6

Multiple Regression Model

In the preceding chapter we have explore d the simple linear regression model in which the dependent variable was a linear function of one independent variable and developed techniques with which to estimate the values of β0 and β_1 and to test hypothesis about the relationship $Y = \beta_0 + \beta_1 X_1 + u$. Simple linear regression models assume that the dependent variable is influenced by only one systematic variable and the random error term (u). In reality, there may be several independent variables that influence the dependent variable. In such cases, it may be worthwhile to formulate a model that allows us to consider the relation of our variable of interest with a set of independent variables. When several independent variables are included in the model, the estimation technique is called multiple regression. In this chapter we extend our analysis to deal with more than one independent variable.

Recall the Keynesian consumption function $C = \beta_0 + \beta_1 Y_d + u$, which says that consumption spending depends only on the level of disposable income; the higher is a family's level of disposable income, the higher should be its level of spending on consumption. However, in economic reality each variable is influenced by a very large number of factors. For instance, the level of consumption expenditure depends not only on current disposable income but also on a host of other factors such as wealth of the population (W) Liquid assets (L), interest rates(R), the age structure of the population (A), distribution of income (D), past incomes (PY), expected future incomes (FY), the level of advertisement expenditure in the economy (LA), previous consumption (PC) and so on.

Note that in the above equation consumption depends not only on a single variable, Y_d but also on the values of a whole set of independent variables. On the basis of this information we now have a more realistic consumption function as follows.

$$Ct = \beta_0 + \beta_1 Y_d + \beta_2 W + \beta_3 L + \beta_4 R + \beta_5 A + \beta_6 D + \beta_7 PY + \beta_8 FY + \beta_9 LA + \beta_{10} PC + u \quad \ldots(6.1)$$

A more realistic formulation of the model requires specification of several

variables in each relation. However, not all the factors influencing a certain variable can be included for various reasons [see omission of variables from the function]. Moreover one must remember that each variable that is included in the model tends to make the computation more tedious but at the same time omission of important variable makes analysis less accurate. The method of multiple regression is an extension of simple regression model to the case of two or more explanatory variables or predictors. In general terms, we consider the following multiple regression model.

The population multiple regression model with k independent variable is expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_k X_k + u.$$

Where β_i ,s are the regression coefficients; Y is the dependent variable, the variable we wish to predict; X_is are the independent variables; and u is the error term.

The parameters β_i, like most parameters, will remain unknown and can be estimated only with sample data. linear model based on sample data is expressed as

$$\mathbf{Y = b_0 + b_1 X_1 b_2 X_2 + b_3 X_3 + \ldots\ldots + b_k X_k + e}$$

Where b_i are the estimates of b_i and e is the random term. Customarily referred to as the residuals when using sample data. However, since e is random, Y can only be estimated.

The estimated multiple linear regression model using sample data takes the form.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ,\ldots . + b_k X_k$$

Where $\hat{Y}$ is the estimated value for the dependent variable and bi are the estimates for the population coefficients b_i. The bi are called the partial (or net) regression co–efficients and carry the same interpretation as in simple regression. Thus, b_i is the amount by which Y_i will change if X_i changes by one unit, assuming all other independent variables are held constant. This assumption was not necessary under simple regression because there were no other independent variables to hold constant.

Multiple regression involves the same assumptions cited in previous chapters for simple regression, plus two others. The first assumption requires that the number of observations, n exceed the number of independent variables, k by at least 2. In multiple regression there are k + 1 parameters to be estimated: coefficient for the k independent variables plus the intercept term. Therefore, the degrees of freedom associated with the model are df = n – (k + 1). If we retain

even one degree of freedom, n must exceed k by at least 2, so that n–(k + 1) is at least 1. The second assumption involves the relationship between the independent variables. It requires none of the independent variables is a linear combination of others. For example,

$$X_1 = X_2 + X_3, \textbf{ (or) } X_1 = 0.5\, X_2 \textbf{ or}$$
$$X_1 = 3 - 2X_2 + 17\, X_3 \textbf{ (or) } X_4 = (X_1 + X_2 + X_3)/3,$$

then a linear relationship would exist between two or more independent variables and a serious problems would arise in the estimation. This problem is known as multicollinearity. Thus, multicollinearity exists if two or more of the independent variables are linearly related. Multicollinearity may cause the algebraic signs of the coefficients to be the opposite of what logic may dictate, while greatly increasing the standard error of the coefficients. A more through discussion of multicollinearity follows later (chapter–**12**).

