Regression Model

How it is done?

1. Relationship between one dependent variable and explanatory variable(s)
2. Use equation to set up relationship
   a. Numerical Dependent (Response) Variable
   b. 1 or More Numerical or Categorical Independent (Explanatory) Variables
3. Used Mainly for Prediction & Estimation

Modeling Steps

1. **Hypothesize Deterministic Component**
   - Define the dependent variable and independent variable
   - To estimate Unknown Parameters
   - Expected Effects (i.e., Coefficients' Signs)
   - Functional Form (Linear or Non-Linear)
   - Interactions
2. **Specify Probability Distribution of Random Error Term.**
   - To estimate Standard Deviation of Error
3. **Evaluate the fitted Model.**
4. **Use Model for Prediction & Estimation.**

Regression Model

- **Linear Regression**
  - Relationship Between Variables Is a Linear Function

Population Y-Intercept **Population Slope** **Random Error**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable **Independent (Explanatory) Variable**

(Input variable) X_i
Explanatory: Experience

Regression Model

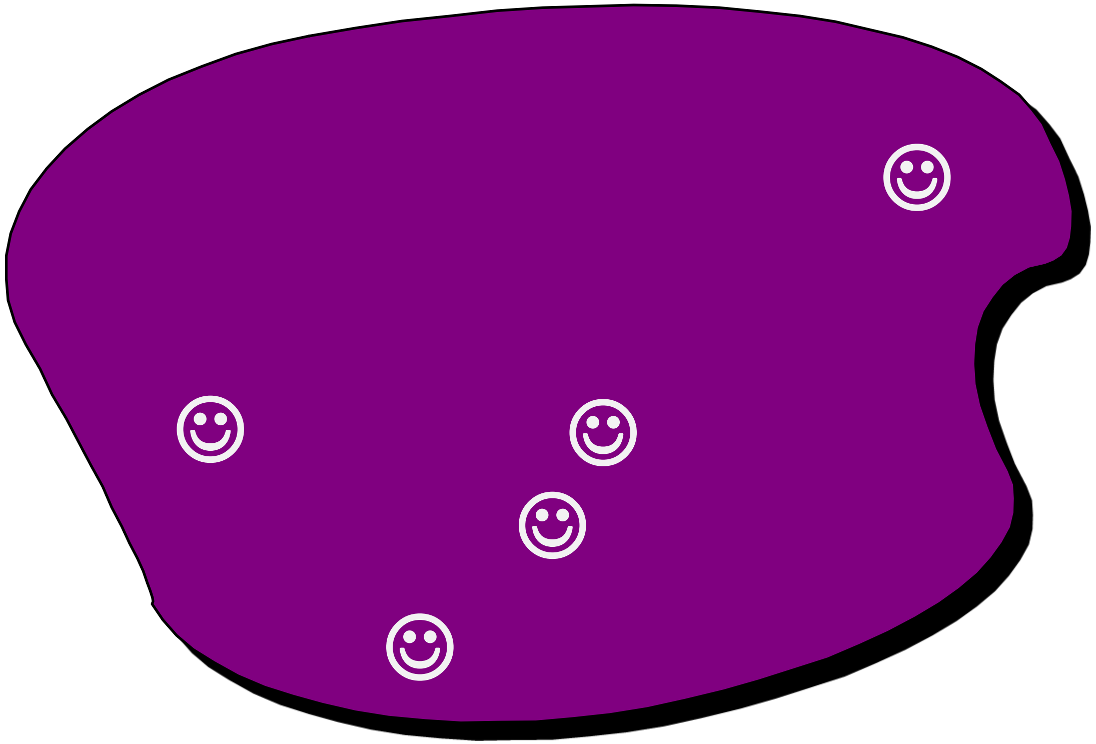
(Output variable) Y_i
Response: Salary

Experience (in years)	Salary (Rs)
1.1	23789.00
1.3	34678.00
1.4	44789.00
1.5	54890.00
2.2	60897.00
2.5	61345.00
2.3	64789.00
2.5	68908.00

