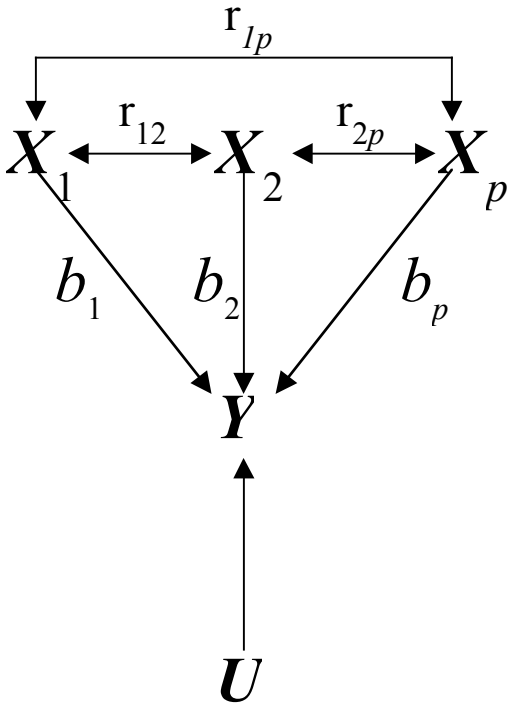
The Mathematical Theory of Regression

The basic formulation for multiple regression is diagramed in Figure 1. Here, there are a series of *predictor* variables, termed $x_1, x_2, \ldots x_n$. There are p of these predictor variables. Many statisticians refer to the predictor variables as *independent* variables or *IV*s. There is a single *predicted* variable, y, also known as the *dependent* variable or *DV*. Variable u is a *residual*, sometimes called an *error* variable or a *disturbance*. The b_is are *partial regression coefficients*; b_i gives the predictive influence of the ith *IV* on the *DV* controlling for all the other *IV*s. The term r_{ij} is the correlation between the ith and jth *IV* and a is a constant reflecting the scale of measurement for y.

Now consider the kth indidivual in a sample. Let y_k denote the kth individual's score on the *DV* and let x_{ki} denote the kth individual's score on the ith *IV*. The multiple regression model writes the kth individual's *DV* as a weighted linear combination of the kth individual's scores on the p *IV*s plus the kth individual's score on the residual. In algebra, we may write

$$Y_k = a + b_1 X_{k1} + b_2 X_{k2} + \ldots b_p X_{kp} + U_k \qquad (1)$$

Figure 1. Pictorial or path diagram of the multiple regression model.



It is convenient to express the multiple regression model in matrix form. Let us write out the structural equations for all N individual's in a sample:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \cdots & X_{Np} \end{pmatrix} \begin{pmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix} \tag{2}$$

Let **y** denote the column vector of Y_is; let **X** denote the (N by p+1) matrix of the X_{ij}s; let **b** denote the column vector consisting of a and the partial regression coefficients b_1 through b_p; and let **u** denote the column vector of residuals. Equation (2) may be parismoniously written in matrix form as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{u} \tag{3}$$

form as

Another way to write Equation (3) is to note that the observed y score for an individual may be written as the sum of the *predicted y* plus error (u). For the kth individual we may write

$$Y_k = \hat{Y}_k + U_k \tag{4}$$

Let **y** denote the (N by 1) column vector of predicted scores. Then another way to write Equation (3) is

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{u} \tag{5}$$

and with a little algebraic manipulation, we can prove that

$$\hat{\mathbf{y}} = \mathbf{Xb}. \tag{6}$$

Estimation of the Parameters

We now want to find a solution for vector **b**. To do this we require some criteria for picking reasonable values for the b_is (and a) from all the possible values that the b_is can take. (Mathematically, when there are more observations than *IV*s, then there are more structural equations than there are unknowns and the solution is *overdetermined*.) The traditional criterion for selecting the b_is is that we want those that minimize the *square prediction errors*. That is, we want to minimize

$$\mathbf{u}'\mathbf{u} = \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (y - \hat{y})^2 \tag{7}$$The process of minimizing a function is d

$$\frac{\partial \mathbf{u}'\mathbf{u}}{\partial \mathbf{b}} = \frac{\partial (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{\partial \mathbf{b}} = \mathbf{0} \tag{8}$$

We need not be concerned with the mechanics of reducing Equation (8), so set us state the result:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y} \tag{9}$$

yielding

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{Xy}. \tag{10}$$

Recall that (**X'X**) is the *sum of squares and cross products* matrix uncorrected for the mean or the USSCP matrix for the x_{ij}s (including the column vector of 1's in the first

column of **X**). The vector (**X'y**) is the *sum of cross products* between matrix **X'** and vector **y**. Hence vector **b** is estimated as the inverse of the USSCP matrix postmultipled by the sum of cross products vector between the independent variables and the dependent variable.

