CHAPTER 6

MULTIPLE REGRESSION ANALYSIS

6.1 MULTIPLE REGRESSION MODEL

It is generally the case in all scientific investigations that there is no single cause of a given phenomenon or outcome. The inherent complexity of most real world problems suggests that we can more accurately describe, predict, and control an outcome variable by using a regression model that employs more than one independent variable. Such a model is called a **multiple regression model** in contrast to linear regression model. The demand for a commodity, for example, is likely to be dependent not only on its own price, but also on the prices of other competing or complementary goods, quality of the goods, income of consumers, their taste and the like. The yield of wheat may depend on such factors as fertilizer, rainfall, soil fertility, irrigation etc. In such situations, our two-variable regression analysis with one independent variable, as discussed earlier, appears inadequate. This leads to the consideration of multiple regression models that employ more than two independent variables. We begin our study of these models by considering the following example.

Suppose an economist wishes to predict income of individuals (y) on the basis of their level of education (x_1) and age (x_2) both measured in years. When plotted as a scatter diagram, the plot shows a straight-line relationship between y and x_1. This suggests that if we wish to predict y on the basis of x_1 alone, the simple linear regression model of the following form relates y to x_1:

$$y = \alpha + \beta_1 x_1 + \varepsilon \quad \ldots (6.1)$$

Further suppose that the scatter plot of y versus x_2 also shows a straight-line relationship between y and x_2. This suggests that if we wish to predict y on the basis of x_2 alone, the simple linear regression model of the following form relates y to x_2:

$$y = \alpha + \beta_1 x_2 + \varepsilon \quad \ldots (6.2)$$

Since we wish to predict y on the basis of both x_1 and x_2, it seems reasonable to combine the models (6.1) and (6.2) to build a model of the following form to relate y to x_1 and x_2 simultaneously

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad \ldots (6.3)$$

The model (6.3) is interpreted as follows:

(1) $\alpha + \beta_1 x_1 + \beta_2 x_2$ is the mean value of y when an individual has x_1 years of education and is of age x_2 years, i.e. $\mu_{y|x_1, x_2} = \alpha + \beta_1 x_1 + \beta_2 x_2$.

(2) α, β_1 and β_2 are the regression parameters of the model relating the mean value of y to x_1 and x_2.

(3) ε is an error term that describes the effects on y of all factors other than x_1 and x_2.

(4) α is the intercept of the regression model.

We call (6.3) a linear regression model because the expression $\alpha + \beta_1 x_1 + \beta_2 x_2$ expresses the mean value of y as a linear function of the parameters α, β_1, and β_2.

The general form of a multiple regression model expresses the dependent variable y as a function of k independent variables $x_1, x_2, \ldots x_k$. The model is of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$
$$= \mu_{y|x_1, x_2, x_3 \ldots\ldots, x_k} + \varepsilon \quad \ldots (6.4)$$

Here we assume that we have obtained n observations with each observation consisting of observed value of y and corresponding observed values of $x_1, x_2, \ldots$, and x_k.

The assumptions in the multiple regression set-up are similar to those of simple regression model with one independent variable. Thus in multiple regression analysis, we assume that

(1) The observed values of $x_1, x_2, \ldots, x_k$ are held fixed.
(2) For any given combination of the values of x_1 and x_2, the dependent variable y has a normal distribution about the conditional mean with constant variance.
(3) The value of the error term ε corresponding to an observed value of y is statistically independent of the error term corresponding to any other observed value of y.

6.6.1 Interpreting the Parameters of the Model

How do we interpret the parameters α, β_1, and β_2 in the model? Let us consider the case in which two independent variables, with x_1 as education and x_2 as age are involved, the dependent variable being the income labeled y. First suppose that x_1=0 and x_2=0. Then

$$\alpha + \beta_1 x_1 + \beta_2 x_2 = \alpha$$

This implies that α is the mean income in the population for persons who are 0 years old and have no education. We wonder whether α has any practical interpretation in this particular instance because we cannot think of level of education of a newborn and consequently his/her income. Indeed, sometimes the parameter α and other parameters in a regression analysis do not have practical interpretations because the situations related to the interpretations would not be likely to occur in practice.

