CHAPTER 19

The Coefficient of Determination in Multiple Regression

In the case of simple regression analysis, the coefficient of determination measures the proportion of the variance in the dependent variable explained by the independent variable. This coefficient is computed using either the variance of the errors of prediction or the variance of the predicted values in relation to the variance of the observed values on the dependent variable as follows:

$$r_{yx}^2 = \left(\frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)} \right) = 1 - \left(\frac{\mathrm{Var}(e)}{\mathrm{Var}(y)} \right)$$

This equation for the coefficient of determination in simple regression analysis can easily be extended to the case of multiple regression analysis. The variances of the predicted values and the errors of prediction in simple regression have direct counterparts in multiple regression. In the case of two independent variables, for example, the following equations obtain:

$$\hat{\mathbf{y}} = \mathbf{u}a + \mathbf{x}_1 b_{y1.2} + \mathbf{x}_2 b_{y2.1}$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

In short, the addition of independent variables to the regression model does not affect the equations for computing either the predicted values or the errors of prediction.

Moreover, the fundamental relationship between the variance of the dependent variable, y, the variance of the predicted values, $\hat{y}$, and the variance of the errors of prediction, e, remains the same, such that:

$$\mathrm{Var}(y) = \mathrm{Var}(\hat{y}) + \mathrm{Var}(e)$$

Therefore, the coefficient of determination in multiple regression analysis has exactly the same definition as it does in simple regression analysis, such that:

$$R^2_{y.12} = \left(\frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}\right) = 1 - \left(\frac{\mathrm{Var}(e)}{\mathrm{Var}(y)}\right)$$

In other words, the interpretation of the coefficient of determination remains the same regardless of how many variables there are in the regression equation.

It will be noticed that the notation used to denote the coefficient of determination is different in the case of multiple regression. The lower case r is replaced with an upper case R and the subscript contains a dot that separates the dependent variable from the independent variables. For example, $R^2_{y.12}$ refers to the proportion of the variance in the dependent variable, y, explained by two independent variables, x_1 and x_2. It will be recalled that the coefficient of determination is an important goodness-of-fit statistic. As such, it measures how well a regression model "fits" the data. It ranges from zero, when there is no relationship between the dependent variable and a linear function of the independent variables, to one, when there is an exact relationship between the dependent variable and a linear function of the independent variables. This goodness-of-fit statistic is very useful because it enables us to compare different regression equations in terms of how well each of them fits the data. Other things being equal, we will normally prefer the regression model with the largest coefficient of determination.

The computations required to obtain the coefficient of determination in a multiple regression model can be demonstrated using the example of the regression of public expenditures, y, on economic openness, x_1, and labor organization, x_2, where the equation for the predicted values of the dependent variable is given by:

$$\hat{y} = \mathbf{u}\,(25.05) + \mathbf{x_1}\,(0.452) + \mathbf{x_2}\,(0.295)$$