From
Hair et al.
5th Ed.

Studentized residual Most commonly used form of *standardized residual.* It differs from other standardization methods in calculating the standard deviation employed. To minimize the effect of a single *outlier,* the standard deviation of residuals used to standardize the *i*th residual is computed from regression estimates omitting the *i*th observation. This is done repeatedly for each observation, each time omitting that observation from the calculations. This approach is similar to the *deleted residual,* although in this situation the observation is omitted from the calculation of the standard deviation.

Tolerance Commonly used measure of *collinearity* and *multicollinearity.* The tolerance of variable *i* (TOL_i) is $1 - R_i^{*2}$, where R_i^{*2} is the coefficient of determination for the prediction of variable *i* by the other predictor variables. Tolerance values approaching zero indicate that the variable is highly predicted (collinear) with the other predictor variables.

Variance inflation factor (VIF) Measure of the effect of other predictor variables on a regression coefficient. VIF is inversely related to the *tolerance* value ($VIF_i = 1 \div TOL_i$). Large VIF values (a usual threshold is 10.0, which corresponds to a tolerance of .10) indicate a high degree of *collinearity* or *multicollinearity* among the independent variables.

Assessing Multicollinearity

As discussed in chapter 4, **collinearity** and **multicollinearity** can have several harmful effects on multiple regression, both in the interpretation of the results and in how they are obtained, such as stepwise regression. The use of several variables as predictors makes the assessment of multiple correlation between the independent variables necessary to identify multicollinearity. But this is not possible by examining only the correlation matrix (which shows only simple correlations between two variables). We now discuss a method developed specifically to diagnose the amount of multicollinearity present and the variables exhibiting the high multicollinearity. All major statistical programs have analyses providing these collinearity diagnostics.

A Two-Part Process

The method has two components. First is the **condition index,** which represents the collinearity of combinations of variables in the data set (actually the relative size of the **eigenvalues** of the matrix). The second is the **regression coefficient variance–decomposition matrix,** which shows the proportion of variance for each regression coefficient (and its associated variable) attributable to each condition index (eigenvalue). We combine these in a two-step procedure:

1. Identify all condition indices above a threshold value. The threshold value usually is in a range of 15 to 30, with 30 the most commonly used value.
2. For all condition indices exceeding the threshold, identify variables with variance proportions above 90 percent. A collinearity problem is indicated when a condition index identified in step 1 as above the threshold value accounts for a substantial proportion of variance (.90 or above) for *two or more* coefficients.

The example shown in Table 4A.1 illustrates the basic procedure and shows both the condition indices and variance decomposition values. First, a threshold of 30

TABLE 4A.1 Hypothetical Coefficient Variance–Decomposition Analysis with Condition Indices

Condition Index (u_i)	*Proportion of variance of coefficient:* b_1	b_2	b_3	b_4	b_5	b_6
1.0 u_1	.003	.001	.000	.003	.000	.000
4.0 u_2	.000	.021	.005	.003	.000	.000
16.5 u_3	.000	.012	.003	.010	.000	.001
45.0 u_4	.001	.963	.003	.972	.983	.000
87.0 u_5	.003	.002	.000	.009	.015	.988
122.0 u_6	.991	.001	.987	.003	.002	.011

for the condition index selects three condition indices (u_4, u_5, and u_6). Second, coefficients exceeding the .90 threshold for these three condition indices are b_1 and b_3 with u_6; b_2, b_4, and b_5 with u_4; and b_6 with u_5 (see the underlined values in Table 4A.1). However, u_5 has only a single value (b_6) associated with it; thus no collinearity is shown for this coefficient. As a result, we would attempt to remedy the significant correlations among two sets of variables: (1) V_1, V_3 and (2) V_2, V_4, V_5.

An Illustration of Assessing Multicollinearity

In chapter 4, we discussed the use of multiple regression in predicting the usage level (X_9) for HATCO customers. The stepwise procedure identified three statistically significant predictors: X_3, X_5, and X_6. However, before we accept these regression results as valid, we must examine the degree of multicollinearity and its effect on the results. To do so, we employ the two-part process (condition indices and the decomposition of the coefficient variance) and make comparisons with the conclusions drawn from the **variance inflation factor (VIF)** and **tolerance** values.

As discussed in chapter 4 and also presented in Table 4A.2 (p. 222), the VIF and tolerance values indicate inconsequential collinearity. No VIF value exceeds 10.0, and the tolerance values show that collinearity does not explain more than 10 percent of any independent variable's variance. This conclusion is supported when we employ the two-step procedure. First, examine the condition indices. We fail to pass the first step, as no condition index is greater than 30.0. Even if we were to proceed to the second step by using a threshold value of 15 for the condition index, we would select only a single condition index (u_4), where only one coefficient (the intercept) loads highly. Thus, we can find no support for the existence of multicollinearity in these regression results, just as indicated by the tolerance and VIF measures.

Identifying Influential Observations

In chapter 4, we examined only one approach to identifying **influential observations**, that being the use of studentized residuals to identify outliers. As noted then, however, observations may be classified as influential even though they are not recognized as outliers. In fact, many times an influential observation will not be identified as an outlier because it has influenced the regression estimation to