UNIT 4:

SUPERVISED LEARNING: REGRESSION

UNIT – IV

Supervised Learning: Regression

Introduction, Example of Regression, Common Regression Algorithms-Simple linear regression, Multiple linear regression, Polynomial Regression Model, Logistic Regression,

---

Regression

- ✓ Regression focuses on solving problems such as predicting value of real estate, demand forecast in retail, weather forecast, etc.
- ✓ Regression is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variable(s) (X), i.e. to find the function ‘f ’ for the association Y = f (X).

Example of Regression – Real estate price prediction

New City is the primary hub of the commercial activities in the country. In the last couple of decades, with increasing globalization, commercial activities have intensified in New City.

Together with that, a large number of people have come and settled in the city with a dream to achieve professional growth in their lives. As an obvious fall-out, a large number of housing projects have started in every nook and corner of the city. But the demand for apartments has still outgrown the supply.

To get benefit from this boom in real estate business, Karen has started a digital market agency for buying and selling real estates (including apartments, independent houses, town houses, etc.). Initially, when the business was small, she used to interact with buyers and sellers personally and help them arrive at a price quote — either for selling a property (for a seller) or for buying a property (for a buyer). Her long experience in real estate business helped her develop an intuition on what the correct price quote of a property could be — given the value of certain standard parameters such as area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc.

However, with the huge surge in the business, she is facing a big challenge. She is not able to manage personal interactions as well as setting the correct price quote for the properties all alone. She hired an assistant for managing customer interactions. But the assistant, being new in the real estate business, is struggling with price quotations. How can Karen solve this problem?

Fortunately, Karen has a friend, Frank, who is a data scientist with in-depth knowledge in machine learning models. Frank comes up with a solution to Karen’s problem. He builds a model which can predict the correct value of a real estate if it has certain standard inputs such as area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc. Wow, that sounds to be like Karen herself doing the job! Curious to know what model Frank has used? Yes, you guessed it right. He used a regression model to solve Karen's real estate price prediction problem.

COMMON REGRESSION ALGORITHMS

The most common regression algorithms are:

1) Simple linear regression
2) Multiple linear regression
3) Polynomial regression
4) Multivariate adaptive regression splines
5) Logistic regression
6) Maximum likelihood estimation (least squares)

Simple linear regression:-

As the name indicates, simple linear regression is the simplest regression model which involves only one predictor. This model assumes a linear relationship between the dependent variable and the predictor variable as shown in Figure below.

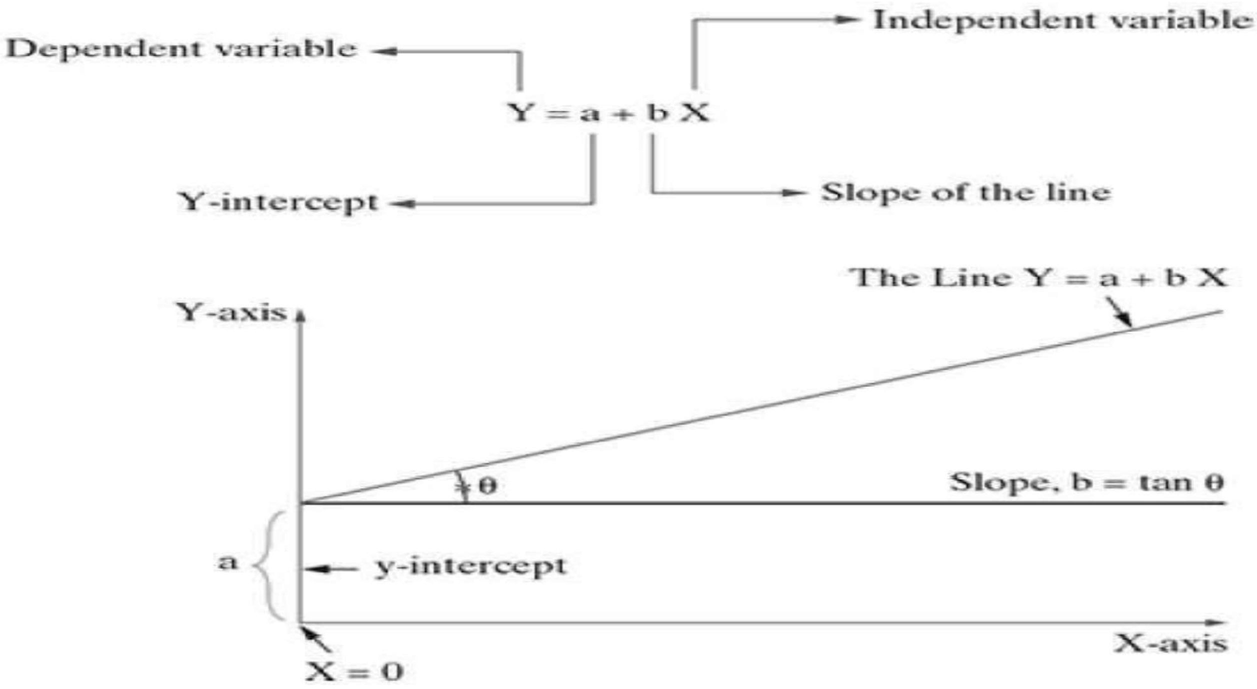
In the context of Karen's problem, if we take Price of a Property as the dependent variable and the Area of the Property (in sq. m.) as the predictor variable, we can build a model using simple linear regression.

$$\text{Price}_{\text{Property}} = f(\text{Area}_{\text{Property}})$$

Assuming a linear association, we can reformulate the model as –

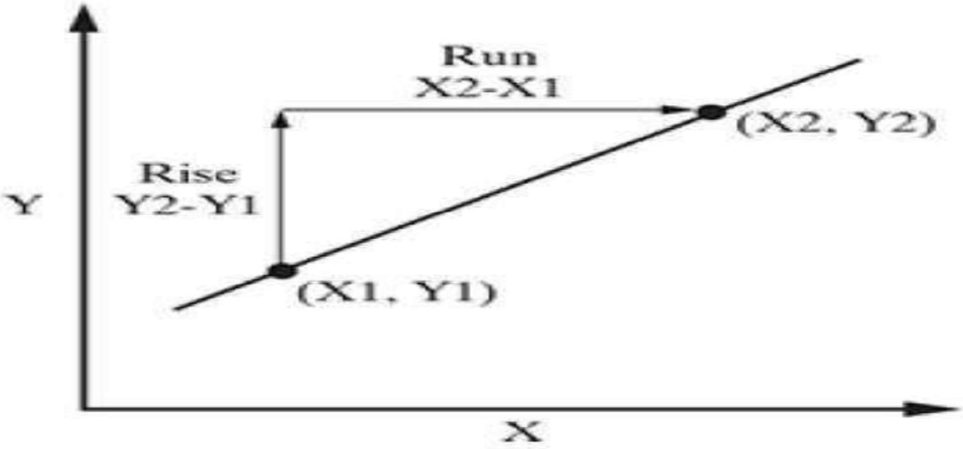
$$\text{Price}_{\text{Property}} = a + b.\,\text{Area}_{\text{Property}}$$

where 'a' and 'b' are intercept and slope of the straight line, respectively.



