Chapter 5

Regression Analysis

"Regression analysis is the hydrogen bomb of the statistics arsenal."
— Charles Wheelan, Naked Statistics: Stripping the Dread from the Data

Regression analysis is a supervised learning method for predicting continuous variables. The difference between classification and regression analysis is that regression methods are used to predict qualitative variables or continuous numbers unlike categorical variables or labels. It is used to predict linear or non-linear relationships among variables of the given dataset. This chapter deals with an introduction of regression and its various types.

Learning Objectives

- Understand the basics of regression analysis
- Introduce concepts of correlation and causation
- Learn about linear regression and its validation techniques
- Discuss about multiple linear regression
- Introduce logistic regression
- Study about the concept of regularization
- Study popular regression methods like Ridge, Lasso, and Elastic Net

5.1 INTRODUCTION TO REGRESSION

Regression analysis is the premier method of supervised learning. This is one of the most popular and oldest supervised learning technique. Given a training dataset D containing N training points (x_i, y_i), where $i = 1...N$, regression analysis is used to model the relationship between one or more independent variables x_i and a dependent variable y_i. The relationship between the dependent and independent variables can be represented as a function as follows:

$$y = f(x) \quad (5.1)$$

Here, the feature variable x is also known as an explanatory variable, exploratory variable, a predictor variable, an independent variable, a covariate, or a domain point. y is a dependent variable. Dependent variables are also called as labels, target variables, or response variables.

Regression analysis determines the change in response variables when one exploration variable is varied while keeping all other parameters constant. This is used to determine the relationship each of the exploratory variables exhibits. Thus, regression analysis is used for prediction and forecasting

Regression is used to predict continuous variables or quantitative variables such as price and revenue. Thus, the primary concern of regression analysis is to find answer to questions such as:

1. What is the relationship between the variables?
2. What is the strength of the relationships?
3. What is the nature of the relationship such as linear or non-linear?
4. What is the relevance of the attributes?
5. What is the contribution of each attribute?

There are many applications of regression analysis. Some of the applications of regressions include predicting:

1. Sales of a goods or services
2. Value of bonds in portfolio management
3. Premium on insurance companies
4. Yield of crops in agriculture
5. Prices of real estate

5.2 INTRODUCTION TO LINEARITY, CORRELATION, AND CAUSATION

The quality of the regression analysis is determined by the factors such as correlation and causation.

Regression and Correlation

Correlation among two variables can be done effectively using a Scatter plot, which is a plot between explanatory variables and response variables. It is a 2D graph showing the relationship between two variables. The x-axis of the scatter plot is independent, or input or predictor variables and y-axis of the scatter plot is output or dependent or predicted variables. The scatter plot is useful in exploring data. Some of the scatter plots are shown in Figure 5.1. The Pearson correlation coefficient is the most common test for determining correlation if there is an association between two variables. The correlation coefficient is denoted by r. Correlation is discussed in Chapter 2 of this book. The positive, negative, and random correlations are given in Figure 5.1. In positive correlation, one variable change is associated with the change in another variable. In negative correlation, the relationship between the variables is reciprocal while in random correlation, no relationship exists between variables.

