Regression

F. M. Arifur Rahman

Senior Lecturer, Department of Mathematical & Physical Sciences

Content

- Simple linear regression
  - Regression Model
  - Analysis procedure
  - Goodness of fit
  - Prediction
  - Interpretations
- Different types of regression

Regression

- Regression analysis is a technique that studies the cause and effect relationship between two or more variables
- Assume or suspect a cause and effect relationship between variables-
  - causal variables as independent variables
  - affected variables as dependent variables
- Regression analysis explains and predicts the changes in the magnitudes of dependent variable(s) in terms of independent variable(s).

Regression

Example 1:

- We know that, there is a positive relationship between income and expenditure, i.e. an increase in income increases expenditures.
- As increase in income causes an increase in expenditures, we took in come as independent variable (X) and expenditures as dependent variable (Y).
- And found a fitted regression model-

$$\hat{Y} = a + bX = 15000 + .78X$$

Where, Y= expenditure and X=income

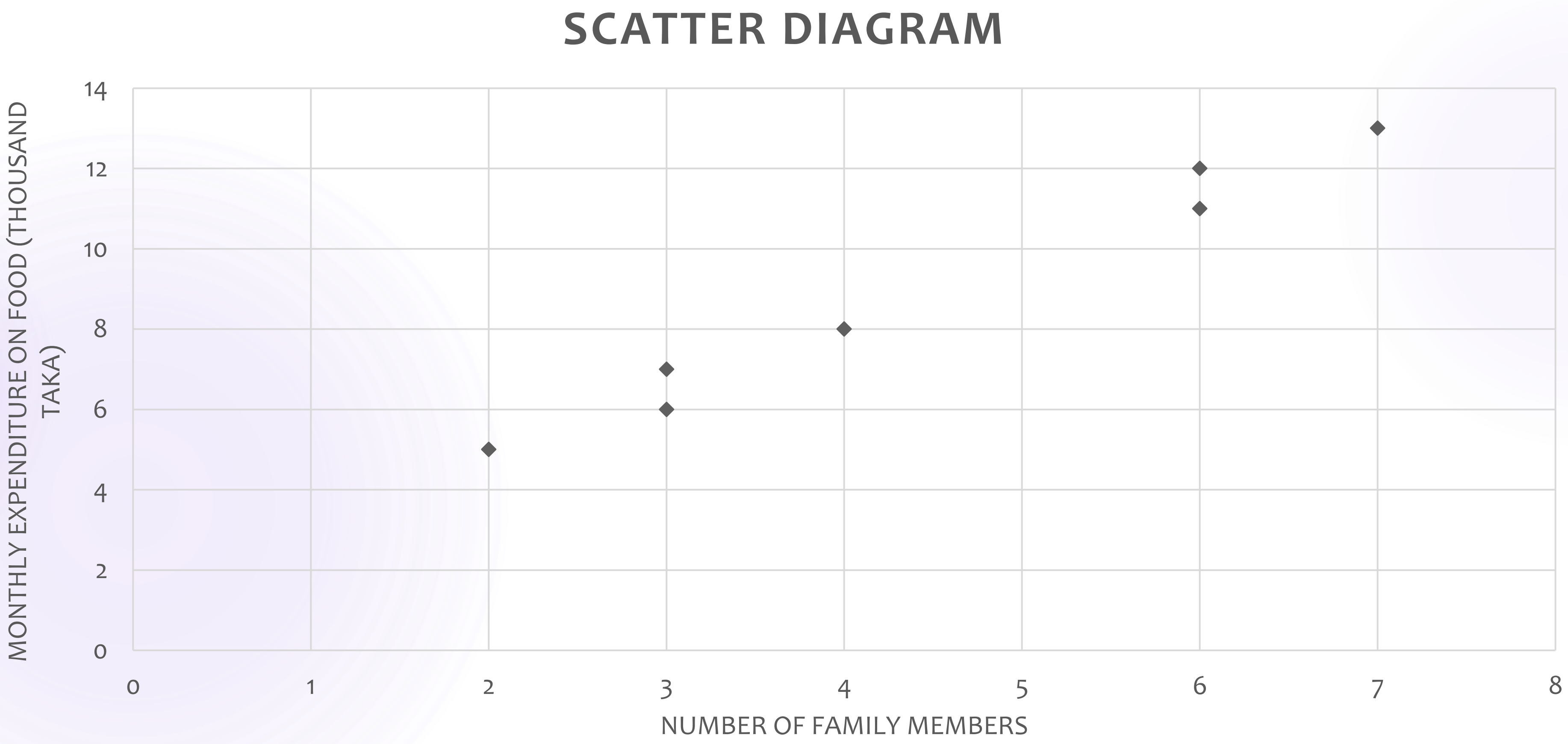
Regression

Example 2:

