21 pg

4

Chapter

MULTIPLE CORRELATION AND REGRESSION

CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Multiple and partial correlation
- Introduction of multiple linear regression, Hypothesis testing of multiple regression, Test of significance of regression, Test of individual regression coefficient
- Model adequacy tests
- Problems and illustrative examples related using software.

Partial Correlation

It is the relationship between two variables keeping all the other remaining variables involved constant. The correlation between two variables keeping one other variable constant is called first order correlation. The correlation between two variables keeping other two variables constant is called second order correlation and so on.

We are interested to study the relationship of production of wheat with seeds, fertilizer, irrigation etc. If we study the relationship between production of wheat with seeds keeping fertilizer and irrigation condition constant is the case of partial correlation. Similarly the study of relationship between production of wheat with fertilizer keeping seeds and irrigation constant, the study of relationship between production of wheat with irrigation keeping seeds and fertilizer constant are the case of partial correlation.

Let us consider three variables X_1, X_2 and X_3 then the partial correlation coefficient between X_1 and X_2 keeping X_3 constant is denoted by $r_{12\cdot3}$ and is given by $r_{12\cdot3} = \dfrac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\,.\sqrt{1 - r_{23}^2}}$

Similarly, the partial correlation coefficient between X_1 and X_3 keeping X_2 constant is denoted by $r_{13\cdot2}$ and is given by $r_{13\cdot2} = \dfrac{r_{13} - r_{12}\, r_{32}}{\sqrt{1 - r_{12}^2}\,.\sqrt{1 - r_{32}^2}}$

Also, the partial correlation coefficient between X_2 and X_3 keeping X_1 constant is denoted by $r_{23\cdot1}$ and is given by $r_{23\cdot1} = \dfrac{r_{23} - r_{21}\, r_{31}}{\sqrt{1 - r_{21}^2}\,.\sqrt{1 - r_{31}^2}}$.

Remarks:

1. (i) $r_{12} = r_{21}$ (ii) $r_{13} = r_{31}$ (iii) $r_{23} = r_{32}$
2. (i) $r_{12.3} = r_{21.3}$ (ii) $r_{13.2} = r_{31.2}$ (iii) $r_{23.1} = r_{32.1}$
3. (i) $-1 \le r_{12\cdot3} \le 1$ (ii) $-1 \le r_{13.2} \le 1$ (iii) $-1 \le r_{23.1} \le 1$
4. r_{12}, r_{13}, r_{23} are zero order correlation coefficients.
5. $r_{12.3}$, $r_{13.2}$, $r_{23.1}$ are first order correlation coefficients.
6. $r_{12.34}$, $r_{23.14}$, $r_{13.24}$, $r_{14.23}$, $r_{24.13}$, $r_{34.12}$ are second order correlation coefficients.

Coefficient of Partial Determination

It is the square of partial correlation coefficient. It is used to measure variation in one variable is explained by other variable keeping next variable constant.

If $r_{12.3} = 0.8$ then coefficient of partial determination is $r_{12.3}^2 = (0.8)^2 = 0.64 = 64\%$. It means 64% of the total variation in X_1 has been explained by variable X_2 when the next variable X_3 is held constant.

Example 1: If $r_{12} = 0.8$, $r_{13} = -0.4$ and $r_{23} = -0.58$ find $r_{12.3}$.

Solution:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\,.\sqrt{1 - r_{23}^2}} = \frac{0.8 - (0.4) \times (-0.58)}{\sqrt{1 - (0.4)^2}\sqrt{1 - (0.58)^2}}$$

$$= \frac{0.568}{\sqrt{0.84}\sqrt{0.6636}} = \frac{0.568}{\sqrt{0.5574}} = 0.76$$

Example 2: If $r_{12} = 0.4$, $r_{23} = 0.5$ and $r_{13} = 0.6$. Find (i) $r_{23.1}$ (ii) $r_{23.1}^2$ and interpret.

Solution:

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{1 - r_{21}^2}\cdot\sqrt{1 - r_{31}^2}}$$

$$= \frac{0.5 - 0.4 \times 0.6}{\sqrt{1-(0.4)^2}\sqrt{1-(0.6)^2}} = \frac{0.26}{\sqrt{0.84}\sqrt{0.64}} = \frac{0.26}{\sqrt{0.5376}} = 0.35$$

$$r_{23.1}^2 = (0.35)^2 = 0.1225 = 12.25\%$$

It means 12.25% variation in variable X_2 is explained by variable X_3 keeping variable X_1 constant.

Example 3: Are the following data consistent; $r_{12} = -0.8$, $r_{13} = 0.3$ and $r_{23} = 0.4$.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\cdot\sqrt{1 - r_{23}^2}}$$

$$= \frac{-0.8 - (0.3)\times(0.4)}{\sqrt{1-(0.3)^2}\sqrt{1-(0.4)^2}} = \frac{-0.92}{\sqrt{0.91}\sqrt{0.84}} = \frac{-0.92}{\sqrt{0.7644}} = \frac{-0.92}{0.874} = -1.052$$

Since $r_{12.3}$ should lie between -1 and +1, here $r_{12.3} = 1.051 > 1$. Hence the given data are inconsistent.

Multiple Correlation

The relationship among three or more variables simultaneously (at the same time) is called multiple correlation. In this case relationship of a variable with two or more variables is studied at a time.

We are interested to study the relationship of production of paddy with seeds, fertilizer and irrigation etc. If we study the relationship of production of paddy with seeds, fertilizer and irrigation jointly is called multiple correlation.

Let us consider three variables X_1, X_2 and X_3 the multiple correlation coefficient of X_1 with X_2 and X_3 is denoted by $R_{1.23}$ and is given by $R_{1.23} = \sqrt{\dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$

Similarly, the multiple correlation coefficient of X_2 with X_1 and X_3 is denoted by $R_{2.13}$ and is given by $R_{2.13} = \sqrt{\dfrac{r_{21}^2 + r_{23}^2 - 2r_{21}\, r_{23}\, r_{13}}{1 - r_{13}^2}}$

Also multiple correlation coefficient of X_3 with X_1 and X_2 is denoted by $R_{3.12}$ and is given by

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}\, r_{32}\, r_{12}}{1 - r_{12}^2}}$$

Properties of Multiple Correlation Coefficient

1. Multiple correlation coefficient lies between 0 and 1
   (i) $0 \le R_{1.23} \le 1$ (ii) $0 \le R_{2.13} \le 1$ (iii) $0 \le R_{3.12} \le 1$
2. Multiple correlation coefficient is not less than zero order correlation coefficient (simple correlation coefficient)
   (i) $R_{1.23} \ge r_{12}, r_{13}, r_{23}$ (ii) $R_{2.13} \ge r_{21}, r_{23}, r_{13}$ (iii) $R_{3.12} \ge r_{31}, r_{32}, r_{12}$

3. (i) If $R_{1.23} = 0$ then $r_{12} = 0$ and $r_{13} = 0$ (ii) If $R_{2.13} = 0$ then $r_{21} = 0$ and $r_{23} = 0$.
   iii) If $R_{3.12} = 0$ then $r_{31} = 0$ and $r_{32} = 0$.
4. i) $R_{1.23} = R_{1.32}$ (ii) $R_{2.13} = R_{2.31}$ (iii) $R_{3.12} = R_{3.21}$

Coefficient of Multiple Determination

It is the square of multiple correlation coefficient. It is used to measure in variation of one variable as explained by two remaining variables.

If $R_{1.23} = 0.7$ then coefficient of multiple determination is $R_{1.23}^2 = 0.49 = 49\%$. It means 49% variation in variable X_1 is explained by two other variables X_2 and X_3 and remaining 51% is due to the effect of other factors.

Example 4: If $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$ find $R_{1.23}$.

Solution:

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

$$= \sqrt{\frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - (0.52)^2}}$$

$$= \sqrt{0.7334} = 0.8564$$

Example 5: If $r_{12} = 0.7$, $r_{23} = r_{31} = 0.5$ find (i) $R_{1.23}$ (ii) $R_{1.23}^2$ and interpret

Solution:

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

$$= \sqrt{\frac{(0.7)^2 + (0.5)^2 - 2 \times 0.7 \times 0.5 \times 0.5}{1 - (0.5)^2}} = \sqrt{0.57} = 0.721$$

Now $R_{1.23}^2 = (0.721)^2 = 0.52 = 52\%$.

It means 52% variation in X_1 has been explained by X_2 and X_3.

Example 6: Show that the values $r_{12} = 0.6$, $r_{13} = -0.4$ and $r_{23} = 0.7$ are inconsistent.

Solution:

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

$$= \sqrt{\frac{(0.6)^2 + (-0.4)^2 - 2 \times 0.6 \times (-0.4) \times 0.7}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (-0.4)^2 - 2 \times 0.6 \times (-0.4) \times 0.7}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.856}{0.51}}$$

$$= 1.29$$

Here $R_{1.23} = 1.29 > 1$

Since $R_{1.23}$ should lie between 0 and 1. Hence inconsistent in the given values.

Example 7: A sample of 10 values of three variables X_1, X_2 and X_3 were obtained as, $\Sigma X_1 = 10$, $\Sigma X_2 = 20$, $\Sigma X_3 = 30$, $\Sigma X_1X_2 = 10$, $\Sigma X_1X_3 = 15$, $\Sigma X_2X_3 = 64$, $\Sigma X_1^2 = 20$, $\Sigma X_2^2 = 68$, $\Sigma X_3^2 = 170$. (i) Find the partial correlation coefficient between X_1 and X_3 eliminating the effect of X_2. (ii) Find the multiple correlation coefficient of X_1 with X_2 and X_3.

