(Lect. 39)
11.1-11.3 Regression Line

In this article we will discuss some methods of dealing with paired data on two variables. And it is limited to linear relationship between two variables only. It is a straight line relationship. Linear regression deals with methods of fitting a straight line called the regression line on a set of sample paired data on two variables. A reasonable form of relationship between the response Y and the regressor x is the linear relationship

$$Y = \beta_0 + \beta_1 x$$

where β_0 is the intercept and β_1 is the slope. The concept of regression analysis deals with finding the best relationship between Y and x, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressor x.

1 The Fitted Regression Line

An important aspect of regression analysis is, very simply, to estimate the parameters β_0 and β_1 (i.e., estimate the so-called regression coefficients). Suppose we denote the estimates b_0 for β_0 and b_1 for β_1. Then the estimated or fitted regression line is given by

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ is the predicted or fitted value.

1.1 Estimating the Regression Coefficients

Given the sample (x_i, y_i); $i = 1, 2, ..., n$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}$$

Question No.2

The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Estimate the linear regression line.
(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.

Solution

x	y	xy	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
77	82	6314	-1.55	8.89	-13.7795	2.4025
50	66	3300	-28.55	-7.11	202.9905	815.1025
71	78	5538	-7.55	4.89	-36.9195	57.0025
72	34	2448	-6.55	-39.11	256.1705	42.9025
81	47	3807	2.45	-26.11	-63.9695	6.0025
94	85	7990	15.45	11.89	183.7005	238.7025
96	99	9504	17.45	25.89	451.7805	304.5025
99	99	9801	20.45	25.89	529.4505	418.2025
67	68	4556	-11.55	-5.11	59.0205	133.4025
$\sum_{i=1}^{9} x_i = 707$	$\sum_{i=1}^{9} y_i = 658$				1568.4445	2018.2225

(1.1)

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 1568.4445$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 2018.2225$$

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i}{9} = \frac{707}{9} = 78.55$$

$$\bar{y} = \frac{\sum_{i=1}^{9} y_i}{9} = \frac{658}{9} = 73.11$$

(a)Regression line is

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1568.4445}{2018.2225} = 0.7771415$$

$$b_0 = \bar{y} - b_1\bar{x} = 73.11 - 0.7771415(78.55) = 12.06553518$$

Now Regression line is

$$\hat{y} = b_0 + b_1 x = 12.06553518 + 0.7771415x$$

(b) The final examination grade of a student who received a grade of 85 on the midterm report is

$$\hat{y} = b_0 + b_1 x = 12.06553518 + 0.7771415x \ for \ x = 85$$

Now for x=85,

$$\hat{y} = b_0 + b_1 x = 12.06553518 + 0.7771415(85) = 78.1225$$

Question No.5

A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature, $\boldsymbol{x}$	Converted Sugar, $\boldsymbol{y}$
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

(a) Estimate the linear regression line.

(b) Estimate the mean amount of converted sugar produced when the coded temperature is 1.75

Solution

x	y	xy	x^2
1.0	8.1	8.1	1
1.1	7.8	8.85	1.21
1.2	8.5	10.2	72.25
1.3	9.8	12.71	1.69
1.4	9.5	13.3	1.96
1.5	8.9	13.35	2.25
1.6	8.6	13.76	2.56
1.7	10.2	17.34	2.89
1.8	9.3	16.74	3.24
1.9	9.2	17.48	3.61
2.0	10.5	21	4
$\sum_{i=1}^{11} x_i = 16.5$	$\sum_{i=1}^{11} y_i = 100.4$	$\sum_{i=1}^{11} x_i y_i = 152.59$	$\sum_{i=1}^{11} {x_i}^2 = 25.85$

(1.2)

Therefore,

$$\bar{x} = \frac{\sum_{i=1}^{11} x_i}{11} = \frac{16.5}{11} = 1.5$$

$$\bar{y} = \frac{\sum_{i=1}^{11} y_i}{11} = \frac{104.4}{11} = 9.4909$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{(11)(152.59) - (16.5)(100.4)}{(11)(25.85) - (16.5)^2} = 1.8091 \qquad (1.3)$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \frac{100.4 - (1.8091)(16.5)}{11} = 6.4136$$

Hence

$$\overset{\Lambda}{y} = 6.4136 + 1.8091x$$

(b) For x = 1.75,

$$\overset{\Lambda}{y} = 6.4136 + (1.8091)(1.75) = 9.580$$

Question No.7

The following is a portion of a classic data set called the "pilot plot data" in Fitting Equations to Data by Daniel and Wood, published in 1971 . The response y is the acid content of material produced by titration, whereas the regressor x is the organic acid content produced by extraction and weighing.

y	x	y	x
76	123	70	109
62	55	37	48
66	100	82	138
58	75	88	164
88	159	43	28

Fit a simple linear regression; estimate a slope and intercept.

Solution

x	y	xy	x^2
123	76	9348	15129
55	62	3410	3025
100	66	6600	10000
75	58	4350	5625
159	88	13992	25281
109	70	7630	11881
48	37	1776	2304
138	82	11316	19044
164	88	14432	26896
28	43	1204	784
$\sum_{i=1}^{10} x_i = 999$	$\sum_{i=1}^{10} y_i = 670$	$\sum_{i=1}^{10} x_i y_i = 74,058,$	$\sum_{i=1}^{11} {x_i}^2 = 119,969$

(1.4)

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{(10)(74,058) - (999)(670)}{(10)(119,969) - (999)^2} = 0.3533 \tag{1.5}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} = \frac{670 - (0.3533)(999)}{10} = 31.71$$

Hence

$$\hat{y} = 31.71 + 0.3533x.$$