The covariance matrix for the elements of **b** is

$$Var(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2_{error}$$

where σ^2_e is the error variance. The error variance is estimated by the *mean square error* which is the *sum of squares* for the residuals divided by its degrees of freedom. Let s^2 denote this estimate of σ^2_e. Then,

$$s^2 = \hat{\sigma}^2_{error} = MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{\mathbf{u}'\mathbf{u}}{N - p - 1} = \frac{\sum_{i=1}^{N}(y - \hat{y})^2}{N - p - 1}$$

Multiple regression programs estimate the variance-covariance matrix by replacing σ^2_e with s^2 giving

The standard error for the *i*th parameter is simply the square root of the *i*th diagonal element of the matrix

$$\hat{Var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}s^2$$

The magnitude of a b_i is determined by three factors: (1) the magnitude of the direct predictability of X_i on Y; (2) the scale of measurement for Y; and (3) the scale of measurement for X_i. When the scale of measurement for the variables is arbitrary, it is customary to express the partial regression coefficients as *standardized partial regression coefficients*. The standardized coefficients are the *b* weights that one would obtain if one were to standardize every variable to a mean of 0.0 and a variance of 1.0 prior to the regression analysis. That is, the standardized coefficients are the coefficients in the equation

$$Z_y = \beta_1 Z_{x1} + \ldots + \beta_p Z_{xp} + \beta_u Z_u$$

The intercept term is not written because it is 0 when the variables are standardized. Similarly, a standardized weight (β_u) is added to the standardized residual. To convert unstandardized regression coefficients to standardized coefficients, use the following equation

$$\beta_i = b_i \frac{\sigma_{X_i}}{\sigma_Y}$$

where σ_{Xi} is the standard deviation of the *i*th *IV* and σ_Y is the standard deviation of the dependent variable.

Standardized regression coefficients are directly comparable to one another. That is, a β of .20 indicates a stronger direct predictive relationship than a β of .15.

Regression, Semipartial, and Partial Correlation

In this section, we want to obtain a deeper appreciation for the terms *direct predictive effects*, *indirect predictive effects,* and *total predictive effects* that were used rather glibly above. To do this, consider a regression model with two *IV*s. For convenience, we will assume that all variables are standardized.

The relationship among the variances and covariances of the three variables may be depicted by the Venn diagram in Figure 2, adapted from Cohen and Cohen (1983). The area of each circle is the variance for the appropriate variable, 1.0 in this case because the variables have been standardized. The variance that variables X_1 and Y have in common is the area $a + c$, the variance in common between variables X_2 and Y is area $b + c$, and the variance in common between variables X_1 and X_2 is area $c + d$. Areas e, f, and g denote the variance unique to respectively variables x_1, y, and x_2. Hence, area e is equivalent to the variance of the residual.

Recall that the square of the correlation between two variables gives the proportion of variance in one variable predicted by the other variable. Hence, the squared correlation between variables Y and X_1 is $a + c$ and the squared correlation between Y and X_2 is $b + c$. The area $a + b + c$ is the proportion of variance in Y that is predictable from X_1 and X_2 jointly. This is known as the square of the *multiple correlation*, usually denoted as R^2. In the case of two *IV*s, it is simple to obtain an estimate of R^2. It is simply

$$R^2 = (a + b + c)$$

In Figure 2, the squared correlation between one *IV*, say X_1, and Y is the sum of two areas, $a + c$ in this case. Area c is the proportion of variance in Yexplained by both X_1 and X_2. Area a is the proportion of variance in Y explained by X_1 uniquely. Similarly, area b is the proportion of variance in Y explained by X_2 uniquely. These unique areas are the squared *semipartial* correlations between the respective variables and the *DV*, sometimes known as the *part* correlation. The semipartial correlation gives the increment in R^2 that occurs when the variable is added to the regression equation. Consider a simple univariate regression of Yon X_1. The squared multiple correlation in this case is $(a + c)$. If we now add variable X_2, we increase the preditability of Y by area b. This correlation is called *semi*partial because it is the correlation between two variables when the variance of a third variable is removed from one and only one variable. The semipartial correlations may be computed directly from the first-order correlations:

$$r_{Y(1.2)} = \sqrt{a} = \frac{r(X_1, Y) - r(X_2, Y)r(X_1, X_2)}{\sqrt{1 - r(X_1, X_2)^2}}$$