Solution:

Here

$$r_{12} = \frac{n\Sigma X_1X_2 - \Sigma X_1\Sigma X_2}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{nX_2^2 - (\Sigma X_2)^2}}$$

$$= \frac{10 \times 10 - 10 \times 20}{\sqrt{10 \times 20 - (10)^2}\sqrt{10 \times 68 - (20)^2}}$$

$$= \frac{-100}{\sqrt{100}\sqrt{280}}$$

$$= -0.59$$

$$r_{13} = \frac{n\Sigma X_1X_3 - \Sigma X_1\Sigma X_3}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{nX_3^2 - (\Sigma X_3)^2}}$$

$$= \frac{10 \times 15 - 10 \times 30}{\sqrt{10 \times 20 - (10)^2}\sqrt{10 \times 170 - (30)^2}}$$

$$= \frac{-150}{\sqrt{100}\sqrt{800}} = -0.53$$

$$r_{23} = \frac{n\Sigma X_2X_3 - \Sigma X_2\Sigma X_3}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{nX_3^2 - (\Sigma X_3)^2}}$$

$$= \frac{10 \times 64 - 20 \times 30}{\sqrt{10 \times 68 - (20)^2}\sqrt{10 \times 170 - (30)^2}}$$

$$= \frac{40}{\sqrt{280}\sqrt{800}} = 0.085$$

i) Partial correlation coefficient between X_1 and X_3 eliminating the effect of X_2 is

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}}$$

$$= \frac{(-0.53) - (-0.598) \times 0.085}{\sqrt{1 - (-0.598)^2}\sqrt{1 - (0.085)^2}}$$

$$= 0.727$$

ii) Multiple correlation coefficient of X_1 with X_2 and X_3 is

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{32}^2}}$$

$$= \sqrt{\frac{(-0.598)^2 + (-0.53)^2 - 2 \times (-0.598) \times (-0.53) \times 0.085}{1 - (0.085)^2}}$$

$$= 0.767$$

Example 8: The height and weight of 10 individuals of different ages are given below:

Age (x_1)	11	10	6	10	8	9	10	7	11	8
Height(X_2)	60	67	53	56	64	57	71	58	67	57
Weight(X_3)	57	55	49	52	57	48	59	50	62	51

Find $r_{12.3}$, $r_{13.2}$, $R_{1.23}$.

Solution:

Age(X_1)	Ht(X_2)	Wt(X_3)	$u_1 = X_1 - 10$	$u_2 = X_2 - 60$	$u_3 = X_3 - 50$	u_1^2	u_2^2	u_3^2	u_1u_2	u_1u_3	u_2u_3
11	60	57	1	0	7	1	0	49	0	7	0
10	67	55	0	7	5	0	49	25	0	0	35
6	53	49	-4	-7	-1	16	49	1	28	4	7
10	56	52	0	-4	2	0	16	4	0	0	-8
8	64	57	-2	4	7	4	16	49	-8	-14	28
9	57	48	-1	-3	-2	1	9	4	3	2	6
10	71	59	0	11	9	0	121	81	0	0	99
7	58	50	-3	-2	0	9	4	0	6	0	0
11	67	62	1	7	12	1	49	144	7	12	84
8	57	51	-2	-3	1	4	9	1	6	-2	-3
			$\Sigma u_1 = -10$	$\Sigma u_2 = 10$	$\Sigma u_3 = 40$	$\Sigma u_1^2 = 36$	$\Sigma u_2^2 = 322$	$\Sigma u_3^2 = 358$	$\Sigma u_1u_2 = 42$	$\Sigma u_1u_3 = 9$	$\Sigma u_2u_3 = 248$

Here

$$r_{12} = \frac{n\Sigma u_1u_2 - \Sigma u_1\Sigma u_2}{\sqrt{n\Sigma u_1^2 - (\Sigma u_1)^2}\sqrt{n\Sigma u_2^2 - (\Sigma u_2)^2}}$$

$$= \frac{10 \times 42 - (-10) \times 10}{\sqrt{10 \times 36 - (-10)^2}\sqrt{10 \times 322 - (10)^2}}$$

$$= 0.577$$

$$r_{13} = \frac{n\Sigma u_1u_3 - \Sigma u_1\Sigma u_3}{\sqrt{n\Sigma u_1^2 - (\Sigma u_1)^2}\sqrt{n\Sigma u_3^2 - (\Sigma u_3)^2}}$$

$$= \frac{10 \times 9 - (-10) \times 40}{\sqrt{10 \times 36 - (-10)^2}\sqrt{10 \times 358 - (40)^2}}$$

$$= 0.683$$

$$r_{23} = \frac{n\Sigma u_2u_3 - \Sigma u_2\Sigma u_3}{\sqrt{n\Sigma u_2^2 - (\Sigma u_2)^2}\sqrt{n\Sigma u_3^2 - (\Sigma u_3)^2}}$$

$$= \frac{10 \times 248 - 10 \times 40}{\sqrt{10 \times 322 - (-10)^2}\sqrt{10 \times 358 - (40)^2}}$$

$$= 0.836$$

Now,

$$r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.577 - 0.683 \times 0.836}{\sqrt{1 - (0.683)^2}\sqrt{1 - (0.836)^2}}$$

$$= 0.014$$

$$r_{13\cdot2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}}$$

$$= \frac{0.683 - 0.577 \times 0.836}{\sqrt{1 - (0.577)^2}\sqrt{1 - (0.836)^2}}$$

$$= 0.447$$

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

$$= \sqrt{\frac{(0.577)^2 + (0.683)^2 - 2 \times 0.577 \times 0.683 \times 0.836}{1 - (0.836)^2}}$$

$$= \sqrt{\frac{0.14}{0.3001}} = \sqrt{0.4665} = 0.683$$

Multiple Linear Regression

It is a linear function of one dependent variable with two or more independent variables. With the help of two or more independent variables the value of dependent variable is predicted. For example, if we wish to test the hypothesis that whether or not the 'pass grade' of students depends on many causes such as previous test mark, study hours, IQ, ...then we can test a regression of cause (pass grade) with effect variables. This test will give us which causes are really significant in generating effect variable and among the significant cause variables their relative value responsible to generate the effect variable. If we assume more than one causes (called X or independent variable) responsible for one effect (also called Y or dependent variable), it is known as multiple regression. If we assume that the relation between Y and X's is linear it is called multiple linear regression. However, there can be nonlinear relationship between Y and X's. For example, population growth (Y) is generally considers to have exponential relation with time and other cause variables.

Regression is used for two purpose. To get predicted value of Y for hypothetic X values. This is called prediction method and is more used for time dependent variables. For example, the future value of national income under similar conditions as existing. The other use of regression is to understand the role of cause variables on the generation of effect. It is called exploratory analysis and is more used for special data for example, the district data.

Let us consider three variables Y, X_1 and X_2 in which Y is dependent variable, X_1 and X_2 are independent variables, then the mathematical form of the linear relationship of Y with X_1 and X_2 is expressed as

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

Where,

Y = Dependent variable

X_1 and X_2 = Independent variable or explanatory variable or regressors

b_0 = Intercept and is called average value of Y when X_1 and X_2 are zero.

b_1 = Regression coefficient of Y on X_1 keeping X_2 constant. It measures the amount of change in Y per unit change in X_1 holding the X_2 constant.

b_2 = Regression coefficient of Y on X_2 keeping X_1 constant. It measures the amount of change in Y per unit change in X_2 holding the X_1 constant.

ε = Random error.

Random error (ε) is not created from mistake. It is a technical term that denoted the excess of value from real by model estimation. Error is also called Residual.

So, error = true value - estimated value from regression. Mathematically, $\varepsilon = Y - \hat{Y}$, where Y is the true value and $\hat{Y}$ is the estimate from regression. If we have 20 observations we will have 20 error values. By analyzing error or residual we can understand how the regression model fit to the given data, if assumptions such as linear is really usable, and other problems of the cause and effect variables. Such analysis is called Residual Analysis and is very useful diagnostic for regression.

Assumptions of Linear Regression

Theory of regression assumes that certain assumptions should hold for a reliable and acceptable regression analysis. If one or more assumptions are not satisfied or violated the regression will have specific problem. The major assumptions are as described below.

Let us consider multiple regression model

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

There are certain assumptions about the model. The assumptions are based on relation between error ε and explanatory variables x_i's.

i. Regression model is linear in parameters.

ii. ε is random real variable

iii. The random errors ε have zero mean, i.e. $E(\varepsilon) = 0$

iv. The random errors ε has constant variance ie. $E(\varepsilon) = \sigma^2$ (Noheteroscedaticity).

v. The random variable ε is normally distributed. i.e. $\varepsilon \sim N(0, \sigma^2)$

vi. The random errors ε are independent i.e. $E(\varepsilon_i\varepsilon_j) = 0 : i \neq j$. (No autocorrelation).

vii. X are uncorrelated to the error term ε, ie. $E(X\varepsilon) = 0$ (uniformity of X over samples

viii. The explanatory variables x_i's are measured without error.

ix. The number of observations must be greater than the number of explanatory variables.

x. The explanatory variables X_i's are not perfectly linearly correlated(No multicollinearity)

Estimation of Coefficients in Multiple Linear Regression

The linear relationship of dependent variable Y with explanatory variables X_1 and X_2 is given by

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

Here b_0, b_1 and b_2 are called parameters of the three variable multiple regression equation.

Error $(e_i) = Y - b_0 - b_1X_1 - b_2X_2$ then $\Sigma e_i^2 = \Sigma(Y - b_0 - b_1X_1 - b_2X_2)^2$

By using the principle of least square by minimizing error sum of square, normal equations to estimate b_0, b_1 and b_2 are

$$\Sigma Y = nb_0 + b_1\Sigma X_1 + b_2\Sigma X_2 \quad \text{......(i)}$$

$$\Sigma YX_1 = \Sigma b_0X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \quad \text{......(ii)}$$

$$\Sigma YX_2 = \Sigma b_0X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \quad \text{......(iii)}$$

Solving i, ii and iii we get, b_0, b_1 and b_2 then substitute values to get multiple regression equation.