No. of family members, x	Monthly expenditure on food (thousand taka), y
2	5
3	7
6	11
4	8
7	13
3	6

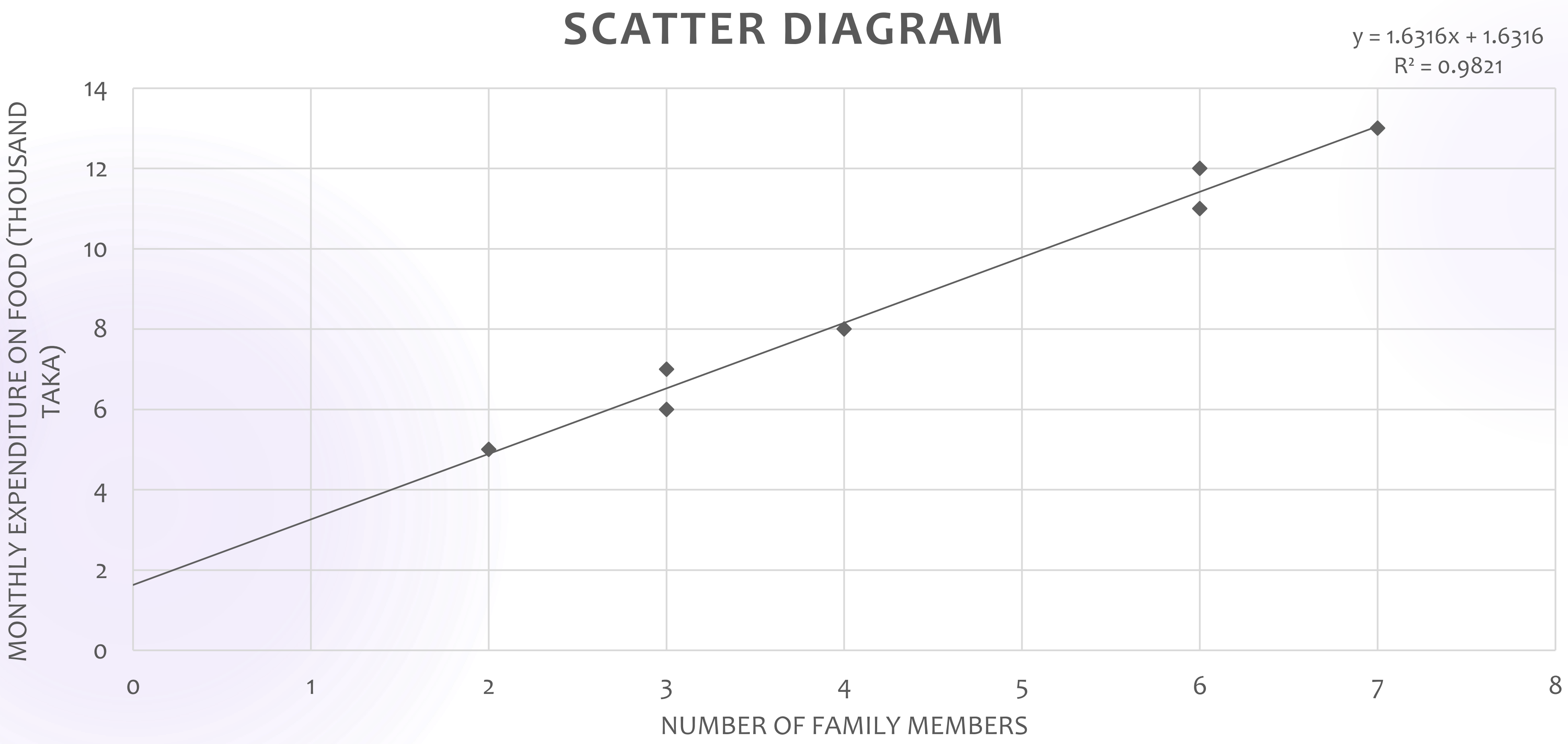
Fit a regression line of **y** on **x**. Interpret the estimates of the parameters. Find the value of R-square. Comment on your result. Estimate that how much monthly expenditure on food would occur if number of family members is 10.