Regression Model

An example of a Population and Sample

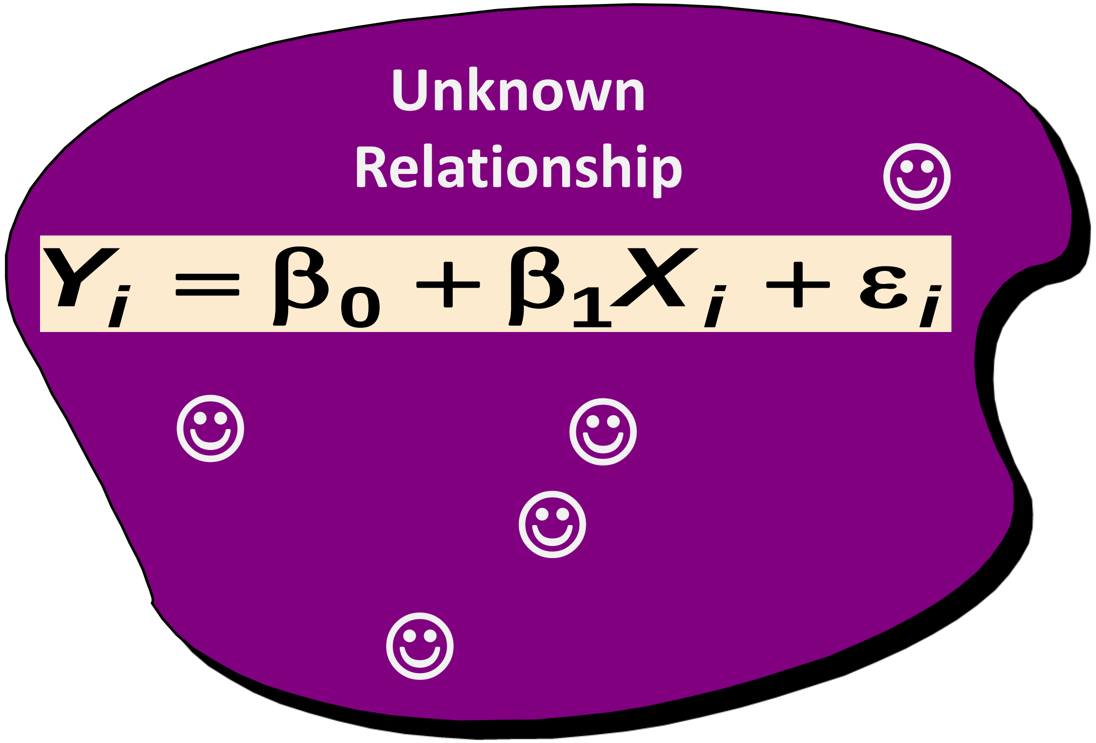
Population



Regression Model

An example of a Population and Sample

Population

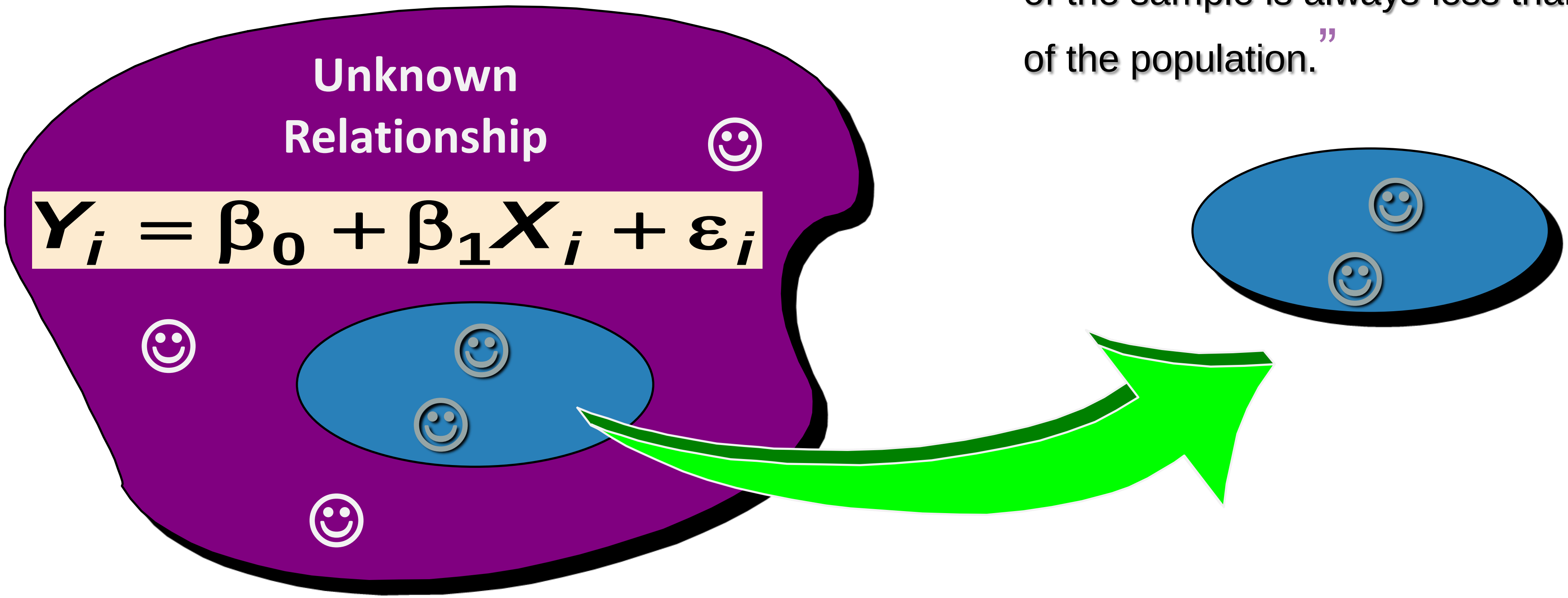


Regression Model

An example of a Population and Sample

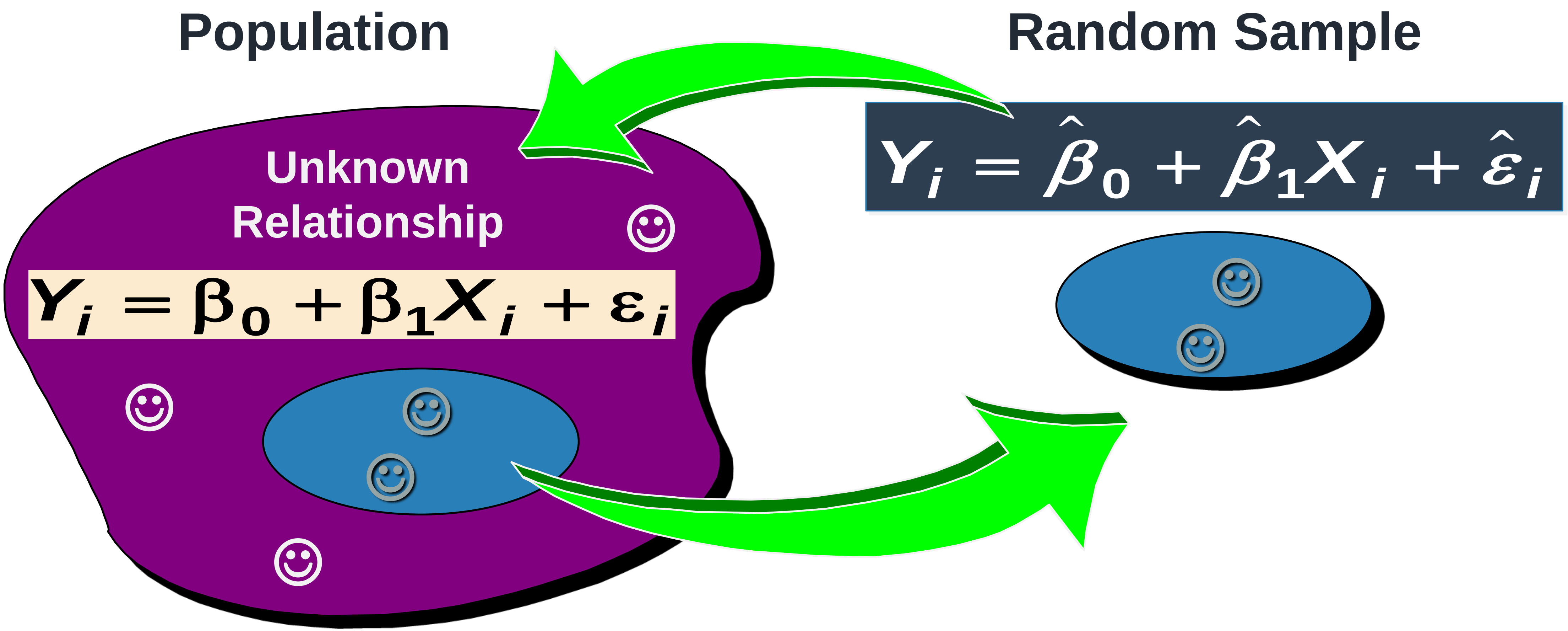
Population: "The entire group that you want to draw conclusions about."

Random Sample: "The specific group that you will collect data from. The size of the sample is always less than the total size of the population."

