Chapter 5 Notes

Analysis of Variance with Unbalanced Data

In Chapter 4, we worked with a balanced data set (same number of observations in each cell). When analyzing a balanced data set, we can use **proc anova**. With balanced data, the effects are orthogonal and there is a unique partitioning of the sums of squares. So for instance, the same amount of variation will be attributed to the **diet** effect in the blood pressure model from chapter 4 regardless of whether the **drug** and **biofeed**back are present or not.

When the data are not balanced, we no longer have orthogonality (in general) and we can no longer uniquely partition the sum of squares. Since we will lose uniqueness, we will need to talk about how we might meaningfully partition the variation. We will also need to discuss a new procedure (**proc glm**) for performing the analysis.

Types of Sums of Squares

Let's introduce some notation.

SS (A|B C) means the additional contribution to the model sum of squares when term **A** is added to the model containing terms **B** and **C**.

R(A|B C) is the reduction in the residual sum of squares when factor **A** is removed from the model containing factors **A**, **B** and **C**.

These are really the same thing. The **SS** notation tells us the additional amount of variation described, while the **R** notation tells us the reduction in residual sum of squares.

There are 4 types of sums of squares. The two most important types (at least for now) are **type I** (sequential sums of squares) and **type III** (partial sums of squares).

Type I: Sum of squares for a term is the amount of additional variation explained by the model when that term is added to the model.

Type III: Sum of squares for a term is the amount of variation that term adds to the model when all other terms are included, or, equivalently, the amount of explained variation lost if we drop that term from the model

Type II and **Type IV** are less common for ANOVA models in practice.

Type II adjusts the sum of squares by leaving out any terms containing the one of interest (so, for instance, the Type II sum of squares for a factor **A** would not include interactions with **A** when computing the sums of squares).

Type IV is like Type III, but takes into account emptiness of empty cells when there are empty cells.

The GLM Procedure

For unbalanced data, we will want to use **proc glm** rather than **proc anova**. Here **glm** stands for **general linear model** (Sometimes glm means *generalized* linear model. We will see generalized linear models later and use **proc genmod** for them.). The **glm** procedure can be used for far more complicated models than the analysis of variance models we are currently considering.

The most important statements for now will be the **class**, **model**, and **lsmeans** statements. This is similar to **proc anova**, but we want to work with the least squares means rather than the sample means. Least squares means are the appropriate estimate if we want to give each observation equal weight in the estimation, and this is consistent with an assumption of iid normal errors. Cell means would give each cell equal weight in the estimation, so observations in cells that have fewer observations would have a greater impact on the estimation, and that is typically not what we want.

Examples and Exercises

The Data Set

In this chapter we will be working with the **ozkids** data set from Quine. The data set contains a continuous response:

- **days**: days absent

and 4 categorical predictors:

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

We will ultimately want to look at the impact of these categorical factors on the number of days students miss from school.

Cross-Tabulation of Means and Counts

Let's start with some cross-tabulations of the data (like we did for the hypertension data in chapter 4).

- Tabulate means and counts for **days** absent categorized by **cell**.
- Tabulate means and counts for **days** absent by the products of the 4 predictor variables.
- Do we notice anything interesting/relevant in the tables?

Some Two-Way Model Examples

Let's consider models that only contain the effects of **origin** and **grade**.

- Obtain the Type I and Type III sums of squares for the main effects model with **origin** and **grade** (in that order).
- What do we notice about the sums of squares?

- Now reverse the order of the terms (e.g. **grade** first and **origin** second). What changes do we notice?
- If grade and origin are both significant, add an interaction between them to the model and interpret the results.
- Will the impact of the interaction term change if we reverse the order of the main effects?

Determining Terms to Keep in the Model

The authors of our textbook mention that there is debate over whether it is better to use Type III, or to use Type I sums of squares and change the order in which the categorical predictors enter the model. The authors prefer the re-ordered Type I approach (and I tend to agree with them). The Type III sums of squares tells us how much explained variation we lose by removing the term, but other terms still in the model may be insignificant making interpretation difficult in some cases.

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

- Obtain the Type III sums of squares for the four-way main effects model.
- What would we conclude about which main effects to keep in the model?
- Perform all four of the one-way analyses of variance.
- What would the sums of squares for these one-way models suggest as main effects we might want to include in the model?
- Obtain the Type I sums of squares for each of the orderings of the terms we would want to keep.
- Based on these results could we further reduce the main effects we would want to keep in the model?

Multiple Comparisons for Least Squares Means

In the previous exercise we chose main effects to use in a model.

- Fit the model with those main effects and all interactions between them to determine which model would be best in this case.
- Perform the Tukey multiple comparison on the least squares means (we'll need the **lsmeans** statement) for the main effects and interactions to determine significantly different groups.
- Perform the Tukey multiple comparison on the means for the main effects and note differences.

Type I and Type III Sums of Squares with All Interactions

For this model we will include the main effects in the order from most significant to least significant one-way ANOVA p-values.

- Perform the analysis of variance on the model with all 4 main effects and all interactions.
- What useful insights can we gain from the Type I sums of squares?
- What insights can we gain from the Type III sums of squares?