Slope of the simple linear regression model

Slope of a straight line represents how much the line in a graph changes in the vertical direction (Y-axis) over a change in the horizontal direction **(X-axis).**



Rise and run representation

Slope = Change in Y/Change in X

$$\text{Slope} = \frac{\text{Rise}}{\text{Run}} = \frac{Y2 - Y1}{X2 - X1}$$

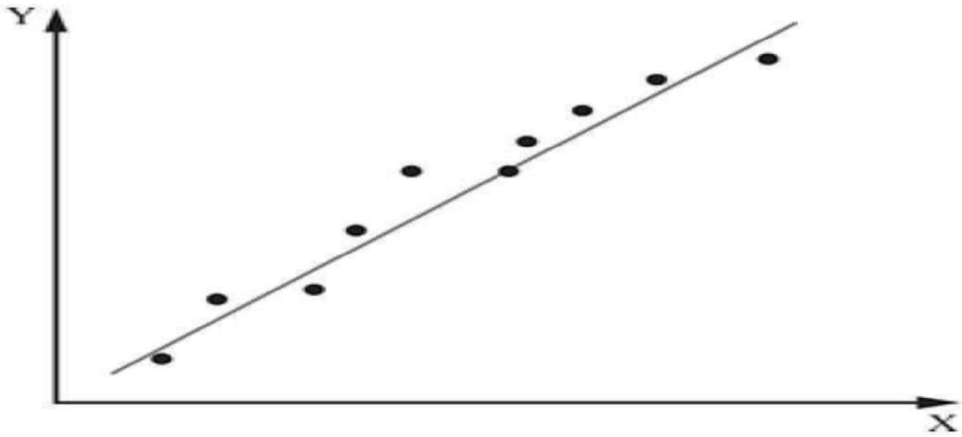
Types of slopes in a Linear Regression

There can be two types of slopes in a linear regression model: positive slope and negative slope. Different types of regression lines based on the type of slope include

1) Linear positive slope
2) Curve linear positive slope
3) Linear negative slope
4) Curve linear negative slope

1) Linear Positive Slope

A positive slope always moves upward on a graph from left to right

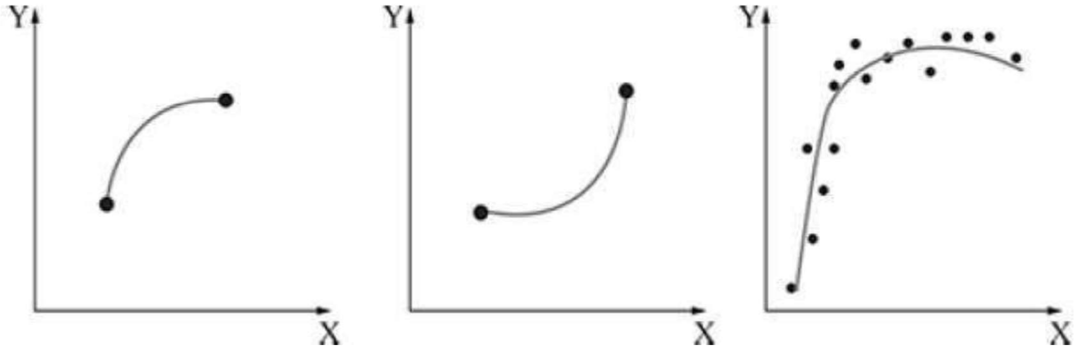


Slope = Rise/Run = $(Y_2 - Y_1) / (X_2 - X_1)$ = Delta (Y) / Delta(X)

2) Curve Linear Positive Slope

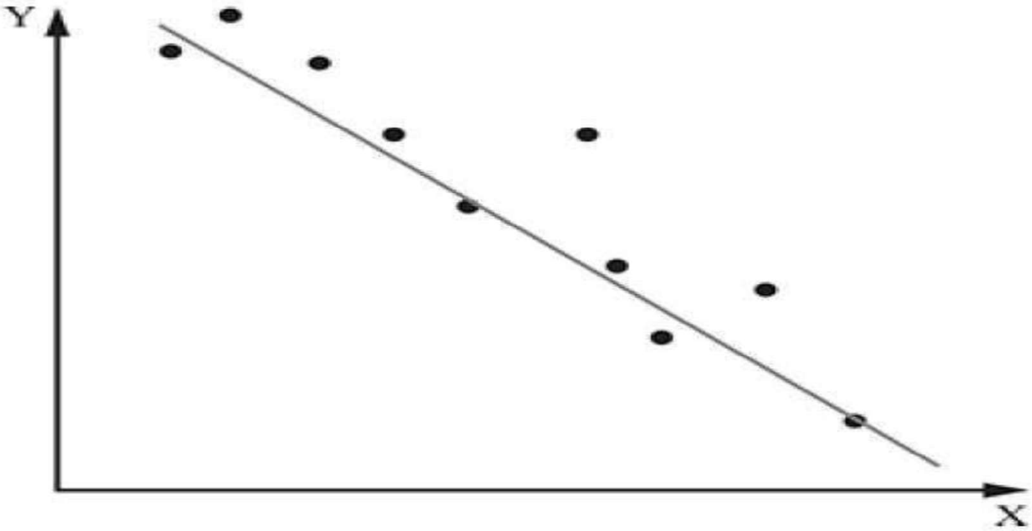
Curves in these graphs slope upward from left to right.