Assumptions: To complete the specification of our multiple linear regression model. We need some assumptions about the random variable u. Multiple regression model is essentially very similar to the two variable model. The assumption are the same as in the two variable case (for details, see assumption of OLS method). It describes a linear functional relationship by which the values of a set of independent variables determine the value of a dependent variable. We must develop a method to estimate the values of parameters in this relationship.

Model with Two Explanatory Variables

Suppose we are using to estimate of regression equation in which the dependent variable Y is linearly related to two explanatory variables X_1 and X_2;

The population regression model for three variable case (two independent variables) can be expressed as

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u}$$

where β_0, β_1 and β_2 are the population parameters and u is the deviation of the observed value of Y from the one predicted by the true equation.

The estimated multiple regression model is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

where $\hat{Y}$ is the predicted value of Y, the value lying on the estimated regression surface. The terms b_i, i = 0, 1 and 2, of the least squares estimates of the population regression parameters β_is.

Then from the above regression we obtain the least squares estimates of population parameters β_0, β_1, and β_2, which generate the predicted values,

$\hat{Y}$. We then generate the residual, e, as the deviation between the observed Y and predicted.

where $Y: e = (Y - \hat{Y})$,

$Y = b_0 + b_1 X_1 + b_2 X_2$ and $Y = b_0 + b_1 X_1 + b_2 X_2 + e$

$$\Sigma e^2 = \Sigma (Y_i - \hat{Y}_i)^2 = \Sigma (Y_i - b_0 - b_1 X_1 - b_2 X_2)^2$$

Where bis are the regression co–efficients and u is the error term.

Suppose we are given a sample of n observation, assuming that the level of consumption expenditure Y depends not only on disposable income X_1 but also on the family size X_2. Hence we try to determine the relationship between consumption expenditure as the dependent variable on the one hand, disposable income and family size as the independent variables on the other.

Y	X_1	X_2
Y_1	X_{11}	X_{21}
Y_2	X_{12}	X_{22}
Y_3	X_{13}	X_{23}
Y_n	X_{1n}	X_{2n}

Given n observations on Y, X_1 and X_2 our problem is to estimate β_0, β_1 and β_2 from the sample data. As in the case of simple regression, the values of the co–efficients b_0, b_1, and b_2 or the estimates of the population parameters β_0, β_1 and β_2 are obtained by minimizing the sum of the squared errors: We then generate e as follows

$$e_i = \hat{Y}_i - Y_i = Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}$$

We have to minimize the function

$$\sum_{i=1}^{n} e_i^2 = \Sigma [Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}]^2 \quad \text{....(6.2)}$$

A necessary condition for this expression to assume a minimum value is that its partial derivatives with respect b_0, b_1 and b_2 be equal to zero.

Partial derivatives with respect to b_0

$$\frac{\partial [\Sigma e_i^2]}{\partial b_0} = \frac{\partial [(\Sigma Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})]^2}{\partial b_0} = 0$$

$$= 2 \Sigma (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}) (-1) = 0$$

$$= -2 \Sigma (Y_i - b_0 - b_1 X_{1i} - b_1 X_{2i}) = 0 \quad \text{.....(6.3)}$$

