STAT-1302; Lecture 17; π-day, 2024

Ch. 13 ...

The Simple Linear Regression Model

$$Y_i = A + BX_i + \varepsilon_i \quad , \quad i = 1, \ldots, n$$

where $\varepsilon_i \overset{\text{independent}}{\sim} N(0, \sigma^2)$

Epsilon

Parameters :

A = intercept

B = Slope

σ^2 = noise or error variance

Notes:

1. X_i (independent variable) is a fixed variable (ie. no randomness). It is part of the data collected.

2. Y_i (dependent or response variable) is a

random variable. It is also part of the data collected.

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

3. ε_i is a random variable that is meant to capture any variation in Y_i that has not been captured by x_i.

Interpretation of B:

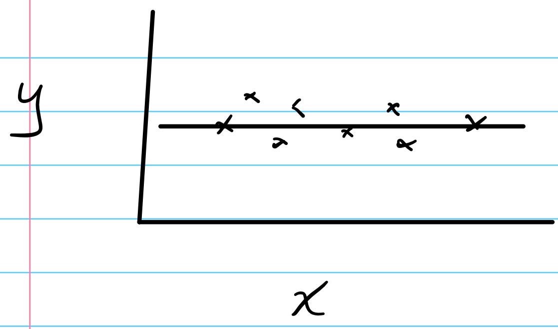
$B > 0 \Rightarrow$ a positive linear relationship between X and Y.

(ie. as X increases, then so does Y).

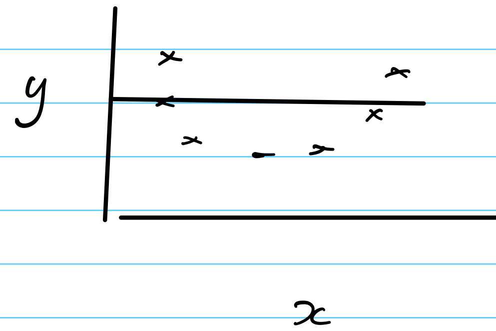
$B < 0 \Rightarrow$ a negative linear relationship between X and Y.

(ie. as X increases, then Y decreases.)

$B = 0 \Rightarrow$ no linear relationship between X and Y.



$B = 0$



$B = 0$

no linear relationship between X and y

More on Interpretation of B:

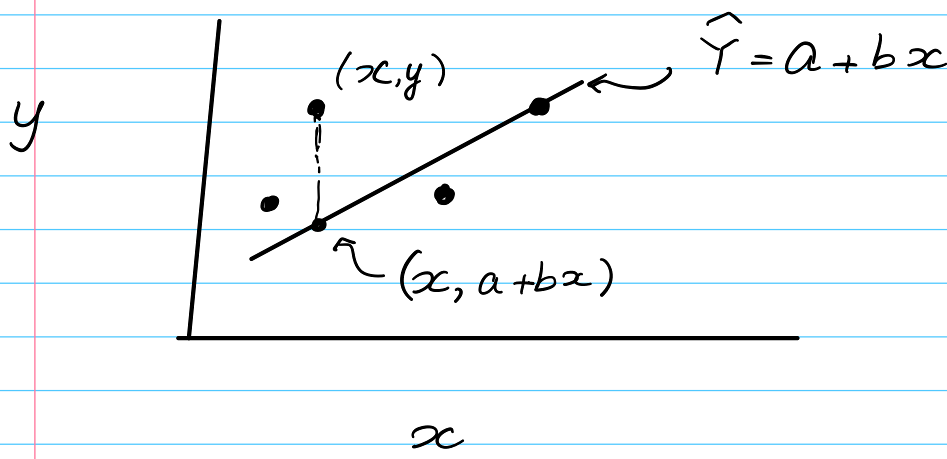
MCQ

- B is the change in the mean of Y when X is increased by one unit.

Fitted/Estimate Simple linear regression model:

$$\hat{Y} = a + bX$$

where "a" and "b" are the least squares estimators of A and B, respectively.



The least squares regression line is found by minimizing $\sum_{i=1}^{n} (Y_i - (a+bX_i))^2$ with respect to a and b. (Calculus Problem).

$$b = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}.$$ Given on formula sheet.