Curve linear positive slope



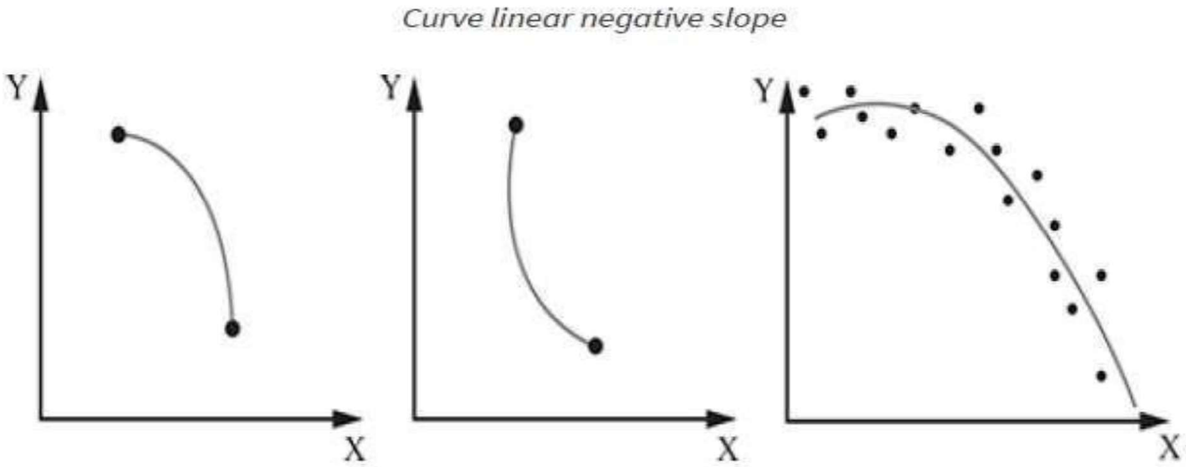
3) Linear Negative Slope

A negative slope always moves downward on a graph from left to right. As X value (on X-axis) increases, Y value decreases



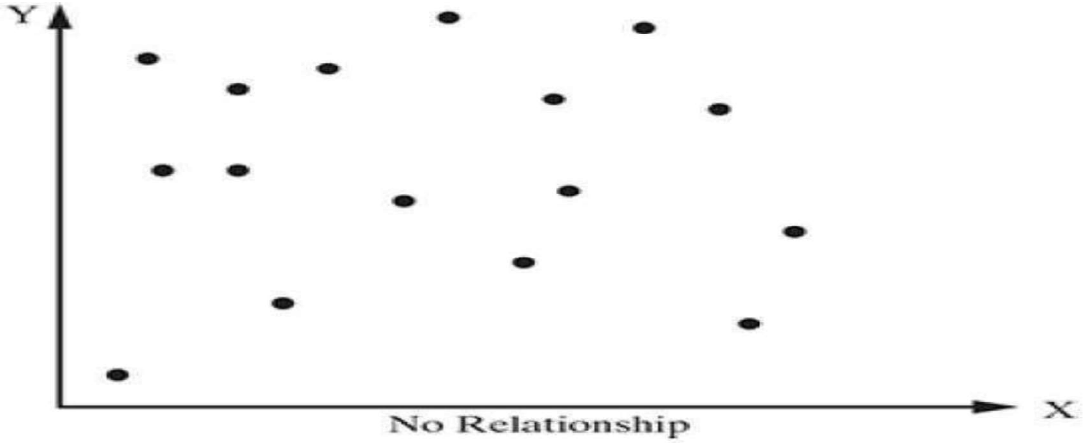
4) Curve Linear Negative Slope

Curves in these graphs slope downward from left to right.



No relationship graph

Scatter graph shown in below Figure indicates 'no relationship' curve as it is very difficult to conclude whether the relationship between X and Y is positive or negative.

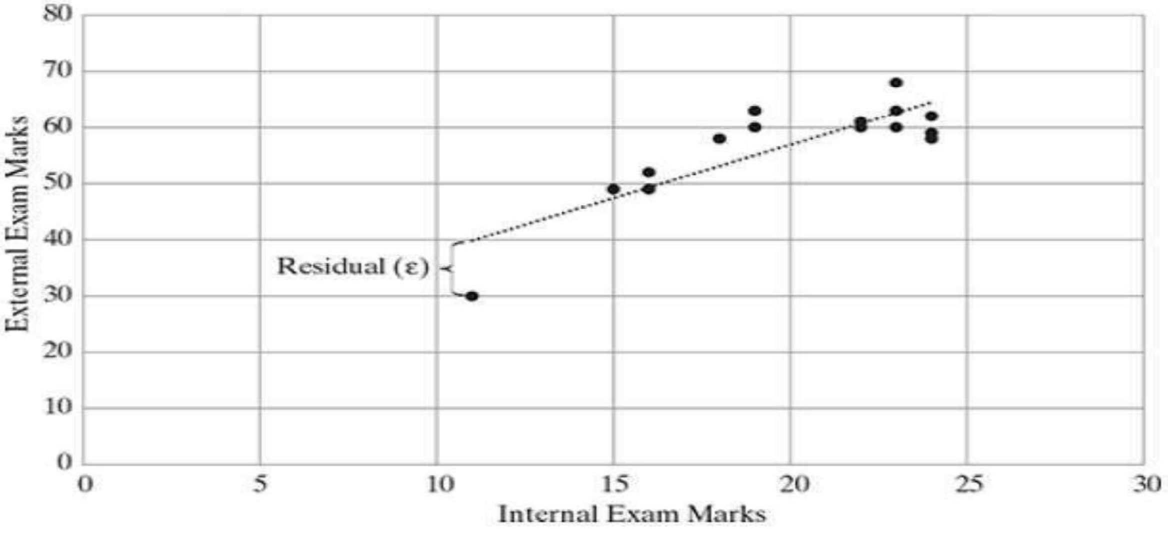


Error in simple regression (Marginal or Residual error)

- ✓ The regression equation model in machine learning uses the above slope–intercept format in algorithms. X and Y values are provided to the machine, and it identifies the values of a (intercept) and b (slope) by relating the values of X and Y.
- ✓ However, identifying the exact match of values for a and b is not always possible. There will be some error value (ε) associated with it. This error is called marginal or residual error.
- ✓ Residual is the distance between the predicted point (on the regression line) and the actual point

$$Y = (a + bX) + \varepsilon$$

Example for Residual error



Ordinary Linear Squares (OLS) Algorithm

Ordinary Least Squares (OLS) is the technique used to

- build a simple linear regression model for a given problem
- estimate a line that will minimize the error (ε)

In Y = a + bX, b value can be calculated with the below formula so that Sum of the Squares of the Errors is least.

