

Chapter 5

Regression Analysis and Multiple Correlations: For Estimating a Measurable Phenomenon

Learning Objectives

After completing this chapter, you should be able to do the following:

- Explain the use of regression analysis and multiple correlation in research.
- Interpret various terms involved in regression analysis.
- Learn to use SPSS for doing regression analysis.
- Understand the procedure of identifying the most efficient regression model.
- Know the method of constructing the regression equation based on the SPSS output.

Introduction

Regression analysis deals with estimating the value of dependent variable on the basis of one or more independent variables. To do so, an equation is developed between dependent and independent variables by means of least square method. When the estimation is done on the basis of one independent variable, the procedure is known as simple regression, and if the estimation involves more than one independent variable, it is referred to as multiple regression analysis.

In multiple regression analysis, the dependent variable is referred to as Y , whereas independent variables are denoted as X . The dependent variable is also known as criterion variable. The goal is to develop an equation that will determine the Y variable in a linear function of corresponding X variables. The regression equation can be either linear or curvilinear, but our discussion shall be limited to linear regression only.

In regression analysis, a regression model is developed by using the observed data obtained on dependent variable and several independent variables. During the process, only those independent variables are picked up for developing the model which shows significant relationship with dependent variable. Therefore, the researcher must be careful in identifying the independent variables in regression

analysis study. It may be quite possible that some of the important independent variables might have been left in the study, and, therefore, in spite of the best possible effort, the regression model so developed may not be reliable.

Multiple regression analysis can be used in many applications of management and behavioral researches. Numerous situations can be listed where the use of this technique can provide an edge to the decision makers for optimum solutions.

For example, in order to evaluate and reform the existing organization and make them more responsive to the new challenges, the management may be interested to know; the factors responsible for sale to plan the business strategy, the estimated number of inventory required in a given month, and the factors affecting the job satisfaction of the employees. They may also be concerned in developing the model for deciding the pay packets of an employee, factors that motivate people to work, or parameters that affect the productivity of work. In all these situations, regression model may provide the input to the management for strategic decision-making. The success of the model depends upon the inclusion of relevant independent variables in the study. For instance, a psychologist may like to draw up variables that directly affect one's mental health causing abnormal behavior. Therefore, it is important for the researchers to review the literature thoroughly for identifying the relevant independent variables for estimating the criterion variable.

Besides regression analysis, there are other quantitative and qualitative methods used in performance forecasting. But the regression analysis is one of the most popularly used quantitative techniques.

In developing a multiple regression equation, one needs to know the efficiency in estimating the dependent variable on the basis of the identified independent variables in the model. The efficiency of estimation is measured by the coefficient of determination (R^2) which is the square of multiple correlation. The coefficient of determination explains the percentage of variance in the dependent variable by the identified independent variables in the model. The multiple correlation explains the relationship between the group of independent variables and dependent variable. Thus, high multiple correlation ensures greater accuracy in estimating the value of dependent variable on the basis of independent variables. Usually multiple correlation, R is computed during regression analysis to indicate the validity of regression model. It is necessary to show the value of R^2 along with regression equation for having an idea about the efficiency in prediction.

Any regression model having larger multiple correlation gives better estimates in comparison to that of other models. We will see an explanation of the multiple correlation while discussing the solved example later in this chapter.

Terminologies Used in Regression Analysis

In order to use the regression analysis effectively, it is essential to know different terminologies involved in it. These terms are discussed in the following sections.

Multiple Correlation

Multiple correlation is a measure of relationship between a group of independent variables and a dependent variable. Since multiple correlation provides the strength of relationship between dependent variable and independent variables, it is used to determine the power of regression models also. The multiple correlation is represented by “ R ” and is computed by the following formula:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2}} \quad (5.1)$$

If the number of independent variables is more than two, then the multiple correlation is computed from the following formula:

$$R_{1.2345\dots n} = \sqrt{1 - (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)\dots(1 - r_{1n.23\dots(n-1)}^2)} \quad (5.2)$$

where $r_{13.2}$, $r_{14.23}$, and $r_{1n.234\dots(n-1)}$ are partial correlations.

The multiple correlation R can have the value in between 0 and +1. Since multiple correlation is computed with the help of product moment correlation coefficients, therefore it also measures the linear relationship only. Further, the order of the multiple correlation is defined by $n - 2$, where n is the number of variables involved in the computation of multiple correlation. Thus, the order of the multiple correlation $R_{12.345\dots n}$ is $n - 2$. The value of R closer to 1 indicates that the independent variables explain most of the variations in the dependent variable. On the other hand, if the value of R is closer to 0, it signifies that independent variables are not capable of explaining the variation in the dependent variable. Thus, multiple correlation can be considered to be the yardstick of efficiency in estimating the value of dependent variable on the basis of the values of independent variables.

Properties of Multiple Correlation

1. The multiple correlation can never be lower than the highest correlation between dependent and any of the independent variables. For instance, the value of $R_{1.234}$ can never be less than the value of any of the product moment correlations r_{12} , r_{13} , or r_{14} .
2. Sometimes, an independent variable does not show any relationship with dependent variable, but if it is combined with some other variable, its effect becomes significant. Such variable is known as suppression variable. These suppression variables should be handled carefully. Thus, if the independent variables are identified on the basis of their magnitude of correlations with the dependent variable for developing regression line, some of the suppression variable might

Table 5.1 Correlation matrix of psychological variables

	X_1	X_2	X_3	X_4
X_1	1	-0.5	0.3	0.6
X_2		1	0.4	0.3
X_3			1	0.4
X_4				1

X_1 : Memory retention, X_2 : Age, X_3 : IQ, X_4 : Stress level

be ignored. To handle this problem, SPSS provides the stepwise regression method.

3. In using the stepwise regression, the variables are picked up one by one depending upon their relative importance. Every time one variable is included, there is an increase in multiple correlation. But the increase in multiple correlation keeps on decreasing with the inclusion of every new variable. This is known as *law of diminishing return*.

Example 5.1 Following is the correlation matrix obtained on the psychological variables. Compute $R_{1.23}$ and $R_{1.234}$ and interpret the findings (Table 5.1):

Solution

(i)

$$\therefore R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Substituting values of these correlations from the correlation matrix,

$$\begin{aligned} R_{1.23} &= \sqrt{\frac{(-0.5)^2 + 0.3^2 - 2 \times (-0.5) \times 0.3 \times 0.4}{1 - 0.4^2}} \\ &= \sqrt{\frac{0.46}{0.84}} = 0.74 \end{aligned}$$

(ii)

$$\therefore R_{1.234} = \sqrt{1 - [(1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)]}$$

To compute $R_{1.234}$, we need to first find the values of first-order partial correlations: $r_{13.2}$, $r_{14.2}$, and $r_{43.2}$.

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}} = \frac{0.3 - (-0.5) \times 0.4}{\sqrt{1 - (-0.5)^2}\sqrt{1 - 0.4^2}} = \frac{0.50}{0.79} = 0.63$$

$$r_{14.2} = \frac{r_{14} - r_{12}r_{42}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{42}^2}} = \frac{0.6 - (-0.5) \times 0.3}{\sqrt{1 - (-0.5)^2}\sqrt{1 - 0.3^2}} = \frac{0.75}{0.83} = 0.90$$

$$r_{43.2} = \frac{r_{43} - r_{42}r_{32}}{\sqrt{1 - r_{42}^2}\sqrt{1 - r_{32}^2}} = \frac{0.4 - 0.3 \times 0.4}{\sqrt{1 - 0.3^2}\sqrt{1 - 0.4^2}} = \frac{0.28}{0.87} = 0.32$$

After substituting these partial correlations, the second-order partial correlation $r_{14.23}$ shall be computed; this in turn shall be used to compute the value of $R_{1.234}$.

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}r_{43.2}}{\sqrt{1 - r_{13.2}^2}\sqrt{1 - r_{43.2}^2}} = \frac{0.90 - 0.63 \times 0.32}{\sqrt{1 - 0.63^2}\sqrt{1 - 0.32^2}} = \frac{0.6984}{0.7358} = 0.95$$

Thus, substituting the values of r_{12} , $r_{13.2}$, and $r_{14.23}$ in the following equation:

$$\begin{aligned} R_{1.234} &= \sqrt{1 - [(1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2)]} \\ &= \sqrt{1 - [(1 - (-0.5)^2)(1 - 0.63^2)(1 - 0.95^2)]} \\ &= \sqrt{1 - [0.75 \times 0.603 \times 0.10]} \\ &= 0.976 \end{aligned}$$

Interpretation

Taking n equals to 2 and substituting the value of r_{12} in Eq. (5.2),

$$R_{1.2} = \sqrt{1 - [1 - (-0.5)^2]} = \sqrt{0.25} = 0.5$$

Now let us have a look on the following values:

$$R_{1.2} = 0.5$$

$$R_{1.23} = 0.74$$

$$R_{1.234} = 0.976$$

It can be seen that the multiple correlation increases with the increase in the independent variable. Further, the increase in multiple correlation is larger when the third independent variable (X_3) is included in the model and after that the increase has reduced when one additional independent variable (X_4) is introduced.

