Simple Linear Regression

Example

A firm has a chain of pizza restaurants around the country. To see the effectiveness of its advertising activities, it has collected the data from 19 randomly selected metropolitan regions. There are two variables in the data:

- Promote: Promotional Expenditure in thousand rupees
- Sales in thousand rupees

We are interested in building the relationship between the two variables.

Simple Linear Regression Model

$$\textbf{\textit{Sales}} = \beta_0 + \beta_1 \textbf{\textit{Promote}} + \varepsilon$$

Sales is a linear function of Promote plus ε

β_0 and β_1 are parameters of the model, ε is a random variable.

The linear term of $\beta_0 + \beta_1$ ***Promote*** is the **variations in Sales that can be explained by Promote**

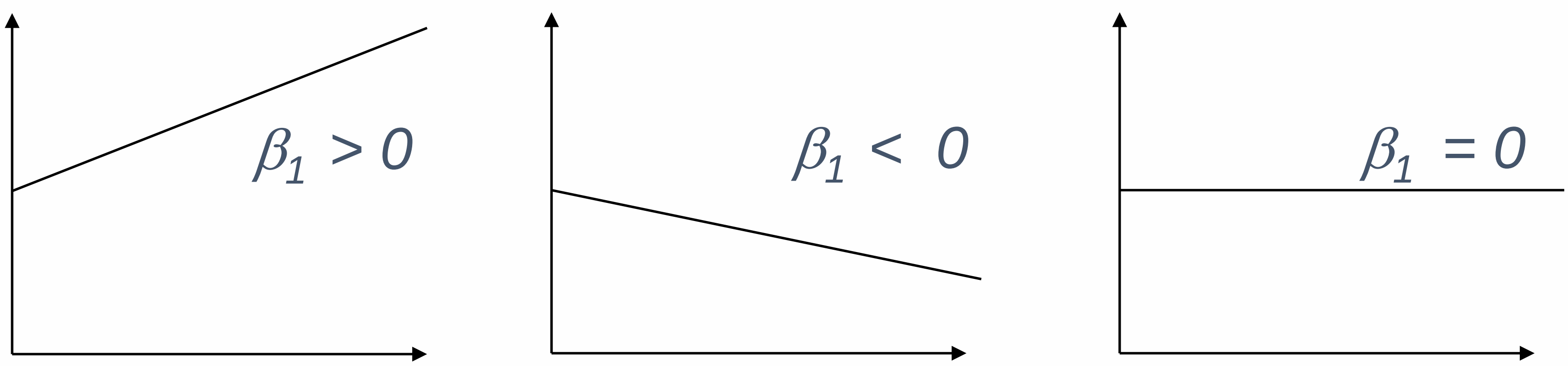
The error term of ε *is* variations in *Sales* that **can not be explained** by the **liner relationship** between *Promote* and *Sales.*