$\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2$, where $\hat{b}_0$, $\hat{b}_1$ and $\hat{b}_2$ are estimated value of b_0, b_1 and b_2 respectively.

Regression equation of X_1 on X_2 and X_3:

Let X_1 be the dependent variable, X_2 and X_3 be the independent variables then the regression equation of X_1 on X_2 and X_3 be

$$X_1 = a + b_2X_2 + b_3X_3$$

By using the principle of least square by minimizing error sum of square, normal equations to estimate a, b_2 and b_3 are

$$\Sigma X_1 = na + b_2\Sigma X_2 + b_3\Sigma X_3 \quad \text{.......(i)}$$

$$\Sigma X_1X_2 = a\Sigma X_2 + b_2\Sigma X_2^2 + b_3\Sigma X_2X_3 \quad \text{.......(ii)}$$

$$\Sigma X_1X_3 = a\Sigma X_3 + b_2\Sigma X_2X_3 + b_3\Sigma X_3^2 \quad \text{.......(iii)}$$

Solving i, ii and iii get a, b_2 and b_3 and substitute values to get multiple regression equation.

Regression equation of X_2 on X_1 and X_3:

Let X_2 be the dependent variable, X_1 and X_3 be the independent variables then the regression equation of X_2 on X_1 and X_3 be

$$X_2 = a + b_1X_1 + b_3X_3$$

By using the principle of least square by minimizing error sum of square, normal equations to estimate a, b_1 and b_3 are

$$\Sigma X_2 = na + b_1\Sigma X_1 + b_3\Sigma X_3 \quad \text{.......(i)}$$

$$\Sigma X_1X_2 = a\Sigma X_1 + b_1\Sigma X_1^2 + b_3\Sigma X_1X_3 \quad \text{.......(ii)}$$

$$\Sigma X_2X_3 = a\Sigma X_3 + b_1\Sigma X_1X_3 + b_3\Sigma X_3^2 \quad \text{.......(iii)}$$

Solving i, ii and iii get a, b_1 and b_3 and substitute values to get multiple regression equation

Regression equation of X_3 on X_1 and X_2:

Let X_3 be the dependent variable, X_1 and X_2 be the independent variables then the regression equation of X_3 on X_1 and X_2 be

$$X_3 = a + b_1X_1 + b_2X_2$$

By using the principle of least square by minimizing error sum of square, normal equations to estimate a, b_1 and b_2 are

$$\Sigma X_3 = na + b_1\Sigma X_1 + b_2\Sigma X_2 \quad \text{.......(i)}$$
$$\Sigma X_1X_3 = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \quad \text{.......(ii)}$$
$$\Sigma X_2X_3 = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \quad \text{.......(iii)}$$

Solving i, ii and iii get a, b_1 and b_2 and substitute values to get multiple regression equation.

Example 9: Consider the following results obtained from a sample of 6:

$\Sigma x_1 = 487$, $\Sigma x_2 = 40$, $\Sigma y = 192$, $\Sigma x_1x_2 = 3346$, $\Sigma yx_1 = 15995$, $\Sigma yx_2 = 1390$, $\Sigma x_1^2 = 39901$, $\Sigma x_2^2 = 296$. Find the regression equation of y on x_1 and x_2. Estimate y when $x_1 = 83$ and $x_2 = 7$.

Solution:

Regression equation of y on x_1 and x_2 is $y = b_0 + b_1x_1 + b_2x_2$ ...... (i)

To estimate b_0, b_1 and b_2

$$\Sigma y = nb_0 + b_1\Sigma x_1 + b_2\Sigma x_2$$

or $$192 = 6b_0 + 487b_1 + 40b_2 \quad \text{....(ii)}$$

$$\Sigma yx_1 = b_0\Sigma X_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2$$

or $$15995 = 487b_0 + 39901b_1 + 3346b_2 \quad \text{....(iii)}$$

$$\Sigma yx_2 = b_0\Sigma x_2 + b_1\Sigma x_1x_2 + b_2\Sigma x_2^2$$

or $$1390 = 40b_0 + 3346b_1 + 296b_2 \quad \text{....(iv)}$$

Using Cramer's rule

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constant
6	487	40	192
487	39901	3346	15995
40	3346	296	1390

$$D = \begin{vmatrix} 6 & 487 & 40 \\ 487 & 39901 & 3346 \\ 40 & 3346 & 296 \end{vmatrix}$$

$= 6(39901 \times 296 - 3346 \times 3346) - 487(487 \times 296 - 40 \times 3346) + 40(487 \times 3346 - 40 \times 39901)$

$= 6416$

$$D_1 = \begin{vmatrix} 192 & 487 & 40 \\ 15995 & 39901 & 3346 \\ 1390 & 3346 & 296 \end{vmatrix}$$

$= -352100$

$$D_2 = \begin{vmatrix} 6 & 192 & 40 \\ 487 & 15995 & 3346 \\ 40 & 1390 & 296 \end{vmatrix}$$

$= 6776$

$$D_3 = \begin{vmatrix} 6 & 487 & 192 \\ 487 & 39901 & 15995 \\ 40 & 3346 & 1390 \end{vmatrix}$$

$= 1114$

Now,

$$b_0 = \frac{D_1}{D} = \frac{-352100}{6416} = -54.878$$

$$b_1 = \frac{D_2}{D} = \frac{6776}{6416} = 1.056$$

$$b_2 = \frac{D_3}{D} = \frac{1114}{6416} = 0.173$$

Substitute value in equation I we get

$y = -54.878 + 1.056x_1 + 0.173x_2$

When $x_1 = 83$ and $x_2 = 7$

$y = -54.878 + 1.056 \times 83 + 0.173 \times 7 = 33.981$

Example 10: The following information has been gathered from a random sample of apartment renters in a city. We are trying to predict rent (in dollars per month) based on the size of apartment (number of rooms) and the distance from downtown (in miles)

Rent (Dollar)	360	1000	450	525	350	300
Number of rooms	2	6	3	4	2	1
Distance from downtown	1	1	2	3	10	4

(i) Obtain the multiple regression models that best relate these variables (ii) Interpret the obtained regression coefficients. (iii) If some one is looking for a two bed apartment 2 miles from down town, what rent should he expect to pay?

Solution:

Here, Rent depends upon the number of rooms and distance from downtown.

Let rent = y, number of rooms = x_1 and distance from down town = x_2 then we have to find the regression equation of y on x_1 and x_2.

Rent (y) Dollar	No of rooms (x_1)	Distance (x_2)	x_1^2	x_2^2	yx_1	yx_2	x_1x_2
360	2	1	4	1	720	360	2
1000	6	1	36	1	6000	1000	6
450	3	2	9	4	1350	900	6
525	4	3	16	9	2100	1575	12
350	2	10	4	100	700	3500	20
300	1	4	1	16	300	1200	4
$\Sigma y=2985$	$\Sigma x_1=18$	$\Sigma x_2=21$	$\Sigma x_1^2=70$	$\Sigma x_2^2=131$	$\Sigma yx_1=11170$	$\Sigma yx_2=8535$	$\Sigma x_1x_2=50$

To fit $y = b_0 + b_1x_1 + b_2x_2$

$\Sigma y = nb_0 + b_1\Sigma x_1 + b_2\Sigma x_2$

$2985 = 6b_0 + 18b_1 + 21b_2$ ....(i)

$\Sigma yx_1 = b_0\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2$

$11170 = 18b_0 + 70b_1 + 50b_2$ ....(ii)

$\Sigma yx_2 = b_0\Sigma x_2 + b_1\Sigma x_1x_2 + b_2\Sigma X_2^2$

$8535 = 21b_0 + 50b_1 + 131b_2$ ....(iii)

Using Cramer's rule

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constant
6	18	21	2985
18	70	50	11170
21	50	131	8535

Now,

$$D = \begin{vmatrix} 6 & 18 & 21 \\ 18 & 70 & 50 \\ 21 & 50 & 131 \end{vmatrix}$$

$$= 6(9170-2500) - 18(2358-1050) + 21(900-1470) = 4506$$

$$D_1 = \begin{vmatrix} 2985 & 18 & 21 \\ 11170 & 70 & 50 \\ 8535 & 50 & 131 \end{vmatrix}$$

$$= 2985(9170 - 2500) - 18(1463270 - 426750) + 21(558500 - 597450) = 434640$$

$$D_2 = \begin{vmatrix} 6 & 2895 & 21 \\ 18 & 11170 & 50 \\ 21 & 8535 & 131 \end{vmatrix}$$

$$= 6(1463270 - 426750) - 2985(2358 - 1050) + 21(153630 - 234570) = 615000$$

$$D_3 = \begin{vmatrix} 6 & 18 & 2985 \\ 18 & 70 & 11170 \\ 21 & 50 & 8535 \end{vmatrix}$$

$$= 6(597450 - 558500) - 18(153630 - 234570) + 2985(900 - 1470) = -10830$$

$$b_0 = \frac{D_1}{D} = \frac{434640}{4506} = 96.458,$$

$$b_1 = \frac{D_2}{D} = \frac{615000}{4506} = 136.484,$$

$$b_2 = \frac{D_3}{D} = \frac{-10830}{4506} = -2.403$$

Substituting values in regression equation

(i) $y = 96.458 + 136.484x_1 - 2.403x_2$

(ii) $b_1 = 136.484$ means on average rent is increased by 136.484 when room is increased by 1 holding the effect of distance from down town constant.

$b_2 = -2.403$ means average rent is decreased by 2.403 when the distance from downtown is increased by 1 holding the effect of number of room constant.

(iii) When $x_1 = 2$ and $x_2 = 2$,

$y = 96.458 + 136.484x_1 - 2.403x_2$

$= 96.458 + 136.484 \times 2 - 2.403 \times 2 = 364.62$

Expected rent for two bed room apartment 2 miles from downtown is 364.62 dollar.

