Multiple Regression Analysis

Partial Correlation Coefficient: It is the statistical technique of studying relationship between a dependent variable and an independent variable by keeping the effect of rest of the independent variable constant.

- **Zero order partial correlation**: A zero order partial correlation or simple correlation is the relationship between two variables but without controlling the effect of other variables on both correlating variables.
- **First order partial correlation**: A first order partial correlation is a technique that quantifies the degree of association between two variables after statistically removing the association of a third variable with both of those two correlating variables. It studies the relationship between tow variables after removing the overlap with a third variable completely from both variables.
- **Higher order partial correlation**: A higher order partial correlation studies the degree of association between two variables after controlling for the effect of two or more other variables.

First order partial correlation coefficient

Consider the study of three variables Y, X_1 and X_2, where Y being a dependent variable, X_1 and X_2 being independent variables. We may be interested to know the correlation between Y and X_1 or Y and X_2. But correlation between any two variables may be partly due to the correlation of third variable with both correlating variables. In such situation we control the effect of third variable on both correlating variables. To control the effect of third variable, we can either control the data by taking cross-section of data or use statistical control. For example, the SBP of a person may be related with age and stress level. But these variables are correlating each other. If we want to see the relationship between SBP and stress scores, we have to control for age. Using the data control, we have to see the relationship for each age level or age-group. Rather we use statistical technique to see the relationship.

$r_{Y1.2}$= Partial correlation coefficient between Y and X_1 after controlling for the effect of X_2

$$= \frac{r_{Y1} - r_{Y2} r_{12}}{\sqrt{1 - r_{Y2}^2}\sqrt{1 - r_{12}^2}}$$

$r_{Y2.1}$= Partial correlation coefficient between Y and X_2 after controlling for the effect of X_1

$$= \frac{r_{Y2} - r_{Y1} r_{12}}{\sqrt{1 - r_{Y1}^2}\sqrt{1 - r_{12}^2}}$$

$r_{12.Y}$= Partial correlation coefficient between X_1 and X_2 after eliminating the effect of Y

$$= \frac{r_{12} - r_{Y1} r_{Y2}}{\sqrt{1 - r_{Y1}^2}\sqrt{1 - r_{Y2}^2}}$$

where,

r_{Y1} = Simple correlation coefficient between Y and X_1

r_{Y2} = Simple correlation coefficient between Y and X_2

r_{12} = Simple correlation coefficient between X_1 and X_2

In general,

$$r_{AB.C} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{1 - r_{AC}^2}\sqrt{1 - r_{BC}^2}}$$

The partial correlation coefficient ranges from – 1 to + 1, so the interpretation of these coefficients is same as that of simple correlation coefficient.

$$-1 \leq r_{Y1.2} \leq +1, -1 \leq r_{Y2.1} \leq +1 \text{ and } -1 \leq r_{12.Y} \leq +1$$

Assumptions of partial correlation coefficient

The major assumptions of partial correlation coefficients are:

- **Assumption #1:** You have **one (dependent) variable** and **one (independent) variable** and these are both measured on a **continuous** scale (i.e., they are measured on an **interval** or **ratio** scale).
  Examples of **continuous variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100),
- **Assumption #2:** You have **one or more control variables**, also known as **covariates** (i.e., control variables are just variables that you are using to adjust the relationship between the other two variables; that is, your dependent and independent variables). These control variables are also measured on a **continuous** scale (i.e., they are **continuous variables**).
- **Assumption #3:** There needs to be a **linear relationship** between all three variables. That is, all possible pairs of variables must show a linear relationship. This is often accomplished by visually inspecting a scatterplot.
- **Assumption #4:** There should be **no significant outliers**. Outliers are simply single data points within your data that do not follow the usual pattern. Partial correlation is sensitive to outliers, which can have a very large effect on the line of best fit and the correlation coefficient, leading to incorrect conclusions regarding your data. Therefore, it is best if there are no outliers or they are kept to a minimum.
- **Assumption #5:** Your variables should be **approximately normally distributed**. In order to assess the statistical significance of the partial correlation, you need to have bivariate normality for each pair of variables, but this assumption is difficult to assess, so a simpler method is more commonly used whereby the distribution for each variable individually is tested.

Example 1

The salaries of workers are expected to be dependent on the number of years they have spent in college (school) and their work experience. The following table gives information on the annual salaries (in thousands of dollars) for 12 persons, the number of years each of them spent in school, and the total number of years of experiences.

S.N.	1	2	3	4	5	6	7	8	9	10	11	12

Salary (‘000 $)	52	44	48	77	68	48	59	83	28	61	27	69
Years of schooling	16	12	13	20	18	16	14	18	12	16	12	16
Years of experience	6	10	15	8	11	2	12	4	6	9	2	18

Example 2:

There is at present a debate among educators and policy makers about the use of aptitude and achievement tests as part of college admissions. Some say aptitude tests should be used because they are minimally influenced by formal education. Thus, they tend to level the playing field and account for differences among schools in grade inflation. Other say that achievement tests should be used because they show what people actually know or can do, and they would provide motivation for students to progress beyond basics. There are many complicated arguments that have some merit on both sides. Let's set all that to one side for a moment and think about the utility of such measures for a moment. Suppose what we want to do is to make good admissions decisions in the sense that we want to maximize our prediction of achievement in college from what we know from the end of high school in the area of mathematics.