Simple Linear Regression Equation

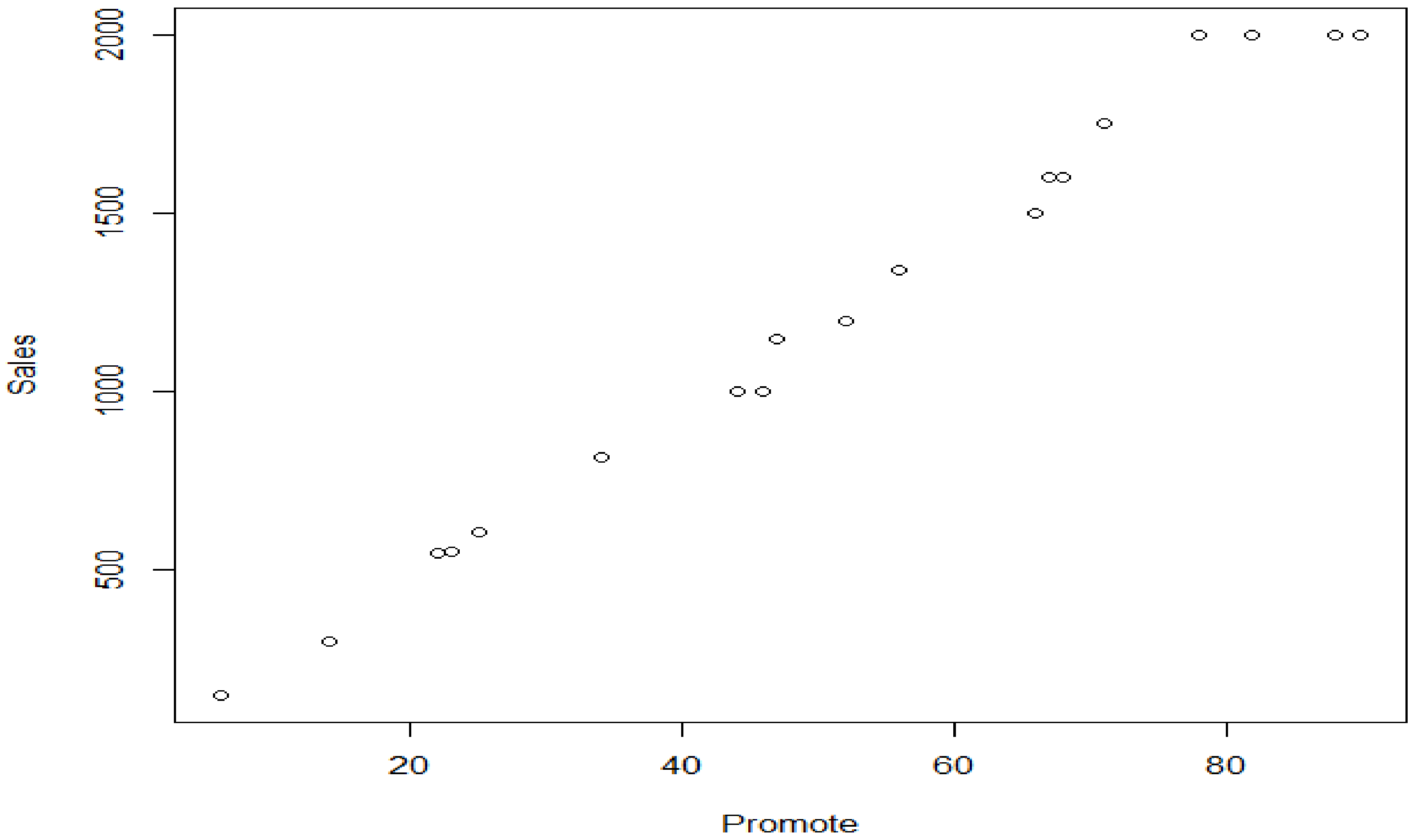
For the time being let us forget ε. The following equation describes **how the mean value of Sales is related to Promote.**

$$\textit{Expected Sales} = \beta_0 + \beta_1 \textit{Promote}$$

β_0 is the intersection with y axis, β_1 is the slope.



Scatter Diagram



Estimated Linear Regression Equation

We want to estimate the relationship between

?

