BUSINESS STATISTICS (ISD 152)

Lecturer: Emmanuel Quansah

Dept: Supply Chain and Information Systems Dept.

Office: TF 32, KSB Undergraduate Block

CORRELATION & REGRESSION

Overview

- Correlation
- Some Terminology
- Scatter Plots and Correlation
  - Characteristics of Correlation
  - Correlation Coefficient
- Regression
  - Correlation and Regression
  - Simple Regression Analysis
  - Linear Regression Model/Equation

Correlation

Correlation is a statistical method used to determine whether a linear relationship exists between any two variables. Correlation is used to answer questions like;

- Are two or more variables related?
- If so what is the strength of the relationship?

Some Terminology 1

- **Dependent Variable -** A dependent or outcome variable is any variable that changes its value in response to another variable known as the independent variable. It is also known as the response variable or resultant variable or outcome variable. Example; student performance in exams (or exam scores)
- **Independent Variable -** An independent variable is a variable that is adjusted to predict another variable known as the dependent variable. In other words it influences another variable. Another name for the independent variable is explanatory variable or predictor variable. Example, amount of time students spend studying (independent variable) influences the students performance (dependent variable)

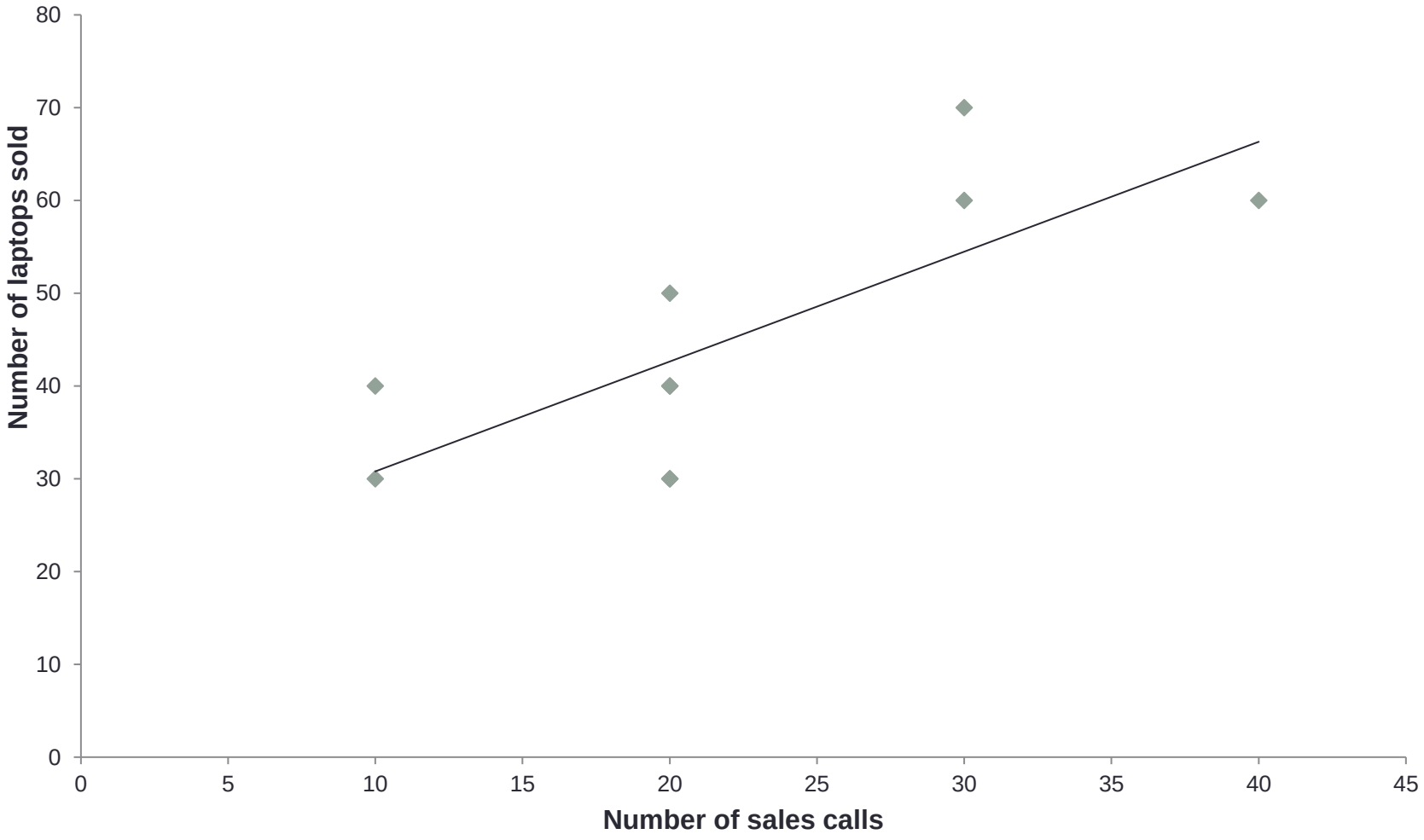
Scatterplots and Correlation

- Independent and dependent variables can be on a graph called a **scatter plot.**
- The independent variable “x” is plotted on the horizontal axis and the dependent variable “y” is plotted on the vertical axis.
- When a scatter plot is drawn and the values of one variable are seen to be related to the value of another variable, the two variables are said to be correlated.

Example 5.1: The data below shows the no. of laptops sold by a sales representative and the number of calls made by the sales representative

No of calls made	20	40	20	30	10	10	20	20	20	30
No. of laptops sold	30	60	40	60	30	40	40	50	30	70

Scatter Plot of Example 5.1



Characteristics of Correlation

