Analysis of Variance

One-way analysis of variance

Previously, we used a dummy variable model for one-way analysis of variance (here assuming 3 groups):

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i,$$

where $D_{i1} = 1$ for observations from group 1 (otherwise is 0), and $D_{i2} = 1$ for observations from group 2 (otherwise is 0). Note that in this model there are as many parameters $(\alpha, \gamma_1, \gamma_2)$ as groups. Now we will switch to a model for one-way ANOVA that is more generalizable to other designs:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where Y_{ij} is the ith observation in group j, μ is the grand mean, α_j is a group-specific effect, and ε_{ij} is the error for the ith observation in group j. Under the usual assumptions about the errors ε_{ij}, taking expected values we have $\mu_j = \mu + \alpha_j$, so that we have one more parameter than we have groups. It is possible to have an overparametrized model and use a generalized matrix inverse to analyze the data, but we will instead impose the constraint:

$$\sum_{j=1}^{m} \alpha_j = 0,$$

sometimes called a sum-to-zero constraint. One consequence of this constraint is that of the m parameters, only $m-1$ of them may freely vary, which we recognize by saying that there are $m-1$ degrees of freedom for the group effect. The sum-to-zero constraint produces the following solution for the parameters:

$$\mu = \frac{\sum \mu_j}{m} \equiv \mu_{\cdot}\,,\ \alpha_j = \mu_j - \mu_{\cdot}$$

We can create a multiple regression model to estimate these parameters by using deviation regressors:

$$S_j = \left\{ \begin{array}{cc} 1 & \text{for observations in group } j \\ -1 & \text{for observations in group } m \\ 0 & \text{for observations in all other groups} \end{array} \right\}.$$

We can either fit this multiple regression model and use the results as shown in Chapter 6 to obtain the ANOVA F test of $H_0 : \mu_1 = \mu_2 = ... = \mu_m$ (or equivalently $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$) or we can directly calculate the ANOVA table using the sums of squares formulas shown in the text.

Two-way analysis of variance

Given two different treatments that occur in a factorial arrangement (so that each level of factor 1 is present with each level of factor 2), a two-way analysis of variance can be used to analyze the data. The basic layout of the data is shown in the text on page 149, and the series of Figures 8.2-8.4 illustrate different kinds of effects, from no effects of either factor to various types of interaction between factors. The model for two-way ANOVA is:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk},$$

where Y_{ijk} is the ith observation in the jk combination of factors 1 and 2, μ is the grand mean, α_j and β_k are a group-specific (main) effects, γ_{jk} is an interaction term for the jk combination of factors 1 and 2, and ε_{ijk} is the error term. Using the notation that μ_{jk} is the mean of the combination of level j of factor 1 and level k of factor 2, we can specify three hypotheses of interest:

$H_0 : \mu_{1.} = \mu_{2.} = ... = \mu_{r.}$ (main effect of factor 1)
$H_0 : \mu_{.1} = \mu_{.2} = ... = \mu_{.c}$ (main effect of factor 2)
$H_0 : \mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'}$ for all j, j' and k, k' (interaction of factors 1 and 2).

One way to understand the meaning of the interaction null hypothesis is that all differences between levels j and j' of factor 1 remain the same regardless of the level of factor 2. Just as we saw for one-way ANOVA, the model above for two-way ANOVA is overparametrized. For a two-factor model where factor 1 has r levels and factor 2 has c levels, there are a total of rc distinct combination means, giving $rc - 1$ total degrees of freedom. Given usual sum-to-zero constraints on the main effect parameters for factors 1 and 2, there are a total of $rc-1-(r-1)-(c-1) = (r-1)(c-1)$ degrees of freedom left for the interaction parameters γ_{jk}. The text illustrates application of the sum-to-zero constraints for a 2 x 3 two factor model to show that, in that case, only two γ_{jk} parameters are included (have regression covariates defined) in the regression model. The text also discusses the use of deviation-coded regressors for main effects and interactions in the regression model.

Testing hypotheses in two-way ANOVA

With two-way ANOVA and more complicated models, the question arises of how to test for effects, since there are several ways to test for a given effect. We use the notation from the text where, for example, $SS(\alpha, \beta, \gamma)$ is the regression sum of squares for the model including both main effects and the interaction, and $SS(\alpha|\beta, \gamma) = SS(\alpha, \beta, \gamma) - SS(\beta, \gamma)$. There are then several ways, for example, to test for the main effect of factor 1, by using $SS(\alpha|\beta)$, $SS(\alpha|\beta, \gamma)$, or even $SS(\alpha)$. The book notes that there is disagreement among statisticians over whether $SS(\alpha|\beta)$ (the Type II SS) or $SS(\alpha|\beta, \gamma)$ (the Type III SS) is the better approach, although all would agree that $SS(\alpha)$ is almost never appropriate. As the text notes, the approach based on $SS(\alpha|\beta, \gamma)$ avoids making assumptions about the interaction, but violates the principle of marginality; while the approach based on $SS(\alpha|\beta)$ assumes the absence of interaction but does not violate the principle of marginality. Also note that when the design is balanced so that there are equal frequencies for every factor combination, then all of these sums of squares for a given effect are equal, such as $SS(\alpha|\beta) = SS(\alpha|\beta, \gamma) = SS(\alpha)$. The authors also remark that for models that violate the principle of marginality, the sums of squares obtained from regression models using dummy coding will not equal the sums of squares obtained by using deviation coding (and are incorrect), so one advantage of the Type II approach is the lack of dependence on the regression coding scheme used.