Chapter 4
Introduction to Multiple Linear Regression

Multiple Linear Regression Model, Partial Regression Coefficients, Testing Significance overall significance of Overall fit of the model, Testing for Individual Regression Coefficients(8 hours)

- In Chapter 3 we studied regression and correlation of two variables.
- scientific, social and economics phenomena do not scope to two variables only.
- we often need to give actual relationship between three or more variables and to explain the strength of association between them.
- For this multivariate regression and correlation are strong tools.

Example:

1. the cost of production of a product mainly depends on the cost of raw material, labour charges and cost of energy.
2. cost of crop mainly depends on the cost of seeds, fertilizer, irrigation etc.

MULTIPLE REGRESSION

Multiple regression is an extension of simple linear regression.

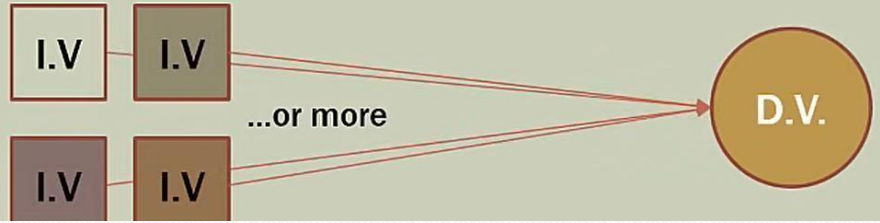
Simple linear regression
one-to-one

Multiple regression
many-to-one

MULTIPLE REGRESSION

Multiple regression is an extension of simple linear regression.

Simple linear regression
one-to-one



Multiple regression
many-to-one

NEW CONSIDERATIONS

- Adding more independent variables to a multiple regression procedure does not mean the regression will be “better” or offer better predictions; in fact it can make things worse. This is called OVERFITTING.
- The addition of more independent variables creates more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially *related to each other*. When this happens, it is called MULTICOLLINEARITY.

NEW CONSIDERATIONS

- Adding more independent variables to a multiple regression procedure does not mean the regression will be “better” or offer better predictions; in fact it can make things worse. This is called OVERFITTING.
- The addition of more independent variables creates more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially *related to each other*. When this happens, it is called MULTICOLLINEARITY.
- The ideal is for all of the independent variables to be correlated with the dependent variable but NOT with each other.

NEW CONSIDERATIONS

- Because of multicollinearity and overfitting, there is a fair amount of prep-work to do BEFORE conducting multiple regression analysis if one is to do it properly.

NEW CONSIDERATIONS

- Because of multicollinearity and overfitting, there is a fair amount of prep-work to do BEFORE conducting multiple regression analysis if one is to do it properly.
  - Correlations
  - Scatter plots
  - Simple regressions

MORE RELATIONSHIPS

Independent variables

milesTraveled, (x_1)

numDeliveries, (x_2)

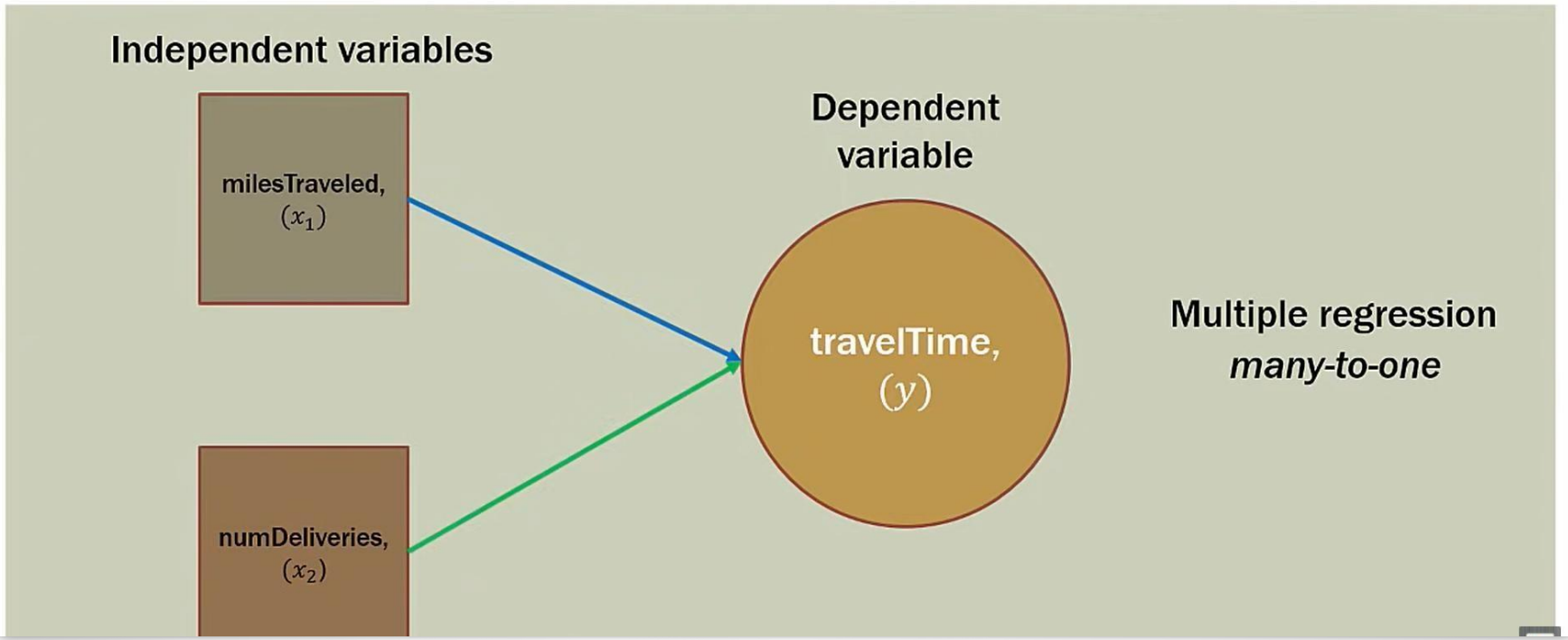
Dependent variable

travelTime, (y)

