8. Analysis of Variance for Unbalanced Two-Factor Experiments

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:

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

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

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 $y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad (i = 1, 2;\ j = 1, 2;\ k = 1, \ldots, n_{ij})$, 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 $\sigma^2 > 0$.

Model in Matrix and Vector Form

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

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

Model in Matrix and Vector Form

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

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

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

1. $E(y_{ijk}) = \mu$
2. $E(y_{ijk}) = \mu + \alpha_i$
3. $E(y_{ijk}) = \mu + \alpha_i + \beta_j$
4. $E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \Longleftrightarrow E(y_{ijk}) = \mu_{ij}$

Matrices with Nested Column Spaces

$$X_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \; X_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \; X_3 = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \text{ and}$$

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

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
```

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

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

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

We have the following equivalent conditions

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

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 $\boldsymbol{\beta}$

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

Time$|_1$ ANOVA Test

```
> 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
```

Time$|1$ ANOVA Test

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

where

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

Time$|1$ ANOVA Test $\neq$ Time Main Effect Test

Null for Time$|1$ ANOVA test:

$$\frac{2}{5}\mu_{11} + \frac{3}{5}\mu_{12} = \frac{4}{5}\mu_{21} + \frac{1}{5}\mu_{22}$$

Null for Time main effect test:

$$\frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{12} \quad = \quad \frac{1}{2}\mu_{21} + \frac{1}{2}\mu_{22}$$

i.e.

$$\bar{\mu}_{1\cdot} \quad = \quad \bar{\mu}_{2\cdot}$$

A Closer Look at the Time$|1$ ANOVA Test

The ANOVA Time$|1$ test is comparing the averages for the two storage times, ignoring storage temperature.

	Storage Temperature		
Storage Time	20°	30°	NOT LSMEANS
3 months	3 5	11 13 15	$\frac{2}{5}\hat{\mu}_{11} + \frac{3}{5}\hat{\mu}_{12}$
6 months	5 6 6 7	16	$\frac{4}{5}\hat{\mu}_{21} + \frac{1}{5}\hat{\mu}_{22}$

A Closer Look at the Time$|1$ ANOVA Test

The ANOVA Time$|1$ test is comparing the averages for the two storage times, ignoring storage temperature.

Storage Time	Storage Temperature 20°	30°	NOT LSMEANS
3 months	3 5	11 13 15	$\frac{2}{5}\left(\frac{3+5}{2}\right)+\frac{3}{5}\left(\frac{11+13+15}{3}\right)$
6 months	5 6 6 7	16	$\frac{4}{5}\left(\frac{5+6+6+7}{4}\right)+\frac{1}{5}\left(\frac{16}{1}\right)$

A Closer Look at the Time$|1$ ANOVA Test

The ANOVA Time$|1$ test is comparing the averages for the two storage times, ignoring storage temperature.

Storage Time	Storage Temperature 20°	30°	NOT LSMEANS
3 months	3 5	11 13 15	$\left(\frac{3+5+11+13+15}{5}\right) = 9.4$
6 months	5 6 6 7	16	$\left(\frac{5+6+6+7+16}{5}\right) = 8.0$

The Test for Time Main Effects

The test for time main effects is based on LSMEANS.

Storage Time	Storage Temperature 20°	30°	LSMEANS
3 months	3 5	11 13 15	$\frac{1}{2}\hat{\mu}_{11} + \frac{1}{2}\hat{\mu}_{12}$
6 months	5 6 6 7	16	$\frac{1}{2}\hat{\mu}_{21} + \frac{1}{2}\hat{\mu}_{22}$

The Test for Time Main Effects

The test for time main effects is based on LSMEANS.

Storage Time	Storage Temperature 20°	30°	LSMEANS
3 months	3 5	11 13 15	$\frac{1}{2}\left(\frac{3+5}{2}\right)+\frac{1}{2}\left(\frac{11+13+15}{3}\right)$
6 months	5 6 6 7	16	$\frac{1}{2}\left(\frac{5+6+6+7}{4}\right)+\frac{1}{2}\left(\frac{16}{1}\right)$

The Test for Time Main Effects

The test for time main effects is based on a comparison of LSMEANS.

Storage Time	Storage Temperature 20°	30°	LSMEANS
3 months	3 5	11 13 15	$\frac{1}{2}4 + \frac{1}{2}13 = 8.5$
6 months	5 6 6 7	16	$\frac{1}{2}6 + \frac{1}{2}16 = 11.0$

Temp$|1$,Time ANOVA Test

