CHAPTER 3

UNIT III

Regression

Syllabus

Linear Regression : Linear Models, A bi-dimensional example, Linear Regression and higher dimensionality, Ridge, Lasso and ElasticNet, Robust regression with random sample consensus, Polynomial regression, Isotonic regression.

Logistic Regression : Linear classification, Logistic regression, Implementation and Optimizations, Stochastic gradient descendent algorithms, Finding the optimal hyper-parameters through grid search, Classification metric, ROC Curve.

3.1 Introduction to Regression Analysis

- Regression analysis is one of the most widely used statistical techniques.
- It estimates relationships among variables, a dependent (target, response) and independent variable (s) (predictor, explanatory)
- It is a form of predictive modeling technique and also used for analyzing data.
- Some of the regression analysis techniques include forecasting, time series modelling.
- A curve or a line is fit to the available data, such that the differences between the distances of data points from the curve or line is minimum.

Uses of Regression Analysis

1. Modeling relationship between variables
2. Prediction of the target variable (forecasting)
3. Testing of hypothesis

Regression analysis can be categorized based on three metrics

1. Number and nature of Independent variable(s)
2. Number and nature of Dependent variable(s)
3. Shape of regression line

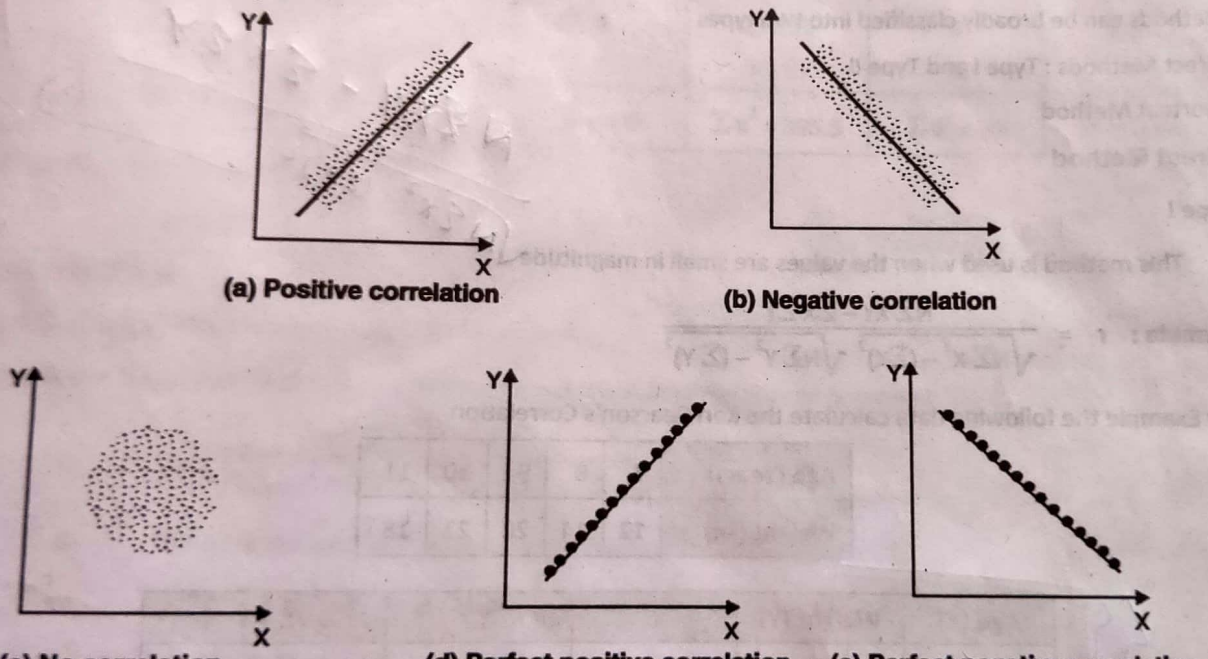
3.1.1 Steps in Conducting a Regression Analysis

- Analyzing the correlation (strength and directionality of data).
- Fitting the regression or least squares line or best fit line.
- Evaluating the model's validity and its usefulness.

(I) Analyzing the correlation (strength and directionality of data)

Definition of Correlation : Correlation, also known as correlation analysis, is a term used to denote the association or relationship between two (or more) quantitative variables.

- This analysis is based on the idea of a best fit line or a straight line, a linear relationship between two quantitative variables. The "strength" or the "extent" of an association between the variables and its direction is measured.
- The correlation analysis results in a correlation coefficient whose values range from – 1 to + 1.
- A value of + 1 indicates that the two variables are related in a positive (linear) manner.
  - As X is increasing, Y is increasing.
  - As X is decreasing, Y is decreasing.
- A value of – 1 indicates that the two variables are related in a negative (linear) manner.
  - As X is increasing, Y is decreasing.
  - As X is decreasing, Y is increasing.
- Whereas a value of 0 (zero) indicates that there is no relationship between the two variables.
- A correlation analysis is shown in a scatter plot or scatter diagram (a graphical representation) with one variable on x-axis and the other on y-axis.



(a) Positive correlation (b) Negative correlation

(c) No correlation (d) Perfect positive correlation (e) Perfect negative correlation

Fig. 3.1.1

Fig. 3.1.1 Scatter plot between two variables (a) shows positive correlation (b) shows negative correlation (c) shows no correlation (d) Shows perfect positive correlation (e) shows perfect negative correlation

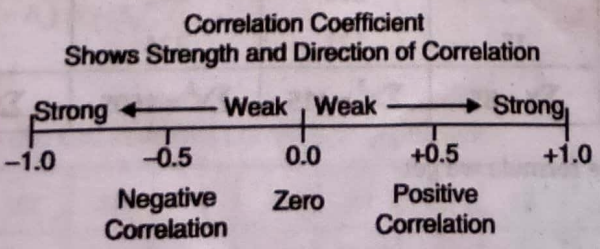


Fig. 3.1.2 : Spectrum of the correlation coefficient

Calculating Correlation coefficients - Karl Pearson's Correlation Coefficient r and Spearman's Correlation Coefficient rho (ρ)

- A correlation coefficient is a value that determines a relationship between the two variables.
- Two methods can be used to calculate the correlation coefficient viz. The Karl Pearson's product moment correlation coefficient r and the Spearman's rank correlation coefficient rho ρ.
- The Pearson's Correlation Coefficient determines the relationship between two variables based on three assumptions.
  1. Linear Relationship
  2. Independent variables
  3. Normal distribution of variables
- When the assumptions of Pearson's coefficient are not met, Spearman's rho method may be used. This method is based on the ranks given to the observations and not on their actual values
- Spearman's coefficient is a non-parametric equivalent of the Pearson's coefficient.
- Spearman's coefficient is robust coefficient and can also be used when one of the variables is ordinal

Karl Pearson's Correlation Coefficient r

Methods can be broadly classified into two types

1. Direct Methods : Type I and Type II
2. Shortcut Method

1. Direct Method

(i) Type I

This method is used when the values are small in magnitude

Formula : $$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

For Example the following data calculate the Karl Pearson's Correlation

Age (Years)	7	8	9	10	11
Weight (kg)	12	14	20	21	18

Age (X)	Weight (Y)	X^2	Y^2	X * Y
7	12	49	144	84
8	14	64	196	112
9	20	81	400	180
10	21	100	441	210
11	18	121	324	198
$\Sigma X = 45$	$\Sigma Y = 85$	$\Sigma X^2 = 415$	$\Sigma Y^2 = 1505$	$\Sigma X * Y = 784$

Substituting the values in above formula we get

$$r = 0.7756$$

(ii) Type II

Type II can be used when $\bar{X}$ and $\bar{Y}$ is not in fraction.

Formula is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

where x is the deviation of X from $\bar{X}$, and y is the deviation of Y from $\bar{Y}$.

xy is the product of the two deviations, x^2 and y^2 is the square of the deviations

Example :

Birth rate (X)	death rate (Y)	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	x * y
20	12	−8.5	−6	72.25	36	51
25	16	−3.5	−2	12.25	4	7
27	17	−1.5	−1	2.25	1	1.5
31	20	2.5	2	6.25	4	5
45	21	16.5	3	272.25	9	49.5
23	22	−5.5	4	30.25	16	−22
$\bar{X} = 28.5$	$\bar{Y} = 18$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 395.5$	$\Sigma y^2 = 70$	$\Sigma xy = 92$

Substituting in the above formula we get

$$r = 0.5529$$

2. Shortcut Method

This method can be used when the mean is in fractions.

Deviations are calculated from the assumed mean and the following formula is applied

$$r = \frac{N\Sigma dx\,dy - \Sigma dx\,\Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\,\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

where Σdx = Sum of deviations of X series from its Assumed Mean i.e. $\Sigma (X - A_x)$

Σdy = Sum of deviations of Y series from its Assumed Mean i.e. $\Sigma (Y - A_y)$

Σdx^2 = Sum of squared deviations of X series from its Assumed Mean i.e. $\Sigma (X - A_x)^2$

Σdy^2 = Sum of squared deviations of Y series from its Assumed Mean i.e. $\Sigma (Y - A_y)^2$

$\Sigma dx\,dy$ = Sum of products of deviations of X and Y series from their respective assumed means.

$$\Sigma dx\,dy = \Sigma (X - A_x)(Y - A_y)$$

N = Number of pairs

Example : Given X and Y calculate the Karl Pearsons Correlation Coefficient

x	32	36	22	21	24	23	18	19
y	23	22	15	16	14	15	15	21

The assumed mean for X is $A_x = 15$ and $A_y = 14$

x	y	$x - A_x = dx$	$y - A_y = dy$	dx^2	dy^2	dxdy
32	23	17	9	289	81	153
36	22	21	8	441	64	168
22	15	7	1	49	1	7
21	16	6	2	36	4	12
24	14	9	0	81	0	0
23	15	8	1	64	1	8
18	15	3	1	9	1	3
19	21	4	7	16	49	28
		$\Sigma dx = 75$	$\Sigma dy = 29$	$\Sigma dx^2 = 985$	$\Sigma dy^2 = 201$	$\Sigma dx\,dy = 379$

Substituting in the above formula we get

$$r = \frac{N\Sigma dx\,dy - \Sigma dx \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$r = \frac{(8 * 379) - (75 * 29)}{\sqrt{8 * 985 - (75 * 75)} * \sqrt{(8 * 201) - (29 * 29)}}$$

$$r = \frac{857}{1315.13}$$

$$r = 0.6516$$

Spearman's Correlation Coefficient rho (ρ)

- In this method variables of both the series are provided ranks.
- These ranks are used for calculating the coefficient of correlation.
- Ranks of X series are denoted by R_1 and Ranks for Y series are denoted as R_2.
- The difference between R_1 and R_2 is calculated and denoted as D.
- Differences are squared up and denoted as D^2.
- The total of D^2 is obtained and expressed as ΣD^2.

