UNBALANCED DATA

Unbalanced data

'Unbalanced' means not all cells (combinations of treatment levels) have the same number of observations. This could happen by accident, or by design (e.g., if certain treatment combinations are extraordinarily expensive).

Consider hypothetical factors A and B, each with two levels. The unbalanced data set is available as `unbal2.txt` on Blackboard. To retrieve the balanced data set, use `bal2 <- rbind(unbal2, data.frame(y=0, A=1, B=2))`.

Balanced:

	B1	B2
A1	2, 3	0, 1
A2	4, 5	1, 2

Unbalanced:

	B1	B2
A1	2, 3	1
A2	4, 5	1, 2

"Unbalanced" means not all cells (combinations of treatment levels) have the same number of observations. This could happen by accident, or by design (e.g., if certain treatment combinations are extraordinarily expensive).

Consider hypothetical factors A and B, each with two levels. The unbalanced data set is available as unbal2.txt (on Canvas not Blackboard).

To retrieve the balanced data set, use the R command listed here.

Ambiguity about how to calculate marginal means.

Notice how, when the data are unbalanced, there is a fundamental ambiguity about how to calculate marginal means.

With balance data:

$$\overline{y}_{1\cdot\cdot} = (2 + 3 + 0 + 1)/4 = 1.5 = (\overline{y}_{11\cdot} + \overline{y}_{12\cdot})/2.$$

With unbalanced data:

$\overline{y}_{1\cdot\cdot} = (2 + 3 + 1)/2$, but $(\overline{y}_{11\cdot} + \overline{y}_{12\cdot})/2 = (2.5 + 1)/2 = 1.75.$

Notice how, when the data are unbalanced, there is a fundamental ambiguity about how to calculate marginal means:

With balance data, the row mean can be computed either as the avarage of all observations in the row or the average of two cell means in the row.

With unbalanced data, the avarage of all observations in the row will be different from the average of two cell means in the row.

We will have the same issue when computing the column means.

Two types of sums of squares

Related to this ambiguity, two types of sums of squares may be reported in the analysis of variance:

- *Type I sum of squares*: the sum of squares for a factor, after adjustment for all factors higher up in the table.
- *Type III sum of squares*: the sum of squares for a factor, after adjustment for all other factors in the table.

Related to this ambiguity, two types of sums of squares may be reported in the analysis of variance:

Type I sum of squares is the sum of squares for a factor, after adjustment for all factors higher up in the table.

Type III sum of squares is the sum of squares for a factor, after adjustment for all other factors in the table.

Example

As an example, for the “unbal2” data, consider models with the main effects of A and B, but no interaction term .

As an example, for the “unbal2” data, consider models with the main effects of A and B, but no interaction term.

Balanced data: Type I and Type III sums of sqaures are the same

. For balanced data, the Type I and Type III sums of squares are the same, and they don't depend on the order in which the factors are entered.

Source	SS	d.f.
A	4.5	1
B	12.5	1
Error	2.5	5
Total	19.5	7

Source	SS	d.f.
B	12.5	1
A	4.5	1
Error	2.5	5
Total	19.5	7

Note also that the sums of squares are additive, or orthogonal: SSA + SSB + SSE = 4.5 + 12.5 + 2.5 = 19.5 = SST.

For balanced data, the Type I and Type III sums of squares are the same, and they don't depend on the order in which the factors are entered.

Note also that the sums of squares are additive, or orthogonal: SST = SSA + SSB + SSE.

For unbalanced data, the Type I sums of squares depend on the order in which the factors are entered

Source	Type I SS	d.f.
A	1.71	1
B	9.6	1
Error	2.4	4
Total	13.71	6

Source	Type I SS	d.f.
B	8.05	1
A	3.27	1
Error	2.4	4
Total	13.71	6

For unbalanced data, the Type I sums of squares depend on the order in which the factors are entered.

The Type III sum of squares are not additive

Also, for unbalanced data, the Type III sums of squares are not additive (orthogonal): $SSA + SSB + SSE = 3.27 + 9.6 + 2.4 = 15.27 \neq 13.71 = SST$.

Source	Type III SS	d.f.
A	3.27	1
B	9.60	1
Error	2.4	4
Total	13.71	6

Also, for unbalanced data, the Type III sums of squares are not additive (orthogonal): SSA + SSB + SSE does not equal to SST .

Recall that in the sketch of the prove for the decomposition of SST in the previous lecture, a key step is that the “cross terms” arising from squaring the right hand side terms sum to 0. This is not true for unbalanced data!