A related index of association between two variables is the *partial* correlation. The squared partial correlation is the proportion of variance in Y that is predictable from X_1 and that is *not* predictable from X_2. Consider the squared partial correlation between X_1 and Y controlling for X_2. We first remove the variance in Y that is predictable from X_2 or area $(b + c)$. The remaining variance in Y is $(a + e)$. The proportion of this variance that is predictable from X_1 is a. Hence, the squared partial correlation is

$$r^2_{y1.2} = \frac{a}{a + e}$$

Just as the semipartial correlation may be derived from the first-order correlations, so may the partial correlation. The formula is

$$r_{Y1.2} = \frac{r(X_1, Y) - r(X_2, Y)r(X_1, X_2)}{\sqrt{(1 - r(X_2, Y)^2)(1 - r(X_1, X_2)^2)}}$$

Another meaning of the partial correlation is that it is the correlation between the residuals of two variables that have been regressed on the third variable. Suppose that we regressed X_1 on X and calculated the residuals. Call the residuals U_1. Now suppose that we regressed Y on X_2 and calculated the residuals, say variable U_Y. The correlation

between U_1 and U_Y is equivalent to the partial correlation between X_1 and Y controlling for X_2. Hence, the partial correlation squared is the proportion of common variance between two variables when the variance predictable from a third variable has been removed from *both* variables.

The Squared Multiple Correlation

There are several equivalent formulas for the squared multiple correlation. One classic derivation is to express the total sum of squares for *y* or SS_t in terms of the sums of squares due to prediction (the *sum of squares for regression* or SS_r) and the sum of squares due to lack of prediction (the *sum of squares error* or SS_e. In this case the term *sum of squares* refers to the sum of squared deviations from the mean. Let us consider the regression equation for the *i*th individual, or

$$Y_i = \hat{Y}_i + U_i \quad \text{(X)}$$

Now subtract the *y* mean from both sides of (21), giving

$$Y_i - \bar{Y} = (\hat{Y} - \bar{Y}) + U_i \quad \text{(X)}$$

Now square both sides

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + 2(\hat{Y}_i - \bar{Y})U_i + U_i^2 \quad \text{(X)}$$

and sum over the the *N* individuals,

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})U_i + \sum_{i=1}^{N}U_i^2 \quad \text{(X)}$$

One can demonstrate (though we shall not do so here) that the residuals are uncorrelated with the predicted *y*. Hence

$$\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})U_i = 0 \quad \text{(X)}$$

and Equation 24 becomes

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}U_i^2 \quad \text{(X)}$$

or

$$SS_t = SS_r + SS_e \quad \text{(X)}$$

The squared multiple correlation is the sum of squares for regression divided by the total sum of squares, or

$$R^2 = \frac{SS_r}{SS_r + SS_e} \quad \text{(X)}$$

Another formula for the squared multiple correlation may be written in terms of the b_is and the covariances among the *IV*s,

$$R^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}b_i b_j cov(X_i, X_j)}{\sigma_y^2}$$

An analogous expression may be written for the standardized case:

$$R^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\beta_i \beta_j r(X_i, X_j).$$

Yet another way to look at the multiple correlation is the correlation between observed *Y* and predicted *Y*, or $R^2 = (Y, \hat{Y})$.

Adjusted R^2

Any two data points may be connected by a straight line. Hence, if there are only two observations for *X* and *Y*, the two variables will be perfectly correlated, even if the population correlation between the two variables is 0.0. In general, it turns out that any (*p* + 1) data points can be perfectly explained by a *p* dimensional surface. The number of predictors in a regression equation corresponds to the dimensionality of a linear problem and the number of subjects corresponds to the number of data points. Hence, if *N*, the number of subjects is, say, 5, their score on one variable can be perfectly predicted by regressing that variable on 4 other variables, even if the population correlation between these four *IV*s and the *DV* is 0.0. Obviously this introduces a bias into regression analysis. The effect of this bias depends upon the ratio of the number of subjects to the number of variables. When this ratio is small, *R* tends to be increasingly biased. As this ratio becomes larger, the bias tends to become smaller.