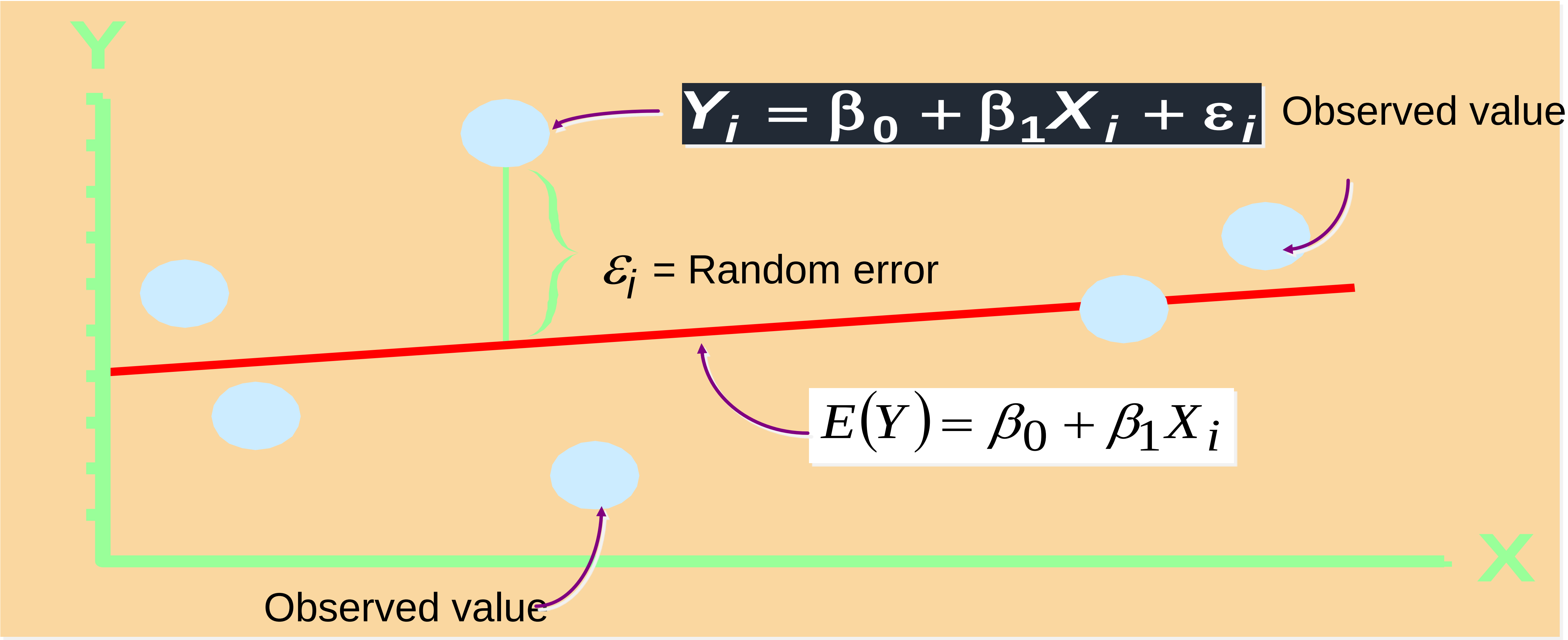


Regression Model

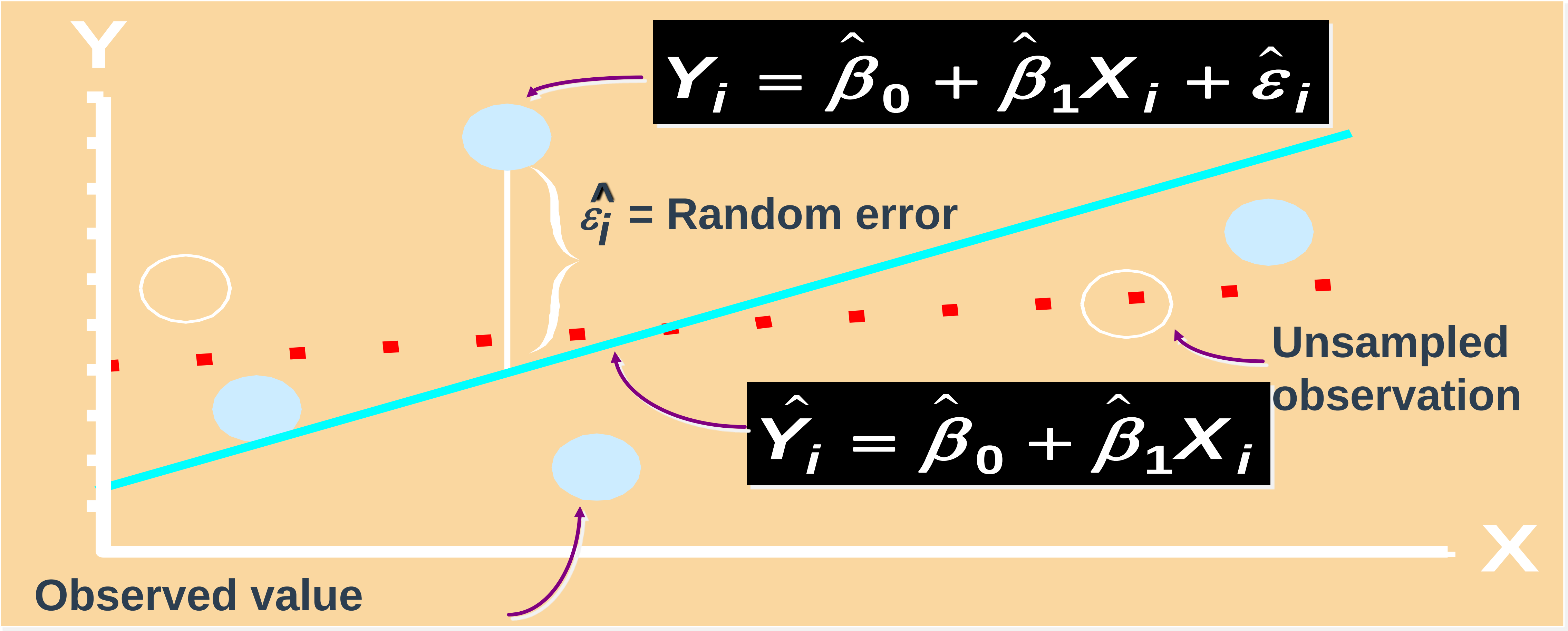
An example of a Population and Sample



Linear Regression Model - Population



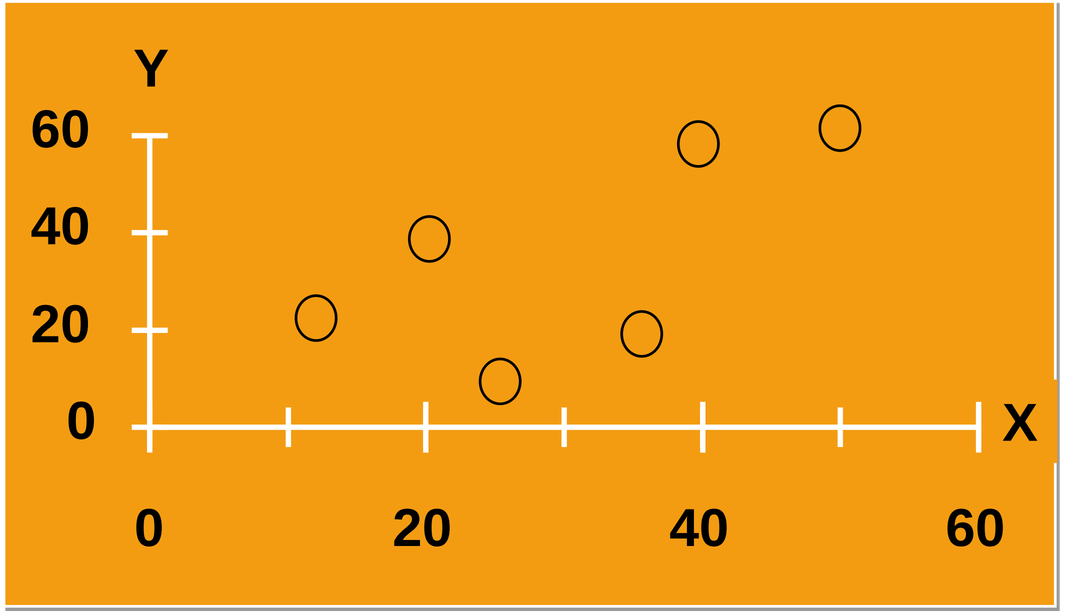