To examine the interpretation of β_1, suppose we want to predict the average income of persons who are n years old having completed m years of education. The mean income of all such persons is

$$\mu_1 = \alpha + \beta_1(m) + \beta_2(n) \qquad \ldots (6.4a)$$

Suppose now that the level of education is increased by 1 year and is equal to $(m+1)$ years for those who are still n years old so that the mean income is

$$\mu_2 = \alpha + \beta_1(m+1) + \beta_2(n) \qquad \ldots (6.4b)$$

Then clearly, the difference between (a) and (b)

$$\mu_D = \mu_2 - \mu_1 = \beta_1 \qquad \ldots (6.4c)$$

Thus, β_1 can be interpreted as the change in mean income that is associated with one-year increase in the level of education when the age does not change (i.e. when age is held constant). More explicitly, for a three-variable regression as in (6.3), β_1 measures the change in the mean value of y i.e. $\mu_{y|x_1,x_2}$ per unit change in x_1, holding x_2 constant. In other words, it gives the "direct" or "net" effect of a unit change in x_1 on the mean value of y, net of x_2. Likewise, β_2 measures the change in the mean value of y i.e. $\mu_{y|x_1,x_2}$ per unit change in x_2, holding x_1 constant. In other words, it gives the "direct" or "net" effect of a unit change in x_2 on the mean value of y, net of x_1.

The β values in the regression set-up are referred to as **partial regression coefficients**. They are called partial because each of the coefficients indicates only the effects of its respective independent variables on the dependent variable, with the effects of all other independent variables in the equation statistically controlled for. Thus the equation is an additive one, with each regression coefficient indicating an independent and unique effect not shared by other variables in the equation.

The regression parameters $\alpha, \beta_1, \beta_2, ..., \beta_k$ in the linear model (6.4) are all unknown. Therefore, they must be estimated from data (observations of y, $x_1, x_2, .. x_k$). These estimates are denoted by a, b_1, b_2, ...,b_k. With these estimates, the estimated regression function is of the following form:

$$\hat{y} = a + b_1x_1 + b_2x_2 + ... + b_kx_k \quad ... (6.5)$$

where $\hat{y}$ is the estimated value of y. We denote the difference between y and $\hat{y}$ by a quantity e, called **residual**.

For the estimation purpose, we can follow principles of least-squares as described in earlier chapter. We can interpret the estimated coefficient a as the y intercept of the estimated regression function and b_1, b_2, ...,b_k as slopes of the independent variables. However, the estimated regression function can no longer be represented by a line in two-dimensional co-ordinate system (x, y). Instead the estimated function (6.5) represents a plane in a (k+1)–dimensional co-ordinate system (x_1, x_2, ...,x_k,, y).

6.2 ESTIMATING THE PARAMETERS IN THE MODEL

The estimation of the parameters in the multiple regression set-up can be accomplished through solving the normal equations generated through the

application of the least-squares procedure to the observed data. For a thr variate case (y, x_1, x_2) as in (6.3), with y as the dependent variable, let define the residual sum of squares as under:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2 \quad \ldots (6.$$

where a, b_1 and b_2 are respectively the estimates of the unknown paramet α, β_1 and β_2 of the regression function (6.3).

Our aim now is to choose a, b_1 and b_2 so as to minimize the residual sum squares Σe_i^2. For this purpose, the partial derivatives of Σe_i^2. in (6.5) wi respect to a, b_1 and b_2 are obtained:

$$\frac{\partial \sum e_i^2}{\partial a} = -\sum (y_i - a - b_1 x_{1i} - b_2 x_{2i})$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = -\sum x_{1i} (y_i - a - b_1 x_{1i} - b_2 x_{2i})$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = -\sum x_{2i} (y_i - a - b_1 x_{1i} - b_2 x_{2i})$$

Setting these derivatives to zero, we have the following normal equations i a, b_1 and b_2:

$$\sum y_i = na + b_1 \sum x_{1i} + b_2 \sum x_{2i} \quad \ldots (a)$$

$$\sum x_{1i} y_i = a \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i} \quad \ldots (b)$$

$$\sum x_{2i} y_i = a \sum x_{2i} + b_1 \sum x_{1i} x_{2i} + b_2 \sum x_{2i}^2 \quad \ldots (c)$$