```
> fractions((p3-p2)%*%x)
   b20:a3 b30:a3 b20:a6 b30:a6
1    9/25  -9/25   6/25  -6/25
2    9/25  -9/25   6/25  -6/25
3   -6/25   6/25  -4/25   4/25
4   -6/25   6/25  -4/25   4/25
5   -6/25   6/25  -4/25   4/25
6    3/25  -3/25   2/25  -2/25
7    3/25  -3/25   2/25  -2/25
8    3/25  -3/25   2/25  -2/25
9    3/25  -3/25   2/25  -2/25
10 -12/25  12/25  -8/25   8/25

> fractions((25/15)*(p3-p2)%*%x)[1,]
b20:a3 b30:a3 b20:a6 b30:a6
   3/5   -3/5    2/5   -2/5
```

Temp$|1$,Time ANOVA Test

$$(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \Longleftrightarrow \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0},$$

where

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

This is not the test for a storage temperature main effect.

Time$\times$Temp$|1$,Time,Temp ANOVA Test

```
> fractions((p4-p3)%*%x)
   b20:a3 b30:a3 b20:a6 b30:a6
1    6/25  -6/25  -6/25   6/25
2    6/25  -6/25  -6/25   6/25
3   -4/25   4/25   4/25  -4/25
4   -4/25   4/25   4/25  -4/25
5   -4/25   4/25   4/25  -4/25
6   -3/25   3/25   3/25  -3/25
7   -3/25   3/25   3/25  -3/25
8   -3/25   3/25   3/25  -3/25
9   -3/25   3/25   3/25  -3/25
10  12/25 -12/25 -12/25  12/25

> fractions((25/6)*(p4-p3)%*%x)[1,]
b20:a3 b30:a3 b20:a6 b30:a6
     1     -1     -1      1
```

Time$\times$Temp$|1$,Time,Temp ANOVA Test

$$(\boldsymbol{P}_4 - \boldsymbol{P}_3)\boldsymbol{X\beta} = \boldsymbol{0} \iff \boldsymbol{C\beta} = \boldsymbol{0},$$

where

$$\begin{aligned} \boldsymbol{C\beta} &= [1 \quad -1 \quad -1 \quad 1] \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} \\ &= \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}. \end{aligned}$$

This is the test for Time $\times$ Temp interaction.

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

1. $E(y_{ijk}) = \mu$
2. $E(y_{ijk}) = \mu + \beta_j$
3. $E(y_{ijk}) = \mu + \alpha_i + \beta_j$
4. $E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \Longleftrightarrow E(y_{ijk}) = \mu_{ij}$

```
> 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+b))
> x2
     b20 b30
1  1   1   0
2  1   1   0
3  1   0   1
4  1   0   1
5  1   0   1
6  1   1   0
7  1   1   0
8  1   1   0
9  1   1   0
10 1   0   1
```

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

ANOVA Table

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

```
> 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(
+ "Temp|1",
+ "Time|1,Temp",
+ "Temp x Time|1,Temp,Time",
+ "Error",
+ "C. Total")
```

```
> data.frame(Source,SumOfSquares)
                   Source SumOfSquares
1                  Temp|1    170.01667
2             Time|1,Temp     11.60333
3 Temp x Time|1,Temp,Time      0.48000
4                   Error     12.00000
5               C. Total     194.10000
>
> anova(lm(y~temp+time+temp:time,data=d))

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1 170.017 170.017 85.0083 9.185e-05 ***
time       1  11.603  11.603  5.8017   0.05267 .
temp:time  1   0.480   0.480  0.2400   0.64160
Residuals  6  12.000   2.000
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Temp|1 ANOVA Test

```
> x=x4
> fractions((p2-p1)%*%x)
    b20:a3 b30:a3 b20:a6 b30:a6
1     2/15  -3/10   4/15  -1/10
2     2/15  -3/10   4/15  -1/10
3     -1/5   9/20   -2/5   3/20
4     -1/5   9/20   -2/5   3/20
5     -1/5   9/20   -2/5   3/20
6     2/15  -3/10   4/15  -1/10
7     2/15  -3/10   4/15  -1/10
8     2/15  -3/10   4/15  -1/10
9     2/15  -3/10   4/15  -1/10
10    -1/5   9/20   -2/5   3/20

> fractions((30/12)*(p2-p1)%*%x)[1,]
b20:a3 b30:a3 b20:a6 b30:a6
   1/3   -3/4    2/3   -1/4
```

