8. Analysis of Variance for Unbalanced Two-Factor Experiments

order of the sequential SS matters!

ANOVA for Unbalanced Two-Factor Experiments

When data are unbalanced, the Type I ANOVA test for two-way interactions is the same as the test for two-way interactions discussed previously.

However, the Type I ANOVA tests for individual factors are not the tests for main effects discussed previously.

Furthermore, the Type I results for individual factors depend on the order that the factors appear in the Type I ANOVA table.

Example Unbalanced Two-Factor Experiment

An experiment was conducted to study the effect of storage time and storage temperature on the amount of active ingredient in a drug lost during storage. A total of 16 vials of the drug, each containing approximately 30 mg/mL of active ingredient, were assigned (using a completely randomized design) to the following treatments:

y

1. Storage for 3 months at 20° C
2. Storage for 3 months at 30° C
3. Storage for 6 months at 20° C
4. Storage for 6 months at 30° C

levels of treatment-combinations

Example Unbalanced Two-Factor Experiment

6 of the 16 vials were damaged during shipment to the lab where the active ingredient was measured. The amount of active ingredient was measured only for the 10 undamaged vials. The table below shows the amount of active ingredient lost during storage (in tenths of mg/mL) for each of the undamaged vials.

	Storage Temperature	
Storage Time	20°	30°
3 months	3 5	11 13 15
6 months	5 6 6 7	16

as temperature ↑ y ↑

as length ↑ y ↑

A Cell Means Model for the Data

Let y_{ijk} denote the amount of active ingredient lost from the k^{th} vial treated with the i^{th} storage time and j^{th} temperature.

Let n_{ij} denote the number of vials measured for the i^{th} storage time and j^{th} temperature.

Suppose $\boxed{y_{ijk} = \mu_{ij} + \epsilon_{ijk}}$ $(i = 1, 2;\ j = 1, 2;\ k = 1, \ldots, n_{ij})$, where $\mu_{11}, \mu_{12}, \mu_{21},$ and μ_{22} are unknown real-valued parameters and the ϵ_{ijk} terms are $i.i.d.$ normal random variables with mean 0 and some unknown variance $\sigma^2 > 0$.

GMM NE

Model in Matrix and Vector Form

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{214} \\ y_{221} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{123} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{213} \\ \epsilon_{214} \\ \epsilon_{221} \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

Model in Matrix and Vector Form

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{214} \\ y_{221} \end{bmatrix} = \begin{bmatrix} \mu_{11} \\ \mu_{11} \\ \mu_{12} \\ \mu_{12} \\ \mu_{12} \\ \mu_{21} \\ \mu_{21} \\ \mu_{21} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{123} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{213} \\ \epsilon_{214} \\ \epsilon_{221} \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

We could consider a sequence of progressively more complex models for the response mean that lead up to our full cell means model.

1. $\mathrm{E}(\mathrm{y}_{\mathrm{ijk}}) = \mu$

add time

2. $\mathrm{E}(\mathrm{y}_{\mathrm{ijk}}) = \mu + \alpha_{\mathrm{i}}$

add time & temp

3. $\mathrm{E}(\mathrm{y}_{\mathrm{ijk}}) = \mu + \alpha_{\mathrm{i}} + \beta_{\mathrm{j}}$

4. $\mathrm{E}(\mathrm{y}_{\mathrm{ijk}}) = \mu + \alpha_{\mathrm{i}} + \beta_{\mathrm{j}} + \gamma_{\mathrm{ij}} \iff \mathrm{E}(\mathrm{y}_{\mathrm{ijk}}) = \mu_{\mathrm{ij}}$

time, temp, interaction

Matrices with Nested Column Spaces

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{X}_3 = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \text{ and}$$

Matrices with Nested Column Spaces

$$\boldsymbol{X}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

ANOVA Table

Source	Sum of Squares	DF
$\text{Time}\|1$	$\boldsymbol{y}^\top(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$2 - 1 = 1$
$\text{Temp}\|1, \text{Time}$	$\boldsymbol{y}^\top(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$	$3 - 2 = 1$
$\text{Time} \times \text{Temp}\|1, \text{Time}, \text{Temp}$	$\boldsymbol{y}^\top(\boldsymbol{P}_4 - \boldsymbol{P}_3)\boldsymbol{y}$	$4 - 3 = 1$
Error	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_4)\boldsymbol{y}$	$10 - 4 = 6$
C. Total	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$10 - 1 = 9$

