Creating an ANCOVA Model Using PROC GLM

Let's see how to code the ANCOVA model in PROC GLM. The code shown here is for the clinical trial scenario. The CLASS statement specifies one or more classification (or grouping) variables—in this example, the categorical predictor **Treatment**—and creates corresponding indicator variables in the model.

In the MODEL statement, the predictors include any interaction terms that involve the classification variable. This MODEL statement specifies three predictors—the individual variables **Treatment** and **BaselineBP**, and the interaction of **Treatment** and **BaselineBP**. Notice that the interaction variable is specified by an asterisk between the two variable names. The classification variable **Treatment** has three values: *Approved Drug*, *New Drug*, and *Placebo*. Of course, SAS cannot perform calculations on non-numeric values like these.

To enable the values of classification variables to be used in the analysis, PROC GLM creates a numeric indicator variable—that is, a design variable—for each level of each classification variable specified in the CLASS statement. This chart shows the three design variables that are created for the three levels of **Treatment**, and the possible values of each. Notice that design variables can have one of two values—0 or 1. The design variable for *Approved Drug* has the value 1 if the subject is given the approved drug; otherwise, it has the value 0. Similarly, the variables for *New Drug* and *Placebo* have the value 1 if the subject is given the new drug or a placebo, respectively; otherwise, they have the value 0.

So, how does the model defined in PROC GLM relate to the mathematical model of ANCOVA? PROC GLM uses the ordinary least squares method to fit the general linear model $Y = X\beta + \varepsilon$ to your data. This shorter expression of the model is the matrix algebra version. The Y matrix represents the observed values of the response variable **BPChange** for the 93 observations in the data set. The first subject had a blood pressure change of -20.4, the second subject had a change of -5.7, and the last subject had a change of -1.8.

Here are the two matrices that correspond to the second term in the general linear model, $X\beta$. The X matrix is the design matrix that was mentioned earlier, which PROC GLM creates. This matrix has one row for each observation in the data set (that is, one row for each of the 93 subjects in the clinical trial) and eight design columns. For example, the first row corresponds to the first subject, and it contains the coefficients of the parameters (shown in the beta matrix) for the first subject.

The eight design columns are as follows:

- The first column, which contains *1*s, corresponds to the intercept.
- The next three columns, which contain *0*s and *1*s, correspond to the three levels of the classification variable **Treatment** (*Approved Drug*, *New Drug*, and *Placebo*). Note that the default order of the columns for the design variables is the sort order of the values of their levels.
- The fifth column corresponds to the value of **BaselineBP**.
- The last three columns correspond to the interaction terms between the three levels of the class variable **Treatment** and the covariate **BaselineBP**.

The β matrix contains a single column of all eight of the parameters whose values are to be estimated using the ordinary least squares method. These parameters include the following:

- μ represents the intercept τ_1 represents the effect of treatment 1 (*Approved Drug*)
- τ_2 represents the effect of treatment 2 (*New Drug*)
- τ_3 represents the effect of treatment 3 (*Placebo*)
- β represents the slope of **BaselineBP**
- ϕ_1 represents the slope effect for *Approved Drug*
- ϕ_2 represents the slope effect for *New Drug*
- ϕ_3 represents the slope effect for *Placebo*

In the X matrix, the first four columns are linearly dependent. This means that PROC GLM fits an over-parameterized model. The last four columns are also linearly dependent. Keep in mind that there are many ways to code classification variables with design variables (which are also known as parameterizations). However, only a few of these coding methods are easy to interpret. In fact, the predictions and many of the statistics do not change based on the choice of parameterization method. The ε matrix represents the random error terms (that is, the unexplained variability) for the 93 observations in the data set. Specifically, the error terms are for each of the three treatment levels. Residuals, which are the differences between the observed change and the predicted change in blood pressure for each subject, are estimates of these error terms.

Close