Regression


SCATTER DIAGRAM
MONTHLY EXPENDITURE ON FOOD (THOUSAND TAKA)
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6
7
8
NUMBER OF FAMILY MEMBERS

Regression line

Estimated regression equation, $\hat{y} = a + bx$



Regression

Example 2:

No. of family members, X	Monthly expenditure on food (thousand taka), Y
2	5
3	7
6	11
4	8
7	13
3	6
6	12

Example 2

No. of family members, x	Monthly expenditure on food (thousand taka), y	x^2	xy
2	5		
3	7		
6	11		
4	8		
7	13		
3	6		
6	12		

Example 2

No. of family members, x	Monthly expenditure on food (thousand taka), y	x^2	xy
2	5	4	10
3	7	9	21
6	11	36	66
4	8	16	32
7	13	49	91
3	6	9	18
6	12	36	72
$\sum x = 31$	$\sum y = 62$	$\sum x^2 = 159$	$\sum xy = 310$

Example 2

Estimates of the parameters:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{7 * 310 - 31 * 62}{7 * 159 - 31^2} = 1.63$$

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{62}{7} - 1.63 * \frac{31}{7} = 1.64$$

Estimated regression line:

$$\hat{y} = 1.64 + 1.63\,x$$

Example 2

Estimated regression line:

$$\hat{y} = 1.64 + 1.63\,x$$

Interpretation:

a = 1.64 means, monthly expenditure on food (Y) is 1.64 (thousand taka) when no. of family members, i.e. X=0

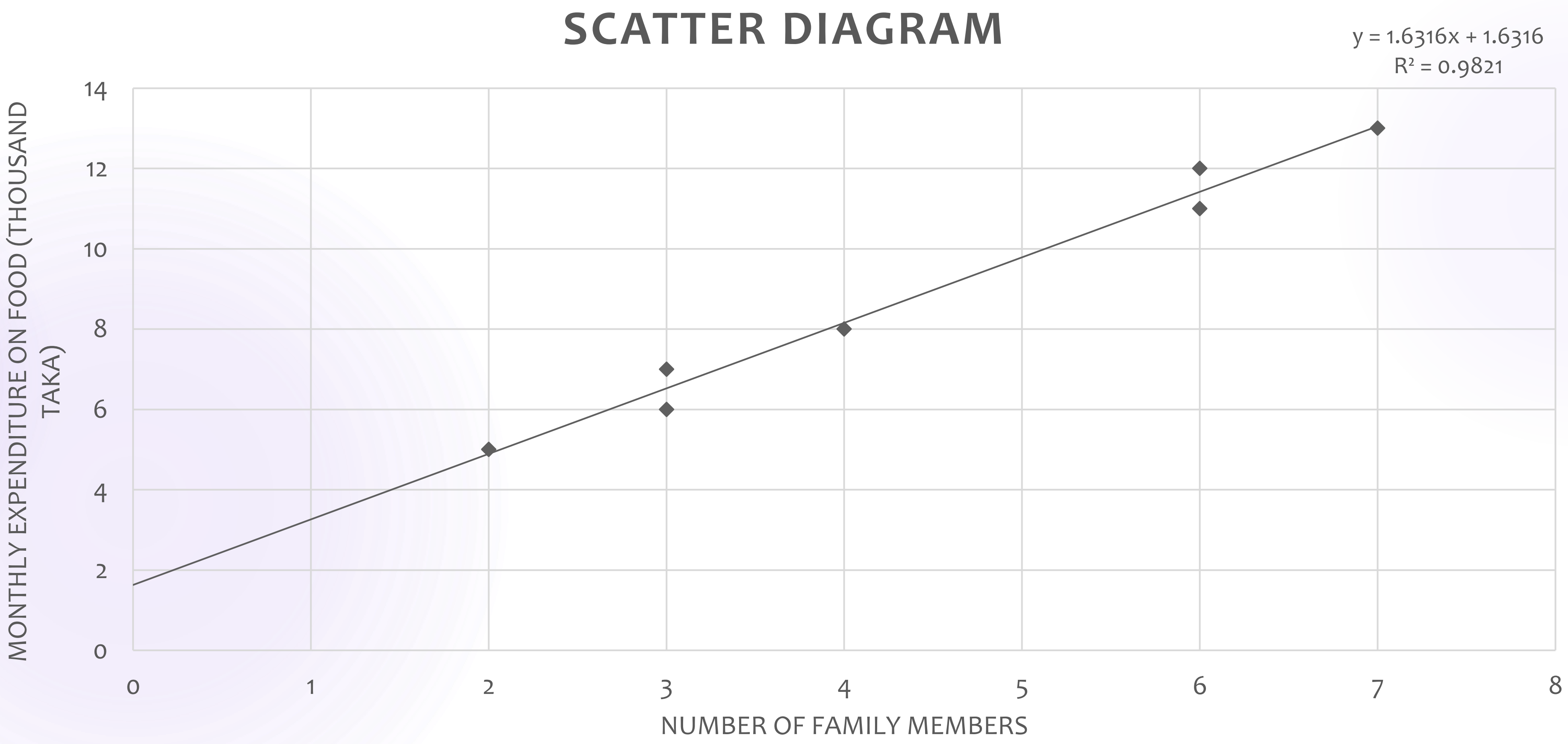
b= 1.63 means, if number of family members is increased by 1 member (i.e. if 1 member is added), on average, monthly expenditure on food will increase by 1.63 (thousand taka)