Coefficient of Determination

It can be defined as the variance explained in the dependent variable on the basis of the independent variables selected in the regression model. It is the square of multiple correlation and is represented by R^2 . Thus, in regression analysis R^2 is

used for assessing the efficiency of the regression model. If for a particular regression model R is 0.8, it means that 64% of the variability in the dependent variable can be explained by the independent variables selected in the model.

The Regression Equation

The equation is said to be simple regression if the value of dependent variable is estimated on the basis of one independent variable only. If Y is the dependent variable and X is the independent variable, then the regression equation of Y on X is written as

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X}) \quad (5.3)$$

Equation 5.3 can be used to predict the value of Y if the value of X is known. Similarly to estimate the value of X from the value of Y , the regression equation of X on Y shall be used which is shown in Eq. 5.4.

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y}) \quad (5.4)$$

where \bar{X} and \bar{Y} are the sample means of X and Y , respectively, and b_{yx} and b_{xy} are the regression coefficients. These regression coefficients can be computed as

$$b_{yx} = r \frac{\sigma_Y}{\sigma_X} \quad (5.5)$$

$$b_{xy} = r \frac{\sigma_X}{\sigma_Y} \quad (5.6)$$

After substituting the value of b_{yx} in Eq. (5.3) and solving, we get

$$Y = r \frac{\sigma_Y}{\sigma_X} X + (\bar{Y} - r \frac{\sigma_Y}{\sigma_X} \bar{X}) \quad (5.7)$$

$$\Rightarrow Y = BX + C \quad (5.8)$$

where B is equal to $r \frac{\sigma_Y}{\sigma_X}$ and C is $(\bar{Y} - r \frac{\sigma_Y}{\sigma_X} \bar{X})$. The coefficients B and C are known as unstandardized regression coefficient and regression constant respectively.

Remark Reproduce $r \frac{\sigma_Y}{\sigma_X}$ and $(\bar{Y} - r \frac{\sigma_Y}{\sigma_X} \bar{X})$ in equation format. \bar{Y} is the mean of Y and \bar{X} is the mean of X

After substituting the values of b_{yx} and b_{xy} in the regression equations (5.3) and (5.4), we get

$$(Y - \bar{Y}) = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

$$(X - \bar{X}) = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

After rearranging the above equations,

$$\frac{(Y - \bar{Y})}{\sigma_y} = r \frac{(X - \bar{X})}{\sigma_x}$$

$$\frac{(X - \bar{X})}{\sigma_x} = r \frac{(Y - \bar{Y})}{\sigma_y}$$

The above two equations can be rewritten as

$$Z_y = \beta_x Z_x \quad (5.9)$$

$$Z_x = \beta_y Z_y \quad (5.10)$$

The Eqs. (5.9) and (5.10) are known as regression equations in standard score form, and the coefficients β_x and β_y are known as beta coefficients and are referred to as standardized regression coefficients.

Conditions of Symmetrical Regression Equations

The two regression equations (5.3) and (5.4) are different. Equation (5.3) is known as regression equation of Y on X and is used to estimate the value of Y on the basis of X , whereas Eq. (5.4) is known as regression equation of X on Y and is used for estimating the value of X if Y is known. These two equations can be rewritten as follows:

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

$$(Y - \bar{Y}) = \frac{1}{b_{xy}}(X - \bar{X})$$

These two regression equations can be same if the expressions in the right-hand side of these two equations are same.

That is,

$$\begin{aligned}
b_{yx}(X - \bar{X}) &= \frac{1}{b_{xy}}(X - \bar{X}) \\
\Rightarrow b_{yx} \times b_{xy} &= 1 \\
\Rightarrow r \frac{\sigma_y}{\sigma_x} \times r \frac{\sigma_x}{\sigma_y} &= 1 \\
\Rightarrow r^2 &= 1 \\
\Rightarrow r &= \pm 1
\end{aligned}$$

Hence, the two regression equations shall be similar if there is a perfect positive or perfect negative correlation between them. In that situation, same regression equation can be used to estimate the value of Y or value of X .

Computation of Regression Coefficient

The regression coefficient can be obtained for the given set of data by simplifying the formula:

$$\begin{aligned}
\therefore B &= r \frac{\sigma_y}{\sigma_x} \\
&= \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}} \times \frac{\sqrt{\frac{1}{N} \sum Y^2 - \left(\frac{\sum Y}{N}\right)^2}}{\sqrt{\frac{1}{N} \sum X^2 - \left(\frac{\sum X}{N}\right)^2}}
\end{aligned}$$

After solving,

$$B = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2} \quad (5.11)$$

From Eqs. (5.7) and (5.8),

$$C = \bar{Y} - r \frac{\sigma_y}{\sigma_x} \bar{X}$$

After substituting the value of $r \frac{\sigma_y}{\sigma_x} = B$ from Eq. (5.11), we get

$$C = \frac{\sum Y}{N} - \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2} \times \frac{\sum X}{N}$$

After simplification,

$$C = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N \sum X^2 - (\sum X)^2} \quad (5.12)$$

Thus, by substituting the value of B and C in Eq. (5.8), regression equation can be developed.

Example 5.2 Consider the two sets of scores on job satisfaction (X) and autonomy (Y) as shown below. Compute the regression coefficient “ B ” and constant “ C ” and develop regression equation.

Autonomy (Y)	: 15 13 7 11 9
Job satisfaction (X)	: 9 8 5 8 6

Solution The regression equation given by $Y = BX + C$ can be constructed if the regression coefficient B and constant C are known. These can be obtained by the following formula (Table 5.2):

$$\therefore B = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2} \quad C = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N \sum X^2 - (\sum X)^2}$$

To compute “ B ” and “ C ,” we shall first compute $\sum X$, $\sum Y$, $\sum X^2$, and $\sum XY$.

$$B = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2} = \frac{5 \times 416 - 55 \times 36}{5 \times 645 - 55 \times 55} = 0.5$$

$$C = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N \sum X^2 - (\sum X)^2} = \frac{36 \times 645 - 55 \times 416}{5 \times 645 - 55 \times 55} = 1.7$$

Substituting the values of B and C , the regression equation becomes

$$Y(\text{Job satisfaction}) = 0.5X(\text{Autonomy}) + 1.7$$

These values can be obtained from the SPSS output discussed in the solved Example 5.1. The SPSS produces these outputs on the basis of least square methods. The method of least square has been discussed later in this chapter.

Properties of Regression Coefficients

1. The square root of the product of two regression coefficients is equal to the correlation coefficient between X and Y . The sign of the correlation coefficient is equal to the sign of the regression coefficients. Further, the signs of the two regression coefficients are always same.

Table 5.2 Computation for regression coefficients

Scores on			
Autonomy	Job satisfaction		
(X)	(Y)	X ²	XY
15	9	225	135
13	8	169	104
7	5	49	35
11	8	121	88
9	6	81	54
<u>ΣX = 55</u>	<u>ΣY = 36</u>	<u>ΣX² = 645</u>	<u>ΣXY = 416</u>

$$\begin{aligned}\therefore b_{yx} \times b_{xy} &= r \frac{\sigma_y}{\sigma_x} \times r \frac{\sigma_x}{\sigma_y} = r^2 \\ \Rightarrow r &= \pm \sqrt{b_{yx} \times b_{xy}}\end{aligned}$$

To prove that the sign of the correlation coefficient between X and Y and both the regression coefficients are same, consider the following formula:

$$r_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{5.13}$$

$$b_{yx} = \frac{\text{Cov}(X, Y)}{\sigma_x^2} \tag{5.14}$$

$$b_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_y^2} \tag{5.15}$$

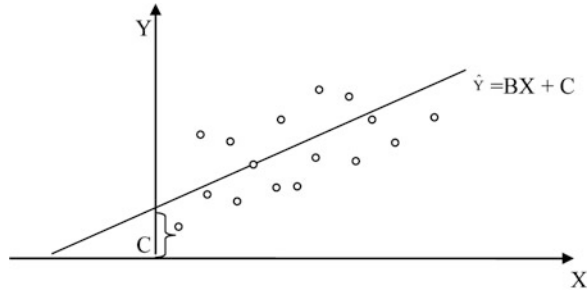
Since σ_x and σ_y are always positive, the values of r_{xy} , b_{yx} , and b_{xy} will be same and will depend upon the sign of $\text{Cov}(X,Y)$.

2. If one of the regression coefficients is greater than 1, the other will have to be less than 1. Thus, in other words, both the regression coefficients can be less than 1 but can never be greater than 1.