Suppose admit people to college without looking at the data, which are test scores for people on the SAT-Q (quantitative or math aptitude), and scores on a math CLEP test (math achievement) and we look at grades in the standard first year math sequence (differential and integral calculus). We want to know about the prediction of math grades from the two tests.

Our data might look like this:

Person	SAT-Q	CLEP	Math GPA
1	500	30	2.8
2	550	32	3.0
3	450	28	2.9
4	400	25	2.8
5	600	32	3.3
6	650	38	3.3
7	700	39	3.5
8	550	38	3.7
9	650	35	3.4
10	550	31	2.9

Example: In a study, three variables were studied which include, Y = SBP (mm Hg), X_1 = Age (years) and X_2 = Cholesterol Concentration (mg/100ml) by taking a random of 142 older women. Following simple correlation coefficient were obtained.

r_{Y1}= 0.3332, r_{Y2}= 0.2495 and r_{12}= 0.5029

We expect there is high degree of correlation between SBP and cholesterol level. But it is evident that both SBP and cholesterol concentration increase with age. The questions are: Aare they related because of their common association with age? or Is there a relation at every age? It might be interested to know $r_{Y2.1}$

$$r_{Y2.1}=\frac{r_{Y2}-r_{Y1}r_{12}}{\sqrt{1-r_{Y1}^2}\sqrt{1-r_{12}^2}}=\frac{0.2495-(0.3332)(0.5029)}{\sqrt{1-(0.3332)^2}\sqrt{1-(0.5029)^2}}=0.1005$$

Multiple linear regression model

In multiple linear regression model, we assume that a linear relationship exists between dependent variable Y and p independent variables, X_1, X_2, …, X_p. The independent variables are also called as explanatory variables or predictor or regressors.The main advantage of multiple regression is that it allows us to use more information available to us to estimate the values of dependent variables. We can use several explanatory variables to predict the value of a dependent variable.

The multiple linear regression model with p explanatory variables is given by,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e$$

where,

Y = Dependent variable

X_j = Independent variable, j =1, 2, …, p

β_j = regression coefficients. j = 1, 2, …, p

The regression coefficient $\beta_j$represent the slope of Y with Xj holding remaining variables constant. It represents the average change or expected change in Y for per unit change in Xj keeping all remaining regressors variables constant. $\beta_j$are also called partial regression coefficients. The coefficient β_0 represent the Y-intercept. It measures the average value of Y when all regressors have zero values.

Assuming E (e) = 0, the response function for regression model is,

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Multiple linear regression with two independent variables

The model is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

where,

β_0= Population Y- intercept.

β_1= Population slope of Y on X_1 holding X_2 constant.

β_2= Population slope of Y on X_2 holding X_1 constant.

The sample linear multiple regression equation with two independent variables is,

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2}$$

where,

$\hat{Y}_i$ = estimated value of Y for observation i (i = 1, 2, …, n)

X_{i1} = Value of independent variable X_1 for observation i

X_{i2} = Value of independent variable X_2 for observation i

b_0 = Sample Y-intercept.

b_1 = Partial regression coefficient of Y on X_1 or sample slope of Y with X_1 holding X_2 constant.

b_2= Partial regression coefficient of Y on X_2 or sample slope of Y with X_2 holding X_1 constant.

Interpretation of regression coefficients

Sample Y intercept b_0

It measures estimated average value of Y when both independent variable takes zero values. It provides the unbiased estimate of population Y-intercept β_0 i.e. $E(b_0) = \beta_0$.

Sample slope b_1

It measures the expected or average change in Y for per unit change in X_1 holding X_2 constant. It provides the unbiased estimate for population regression coefficient β_1i.e. $E(b_1) = \beta_1$.

Sample slope b_2

It provides the unbiased estimate for population regression coefficient β_2. It measures the expected or average change in Y for per unit change in X_2 holding X_1 constant i.e., $E(b_2) = \beta_2$

Estimating regression coefficients

The values of regression coefficients β_0, β_1 and β_2 are generally not known, and they are to be estimated from the sample data. To develop a predicting equation for given sample set, we have to find the estimated value of these parameters.

The two common methods of finding the coefficients are:

1. Ordinary Least Square Method (OLS method)
2. Maximum likelihood Method (MLE method)

Ordinary least square method

The principle behind the least square method is to find the optimum value of b_0, b_1 and $\widehat{\ }_2$ so that sum of square of residuals or errors is minimum. We can't minimize the sum of residuals because $\sum_{i=1}^{n} e_i$= 0.

