Chapter - r

CH - 8

Least - Squares Estimates

It is a statistical method used to find the best - fitting line or curve for a set of data points by minimizing the sum of the squares of the residuals (the differences between observed & predicted values)

→ For a simple linear regression eqn, the model is

$$\boxed{y = \beta_0 + \beta_1 x + \epsilon} \quad \text{—— (1)}$$

where β_0 & β_1 are the parameters to estimate. (ie estimate parameters)

ϵ → error term

The residuals ($y_i - \hat{y}$) are estimates of the error term ϵ_i, $i = 1, \cdots, n$

→ Now,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \cdots, n$$

→ The sum of squared errors

$$SSE_p = \sum_{i=1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad \text{—— (2)}$$

find the values of β_0 & β_1 that minimizes the ϵ_i. So we have to use Partial derivatives of equⁿ $\textcircled{2}$ w.r.to. β_0 & β_1 . β_0

$$\left.\begin{aligned} \frac{\partial SSE_p}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i) \\ \frac{\partial SSEp}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i(y_i-\beta_0-\beta_1 x_i) \end{aligned}\right\} \longrightarrow \textcircled{3}$$

The values for the estimates bo & b1, set the equⁿ $\textcircled{3}$ equal to zero.

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$$

Distributing the summation gives us

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \left.\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\} \longrightarrow \textcircled{4}$$

Solving eqn (4) for b_1 & b_0 we have

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i)(\sum y_i)]/n}{\sum x_i^2 - (\sum x_i)^2/n} \quad \text{--- (5)}$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{--- (6)}$$

Q ~~$b_0 = 59.853$~~ , ~~$b_1 = -2.4614$~~

Fit the regression model

$$y_i = b_0 + b_1 x_i + e_i$$

X	Y
1	2
2	4
3	5
4	4
5	6

S-1 $b_1 = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

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

$\bar{x} = 3$, $\bar{y} = 4.2$

$b_1 = \frac{8}{10} = 0.8$, $b_0 = 4.2 - (0.8)3$
$= 1.8$

So, $\boxed{\hat{y} = 1.8 + 0.8x}$

S-2 compute predicted values & Residuals

$$e_i = y_i - \hat{y}_i$$

x	y	$\hat{y} = 1.8 + 0.8(X)$	Residual (e_i)	(e-4) $x_i e_i$
1	2	2.6	-0.6	-0.6
2	4	3.4	0.6	1.2
3	5	4.2	0.8	2.4
4	4	5.0	-1.0	-4.0
5	6	5.8	0.2	1.0

S-3 Sum of residuals = 0

$\sum e_i = (-0.6) + 0.6 + 0.8 - 1.0 + 0.2 = 0$

42/59

S-4 Sum of $x_i e_i$ = 0

$\sum x_i e_i = -0.6 + 1.2 + 2.4 - 4.0 + 1.0 = 0$

x	ei	e_i^2
1	-0.6	0.36
2	0.6	0.36
3	0.8	0.64
4	-1.0	1.00
5	0.2	0.04

$SSE = \sum e_i^2$

$= 0.36 + 0.36 + 0.64 + 1.00 + 0.04$

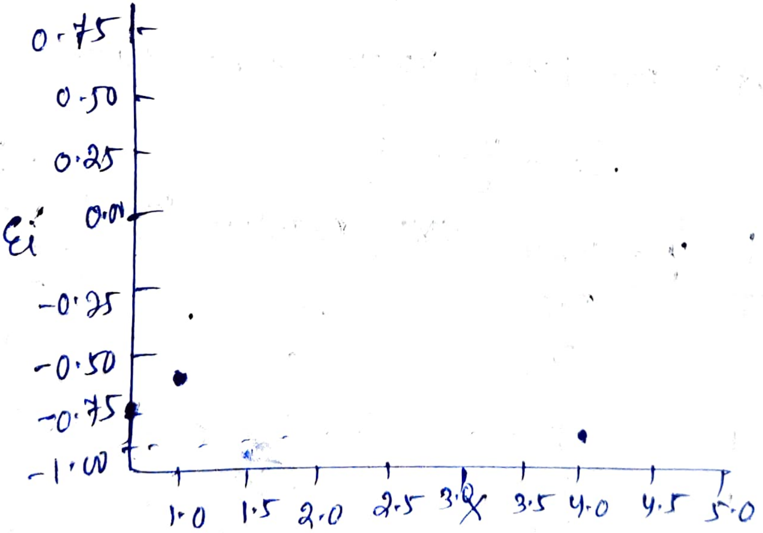
total prediction error of the regression model $= 2.40$

Interpretation

Smaller SSE → better fit

SSE = 0 → Perfect Prediction

* Residual plot for simple Linear Regression



2 Extrapolation

It is a Statistical and analytical technique used to predict values beyond the range of Observed data, based on the existing trend or Pattern.