We have

$$\begin{aligned}b_{yx} \times b_{xy} &= r^2 \text{ and } -1 \leq r \leq 1 \\ \therefore b_{yx} \times b_{xy} &\leq 1 \\ \Rightarrow \text{If } b_{yx} > 1, &\text{ then } b_{xy} < 1 \\ \text{Or if } b_{xy} > 1, &\text{ then } b_{yx} < 1\end{aligned}$$

Fig. 5.1 Plotting of data and the line of best fit



Hence, the two regression coefficients cannot be simultaneously greater than one.

3. The average of the two regression coefficients is always greater than the correlation coefficient.

$$\frac{b_{yx} + b_{xy}}{2} > r$$

Least Square Method for Regression Analysis

The simple linear regression equation (5.8) is also known as least squares regression equation. Let us plot the paired values of X_i and Y_i for n sets of data; the scattergram shall look like Fig. 5.1.

The line of best fit can be represented as

$$\hat{Y} = BX + C$$

where B is the slope of the line and C is the intercept on Y axis. There can be many lines passing through these points, but the line of best fit shall be the one for which the sum of the squares of the residuals should be least. This fact can be explained as follows:

Each sample point has two dimensions X and Y . Thus, for i th point, Y_i is the actual value and \hat{Y}_i is the estimated value obtained from the line. We shall call the line as the line of best fit if the total sum of squares is least for all these points.

$$\sum (Y_i - \hat{Y}_i)^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + (Y_3 - \hat{Y}_3)^2 + \dots + (Y_n - \hat{Y}_n)^2$$

Since the criterion used for selecting the best fit line is based upon the fact that the squares of the residuals should be least, the regression equation is known as least square regression equation. This method of developing regression equation is known as ordinary least square method (OLS) or simply least square method.

Computation of Regression Coefficients by Least Square Methods

Least square means that the criterion used to select the best fitting line is that the sum of the squares of the residuals should be least.

In other words, the least squares regression equation is the line for which the sum of squared residuals $\sum (Y_i - \hat{Y}_i)^2$ is least.

The line of best fit is chosen on the basis of some algebra based on the concept of differentiation and solving the normal equations. We can compute the regression coefficient B and regression constant C so that the sum of the squared residuals is minimized. The procedure is as follows:

Consider a set of n data points $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, then the regression line is

$$\hat{Y}_i = BX_i + C \quad (5.16)$$

and the actual value of Y_i can be obtained by the model

$$Y_i = \hat{Y}_i + \varepsilon_i \quad (5.17)$$

where Y_i is the actual value and \hat{Y}_i is the estimated value obtained from the regression line shown in Eq. (5.16). The ε_i is the amount of error in estimating Y_i .

Our effort is to minimize the error $\varepsilon_i \forall i$, so as to get the best fit of the regression line. This can be done by minimizing the sum of the squared deviation S^2 as shown below:

$$S^2 = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - BX_i - C)^2 \quad (5.18)$$

The coefficients B and C are so chosen that S^2 is minimized. This can be done by differentiating equation (5.18) first with respect to B and then with respect to C and equating the results to zero.

Thus,

$$\frac{\partial S^2}{\partial B} = -2 \sum_{i=1}^n X_i (Y_i - BX_i - C) = 0$$

$$\text{and } \frac{\partial S^2}{\partial C} = -2 \sum_{i=1}^n (Y_i - BX_i - C) = 0$$

Solving these equations, we get

$$\sum_{i=1}^n X_i (Y_i - BX_i - C) = 0$$

and $\sum_{i=1}^n (Y_i - BX_i - C) = 0$

Taking the summation inside the bracket, the equations become

$$B \sum_{i=1}^n X_i^2 + C \sum_{i=1}^n X_i = \sum_{i=1}^n X_i Y_i \quad (5.19)$$

$$B \sum_{i=1}^n X_i + nC = \sum_{i=1}^n Y_i \quad (5.20)$$

The above two equations are known as normal equations having two unknowns B and C .

After solving these equations for B and C ,

we get

$$B = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}$$

and

$$C = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N \sum X^2 - (\sum X)^2}$$

Assumptions Used in Linear Regression

In using the linear regression model, the following assumptions must be satisfied:

1. Both the variables X and Y must be measured on either interval or ratio scale.
2. The regression model is linear in nature.
3. Error terms in estimating the dependent variable are independent and normally distributed.
4. Error distribution in predicting the dependent variable is constant irrespective of the values of X .

Multiple Regression

Estimating a phenomenon is always a complex procedure and depends upon numerous factors. Therefore, complex statistical techniques are needed which can deal with interval or ratio data and can forecast for future outcomes. Ordinary least square method which is widely used in case of simple regression is also most widely used in case of predicting the value of dependent variable from the values of two or more independent variables. Regression equation in which dependent variable is estimated by using two or more independent variables is known as

multiple regression. Multiple regression equation having four independent variables looks like

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

where

Y is a dependent variable

$X_1, X_2, X_3,$ and X_4 are the independent variables

a represents regression constant

$b_1, b_2, b_3,$ and b_4 are the unstandardized regression coefficients

Procedure in Multiple Regression

In developing the multiple regression equation, independent variables must be carefully chosen. Only those variables should be included in the study which is supposed to explain some variation in the dependent variable. Further, one must ensure that high degree of correlations does not exist among the independent variables so as to avoid the multicollinearity.

The following steps are used in multiple regression analysis:

1. Compute descriptive statistics like mean, standard deviation, skewness, kurtosis, frequency distribution, etc., and check the distribution of each variable by testing the significance of skewness and kurtosis.
2. Assess the linearity of each independent variable with the dependent variable by plotting the scatter diagram.
3. Check for multicollinearity among the independent variables by computing the correlation matrix among the independent variables. If multicollinearity exists between the independent variables then one of the independent variables must be dropped as it does not explain additional variability in the dependent variable.
4. Develop a regression equation by using the unstandardized regression coefficients (B coefficients).
5. Test the significance of the regression coefficients by using the t -test. As a rule of thumb, a t -value greater than 2.0 is usually statistically significant but one must consult a t -table to be sure.
6. Test the significance of the regression model by using the F -test. The F -value is computed by dividing the explained variance by the unexplained variance. In general, an F -value of greater than 4.0 is usually statistically significant, but one must consult an F -table to be sure.
7. Compute R^2 and adjusted R^2 to know the percentage variance of the dependent variable as explained by all the independent variables together in the regression model.

Limitations of Multiple Regression

There are certain limitations of multiple regression which are as follows:

1. Like simple regression, multiple regression also will not be efficient if the independent variables are not linearly related with dependent variable.
2. Multiple regression can be used only if the variables are either measured on interval or ratio scale. In case the data is measured on some other scale, other methods should be used for estimation.
3. Simple regression having one dependent and one independent variable usually requires a minimum of 30 observations. In general, add minimum of at least 10 observations for each additional independent variable added in the study.

What Happens If the Multicollinearity Exists Among the Independent Variables?

While doing multiple regression if multicollinearity exists, the following things may happen:

1. The F -test for the multiple regression equation shows significance, but none of the t -ratios for the regression coefficients will be statistically significant.
2. By adding any additional variable in the equation, the size or the sign of the regression coefficients of other independent variables may radically change.

In case the multicollinearity is noted between any two independent variables, one may either drop one of the two independent variables or simply show in their findings that the multicollinearity is present.

Unstandardized and Standardized Regression Coefficients

Unstandardized regression coefficients are usually known as B coefficients, whereas standardized regression coefficients are denoted as β (beta) coefficients. B coefficient explains the slopes of the regression lines. It indicates the *amount of change* in the dependent variable (Y) that is associated with a change in one unit of the independent variable (X). All B coefficients are known as unstandardized coefficients because the magnitude of their values is relative to the means and standard deviations of the independent and dependent variables in the equation. In other words, the slopes can be interpreted *directly* in terms of the raw values of X and Y . Because the value of a B coefficient depends on the scaling of the raw data, therefore it varies if the unit of the independent variable varies. For example, the magnitude of B coefficient keeps changing if the unit of the independent variable time changes as days, hours, minutes, etc. Since B coefficients depend upon the units of the independent variables, it cannot be easily compared within a regression equation. Thus, unstandardized regression coefficients cannot be used to find the relative importance of the independent variables in explaining the variability of the dependent variable in the model.

In order to compare the relative contribution of the independent variables in the regression model, another type of regression coefficient, beta (β), is used. These beta coefficients are standardized coefficients such that it adjusts for the different means and variances of the independent variables in the regression model. The standardized regression coefficients in any regression equation are measured on the same scale on 0 to 1. Thus, these standardized regression coefficients can be directly compared to one another, with the largest coefficient indicating the corresponding independent variable having the maximum influence on the dependent variable.

Procedure of Multiple Regression in SPSS

In using SPSS for regression analysis, the regression coefficients are computed in the output. Significance of these regression coefficients are tested by means of *t*-test. The regression coefficient becomes significant at 5% level if its significance value (*p* value) provided in the output is less than .05. Significance of regression coefficient indicates that the corresponding variable significantly explains the variation in the dependent variable and it contributes to the regression model. *F*-test is computed in the output to test the significance of overall model whereas R^2 and adjusted R^2 show the percentage variability in the dependent variable as explained by all the independent variables together in the model. Further, standardized regression coefficients are computed in the output to find the relative predictability of the independent variables in the model.

Methods of Regression Analysis

While doing regression analysis, the independent variables are selected either on the basis of literature or some known information. In conducting a regression study, a large number of independent variables are selected, and, therefore, there is a need to identify only those independent variables which explain the maximum variation in the dependent variable. This can be done by following any of the two methods, namely, “stepwise regression” or “Enter” method in SPSS.

Stepwise Regression Method

This method is used in exploratory regression analysis where a larger number of independent variables are investigated and the researcher does not have much idea about the relationship of these variables with that of the dependent variable. In stepwise regression analysis, the independent variables are selected one by one depending upon the relative importance in the regression model. In other words, the first entered variable in the model has the largest contribution in explaining variability in the dependent variable. A variable is included in the model if its regression coefficient is significant at 5% level. Thus, if the stepwise regression method is selected for regression analysis, the variables are selected one by one and finally the regression coefficients of the retained variables are generated in the output. These regression coefficients are used to develop the required regression equation.

Enter Method

This method is used in confirmatory regression analysis in which an already developed regression model is tested for its validity on the similar sample group for which it was earlier developed. In this procedure a regression model is developed by selecting all the independent variables in the study. The computed value of R^2 is used to assess whether the developed model is valid for the population for which it is tested.

Application of Regression Analysis

The main focus of any industry is to maximize the profits by controlling different strategic parameters. Optimum processes are identified, employees are motivated, incentives are provided to sales force, and human resources are strengthened to enhance the productivity and improve profit scenario. All these situations lead to an exploratory study where the end result is estimated on the basis of certain independent parameters. For instance, if one decides to know what all parameters are required to boost the sales figure in an organization, then a regression study may be planned. The parameters like employee's incentives, retailer's margin, user's schemes, product info, advertisement expenditure, and socioeconomic status may be studied to develop the regression model. Similarly, regression analysis may be used to identify the parameters responsible for job satisfaction in the organization. In such case, parameters like employee's salary, motivation, incentives, medical facility, family welfare incentives, and training opportunity may be selected as independent variables for developing regression model for estimating the job satisfaction of an employee. Regression analysis may identify independent variables which may be used for developing strategies in production process, inventory control, capacity utilization, sales criteria, etc. Further, regression analysis may be used to estimate the value of dependent variable at some point of time if the values of independent variables are known. This is more relevant in a situation where the value of dependent variable is difficult to know. For instance, in launching a new product in a particular city, one cannot know the sales figure, and accordingly it may affect the decision of stock inventory. By using the regression model on sales, one can estimate the sales figure in a particular month.

Solved Example of Multiple Regression Analysis Including Multiple Correlation

Example 5.3 In order to assess the feasibility of a guaranteed annual wage, the Rand Corporation conducted a study to assess the response of labor supply in terms of average hours of work (Y) based on different independent parameters. The data were drawn from a national sample of 6,000 households with male head earnings less than \$15,000 annually. These data are given in Table 5.3. Apply regression

analysis by using SPSS to suggest a regression model for estimating the average hours worked during the year based on identified independent parameters.

Solution To develop the regression model for estimating the average hours of working during the year for guaranteed wages on the basis of socioeconomic variables, do the following steps:

- (i) Choose the “stepwise regression” method in SPSS to get the regression coefficients of the independent variables identified in the model for developing the regression equation.
- (ii) Test the regression coefficients for its significance through *t*-test by using its significance value (*p* value) in the output.
- (iii) Test the regression model for its significance through the *F*-value by looking to its significance value (*p* value) in the output.
- (iv) Use the value of R^2 in the output to know the amount of variance explained in the dependent variable by the identified independent variables together in the model.

Steps involved in getting the output of regression analysis by using SPSS have been explained in the following sections.

Computation of Regression Coefficients, Multiple Correlation, and Other Related Output in the Regression Analysis

(a) *Preparing Data File*

Before using the SPSS commands for different output of regression analysis, the data file needs to be prepared.

The following steps will help you to prepare the data file:

- (i) *Starting SPSS*: Use the following command sequence to start SPSS:

Start → All Programs → IBM SPSS Statistics → IBM SPSS Statistics 20

After clicking the **Type in Data**, you will be taken to the **Variable View** option for defining the variables in the study.

- (ii) *Defining variables*: There are nine variables in this exercise which need to be defined in SPSS first. Since all these variables were measured on interval scale, they will be defined as “Scale” variable in SPSS. The procedure of defining the variables in SPSS is as follows:
 1. Click **Variable View** to define variables and their properties.
 2. Write short name of these variables, that is, *Hours*, *Rate*, *Ers*, *Erno*, *Nein*, *Assets*, *Age*, *Dep*, and *School* under the column heading **Name**.
 3. Full name of these variables may be defined as *Average hours worked during the year*, *Average hourly wage in dollars*, *Average yearly*

Table 5.3 Data on average yearly hour and other socioeconomic variables

S.N.	Hours (X_1)	Rate (X_2)	ERSP (X_3)	ERNO (X_4)	NEIN (X_5)	Assets (X_6)	Age (X_7)	DEP (X_8)	School (X_9)
1	2,157	2.905	1,121	291	380	7,250	38.5	2.340	10.5
2	2,174	2.970	1,128	301	398	7,744	39.3	2.335	10.5
3	2,062	2.350	1,214	326	185	3,068	40.1	2.851	8.9
4	2,111	2.511	1,203	49	117	1,632	22.4	1.159	11.5
5	2,134	2.791	1,013	594	730	12,710	57.7	1.229	8.8
6	2,185	3.040	1,135	287	382	7,706	38.6	2.602	10.7
7	2,210	3.222	1,100	295	474	9,338	39.0	2.187	11.2
8	2,105	2.493	1,180	310	255	4,730	39.9	2.616	9.3
9	2,267	2.838	1,298	252	431	8,317	38.9	2.024	11.1
10	2,205	2.356	885	264	373	6,789	38.8	2.662	9.5
11	2,121	2.922	1,251	328	312	5,907	39.8	2.287	10.3
12	2,109	2.499	1,207	347	271	5,069	39.7	3.193	8.9
13	2,108	2.796	1,036	300	259	4,614	38.2	2.040	9.2
14	2,047	2.453	1,213	297	139	1,987	40.3	2.545	9.1
15	2,174	3.582	1,141	414	498	10,239	40.0	2.064	11.7
16	2,067	2.909	1,805	290	239	4,439	39.1	2.301	10.5
17	2,159	2.511	1,075	289	308	5,621	39.3	2.486	9.5
18	2,257	2.516	1,093	176	392	7,293	37.9	2.042	10.1
19	1,985	1.423	553	381	146	1,866	40.6	3.833	6.6
20	2,184	3.636	1,091	291	560	11,240	39.1	2.328	11.6
21	2,084	2.983	1,327	331	296	5,653	39.8	2.208	10.2
22	2,051	2.573	1,194	279	172	2,806	40.0	2.362	9.1
23	2,127	3.262	1,226	314	408	8,042	39.5	2.259	10.8
24	2,102	3.234	1,188	414	352	7,557	39.8	2.019	10.7
25	2,098	2.280	973	364	272	4,400	40.6	2.661	8.4
26	2,042	2.304	1,085	328	140	1,739	41.8	2.444	8.2
27	2,181	2.912	1,072	304	383	7,340	39.0	2.337	10.2
28	2,186	3.015	1,122	30	352	7,292	37.2	2.046	10.9
29	2,188	3.010	990	366	374	7,325	38.4	2.847	10.6
30	2,077	1.901	350	209	951	370	37.4	4.158	8.2
31	2,196	3.009	947	294	342	6,888	37.5	3.047	10.6
32	2,093	1.899	342	311	120	1,425	37.5	4.512	8.1
33	2,173	2.959	1,116	296	387	7,625	39.2	2.342	10.5
34	2,179	2.971	1,128	312	397	7,779	39.4	2.341	10.5
35	2,200	2.980	1,126	204	393	7,885	39.2	2.341	10.6

Source: D. H. Greenberg and M. Kosters, Income Guarantees and the Working Poor, The Rand Corporation, R-579-OEO, December 1970.