Partial derivative with respect to b_1

$$\frac{\partial [\Sigma e_j^2]}{\partial b_1} = \frac{\partial [(\Sigma Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})]^2}{\partial b_1}$$

$$= 2\Sigma(Y_i - b_0 - b_1X_{1i} - b_2X_{2i})(-X_{1i}) = 0$$
$$= -2\Sigma(Y_i - b_0 - b_1X_{1i} - b_2X_{2i}) = 0 \quad \text{.....(6.4)}$$

Partial derivatives with respect to b_2

$$\frac{\partial\left[\Sigma e_i^2\right]}{\partial b_2} = \frac{\partial\left[(\Sigma Y_i - b_0 - b_1X_{1i} - b_2X_{2i})\right]^2}{\partial b_2}$$

$$= 2\Sigma(Y_i - b_0 - b_1X_{1i} - b_2X_{2i})(-X_{1i}) = 0$$
$$= -2\Sigma X_{2i}(Y_i - b_0 - b_1X_{1i} - b_2X_{2i}) = 0 \quad \text{....(6.5)}$$

Multiplying both sides of each equation (6.3, 6.4 and 6.5) by – ½ and simplifying, we obtain

$$= \Sigma(Y_i - b_0 - b_1X_{1i} - b_2X_{2i}) = 0$$
$$= \Sigma X_{1i}(Y_i - b_0 - b_1X_{1i} - b_2X_{2i}) = 0$$
$$= \Sigma X_{2i}(Y_i - b_0 - b_1X_{1i} - b_2X_{2i}) = 0$$

We now apply the Σ operator to the terms inside the brackets to give.

$$= \Sigma Y_i - n\,b_0 - b_1\Sigma X_{1i} - b_2\Sigma X_{2i} = 0 \quad \text{....(6.6)}$$
$$= \Sigma X_{1i}Y_i - b_0\Sigma X_{1i} - b_1\Sigma X_{1i}^2 - b_2\Sigma X_{1i}X_{2i} = 0 \quad \text{....(6.7)}$$
$$= \Sigma X_{2i}Y_i - b_0\Sigma X2_1 - b_1X_{1i}Y_{2i} - b_2\Sigma X_{2i}^2 = 0 \quad \text{....(6.8)}$$

On rearranging these equations (6.6 , 6.7,and 6.8), we obtain the following three normal equations of the least squares method.

$$\Sigma Y_i = nb_0 + b_1\Sigma X_{1i} + b_2\Sigma X_{2i} \quad \text{....(6.9)}$$
$$\Sigma X_{1i}Y_i = b_0\Sigma X_{1i} + b_1\Sigma X_{1i}^2 - b_2\Sigma X_{1i}X_{2i} \quad \text{.....(6.10)}$$
$$\Sigma X_{2i}Y_i = b_0\Sigma X_{2i} + b_1X_{1i}X_{2i} - b_2\Sigma X_{2i}^2 \quad \text{....(6.11)}$$

The three resulting equations are the normal equations for a regression with the two explanatory variables. The sums and products associated with b_0, b_1 and b_2 in the normal equation can be computed from sample observations (data). Once the sums and products have been computed, the three equations can be solved for b_0, b_1 and b_2.

Deviation Method

The following formulae, in which the variables are expressed in deviation from their mean, may also be used for obtaining the value of the coefficients b_0, b_1 and b_2.

Solving (6.9) for b_0 gives

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 \quad \text{.....(6.12)}$$

and substituting this expression for b_0 in to equations 6.10 and 6.11 yields

$$\Sigma X_{1i}Y_i = (\overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2)\Sigma X_{1i} + b_1\Sigma X^2{}_{1i} + b_2\Sigma X_{1i}X_{2i}$$
$$\Sigma X_{2i}Y_i = (\overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2)\Sigma X_{2i} + b_1\Sigma X_{1i}X_{2i} + b_2\Sigma X^2{}_{2i}$$