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here ΣD^2 = The total of the squares of differences of corresponding ranks

N = Number of pairs of observations

r_k = Coefficient of correlation.

(I) Fitting the regression or least squares line

In Mathematics the equation of Line is given by

$$y = mx + c$$

Where m = slope or gradient of the line and c=intercept or where the line cuts the y axis.

In regression this equation is given by

$$y = b_0 + b_1 x$$

Where B_0 represents the intercept is that value of Y or the dependent variable, when the value of predictor variable is zero and B_1 represents the slope

(II) Evaluating the validity and usefulness of the model

Validation techniques can be broadly classified into two types

(i) Numerical and

(ii) Graphical

(i) In Numerical technique the value of R^2 (coefficient of Determination) is to be analyzed

R^2 (coefficient of Determination) predicts the extent of variability in the dependent variable. This variability can be explained by the independent variable

(ii) A graphical technique known as Graphical analysis of residuals can be used for validation, which uses graphs to visually inspect the data's robustness

3.2 Linear Regression : A Bi-dimensional Example

- Linear Regression is a part of a special area of statistics called as **Bivariate statistics.**
- Bivariate means two variables (One independent and one dependent)

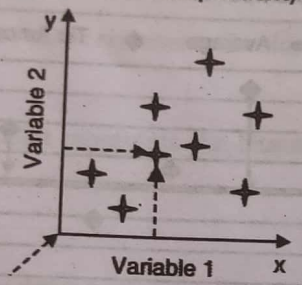


Fig. 3.2.1

- The value of one variable is a function of the other variable.
- The value of y is a function of x; $y = f(x)$.
- The value of the dependent variable is a function of the independent variable.

Let us start with a simple example. The problem statement is give below.

Example : A restaurant owner would like to develop a model that will allow you to make predictions about the tip amount that will be given for a particular bill amount.

- First we will use only a single variable (tip amount).
- Here the meal number has no significance with respect to tip amount column. It just represents the meal number.

Table 3.2.1 : Meal number and Tip amount

Meal	Tip Amount (Rs.)
1	5
2	17
3	11
4	8
5	14
6	5

- The Table 3.2.1 shows the meal number and the amount of tip offered by the customer in Rs.
- With only one variable (tip amount) and no other information, the best prediction for next measurement is the mean of the sample itself.

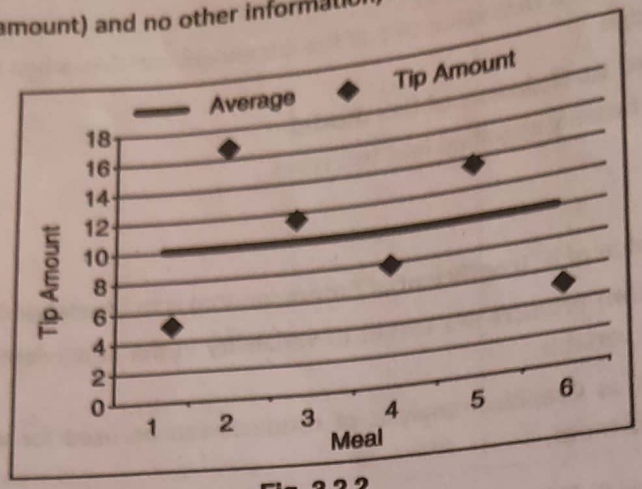


Fig. 3.2.2

So based on Fig. 3.2.2 we can say that for meal number 7, a tip amount of Rs. 10 will be given.

Goodness of fit for the tips

Calculate the residual error (Difference of Tip amount from the average value).

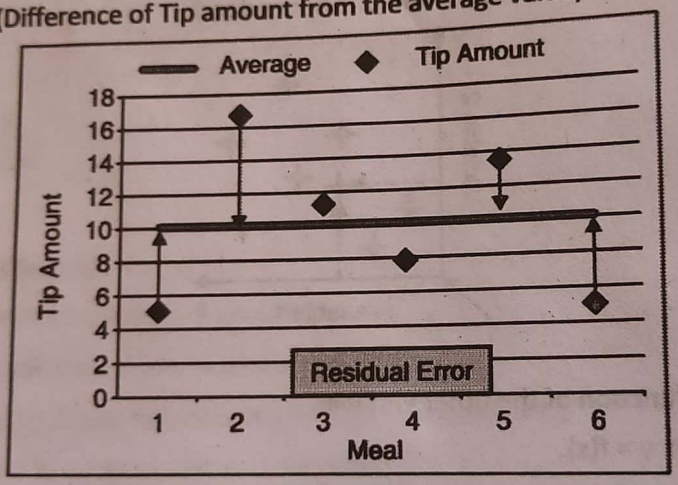


Fig. 3.2.3

Meal	Residual	Residual^2
1	−5	25
2	7	49
3	1	1
4	−2	4
5	4	16
6	−5	25

Residual^2 : Makes the deviation positive and it also emphasizes the larger deviation.

Sum of Squared error(SSE) = (25 + 49 + 1 + 4 + 16 + 25) = 120

- **Goal of Simple Linear Regression :** To create a linear model that minimizes the sum of squares of residuals / error (SSE).
- Let us now consider an independent variable (Bill amount) along with dependent variable (Tip amount).
- Before we begin with linear regression, let us understand some basics of the Equation of a Line.

3.2.1 The Equation of Line

The equation of Line can be represented in three different forms :

1. Slope-Intercept form : $y = mx + b$
2. Point Slope form : $y - y_1 = m(x - x_1)$
3. Standard form : $A_x + B_y = C$

We will use Slope Intercept form $y = mx + b$

Where **m** : slope of the line

b : Y-intercept

For Example given an equation of Line $y = -\frac{6}{5}x + 2$

This equation can be represented as follows (– 6/5 : represents the slope (rise = – 6, run = 5) and y-intercept = 2)

Procedure to draw the line

- To draw a line we need two points(one point represents the intercept and the other we obtain by marking the slope value)
- First Mark the y-intercept on a XY plane
- Then from intercept rise = – 6 (that means go down by 6 units)
- Next run = 5 (go right from the above step and mark a point)
- The dark dot indicates the two points which represents the lines

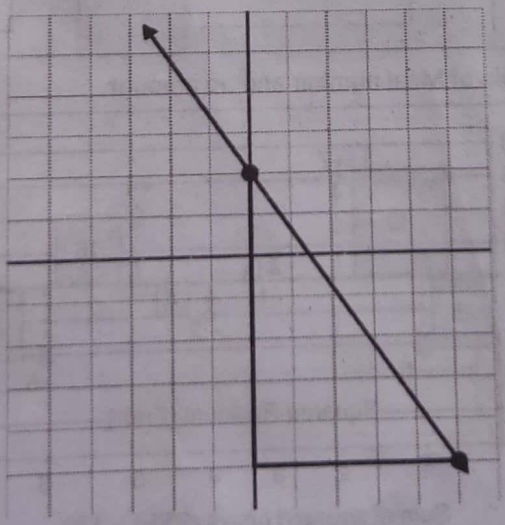


Fig. 3.2.4

3.2.2 Simple Linear Regression Model

- We know that slope intercept form of Line is represented as

$$y = mx + b$$

- The overall regression model for entire population is represented as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where, β_0 : y-intercept population parameter

β_1 : Slope Population Paramter

ε : Error term, unexplained variation in y

- A simple linear regression model is represented as

$$E(y) = \beta_0 + \beta_1 x$$

Where, **E(y) :** Mean or Expected value of y for a given value of x

- If we actually knew the population parameters β_0 and β_1, we could simply use the simple linear regression equation
- In reality we almost never have the population parameters. Therefore we will estimate them using sample data. When using sample data , we have to change our equation a little bit

$$\hat{Y} = b_0 + b_1 x$$

- Where y-hat is the point estimator of E(y), and is the mean of y for a given value of x

Example : Pattern Matching to General regression Models

$\hat{Y} = 14.87 - 0.014x$	$\hat{Y} = 48 + 7.8x$	$\hat{Y} = 0.3 - 3.3x$
E(y) Slope β_1 is/near 0 x	E(y) Slope β_1 is + x	E(y) Slope β_1 is – x

- Let us consider the previous example of Meal number and tip amount.

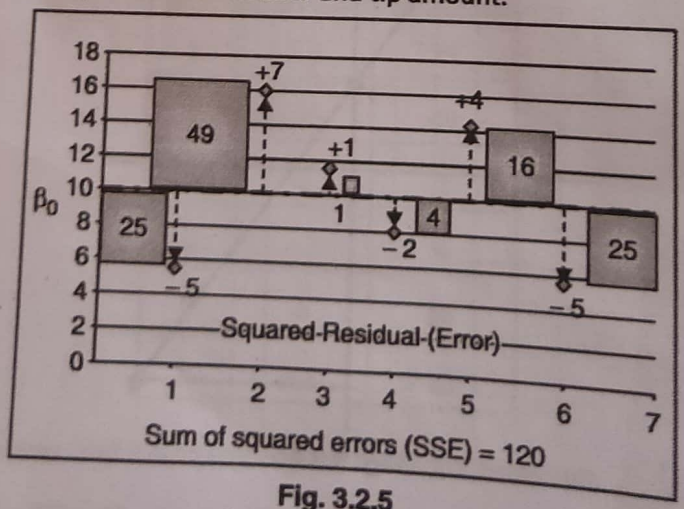


Fig. 3.2.5

- Using the equation of Linear regression model for every value of x the y estimate is given by ($b_1 = 0$ as we can see from the above slope of the line is zero).

$$\hat{Y} = b_0 + b_1 x$$

$$\hat{Y} = b_0 + (0)\, x$$

$$\hat{Y} = b_0$$

$$\hat{Y} = 10$$

- With only one dependent variable the only sum of squares is due to error. Therefore it is also the total and maximum sum of squares for the data under analysis

$$SSE = 120$$

$$SSE = SST \text{ (sum of squares total)}$$

$$\text{Therefore } SST = 120$$

3.2.2(A) Least Squares Method

- The Least squares method uses the Least squares criterion which is given by

$$\text{Min } \Sigma (y_i - \hat{y}_i)^2$$

Where, y_i : Observed value of Dependent variable

$\hat{y}_i$: Estimated (predicted) value of the dependent variable (predicted tip amount)

- The Goal is to minimize the sum of the squared difference between the observed value for the dependent variable and the estimated / predicted value of dependent variable that is provided by the regression line, Sum of the squared residuals.

