Regression Analysis

- It is a statistical tool to find the relationship between one dependent variable and one or more independent variables.
- Example: Consider a Sales Company Dataset and you are a Marketing Analyst of the company.
- Let the dataset has attributes like Adv. Cost (Rs.) and Sales Amount (Quantity)

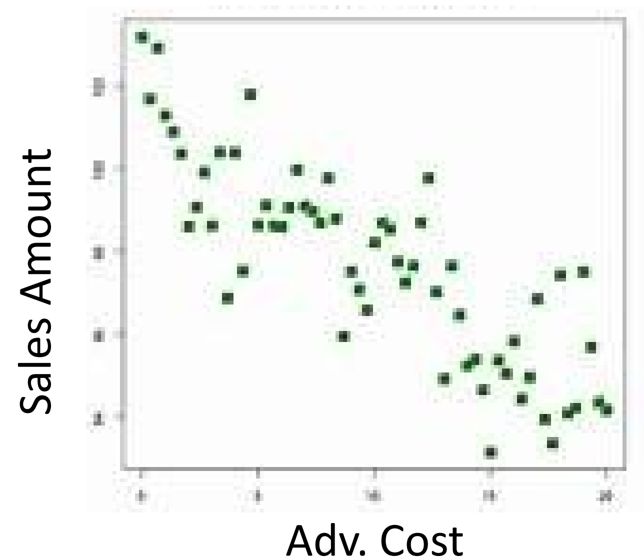
Adv. Cost (Rs.)	Sales Amount
1	1
2	1
3	2
4	2
5	4

Regression Analysis

- Here, we may decide how much money we can spent for advertising. So, amount of money spent is the controlled variable .
- But, we can’t control the Sales amount, so it is not a controlled variable. This is dependent variable depends on Advertising Cost.
- This is not only the factor, Sales amount may also dependent on some other factors like no. of persons working, etc.
- Adv. Cost : X→ Independent or Regressor Variable
- Sales Amount: Y →Dependent or Response or Random Variable

Regression Analysis

- Scatter plot is a mathematical diagram to display values of two variables for a set of data.
- It is used to investigate the possible relationship between the variables.
- If it indicates the linear relation then linear regression is consider; otherwise polynomial regression, and so on are considered.



Simple Linear Regression (SLR) Model

- It is a model with a single regressor variable X that has linear relationship with a response variable Y.
- The simple linear regression model is:

