Correlation and Linear Regression

Chapter 13

Dr. Sultana Parveen, Professor, IPE Dept, BUET

Learning Objectives

LO1 Define the terms *dependent* and *independent variable*.

LO2 Calculate, test, and interpret the relationship between two variables using the c*orrelation* coefficient.

LO3 Apply regression analysis to estimate the linear relationship between two variables

LO4 Interpret the regression analysis.

LO5 Evaluate the significance of the slope of the regression equation.

LO6 Evaluate a regression equation to predict the dependent variable.

LO7 Calculate and interpret the coefficient of determination.

LO8 Calculate and interpret confidence and prediction intervals.

LO1 Define the terms *dependent* and *independent variable*.

Regression Analysis - Introduction

- Recall in Chapter 4 the idea of showing the relationship between *two* variables with a scatter diagram was introduced.
- In that case we showed that, as the age of the buyer increased, the amount spent for the vehicle also increased.
- In this chapter we carry this idea further. Numerical measures to express the strength of relationship between two variables are developed.
- In addition, an equation is used to express the relationship between variables, allowing us to estimate one variable on the basis of another.

EXAMPLES

1. Is there a relationship between the amount Healthtex spends per month on advertising and its sales in the month?
2. Can we base an estimate of the cost to heat a home in January on the number of square feet in the home?
3. Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
4. Is there a relationship between the number of hours that students studied for an exam and the score earned?

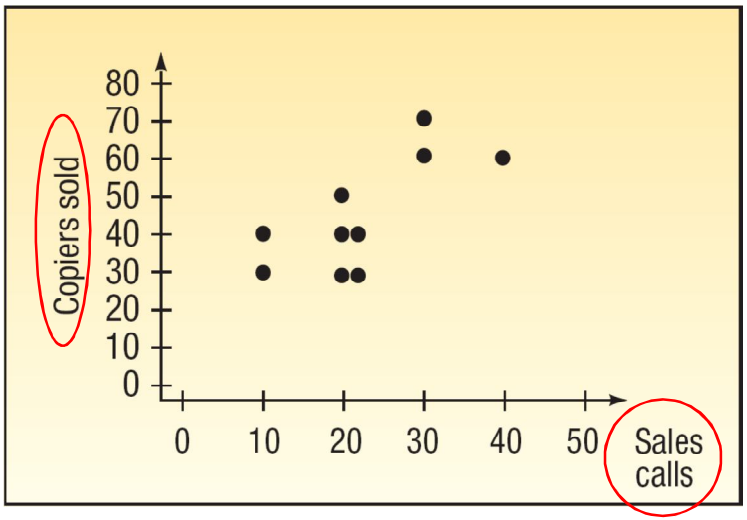
- **Correlation Analysis** is the study of the relationship between variables. It is also defined as group of techniques to measure the association between two variables.
- **Scatter Diagram** is a chart that portrays the relationship between the two variables. It is the usual first step in correlations analysis
- The **Dependent Variable** is the variable being predicted or estimated.
- The **Independent Variable** provides the basis for estimation. It is the predictor variable.

LO1

Scatter Diagram Example

The sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a **relationship between** the **number of sales calls made** in a month and the **number of copiers sold that month**. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made last month and the number of copiers sold.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

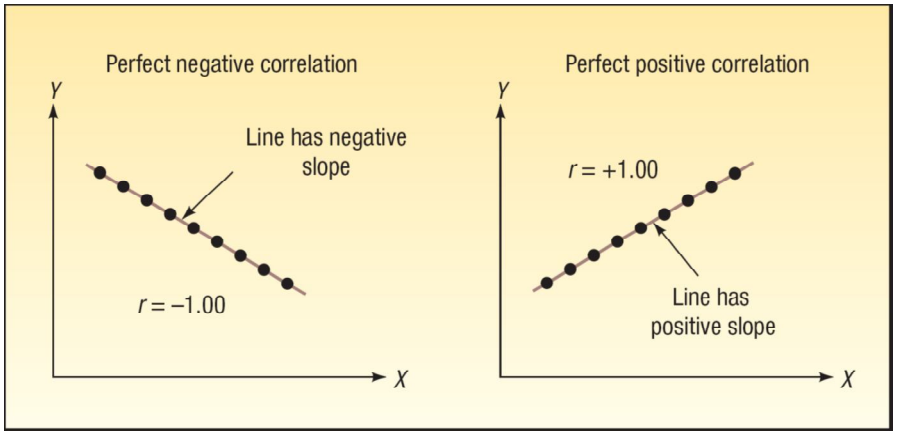


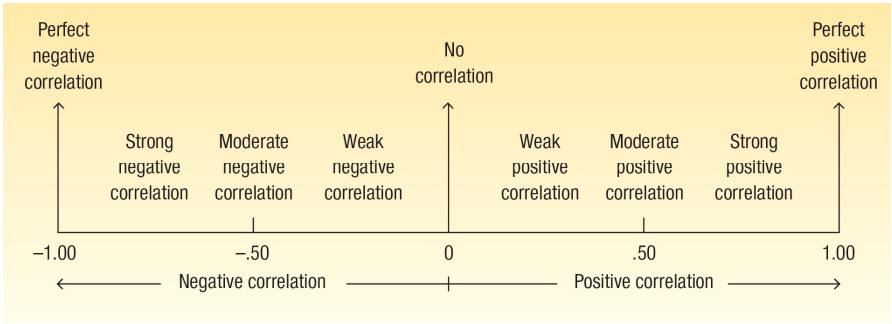
LO2 Calculate, test, and interpret the relationship between two variables using the *correlation* coefficient.