3.3 Step by Step Procedure for Linear Regression

Let us now consider two variables, a dependent variable (Tip amount) and a independent variable(Bill amount)

Table 3.3.1 : Bill Amount and Tip amount

Bill Amount (Rs.)	Tip Amount (Rs.)
34	5
108	17
64	11
88	8
99	14
51	5

Step 1 : Scatter Plot

- This is used to observe a pattern in the data.
- It will also help you in looking for any outliers present in your data.
- Also see that your graph is scaled correctly and proportionately.

- Since the smallest bill amount was Rs.34, So started the x-axis at Rs. 20. Similarly for y-axis the smallest tip amount was Rs.5 so the y-axis is started at Rs. 4.
- For our example the scatter plot is shown in Fig. 3.3.1.

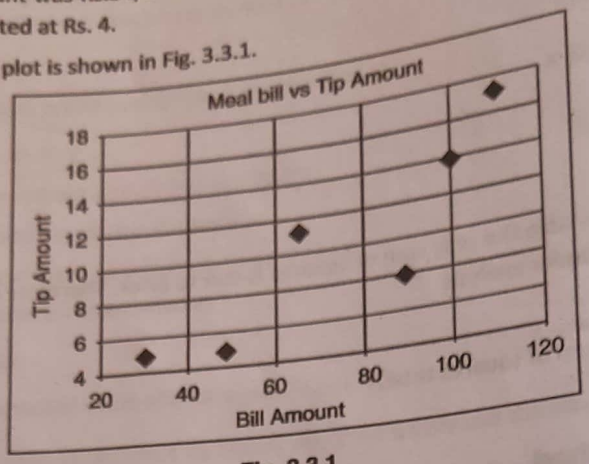


Fig. 3.3.1

Step 2 : Look for a Visual Line

- Look for a rough visual line passing through the data.
- We have to check whether the data seems to fall along a line.
- As we can see from Fig. 3.3.2 that the data points do seem to fall on a line.
- If your data does not seem to fall on a line, just stop the process.

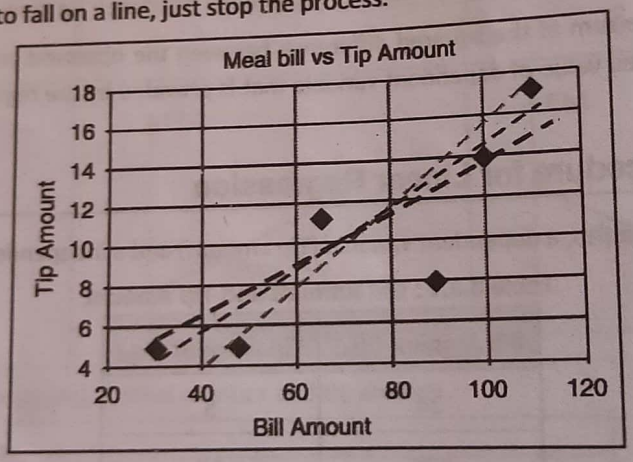


Fig. 3.3.2

Step 3 : Correlation

- To find the strength of relationship between the dependent variable and independent variable, find the correlation.
- The formula to find the correlation is using Pearson correlation.

$$r = \frac{\text{covariance (x, y)}}{\text{standard dev (x)} * \text{standard dev (y)}}$$

- Covariance of a sample and population can be calculated as follows.
- The only difference between the two terms is in sample covariance denominator has n-1 and in population covariance the denominator has N.

Sample Covariance

$$S_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Population Covariance

$$\sigma_{xy} = \frac{\Sigma (x_i - \mu_x)(y_i - \mu_y)}{N}$$

- The standard deviation can be calculated as

$$\sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

- Using the above formulas , let us calculate the correlation coefficient

Bill Amount (x)	Tip Amount (y)	Bill Deviation $(x_i - \bar{x})$	Tip Deviation $(y_i - \bar{y})$	Deviation Products $(x_i - \bar{x}) * (y_i - \bar{y})$	Bill deviation squared $(x_i - \bar{x})^2$	Tip Deviation squared $(y_i - \bar{y})^2$
34	5	−40	−5	200	1600	25
108	17	34	7	238	1156	49
64	11	−10	1	−10	100	1
88	8	14	−2	−28	196	4
99	14	25	4	100	625	16
51	5	−23	−5	115	529	25
$\bar{x} = 74$	$\bar{y} = 10$			$\Sigma (x_i - \bar{x}) * (y_i - \bar{y})$ = 615	$\Sigma (x_i - \bar{x})^2$ = 4206	$\Sigma (y_i - \bar{y})^2$ = 120

- Calculating the sample covariance

$$S_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{615}{5}$$

$$S_{xy} = 123$$

- Calculating the standard deviation

$$\sigma_x = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

$$\sigma_x = \sqrt{\frac{4206}{5}}$$

$$\sigma_x = 29$$

- Similarly standard deviation for y

$$\sigma_y = \sqrt{\frac{4206}{5}}$$

$$\sigma_y = 4.89$$

$$r = \frac{\text{covariance (x, y)}}{\text{standard dev (x) * standard dev (y)}}$$

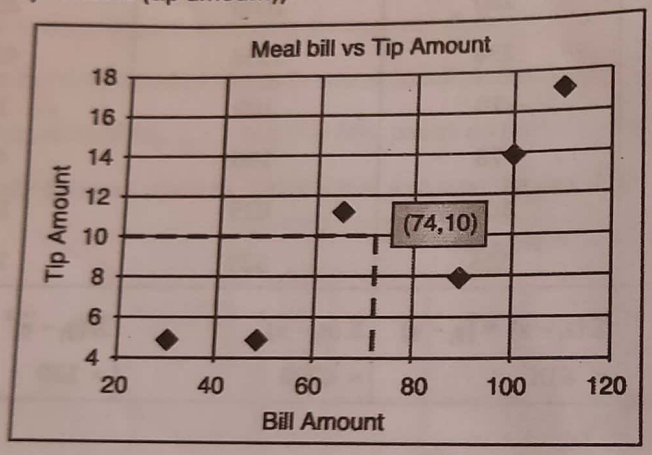
$$r = \frac{123}{29 * 4.89}$$

$$r = 0.8656$$

- The value of correlation coefficient indicates a strong positive linear relationship between the dependent variable(Tip amount) and the independent variable(bill amount)

Step 4 : Descriptive statistics/ Centroid

- Find the mean of each variable
- As we can see the mean of bill amount is 74 and mean of tip amount is 10
- Plot these mean on the graph as shown in Fig. 3.3.3
- This point is very important and is called as **centroid.**
- The best fit for regression line will or must pass through this centroid (mean of x variable(bill amount) and mean of y variable (tip amount)).



Bill Amount	Tip Amount
34	5
108	17
64	11
88	8
99	14
51	5
$\bar{x} = 74$	$\bar{y} = 10$

Fig. 3.3.3

Step 5 : Calculations

- We know the regression model

$$\hat{Y} = b_0 + b_1 x$$

where b_0 represents the intercept and b_1 represents the slope.

- The formula to calculate b_0 and b_1 is shown below :

$$b_0 = \bar{y} - b_1 x$$

where $\bar{x}$ is the mean of the independent variable(bill amount) and $\bar{y}$ is the mean of the dependent variable(tip amount)

$$b_1 = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Where x_i is the value of independent variable and y_i is the value of dependent variable.

- Calculating the values for b_0 and b_1 as shown in the Table 3.3.2.

Table 3.3.2

Bill Amount (x)	Tip Amount (y)	Bill Deviation $(x_i - \bar{x})$	Tip Deviation $(y_i - \bar{y})$	Deviation Products $(x_i - \bar{x}) * (y_i - \bar{y})$	Bill deviation squared $(x_i - \bar{x})^2$
34	5	−40	−5	200	1600
108	17	34	7	238	1156
64	11	−10	1	−10	100
88	8	14	−2	−28	196
99	14	25	4	100	625
51	5	−23	−5	115	529
$\bar{x} = 74$	$\bar{y} = 10$			$\Sigma (x_i - \bar{x}) * (y_i - \bar{y}) = 615$	$\Sigma (x_i - \bar{x})^2 = 4206$

As the value of b_0 is dependent on b_1 we shall first calculate b_1 and then b_0

$$b_1 = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$b_1 = \frac{615}{4206}$$

$$b_1 = 0.1462$$

After calculating b_1, now let us calculate b_0

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_0 = 10 - (0.1462 * 74)$$

$$b_0 = -0.8188$$

Therefore the Linear regression equation becomes

$$\hat{Y} = b_0 + b_1 x$$

Intercept

$$b_0 = -0.8188$$

Slope

$$b_1 = 0.1462$$

$$\hat{y}_i = -0.8188 + 0.1462\,x \quad \text{OR} \quad \hat{y}_i = -0.1462\,x - 0.8188$$

Using the above equation we can now draw the regression line

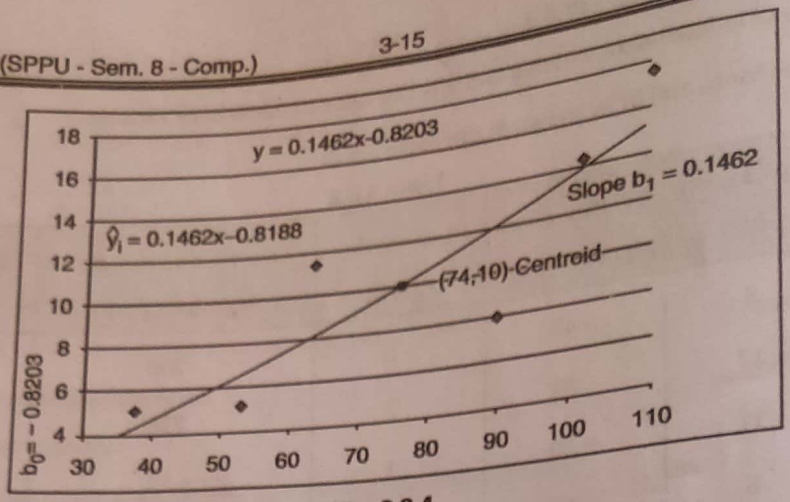


Fig. 3.3.4

Interpretation

- Based on the regression equation obtained following interpretation can be obtained.