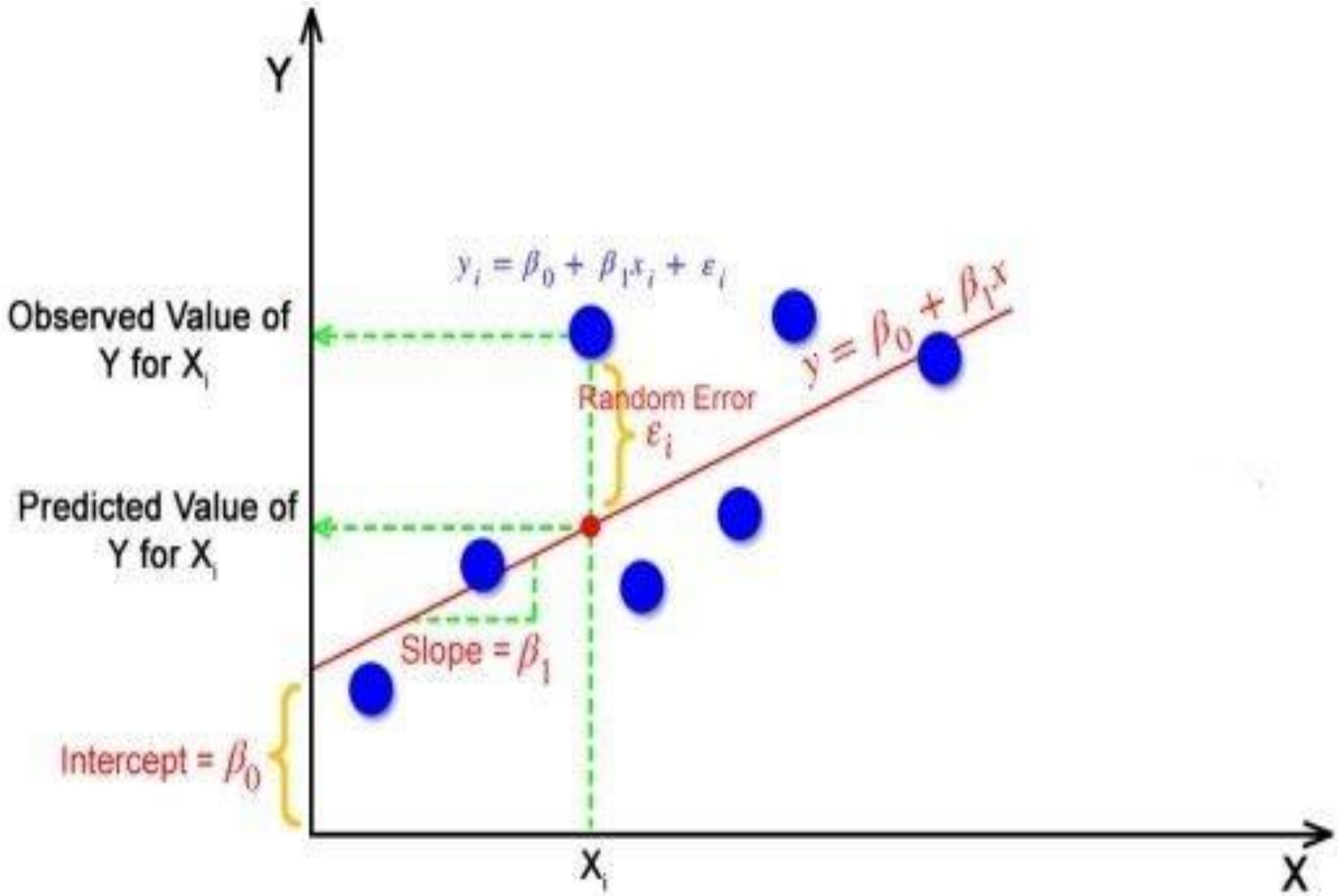
  $Y = a + cX + \epsilon$, where a→ intercept, c→ Slope, and ϵ is a random error component.

=> For a given X, the corresponding observation Y consists of the value $a + cX + \epsilon$.

- The same model may be written as:

  $y_i = a + cx_i + \epsilon_i$ for $i = 1, 2, \ldots, n$; n is the no. of observations.

Simple Linear Regression (SLR) Model



Simple Linear Regression (SLR) Model

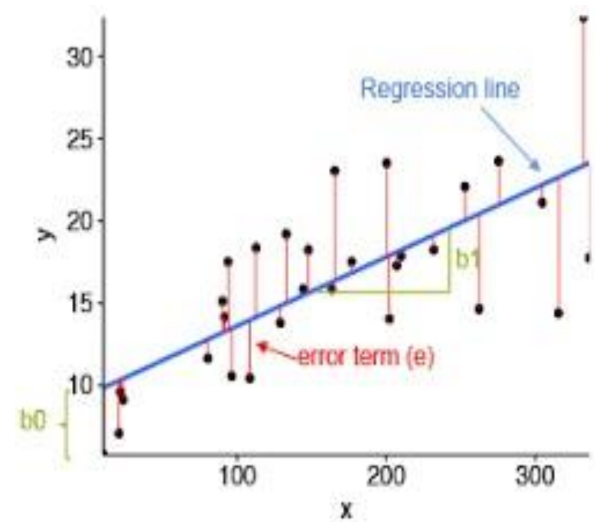
- Let the best fitted model is:

$\hat{Y} = \hat{a} + \hat{c}X$ or

$\hat{y}_i = \hat{a} + \hat{c}x_i$; $i=1,2,\ldots,n$

- The line fitted by Least Square Method (LSM) which makes the sum of square of all vertical discrepancies as small as possible.
- The LSM estimates the parameters a and c using the dataset <X, Y> as â and ĉ.