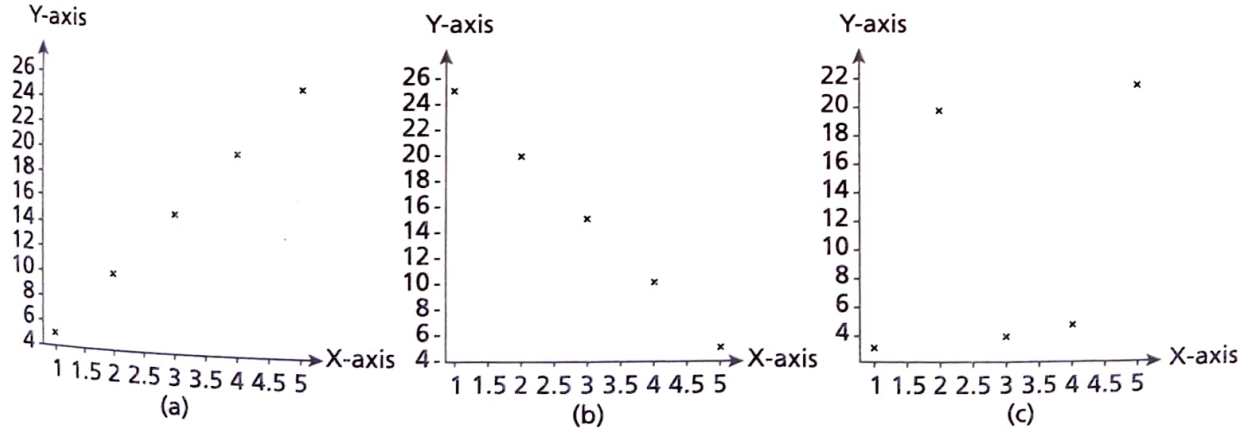

Figure 5.1: Examples of (a) Positive Correlation (b) Negative Correlation (c) Random Points with No Correlation

While correlation is about relationships among variables, say x and y, regression is abo predicting one variable given another variable.

Regression and Causation

Causation is about causal relationship among variables, say x and y. Causation means knowi whether x causes y to happen or vice versa. x causes y is often denoted as x implies y. Correlati and Regression relationships are not same as causation relationship. For example, the correlati between economical background and marks scored does not imply that economic backgrou causes high marks. Similarly, the relationship between higher sales of cool drinks due to a ri in temperature is not a causal relation. Even though high temperature is the cause of cool drin sales, it depends on other factors too.

Linearity and Non-linearity Relationships

The linearity relationship between the variables means the relationship between the depende and independent variables can be visualized as a straight line. The line of the form, $y = ax +$ can be fitted to the data points that indicate the relationship between x and y. By linearity, it meant that as one variable increases, the corresponding variable also increases in a linear manne A linear relationship is shown in Figure 5.2 (a). A non-linear relationship exists in functions such exponential function and power function and it is shown in Figures 5.2 (b) and 5.2 (c). Here, x-ax is given by x data and y-axis is given by y data.

