The University of Georgia

The General Linear Model and Unbalanced Designs

ERSH 8310

Keppel and Wickens Chapter 14

Today's Class

- The General Linear Model
- The Two-Factor Analysis
- Averaging of Groups and Individuals
- Contrasts and Other Analytical Analyses

THE GENERAL LINEAR MODEL

The General Linear Model

- The ideal for most research designs is to have an equal number of subjects in each group

- Unequal sample sizes are often the reality
  - Even in studies that were planned to have equal sample sizes

- The use of classification factors whose levels can only be determined after the sampling has occurred almost always leads to unequal groups
  - e.g., gender, college major, race, occupation, etc…

The General Linear Model

- To analyze experiments with unequal samples, the statistical procedures of the analysis of variance must be given in a more general form

- Statisticians have observed that most varieties of the analysis of variance can be expressed in a common way and that this representation also includes related techniques such as multiple regression

- This approach is known as the general linear model

A Numerical Example

- The one-way data presented in Section 3.5 will be re-analyzed (see Table 14.1)

Table 14.1: Data from three groups (from Table 3.7, p. 57) with descriptions fitted under the null hypothesis (first block of three columns) and its alternative (second block).

		Null hypothesis true			Null hypothesis false		
Group	Y_{ij}	$\bar{Y}_T$	$Y_{ij} - \bar{Y}_T$	$(Y_{ij} - \bar{Y}_T)^2$	$\bar{Y}_j$	$Y_{ij} - \bar{Y}_j$	$(Y_{ij} - \bar{Y}_j)^2$
	5	7.77	−2.77	7.67	7.00	−2.00	4.00
a_1	9	7.77	1.23	1.51	7.00	2.00	4.00
	7	7.77	−0.77	0.59	7.00	0.00	0.00
	12	7.77	4.23	17.90	10.00	2.00	4.00
	10	7.77	2.23	4.97	10.00	0.00	0.00
a_2	10	7.77	2.23	4.97	10.00	0.00	0.00
	8	7.77	0.23	0.05	10.00	−2.00	4.00
	11	7.77	3.23	10.43	10.00	1.00	1.00
	9	7.77	1.23	1.51	10.00	−1.00	1.00
	3	7.77	−4.77	22.75	5.00	−2.00	4.00
a_3	6	7.77	−1.77	3.13	5.00	1.00	1.00
	5	7.77	−2.77	7.67	5.00	0.00	0.00
	6	7.77	−1.77	3.13	5.00	1.00	1.00
Sum of squared deviations:				86.28			24.00

A Numerical Example

- When the null hypothesis is true, the unexplained sum of squares is:

$$SS_{unexp}^{H_0} = \sum \left(Y_{ij} - \bar{Y}_T \right)^2 = 86.28$$

- When the alternative hypothesis is true, the unexplained sum of squares is

$$SS_{unexp}^{H_1} = \sum \left(Y_{ij} - \bar{Y}_j \right)^2 = 24.00$$

Variability

- The variability produced by the treatment effects can be measured by taking the difference between the two unexplained sums of squares:

$$
\begin{aligned}
SS_A &= SS_{unexp}^{H_0} - SS_{unexp}^{H_1} \\
&= (\text{treatment effects} + \text{error}) - (\text{error}) \\
&= 86.28 - 24.00 = 62.28
\end{aligned}
$$

- Don’t trust me? How about we confirm this with SPSS?

So…there's an error in the book…

Tests of Between-Subjects Effects

Dependent Variable:DV

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	62.308[a]	2	31.154	12.981	.002
Intercept	645.333	1	645.333	268.889	.000
IV	62.308	2	31.154	12.981	.002
Error	24.000	10	2.400		
Total	871.000	13			
Corrected Total	86.308	12			

a. R Squared = .722 (Adjusted R Squared = .666)

SS_A

$SS_{unexp}^{H_1}$

$SS_{unexp}^{H_0}$

Degrees of Freedom

- The degrees of freedom are equal to the different values used in the predictions (e.g., one for the grand mean μ_T in the null hypothesis, and three for the a = 3 group means μ_j in the alternative hypothesis)

$$df_{unexp}^{H_0} = 13 - 1 = 12 \qquad df_{unexp}^{H_1} = 13 - 3 = 10$$

- The degrees of freedom for the effect equal the difference between the two values:

$$df_A = df_{unexp}^{H_0} - df_{unexp}^{H_1} = 12 - 10 = 2$$

F-Ratio

- We can obtain the F ratio using:

$$F = \frac{MS_A}{MS_{error}}$$

- where
  - $MS_A = SS_A / df_A$
  - $MS_{error} = SS_{error} / df_{error}$

- All of this is as it was before…

The Linear Model

- For the one-way design, the linear model is:

$$Y_{ij} = \mu_T + \alpha_j + E_{ij}$$

  - This is the alternative-hypothesis model

- The null-hypothesis model has all $\alpha_j = 0$

$$Y_{ij} = \mu_T + E_{ij}$$

- The two models form a hierarchical pair, in which the null-hypothesis model is a special case of the other

