Performing Multicollinearity Diagnostics

This demonstration shows how to identify multicollinearity by using diagnostics that are produced by PROC CORR and PROC REG.

First, we use PROC CORR to examine the pairwise correlations between the independent variables. The input data set is **d_paper**, which PROC GLMSELECT created in the last demonstration. In the PROC CORR statement, the NOSIMPLE option suppresses the printing of simple descriptive statistics for each variable in the results. PLOTS=MATRIX requests a scatter plot matrix for the variables. The VAR statement specifies &_GLSMOD (the automatic macro variable that PROC GLMSELECT created in the last demonstration), which stores the list of effects in the final model (**Amount**, **Amount^2**, and **Amount^3**).

We run the code.

```
title 'Collinearity Diagnosis for the Cubic Model';
proc corr data=d_paper nosimple plots=matrix;
   var &_GLSMOD;
run;
```

In the table at the top of the results, the correlation coefficients for all variables are close to 1. It is obvious that very high correlations exist between **Amount** and **Amount^2** (r=0.98), **Amount** and **Amount^3** (r=0.95), and **Amount^2** and **Amount^3** (r=0.99). The scatter plots show a slight curvilinear relationship among these variables.

Now let's use PROC REG to generate the variance inflation factor (VIF) and condition index values. In this PROC REG statement, we again specify **d_paper** as the input. We don't need plots, so we specify the option PLOTS=NONE. The MODEL statement specifies the dependent variable (**Strength**) and the &_GLSMOD macro variable represents the independent variables. The following options request the diagnostics: VIF specifies the variance inflation factor, and COLLIN and COLLINOINT specify the condition index values.

We run the code.

```
proc reg data=d_paper plots=none;
   model strength=&_GLSMOD / vif collin collinoint;
run;
quit;
title;
```

Let's look at the results. The Parameter Estimates table shows the variance inflation factors (VIFs). Remember that values greater than 10 indicate multicollinearity. Here, the VIFs are quite large for all three independent variables. **Amount^2** has the highest value but removing this variable would not be hierarchically sound.

Next, the two Collinearity Diagnostics tables show the collinearity diagnostics generated by the COLLIN option and the COLLINOINT option, respectively. The collinearity analysis includes eigenvalues, condition indices, and decomposition of the variances of the estimates with respect to each eigenvalue. In the second table, which is produced by COLLINOINT, the intercept variable is adjusted out (that is, not included in the diagnostics).

Notice that the analysis in PROC REG is reported with eigenvalues of X'X rather than singular values of X. The eigenvalues of X'X are the squares of the singular values of X. The condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue. The largest condition index is the condition number of the scaled X matrix.

For each variable, PROC REG produces the proportion of the variance of the estimate accounted for by each principal component. Variance proportions greater than 0.50 associated with a large condition index identify the subsets of the predictor variables that are multicollinear.

Let's look at a process for interpreting the COLLIN guidelines in conjunction with the COLLINOINT statistics and evaluating the severity of the collinearity with the intercept adjusted out. To start, look at the last row of the first Collinearity Diagnostics table and find the Condition Index (CI) value. Remember that a CI greater than 100 indicates strong collinearity. If the CI is less than 100, then there is no indication of strong collinearity and no further investigation is required.

In our results, is the CI greater than 100? Yes, it is 192, which indicates a strong collinearity. If the CI is greater than 100, we look at the Proportion of Variation value for the Intercept in that row. In our results, is the value greater than 0.5? Yes, it is 0.8. If the intercept's variance proportion value is greater than 0.5, then we conclude that the intercept is involved in collinearity, and we look at values in the last row of the next table (the intercept-adjusted table). If any variables in that row have a variance proportion value greater than 0.5, they are the collinear variables.

In our results, we look at the variance proportion values for the three independent variables in the last row of the bottom table. All of these values are greater than 0.5, so all of these variables are involved in the collinearity. We will need to figure out how to deal with this collinearity.

Before we conclude this demonstration, let's quickly finish our discussion of interpreting the collinearity diagnostics. Back in the first Collinearity Diagnostics table, in the row in which the CI value is greater than 100, if the variance proportion for the intercept is less than 0.5, then the collinear variables are the ones in that row whose variance proportion is greater than 0.5. In this situation, you do not have to look at the second table at all.

Close