Fitting a Linear Model with Nested Classification

We're ready to analyze the data for the school study that examines the effect of four different teaching materials on student test scores. This demonstration shows how to perform an exploratory analysis of the data and then fit a linear mixed model.

To show the distribution of scores by materials, we use this PROC SGPANEL step to create a panel of box plots. In the PROC SGPANEL statement, DATA= specifies the data set **mydata.scores**. The PANELBY statement specifies **Material** as the classification variable. The COLUMNS= option specifies that the four plots (one for each value of **Material**) will be placed side-by-side in four columns. The VBOX statement requests a vertical box plot of **Score**.

Let's run this code.

```
title1 "Distribution of Scores by Materials";
proc sgpanel data=mydata.scores;
   panelby material / columns=4;
   vbox score;
run;
```

In the results, notice that this panel of plots has a shared Y axis for **Score**, and each individual plot has the type of material labeled at the top. These vertical box plots show some basic descriptive statistics about the outcome of using each instructional material. Box plots display the distribution of data by using a rectangular box and whiskers. Inside each box, the horizontal line indicates the median and the diamond indicates the mean. The top line of the box indicates the third quartile and the bottom line indicates the first quartile. The lines that extend in both directions from the box are the whiskers, which indicate a data range outside of the box.

Based on these plots, there appear to be some differences among the four treatment means. However, we can't tell whether the difference is due to the material effect or random errors. Material A seems to be less variable than materials B and C.

As a next step, let's determine whether there is a significant difference in the mean test scores using the four different teaching materials. This PROC GLIMMIX code analyzes the **mydata.scores** data. The CLASS statement specifies both **Material** (a fixed effect) and **Teacher** (a random effect) as classification variables. The MODEL statement specifies **Score** as the dependent variable and only the fixed independent variable **Material**. The RANDOM statement specifies **Teacher(Material)** as a random effect. The COVTEST statement specifies a label followed by the GLM keyword. This statement tests the model against a null model of complete independence. All G-side covariance parameters are eliminated and the R-side covariance structure is reduced to a diagonal structure.

Let's run this code.

```
proc glimmix data=mydata.scores;
   class material teacher;
   model score=material;
   random teacher(material);
   covtest 'Test Need for Random Effect' glm;
run;
```

In the PROC GLIMMIX results, the Model Information table displays basic information about the fitted model, such as the link and variance functions, the distribution of the response, and the data set. The default estimation technique for the normal distribution is Restricted Maximum Likelihood. The row in this table labeled Degrees of Freedom Method lists the method used for estimating the denominator degrees of freedom for the fixed effect. Five possibilities for this row are Containment, Between-Within, Residual, Satterthwaite, Kenward-Roger, and none. Containment is the default method when the RANDOM statement is specified. To specify other methods, you can use the DDFM= option in the MODEL statement.

The Class Level Information table lists the levels of every variable specified in the CLASS statement. You should check this information to make sure that the data is correct. You can adjust the order of the CLASS variable levels by using the ORDER= option in the PROC GLIMMIX statement. The default order is alphanumeric.

The Number of Observations table shows the total number of observations in the data set and how many are used in fitting the model. All 120 observations in the data set were used for the analysis.

The Dimensions table lists the sizes of relevant matrices. This table can be useful in determining CPU time and memory requirements. The G-side covariance parameter corresponds to σ^2_t. The R-side covariance parameter is σ^2. The five columns in the X matrix correspond to the intercept and four design columns for the classification variable Material, and the 20 columns in the Z matrix correspond to the 20 teachers.

As shown in the Optimization Information table, the optimization is performed using a Dual Quasi-Newton algorithm, and the rows of this table describe the iterations that this algorithm takes in order to minimize the objective function. Other algorithms are available by using the NLOPTIONS statement in PROC GLIMMIX. The Iteration History table describes the optimization of the restricted log-likelihood function.

The Fit Statistics table provides information for goodness of fit. All information criteria consider both the fit of the model and the model complexity. The model with more parameters receives a larger penalty. The Bayesian Information Criterion (BIC) tends to produce a larger penalty than both Akaike Information Criterion (AIC) and finite-sample corrected Akaike Information Criterion (AICC) for the same model. For all information criteria, smaller values indicate a better model.

The Covariance Parameter Estimates table shows the following variance component estimates: $\hat{\sigma}^2_t$ for **Teacher(Material)** is 34.6596 and $\hat{\sigma}^2$ is 17.2550.

The Type III Tests of Fixed Effects table contains the F test for **Material** (F = 0.54), with a *p*-value of 0.6647. Therefore, there is not enough evidence to conclude that the average test scores for **Material** across all teachers are statistically significantly different at a 5% significance level.

For details about how PROC GLIMMIX computes the F statistic for fixed effects, click the Information button.

The null hypothesis for the COVTEST statement is that a model that assumes independence (that is, with no RANDOM statement) fits as well as the current model with the RANDOM statement. In the table named "Tests of Covariance Parameters Based on the Restricted Likelihood," the low *p*-value (<.0001) provides evidence to reject the null hypothesis and conclude that the RANDOM statement is needed. The note below the table indicates that the p-value is computed based on a mixture of chi-squares.

What would happen if you incorrectly specified **Teacher(Material)** as a fixed effect? Let's find out. In this second version of the PROC GLIMMIX step, the random effect **Teacher(Material)** has been added to the MODEL statement. The COVTEST statement is not needed and is removed. An OUTPUT statement is added to request that the residual variance estimate be output to the temporary **checkvar** data set. The PROC PRINT step then prints **checkvar**.

Let's run this code.

```
title 'Random Effect is Incorrectly Specified as Fixed Effect';
proc glimmix data=mydata.scores;
   class material teacher;
   model score=material teacher(material);
   output out=checkvar variance=ResidualVariance;
run;

proc print data=checkvar (obs=1);
   var ResidualVariance;
run;

title;
```

In the PROC PRINT results, we see that the estimate for the residual variance is the same as the one obtained from the previous model. (Remember that the residual variance estimate for the original model was shown in the Covariance Parameter Estimates table.) These estimates are the same in this case, because the data is balanced.

In the new PROC GLIMMIX results, we find the Type III Tests of Fixed Effects table. This time, the F value for the **Material** effect is 6.99 with the denominator degrees of freedom 100. The *p*-value is 0.0003. Based on these results, you would incorrectly conclude that there is a significant difference in the average test scores among the four teaching materials.

Close