Print

Practice 6.1 (Level 1): Analyzing Nested Data

Task

In this practice, you analyze nested data using PROC GLIMMIX.

Recall that data were collected by a school district to assess the reading skill progress of students in their first year of formal schooling. The data are stored in the **mydata.school** data set. In a previous lesson, you used ANOVA models to evaluate the significance of some factors in explaining the difference in the average **Reading3** test scores. The data set has another variable, **Teacher**. Teachers are nested within schools. Factors of interest are **School** and **Gender**.

Reminder: Make sure you've defined the **mydata** library.

1. For an initial data exploration, do the following:
   - Use PROC PRINT to look at the first 25 observations in **mydata.school**.
   - Use PROC FREQ to examine whether the data in **mydata.school** is balanced for **School** and **Gender** combinations.

```
proc print data=mydata.school(obs=25);
run;

proc freq data=mydata.school;
   tables School*Gender / nocol norow nopercent;
run;
```

   As shown in the results, the data is unbalanced.

2. Write a PROC SGPANEL step to generate side-by-side box plots of **Reading3** versus **School** and **Gender** combinations. (Hint: Put both **School** and **Gender** in the PANELBY statement.) Use a WHERE statement to make sure that the graph uses the data values only for observations in which all of **Reading3**, **Teacher**, **School**, and **Gender** are present. Why do you see graphs for only three out of four schools? Do any groups seem to be different from other groups?

```
title1 "Distribution of Scores by School and Gender";
proc sgpanel data=mydata.school;
   where Reading3 is not missing
         and Teacher is not missing
         and School is not missing
         and Gender is not missing;
   panelby School Gender / layout=lattice;
   vbox Reading3;
run;
```

   Examine the results. You do not see data from *Dogwood* because all values for **Teacher** are missing for this school. Data from *Dogwood* are not included in the analysis. The test scores for female students at *Cottonwood* seem to be different from other groups.

3. Are teachers crossed or nested within schools? Should you consider **Teacher** as a fixed effect or a random effect, and how does that affect your analysis?

   Teachers are nested within schools. **Teacher**(**School**) should be considered a random effect. Specifying **Teacher(School)** as a random effect enables you to do the following:
   - estimate the variance in **Reading3** scores among the teachers
   - apply your conclusions about the material effect over a population of teachers

4. Use PROC GLIMMIX to determine whether there is a difference in the mean **Reading3** test scores among **School**, **Gender**, and **School*Gender**. What are the estimates of the variance components? What do you conclude about the fixed effects?

```
title;
proc glimmix data=mydata.school;
   class School Gender Teacher;
   model Reading3=School Gender School*Gender;
   random Teacher(School);
run;
```

Examine the results. As shown in the Covariance Parameter Estimates table, the estimated teacher variance is 51.97 and the estimated residual variance is 1445.28.

In the Type III Tests of Fixed Effects table, you can see that the *F* value for **School*Gender** is 5.16 and the *p*-value is 0.0070. Averaged across all teachers, there is a significant effect of **School*Gender** on the average **Reading3** test scores.

Interpreting main effects in the presence of a significant interaction might be misleading. Instead, simple effects can be explored by using the LSMEANS statement with the SLICE option. However, for comparison, notice that the *p*-values for both the **School** and **Gender** effects are higher than 0.05.

5. Use an LSMESTIMATE statement to estimate the difference in **Reading3** test scores between *Cottonwood* girls and all other students. What do you conclude?

```
proc glimmix data=mydata.school;
   class School Gender Teacher;
   model Reading3=School Gender School*Gender;
   random Teacher(School);
   lsmestimate School*Gender 'Cottonwood Girls vs. All Others'
               5 -1 -1 -1 -1 -1 / divisor=5 elsm;
run;
```

Examine the results. As shown in the Least Squares Means Estimate Coefficients table, because *Dogwood* is missing the value of **Teacher** for all observations, only six of the eight **School** by **Gender** by **Teacher** combinations have estimates. Thus, you need only six coefficients in the LSMESTIMATE statement. The ELSM option shows you how the coefficients are applied to the least squares means, after the DIVISOR=5 option is applied.

In the Least Squares Means Estimates table, the average **Reading3** test score for *Cottonwood* girls is 34.06 below the average scores for all other students. The standard error for the mean difference is 10.91. The *p*-value for the difference is 0.0022. Presuming an alpha level of 0.05, you would reject the null hypothesis and conclude that, on average, *Cottonwood* girls score significantly lower than all other students.

6. What happens if you incorrectly specify the random effect as a fixed effect? Try it.

```
proc glimmix data=mydata.school;
   class School Gender Teacher;
   model Reading3=School Gender School*Gender Teacher(School);
   output out=checkvar variance=ResidualVariance;
run;

proc print data=checkvar(obs=1);
   var ResidualVariance;
title 'Check Residual Variance';
run;
title;
```

In the results, the PROC PRINT output shows that the estimated residual variance is 1437.06. This is slightly different from the estimate from the previous model. Recall that you do not have balanced data.

In the Type III Tests of Fixed Effects table, you can see that the *F* value for **School*Gender** is 4.36 and the *p*-value is 0.0148. There is a significant effect of **School*Gender** on the average **Reading3** test scores at a significance level of 0.05.

Interpreting the main effects of **School** and **Gender** in the presence of the significant interaction might be misleading. However, for comparison, notice that the *F* value for **School** is 5.20, the denominator degree of freedom for **School** is 123 (compared with 5 from the previous model), and the *p*-value is 0.0068. The *F* values and *p*-values for the effects are different from the ones obtained from the previous model. Specifying a random effect incorrectly as a fixed effect jeopardizes your conclusions about the treatment effects.

Hide Solution

Close