Linear Regression Model - Sample

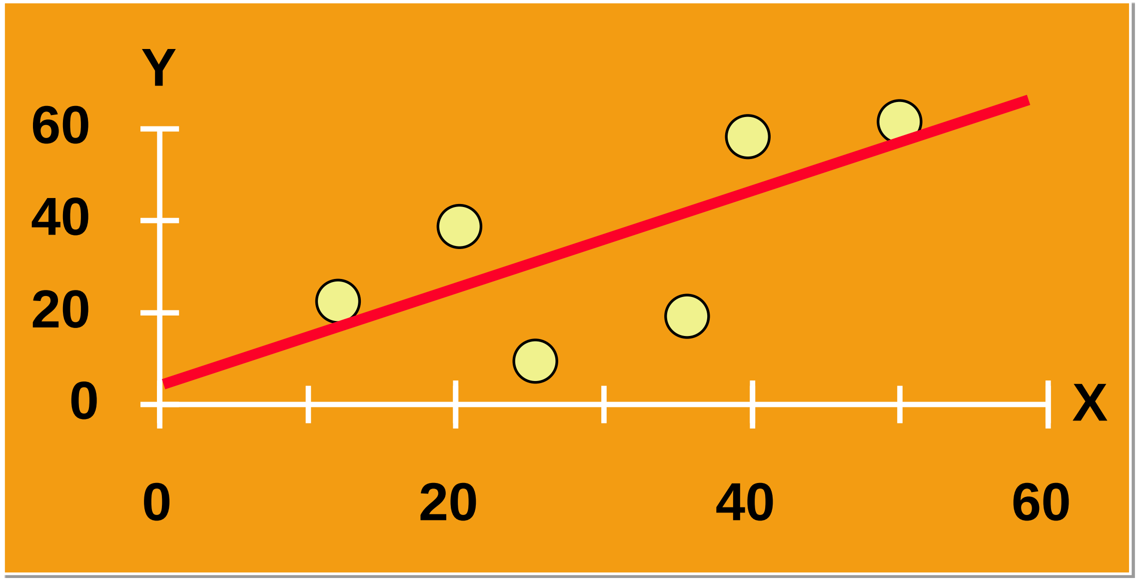


Estimating Parameters : Least Square Method

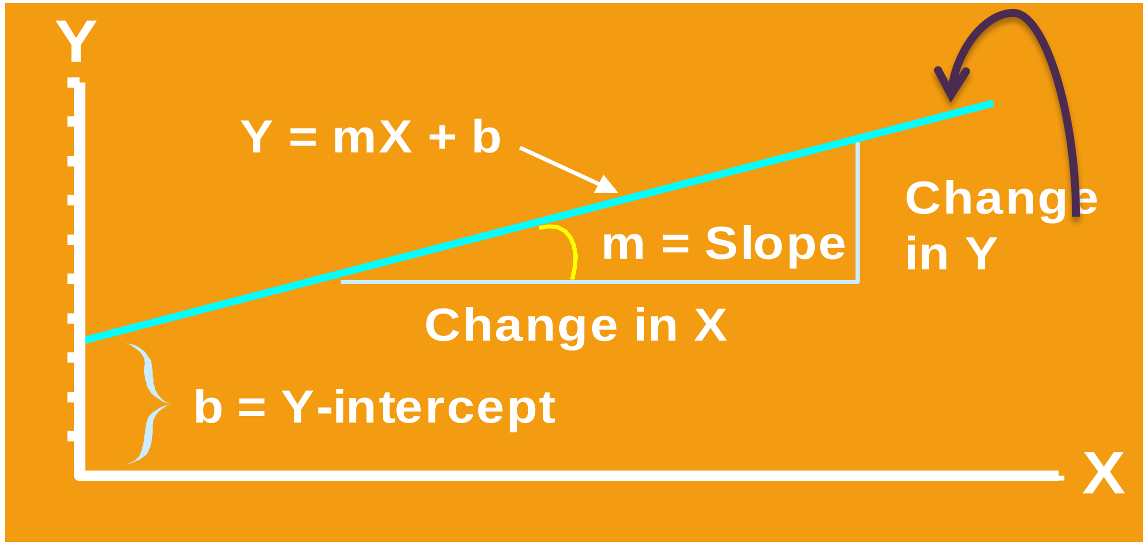
How would you draw a line through the points? How do you determine which line 'fits best'?