Measures of Variation

In regression model value of dependent variable are estimated on the basis of independent variables. In regression analysis total variation is divided into explained variation (sum of square due to regression) and unexplained variation (sum of square due to error). Hence according to Fisher total sum of square is decomposed into sum of square due to regression and sum of square due to error (residual).

Total sum of square (TSS) = Sum of square due to regression (SSR) + Sum of square due to error (SSE)

For regression model $Y = b_0 + b_1X_1 + b_2X_2$, where Y is dependent variable, X_1 and X_2 are independent (explanatory) variables

$TSS = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$

$SSE = \Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma YX_1 - b_2\Sigma YX_2$

$SSR = TSS - SSE$

For regression model $x_1 = a + b_2x_2 + b_3x_3$, where x_1 is dependent variable and x_2, x_3 are independent variables

$TSS = \Sigma(x_1 - \bar{x}_2)^2 = \Sigma x_1^2 - n\bar{x}_1^2$

$SSE = \Sigma(x_1 - \hat{x}_1)^2 = \Sigma x_1^2 - a\Sigma x_1 - b_2\Sigma x_1x_2 - b_3\Sigma x_1x_3$

$SSR = TSS - SSE$

For regression model $x_2 = a + b_1x_1 + b_3x_3$, where x_2 is dependent variable and x_1, x_3 are independent variables

$TSS = \Sigma(x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - n\bar{x}_2^2$

$SSE = \Sigma(x_2 - \hat{x}_2)^2 = \Sigma x_2^2 - a\Sigma x_2 - b_1\Sigma x_1x_2 - b_3\Sigma x_2x_3$

$SSR = TSS - SSE$

For regression model $x_3 = a + b_1x_1 + b_2x_2$, where x_3 is dependent variable and x_1, x_2 are independent variables

$TSS = \Sigma(x_3 - \bar{x}_3)^2 = \Sigma x_3^2 - n\bar{x}_3^2$

$SSE = \Sigma(x_3 - \hat{x}_3)^2 = \Sigma x_3^2 - a\Sigma x_3 - b_1\Sigma x_1x_3 - b_2\Sigma x_2x_3$

$SSR = TSS - SS\,E$

ANOVA table of regression analysis

Source of variation(S.V.)	Degree of freedom (df)	Sum of square (SS)	Mean square(MS) (Variance)
Regression	k(no of independent variable)	SSR	MSR = SSR/k
Error	n-k-1	SSE	MSE = SSE/n-k-1
Total	n-1	TSS	

Standard Error of the Estimate

Standard error is the Square root of the variance computed from sample data. The standard error of the estimate measures the average variation or scatterness of the observed data point around regression line. Standard error of the estimate is used to measure the reliability of the regression equation. Regression line having less standard error of estimate is more reliable than regression line having more standard error of estimate.

It is given by $S_e = \sqrt{\frac{SSE}{n - k - 1}}$

SSE = sum of square due to error

k = number of independent variable in regression model

n = number of observations.

When $S_e = 0$, there is no variation of observed data around regression line. In such case regression line is perfect for estimating the dependent variable.

Coefficient of Determination

It measures the proportion of variation in dependent variable that is explained by the set of independent variables .It is the measure based upon measure of variation and is used to determine the fitness of the data to the model. The regression line is reliable if the sum of square due to regression is much greater than sum of square due to error. It is the ratio of sum of square due to regression to the total sum of square. It is denoted by R^2 and is given by, $R^2 = \frac{SSR}{TSS}$

It is also obtained by simply squaring the correlation coefficient i.e., $R^2 = r^2$. Higher the value of R^2 the more reliable is the fitted equation .It lies between 0 and 1.

R^2 can never decrease when another independent variable is added to a regression. R^2 will usually increase with increase in number of independent variables.

It is suggested that the adjusted R^2 should be used in place of R^2 in multiple regression model. Adjusted R^2 is simply a R^2 adjusted by its degree of freedom and reflects both the number of independent variables and sample size used in the model. Adjusted R^2 is considered as an important measure for the comparising of two or more regression models that predict same dependent variable with different number of independent variables.

$R^2_{adjusted} (\bar{R}^2) = 1 - \frac{(n - 1)}{(n - k - 1)} [1 - R^2]$; where n = no of pair of observations, k = no of independent variables.

Example 11: A health research team collects data on ten communities. Measurement are obtained on the following variable.

y = Health care facility utilization

x_1 = Median family income

x_2 = Proportion of worker with health insurance

x_3 = Doctor population ratio.

Source of variation	Sum of square	df
Regression	?	3
Error	88.66	?
Total	476.9	9

(i) Complete the table (ii) Compute R^2 and interpret

(iii) Compute adjusted R^2 (iv) Compute standard error of estimate.

Solution:

Here

$SSE = 88.66$, $TSS = 476.9$, $k = 3$, $n-1 = 9$

df for error = n-k-1 = 9-3 =6

$SSR = TSS - SSE$

$= 476.9 - 88.66 = 388.24$

$$R^2 = \frac{SSR}{TSS}$$

$$= \frac{388.24}{476.9} = 0.814 = 81.4\%$$

It means 81.5% of the total variation in health care facility utilization can be explained by the variation in median family income, proportion of worker with health insurance and doctor population ratio.

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{(n-k-1)}[1 - R^2]$$

$$= 1 - \frac{9}{6}(1 - 0.814)$$

$$= 1 - 0.279 = 0.721$$

$$S_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{88.66}{6}} = 3.84.$$

Example 12: Find $S_{e(1.23)}$, $R_{1.23}^2$ on the basis of following information:

$\Sigma x_1 = 272$, $\Sigma x_2 = 441$, $\Sigma x_3 = 147$, $\Sigma x_1x_2 = 12005$, $\Sigma x_1x_3 = 4013$, $\Sigma x_2x_3 = 6485$, $\Sigma x_1^2 = 7428$, $\Sigma x_2^2 = 19461$, $\Sigma x_3^2 = 2173$, $n = 10$.

Solution:

We have to find the regression equation of x_1 on x_2 and x_3

$x_1 = a + b_2x_2 + b_3x_3$

To estimate a, b_2 and b_3

$\Sigma x_1 = na + b_2\Sigma x_2 + b_3\Sigma x_3$

or $272 = 10a + 441b_2 + 147b_3$ ......(i)

$\Sigma x_1x_2 = a\Sigma x_2 + b_2\Sigma x_2^2 + b_3\Sigma x_2x_3$

or $12005 = 441a + 19461b_2 + 6485b_3$ ......(ii)

$\Sigma x_1x_3 = a\Sigma x_3 + b_2\Sigma x_2x_3 + b_3\Sigma x_3^2$

$4013 = 147a + 6485b_2 + 2173b_3$ ......(iii)

To find a, b_2 and b_3 using Cramer's rule

Coefficient of a	Coefficient of b_2	Coefficient of b_3	Constant
10	441	147	272
441	19461	6485	12005
147	6485	2173	4013

$$\text{Now, } D = \begin{vmatrix} 10 & 441 & 147 \\ 441 & 19461 & 6485 \\ 147 & 6485 & 2173 \end{vmatrix}$$

$$= 10(42288753 - 42055225) - 441(958293 - 953295) + 147(2859885 - 2860767)$$

$$= 1508$$

$$D_1 = \begin{vmatrix} 272 & 441 & 147 \\ 12005 & 19461 & 6485 \\ 4013 & 6485 & 2173 \end{vmatrix}$$

$$= 272(42288753 - 42055225) - 441(26086865 - 26024305) + 147(77852425 - 78096993)$$

$$= -20840$$

$$D_2 = \begin{vmatrix} 10 & 272 & 147 \\ 441 & 12005 & 6485 \\ 147 & 4013 & 2173 \end{vmatrix}$$

$$= 10(26086865 - 26024305) - 272(958293 - 953295) + 147(1769733 - 1764735)$$

$$= 850$$

$$D_3 = \begin{vmatrix} 10 & 441 & 272 \\ 441 & 19461 & 12005 \\ 147 & 6485 & 4013 \end{vmatrix}$$

$$= 10(78096993 - 77852425) - 441(1769733 - 1764735) + 272(2859885 - 2860767)$$

$$= 1658$$

Now

$$a = \frac{D_1}{D} = \frac{-20840}{1508} = -13.819$$

$$b_2 = \frac{D_2}{D} = \frac{850}{1508} = 0.563$$

$$b_3 = \frac{D_3}{D} = \frac{1658}{1508} = 1.099$$

$$\bar{x}_1 = \frac{\Sigma X_1}{n} = \frac{272}{10} = 27.2$$

$$SSE(X_{1.23}) = \Sigma x_1^2 - a\Sigma x_1 - b_2\Sigma x_1x_2 - b_3\Sigma x_1x_3$$

$$= 7428 - (-13.819) \times 272 - 0.563 \times 12005 - 1.099 \times 4013 = 17.661$$

$$TSS = \Sigma x_1^2 - n\bar{x}_1^2$$

$= 7428 - 10 \times (27.2)^2 = 29.6$

SSR $= TSS - SSE$

$= 29.6 - 17.661 = 11.939$

$$S_{y(1,2)} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{17.661}{10-2-1}} = 1.58$$

$$R^2_{y.12} = \frac{SSR}{TSS} = \frac{11.939}{29.6} = 0.403$$

Test of Significance for Regression Coefficients

To test the significance of the individual regression coefficients t test is used. It helps to determine whether there is significant linear relationship between dependent variable and independent variable.

Let us consider regression equation

$y = b_0 + b_1x_1 + b_2x_2$, for multiple regression equation of three variables. Where y is dependent variable; x_1, x_2 are independent variables, b_0 constant value, b_1 is regression coefficient of y on x_1 keeping x_2 constant, b_2 is regression coefficient of y on x_2 keeping x_1 constant.

Let β_1 and β_2 be the population regression coefficients of the sample regression equation:

$y = b_0 + b_1 x_1 + b_2 x_2$.

Different steps in the test are

Problem to test

$H_0 : \beta_i = 0$ (There is no linear relationship between dependent variable y and independent variable x_i, $i = 1, 2$).