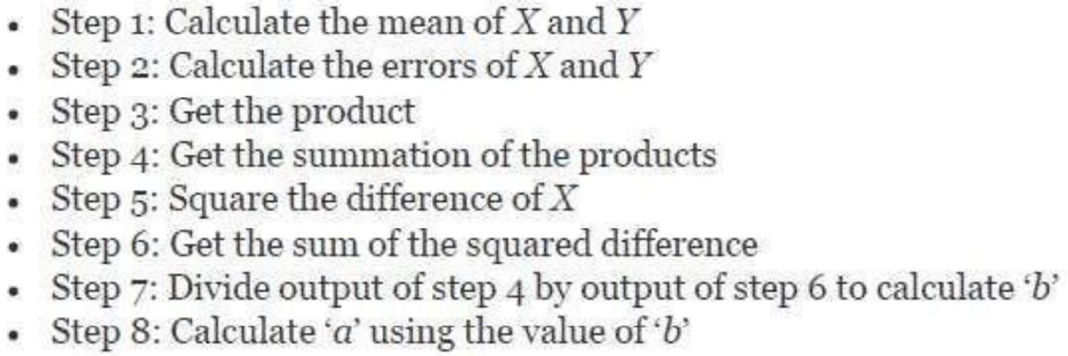
$$b = \frac{\Sigma_i (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma_i (X_i - \overline{X})^2} = \frac{\mathrm{cov}(X,\ Y)}{\mathrm{Var}(X)}$$

The corresponding value of ‘a’ calculated using the above value of ‘b’ is

$$a = \overline{\mathrm{Y}} - b\overline{\mathrm{X}}$$

Where $\overline{\mathrm{Y}}$ is the mean of Y and $\overline{\mathrm{X}}$ is the mean of X,

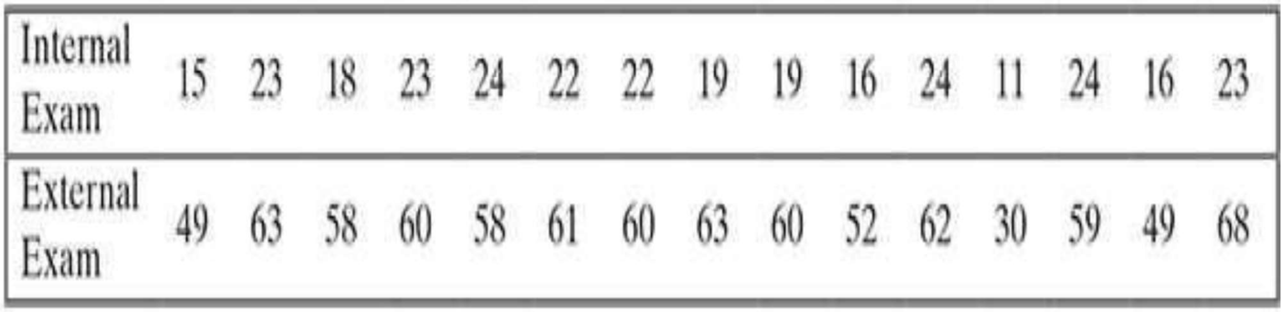
Steps in OLS algorithm



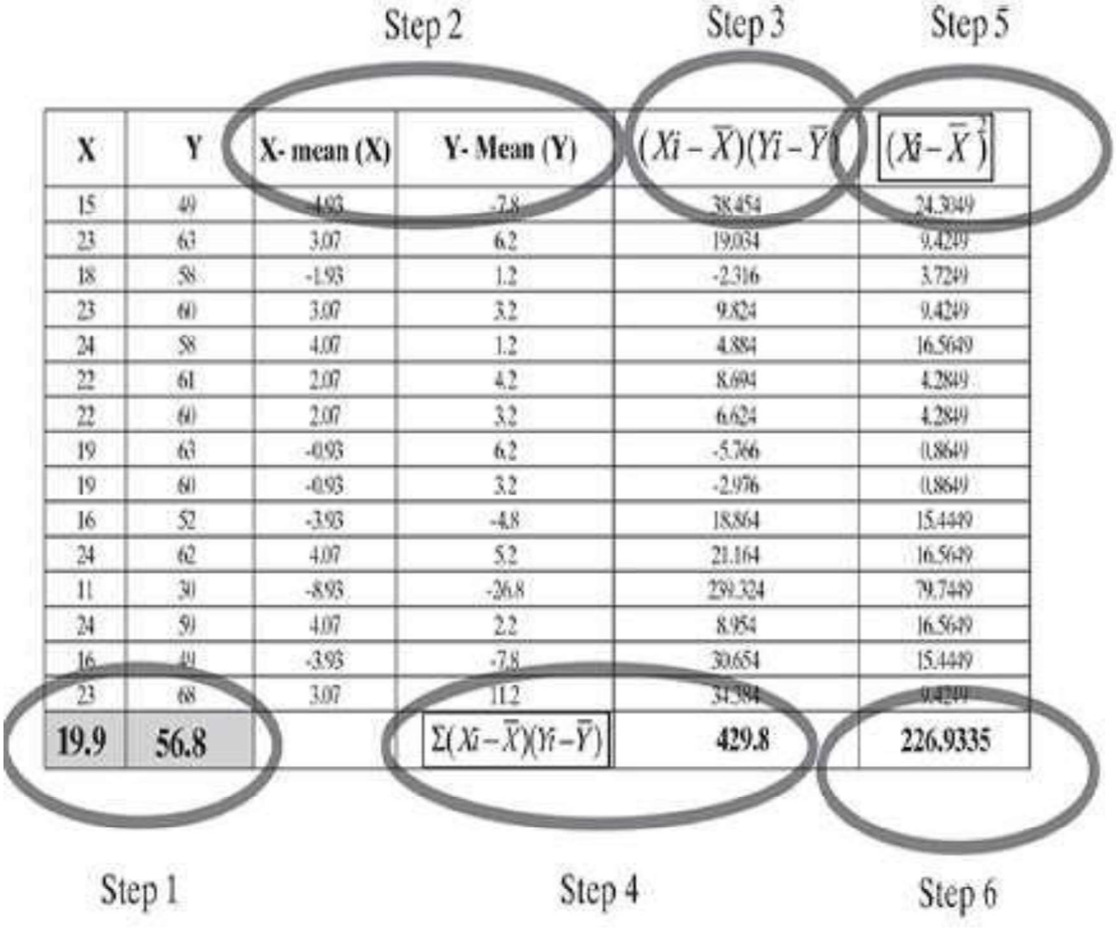
- Step 1: Calculate the mean of *X* and *Y*
- Step 2: Calculate the errors of *X* and *Y*
- Step 3: Get the product
- Step 4: Get the summation of the products
- Step 5: Square the difference of *X*
- Step 6: Get the sum of the squared difference
- Step 7: Divide output of step 4 by output of step 6 to calculate ‘*b*’
- Step 8: Calculate ‘*a*’ using the value of ‘*b*’