Rearranging the terms and using the result that

$$\Sigma X^2_i - \overline{X}\ \Sigma X_i = \Sigma (X_i - \overline{X})^2$$

gives [for details ,see simple linear regression equation]

$$\Sigma (X_{1i} - \overline{X}_1)(Yi - \overline{Y}) = b_1 \Sigma (X_{1i} - \overline{X}_1)^2 + b_2 \Sigma (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)$$

$$\Sigma (X_{2i} - \overline{X}_2)(Yi - \overline{Y}) = b_1 \Sigma (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) + b_2 \Sigma (X_{2i} - \overline{X}_2)^2$$

or using lower case letters to denote deviations from sample means

$$\Sigma x_1 y = b_1 \Sigma x^2_1 + b_2 \Sigma x_1 x_2 \quad \ldots(6.13)$$

$$\Sigma x_2 y = b_1 \Sigma x_1 x_2 + b_2 \Sigma x^2_2 \quad \ldots(6.14)$$

Solving 6.13 and 6.14 for b_1 and, b_2 we obtain

$$b_1 = \frac{(\Sigma x_{1i} y_i)(\Sigma x^2_{2i}) - (\Sigma x_{2i} y_i)(\Sigma x_{1i} x_{2i})}{(\Sigma x^2_{1i})(\Sigma x^2_{2i}) - (\Sigma x_{1i} x_{2i})^2} \quad \ldots(6.15)$$

$$b_2 = \frac{(\Sigma x_{2i} y_i)(\Sigma x^2_{1i}) - (\Sigma x_{1i} y_i)(\Sigma x_{1i} x_{2i})}{(\Sigma x^2_{1i})(\Sigma x^2_{2i}) - (\Sigma x_{1i} x_{2i})^2} \quad \ldots(6.16)$$

Statistical Properties of the Least Squares Estimates Multiple Linear Regression Model

In order to investigate the statistical properties of the least square estimates (given in (6.12), (6.15), and (6.16) of the multiple regression model it is necessary to make some assumptions concerning the random errors, u_i in equation, 6.17.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad \ldots(6.17)$$

The assumptions made here will be the same as in the simple linear regression model developed in the last chapter, with one modification. As before, we shall assume that the expected value of the u_i 's is zero, that they are drawn than a population with variance σ^2, and that they are uncorrelated among themselves. that is,

$$Eu_i = 0; i = 1,\ldots..,n \quad \ldots(6.18)$$

$$E(u_i^2) = \sigma^2$$

$$E(u_i\ u_j) = 0, \text{ when } i \neq j.$$

In multiple linear regression, the assumption of simple linear regression model, $E(X_i u_j) = X_i E(u_j) = 0$, for all $i,j = 1,\ldots\ldots\ldots\ldots n$) has to be modified, since we now have two independent variasses, X_1 and X_2. Our new assumption is that both X_{1i} and X_{2i} may be treated as constants, so that

$$E(X_{1i}\ u_j) = E(X_{2i} u_j) = 0 \quad \text{for all i,j}$$

This condition is automatically fulfilled if we assume that the values of the X's are a set of fixed numbers in all hypothetical samples.

By virtue of $E(u_i) = 0$, $i = 1,\ldots\ldots n$) in simple linear regression.

On the basis of these assumptions, let us show that the method of least squares yields estimates, which are best linear and unbiased.

$$= \frac{}{4600-1600}$$

$$= \frac{1830}{3000}$$

$$b_1 = 0.61$$

$$b_2 = \frac{(\Sigma x_2 y)(\Sigma x_1^2) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)}$$

$$= \frac{(40)(100) - (5)(-40)}{(100)(46) - (40)^2}$$

$$= \frac{4000 - (-200)}{4600 - 1600}$$

$$= \frac{4200}{3000}$$

$$b_2 = 1.4$$

$$b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$
$$= 15 - 0.61(5) - 1.4(5)$$
$$= 15 - 3.05 - 7$$
$$= 15 - 10.05$$
$$= 4.95$$

[Hereafter we use this formula to compute b0]The coefficients obtained from the deviation method are similar to the above one.Hence the estimated equation is:

$$Y = 4.95 + 0.61 X_1 + 1.4 X_2$$

Hereafter we use the deviation method for computing the values of b_0 b_1 and b_3.

