These tables give sequential (Type I) sums of squares. Construct a Type II ANOVA table. What would you conclude about which effects and interactions are needed?

	DF	SS	MS
a	1	1.9242	1.9242
b	2	1584.2	792.1
a.b	2	19.519	9.7595
c	1	1476.7	1476.7
a.c	1	17.527	17.527
b.c	2	191.84	95.92
a.b.c	2	28.567	14.284
Error	11	166.71	15.155

(Figure 1)

	DF	SS	MS
b	2	1573	786.49
c	1	1428.7	1428.7
b.c	2	153.62	76.809
a	1	39.777	39.777
b.a	2	69.132	34.566
c.a	1	27.51	27.51
b.c.a	2	28.567	14.284
Error	11	166.71	15.155

(Figure 2)

	DF	SS	MS
c	1	1259.3	1259.3
a	1	9.0198	9.0198
c.a	1	0.93504	0.93504
b	2	1776.1	888.04
c.b	2	169.92	84.961
a.b	2	76.449	38.224
c.a.b	2	28.567	14.284
Error	11	166.71	15.155

(Figure 3)

題目模型：

$$y_{ijk} = a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk} + abc_{ijk} + \epsilon_{ijk}$$

$$i = 1,2; j = 1,2,3; k = 1,2$$

其中：

$$\sum_{i=1}^{2} a_i = \sum_{j=1}^{3} b_j = \sum_{k=1}^{2} c_k = 0$$

$$\sum_{i=1}^{2} ab_{ij} = \sum_{j=1}^{3} ab_{ij} = \sum_{i=1}^{2} ac_{ik} = \sum_{k=1}^{2} ac_{ik} = \sum_{j=1}^{3} bc_{jk} = \sum_{k=1}^{2} bc_{jk} = 0$$

$$\sum_{i=1}^{2} abc_{ijk} = \sum_{j=1}^{3} abc_{ijk} = \sum_{k=1}^{2} abc_{ijk} = 0 \quad \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

題目所給的表(Figure1~3)為 Type I 的 Sums of square，由於本題並無截距項，因此在計算上不考慮截距的部分。例如 Figure 1 中的 a.b 指得是 SS(ab) = SS(ab | a, b)。

其中

$$SS(ab) = SS(ab \mid a, b) = SS(ab, a, b) - SS(a, b)$$

而題目欲計算的部分為 Type II 的 Sums of square，由 Type I 的 Sums of square 去計算。

Type II 的 Sums of square 的計算模式為，使用 a.b 為例

$$SS(ab) = SS(ab \mid a, b, c, ac, bc) = SS(a, b, c, ab, ac, bc) - SS(a, b, c, ac, bc) \Rightarrow \text{Figure 3 的 a.b}$$

透過 Figure 1~3 計算 Type II Sums of square 得到下表

	DF	**SS**	**MS**
SS(a)	1	39.777	39.777
SS(b)	2	1776.1	888.04
SS(a.b)	2	76.449	38.224
SS(c)	1	1476.7	1476.7
SS(a.c)	1	27.51	27.51
SS(b.c)	2	191.84	95.92
SS(a.b.c)	2	28.567	14.284
Error	11	166.71	15.155
Total	22	3486.9872	-

(Figure 4)

從 Type II 表(Figure 4)中發現各因子與交互作用項的 MS 有大有小，而我認為 MS 越大可能對模型影響越大，即因子 b,c 以及交互作用項 bc 對於模型而言有較大的影響。

※ Exercise 10.1

In a study of patient confidentiality, a large number of pediatricians was surveyed. Each pediatrician was given a "fable" about a female patient less than 18 years old. There were sixteen different fables, the combinations of the factors complaint (C: 1—drug problem, 2—venereal disease), age (A: 1—14 years, 2—17 years), the length of time the pediatrician had known the family (L: 1—less than 1 year, 2—more than 5 years), and the maturity of patient (M: 1—immature for age, 2—mature for age). The response at each combination of factor levels is the fraction of doctors who would keep confidentiality and not inform the patient's parents (data modeled on Moses 1987). Analyze these data to determine which factors influence the pediatrician's decision.