To avoid this bias, most computer packages also calculate an *adjusted* R^2. To see how this adjustment works, write the squared multiple correlation as

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

where σ_e^2 is the variance due to error. With *N* subjects and *p* predictors, an unbiased estimate of σ_e^2 is the sum of squares for error divided by its degrees of freedom, or

$$\hat{\sigma}_e^2 = \frac{SS_e}{N - p - 1}$$

Similarly, $SS_y/(N - 1)$ provided an unbiased estimate of σ_Y^2. Substitute both of these quantities into Equation (31). Now the sample squared multiple correlation may be written as $R^2 = 1 - SS_e/SS_Y$, so $SS_e = (1 - R^2)SS_Y$. Substituting this expression for SS_e into the equation gives an adjusted R^2:

$$R^2_{adj} = 1 - \frac{N - 1}{N - p - 1}(1 - R^2)$$

A disadvantage with this adjustment is that it may yield negative R^2. In such cases, it is recommended that the reported R^2 be 0.

Testing for Significance

There are surprisingly few assumptions required for fitting a regression model to data. About the only mathematical restrictions are that the elements of vector **b** be free and that the observations (i.e., the rows in matrix **X**) be independent. Another restriction usually applied is that no single *IV* be a perfect linear combination of the other *IV*s. If one variable is a perfect linear combination of the other variables then the matrix **X'X** is singular and cannot be inverted, preventing use of equation (11) to solve for the parameter estimates.

However, there are several important assumptions required to test the *significance* of a regression model. Primary among them are two major assumptions about the residuals. First, for every value of Y, the residuals, $Y - \hat{Y}$, are assumed to be normally distributed. Second is the assumption of homoscedasticity. That is, for every value of Y, it is assumed that the residuals have a constant variance, σ_e^2.

A second set of assumptions involved the elements in matrix **X**. It is assumed that these elements are *fixed* or not subject to sampling fluctuation, as if one were performing an experiment and fixing the values in **X**. It is also assumed that these elements are not subject to measurement error. Because of these assumptions, the elements of **X** may be treated as constants from one replication of the study to another replication.

When these assumptions are met, then an F test is usually employed to test the whole regression model. That is, if the true population values for all the b_is is 0, then the mean squares for the regression divided by the mean square error will be distributed as an F statistic:

$$F = \frac{MS_r}{MS_e} = \frac{\frac{SS_r}{p}}{\frac{SS_e}{N - p - 1}}$$

with p degrees of freedom for the numerator and $N - p - 1$ degrees of freedom for the denominator. The F statistic may also be written as a function of R^2,

$$F = \frac{R^2(N - p - 1)}{(1 - R^2)p}$$

Note that this is a test for the regression model *as a whole* and is not a test of its individual components.

In addition to assessing the adequacy of the regression model, it is customary to test each regression coefficient against its standard error. The formula for the standard errors of the b_is was developed above and need not concern us here. What is important is that if the assumptions for significance testing are met and, in addition, if the true population value of b_i is 0, then b_i divided by its standard error will be distributed as a t with $N - p - 1$ degrees of freedom.

Simultaneous, Hierarchical, and Stepwise Regression

In the formulations above, it was assumed that a *DV* is regressed upon a set of p variables and interest is only in that set of p variables. This may be termed *simultaneous* multiple regression. There are cases in which one wants to compare a set of q IVs with a subset or superset of p variables, $p \neq q$, to test such problems as whether one gets a significant increment in prediction by adding several variables. A second major reason for interest in sub- or supersets of variables is in statistical control. Suppose, for example, one wanted to predict WAIS performance IQ in the elderly as a function of social support systems. It might be that age is a confounding factor that influences IQ. One strategy would be to regress performance IQ on age for the sample, output the residuals, and then correlated those residuals with the index of social support. Another and equivalent way is to regress WAIS IQ on age and social support but enter age into the equation first. This is

an example of *hierarchical* multiple regression in which one set of variables is first entered into the equation (age, in this case) to statistically control for their predictive effect before another set of variables (social support) is assessed.