Let's define the least square function S as follows,

S = Sums of square due to error

= $\sum_{i=1}^{n} e_i^2$ =$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

S = $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2})^2$

The function S must be minimized with respect to b_0, b_1 and b_2, so that overall squared error is minimum.

The least square estimator of b_0 must satisfies following,

$$\frac{\partial S}{\partial b_0} = 0$$

$$\frac{\partial}{\partial b_0}\left\{\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2})^2\right\} = 0$$

$$\frac{\partial}{\partial b_0}\{(y_1 - b_0 - b_1X_{11}) + \cdots + (y_n - b_0 - b_1X_{n1})\} = 0$$

…

$$\Rightarrow -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = nb_0 + b_1\sum_{i=1}^{n} X_{i1} + b_2\sum_{i=1}^{n} X_{i2}$$

Similarly,

$$\frac{\partial S}{\partial b_1} = 0$$

$$\frac{\partial}{\partial b_1}\left\{\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2})^2\right\} = 0$$

…

$$\Rightarrow -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2})\ X_{i1} = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i\,X_{i1} = b_0\sum_{i=1}^{n} X_{i1} + b_1\sum_{i=1}^{n} X_{i1}^2 + b_2\sum_{i=1}^{n} X_{i1}X_{i2}$$

And,

$$\frac{\partial S}{\partial b_2} = 0$$

$$\frac{\partial}{\partial b_2}\left\{\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2})^2\right\} = 0$$

…

$$\Rightarrow -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i2})\ X_{i2} = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i\,X_{i2} = b_0\sum_{i=1}^{n} X_{i2} + b_1\sum_{i=1}^{n} X_{i1}X_{i2} + b_2\sum_{i=1}^{n} X_{i1}^2$$

Hence, the normal equations for estimating regression coefficients are:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1\sum_{i=1}^{n} X_{i1} + b_2\sum_{i=1}^{n} X_{i2}$$

$$\sum_{i=1}^{n} Y_i\,X_{i1} = b_0\sum_{i=1}^{n} X_{i1} + b_1\sum_{i=1}^{n} X_{i1}^2 + b_2\sum_{i=1}^{n} X_{i1}X_{i2}$$

$$\sum_{i=1}^{n} Y_i\,X_{i2} = b_0\sum_{i=1}^{n} X_{i2} + b_1\sum_{i=1}^{n} X_{i1}X_{i2} + b_2\sum_{i=1}^{n} X_{i2}^2$$

Matrix form of normal equations.

$$\begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i X_i \\ \sum_{i=1}^{n} Y_i X_i \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} \\ \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i2}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

or, Y = X . $\hat{\beta}$

Regression coefficients can be calculated directly by following formulas

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

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

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

Where,

$$x_1 = X_1 - \overline{X}_1, x_2 = X_2 - \overline{X}_2 \text{ and } y = Y - \overline{Y}$$

Estimation of regression equations using simple correlation coefficients and standard deviations

Let W = $\begin{bmatrix} 1 & r_{Y1} & r_{Y2} \\ r_{Y1} & 1 & r_{12} \\ r_{Y2} & r_{Y1} & 1 \end{bmatrix}$

$$b_1 = \frac{s_Y}{s_1}\frac{\begin{bmatrix} r_{Y1} & r_{12} \\ r_{Y2} & 1 \end{bmatrix}}{\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}} \qquad b_2 = \frac{s_Y}{s_2}\frac{\begin{bmatrix} 1 & r_{Y1} \\ r_{12} & r_{Y2} \end{bmatrix}}{\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}}$$

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

The regression coefficient of Y on X_1 and X_2 can also be given by,

$$W_{11}\frac{\hat{Y} - \overline{Y}}{s_Y} + W_{12}\frac{X_1 - \overline{X}_1}{s_1} + W_{13}\frac{X_2 - \overline{X}_2}{s_2} = 0$$

Where,

$$W_{11} = \begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix} \qquad W_{12} = -\begin{vmatrix} r_{Y1} & r_{12} \\ r_{Y2} & 1 \end{vmatrix} \qquad W_{13} = \begin{vmatrix} r_{Y1} & 1 \\ r_{Y2} & r_{12} \end{vmatrix}$$

Matrix approach to solve the normal equations

It is simpler to solve the normal equations if they are expressed in matrix notation.

Let Y = $\begin{bmatrix} Y1 \\ Y2 \\ \dots \\ Yn \end{bmatrix}_{n \text{ X } 1}$ X = $\begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} \end{bmatrix}_{n \text{ X } 3}$ $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3 \text{ X } 1}$ $e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}_{n \text{ X } 1}$

The least square estimator of β is given by,

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Evaluation of Multiple Linear Regression

1. **Multiple correlation coefficient**: The closeness of association between observed value of dependent variable Y and its estimated value $\hat{Y}$ as given by the regression equation of Y on X_1 and X_2 i.e. $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$is called multiple correlation coefficient. It is denoted by $R_{Y.12}$. It is actually Pearson's correlation coefficient between Y and $\hat{Y}$.