Example

A college professor believes that if the grade for internal examination is high in a class, the grade for external examination will also be high. A random sample of 15 students in that class was selected, and the data is given below:



Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68



Step 2

Step 3

Step 5

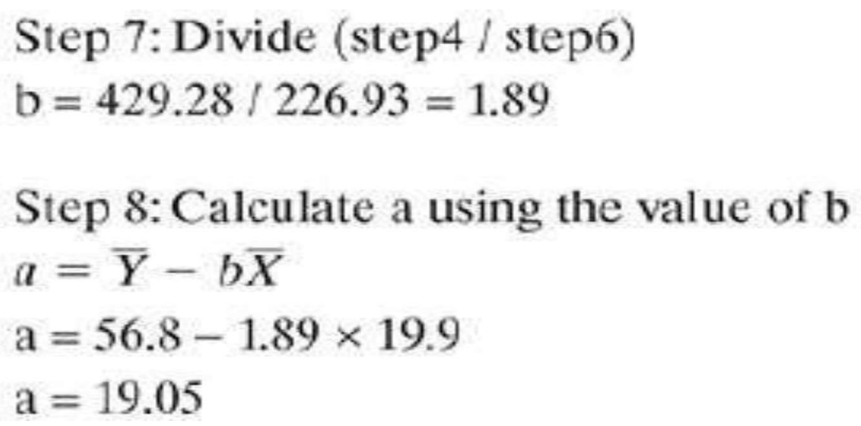
X	Y	X- mean (X)	Y- Mean (Y)	$(Xi-\overline{X})(Yi-\overline{Y})$	$(Xi-\overline{X})^2$
15	49	-4.93	-7.8	38.454	24.3049
23	63	3.07	6.2	19.034	9.4249
18	58	-1.93	1.2	-2.316	3.7249
23	60	3.07	3.2	9.824	9.4249
24	58	4.07	1.2	4.884	16.5649
22	61	2.07	4.2	8.694	4.2849
22	60	2.07	3.2	6.624	4.2849
19	63	-0.93	6.2	-5.766	0.8649
19	60	-0.93	3.2	-2.976	0.8649
16	52	-3.93	-4.8	18.864	15.4449
24	62	4.07	5.2	21.164	16.5649
11	30	-8.93	-26.8	239.324	79.7449
24	59	4.07	2.2	8.954	16.5649
16	49	-3.93	-7.8	30.654	15.4449
23	68	3.07	11.2	34.384	9.4249
19.9	**56.8**		$\Sigma(Xi-\overline{X})(Yi-\overline{Y})$	**429.8**	**226.9335**

Step 1

Step 4

Step 6

<u>The Calculation summary is</u>



Step 7: Divide (step4 / step6)

b = 429.28 / 226.93 = 1.89

Step 8: Calculate a using the value of b

$a = \overline{Y} - b\overline{X}$

a = 56.8 – 1.89 × 19.9

a = 19.05

Sum of $X = 299$

Sum of $Y = 852$

Mean X, $M_X = 19.93$

Mean Y, $M_Y = 56.8$

Sum of squares $(SS_X) = 226.9333$

Sum of products $(SP) = 429.8$

Regression equation $= \hat{y} = bX + a$

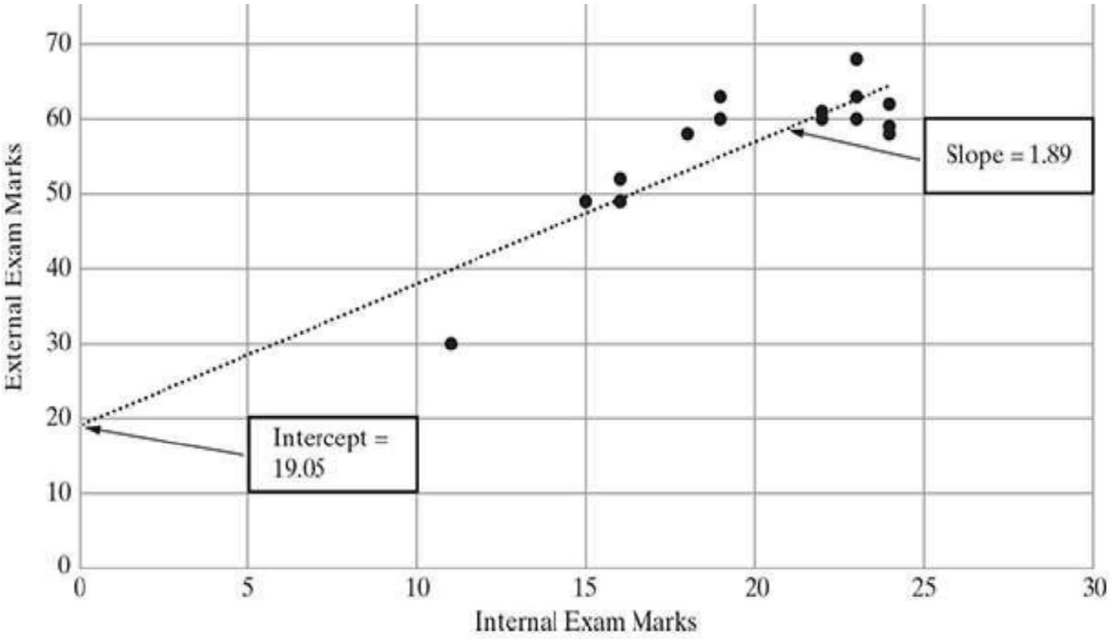
$$b = \frac{SP}{SS_X} = \frac{429.8}{226.93} = 1.89395$$

$a = M_Y - bM_X = 56.8 - (1.89 \times 19.93) = 19.0473$

$\hat{y} = 1.89395X + 19.0473$

Linear Regression model is

$$M_{\mathrm{Ext}} = \mathbf{19.04 + 1.89} \times M_{\mathrm{Int}}$$



Multiple Linear Regression

In a multiple regression model, two or more independent variables, i.e. predictors are involved in the model.