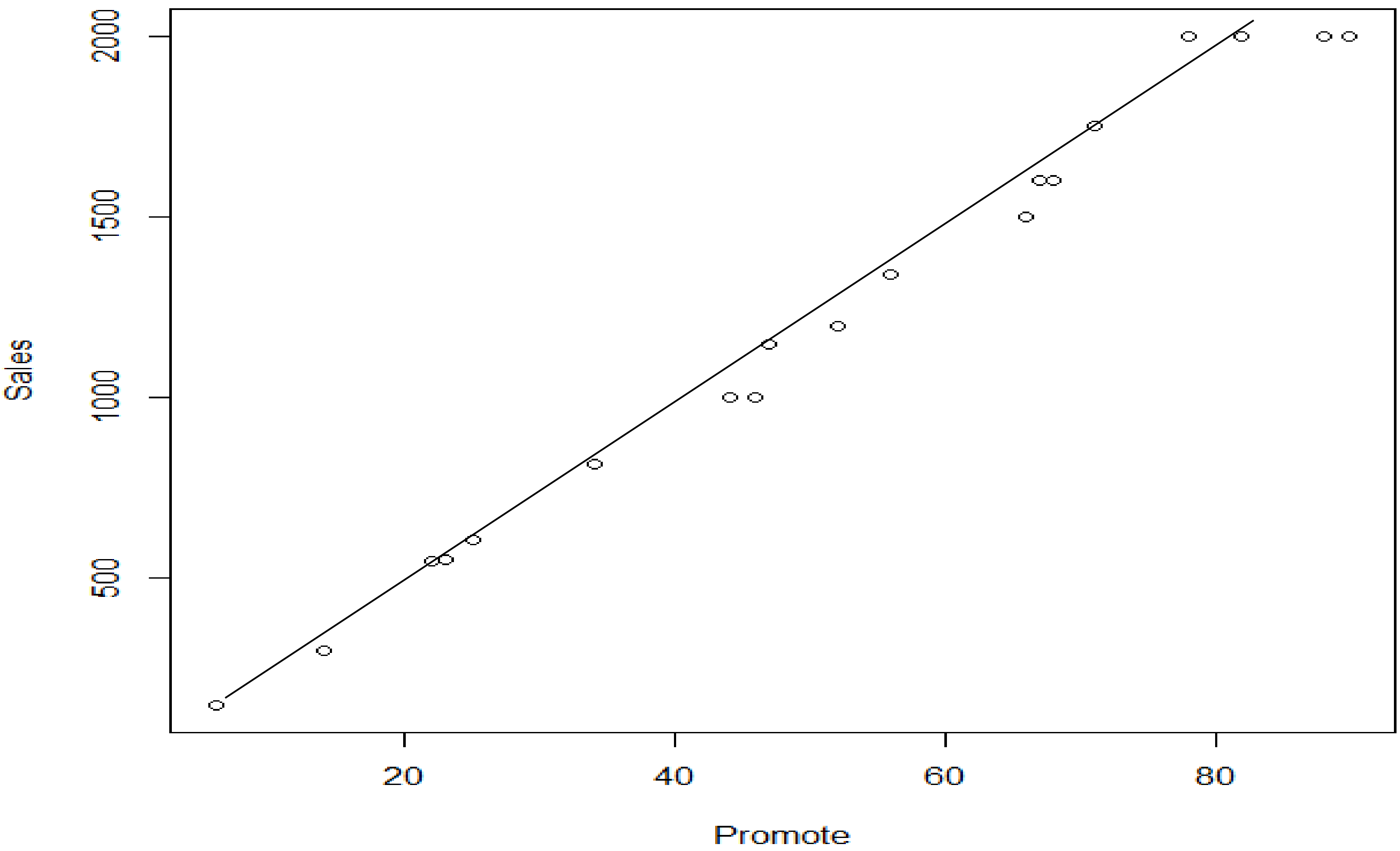
?

