

# Linear Regression

# Introduction

Many situations arise in which two things appear to be related. For example, it is reasonable to speculate that there is a relationship between:

- ▶ The number of litres of gasoline sold at service stations and the volume of traffic passing their locations;
- ▶ The number of meals sold in a company cafeteria and the number of company employees;
- ▶ Advertising expenditures and sales revenue.

We already know that in the analysis of the relationship, the two things (variables) are identified by the symbols  $x$  and  $y$ . The variable denoted by  $y$  is called the dependent variable. The variable denoted by  $x$  is referred to as the independent variable. The use of these two labels implies that the value of the variable  $y$  depends on the value of the variable  $x$ .



# Introduction

**Regression** is an analytic technique for determining the relationship between a dependent variable and an independent variable. When the two variables have a linear correlation, we can develop a simple mathematical model of the relationship between the two variables by finding a line of best fit.

We used a **scatter plot** to show the joint distribution of two variables in which each point on the graph represents a pair of values. Now we will estimate the line of best fit on a scatter plot.

# Least – Squares Method

It is fairly easy to “eyeball” a good estimate of the line of best fit on a scatter plot when the linear correlation is strong. However, an analytic method using a **least-squares method** gives more accurate results, especially for weak correlations.

**The least squares method** provides a mathematical procedure for determining the equation of a straight line that best fits the data. The equation is referred to as the regression equation.

The graphical representation of the equation is known as the regression line, or line of best fit. This line is called the line of best fit since the procedure minimizes the sum of the vertical deviations (residuals) of the data points about the line.



# Least – Squares Method

The general form of the least squares equation is given by:

$$y_p = ax + b$$

$a$  is the slope of the regression line and indicates the change in the dependent variable  $y$  for a change of one unit in the independent variable  $x$ ;

$b$  is the value of  $y$  when  $x = 0$ ; that is,  $b$  is the  $y$  intercept;

$x$  is a selected value of the independent variable;

$y_p$  is the predicted (or computed) value of the dependent variable for a given value of  $x$ .

# Least – Squares Method

The regression equation for a specific set of data can be uniquely determined by computing the values of the regression coefficients a and b from the following formulas:

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{\left[ n \sum x^2 - (\sum x)^2 \right]}$$

$$b = \frac{\sum y}{n} - a \frac{\sum x}{n}$$



# Least – Squares Method

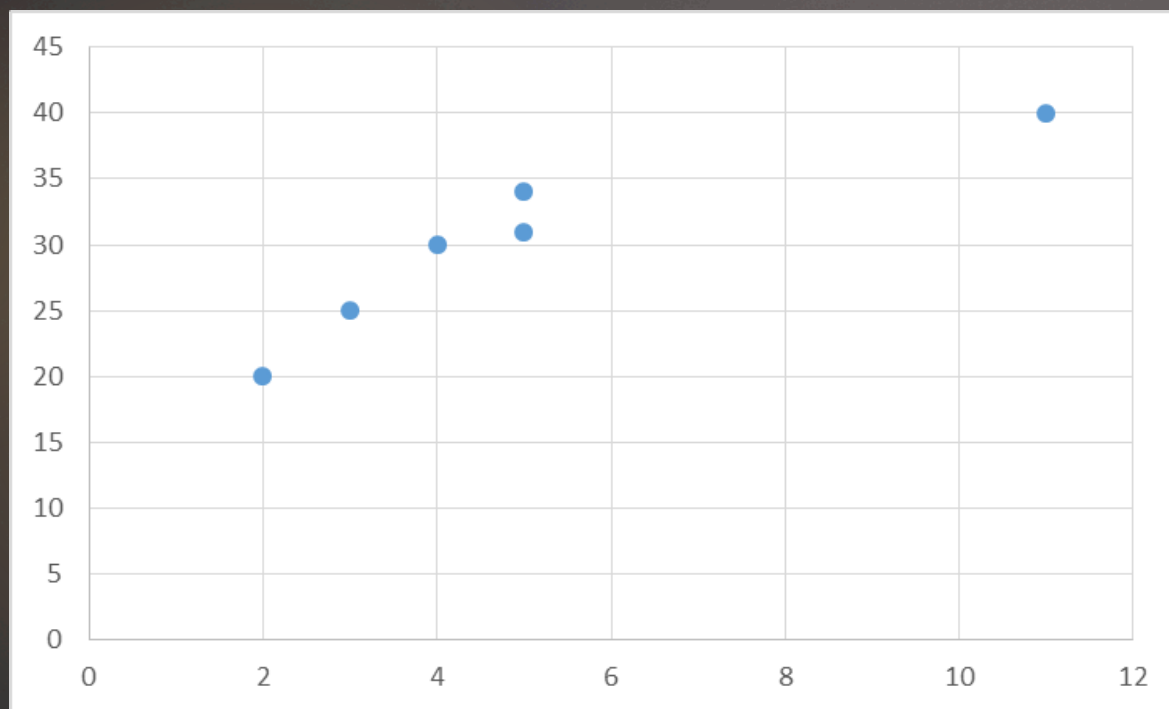
**Example 1:** Many companies involved in assembly operations use aptitude tests on potential employees before hiring them and on current employees before promoting them to more demanding tasks. To obtain a reading on the usefulness of a particular test, the personnel department of Mega Tech, Inc., has administered the aptitude test to a group of employees. The resulting test scores matched to output data are given below. Determine the regression equation.

Employee	Output	Test Score
A	31	5
B	40	11
C	30	4
D	34	5
E	25	3
F	20	2

# Least – Squares Method

## Example 1:

The scatter diagram is shown below (later in this activity, we will learn how to use an Excel spreadsheet to create scatter plots):



The scatter diagram indicates that there is a reasonable linear relationship between the two variables. Use of the least squares method is appropriate to create the regression line.



# Least – Squares Method

## Example 1:

To determine the specific regression equation we must compute the values of the regression coefficients  $a$  and  $b$ . To do so we first need to determine:

$\sum x, \sum y, \sum xy$  and  $\sum x^2$ .

$$\sum x = 5 + 11 + 4 + 5 + 3 + 2 = 30$$

$$\sum y = 31 + 40 + 30 + 34 + 25 + 20 = 180$$

$$\sum xy = 155 + 440 + 120 + 170 + 75 + 40 = 1000$$

$$\sum x^2 = 25 + 121 + 16 + 25 + 9 + 4 = 200$$

Employee	Output (y)	Test Score (x)
A	31	5
B	40	11
C	30	4
D	34	5
E	25	3
F	20	2

# Least – Squares Method

## Example 1:

The values of the regression coefficients can now be found by substituting the appropriate values into the general form of the least squares equation.

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{[n \sum x^2 - (\sum x)^2]}$$

$$a = \frac{6(1000) - (30)(180)}{6(200) - (30)^2}$$

$$a = 2$$

$$b = \frac{\sum y}{n} - a \frac{\sum x}{n}$$

$$b = \frac{180}{6} - 2 \frac{30}{6}$$

$$b = 20$$



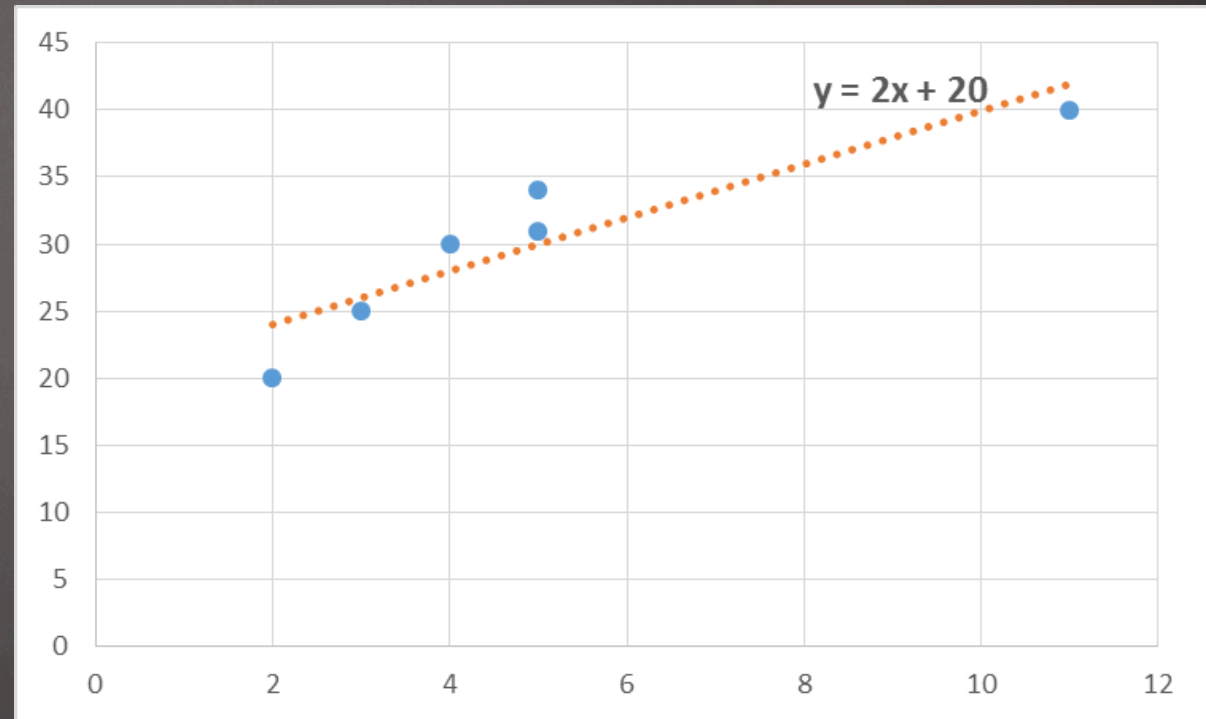
# Least – Squares Method

## Example 1:

Thus the regression equation is given by:

$$y_p = ax + b$$

$$y_p = 2x + 20$$



# Least – Squares Method

## Example 1:

Follow up – In this example, we can also determine the coefficient of correlation for the data given in the table.

Recall, we already determined:

$$\sum x = 30$$

$$\sum y = 180$$

$$\sum xy = 1000$$

$$\sum x^2 = 200$$

However, now we also need:

$$\sum y^2 = 31^2 + 40^2 + 34^2 + 25^2 + 20^2$$

$$\sum y^2 = 5642$$



# Least – Squares Method

## Example 1:

So, the coefficient of correlation is:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{\left[ n \sum x^2 - (\sum x)^2 \right] \left[ n \sum y^2 - (\sum y)^2 \right]}}$$

$$r = 0.9091$$

Therefore, there is a strong positive correlation between test score and output. The higher scores on the test, the higher the output will be.

# Least – Squares Method

## Example 2:

This table shows data for the full-time employees of a small company.

**a)** Use a scatter plot to classify the correlation between age and income.

**b)** Find the equation of the line of best fit analytically.

**c)** Predict the income for a new employee who is 21 and an employee retiring at age 65.

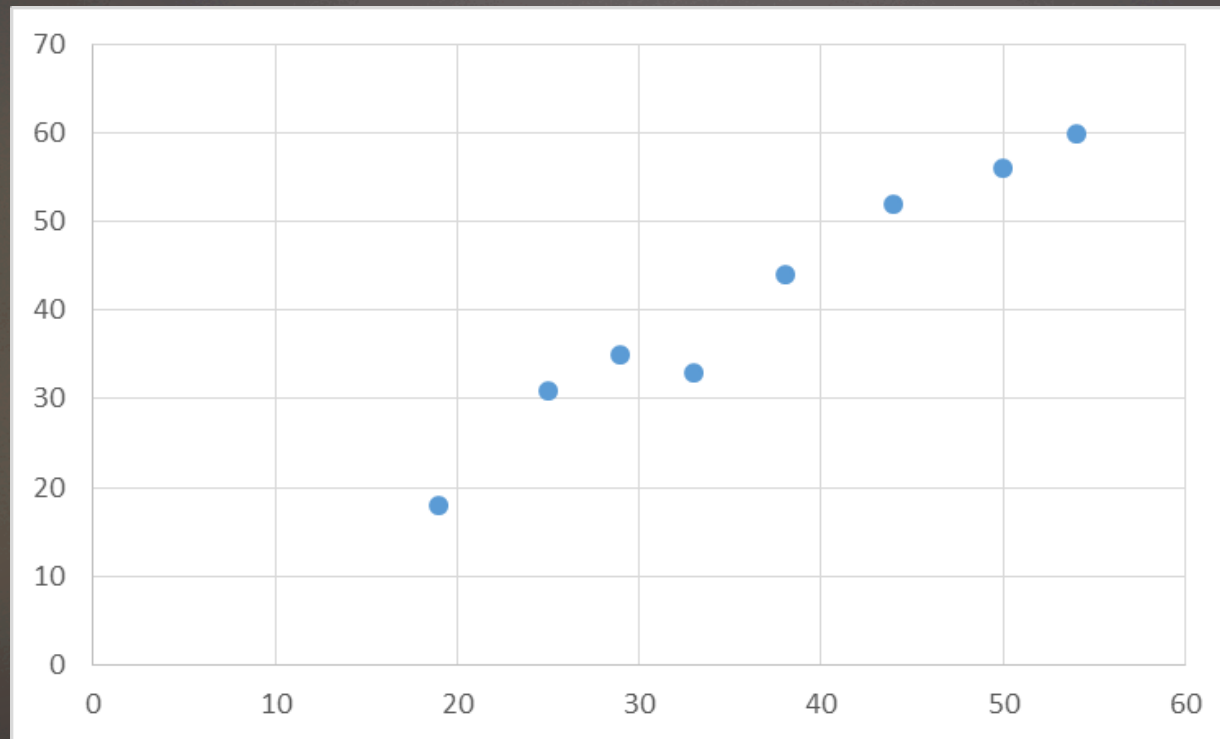
Age (years)	Annual Income (\$000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35



# Least – Squares Method

## Example 2:

**a)** The scatter plot suggests a strong positive linear correlation between age and income level.



# Least – Squares Method

## Example 2:

b) As we did in the previous example. We first need to determine:

$\sum x$ ,  $\sum y$ ,  $\sum xy$  and  $\sum x^2$ .

$$\sum x = 292$$

$$\sum y = 329$$

$$\sum xy = 13221$$

$$\sum x^2 = 11712$$

Age (years)	Annual Income (\$000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35



# Least – Squares Method

## Example 2:

**b)** The values of the regression coefficients can now be found by substituting the appropriate values into the general form of the least squares equation.

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{[n \sum x^2 - (\sum x)^2]}$$

$$a = \frac{8(13221) - (292)(329)}{8(11712) - (292)^2}$$

$$a = 1.15$$

$$b = \frac{\sum y}{n} - a \frac{\sum x}{n}$$

$$b = \frac{329}{8} - (1.15) \frac{292}{8}$$

$$b = -0.85$$

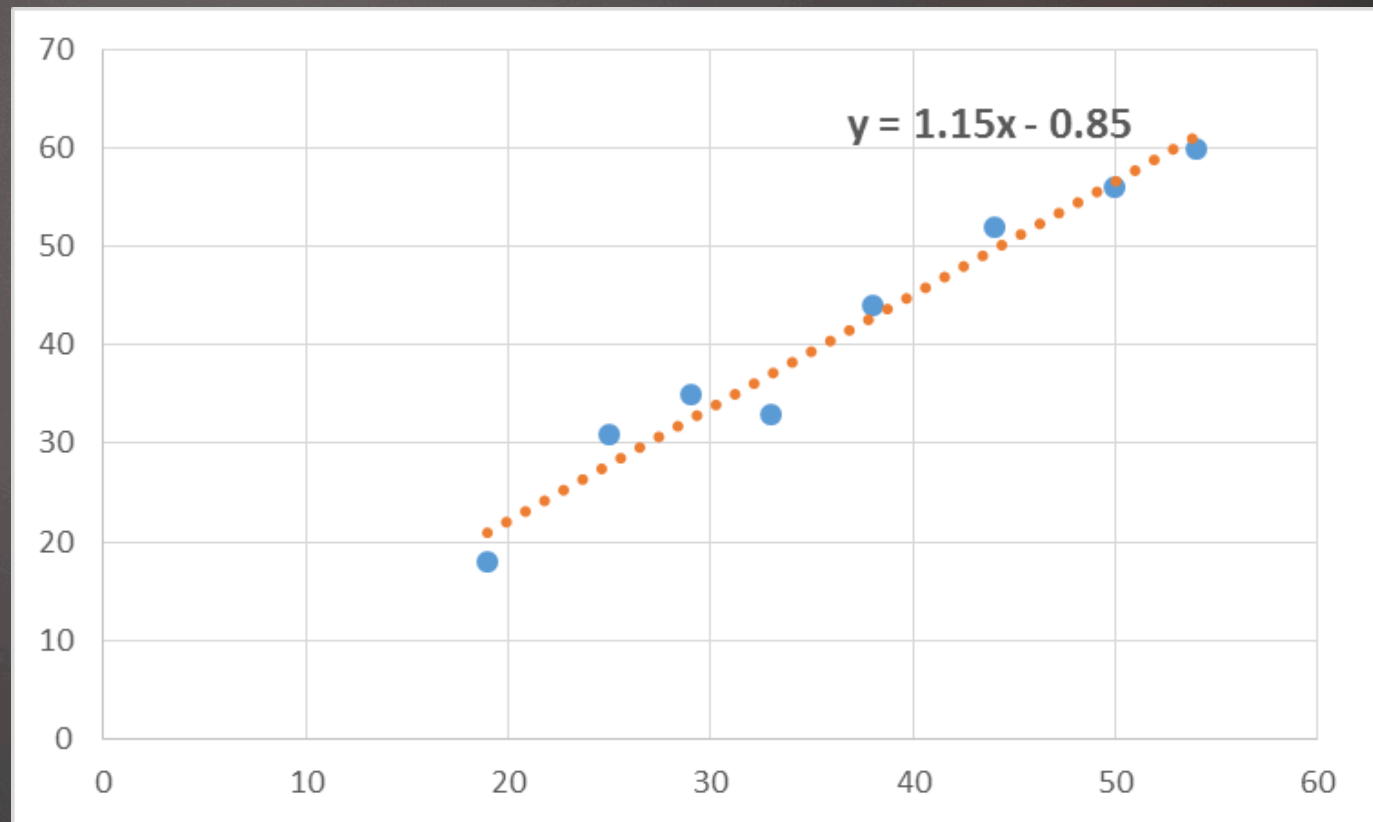
# Least – Squares Method

**Example 2:**

**b)** Thus the regression equation is given by:

$$y_p = ax + b$$

$$y_p = 1.15x - 0.85$$





# Least – Squares Method

## Example 2:

c) Use the equation of the line of best fit as a model.

For a 21-year-old employee,

$$y_p = ax + b$$

$$y_p = 1.15x - 0.85$$

$$y_p = 1.15(21) - 0.85$$

$$y_p = 23.3$$

For a 65-year-old employee,

$$y_p = ax + b$$

$$y_p = 1.15x - 0.85$$

$$y_p = 1.15(65) - 0.85$$

$$y_p = 73.9$$

Therefore, you would expect the new employee to have an income of about \$23,300 and the retiring employee to have an income of about \$73,900.

# Least – Squares Method

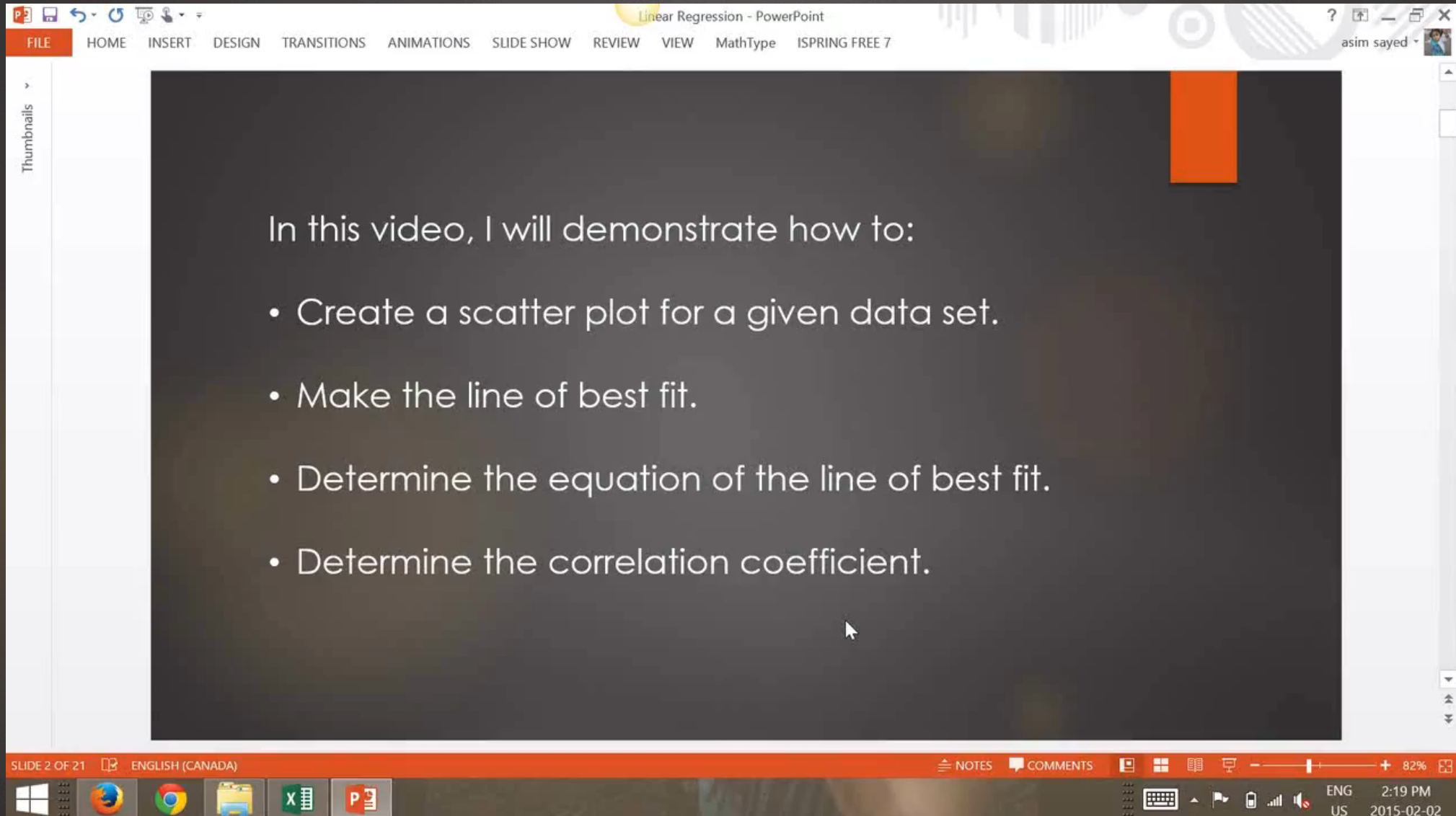
- ▶ Note that the slope  $a$  indicates only how  $y$  varies with  $x$  on the line of best fit.
- ▶ The slope does not tell you anything about the strength of the correlation between the two variables.
- ▶ It is quite possible to have a weak correlation with a large slope or a strong correlation with a small slope.



# Brief Summary

- ▶ Linear regression provides a means for analytically determining a line of best fit.
- ▶ In the least-squares method, the line of best fit is the line which minimizes the sum of the squares of the residuals while having the sum of the residuals equal zero.
- ▶ We can use the equation of the line of best fit to predict the value of one of the two variables given the value of the other variable.
- ▶ The correlation coefficient is a measure of how well a regression line fits a set of data.

# Tutorial – MS Excel 2013



Linear Regression - PowerPoint

FILE HOME INSERT DESIGN TRANSITIONS ANIMATIONS SLIDE SHOW REVIEW VIEW MathType ISPRING FREE 7

asim sayed

Thumbnails

In this video, I will demonstrate how to:

- Create a scatter plot for a given data set.
- Make the line of best fit.
- Determine the equation of the line of best fit.
- Determine the correlation coefficient.

SLIDE 2 OF 21 ENGLISH (CANADA)

NOTES COMMENTS

82%

ENG US 2:19 PM 2015-02-02