Example 2

x	y	x^2	xy	$\hat{y}$
2	5	4	10	=1.64+1.63*2 = 4.9
3	7	9	21	=1.64+1.63*3 = 6.53
6	11	36	66	=1.64+1.63*6 = 11.42
4	8	16	32	8.16
7	13	49	91	13.05
3	6	9	18	6.53
6	12	36	72	11.42
$\sum x = 31$	$\sum y = 62$	$\sum x^2 = 159$	$\sum xy = 310$	

Example 2

Estimated regression equation, $\hat{y} = 1.64 + 1.63\ x$



Goodness of fit

R-square:

$$n\,Var(Y) = \sum(\mathrm{Y_i} - \overline{\mathrm{Y}})^2 = \sum\left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum\left(Y_i - \widehat{Y}_i\right)^2$$

Or, Total variation = Explained variation + Unexplained variation

Or, Total Sum of Squres $(SST) = Regression\ Sum\ of\ Squares\ (SSR) + Error\ Sum\ of\ Squares\ (SSE)$

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum e_i^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}$$

Goodness of fit

R-square interpretation:

Range: $0 \leq R^2 \leq 1$

If $\boldsymbol{R^2} \rightarrow \mathbf{0}$: Poor fit i.e. the model is not strong or effective enough

If $\boldsymbol{R^2} \rightarrow \mathbf{1}$: Good fit i.e. the model is strong or effective enough

R^2% variation in dependent variable (Y) can be explained by the variation in independent variable (X).

Example

x	y	y^2	$\hat{y}$	$e_i = (y - \hat{y})$	e_i^2
2	5				
3	7				
6	11				
4	8				
7	13				
3	6				
6	12				

Example

x	y	y^2	$\hat{y}$	$e_i = (y - \hat{y})$	e_i^2
2	5	25	=1.64+1.63*2 = 4.9	=5-4.9=.10	0.01
3	7	49	=1.64+1.63*3 = 6.53	=7-6.53=0.47	.2209
6	11	121	=1.64+1.63*6 = 11.42	=11—11.42=-.42	0.1764
4	8	64	8.16	-0.16	0.0256
7	13	169	13.05	-0.05	0.0025
3	6	36	6.53	-0.53	0.2809
6	12	144	11.42	0.58	0.3364
$\sum x = 31$	$\sum y = 62$	$\sum y^2 = 608$		$\sum e_i = 0$	$\sum e_i^2$ =1.05 27

Example

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}$$

$$= 1 - \frac{1.0527}{608 - \frac{62^2}{7}}$$

$$= 0.9821$$

Notice that, $r^2 = R^2$.

Interpretation:

98.21% variation in monthly expenditure on food (Y) can be explained by the variation in no. of family members (X).

That means, the fitted model has a good fit to the data and capable of explaining almost all variation in the dependent variable Y.

Prediction (For example data)

For x= 10 (if number of family members is 10), then the estimated monthly expenditure on food -

$\hat{y} = 1.64 + 1.63 \, * x = 1.64 + 1.63 \, * 10 = 17.93 \; (thousand \; taka)$

Simple Linear Regression

Simple Linear Regression Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i \qquad ; \qquad i = 1, 2, \dots, n$$

Simple Linear Regression

Simple Linear Regression Model:

$$Y_i = \alpha + \beta X_i + \epsilon_i \qquad ; \qquad i = 1, 2, \ldots, n$$

Where,

Y= dependent variable

X= independent variable

α= Intercept

β= Slope

Regression coefficients (Parameters)

Є= Error term (unexplained factor)

Simple Linear Regression

Simple Linear Regression Model: (for sample)

$$y_i = a + bx_i + e_i \qquad ; \qquad i = 1, 2, \dots, n$$

Estimated regression line-

$$E(Y_i|X_i) = \hat{y}_i = a + bx_i \qquad ; \qquad i = 1, 2, \dots, n$$

$$\therefore e_i = y_i - (a + bx_i) = y_i - \hat{y}_i$$

Estimation of Regression parameters

Least Square Method:

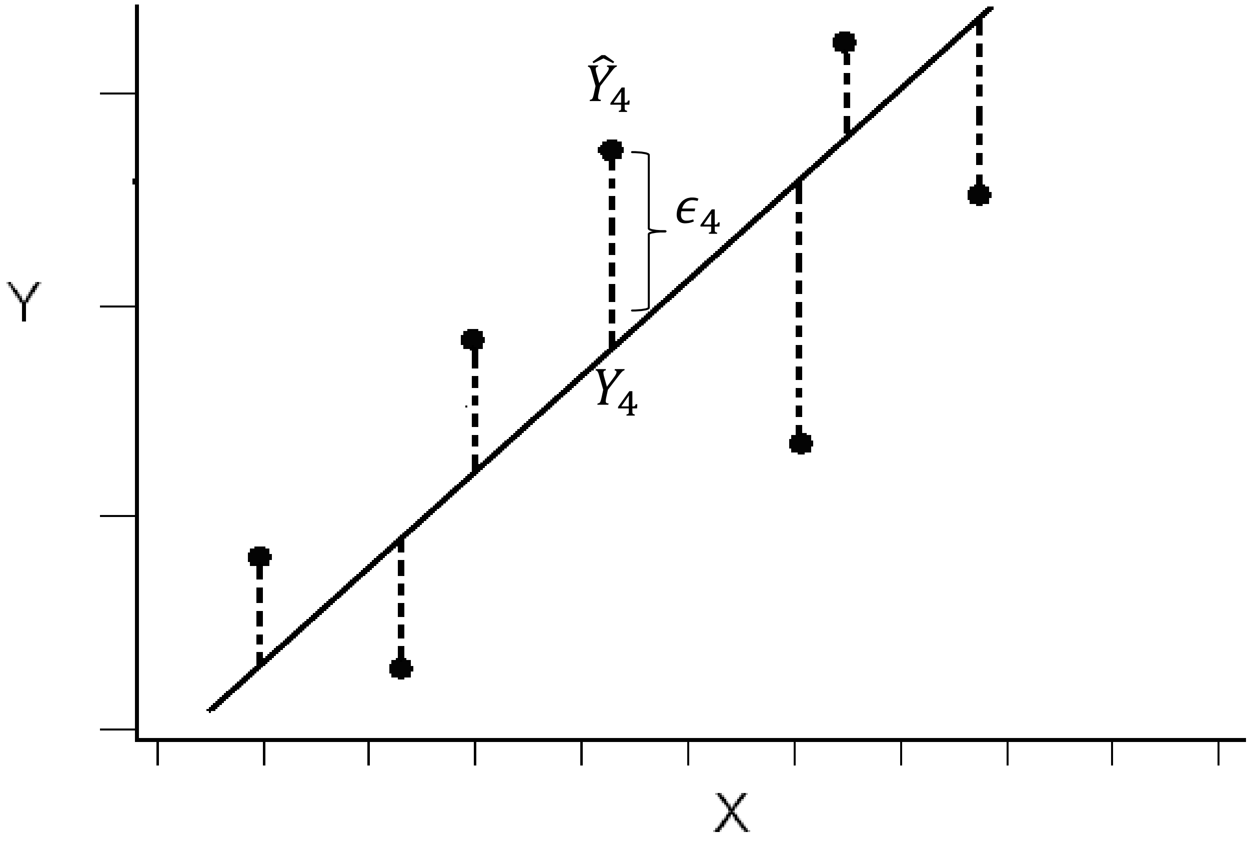
Principle: Determining regression equation i.e. estimating regression parameters such that the sum of squares of the vertical distances between the actual Y values and the predicted Y values i.e. sum of squares of errors ($\sum_i \epsilon_i^2$) is minimized.

Estimation of Regression parameters

Least Square Method:

Minimize,

$$\sum \epsilon_i^2 = \sum (Y_i - \hat{Y}_i)^2$$



Estimation of Regression parameters

Least Square Estimates (LSE) of the parameters:

Let ***a*** and ***b*** are the least square estimates of **α** and **β** respectively, then-

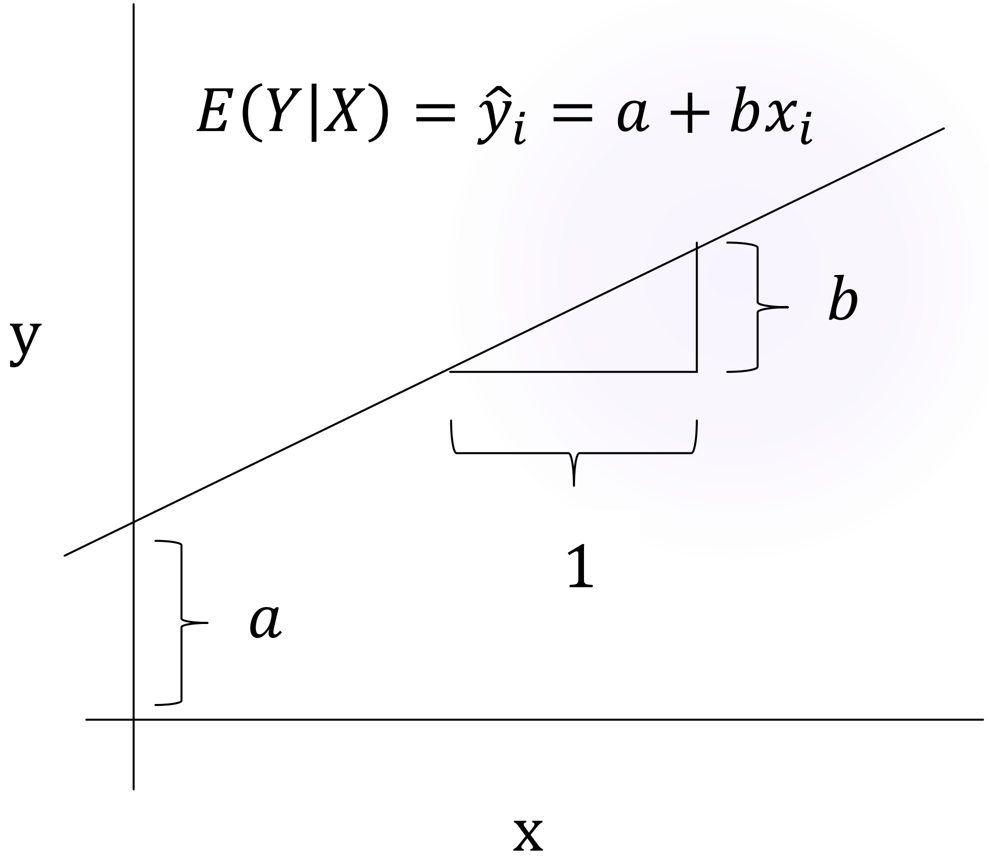
$$\hat{\boldsymbol{\beta}} = b = \frac{cov(X,Y)}{v(X)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