$$R_{Y.12} = \frac{Cov\,(Y\,,\hat{Y})}{\sqrt{Var(Y)}\sqrt{Var(\hat{Y})}} = \frac{n\sum Y.\hat{Y} - \sum Y \sum \hat{Y}}{\sqrt{n\sum Y^2 - (\sum Y)^2}\sqrt{n\sum \hat{Y}^2 - (\sum \hat{Y})^2}}$$

Alternatively,

$$R_{Y.12} = \sqrt{\frac{r_{Y1}^2 + r_{Y2}^2 - 2\, r_{Y1}\, r_{Y2}\, r_{12}}{1 - r_{12}^2}}$$

Since, $R_{Y.12}$is a correlation coefficient, we expect it to lie between -1 and +1, but in practice it is not negative. So, $R_{Y.12}$lies between 0 and 1. The multiple correlation coefficient $R_{Y.12}$is a good index for the goodness of fit of the multiple linear equation. If $R_{Y.12}$= 1, then the predicting equation is 100 % reliable which means all the regression residuals are 0 and σ^2= 0. If $R_{Y.12}$= 0, then Y is completely uncorrelated with independent variables. Closer the value of $R_{Y.12}$towards 1, better is the fit, and lesser is the value of $R_{Y.12}$towards 0, poorer is the fit and suggest independence of variables.

2. **Coefficient of multiple determination:** The one-way to evaluate reliability of the fitted regression equation is to compare the scatterness of observations on response variable Y from the fitted value with scatterness of points about its mean value.

Here,

SST = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ = Sum of squares due to total

SST measures total variation of Y values around their mean $\overline{Y}$ and it can be split into two components (i) Regression SS and (ii) Residual SS

SSR = $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$= Sum of squares due to regression

SSR measures variation among regression estimates. It measures the portion of variation in Y due to regression. It is also called explained variation.

SSE = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ = Sum of squares due to error

SSE measures the variation of the observed values Y about the regression line. It represents the portion of total variation in Y unexplained by regression equation.

SST = SSR + SSE

The ratio of SSR to SST is called coefficient of multiple determination and is given by,

$$R_{Y.12}^2 = \frac{SSR}{SST}$$

Equivalently,

$$R_{Y.12}^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

If sample data fits well then greater portion of SST is due to SSR. Hence the ratio measures the goodness of fit of the regression equation. The value of $R_{Y.12}^2$lies between 0 and 1. If $R_{Y.12}^2$= 1, then

the regression equation is 100 % reliable (perfect fit) and becomes a deterministic model. If $R^2_{Y.12}$= 0, then there is no relationship between variables. If value of $R^2_{Y.12}$approaches 1, then it indicates a good fit and if it approaches 0, then it indicates poor fit.

$R^2_{Y.12}$indicates the proportion of variation in Y values due to linear relationship between Y, X_1 and X_2 i.e. explained by the sample linear regression equation $\hat{Y}_i = b_0 + b_1X_{i1} + b_2X_{i2}$.

Alternatively, we can calculate $R^2_{Y.12}$by following formula.

$$R^2_{Y.12} = \frac{r^2_{Y1}+r^2_{Y2}-2\, r_{Y1}\, r_{Y2}\, r_{12}}{1-r^2_{12}}$$

Adjusted Coefficient of Multiple Determination

The adjusted coefficient of multiple determination $R^2_{Y.12}$ (adj.) is given by,

$$R^2_{Y.12}(adj) = 1 - \frac{MSE}{MST}$$

$$= 1 - \frac{\frac{SSE}{n-p-1}}{SST/n-1}$$

$$= 1 - \frac{n-1}{n-p-1}\frac{SSE}{SST}$$

$$= 1 - \left[(1 - R^2_{Y.12})\frac{n-1}{n-p-1}\right]$$

Unadjusted will increase as we add more and more explanatory variables to the model, this would artificially "improve" the model. Unadjusted R^2 may show high figure for a model even if it doesn't fit well.

The unadjusted R^2 gives the percentage of explained variation by the regression equation as if all independent variables in the model affect the dependent variable Y, where as adjusted R^2 gives the percentage variation explained by only those independent variables that in reality affect the dependent variable R^2.

Main differences between unadjusted R^2 and adjusted R^2

- Unadjusted R^2 shows the proportion of variation in Y explained by independent variables X_1 and X_2 in a linear model, irrespective of how well they are correlated to dependent variable Y. The adjusted R^2 provides an adjustment to the R^2 statistic such that an independent variable that has a strong correlation to Y increases the adjusted R^2 and any variable without a strong correlation will make adjusted R^2 decreases.
- Unadjusted R^2 assumes that every independent variable in the model explains the variation in Y. So, it assumes every predictor added to a model increases R^2 and never decrease it. Thus, a model with more terms may seem to have better fit just because it has more predictors. Adjusted R^2 compensates for the addition of variables and only increases the value of statistic if new independent variable enhances the model.