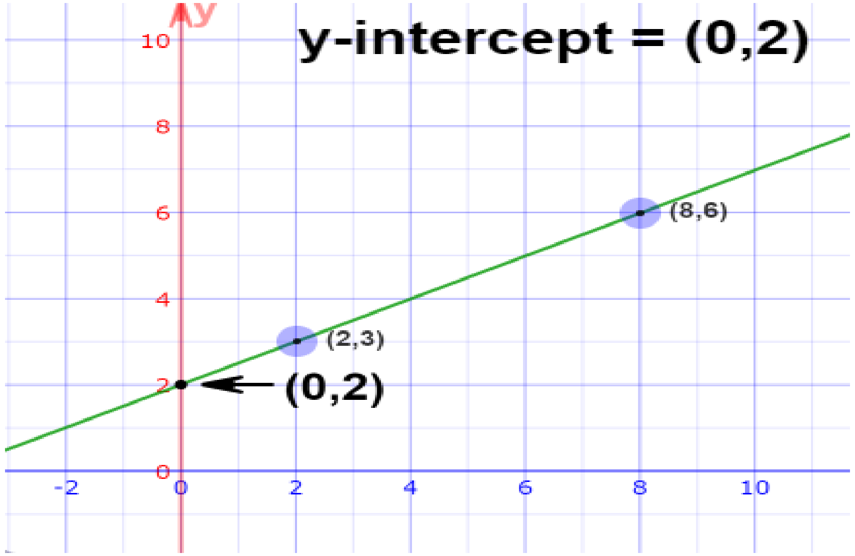
1. Plot of All (X_i, Y_i) Pairs
2. Suggests How Well Model Will Fit



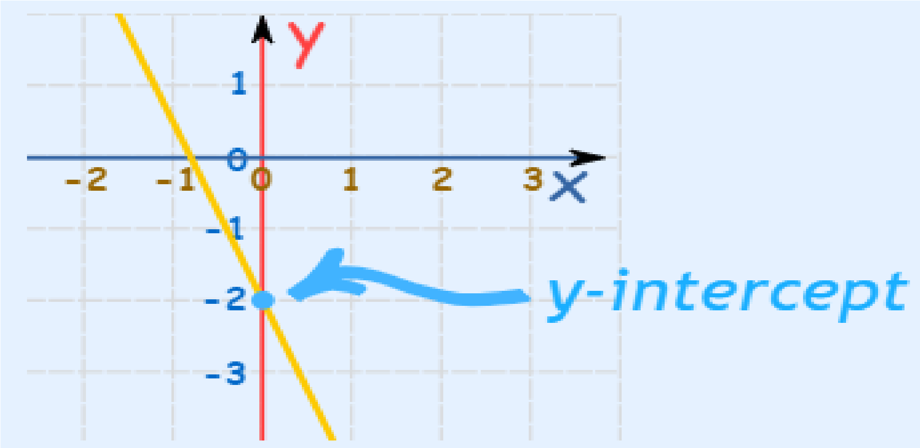


Linear Equation – a review

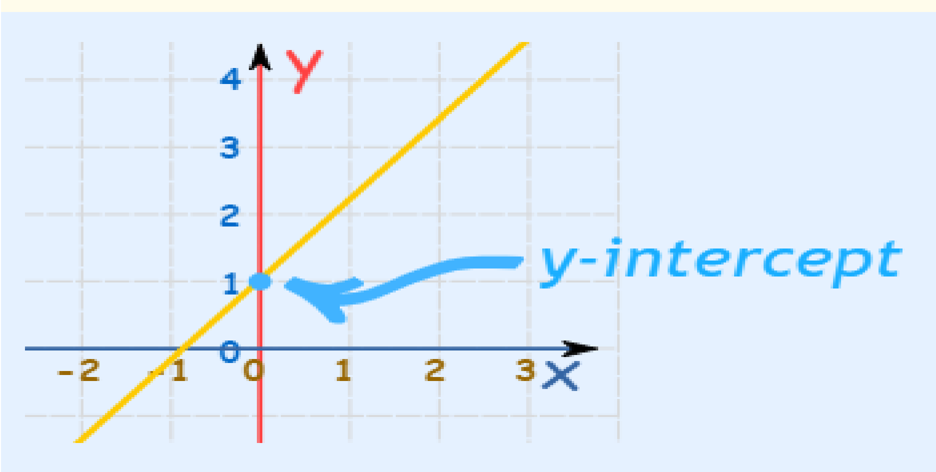




The y-intercept is an (x, y) point with x=0,



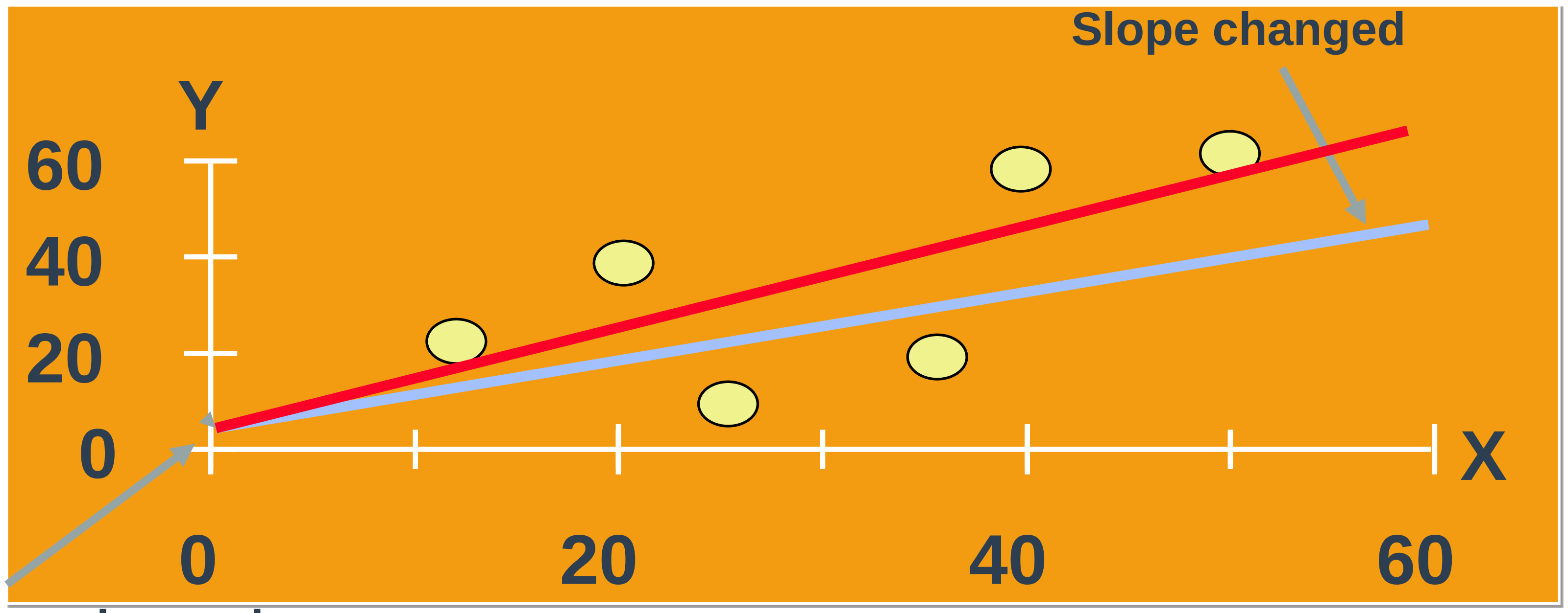
Where the line crosses the Y-axis, just find the value of y when x equals 0.The line crosses the Y-axis at y=1, and y=-2



Thinking Challenge

How would you draw a line through the points?