To examine hierarchial regression, assume that we are interested in controlling for m variables (termed the "control" variables) and testing for the predictive effects of q variables (termed the "interest" variables), so that the total number of *IV*s in the regression equation is $p = m + q$. If we were to enter all p variables into the regression equation, we would be interested in the set of m variables, controlling for the set of q variables, and in the set of q variables, controlling for the set of m variables. But that is not what we hypothesize. We want to control for the set of m variables first and then examine the effect of the q variables. In a sense, we want to perform two regressions. The first regressing y on the set of m control variables. The second, to test the predictive effects of the q variables on the residuals from the first regression.

In hierarchical analysis the regression sum of squares for y is decomposed into three additive parts: (1) SS_m or the sum of squares due to the m control variables; (2) $SS_{q|m}$ or the sum of squares due to the q "interest" variables given the m control variables; (3) SS_e or sum of squares error. Thus,

$$SS_y = SS_m + SS_{q|m} + SS_e$$

It is important to note that $SS_{q|m}$ is *not* equivalant to the sum of squares due to the q variables that would be calculated if all p variables were entered into the equation simultaneously

Similarly, we may decompose the R^2 for the total set of p variables into two parts. The first part due to the m control variables and the second part due to the q "interest" variables controlling for the m variables, or

$$R^2_p = R^2_m + R^2_{q|m} \tag{1}$$

The term $R_{q|m}{}^2$ is referred to as the *increment* in R^2 accounted for by the addition of the q interest variables into the equation.

The main interest in hierarchical multiple regression is in determining whether the q "interest" variables add significant predictability of y after the predictability of the m control variables has been taken into account. To do this one may construct an F *increment* or an F statistic that tests the increment in predictability given by adding the q "interest" variables to the equation:

$$F_{inc} = \frac{R^2_p - R^2_m}{1 - R^2_p} \bullet \frac{N - m - q - 1}{q} \tag{1}$$

where $R_p{}^2$ is the R^2 for all p variables and $R^2{}_m$ is the R^2 for the m control variables. The degrees of freedom for the numerator of the F ratio equals q and the degrees of freedom for the demonimator equals $(N - m - q - 1)$.

In hierarchical regression, variables are entered into the regression equation in the order of interest, theory, or hypothesis. *Stepwise* regression, in contrast, enters the variables in order of their *statistical predictability*. There are three modes to stepwise regression, *forward selection*, *backward elimination*, and *pure* stepwise, and the three do not necessarily agree all the time. In forward stepwise regression, the *IV* with the highest correltion with y is first entered into the equation. The second entry is the variable with the highest semipartial correlation given the first variable. Of all the $(p - 1)$ variables remaining, this is the variable that would generate the highest R^2 if it were entered. Of all

the (p - 2) remainging variables, the one that would result in the highest R^2 is entered next, and so on. Hence, all p variables may be ordered in terms of their predictability. Alternately, the order may be stopped at some arbitrary criterion, frequently whether the probability of F_{inc} is greater than .05.

In backward stepwise regression, the order is reversed. The first step is the entry of all p variables. Then the one responsible for the smallest increment in R^2 is dropped. The (p - 1) variables are entered and the one giving the smallest R^2 increment is dropped, etc.

Pure stepwise regression combines elements of both forward selection and backward elimination. The procedure is similar to forward selection but differs from it in that at each stage of entry, variables are assessed for their contribution to R^2. If a variable does not make a significant contribution, then it is dropped from the equation. In this way, a variable may be entered on, say, step 1 but dropped on step 5.

The aim of stepwise regression is to obtain an approximation to the best linear combination of *IV*s that explain the variance of the *DV*. The method is essentially atheoretical. It is useful only when one requires predictability irrespective of justification of theory.

Diagnostics

There are three major areas used to diagnose the adequacy of a multiple regression model. The first consists of the magnitude of the correlation among the *IV*s. The second is an examination of the residuals. The third is the analysis of influential data points. We deal with each in turn.

As mentioned above, when one *IV* is a perfect linear combination of the other *IV*s, the matrix $\mathbf{X'X}$ does not have a unique inverse and the parameters cannot be estimated. This is, in effect, the limiting (i.e., most extreme) case of a condition known as *multicollinearity*. As the multiple correlation between any one *IV* and the remaining *IV*s approaches 1.0, multicollinearity becomes a problem.