$$\hat{y}_i = 0.1462\,x - 0.8188$$

- For every Rs. 1 the bill amount (x) increase, the tip amount would increase by Rs. 0.1462
- If the bill amount is zero, then the expected/predicted tip amount is Rs. 0.8188(this really makes no sense). The intercept may or may not make sense in the real world

3.3.1 Is the Regression Line Really Good ?

- Using the Regression Line obtained let us calculate the estimated (predicted) regression values for y.
- The details of calculation is shown below in the table

$$\hat{y} = 0.1462(x) - 0.8188$$

Substitute the value if x = 34

$$\hat{y} = 0.1462(34) - 0.8188$$

$$\hat{y} = 4.152$$

Similarly other values for $\hat{y}$ may be obtained

Bill Amount	Tip Amount	Predicted value
x	y	$\hat{y}_i = 0.1462(x) - 0.8188$
34	5	4.152
108	17	14.9708
64	11	8.538
88	8	12.0468
99	14	13.655
51	5	6.6374

After plotting these estimated values we get

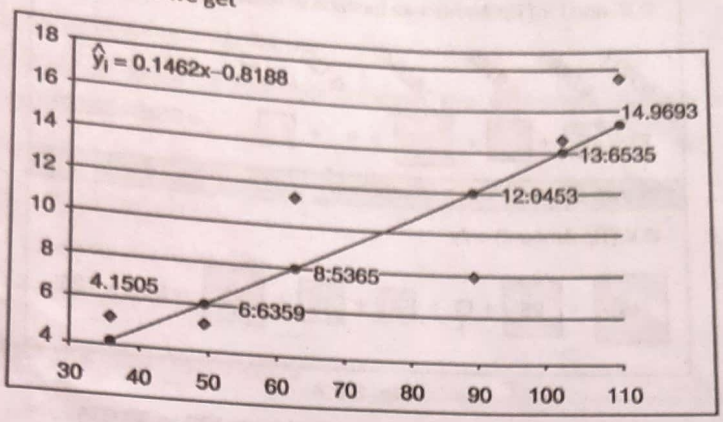


Fig. 3.3.5

- Next we need to find the residuals.
- Residual is the difference between the predicted value and the actual value of the tip amount as shown in the Fig. 3.3.6

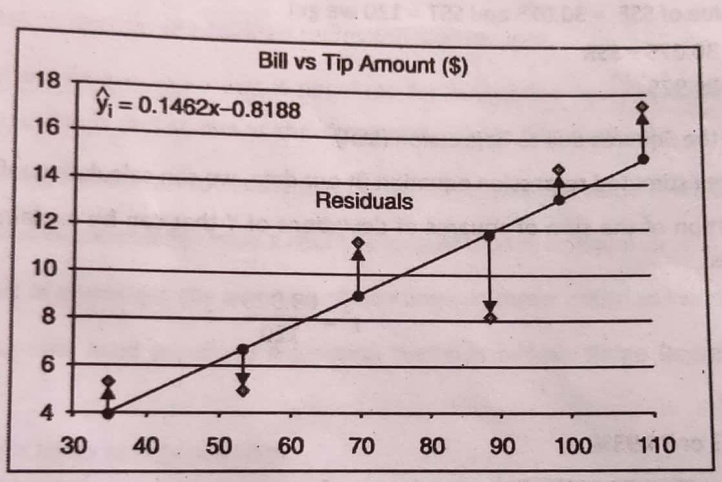


Fig. 3.3.6

Bill Amount	Tip Amount	Predicted Tip Amount	Error $(y - \hat{y}_i)$	Squared error $(y - \hat{y}_i)^2$
x	y	$\hat{y}_i = 0.1462(x) - 0.8188$		
34	5	4.152	0.848	0.7191
108	17	14.9708	2.0292	4.1177
64	11	8.538	2.462	6.0614
88	8	12.0468	−4.0468	16.3766
99	14	13.655	0.345	0.1190
51	5	6.6374	−1.6374	2.6811
				SSE Σ = 30.0749

Let us compare this case with our previous case where we have considered only one variable (dependent) i.e. tip amount.

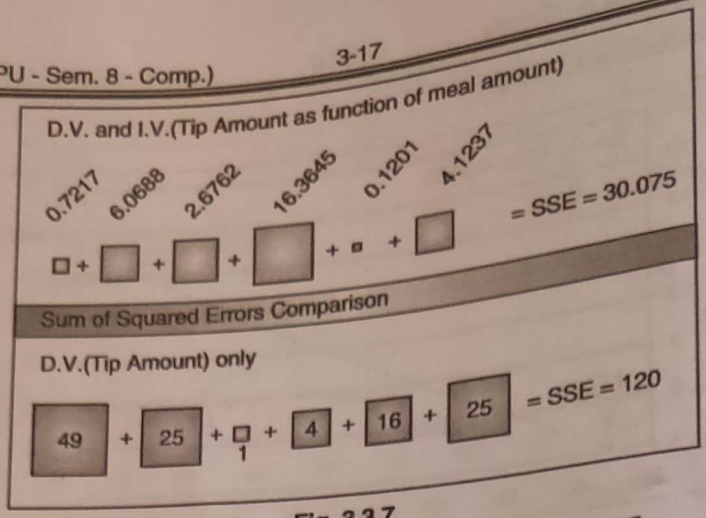


Fig. 3.3.7

- As we can see from the Fig. 3.3.7 the value of SSE has decreased from 120 to 30.075.
- We know that in our previous case SSE = 120 and SST = SSE = 120.
- Now in this case we have SSE = 30.075 and we know that

$$SST = SSE + SSR$$

Substituting the value of SSE = 30.075 and SST = 120 we get

$$120 = 30.075 + SSR$$

$$SSR = 89.925$$

This is called as Sum of the Squares due to Regression (SSR)

- Now to find out how the estimated regression equation fit our data, we can calculate coefficient of Determination r^2.
- r^2 measures the proportion of the sum of squares of deviations of Y that can be explained by the relationship fitted using predictor variables

$$r^2 = \frac{SSR}{SST} \qquad r^2 = \frac{89.925}{120}$$

Therefore

$$r^2 = 0.7493 \text{ or } 74.93\%$$

- We can conclude stating that 74.93% of the total sum of squares can be explained by the estimated regression equation to predict the tip amount. The remainder is an error. Therefore the regression line is a **GOOD FIT.**
- We can finally represent the SST , SSR and SSE terms graphically as shown Fig. 3.3.8.

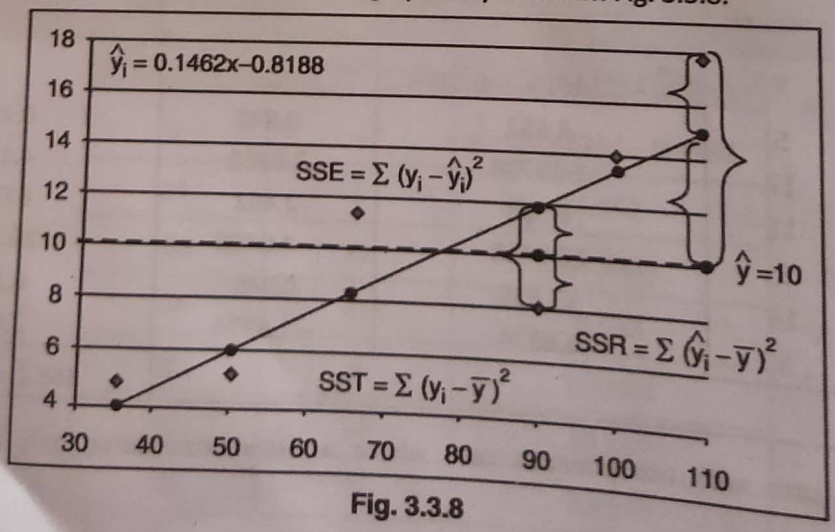


Fig. 3.3.8

Limitations of Linear regression

- Linear regression is limited to Linear relationships :
  - Linear regression looks at linear relationship between the dependent variable and independent variable. Sometimes this may not be true.
- Linear regression looks at the mean of the dependent variable.
- Linear regression is sensitive to outliers :
  - Outliers can have a effect on the regression
- Linear regression assumes the data are independent.

3.4 Linear and Higher Dimensionality

- We know that Linear regression is a statistical technique used for examining a relationship between a dependent variable Y and one or more independent variable x.
- The performance of Standard linear model (or the ordinary least squares method) degrades when a large multivariate data set containing a number of variables superior to the number of samples are given.
- To overcome this alternative strategy of penalized regression may be used.
- In this strategy, a linear model is created that is penalized for having too many variables in the model by adding a constraint in the equation. This is also known as **shrinkage** or **regularization methods**.
- The significance of imposing a penalty is to reduce (i.e. shrink) the coefficient value towards zero.
- This result in less contributive variables to have a coefficient close to zero or equal to zero.
- To determine the amount of shrinkage, the selection of a tuning parameter called as lambda is important.
- Some of the most commonly used penalized regression methods include Ridge Regression, Lasso Regression and Elastic Net Regression.
- Following are the different types of regularization.

1. **L1 Regularization**
   - This regularization is also known as Lasso Regularization.
   - This adds regularization terms in the model, which are functions of absolute value of coefficients.
   - Hence this technique can be used for feature selection.
2. **L2 Regularization**
   - This regularization is also known as **Ridge Regularization**.
   - This adds regularization terms in the model, which are functions of square of coefficients of parameters.
   - The coefficients of parameter can approach zero but never become zero.
3. **Combination of the above two**
   - This regularization is a combination of L1 and L2 regularization technique .
   - It known as Elastic Net Regularization.
   - This technique adds regularization terms in the model, which are combination of both L1 and L2.

Need for Regularization techniques in Generalized Linear Models (GLM)

- No particular distribution is assumed for the dependent variable. The dependent variable may follow distributions like normal, binomial, Poisson.
- The variance –bias tradeoff is addressed, it generally lowers the variance.
- The technique is more robust to handle multicollinearity.
- Sparse data is handled better.
- Natural feature selection.
- Overfitting on the trained data is minimized which results in more accurate predictions.