Standard error of estimation

The unbiased estimator of variance of errorσ^2 is given by

$\hat{\sigma}^2 = \frac{SSE}{n-p-1}$ = MSE = Est. mean sum of squares due to error

The standard error of estimate is the square root of MSE. The mean squared error (MSE) is basically measures the average squared difference between the actual values (y) and the predicted values (ŷ) in a regression model.

$$MSE = \frac{SSE}{n-p-1}$$

$$S.E. or\ S_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}}$$

MSE is closely related to the variance of the error distribution (σ^2). In fact, MSE is an unbiased estimator of the error variance (σ^2) if n is large. The relationship between MSE and error variance relies on assumption that errors are normally distributed random variable.

For two independent variables, the standard error of estimate is,

$$S.E. = \sqrt{\frac{SSE}{n-3}}$$

MSE (or SSE) is a crucial metric in regression analysis. It provides insights into the model's predictive accuracy and helps assess how well it fits the data. MSE is often used to compare the performance of different regression models. A model with a lower MSE is generally preferred. A lower MSE indicates that the model's predictions are closer to the actual values, suggesting a better fit to the data.

Overall or global test of model accuracy (F-test for whole model)

It is important to test if there is a linear relationship between response variable Y and a set of p independent variables namely X_1, X_2, …, X_p . The test hypothesis is,

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1: $\beta_j \neq 0$ for one or more j = 1, 2, …, p

Rejection of this null hypothesis implies that at least one of the independent variables contribute significantly to the model.

Test Statistic

The test statistic is,

$$F = \frac{MSR}{MSE} = \frac{Variance\ due\ to\ regression}{Variance\ due\ to\ error}$$

ANOVA Table

Source of variation	Degrees of freedom (D. f.)	Sums of Squares (SS)	Mean Sum of Square (MS)	Cal F
Regression	p	SSR = $\sum(\hat{Y} - \overline{Y})^2$	MSR = $\frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	n – p – 1	SSE = $\sum(Y - \hat{Y})^2$	MSE = $\frac{SSE}{n-p-1}$	
Total	n - 1	SST	----	

SSR = Regression sum of squares

= $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$

SST = Total sum of squares

$= \sum_{i=1}^{n}(Y_i - \overline{Y})^2$

SSE = SST – SSR

Decision Rule: If calculate F ≥ $F_\alpha\{p, n-p-1\}$ we reject H_0 at α level of significance. It means that at least one independent variable has significant linear relationship with dependent variable Y. To know which of the p independent variables have significant relationship with Y, we have to perform t-test for individual slope.

Inference on individual slope

If F-test of whole model favours alternative hypothesis i.e. at least one of the independent variables is important, then we have to test the significance of every individual regression coefficient for their significance. To know that we have to test following hypothesis,

H_0: $\beta_j = 0$ (There is no linear relationship between Y and X_j)

H_1: $\beta_j \neq 0$ (There is linear relationship between Y and X_j)

j = 1, 2, …., p

Test for the Y-intercept $\boldsymbol{\beta_0}$

H_0: $\beta_0 = 0$ (Y-intercept is insignificant)

H_1: $\beta_0 \neq 0$ (Y-intercept is significant)

Test statistic in this case is,

$$t = \frac{b_0}{S_{b_0}}$$

The standard error of b_0 is given by the following formula

S.E.($\boldsymbol{b_0}$) = $S_{b_0} = \frac{\sqrt{MSE}}{\sqrt{SSY}}$

Decision: Reject H_0 if $\lceil cal\ t \rceil > critical\ t\ i.e. t_{\alpha/2}(n-2)$

Test for the significance of the slope $\boldsymbol{\beta_1}$

H_0: $\beta_1 = 0$ (There is no linear relationship between Y and X_1 i.e. slope β_1 is insignificant

H_1: $\beta_1 \neq 0$ (There is linear relationship between Y and X_1 i.e. slope β_1 is significant

Test statistic in this case is,

$$t = \frac{b_1}{S_{b_1}}$$

The standard error of $\boldsymbol{b_1}$ is given by following formula

S.E.($\boldsymbol{b_1}$) = $S_{b_1} = \frac{\sqrt{MSE}}{\sqrt{SSX_1}}$

Decision: Reject H_0 if $\lceil cal\ t \rceil > critical\ t\ i.e. t_{\alpha/2}(n-2)$

Test for the significance of the slope $\boldsymbol{\beta_2}$

H_0: $\beta_2 = 0$ (There is no linear relationship between Y and X_2 i.e. slope β_2 is insignificant

H_1: $\beta_2 \neq 0$ (There is linear relationship between Y and X_2 i.e. slope β_2 is significant

Test statistic in this case is,

$$t = \frac{b_2}{S_{b_2}}$$

Standard error of b_2 is given by,

S.E.($\boldsymbol{b_2}$) = $S_{b_2} = \frac{\sqrt{MSE}}{\sqrt{SSX_2}}$

Numerical Example:

The assessed value (in lakh rupees), the size of the house (thousands square feet) and the age of houses (in years) is given in following table.

