In simple linear regression, the model is between a single dependent variable (explanatory variable) and a single independent (study) variable.

Let us denote the independent variable (IV) by 'x' and the dependent variable (DV) by 'y'.

We collect paired observations $(x_1, y_1), (x_2, y_2) \cdots (x_n y_n)$ on the variables X and Y.

The relationship between these two variables x and y can be used to predict one (y) when the other (x) is known.

The basic idea is to draw a line through the points of the scatterplot (fig 1), in such a way that this line best approximates the relationship between the two variables (x, y).

This line is then used for prediction.

## Least Squares fit

The objective is to find the line through points $(x_i, y_i)$ that minimizes the sum of squares of the difference between each point and the line in the vertical direction (fig 2)

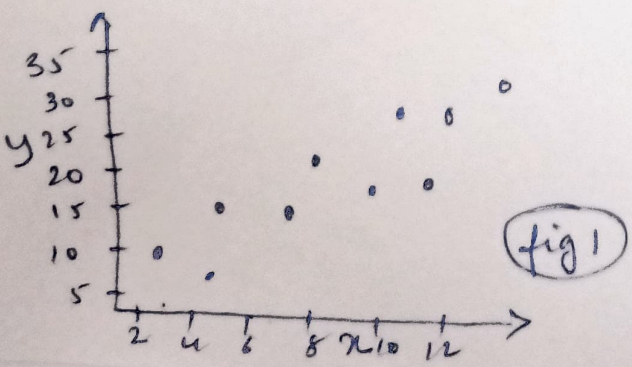


fig 1

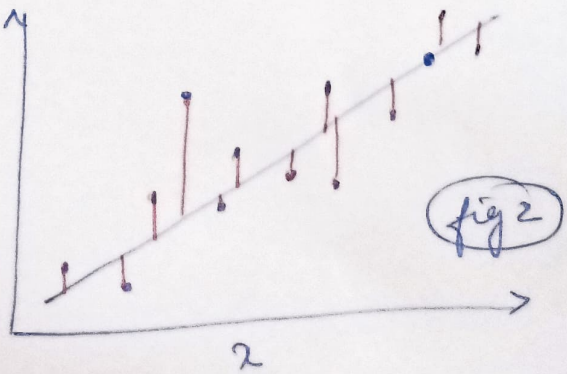


fig 2

The following is the relation between x and y using a straight line.

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

$\beta_0$ – y intercept of regression line. i.e value of y when $x = 0$.

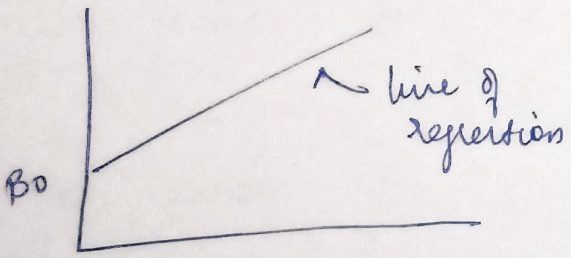
$\beta_1$ – slope of regression line

$\epsilon$ – error term (represents the difference in the linear model and a particular observed value for y.

e.g Suppose we want build a linear regression model that estimates a weight gain as a function of number of hours of television viewing.

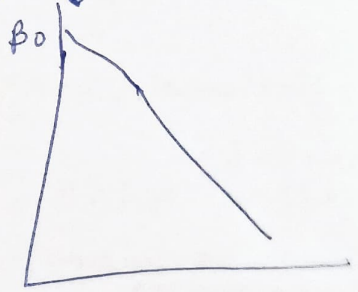
The model would be expressed as :-

weight gain = $\beta_0 + \beta_1$ (hours of tv. viewing) $+ \epsilon$.



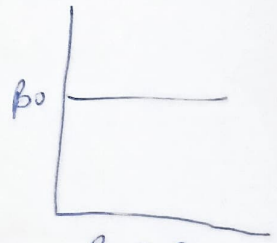

$\beta_1$ is +ve (slope)

∴ increasing relationship




$\beta_1$ → slope is -ve.

as x ↑ es, y ↓ es.




$\beta_1 = 0$

No relationship.

Let $b_0$ and $b_1$ be sample statistics used to estimate $\beta_0$ and $\beta_1$ respectively.

where $\beta_0$ and $\beta_1$ are population parameters.