A correlational relationship has specific characteristics which enable us to determine the type of correlation as well as the strength of the correlation.

a. Positive correlation: As one variable increases so does the other. If one variable reduces the other variable also reduces. This is indicated on the scatter plot by an upward sloping straight line.

Characteristics of Correlation cont'd

b. Negative correlation: In a negative correlation, as one variable increases, the other decreases. Thus, high value of one variable are associated with low value of the other variable. This is indicated on the scatter plot by a downward sloping straight line.

Characteristics of Correlation cont'd

c. Perfect correlation: If the relationship or correlation between the two variables is perfect, the points in the scatter diagram will lie exactly in a straight line, thus an exact linear relationship exists between the two variables. If this perfect correlation is positive, the line will slope upward and when it is negative, the line will slope downward.

Characteristics of Correlation cont'd

d. Partial correlation: In reality, the relationship between two variables is rarely perfect. Thus in most cases, the points (variables) will lie close to the straight line but not lie perfectly on it. If the points are clustered close to the straight line the relationship is considered as quite strong. As the points get more spread out from the line, the relationship gets weaker. A partial correlation may be positive or negative.

Characteristics of Correlation cont'd

e. No correlation: There is no correlation (or relationship) between the two variables if the scatter plot displays a random scatter of points which are neither going up or down. This means that a linear relationship does exist not exist between these variables.

Product Moment Correlation Coefficient (r)

The product moment correlation coefficient (r) provides a numerical method of determining the type of relationship between two variables (positive and negative) and how strong this relationship is. It measures the strength of a linear relationship.

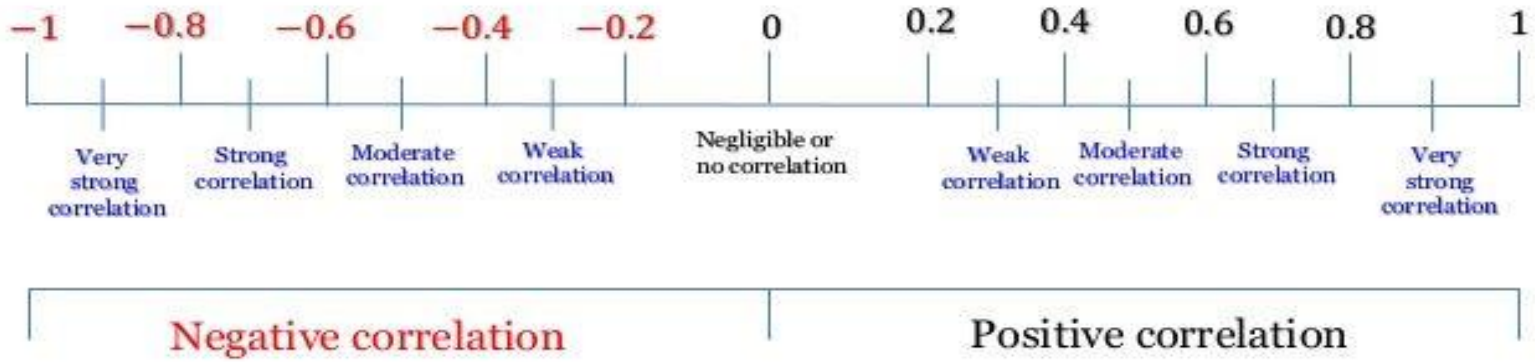
Range of “r”

