# Assignment No. 03

Q.) Explain simple linear regression.

→ Simple linear regression :-

Simple linear regression is a statistical method for establishing the relationship betn two variables using a straight line.

The line is drawn by finding the slope and intercept which define the line and minimize regression errors.

The simple form of simple linear regression has only one x variable and one y variable. The x variable is the independant ~~of~~ variable because it is independant of what you try to predict the dependant variable. They variable is the dependant variable because it depends on what you try to predict.

$y = \beta_0 + \beta_1 x + \epsilon$ is the formula used for simple linear regression.

- y is the predicted value of the dependant variable (y) for any given value of the independant variable (x).
- $\beta_0$ is the intercept, the predicted value of y when the x is 0.
- $\beta_1$ is the regression coefficient - how much we expect y to change as x increases.
- x is the independant variable.
- e is the error of the estimate.

Simple linear regression establishes a line that fits your data but it does not guarantee that the line is good enough.

for example, if your data points have an upward far end and are very far apart, then simple linear regression will give you a downward sloping line, which will not match your data.

* Assumptions of simple linear regression-

• linearity -
The relationship between x and y should be linear. It means that, is one value increases the other increases correspondingly. The scatter plot should show this linearity.

• Independant errors:-
It is essential to check if you data are independant of errors. If there is a relationship between the residuals and the variable this could cause problems with your model.

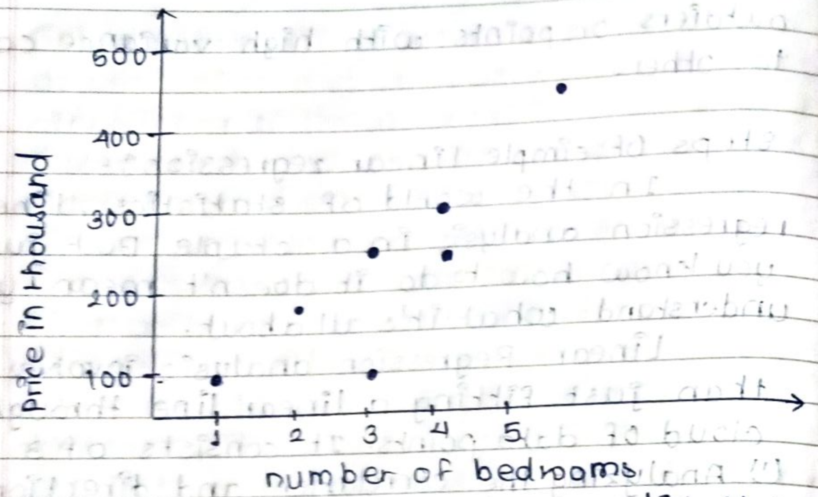
• Normal Distribution:-
It is also essential to check if your data are normally distributed. It will be also show the most of your observations are also close to 0 or 1 (maximum value). It will help you make sure your model is accurate and reliable.

• variance equality:-
Finally, it is essential to check if your data have equal variances. If there are outliers or points with high variance compared to others.

Steps of simple linear regression:-
In the world of statistics, linear regression analysis is a staple. But just becaus you know how to do it doesn't mean you understand what it's all about.
Linear Regression analysis involves more than just fitting a linear line through a cloud of data points. It consists of 3 phases:-
(1) Analyzing the correlation and directionality of the data.
(2) Estimating the model i.e. fitting the line
(3) Evaluating the validity and usefullness of the model.

Q.2 Explain gradient descent for simple linear regression.
→ * Gradient descent for simple linear Regression:
Gradient descent is one of the most focus techniques in machine learning and used for training all sorts of neural network.

Formulating the problem:-
-We have a small dataset where we are given the number of bedrooms in a certain house and we want to predict the price of said house. House prices




- We want to somehow predict the price of any new house enters our dataset.
- To do, this we create a linear function.

$F(x) = b + mx$ that has a minimal mean squared error (or MSE) with regard to our data points.

- once a new point enters our dataset, simply plug in the number of bedrooms of house into function and receive the predicted price for that dataset. But we can't just magically come up with our ideal function. We have to compute it.

Gradient descent algorithms offer a very good perspective for solving problems related to data analysis. They are a class of algorithm that is used to minimize funn.

Here, we would like to minimize the cost function. Gradient descent repeatedly updates each parameters ($\theta_j$). Using a learning rate ($\alpha$). The algorithm work as follows:

Repeat until convergence {

$$\theta_j \cdot \theta_j - a \sum_{i=1}^{m} (h_\theta(x^i) - y^i)\, x_j^i$$

}

If we have n features we have to update all the $\theta$ values ($\theta_0$ to $\theta_n$) simultaneously. This can also be written as:

Repeat until covergence {

$$\theta_j := \theta_j - a \frac{\partial}{\partial \theta_j} J(\theta)$$

}

Here $\alpha$ is the learning rate and controls how big a step the algorithm can take if $\alpha$ is big gradient descent takes large steps, else it takes tiny steps.