And $$\hat{\boldsymbol{\alpha}} = a = \bar{y} - b\bar{x} = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

Interpretation of estimated parameters

a (intercept): when x=0, the baseline value of y is ***a*** units

b (Slope): For 1 unit increase in x, the average or expected increase (if, b>0) or decrease (if, b<0) in y is ***b*** units.

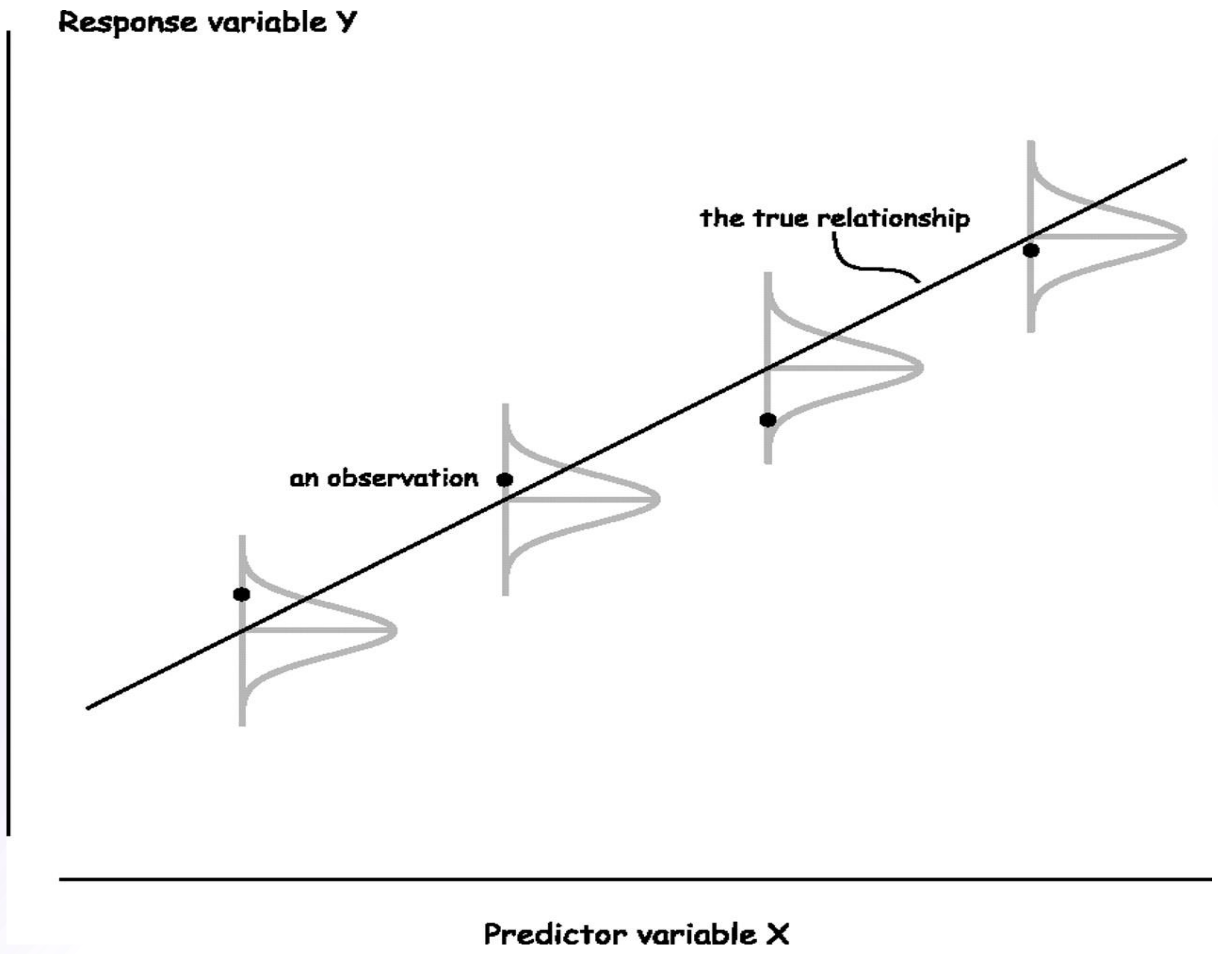


Assumptions of Regression

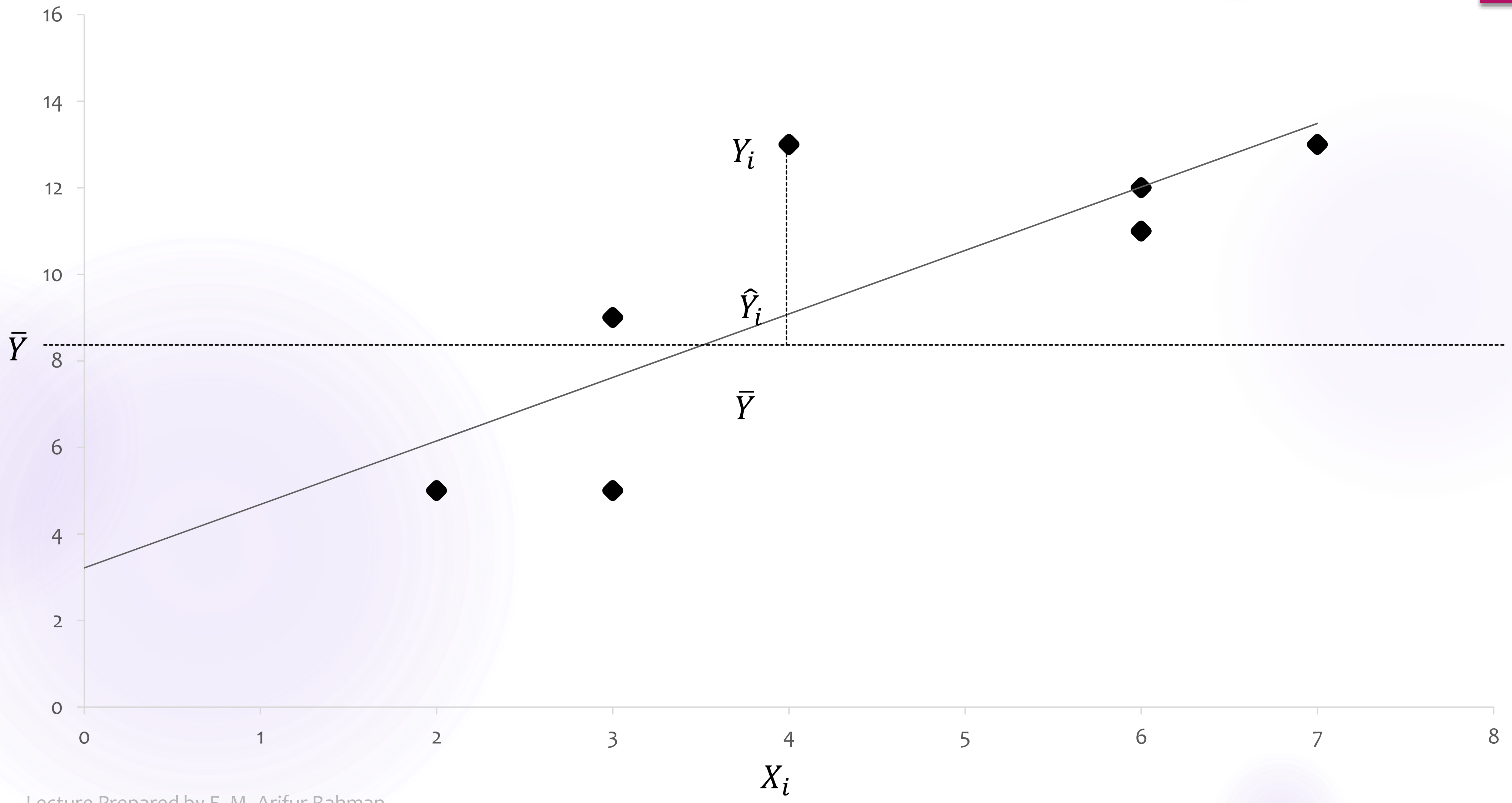
Assumptions of Simple Linear Regression Model:

