37 Manav Pahilwan

SAIDS

Assignment - 6

Q1.

(x) Hours	(y) Score
0.5	57
0.75	64
1	59
1.25	68
1.5	74
1.75	76
2	79
2.25	83
2.5	85
2.75	86
3	88
3.25	89
3.5	90
3.75	94
4	96

Simple linear regression

$y = b_0 + b_1 x_1$

$$b_1 = \frac{\Sigma (x_1 - \bar{x})(y_1 - \bar{y})}{\Sigma (\hat{x}_1 - \bar{x})^2}$$

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

$\bar{x} = \frac{33.75}{15} = 2.25$ $\quad \bar{y} = 79.2$

$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
-1.75	-22.2	38.85	3.0625
-1.5	-15.2	22.8	2.25
-1.25	-20.2	25.25	1.5625
-1	-11.2	11.2	1
-0.75	-5.2	3.9	0.5625
-0.5	-3.2	1.6	0.025
-0.25	-0.2	0.05	0.0625
0	3.8	0	0
0.25	5.8	1.45	0.0625
0.5	6.8	3.4	0.025
0.75	8.8	6.6	0.5625
1	9.8	9.8	1
1.25	13.5	13.5	1.5625
1.5	22.2	22.2	2.25
1.75	29.4	29.4	3.0025

$\Sigma = 170$ $\quad\quad$ $\Sigma = 17.05$

$b_1 = \frac{170}{17.05}$

$b_1 = 9.97$ — (I)

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

$= 79.2 - (9.97)(2.25)$

$b_0 = 56.7675$ — (II)

$y = 56.7675 + 9.97x$ — (III)

coefficient of determination

$r^2 = \frac{SSt}{SST}$

$SSt = \Sigma(y_i - \hat{y}_i)^2$

$y = 56.7675 + 9.97(0.5) = 61.7525$

$y = 56.7675 + 9.97(0.75) = 64.255$

$y = 56.7675 + 9.97(1) = 66.7375$

$y = 56.7675 + 9.97(1.25) = 69.23$

$y = 56.7675 + 9.97(1.5) = 71.72$

$y = 56.7675 + 9.97(1.75) = 74.21$

$y = 56.7675 + 9.97(2) = 76.70$

$y = 56.7675 + 9.97(2.25) = 79.2$

$y = 56.7675 + 9.97(2.75) = 84.18$

⋮

$y = 56.7675 + 9.97(4) = 96.6475$

$$SSE = (57-61.75)^2 + (64-64.24)^2 + (39-66.75)^2 + (68-69.23)^2 + (75-71.72)^2 + (76-74.21)^2 + (79-76.70)^2 + (83-79.2)^2 + (85-81.69)^2 + (86-84.18)^2 + (88-86.67)^2 + (89-89.17)^2 + (96-91.66)^2 + (95-94.15)^2 + (96-96.65)^2$$

$$= 131.22 \longrightarrow 4$$

$$SST = \Sigma(y_i - \bar{y})^2$$
$$= 1980.28$$

$$r^2 = \frac{131.212}{1980.28}$$
$$r^2 = 0.06225$$
$$r = 0.257.$$

Q2.

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
5	8	-3.66	-6	21.96	13.39
7	9	-1.66	-5	8.3	2.755
4	12	-4.66	-2	9.32	21.715
15	26	6.34	12	76.08	40.195
12	16	3.34	2	6.68	11.155
9	13	0.34	-1	-0.34	0.115

$\bar{X} = 8.66$ $\bar{Y} = 14$ $\Sigma = 122$ $\Sigma = 89.336$

Slope $b_1 = \frac{122}{89.336}$

$b_1 = 1.3657$ — (I)

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

$= 14 - (1.3657)(8.66)$

$b_0 = 2.173$ — (II)

$SSE = \Sigma(\hat{y}_i - \hat{y}_1)^2$

$\therefore SSE = 47.363$ — (III)

$SST = \Sigma(y_1 - \bar{y})^2$

$SST = 214$ — (IV)

$SSR = SST - SSE$

$= 214 - 47.363$

$= 166.63$ — (V).

$SSR = 166.63$.

Q3. Multicollinearity happens when independent variables in regression model are highly corelated to each other. Independent variable can be predicted from another independent variable in a regression model. We would not be able to distinguish between the individual effects of independent variables on the dependent variable. It may not affect the accuracy of the model as much we might lose reliablity in determining the effect.

Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This occurs when the model is too complex. Model performs better than on the training set than on the test set. It happens when the model learns the detail and noise in the training data to extent that it negatively impacts the performance.

Q4. The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of offsets or residuals of points from the plotted curve.

Sum of squares of error should be less when there was only variable.