Hypothesis test

If the data are unbalanced, certain least-squares estimators have to be modified (see Kuehl, p. 212-213).

To test hypotheses, I recommend using extra-sum-of-squares F-tests based on SSE's from two nested regression models. (Type I and Type III SSE will be the same, since the error term is always the last term.)

That is, we formulate the null as hypothesis as the difference between a full model and a reduced model and use the F-test for nestted regression models to test the hypothesis.

In R, ANOVA tables show Type I sums of squares, but the t-tests in the regression output are based on Type III sums of squares.

If the data are unbalanced, certain least-squares estimators have to be modified.

To test hypotheses, I recommend using extra-sum-of-squares F-tests based on SSE's from two nested regression models. Note that SSE will be the same in both Type I and Type III sum of squares, since the error term is always the last term.

That is, we formulate the null as hypothesis as the difference between a full model and a reduced model and use the F-test for nestted regression models to test the hypothesis.

In R, ANOVA tables show Type I sums of squares, but the t-tests in the regression output are based on Type III sums of squares.

Example

Using the unbalanced data (unbal2), test H_0: no effect of factor B in a model already containing A.

Strategy:

Compare the full model with A and B to a reduced model with just A.

Let's see an example.

Using the unbalanced data (unbal2), suppose we want to test the null hypothesis of no effect of factor B in a model already containing A.

To test this null hypothesis, we can compare the full model with A and B to a reduced model with just A.

Example (continued)

```
> anova(lm(y ~ A, data=unbal2))      # SSE(r) = 12.0, df(r) = 5
> anova(lm(y ~ A + B, data=unbal2))  # SSE(f) = 2.40, df(f) = 4
```

$$F^* = \frac{\text{SSE}_r - \text{SSE}_f}{df_r - df_f} \div \frac{\text{SSE}_f}{df_f} = \frac{12.0 - 2.4}{5 - 4} \div 2.4/4 = 16$$

$$P\text{-value} = \Pr(F_{1,4} > 16) = 0.0161 \ .$$

Using R, we fit the full model with A and B and the reduced model with only A to the unbalanced data.

The F-test for nested models give a p-value of 0.0161. There is strong evidence that factor B is significant.

Using regression output

Fitting a model with A and B for the unbalanced data set:

```
> summary(lm(y ~ A + B, data=unbal2))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8000     0.4899   5.715  0.00464
A2            1.4000     0.6000   2.333  0.07996
B2           -2.4000     0.6000  -4.000  0.01613
```

Note that the square of the t-statistic for $B2$, $(-4)^2 = 16$, is equal to the F-statistic from the above extra-sum-of-squares test. The P-values in the regression output are based on Type III sums of squares. □

If we use R function "summary" to summarize the regression output of the fitted full model.

We see that the square of the t-statistic for B2, (-4) squared equals 16, is equal to the F-statistic from the F-test on the previous slide. The p-values in the regression output are based on Type III sums of squares.

In fact, in this case, the t-test is equivalent to the F-test. But note that the t-test will only work when the null hypothesis corresonds to setting one regression coefficient to 0.

The F-test can test more general null hypotheses. For example, if the factor B has more than two levels, then we can no longer use a t-test to test the null hypothesis that there is no B effects in model already containt A.

Get Type III sums of squares in R

```
## Load the cars data
library(car);
## Loading required package: carData
## Fit a regression model
m = lm(dist ~ speed, data=cars);
## Summarize the results
Anova(m, type="III");
## Anova Table (Type III tests)
##
## Response: dist
##              Sum Sq Df F value    Pr(>F)
## (Intercept)  1600.3  1  6.7655  0.01232 *
## speed       21185.5  1 89.5671 1.49e-12 ***
## Residuals   11353.5 48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we ever need to compute Type III sums of squares in R, we can use the Anova function with a captial “A” from the add-on package “car”.

Summary

1. For unbalanced data, two types of sums of squares may be reported: Type I and Type III.
2. Type I and Type III sum of square are the same for balanced data, but different for unbalanced data.
3. For hypothesis test, we can use the F-test for nested models to avoid confussion or ambiguity.
4. The t-test from regression output is equivalent to an F-test using Type III sum of squares.
5. Study the R code and notes in “unbal2.html”.

For unbalanced data, two types of sums of squares may be reported: Type I and Type III.

Type I and Type III sum of square are the same for balanced data, but different for unbalanced data.

For hypothesis test, we can use the F-test for nested models to avoid confussion or ambiguity.

The t-test from regression output is equivalent to an F-test using Type III sum of squares.

Finaly, make sure to study the R code and notes in “unbal2.html”.