For Example, we consider Price of a Property (in $) as the dependent variable and Area of the Property (in sq. m.), location, floor, number of years since purchase and amenities available as the independent variables, we can form a multiple regression equation as shown below:

$$\text{Price}_{\text{Property}} = f\,(\text{Area}_{\text{Property},}\ \text{location, floor, Ageing, Amenities})$$

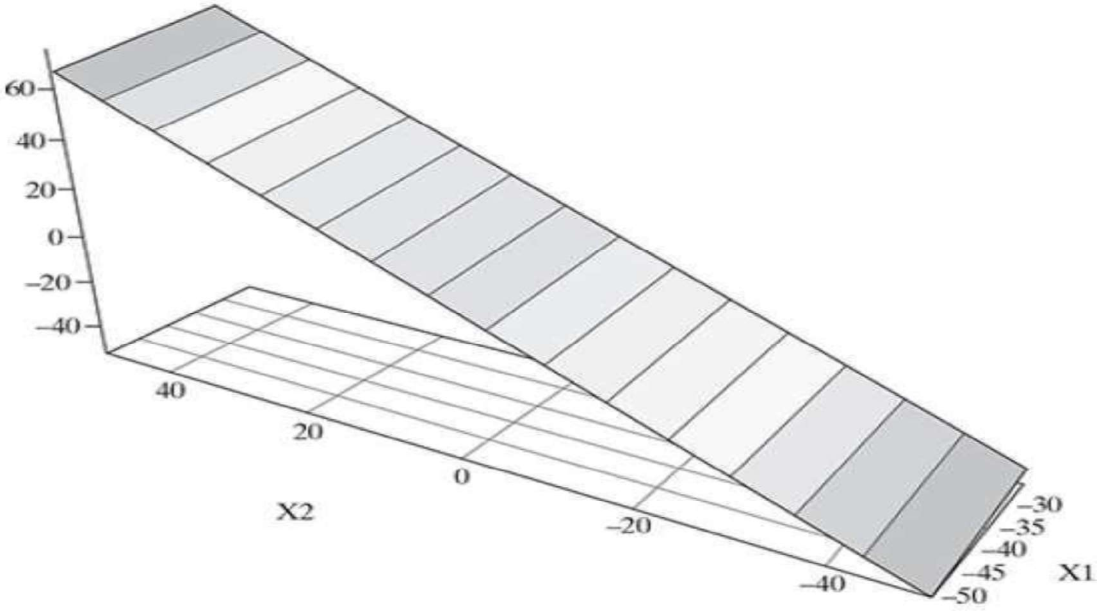
The following expression describes the equation involving the relationship with two predictor variables, namely X1 and X2

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

✓ The model describes a plane in the three-dimensional space of Ŷ, X1, and X2. Parameter ‘a’ is the intercept of this plane. Parameters ‘b1’ and ‘b2’ are referred to as partial regression coefficients.

✓ Parameter b1 represents the change in the mean response corresponding to a unit change in X1 when X2 is held constant.

✓ Parameter b_2 represents the change in the mean response corresponding to a unit change in X_2 when X_1 is held constant.



$$\hat{Y} = 22 + 0.3X_1 + 1.2X_2$$

Above figure shows the sample graph for

Assumptions in Regression Analysis

1. The dependent variable (Y) can be calculated / predicated as a linear function of a specific set of independent variables (X's) plus an error term (ε).
2. The number of observations (n) is greater than the number of parameters (k) to be estimated,

i.e. n > k.

3. Relationships determined by regression are only relationships of association based on the data set and not necessarily of cause and effect of the defined class.
4. Regression line can be valid only over a limited range of data. If the line is extended (outside the range of extrapolation), it may only lead to wrong predictions.
5. If the business conditions change and the business assumptions underlying the regression model are no longer valid, then the past data set will no longer be able to predict future trends.
6. Variance is the same for all values of X (**homoskedasticity**).
7. The error term (ε) is normally distributed. This also means that the mean of the error (ε) has an expected value of 0.
8. The values of the error (ε) are independent and are not related to any values of X. This means that there are no relationships between a particular X, Y that are related to another specific value of X, Y.

Main Problems in Regression Analysis

In multiple regressions, there are two primary problems: **multicollinearity** and heteroskedasticity.

1) Multicollinearity

- ✓ Two variables are perfectly collinear if there is an exact linear relationship between them.
- ✓ Multi collinearity is the situation in which the degree of correlation is not only between the dependent variable and the independent variable, but there is also a strong correlation within (among) the independent variables themselves.
- ✓ A multiple regression equation can make good predictions when there is multicollinearity.

2) Heteroskedasticity

Heteroskedasticity refers to the changing variance of the error term. If the variance of the error term is not constant across data sets, there will be erroneous predictions. In general, for a regression equation to make accurate predictions, the error term should be independent, identically (normally) distributed (iid).

Mathematically, this assumption is written as

$$\mathrm{var}(u_i) = \sigma^2 \quad \text{and}$$
$$\mathrm{cov}(u_i u_j) = 0 \quad \text{for } i \neq j.$$

- ✓ '*u*' represents the error terms
- ✓ 'var' represents the variance
- ✓ 'cov' represents the covariance

Improving the accuracy of Linear Regression Model

Bias and Variance

Accuracy refers to how close the estimation is near the actual value, whereas prediction refers to continuous estimation of the value.

High bias = low accuracy (not close to real value)

High variance = low prediction (values are scattered)

Low bias = high accuracy (close to real value)

Low variance = high prediction (values are close to each other)

- ✓ Let us say we have a regression model which is highly accurate and highly predictive; therefore, the overall error of our model will be low, implying a low bias (high accuracy) and low variance (high prediction). This is highly preferable.
- ✓ Similarly, we can say that if the variance increases (low prediction), the spread of our data points increases, which results in less accurate prediction.
- ✓ As the bias increases (low accuracy), the error between our predicted value and the observed values increases.
- ✓ Therefore, balancing out bias and accuracy is essential in a regression model.

Methods to improve accuracy of Linear Regression Model

Accuracy of linear regression can be improved using the following three methods:

1) Shrinkage Approach
2) Subset Selection
3) Dimensionality (Variable) Reduction

1) Shrinkage (Regularization) approach

By limiting (shrinking) the estimated coefficients, we can try to reduce the variance at the cost of a negligible increase in bias. This can in turn lead to substantial improvements in the accuracy of the model.