Hours(X_1): average hours worked during the year

Rate(X_2): average hourly wage (dollars)

ERSP(X_3): average yearly earnings of spouse (dollars)

ERNO(X_4): average yearly earnings of other family members (dollars)

NEIN(X_5): average yearly non-earned income

Assets(X_6): average family asset holdings (bank account) (dollars)

Age(X_7): average age of respondent

Dep(X_8): average number of dependents

School(X_9): average highest grade of school completed

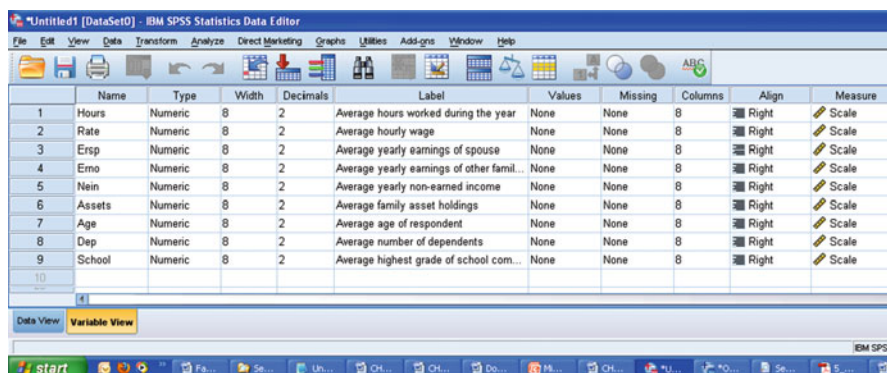


Fig. 5.2 Defining variables along with their characteristics

*earnings of spouse in dollars, Average yearly earnings of other family members in dollars, Average yearly non-earned income, Average family asset holdings (bank account, etc.) in dollars, Average age of respondent, Average number of dependents, and Average highest grade of school completed under the column heading **Label**.*

4. Under the column heading **Measure**, select the option “Scale” for all these variables.
5. Use default entries in all other columns.

After defining these variables in variable view, the screen shall look like Fig. 5.2.

- (iii) **Entering data:** After defining these variables in the **Variable View**, click **Data View** on the left bottom of the screen to enter data. For each variable, enter the data column wise. After entering data, the screen will look like Fig. 5.3. Save the data file in the desired location before further processing.

(b) **SPSS Commands for Computing Correlation Coefficient**

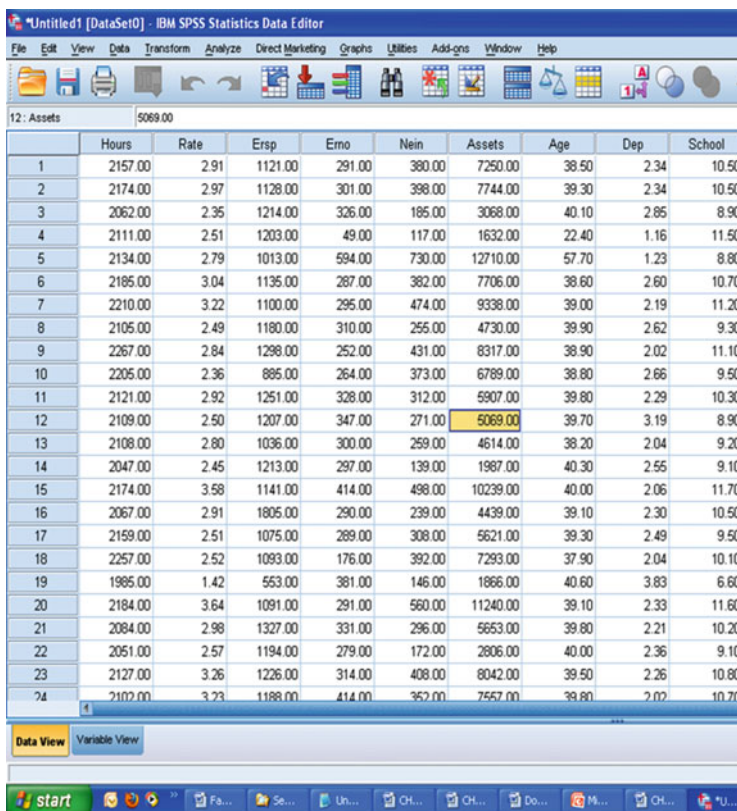
After preparing the data file in data view, take the following steps for regression analysis. Data file in SPSS can also be prepared by transporting the data from the other format like EXCEL or ASCII. The procedure of transporting data from other formats has been explained in Chap. 1.

- (i) **Initiating the SPSS commands for regression analysis:** In data view, choose the following commands in sequence:

Analyze → Regression → Linear

The screen shall look like Fig. 5.4.

- (ii) **Selecting variables for regression analysis:** After clicking the **Linear** option, you will be taken to the next screen as shown in Fig. 5.5 for selecting the variables for regression analysis. Select the variable *Average hours worked during the year* (dependent variable) from left panel to the “Dependent” section of the right panel. Select all independent variables from left panel to the “Independent(s)” section of the right panel.



12: Assets 5069.00

	Hours	Rate	Ersp	Erno	Nein	Assets	Age	Dep	School
1	2157.00	2.91	1121.00	291.00	380.00	7250.00	38.50	2.34	10.50
2	2174.00	2.97	1128.00	301.00	398.00	7744.00	39.30	2.34	10.50
3	2062.00	2.35	1214.00	326.00	185.00	3068.00	40.10	2.85	8.90
4	2111.00	2.51	1203.00	49.00	117.00	1632.00	22.40	1.16	11.50
5	2134.00	2.79	1013.00	594.00	730.00	12710.00	57.70	1.23	8.80
6	2185.00	3.04	1135.00	287.00	382.00	7706.00	38.60	2.60	10.70
7	2210.00	3.22	1100.00	295.00	474.00	9338.00	39.00	2.19	11.20
8	2105.00	2.49	1180.00	310.00	255.00	4730.00	39.90	2.62	9.30
9	2267.00	2.84	1298.00	252.00	431.00	8317.00	38.90	2.02	11.10
10	2205.00	2.36	885.00	264.00	373.00	6789.00	38.80	2.66	9.50
11	2121.00	2.92	1251.00	328.00	312.00	5907.00	39.80	2.29	10.30
12	2109.00	2.50	1207.00	347.00	271.00	5069.00	39.70	3.19	8.90
13	2108.00	2.80	1036.00	300.00	259.00	4614.00	38.20	2.04	9.20
14	2047.00	2.45	1213.00	297.00	139.00	1987.00	40.30	2.55	9.10
15	2174.00	3.58	1141.00	414.00	498.00	10239.00	40.00	2.06	11.70
16	2067.00	2.91	1805.00	290.00	239.00	4439.00	39.10	2.30	10.50
17	2159.00	2.51	1075.00	289.00	308.00	5621.00	39.30	2.49	9.50
18	2257.00	2.52	1093.00	176.00	392.00	7293.00	37.90	2.04	10.10
19	1985.00	1.42	553.00	381.00	146.00	1866.00	40.60	3.83	6.60
20	2184.00	3.64	1091.00	291.00	560.00	11240.00	39.10	2.33	11.60
21	2084.00	2.98	1327.00	331.00	296.00	5653.00	39.80	2.21	10.20
22	2051.00	2.57	1194.00	279.00	172.00	2806.00	40.00	2.36	9.10
23	2127.00	3.26	1226.00	314.00	408.00	8042.00	39.50	2.26	10.80
24	2102.00	3.23	1188.00	414.00	352.00	7552.00	39.80	2.02	10.70

Fig. 5.3 Screen showing entered data for all the variables in the data view

Either the variable selection is made one by one or all at once. To do so, the variable needs to be selected from the left panel, and by arrow command, it may be brought to the right panel. After choosing the variables for analysis, the screen shall look like Fig. 5.5.

- (iii) *Selecting the options for computation:* After selecting the variables, option needs to be defined for the regression analysis. Take the following steps:
- In the screen shown in Fig. 5.5, click the tag **Statistics**; you will get the screen as shown in Fig. 5.6.
 - Check the box “*R* squared change,” “Descriptive,” and “Part and partial correlations.”
 - By default, the options “Estimates” and “Model fit” are checked. Ensure that they remain checked.
 - Click **Continue**. You will now be taken back to the screen shown in Fig. 5.5.

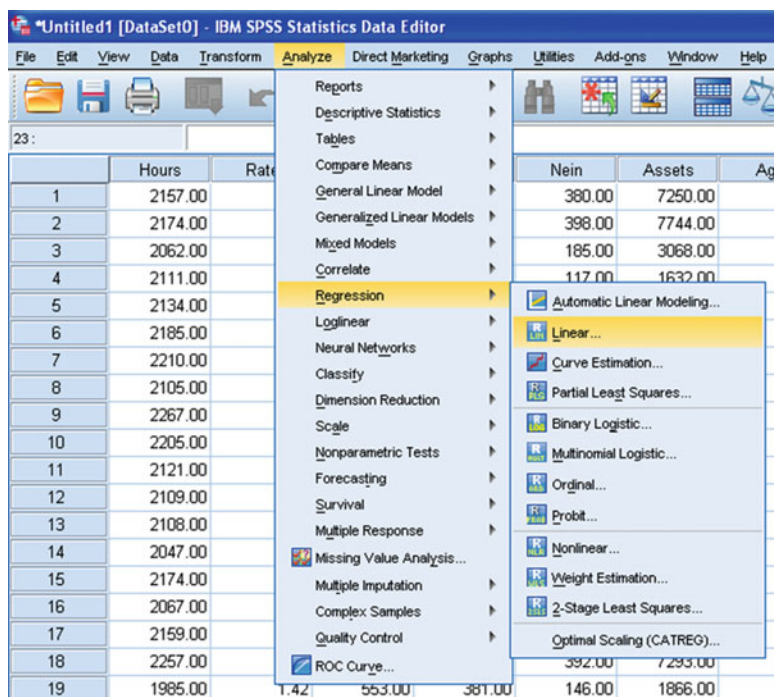


Fig. 5.4 Screen showing SPSS commands for regression analysis

By checking the option “*R* squared change” the output shall include the values of R^2 and adjusted R^2 . Similarly by checking the option “Descriptive” the output will provide the values of mean and standard deviations along with correlation matrix of all the variables, whereas checking the option “Part and partial correlations” shall provide the partial correlations of various orders between *Average hours worked during the year* and other variables. Readers are advised to try other options and see what changes they are getting in their outputs.