To compute "b" use the short-cut formula given on formula sheet:

$$b = \frac{SS_{xy}}{SS_{xx}}$$ where

Given

$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

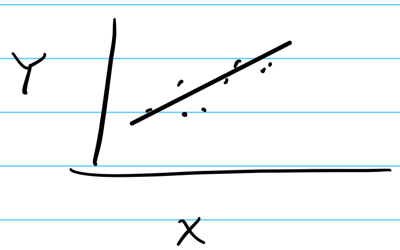
$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

n = total no. of observations :

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Ex. Y = food expenditure

X = monthly income (in hundreds of dollars)



Summary Statistics:

$\sum x = 212$, $\sum y = 64$, $\sum xy = 2150$, $\sum x^2 = 7222$.

$n = 7$

a) Find the least squares regression line where food expenditure (y) is regressed against monthly income (x).

$$b = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum_{i=1}^{7} x_i \cdot y_i - \frac{\left(\sum_{i=1}^{7} x_i\right)\left(\sum_{i=1}^{7} y_i\right)}{7}$$

$$= 2150 - \frac{212(64)}{7} = 211.7143$$

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{7} \qquad n = 7$$

$$= 7222 - \frac{212^2}{7} = 801.4286$$

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{211.7143}{801.4286} = 0.2642$$

$$a = \overline{Y} - b\overline{X} = \frac{\sum y}{7} - 0.2642 \times \frac{\sum x}{7}$$

$$= \frac{64}{7} - 0.2642 \times \frac{212}{7} = 1.1414$$

The estimated least Squares regression line is

$$\hat{Y} = 1.1414 + 0.2642 \times X$$

must write "^" in least Squares regression line

Or,

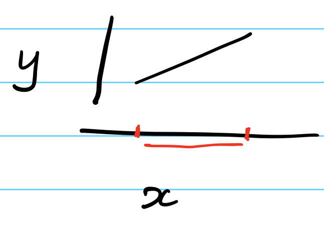
$$\widehat{\text{Expenditure}} = 1.1414 + 0.2642 \times \text{Income}$$

b) What is the predicted monthly food expenditure for a household whose monthly income is 10 (i.e. \$1,000) ?

$$\widehat{\text{Expenditure}} = 1.1414 + 0.2642 \times 10$$

$$= 3.7834$$

i.e. \$378.34. (x is in hundreds of dollars).



Residual:

$$e_i = Y_i - \hat{Y}_i$$

Y_i: Observed response

$\hat{Y}_i$: predicted response for regression line.

Ex. c) What is the residual when $X = 35$ and $y = 9$?

$$\hat{Y} = 1.1414 + 0.2642(35) = 10.3884$$

$$e = Y - \hat{Y} = 9 - 10.3884 = -1.3884.$$

That is, the expenditure is overestimated by \$138.84 when monthly income is \$3500.

[Extra: STAT_3103.

We use e_i's to assess the adequacy to the fitted regression line.

$$\sum e_i \cong 0 \Rightarrow E(\varepsilon_i) = 0 \text{ assumption holds.}$$]

Interpretation of a:

"a" is the mean of the response variable when $X = 0$.

Assumptions of the Simple linear regression model:

$$Y = A + BX + \mathcal{E}$$

1. $\mathcal{E}$ is a random variable that has mean zero at each x.

2. $\mathcal{E}_1, \dots, \mathcal{E}_n$, the errors associated with each of $x_1, \dots, x_n$, respectively are independent.

3. For any given x, the errors $\mathcal{E}_1, \dots, \mathcal{E}_n$, are normally distributed.

4. $\mathcal{E}_1, \dots, \mathcal{E}_n$ have Constant variance σ^2.

[i.e. $Y_i = A + BX_i + \mathcal{E}_i$ where

$\mathcal{E}_i \overset{ind}{\sim} N(0, \sigma^2)$ for $i = 1, \dots, n$.]

Ex. Data for understanding how the Price of a particular Car model depreciates with age are as follows.

Age	8	3	6	9	2	5	6	3
Price	45	210	100	33	267	134	109	235

Age (x): in years
Price (y): in hundreds of dollars.

a) Determine the estimated least squares regression line when price is regressed against age.

Given: $\sum xy = 4549$, $\sum x^2 = 264$, $\sum y = 1144$, $\sum x = 42$; $n = 8$.

$$b = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$= 4549 - 42(1144)/8 = -1457$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 264 - \frac{42^2}{8} = 43.5$$

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{-1457}{43.5} = -33.49$$

$$a = \bar{Y} - b\bar{X} = \frac{1144}{8} - (-33.49)\frac{42}{8}$$

$$= 318.83$$

$$\widehat{Price} = 318.83 - 33.49 \times Age$$

b) Give a brief interpretation of "a" and "b" in part (a).