Chapter 5

Analysis of Variance
(Unbalanced Case)

Review: Balanced Case

In Chapter 4:

- Breaking up variation of data based on source
- Balanced data (or special structure like one-way model)
- Terms orthogonal
- Can uniquely decompose variation
- Could use **proc anova**
- Could compare effect means

Unbalanced Case

- Cell counts vary
- Still want to decompose variation
- Effects are not orthogonal
- Decomposition is not unique
- Consider other types of sums of squares
- Use **proc glm**
- Compare least squares means instead

Types of Sums of Squares

Notations:

- **SS (C|A B)** - additional contribution when C is added to model containing A and B
- **R(C|A B)** - increase in residual sum of squares when C is removed from model containing A, B, and C

Type I

- Sequential sum of squares
- Additional variation explained by the model when that term is added to terms already in

Type III

- Partial sums of squares
- Explained variation that term adds when all other terms are already included
- Explained variation we would lose if term is removed from full model

Type II

- Adjusts sum of squares by leaving out any terms containing the one of interest
- E.g. Type II sum of squares for A would not include interactions with A in the full model
- Makes sense logically
- Tougher for direct comparisons

Type IV

- Same as Type III when no cells are empty
- Accounts for emptiness of cells when cells are empty

Example: ozkids data

Continuous response:

- **days**: days absent

Categorical predictors:

- **origin**: Aboriginal or not
- **sex**: male or female
- **grade**: level in school
- **type**: type of learner

Exercise: Mean Tabulations

- Get mean and count tabulation for **days** within each **cell**
- Get mean and count cross-tabulation for **days** for **origin**, **sex**, **grade**, and **type**
- Issues with data?
- Any apparent differences in days absent by category?

proc glm

- For general linear model
- Allows us to do ANOVA with unbalanced data
- Also handles much broader class of models

Example: **origin** and **grade** Models

- Days absent with **origin** and **grade** predictors
- See Type I and Type III sums of squares
- Reverse order of terms (e.g. **grade** first and then **origin**)
- What stays the same, and what changes?
- Does adding an interaction make sense?
- Impact of reversing main effects on interaction model?

Exercise: Type III Analysis in Four-Way Main Effects Model

- Type III SS for the four-way main effects model
- Conclusions about main effects to keep in the model?
- Perform all four of the one-way analyses of variance
- Conclusions of these 4 models?
- Get Type I SS for each of the orderings of the terms we might want to keep
- Could we further reduce the main effects we would want to keep in the model?

Multiple Comparisons Revisited

- Estimates are from least squares means
- Can make comparisons of least squares means like we did for cell means
- Use the **lsmeans** statement
- Can test differences of least squares means for main effects and interactions
- Can still use **means** statement for group means if desired, but that gives equal weight to cells rather than observations

Exercise: Multiple Comparisons

- Fit model with previous main effects and all interactions between them.
- Which main effects and interactions kept?
- Do Tukey multiple comparison on the ls means for the main effects and interactions.
- Significantly different groups?
- Do Tukey multiple comparison on the means for the main effects and note differences.

Type I and Type III with All Interactions

- Analysis of variance on the model with all 4 main effects and all interactions
- Order main effects based on one-way p-values (smallest p-value first)
- Insights from the Type I sums of squares?
- Insights from the Type III sums of squares?