X	Y
x_1	y_1
x_2	y_2
x_i	y_i
x_n	y_n



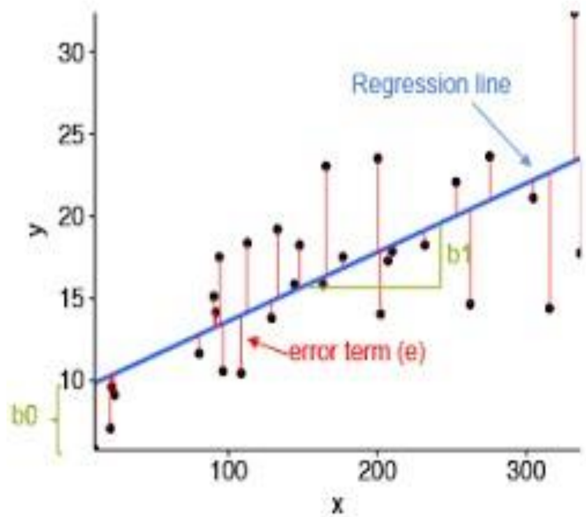
Simple Linear Regression (SLR) Model

- The cost function or error function of the LSM is:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- We estimate a and c so that sum of square of all the differences between the observed y_i and the predicted $\hat{y}_i$ is minimum, i.e., S is minimum.
- This S is called Sum of Square Residual, i.e., SS_{Res} = $S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Residual = Deviation between actual and predicted value
- Error = Deviation between actual value and mean of population.

X	Y
x_1	y_1
x_2	y_2
x_i	y_i
x_n	y_n



Parameter estimation using LSM

$$SS_{Res} = S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{c}x_i)^2$$

- By taking the partial derivatives = 0 and solving, we get the best fitted model is:

  $\hat{Y} = \hat{a} + \hat{c}X$, where

$$\hat{a} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{c} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Parameter estimation using LSM

The least square estimator of a & b (i.e. $\hat{a}$ & $\hat{b}$) must satisfies: $\left.\frac{\partial S}{\partial a}\right|_{\hat{a},\hat{b}} = 0$ and $\left.\frac{\partial S}{\partial b}\right|_{\hat{a},\hat{b}} = 0$

$$\therefore SS_{res} = S = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - \hat{a} - \hat{b} x_i)^2$$

$$\therefore \frac{\partial S}{\partial a} = 0 \ \& \ \frac{\partial S}{\partial b} = 0$$

$$\therefore -2 \sum (y_i - \hat{a} - \hat{b} x_i) = 0 \quad (1) \quad \& \quad -2 \sum x_i (y_i - \hat{a} - \hat{b} x_i) = 0 \quad (2)$$

(1) $\Rightarrow$ $\sum (y_i - \hat{a} - \hat{b} x_i) = 0$

$\Rightarrow \sum y_i - \sum \hat{a} - \sum \hat{b} x_i = 0$

$\Rightarrow \sum y_i - n\hat{a} - \hat{b} \sum x_i = 0$

$\Rightarrow \hat{a} = \frac{\sum y_i}{n} - \hat{b} \frac{\sum x_i}{n}$

$= \bar{y} - \hat{b} \bar{x}$ where $\bar{y} = \frac{\sum y_i}{n}$ & $\bar{x} = \frac{\sum x_i}{n}$ — (3)

Parameter estimation using LSM

05.00

(2) $\Rightarrow \sum x_i(y_i - \hat{a} - \hat{b} x_i) = 0$

06.00

$\Rightarrow \sum x_i(y_i - \bar{y} + \hat{b}\bar{x} - \hat{b} x_i) = 0$ [using (3)]

$\Rightarrow \sum x_i(y_i - \bar{y}) + \hat{b} \sum x_i(\bar{x} - x_i) = 0$

$\Rightarrow \hat{b} \sum (x_i - \bar{x}) x_i = \sum (y_i - \bar{y}) x_i$

$$\Rightarrow \hat{b} = \frac{\sum (y_i - \bar{y}) x_i}{\sum (x_i - \bar{x}) x_i} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})(x_i - \bar{x})} \quad (4)$$

if $\sum (y_i - \bar{y}) \bar{x} = 0$

and $\sum (x_i - \bar{x}) \bar{x} = 0$

Important :