3.5 Ridge Regression

- Ridge regression is an extension to Linear Regression.
- This technique shrinks the regression coefficients, which results in variables with minor contribution resulting in their coefficients close to zero.
- The shrinkage is achieved by a term called as L2-Norm which is used in penalizing the regression model, which is the sum of squared coefficients.
- The amount of penalty can be fine tuned with lambda (λ) a constant.
- Selection of a good value for lambda is important.
- When λ=0, the term penalty will have no effect, and ridge regression will be equivalent to ordinary least squares coefficients.
- As λ increases to a large infinite value the shrinkage penalty grows and the ridge regression coefficients will get close to zero.
- As compared to Ordinary Least squares regression, Ridge regression is highly sensitive to the scale of predictors.
- Hence it is better to standardize the scale of the predictors before applying the Ridge regression so that all the predictors are on the same scale.
- This technique shrinks the coefficients close to zero but it will not set the coefficients exactly to zero.
- This can be overcome using Lasso regression.
- The least squares criterion or the cost function is given as

$$\sum_{i=1}^{M} \left(y_i - \hat{y}_i\right)^2$$

This can be rewritten as

$$\sum_{i=1}^{M} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} \omega_j \times x_{ij}\right)^2$$

M represents the instances

p- number of features

- In Ridge regression , the above cost function is altered by adding a penalty term which is square of the magnitude of the coefficients

$$\sum_{i=1}^{M}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}\omega_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}\omega_j^2$$

penalty

- The constraint on Ridge regression coefficients is given below

For some $c>0$, $\sum_{j=0}^{p}\omega_j^2<c$

- Ridge regression places a constraint on the coefficients (ω).
- Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multicollinearity.
- This technique performs better when the outcome is a function of many predictors, where all the coefficients are of roughly equal size.

3.5.1 Lasso Regression

- Lasso stands for Least Absolute Shrinkage and Selection Operator.
- This technique shrinks the regression coefficients towards zero by penalizing the model with a penalty term called as L1-norm.
- The penalty term is the sum of absolute coefficients.
- The coefficients with a minor contribution to the model are made exactly to zero by the penalty term.
- Lasso can be seen as a subset or feature selection method, which can reduce the complexity of the model.
- Selection of a good value for lambda λ is important.
- Lasso when compared to Ridge regression model produces more simple and interpretable models, that uses only a reduced set of predictors.
- Lasso regression performs better in situations where some of the predictors have large coefficients and the remaining have small coefficients.
- To make a choice between the two techniques (Ridge and Lasso regression), cross validation methods may be used for identifying which of the technique performs better on a particular data set.
- The cost function for Lasso Regression can be given as follows :

$$\sum_{i=1}^{M}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}\omega_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|\omega_j|$$

- The constraint given by the technique is for some $t>0$, $\sum_{j=0}^{p}|\omega_j|<t$
- This technique leads to zero coefficients, meaning some of the coefficients are completely neglected for evaluating the output.
- The amount of penalty can be fine tuned with lambda (λ) a constant.
- So Lasso regression helps in reducing the over fitting and also helps in feature selection.

3.5.2 ElasticNet Regression

- The Lasso regression technique sometimes does not perform well with highly correlated data and often performs worse than ridge in prediction.
- To overcome this drawback, a penalty that combines l1 norm and l2 norm is developed.
- The result of this is to effectively shrink the coefficients (like in ridge regression) and to set come coefficients to zero like lasso regression.
- The penalty for Elastic net is given as

$$\lambda\left[\frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right]$$

- Here $\alpha \in [0,1]$ is called the **mixing parameter** and λ has the same **interpretation.**
- For $\alpha = 0$ we have the Ridge Regression and for $\alpha = 1$ we have the Lasso regression.

3.6 Robust Regression with Random Sample Consensus

- Robust regression with **R**andom **S**ample **C**onsensus (**RANSAC**) is a general algorithm which can be used with other parametric estimation methods to obtain robust models.
- **RANSAC** is aimed to determine function parameters when the data has gross erroneous samples that can mislead the parameter estimation.
- The assumption RANSAC makes is that the training data consists of inliers that can be explained with the model and outliers that are gross erroneous samples that do not fit the model at all.
- If this assumption does not hold, RANSAC will not harm the parameter estimation as in this condition it will consider the whole dataset as inliers.
- It trains the model using only the inliers while ignoring the outliers.
- To reject the outliers, RANSAC uses small set of samples to train a model instead of using the whole of the data. It then enlarges the set with appropriate samples.
- Steps followed in RANSAC(General version)
  - A subset of data is sampled uniformly at random (i.e. the minimum number of points needed to estimate the model).
  - Using the sampled subset, estimate the parameters for the model of choice.
  - Error is calculated for all the remaining samples using an error function.
  - The number of inliers are calculated (i.e. all samples below a threshold error).
  - Recompute the Model using all inliers and hypothesis if the number of inliers are above a given threshold.
  - Repeat the above to find the best model.

3.6.1 Advantages and Disadvantages of RANSAC

Advantages

1. The method is simple and general.
2. It can be applied to many different problems.
3. It often works well in practice.

Disadvantages

1. Tuning of parameters is required.
2. Too many iterations are needed in some cases.
3. The method can fail if there is very low inlier ratio.

3.7 Polynomial Regression

- **Polynomial regression** is a form of linear regression.
- An n^{th} order Polynomial is used to model the relationship between the independent variable x and the dependent variable y.
- It fits a non-linear relationship between the value of x and conditional mean of y represented as $E(y|x)$.
- It is considered to be a special case of multiple linear regression.
- We know that the Linear regression model is represented as

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

- This model can be used for fitting any relationship that is linear in the unknown parameters β_1.
- In many cases such conditions may not hold, the relationship may be curvilinear between x and y. In such a cases an important class of Polynomial regression models may be used.

  For example the second order polynomial in one variable

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- In two variables the second order polynomial can be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

- Polynomial regression models are used when the response is curvilinear.
- For example consider some randomly generated data that has the following form.

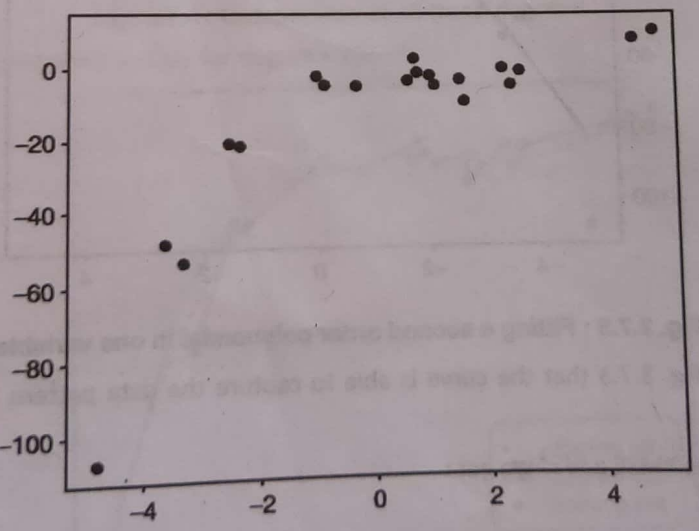


Fig. 3.7.1 : Randomly Generated data

- If we apply a simple Linear regression model to this data, it would look like the following :

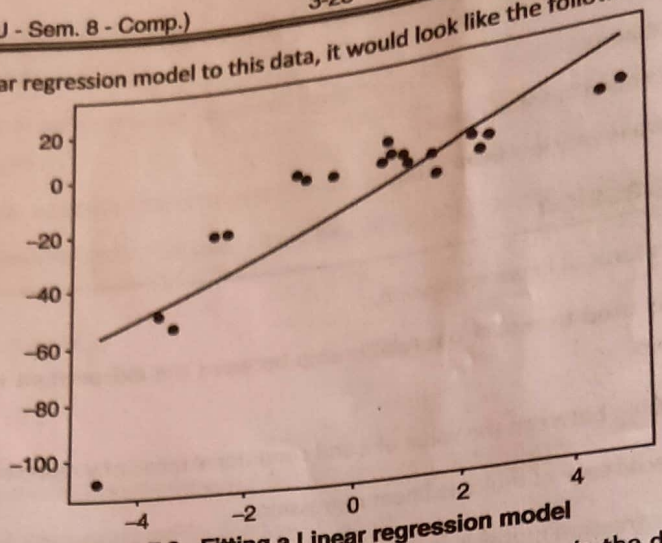


Fig. 3.7.2 : Fitting a Linear regression model

- As can be seen from Fig. 3.7.2 a straight line is unable to capture that pattern in the data. These can an underfitting case.
- If we calculate RMSE(Root Mean Squared Error) and r^2score of the above line we would get the following values :

  RMSE : 15.90

  r^2 : 0.6386

- To overcome the underfitting problem we can increase the complexity of the model. So that the nature of the curve would be quadratic but the underlying model will be still to be considered as linear (Applying second order polynomial in one variable).

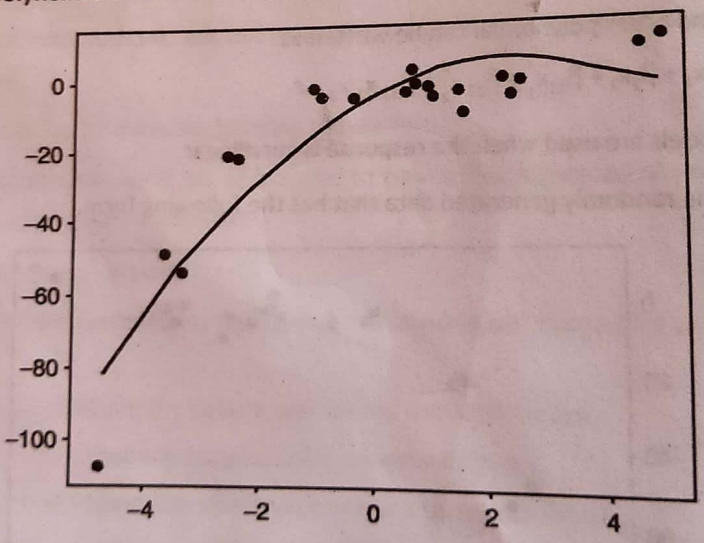


Fig. 3.7.3 : Fitting a second order polynomial in one variable

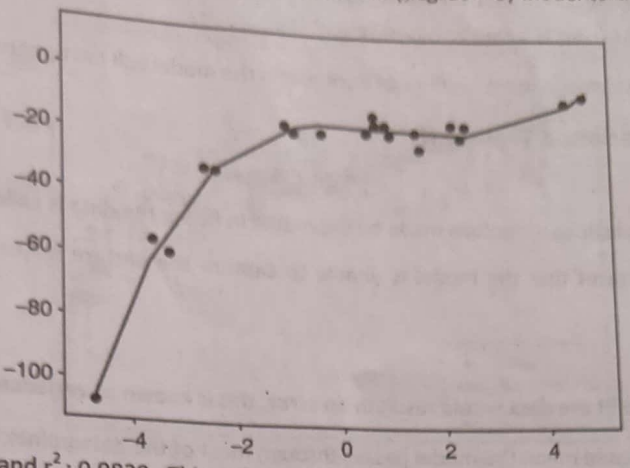
- As we can see from the Fig. 3.7.3 that the curve is able to capture the data pattern better than our Simple Linear regression model.
- Now if we have to calculate RMSE and r^2 we get :

  RMSE: 10.12

  r^2 : 0.8537

- We can see that our error has decreased and r^2score has increased.