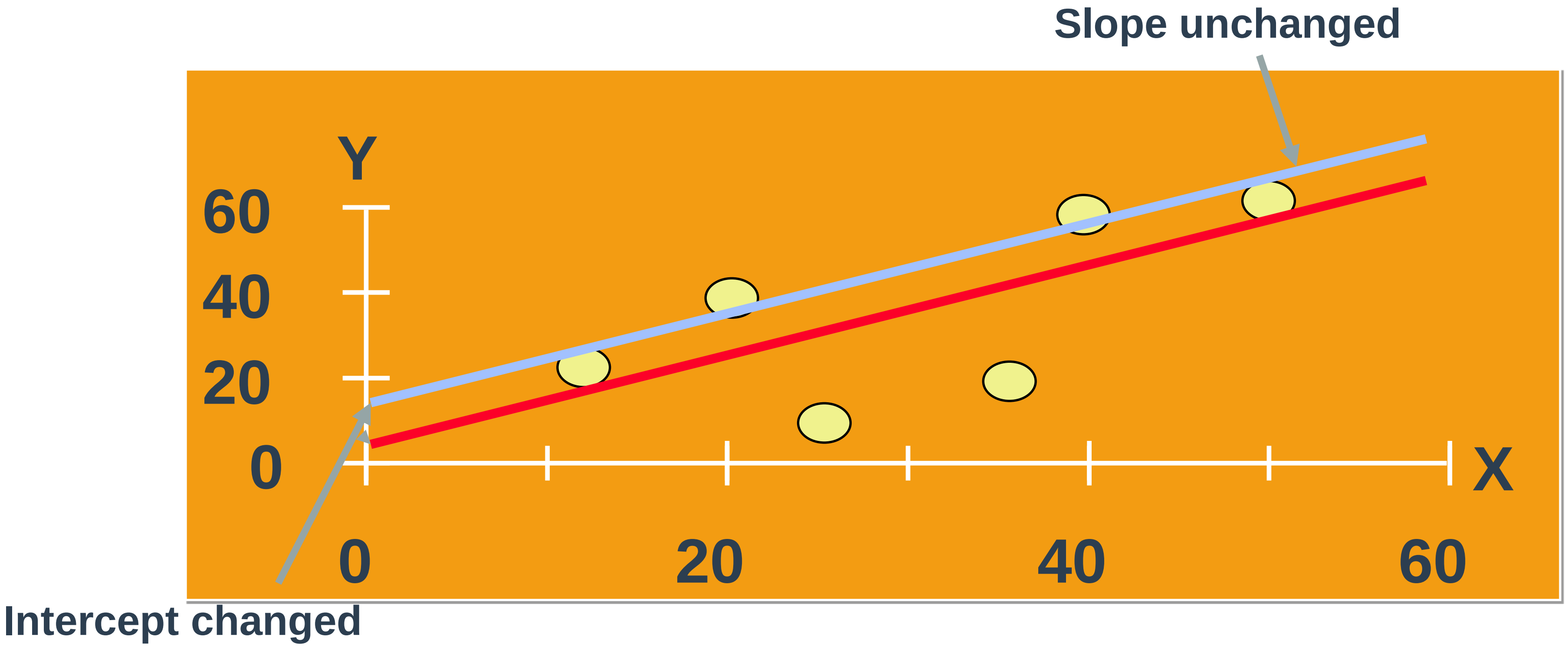
How do you determine which line ‘fits best’?



Thinking Challenge

How would you draw a line through the points?

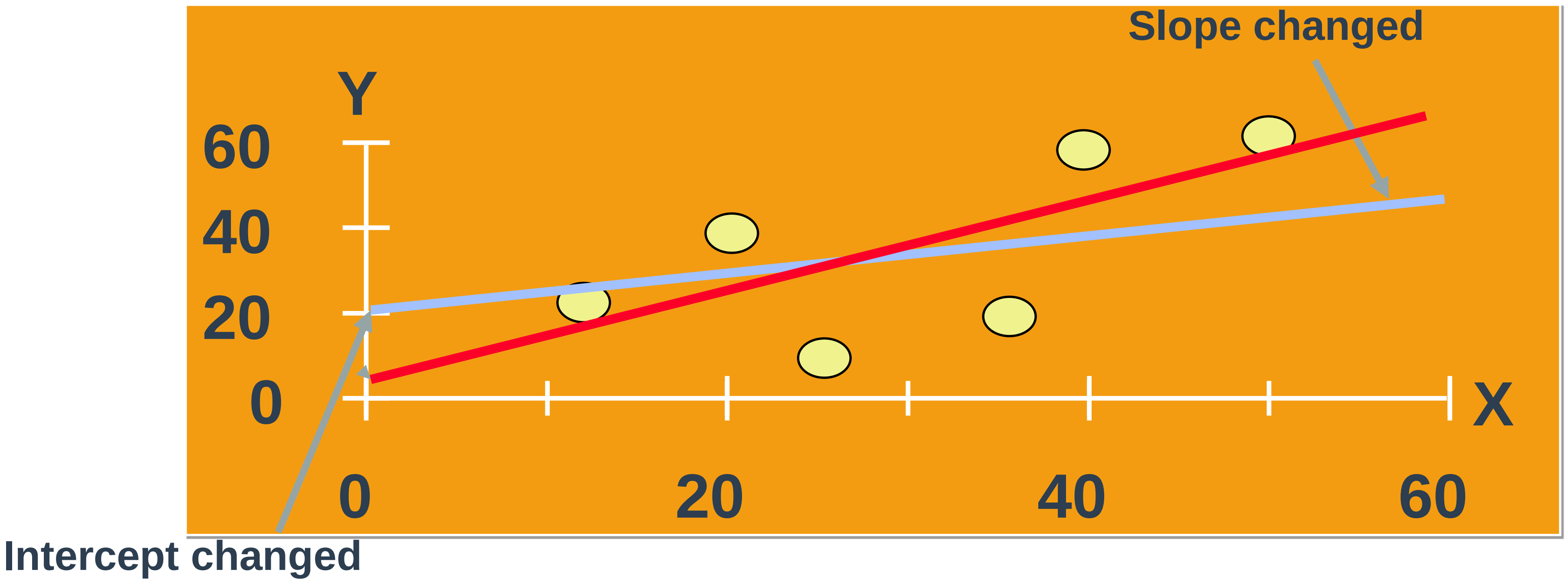
How do you determine which line ‘fits best’?