$H_1: \beta_i \neq 0$.

Test statistic

$t = \frac{b_i}{Sb_i} \sim$ t distribution with n-k-1 degree of freedom, n = no of observation and k = no of independent variables

Where b_i = sample regression coefficient and Sb_i = Standard error of regression coefficient

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of t is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision:

Reject H_0 at α level of significance if $|t| > t_{tabulated}$, accept otherwise.

Confidence interval for regression coefficient

At α% level of significance for n-k-1 degree of freedom the critical value of t is $t_{\alpha/2\,(n-k-1)}$, then $(100 - \alpha\%)$ confidence or fudicial limits for regression coefficient β_i is given by $b_i \pm t_{\alpha/2\,(n-k-1)} Sb_i$.

Example 13: To study the effect of age (x_1 in years) and weight (x_2 in lbs) on systolic blood pressure (y mm in Hg), the data were recorded for a sample of 15 adult males. The estimated regression model based on data is described below where figures within parenthesis are standard error of estimate.

$$y = 27.4 + 0.221\, x_1 + 0.56\, x_2$$
$$(24.68) \quad (0.248) \quad (0.155)$$

Test the significance of regression coefficients at 1% level of significance.

Solution:

Here,

Sample size (n) = 15, Number of independent variable (k) = 2, b_0 = 27.4, b_1 = 0.221, b_2 = 0.56, Sb_0 = 24.68, Sb_1 = 0.248, Sb_2 = 0.115, α = 1%.

Let β_1 and β_2 be the population regression coefficients.

For the first regression coefficient

Problem to test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Test statistic

$$t = \frac{b_1}{Sb_1} = \frac{0.221}{0.248} = 0.89$$

Critical value

At α = 0.01 level of significance, critical value for two tailed test is

$t_{tabulated} = t_{\alpha/2(n-k-1)} = 3.055.$

Decision

$t = 0.89 < t_{tabulated} = 3.055$, accept H_0 at 5% level of significance.

Conclusion

There is no significant linear relationship between y and x_1.

For the second regression coefficient

Problem to test

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

Test statistic

$$t = \frac{b_2}{Sb_2} = \frac{0.56}{0.115} = 4.869.$$

Critical value

At α = 0.01 level of significance, critical value for two tailed test is

$t_{tabulated} = t_{\alpha/2(n-k-1)} = 3.055$

Decision

$t = 4.869 > t_{tabulated} = 3.055$, reject H_0 at 5% level of significance.

Conclusion

There is a significant linear relationship between y and x_2.

Test of Overall Significance of the Regression Coefficients

To test the significance of over all regression coefficients F test is used. It helps to determine whether there is significant linear relationship between the dependent variable and the set of independent variables.

Let us consider regression equation

$y = b_0 + b_1 x_1 + b_2 x_2$, for multiple regression equation of three variables. Where y is dependent variable; x_1, x_2 are independent variables, b_0 constant value, b_1 is regression coefficient of y on x_1 keeping x_2 constant, b_2 is regression coefficient of y on x_2 keeping x_1 constant.

Let β_1 and β_2 be the population regression coefficients of the sample regression equation $y = b_0 + b_1 x_1 + b_2 x_2$.

Different steps in the test are

Problem to test

$H_0: \beta_1 = \beta_2 = 0$ (There is no linear relationship between dependent variable y and independent variables)

H_1: At least one β_i is different from zero (i = 1, 2)

(There is linear relationship between the dependent variable and at least one independent variable)

Test statistic

$F = \frac{MSR}{MSE} \sim$ F distribution with (k, n-k-1) degree of freedom, where k = no of independent variables

MSR = mean sum of square due to regression and

MSE = mean sum of square due to error

ANOVA table for regression analysis

Source of Variation (SV)	Degree of freedom (df)	Sum of Squares (SS)	Mean Squares (MS)	F	$F_{tabulated}$
Regression	k	SSR	MSR	F_R=MSR/MSE	$F_{\alpha (k, n-k-1)}$
Error	n-k-1	SSE	MSE		
Total	n-1	TSS			

$TSS = \Sigma(y - \bar{y})^2$, $SSE = \Sigma (y - \hat{y})^2$, SSR = TSS - SSE.

Level of significance

Let α be the level of significance. Usually we take α = .05 unless we are given.

Critical value

Critical or tabulated value of F is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $F > F_{tabulated}$, accept otherwise.

Relationship between F and R^2

We know,

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{(n-k-1)}{k} \times \frac{SSR}{SSE}$$

$$= \frac{(n-k-1)}{k} \times \frac{\frac{SSR}{TSS}}{\frac{SSE}{TSS}}$$

$$= \frac{(n-k-1)}{k} \times \frac{\frac{SSR}{TSS}}{\frac{(TSS-SSR)}{TSS}}$$

$$= \frac{(n-k-1)}{k} \times \frac{\frac{SSR}{TSS}}{\frac{TSS}{TSS}-\frac{SSR}{TSS}}$$

$$= \frac{(n-k-1)}{k} \times \frac{R^2}{1-R^2}$$

Example 14: The following ANOVA summary table was obtained from a multiple regression model with two independent variables.

Source of variation	Degree of freedom	Sum of Square
Regression	2	30
Error	10	120
Total	12	150

Test the overall fit of the model at 0.05 level of significance.

Solution:

Here, $n-k-1 = 12$, $k = 2$ $SSR = 30$, $SSE = 120$, $TSS = 150$, $\alpha = 0.05$

$n = 12 + k + 1 = 12 + 2 + 1 = 15$

$MSR = \frac{SSR}{k} = \frac{30}{2} = 15$, $MSE = \frac{SSE}{n-k-1} = \frac{120}{10} = 12.$

Problem to test

$H_0 : \beta_1 = \beta_2 = 0$

H_1 : At least one β_i is different from 0, $i = 1, 2$

Test statistic

$$F = \frac{MSR}{MSE} = \frac{15}{12} = 1.25$$

Critical value

At $\alpha = 0.05$ level of significance for one tailed test the critical value is $F_{\alpha(k,\, n-k-1)} = 3.89$

Decision: $F = 1.25 < F_{tabulated} = 3.89$, accept H_0 at 0.05 level of significance.

Conclusion: There is no significant relationship between dependent variable and two independent variables.

Example 15: To study the effect of age (x_1 in years) and weight (x_2 in lbs) on systolic blood pressure (y mm in Hg), the data were recorded for a sample of 15 adult males. The estimated regression model based on data is described below:

$$y = 27.4 + 0.221x_1 + 0.56x_2$$

Further computation shows that $\Sigma(y - \bar{y})^2 = 1835.7$ and $\Sigma(y - \hat{y})^2 = 1101.3$.

Carry out the overall goodness of fit test of the model at 5% level of significance.

Solution:

Here, Sample size (n) = 15, Number of independent variables (k) = 2

$b_0 = 27.4$, $b_1 = 0.221$, $b_2 = 0.56$, Level of significance (α) = 5%

$$TSS = \Sigma(y - \bar{y})^2 = 1835.7$$

$$SSE = \Sigma(y - \hat{y})^2 = 1101.3$$

$$SSR = TSS - SSE = 1835.7 - 1101.3 = 734.4,$$

$$MSR = \frac{SSR}{k} = \frac{734.4}{2} = 367.2$$

$$MSE = \frac{SSE}{n - k - 1} = \frac{1101.3}{12} = 91.775$$

Problem to test

$H_0: \beta_1 = \beta_2 = 0$

H_1: At least one β_i is different from zero, i = 1, 2

Test statistic

$$F = \frac{MSR}{MSE} = \frac{367.2}{91.775} = 4.001$$

Critical value

At $\alpha = 0.05$ level of significance, critical value is $F_{\alpha(k, n-k-1)} = 3.89$.

Decision

$F = 4.001 > F_{tabulated} = 3.89$, reject H_0 at 5% level of significance.

Conclusion

There is linear relationship of dependent variable y with both the independent variables x_1 and x_2.

EXERCISE

1. What do you mean by partial correlation? Write down the relationship between partial and simple correlation coefficients.
2. What do you mean by multiple correlation? Write down the relationship between multiple correlation coefficient and simple correlation coefficients.
3. Write down the properties of multiple correlation coefficient.
4. Differentiate between partial and multiple correlation coefficient.
5. What is multiple regression? Write down the method of obtaining multiple regression line.
6. What are underlying assumptions of linear regression model?
7. What do you mean by standard error of estimate? Write down role of it in regression analysis.
8. What do you mean by coefficient of determination? How is it different from correlation coefficient?
9. If $r_{12} = 0.5$, $r_{23} = 0.1$ and $r = 0.4$ compute $r_{12.3}$ and $r_{13.2}$. **Ans:** 0.5, 0.4
10. For a trivariate distribution $r_{12} = 0.4$, $r_{23} = 0.5$ and $r_{13} = 0.6$. Find (i) $R_{1.23}$ (ii) $r_{23.1}$ (iii) $R_{1.23}^2$ (iv) $r_{23.1}^2$ and comment. **Ans:** 0.611, 0.35, 0.37, 0.125
11. Are the following data consistent; $r_{23} = 0.8$, $r_{31} = -0.5$, $r_{12} = 0.6$. **Ans:** inconsistent
12. From the data related to the yield of dry bark (x_1), height (x_2) and girth (x_3) for 18 cinchona plants the following correlation coefficient were obtained $r_{12} = 0.77$, $r_{13} = 0.72$, $r_{23} = 0.52$. Find the partial correlation coefficients. **Ans:** 0.63, 0.85, - 0.07
13. Suppose a computer has found for a given set of values x_1, x_2 and x_3: $r_{12} = 0.91$, $r_{13} = 0.33$, $r_{32} = 0.81$. Examine whether the computations may be said to be free from error? **Ans:** No
14. The following are zero order correlation coefficients $r_{12} = 0.8$, $r_{13} = 0.44$, $r_{23} = 0.54$. Calculate the partial correlation coefficient between first and third variables keeping the effect of second variable constant. **Ans:** 0.0158
15. Consider the following results obtained from a sample of 10 and x_1, x_2 and x_3 are measured in arbitrary unit $\Sigma x_1 = 10$, $\Sigma x_2 = 20$, $\Sigma x_3 = 30$, $\Sigma x_1^2 = 20$, $\Sigma x_2^2 = 68$, $\Sigma x_3^2 = 170$, $\Sigma x_1x_2 = 10$, $\Sigma x_1x_3 = 15$, $\Sigma x_2x_3 = 64$. Compute $r_{12.3}$ and $R_{1.23}$. **Ans:** -0.65, 0.76
16. From the information given below calculate $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$. **Ans:** 0.86, 0.62, -0.177

x_1	6	8	9	11	12	14
x_2	14	16	17	18	20	23
x_3	21	22	27	29	31	32

17. Given the following information from a multiple regression analysis;

$n = 20$, $b_1 = 4$, $b_2 = 3$, $Sb_1 = 1.2$, $Sb_2 = 0.8$. At 0.05 level of significance, determine whether each of explanatory (dependent) variable makes a significant contribution to the regression model.