A Measure of Goodness of Fit – Multiple Linear Regression Model

In the last chapter, we have defined r^2 and this measure extends easily to the multiple regression models. We have already seen that the coefficient of determination r^2, measures the proportion of variation in the dependent variable (Y) explained by variations in independent variable. For a multiple regression, the coefficient of multiple determination, R^2, measures the proportion of variation in the dependent variable (Y) which is explained by variation in two or more independent variables. It is computed as the ratio of the explained variation to the total variation about Y:

We define coefficient of multiple determination (R^2) as.

$$R^2 = \frac{\text{Total variation - unexplained variation}}{\text{Total variation}}$$

$$R^2 = \frac{\Sigma(Y-\bar{Y})^2 - \Sigma(Y-\hat{Y})^2}{\Sigma(Y-\bar{Y})^2} \text{ (or)}$$

$$R^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

$$R^2 = 1 - \frac{\Sigma(Y-\hat{Y})^2}{\Sigma(Y-Y)^2}$$

$$R^2 = \frac{\Sigma e_i^2}{\Sigma y_i^2}$$

$$R^2 = \frac{\Sigma y_i^2 - \Sigma e_i^2}{\Sigma y_i^2} \quad \text{....6.2}$$

we know, $\Sigma e_i^2 = \Sigma(Y_i - b_0 - b_1X_1 - b_2X_2)^2$

$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$

$\Sigma e_i^2 = \Sigma(Y_i - (\bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2) - b_1X_1 - b_2 X_2)^2$,

[Since $b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2$

$= \Sigma(Y_i - \bar{Y} + b_1 \bar{X}_1 + b_2 \bar{X}_2 - b_1X_1 - b_2X_2)^2$

$\Sigma e_i^2 = \Sigma(Y_i - \bar{Y}) - b_1(X_1 - \bar{X}_1) - b2(X_2 - \bar{X}_2)^2$

$= \Sigma(y - b_1 x_{1i} - b_2 x_{2i})^2$

$= \Sigma e_i (y_i - b_1x_{1i} - b_2x_{2i})$

[since $e_i = (y_i - b_1x_{1i} - b_2x_2$

$= \Sigma e_i y_i - b_1 \Sigma e_i x_{1i} - b_2 \Sigma e_i x_{2i}$ .....(6.2

Now $\Sigma e_i x_{1i}$ and $\Sigma e_i x_{2i}$ were set equal to zero in the equation (6.28)

$= \Sigma e_i y_i$

[Since $\Sigma e_i x_{1i} = \Sigma e_i x_{2i} =$

Substituting for e_i in the above equation yields the residual variation, Σ as follows:

$\Sigma e_i^2 = \Sigma y_i (y_i - b_1 x_{1i} - b_2 x_{2i})$ [wher $e_i = y_i - b_1x_{1i} - b_2x_2$

$= \Sigma y_i^2 - b_1 \Sigma x_{1i} y_i - b_2 \Sigma x_{2i} y_i$ .....(6.2

Rearranging the terms in equation (6.29), we obtain

$$\Sigma y_i^2 = b_1 \Sigma x_{1i} + b_2 \Sigma x_{2i} y_i + \Sigma e_i^2$$

Total variation = explained variation + residual variation. We know,

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{Total variation} - \text{residual variation}}{\text{Total variation}}$$

$$R^2 = \frac{\Sigma y_i^2 - (\Sigma y_i^2 - b_1 \Sigma x_{1i} y_i - b_2 \Sigma x_{2i} y_i)}{\Sigma y_i^2}$$

$$= \frac{\Sigma y_i^2 - \Sigma y_i^2 + b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i}{\Sigma y_i^2}$$

$$R^2 = \frac{b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i}{\Sigma y_i^2} \quad \text{....(6.30)}$$

where R^2 called the coefficient of multiple determination. As with simple linear regression, the value of R^2 lies between zero and unity. The closer to unity is R^2 the grater the explanatory power of the estimated regression equation. At the other extreme, where the estimated equation explains less variation of the dependent variable, R^2 is closer to zero.