The two best-known techniques for shrinking the regression coefficients towards zero are

a) ridge regression

b) lasso (Least Absolute Shrinkage Selector Operator)

a) Ridge regression

- ✓ Ridge regression performs L2 regularization, i.e. it adds penalty equivalent to square of the magnitude of coefficients

Minimization objective of ridge = LS Obj + α × (sum of square of coefficients)

- ✓ Ridge regression (include all k predictors in the final model) is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity

b) Lasso regression

Lasso regression performs L1 regularization, i.e. it adds penalty equivalent to the absolute value of the magnitude of coefficients.

Minimization objective of ridge = LS Obj + α × (absolute value of the magnitude of coefficients)

- ✓ The lasso yields sparse models (involving only subset) that are simpler as well as more interpretable.
- ✓ The lasso can be expected to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or equal to zero.

2) Subset selection

Identify a subset of the predictors that is assumed to be related to the response and then fit a model using OLS on the selected reduced subset of variables. There are two methods in which subset of the regression can be selected:

a) Best subset selection (considers all the possible (2^k))

b) Stepwise subset selection

i. Forward stepwise selection (0 to k)

ii. Backward stepwise selection (k to 0)

a) Best subset selection (considers all the possible (2k))

In best subset selection, we fit a separate least squares regression for each possible subset of the k predictors. For computational reasons, best subset selection cannot be applied with very large value of predictors (k). The best subset selection procedure considers all the possible (2^k) models containing subsets of the p predictors.

b) Stepwise subset selection

i. Forward stepwise selection (0 to k)

✓ Forward stepwise selection begins with a model containing no predictors, and then, predictors are added one by one to the model, until all the k predictors are included in the model.

✓ In particular, at each step, the variable (X) that gives the highest additional improvement to the fit is added.

ii. Backward stepwise selection (k to 0)

Backward stepwise selection begins with the least squares model which contains all k predictors and then iteratively removes the least useful predictor one by one.

3) Dimensionality reduction (Variable reduction)

✓ The earlier methods, namely subset selection and shrinkage, control variance either by using a subset of the original variables or by shrinking their coefficients towards zero. In dimensionality reduction, predictors (X) are transformed, and the model is set up using the transformed variables after dimensionality reduction.

✓ The number of variables is reduced using the dimensionality reduction method. Principal component analysis is one of the most important dimensionality (variable) reduction techniques.

Polynomial Regression Model

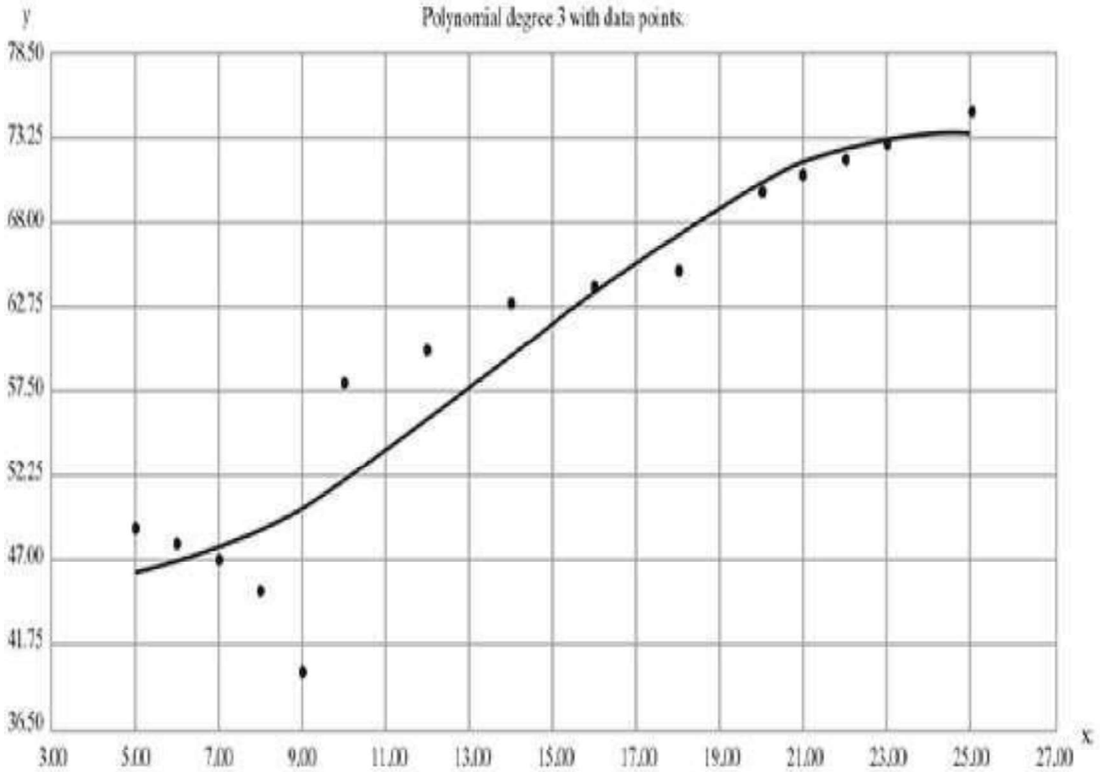
✓ Polynomial regression model is the extension of the simple linear model by adding extra predictors obtained by raising (squaring) each of the original predictors to a power.

✓ For example, if there are three variables, X, X^2, and X^3 are used as predictors. This approach provides a simple way to yield a non-linear fit to data.

$$f(x) = c_0 + c_1.X^1 + c_2.X^2 + c_3.X^3$$

In the above equation, c0 , c1, c2 , and c3 are the coefficients.

Example: Let us use the below data set of (*X*, *Y*) for degree 3 polynomial.



Polynomial regression degree 3

Logistic Regression

- ✓ Logistic regression is both classification and regression technique depending on the scenario used.
- ✓ Logistic regression (logit regression) is a type of regression analysis used for predicting the outcome of a categorical dependent variable similar to OLS regression.
- ✓ In logistic regression, dependent variable (Y) is binary (0,1) and independent variables (X) are continuous in nature.
- ✓ The goal of logistic regression is to predict the likelihood that Y is equal to 1 (probability that Y = 1 rather than 0) given certain values of X. That is, if X and Y have a strong positive linear relationship, the probability that a person will have a score of Y = 1 will increase as values of X increase.
- ✓ So, we are predicting probabilities rather than the scores of the dependent variable.