Ans: $t = 3.33$, Sig. $t = 3.75$, Sig.

6. In order to establish the functional relationship between annual salaries(y), years of educated high school (x_1) and years of experience with the firm (x_2), data on these three variables were collected from a random sample of 10 persons working in a large firm. Analysis of data produces the following results. Total sum of squares $\Sigma(y - \bar{y})^2 = 397.6$.

Sum of squares due to error $\Sigma(y - \hat{y})^2 = 23.5$. Test the over all significance of regression coefficients at 5% level of significance.

Ans: $F = 55.83$, Sig.

7. Suppose you are given following information:

Multiple regression model $y = 5 + 18x_1 + 20x_2$, sample size $n = 28$

Total sum of squares (TSS) = 250

Sum of square due to error (SSE) = 100

Standard error of regression coefficient of x_1 (Sb_1) = 3.2

Standard error of regression coefficient of x_2 (Sb_2) = 5.5

Test the significance of regression coefficient of x_2 at 1% level of significance

Also test the over all significance of regression coefficients at 5% level of significance.

Ans: $t = 3.63$, Significant, $F = 18.75$, Significant

3. From following information of variables X_1, X_2 and X_3

$\Sigma X_1 = 13$, $\Sigma X_2 = 11$, $\Sigma X_3 = 51$, $\Sigma X_1^2 = 63$, $\Sigma X_2^2 = 95$, $\Sigma X_1X_3 = 77$, $\Sigma X_2X_3 = 136$, $\Sigma X_1X_2 = -240$, $n = 10$, $\Sigma X_3 = 450$ and interpret the regression coefficients.

(i) Find the regression equation of X_3 on X_1 and X_2 and interpret the regression coefficients.

(ii) Predict X_3 when $X_1 = 1$ and $X_2 = 4$.

(iii) Compute TSS, SSR and SSE

(iv) Compute standard error of estimate

(v) Compute the coefficient of multiple determination and interpret.

Ans: $X_3 = 1.008 + 1.676X_1 + 1.738X_2$, 9.636, 189.9, 156.72, 33.17, 2.17, 0.82

2. From the following information of three variables Y, X_1 and X_2

$\Sigma(y - \bar{y})^2 = 3450$, $\Sigma(y - \hat{Y})^2 = 365.7$, $\Sigma x_1x_2 = 5779$, $\Sigma yx_2 = 6796$, $\Sigma yx_1 = 40830$,

$\Sigma y^2 = 48139$, $\Sigma x_1^2 = 3483$, $\Sigma x_2^2 = 976$, $\Sigma y = 753$, $\Sigma x_1 = 643$, $\Sigma x_2 = 106$, $n = 12$

(i) Find the least square regression of y on x_1 and x_2

(ii) Find the standard error of estimate.

(iii) find the coefficient of multiple determination.

Ans: $y = 30.69 - 0.0038x_1 + 3.652x_2$, 6.37, 0.89

22. The table shows the corresponding values of the three variables X_1, X_2 and X_3

X_1:	5	7	8	6	10	9
X_2:	12	20	30	40	33	25
X_3:	51	55	58	60	70	66

Find the regression equation of X_1 on X_2 and X_3. Estimate X_1 when $X_2 = 50$ and $X_3 = 100$. Where X_1 represents pull strength, X_2 represents wire length and X_3 represents die height.

Ans: $X_1 = -7.862 - 0.048X_2 + 0.277X_3$, 19.78

23. From the following set of data (i) find the multiple regression equation (ii) Interpret the regression coefficients (iii) Predict y when $X_1 = -10$ and $X_2 = 4$.

Y:	6	10	9	14	7	5
X_1:	1	3	2	-2	3	6
X_2:	3	-1	4	7	2	-4

Ans: $Y = 12.425 - 1.487X_1 - 0.383X_2$, 25.76

24. A developer of food for pig would like to determine what relationship exists among the age of a pig when it starts receiving a newly developed food supplement, the initial weight of the pig and the amount of weight it gains in a week period with the food supplement. The following information is the result of study of eight piglets.

Piglet number	Initial weight (pounds) x_1	Initial age (weeks) x_2	Weight gain y
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

(i) Calculate the least square equation that best describes these three variables.

(ii) Calculate the standard error of estimate.

(iii) How much might we expect a pig to gain weight in a week with the food supplement if it were 9 weeks old and weighted 48 pounds?

Ans: $Y = -3.66 + 0.105X_1 + 0.732x_2$, 1.25, 8

Using Software

Regression Analysis

A developer of food for pig wish to determine what relationship exists among 'age of a pig' when it starts receiving a newly developed food supplement, the initial weight of the pig and the amount of weight it gains in a week period with the food supplement. The following information is the result of study of eight piglets.

Piglet number	Initial weight(pounds) x_1	Initial age (weeks) x_2	Weight gain y
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

I. Determine the least square equation that best describes these three variables.

II. Calculate the standard error.

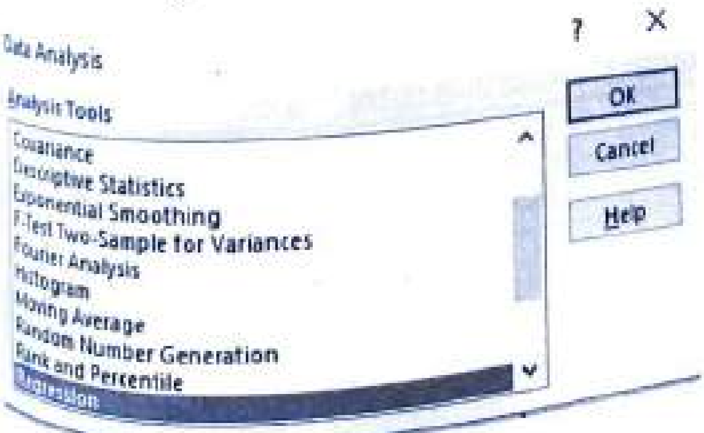
III. How much gain in weight of a pig in a week can we expect with the food supplement if it were 9 weeks old and weighed 48 pounds?

IV. Test the significance of regression coefficients and overall fit of the regression equation

V. Conduct the residual analysis

VI. Determine partial correlations, multiple correlation and coefficient of multiple determination. Interpret.

Using **data analysis tool**



	A	B	C	D
1	Piglet number	Initial weight(pounds)x_1	Initial age (weeks) x_2	Weight gain y
2	1	39	8	7
3	2	52	6	6
4	3	49	7	8
5	4	46	12	10
6	5	61	9	9
7	6	35	6	5
8	7	25	7	3
9	8	55	4	4

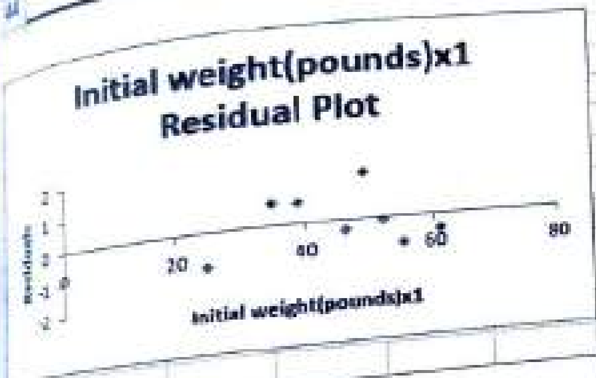
	A	B	C	D	E	F	G
12	SUMMARY OUTPUT						
13							
14	*Regression Statistics*						
15	Multiple R	0.93870818					
16	R Square	0.88117304					
17	Adjusted R Square	0.83364226					
18	Standard Error	0.99907279					
19	Observations	8					
20							
21	ANOVA						
22		*df*	*SS*	*MS*	*F*	*Significance F*	
23	Regression	2	37.00927	18.50463	18.539	0.004867292	
24	Residual	5	4.990732	0.998146			
25	Total	7	42				
26							
27		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
28	Intercept	-4.1917094	1.888119	-2.22004	0.077124	-9.045274309	0.66185550
29	Initial weight(pounds)x1	0.10483433	0.032291	3.246502	0.022784	0.021826458	0.18784219
30	Initial age (weeks) x2	0.80650253	0.158237	5.096815	0.00378	0.399742475	1.21326258
31							