- Now we can still increase the order of our model and try a cubic (degree = 3) model to the randomly generated data given above and see.



- The Value for RMSE: 3.44 and r^2 : 0.9830 . This reduces the error further and increases the r^2 score
- A comparison of all three degrees (Degree = 1,2, and 3)

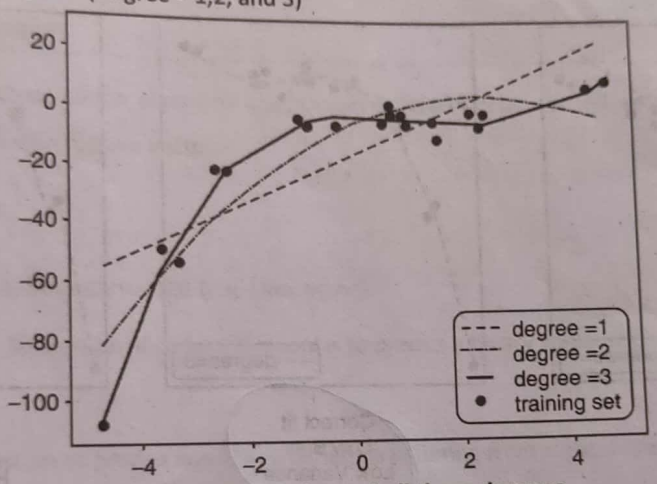


Fig. 3.7.4 : Comparison of all three degrees

- We can further increase the degree and try for degree = 20.

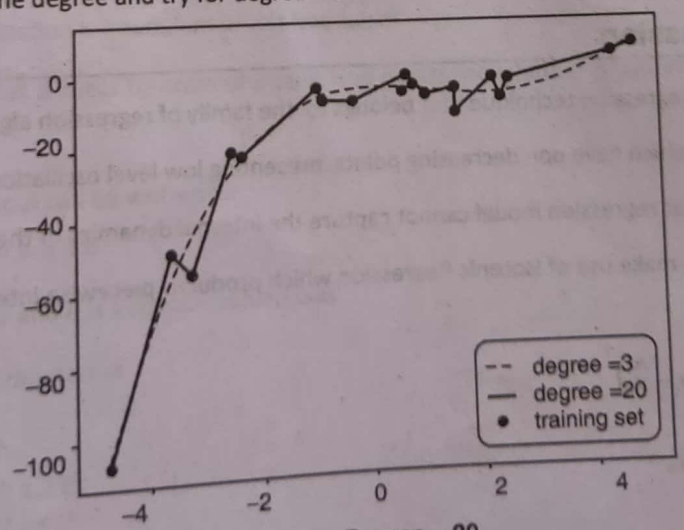


Fig. 3.7.5 : Degree = 20

- As we can see from the Fig. 3.7.5 that the curve passes through more data points. But the model has now started capturing noise in the data, this is a case of overfitting.
- Even if the model passes through more number of data points the model will fail to generalize on unseen data.

3.7.1 The Bias vs. Variance Trade-off

Bias

- The error due to the simplistic assumptions made by the model in fitting the data is called **Bias.**
- A high value of bias indicates that the model is unable to capture the pattern in the data, this is known as **under fitting.**

Variance

- A complex model trying to fit the data would result in an error, this is known as **variance.**
- A high value of variance would mean the model passes through most of the data points, but it results in overfitting.
- A graphical representation of the same is shown in Fig. 3.7.6

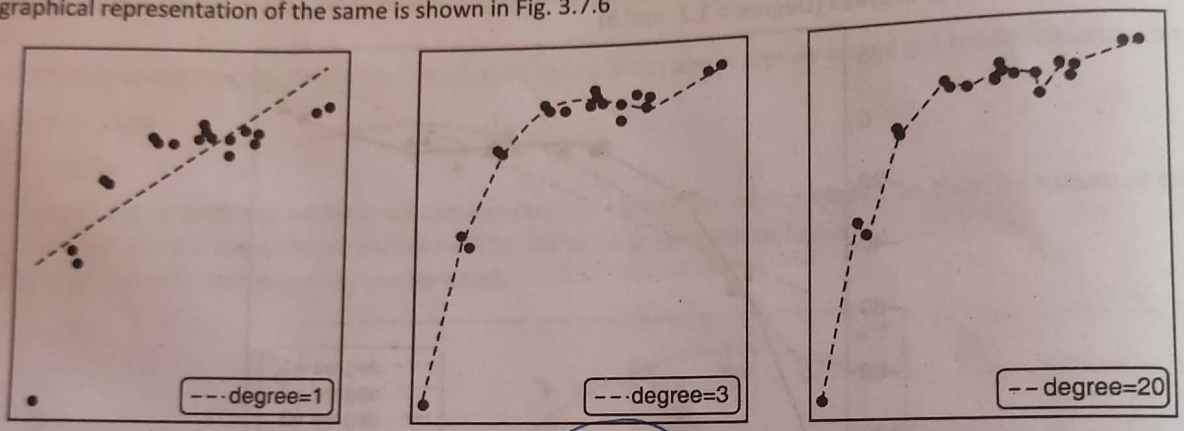


Fig. 3.7.6

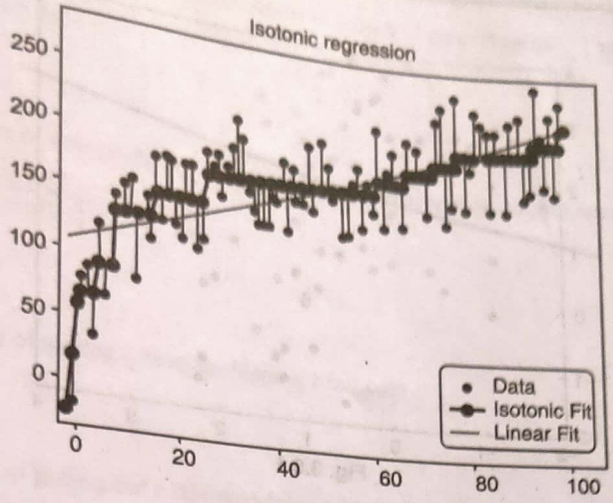
3.8 Isotonic Regression

- Isotonic Regression is a regression technique that belongs to the family of regression algorithms
- There may be datasets, which have non decreasing points, presenting low level oscillation (noise).
- In such a situation, a linear regression model cannot capture the internal dynamics of the data.
- In such situations we can make use of Isotonic Regression which produces piecewise interpolating function minimizing the functional

$$f(x) = \sum_i w_i \left(y_i - \hat{y}_i\right)^2$$

where $y_0 \leq y_1 \leq y_2 \leq \ldots\ldots \leq y_n$

Example :



3.9 Linear Classification

- Linear classification is a classification algorithm that makes its classification based on a Linear predictor function, combining a set of weights with feature vector.

$$y = f\left(\sum_j w_i x_j\right)$$

- In this classifier the decision boundary is flat (e.g. Line, plane).

Let us consider the task of Binary Classification. The goal is to predict a binary value target.

- Example includes :
  - A medical diagnosis system to predict whether a patient is suffering from a given disease.
  - Whether a given email is spam or not spam.
  - Whether a given transaction is fraudulent or not fraudulent.
- A binary classifier computes a linear function of inputs, and determine whether or not the value is larger than some threshold th.
- A linear function of the input can be written as

$$w_1 x_1 + \ldots + w_D x_D + b = w^T x + b$$

where w is a weight vector and b is a scalar – valued bias

- The prediction y can be computed as

$$z = w^T x + b$$

$$y = \begin{cases} 1 & \text{if } z \geq th \\ 0 & \text{if } z < th \end{cases}$$

- Example of Linear Classifier

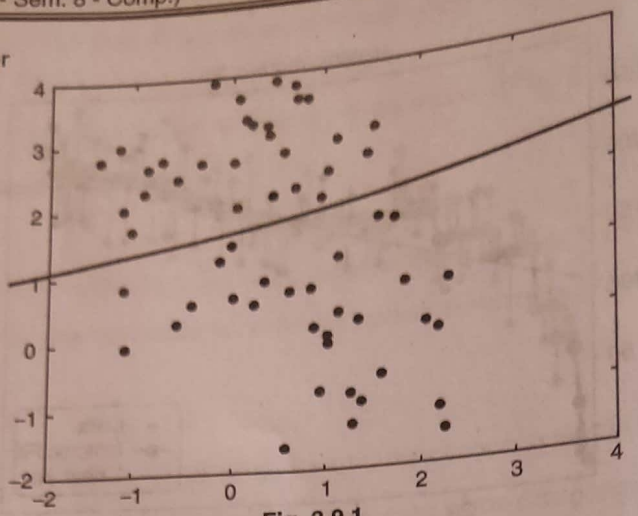


Fig. 3.9.1

3.10 Logistic Regression

- **Logistic Regression** is a statistical technique for analyzing a dataset that predicts the probability of an outcome that can only have two values (i.e. a dichotomy).
- The goal is to find the best fitting model to describe relationship between the dichotomous characteristics of interest (dependent variable) and a set of independent variable(predictor or explanatory variable)
- Logistic regression seeks to
  - Model the probability of an event occurring depending on the values of the independent variables , which can be categorical or numerical.
  - Estimate the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur.
  - Predict the effect of a series of variables on a binary response variable.
  - Classify observations by estimating the probability that an observation is in particular category (such as approved or not approved)
- Linear Regression method has some major problem
  - Binary data does not have a normal distribution, which is a condition needed for most other types of regression.
  - Predicted values of the dependent variable can be beyond 0 and 1, which violates the definition of probability.
- In logistic regression the predicted value has to be between 0 and 1.

