Text Version

Expand Print

Multicollinearity Diagnostics

To determine whether multicollinearity is present in a model that has two or more explanatory variables, you can examine the following diagnostic measures: correlation statistics, variance inflation factors (or VIFs), and condition index values.

Correlation is a measure of the degree of linear relationship between two variables. You can use correlation statistics to check for bivariate correlations between pairs of independent variables. The Pearson correlation coefficient is calculated as shown here. To calculate the correlation statistic between the two variables x_1 and x_2, you divide the covariance of x_1 and x_2 by the square root of the product of the variance of x_1 and the variance of x_2. You can use PROC CORR to produce correlation statistics.

VIFs are useful in determining whether multicollinearity exists or not, and which variables might be involved in the multicollinearity. To produce the VIF, you specify the VIF option in the MODEL statement of PROC REG. The VIF for the i^{th} independent variable is calculated as 1 divided by the quantity ($1-R_i^2$), where R_i^2 is the coefficient of determination for the regression of the i^{th} independent variable on all other independent variables. In other words, R^2 is calculated for each predictor from a model that considers that predictor as the response variable and regresses it on all the other predictors in the model. As the multicollinearity of an x variable with other regressors increases, the VIF for that x variable also increases and can have an infinite limit.

To diagnose multicollinearity, you can also use condition index values, which are computed based on eigenanalysis. You can examine the eigenvalues of the sums of squares and cross-product matrix (or SSCP matrix) X'X. Condition indices are the square roots of the ratios of the largest eigenvalues (the lambdas) to the individual i^{th} eigenvalues. Later in this lesson, you learn how to obtain the condition index values using PROC REG.

If you're not familiar with eigenanalysis, looking at a graph might be helpful. Principal components are obtained by computing the eigenvalues and eigenvectors of the SSCP matrix (X'X). The eigenvalues (the lambdas in the graph) are the variances of the components. If the correlation matrix has been used, the variance of each input variable is one; the sum of variances (eigenvalues) of the component variables is equal to the number of variables. Eigenvectors are the coefficients of the linear equations that relate the component variables (w_1 and w_2) to the original variables. A very small eigenvalue that is close to zero implies severe multicollinearity. (In this example, because the correlation is between only two variables, the more specific term collinearity is displayed.) A zero eigenvalue indicates perfect multicollinearity among independent variables. Therefore, the ratio of the eigenvalues can be useful for examining multicollinearity. Eigenvalues of relatively equal magnitudes indicate that there is little multicollinearity, and a wide variation in magnitudes indicates severe multicollinearity. For a more complete discussion of eigenanalysis and multicollinearity, click the Information button.

Close