Thinking Challenge

How would you draw a line through the points?

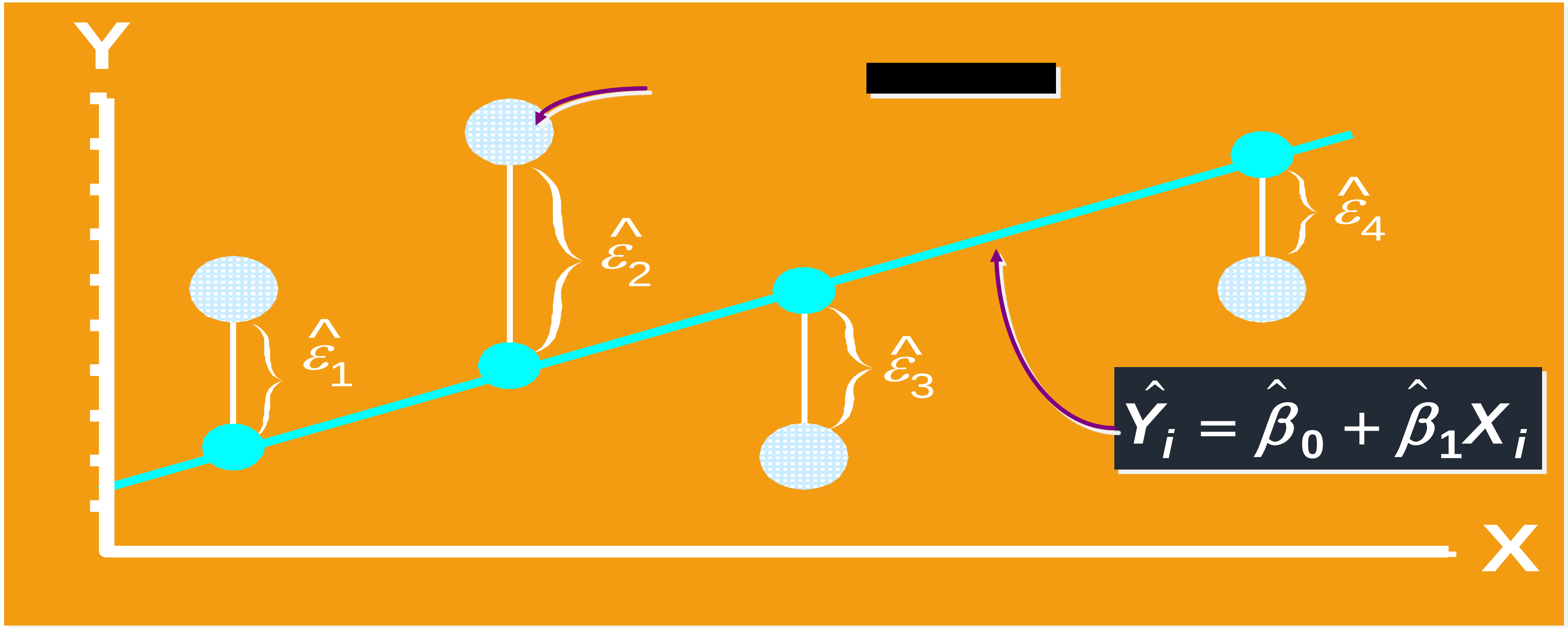
How do you determine which line ‘fits best’?



Least Squares

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$



1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!
2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Derivation of Parameters

- **Least Squares (L-S): Minimize squared error**

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0}$$

$$= -2\left(n\bar{y} - n\beta_0 - n\beta_1 \bar{x} \right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters

- **Least Squares (L-S):** Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_1}$$

$$= -2 \sum x_i \left(y_i - \beta_0 - \beta_1 x_i \right)$$

$$= -2 \sum x_i \left(y_i - \overline{y} + \beta_1 \overline{x} - \beta_1 x_i \right)$$

$$\beta_1 \sum x_i \left(x_i - \overline{x} \right) = \sum x_i \left(y_i - \overline{y} \right)$$

$$\beta_1 \sum \left(x_i - \overline{x} \right)\left(x_i - \overline{x} \right) = \sum \left(x_i - \overline{x} \right)\left(y_i - \overline{y} \right)$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Computation Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Interpretation of Coefficient

- Slope ($\hat{\beta}_1$)
  - Estimated Y Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in X
    - If $\hat{\beta}_1 = 2$, then Y Is Expected to Increase by 2 for Each 1 Unit Increase in X.

- Y-Intercept ($\hat{\beta}_0$) - Average Value of Y When $X = 0$

If $\hat{\beta}_0 = 4$, then Average Y Is Expected to Be 4 When X Is 0

Example : Parameter Estimation

Food (lb.)	Milk yield (lb.)
4	3.0
6	5.5
10	6.5
12	9.0

What is the **relationship** between cows' food intake and milk yield?

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
4	3.0	16	9.00	12
6	5.5	36	30.25	33
10	6.5	100	42.25	65
12	9.0	144	81.00	108
32	24.0	296	162.50	218