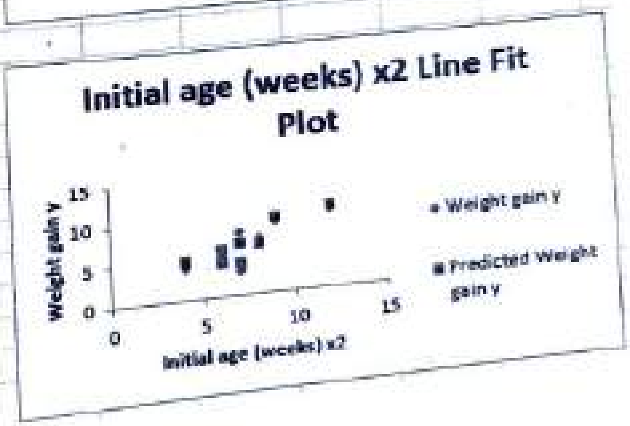
RESIDUAL OUTPUT

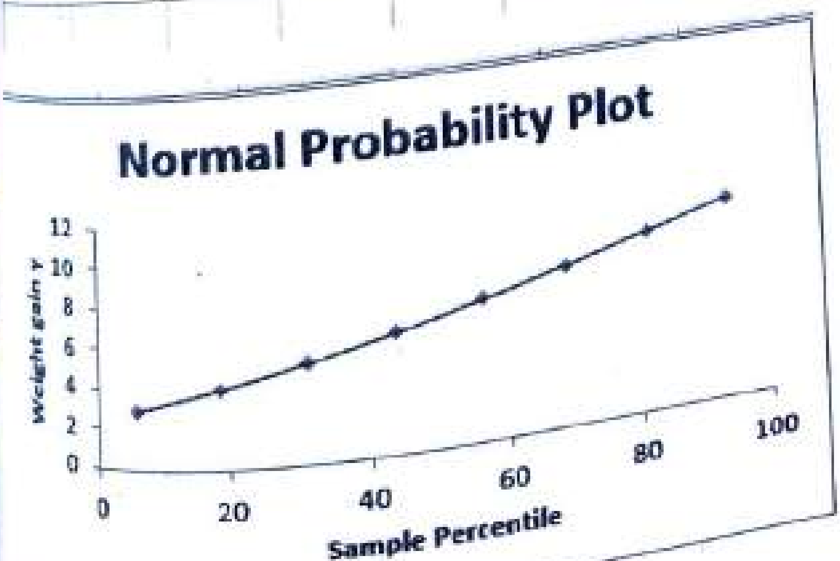
Observation	Predicted Weight gain y	Residuals
1	6.34884955	0.65115
2	6.09869072	-0.09869
3	6.59069027	1.40931
4	10.3087	-0.3087
5	9.46170724	-0.46171
6	4.31650718	0.683493
7	4.07466646	-1.07467
8	4.80018863	-0.80019

PROBABILITY OUTPUT

Percentile	Weight gain y
6.25	3
18.75	4
31.25	5
43.75	6
56.25	7
68.75	8
81.25	9
93.75	10







Here:

1. The regression equation of weight gain on Initial weight(pounds) and Initial age (weeks) is:

   V = (-4.1917) + (0.1048)x1 + (0.8065)x2

2. Standard error = 0.9991

3. Weight gain is 35.4639 units

4. For testing null hypothesis B0 = 0, since p value = 0.077. It is insignificant

   For testing null hypothesis B1 = 0, since p value = 0.023. It is significant

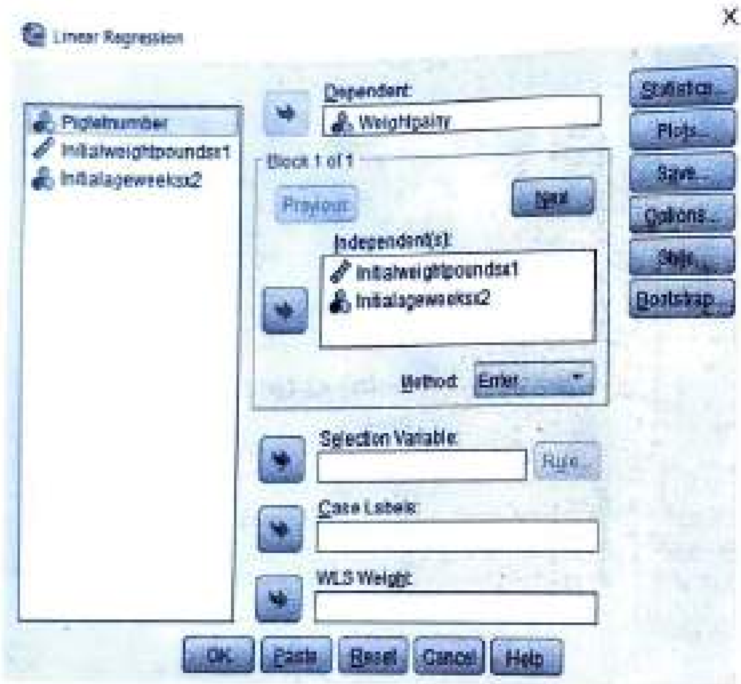
   For testing null hypothesis B2 = 0, since p value = 0.004. It is significant

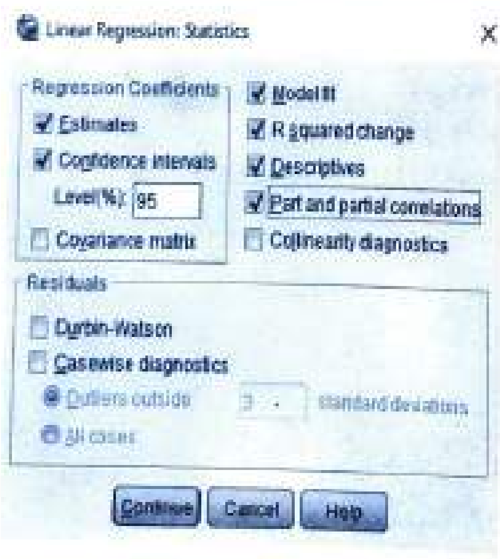
   For testing null hypothesis: overall fit of the regression coefficients =0, since here the p value = 0.0048 for F test, that indicates overall fit is significant

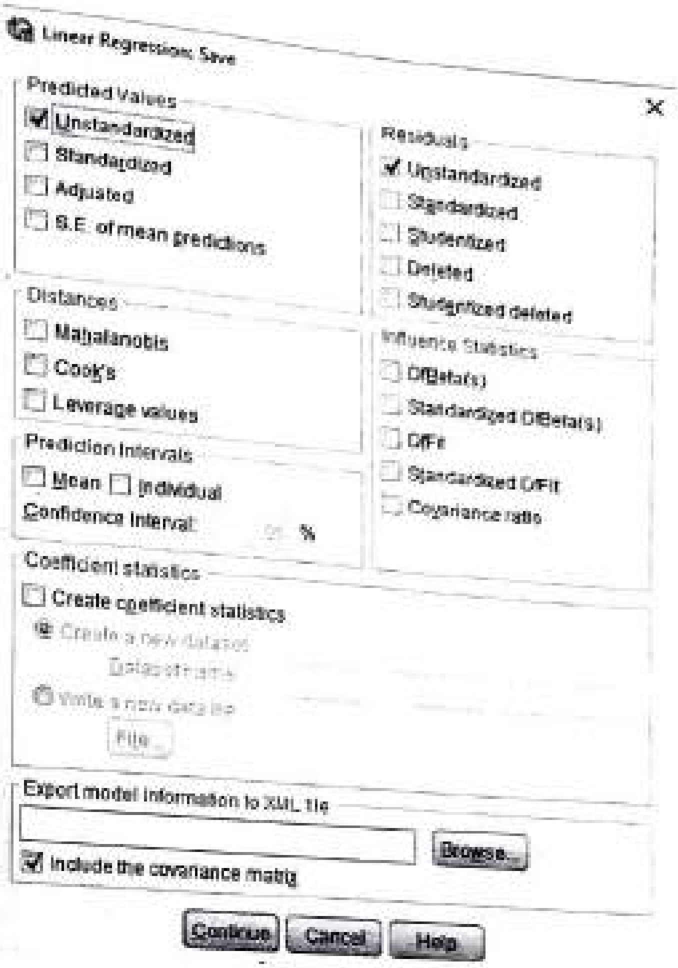
5. Here

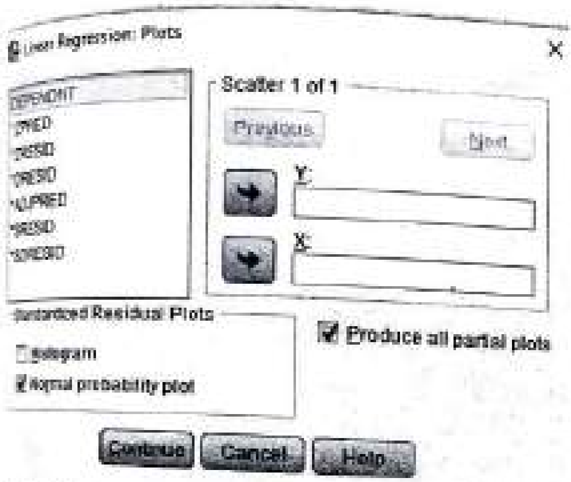
   Adj R2= 0.8336. That indicates this regression equation can represent 83.36% of the true observations.