Then, the estimated line of regression is

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

$\hat{y}$ - predicted value of $y$ given $x$

$b_0$ - $y$ intercept

$b_1$ - slope of the line.

Best fit line

The distance between $\hat{y}_i$ and $y_i$ should be minimized.

i.e $\min \Sigma (y_i - \hat{y}_i)$

Consider the following data $(x, y)$

| $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 2 | 69 | −2.8 | −8.8 | 24.64 | 7.84 |
| 9 | 98 | 4.2 | 20.2 | 84.84 | 17.64 |
| 5 | 82 | 0.2 | 4.2 | 0.84 | 0.04 |
| 5 | 77 | 0.2 | −0.8 | −0.16 | 0.04 |
| 3 | 71 | −1.8 | −6.8 | 12.24 | 3.24 |
| 7 | 84 | 2.2 | 6.2 | 13.64 | 4.84 |
| 1 | 55 | −3.8 | −22.8 | 86.64 | 14.44 |
| 8 | 94 | 3.2 | 16.2 | 51.84 | 10.24 |
| 6 | 84 | 1.2 | 6.2 | 7.44 | 1.44 |
| 2 | 64 | −2.8 | −13.2 | 38.64 | 7.84 |

$\Sigma x = 48$

$\bar{x} = 48/10 = 4.8$

$\Sigma y = 778$

$\bar{y} = \frac{778}{10} = 77.8$

$\Sigma (x - \bar{x})(y - \bar{y}) = 320.6$

$\Sigma (x_i - \bar{x})^2 = 67.6$

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

$$= \frac{320.6}{67.6} = 4.74$$

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

$$= 77.8 - (4.74)(4.8)$$

$$= 55.048$$

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

$\hat{y} = 55.048 + 4.74x$ is the estimated regression line.

a) Correlation coefficient

$$r = \frac{S_x}{S_y} \times b_1$$

$$S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \qquad S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$$

$$= \sqrt{\frac{67.6}{10-1}} = 2.6 \qquad = 12.647$$

$$r = \frac{2.6}{12.647} \times 4.74. = 0.974.$$

correlation between $x$ and $y$ is 97.4%.

b) Predict $y$ for given $x$.

$x = 3$

$$\hat{y} = 55.048 + 4.74 \times 3$$

$$= 69.268$$

c) Coefficient of determination

- we use $x_i$ to explain as much variation in $y_i$ as possible.
- how well does the regression line fit the data.

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

- SSR (sum of squares of regression)
  - $\Sigma(\hat{y}_i - \bar{y})^2$
  - that part of the total variation in $y$ about its sample mean that is explained by the fitted line.

SST (total sum of squares)

- $\Sigma (y_i - \bar{y})^2$
- a measure of the total variation in y around its sample mean.

SSE (Error sum of squares)

- $\Sigma (y_i - \hat{y})^2$
- that part of total variation in y about its sample mean that is not explained by a fitted line.

$$SST = SSR + SSE$$

Note.

$$SST = SSR + SSE$$

$$\therefore SSR = SST - SSE$$

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

| $x_i$ | $y_i$ | $\hat{y}_i$ | $y_i - \hat{y}_i$ | $(y_i - \hat{y})^2$ | $y_i - \bar{y}$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 2 | 69 | 64.528 | 4.472 | 19.9988 | -8.8 | 77.44 |
| 9 | 98 | 97.708 | 0.292 | 0.0852 | 20.2 | 408.04 |
| 5 | 82 | 78.748 | 3.252 | 10.5755 | 4.2 | 17.64 |
| 5 | 77 | 78.748 | -1.748 | 3.0555 | -0.8 | 0.64 |
| 3 | 71 | 69.268 | 1.732 | 2.9998 | -6.8 | 46.24 |
| 7 | 84 | 88.228 | -4.228 | 17.8759 | 6.2 | 38.44 |
| 1 | 55 | 59.788 | -4.788 | 22.9249 | -22.8 | 519.84 |
| 8 | 94 | 92.968 | 1.032 | 1.0650 | 16.2 | 262.44 |
| 6 | 84 | 83.468 | 0.612 | 0.2621 | 6.2 | 38.44 |
| 2 | 64 | 64.528 | 0.528 | 0.2788 | -13.8 | 190.44 |

$SSE = 79.1215$ $\qquad SST = 1599.6$

$$\hat{y} = 55.048 + 4.74(x)$$

$$r^2 = SSR/SST = (SST - SSE)/SST$$

$$= \frac{1599.6 - 79.1215}{1599.6}$$

$$= \frac{1520.4785}{1599.6}$$

$$= 0.9505$$

- $r^2$ measures the % of variability in y that can be explained by variable x.