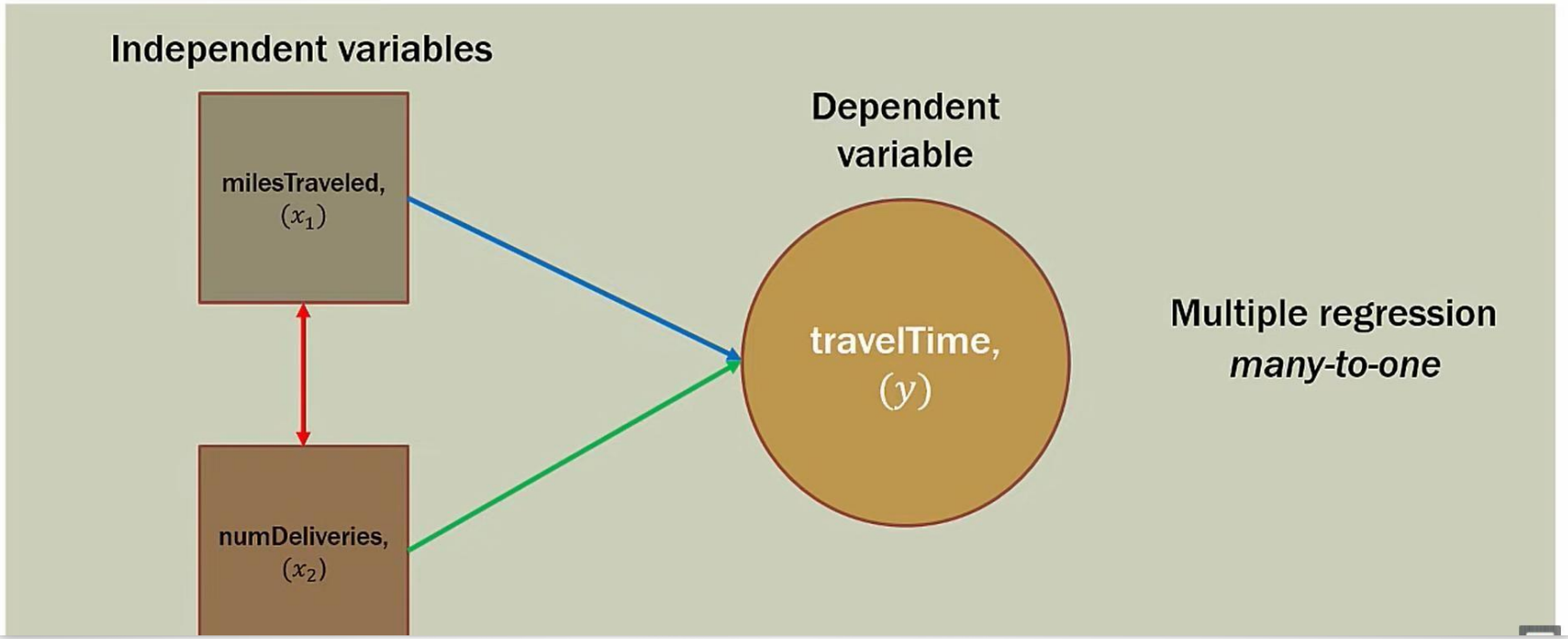
Multiple regression

many-to-one

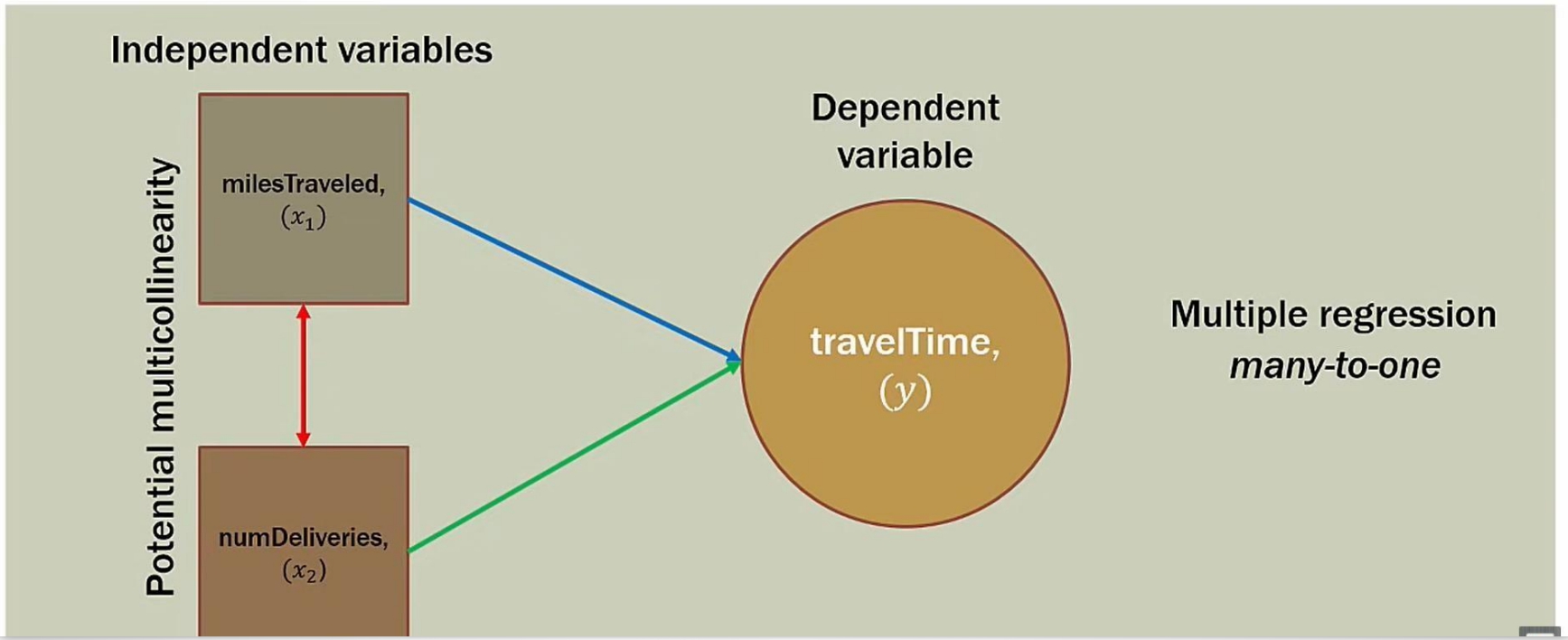
MORE RELATIONSHIPS



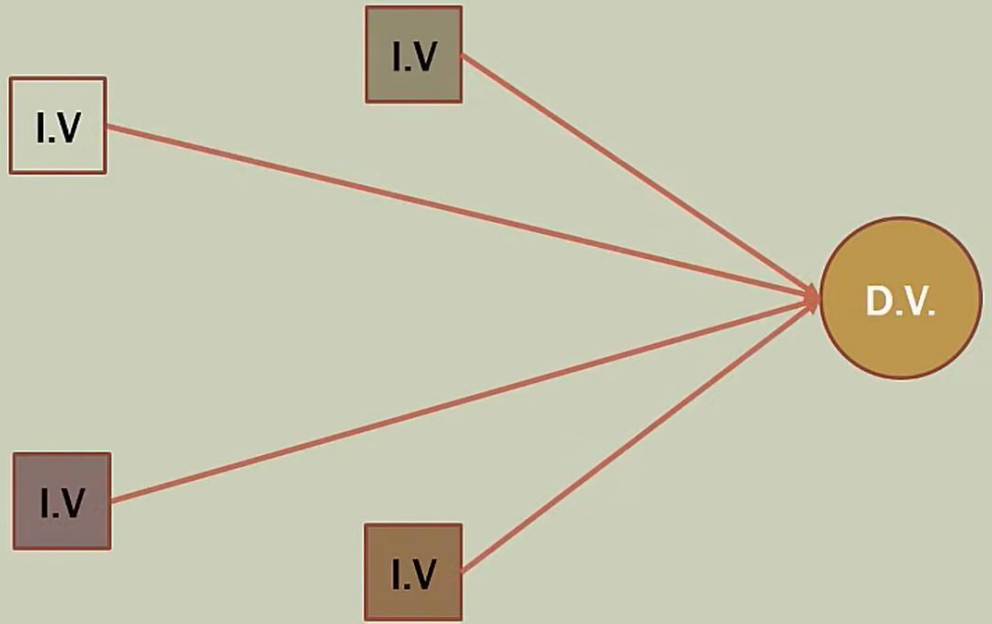
MORE RELATIONSHIPS



MORE RELATIONSHIPS

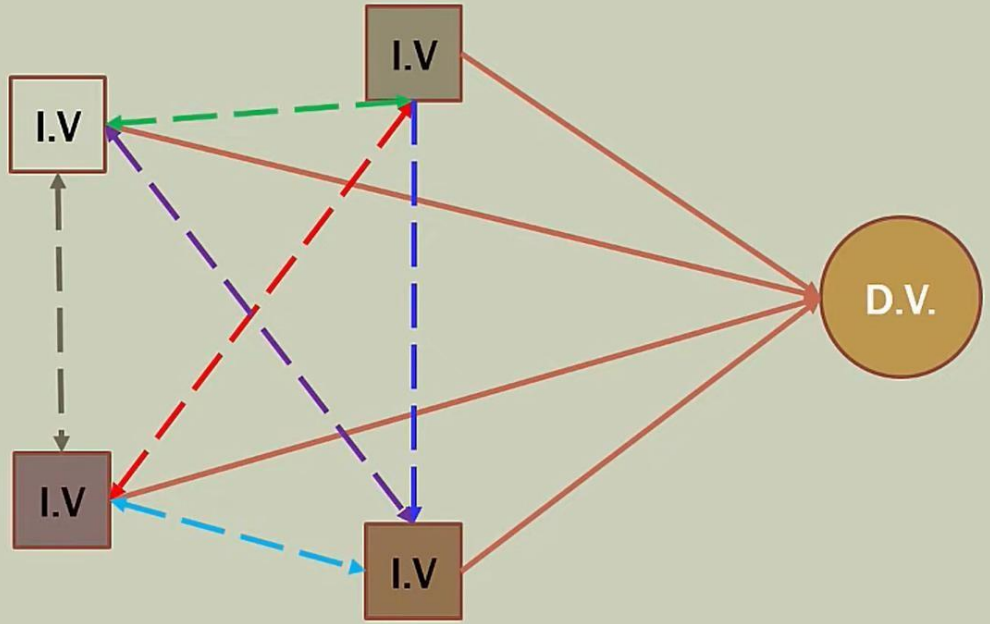


MANY RELATIONSHIPS



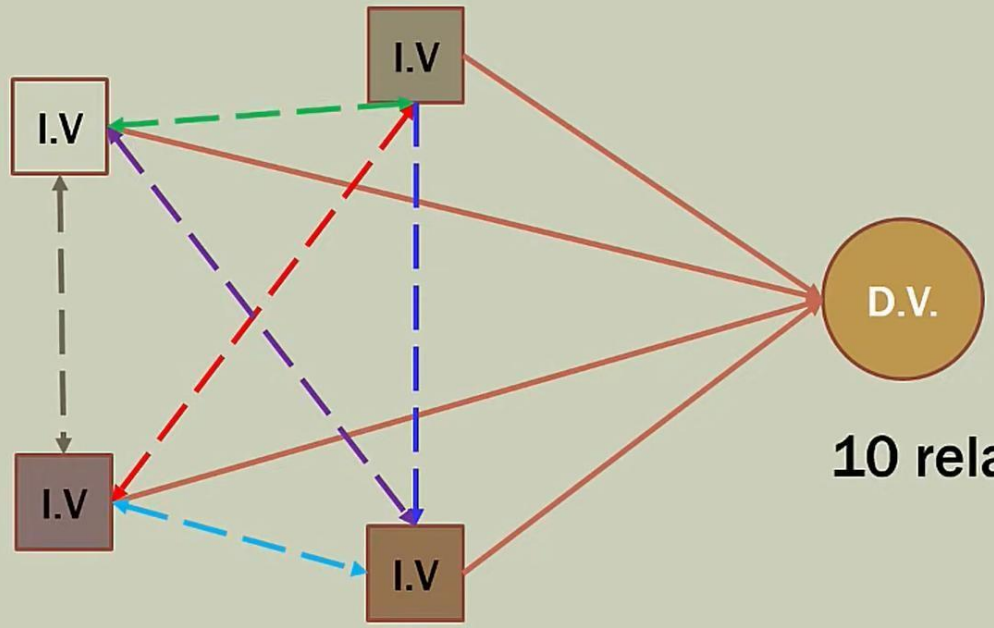
Multiple regression
many-to-one

MANY RELATIONSHIPS



Multiple regression
many-to-one

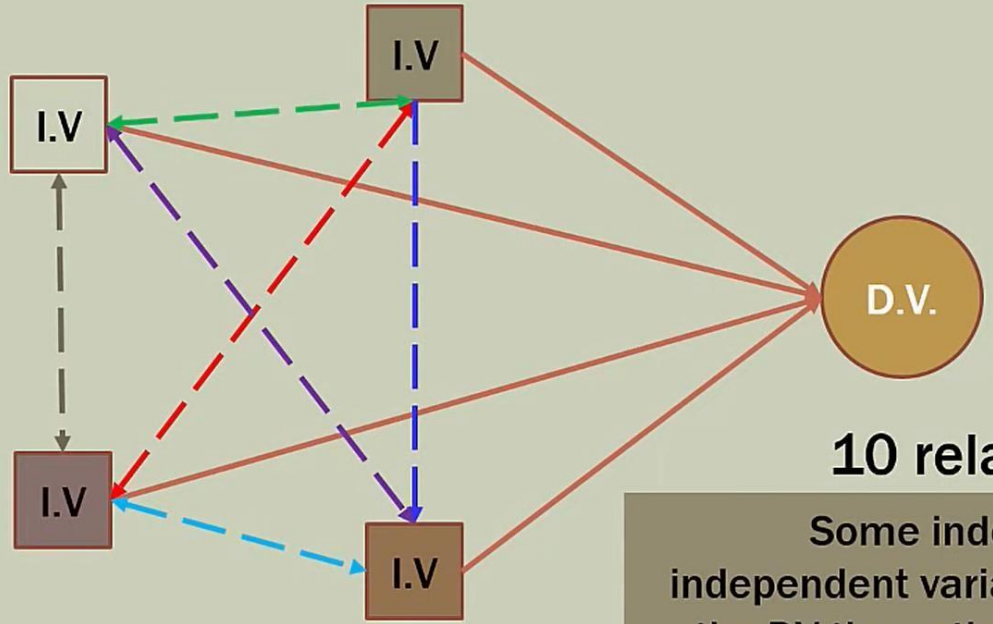
MANY RELATIONSHIPS



Multiple regression
many-to-one

10 relationships to consider!

MANY RELATIONSHIPS

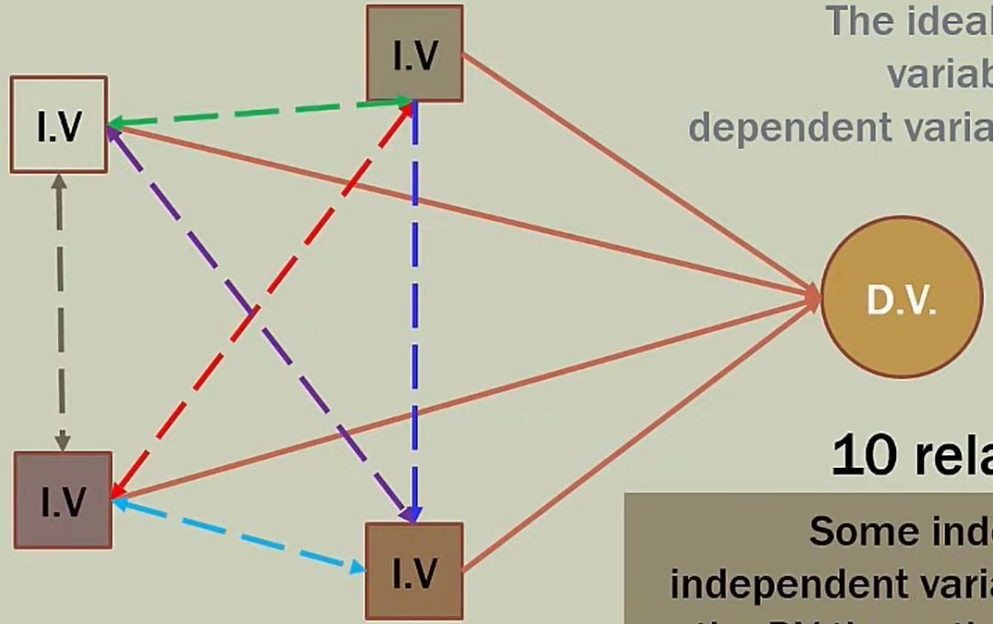


Multiple regression
many-to-one

10 relationships to consider!

Some independent variables, or sets of independent variables, are better at predicting the DV than others. Some contribute nothing.

MANY RELATIONSHIPS



The ideal is for all of the independent variables to be correlated with the dependent variable but NOT with each other.

Multiple regression
many-to-one

10 relationships to consider!

Some independent variables, or sets of independent variables, are better at predicting the DV than others. Some contribute nothing.

MULTIPLE REGRESSION MODEL

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

MULTIPLE REGRESSION MODEL

Multiple Regression Model

$$y = \boxed{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p} + \epsilon$$

linear parameters

MULTIPLE REGRESSION MODEL

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters

error

MULTIPLE REGRESSION MODEL

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters — error

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

error term assumed to be zero

MULTIPLE REGRESSION MODEL

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters — error

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

error term assumed to be zero

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_p x_p$$

$b_0, b_1, b_2, \ldots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

$\hat{y}$ = predicted value of the dependent variable

ESTIMATED MULTIPLE REGRESSION EQUATION

Example

$$\hat{y} = 6.211 + 0.014x_1 + 0.383x_2 - 0.607x_3$$

ESTIMATED MULTIPLE REGRESSION EQUATION

Example

$$\hat{y} = 6.211 + 0.014x_1 + 0.383x_2 - 0.607x_3$$

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

ESTIMATED MULTIPLE REGRESSION EQUATION

variables

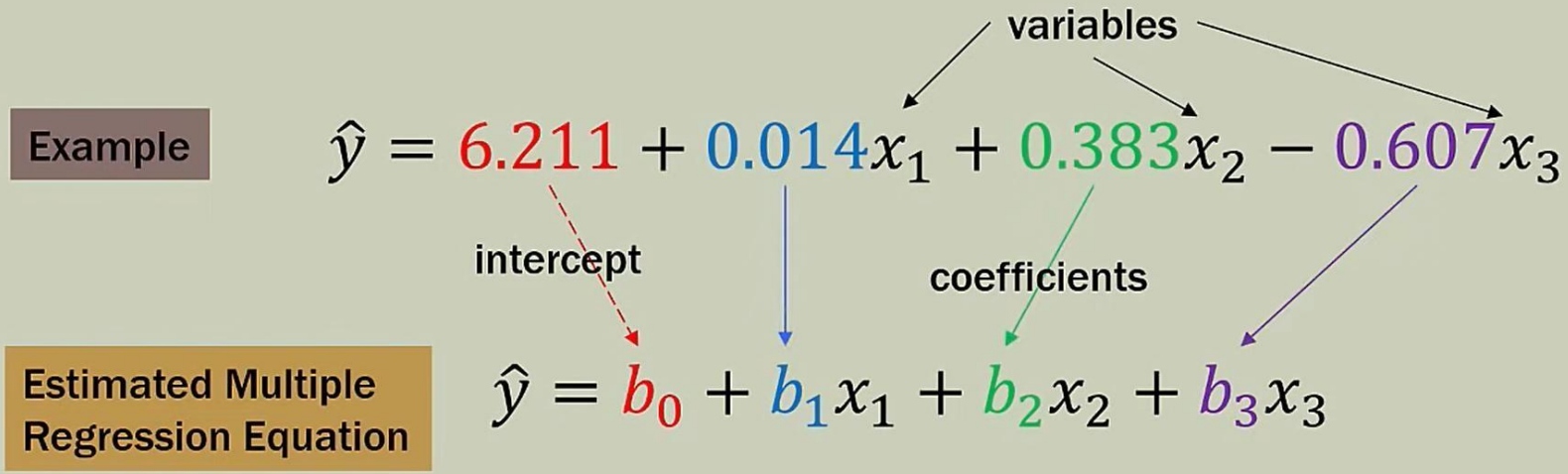
Example

$$\hat{y} = 6.211 + 0.014x_1 + 0.383x_2 - 0.607x_3$$

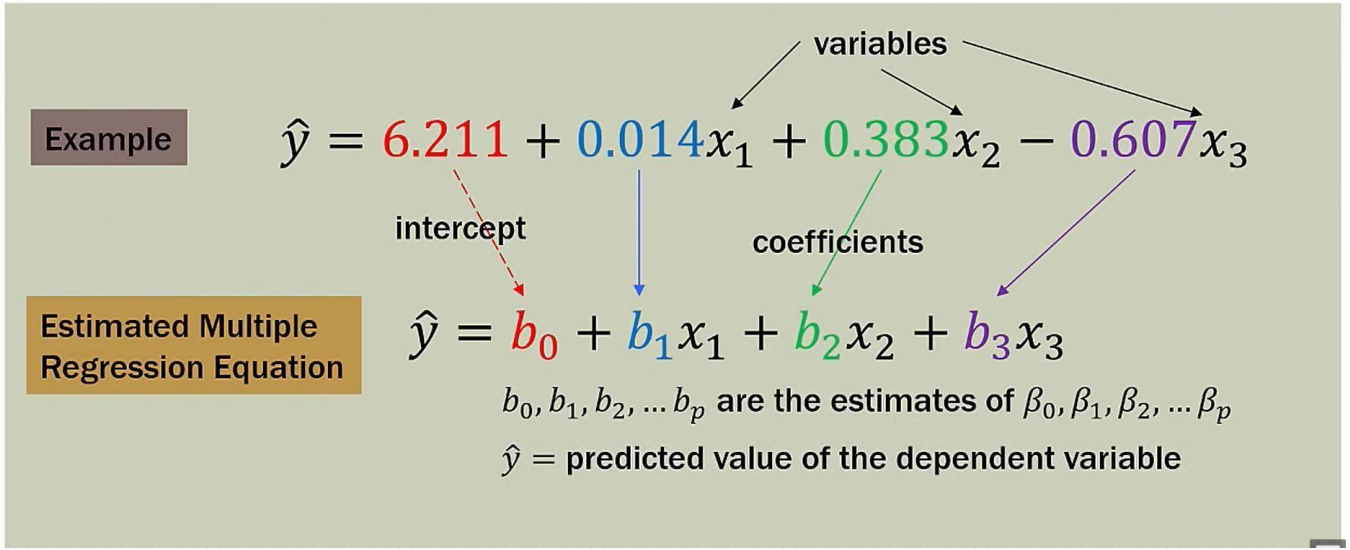
Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

ESTIMATED MULTIPLE REGRESSION EQUATION



ESTIMATED MULTIPLE REGRESSION EQUATION



What is partial regression coefficient?

Partial regression coefficients are the most important parameters of the multiple regression model. They **measure the expected change in the dependent variable associated with a one unit change in an independent variable holding the other independent variables constant**.

Or

It is often important to measure the correlation between a dependent variable and particular independent variable when all other variables involved are kept constant i.e, when the effects of all other variables are removed. This can be obtained by calculating coefficient of partial correlation.

Partial Correlation Coefficient

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variables with the effect of most of the variables eliminated.

If we denote by $\mathbf{r_{12.3}}$ the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, we find that

$$\mathbf{r_{12.3}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)}\sqrt{(1-r_{23}^2)}}$$

similarly,

$$\mathbf{r_{13.2}} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)}\sqrt{(1-r_{23}^2)}}$$

where $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$\mathbf{r_{23.1}} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)}\sqrt{(1-r_{13}^2)}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3 keeping X_1 constant.

Zero Order, First Order and Second Order Coefficients

- Simple coefficients (correlation between two variables only) are called **zero order coefficients**, since no variables are held constant.

- Partial coefficients such as $\mathbf{r_{12.3}}$, $\mathbf{r_{13.2}}$ are often referred to as **first order coefficients,** since one variable has been held constant.

- $\mathbf{r_{12.34}}$, $\mathbf{r_{13.24}}$, etc., are called **second order coefficients** since two variables are kept constant.

- The order of designation indicates the number of variables that have been held constant statistically.

Example: On the basis of the following information compute.

i) $r_{23.1}$ ii) $r_{13.2}$ and iii) $r_{12.3}$

$r_{12} = 0.70$ $r_{13} = 0.61$ $r_{23} = 0.40$

Example: On the basis of observations made on 39 cotton plants, the total correlation of yield of cotton X_1, number of seed vessels X_2 and height X_3 are found to be.

$r_{12} = 0.8$ $r_{13} = 0.65$ $r_{23} = 0.7$

Comment on the partial correlation between yield of cotton and the number of seed vessels, eliminating the effect of height.

Coefficient of Multiple Correlation

The coefficient of multiple linear correlation is represented by R_1 and it is common to add subscripts designating the variables involved. Thus, $R_{1\cdot234}$ would represent the coefficient of multiple linear correlation between X_1, on the one hand and X_2, X_3, and X_4, on the other. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiple correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows.

$$R_{1\cdot23} = \sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23} \;/\; 1 - r^2_{23}}$$

$$R_{2\cdot13} = \sqrt{r^2_{12} + r^2_{23} - 2\, r_{12}\, r_{13}\, r_{23} \;/\; 1 - r^2_{13}}$$

$$R_{3\cdot12} = \sqrt{r^2_{13} + r^2_{23} - 2\, r_{12}\, r_{13}\, r_{23} \;/\; 1 - r^2_{12}}$$

It should be noted that $R_{1\cdot23}$ is the same as $R_{1\cdot32}$. C.M.C **$R_{1\cdot23}$ lies between 0 and 1.**

Example: The following zero-order correlation coefficients are given

$r_{12} = 0.98 \quad r_{13} = 0.44 \quad r_{23} = 0.54$

Calculate multiple correlation coefficient treating 1st variable as dependent and 2nd and 3rd variable as independent.

Advantages of Multiple Correlation Analysis and

The coefficient of multiple correlation serves the following purposes:

1. It serves as a measure of the degree of association between one variable taken as the dependent variable and a group of other variables taken as the independent variables.
2. It also serves as a measure of goodness of fit of the calculated plane of regression and consequently as a measure of the general degree of accuracy of estimates made by reference to equation for the plane of regression.

MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis represents a logical extension of two-variable regression analysis. Instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable.

The following are the three main objectives of multiple regression analysis:

1. To derive an equation which provides estimates of the dependent variable from values of the two or more independent variables.
2. To obtain a measure of the error involved in using this regression equation as a basis for estimation.
3 To obtain a measure of the proportion of variance in the dependent variable accounted for the independent variables.

- The first purpose is accomplished by deriving an appropriate regression equation by the method of least squares.
- The second purpose is achieved through the calculation of a standard error of estimate.
- The third purpose is accomplished by computing the multiple coefficient of determination.

The normal equation for the Least Square Regression Plane:

X_1 on X_2 and X_3 has the equation where $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously the normal equation.

The normal equation are:

$$\sum X_1 = Na_{1.23} + b_{12.3} \sum X_2 + b_{13.2} \sum X_3$$

$$\sum X_1 X_2 = a_{1.23} \sum X_2 + b_{12.3} \sum X^2_2 + b_{13.2} \sum X_2 X_3$$

$$\sum X_1 X_3 = a_{1.23} \sum X_3 + b_{12.3} \sum X_2 X_3 + b_{13.2} \sum X^2_3$$

Deviation taken from Actual Means:

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$$

$$x_1 = (x_1 - \overline{x_1}) \;, \qquad x_2 = (x_2 - \overline{x_2}) \;, \qquad x_3 = (x_3 - \overline{x_3})$$

$$\sum x_1 x_2 = b_{12.3} \sum x^2_2 + b_{13.2} \sum x_2 x_3$$

$$\sum x_1 x_3 = b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x^2_3$$

The value of $b_{12.3}$ **and** $b_{13.2}$ **can also be obtained as follows:**

$$b_{12.3} = r_{12.3} \times \sigma_{1.23} / \sigma_{2.13}$$

$$b_{13.2} = r_{13.2} \times \sigma_{13.2} / \sigma_{3.12}$$

Remember

(1) If the variables X_1, X_2, X_3 are not measured from their means, then the equation of plane of regression of X_1 on (X_2, X_3) is given by

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

Similarly the plane of regression of X_2 on (X_1, X_3) is given

$$X_2 - \bar{X}_2 = b_{21.3}(X_1 - \bar{X}_1) + b_{23.1}(X_3 - \bar{X}_3)$$

and the plane of regression of X_3 on (X_1, X_2) is given

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

The equation of the plane of regression of X1 on X2 and X3

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3$$

$$b_{12.3} = \sigma_1 / \sigma_2 \times r_{12} - r_{13} r_{23} / 1 - r^2_{23}$$

$$b_{13.2} = \sigma_1 / \sigma_2 \times r_{13} - r_{12} r_{23} / 1 - r^2_{23}$$

The equation of the plane of regression of X_2 on X_1 and X_3

$$X_2 = a + b_{21.3} x_1 + b_{23.1} x_3$$

$$b_{21.3} = \frac{\sigma_2}{\sigma_1} \times \left(\frac{r_{21} - r_{13} \cdot r_{23}}{1 - r_{13}^2} \right)$$

$$b_{23.1} = \frac{\sigma_2}{\sigma_3} \times \left(\frac{r_{23} - r_{21} \cdot r_{31}}{1 - r_{13}^2} \right)$$

The equation of the plane of regression of x_3 on x_1 and x_2 is

$$x_3 = a + b_{31.2}x_1 + b_{32.1}x_2$$

$$b_{31.2} = \frac{\sigma_3}{\sigma_1}\left(\frac{r_{31} - r_{32}\cdot r_{12}}{1 - r_{12}^2}\right)$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2}\left(\frac{r_{32} - r_{31}\cdot r_{21}}{1 - r_{21}^2}\right)$$

The regression equation of X_1 on X_2 and X_3 can be expressed as follows:

$$(X_1 - \bar{X}_1) = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_3}\right)(X_3 - \bar{X}_3)$$

The regression equation of X_3 on X_2 and X_1 can be written as follows:

$$(X_3 - \bar{X}_3) = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{S_3}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \bar{X}_1)$$

Example: given the following determine the regression equation of:

i) X1 on X2 and X3 ii) X2 on X1 and X3

$r_{12} = 0.8$ $r_{13} = 0.6$ $r_{23} = 0.5$

$\sigma_1 = 10$ $\sigma_2 = 8$ $\sigma_3 = 5$

Example: in a trivariate distribution

$r_{12} = 0.7$ $r_{13} = r_{23} = 0.5$

$\sigma_1 = 2$ $\sigma_2 = \sigma_3 = 3$

Find i) $b_{12.3}$ ii) $b_{13.2}$

Example: Find the multiple linear regression equation of X1 on X2 and X3 from the data relating to three variables given below:

X1 =	4	6	7	9	13	15
X2 =	15	12	8	6	4	3
X3 =	30	24	20	14	10	4

YULE'S NOTATION

Let x1,x2 and x3 be three variables :

- regression equation of x1 on x2 and x3 becomes $x1=b12.3\ x2+b13.2\ x3$
- regression equation of x2 on x1 and x3 becomes $x2=b21.3\ x1+b23.1\ x3$
- regression equation of x3 on x1 and x2 becomes $x3=b31.2\ x1+b32.1\ x2$

Some Useful Relations

(*i*) In a trivariate distribution with variables X_1, X_2 and X_3 having known the means $\bar{x}_1, \bar{x}_2, \bar{x}_3$ and standard deviations s_1, s_2, s_3 of the three variables respectively and the simple correlation coefficients r_{12}, r_{13} and r_{23}. The multiple linear regression equation of X_1 on X_2 and X_3 can be obtained by the following determinant equation,

$$\begin{bmatrix} \frac{X_1 - \bar{x}_1}{s_1} & \frac{X_2 - \bar{x}_2}{s_2} & \frac{X_3 - \bar{x}_3}{s_3} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = 0 \qquad \ldots(16.39)$$

$$\frac{(X_1 - \bar{x}_1)}{s_1}(1 - r_{23}^2) - \frac{(X_2 - \bar{x}_2)}{s_2}(r_{21} - r_{31}r_{23}) + \frac{(X_3 - \bar{x}_3)}{s_3}(r_{21}r_{32} - r_{31}) = 0 \qquad \ldots(16.39.1)$$

(*ii*) The partial regression coefficients $b_{12.3}$ and $b_{21.3}$ can be obtained by the following formulae:

$$b_{12.3} = \frac{r_{12} - r_{23}r_{13}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2} \qquad \ldots(16.40)$$

$$b_{21.3} = \frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2} \cdot \frac{s_2}{s_1} \qquad \ldots(16.41)$$

(*iii*) The relation between two partial regression coefficients and partial correlation coefficient is,

$$b_{12.3} \times b_{21.3} = r_{12.3}^2 \qquad \ldots(16.42)$$

Example 16.2. In a trivariate population of random variables X_1, X_2, X_3, following lts about means, S.D. and correlation coefficients were found in a sample of size 20.

$$\bar{x}_1 = 40,\ \bar{x}_2 = 50,\ \bar{x}_3 = 60;$$

$$s_1 = 3,\ s_2 = 4,\ s_3 = 5;$$

$$r_{12} = 0.6,\ r_{13} = 0.5,\ r_{23} = 0.4.$$

Making use of the given information, we would, (*i*) find regression equation of X_1 on nd X_3, (*ii*) estimate X_1 when $X_2 = 70$ and $X_3 = 75$, (*iii*) calculate $b_{12.3}$, (*iv*)work out $r_{12.3}$.

(*i*) Regression equation of X_1 on X_2 and X_3 by the

i) Regression equation of x_1 on x_2 & x_3 by the equation.

$$\frac{(x_1-\bar{x}_1)}{s_1}(1-r_{23}^2)-\frac{(x_2-\bar{x}_2)}{s_2}(r_{21}-r_{31}r_{23})$$

$$+\frac{(x_3-\bar{x}_3)}{s_3}(r_{21}r_{32}-r_{31})=0$$

$$\frac{x_1-40}{3}(1-.4^2)-\frac{(x_2-50)}{4}(.6-.5\times.4)$$

$$+\frac{x_3-60}{5}(.6\times.4-.5)=0$$

$$\frac{x_1-40}{3} \times 0.84 - \frac{x_2-50}{4} \times 0.4 + \frac{x_3-60}{5} \times (-0.26) = 0$$

$$0.28(x_1-40) - 0.10(x_2-50) - 0.052(x_3-60) = 0$$

$$0.28x_1 - 0.10x_2 - 0.052x_3 = 3.08.$$

— eqn (1)

2) putting $x_2 = 70$ & $x_3 = 75$ in eqn (1) we get estimated value of x_1 as follows.

$$0.28\hat{x}_1 = 0.10 \times 70 + 0.052 \times 75 + 3.08$$

$$= 13.98$$

$\therefore \hat{x}_1 = 49.93$

3) $b_{12.3} = \frac{r_{12} - r_{23} r_{13}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2}$

$= \frac{.6 - .4 \times .5}{1 - .4^2} \times \frac{3}{4}$

$= \frac{.40}{.80} \times \frac{3}{4}$

$b_{12.3} = 0.357$

$$b_{21\cdot3} = \frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2} \times \frac{s_2}{s_1}$$

$$= \frac{.6 - .4 \times .5}{1 - .5^2} \times \frac{4}{3}$$

$$= \frac{.40}{.75} \times \frac{4}{3}$$

$$= 0.711$$

4) $r^2_{12\cdot3} = b_{12\cdot3} \times b_{21\cdot3}$

$= 0.357 \times 0.711$

$= 0.254$

$r_{12\cdot3} = \sqrt{0.254}$

$= 0.504.$

3 : From heights (X_1) in inches, weights (X_2) in kg., and ages (X_3) in years of a group of students, the following means, variances and correlation coefficients were obtained :

$$\bar{X}_1 = 40, \quad \bar{X}_2 = 50, \quad \bar{X}_3 = 20 ;$$

$$S_1 = 3, \quad S_2 = 4, \quad S_3 = 2 ;$$

$$r_{12} = 0.4 , \quad r_{23} = 0.5, \quad r_{13} = 0.25$$

where $\bar{X}_i$ is the mean of X_i, S_i^2 is the variance of X_i and r_{ij} is correlation coefficient between X_i and X_j for i, j = 1, 2, 3. Find the multiple regressive equation of X_3 (on X_1 and X_2) and estimate the value of X_3 when $X_1 = 43$ inches, $X_2 = 54$ kg.

$X_2 = 64\ X_3$.

Soln. :

Step I : The multiple regression equation of X_3 on X_1 and X_2 is given by :

$$X_3 - \bar{X}_3 = b_{31.2}\left(X_1 - \bar{X}_1\right) + b_{32.1}\left(X_2 - \bar{X}_2\right)$$

Given data : $\bar{X}_1 = 40$, $\bar{X}_2 = 50$, $\bar{X}_3 = 20$;

S_1 = Std. deviation = 3, $S_2 = 4$, $S_3 = 2$

$r_{12} = 0.4$, $r_{23} = 0.5$, $r_{13} = 0.25$

Step II :

Now, $b_{31.2} = \frac{S_3}{S_1}\left(\frac{r_{31} - r_{32}\, r_{12}}{1 - r_{12}^2}\right)$ and $b_{32.1} = \frac{S_3}{S_2}\left(\frac{r_{32} - r_{31}\, r_{21}}{1 - r_{21}^2}\right)$ ...(iii)

Substituting the given values of r_{ij}'s and S_i's in Equation (iii) from Equation (ii) and noting $r_{ij} = r_{ji}$; we get

$$b_{31.2} = \frac{2}{3}\left[\frac{0.25 - 0.5 \times 0.4}{1 - (0.4)^2}\right] = \frac{2}{3}\left[\frac{0.25 - 0.20}{1 - 0.16}\right] = \frac{2 \times 0.05}{3 \times 0.84} = 0.0397 = 0.04$$

and $$b_{312.1} = \frac{2}{4}\left[\frac{0.5 - 0.25 \times 0.4}{1 - (0.4)^2}\right] = \frac{2}{4}\left[\frac{0.5 - 0.1}{1 - 0.16}\right] = \frac{1 \times 0.4}{2 \times 0.84} = 0.2381 = 0.24$$

(i) the required equation

Step III : Substituting these values in Equation (i), the required equation of regression of X_3 on X_1 and X_2 becomes :

$$X_3 - 20 = 0.04\,(X_1 - 40) + 0.24\,(X_2 - 50)$$

$$\therefore \quad X_3 = 0.04\,X_1 + 0.24\,X_2 + (20 - 0.04 \times 40 - 0.24 \times 50)$$

$$\therefore \quad X_3 = 0.04\,X_1 + 0.24\,X_2 + 6.4$$

The estimated value of X_3 when $X_1 = 43$ inches and $X_2 = 54$ kg is given by

$$\hat{X}_3 = 0.04 \times 43 + 0.24 \times 54 + 6.4 = 17.2 + 12.9 + 6.40 = \mathbf{21.08\ years}$$

$$= \hat{X}_3 = 21.08 \text{ years}$$

Given $\bar{x_1} = 48, \sigma_1 = 3, r_{12} = 0.6, \bar{x_2} = 40, \sigma_2 = 4, r_{13} = 0.7, \bar{x_3} = 62, \sigma_3 = 5, r_{23} = 0.8$, estimate the value of x_3 when $x_1 = 30$ and $x_2 = 50$.

Solution:

Multiple Regression of x_3 on x_1 and x_2 will give the value of x_3 for $x_1 = 30$ and $x_2 = 50$, this equation is.

$$x_3 = a_{3.12} + b_{31.2}x_1 + b_{32.1}x_2$$

The values of regression coefficients:

$$b_{31.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} \times \frac{\sigma_3}{\sigma_1}$$

$$= \frac{0.7 - (0.6)(0.8)}{1 - (0.6)^2} \times \frac{5}{3} = 0.57$$

$$b_{32.1} = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2} \times \frac{\sigma_3}{\sigma_2}$$

$$= \frac{0.8 - (0.6)(0.7)}{1 - (0.6)^2} \times \frac{5}{4} = 0.742$$

Now

$$a_{3.12} = \bar{x_3} - b_{31.2}\bar{x_1} - b_{32.1}\bar{x_2}$$
$$= 62 - (0.57 \times 48) - (0.74 \times 40) = 5.04$$

Substitute the value of $a_{3.12}$ in equation 9.6

$$x_3 = 5.04 + (0.57 \times 30) + (0.74 \times 50) = 59.14$$

$\therefore$ Thus estimated value of $x_3 = 59.14$

Given the following, determine the regression equation of:

(i) X_1 on X_2 and X_3 (Ans: $X_1 = 0.833X_2 + 0.533X_3$)

(ii) X_2 and X_1 and X_3 (Ans: $X_2 = 0.625X_1 + 0.05X_3$)

$r_{12} = 0.8, r_{13} = 0.6, r_{23} = 0.5, \sigma_1 = 10, \sigma_2 = 8, \sigma_3 = 5$

Given $r_{12} = 0.28$, $r_{23} = 0.49$, $r_{31} = 0.51$, $\sigma_1 = 2.7, \sigma_2 = 2.4, \sigma_3 = 2.7$. Find the regression equation of X_3 on X_1 and X_2.

Ans) $X_3 = 0.4045X_1 + 0.4238X_2$

$$r_{12.3} = \sqrt{b_{12.3} \times b_{21.3}} \qquad r_{13\cdot2} = \sqrt{b_{13\cdot2} \times b_{31\cdot2}}$$

Problems

9.22 If $r_{12} = 0.28, r_{23} = 0.49$ and $r_{13} = 0.51$ calculate $r_{12.3}$ and $r_{13.2}$

9.23 Given $r_{12} = 0.28$, $r_{23} = 0.49$, $r_{13} = 0.51$ $\sigma_1 = 2.7$, $\sigma_2 = 2.4$, $\sigma_3 = 0.27$ Find the regression equation of x_3 on x_1 and x_2. If the variables have been measured from their actual means.

9.24 From the following data, find regression equation:

Wheat yield (x_1)(per hectare Quintals)	40	45	50	65	70	70	80
Use of Fertilizers (x_2)(Kg. per hectare)	10	20	30	40	50	60	70
Rainfall (x_3) (inches)	36	33	37	37	34	32	36

Also calculate the value of x_3 when $x_1 = 45$ and $x_1 = 30$

- From the data given below, estimate the value of x_3 when $x_1 = 58$ and $x_2 = 52.5$

$$\bar{x_1} = 55.95\ \sigma_1 = 2.26,\ r_{12} = 0.578$$

$$\bar{x_2} = 51.48\ \sigma_2 = 4.39,\ r_{13} = 0.581$$

$$\bar{x_3} = 56.03\ \sigma_3 = 4.41,\ r_{23} = 0.974$$

- Find the multiple linear regression of x_1 on x_2 and x_3 from the data relating to three variables given below:

x_1	4	6	7	9	13	15
x_2	15	12	8	6	4	3
x_3	30	24	20	14	10	4

What is a Hypothesis?

A hypothesis is a calculated prediction or assumption about a population parameter based on limited evidence. The whole idea behind hypothesis formulation is testing—this means the calculated assumption to a series of evaluations to know whether they are true or false.

Or

Hypothesis testing can be defined as tests performed to evaluate whether a claim or theory about something is true or otherwise.

What are the Types of Hypotheses?

1. **Null Hypothesis**
2. **Alternative Hypothesis**

The procedure that enable us to decide whether to accept or reject the hypotheses are called test of significance.

Null Hypothesis

"There's no relationship between the variables in an observation"

- independent variable have no effect on the dependent variable.

Examples of Null Hypothesis

- This is no significant change in a student's performance if they drink coffee or tea before classes.
- There's no significant change in the growth of a plant if one uses distilled water only or vitamin-rich water

Null hypothesis symbol:

- The symbol for the null hypothesis is $\mathbf{H_0}$, and it is read as **H-null, H-zero, or H-naught.**
- The null hypothesis is usually associated with just **= 'equals to'** sign as a null hypothesis can either be accepted or rejected.

$H_0: \mu_1 = \mu_2$ H_0= null hypothesis μ_1= mean of A μ_2= mean of B

Alternative Hypothesis

An alternative hypothesis is **an opposing theory to the null hypothesis**. For example, if the null hypothesis predicts something to be true, the alternative hypothesis predicts it to be false.

independent variable have effect on the dependent variable.

Alternative hypothesis symbol

The symbol of the alternative hypothesis is either **H_1 or H_a**

while using **< (less than), > (greater than) or ≠ not equal signs.**

Examples of Alternative Hypotheses

- Starting your day with a cup of tea instead of a cup of coffee can make you more energetic in the morning.
- If a researcher is assuming that the bearing capacity of a bridge is more than 10 tons, then the hypothesis under this study will be:

Null hypothesis H_0: $\mu = 10$ tons

Alternative hypothesis H_a: $\mu > 10$ tons

Type 1 error

- Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a **null hypothesis** when it is true.
- Type 1 error is caused when the hypothesis that should have been accepted is rejected.
- Type I error is denoted by α (alpha), known as an error, also called the level of significance of the test.
- This type of error is a false positive error where the null hypothesis is rejected based on some error during the testing.
- Type 1 error occurs when the null hypothesis is rejected even when there is no relationship between the variables.
- As a result of this error, the researcher might believe that the hypothesis works even when it doesn't.

Type II error

- Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
- In simple words, Type II error means accepting the hypothesis when it should not have been accepted.
- The type II error results in a false negative result.
- The Type II error is denoted by β (beta) and is also termed the beta error.
- Type II error occurs when the null hypothesis is acceptable considering that the relationship between the variables is because of chance, and even when there is a relationship between the variables.
- As a result of this error, the researcher might believe that the hypothesis doesn't work even when it should.

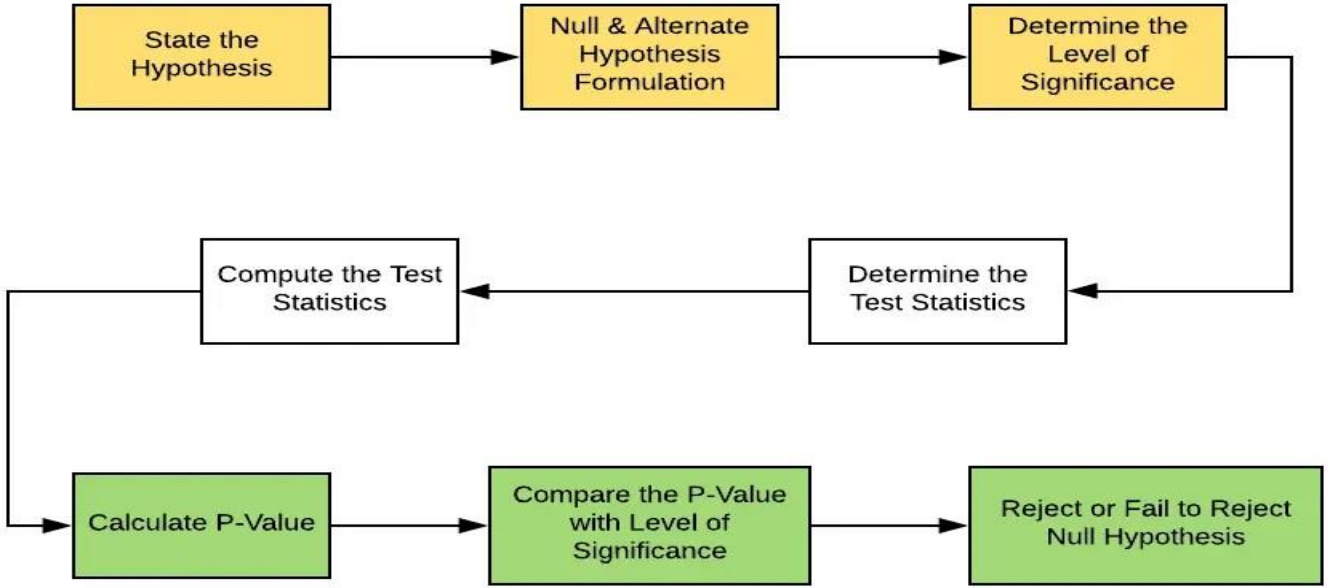
	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

- **Type I Error**: When one rejects the Null Hypothesis (H0 – Default state of being) given that H0 is true, one commits a Type I error. It can also be termed as false positive.
- **Type II Error**: When one fails to reject the Null hypothesis when it is actually false or does not hold good, one commits a Type II error. It can also be termed as a false negative.
- In other cases when one rejects the Null Hypothesis when it is false or not true, and when fails to reject the Null hypothesis when it is true is the **correct decision**.

How is Hypothesis testing related to significance level?

In order to perform hypothesis tests, the following steps need to be taken:

- Hypothesis formulation: Formulate the null and alternate hypothesis
- Data collection: Gather the sample of data
- Statistical tests: Determine the statistical test and test statistics. The statistical tests can be z-test or t-test depending upon the number of data samples and/or whether the population variance is known otherwise.
- Set the level of significance
- Calculate the p-value
- Draw conclusions: Based on the value of p-value and significance level, reject the null hypothesis or otherwise.



Hypothesis Testing Workflow

Why does one need a level of significance?

In hypothesis tests, if we do not have some sort of threshold by which to determine whether your results are statistically significant enough for you to reject the null hypothesis, then it would be tough for us to determine whether your findings are significant or not.

This is why we take into account levels of significance when performing hypothesis tests and experiments.

Since hypothesis testing helps us in making decisions about our data, having a level of significance set up allows one to know what sort of chances their findings might have of actually being due to the null hypothesis. If you set your level of significance at 0.05 for example, it would mean that there's only a five percent chance that the difference between groups (assuming two groups are tested) is due to random sampling error.

TESTING THE SIGNIFICANCE OF THE OVERALL MODEL

The population model for the multiple regression equation is

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk + \epsilon$$

where x1,x2,…,xk are the independent variables, β0,β1,…,βk are the population parameters of the regression coefficients, and ϵ is the error variable.

- The error variable ϵ accounts for the variability in the dependent variable that is not captured by the linear relationship between the dependent and independent variables.
- The value of ϵ cannot be determined, but we must make certain assumptions about ϵ and the errors/residuals in the model in order to conduct a hypothesis test on how well the model fits the data.

Testing the Overall Model

We want to test if there is a relationship between the dependent variable and the **set** of independent variables. In other words, we want to determine if the regression model is valid or invalid.

Invalid Model.

There is no relationship between the dependent variable and the set of independent variables.

In this case, all of the regression coefficients βi in the population model are zero.

This is the claim for the null hypothesis in the overall model test: **H0:β1=β2=⋯=βk=0.**

Valid Model.

There is a relationship between the dependent variable and the set of independent variables.

In this case, at least one of the regression coefficients βi in the population model is not zero.

This is the claim for the alternative hypothesis in the overall model test:

Ha: at least one βi ≠ 0: at least one .

The logic behind the overall model test is based on two independent estimates of the variance of the errors:

•One estimate of the variance of the errors, MSR, is based on the mean amount of explained variation in the dependent variable y.

•One estimate of the variance of the errors, MSE, is based on the mean amount of unexplained variation in the dependent variable y.

Definitions for Regression with Intercept

n is the number of observations; k is the number of regression parameters

Corrected Sum of Squares for Model: also called sum of squares for regression.

$$\text{(SSR) SSM} = \sum (\hat{y} - \bar{y})^2$$

Sum of Squares for Error: also called sum of squares for residuals.

$$\text{SSE} = \sum (y_i - \hat{y})^2,$$

Corrected Sum of Squares Total: $\text{SST} = \Sigma\, (y_i - \bar{y})^2$

This is the sample variance of the y-variable multiplied by n - 1.

- For multiple regression models, we have this remarkable property:

$$\text{SST} = \text{SSM} + \text{SSE}$$

Corrected Degrees of Freedom for Model: DFM = k – 1

Degrees of Freedom for Error: DFE = n – k

Corrected Degrees of Freedom Total: DFT = n - 1

Subtract 1 from n for the corrected degrees of freedom

Horizontal line regression is the null hypothesis model.

For multiple regression models with intercept, DFM + DFE = DFT.

Mean of Squares for Model: MSM = SSM / DFM

Mean of Squares for Error: MSE = SSE / DFE

The sample variance of the residuals.

Mean of Squares Total: MST = SST / DFT

The sample variance of the y-variable.

In general, a researcher wants the variation due to the model (MSM) to be large with respect to the variation due to the residuals (MSE).

The **F-test for linear regression** tests whether any of the independent variables in a multiple linear regression model are significant.

Or

The F-Test of overall significance in regression is a test of whether or not your linear regression model provides a better fit to a dataset than a model with no predictor variables.

The F-test

- For a multiple regression model with intercept, we want to test the following null hypothesis and alternative hypothesis:

$$\mathbf{H_0: \ \beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0}$$

$$\mathbf{H_1: \ \beta_k \neq 0, \text{ for at least one value of } k}$$

This test is known as the overall **F-test for regression**.

- Here are the five steps of the **overall F-test for regression**

  State the null and alternative hypotheses:

  H_0: $\beta_1 = \beta_2 = ... = \beta_{k-1} = 0$; H_1: $\beta_k \neq 0$, for at least one value of k

  Compute the test statistic assuming that the null hypothesis is true:

  F = MSM / MSE = (explained variance) / (unexplained variance)

  - Find a (1 - α)100% confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.
  - Accept the null hypothesis if F ∈ I; reject it if F ∉ I.
  - Use statistical software to determine the p-value.

Practice Problem:
For a multiple regression model with 35 observations and 9 independent variables (10 parameters), SSE = 134 and SSM = 289, test the null hypothesis that all of the regression parameters are zero at the 0.05 level.

Solution:

DFE = n - p = 35 - 10 = 25 and DFM = p - 1 = 10 - 1 = 9.

Here are the five steps of the test of hypothesis:

1. State the null and alternative hypothesis:

H_0: $\beta_1 = \beta_2 = , \ldots , = \beta_{p-1} = 0$

H_1: $\beta_j \neq 0$ for some j

2. Compute the test statistic:

F = MSM / MSE = (SSM / DFM) / (SSE / DFE)

= (289 / 9) / (134 / 25) = 32.111 / 5.360

= 5.991

3. Find a (1 - 0.05) × 100% confidence interval for the test statistic.
Look in the F-table at the 0.05 entry for 9 df in the numerator and 25 df in the denominator. This entry is 2.28, so the 95% confidence interval is [0, 2.34].
This confidence interval can also be found using the R function call qf(0.95, 9, 25).

4. Decide whether to accept or reject the null hypothesis: $5.991 \notin [0, 2.28]$, so reject H_0.

5. Determine the p-value.
However, we can find a rough approximation to the p-value by examining the other entries in the F-table for (9, 25) degrees of freedom:

Level	Confidence Interval	F-value
0.100	[0, 0.900]	1.89
0.050	[0, 0.950]	2.28
0.025	[0, 0.975]	2.68
0.010	[0, 0.990]	2.22
0.001	[0, 0.999]	4.71

The F-value is 5.991, so the p-value must be less than 0.005.

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588

Testing the Regression Coefficients

For an individual regression coefficient, we want to test if there is a relationship between the dependent variable y and the independent variable xi.

- **No Relationship**. There is no relationship between the dependent variable y and the independent variable xi.
- In this case, the regression coefficient βi is zero.
- This is the claim for the null hypothesis in an individual regression coefficient test: **H0: βi=0.**
- **Relationship.** There is a relationship between the dependent variable y and the independent variable xi.
- In this case, the regression coefficients βi is not zero.
- This is the claim for the alternative hypothesis in an individual regression coefficient test: Ha:βi≠0.
- We are not interested if the regression coefficient βi is positive or negative, only that it is not zero.
- This makes the test on a regression coefficient a two-tailed test.

T-test definition

The t-test is a test in statistics that is used for testing hypotheses regarding the mean of a small sample taken population when the standard deviation of the population is not known.

- The t-test is used to determine if there is a significant difference between the means of two groups.
- The t-test is used for hypothesis testing to determine whether a process has an effect on both samples or if the groups are different from each other.
- Basically, the t-test allows the comparison of the mean of two sets of data and the determination if the two sets are derived from the same population.
- After the null and alternative hypotheses are established, t-test formulas are used to calculate values that are then compared with standard values.
- Based on the comparison, the null hypothesis is either rejected or accepted.
- The T-test is similar to other tests like the z-test and f-test except that t-test is usually performed in cases where the sample size is small ($n \leq 30$).

The formula for the manual calculation of t-value is given below:

$$t = \frac{\bar{x} + \mu}{\frac{\sigma^2}{\sqrt{n}}}$$

where $\bar{x}$ is the mean of the sample,
μ is the assumed mean,
σ is the standard deviation,
n is the number of observations

T-test for the difference in mean:

$$t = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{(n1-1)\sigma1^2 + (n2-1)\sigma2^2}{n1+n2-2}} \times \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

where $\bar{x}1$ and $\bar{x}2$ are the mean of two samples.
$\sigma1$ and $\sigma2$ is the standard deviation of two samples.
$n1$ and $n2$ are the numbers of observation of two samples.

T-test example

If a sample of 10 copper wires is found to have a mean breaking strength of 527 kgs, is it feasible to regard the sample as a part of a large population with a mean breaking strength of 578 kgs and a standard deviation of 12.72 kgs? Test at 5% level of significance.

Taking the null hypothesis that the mean breaking strength of the population is equal to 578 kgs, we can write:

$H_0 : \mu = 578$ kgs

$H_a : \mu \neq 578$ kgs

$\bar{x} = 527$ kgs , $\sigma = 12.72$, $n = 10$.

Based on the assumption that the population to be normal, the formula for the test statistic ***t*** can be written as:

$$t = \frac{\bar{x} + \mu}{\frac{\sigma^2}{\sqrt{n}}}$$

t = (527+578) / ($12.72^2/\sqrt{10}$)

t = 21.597

As H_a is two-sided in the given question, a two-tailed test is to used for the determination of the rejection regions at a 5% level of significance which comes to as under, using normal curve area table:

***R* : | *t* | > 1.96**

The observed value of *t* is -1.488 which is in the acceptance region since *R*: | *t* | > 1.96, and thus, H_0 is accepted.

Find the t-test value for the following given two sets of values: 7, 2, 9, 8 and 1, 2, 3, 4 ?

.

Steps to Conduct a Hypothesis Test on a Regression Coefficient

1. Write down the null hypothesis that there is no relationship between the dependent variable y and the independent variable xi:

$$H0: \beta i = 0$$

2. Write down the alternative hypotheses that is a relationship between the dependent variable y and the independent variable xi:

$$Ha: \beta i \neq 0$$

3. Collect the sample information for the test and identify the significance level α.

4. The *p*-value is the sum of the area in the tails of the t-distribution.

5. The t-score and degrees of freedom are

$$t = \frac{b_i - \beta i}{S_{b_i}}$$

$$df = n - k - 1$$

6. Compare the *p*-value to the significance level and state the outcome of the test:

- If *p*-value≤α≤, reject H0 in favour of Ha.
  - The results of the sample data are significant. There is sufficient evidence to conclude that the null hypothesis H0 is an incorrect belief and that the alternative hypothesis Ha is most likely correct.
- If *p*-value>α, do not reject H0.
  - The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis Ha may be correct.

Simple Linear Regression Analysis - Let's Practice

14-25. A regional retailer would like to determine if the variation in average monthly store sales can, in part, be explained by the size of the store measured in square feet (continuation of 14-12)

a) Compute the simple linear regression model using the sample data to determine whether variation in average monthly sales can be explained by store size. Interpret the slope and intercept coefficients.

b) Test for the significance of the slope coefficient of the regression model. Use a level of significance of 0.05.

c) Based on the estimated regression model, what percentage of the total variation in average monthly sales can be explained by store size?

Test for Significance of the Regression Slope Coefficient

- Hypotheses:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

A slope of 0 implies there is NO LINEAR RELATIONSHIP between x and y, and that x in its linear form is of no use in explaining the variation in y.

- Testing Approaches: Critical Value and p-value

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad df = n - 2$$

b_1 - Sample regression slope coefficient
β_1 - Hypothesized slope (usually $\beta_1 = 0$)
s_{b_1} - Estimator of the standard error of the slope

If p-value < $\alpha / 2$, reject H_0

Simple Linear Regression Analysis - Let's Practice

b) $H_0 : B_1 = 0.0$ (no linear relationship; not significant)

$H_a : B_1 \neq 0.0$ (there is a linear relationship; significant)

$\alpha = 0.05$ Degrees of Freedom = n – 2 = **19**

=T.INV.2T(α, df) Critical t (Appendix F) = **± 2.093**

		Coefficients	*Standard Error*	*t Stat*	*P-value*
b_0	Intercept	171205.8279	59846.1252	2.8608	0.0100
b_1	Store Size (Sq. Ft.)	25.3160	3.5767	7.0780	0.0000

$$t = \frac{b_1 - B_1}{s_{b_1}} = \frac{25.316 - 0}{3.5767} = 7.08$$

Conclusion: Since 7.08 is greater than CV 2.093, Reject H_0 and conclude that the population slope coefficient is significant, there is a linear relationship.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745

Test Statistic for Significance of the Coefficient of Determination

$H_o : \rho^2 = 0.0$

$H_A : \rho^2 > 0.0$

$\alpha = 0.05$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8515					
5	R Square	0.7250					
6	Adjusted R Square	0.7106					
7	Standard Error	40513.9954					
8	Observations	21					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	82230575305	82230575305	50.0983	0.0000	
13	Residual	19	31186292697	1641383826			
14	Total	20	113416868003				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	171205.8279	59846.1252	2.8608	0.0100	45946.4483	296465.2075
18	Store Size (Sq. Ft.)	25.3160	3.5767	7.0780	0.0000	17.8299	32.8022

The F-ratio and p-value for testing whether the regression slope = 0.0

Test Statistic $F = \dfrac{MSR}{MSE}$

$\dfrac{82230575305}{1641383826} = 50.0983$

Since F = 50.0983 > $F_{critical.0.05}$ = 4.381 (Appendix H), reject the H_0

Example 2

The data in [Table 8] are taken from a clinical trial to compare two hypotensive drugs used to lower the blood pressure during operations. The dependent variable, y, is the recovery time (in minutes) elapsing between the time at which the drug was discontinued and the time at which the systolic blood pressure had returned to 100 mmHg. The two predictors are quantity of drugs used in mg (x_1) and mean level of systolic blood pressure during hypotension in mmHg (x_2).

Table 8: Data on use of hypotensive drugs

Y	X_1	X_2
2.45	84	15
1.72	66	8
2.37	68	46
2.23	65	24
1.92	69	12
1.99	72	25
1.99	63	45
2.35	56	72

Test the following hypotheses at $\alpha=0.05$

Regression equation is predicted $Y = 58.603 + 53.688\ X_1 - 2.091\ X_2$.