The range of the correlation coefficient is between **-1 and +1.** When the value of r is near + 1, there is a strong positive linear relationship. When r is near -1, there is a strong negative relationship. When the value of r is near 0, the linear relationship is weak or non existent.

Range of "r"

Correlation Coefficient Interpretation Guideline

The correlation coefficient (r) ranges from -1 (a perfect negative correlation) to 1 (a perfect positive correlation). In short, $-1 \leq r \leq 1$.



Product Moment Correlation Coefficient (r)

- **The Correlation Coefficient**: A single summary number that tells you whether a relationship exists between two variables, how strong that relationship is and whether the relationship is positive or negative (indicated as **R**).

- $$\mathbf{R} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2 \times \sum(y-\bar{y})^2}}$$

- **The Coefficient of Determination**: A single summary number that tells you how much variation in one variable is directly related to variation in another variable (indicated as $\mathbf{R^{2)}}$

Example 5.2: Using the data below, find the correlation coefficient (r).

No of calls made	**20**	**40**	**20**	**30**	**10**	**10**	**20**	**20**	**20**	**30**
No. of laptops sold	**30**	**60**	**40**	**60**	**30**	**40**	**40**	**50**	**30**	**70**

Example 5.2: Solution

No of Calls (x)	No of laptops sold (y)	$(y-\bar{y})$	$(x-\bar{x})$	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})^2$
20	30	-15	-2	30	225	4
40	60	15	18	270	225	324
20	40	-5	-2	10	25	4
30	60	15	8	120	225	64
10	30	-15	-12	180	225	144
10	40	-5	-12	60	25	144
20	40	-5	-2	10	25	4
20	50	5	-2	-10	25	4
20	30	-15	-2	30	225	4
30	70	25	8	200	625	64
Mean of x = 22 Mean of y = 45				**900**	**1850**	**760**

Solution:

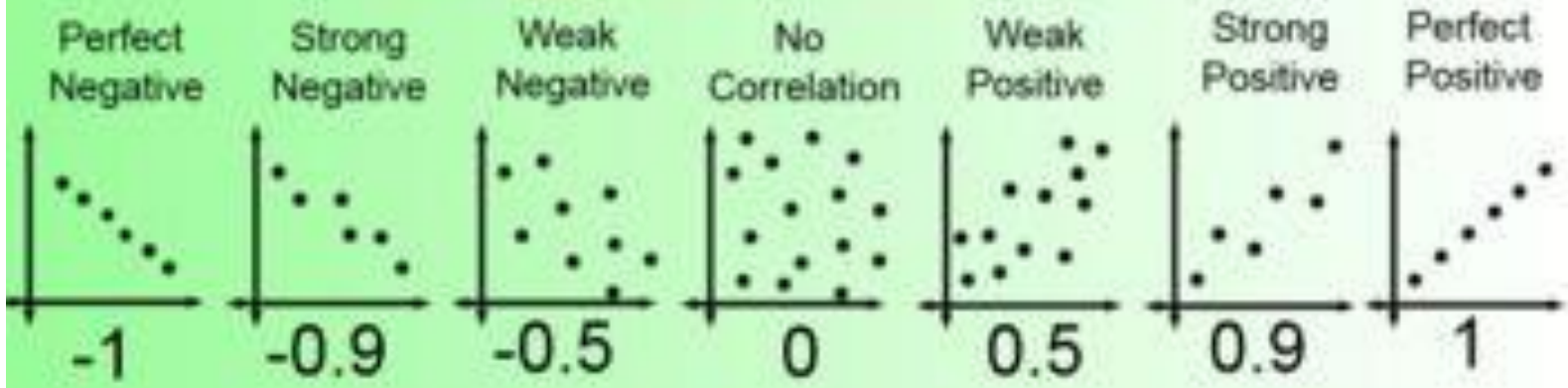
$$R = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2 \times \sum(y-\bar{y})^2}} \qquad R = \frac{900}{\sqrt{760\ x\ 1850}} = 0.7590$$ (strong positive correlation)

Scatter Plots & Correlation Coefficient Values

Correlation Coefficient

The measure of the strength of the line of best fit.