Temp|1 ANOVA Test

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

where

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

Time$|1$,Temp ANOVA Test

```
> fractions((p3-p2)%*%x)
   b20:a3  b30:a3  b20:a6  b30:a6
1    32/75    6/25  -32/75   -6/25
2    32/75    6/25  -32/75   -6/25
3     4/25   9/100   -4/25  -9/100
4     4/25   9/100   -4/25  -9/100
5     4/25   9/100   -4/25  -9/100
6   -16/75   -3/25   16/75    3/25
7   -16/75   -3/25   16/75    3/25
8   -16/75   -3/25   16/75    3/25
9   -16/75   -3/25   16/75    3/25
10  -12/25 -27/100   12/25  27/100

> fractions((3/2)*(p3-p2)%*%x)[1,]
b20:a3 b30:a3 b20:a6 b30:a6
 16/25   9/25 -16/25  -9/25
```

Time$|1$,Temp ANOVA Test

$$(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \iff \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0},$$

where

$$\begin{aligned} \boldsymbol{C}\boldsymbol{\beta} &= \begin{bmatrix} \frac{16}{25} & \frac{9}{25} & -\frac{16}{25} & -\frac{9}{25} \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} \\ &= \left(\frac{16}{25}\mu_{11} + \frac{9}{25}\mu_{12} \right) - \left(\frac{16}{25}\mu_{21} + \frac{9}{25}\mu_{22} \right). \end{aligned}$$

Temp$\times$Time$|1$,Temp,Time ANOVA Test

```
> fractions((p4-p3)%*%x)
   b20:a3 b30:a3 b20:a6 b30:a6
1    6/25  -6/25  -6/25   6/25
2    6/25  -6/25  -6/25   6/25
3   -4/25   4/25   4/25  -4/25
4   -4/25   4/25   4/25  -4/25
5   -4/25   4/25   4/25  -4/25
6   -3/25   3/25   3/25  -3/25
7   -3/25   3/25   3/25  -3/25
8   -3/25   3/25   3/25  -3/25
9   -3/25   3/25   3/25  -3/25
10  12/25 -12/25 -12/25  12/25

> fractions((25/6)*(p4-p3)%*%x)[1,]
b20:a3 b30:a3 b20:a6 b30:a6
     1     -1     -1      1
```

Temp$\times$Time$|1$,Temp,Time ANOVA Test

$$(\boldsymbol{P}_4 - \boldsymbol{P}_3)\boldsymbol{X\beta} = \boldsymbol{0} \iff \boldsymbol{C\beta} = \boldsymbol{0},$$

where

$$\begin{aligned} \boldsymbol{C\beta} &= [1 \ \ -1 \ \ -1 \ \ \ 1] \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} \\ &= \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}. \end{aligned}$$

```
> test=function(lmout,C,d=0){
+   b=coef(lmout)
+   V=vcov(lmout)
+   dfn=nrow(C)
+   dfd=lmout$df
+   Cb.d=C%*%b-d
+   Fstat=drop(t(Cb.d)%*%solve(C%*%V%*%t(C))%*%Cb.d/dfn)
+   pvalue=1-pf(Fstat,dfn,dfd)
+   list(Fstat=Fstat,pvalue=pvalue)
+ }
```

```
> o=lm(y~0+temp:time)
>
> #Test for time main effect
>
> C=matrix(c(
+ .5,.5,-.5,-.5
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 6

$pvalue
[1] 0.04982526
```

```
> #ANOVA Test for time|1
>
> C=matrix(c(
+ 2/5,3/5,-4/5,-1/5
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 2.45

$pvalue
[1] 0.1685623
```

```
> #ANOVA Test for time|1,temp
>
> C=matrix(c(
+ 16/25,9/25,-16/25,-9/25
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 5.801667

$pvalue
[1] 0.05266955
```

```
> #Test for temp main effect
>
> C=matrix(c(
+ .5,-.5,.5,-.5
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 86.64

$pvalue
[1] 8.704602e-05
```

```
> #ANOVA Test for temp|1
>
> C=matrix(c(
+ 1/3,-3/4,2/3,-1/4
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 85.00833

$pvalue
[1] 9.185462e-05
```

```
> #ANOVA Test for temp|1,time
>
> C=matrix(c(
+ 3/5,-3/5,2/5,-2/5
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 88.36

$pvalue
[1] 8.233372e-05
```

```
> #Test for interactions
>
> C=matrix(c(
+ 1,-1,-1,1
+ ),nrow=1,byrow=T)
>
> test(o,C)
$Fstat
[1] 0.24

$pvalue
[1] 0.6416021
```

Different Types of Sums of Squares

Source	Type I	Type II	Type III
A	$SS(A\|1)$	$SS(A\|1,B)$	$SS(A\|1,B,AB)$
B	$SS(B\|1,A)$	$SS(B\|1,A)$	$SS(B\|1,A,AB)$
AB	$SS(AB\|1,A,B)$	$SS(AB\|1,A,B)$	$SS(AB\|1,A,B)$
Error	SSE	SSE	SSE
C. Total	$SSTotal$	?	?