DECEMBER 2018								JANUARY 2019							
WK	S	M	T	W	T	F	S	WK	S	M	T	W	T	F	S
48	30	31					1	1			1	2	3	4	5
49	2	3	4	5	6	7	8	2	6	7	8	9	10	11	12
50	9	10	11	12	13	14	15	3	13	14	15	16	17	18	19
51	16	17	18	19	20	21	22	4	20	21	22	23	24	25	26
52	23	24	25	26	27	28	29	5	27	28	29	30	31		

Parameter estimation using LSM

$$\hat{b} = \cancel{\Sigma x_i y_i - \Sigma}\ \frac{\Sigma(x_i y_i) - \Sigma(x_i \bar{y})}{\Sigma x_i^2 - \Sigma(x_i \bar{x})}$$

$$= \frac{\Sigma x_i y_i - \bar{y}\,\Sigma x_i}{\Sigma x_i^2 - \bar{x}\,\Sigma x_i}$$

$$= \frac{\Sigma x_i y_i - \frac{\Sigma y_i}{n}\,\Sigma x_i}{\Sigma x_i^2 - \frac{\Sigma x_i}{n}\,\Sigma x_i} = \frac{n\,\Sigma x_i y_i - \Sigma x_i \,\Sigma y_i}{n\,\Sigma x_i^2 - (\Sigma x_i)^2} \quad (5)$$

ortant :

Parameter estimation using LSM

from (3)

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

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

$$= \frac{\sum y_i}{n} - \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}\cdot \frac{\sum x_i}{n}$$

$$= \frac{\sum y_i \{n\sum x_i^2 - (\sum x_i)^2\} - \sum x_i (n\sum x_i y_i - \sum x_i \sum y_i)}{n\{n\sum x_i^2 - (\sum x_i)^2\}}$$

$$= \frac{n\sum x_i^2 \sum y_i \cancel{- (\sum x_i)^2 \sum y_i} - n\sum x_i \sum x_i y_i \cancel{+ (\sum x_i)^2 \sum y_i}}{n\{n\sum x_i^2 - (\sum x_i)^2\}}$$

$$= \frac{\cancel{n}\{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i\}}{\cancel{n}\{n\sum x_i^2 - (\sum x_i)^2\}}$$

$$= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} \quad \text{--- (6).}$$

Parameter estimation using LSM

2019 JANUARY 17 THURSDAY 3rd Week • 017-348

two thousand Nineteen

∴ The simple regression model is

$$Y = \hat{a} + \hat{b} X$$ where

09.00

$$\therefore \hat{a} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

10.00

$$\& \quad \hat{b} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

11.00

$\epsilon \leftarrow$ random error variable

During computation some error may occur due to instrument.

Basic Assumptions on SLR Model

- In simple LRM, $y_i = a + cx_i + \epsilon_i$ for $i = 1, 2, \ldots, n$:

i) ϵ_i is a random variable with zero mean and variance σ^2 (Unknown), i.e., $E(\epsilon_i)=0$ and $Var(\epsilon_i)= \sigma^2$

ii) ϵ_i and ϵ_j are uncorrelated, $i \neq j$, i.e., $COV(\epsilon_i,\epsilon_j) = 0$

iii) ϵ_i is a normally distributed random variable with zero mean and variance σ^2. i.e., $\epsilon_i \sim N(0, \sigma^2)$

So, ϵ_i's are normally distributed and uncorrelated => ϵ_i's are independent.

Consequences in terms of y_i

- $y_i = a + cx_i + \epsilon_i$ for $i = 1, 2, \ldots, n$:

i) $E(y_i) = E(a + cx_i + \epsilon_i)$

$= E(a) + E(cx_i) + E(\epsilon_i)$

$= a + x_i E(c) + 0$ [as x is controlled variable, and $E(\epsilon_i)=0$]