Correlation Coefficient Values



1 is a perfect positive correlation

0 is no correlation (the values don't seem linked at all)

-1 is a perfect negative correlation

Regression Analysis

Correlation & Regression

- **Correlation -** is a statistical method used to ***determine*** whether a ***linear relationship*** exists between any two variables. It allows you to find out if there is a statistically significant relationship between TWO variables

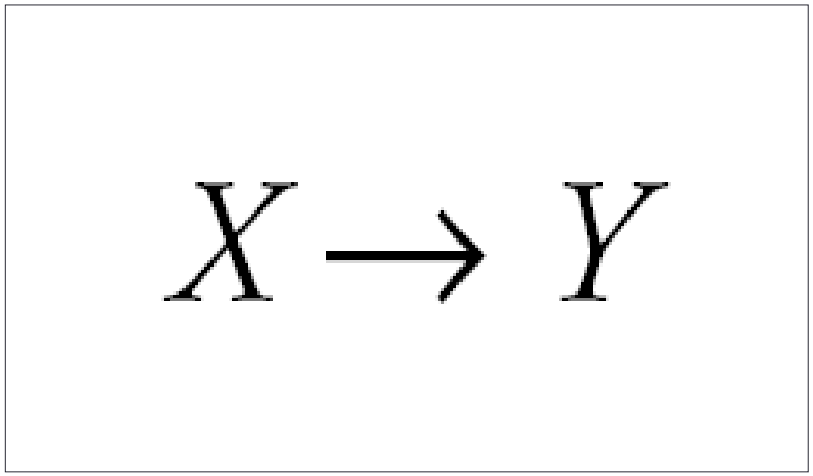
- **Regression -** is a statistical method used to ***describe the nature of the relationship*** (causation) between any two variables. It allows you to make predictions based on the relationship that exists between two variables.
- Correlation shows there is a linear relationship. Correlation however, does not mean causation.
- To investigate causation, we use regression analysis.

Simple Regression Analysis

- In a simple regression analysis, there is a simple relationship between two variables (an independent variable and a dependent variable).
- The independent variable is needed to predict the depended variable.
- Example: the relationship between the year of experience of a salesman and the amount of sale he makes involves a simple relationship between two variables – years of experience and amount of sales.

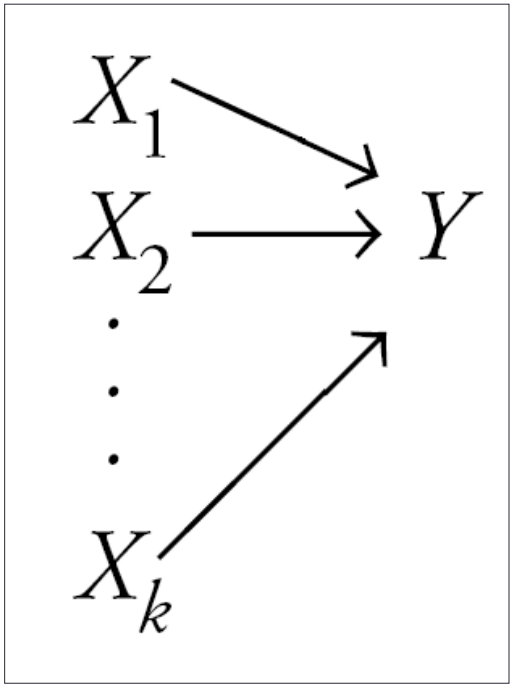
The General Idea

- **Simple Regression Analysis**
  - Considers the relationship between two variables (an independent variable (X) and a dependent variable (Y)) OR between a single explanatory variable and response variable



