STAT22200 Chapter 10

Yibi Huang

10.1 Unbalanced Data

What Happens If Factorial Data Become Unbalanced

With unbalanced data (but no empty cell), what are **changed**?

- no simple formulae for parameter estimates and SS.
- the parameter estimates and SS of a term will depend on the presence of other terms in the model, e.g., the estimates for α_i's might be different in the following 3 models

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$
$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$
$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$$

- need to rely on statistical software for computation
- there are 3 variations of SS

With unbalanced data (but no empty cell), what are **unchanged**?

- the means model and the main-effect-interaction model can still be used,
- for the means model (e.g., $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$), the estimate for μ_{ij} remains to be the sample group mean $\overline{y}_{ij\bullet}$ for that group
- one can still make interaction plots and using them the visualize the main effect and interactions
- one can check model assumptions as usual

Notation for Models

In the following, we denote various models by listing the included effect. For example,

- $(1, A, B, AB)$ denotes the model $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$
- $(1, A, B)$ denotes the model $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
- $(1, A, B, C, AB, AC)$ denotes the model

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \varepsilon_{ijkl}$$

Here the "$\mathbf{1}$" stands for the grand mean μ.

In the following SSE(model) denotes the SSE of that model, e.g., SSE$(1, A, B, AB)$ means the SSE of the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}.$$

For unbalanced data, there is no simple formula to compute the SSE. One must write the model as a regression model and use statistical software to compute the SSE.

Adjusted Sum of Squares (1)

The **adjusted sum of squares** for main effects B adjusted for A is defined as

$$SS(B|1, A) = \text{SSE}(1, A) - \text{SSE}(1, A, B).$$

- $SS(B|1, A) \geq 0$ since the model $(1, A)$ is nested in the model $(1, A, B)$ and hence the latter always has a smaller SSE
- $SS(B|1, A)$ is the reduction in SSE after B is included in the model
- $SS(B|1, A)$ describes the effect of B adjusted for A since with consider two models that A is present in both and the two models only differ by B

Adjusted Sum of Squares (2)

Likewise, the **adjusted sum of squares** for main effects B adjusted for A, C, and AC is

$$SS(B|1, A, C, AC) = \text{SSE}(1, A, C, AC) - \text{SSE}(1, A, B, C, AC).$$

In general, the **adjusted sum of squares** for a term adjusted for some other terms is

$$\begin{aligned} &SS(\text{a term}|\text{some other terms}) \\ = &\text{SSE}(\text{some other terms}) - \text{SSE}(\text{a term, some other terms}) \end{aligned}$$

For balanced data, the adjusted SS is identical to the unadjusted SS,

$$SS(A|1, B) = SS(A|1, B, C) = SS(A|1, B, C, BC) = SS(A|1).$$

Sequential Sum of Squares (aka. Type I SS)

For a specified model, the sequential SS for any term is adjusted for those terms that precede it in the model.

- E.g, the sequential SS's for the model (1, A, B, AB, C) are

Source	d.f.	Sequential SS
A	$a-1$	$SS(A\|1)$
B	$b-1$	$SS(B\|1, A)$
AB	$(a-1)(b-1)$	$SS(AB\|1, A, B)$
C	$c-1$	$SS(C\|1, A, B, AB)$

Sequential SS's depend on how the terms are **ordered** in a model:

- E.g, if the terms in the model (1, A, B, AB, C) is reshuffled as (1, C, A, B, AB), then the sequential SS's become

Source	d.f.	Sequential SS
C	$c-1$	$SS(C\|1)$
A	$a-1$	$SS(A\|1, C)$
B	$b-1$	$SS(B\|1, A, C)$
AB	$(a-1)(b-1)$	$SS(AB\|1, A, B, C)$

Sequential Sum of Squares and SSE Add Up to SST (1)