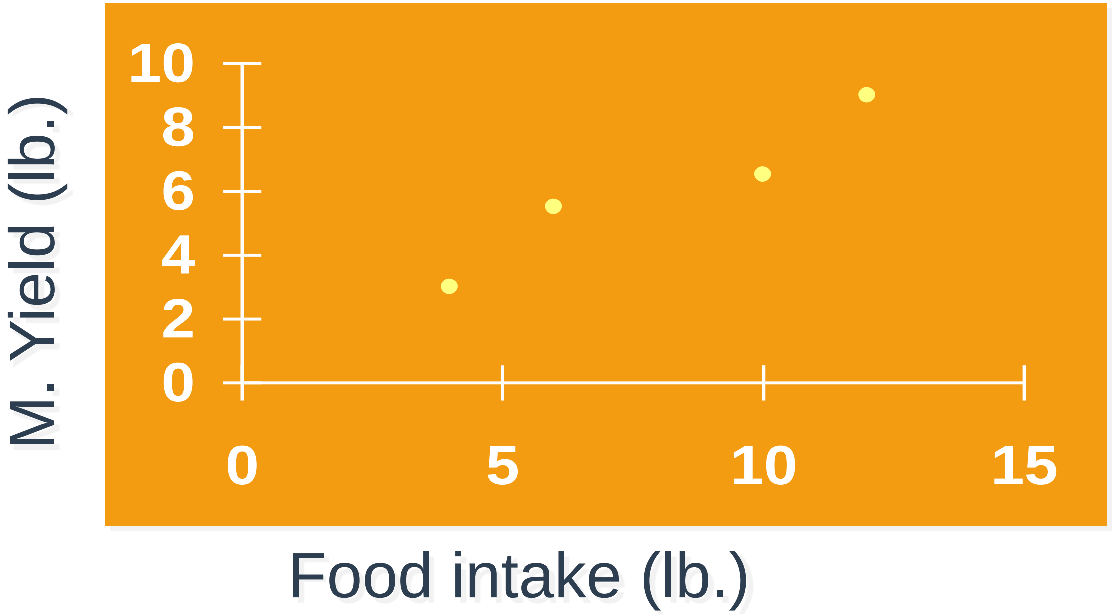


Figure. Scatter plot Food(lb) Versus Milk yield(lb.)

Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = 0.65$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 6 - (0.65)(8) = 0.80$$

Interpretation of Coefficient

1. Slope ($\hat{\beta}_1$)

   Milk Yield (Y) Is Expected to Increase by .65 lb. for Each 1 lb. Increase in Food intake (X)

2. Y-Intercept (β_0)

   Average Milk yield ($\hat{Y}$) Is Expected to Be 0.8 lb. When Food intake (X) Is 0

1. Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

$$\overline{X} = \frac{\sum X}{N} \qquad \overline{Y} = \frac{\sum Y}{N}$$

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X$ =28	$\sum Y$ =77	$\sum X^2$ =140	$\sum Y^2$ =875	$\sum XY = 334$

Regression coefficient of X on Y

$$b_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2}$$

Regression Equation of X on Y

$$(X - \overline{X}) = b_{xy}(Y - \overline{Y})$$

Regression coefficient of Y on X

$$b_{yx} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

Regression Equation of Y on X

$$(Y - \overline{Y}) = b_{yx}(X - \overline{X})$$

1. Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

$$\bar{X} = \frac{\sum X}{N} = \frac{28}{7} = 4 \quad \bar{Y} = \frac{\sum Y}{N} = \frac{77}{7} = 11$$

X	Y	X2	Y2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X$ =28	$\sum Y = 77$	$\sum X^2 = 140$	$\sum Y^2$ =875	$\sum XY$ =334

Regression coefficient of X on Y

$$b_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \frac{7(334) - (28)(77)}{7(875) - 77^2} = 0.929$$

Regression Equation of X on Y

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y}) = (X - 4) = 0.929(Y - 11)$$

X=0.929Y – 6.219

Regression coefficient of Y on X

$$b_{yx} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{7(334) - (28)(77)}{7(140) - 28^2} = 0.929$$

Regression Equation of Y on X

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X}) = (Y - 11) = 0.929(X - 4) = 0.929X + 7.284$$

2. Calculate the regression coefficient and obtain the lines of regression from the data given below, taking deviations from a actual means of X and Y. Estimate likely demand when price is Rs. 25.

Price (Rs.)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

$$\bar{X} = \frac{\sum X}{N} \qquad \bar{Y} = \frac{\sum Y}{N}$$

X	x=(X-13)	Y	y=(Y-41)	x^2	Y^2	xy
10	-3	40	-1	9	1	3
12	-1	38	-3	1	9	3
13	0	43	2	0	4	0
12	-1	45	4	1	16	-4
16	3	37	-4	9	16	-12
15	2	43	2	4	4	4
$\sum X =$	$\sum x = 0$	$\sum Y =$	$\sum y = 0$	$\sum x^2 = 24$	$\sum y^2 = 50$	$\sum xy = -6$

Regression coefficient of X on Y

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2}$$

Regression Equation of X on Y

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y})$$

Regression coefficient of Y on X

$$b_{yx} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

Regression Equation of Y on X

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

2. Calculate the regression coefficient and obtain the lines of regression from the data given below, taking deviations from actual means of X and Y. Estimate likely demand when price is Rs. 25.

Price (Rs.)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

$\bar{X} = \frac{\sum X}{N} = \frac{78}{6}$ $\bar{Y} = \frac{\sum Y}{N} = \frac{246}{6}$

X	x=(X-13)	Y	y=(Y-41)	x^2	y^2	xy
10	-3	40	-1	9	1	3
12	-1	38	-3	1	9	9
13	0	43	2	0	4	0
12	-1	45	4	1	16	-4
16	3	37	-4	9	16	-12
15	2	43	2	4	4	4
$\sum X$ =78	$\sum x = 0$	$\sum Y = 246$	$\sum y$ =0	$\sum x^2 = 24$	$\sum y^2$ =50	$\sum xy = -6$

Regression coefficient of X on Y

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2}$$

Regression coefficient of Y on X

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = -\frac{6}{24} = -0.25$$

Regression Equation of X on Y $(X-\bar{X}) = b_{xy}(Y-\bar{Y}) = X = -0.12Y + 17.92$

When price is Rs. 25, the likely demand is

Regression Equation of Y on X $(Y-\bar{Y}) = b_{yx}(X-\bar{X}) = Y = -0.25X + 44.25 = -0.25(25) + 44.25 = 38$

3. The following table shows the sales and advertisement expenditure of a form

	Sales	Advertisement (Rs. Crores)
Mean	40	6
Standard Deviation	10	1.5

Coefficient of correlation r = 0.9. Estimate the likely sales for a proposed advertisement expenditure of Rs. 10 crores.

3. The following table shows the sales and advertisement expenditure of a form

	Sales	Advertisement (Rs. Crores)
Mean	40	6
Standard Deviation	10	1.5

Coefficient of correlation r = 0.9. Estimate the likely sales of units for a proposed advertisement expenditure of Rs. 10 crores.

Solution:

Given $\overline{X}=40, \overline{Y}=6, \sigma_x=10, \sigma_y=1.5$ and $r=0.9$

Equation of line of regression x on y is

$$X-\overline{X} = r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$X-40 = (0.9)\frac{10}{1.5}(Y-6)$$

$$X-40 = 6Y-36$$

$$X = 6Y+4$$

When advertisement expenditure is 10 crores i.e., Y=10 then sales X=6(10)+4=64 which implies sales is 64 units