- In the option **Method** shown in Fig. 5.5, select “Stepwise.”
- Click **OK**.

(c) *Getting the Output*

Clicking the **OK** tag in Fig. 5.5 will lead you to the output window. In the output window of SPSS, the relevant outputs can be selected by using the right click of the mouse and may be copied in the word file. The output panel shall have the following results:

1. Mean and standard deviation
2. Correlation matrix along with significance value
3. Model summary along with the values of R , R^2 and adjusted R^2
4. ANOVA table showing F -values for all the models

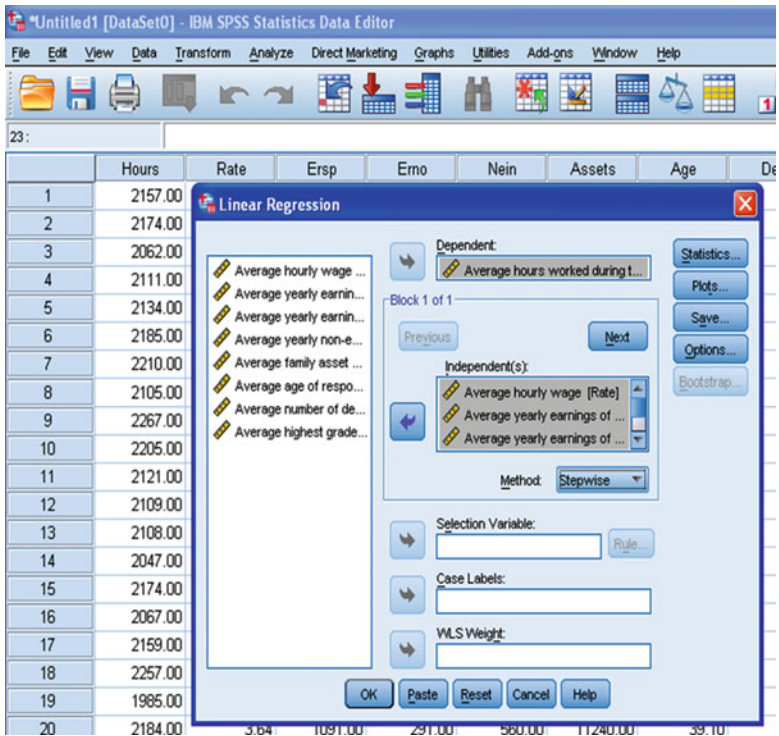


Fig. 5.5 Screen showing selection of variables for regression analysis

5. Standardized and unstandardized regression coefficients of selected variables in different models along with their *t*-values and partial correlations

In this example, all the outputs so generated by the SPSS have been shown in Tables 5.4, 5.5, 5.6, 5.7, and 5.8.

Interpretation of the Outputs

Different outputs generated in the SPSS are shown below along with their interpretations.

1. The values of mean and standard deviation for all the variables are shown in Table 5.4. These values can be used for further analysis in the study. By using the procedure discussed in Chap. 2, a profile chart may be prepared by computing other descriptive statistics for all the variables.
2. The correlation matrix in Table 5.5 shows the correlations among the variables along with their significance value (*p* value). Significance of these correlations has been tested for one-tailed test. The correlation coefficient with one asterisk

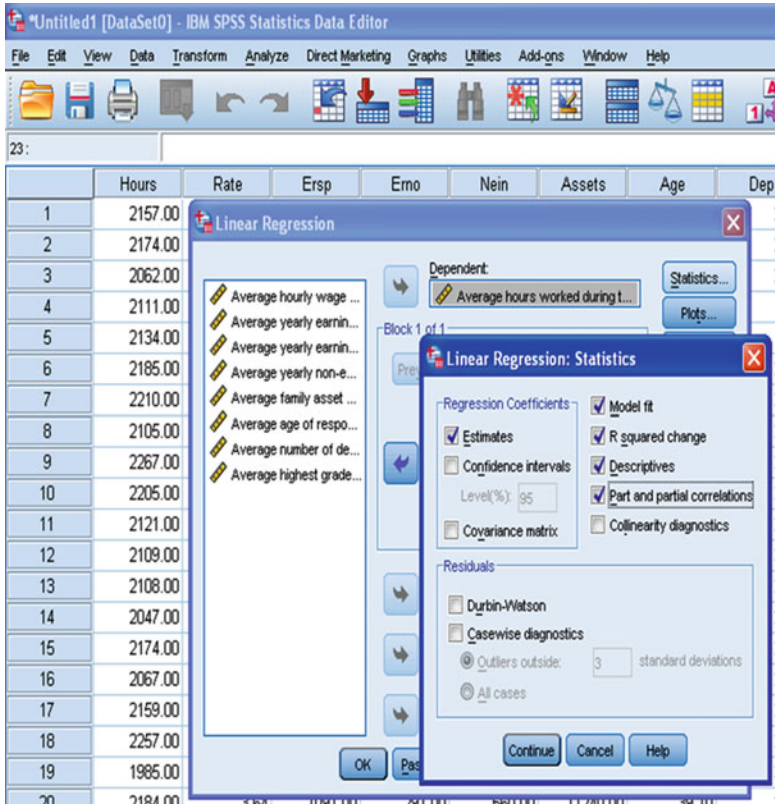


Fig. 5.6 Screen showing options for computing various components of regression analysis

Table 5.4 Descriptive statistics for different variables

Variables	Mean	SD	N
Average hours worked during the year	2,137.09	64.12	35
Average hourly wage	2.74	.46	35
Average yearly earnings of spouse	1,083.66	256.78	35
Average yearly earnings of other family members	298.23	94.99	35
Average yearly non-earned income	348.23	167.59	35
Average family asset holdings	6,048.14	2,921.40	35
Average age of respondent	39.24	4.40	35
Average number of dependents	2.49	.66	35
Average highest grade of school completed	9.92	1.17	35

mark (*) indicates its significance at 5% level. The asterisk mark (*) is put on the correlation coefficient if its value is more than the required value of correlation coefficients for its significance at 5% level which is .284. For one-tailed test, the required value of “r” for significance with 33 ($N - 2$) df can be seen from Table A.3 in the [Appendix](#).

Table 5.5 Correlation matrix for different variables along with significance level

	Hours	Rate	Ersp	Erno	Nein	Assets	Age	Dep	School
<i>Pearson correlation</i>									
<i>Hours</i>	1.000	.556**	.124	−.245	.413**	.716**	−.077	−.339*	.681**
<i>Rate</i>	.556**	1.000	.572**	.059	.297*	.783**	.044	−.601**	.881**
<i>Ersp</i>	.124	.572**	1.000	−.041	−.238	.298*	−.015	−.693**	.549**
<i>Erno</i>	−.245	.059	−.041	1.000	.152	.296*	.775**	.050	−.299*
<i>Nein</i>	.413**	.297*	−.238	.152	1.000	.512**	.347*	−.045	.219
<i>Assets</i>	.716**	.783**	.298*	.296*	.512**	1.000	.414**	−.530**	.634**
<i>Age</i>	−.077	.044	−.015	.775**	.347*	.414**	1.000	−.048	−.331
<i>Dep</i>	−.339*	−.601**	−.693**	.050	−.045	−.530**	−.048	1.000	−.603**
<i>School</i>	.681**	.881**	.549**	−.299*	.219	.634**	−.331*	−.603**	1.000
<i>Sig. (1-tailed)</i>									
<i>Hours</i>		.000	.239	.078	.007	.000	.330	.023	.000
<i>Rate</i>	.000	.	.000	.368	.041	.000	.401	.000	.000
<i>Ersp</i>	.239	.000	.	.408	.084	.041	.465	.000	.000
<i>Erno</i>	.078	.368	.408	.	.192	.042	.000	.387	.041
<i>Nein</i>	.007	.041	.084	.192	.	.001	.021	.398	.103
<i>Assets</i>	.000	.000	.041	.042	.001	.	.007	.001	.000
<i>Age</i>	.330	.401	.465	.000	.021	.007	.	.391	.026
<i>Dep</i>	.023	.000	.000	.387	.398	.001	.391	.	.000
<i>School</i>	.000	.000	.000	.041	.103	.000	.026	.000	.