Limitations of Gradient descent:-

(1) The gradient descent algorithm takes time to converge to the global optima based on the value of $\alpha$ choosen. The learning rate has to be chosen manually. The algorithm does

not update the learning rate automatically.

(2) The number of iterations should be chosen experimentally in the gradient descent algorithm and partial derivatives of the cost function must be calculated for each iteration.

Q.3) What is hypothesis function for simple linear regression?

→ Hypothesis function for simple Linear regression:-

1) It is a regression algorithm that models the relationship between a dependant (y) and independant variable (x) as $n^{th}$ degree polynomial.

2) equation:-

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_2 x_1^3 + \dots + b_n x_1^n$$

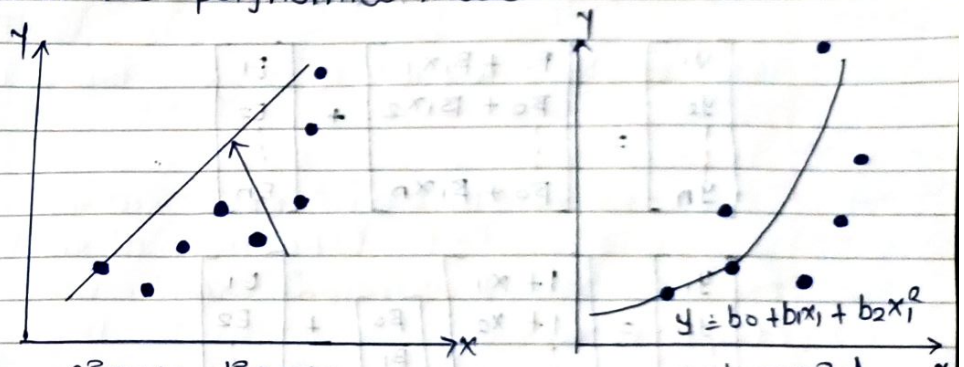
3) It is a linear model with some modification in order to increase the accuracy.

4) The dataset used to in polynomial regression for training of non-linear nature.

5) The dataset used in polynomial regression for training of non-linear nature.

6) Hence, In polynomial Regression the original features are converted into polynomial features of required degree (2, 3, ..., n) &

7) In the above image, we have taken a dataset which is arranged non-linear so if. We try to cover it with a linear model then we can clearly see that it hardly covers any data point on the other hand a curve is suitable to coverst most of the data points which is of the polynomial model.




simple linear                    polynomial

8) Hence, if the datasets are arranged in a non-linear fashion, then we should use the polynomial regression model instead of simple linear regression.

Q.4 Explain simple Regression in matrix form.

→ Simple linear Regression in matrix form

Simple linear Regression in matrix form.

The SLR model in scalar form.

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

Consider, writting as eqn for each observation

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2$$
$$y_3 = \beta_0 + \beta_1 x_3 + \epsilon_3$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$

The SLR model in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 + x_1 \\ 1 + x_2 \\ \vdots \\ 1 + x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The design matrix

$$X_{n \times 2} = \begin{bmatrix} 1 + x_1 \\ 1 + x_2 \\ \vdots \\ 1 + x_n \end{bmatrix}$$

Vector of parameter

$$\beta_2 = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vector of Responses

$$y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Thus,

$$y = X\beta + \epsilon$$

$$\therefore \; y_{n \times 1} = X_{n \times 2} + \beta_{2 \times 1} + \epsilon_{n \times 1}$$

Q.5 Explain Least squares in matrix Form.

→ Least squares in matrix form

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems by minimizing the sum of the residuals made in the results of each individual equation.

Residuals are $\epsilon = y - X\beta$ want to minimize sum of squared residuals.

$$\sum \epsilon_i^2 = [\epsilon_1, \epsilon_2, \epsilon_3, \cdots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon'\epsilon$$

We want to minimize $\epsilon'\epsilon = (y - X\beta)'(y - X\beta)$ where the "prime ( ' )" denotes the transpose of the matrix. (exchange the rows & columns).

We take the derivative with respect to the vector $\beta$. This is like a quadratic function : think "$(y-X\beta)^2$"

The derivative works out to 2 times the derivation of $(y-X\beta)'$ with respect to $\beta$. That is $\frac{d}{d\beta}(y-X\beta)'(y-X\beta) = -2X'(y-X\beta)$.

We set this equal to 0. & solve for $\beta$.