Promotional Expenditure   Mean value of sales

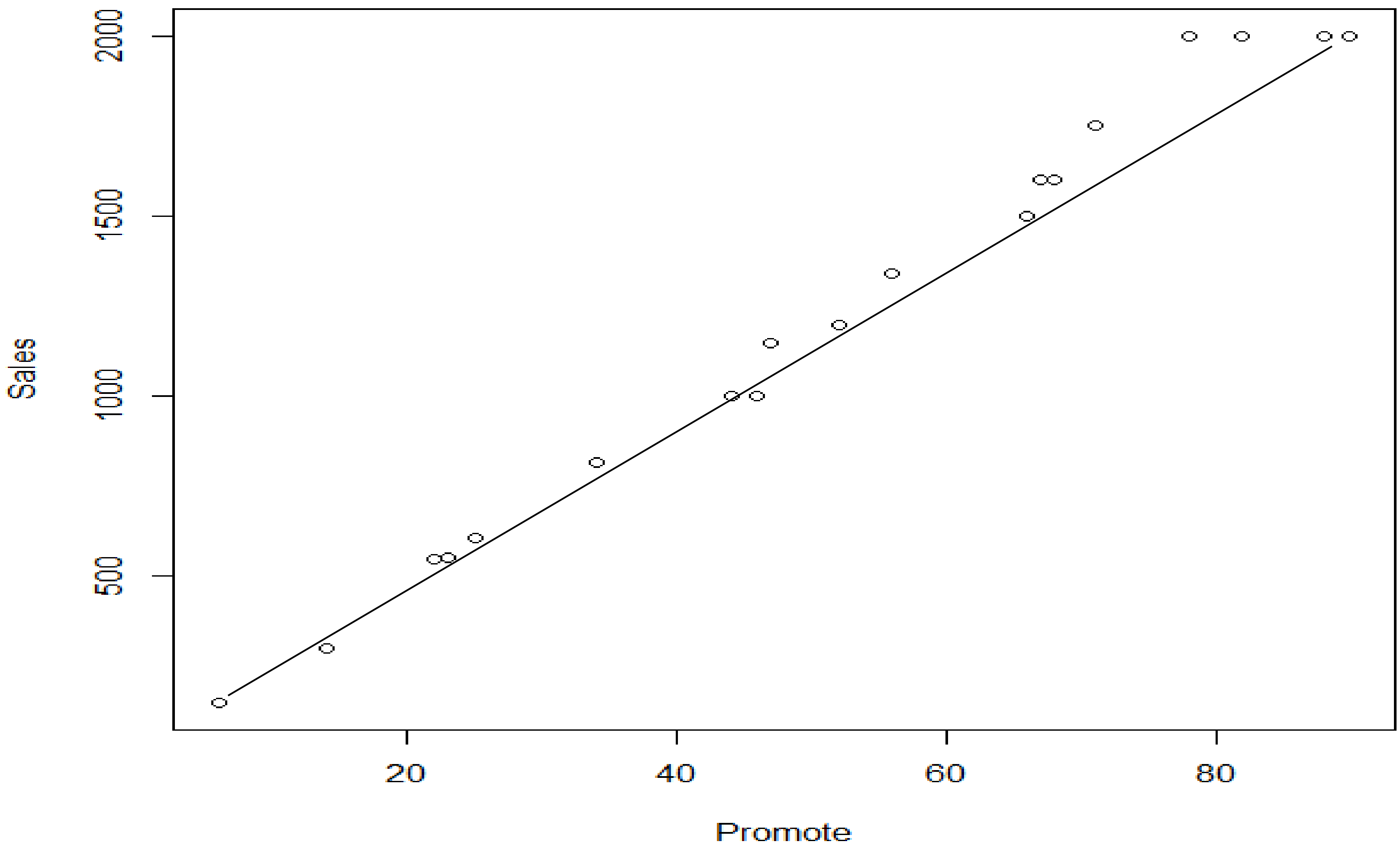
We may rely on own judgement, and draw a line to fit them.

Then we measure the intersection with y axis and that is b_0, and the slope is b_1