Testing by Model Comparison

- The general linear model uses a hierarchical pair of linear models that differ in whether the effect to be tested is included or not
- The sum of squared deviations has the form:

$$SS_{un\exp} = \sum_{ij}\left[(\text{data}) - (\text{fitted value})\right]^2 = \sum_{ij}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

- In the one-way ANOVA:

$$\hat{Y}_{ij} = \overline{Y}_T \ \ for \quad H_0$$

$$\hat{Y}_{ij} = \overline{Y}_j \ \ for \quad H_1$$

Testing by Model Comparison

- The difference between two sums of squares is the sum of squares for the test of a null hypothesis,

$$SS_{effect} = SS_{unexp}^{H_0} - SS_{unexp}^{H_1}$$

- The degrees of freedom for the unexplained variability are calculated via

$$df_{unexp} = (\text{observations}) - (\text{parameters})$$

Testing by Model Comparison

- For the data in Table 14.1, the null-hypothesis model has only one parameter and the alternative-hypothesis model has three parameters (i.e., μ_T and two α_j due to the constraint that $\Sigma_j a_j = 0$)

- The degrees of freedom for an effect are calculated by:

$$df_{effect} = df_{unexp}^{H_0} - df_{unexp}^{H_1}$$

THE TWO-FACTOR ANALYSIS

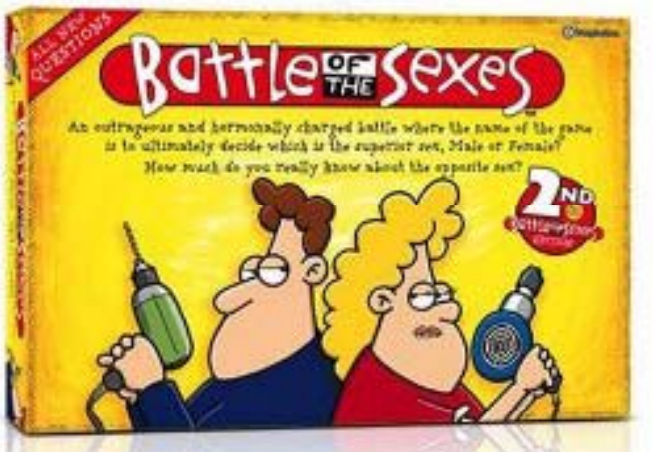
The Two-Factor Analysis

- Table 14.2 contains memory retention scores
  - Two Factors (A and B)

- There were three treatment levels for factor A:
  - Subjects in the first session received a_1
  - Subjects in the second session received a_2
  - Subjects in the third session received a_3

- There were two levels for factor B
  - Men (b_1)
  - Women (b_2)

It’s all about memory…and…



The General Linear Model in a Factorial Design

- The standard two-factor linear model is:

$$Y_{ij} = \mu_T + \alpha_j + \beta_k + (\alpha\beta)_{jk} + E_{ijk}$$

- It is sometimes called the full linear model or the general alternative-hypothesis model (i.e., H_1)

- We may create the three null-hypothesis models by deleting different parameters from this model
  - One for each type of effect in the two-way ANOVA
    - Main effect for Factor A
    - Main effect for Factor B
    - Interaction effect

Two-Factor Null Hypothesis Models

- H_0 for Main effect for Factor A

$$Y_{ij} = \mu_T + \beta_k + (\alpha\beta)_{jk} + E_{ijk}$$

- H_0 for Main effect for Factor B

$$Y_{ij} = \mu_T + \alpha_j + (\alpha\beta)_{jk} + E_{ijk}$$

- H_0 for Interaction effect

$$Y_{ij} = \mu_T + \alpha_j + \beta_k + E_{ijk}$$

Let's Do It Now…

- See the Excel SpreadSheet on ELC for the calculations

$$SS_{unexp}^{H_0(A)} = 952.990$$

$$SS_{unexp}^{H_0(B)} = 372.190$$

$$SS_{unexp}^{H_0(AxB)} = 410.990$$

$$SS_{unexp}^{H_1} = 353.295$$

Constructing the SS part of the ANOVA Table

- Remember: $SS_{effect} = SS_{unexp}^{H_0} - SS_{unexp}^{H_1}$

- SO…

$$SS_A = SS_{unexp}^{H_0(A)} - SS_{unexp}^{H_1} = 952.990 - 353.295 = 599.695$$

$$SS_B = SS_{unexp}^{H_0(B)} - SS_{unexp}^{H_1} = 372.190 - 353.295 = 18.895$$

$$SS_{AxB} = SS_{unexp}^{H_0(AxB)} - SS_{unexp}^{H_1} = 410.990 - 353.295 = 57.694$$

$$SS_{error} = SS_{unexp}^{H_1} = 353.295$$

Sums of Squares

- The sums of squares we just constructed are the so-called Type III sums of squares.