The General Idea

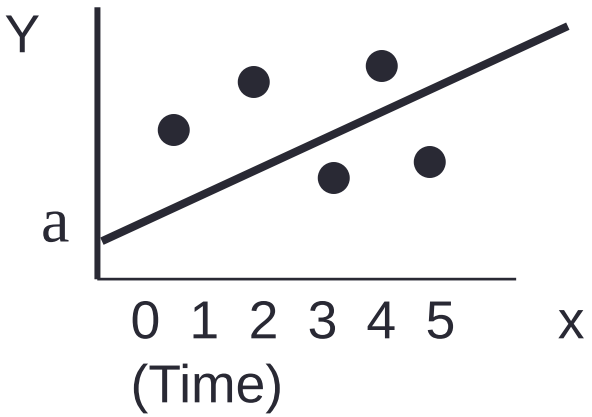
- **Multiple regression** simultaneously considers the influence of multiple explanatory variables (X1, X2…Xk) on a response variable Y



The intent is to look at the independent effect of each variable as well as the combined effect

Simple Linear Regression Model

A regression analysis can be used to develop an equation showing how the dependent variable ***(y)*** is related to the independent variable ***(x)***. A simple linear regression equation can be expressed as:



$$Y' = a + bx$$

Y' is the dependent variable in the model,

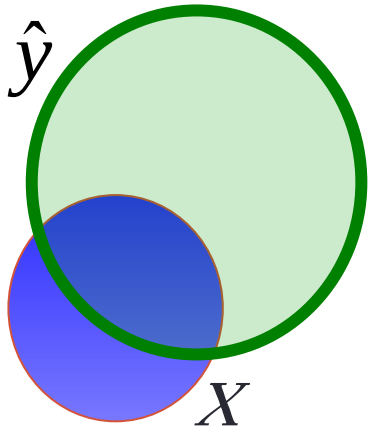
x is independent variable in the model

a is the intercept value of the regression line, and

b is the slope of the regression line.

Simple Linear Regression Formulas for Calculating "a" and "b"

$$Y^{l} = a + bx$$



a= $\bar{y}$-b$\bar{x}$

b= $\frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sum(x-\bar{x})^2}$

(Formula 1)

b= $\frac{\sum xy-n\,(\bar{y})(\bar{x})}{\sum x^2-n(\bar{x})^2}$

(Formula 2)

Where;

xi/x = values of independent var.

yi/y = values of dependent var.

$\bar{y}$ = mean of the dependent var.

$\bar{x}$ = mean of independent var.

n = total no. of observations

Simple Linear Regression Example

Example 5.3: Given the data below, what is the simple linear regression model that can be used to predict sales?

Week	Sales
1	150
2	157
3	162
4	166
5	177

First, using the linear regression formulas, we can compute “a” and “b”.