Judgmental Solution 1



Judgmental Solution 2

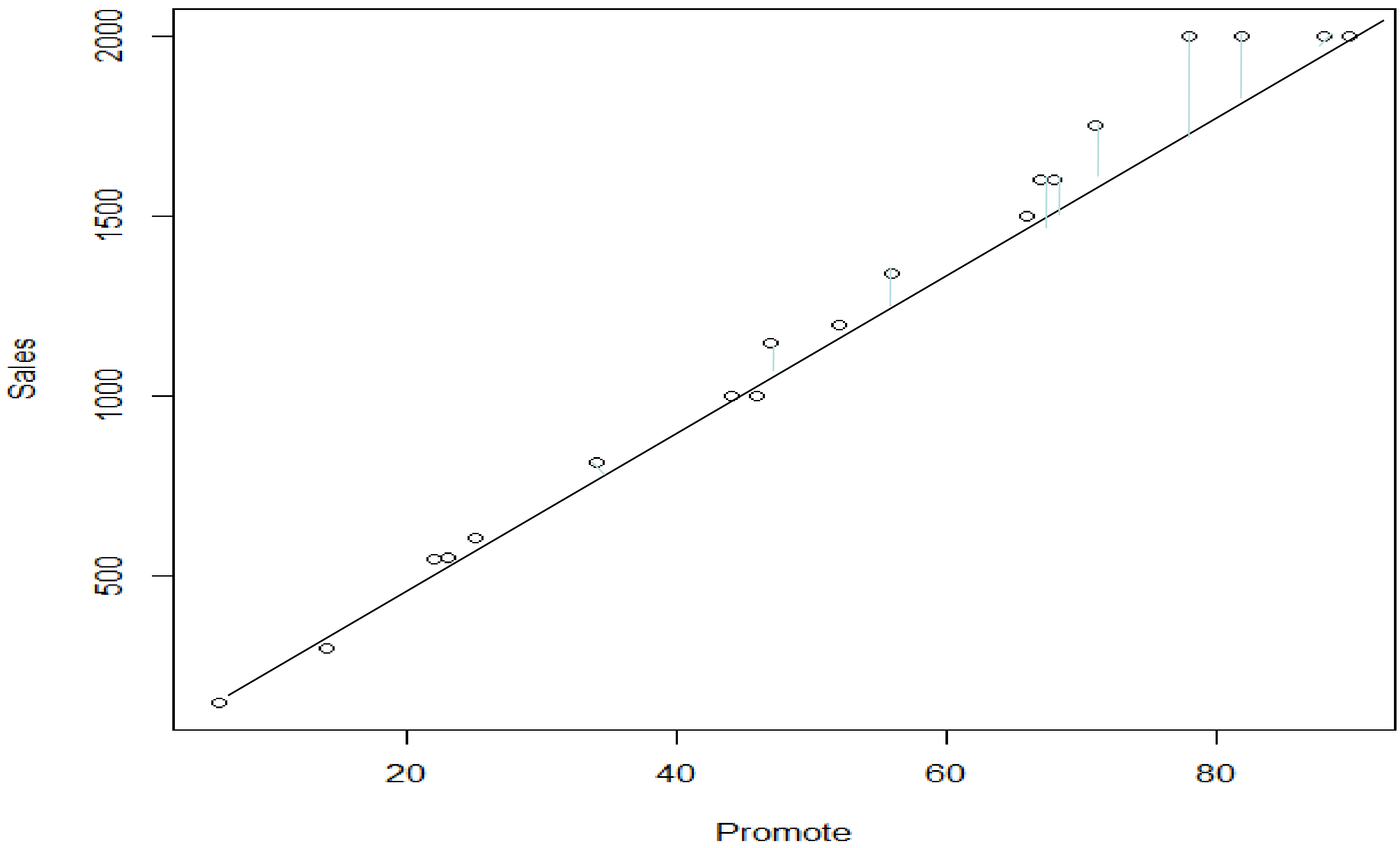


The Least Square Method

Judgmental Solution can’t be a standard approach as it may be person dependent.

We need to use algebra and calculus for correctly calculating the optimal line.

Hence we follow **The Least Square Method** approach.


Sales
500
1000
1500
2000
20
40
60
80
Promote

The Least Square Method

y_i	x_i	$\hat{y}_i$
y_1	x_1	$b_0 + b_1 x_1$
y_2	x_2	$b_0 + b_1 x_2$
y_3	x_3	$b_0 + b_1 x_3$
.	.	.
.	.	.
y_n	x_n	$b_0 + b_1 x_n$

$$Min \qquad Z = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$Min \qquad Z = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Classic Minimization

$$Min \qquad Z = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

We want to minimize this function with respect to b_0 and b_1
This is a optimization problem.

We may remember from high school algebra that to find the minimum value we should get the derivative and set it equal to zero.