Different Types of Sums of Squares for Three Factors

Type I	Type II	Type III
$SS(A\|1)$	$SS(A\|1,B,C,BC)$	$SS(A\|1,B,C,AB,AC,BC,ABC)$
$SS(B\|1,A)$	$SS(B\|1,A,C,AC)$	$SS(B\|1,A,C,AB,AC,BC,ABC)$
$SS(C\|1,A,B)$	$SS(C\|1,A,B,AB)$	$SS(C\|1,A,B,AB,AC,BC,ABC)$
$SS(AB\|1,A,B,C)$	$SS(AB\|1,A,B,C,AC,BC)$	$SS(AB\|1,A,B,C,AC,BC,ABC)$
$SS(AC\|1,A,B,C,AB)$	$SS(AC\|1,A,B,C,AB,BC)$	$SS(AC\|1,A,B,C,AB,BC,ABC)$
$SS(BC\|1,A,B,C,AB,AC)$	$SS(BC\|1,A,B,C,AB,AC)$	$SS(BC\|1,A,B,C,AB,AC,ABC)$
$SS(ABC\|1,A,B,C,AB,AC,BC)$	$SS(ABC\|1,A,B,C,AB,AC,BC)$	$SS(ABC\|1,A,B,C,AB,AC,BC)$
SSE	SSE	SSE
$SSTotal$	?	?

Sums of Squares for Balanced Data

For balanced data, the three types of sums of squares are identical: Type I=Type II=Type III.

This equality in not obvious (at least to most normal humans), but it is true. We will not attempt to prove this in 510.

The ANOVA F tests in *the* ANOVA table can be used to test for factor main effects and interactions.

Sums of Squares for Unbalanced Data

For unbalanced data, the types of sums of squares differ.

Type I sums of squares always add to the total sum of squares, even when data are unbalanced.

Type II and III sums of squares do not add to anything special when data are unbalanced.

The ANOVA F tests in the Type III ANOVA table can be used to test for factor main effects and interactions.

Type I and II ANOVA F tests do not, in general, test for factor main effects or interactions (except for the F test for the highest-order interactions, which is the same for all three types).

SAS Code and Output

```
proc glm;
  class time temp;
  model y=time temp time*temp / ss1 ss2 ss3;
run;

The GLM Procedure

   Class Level Information

Class          Levels    Values
time                2    3 6
temp                2    20 30

Number of Observations Read          10
Number of Observations Used          10
```

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	182.1000000	60.7000000	30.35	0.0005
Error	6	12.0000000	2.0000000		
Corrected Total	9	194.1000000			

R-Square	Coeff Var	Root MSE	y Mean
0.938176	16.25533	1.414214	8.700000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
time	1	4.9000000	4.9000000	2.45	0.1686
temp	1	176.7200000	176.7200000	88.36	<.0001
time*temp	1	0.4800000	0.4800000	0.24	0.6416

Source	DF	Type II SS	Mean Square	F Value	Pr > F
time	1	11.6033333	11.6033333	5.80	0.0527
temp	1	176.7200000	176.7200000	88.36	<.0001
time*temp	1	0.4800000	0.4800000	0.24	0.6416

Source	DF	Type III SS	Mean Square	F Value	Pr > F
time	1	12.0000000	12.0000000	6.00	0.0498
temp	1	173.2800000	173.2800000	86.64	<.0001
time*temp	1	0.4800000	0.4800000	0.24	0.6416

Type IV Sums of Squares

In addition to computing Type I, II, and III sums of squares, SAS can compute Type IV sums of squares.

Type IV sums of squares are only relevant for factorial designs with missing cells.

When cells are missing, I recommend determining the linear combinations of the estimable cell means that are of scientific interest, and then conducting the corresponding tests as tests of $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$.

Calculation of Type I, II, and III Sums of Squares

Every Type I, II, or III sum of squares is the error sums of squares for a reduced model minus the error sum of squares for a model that adds one term to the reduced model:

$$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced}}})\boldsymbol{y} - \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced+term}}})\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{P}_{\boldsymbol{X}_{\text{reduced+term}}} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced}}})\boldsymbol{y},$$

where $\mathcal{C}(\boldsymbol{X}_{\text{reduced}}) \subset \mathcal{C}(\boldsymbol{X}_{\text{reduced+term}}) \subseteq \mathcal{C}(\boldsymbol{X})$.