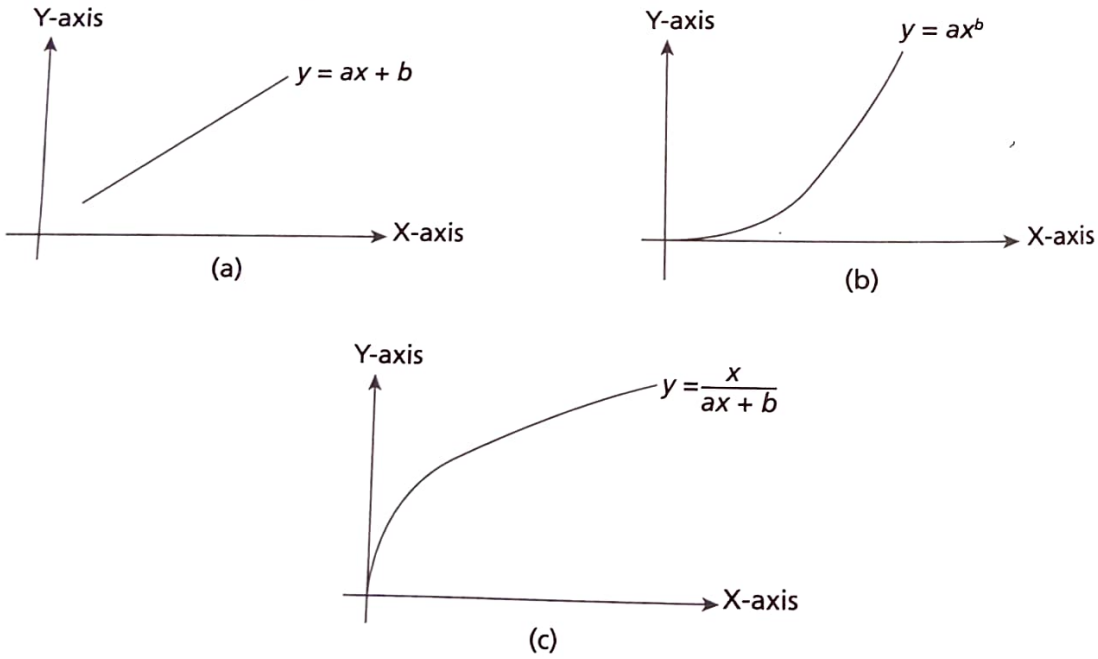

Figure 5.2: (a) Example of Linear Relationship of the Form $y = ax + b$ (b) Example of a Non-linear Relationship of the Form $y = ax^b$ (c) Examples of a Non-linear Relationship $y = \frac{x}{ax + b}$

The functions like exponential function ($y = ax^b$) and power function $\left(y = \frac{x}{ax + b}\right)$ non-linear relationships between the dependent and independent variables that cannot be fitte a line. This is shown in Figures 5.2 (b) and (c).

Types of Regression Methods

The classification of regression methods is shown in Figure 5.3.

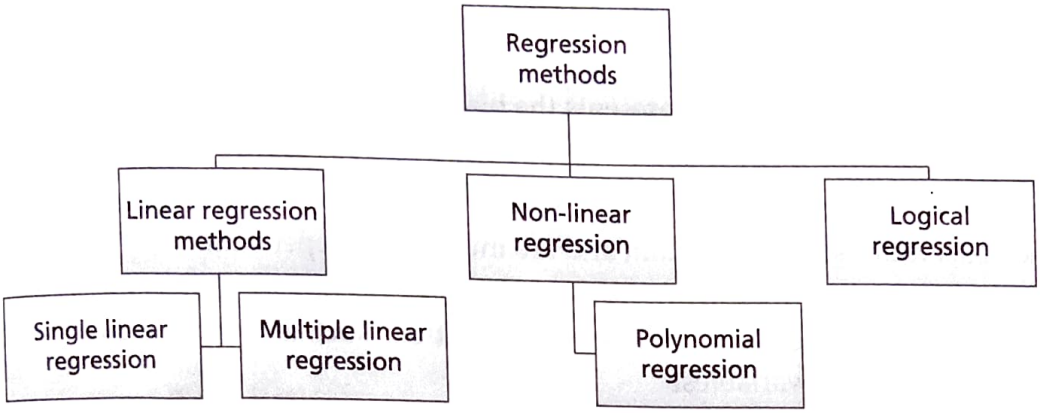

Figure 5.3: Types of Regression Methods

Linear Regression It is a type of regression where a line is fitted upon given data for finding the linear relationship between one independent variable and one dependent variable to describe relationships.

Multiple Regression It is a type of regression where a line is fitted for finding the linear relationship between two or more independent variables and one dependent variable to describe relationships among variables.

Polynomial Regression It is a type of non-linear regression method of describing relationships among variables where N^{th} degree polynomial is used to model the relationship between one independent variable and one dependent variable. Polynomial multiple regression is used to model two or more independent variables and one dependant variable.

Logistic Regression It is used for predicting categorical variables that involve one or more independent variables and one dependent variable. This is also known as a binary classifier.

Lasso and Ridge Regression Methods These are special variants of regression method where regularization methods are used to limit the number and size of coefficients of the independent variables.

Limitations of Regression Method

1. Outliers – Outliers are abnormal data. It can bias the outcome of the regression model, as outliers push the regression line towards it.
2. Number of cases – The ratio of independent and dependent variables should be at least 20 : 1. For every explanatory variable, there should be at least 20 samples. Atleast five samples are required in extreme cases.
3. Missing data – Missing data in training data can make the model unfit for the sampled data.
4. Multicollinearity – If exploratory variables are highly correlated (0.9 and above), the regression is vulnerable to bias. Singularity leads to perfect correlation of 1. The remedy is to remove exploratory variables that exhibit correlation more than 1. If there is a tie, then the tolerance (1 – R squared) is used to eliminate variables that have the greatest value.