Adjusted Co–efficient of Multiple Determination

The above formula for R^2 does not take in to account the loss of degrees of freedom from the inclusion of additional explanatory variable in the regression equation. To correct this, a new measure of fit of a multiple linear regression model must be introduced. This measure is appropriately named the Adjusted or Corrected Coefficient of Multiple Determination, and is designated as $\overline{R}^2$.

This new statistic $\overline{R}^2$ takes explicit account of the number of explanatory variables used in the function. It is useful for comparing the fit of specifications that differ in the inclusion or exclusions of explanatory variables. The unadjusted R^2 will never decrease with the addition of any explanatory variable. If the added variable is totally irrelevant the ESS (explained sum of squares) simply remains constant and will usually raise it. That is, if the number of irrelevant explanatory variable is increased, R^2 will never decrease, it will generally increase. The adjusted coefficient does not always increase as new variables are entered in to our regression equation. When $\overline{R}^2$ does increase as a new variable entered in to the regression equation, it may be worthwhile to include the variable in the equation. The adjusted coefficient of multiple determination may decrease with the addition of variable of low explanatory power. As Theil points: It is a good practice of use $\overline{R}^2$ rather than R^2 because R^2 tends to give an overly optimistic

picture of the fit of the regression, particularly when the number of explanatory variables is not very small compared with the number of observations. The formula for $\bar{R}^2$ is.

$$\bar{R}^2 = 1-\left(1-R^2\right)\left(\frac{(n-1)}{(n-k)}\right)$$

(or)

$$\bar{R}^2 = 1-\left(\frac{\Sigma e_i^2 / n-k}{\Sigma y^2 / n-1}\right)$$

Where n is the number of observations, k is the number of parameters estimated from the sample and R^2 is the(un adjusted) coefficient of multiple determination.

Computer output for multiple regression analysis usually includes the adjusted coefficient of multiple determination ($\bar{R}^2$).Let us note the following properties of $\bar{R}^2$.

1. Unless R^2 is equal to 1, $\bar{R}^2$ is always less than R^2.
2. If the number of data points (observations) is relatively large compared to the number of regressors or explanatory variables, R^2 and $\bar{R}^2$ are close to each other in value.
3. If the number of data points is relatively small in relation to the regressors , $\bar{R}^2$ is much smaller than R^2 and can even take negative values, in which case $\bar{R}^2$ should be interpreted as zero ($\bar{R}^2$ can be < 0)
4. If $R^2 = 0$, $\bar{R}^2 = (1-k)/(n-k)$, in which case $\bar{R}^2$ can be negative if $k > 1$.
5. If the change in R^2 is smaller than the percentage change in the degrees of freedom when additional variable is added to the regression equation, $\bar{R}^2$ will fall.

Testing of Significance of Individual Regression Coefficients

Since the hypotheses to be tested are H_0: $\beta_i = 0$ and H_1 : $\beta_i \neq 0$, it may be shown that the t – ratio for each b_i s as follows.

$$t^* = \frac{b_i - \beta_i}{Sb_i} \qquad i = 0,1,2 \qquad \text{....(6.32)}$$

[Which follows the t– distribution with (n–k) degrees of freedom. It is noted that this is a test for the population regression coefficient.]

More specifically,

$$t = \frac{b_0 - \beta_0}{(Sb_0)}$$

$$t = \frac{b_1 - \beta_1}{(Sb_1)} \quad \text{and}$$

$$t = \frac{b_2 - \beta_2}{(Sb_2)}$$

Note that the most important statistical test in linear regression model is the test of whether the slope of the (true) parameter âi is equal to zero. If we assume that the slope of the true but unknown population regression line is zero, there is no relationship between the independent variables and Y. The hypothesis makes statement about the true parameter value, β_i and not the estimate, b_i, which was obtained from the regression equation and is known for certain.