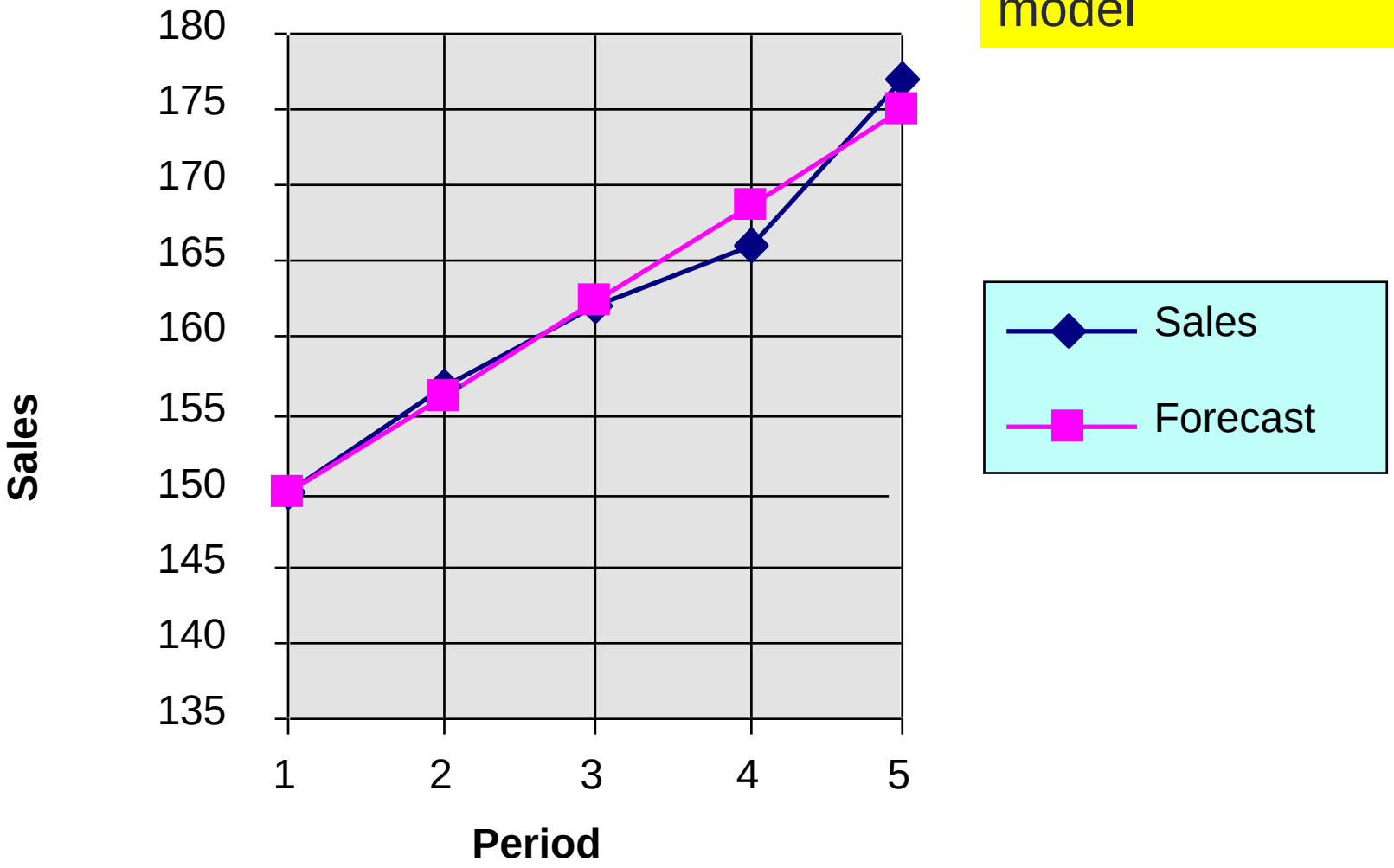
Week (x)	x squared	Sales (y)	xy
1	1	150	150
2	4	157	314
3	9	162	486
4	16	166	664
5	25	177	885
Avg of x = 3	**Sum of x squared = 55**	**Avg of y = 162.4**	**Sum of xy = 2499**

$$b = \frac{\sum xy - n(\bar{y})(\bar{x})}{\sum x^2 - n(\bar{x})^2} = \frac{2499 - 5(162.4)(3)}{55 - 5(9)} = \frac{63}{10} = \mathbf{6.3}$$

$$a = \bar{y} - b\bar{x} = 162.4 - (6.3)(3) = \mathbf{143.5}$$

$Y^{I} = 143.5 + 6.3x$

Resulting regression model



Now if we plot the regression generated forecasts against the actual sales we obtain the following chart:

Example 5.4:

- In the previous example, the manager after determining that there is a relationship between sales calls and sales volume, wants to predict future sales based on sales calls.
- A regression analysis can be used to develop an equation showing how the dependent variable(y) is related to the independent variable (x).

Example 5.4:

i) Determine the simple linear equation for this data.

ii) Assuming 35 calls are made, what will be the predicted number of laptops sold?

No of calls made	**20**	**40**	**20**	**30**	**10**	**10**	**20**	**20**	**20**	**30**
No. of laptops sold	**30**	**60**	**40**	**60**	**30**	**40**	**40**	**50**	**30**	**70**

Solve this on your own

Multiple Regression Analysis

- In **a multiple regression analysis;**
  - there are two or more independent variables that are used to predict on dependent variable.
  - For example a persons health condition may depend on the number of factors such as genetics/hereditary, lifestyle (smoking, drinking etc.), poverty etc.
  - Investigating this situation/study will demand the use of multiple regressions since it involves several variables.
  - The concept of multiple regression will be dealt with in detail in a later course

END OF SESSION