5.3 INTRODUCTION TO LINEAR REGRESSION

In the simplest form, the linear regression model can be created by fitting a line among the scattered data points. The line is of the form given in Eq. (5.2).

$$y = a_0 + a_1 \times x + e \tag{5.2}$$

Here, a_0 is the intercept which represents the bias and a_1 represents the slope of the line. These are called regression coefficients. e is the error in prediction.

The assumptions of linear regression are listed as follows:

1. The observations (y) are random and are mutually independent.
2. The difference between the predicted and true values is called an error. The error is also mutually independent with the same distributions such as normal distribution with zero mean and constant variables.
3. The distribution of the error term is independent of the joint distribution of explanatory variables.
4. The unknown parameters of the regression models are constants.

The idea of linear regression is based on Ordinary Least Square (OLS) approach. This method is also known as ordinary least squares method. In this method, the data points are modelled using a straight line. Any arbitrarily drawn line is not an optimal line. In Figure 5.4, three data points and their errors (e_1, e_2, e_3) are shown. The vertical distance between each point and the line (predicted by the approximate line equation $y = a_0 + a_1x$) is called an error. These individual errors are added to compute the total error of the predicted line. This is called sum of residuals. The squares of the individual errors can also be computed and added to give a sum of squared error. The line with the lowest sum of squared error is called line of best fit.

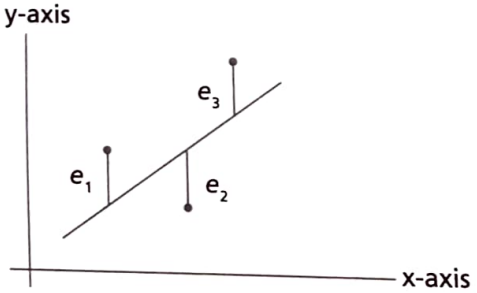

Figure 5.4: Data Points and their Errors

In another words, OLS is an optimization technique where the difference between the data points and the line is optimized.

Mathematically, based on Eq. (5.2), the line equations for points (x_1, x_2, …, x_n) are:

$$y_1 = (a_0 + a_1x_1) + e_1$$
$$y_2 = (a_0 + a_1x_2) + e_2$$
$$\vdots$$
$$y_n = (a_0 + a_1x_n) + e_n$$

In general, the error is given as: $e_i = y_i - (a_0 + a_1x_i)$

This can be extended into the set of equations as shown in Eq. (5.3).

Here, the terms $(e_1, e_2, \ldots, e_n)$ are error associated with the data points and denote the difference between the true value of the observation and the point on the line. This is also called as residuals. The residuals can be positive, negative or zero.

A regression line is the line of best fit for which the sum of the squares of residuals is minimum. The minimization can be done as minimization of individual errors by finding the parameters a_0 and a_1 such that:

$$E = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i)) \tag{5.5}$$

Or as the minimization of sum of absolute values of the individual errors:

$$E = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |(y_i - (a_0 + a_1 x_i))| \tag{5.6}$$

Or as the minimization of the sum of the squares of the individual errors:

$$E = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2 \tag{5.7}$$

Sum of the squares of the individual errors, often preferred as individual errors (positive and negative errors), do not get cancelled out and are always positive, and sum of squares results in a large increase even for a small change in the error. Therefore, this is preferred for linear regression.

Therefore, linear regression is modelled as a minimization function as follows:

$$\begin{aligned} J(a_1, a_0) &= \sum_{i=1}^{n} [y_i - f(x_i)]^2 \\ &= \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]^2 \end{aligned} \tag{5.8}$$

Here, $J(a_0, a_1)$ is the criterion function of parameters a_0 and a_1. This needs to be minimized. This is done by differentiating and substituting to zero. This yields the coefficient values of a_0 and a_1. The values of estimates of a_0 and a_1 are given as follows:

$$a_1 = \frac{(\overline{xy}) - (\overline{x})(\overline{y})}{(\overline{x_i^2}) - (\overline{x})^2} \tag{5.9}$$

And the value of a_0 is given as follows:

$$a_0 = (\overline{y}) - a_1 \times \overline{x} \tag{5.10}$$

Let us consider a simple problem to illustrate the usage of the above concept.