The Coefficient of Correlation, *r*

The **Coefficient of Correlation** (*r*) is a measure of the strength of the relationship between two variables.

- It shows the direction and strength of the linear relationship between two interval or ratio-scale variables
- It can range from -1.00 to +1.00.
- Values of -1.00 or +1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an **inverse** relationship and positive values indicate a **direct** relationship.





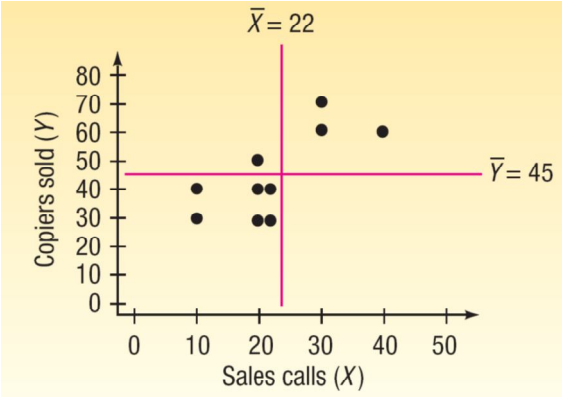
LO2

Correlation Coefficient - Example

EXAMPLE

Using the Copier Sales of America data which a scatterplot is shown below, compute the correlation coefficient and coefficient of determination.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70



Using the formula:

CORRELATION COEFFICIENT

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y}$$

Sales Representative	Calls, Y	Sales, X	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramirez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

How do we interpret a correlation of 0.759?
First, it is positive, so we see there is a direct relationship between the number of sales calls and the number of copiers sold. The value of 0.759 is fairly close to 1.00, so we conclude that the association is strong. However, does this mean that more sales calls ***cause*** more sales? No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are related.

LO3 Apply regression analysis to estimate the linear relationship between two variables.

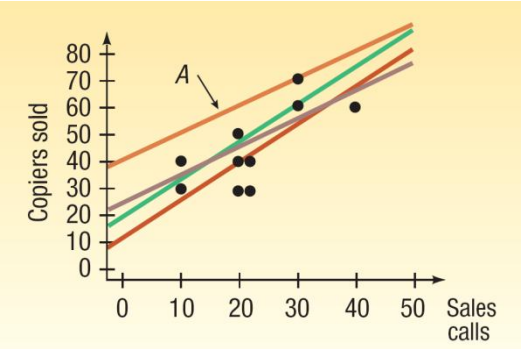
Regression Analysis

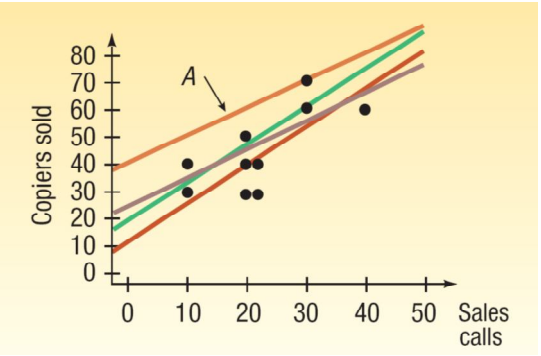
In regression analysis we use the independent variable (*X*) to estimate the dependent variable (*Y*).

- The relationship between the variables is linear.
- Both variables must be at least interval scale.
- The least squares criterion is used to determine the equation.

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual *Y values and* the predicted values of *Y.*





$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

LO3

Linear Regression Model

GENERAL FORM OF LINEAR REGRESSION EQUATION $\hat{Y} = a + bX$

where

$\hat{Y}$ read Y hat, is the estimated value of the Y variable for a selected X value.

a is the Y-intercept. It is the estimated value of Y when $X = 0$. Another way to put it is: a is the estimated value of Y where the regression line crosses the Y-axis when X is zero.

b is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit (either increase or decrease) in the independent variable X.

X is any value of the independent variable that is selected.

SLOPE OF THE REGRESSION LINE $b = r\dfrac{s_y}{s_x}$

where

r is the correlation coefficient.

s_y is the standard deviation of Y (the dependent variable).

s_x is the standard deviation of X (the independent variable).

***Y*-INTERCEPT** $a = \overline{Y} - b\overline{X}$

where

$\overline{Y}$ is the mean of Y (the dependent variable).

$\overline{X}$ is the mean of X (the independent variable).

Regression Equation - Example

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. Use the least squares method to determine a linear equation to express the relationship between the two variables.

What is the expected number of copiers sold by a representative who made 20 calls?

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Step 1 – Find the slope (b) of the line

$$b = r\left(\frac{s_y}{s_x}\right) = .759\left(\frac{14.337}{9.189}\right) = 1.1842$$

Step 2 – Find the y-intercept (a)

$$a = \bar{Y} - b\bar{X} = 45 - 1.1842(22) = 18.9476$$

The regression equation is :

$$\hat{Y} = a + bX$$

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(20)$$

$$\hat{Y} = 42.6316$$