Source	Sequential SS				
A	$SS(A\|1)$	$=$	$\text{SSE}(1)$	$-$	$\text{SSE}(1, A)$
B	$SS(B\|1, A)$	$=$	$\text{SSE}(1, A)$	$-$	$\text{SSE}(1, A, B)$
AB	$SS(AB\|1, A, B)$	$=$	$\text{SSE}(1, A, B)$	$-$	$\text{SSE}(1, A, B, AB)$
C	$SS(C\|1, A, B, AB)$	$=$	$\text{SSE}(1, A, B, AB)$	$-$	$\text{SSE}(1, A, B, AB, C)$
Error	$\text{SSE}(1, A, B, AB, C)$				
Sum	$\text{SSE}(1) = \text{SST}$				

SSE(1) is the SSE for the model $y_{ijk\ell} = \mu + \varepsilon_{ijk\ell}$, of which the optimal (least square) estimate for μ is the overall mean $\bar{y}_{\bullet\bullet\bullet\bullet}$. Hence,

$$\text{SSE}(1) = \sum_{ijk\ell}(y_{ijk\ell} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 = \text{SST}.$$

Sequential Sum of Squares and SSE Add Up to SST (2)

If the order of terms in the model $(1, A, B, AB, C)$ is changed to $(1, C, A, B, AB)$,

- the sequential SS's are changed;
- $\text{SSE}(1, A, B, AB, C) = \text{SSE}(1, C, A, B, AB)$ is not affected by the order of terms;
- the sequential SS's and the SSE always add up to SST.

Source	Sequential SS
C	$SS(C\|1) = \text{SSE}(1) - \text{SSE}(1, C)$
A	$SS(A\|1, C) = \text{SSE}(1, C) - \text{SSE}(1, C, A)$
B	$SS(B\|1, C, A) = \text{SSE}(1, C, A) - \text{SSE}(1, C, A, B)$
AB	$SS(AB\|1, A, B, C) = \text{SSE}(1, C, A, B) - \text{SSE}(1, A, B, C, AB)$
Error	$\text{SSE}(1, C, A, B, AB)$
Sum	$\text{SSE}(1) = \text{SST}$

Example 8.10 Amylase data Revisit

$8 \times 2 \times 2$ design with 3 replicates $\Rightarrow$ Balanced Data!

		Analysis Temperature							
GT	Var.	40	35	30	25	20	15	13	10
25	B73	391.8	427.7	486.6	469.2	383.1	338.9	283.7	269.3
		311.8	388.1	426.6	436.8	408.8	355.5	309.4	278.7
		367.4	468.1	499.8	444.0	429.0	304.5	309.9	313.0
	O43	301.3	352.9	376.3	373.6	377.5	308.8	234.3	197.1
		271.4	296.4	393.0	364.8	364.3	279.0	255.4	198.3
		300.3	346.7	334.7	386.6	329.2	261.3	239.4	216.7
13	B73	292.7	422.6	443.5	438.5	350.6	305.9	319.9	286.7
		283.3	359.5	431.2	398.9	383.9	342.8	283.2	266.5
		348.1	381.9	388.3	413.7	408.4	332.2	287.9	259.8
	O43	269.7	380.9	389.4	400.3	340.5	288.6	260.9	221.9
		284.0	357.1	420.2	412.8	309.5	271.8	253.6	254.4
		235.3	339.0	453.4	371.9	313.0	333.7	289.5	246.7

```
read.table("amylaze.txt", h=T)
amyl$at = as.factor(amyl$atemp)
amyl$gt = as.factor(amyl$gtemp)
amyl$v  = as.factor(amyl$variety)
```

```
> anova(lm(log(y) ~  at+gt+v+gt:v, data = amyl))

          Df  Sum Sq Mean Sq F value    Pr(>F)
at         7 3.01613 0.43088 72.3869 < 2.2e-16 ***
gt         1 0.00438 0.00438  0.7358 0.3934347
v          1 0.58957 0.58957 99.0475 6.413e-16 ***
gt:v       1 0.08599 0.08599 14.4468 0.0002704 ***
Residuals 85 0.50595 0.00595
```

```
> anova(lm(log(y) ~  gt+v+at+gt:v, data = amyl))

          Df  Sum Sq Mean Sq F value    Pr(>F)
gt         1 0.00438 0.00438  0.7358 0.3934347
v          1 0.58957 0.58957 99.0475 6.413e-16 ***
at         7 3.01613 0.43088 72.3869 < 2.2e-16 ***
gt:v       1 0.08599 0.08599 14.4468 0.0002704 ***
Residuals 85 0.50595 0.00595
```

For balanced data, SS's are not affected by the order of the terms.