Testing of Individual Coefficients : One tailed test:

Step 1:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i > 0 \text{ or } \beta_i < 0$$

Step 2 : Apply t–statistic:

$$t = \frac{\beta_i - 0}{s.e.(b_i)}; (n-k) d.f.$$

where k is the number or parameter estimates (including the intercapt b_0).

Step 3 : Reject H_0 if the absolute value of 't' is greater than the tabular value of 't'.

$$|t_{cal}| > t_{tab}$$

Testing of Individual Coefficients: Two tailed Test

Step 1:

$$H_o : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Step 2: Construct same t–statistic as one tail test i.e.

$$t = \frac{\beta - 0}{s.e.(b_i)}; (n-k) d.f.$$

Step 3: Reject H_0 if the absolute value of 't' is greater than the tabular value of 't'.

$$|t_{cal}| > t_{tab}$$

Evaluating the Model as a Whole : In the last chapter we developed a technique for the goodness of fit which involved the F distribution. This test may also be extended in the content of multiple linear regression model. ANOVA

procedure (F–test) for goodness of fit of the regression of Y on X_1 and X_2 as shown in Table [6.1]. ANOVA procedure tests whether any of the independent variables has a relationship with the dependent variable. ANOVA procedure tests the null hypothesis that all the β–values are zero against the alternative hypothesis that at least one b is not zero. That is

$$H_0 : \beta_1 = \beta_2 = \beta_3, \ldots, = \beta_k = 0$$

H_1 : At least one β is not zero (or) Not all the β–values are zero.

If the null hypothesis is accepted, there is no linear relationship between Y and any of the independent variables in the regression equation. On the other hand, if the null hypothesis is rejected, there is a linear relationship between Y and at least one of the independent variables. Table 6.1 provides the general format for ANOVA Table for multiple linear regression.

TABLE 6.1. ANALYSIS OF VARIANCE TABLE FOR MULTIPLE LINEAR REGRESSION

Source of Variation	Sum of Squares	Degrees of Freedom	Mean sum of Squares	F* ratio
Due to Regression	$\Sigma\hat{y}_i^2$	k–1	ESS/k–1 $= S_1^2$	$\frac{S_2^2}{S_2^2} = \frac{(ESS/k-1)}{RSS/(n-k)}$
Due to Residual/Error	RSS Σe_i^2	n–k	RSS/(n–k) $= S_2^2$	
Total	TSS Σy_i^2	n–1		

Let as recall the analysis variance technique to test the goodness of fit in the simple linear regression analysis. In this case we may set up our test for the goodness of fit of the regression of Y on X_1 and X_2 as shown in the Table [6.1]

$$\Sigma e_i^2 = \Sigma y_i^2 - b_1 \Sigma X_{1i} y_i - b_2 \Sigma x_{2i} \quad \text{....(6.33)}$$

[Refer equation 6.29]

Rearranging the equation, we have

$$\Sigma y_i^2 = b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i + \Sigma e_i^2$$

Total Variation = Explained variation + Residual variation

The above equation may be expressed in several other forms. For instance

$$\Sigma y_i^2 = \Sigma (b_1 x_{1i} + b_2 x_{2i})^2 + \Sigma e_i^2$$

Notice that the degrees of freedom for ESS is equal to k–1, where k is the number of parameter estimates includng the intercept b_o, while the degree of freedom for RSS is n–k. The F–test follows the F distribution with k–1 (number of explanatory variables) and n–k degrees of freedom. This forms the basis for testing $R^2 = 0$.