Formula for Logistic Regression

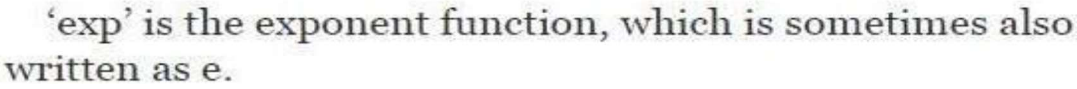
- ✓ An explanation of logistic regression begins with an explanation of the logistic function, which always takes values between zero and one. The logistic formulae are stated in terms of the probability that Y = 1, which is referred to as P. The probability that Y is 0 is 1 − P.

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\ln(p/1 - p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

- ✓ Probability (P) can also be computed from the regression equation. So, if we know the regression equation, we could, theoretically, calculate the expected probability that Y = 1 for a given value of X.

$$P = \frac{\exp\left(a + bX\right)}{1 + \exp\left(a + bx\right)} = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

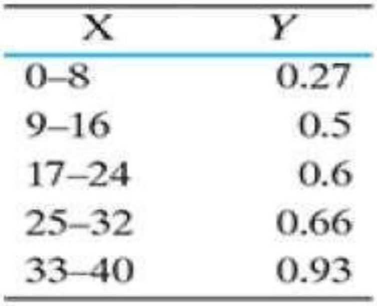


'exp' is the exponent function, which is sometimes also written as e.

Example

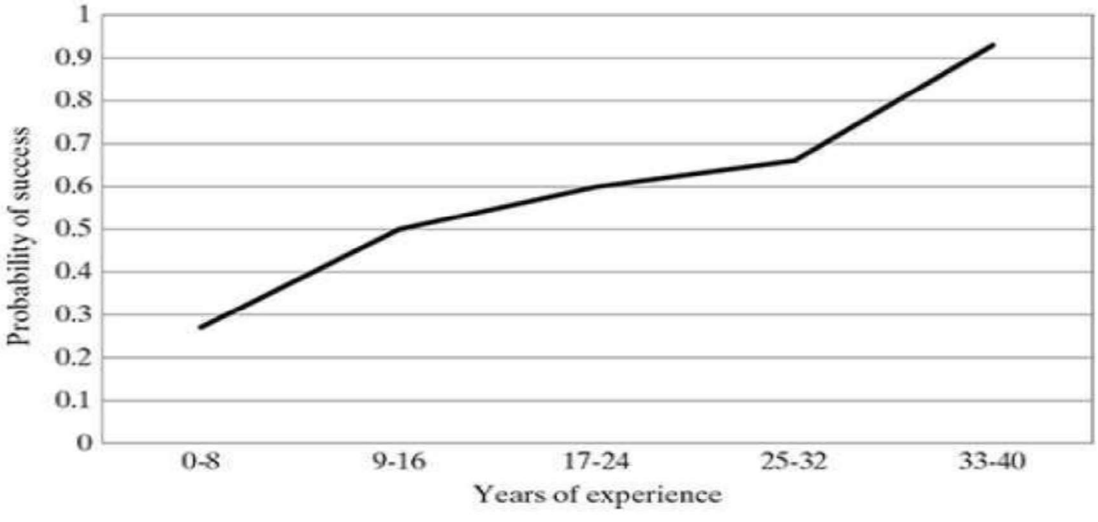
We might try to predict whether or not a small project will succeed or fail on the basis of the number of years of experience of the project manager handling the project. We presume that those project managers who have been managing projects for many years will be more likely to succeed. This means that as X (the number of years of experience of project manager) increases, the probability that Y will be equal to 1 (success of the new project) will tend to increase.

To illustrate this, it is convenient to segregate years of experience into categories (i.e. 0–8, 9– 16, 17–24, 25–32, 33–40). If we compute the mean score on Y (averaging the 0s and 1s) for each category of years of experience, we will get something like



X	*Y*
0–8	0.27
9–16	0.5
17–24	0.6
25–32	0.66
33–40	0.93

When the graph is drawn for the above values of X and Y, it appears like the graph in below Figure



Logistic regression

Assumptions in logistic regression

- ✓ The following assumptions must hold when building a logistic regression model:
- ✓ There exists a linear relationship between logit function and independent variables
- ✓ The dependent variable Y must be categorical (1/0) and take binary value, e.g. if pass then Y = 1; else Y = 0
- ✓ The data meets the 'iid' criterion, i.e. the error terms, ε, are independent from one another and identically distributed
- ✓ The error term follows a binomial distribution [n, p]
    - o n = # of records in the data
    - o p = probability of success (pass, responder)

Maximum Likelihood Estimation

The coefficients in a logistic regression are estimated using a process called Maximum Likelihood Estimation (MLE).

what is likelihood function

A fair coin outcome flips equally heads and tails of the same number of times. If we toss the coin 10 times, it is expected that we get five times Head and five times Tail.

Let us now discuss about the probability of getting only Head as an outcome; it is 5/10 = 0.5 in the above case. Whenever this number (P) is greater than 0.5, it is said to be in favour of Head. Whenever P is lesser than 0.5, it is said to be against the outcome of getting Head.

Let us represent ‘n’ flips of coin as X_1 , X_2 , X_3 ,…, X_n . Now X_i can take the value of 1 or 0.

$X_i = 1$ if Head is the outcome $X_i = 0$ if Tail is the outcome

When we use the Bernoulli distribution represents each flip of the coin:

$$f(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$$

Each observation X is independent and also identically distributed (iid), and the joint distribution simplifies to a product of distributions.

$$f(x_1, \ldots, x_n|\theta){}_{i=1}^{n}f(x_i|\theta) = \theta^{x_1}(1 - \theta)^{1-x_1} \ldots \theta^{x_n}(1 - \theta)^{1-x_n} = \theta^{\#H}(1 - \theta)^{n-\#H},$$

where $\#H$ is the number of flips that resulted in the expected outcome (heads in this case).

The likelihood equation is

$$L(\theta|x) = {}_{i=1}^{n}f(x_i|\theta)$$

MLE is about predicting the value for the parameters that maximizes the likelihood function.

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

-------------------END-------------------

TEXT BOOKS:

1. Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, Machine Learning, Pearson, 2019.

REFERENCE BOOKS:

1. Ethern Alpaydin, — Introduction to Machine Learning, MIT Press, 2004.
2. Stephen Marsland, — Machine Learning - An Algorithmic Perspective, Second Edition, Chapman and Hall / CRC Machine Learning and Pattern Recognition Series, 2014