So,

$$-2X'(y-X\beta) = 0 \quad \text{or}$$

$$X'y = X'X\beta \quad \text{(the normal eqn)}$$

Normal $eq^n$

$$X'y = (X'X)\beta$$

Solving this $eq^n$ for $\beta$ gives the least squares $sol^n$ for $b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$ multiply on the left by the inverse of the matrix $X'X$.

$$\boxed{b = (X'X)^{-1}X'y}$$

Assignment No. 04

Q.1 Explain sampling Distribution of Estimators.

→ The sampling distribution of estimators depends on the sampling size. The effect of change of the same size to be determined. An estimate has single numerical value and hence they are called point estimates. There are various estimates like sample mean, sample standard derivation proportion, variance, range etc.

Sampling distribution of the mean:- It is the population mean which the samples are drawn. For all the sample sizes. It is like be normal. The population mean is equal to the mean of the sampling distribution of the mean. Sampling distribution of the mean has the standard deviation

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

where, $\sigma_m$ is standard deviation of the sampling mean, $\sigma$ is the population standard deviation and n is the sample size.

As the size of the sample increases the speed of the sampling distribution of the mean decreases.

But the mean of the distribution remains the same and it is not affected by the sample size.

The sampling distribution of the standard deviation is the standard error of the standard deviation.

$$\sigma_s = \frac{\sigma}{\sqrt{2n}}$$

Here,

$\sigma_s$ is the sampling distribution of the standard deviation.

Q.2) Explain multivariate linear regression.

→ Multivariate linear regression resembles simple linear regression except that in multivariate linear regression, multiple independant variables contribute to the dependant variables and so multiple coefficient are used in the computation.

- It is used to derive a mathematical relationship amongst multiple random variables. It explains are associated with one dependent variable.

- The details of the multiple independent variables are used to make an accurate prediction of the influence they have on the outcome variable.

- Multivariate linear regression model generates a relationship in a linear form (a form of straight line) with the best approximation of each datapoint.

- The equation of the multivariate linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \sigma(y),\ sd(y) = \sigma \text{ (independent of } x's)$$

- y has normal distribution with mean,

- The model parameters $\beta_0 + \beta_1 + \cdots + \beta_p$ and $\delta$ must be estimated from data.

$\beta_0$ = intercept

$\beta_1 \beta_p$ = regression coefficients

$\delta = \delta_{res}$ = residual standard deviation

- Assumptions :-

① The dependent and the independent variables have a linear relationship.

② The independent variables don not have a strong correlation among themselves.

③ The observations of y are chosen randomly and individually from the population.

* Advantages of multivariate reg :-

1) Multivariate regression helps us to study the relationships among multiple variables in the dataset.

2) The correlation betn dependent & independent variables helps in predicting the outcome.

3) It is one of the most convienient & popular algorithms used in machine learning.

* Disadvantages of multivariate reg :-

① complexity of multivariate tech. requires complex.

② not easy to interpret the o/p since loss & error o/p's.

③ cannot be applied to smaller datasets.

Q.3 What is hypothesis function for multivariate linear regression?

→

Hypothesis function for multivariate linear reg :-

A real world dataset always has more than one variable or feature, when a regression problem has more than one feature / variable to consider for the outcome, then it is called multivariate regression.

The hypothesis function of a single variable is,

$$h_\theta(x) = \theta_0 + \theta_1 x$$

In the case of multiple variables n, $x$ becomes a vector or a column containing n variables as vector elements.

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

For n features, the parameters also become a vector,

$$\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

The hypothesis function of multiple variables.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

For our convenience, we add a zeroth features of $x$ which is always 1.

Hence, the hypothesis function becomes

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

In vector form, the hypothesis function can be written as,

$$h_\theta(x) = [\theta_0, \theta_1, \dots \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \vec{\theta}^T \vec{x}$$