As usual, $\boldsymbol{X}$ represents the model matrix for the most complex model under consideration (a.k.a., the full model).

For all Type III sums of squares, the reduced+term model is the full model.

For example, $SS(\text{temp}|1, \text{time})$ is

$$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced}}})\boldsymbol{y} - \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced+term}}})\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{P}_{\boldsymbol{X}_{\text{reduced+term}}} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced}}})\boldsymbol{y},$$

where we can choose

$$\boldsymbol{X}_{\text{reduced}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{X}_{\text{reduced+term}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

How does this apply to a Type *III* Sum of Squares like $SS(\text{time}|1, \text{temp}, \text{time} \times \text{temp})$?

Here the reduced+term model is actually the full cell means model that includes an intercept, time and temperature main effects, and time $\times$ temperature interaction.

The reduced model has an intercept, temperature main effect, and time $\times$ temperature interaction but no time main effect.

How do we fit such a reduced model?

Code like `lm(y~b+a:b)` or `model y=b a*b;` does not fit the reduced model with no `a` main effects.

```
> model.matrix(~temp+time:temp)
   (Intercept) temp30 temp20:time6 temp30:time6
1            1      0            0            0
2            1      0            0            0
3            1      1            0            0
4            1      1            0            0
5            1      1            0            0
6            1      0            1            0
7            1      0            1            0
8            1      0            1            0
9            1      0            1            0
10           1      1            0            1
```

This model matrix has rank 4 and the same column space as $\boldsymbol{X}$.

Removing Time Main Effect from Cell Means Model

Time	Temp 20°	30°	Marginal Mean
3 months	μ_{11}	μ_{12}	$\bar{\mu}_{1\cdot} = (\mu_{11} + \mu_{12})/2$
6 months	μ_{21}	μ_{22}	$\bar{\mu}_{2\cdot} = (\mu_{21} + \mu_{22})/2$

The parameters μ_{11}, μ_{12}, μ_{21}, and μ_{22} can be anything as long as

$$\bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} \iff \mu_{22} = \mu_{11} + \mu_{12} - \mu_{21}.$$

A Model Matrix for Model with $1, \text{temp}, \text{time} \times \text{temp}$

$$\boldsymbol{X}_{\text{reduced}}\boldsymbol{\beta}_{\text{reduced}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \end{bmatrix} = \begin{bmatrix} \mu_{11} \\ \mu_{11} \\ \mu_{12} \\ \mu_{12} \\ \mu_{12} \\ \mu_{21} \\ \mu_{21} \\ \mu_{21} \\ \mu_{21} \\ \mu_{11} + \mu_{12} - \mu_{21} \end{bmatrix}$$

```
#Type III Sum of Squares for SS(time|1,temp,time x temp)
> x0=x[,1:3]
> x0[10,]=c(1,1,-1)
> x0
1        1       0       0
2        1       0       0
3        0       1       0
4        0       1       0
5        0       1       0
6        0       0       1
7        0       0       1
8        0       0       1
9        0       0       1
10       1       1      -1
> anova(lm(y~0+x0),lm(y~0+x))
Analysis of Variance Table
Model 1: y ~ 0 + x0
Model 2: y ~ 0 + x
  Res.Df RSS Df Sum of Sq  F  Pr(>F)
1      7  24
2      6  12  1        12  6 0.04983 *
```

Alternative Computation of Sums of Squares

Let $SS = \boldsymbol{y}'(\boldsymbol{P}_{\boldsymbol{X}_{\text{reduced+term}}} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced}}})\boldsymbol{y}$ represent any Type I, II, or III sum of squares.

Let $q = \text{rank}(\boldsymbol{X}_{\text{reduced+term}}) - \text{rank}(\boldsymbol{X}_{\text{reduced}})$ be the degrees of freedom associated with SS.

Let $\boldsymbol{C}$ be any $q \times p$ matrix whose rows are a basis for the row space of $(\boldsymbol{P}_{\boldsymbol{X}_{\text{reduced+term}}} - \boldsymbol{P}_{\boldsymbol{X}_{\text{reduced}}})\boldsymbol{X}$.

Alternative Computation of Sums of Squares

Then the ANOVA F statistic

$$\frac{SS/q}{MSE} = \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}']^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}/q}{\hat{\sigma}^2}.$$

Thus, any SS can be computed as $\hat{\boldsymbol{\beta}}'\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}']^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}$ for an appropriate matrix $\boldsymbol{C}$.