Dummy variable regression (ANCOVA)

The discussion in this chapter starts with a simple single-variable two-group model without interaction:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i$$

where X_i is a measured covariate and D_i is a dummy variable (also called an indicator variable). This model assumes that the relationship between X_i and Y_i does not change with group, hence the two lines are parallel, as illustrated in Figures 7.2 and 7.3. They next generalize to the case where there can be more covariates and more groups, with this model as an example:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i.$$

In this example we are fitting three planes, again all assumed parallel, as shown in Figure 7.6. This example raises the point that if there are m groups, then we only need $m - 1$ dummy variables, so that the group that is uncoded has an intercept of α, while the other groups have intercepts of the form $\alpha + \gamma_i$. A nice way to understand this model is to write the group-specific models separately as shown for the prestige example in the text. If we have one group of special interest we can leave it as the uncoded group, so that the γ_i terms can then be used to test equality of intercepts between the special interest group and any other group. How do we test for equality of intercepts between two groups that both have dummy variables? We then need to calculate the standard error of the difference $\widehat{\gamma}_i - \widehat{\gamma}_j$'s via:

$$SE(\widehat{\gamma}_i - \widehat{\gamma}_j) = \sqrt{\widehat{V}(\widehat{\gamma}_i) + \widehat{V}(\widehat{\gamma}_j) - 2\widehat{Cov}(\widehat{\gamma}_i, \widehat{\gamma}_j)}.$$

Unfortunately the covariance terms are usually not available from published results, so it can be difficult to address these questions using other people's published work. The text introduces an idea, called coefficient quasi-variances $\widetilde{V}(\widehat{\gamma}_i)$, that can be used instead. The idea is that if authors publish these coefficient quasi-variances, then $\widehat{SE}(\widehat{\gamma}_i - \widehat{\gamma}_j)$ can be approximated with:

$$SE(\widehat{\gamma}_i - \widehat{\gamma}_j) \approx \sqrt{\widetilde{V}(\widehat{\gamma}_i) + \widetilde{V}(\widehat{\gamma}_j)}.$$

We will return to this topic and how coefficient quasi-variances are calculated in Chapter 15.

Modeling interactions

The models above specify parallel lines or planes, which may not be a realistic assumption. By adding interaction terms (crossproducts) between dummy variables and covariates, we can allow for the possibility of different slopes. The example model above is generalized in the text into:

$$\begin{aligned} Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i, \end{aligned}$$

which allows both intercepts and slopes to differ in each of the three groups. Again, one way to understand this model is by constructing group-specific regression models as shown on page 145 of the text, and Figure 7.11 shows a way to graphically examine the implications of this model.

The Principle of Marginality

The text's definition: *The principle of marginality specifies that a model including a high-order term (such as an interaction) should normally also include the "lower-order relatives" of that term (the main effects that "compose" the interaction).* This principle is extremely important for understanding how to perform hypothesis tests for many statistical models, for example, for ANOVA, ANCOVA, and some approaches to model selection in multiple regression.

Hypothesis Tests

To conduct tests about main effects and interactions with these ANCOVA models, we use the same approach as illustrated in the previous chapter, setting up a complete and a reduced model for a particular null hypothesis. In doing so we need to use the principle of marginality in selecting appropriate models for comparison. This strategy is illustrated in Tables 7.1 - 7.3. When using SAS procedures to duplicate the analyses in Table 7.2, it helps to understand the distinction between the three types of sums of squares used:

Type I sums of squares are sequential, so that all terms specified previous to the current term define the reduced model.

Type III sums of squares for a given term set the reduced model to include all terms except the term being tested. This approach is very useful for multiple regression models with linear terms, because then a test for a given term measures the increase in sums of squares after all other terms are in

the model. On the other hand, for models with interaction terms, the Type III sums of squares may define a reduced model that violates the principle of marginality.

Type II sums of squares for a given term set the reduced model to include as many terms as possible without violating the principle of marginality.

There is a fourth type of sums of squares, which is used in situations with missing data - for example for factorial designs with factor-combinations that have no observations (empty cells).

More cautions for standardized coefficients

A final note in this chapter mentions that it is not appropriate to standardize either dummy regression terms or interaction terms.