Dividing (a) throughout by n

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \quad \ldots (6.7)$$

Substituting the value of a thus obtained in (b), we get

$$\sum x_{1i} y_i - \frac{\sum x_{1i} \sum y_i}{n} = b_1 \left[\sum x_{1i}^2 - \frac{\left(\sum x_{1i} \right)^2}{n} \right] + b_2 \left[\sum x_{1i} x_{2i} - \frac{\sum x_{1i} \sum x_{2i}}{n} \right]$$

Or

$$\sum (x_{1i} - \bar{x}_1)(y_i - \bar{y}) = b_1 \sum (x_{1i} - \bar{x}_1)^2 + b_2 \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

The above expression can be written in a more reduced form as follows:

$$S_{1y} = b_1 S_{11} + b_2 S_{12} \qquad \ldots (6.8a)$$

Similarly, on substituting the value of a in (c), we obtain

$$\sum x_{2i} y_i - \frac{\sum x_{2i} \sum y_i}{n} = b_1 \left[\sum x_{1i} x_{2i} - \frac{\sum x_{1i} \sum x_{2i}}{n} \right] + b_2 \left[\sum x_{2i}^2 - \frac{\left(\sum x_{2i} \right)^2}{n} \right]$$

Cr

$$\sum (x_{2i} - \bar{x}_2)(y_i - \bar{y}) = b_1 \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + b_2 \sum (x_{2i} - \bar{x}_2)^2$$

from which

$$S_{2y} = b_1 S_{12} + b_2 S_{22} \qquad \ldots (6.8b)$$

The equations (6.8a) and (6.8b) are known as the reduced normal equations.

The simultaneous solutions of the above equations for the unknown constants b_1 and b_2 are

$$b_1 = \frac{S_{22} S_{1y} - S_{12} S_{2y}}{S_{11} S_{22} - S_{12}^2} \text{ and } b_2 = \frac{S_{11} S_{2y} - S_{12} S_{1y}}{S_{11} S_{22} - S_{12}^2} \qquad \ldots (6.9)$$

Having obtained b_1 and b_2 from the observed data, the value of a can be obtained from (6.7), and finally the resulting equation in multiple regression set-up with two independent variables is of the form

$$\hat{y} = a + b_1 x_1 + b_2 x_2 \qquad \ldots (6.10a)$$

We now show that the mean value of $\hat{y}_i$ is equal to the mean value of the observed y. That is $\bar{\hat{y}} = \bar{y}$. To show this, we proceed as follows:

$$\begin{aligned} \hat{y}_i &= a + b_1 x_{1i} + b_2 x_{2i} \\ &= (\bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2) + b_1 x_{1i} + b_2 x_{2i} \\ &= \bar{y} + b_1 (x_{1i} - \bar{x}_1) + b_2 (x_{2i} - \bar{x}_2) \end{aligned} \qquad \ldots (6.10b)$$

Summing and dividing throughout by n, it follows that $\bar{\hat{y}} = \bar{y}$, since $b_1 \sum (x_{1i} - \bar{x}_1) = 0$ and $b_2 \sum (x_{2i} - \bar{x}_2) = 0$

6.3 SOME PROPERTIES OF THE ESTIMATORS

Specifically, the properties of the least squares estimators of regression model with k independent variables are as follows:

$$\begin{aligned}\sum(\hat{y}_i-\bar{y})^2 &= \sum(b_1X_{1i}+b_2X_{2i})^2\\ &= \sum(b_1X_{1i}+b_2X_{2i})(b_1X_{1i}+b_2X_{2i})\\ &= b_1\left[b_1\sum X_{1i}^2+b_2\sum X_{1i}X_{2i}\right]+b_2\left[b_1\sum X_{1i}X_{2i}+b_2\sum X_{2i}^2\right]\\ &= b_1(b_1S_{11}+b_2S_{12})+b_2(b_1S_{12}+b_2S_{22})\end{aligned}$$

Comparing these expressions inside the parentheses with the normal equations developed before (see Section 6.2), we find that

$$b_1S_{11}+b_2S_{12}=S_{1y} \text{ and } b_1S_{12}+b_2S_{22}=S_{2y}$$

Hence

$$\text{SSR}=\sum(\hat{y}_i-\bar{y})^2=b_1S_{1y}+b_2S_{2y} \quad \text{(Proved)}$$

Example 6.1: For 10 families, the following data were available on their income, expenditure (both in '000 taka) and family size.