C	A	L	M	Response	C	A	L	M	Response
1	1	1	1	.445	2	1	1	1	.578
1	1	1	2	.624	2	1	1	2	.786
1	1	2	1	.360	2	1	2	1	.622
1	1	2	2	.493	2	1	2	2	.755
1	2	1	1	.513	2	2	1	1	.814
1	2	1	2	.693	2	2	1	2	.902
1	2	2	1	.534	2	2	2	1	.869
1	2	2	2	.675	2	2	2	2	.902

從資料來看，共有 4 個因子，分別為 complaint (C), age (A), the length of time (L), maturity of patient (M), 以及一組對應的資料(Response)

Complaint : 病患是否有毒癮或性病：無→ 1， 有→ 2

Age : 病患屬於哪個年齡區間：1-14 歲→ 1，2-17 歲→ 2

Length : 小兒科醫生熟識病患家庭的時間：少於一年→1，超過 5 年→ 2

Maturity：病患心智是否成熟：不成熟→1，成熟→2

Response：醫生是否對病患家屬保密病患情況的比率

而在實驗設計中至少需要2^4組樣本去做分析，樣本分配的狀況參考下圖

I	A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	ABCD
1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1
1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1
1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

本題剛好提供了 16 組樣本，而且也依循上表(-1⇒1, +1⇒ 2)的方式去得出 Response。若是要完全去分析每個因子以及交互作用項的是否顯著，依照本題給的樣本數會造成 ANOVA table 中的 SSE (Sum of square of error) 的自由度為 0，進而無法分析。因此，我在使用 ANOVA 分析上，剔除最高交互項，即四因子的交互作用項，讓 SSE 的自由度為 1，使得可以使用 ANOVA 進行分析。

ANOVA 檢定

H_0: 因子或交互作用項不顯著

H_1: 因子或交互作用項顯著

	Df	Sum Sq	Mean Sq	F	p-value
C	1	0.223493	0.223493	21159.0592	0.004376
A	1	0.095945	0.095945	9083.5562	0.006679
L	1	0.001314	0.001314	124.4083	0.056924
M	1	0.074939	0.074939	7094.8225	0.007558
C:A	1	0.004001	0.004001	378.7515	0.032683
C:L	1	0.004935	0.004935	467.2249	0.029431
A:L	1	0.004258	0.004258	403.0828	0.031683
C:M	1	0.001828	0.001828	173.0237	0.048305
A:M	1	0.002783	0.002783	263.4379	0.039174
L:M	1	0.002889	0.002889	273.5207	0.038446
C:A:L	1	0.001958	0.001958	185.3787	0.046674
C:A:M	1	0.003278	0.003278	310.3018	0.036101
C:L:M	1	0.000127	0.000127	11.9822	0.179038
A:L:M	1	0.000046	0.000046	4.3136	0.285666
Residuals	1	0.000011	0.000011	-	-

(table 1)

在給定顯著水準為 0.05 下，僅有因子 L，以及交互作用項 CLM、ALM 呈現不拒絕H_0的結果，顯示這三個不顯著，但其餘與 L 有關的交互作用項都呈現拒絕H_0的假設，顯示該交互作用項為顯著，加上因子 L 的 p-value 非常接近 0.05。因此我僅將交互作用項 CLM 與 ALM 兩項排除，並重新進行 ANOVA 分析，並得到下表(table 2)。

ANOVA 檢定

H_0: 因子或交互作用項不顯著

H_1: 因子或交互作用項顯著

	Df	Sum Sq	Mean Sq	F	p-value
C	1	0.223493	0.223493	3670.08	0.00000991
A	1	0.095945	0.095945	1575.56	0.00003518
L	1	0.001314	0.001314	21.579	0.018808
M	1	0.074939	0.074939	1230.611	0.00005094
C:A	1	0.004001	0.004001	65.695	0.003925
C:L	1	0.004935	0.004935	81.041	0.002894
A:L	1	0.004258	0.004258	69.915	0.003587
C:M	1	0.001828	0.001828	30.011	0.011961
A:M	1	0.002783	0.002783	45.694	0.006614
L:M	1	0.002889	0.002889	47.443	0.006269
C:A:L	1	0.001958	0.001958	32.154	0.010864
C:A:M	1	0.003278	0.003278	53.822	0.005233
Residuals	3	0.000183	0.000061		

(table 2)

在給定顯著水準為 0.05 下，table 2 的各項因子與交互作用項都拒絕H_0假設，都顯示有顯著的效果。因此此題最終線性模型為

$$\text{Response}_i = 0.4430625 + 0.1402500\text{C}_i + 0.0662500\ \text{A}_i - 0.0811250L_i + 0.1828750M_i + 0.1647500\ C_iA_i + 0.1145000\ C_iL_i + 0.1095000A_iL_i + 0.0145000\ C_iM_i + 0.0045000A_iM_i - 0.0537500L_iM_i - 0.0885000\ C_iA_iL_i - 0.1145000\ C_iA_iM_i$$