Multicollinearilty affects the *interpretation* of the b_is, not the magnitude of R^2. As *IV*s become increasingly correlated the diagonal elements of $(\mathbf{X'X})^{-1}$ become larger, increasing the standard errors of the parameter estimates (see Equation (14)). Hence, the confidence interval of the b_is increases, making it more difficult to test for significance.

In addition to large standard errors for the parameters, problems with multicollinearity may be signalled by the statistic *tolerance* which many computer programs output for each *IV* . Tolerance for the ith *IV* is (1 - R_i^2) where R_i^2 is the squared muliple correlation of the *i*th variable regressed upon the (p - 1) remaining *IV*s. Other statisticians prefer to interpret the *VIF* or *variance inflation factor* which is the reciprocal of tolerance. Low tolerances or high *VIF*s are symptomatic of multicollinearity problems.

What does one do when faced with multicollinearity? One easy solution is simply to drop the offending variable from the equation. More advanced statistical techniques such as ridge regression or regression on principal components may also be employed.

A second major area for diagnostics is an analysis of the residuals. Because assumptions about the residuals underly the tests for statistical significance, many statisticians suggest that a thorough analysis of the residuals accompany a regression.

Many computer packages output plots of the residuals as a function of the predicted *y* so that one may assess the assumptions of normality of the residuals, linearity, and homoscedasticity.

In addition, several packages have the option of printing out each case with its residual and several different statistcs that identify two problems: (1) those that are *outliers*; and (2) *influential datapoints*, i.e., those observations that have an inordinate influence on the regression results.

In bivariate regression, outliers may be readily identified by examination of the plot of *x* and *y*. The situation is more complicated with more than a single *IV*.There are several different measures for outliers. One set of indices test the extent to which an observation differs from the center of the distribution on the *p IV*s. One of these indices is the *Mahalanobis distance* which seeks to find a multivairate outlier. For the *i*th observation, the Mahalanobis distance is defined as

$$D_i = (\mathrm{x}_i - \bar{\mathrm{x}})' S^{-1} (\mathrm{x}_i - \bar{\mathrm{x}}) \quad (1)$$

where $\mathbf{x}_i$ is the vector of values for the *i*th observation on the *IV*s, **x** is the vector of means, and **S** is the unbiased estimate of the covariance matrix. Sometimes this value is expressed in terms of *leverage*:

$$h_i = \frac{1}{N} + (N - 1) D_i \quad (1)$$

Both the Mahalanobis distance and leverage index the extent to which the *i*th observation differs from the center of the distribution on the **x** vector. It is also possible to be an outlier in terms of the residuals. It is customary to standardize the residuals by dividing the residual by its standard error to aid inspection of their values. The standard error for the residual on the *i*th observation may be defined as

$$se(u_i) \quad (1)$$

where h_i is the leverage for the *i*th subject as defined by Equation (40). The standard error may be estimated in one of two ways. First, the square root of the mean square error from the regression model may be used. In this case the residual divided by the estimated standard error is termed an *internalized studentized residual*. The second estimate of σ_e deletes the *i*th observation from its calculation. The residual divided by this standard error is called an *externalized studentized residual*. Unfortumately, there is some confusion in the literature and in software packages over what constitutes a "studentized residual," either the internalized or esternalized residual, but the difference between the two is generally small when *N* is relatively large.

An outlier need not have an undue influence on the regression coefficients or R^2. Cook's distance (*d*) is generally used to identify an observation that may be highly influential in a regression analysis. Cook's *d* for the *i*th observation is defined as

$$d_i = \frac{\sum_{j=1}^{N} (\hat{y}_j^{(-i)} - \hat{y}_j)^2}{(p + 1)\sigma_e^2} \quad (1)$$

where $y_j^{(-i)}$ is the predicted value for the *j*th observation if the *i*th observation has been omitted from the regression analysis. A Cook's distance of 1 or greater is generally regarded as large. Cook's distance may also be written as a function of both the leverage and the internalized studentized residual. Hence, an observation may have a large Cook's

distance by having a large leverage (i.e., being an outlier in terms of the *IV*s) or a large residual (being an outlier in terms of the regression model).