Example 5.1: Let us consider an example where the five weeks' sales data (in Thousands) is given as shown below in Table 5.1. Apply linear regression technique to predict the 7th and 9th month sales.

Table 5.1: Sample Data

x_i (Week)	y_i (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Solution: Here, there are 5 items, i.e., $i = 1, 2, 3, 4, 5$. The computation table is shown below (Table 5.2). Here, there are five samples, so i ranges from 1 to 5.

Table 5.2: Computation Table

x_i	y_i	$(x_i)^2$	$x_i \times y_i$
1	1.2	1	1.2
2	1.8	4	3.6
3	2.6	9	7.8
4	3.2	16	12.8
5	3.8	25	19
Sum = 15 Average of (x_i) $= \bar{x} = \frac{15}{5}$ $= 3$	Sum = 12.6 Average of (y_i) $= \bar{y} = \frac{12.6}{5}$ $= 2.52$	Sum = 55 Average of (x_i^2) $= \overline{x_i^2} = \frac{55}{5}$ $= 11$	Sum = 44.4 Average of $(x_i \times y_i)$ $= \overline{xy} = \frac{44.4}{5}$ $= 8.88$

Let us compute the slope and intercept now using Eq. (5.9) as:

$$a_1 = \frac{8.88 - 3(2.52)}{11 - 3^2} = 0.66$$

$$a_0 = 2.52 - 0.66 \times 3 = 0.54$$

The fitted line is shown in Figure 5.5.

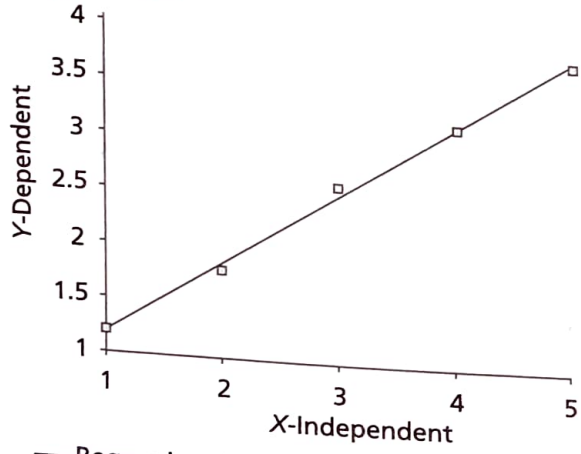

Figure 5.5: Linear Regression Model Constructed

Let us model the relationship as $y = a_0 + a_1 \times x$. Therefore, the fitted line for the above data is: $y = 0.54 + 0.66 \times x$.

The predicted 7th week sale would be (when $x = 7$), $y = 0.54 + 0.66 \times 7 = 5.16$ and the 12th month, $y = 0.54 + 0.66 \times 12 = 8.46$. All sales are in thousands.

Scan for *'Additional Examples'*

Linear Regression in Matrix Form

Matrix notations can be used for representing the values of independent and dependent var ables. This is illustrated through Example 5.2.

The Eq. (5.3) can be written in the form of matrix as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{5.11}$$

This can be written as:

$Y = Xa + e$, where X is an $n \times 2$ matrix, Y is an $n \times 1$ vector, a is a 2×1 column vector and e is an $n \times 1$ column vector.

Example 5.2: Find linear regression of the data of week and product sales (in Thousands) given in Table 5.3. Use linear regression in matrix form.