Probability Review for Logistic Regression

(I) Probability

- The probability is given as

$$p = \frac{\text{Outcomes of Interest}}{\text{all possible outcomes}}$$

- For example Probability of getting a head from a Fair coin Flip is given as

$$p(heads) = \frac{1}{2} = 0.5$$

- For example Probability of pulling out a diamond from a Deck of playing cards

$$p(\text{diamond}) = \frac{13}{52} = \frac{1}{4} = 0.25$$

(II) **Odds**

- Odds is the probability of something occurring divided by the probability of not occurring and is given as

$$\text{odds} = \frac{P(\text{occuring})}{P(\text{not occuring})}$$

$$\text{odds} = \frac{p}{1-p}$$

- For example the odds of getting a head for flipping a fair coin

$$\text{odds(heads)} = \frac{0.5}{0.5} = 1$$

- For example the odds of pulling out a diamond from a Deck of playing cards

$$\text{odds(heads)} = \frac{0.25}{0.75} = 0.333$$

Odds Ratio

- Odds ratio is ratio of two odds and is given as

$$\text{odds(heads)} = \frac{\text{odds}_1}{\text{odds}_0}$$

Note : Odds Ratio in Logistic Regression represents how the odds change with a 1 unit increase in that variable holding all other variables constant

- The dependent variable is binary its either 0 or 1. We need to link the probability between 0 and 1 back to our independent variables.
- The dependent variable in Logistic Regression follows the Bernoulli distribution having an unknown probability p.
- Bernoulli distribution is a special case of the Binomial distribution

  Where, n = 1 (just one trial) ; Success is 1 and failure is 0
- So the probability of success is p and failure is q = 1 – p.
- In logistic regression we are estimating an unknown p for any given linear combination of the independent variables.
- So we need to link the independent variables to Bernoulli distribution, that link is called as **Logit**.
- In Logistic Regression we do not know the value of p. So the Goal of Logistic Regression is to estimate p for a linear combination of the independent variables. Estimate of p is $\hat{p}$ (p-hat)
- The natural log of the odds ratio, the Logit is that Link function that will map the Linear combination of variables to Bernoulli distribution that could result in a value between 0 and 1. This is given by

$$ln(\text{odds}) = ln\left(\frac{p}{1-p}\right) \text{ is the } logit\,(p) \quad \textbf{OR} \quad ln(p) - ln(1-p) = logit(p)$$

Note : $\log_e(x) = ln(x)$

- This can be graphically represented as shown in Fig. 3.10.1

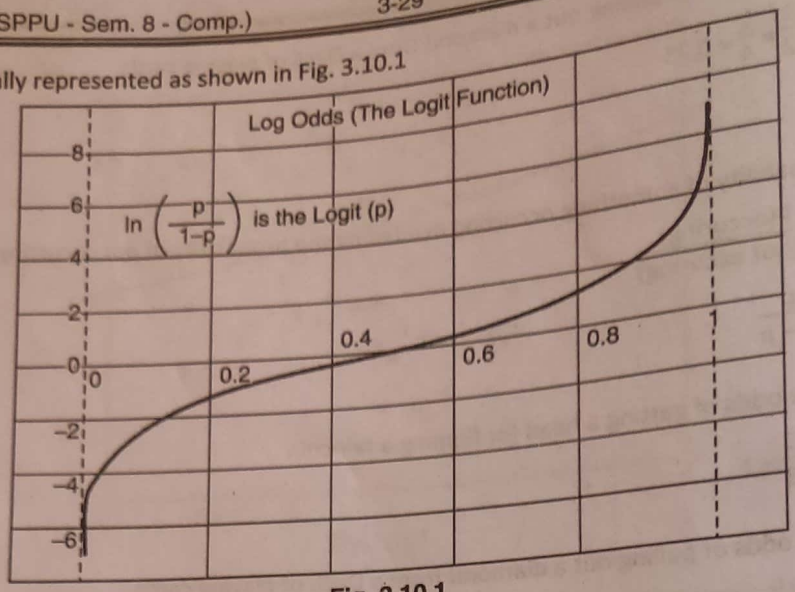


Fig. 3.10.1

When p = 0, *ln*(0) is undefined

When p = 1, is again going to undefined

- But when p = 0.5, *ln*(1) = 0, that is what we can see graphically this shows a sigmoid curve.

Inverse Logit

- In the above graphical representation our probabilities are shown on x-axis , we want them on y-axis . we can get this by taking the inverse of the logit function
- We know that

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

where p is between 0 and 1

when we take the inverse of logit(p) we get

$$logit^{-1}(p) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}} \qquad \alpha = \text{some number}$$

- We can represent this graphically as shown in Fig. 3.10.2

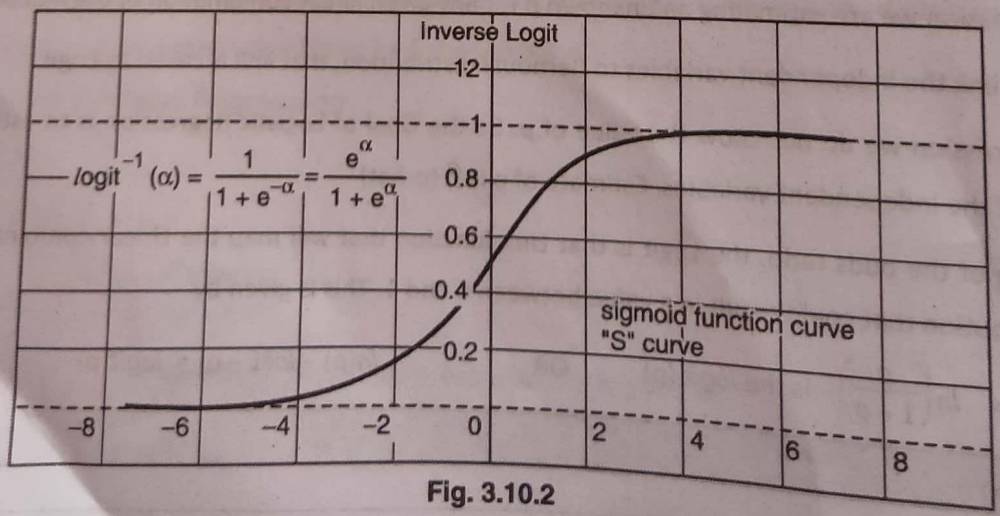


Fig. 3.10.2

The Estimated Regression equation

- The antilog of the logit function allows us to find the estimated regression equation.
- We know that

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

antilog will give us

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1}$$

$$p = e^{\beta_0+\beta_1 x_1}(1-p) = e^{\beta_0+\beta_1 x_1} - e^{\beta_0+\beta_1 x_1} * p$$

$$p + e^{\beta_0+\beta_1 x_1} * p = e^{\beta_0+\beta_1 x_1}$$

$$p(1 + e^{\beta_0+\beta_1 x_1}) = e^{\beta_0+\beta_1 x_1}$$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{(1 + e^{\beta_0+\beta_1 x_1})}$$

This is the estimated regression equation.

3.11 Implementation and Optimizations

- An example on Logistic Regression using scikit learn Library in python.
- You can download the data set from https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Step 1 : Loading the dataset

```
#import pandas
import pandas as pd
col_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']
# load dataset
pima = pd.read_csv("pima-indians-diabetes.csv", header=None, names=col_names)
print(pima.head())
```

Step 2 : Selecting Features

Divide the data into dependent and independent variables

```
#split dataset in features and target variable
feature_cols = ['pregnant', 'insulin', 'bmi', 'age','glucose','bp','pedigree']
X = pima[feature_cols] # Features
y = pima.label # Target variable
```

Step 3 : Splitting the data

```
# split X and y into training and testing sets
from sklearn.cross_validation import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)
```

Step 4 : Fit your model

```
# import the class
from sklearn.linear_model import LogisticRegression
# instantiate the model (using the default parameters)
logreg = LogisticRegression()
# fit the model with data
logreg.fit(X_train,y_train)
y_pred=logreg.predict(X_test)
```

Step 5 : Model Evaluation

```
# import the metrics class
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
print(cnf_matrix)
```

Step 6 : Confusion Matrix Evaluation Metrics

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

Output

```
Accuracy: 0.8072
Precision: 0.7659
Recall: 0.58064
```

3.12 Stochastic Gradient Descent Algorithms

3.12.1 Introduction to Gradient Descent

- Optimization is the task of minimizing / maximizing an objective function f(x) parameterized by x
- In machine learning it refers to the task of minimizing the cost/loss function, parameterized by model's parameters.
- The goal of optimization algorithms is to have one of the following
  - Finding the global minimum of the convex objective function
  - Finding the lowest possible value of the non convex objective function within its neighborhood
- Gradient descent is the most common optimization algorithm in machine learning
- Machine Learning models have parameters (e.g. weights and biases) and a cost function to evaluate the goodness of particular set of parameters.
- During training the goal is to find the predicted values close to the target values so that cost J(W) is minimum.

Gradient descent is an iterative method used to minimize the cost function J(W) parameterized by a model parameters W.

- The gradient or derivative tells the slope/cost of the function, In order to reduce it, an opposite direction is chosen.
- Let us consider for e.g. a logistic regression model having two parameters weight w and bias b

Step 1 : Initialize the weight w and bias b to random values.

Step 2 : Pick a value for the learning rate α.

- Learning Rate determines the step size for each iteration :
  - If α is small, it would take longer to converge and also it would be computationally expensive.
  - If α is large, it may fail to converge.

Step 3 :

On each iteration take the partial derivate of the cost function J(W) with respect to each parameter e.g. in this case weight and bias

$$\frac{\partial}{\partial w} J(w) = \nabla_w J \quad \text{and} \quad \frac{\partial}{\partial b} J(w) = \nabla_b J$$

Step 4 : Update the Equations

$$w = w - \alpha \nabla_w J$$

$$b = b - \alpha \nabla_b J$$

- Let us ignore say bias, if the slope of the current value of $w > 0$, this means we are to the right of the optimal w^*. Therefore update will be negative and start going close to optimal value.
- However if the value of $w < 0$ then the update will be positive and this will increase the value of w and converge to the optimal value of w^*. This is shown in the Fig. 3.12.1.