$= a + cx_i$

ii) $Var(y_i) = Var(a + cx_i + \epsilon_i)$

$= Var(a) + Var(cx_i) + Var(\epsilon_i)$

$= 0 + 0 + \sigma^2 = \sigma^2$

Since, ϵ_i follows normal distribution with zero mean and σ^2 variance, so the consequence of y_i is that y_i follows normal distribution with $a + cx_i$ mean and σ^2 variance,

i.e., $y_i \sim N(a + cx_i, \sigma^2)$

Consequences in terms of y_i

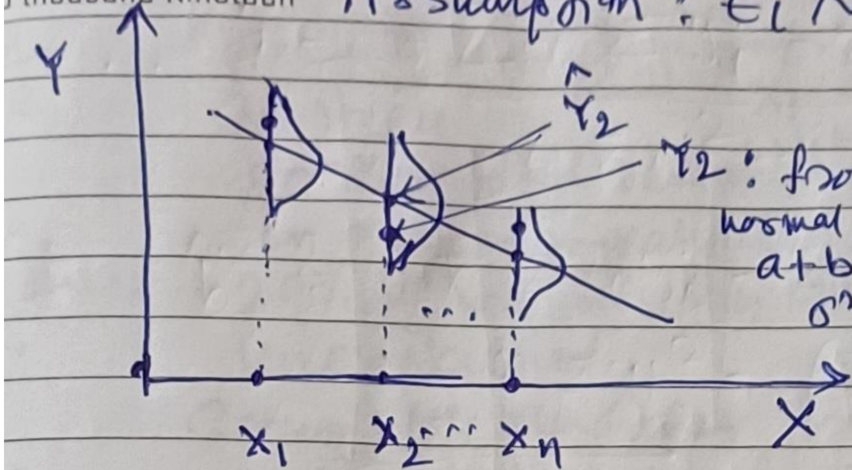
- So, we assume that, i-th observation, y_i is from normal distribution with mean = $a + cx_i$ and Var = σ^2 (Also, y_i's are uncorrelated and independent)
- Therefore, given a set of data, the dataset must satisfy this assumptions.
- If the assumptions are not hold, then we should not apply this regression analysis. This is verify using topic modeling adequacy checking (study in your own, if interested).

Graphical interpretation of Assumptions in SLR Model

Graphical Representation

Assumption: $\epsilon_i \overset{\text{independent}}{\sim} N(0, \sigma^2)$ &

$Y_i \overset{\text{ind}}{\sim} N(a + b x_i, \sigma^2)$



Y_2: from normal distribution
$a + b x_2$ mean &
$\sigma^2 \to$ variance

X	Y
x_1	Y_1
x_2	Y_2
$\vdots$	
x_n	Y_n

The line is $Y = \hat{a} + \hat{b} x$

i-th value of the response variable Y_i is coming from normal distribution with $a + b x_i \to$ mean & $\sigma^2 \to$ vari.

Multiple Linear Regression (MLR) Model

- Consider the same Company Sales Dataset.
- In case of SLR, we assume that the response variable “Sales Amount” is fully explained by the regressor variable “Adv. Cost”
- But in reality, it may be say, 80% explained by “Adv. Cost”. Remaining 20% may be explained by other factor, say “No. of sales person” employed.
- In practice, there are more than one regressor variables, in that case, we consider MLR.

MLR Model

MLR:

MLR Model: more than one regressor variable,
say k−1 vari; $x_1, x_2, \cdots x_{k-1}$

General form of MLR is: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_{k-1} x_{k-1} + \epsilon$

or $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i,k-1} + \epsilon_i$
for $i = 1, 2, \cdots n$

Here we have more than 1 regressor variable so called multiple regression and the model is linear in terms ~~of unknown~~ regressor variables. ~~parameters $\beta_0, \beta_1, \cdots \beta_{k-1}$ (not in $x_1, x_2 \cdots x_{k-1}$)~~

∴ The model is called MLR

β_0

	β_1	β_2		β_{k-1}		
	x_1	x_2	$\cdots$	x_{k-1}	Y	ϵ
	x_{11}	x_{12}	$\cdots$	$x_{1,k-1}$	Y_1	ϵ_1
	x_{21}	x_{22}		$x_{2,k-1}$	Y_2	ϵ_2
	$\vdots$					$\vdots$
i-th observation →	x_{i1}	x_{i2}	$\cdots$	$x_{i,k-1}$	Y_i	ϵ_i
	$\vdots$					$\vdots$
	x_{n1}	x_{n2}	$\cdots$	$x_{n,k-1}$	Y_n	ϵ_n