The Least Square Method

Note : Our unknowns are b_0 and b_1 .
x_i and y_i are known. They are our data.

y_i	x_i	$\hat{y}_i$
y_1	x_1	$b_0 + b_1 x_1$
y_2	x_2	$b_0 + b_1 x_2$
y_3	x_3	$b_0 + b_1 x_3$
.	.	.
.	.	.
y_n	x_n	$b_0 + b_1 x_n$

$$Z = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Find the derivative of Z with respect to b_0 and b_1 and set them equal to zero

Derivatives

$$Z=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$$

$$\frac{\partial Z}{\partial b_0}=\sum_{i=1}^{n}2(-1)(y_i-b_0-b_1x_i)=0$$

$$\frac{\partial Z}{\partial b_1}=\sum_{i=1}^{n}2(-x_i)(y_i-b_0-b_1x_i)=0$$

b_0 and b_1

$$b_1 = \frac{\sum xy - (\sum x \sum y)/n}{\sum x^2 - (\sum x)^2/n}$$

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

Example

Promote(X)	Sales (Y)	XY	X square	Y Square
23	554	12742	529	306916
56	1339	74984	3136	1792921
34	815	27710	1156	664225
25	609	15225	625	370881
67	1600	107200	4489	2560000
82	2000	164000	6724	4000000
46	1000	46000	2116	1000000
14	300	4200	196	90000
6	150	900	36	22500
47	1150	54050	2209	1322500
52	1200	62400	2704	1440000
88	2000	176000	7744	4000000
71	1750	124250	5041	3062500
78	2000	156000	6084	4000000
66	1500	99000	4356	2250000
44	1000	44000	1936	1000000
68	1600	108800	4624	2560000
90	2000	180000	8100	4000000
22	550	12100	484	302500

Totals	**979**	**23117**	**1469561**	**62289**	**34744943**

b_1

$$b_1 = \frac{\sum xy - (\sum x \sum y)/n}{\sum x^2 - (\sum x)^2/n}$$

$$b_1 = 23.506$$

b_0

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

$$\bar{y} = \frac{23117}{20} = 1155.85$$

$$\bar{x} = \frac{979}{20} = 48.95$$

$$1155.85 = b_0 + 23.506(48.95)$$

$$b_0 = 5.48$$

Estimated Regression Equation

$$y = 5.48 + 23.51x$$

Now we can predict.
For example, if one of restaurants of this Pizza Chain is having an expenditure of 72
We predict the mean of its quarterly sales is

$$y = 5.48 + 23.51(72)$$

$$y = 1697.95 \qquad thousand\ \ rupees$$

Summary : The Simple Linear Regression Model

- Simple Linear Regression Model

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

- Simple Linear Regression Equation

$$\mathrm{E}(y) = \beta_0 + \beta_1 x$$

- Estimated Simple Linear Regression Equation

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

Summary : The Least Square Method

- Least Squares Criterion

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

where

y_i = observed value of the dependent variable for the i th observation

$\hat{y}_i$ = estimated value of the dependent variable for the i th observation

Summary : The Least Square Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

- *y* -Intercept for the Estimated Regression Equation

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

x_i = value of independent variable for *i* th observation

y_i = value of dependent variable for *i* th observation

x = mean value for independent variable

y = mean value for dependent variable

n = total number of observations

Sane's STATS Academy of Statistics

Standard Error of the Estimate

- Residual: Difference between the observed value and fitted value. Denoted by e_i
- It is a measure of goodness of fit.
- Lesser the Standard Error, better is the model fit
- Standard Error is calculated by the formula:

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Coefficient of Determination

- It is the fraction of variation of the dependent variable explained by the regression line.
- It is another measure of goodness of fit
- Bigger the R^2, better is the model fit
- Its formula is

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \overline{Y})^2} \qquad 0 \le R^2 \le 1$$

Dummy Variables

- Categorical Variables can be converted into indicators called dummy variables.
- e.g.
  - Gender having values 1 and 0
  - Quarter: four dummy variables Q1-Q4 with 1s and 0s

Multiple linear regression

Multiple Linear Regression

- Instead of fitting a line we fit a plane.
- General Form is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Assumptions of Linear Regression

- Normality: Errors are Normally Distributed with mean zero
- Independence: Errors are independent
- Linearity: Mean of dependent variable Y is linearly related to Xis
- Homoscedasticity: Errors have constant variance