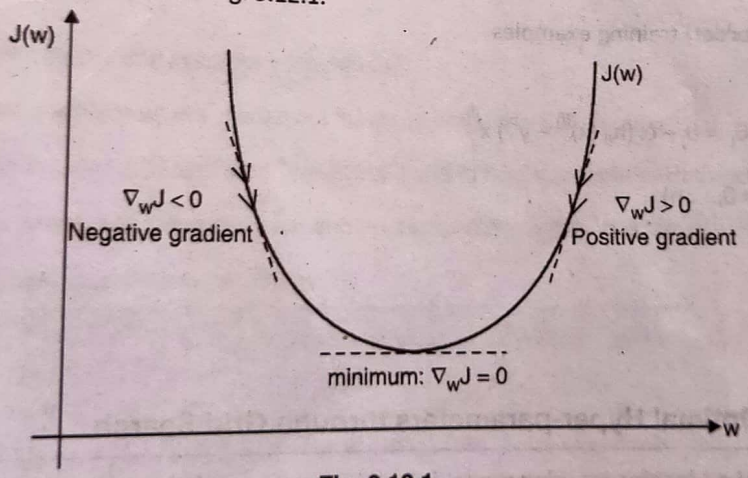


Fig. 3.12.1

Step 5 : Continue the process until convergence

Example of Linear Regression with Gradient descent

$$h_\theta(x) = \sum_{j=0}^{n} \theta_j x_j$$

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Repeat {

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(for every j = 0,....,n)

where h(θ) is the hypothesis function.

J_{train} is the cost function and

Repeat {} block is to update the θ parameter.

3.12.2 Stochastic Gradient Descent

- When there is a large training data set, Gradient descent becomes computationally expensive.
- In such cases a modification to gradient descent algorithm called as stochastic gradient descent can be used.
- In Stochastic gradient descent unlike gradient descent, all of training data is not used, only a single example is used.
- Example of Linear Regression with Stochastic Gradient descent is given below :

$$cost(\theta, x^{(i)}, y^{(i)}) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{train}(\theta) = \frac{1}{m} \sum_{i=1}^{m} cost(\theta, (x^{(i)}, y^{(i)}))$$

Steps in Stochastic Gradient descent

1. Randomly shuffle (reorder) training examples
2. Repeat {

   for i : = 1,....,m { $\theta_j := \theta_j - \alpha (h_\theta(x)^{(i)} - y^{(i)}) x_j^{(i)}$

   (for every j = 0,....,n)

   }

   }

3.13 Finding the Optimal Hyper-parameters through Grid Search

- To tune the behavior of a Machine learning model for a given problem, these models are parameterized
- The Machine learning models may have many parameters, a best combination of these parameters is considered to be a search problem and is very important for better performance of a machine learning model.
- A parameter that is internal to the model and whose value can be estimated from the given data is called as ***model parameter.***
- On the other hand, a parameter that is external to the model and whose value cannot be estimated from data is called as **Model Hyper Parameter.**

- Example of Hyper parameter includes :
  - K in K-nearest Neighbor
  - Learning rate for training a Neural Network
  - The C and sigma parameter for Support vector machines
- Grid Search is a simple strategy for optimizing/tuning the hyper parameters.
- It is an approach that will build and evaluate a model for each combination of parameters specified in the grid.
- For example Consider a Machine Learning Model X takes hyper parameters a_1 and a_2.
- In Grid Search, first define the range of values for each of the hyper parameter.
- The technique will construct many versions of model X with all possible combinations of Hyper parameter.
- For example let the grid values be

  $a_1 = [1, 5, 6]$ and $a_2 = [10, 20, 30]$.
- The grid search will start with a combination of 1 and 10 and go through all intermediate combinations between these values and end at 6 and 30.

3.14 Classification Metric

- Evaluating machine learning model is essential.
- Different performance metrics used to evaluate classification algorithms are discuss as follows.

Confusion Matrix

- It is used to find the correctness and accuracy of the model.
- It is used in classification problem where the output is two or more types of classes.
- It is a table with two dimensions ("Actual" and "Predicted") and sets of classes in both dimensions.
- The Actual dimension is presented along the rows and Predicted dimension along the columns.
- **Example :** Consider a Binary classification problem

	Predicted class C_1 (YES)	**Predicted class $\neg C_1$ (NO)**
Actual class C_1 (YES)	True Positives (TP)	False Negatives (FN)
Actual class $\neg C_1$ (NO)	False Positives (FP)	True Negatives (TN)

The terms in the confusion matrix are :

1. **True Positives (TP)** : The cases in which the prediction was YES and the actual output was also YES.
2. **True Negatives (TN)** : The cases in which the prediction was NO and the actual output was NO.
3. **False Positives (FP)** : The cases in which the prediction was YES and the actual output was NO.
4. **False Negatives (FN)** : The cases in which the prediction was NO and the actual output was YES.

Accuracy

- Accuracy is the percentage of test tuples that are correctly classified.
- It is calculated as follows :

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total number of Samples}}$$

Error Rate

The Error rate may be calculated as follows :

Error Rate= 1 – Accuracy **OR** $$\text{Error rate} = \frac{\text{False Positive + False Negative}}{\text{Total number of Samples}}$$

Sensitivity

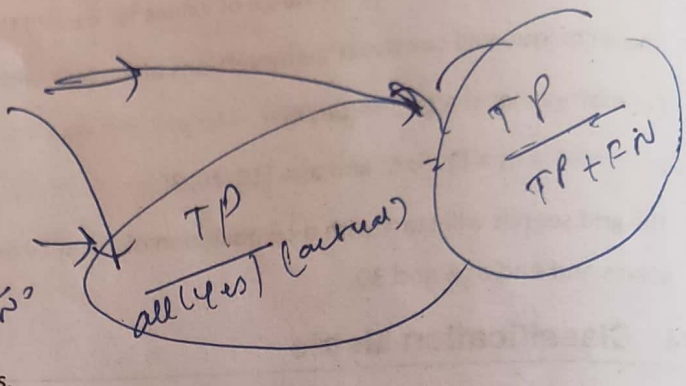
True positive recognition rate

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{Total Number of Positive samples}}$$

Specificity

True positive recognition rate

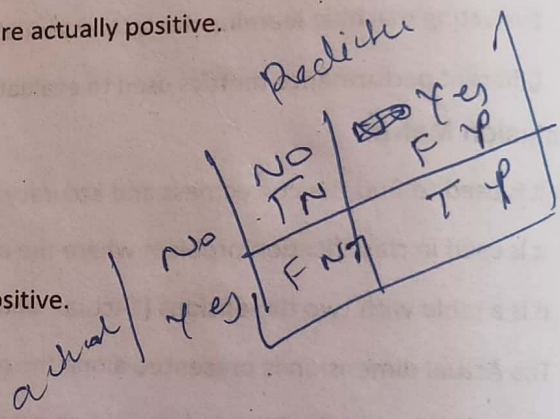
$$\text{Specificity} = \frac{\text{True Negative}}{\text{Total Number of Negative samples}}$$



Precision

- It is the exactness that the classification algorithm gives.
- It is the percentage of tuples that the classifier labeled as positive are actually positive.
- The formula to calculate Precision is as follows :

$$\text{Precision} = \frac{\text{True Negative}}{\text{True Positive + False positive}}$$



Recall

- Recall is also the completeness.
- It is the percentage of positive tuples did the classifier labeled as positive.
- The formula to calculate Recall is as follows :

$$\text{Recall} = \frac{\text{True Negative}}{\text{True Positive + False positive}}$$

F-measure (F_1 or F-Score)

It is the harmonic mean of precision and recall.

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision + recall}}$$

Example for Classification Metric

Actual Class\Predicted class	cancer = yes	Cancer = no	Total	Recognition(%)
cancer = yes	90	210	300	30.00 (sensitivity)
cancer = no	140	9560	9700	98.56 (specificity)
Total	230	9770	10000	96.40 (accuracy)

$$\text{Precision} = \frac{90}{230} = 39.13\% \quad \text{and} \quad \text{Recall} = \frac{90}{300} = 30.00\%$$

3.15 ROC Curve

- ROC curve stands for Receiver Operating Characteristics Curve.
- Using ROC curve one can visually compare classification models.
- ROC curve has its originating roots from signal detection theory.
- A trade-off between the true positive rate and the false positive rate is shown on ROC curve.
- The accuracy of the model can be measured by the area under the ROC curve.
- Vertical axis represents the true positive rate and horizontal axis represents the false positive rate.
- The model with perfect accuracy will have an area of 1.0.

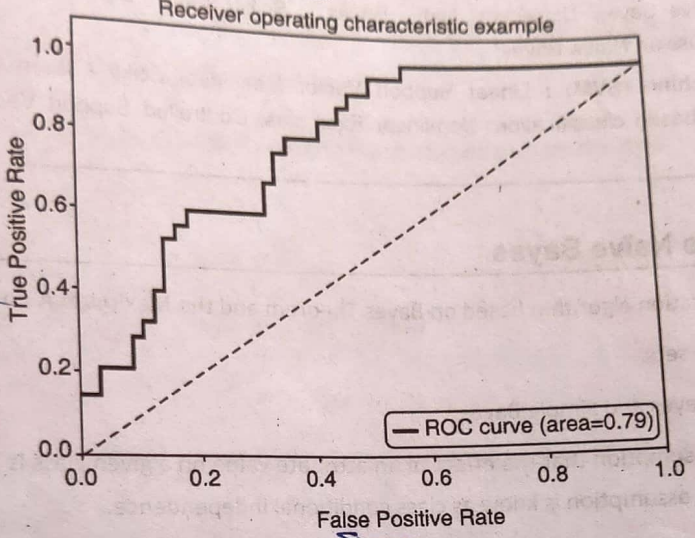


Fig. 3.15.1

Review Questions

Q. 1 Explain different steps in conducting regression analysis.

Q. 2 Explain different methods for calculating correlation coefficient.

Q. 3 Explain Linear regression using Least squares method, support your answer with a suitable example.

Q. 4 Explain the different regularization techniques used with Linear regression to handle multivariate data.

Q. 5 Write short notes on :

(i) Ridge regression (ii) Lasso Regression

(iii) ElasticNet regression (iv) Robust regression with random sample consensus

(v) Polynomial regression (vi) Isotonic regression

Q. 6 Explain Logistic Regression.

Q. 7 Explain Stochastic Gradient descent technique.

Q. 8 Explain different classification metrics.