Important :

FEBRUARY 2019

WK	S	M	T	W	T	F	S
5						1	2
6	3	4	5	6	7	8	9
7	10	11	12	13	14	15	16
8	17	18	19	20	21	22	23

MARCH 2019

WK	S	M	T	W	T	F	S
9	31					1	2
10	3	4	5	6	7	8	9
11	10	11	12	13	14	15	16
12	17	18	19	20	21	22	23

MLR Model

JANUARY 23 WEDNESDAY

4th Week • 023-342

Two thousand Nineteen

Assumption: error $\epsilon_i \sim N(0, \sigma^2)$ and

Define matrices:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}_{k\times 1}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1} \text{. and}$$

vector of observations on Y — vector of parameters — vector of errors

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,k-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{bmatrix}_{n\times k}$$

Dataset

x_1	x_2	$\cdots$	x_{k-1}	Y
x_{11}	x_{12}	$\cdots$	$x_{1,k-1}$	Y_1
$\vdots$				
x_{n1}	x_{n2}	$\cdots$	$x_{n,k-1}$	Y_n

MLR Model

04.00

This model can be expressed as:

05.00

$Y = X\beta + \epsilon$ in matrix form.

06.00

We have to fit this model means we have to estimate the parameters.

X: $n \times k$, β: $k \times 1$
$\Rightarrow X\beta$: $n \times 1$
ϵ: $n \times 1$
$\therefore Y$: $n \times 1$

Estimation of Model parameters (MLR)

Like simple linear Regression (SLR), we will estimate using Least Square ~~error~~ method, where the parameters are ~~estimated~~ determined by minimizing the SS Residual (sum of square residual).

DECEMBER							2018
WK	S	M	T	W	T	F	S
48	30	31					1
49	2	3	4	5	6	7	8
50	9	10	11	12	13	14	15
51	16	17	18	19	20	21	22
52	23	24	25	26	27	28	29

JANUARY							2019
WK	S	M	T	W	T	F	S
1			1	2	3	4	5
2	6	7	8	9	10	11	12
3	13	14	15	16	17	18	19
4	20	21	22	23	24	25	26
5	27	28	29	30	31		

Important :

MLR Model

2019 — two thousand Nineteen — THURSDAY 24 — 4th Week • 024-341

Least square method (LSM) determines the parameters by minimizing SS_{Res} where

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

Let the fitted model is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_{k-1} x_{k-1}$

$$\text{i.e. } SS_{Res} = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_{k-1} x_{i\,k-1}\right)^2$$

We now represent SS_{Res} in matrix form, for that we consider residual vector e as

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad e_i = i\text{-th residual}$$

$\boxed{e_i = y_i - \hat{y}_i \;\; \forall\, i = 1, 2, \ldots n \qquad e' = (e_1, e_2, \cdots e_n)}$

$$\therefore\; e = \underbrace{Y}_{\text{vector of observations}} - \underbrace{\hat{Y}}_{\text{vector of observations for the fitted value}}$$

MLR Model

05.00

$\therefore SS_{Res} = \sum_{i=1}^{n} e_i^2 = e'e$

$e'e = [e_1\ e_2 \cdots e_n]\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$

$= (Y - \hat{Y})'(Y - \hat{Y})$

06.00

$\because Y = XB + e \Rightarrow$

$= (Y - X\hat{\beta})'(Y - X\hat{\beta})$

and $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{k-1} X_{k-1}$

$= X\hat{\beta}$ (in matrix form)

$= (Y' - \hat{\beta}' X')(Y - X\hat{\beta})$

$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$

$=$ $1\times n, n\times 1$ $= 1\times 1$

$1\times n, n\times k, k\times 1 \Rightarrow 1\times 1 \Rightarrow$ scalar quantity

$1\times k, k\times n, n\times 1$, $1\times 1 \Rightarrow$ scalar quantity

$1\times k, k\times n, n\times k, k\times 1 \Rightarrow 1\times 1$

$\Rightarrow$ every term is scalar quantity

Important

Transpose of the scalar is itself.

