Identifying Violations of ANOVA Assumptions

Recall the last model that you fit to the **mydata.school** data set. Let's now use the diagnostic plots that are available in PROC GLM to evaluate the ANOVA model assumptions. We can also output the residuals and use PROC UNIVARIATE to look at the descriptive statistics of the residuals. Let's look at the code.

The PLOTS option in the PROC GLM statement controls the plots produced with ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. The available plots include the following: DIAGNOSTICS (UNPACK) requests that a panel of regression diagnostics for the fit be displayed. The panel displays scatter plots of residuals, absolute residuals, studentized residuals, and observed responses by predicted values; studentized residuals by leverage; Cook's D by observation; a Q-Q plot of residuals; a residual histogram; and a residual-fit spread plot. The UNPACK option unpanels the diagnostic display and produces the series of individual plots that form the paneled display. The OUTPUT statement creates a new SAS data set that saves diagnostic measures that are calculated after fitting the model. All the variables in the original data set are included in the new data set, along with variables created in this statement. These new variables contain the values of a variety of diagnostic measures that are calculated for each observation in the data set. The ODS SELECT statement specifies the selected tables to be displayed in the Output window.

Let's submit the code.

```
proc glm data=mydata.school plots(unpack)=diagnostics;
   class gender semesters school;
   model reading3=gender semesters school gender*school;
   output out=check r=residuals p=predicted;
run;
quit;

goptions reset=all;
ods select moments BasicMeasures GoodnessOfFit;
proc univariate data=check;
   var residuals;
   histogram / normal;
run;
```

Here is the PROC UNIVARIATE output. The residuals are skewed to the right. The normality tests indicate that residuals are not normally distributed. However, the skewness and kurtosis statistics indicate that the departure from normality might be minor. The histogram of the residuals indicates that they are skewed slightly to the right. The normal probability plot also indicates that residuals are skewed slightly to the right. The residual plot shows that the model seems to fit the data fairly well. The variances might not be the same for all groups. Notice that for some groups, the relatively large range of the residuals might be due to the fact that there are more observations in these groups.

The homogeneity of variances tests in PROC GLM are available only for one-way ANOVA. One approach to testing the homogeneity of variance assumption for two-way or higher-ordered ANOVA would be to create a new variable in a DATA step that is the actual treatment group for each observation. Essentially, this creates one factor with 24 levels (2x4x3) for this example, and treats the analysis as a one-way ANOVA. This enables you to assess the equality of variances using the HOVTEST option in the MEANS statement in PROC GLM. Notice the MEANS statement option. HOVTEST performs the default test, which is Levene's squared residuals test for homogeneity (equality) of variances. The null hypothesis for this test is that the variances are equal. Other tests are available including Bartlett's test, O'Brien's test, and the Brown-Forsythe test.

Let's run the code and look at a portion of the output.

```
data school;
   set mydata.school;
   group=compress(gender||school||semesters);
run;

ods select classlevels hovftest means;
proc glm data=school;
   class group;
   model reading3=group;
   means group / hovtest;
```

```
run;
quit;
```

The reason that only 23 levels were created is that there is one empty cell for FCottonwood8. The null hypothesis for tests for homogeneity of variance is that the variances are equal. The alternative hypothesis is that the variances are not all equal. The *p*-value for Levene's test is 0.7420. Therefore, you do not reject the null hypothesis. The variances do not appear to be sufficiently unequal to cause concern for the validity of the ANOVA model under the equal variance assumption. The MEANS statement produces the sample size, arithmetic means, and standard deviations of **Reading3** for each group. Because you have unbalanced data, you might want to examine the results from the LSMEANS statement for means comparisons.

Close