4.5: Computational Aspects of the Effects Model

Model 4 - The Effects Model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{4.5.1}$$

where τ_i are the the deviations of each factor level mean from the overall mean so that $\sum_{i=1}^{T} \tau_i = 0$.

In the effects model that we discussed in chapter 3, the treatment means were not estimated but instead, the τ_i's, or the deviations of treatment means from the overall mean, were estimated. The model must include the overall mean, which is estimated by the intercept, and hence the design matrix to be inputted for IML is:

```
/* The Effects Model */
x={
1    1    0,
1    1    0,
1    0    1,
1    0    1,
1    -1    -1,
1    -1    -1};
```

Here we have another omission of a treatment level, but for a different reason. In the effects model, we have the constraint $\sum \tau_i = 0$. As a result, coding for one treatment level can be omitted.

Running the IML program with this design matrix yields:

Regression Coefficients	
Beta_0	3.5
Beta_1	-2
Beta_2	0

ANOVA				
Treatment	dF	SS	MS	F
	2	16	8	16
Error	3	1.5	0.5	
Total	5	17.5		

The regression coefficient Beta_0 is the overall mean and the coefficients β_1 and β_2 are τ_1 and τ_2, respectively. The estimate for τ_3 is obtained as $-(\tau_1) - (\tau_2) = 2.0$.

In Minitab, if we change the coding now to be Effect coding (-1,0,+1), which is the default setting, we get the following:

Regression Equation

y = 3.500 - 2.000 trt_A - 0.000 trt_B + 2.000 trt_C

The ANOVA table is the same as for the dummy-variable regression model above. We can also observe that the factor level means and General Linear F Statistics values obtained for all 3 representations (cell means, dummy coded regression and effects coded regression) are identical, confirming that the 3 representations are identical.

The intermediates were:

xprimex			check			xprimey
6	0	0	1	0	0	21
0	4	2	0	1	0	-8
0	2	4	0	0	1	-4

xprimexinv		
0.1666667	0	0
0	0.3333333	-0.166667
0	-0.166667	0.3333333

SumY2	CF
89.5	73.5

By coding treatment or factor levels into numerical terms, we can use regression methods to perform the ANOVA.

To state the effects model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{4.5.2}$$

as a regression model, we need to include $\mu, \tau_i, \ldots, \tau_T$ as elements in the parameter vector $\boldsymbol{\beta}$ of the GLM model. Note that, in the case of equal replication at each factor level, the deviations satisfy the following constraint:

$$\sum_{i=1}^{T} \tau_i = 0 \tag{4.5.3}$$

This implies one of the τ_i parameters is not needed since it can be expressed in terms of the other $T-1$ parameters and need not be included in the $\boldsymbol{\beta}$ parameter vector. We shall drop the parameter τ_T from the regression equation, as it can be expressed in terms of the other $T-1$ parameters τ_i as follows:

$$\tau_T = -\tau_1 - \tau_2 - \ldots - \tau_{T-1} \tag{4.5.4}$$

Thus, the $\boldsymbol{\beta}$ vector of the GLM is a $T \times 1$ vector containing only the parameters $\mu, \tau_1, \ldots, \tau_{T-1}$ for the linear model.

To illustrate how a linear model is developed with this approach, consider a single-factor study with $T=3$ factor levels when $n_1 = n_2 = n_3 = 2$. The $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ matrices for this case are as follows:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix} \tag{4.5.5}$$

where β_0, β_1, and β_2 correspond to μ, τ_1, and τ_2 respectively.

Note that the vector of expected values $\mathbf{E}\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta}$ yields the following:

$$\mathbf{E}\{\mathbf{Y}\} = \mathbf{X}\beta \tag{4.5.6}$$

$$\begin{bmatrix} E\{Y_{11}\} \\ E\{Y_{12}\} \\ E\{Y_{21}\} \\ E\{Y_{22}\} \\ E\{Y_{31}\} \\ E\{Y_{32}\} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{4.5.7}$$

$$= \begin{bmatrix} \mu+\tau_1 \\ \mu+\tau_1 \\ \mu+\tau_2 \\ \mu+\tau_2 \\ \mu-\tau_1-\tau_2 \\ \mu-\tau_1-\tau_2 \end{bmatrix} \tag{4.5.8}$$