How to use SPSS for Regression









Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Weight gain y	6.50	2.449	8
Initial weight(pounds)x1	45.25	11.696	8
Initial age (weeks) x2	7.38	2.387	8

Correlations

		Weight gain y	Initial weight (pounds)x1	Initial age (weeks) x2
Pearson Correlation	Weight gain y	1.000	.514	.794
	Initial weight(pounds)x1	.514	1.000	.017
	Initial age (weeks) x2	.794	.017	1.000
Sig. (1-tailed)	Weight gain y	.	.096	.009
	Initial weight(pounds)x1	.096	.	.484
	Initial age (weeks) x2	.009	.484	.
N	Weight gain y	8	8	8
	Initial weight(pounds)x1	8	8	8
	Initial age (weeks) x2	8	8	8

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Initial age (weeks) x2, Initial weight (pounds)x1[b]		Enter

a. Dependent Variable: Weight gain y

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.939[a]	.881	.834	.999	.881	18.539	2	5	.005

a. Predictors: (Constant), Initial age (weeks) x2, Initial weight(pounds)x1

b. Dependent Variable: Weight gain y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37.009	2	18.505	18.539	.005[b]
	Residual	4.991	5	.998		
	Total	42.000	7			

a. Dependent Variable: Weight gain y

b. Predictors: (Constant), Initial age (weeks) x2, Initial weight(pounds)x1

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	Upper Bound	Correlations Zero-order	Partial	Part
1	(Constant)	-4.192	1.888		-2.220	.077	-9.045	.662			
	Initial weight(pounds)x1	.105	.032	.501	3.247	.023	.022	.188	.514	.824	.50
	Initial age (weeks) x2	.807	.158	.786	5.097	.004	.400	1.213	.784	.916	.78

a. Dependent Variable: Weight gain y

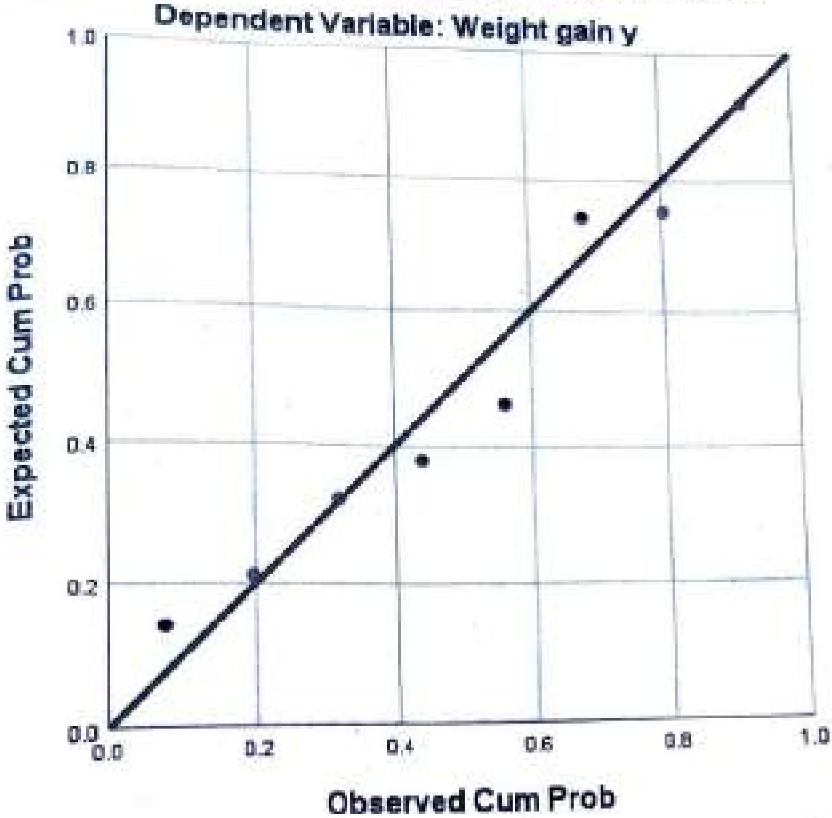
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.07	10.38	6.50	2.299	8
Residual	-1.075	1.409	.000	.844	8
Std. Predicted Value	-1.055	1.658	.000	1.000	8
Std. Residual	-1.076	1.411	.000	.845	8

a. Dependent Variable: Weight gain y

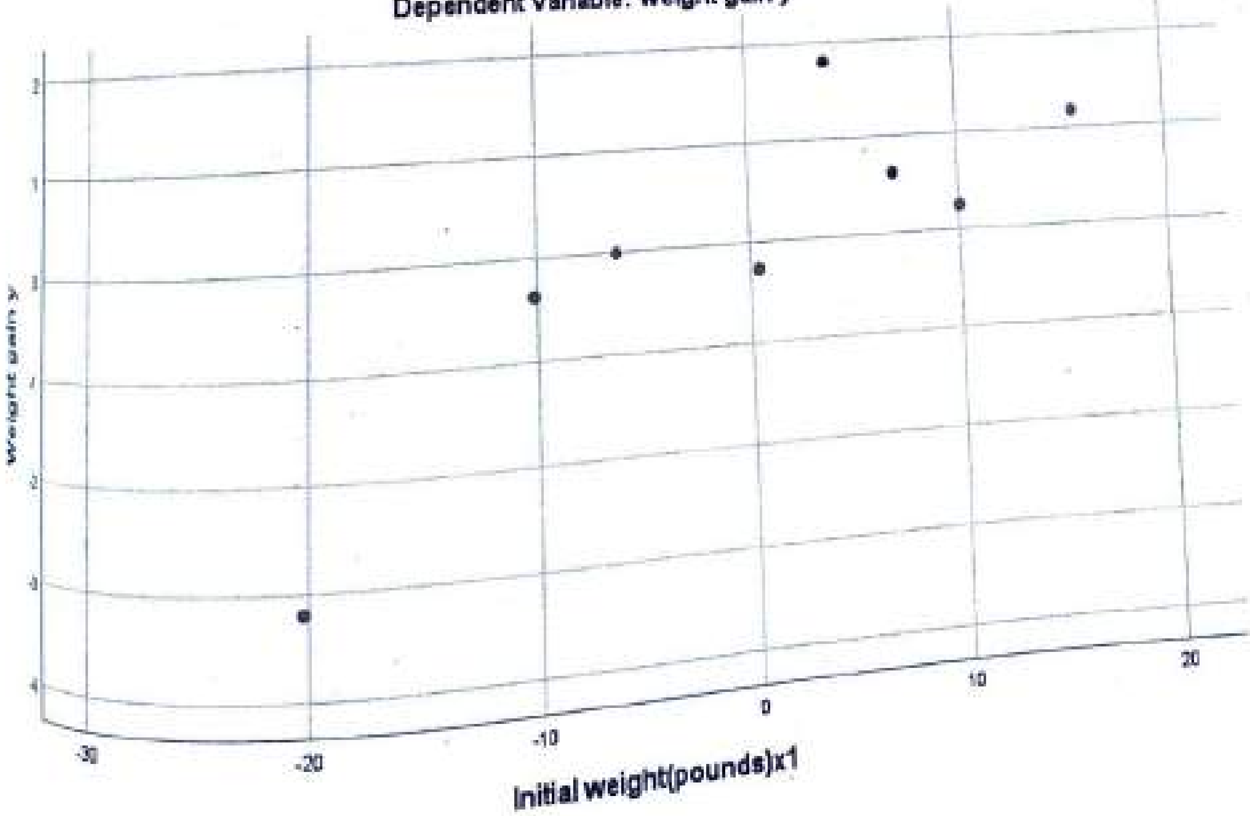
Normal P-P Plot of Regression Standardized Residual

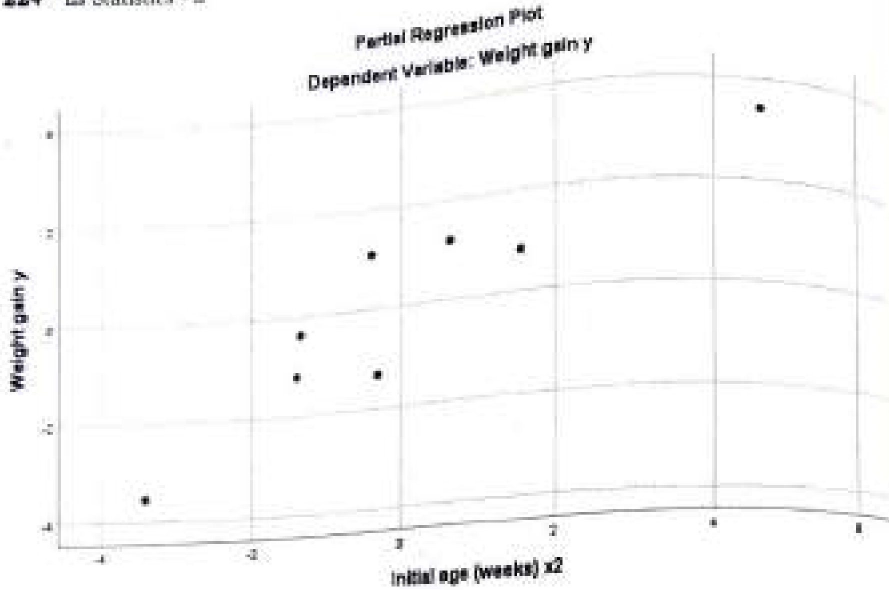
Dependent Variable: Weight gain y



Partial Regression Plot

Dependent Variable: Weight gain y





How to use STATA for Regression

(variable names replaced by y=Weightgainy, x1= Initialweightpoundsx1, x2= Initialageweeksx2)
STATA commands shown in the output display

```
. reg y x1 x2

      Source |       SS           df       MS      Number of obs   =         8
-------------+----------------------------------   F(2, 5)         =     18.54
       Model |  37.0092678         2  18.5046339   Prob > F        =    0.0049
    Residual |  4.99073219         5  .998146438   R-squared       =    0.8812
-------------+----------------------------------   Adj R-squared   =    0.8336
       Total |          42         7           6   Root MSE        =    .99907

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1048343   .0322915     3.25   0.023     .0218265    .1878422
          x2 |   .8065025   .1582366     5.10   0.004     .3997425    1.213263
       _cons |  -4.191709   1.888119    -2.22   0.077    -9.045274    .6618555
------------------------------------------------------------------------------

. display _b[_cons] + _b[x1]*9+ _b[x2]*48
35.463921
```

1. The regression equation of weight gain on Initial weight(pounds) and Initial age (weeks) is
y = (-4.1917) + (0.1048)x1 + (0.8065)x2
2. Standard error (Root MSE)= 0.9991
3. Weight gain is 35.4639 units (the display command)
4. Look at the regression output, P >|t| and for F test, Prob > F
5. Adj R2= 0.8336.

□□□