The *F* statistic represents the ratio of the variance explained by the regression model (regression mean square) to the not explaine variance (residuals mean square). It can be calculated easily using an online calculator in comparison to the manual approach. The F-test o overall significance tests whether all of the predictor variables are jointly significant while the *t*-test of significance for each individua predictor variable merely tests whether each predictor variable is individually significant. Thus, the F-test determines whether or not all of th predictor variables are jointly significant. It is possible that each predictor variable is not significant and yet the F-test says that all of th predictor variables combined are jointly significant.

z-test definition

z-test is a statistical tool used for the comparison or determination of the significance of several statistical measures, particularly the mean in a sample from a normally distributed population or between two independent samples.

- Like t-tests, z tests are also based on normal probability distribution.
- Z-test is the most commonly used statistical tool in research methodology, with it being used for studies where the sample size is large (n>30).
- In the case of the z-test, the variance is usually known.
- Z-test is more convenient than t-test as the critical value at each significance level in the confidence interval is the sample for all sample sizes.
- A z-score is a number indicating how many standard deviations above or below the mean of the population is.

z-test formula

For the normal population with one sample:

$$Z = \frac{\bar{x} + \mu}{\frac{\sigma^2}{\sqrt{n}}}$$

where $\bar{x}$ is the mean of the sample, and **μ** is the assumed mean, **σ** is the standard deviation, and **n** is the number of observations.

z-test for the difference in mean:

where $\bar{x}1$ and $\bar{x}2$ are the means of two samples, **σ** is the standard deviation of the samples, and **n1** and **n2** are the numbers of observations of two samples.

$$z = \frac{\overline{X1} - \overline{X2}}{\sqrt{\sigma^2\left(\frac{1}{n1} + \frac{1}{n2}\right)}}$$

If a sample of 400 male workers has a mean height of 67.47 inches, is it reasonable to regard the sample as a sample from a large population with a mean height of 67.39 inches and a standard deviation of 1.30 inches at a 5% level of significance?

Taking the null hypothesis that the mean height of the population is equal to 67.39 inches, we can write:

$H_0 : \mu = 67.39"$

$H_a: \mu \neq 67.39"$

$\bar{x} = 67.47"$, $\sigma = 1.30"$, $n = 400$

Assuming the population to be normal, we can work out the test statistic *z* as under:

$$Z = \frac{\bar{x} + \mu}{\frac{\sigma^2}{\sqrt{n}}}$$

$$Z = \frac{67.47 + 67.39}{\frac{1.30^2}{\sqrt{400}}}$$

As H_a is two-sided in the given question, we shall be applying a two-tailed test for determining the rejection regions at a 5% level of significance which comes to as under, using normal curve area table:

$R : |z| > 1.96$

The observed value of *t* is 1.231 which is in the acceptance region since $R: |z| > 1.96$, and thus, H_0 is accepted.

Example 1

In estimating output (Y) of physiotherapist from a knowledge of his/her test score on the aptitude test (X_1) and years of experience (X_2) in a hospital, the [Table 1] summarizes the findings of the study.

X_1	X_2	Y
160	5.5	32
80	6.0	15
112	9.5	30
185	5.0	34
152	8.0	35
90	3.0	10
170	9.0	39
140	5.0	26
115	0.5	11
150	1.5	23

Test the following hypotheses at $\alpha=0.05$

Obtaining the regression equati

$$H_0: Y = b_0$$

$$H_1: Y = b_0 + b_1X_1 + b_2X_2$$

The given data are reproduced in [Table 2]. [Table 2] also shows other inputs required for obtaining the regression equation.

Table 2: Obtaining regression equation

Y	X_1	X_2	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2
32	160	5.5	5120	176	880	25600	30.25
15	80	6.0	1200	90	480	6400	36
30	112	9.5	3360	285	1064	12544	90.25
34	185	5.0	6290	170	925	34225	25
35	152	8.0	5320	280	1216	23104	64
10	90	3.0	900	30	270	8100	9
39	170	9.0	6630	351	1530	28900	81
26	140	5.0	3640	130	700	19600	25
11	115	0.5	1265	5.5	57.5	13225	0.25
23	150	1.5	3450	34.5	225	22500	2.25
255	1354	53	37175	1552	7347.5	194128	363

$$\bar{X}_1 = \frac{\sum x_1}{n} = \frac{255}{10} = 25.5, \bar{X}_2 = \frac{\sum x_2}{n} = \frac{1354}{10}$$
$$= 135.4, \bar{Y}_1 = \frac{\sum Y}{n} = \frac{53}{10} = 5.3$$

The general form of multiple equation applicable in this case is:

$Y = b_0 + b_1X_1 + b_2X_2$

Moreover, the required normal equations to find the values of b_0, b_1, and b_2 can be written as under:

$$\sum Y = n\,b_0 + b_1 \sum X_1 + b_2 \sum X_2 \qquad (1)$$

$$\sum X_1 Y = b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 \qquad (2)$$

$$\sum Y X_2 = b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 \qquad (3)$$

Accordingly, the three equations are:

$255 = 10\ b_0 + 1354\ b_1 + 53\ b_2$

$37175 = 1354\ b_0 + 194128\ b_1 + 7374.5\ b_2$

$1552 = 53\ b_0 + 7347.5\ b_1 + 363\ b_2$

Solving the three equations simultaneously, we obtain $b_0 = -13.824567$, $b_1 = 0.212167$, and $b_2 = 1.999461$. Thus, the regression equation of Y on X_1 and X_2 is: $Y_C = -13.824567 + 0.212167\ X_2 + 1.999461\ X_2$.

Table 3: Calculation of total, explained, and unexplained variation

Y	X_1	X_2	Y_c	$(Y-\bar{Y})^2$	$Y-Y_c$	$(Y-Y_c)^2$	$(Y_c-\bar{Y})^2$	Std residual
32	160	5.5	31.119	42.25	0.881	0.776	31.575	0.780
15	80	6.0	15.146	110.25	−0.146	0.022	107.214	−0129
30	112	9.5	28.933	20.25	1.067	1.138	11.786	0.945
34	185	5.0	35.424	72.25	−1.424	2.027	98.479	−1.260
35	152	8.0	34.421	90.25	0.579	0.336	79.576	0.513
10	90	3.0	11.269	240.25	−1.269	1.610	202.526	−1.123
39	170	9.0	40.239	182.25	−1.239	1.536	217.238	−1.097
26	140	5.0	25.876	0.25	0.123	0.0153	0.141	0.110
11	115	0.5	11.574	210.25	−0.574	0.330	193.922	−0.509
23	150	1.5	21.000	6.25	2.000	4.001	20.253	1.771
255	1354	53		974.5		11.791	962.710	

Y_c: Predicted Y, $Y-Y_c$: Residual

Total variation (sum of squares total, SST) $\sum(Y-\bar{Y})^2$ =974.5.

Explained variation (sum of square regression, SSR) $\sum(Y_c-\bar{y})^2$ =962 710

Unexplained variation (sum of squares error, SSE) $\sum(Y-\overline{Y_c})^2$ =11.791

R square (R^2) $= SSR/SST = 962.710/974.5$=0.988, R= 0.984

Mean square regression (MS_R) $= SSR/df = 962.710/2$ =481.355

Mean square error (MS_E) $SSE/df = 11.791/7$ =1.684

F = $MS_R/MS_E = 481.355/1.684$ =285 775

The F test value corresponding with degree of freedom n1=2 and n2=7 is 4.74.since 285.775 we ignore the null hypothesis and conclude that Y = b0 or Y= b0+b1X1+b2X2

Goodness of fit

The *F* table value [Table 4] corresponding with degree of freedom $n_1 = 2$ and $n_2 = 7$ is 4.74. Since 285.775>4.74, we ignore the null hypothesis and conclude that $Y = b_0$ or $Y = b_0 + b_1X_1 + b_2X_2$.

If there is a relationship between the dependent variable and the set of independent variables, then the MSR provides an overestimate of the variance of the errors.

$$\text{SSR (SSM)} = \sum (\hat{y} - \bar{y})^2$$

$$\text{MSR} = \frac{SSR}{k}$$

The MSE always provides an unbiased estimate of the variance of errors, regardless of whether or not there is a relationship between the dependent variable and the set of independent variables.

$$\text{SSE} = \sum (y - \hat{y})^2$$

$$\text{MSE} = \frac{SSE}{N-k-1}$$

Steps to Conduct a Hypothesis Test on the Overall Regression Model

Write down the null hypothesis that there is no relationship between the dependent variable and the set of independent variables:

$$H0:\beta1=\beta2=\cdots=\beta k=0$$

Write down the alternative hypotheses that there is a relationship between the dependent variable and the set of independent variables:

Ha: at least one βi is not 0:at least one is not 0

Collect the sample information for the test and identify the significance level alpha .

The *p*-value is the area in the right tail of the F-distribution. The F-score and degrees of freedom are $F=\frac{MSR}{MSE}$ $df_1 = k$ $df_2 = n-k-1$

smoke coming out of a house. There are two possibilities. Either the smoke is due to some sort of food getting cooking OR alternatively, the house is on fire. Let's state the null hypothesis, H0, that the house is not on the fire and the smoke is mainly due to some food getting cooked. Thus, the alternate hypothesis, Ha, will be that the house is on fire.

A person passing by the house thought that the house is actually burning with fire and thus called the firefighters. However, firefighters after arriving at the spot found that the smoke was actually due to the food being cooked.

Limitations of Partial Correlation Analysis

1. The usefulness of the partial analysis is somewhat limited by the following basic assumptions of the method:

(i) The gross or zero-order correlation must have linear regressions.

(ii) The effects of the independent variables must be additively and not jointly related.

(iii) Because the reliability of partial coefficients decreases as its order increases, the number of observations in gross correlations should be large.

2. The interpretation of the partial correlation results tends to assume that the independent variables have causal effects on dependent variable. This assumption is sometimes true, but more often untrue in varying degrees.

Limitations of Multiple Correlation Analysis

1. Multiple correlation analysis is based on the assumption that the relationship between the variables is linear. In other words, the rate of change in one variable in terms of another is assumed to be constant for all values.

2. A second important limitation is the assumption that effects of independent variables on the dependent variables are separate, distinct and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable regardless of the sizes of the other two independent variables.

3. Linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only a few, well-trained in the method are able to interpret them. However, this lack of understanding and resulting misuse are due to the complexity of the method