If One Observation Is Missing...

		Analysis Temperature							
GT	Var.	40	35	30	25	20	15	13	10
25	B73	391.8	427.7	486.6	469.2	383.1	338.9	283.7	269.3
		311.8	388.1	426.6	436.8	408.8	355.5	309.4	278.7
		367.4	468.1	499.8	444.0	429.0	304.5	309.9	313.0
	O43	301.3	352.9	376.3	373.6	377.5	308.8	234.3	197.1
		271.4	296.4	393.0	364.8	364.3	279.0	255.4	198.3
		300.3	346.7	334.7	386.6	329.2	261.3	239.4	216.7
13	B73	292.7	422.6	443.5	438.5	350.6	305.9	319.9	286.7
		283.3	359.5	431.2	398.9	383.9	342.8	283.2	266.5
		348.1	381.9	388.3	413.7	408.4	332.2	287.9	259.8
	O43	269.7	380.9	389.4	400.3	340.5	288.6	260.9	221.9
		284.0	357.1	420.2	412.8	309.5	271.8	253.6	254.4
		235.3	339.0	453.4	371.9	313.0	333.7	289.5	246.7

Suppose the first observation (the number 391.8 in blue box) is missing. The data become unbalanced.

When the first observation is removed `amyl[-1,]`, the data become unbalanced, and R produces the ANOVA table using the sequential SS.

```
> anova(lm(log(y) ~  at+gt+v+gt:v, data = amyl[-1,]))

          Df  Sum Sq Mean Sq F value    Pr(>F)
at         7 3.06282 0.43755 76.2546 < 2.2e-16 ***
gt         1 0.00140 0.00140  0.2433 0.6231292
v          1 0.55282 0.55282 96.3443 1.361e-15 ***
gt:v       1 0.07554 0.07554 13.1647 0.0004888 ***
Residuals 84 0.48199 0.00574

> anova(lm(log(y) ~  gt+v+at+gt:v, data = amyl[-1,]))

          Df  Sum Sq Mean Sq F value    Pr(>F)
gt         1 0.00244 0.00244   0.426 0.5157603
v          1 0.57061 0.57061  99.444 6.605e-16 ***
at         7 3.04399 0.43486  75.786 < 2.2e-16 ***
gt:v       1 0.07554 0.07554  13.165 0.0004888 ***
Residuals 84 0.48199 0.00574
```

For unbalanced data, the sequential SS's *changes* with the order of terms in the model.

```
> anova(lm(log(y) ~ gt+v+at+gt:v, data = amyl[-1,]))
          Df  Sum Sq Mean Sq F value    Pr(>F)
gt         1 0.00244 0.00244   0.426 0.5157603
v          1 0.57061 0.57061  99.444 6.605e-16 ***
at         7 3.04399 0.43486  75.786 < 2.2e-16 ***
gt:v       1 0.07554 0.07554  13.165 0.0004888 ***
Residuals 84 0.48199 0.00574
```

How is the sequential SS for `v` in the table above computed?

```
> lm1 = lm(log(y) ~ gt, data = amyl[-1,])
> lm2 = lm(log(y) ~ gt+v, data = amyl[-1,])
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: log(y) ~ gt
Model 2: log(y) ~ gt + v
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     93 4.1721
2     92 3.6015  1   0.57061 14.576 0.000244 ***
```

When the `gt:v` interaction is placed before `at` main effects in the model formula, R still includes `at` before `gt:v`.

```
> anova(lm(log(y) ~ gt+v+gt:v+at, data = amyl[-1,]))
          Df  Sum Sq Mean Sq F value    Pr(>F)
gt         1 0.00244 0.00244   0.426 0.5157603
v          1 0.57061 0.57061  99.444 6.605e-16 ***
at         7 3.04399 0.43486  75.786 < 2.2e-16 ***
gt:v       1 0.07554 0.07554  13.165 0.0004888 ***
Residuals 84 0.48199 0.00574
```

R always includes main effects before two-way interactions, and lower-order interactions before higher-order interactions, regardless how they are ordered in the model formula.