Q.4 Using the given data set find the value of y when $x = 10$

$x = \{1, 1, 2, 3, 4, 4, 5, 6, 6, 7\}$

$Y = \{2.1, 2.5, 3.1, 3.0, 3.5, 3.2, 4.3, 3.9, 4.4, 4.8\}$

| | | A | B | A | B | |
|---|---|---|---|---|---|---|
| $x$ | $y$ | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ | A * B |
| 1 | 2.1 | -2.9 | -1.38 | 8.41 | 1.9044 | ~~16.01604~~ 4 |
| 1 | 2.5 | -2.9 | -0.98 | 8.41 | 0.9604 | ~~8.0770~~ 2.84 |
| 2 | 3.1 | -1.9 | -0.38 | 3.61 | 0.1444 | ~~0.5213~~ 0.72 |
| 3 | 3.0 | -0.9 | -0.48 | 0.81 | 0.2304 | ~~0.1866~~ 0.43 |
| 4 | 3.5 | 0.1 | -0.02 | 0.01 | 0.0004 | ~~0.~~ -0.0020 |
| 4 | 3.2 | 0.1 | -0.28 | 0.01 | 0.0784 | ~~0.0008~~ -0.028 |
| 5 | 4.3 | 1.1 | 0.82 | 1.21 | 0.6724 | ~~0.8136~~ 0.902 |
| 6 | 3.9 | 2.1 | 0.42 | 4.41 | 0.1764 | ~~0.7779~~ 0.882 |
| 6 | 4.4 | 2.1 | 0.92 | 4.41 | 0.8464 | ~~3.7326~~ 1.93 |
| 7 | 4.8 | 3.1 | 1.32 | 9.61 | 1.7424 | ~~16.7445~~ 4.0 |
| 39 | 34.80 | 0 | -0.04 | 40.9 | 6.7560 | 15.7640 |

$\bar{x} = \frac{39}{10} = 3.9$

$\bar{y} = \frac{34.80}{10} = 3.48$

$S_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = \sqrt{\frac{40.90}{10-1}} = \sqrt{\frac{40.90}{9}}$

$S_x = 2.1318$

$S_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = \sqrt{\frac{6.7560}{9}} = 0.8664$

$r = \frac{\Sigma (x-\bar{x}) * (y-\bar{y})}{(n-1) * S_x * S_y} = \frac{15.7640}{9 * 2.1318 * 0.8664}$

$= 0.9483$

$b = \frac{r * S_y}{S_x} = \frac{0.9483 * 0.8664}{2.1318} = 0.3854$

$a = \bar{y} - (b)(\bar{x})$

$= 3.48 - (0.3854)(3.9)$

$= 1.9769$

$\hat{y} = a + bx$

$= 1.9769 + 0.3854x$

① $x = 4$

$1.9769 + 0.3854 * 4 = 3.5185$

② $x = 8$

$1.9769 + 0.3854 * 8 = 5.0601$

③ $x = 12$

$1.9769 + 0.3854 * 12 = 6.6017$

④ $x = 16$

$1.9769 + 0.3854 * 16 = 8.1433$

(5) $x = 24$

$1.9769 + 0.3854 * 24 = 11.2265$

(6) $x = 36$

$1.9769 + 0.3854 * 36 = 15.8513$

(7) $x = 20$

$1.9769 + 0.3854 * 20 = 9.6849$

(8) $x = 28$

$1.9769 + 0.3854 * 28 = 12.7681$

(9) $x = 32$

$1.9769 + 0.3854 * 32 = 14.3097$

(10) $x = 40$

$1.9769 + 0.3854 * 40 = 17.3929$

(11) $x = 10$

$1.9769 + 0.3854 * 10 = 5.8309$

Q.5 Using the given data set find value of y when $x = 10$.

$x = \{1, 2, 3, 4, 5, 6\}$

$y = \{25, 35, 42, 50, 55, 60\}$

| $x$ | $y$ | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ | A * B |
|---|---|---|---|---|---|---|
| 1 | 25 | -2.5 | -19.5 | 6.25 | 380.25 | 48.5 |
| 2 | 35 | -1.5 | -9.5 | 2.25 | 90.25 | 14.5 |
| 3 | 42 | -0.50 | -2.5 | 0.25 | 6.25 | 1.25 |
| 4 | 50 | 0.50 | 5.5 | 0.25 | 30.25 | 2.75 |
| 5 | 55 | 1.5 | 10.5 | 2.25 | 110.25 | 15.75 |
| 6 | 60 | 2.5 | 15.5 | 6.25 | 240.25 | ~~40~~ 38.75 |
| 21 | 267 | 0 | 3 | 17.5 | 857.50 | 121.50 |

$\bar{x} = \frac{21}{6} = 3.5$

$\bar{y} = \frac{267}{6} = 44.50$

$S_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = \sqrt{\frac{17.5}{6-1}} = \sqrt{\frac{17.5}{5}} = \sqrt{3.5}$

$= 1.8708$

$S_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = \sqrt{\frac{857.50}{5}} = 13.10958$

$r = \frac{\Sigma(x-\bar{x}) * (y-\bar{y})}{(n-1) * S_x * S_y}$

$= \frac{121.50}{5 * 1.8708 * 13.0958}$

$= 0.9919$

$$b = \frac{r * S_y}{S_x} = \frac{0.9919 * 13.0958}{1.8708}$$

$$= 6.9434$$

$$a = \bar{y} - (b)(\bar{x})$$

$$= 44.50 - 6.9434 * 3.5$$

$$= 20.1981$$

$$\hat{y} = a + bx$$
$$= 20.1981 + 6.9434x$$

$x = 10$

$$\hat{y} = a + bx$$
$$= 20.1981 + 6.9434 \times 10$$
$$= 89.6321$$