Table 5.3: Sample Data for Regression

x_i (Week)	y_i (Product Sales in Thousands)
1	1
2	3
3	4
4	8

Solution: Here, the dependent variable X is be given as:

$$x^T = [1\ 2\ 3\ 4]$$

And the independent variable is given as follows:

$$y^T = [1\ 3\ 4\ 8]$$

The data can be given in matrix form as follows:

$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$. The first column can be used for setting bias.

and $Y = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 8 \end{pmatrix}$

The regression is given as:

$$a = ((X^T X)^{-1} X^T) Y$$

The computation order of this equation is shown step by step as:

1. Computation of $(X^TX) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \times \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}$

2. Computation of matrix inverse of $(X^TX)^{-1} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}^{-1} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix}$

3. Computation of $((X^TX)^{-1}X^T) = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{pmatrix}$

4. Finally, $((X^TX)^{-1}X^T)Y = \begin{pmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \\ 4 \\ 8 \end{pmatrix} = \begin{pmatrix} -1.5 \\ 2.2 \end{pmatrix} \begin{pmatrix} Intercept \\ slope \end{pmatrix}$

Thus, the substitution of values in Eq. (5.11) using the previous steps yields the fitted line $2.2x - 1.5$.

5.4 VALIDATION OF REGRESSION METHODS

The regression model should be evaluated using some metrics for checking the correctness. The following metrics are used to validate the results of regression.

Standard Error

Residuals or error is the difference between the actual (y) and predicted value ($\hat{y}$).

If the residuals have normal distribution, then the mean is zero and hence it is desirable. This is a measure of variability in finding the coefficients. It is preferable that the error be less than the coefficient estimate. The standard deviation of residuals is called residual standard error. If it is zero, then it means that the model fits the data correctly.

Mean Absolute Error (MAE)

MAE is the mean of residuals. It is the difference between estimated or predicted target value and actual target incomes. It can be mathematically defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{5.12}$$

Here, $\hat{y}$ is the estimated or predicted target output and y is the actual target output, and n is the number of samples used for regression analysis.

Mean Squared Error (MSE)

It is the sum of square of residuals. This value is always positive and closer to 0. This is given mathematically as:

$$\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2 \tag{5.13}$$

Root Mean Square Error (RMSE)

The square root of the MSE is called RMSE. This is given as:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{5.14}$$

Relative MSE

Relative MSE is the ratio of the prediction ability of the $\hat{y}$ to the average of the trivial population. The value of zero indicates that the model is perfect and its value ranges between 0 and 1. If the value is more than 1, then the created model is not a good one. This is given as follows:

$$\text{RelMSE} = \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2} \tag{5.15}$$

Coefficient of Variation

Coefficient of variation is unit less and is given as:

$$\text{CV} = \frac{\text{RMSE}}{\bar{y}} \tag{5.16}$$

Example 5.3: Consider the following training set Table 5.4 for predicting the sales of the items.

Table 5.4: Training Item Table

Items x_i	Actual Sales (In Thousands) y_i
I_1	80
I_2	90
I_3	100
I_4	110
I_5	120

Consider two fresh items I_6 and I_7, whose actual values are 80 and 75, respectively. A regression model predicts the values of the items I_6 and I_7 as 75 and 85, respectively. Find MAE, MSE, RMSE, RelMSE and CV.

Solution: The test items' actual and prediction is given in Table 5.5 as:

Table 5.5: Test Item Table

Test Items	Actual Value y_i	Predicted Value $\bar{y}_i$
I_6	80	75
I_7	75	85

Mean Absolute Error (MAE) using Eq. (5.12) is given as:

$$\text{MAE} = \frac{1}{2} \times |80 - 75| + |75 - 85| = \frac{15}{2} = 7.5$$

Mean Squared Error (MSE) using Eq. (5.13) is given as:

$$\text{MSE} = \frac{1}{2} \times |80 - 75|^2 + |75 - 85|^2 = \frac{125}{2} = 62.5$$

Root Mean Square error using Eq. (5.14) is given as:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{62.5} = 7.91$$

For finding RelMSE and CV, the training table should be used to find the average of y.