Family (i)	Expenditure (y_i)	Income (x_{1i})	Family size (x_{2i})
1	7	10	4
2	8	12	5
3	9.5	15	5
4	10	18	8
5	11	18	7
6	15	20	9
7	18	20	8
8	18	19	9
9	20	20	9
10	25	29	10

(a) Fit a regression line of the type $y_i=\alpha+\beta_1x_{1i}+\beta_2x_{2i}$.

(b) Predict the average expenditure of a family with 6 members reporting an income of 17000 taka.

(c) Find the standard error of the estimate using formula (6.15b and 6.15c) and verify if they are equivalent.

(d) Find the variances of the least squares estimates.

Solution: To accomplish the tasks above, we construct the following table:

y_i	x_{1i}	x_{2i}	y_i^2	x_{1i}^2	x_{2i}^2	$x_{1i}y_i$	$x_{2i}y_i$	$x_{1i}x_{2i}$
7	10	4	49	100	16	70	28	40
8	12	5	64	144	25	96	40	60
9.5	15	5	90.25	225	25	142	47	75
10	18	8	100	324	64	180	80	144
11	18	7	121	324	49	198	77	126
15	20	9	225	400	81	300	135	180
18	20	8	324	400	64	360	144	160
18	19	9	324	361	81	342	182	171
20	20	9	400	400	81	400	180	180
25	29	10	625	841	100	725	250	290

Summing the relevant columns

$$\sum y_i = 141.5, \sum x_{1i} = 181, \sum x_{2i} = 74,$$
$$\sum y_i^2 = 2322.25, \sum x_{1i}^2 = 3519$$
$$\sum x_{2i}^2 = 586, \sum x_{1i}y_i = 2813.5,$$
$$\sum x_{2i}y_i = 1143.5, \sum x_{1i}x_{2i} = 1436$$

With the above values

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 2322.25 - \frac{(141.5)^2}{10} = 320.02$$

$$S_{1y} = \sum x_{1i}y_i - \frac{\sum x_{1i}\sum y_i}{n} = 2813.50 - \frac{(181)(141.5)}{10} = 252.35$$

$$S_{2y} = \sum x_{2}y_i - \frac{\sum x_{2i}\sum y_i}{n} = 1143.50 - \frac{(74)(141.5)}{10} = 96.4$$

$$S_{11} = \sum x_{1i}^2 - \frac{\left(\sum x_{1i}\right)^2}{n} = 3519 - \frac{(181)^2}{10} = 242.9$$

$$S_{22} = \sum x_{2i}^2 - \frac{\left(\sum x_{2i}\right)^2}{n} = 586 - \frac{(74)^2}{10} = 38.4$$

$$S_{12} = \sum x_{1i}x_{2i} - \frac{\sum x_{1o}\sum x_{2i}}{n} = 1426 - \frac{(181)(74)}{10} = 86.6$$

hus the estimates of the parameters are

$$b_1 = \frac{S_{22}S_{1y} - S_{12}S_{2y}}{S_{11}S_{22} - S_{12}^2}$$

$$= \frac{(38.4)(252.35) - (86.6)(96.4)}{(242.9)(38.4) - (86.6)^2}$$

$$= \frac{1342}{1827.8} = 0.734$$

$$b_2 = \frac{S_{11}S_{2y} - S_{12}S_{1y}}{S_{11}S_{22} - S_{12}^2}$$

$$= \frac{(242.9)(96.4) - (86.6)(252.35)}{(242.9)(38.4) - (86.6)^2}$$

$$= \frac{1562.05}{1827.8} = 0.855$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = 14.15 - (.734)(18.1) - (.855)(7.4) = -5.462$$

A value of –5.642 for a implies that the expenditure is –5463 taka when both income and family size are zero. Of course, it does not make any sense to have zero income and zero family size and consequently negative expenditure. It is important to keep in mind that a regression equation is not effective outside the range of the sample values for the dependent variable.