- The corresponding degrees of freedoms are (a-1), (b-1), and (a-1)(b-1)
  - Just like the balanced two-way ANOVA we learned the past two weeks

- SO…
  - $df_A = 2$
  - $df_B = 1$
  - $df_{AxB} = 2$
  - $df_T = 37 - 1 = 36$
  - $df_{error} = 36 - 2 - 1 - 2 = 31$

Putting Together the ANOVA Table

Source	SS	df	MS	F
A	599.695	2	299.848	26.310
B	18.895	1	18.895	1.658
AxB	57.694	2	28.847	2.531
Error	353.295	31	11.397	
Total	????			

No Differences Model

- Note that the model of no differences among all treatment cells is:

$$Y_{ij} = \mu_T + E_{ijk}$$

- And its sum of squares is:

$$SS_{total} = \sum_{ijk} \left(Y_{ijk} - \overline{Y}_T \right)^2$$

- With the degrees of freedom N-1
  - N is the total number of observations

$$SS_{unexp}^{H_0(ALL)} = 1178.271$$

Nonorthogonality of the Effects

- The unequal group sizes cause the A, B, and A ×B effects to be partially confounded with each other

- This nonorthogonality is an intrinsic characteristic of unbalanced factorial designs, and it is why they are also called nonorthogonal or unbalanced designs

- The additive sums of squares can be obtained using the Type I sums of squares (i.e., hierarchical sums of squares)
  - Not usually used

Putting Together the ANOVA Table

Source	SS	df	MS	F
A	599.695	2	299.848	26.310
B	18.895	1	18.895	1.658
AxB	57.694	2	28.847	2.531
Error	353.295	31	11.397	
Total	1178.271	36		

$$SS_{Total} \neq SS_A + SS_B + SS_{AxB} + SS_{Error}$$

$$1178.271 \neq 599.695 + 18.895 + 57.694 + 353.295$$

From SPSS

Tests of Between-Subjects Effects

Dependent Variable:retention

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	824.975[a]	5	164.995	14.478	.000
Intercept	6740.110	1	6740.110	591.413	.000
treatment	599.695	2	299.848	26.310	.000
gender	18.895	1	18.895	1.658	.207
treatment * gender	57.694	2	28.847	2.531	.096
Error	353.295	31	11.397		
Total	9265.000	37			
Corrected Total	1178.270	36			

a. R Squared = .700 (Adjusted R Squared = .652)

AVERAGING OF GROUPS AND INDIVIDUALS

Averaging of Groups and Individuals

- The presence of unequal sample sizes presents a dilemma whenever we combine data from two or more groups
- The averages can be obtained either score-based (i.e., considering different sample sizes) or group-based computation
- The group-based averages are often identified as estimated means, adjusted means, or least-squares means
- These two sets of means have different interpretations

Averaging of Groups and Individuals

- The Type III and Type I sums of squares are identical when the sample sizes are equal

- Unless the conclusions depend on the size of the samples, the Type III sums of squares should be used with unequal sample sizes

- You should interpret them using the marginal averages based on the cell means (i.e., the estimated, adjusted, or least-squares means)

Introducing…Least Squares Means

- The means for Gender:

Report

retention

gender	Mean	N	Std. Deviation
Male	15.7143	14	4.30436
Female	14.2174	23	6.45936
Total	14.7838	37	5.72099

- The Least Squares means for Gender

3. gender

Dependent Variable:retention

			95% Confidence Interval	
gender	Mean	Std. Error	Lower Bound	Upper Bound
Male	15.089	.941	13.169	17.009
Female	13.571	.709	12.126	15.017

How Least Squares Means Work…

- Least Squares means are the model predicted means
  - Formed by the effects of the model

- From our model, we have: $\hat{\mu}_T = 14.330$

$$\hat{\beta}_1 = 0.759$$

$$\hat{\beta}_2 = -0.759$$

- The LS means are:

$$\hat{\mu}_{Male} = \hat{\mu}_T + \hat{\beta}_1 = 14.330 + 0.759 = 15.089$$

$$\hat{\mu}_{Female} = \hat{\mu}_T + \hat{\beta}_2 = 14.330 - 0.759 = 13.571$$

But Why⁉

- Least squares means are equivalent to regular means when the sample size is equal across groups

- They will become important entities next week
  - Analysis of Covariance introduces covariates to help partition variance in an ANOVA model
  - The covariates “adjust” the least squares means so that mean differences are more pronounced
    - Provided the covariate is related to the DV

- So relax…and enjoy the LS means

Sensitivity to Assumptions

- Unequal samples do increase the sensitivity of analyses to heterogeneity of variance

- From the book: Try to design your study so that the sizes of the groups are the same
  - And if it happens…you will have won the lottery…

Wrapping Up…

- Today's class, although theoretical, was an important one from a statistical standpoint

- Null and alternative nested models are the basis for most of linear regression
  - And unbalanced ANOVA models

- Further, LS means will become important next week

Up Next…

- GLM in lab
  - Blurring the line between ANOVA and regression
- Homework:
  - Posted tomorrow morning – due next week before class
- Next week:
  - Read Chapter 15 - ANCOVA