Hours: Average hours worked during the year
Rate: Average hourly wage
Ersp: Average yearly earnings of spouse
Erno: Average yearly earnings of other family members
Nein: Average yearly non-earned income
Assets: Average family asset holdings
Age: Average age of respondent
Dep: Average number of dependents
School: Average highest grade of school completed
*Significant at 0.05 level (1-tailed) Significant value of *r* at .05 level with 33 df (1-tailed) = 0.284;
**Significant at 0.01 level (1-tailed) Significant value of *r* at .01 level with 33 df (1-tailed) = 0.392

Similarly for one-tailed test, the significance value for the correlation coefficient at .01 level with 33 (=N − 2) df can be seen as 0.392. Thus, all those correlation coefficients having values more than 0.392 are significant at 1% level. Such correlation coefficients have been shown with two asterisk marks (**).
Readers may also show the correlation matrix by writing the upper diagonal values as has been done in Chap. 4.

3. From Table 5.5, it can be seen that *Hours* (Average hours worked during the year) is significantly correlated with *Rate* (Average hourly wage), *Nein* (Average yearly non-earned income), *Assets* (Average family asset holdings), and *School* (Average highest grade of school completed) at 1% level, whereas with *Dep* (Average number of dependents) at 5% level.

Table 5.6 Model summary along with the values of R and R square

Model	R	R square	Adj R square.	SE of the estimate	Change statistics				
					R square change	F change	df1	df2	Sig. F change
1	.716 ^a	.512	.498	45.44102	.512	34.687	1	33	.000
2	.861 ^b	.742	.726	33.58681	.229	28.405	1	32	.000
3	.879 ^c	.773	.751	32.00807	.031	4.235	1	31	.048

^aPredictors: (Constant), Average family asset holdings

^bPredictors: (Constant), Average family asset holdings, Average yearly earnings of other family members

^cPredictors: (Constant), Average family asset holdings, Average yearly earnings of other family members, Average number of dependents

- The three regression models generated by the SPSS have been presented in Table 5.6. In the third model, the value of R^2 is .773, which is maximum, and, therefore, third model shall be used to develop the regression equation. It can be seen from Table 5.6 that in the third model, three independent variables, namely, *Assets* (Average family asset holdings), *Erno* (Average yearly earnings of other family members), and *Dep* (Average number of dependents), have been identified, and, therefore, the regression equation shall be developed using these three variables only. The R^2 value for this model is 0.773, and, therefore, these three independent variables explain 77.3% variations in *Hours* (Average hours worked during the year) in the USA. Thus, this model can be considered appropriate to develop the regression equation.
- In Table 5.7, F -values for all the models have been shown. Since F -value for the third model is highly significant, it may be concluded that the model selected is highly efficient also.
- Table 5.8 shows the unstandardized and standardized regression coefficients in all the three models. Unstandardized coefficients are also known as “B” coefficients and are used to develop the regression equation whereas standardized regression coefficients are denoted by “ β ” and are used to explain the relative importance of independent variables in terms of their contribution toward the dependent variables in the model. In the third model, t -values for all the three regression coefficients are significant as their significance values (p values) are less than .05. Thus, it may be concluded that the variables *Assets* (Average family asset holdings), *Erno* (Average yearly earnings of other family members), and *Dep* (Average number of dependents) significantly explain the variations in the *Hours* (Average hours worked during the year).

Regression Equation

Using unstandardized regression coefficients (B) of the third model shown in Table 5.8, the regression equation can be developed which is as follows:

$$\text{Hours} = 2064.285 + 0.22 \times (\text{Assets}) - 0.371 \times (\text{Erno}) + 20.816 \times (\text{Dep})$$

Table 5.7 ANOVA table showing *F*-values for all the models^a

Model		Sum of squares	df	Mean square	<i>F</i>	Sig.
1	Regression	71,625.498	1	71,625.498	34.687	.000 ^b
	Residual	68,141.245	33	2,064.886		
	Total	139,766.743	34			
2	Regression	103,668.390	2	51,834.195	45.949	.000 ^c
	Residual	36,098.353	32	1,128.074		
	Total	139,766.743	34			
3	Regression	108,006.736	3	36,002.245	35.141	.000 ^d
	Residual	31,760.007	31	1,024.516		
	Total	139,766.743	34			

^aDependent variable: Average hours worked during the year
^bPredictors: (Constant), Average family asset holdings
^cPredictors: (Constant), Average family asset holdings, Average yearly earnings of other family members
^dPredictors: (Constant), Average family asset holdings, Average yearly earnings of other family members, Average number of dependents

where
Hours: Average hours worked during the year
Assets: Average family asset holdings
Erno: Average yearly earnings of other family members
Dep: Average number of dependents

Thus, it may be concluded that the above regression equation is quite reliable as the value of R^2 is 0.773. In other words, the three variables selected in this regression equation explain 77.3% of the total variability in the *Hour* (Average hours worked during the year), which is quite good. Since the *F*-value for this regression model is highly significant, the model is reliable. At the same time, all the regression coefficients in this model are highly significant, and, therefore, it may be interpreted that all the three variables selected in the model, namely, *Assets* (Average family asset holdings), *Erno* (Average yearly earnings of other family members), and *Dep* (Average number of dependents), have significant predictability in estimating the value of the *Hour* (Average hours worked during the year) in the USA.

Summary of the SPSS Commands For Regression Analysis

1. Start SPSS by using the following commands:
Start → **All Programs** → **IBM SPSS Statistics** → **IBM SPSS Statistics**
2. Create data file by clicking the tag **Type in Data**. Define all the variables and their characteristics by clicking the **Variable View**. After defining the variables, type the data for these variables by clicking **Data View**.

Table 5.8 Regression coefficients of selected variables in different models along with their *t*-values and partial correlations^a

Model	Unstandardized coefficients			Standardized coefficients		Correlations			
	<i>B</i>	Std. error	<i>t</i>	Sig.	Beta	Zero-order	Partial	Part	
1	(Constant) Average family asset holdings	2,042.064 .016	17.869 .003	114.280 5.890	.000 .000	.716 .716	.716 .716	.716 .716	
2	(Constant) Average family asset holdings Average yearly earnings of other family members	2,123.257 .019 -.338	20.162 .002 .063	105.308 9.190 -5.330	.000 .000 .000	.716 .716 -.245	.852 .852 -.686	.826 .826 -.479	
3	(Constant) Average family asset holdings Average yearly earnings of other family members Average number of dependents	2,064.285 .022 -.371 20.816	34.503 .002 .063 10.116	59.828 9.092 -5.933 2.058	.000 .000 .000 .048	.716 .716 -.245 -.339	.853 .853 -.729 .347	.778 .778 -.508 .176	

^aDependent variable: Average hours worked during the year

3. Once the data file is ready, use the following command sequence for selecting the variables for analysis.

Analyze → Regression → Linear

4. Select the dependent variable from left panel to the “Dependent” section of the right panel. Select all other independent variables from left panel to the “Independent(s)” section of the right panel.
5. After selecting the variables for regression analysis, click the tag **Statistics** on the screen. Check the box “*R* squared change,” “Descriptive,” and “Part and partial correlations.” Press **Continue**.
6. In the **Method** option, select “Stepwise,” then press **OK** to get the different outputs for regression analysis.

Exercise

Short-Answer Questions

Note: Write answer to each of the questions in not more than 200 words.

- Q.1. Describe regression analysis. Explain the difference between simple regression and multiple regression models.
- Q.2. What is the difference between stepwise regression and backward regression?
- Q.3. Discuss the role of R^2 in regression analysis. Explain multiple correlation and its order.
- Q.4. Explain an experimental situation where regression analysis can be used.
- Q.5. How will you know that the variables which are selected in the regression analysis are valid?
- Q.6. What is the difference between Stepwise and Enter method in developing multiple regression equation?

Multiple-Choice Questions

Note: For each of the question, there are four alternative answers. Tick mark the one that you consider the closest to the correct answer.

1. The range of multiple correlation R is
 - (a) -1 to 0
 - (b) 0 to 1
 - (c) -1 to 0
 - (d) None of the above
2. SPSS commands for multiple regression analysis is
 - (a) Analyze -> Linear -> Regression
 - (b) Analyze -> Regression -> Linear
 - (c) Analyze -> Linear Regression
 - (d) Analyze -> Regression Linear

3. Choose the most appropriate statement

- (a) R^2 is a measure of multiple correlation.
- (b) R^2 is used for selecting the variables in the regression model.
- (c) R^2 is the amount of variability explained in the dependent variable by the independent variables.
- (d) All above are correct.