Type I ANOVA table

The Type I ANOVA table for unbalanced data are identical to the ANOVA table for balanced data in every aspect except the SSs are replaced by the sequential SS.

Source	d.f.	Sequential SS	MS	F-value
A	$a-1$	$SS(A\|1)$	SS_A/df_A	MS_A/MSE
B	$b-1$	$SS(B\|1,A)$	SS_B/df_B	MS_B/MSE
C	$c-1$	$SS(C\|1,A,B)$	SS_C/df_C	MS_C/MSE
AB	$(a-1)(b-1)$	$SS(AB\|1,A,B,C)$	SS_{AB}/df_{AB}	MS_{AB}/MSE
AC	$(a-1)(c-1)$	$SS(AC\|1,A,B,C,AB)$	SS_{AC}/df_{AC}	MS_{AC}/MSE
BC	$(a-1)(c-1)$	$SS(BC\|1,A,B,C,AB,AC)$	SS_{BC}/df_{BC}	MS_{BC}/MSE
ABC	$(a-1)(b-1)(c-1)$	$SS(ABC\|1,A,B,C,AB,AC,BC)$	SS_{ABC}/df_{ABC}	MS_{ABC}/MSE
Error	$N-abc$	SSE	SSE/df_E	
Total	$N-1$	SST		

Sequential SS's and the SSE always add up to SST.

Why Sequential SS's Are Not Ideal?

Look the 3 F-statistic for the 3 main effects in the previous page.

- The F-statistic for A is unadjusted
- The F-statistic for B is adjusted with A
- The F-statistic for C is adjusted with both A and B

When considering whether a term, say A, is needed in a model, one should look at the *net effect* of A after adjusting for the effect of other terms.

What are the terms that should be accounted for before considering A?

$$1, B, C, BC.$$

Why not adjusting for AB, AC and ABC?

Thus, a more sensible adjusted SS for A is $SS(A|1, B, C, BC)$.

Such adjusted SS's are called the Type II Sum of Squares.

Type II Sum of Squares (Yates' Fitting Constant)

The Type II SS_U of an effect U (U can be a main effect or an interaction) is computed as follows:

- take the biggest hierarchical model without effect U, and then compare it to the model with U added.

Here"biggest hierarchical model" means all the effects that don't include term U. E.g., for the model (1, A, B, C, AB, AC, BC, ABC),

- the Type II SS for AB is $SS(AB|1, A, B, C, AC, BC)$
- the Type II SS for C is $SS(C|1, A, B, AB)$ but not $SS(C|1, A)$ or $SS(C|1, A, AB)$

Unlike Type I SS, Type II SS does NOT depend on the *order* of terms in a model

Type II 3-Way ANOVA table

Source	d.f.	Type II SS	MS	F-value
A	$a-1$	$SS(A\|1,B,C,BC)$	SS_A/df_A	MS_A/MSE
B	$b-1$	$SS(B\|1,A,C,AC)$	SS_B/df_B	MS_B/MSE
C	$c-1$	$SS(C\|1,A,B,AB)$	SS_C/df_C	MS_C/MSE
AB	$(a-1)(b-1)$	$SS(AB\|1,A,B,C,AC,BC)$	SS_{AB}/df_{AB}	MS_{AB}/MSE
AC	$(a-1)(c-1)$	$SS(AC\|1,A,B,C,AB,BC)$	SS_{AC}/df_{AC}	MS_{AC}/MSE
BC	$(a-1)(c-1)$	$SS(BC\|1,A,B,C,AB,AC)$	SS_{BC}/df_{BC}	MS_{BC}/MSE
ABC	$(a-1)(b-1)(c-1)$	$SS(ABC\|1,A,B,C,AB,AC,BC)$	SS_{ABC}/df_{ABC}	MS_{ABC}/MSE
Error	$N-abc$	SSE	SSE/df_E	