A b_1 value of 0.734 implies that for each increase of 1 taka in income, the family expenditure would increase by 0.734 taka regardless of the family size. The b_2 value (=.855) implies that for each additional increase of one member in the family, the family expenditure would go up by 0.855 taka regardless of the family income.

The estimated expenditure of a family with 6 members and an income of 17 thousand is

$$\hat{y}_{(17)} = -5.462 + 0.734(17) + 0.855(6) = 12.15.$$

The standard error of the estimate is

$$s_e^2 = \frac{S_{yy} - b_1S_{1y} - b_2S_{2y}}{n-3}$$

$$= \frac{320.02 - (.734)(252.35) - (.855)(96.4)}{7}$$

$$= 7.48$$

6.6 MULTIPLE CORRELATION

The multiple correlation coefficient is symbolized by R which shows the correlation among more than two variables. As with r^2, R^2 (square of the multiple correlation coefficient) indicates the proportion of variance in the dependent variable that is accounted for by the set of predictors (explanatory variables) included in the regression equation. If $R = 0.5$, then $R^2 = 0.25$ and we would conclude that the predictors being considered account for 25 percent of the total variance in the dependent variable. The multiple correlation coefficient ranges from 0 (when the independent variables in no way help to predict y) to 1 (when the independent variables predict y with complete accuracy).

The multiple correlation coefficient R is usually written with subscripts to indicate the variables being correlated, with a decimal after the dependent variable. For example, if variables 1 through 3 were being correlated and variable 1 were the dependent variable and the other two were independent variables, the multiple correlation coefficient would be written $R_{1.23}$. Thus, for three variables designated x_1, x_2 and x_3, $R_{1.23}$ is a measure of the degree of association between dependent variable x_1 and the independent variables x_2 and x_3 jointly. The most convenient form of expressing the multiple correlation coefficient in terms of simple correlation coefficients for 3-variable case is

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{13}}{1 - r^2_{23}} \qquad \ldots (6.18)$$

where r_{ij} is the **simple** or **zero order correlation coefficient** between x_i and x_j, $i\,j = 1, 2, 3$

Example 6.2: Given the following values of x_1, x_2, and x_3. Compute the multiple correlation coefficient $R_{1.23}$.

x_1:	18	16	30	15	13	26	25
x_2:	14	12	14	10	7	13	8
x_3:	13	14	18	12	11	16	12

Solution: To make use of the formula (6.18), we first compute the following simple correlation coefficients: r_{12}, r_{13} and r_{23}, where r_{12}, r_{13}, and r_{23} are respectively the coefficients of correlation between the pairs (x_1, x_2), (x_1, x_3), and (x_2, x_3). It can be shown that r_{12}, $=0.445$, $r_{13} = 0.778$ and $r_{23} = 0.778$. Thus

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{13}}{1 - r^2_{23}}$$

$$= \frac{.445^2 + .778^2 - 2(.445)(.778)(778)}{1 - .778^2}$$

$$= 0.67$$

Extracting the square root

$$R_{1.23} = \sqrt{0.67} = 0.81$$

6.6.1 An Alternative Method of Computing $R_{1.23}$

The multiple correlation coefficient R can also be computed without any reference to the simple correlation coefficients as in (6.18). We know that the total variability in the dependent variable can be sub-divided into two components: one is attributable to the regression (labeled regression) and the other is not (labeled residual). In terms of sum of squares SST= SSR+ SSE.

Since R^2 is the proportion of the variations in the dependent variable "explained" by the model,

$$R^2_{1.23} = \frac{\text{Sum of squares due to regression}}{\text{Total sum of squares}}$$

$$= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{SSR}}{\text{SST}} \qquad \ldots (6.19)$$

Computation of Total SS from data is straightforward. The SSR can be calculated using the following formula:

$$\text{SSR} = b_1 S_{1y} + b_2 S_{2y} + \ldots + b_i S_{ky} \qquad \ldots (6.20)$$

for k independent variables.