The Time-Temperature Dataset

```
> time=factor(rep(c(3,6),each=5))
> temp=factor(rep(c(20,30,20,30),c(2,3,4,1)))
> a=time
> b=temp
> y=c(3,5,11,13,15,5,6,6,7,16)
```

```
> d=data.frame(time,temp,y)
> d
   time temp  y
1     3   20  3
2     3   20  5
3     3   30 11
4     3   30 13
5     3   30 15
6     6   20  5
7     6   20  6
8     6   20  6
9     6   20  7
10    6   30 16
```

```
> x1=matrix(1,nrow=nrow(d),ncol=1)
> x1
      [,1]
 [1,]    1
 [2,]    1
 [3,]    1
 [4,]    1
 [5,]    1
 [6,]    1
 [7,]    1
 [8,]    1
 [9,]    1
[10,]    1
```

```
> x2=cbind(x1,model.matrix(~0+a))
> x2
      a3 a6
1  1  1  0
2  1  1  0
3  1  1  0
4  1  1  0
5  1  1  0
6  1  0  1
7  1  0  1
8  1  0  1
9  1  0  1
10 1  0  1
```



```
> x3=cbind(x2,model.matrix(~0+b))
> x3
      a3 a6 b20 b30
1   1  1  0   1   0
2   1  1  0   1   0
3   1  1  0   0   1
4   1  1  0   0   1
5   1  1  0   0   1
6   1  0  1   1   0
7   1  0  1   1   0
8   1  0  1   1   0
9   1  0  1   1   0
10  1  0  1   0   1
```

```
> x4=model.matrix(~0+b:a)
> x4
   b20:a3 b30:a3 b20:a6 b30:a6
1       1      0      0      0
2       1      0      0      0
3       0      1      0      0
4       0      1      0      0
5       0      1      0      0
6       0      0      1      0
7       0      0      1      0
8       0      0      1      0
9       0      0      1      0
10      0      0      0      1
```

```
> library(MASS)
> proj=function(x){
+   x%*%ginv(t(x)%*%x)%*%t(x)
+ }
>
> p1=proj(x1)
> p2=proj(x2)
> p3=proj(x3)
> p4=proj(x4)
> I=diag(rep(1,10))
```

```
> SumOfSquares=c(
+ t(y)%*%(p2-p1)%*%y,
+ t(y)%*%(p3-p2)%*%y,
+ t(y)%*%(p4-p3)%*%y,
+ t(y)%*%(I-p4)%*%y,
+ t(y)%*%(I-p1)%*%y)
>
> Source=c(
+ "Time|1",
+ "Temp|1,Time",
+ "Time x Temp|1,Time,Temp",
+ "Error",
+ "C. Total")
```

```
> data.frame(Source,SumOfSquares)
                    Source SumOfSquares
1                   Time|1         4.90
2              Temp|1,Time       176.72
3 Time x Temp|1,Time,Temp          0.48
4                    Error        12.00
5                 C. Total       194.10
> anova(lm(y~time+temp+time:temp,data=d))

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
time       1   4.90    4.90    2.45    0.1686
temp       1 176.72  176.72   88.36 8.233e-05 ***
time:temp  1   0.48    0.48    0.24    0.6416
Residuals  6  12.00    2.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

lack of evidence

overwhelm. evidence

lack of evidence

What do the F-tests in this ANOVA table test?

Recall the null hypothesis for F_j is true if and only if

$$\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} = 0.$$

We have the following equivalent conditions

$$\begin{aligned}
\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} = 0 &\iff \boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)^\top (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} = 0 \\
&\iff \quad || \, (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} \, ||^2 = 0 \\
&\iff \quad (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \\
&\iff \quad \boldsymbol{C}\boldsymbol{\beta} = \mathbf{0},
\end{aligned}$$

where $\boldsymbol{C}$ is any full-row-rank matrix with the same row space as $(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$.

What do the F-tests in this ANOVA table test?

Let's take a look at $(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$ for each test in the ANOVA table.

When computing $(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$, we can use any model matrix $\boldsymbol{X}$ that specifies one unrestricted treatment mean for each of the four treatments.

The entries in any rows of $(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$ are coefficients defining linear combinations of the elements of the parameter vector $\boldsymbol{\beta}$ that corresponds to the chosen model matrix $\boldsymbol{X}$.

Our Choice for $\boldsymbol{X}$ and $\boldsymbol{\beta}$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix}$$

```
> x=x4
> fractions((p2-p1)%*%x)
   b20:a3 b30:a3 b20:a6 b30:a6
1     1/5   3/10   -2/5  -1/10
2     1/5   3/10   -2/5  -1/10
3     1/5   3/10   -2/5  -1/10
4     1/5   3/10   -2/5  -1/10
5     1/5   3/10   -2/5  -1/10
6    -1/5  -3/10    2/5   1/10
7    -1/5  -3/10    2/5   1/10
8    -1/5  -3/10    2/5   1/10
9    -1/5  -3/10    2/5   1/10
10   -1/5  -3/10    2/5   1/10

> fractions(2*(p2-p1)%*%x)[1,]
b20:a3 b30:a3 b20:a6 b30:a6
   2/5    3/5   -4/5   -1/5
```


end
lecture 16
2-26-25

$$(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \iff \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0},$$

where

$$\begin{aligned} \boldsymbol{C}\boldsymbol{\beta} &= \begin{bmatrix} \frac{2}{5} & \frac{3}{5} & -\frac{4}{5} & -\frac{1}{5} \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} \\ &= \left(\frac{2}{5}\mu_{11} + \frac{3}{5}\mu_{12}\right) - \left(\frac{4}{5}\mu_{21} + \frac{1}{5}\mu_{22}\right). \end{aligned}$$