The average of y is $\frac{80 + 90 + 100 + 110 + 120}{5} = \frac{500}{5} = 100.$

RelMSE using Eq. (5.15) can be computed as:

$$\text{RelMSE} = \frac{(80 - 75)^2 + (75 - 85)^2}{(80 - 100)^2 + (75 - 100)^2} = \frac{125}{1025} = 0.1219$$

CV can be computed using Eq. (5.16) as $\frac{\sqrt{62.5}}{100} = 0.08.$

Coefficient of Determination

To understand the coefficient of determination, one needs to understand the total variation c coefficients in regression analysis. The sum of the squares of the differences between the y-valu of the data pair and the average of y is called total variation. Thus, the following variations ca be defined.

The explained variation is given as:

$$= \Sigma(\hat{y}_i - \bar{y}_i)^2 \quad (5.1$$

The unexplained variation is given as:

$$= \Sigma(y_i - \hat{y}_i)^2 \quad (5.1$$

Thus, the total variation is equal to the explained variation and the unexplained variation.

The coefficient of determination r^2 is the ratio of the explained and total variations.

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} \quad (5.1$$

It is a measure of how many future samples are likely to be predicted by the regression mod Its value ranges from 1 to $-\infty$, where 1 is the most optimum value. It also signifies the proporti of variance. Here, r is the correlation coefficient. If $r = 0.95$, then r^2 is given as $0.95 \times 0.95 = 0.90$ This means that 90% of the model can be explained by the relationship between x and y. The re 10% is unexplained and that may be due to various reasons such as noise, chance, or error.

Standard Error Estimate

Standard error estimate is another useful measure of regression. It is the standard deviation of observed values to the predicted values. This is given as:

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)}{n - 2}} \quad (5$$

Here, as usual, y_i is the observed value and $\hat{y}_i$ is the predicted value. Here, n is the num of samples.

Example 5.4: Let us consider the data given in the Table 5.3 with actual and predicted values. Find standard error estimate.

Solution: The observed value or the predicted value is given below in Table 5.6.

Table 5.6: Sample Data

x_i	y_i	Predicted Value	$(y - \hat{y})^2$
1	1.5	1.46	$(1.5 - 1.46)^2 = 0.0016$
2	2.9	2.02	$(2.9 - 2.02)^2 = 0.7744$
3	2.7	2.58	$(2.7 - 2.58)^2 = 0.0144$
4	3.1	3.14	$(3.1 - 3.14)^2 = 0.0016$

The sum of $(y - \hat{y})^2$ for all $i = 1, 2, 3$ and 4 (i.e., number of samples $n = 4$) is 0.792. The standard deviation error estimate as given in Eq. (5.20) is:

$$\sqrt{\frac{0.792}{4-2}} = \sqrt{0.396} = 0.629$$

5.5 MULTIPLE LINEAR REGRESSION

Multiple regression model involves multiple predictors or independent variables and one dependent variable. This is an extension of the linear regression problem. The basic assumptions of multiple linear regression are that the independent variables are not highly correlated and hence multicollinearity problem does not exist. Also, it is assumed that the residuals are normally distributed.

For example, the multiple regression of two variables x_1 and x_2 is given as follows:

$$\begin{aligned} y &= f(x_1, x_2) \\ &= a_0 + a_1 x_1 + a_2 x_2 \end{aligned} \tag{5.21}$$

In general, this is given for 'n' independent variables as:

$$\begin{aligned} y &= f(x_1, x_2, x_3, \ldots, x_n) \\ &= a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + \varepsilon \end{aligned} \tag{5.22}$$

Here, $(x_1, x_2, \ldots, x_n)$ are predictor variables, y is the dependent variable, $(a_0, a_1, \ldots, a_n)$ are the coefficients of the regression equation and ε is the error term. This is illustrated through Example 5.5.

Example 5.5: Apply multiple regression for the values given in Table 5.7 where weekly sales along with sales for products x_1 and x_2 are provided. Use matrix approach for finding multiple regression.

Table 5.7: Sample Data