House No.	Market Price (X Rs 1,00,000)	Size of the house (X 1000 sq. ft.)	Age of house (years)
1	63.0	1.605	35
2	65.1	2.489	45
3	69.9	1.553	20
4	76.8	2.404	32
5	73.9	1.884	25
6	77.9	1.558	14
7	74.9	1.748	8
8	78.0	3.105	10
9	79.0	1.682	28
10	83.4	2.470	30
11	79.5	1.820	2
12	83.9	2.143	6
13	79.7	2.121	14
14	84.5	2.485	9
15	96.0	2.300	19
16	109.5	2.714	4
17	102.5	2.463	5
18	121.0	3.076	7
19	104.9	3.048	3
20	128.0	3.267	6
21	129.0	3.069	10
22	117.9	4.765	11
23	140.0	4.540	8

(a) Fit the multiple regression equation of market price of house on size of the house and age of the house.
(b) Interpret the meaning of the slopes
(c) Predict the market price of house that has a size of 1500 square feet and is 5 years old
(d) Compute the coefficient of multiple determination and interpret the meaning of value
(e) Determine the adjusted R2
(f) Determine whether there is a significant relationship between market price of house and the two independent variables (size and age) at the 0.05 level of significance.

(g) At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model.

Solution:

(a) **Fitting of the regression**

Here,

Y = Price of the house (in lakh rupees)

X1 = Total number of square feet

X2 = Age of the house

The multiple linear equation of Y on X_1 and X_2 is given by,

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} \quad \text{... (i)}$$

The least square equations for estimating b_0, b_1 and b_2 is given below:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_{i1} + b_2 \sum_{i=1}^{n} X_{i2} \quad \text{... (ii)}$$

$$\sum_{i=1}^{n} Y_i X_{i1} = b_0 \sum_{i=1}^{n} X_{i1} + b_1 \sum_{i=1}^{n} X_{i1}^2 + b_2 \sum_{i=1}^{n} X_{i1} X_{i2} \quad \text{... (iii)}$$

$$\sum_{i=1}^{n} Y_i X_{i2} = b_0 \sum_{i=1}^{n} X_{i2} + b_1 \sum_{i=1}^{n} X_{i1} X_{i2} + b_2 \sum_{i=1}^{n} X_{i2}^2 \quad \text{... (iv)}$$

Computational Result:

$\sum Y = 2118.3$, $\sum X_1 = 58.309$, $\sum X_2 = 351$

$\sum X_1^2 = 164.0274$, $\sum X_2^2 = 8441$

$\sum YX_1 = 5706.2739$, $\sum YX_2 = 28971.2$, $\sum X_1X_2 = 816.441$

$SSY = 11050.74$, $SSX_1 = 16.2039$, $SSX_2 = 3084.4348$

SSR = 8189.72301, SSE = 2861.01699, SST = 11050.74

Substituting the values in equation (ii), (iii) and (iv) we get

$2118.3 = 23\ b_0 + 58.309\ b_1 + 351\ b_2$

$5706.2739 = 58.309\ b_0 + 164.0274\ b_1 + 816.441\ b_2$

$28971.2 = 351\ b_0 + 816.441\ b_1 + 8441\ b_2$

On solving these equations we get,

b0 = 57.3507

b1 = 17.7180

b2 = -0.663

The fitted equation of house price on house size and house age is given by,

$$\hat{Y} = 57.3507 + 17.7180\ X_1 - 0.663\ X_2$$

or $Market\ Price = 57.3507 + 17.7180\ x\ House\ Size - 0.663\ x\ House\ Age$

(b) **Meaning of regression coefficient**

b0 = 57.3507

The base price for a house with zero size and age is 57.35 Lakhs

b1 = 17.7180

For every 1000 sq. ft. increase, the market price increases by 17.72 Lakhs, holding age constant.

b2 = - 0.663

For every year increase in age, the market price decreases by 0.67 Lakhs, holding size constant.

(c) **Prediction for market price of house having size of 1500 sq. ft. and is 5 years old**

Put X1 = 1.5 (size is measured in per thousand square feet) and X2 = 5 in the fitted equation.

$$\hat{Y} = 57.3507 + 17.7180\, X_1 - 0.663\, X_2$$

= 57.3507 + 17.7180 x 1.5 – 0.6663 x 5

= Rs 80.60 lakhs

(d) Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{8189.72301}{11050.74} = 0.7411$$

Interpretation: 74.1% of the variability in the market price is explained by the size and age of the house. Remaining 25.9 % of the variability in the market price is explained by the factors other than house size and age.