Type II SS of terms in a model will NOT sum to SST

Computing Type II ANOVA Table in R

The build-in function `anova()` in R gives Type I sums of squares only. To get the Type II SS's, first load the library `car` (which is the short for "Companion to Applied Regression"), and then use the function `Anova()` as follows.

```
> library(car)
> Anova(yourmodel, type=2)
```

Note the first letter `A` in `Anova()` is a capital letter A.

```
> lm2b = lm(log(y) ~ gt+v+at+gt:v, data = amyl[-1,])
> Anova(lm2b,type=2)
Anova Table (Type II tests)

Response: log(y)
           Sum Sq Df F value    Pr(>F)
gt        0.00211  1  0.3673 0.5461286
v         0.55282  1 96.3443 1.361e-15 ***
at        3.03751  7 75.6243 < 2.2e-16 ***
gt:v      0.07554  1 13.1647 0.0004888 ***
Residuals 0.48199 84
```

Exercise 10.3 (p.246)

Here are two sequential (Type I) ANOVA tables for the same data.

	DF	SS
r	3	3.3255
c	3	112.95
r.c	9	0.48787
Error	14	0.8223

	DF	SS
c	3	116.25
r	3	______
c.r	9	0.48787
Error	14	0.8223

Q1: Fill in the 3 blanks in the second ANOVA table.

- SSE doesn't depend on the order of terms $\Rightarrow$ SSE $= 0.8223$
- the Type I SS for interaction are both $SS(rc|1, r, c)$
  $\Rightarrow SS_{r.c} = SS_{c.r} = 0.48787$

Exercise 10.3 (p.246)

	DF	SS
r	3	3.3255
c	3	112.95
r.c	9	0.48787
Error	14	0.8223

	DF	SS
c	3	116.25
r	3	0.0255
c.r	9	0.48787
Error	14	0.8223

Type I SS's and SSE add up to SST in both tables, i.e.,

$$
\begin{aligned}
& SS(r) + SS(c|1,r) + \cancel{SS(r.c|1,r,c)} + \cancel{SSE} \\
= & SS(c) + SS(r|1,c) + \cancel{SS(c.r|1,r,c)} + \cancel{SSE}
\end{aligned}
$$

As the two tables have identical SSE and SS for the interactions, we know

$$SS(r) + SS(c|1,r) = SS(c) + SS(r|1,c).$$

So

$$
\begin{aligned}
SS(r|1,c) &= SS(r) + SS(c|1,r) - SS(c) \\
&= 3.3255 + 112.95 - 116.25 = 0.0255.
\end{aligned}
$$

Exercise 10.3 (p.246)

	DF	SS
r	3	3.3255
c	3	112.95
r.c	9	0.48787
Error	14	0.8223

	DF	SS
c	3	116.25
r	3	0.0255
c.r	9	0.48787
Error	14	0.8223

Q2: What do you conclude about the significance of row effects, column effects, and interactions?

Ans: One should determine using the **Type II SS**.

- The Type I SS for r.c. in the two tables is also the Type II SS = SS(r.c.| 1,r,c).

$$F = \frac{SS(r.c.|1, r, c)/df_{r.c.}}{MSE} = \frac{0.48787/9}{0.8223/14} \approx 0.9229 \sim F_{9,14}$$

The P-value is `pf(0.9229,9,14,lower.tail=F)` ≈ 0.534, not significant

Exercise 10.3 (p.246)

	DF	SS
r	3	3.3255
c	3	112.95
r.c	9	0.48787
Error	14	0.8223

	DF	SS
c	3	116.25
r	3	0.0255
c.r	9	0.48787
Error	14	0.8223

- The Type II SS for the row effect is SS(r | 1,c), which is the Type I SS for r in the right table.

$$F = \frac{SS(r|1,c)/df_r}{MSE} = \frac{0.0255/3}{0.8223/14} \approx 0.1447 \sim F_{3,14}$$

  The P-value is `pf(0.1447,3,14,lower.tail=F)` ≈ 0.93, not significant

- The Type II SS for the column effect is SS(c | 1,r), which is the Type I SS for c in the left table.

$$F = \frac{SS(c|1,r)/df_{c.}}{MSE} = \frac{112.95/3}{0.8223/14} \approx 641.0 \sim F_{3,14}$$

  The P-value is `pf(641.0,3,14,lower.tail=F)` $\approx 3.22 \times 10^{-15}$, highly significant

Exercise 10.1 on p. 245-245

Three ANOVA tables are given for the results of a single experiment. These tables give the Type I sums of squares. Construct a Type II ANOVA table.