TABLE 6.2 . ANOVA TABLE FOR THE REGRESSION Y ON X_1 AND X_2

Source of Variation	Sum of Squares	Degrees of Freedom	Mean sum of Squares	F* ratio
Regression	$b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i$ (or) $= \Sigma (b_1 x_{1i} + b_2 x_{2i})^2$ (ESS)	k–1 = 2	$(b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i)/k-1$ (or) $\Sigma (b_1 x_{1i} + b_2 x_{2i})^2 / k-1$	$F = \frac{S_1^2}{S_2^2}$
Error	Σe_i^2 (RSS)	n–k	$\Sigma e_i^2 / n-k = S_2^2$	
Total	Σy_i^2 (TSS)	n–1		

The F–ratio may be related to R^2 By the following modifications, we know,

$$F = \frac{S_1^2}{S_2^2} \qquad \frac{b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i / k-1}{\Sigma e_i^2 / n-k} \qquad \text{.....(6.34)}$$

$$= \frac{(n-k) b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i}{\Sigma e_i^2 (k-1)} \qquad \text{.....(6.35)}$$

Co–efficient of Multiple Determination and F Value

$$F = \frac{(n-k) b_1 \Sigma x_{1i} y_i + b_2 \Sigma x_{2i} y_i}{\Sigma e_i^2 (k-1)} \qquad \text{.....(6.36)}$$

We know $R^2 = \frac{\Sigma \hat{y}^2}{\Sigma y^2} = \frac{\Sigma (\hat{Y} - \overline{Y})^2}{\Sigma (Y - \overline{Y})^2} = \frac{\Sigma y_i^2 - \Sigma e_i^2}{\Sigma y_i^2}$

$$= 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2}$$

$$= \frac{\Sigma y_i^2 - \Sigma e_i^2}{\Sigma y_i^2} \qquad \text{....(6.37)}$$

$$R^2 \Sigma y_i^2 = \Sigma y_i^2 - \Sigma e_i^2$$

$$\Sigma e_i^2 = \Sigma y_i^2 - R^2 \Sigma y_i^2$$

$$= (1-R^2) \Sigma y_i^2$$

substituting equation [6.38] in equation(6.36), we have the F–value as

$$F = \frac{(n-k)b_1\Sigma x_{1i}y_i + b_2\Sigma x_{2i}y_i}{k-1(1-R^2)\Sigma y^2}$$

$$= \frac{(n-k)R^2}{k-1(1-R^2)} \text{.....Since } R^2 = \frac{b_1\Sigma x_{1i}y_i + b_2\Sigma x_{2i}y_i}{\Sigma y^2} \quad \text{...(6.}$$

$$F = \frac{(n-k)R^2}{k-1(1-R^2)} \quad \text{...(6.}$$

Where k is the number of independent variables and n is the number observations. The F–ratio must be sufficiently larger to reject the null hypothesis that $H_0 : \beta_1 = \beta_2 = \beta_3, \ldots, = \beta_k = 0$; that is, Y is unrelated to X_1 and X_2. A high calculated F F* – Value, suggests that the explained variation is larger relative to the unexplained variation of the dependent variable. In other words, a high F–test signals that the model possesses significant explanatory power.

Let us consider the following empirical examples to illustrate the method of obtaining the multiple linear regression equation, coefficient of multiple determination adjusted coefficient of multiple determinations,' t' test and F-test

EMPIRICAL ILLUSTRATIONS

Example 6.2: The manager of a supermarket would like to determine the relationship between the quantity demanded Y and two independent variables X_1 and X_2, where X_1 represents the price of Y and X_2 represents the income of the customers . Ten customers were selected at random and the values of the variables are presented in the following table.

(a) Given the data below , estimate the parameters of the demand function by least squares method.

(b) Evaluate the results on the basis of 'a priori' criteria and calculate the efficient of multiple determination

(c) Test the significance of the regression coefficients at the 5 percent level of significance.

(d) Find the confidence limits of the regression coefficients for a confidence coefficient at 95 percent.

(e) Is the overall regression significant at the five percent level?