FEBRUARY							2019
WK	S	M	T	W	T	F	S
5						1	2
6	3	4	5	6	7	8	9
7	10	11	12	13	14	15	16
8	17	18	19	20	21	22	23
9	24	25	26	27	28		

MARCH							2019
WK	S	M	T	W	T	F	S
9	31					1	2
10	3	4	5	6	7	8	9
11	10	11	12	13	14	15	16
12	17	18	19	20	21	22	23
13	24	25	26	27	28	29	30

MLR Model

JANUARY 25 FRIDAY 4th Week • 025-340

2019 two thousand Nineteen

$$\underbrace{Y'X\hat{\beta}}_{1\times 1 \text{ matrix}} = (X'Y)'\hat{\beta} = (\hat{\beta}'(X'Y))'$$

For 1×1 matrix A, $A = A'$

$$= \underbrace{(\hat{\beta}'X'Y)'}_{1\times 1 \text{ matrix}} = \hat{\beta}'X'Y$$

$$SS_{Res} = Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - \hat{\beta}'X'Y - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$\therefore$ We have two SS_{Res} in different forms.

For estimation of K-unknowns $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{K-1}$

we get,

$$\left.\begin{aligned} \left.\frac{\partial SS_{Res}}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{K-1}} &= 0 \\ \left.\frac{\partial SS_{Res}}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{K-1}} &= 0 \\ &\vdots \\ \left.\frac{\partial SS_{Res}}{\partial \beta_{K-1}}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{K-1}} &= 0 \end{aligned}\right\}$$

So we have K normal equations with K-unknowns $\Rightarrow$ Solution gives the estimated values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{K-1}$.

MLR Model

Process LSM for both forms;

1. matrix form

$$SS_{Res} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$\frac{\partial SS_{Res}}{\partial \hat{\beta}} = 0 \Rightarrow$$

$$\frac{\partial}{\partial \hat{\beta}}\left(Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right) = 0$$

$$\Rightarrow 0 - 2X'Y + 2X'X\hat{\beta} = 0$$

$$= -2X'Y + 2X'X\hat{\beta} = 0$$

$$\Rightarrow X'Y = X'X\hat{\beta}$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

2nd Form

$$SS_{Res} = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{k-1} x_{i,k-1}\right)^2$$

normal equ.

$$\frac{\partial SS_{Res}}{\partial \hat{\beta}_0} = 0 \Rightarrow$$

$$\sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{k-1} x_{i,k-1}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} e_i = 0$$

MLR Model

$\Rightarrow X'Y = X'X\hat{\beta}$

$\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$

This will give all the unknown parameters.

$\frac{d}{dx}(x^T B) = B$ $\quad \frac{d}{dx}(x^T b) = b$

$\frac{d}{dx}(x^T x) = 2x$

$\frac{d}{dx}(x^T B x) = 2Bx$

$$\frac{\partial SS_{Res}}{\partial \hat{\beta}_1} = 0 \Rightarrow$$

$$\sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{k-1} x_{i,k-1}\right) \cdot x_{i1} = 0$$

$$\Rightarrow \sum (Y_i - \hat{Y}_i) x_{i1} = 0$$

$$\Rightarrow \sum_{i=1}^{n} e_i x_{i1} = 0$$

Similarly we get

$$\sum e_i x_{i,k-1} = 0$$

So we have k-normal equations

$\sum e_i = 0$, $\sum e_i x_{i1} = 0, \cdots, \sum e_i x_{i,k-1} = 0$

All are independent which gives the sol^n of k-unknowns $\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_{k-1}$.

Degree of Freedom (DF) of SS_{Res}

2019 FEBRUARY

SATURDAY 09

6th Week • 040-325

two thousand Nineteen

<u>Linear regression</u>

We may represent SS residual i.e., SS_{res} in many different forms, <u>like</u>

i) $SS_{res} = \sum_{i=1}^{n} e_i^2$

ii) $SS_{res} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$

where $\hat{\beta} = (X'X)^{-1}X'Y$

$$\therefore SS_{res} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'\underbrace{X'X(X'X)^{-1}}_{I}X'Y$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y$$

$$= Y'Y - \hat{\beta}'X'Y$$

*) Now, find degree of freedom (DF) of SS_{res}

$\because SS_{res} = \sum_{i=1}^{n} e_i^2$ where $e_i \sim N(0, \sigma^2)$ and e_i satisfies k-Constraints:

$$\left.\begin{aligned} &\sum e_i = 0 \\ &\sum e_i x_{i1} = 0 \\ &\vdots \\ &\sum e_i x_{ik-1} = 0 \end{aligned}\right\}$$