	DF	SS
a	1	1.9242
b	2	1584.2
a.b	2	19.519
c	1	1476.7
a.c	1	17.527
b.c	2	191.84
a.b.c	2	28.567
Error	11	166.71

	DF	SS
b	2	1573
c	1	1428.7
b.c	2	153.62
a	1	39.777
b.a	2	69.132
c.a	1	27.51
b.c.a	2	28.567
Error	11	166.71

	DF	SS
c	1	1259.3
a	1	9.0198
c.a	1	0.93504
b	2	1776.1
c.b	2	169.92
a.b	2	76.449
c.a.b	2	28.567
Error	11	166.71

Type I SS's for the first table

	DF	SS	Type I SS	
a	1	1.9242	$SS(A\|1)$	
b	2	1584.2	$SS(B\|1, A)$	
a.b	2	19.519	$SS(AB\|1, A, B)$	
c	1	1476.7	$SS(C\|1, A, B, AB)$	also Type II
a.c	1	17.527	$SS(AC\|1, A, B, C, AB)$	
b.c	2	191.84	$SS(BC\|1, A, B, C, AB, AC)$	also Type II
a.b.c	2	28.567	$SS(ABC\|1, A, B, C, AB, AC, BC)$	also Type II

Type I SS's for the second table:

	DF	SS	Type I SS	
b	2	1573	$SS(B\|1)$	
c	1	1428.7	$SS(C\|1, B)$	
b.c	2	153.62	$SS(BC\|1, B, C)$	
a	1	39.777	$SS(A\|1, B, C, BC)$	also Type II
b.a	2	69.132	$SS(AB\|1, A, B, C, BC)$	
c.a	1	27.51	$SS(AC\|1, A, B, C, AB, BC)$	also Type II
b.c.a	2	28.567	$SS(ABC\|1, A, B, C, AB, AC, BC)$	also Type II

Type I SS's for the last table:

	DF	SS	Type I SS	
c	1	1259.3	$SS(C\|1)$	
a	1	9.0198	$SS(A\|1, C)$	
c.a	1	0.93504	$SS(AC\|1, A, C)$	
b	2	1776.1	$SS(B\|1, A, C, AC)$	also Type II
c.b	2	169.92	$SS(BC\|1, A, B, C, AC)$	
a.b	2	76.449	$SS(AB\|1, A, B, C, AC, BC)$	also Type II
c.a.b	2	28.567	$SS(ABC\|1, A, B, C, AB, AC, BC)$	also Type II

Collecting the Type II SS's in the 3 tables above one can construct the Type II ANOVA table.

	DF	SS	Formula	F-value	P-value
a	1	39.777	$SS(A\|1, B, C, BC)$	2.6247	0.13350
b	2	1776.1	$SS(B\|1, A, C, AC)$	58.597	1.36×10^{-06}
c	1	1476.7	$SS(C\|1, A, B, AB)$	97.440	8.41×10^{-07}
a.b	2	76.449	$SS(AB\|1, A, B, C, AC, BC)$	2.5222	0.12542
c.a	1	27.51	$SS(AC\|1, A, B, C, AB, BC)$	1.8152	0.20498
b.c	2	191.84	$SS(BC\|1, A, B, C, AB, AC)$	6.3293	0.01482
a.b.c	2	28.567	$SS(ABC\|1, A, B, C, AB, AC, BC)$	0.94253	0.41897
Error	11	166.71			

Only the main effects `b`, `c` and their interactions `b.c` are significant.