1. X values are fixed
2. The relationship between X and Y is linear
3. $\epsilon_i \sim N(0,\sigma^2)$, i.e. error terms follows normal distribution with mean 0 and variance σ^2.
4. X and ϵ are uncorrelated, i.e. $Corr(X,\epsilon) = r_{X\epsilon} = 0$

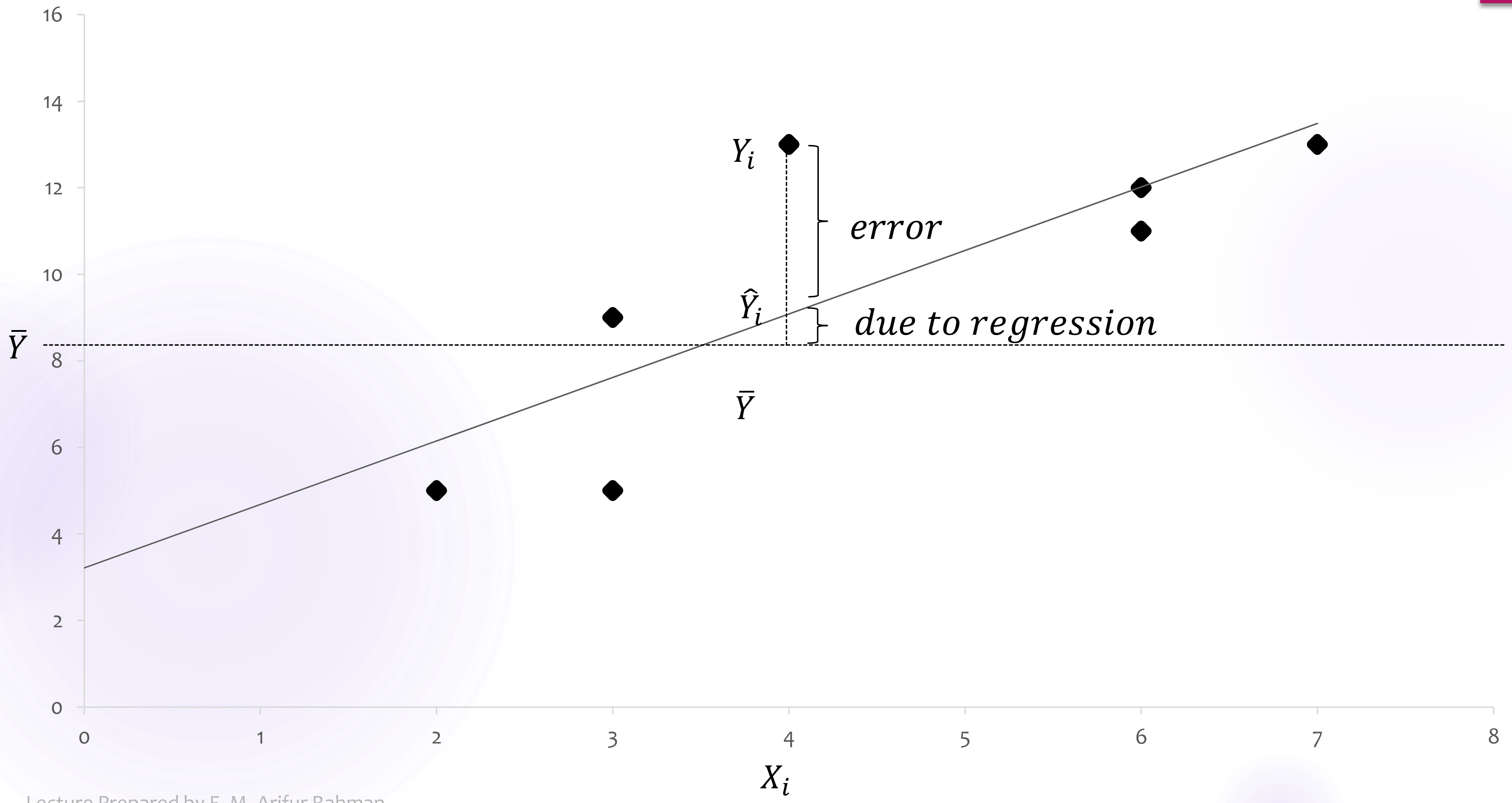
Assumptions of Regression



Goodness of fit



Goodness of fit



Steps of regression

Hypothesize a Model of Relationship

Estimation of Regression Equation

Goodness of fit test of the Model

Prediction

Uses of regression

Uses:

1. Estimate the relationship that exists, on average, between the dependent variable and the independent (explanatory) variable.
2. Determine the effect of each of the explanatory variables on the dependent variable, controlling the effects of all other explanatory variables, if any.
3. Predict the value of the dependent variable for a given or known value of the explanatory variable

Other regression types: