Regression

Md. Mahfuzur Rahman

Senior Lecturer, Statistics

Content

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

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

Simple Linear Regression

Simple Linear Regression Model:

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

Where,

Y= dependent variable

X= independent variable

α= Intercept

β= Slope

Regression coefficients (Parameters)

Є= Error term (unexplained factor)

Simple Linear Regression

Estimating Parameters regression line:

One important objectives of regression analysis is to find estimates for α and β for a given model. There are several methods of estimating the parameters of a regression model. Such as:

- Graphical method
- Least square method

We will discuss here about the most commonly used method that is least square method for estimating the parameters of a regression model.

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

Thus the fitted regression line : $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X$

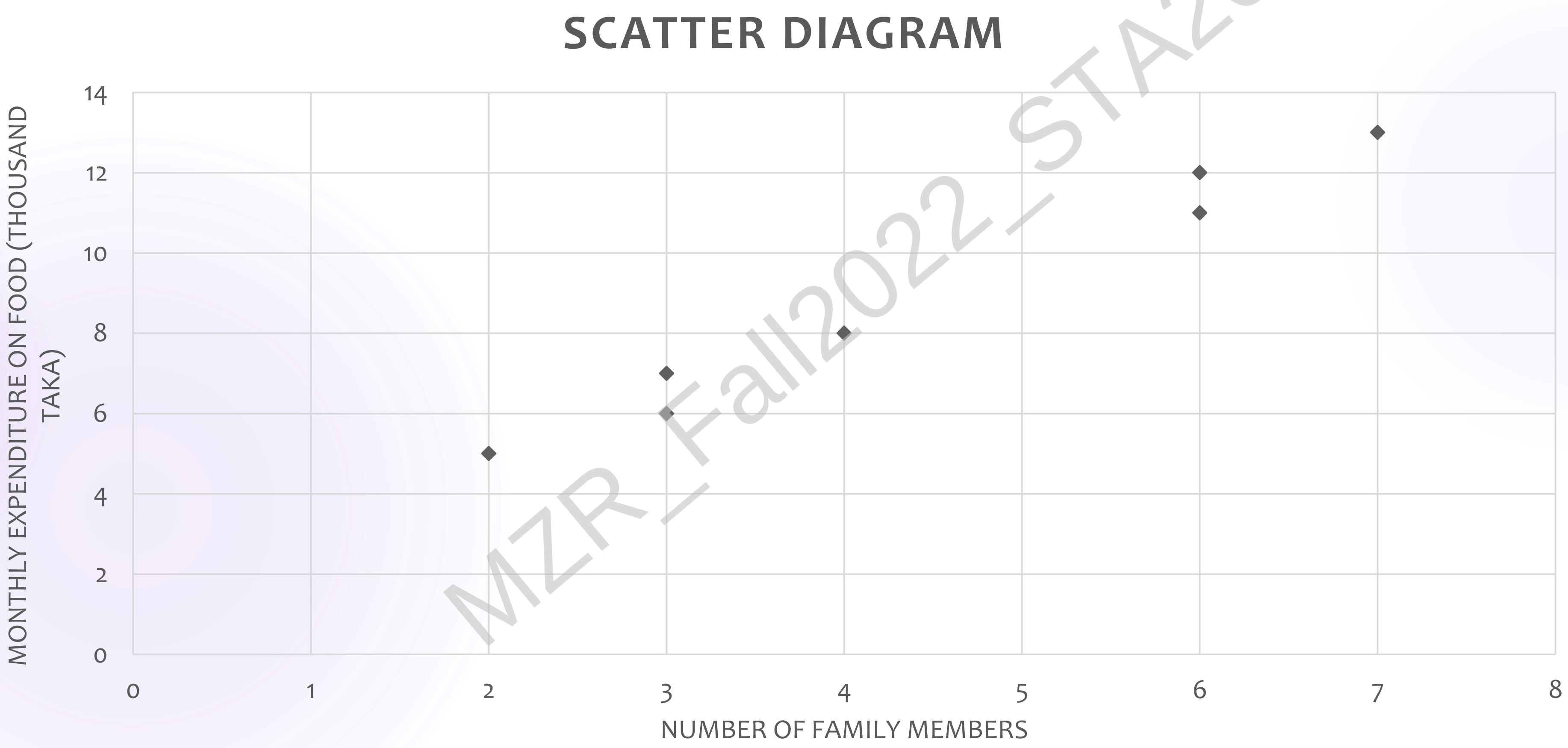
Regression

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

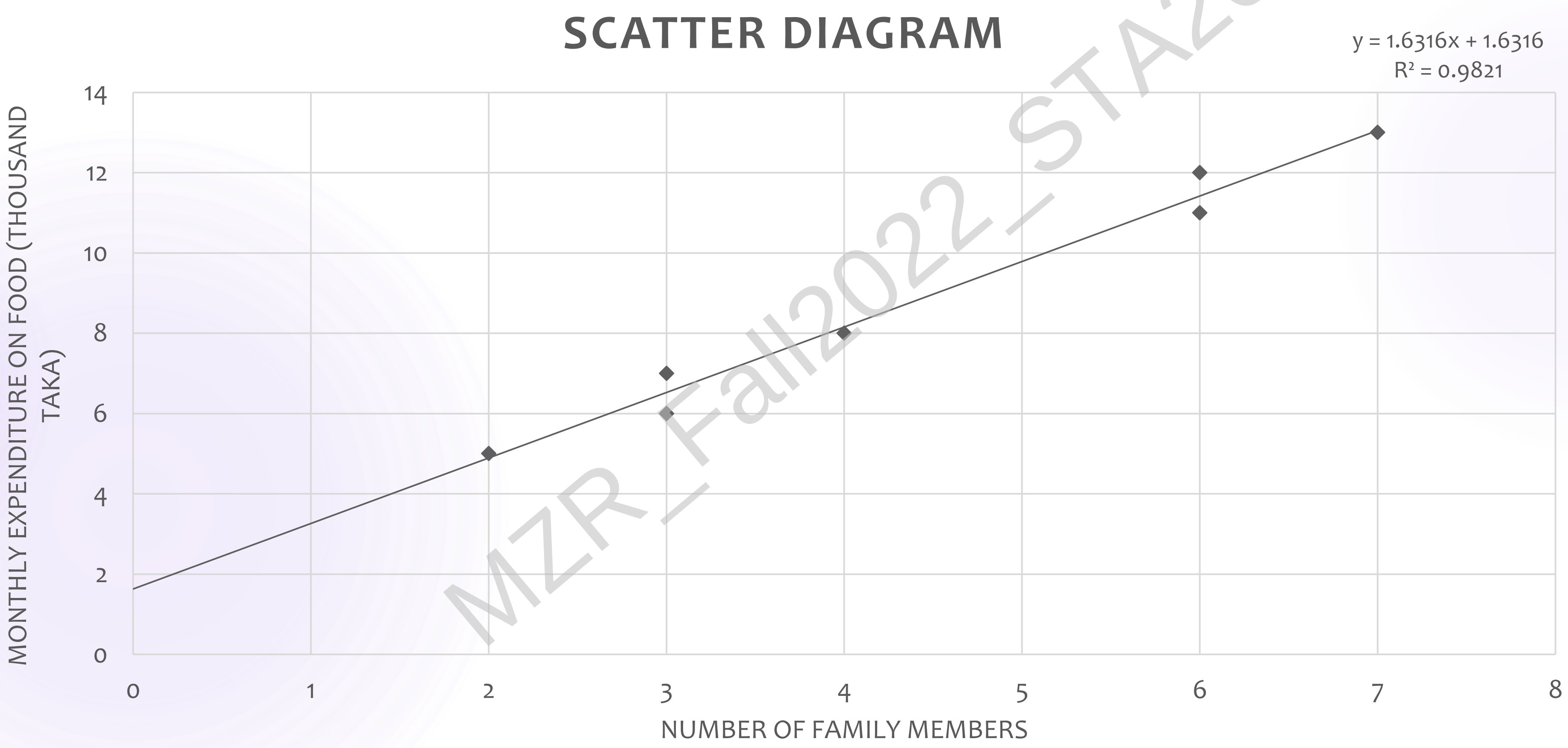
Fit a regression line of **y** on **x**. Interpret the estimates of the parameters. Find the value of R-square. Comment on your result. Estimate that how much monthly expenditure on food would occur if number of family members is 10.

Regression


SCATTER DIAGRAM
MONTHLY EXPENDITURE ON FOOD (THOUSAND TAKA)
14
12
10
8
6
4
2
0
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

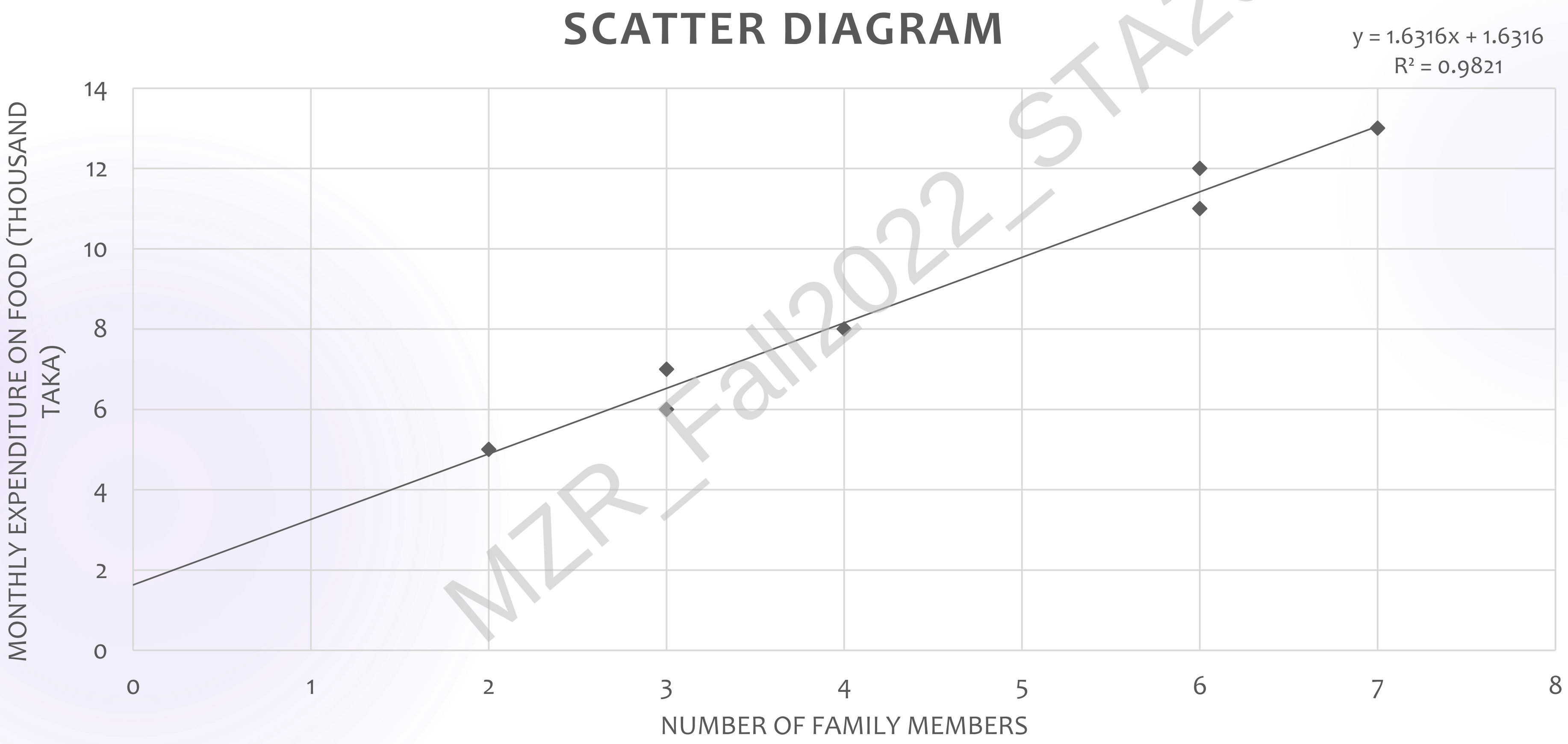
b= 1.63 means, if number of family members is increased by 1 member (i.e. if 1 member is added), on average, monthly expenditure on food will increase by 1.63 (thousand taka)

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



Prediction (For example data)

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

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

Some other relevant calculation

- Calculation of standard error of estimate:

$$s_e = \sqrt{\frac{\sum Y^2 - \hat{\alpha}\sum Y - \hat{\beta}\sum XY}{n-2}}$$

- Determine the standard error of b:

$$s_b = \frac{s_e}{\sqrt{\left(\sum X^2 - n\bar{X}^2\right)}}$$

Problem

For the following data set

X	13	16	14	11	17	9	13	17	18	12
Y	6.2	8.6	7.2	4.5	9.0	3.5	6.5	9.3	9.5	5.7

Answer the following

i. Plot the scatter diagram

ii. Develop the estimating equation that best describe the data/ fit regression line of y on x / Fit regression line of y on x by LSM.

iii. Predict Y for X= 10, 15, 20.

iv. Calculate the standard error of estimate or estimate mean square error (MSE).

v. Determine the standard error of b.

Goodness of fit

R-square:

Total variation = Explained variation + Unexplained variation

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

Notice that, $\boldsymbol{r^2 = R^2}$. *(Coefficient of Determination)*

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.

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. $Corr(X,\epsilon) = r_{X\epsilon} = 0$

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