Ex

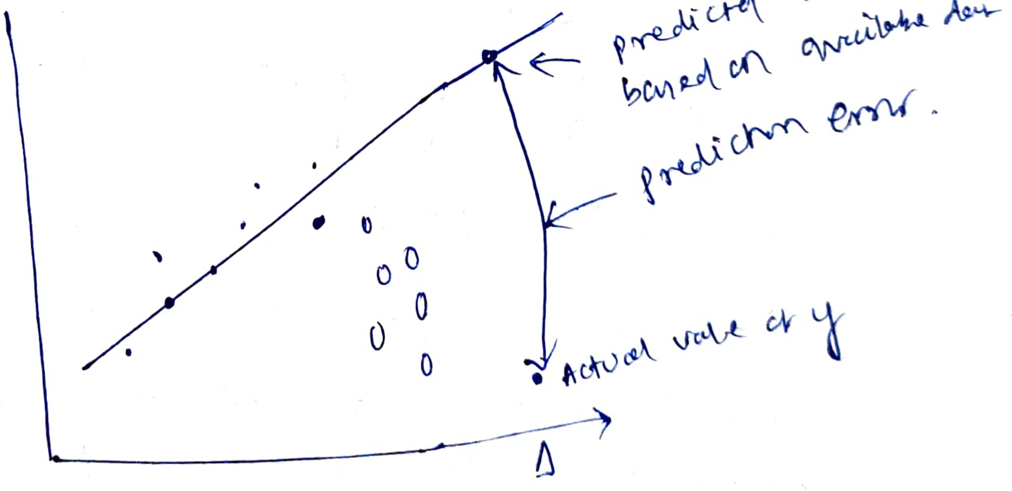
Hours Studied	Exam score
2	40
4	55
6	70

If we predict the score for 8 hours of Study, this is extrapolation. because 8 is outside the observed range.

Eg $\hat{y} = 10x + 20$

For $x = 8$, $\hat{y} = 10 \times 8 + 20 = 100$

This prediction is extrapolated.



→ Extrapolation should be avoided if possible. If predich outside the given range of x must be performed, the end user of the prediction needs to be informed that no x-data is available to support such a prediction.

8.3 Coefficient of Determination, r^2

→ r^2, for measuring the goodness of fit of the regression.

→ r^2 also known as coefficient of determination,

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

where
SSE = sum of squares error
SSR = sum of squares regression.
SST = sum of squares total

$\therefore (y-\bar{y}) = (\hat{y}-\bar{y}) + (y-\hat{y})$

$SST = SSR + SSE$

$$\boxed{\text{ie} \Rightarrow SSR = SST - SSE}$$

$$SST = \sum_{i=1}^{n} (y-\bar{y})^2$$

→ squares both side & summation

$$\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y}_i)^2$$

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

$$SSR = \sum_{i=1}^{n} (\hat{y}-\bar{y})^2$$

$\Rightarrow SST = SSR + SSE$

→ if $SSE = 0$ then $SST = SSR$, so $r^2 = 1$

maximum value of r^2 is 1 which occurs when the regression is a perfect fit.

Subject	x = Time	y = Distance	Predicted score $\hat{y} = 6+2x$	Error in prediction $(y-\hat{y})$	Error in prediction² $(y-\hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	-1	1
5	4	14	14	0	0
6	5	15	16	-1	1
7	6	20	18	+2	4
8	7	18	20	-2	4
9	8	22	22	0	0
10	9	25	24	1	1

$SSE = \sum (y-\hat{y})^2$
$= 12$

8

T-2 Calculate SST (page - 224)

x	y	$\bar{y}$	$(y-\bar{y})$	$(y-\bar{y})^2$
		16	-6	36
		16	-5	25
		16	-4	16
			-3	9
		〃	-2	4
			-1	1
			4	16
			2	4
			6	36
		16	9	81

$SST = \sum (y-\bar{y})^2 = 228$

S-3 $SSR = SST - SSE = 228 - 12 = 216$

S-4 $$R^2 = \frac{216}{228} = 0.947$$

8.4 Standard Error Of the estimate, S

→ R2 statistics measures the goodness of fit of the regression to the dataset.

→ 'S' is standard error of the estimate. is a measure of the accuracy of the estimates produced by the regression.

→ To find the value of 'S' we find mean square error (MSE)

$$MSE = \frac{SSE}{(n-m-1)}$$

where m indicates the no. of predictor variables, which is 1 for simple linear regression greater than 1 for multiple regression case.

the standard error of the estimate is

$$S = \sqrt{MSE} = \sqrt{\frac{SSE}{(n-m-1)}}$$

ex

$$S = \sqrt{MSE} = \sqrt{\frac{12}{(10-1-1)}} = 1.2$$

8.5 correlation coefficient (r)

The correlation coefficient r (also known as the Pearson product moment correlation coefficient) is an indication of the Strength of the linear relationship between two quantitative variables,

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

where S_x & S_y → sample standard deviations of x & y data values respectively.

Q

x	23	23	27	27	39	41	47	49	50
y	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

$n = 9$

S-1 $\bar{x} = \frac{\sum x}{n} = 36.22$, $\bar{y} = 22.74$

S-2 $\sum (x-\bar{x})(y-\bar{y}) = 571.51$

S-3

For Age x

$$S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{1019.51}{8}} = \sqrt{127.44}$$

$$= 11.29$$

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{635.25}{8}} = \sqrt{79.41}$$

$$= 8.91$$

S-4

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

$$= \frac{571.51}{(9-1)(11.29)(8.91)} = \frac{571.51}{8 \times 100.57}$$

$$= \frac{571.51}{804.56}$$

$$\simeq 0.71$$