其中$i = 1 \sim 16$，C,A,L 以及 M 為 1 或 2

影響醫生最大的因子為 Complaint:病人是否有毒癮或性病。

Appendix: http://rpubs.com/YaPi/393000

Code

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

# data
Response = c(0.445,0.624,0.360,0.493,0.513,0.693,0.534,0.675,0.578,0.786,0.622,0.755,
0.814,0.902,0.869,0.902)
C <- factor(rep(c("1","2"), each = 8))
A <- factor(rep(c("1","2"), each = 4, times = 2))
L <- factor(rep(c("1","2"), each = 2, times = 4))
M <- factor(rep(c("1","2"), times = 8))
temp <- cbind(C,A,L,M,Response) %>% as.matrix()
temp

##       C A L M Response
##  [1,] 1 1 1 1    0.445
##  [2,] 1 1 1 2    0.624
##  [3,] 1 1 2 1    0.360
##  [4,] 1 1 2 2    0.493
##  [5,] 1 2 1 1    0.513
##  [6,] 1 2 1 2    0.693
##  [7,] 1 2 2 1    0.534
##  [8,] 1 2 2 2    0.675
##  [9,] 2 1 1 1    0.578
## [10,] 2 1 1 2    0.786
## [11,] 2 1 2 1    0.622
## [12,] 2 1 2 2    0.755
```

```
## [13,] 2 2 1 1    0.814
## [14,] 2 2 1 2    0.902
## [15,] 2 2 2 1    0.869
## [16,] 2 2 2 2    0.902

# Anova table
lm.result1 <- lm(Response ~ C*A*L*M-C:A:L:M)
anova(lm.result1)

## Analysis of Variance Table
##
## Response: Response
##           Df   Sum Sq  Mean Sq    F value   Pr(>F)
## C          1 0.223493 0.223493 21159.0592 0.004376 **
## A          1 0.095945 0.095945  9083.5562 0.006679 **
## L          1 0.001314 0.001314   124.4083 0.056924 .
## M          1 0.074939 0.074939  7094.8225 0.007558 **
## C:A        1 0.004001 0.004001   378.7515 0.032683 *
## C:L        1 0.004935 0.004935   467.2249 0.029431 *
## A:L        1 0.004258 0.004258   403.0828 0.031683 *
## C:M        1 0.001828 0.001828   173.0237 0.048305 *
## A:M        1 0.002783 0.002783   263.4379 0.039174 *
## L:M        1 0.002889 0.002889   273.5207 0.038446 *
## C:A:L      1 0.001958 0.001958   185.3787 0.046674 *
## C:A:M      1 0.003278 0.003278   310.3018 0.036101 *
## C:L:M      1 0.000127 0.000127    11.9822 0.179038
## A:L:M      1 0.000046 0.000046     4.3136 0.285666
## Residuals  1 0.000011 0.000011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# delete non-significant factor
#Anova table
lm.result2 <- lm(Response~C*A*L*M-C:A:L:M-C:L:M-A:L:M)
anova(lm.result2)

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

```
##           Df   Sum Sq  Mean Sq  F value    Pr(>F)
## C          1 0.223493 0.223493 3670.080 9.909e-06 ***
## A          1 0.095945 0.095945 1575.560 3.518e-05 ***
## L          1 0.001314 0.001314   21.579  0.018808 *
## M          1 0.074939 0.074939 1230.611 5.094e-05 ***
## C:A        1 0.004001 0.004001   65.695  0.003925 **
## C:L        1 0.004935 0.004935   81.041  0.002894 **
## A:L        1 0.004258 0.004258   69.915  0.003587 **
## C:M        1 0.001828 0.001828   30.011  0.011961 *
## A:M        1 0.002783 0.002783   45.694  0.006614 **
## L:M        1 0.002889 0.002889   47.443  0.006269 **
## C:A:L      1 0.001958 0.001958   32.154  0.010864 *
## C:A:M      1 0.003278 0.003278   53.822  0.005233 **
## Residuals  3 0.000183 0.000061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm.result2$coefficients
```

```
## (Intercept)         C2         A2         L2         M2      C2:A2
##   0.4430625   0.1402500   0.0662500  -0.0811250   0.1828750   0.1647500
##       C2:L2      A2:L2      C2:M2      A2:M2      L2:M2    C2:A2:L2
##   0.1145000   0.1095000   0.0145000   0.0045000  -0.0537500  -0.0885000
##    C2:A2:M2
##  -0.1145000
```