Since $\tau_3 = -\tau_1 - \tau_2$, as shown above, we see that $E\{Y_{31}\} = E\{Y_{32}\} = \mu + \tau_3$. Thus, the above $\mathbf{X}$ matrix and $\boldsymbol{\beta}$ vector representation provides the appropriate expected values for all factor levels as expressed below:

$$E\{Y_{ij}\} = \mu + \tau_i \tag{4.5.9}$$

Using R: Effects Model

Steps in R

1. Define response variable and design matrix

```
y<-matrix(c(2,1,3,4,6,5), ncol=1)
x = matrix(c(1,1,0,1,1,0,1,0,1,1,0,1,1,-1,-1,1,-1,-1),ncol=3,nrow=6,byrow=TRUE)
```

2. Regression coefficients

```
beta<-solve(t(x)%*%x)%*%(t(x)%*%y)
#  beta
#     [,1]
# [1,]  3.5
# [2,] -2.0
# [3,]  0.0
```

3. Calculate the entries of the ANOVA Table

```
n<-nrow(y)
p<-ncol(x)
J<-matrix(1,n,n)
ss_tot = (t(y)%*%y) - (1/n)*(t(y)%*%J)%*%y #17.5
ss_trt = t(beta)%*%(t(x)%*%y) - (1/n)*(t(y)%*%J)%*%y #16
ss_error = ss_tot - ss_trt #1.5
total_df=n-1 #5
trt_df=p-1 #2
error_df=n-p #3
MS_trt = ss_trt/(p-1) #8
MS_error = ss_error / error_df #0.5
F=MS_trt/MS_error #16
```

4. Creating the ANOVA table

```
ANOVA <- data.frame(
c ("","Treatment","Error", "Total"),
c("DF", trt_df,error_df,total_df),
c("SS", ss_trt, ss_error, ss_tot),
c("MS", MS_trt, MS_error, ""),
c("F",F,"",""),
stringsAsFactors = FALSE)
names(ANOVA) <- c(" ", "  ", " ","","")
```

5. Print the ANOVA table

```
print(ANOVA)
# 1           DF   SS  MS  F
# 2 Treatment  2   16   8 16
# 3     Error  3  1.5 0.5
# 4     Total  5 17.5
```

Copy code

6. Intermediates in the matrix computations

```
xprimex<-t(x)%*%x
#  xprimex
#      [,1] [,2] [,3]
# [1,]    6    0    0
# [2,]    0    4    2
# [3,]    0    2    4
xprimey<-t(x)%*%y
#  xprimey
#      [,1]
# [1,]   21
# [2,]   -8
# [3,]   -4
xprimexinv<-solve(t(x)%*%x)
#  xprimexinv
#           [,1]       [,2]       [,3]
# [1,] 0.1666667  0.0000000  0.0000000
# [2,] 0.0000000  0.3333333 -0.1666667
# [3,] 0.0000000 -0.1666667  0.3333333
check<-xprimexinv%*%xprimex
# check
#      [,1] [,2] [,3]
# [1,]    1    0    0
# [2,]    0    1    0
# [3,]    0    0    1
```

```
SumY2<-t(beta)%*%(t(x)%*%y) #89.5
CF<-(1/n)*(t(y)%*%J)%*%y # 73.5
```

7. Regression Equation and ANOVA table

```
trt_level1<-x[,2]
trt_level2<-x[,3]
model<-lm(y~trt_level1+trt_level2)
```

8. With the command summary(model) we can get the following output:

```
Call:
lm(formula = y ~ trt_level1 + trt_level2)
Residuals:
1    2    3    4    5    6
0.5 -0.5 -0.5  0.5  0.5 -0.5
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.500e+00  2.887e-01  12.124  0.00121 **
trt_level1  -2.000e+00  4.082e-01  -4.899  0.01628 *
trt_level2  -1.282e-16  4.082e-01   0.000  1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.7071 on 3 degrees of freedom
Multiple R-squared:  0.9143,    Adjusted R-squared:  0.8571
F-statistic:    16 on 2 and 3 DF,  p-value: 0.02509
```

From the output we can see the estimates for the coefficients are b0=3.5, b1=-2, b2=0 and the F-statistic is 16 with a p-value of 0.02509.

By using the estimates we can write the regression equation:

> y=3.5-2 trt_level1-0 trt_level2+2 trt_level3

The estimator τ_3 is obtained as $-\tau_1 - \tau_2 = 2$

9. With the command anova(model) we can get the following output:

```
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value  Pr(>F)
trt_level1  1   16.0    16.0      32 0.01094 *
trt_level2  1    0.0     0.0       0 1.00000
Residuals   3    1.5     0.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that R is giving the sequential sum of squares in the ANOVA table.