For a three-variable regression set up of the form $\hat{x}_1 = a + b_{12.3}x_2 + b_{13.2}x_3$

$$R^2_{1.23} = \frac{b_{12.3} S_{12} + b_{13.2} S_{13}}{S_{11}} \qquad \ldots (6.21a)$$

We verify the formula (6.21a) for its consistency by Example 6.2. The necessary computations are shown in the accompanying table.

x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3
18	14	13	324	196	169	252	234	182
16	12	14	256	144	196	192	224	168
30	14	18	900	196	324	420	540	262
15	10	12	225	100	144	150	180	120
13	7	11	169	49	121	91	143	77
26	13	16	676	169	256	338	416	208
25	8	12	625	64	144	200	300	96
143	78	96	3175	918	1354	1643	2037	1103

To compute $R_{1.23}$, we calculate the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ as follows:

$$b_{12.3} = \frac{S_{33}S_{12} - S_{23}S_{13}}{S_{22}S_{33} - S_{23}^2} \text{ and } b_{13.2} = \frac{S_{22}S_{13} - S_{23}S_{12}}{S_{22}S_{33} - S_{23}^2} \quad \text{... (6.21b)}$$

Now we calculate the sum of products and sum of squares to obtain the regression coefficients.

$$S_{11} = \sum x_{1i}^2 - \frac{\left(\sum x_{1i}\right)^2}{n} = 3175 - \frac{(143)^2}{7} = 253.71$$

$$S_{22} = \sum x_{2i}^2 - \frac{\left(\sum x_{2i}\right)^2}{n} = 918 - \frac{(78)^2}{7} = 48.86$$

$$S_{33} = \sum x_{3i}^2 - \frac{\left(\sum x_{3i}\right)^2}{n} = 1354 - \frac{(96)^2}{7} = 49.57$$

$$S_{12} = \sum x_{1i}x_{2i} - \frac{\sum x_{1i}\sum x_{2i}}{n} = 1643 - \frac{143 \times 78}{7} = 49.57$$

$$S_{13} = \sum x_{1i}x_{3i} - \frac{\sum x_{1i}\sum x_{3i}}{n} = 2037 - \frac{143 \times 96}{7} = 75.86$$

$$S_{23} = \sum x_{2i}x_{3i} - \frac{\sum x_{2i}\sum x_{3i}}{n} = 1103 - \frac{78 \times 96}{7} = 33.29$$

Substituting the above values in (6.21a)

$$b_{12.3} = \frac{(37.43)(49.57) - (33.29)(75.86)}{(48.86)(37.43) - (33.29)^2} = \frac{-669.97}{720.61} = -0.93$$

$$b_{13.2} = \frac{(48.86)(75.86) - (33.29)(49.57)}{(48.86)(37.43) - (33.29)^2} = \frac{2056.33}{720.61} = 2.85$$

Hence

$$R^2_{1.23} = \frac{(-.93)(49.57)+(2.85)(75.86)}{253.71} = \frac{170.1}{253.71} = 0.67$$

which is in complete agreement with our previous result.

The multiple correlation coefficient R is always non-negative. This means that unlike an ordinary correlation coefficient r, the coefficient of multiple correlation can vary from 0 to 1. We put this property of R in the following theorem with proof.

***Theorem* 6.3**: *The multiple correlation coefficient R is non-negative i.e.* $R \geq 0$.

Proof: Suppose that the variables in a multiple regression all have been measured from their respective means so that

$$\hat{y} = b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

The coefficient of correlation between the observed y and the estimated y (i.e. $\hat{y}$) represents the multiple correlation coefficient. It seems obvious that if the sum of product of y and its estimate $\hat{y}$ is non-negative, then R will also be non-negative. We show below that $S_{y\hat{y}}$ is indeed non-negative.

By definition

$$\begin{aligned} S_{y\hat{y}} &= \sum y\hat{y} = \sum y(b_1x_1 + b_2x_2 + \ldots + b_kx_k) \\ &= b_1\sum x_1y + b_2\sum x_2y + \ldots + b_k\sum x_ky \\ &= b_1S_{1y} + b_2S_{2y} + \ldots + b_kS_{ky} \\ &= \text{SSR} \geq 0 \quad \text{(Proved)} \end{aligned}$$