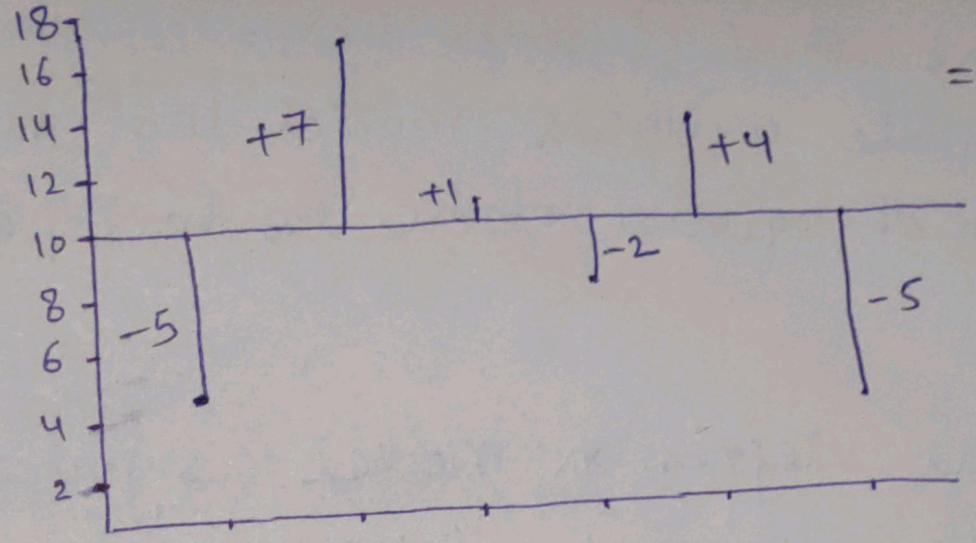
Linear Regression

Problem

Let's assume that you are a small restaurant owner or a very business minded server/waiter at a nice restaurant. The "tips" are a very important part of a waiter's pay. Most of the time the dollar amt. of the tip is related to the dollar amt. of the bill.

As the waiter or owner, you would like to develop a model that will allow you to make a prediction about what amt. of tip to expect for any given bill amt. Therefore one evening, you collect data for six meals.

* If you have only one variable, then the prediction for the next measurement is the mean of the sample itself.
* The variability in the tip amt. can only be explained by the tip themselves (In case of only one variable)
* Dependent variable will always be placed at yaxis.



Take a mean $\frac{5+17+11+8+14+5}{6}$

$= 10$

Mean will always bifurcate your data into two equal partitions.

Now we are calculating the distance of each data points from the mean.

Residuals: The distance from the bestfit lines to the observed value is called residuals / Errors.

1) we square them to make them +ve.
2) Emphasizes larger deviations

Now, Sum of Square Residual / Error = 120

Note: The Goal of simple linear regression is to create a linear model that minimizes the sum of squares of the residuals/ error (SSR)

⇒ If our regression model is significant, it will "eat up" much of the raw SSE we had when we assumed that the independent variable doesn't even exists.

⇒ The regression line should literally "fit" the data better.

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

β_0 = y-intercept population parameter

β_1 = slope population parameter

ε = error term, unexplained variation in y.

