0.1 Addressing Unbalanced Data for Factorial Treatment Designs

Thus far we have assumed that we had balanced factorial data, meaning that there were the same number of replicates for each combination of treatment factors. When we have unbalanced factorial data, some issues arise as to how to conduct tests of hypotheses. Recall in an earlier lecture that we discussed the concept of defining a full and a reduced model as a general way to conduct tests with linear models. With unbalanced factorial data, our sum of squares for a factor can depend upon what models are defined as the full and reduced models for the test. We will use the notation that $SS(A|B)$ is the model sum of squares to test the effect of A while 'adjusting' for effect B. Thus $SS(A|B) = SS(A, B) - SS(B)$, where $SS(A, B)$ is the model sum of squares for the model with both the A and B effects. For a two-factor model with interaction, we will use AB to denote the interaction effect. With this notation we can explain the difference between what are called Type I, II, and III sums of squares. Suppose the model is written as: y = A B AB, then the Type I SS are formed as:

$$\begin{aligned} SS(A) &= SS(A|1) \\ SS(B|A) &= SS(A, B) - SS(A) \\ SS(AB|A, B) &= SS(A, B, AB) - SS(A, B) \end{aligned}$$

The Type II SS are formed as:

$$\begin{aligned} SS(A|B) &= SS(A, B) - SS(B) \\ SS(B|A) &= SS(A, B) - SS(A) \\ SS(AB|A, B) &= SS(A, B, AB) - SS(A, B) \end{aligned}$$

The Type III SS are formed as:

$$\begin{aligned} SS(A|B, AB) &= SS(A, B, AB) - SS(B, AB) \\ SS(B|A, AB) &= SS(A, B, AB) - SS(A, AB) \\ SS(AB|A, B) &= SS(A, B, AB) - SS(A, B) \end{aligned}$$

The pattern that emerges is that Type I SS are sequential, Type III SS adjust for every term in the model, and Type II SS adjust for all terms at the same level or below the term being tested. There is an active debate about whether Type II SS or Type III SS are more appropriate for tests with unbalanced data, but most would agree that Type I SS are often not appropriate. Check the SAS and R code with this lecture to illustrate the computation of these sums of squares.

0.2 Standard errors and interval estimates for means

Recall that for a variable y with $Var(y) = \sigma^2$, we have for the sample mean taken with n observations, $Var(\overline{y}) = \sigma^2/n$. Also recall the expression on page 43 of the text for a generic confidence interval for a mean: unbiased estimate $\pm$ multiplier $\times$ (estimated) standard error of estimate. Since the standard error of a mean is the square root of its variance, once we know how to calculate variances we can calculate standard errors and confidence intervals.

Calculations for least squares means with unbalanced data become slightly more complicated. In the handwritten example from the notes for an unbalanced 2x2 factorial, the least squares mean of level 1 of factor A, $\overline{y}_{1.}$, is the average of the cell means,

$$\overline{y}_{1.} = \frac{\overline{y}_{11} + \overline{y}_{12}}{2} = \frac{5+9}{2} = 7$$

To calculate the (estimated) standard error of this least squares mean, we calculate its (estimated) variance and take a square root. Since the samples from different cells are independent, their variances add. Thus we have

$$\begin{aligned} Var(\overline{y}_{1.}) &= Var(\frac{\overline{y}_{11} + \overline{y}_{12}}{2}) = \frac{Var(\overline{y}_{11}) + Var(\overline{y}_{12})}{4} \\ &= \frac{(\sigma^2/n_{11}) + (\sigma^2/n_{12})}{4} = \frac{(\sigma^2/3) + (\sigma^2/1)}{4} \end{aligned}$$

We estimate σ^2 from the MSE of the ANOVA table, giving us $\widehat{\sigma}^2 = 6.5$. Thus our estimated standard error for $\overline{y}_{1.}$ is:

$$\begin{aligned} s.e.(\overline{y}_{1.}) &= \sqrt{\frac{(\widehat{\sigma}^2/3) + (\widehat{\sigma}^2/1)}{4}} = \sqrt{\frac{(6.5/3) + (6.5/1)}{4}} \\ &= \sqrt{\frac{2.17 + 6.5}{4}} = \sqrt{2.167} = 1.472 \end{aligned}$$

0.3 Power and Sample Size for Factorial Treatment Designs

Power and sample size for factorial models is discussed in section 10.3, along with a listing of the noncentrality parameter for the interaction test in a 2 factor model.