(e) Adjusted R^2

Adjusted $R^2 = 1 - \left[(1 - R^2_{Y.12})\frac{n-1}{n-p-1}\right]$

$= 1 - \left[(1 - 0.7111)\left(\frac{23-1}{23-2-1}\right)\right] = 0.7152$

(f) **F test for whole model**

Hypothesis to test

H0: there is significant linear relationship between market price of house, size and price

(H0: $\beta_1 = \beta_2 = 0$

H_1: At least one independent variable no significant relationship with dependent variable

(H1: $\beta_j \neq 0$ for one or more j = 1, 2)

Test Statistic

The test statistic is,

$$F = \frac{MSR}{MSE}$$

The test statistic has F-distribution with p degrees of freedom in the numerator and n-p-1 degrees of freedom in denominator

Calculation of F statistics

One-way ANOVA Table

	df	*SS*	*MS*	*F*
Regression	2	8189.72301	4094.8615	28.6252
Residual	20	2861.01699	143.0508	
Total	22	11050.74		

Critical F

The test is right sided and chosen level of significance is 0.05

Numerator d.f. = 2

Denominator d.f. = 20

From table of F distribution,

$$F_{0.05}(2, 20) = 3.49$$

Statistical Decision

Since Cal F (28.63) >> Critical F (3.49), we strongly reject H0 in favour of H1.

Conclusion: There is a significant relationship between the market price and at least one of the two predictors (size and age)

(g) t -test for the individual slope

Test for significance of the slope β1

H_0: $\beta_1 = 0$ (there is no significant relationship between house price and house size)

H_1: $\beta_1 \neq 0$ (there is significant relationship between house price and house size)

Test statistic

The test statistic for slope test is,

$$t = \frac{b_1}{s_{b_1}}$$

The test statistic has t-distribution with n-p-1 degrees of freedom

Calculated t

From fitted equation, b_1 = 17.7180 and standard error of b_1 i.e. s_{b_1} = 3.1456 (given in the exam)

Hence,

$$t = \frac{17.7180}{3.1456} = 5.6326$$

Critical t

The test is two-sided and chosen level of significance is 0.05

From table of t distribution

$$t_{0.025}(20) = 2.0$$

Statistical Decision

Since |Cal t| = 5.63 > Critical t (2.09), we reject H0 at 5 % level of significant

Conclusion:

The house size is the significant predictor for the market price of the house.

Test for significance of the slope β2

H_0: $\beta_2 = 0$ (there is no significant relationship between house price and house age)

H_1: $\beta_2 \neq 0$ (there is significant relationship between house price and house age)

Test statistic

The test statistic for slope test is,

$$t = \frac{b_2}{s_{b_2}}$$

The test statistic has t-distribution with n-p-1 degrees of freedom

Calculated t

From fitted equation, b_2 = -0.6663 and standard error of b_1 i.e. s_{b_1} = 0.2280 (given in the exam)

Hence,

$$t = \frac{-0.6663}{0.2280} = -2.9226$$

Critical t

The test is two-sided and chosen level of significance is 0.05

From table of t distribution

$$t_{0.025}(20) = 2.086$$

Statistical Decision

Since |Cal t| = 2.9226 > Critical t (2.09), we reject H0 at 5 % level of significant

Conclusion:

The house age is the significant predictor for the market price of the house.

The t-test indicates that both predictors (house size and house age) make significant contributions to predicting the market price.

Computer Output

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.3507	10.0072	5.7310	0.0000	36.4762	78.2253
X Variable 1	17.7180	3.1456	5.6326	0.0000	11.1564	24.2797
X Variable 2	-0.6663	0.2280	-2.9226	0.0084	-1.1419	-0.1908

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	8189.72301	4094.8615	28.6252
Residual	20	2861.01699	143.0508	
Total	22	11050.74		

Regression Statistics	
Multiple R	0.86087
R Square	0.7411
Adjusted R Square	0.7152
Standard Error	11.96039
Observations	23

Residual Output

Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	62.4660	0.5340	0.0468
2	71.4653	-6.3653	-0.5582
3	71.5399	-1.6399	-0.1438
4	78.6218	-1.8218	-0.1598
5	74.0728	-0.1728	-0.0152
6	75.6266	2.2734	0.1994
7	82.9911	-8.0911	-0.7095
8	105.7018	-27.7018	-2.4292
9	68.4947	10.5053	0.9212
10	81.1239	2.2761	0.1996
11	88.2649	-8.7649	-0.7686
12	91.3224	-7.4224	-0.6509
13	85.6018	-5.9018	-0.5175
14	95.3829	-10.8829	-0.9543
15	85.4416	10.5584	0.9259
16	102.7721	6.7279	0.5900
17	97.6585	4.8415	0.4245
18	107.1870	13.8130	1.2113
19	109.3563	-4.4563	-0.3908
20	111.2375	16.7625	1.4699
21	105.0639	23.9361	2.0990
22	134.4474	-16.5474	-1.4510
23	132.4598	7.5402	0.6612

Multiple linear regression with ordinal response variable

Suppose we want to predict job performance (in the scale of 1 to 5, 1 indicating lowest preference and 5 the highest preference) of mechanics based on mechanical aptitude test scores and test scores from personality test that measures conscientiousness.