- Intercept Slope form of a line

$$y = mx + c$$

m = slope

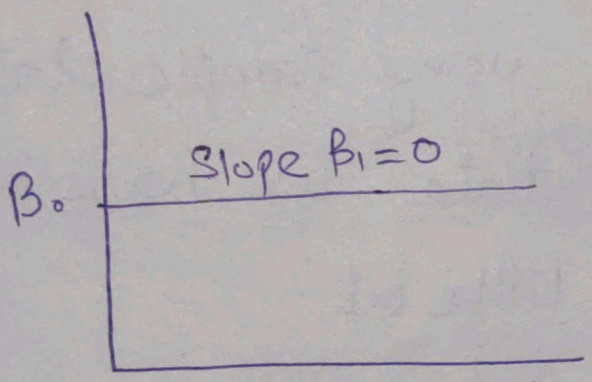
c = intercept

Simple Linear Regression

$$E(y) = \beta_0 + \beta_1 x$$

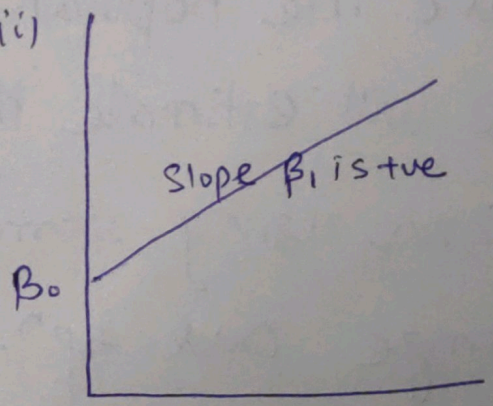
Expected value of y = is the mean or distribution around y for a given value of x.

(i)



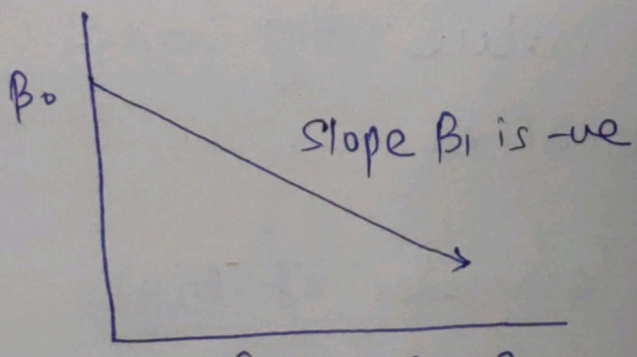
$E(y) = \beta_0 + 0(x)$

(ii)



$E(y) = \beta_0 + \beta_1 x$

(iii)



$E(y) = \beta_0 - \beta_1 x$

Imp:-

If we actually knew the population parameters β_0 and β_1 we could use the simple linear Regression eqn.

$$E(y) = \beta_0 + \beta_1 x$$

However, in reality we almost never have the population paremeters. Therefore we will estimate them using sample data. When using sample data we have to change our eqn. a little bit.

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

$\hat{y}$ = the mean value of y for a given value of x.

↳ is the point estimator of $E(y)$

Hypothesis:-

Tip amount depends on Bill amount. So in general lower bill will resolve a lower tip, and more expensive bill will resolve a higher tip.

Bill ($)	Tip ($)
34	5
108	17
64	11
88	8
99	14
51	5

Least Square Criterion

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

y_i = Observed value of dependent variable (tip amt)

$\hat{y}_i$ = estimated / predicted value of the dependent variable (predicted tip amt.)

The goal is to minimize the sum of squared difference b/w the observed and the estimate value of dependent variable

Calculation :-

Step 1 : Scatter plot

Meal amt V/S Tip amt

Step 2 : Look for a visual line

Does the data seems to fully along a line?

In this case Yes! Proceed.

If NOT, then it would be no linear pattern.

Step 3 : Correlation (optional)

Step 4 : Descriptive Statistics / Centroid

First of all find the meam of each Variable

$$\bar{x} = 74 \qquad \bar{y} = 10$$

The best fit regression line will / must pass through the centroid.

Step 5:- Calculation

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

$$\text{Slope} \Rightarrow \quad b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$\bar{x}$ = mean of the independent variable

x_i = value " " " "

$\bar{y}$ = mean of the dependent variables

y_i = value " " " "

$$\text{Intercept} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Note:- $SST = SSR + SSE$

$SSR = SST - SSE$

Coefficient of determination

$$r^2 = \frac{SSR}{SST} = \frac{89.925}{120}$$

$$= 0.7493$$

or 74.93 %.

We can conclude that 74.93 % of the total sum of square can be explained by using the estimated regression eqn: to predict the tip amt. The remainder is an error.