- ∴ $r^2 = 95.05\%$ means about 95.05% of the variation in y is explained by the variations in x.

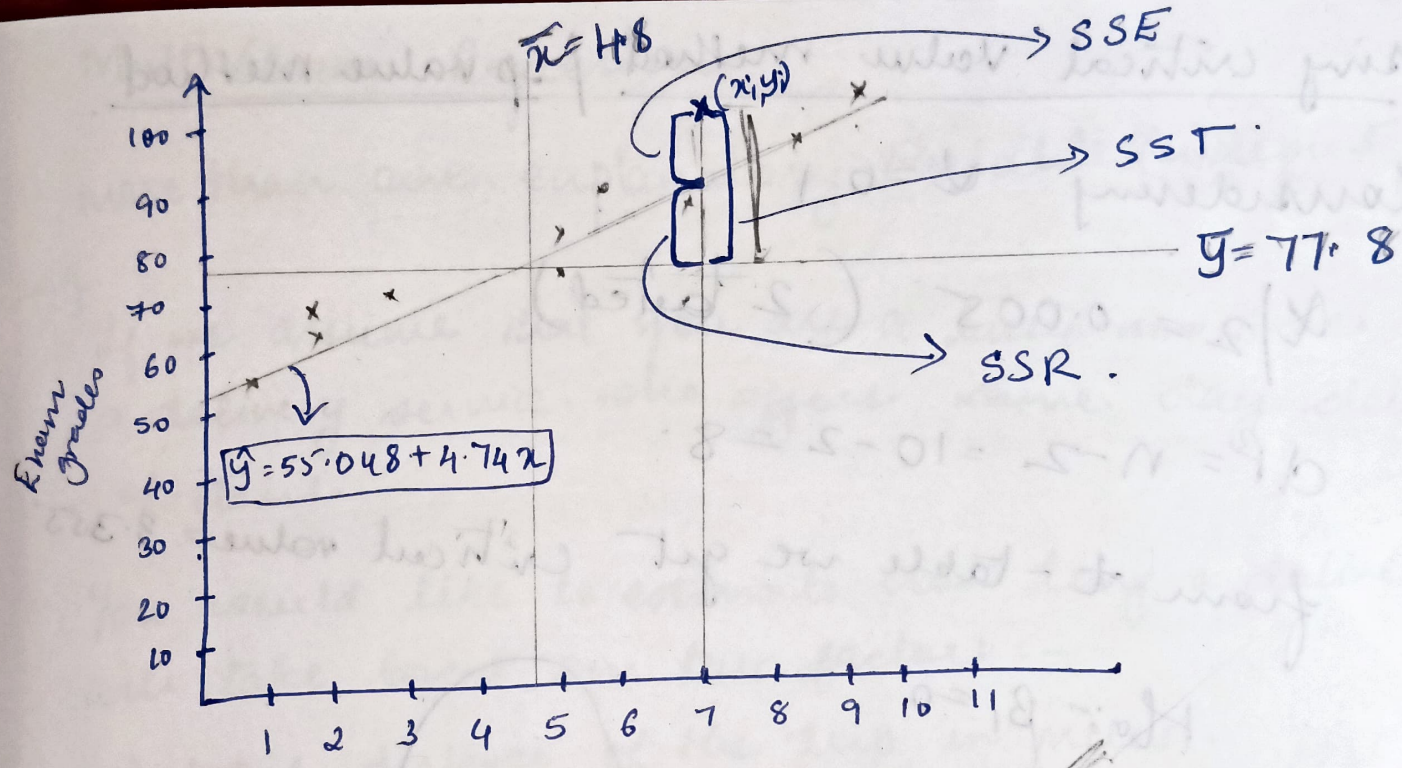
or

x = # hours student studied

y = student grade.

then,

95.05% of the variation in grades is because of variations in hours student studied


$\bar{x} = 4.8$
SSE
$(x_i, y_i)$
SST
$\bar{y} = 77.8$
SSR
$\hat{y} = 55.048 + 4.74x$
Exam grades
100
90
80
70
60
50
40
30
20
10
1 2 3 4 5 6 7 8 9 10 11