4. If p value for the correlation between college GPA and GMAT score is .008, what conclusion can be drawn?

- (a) Correlation is not significant at 1% level.
- (b) Correlation is not significant at 5% level.
- (c) Correlation is significant at 1% level.
- (d) All above statements are wrong.

5. Following are the two statements about the significance of value of r :

Statement I: Correlation coefficient required for significance at 1% is 0.462.

Statement II: Correlation coefficient required for significance at 5% is 0.337.

Choose the most appropriate alternative.

- (a) Statement I is right, but II is wrong.
- (b) Statement I is wrong, but II is right.
- (c) Both statements I and II are wrong.
- (d) Both statements are right.

6. In regression analysis, four models have been developed. Which model in your opinion is the most appropriate?

Models	No. of independent variables	R^2
(a) Model I:	5	0.88
(b) Model II:	4	0.87
(c) Model III:	3	0.86
(d) Model IV:	2	0.65

7. In a regression analysis to estimate the sale of a particular product, the regression coefficients of independent variables were as follows:

Independent variables	B coefficient	p value
Customer's incentive	1.5	.06
Dealer's incentive	2.2	.009
Hours of marketing	3.1	.32
Product price	1.2	.006

Choose the most appropriate statement.

- (a) Both Customer's incentive and Hours of marketing are significant at .05 level in the model.
- (b) Both Hours of marketing and Product price are significant at .05 level in the model.

- (c) Both Dealer's incentive and Product price are significant at .01 level in the model.
- (d) Both Dealer's incentive and Product price are not significant at .05 level in the model.

8. Choose correct statement about B and β coefficients.

- (a) "B" is an unstandardized coefficient and " β " is a standardized coefficient.
- (b) " β " is an unstandardized coefficient and "B" is a standardized coefficient.
- (c) Both "B" and " β " are standardized coefficients.
- (d) Both "B" and " β " are unstandardized coefficients.

Assignments

1. The data on copper industry and its determinants in the US market during 1951–1980 are shown in the following table. Construct a regression model and develop the regression equation by using the SPSS. Test the significance of regression coefficients and explain the robustness of the regression model to predict the price of the copper in the US market.

Determinants of US domestic price of copper					
DPC	GNP	IIP	MEPC	NOH	PA
21.89	330.2	45.1	220.4	1,491.00	19.00
22.29	347.2	50.9	259.5	1,504.00	19.41
19.63	366.1	53.3	256.3	1,438.00	20.93
22.85	366.3	53.6	249.3	1,551.00	21.78
33.77	399.3	54.6	352.3	1,646.00	23.68
39.18	420.7	61.1	329.1	1,349.00	26.01
30.58	442.0	61.9	219.6	1,224.00	27.52
26.30	447.0	57.9	234.8	1,382.00	26.89
30.70	483.0	64.8	237.4	1,553.70	26.85
32.10	506.0	66.2	245.8	1,296.10	27.23
30.00	523.3	66.7	229.2	1,365.00	25.46
30.80	563.8	72.2	233.9	1,492.50	23.88
30.80	594.7	76.5	234.2	1,634.90	22.62
32.60	635.7	81.7	347.0	1,561.00	23.72
35.40	688.1	89.8	468.1	1,509.70	24.50
36.60	753.0	97.8	555.0	1,195.80	24.50
38.60	796.3	100.0	418.0	1,321.90	24.98
42.20	868.5	106.3	525.2	1,545.40	25.58
47.90	935.5	111.1	620.7	1,499.50	27.18
58.20	982.4	107.8	588.6	1,469.00	28.72
52.00	1,063.4	109.6	444.4	2,084.50	29.00
51.20	1,171.1	119.7	427.8	2,378.50	26.67
59.50	1,306.6	129.8	727.1	2,057.50	25.33
77.30	1,412.9	129.3	877.6	1,352.50	34.06
64.20	1,528.8	117.8	556.6	1,171.40	39.79
69.60	1,700.1	129.8	780.6	1,547.60	44.49
66.80	1,887.2	137.1	750.7	1,989.80	51.23

(continued)

(continued)

Determinants of US domestic price of copper					
DPC	GNP	IIP	MEPC	NOH	PA
66.50	2,127.6	145.2	709.8	2,023.30	54.42
98.30	2,628.80	152.5	935.7	1,749.20	61.01
101.40	2,633.10	147.1	940.9	1,298.50	70.87

DPC = 12-month average US domestic price of copper (cents per pound)

GNP = annual gross national product (\$, billions)

IIP = 12-month average index of industrial production

MEPC = 12-month average London Metal Exchange price of copper (pounds sterling)

NOH = number of housing starts per year (thousands of units)

PA = 12-month average price of aluminum (cents per pound)

Note: The data are from the sources such as American Metal Market, Metals Week, and US Department of Commerce publications

Note: The data were collected by Gary R. Smith from sources such as American Metal Market, Metals

2. Data in the following table shows the crime rate in 47 states in the USA in 1960. Develop a suitable regression model for estimating the crime rate depending upon identified socioeconomic variables.

US crime data for 47 states										
S.N.	R	Age	ED	EX0	LF	N	NW	U1	U2	X
1	79.1	151	91	58	510	33	301	108	41	261
2	163.5	143	113	103	583	13	102	96	36	194
3	57.8	142	89	45	533	18	219	94	33	250
4	196.9	136	121	149	577	157	80	102	39	167
5	123.4	141	121	109	591	18	30	91	20	174
6	68.2	121	110	118	547	25	44	84	29	126
7	96.3	127	111	82	519	4	139	97	38	168
8	155.5	131	109	115	542	50	179	79	35	206
9	85.6	157	90	65	553	39	286	81	28	239
10	70.5	140	118	71	632	7	15	100	24	174
11	167.4	124	105	121	580	101	106	77	35	170
12	84.9	134	108	75	595	47	59	83	31	172
13	51.1	128	113	67	624	28	10	77	25	206
14	66.4	135	117	62	595	22	46	77	27	190
15	79.8	152	87	57	530	30	72	92	43	264
16	94.6	142	88	81	497	33	321	116	47	247
17	53.9	143	110	66	537	10	6	114	35	166
18	92.9	135	104	123	537	31	170	89	34	165
19	75	130	116	128	536	51	24	78	34	135
20	122.5	125	108	113	567	78	94	130	58	166
21	74.2	126	108	74	602	34	12	102	33	195
22	43.9	157	89	47	512	22	423	97	34	276
23	121.6	132	96	87	564	43	92	83	32	227
24	96.8	131	116	78	574	7	36	142	42	176

(continued)

(continued)

US crime data for 47 states										
S.N.	R	Age	ED	EX0	LF	N	NW	U1	U2	X
25	52.3	130	116	63	641	14	26	70	21	196
26	199.3	131	121	160	631	3	77	102	41	152
27	34.2	135	109	69	540	6	4	80	22	139
28	121.6	152	112	82	571	10	79	103	28	215
29	104.3	119	107	166	521	168	89	92	36	154
30	69.6	166	89	58	521	46	254	72	26	237
31	37.3	140	93	55	535	6	20	135	40	200
32	75.4	125	109	90	586	97	82	105	43	163
33	107.2	147	104	63	560	23	95	76	24	233
34	92.3	126	118	97	542	18	21	102	35	166
35	65.3	123	102	97	526	113	76	124	50	158
36	127.2	150	100	109	531	9	24	87	38	153
37	83.1	177	87	58	638	24	349	76	28	254
38	56.6	133	104	51	599	7	40	99	27	225
39	82.6	149	88	61	515	36	165	86	35	251
40	115.1	145	104	82	560	96	126	88	31	228
41	88	148	122	72	601	9	19	84	20	144
42	54.2	141	109	56	523	4	2	107	37	170
43	82.3	162	99	75	522	40	208	73	27	224
44	103	136	121	95	574	29	36	111	37	162
45	45.5	139	88	46	480	19	49	135	53	249
46	50.8	126	104	106	599	40	24	78	25	171
47	84.9	130	121	90	623	3	22	113	40	160

Source: W. Vandaele, "Participation in Illegitimate Activities: Erlich Revisited," in A. Blumstein, J. Cohen, and Nagin, D., eds., *Deterrence and Incapacitation*, National Academy of Sciences, 1978, pp. 270–335. 386

- R = crime rate, number of offenses reported to police per million population
Age = number of males of age 14–24 per 1,000 population
S = indicator variable for southern states (0 = no, 1 = yes)
ED = mean number of years of schooling times 10 for persons age 25 or older
EX0 = 1,960 per capita expenditure on police by state and local government
LF = labor force participation rate per 1,000 civilian urban males age 14–24
N = state population size in hundred thousands
NW = number of nonwhites per 1,000 population
U1 = unemployment rate of urban males per 1,000 of age 14–24
U2 = unemployment rate of urban males per 1,000 of age 35–39
X = the number of families per 1,000 earnings 1/2 the median income

Answers to Multiple-Choice Questions

Q.1	b	Q.2	b
Q.3	c	Q.4	c
Q.5	d	Q.6	c
Q.7	c	Q.8	a