Degree of Freedom (DF) of SS_{Res}

6.00

SUNDAY 10

$\therefore$ To compute SS_{res}, we don't have freedom to choose all e_i for $i = 1, 2, \cdots, n$ independently.
We can choose $n-k$ of $e_1, e_2, \cdots, e_n$ independently
i.e., we have the ~~degree~~ of freedom to choose $n-k$ of n e_i's and the remaining k e_i's have to be chosen in such a way that they satisfy above k-constraints.

Important :

$\therefore$ We are loosing k degree of freedom for k-constraints on the residuals.

$\therefore$ $DF(SS_{res}) = n-k$.

MARCH 2019

WK	S	M	T	W	T	F	S
9	31					1	2
10	3	4	5	6	7	8	9
11	10	11	12	13	14	15	16
12	17	18	19	20	21	22	23
13	24	25	26	27	28	29	30

APRIL 2019

WK	S	M	T	W	T	F	S
14		1	2	3	4	5	6
15	7	8	9	10	11	12	13
16	14	15	16	17	18	19	20
17	21	22	23	24	25	26	27
18	28	29	30				

Degree of Freedom (DF) of SS_T

FEBRUARY 11 MONDAY
7th Week • 042-323
two thousand Nineteen

SS_{Total} i.e., SS_T is the variation i.e., variability in the response variable (Y).

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

in matrix form.

$$= \sum(Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2)$$

$$= \sum Y_i^2 - 2\bar{Y}\sum Y_i + \bar{Y}^2\sum 1$$

$$= \sum Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2$$

$$= \sum Y_i^2 - n\bar{Y}^2$$

Here, SS_T = sum of n-terms i.e., $\sum(Y_i - \bar{Y})^2$

To compute SS_T we must satisfy the constraint $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ ← obvious

partial derivative $\sum(Y_i - \bar{Y}) = 0$ ⇒ choose $n-1$ freely.

Degree of Freedom (DF) of SS_T

$\therefore$ To Compute SS_T we don't have freedom to choose all the n-terms independently.
out of all the n-terms $(Y_1-\bar{Y}), (Y_2-\bar{Y}), \ldots (Y_n-\bar{Y})$ we may choose atmost $(n-1)$ terms independently and 1-terms should be chosen in such a way that the constraint $\Sigma(Y_i-\bar{Y})=0$

$\therefore DF(SS_T) = n-1$

NUARY 2019 FEBRUARY 2019 Important :

Degree of Freedom (DF) of SS_{Reg}

SS regressor i.e., SS_{reg} defines how much of the variability in response variable Y is explained by the regressor model.

$$\therefore \; SS_T = SS_{reg} + SS_{res}$$

$SS_T \rightarrow$ Total variability in response variable

$SS_{reg} \rightarrow$ variability in response variable explained by model

$SS_{res} \rightarrow$ variability in Y not explained by model (i.e., error)

$\therefore$ Matrix form of SS_{reg} is

From previous page

$$
\begin{aligned}
SS_{reg} &= SS_T - SS_{res} \\
&= \sum Y_i^2 - n\bar{Y}^2 - (Y'Y - \hat{\beta}'X'Y) \\
&= \cancel{Y'Y} - n\bar{Y}^2 - \cancel{Y'Y} + \hat{\beta}'X'Y \\
&= \hat{\beta}'X'Y - n\bar{Y}^2
\end{aligned}
$$

$$\because \; SS_T = SS_{reg} + SS_{res}$$

$$\therefore \; DF(SS_T) = DF(SS_{reg}) + DF(SS_{res})$$

$$\Rightarrow \; n-1 = DF(SS_{reg}) + n-K$$

$$\Rightarrow \; DF(SS_{reg}) = K-1$$

Thank you