Step 1 :- Plot the graph b/w variables.

Step 2 :- Look for a visual line. There can be more than 1 line, select the line which gives min. residual value.

Derivation

$y = a + bx$

$e = y - (a+bx)$

Minimize

$S = \Sigma e^2 = \Sigma (y-(a+bx))^2$

$ds/da = \Sigma (y-a-bx)^2$

$= \Sigma (y-a-bx)(-1)$

$ds/db = \Sigma (y-a-bx)^2$

$= \Sigma 2(y-a-bx)(-x)$

$ds/da = 0 \qquad ds/db = 0$

$\Sigma (-2)(y-a-bx) = 0$

$\Sigma (-2x)(y-a-bx) = 0$

$$\Sigma y = na + b\Sigma x \quad \text{—(I)}$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{—(II)}$$

multiply eqn (II) by n & eqn (I) by Σx

$$\therefore b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

Divide by n^2

$$b = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

$$= \frac{cov(x,y)}{var(x)}$$

Q.5. Linear regression - It models the relationship between a dependent variable and one or more ~~in~~ explanatory variables using a linear rn

ex: predict rent based on square feet alone.

Multiple regression :- If two or more explanatory variables have a linear relationship with dependent variable. It is broader class of regressions that encompasses linear & non-linear regressions with multiple explanatory variables.

One y & two or more X.

ex. predict rent based on square foot & age of building.

Q6. In simple linear regression, mean square error to calculate error of the model.
Calculated by :-
1] measuring the distance of observed y-values from the predicted y-values at each value of x.
2] squaring each of these distances
3] Calculating mean of each of the squared distance.

$$MSE = \frac{\Sigma(y_i - \hat{y}_i)^2}{n}$$

Coefficient of determination is a statistical measurement that examines how differences in one variables can be explained by the difference in 2nd.

$$R^2 = \frac{SSE}{SST}$$

Q7.

Salary	YOE (x_1)	Age (x_2)	$x_1 \cdot x_2$	x_1^2	x_2^2	$x_1 \cdot y$	$x_2 y$
26315	18	5	90	324	25	473670	
39493	20	7	140	400	49	789860	
37209	22	8	176	484	64	818598	
24380	23	6	138	529	36	560740	
25751	23	7	161	529	49	592273	
44629	25	5	125	625	25	1115725	
37616	2	8	16	4	64	75232	
33305	28	6	168	784	36	932540	
36848	29	5	145	841	25	1068592	

42551	32	7	224	1024	49	1361632
25700	37	9	333	1369	81	950900
37303	41	6	246	1681	36	1529423
24659	46	7	322	2116	49	1134314
32617	49	8	392	2401	64	1598233
35771	53	6	318	2809	36	1895863
Σ 410517	448	100	2994	15920	688	14897585

$x_2 \cdot y$

131575	297857
276451	231300
297672	223818
146280	172613
180257	260936
223145	214626
200928	
199830	
184240	

$\bar{y} = 27367.8$

$\bar{x}_1 = 29.867$

$\bar{x}_2 = 6.667$

$\Sigma x_2 \cdot y = 313358$

$\Sigma (x_1)^2 = \Sigma x_1 \cdot x_1 = \frac{\Sigma x_1 \cdot \Sigma x_1}{N} = 18920 - \frac{(448)(448)}{15}$

$= 2539.73$

$\Sigma (x_2)^2 = 688 - \frac{(100)(100)}{15} = 21.33$

$\Sigma x_1 \cdot y = 14897595 - \frac{(448)(410517)}{15} = 2636820.6$

$\Sigma x_2 \cdot y = 396578$

$\Sigma x_1 \cdot x_2 = 7.33$

$$b_1 = \frac{(\Sigma x_2)^2 (\Sigma x_1 \cdot y) - (\Sigma x_1 \cdot \Sigma x_2)(\Sigma x_2 \cdot y)}{(\Sigma x_1)^2 (\Sigma x_2)^2 - (\Sigma x_1 \cdot x_2)^2}$$

$$= \frac{(21.33)(2636820.67) - (7.33)(396578)}{(2539.73)(21.33) - (7.33)^2}$$

$$= \frac{53336666.66}{54158373.36}$$

$b_1 = 0.984825$ — (I)

Similarly,

$$b_2 = \frac{987873159}{54158373.36}$$

$b_2 = 18.24045$ — (II)

$a = \bar{y} - b_1 \bar{x}_1 - b_2 x_2$

$a = 27367.8 - (0.9848)(29.867) - (18.24)(6.67)$

$a = 27367.8 - (29.4137) - 121.609$

$a = 27216.7773$ — (III)

$$\boxed{y = 27216.7773 + 0.9849\, x_1 + 18.24 x_2.}$$