S.N.	Job preference (Y)	Mechanical aptitude test scores (X1)	Conscientiousness (X2)
1	1	40	25
2	2	45	20
3	1	38	30
4	3	50	30
5	2	48	28
6	3	55	30
7	3	53	34
8	4	55	36
9	4	58	32
10	3	40	34
11	5	55	38
12	3	48	28
13	3	45	30
14	2	55	36
15	4	60	34
16	5	60	38
17	5	60	42
18	5	65	38
19	4	50	34
20	3	58	38

Exercise:

1. Suppose we have following information about three variables X1, X2 and X3.
   $r_{12} = 0.91$, $r_{13} = 0.33$ and $r_{23} = 0.81$

   Check whether this information is consistent or not.

2. The salaries of workers are expected to be dependent on the number of years they have spent in college (school) and their work experience. The following table gives information on the annual salaries (in thousands of dollars) for 12 persons, the number of years each of them spent in school, and the total number of years of experiences.

S.N.	1	2	3	4	5	6	7	8	9	10	11	12
Salary ('000 $)	52	44	48	77	68	48	59	83	28	61	27	69
Years of schooling	16	12	13	20	18	16	14	18	12	16	12	16
Years of experience	6	10	15	8	11	2	12	4	6	9	2	18

   (a) Find the least square equation (fit a multiple linear regression) of annual salary on years of schooling and years of experience.
   (b) Interpret the value of regression coefficient

(c) Predict the annual salary of the person who had spent 16 years in school and had 4 years' job experience.
(d) Find the proportion of variation in annual salary explained by the linear equation.
(e) At a $\alpha = 0.05$, which of the independent variables are significant explanatory variables for sales growth.
(f) Is the overall model significant as a whole?

3. A developer of food for pigs would like to determine, what relationship exists among the age of a pig when it starts receiving a newly developed food supplement (X_1), the initial weight of the pig (X_2), and the amount of weight it gains in a 1-week period with the food supplement (Y). The following information is the result of a study of eight piglets.

Piglet No.	1	2	3	4	5	6	7	8
Initial age (weeks)	8	6	7	12	9	6	7	4
Initial weight (lbs)	39	52	49	46	61	35	25	55
Weight gain (lbs)	7	6	8	10	9	5	3	4

(a) Calculate the least-squares equation that best describes these three variables
(b) How much weight might we expect to gain in a week time with the food supplement if it were 9 weeks old and weighed 48 pounds?

4. An agriculturist is interested in designing a regression model, to relate the damage susceptibility pf peaches to the height at which they are dropped (drop height, measured in mm) and the density of the peach (measured in g/cm2). Goal of model building is to provide a predictive model for peach damage to serve as a guideline for harvesting and post-harvesting operations.

Y	X1	X2
3.62	303.7	0.9
7.27	366.7	1.04
2.66	336.8	1.01
1.53	304.5	0.95
4.91	346.8	0.98
10.36	600	1.04
5.26	369	0.96
6.09	418	1
6.57	269	1.01
4.24	323	0.94
8.04	562.2	1.01
3.46	284.2	0.97
8.5	558.6	1.03
9.34	415	1.01
5.55	349.5	1.04
8.11	462.8	1.02
7.32	333.1	1.05
12.58	502.1	1.1
0.15	311.1	0.91
5.23	351.4	0.96

(a) What is the best fitting regression equation?

(b) What fraction of the variation in peach damage is explained by this regression?
(c) What amount of peach damage is expected if it is dropped from the height of 300 mm and has density of 0.95 g/cm2
(d) At 5 % level of significance determine whether the regression as a whole is significant?
(e) Is drop height significant predictor for peach damage?

5. The following statistics are obtained from 75 measurements on length in mm (X_1), volume in c.c. (X_2) and weight in gm (X_3) of 300 eggs.

Variable	Mean	Standard Deviation
Egg Length	55.95	2.26
Egg Volume	51.48	4.39
Egg Weight	56.03	4.41

Correlation Matrix

	Egg Length	Egg Volume	Egg Weight
Egg Length	-----	0.578	0.581
Egg Volume	0.578	----	0.974
Egg Weight	0.581	0.974	----

(a) Fit a multiple regression to study the dependence of egg weight on egg length and egg volume.
(b) Estimate the mean weight of an egg whose length is 57.5 mm and volume 53 c.c.
(c) Compute partial correlation coefficients
(d) Compute $R_{3.12}$

6. In a study of a random sample of 120 students the following results were obtained regarding three variables; X_1 = percentage of marks in first test, X_2 = percentage of marks in second test and Y = percentage of marks in third test.

Variable	Mean	Variance
Percentage in first test	68	100
Percentage in second test	70	25
Percentage in third test	74	81

Correlation Matrix

	% First exam	% Second exam	% Third exam
% First exam	-----	0.60	0.70
% Second exam	0.60	----	0.65
% Third exam	0.70	0.65	----

(a) Obtain the least square equation of Y on X1 and X2
(b) Estimate the percentage marks of a student in the final examination if he got 60 % in first examination and 67 % in second examination.
(c) Compute partial correlation coefficients
(d) Compute coefficient of multiple determination for the fitted equation