

# **Statistical and Mathematical Methods for Data Analysis**

**Dr. Faisal Bukhari**

**Punjab University College of Information Technology  
(PUCIT)**

# Textbooks

- ❑ **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- ❑ **Elementary Statistics: Picturing the World**, 6<sup>th</sup> Edition, Ron Larson and Betsy Farber
- ❑ **Elementary Statistics**, 13<sup>th</sup> Edition, Mario F. Triola

# Reference books

- ❑ **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- ❑ **Probability Demystified,** Allan G. Bluman
- ❑ **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- ❑ **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- ❑ **Python for Probability, Statistics, and Machine Learning,** José Unpingco

# References

- ❑ Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- ❑ Elementary Statistics, Tenth Edition, Mario F. Triola

These notes contain material from the above resources.

# Correlation

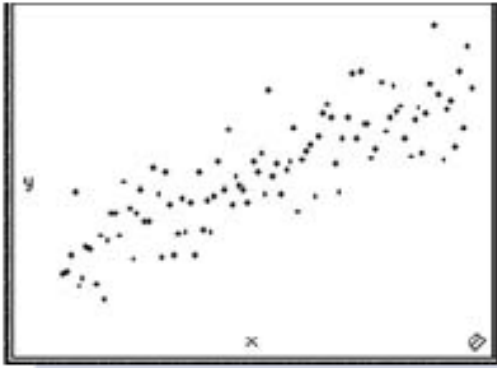
A **correlation** exists between **two variables** when **one** of them is **related** to the other in some way.

# Exploring the Data

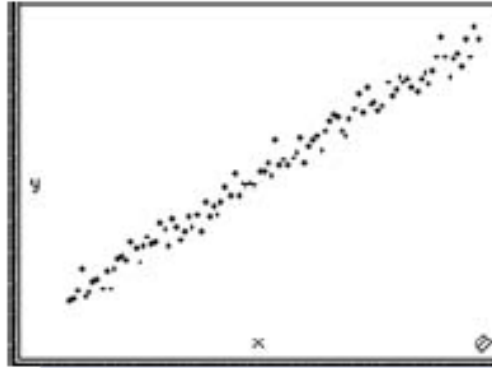
We can often see a **relationship between two variables** by constructing a **scatterplot**. When we examine a **scatterplot**, we should study the **overall pattern** of the plotted points. If there is a pattern, we should note its **direction**.

- ❑ An **uphill direction** suggests that as **one variable increases**, the **other also increases**.
- ❑ A **downhill direction** suggests that as **one variable increases**, the **other decreases**.
- ❑ We should look for **outliers**, which **are points that lie very far away** from all of the **other points**.

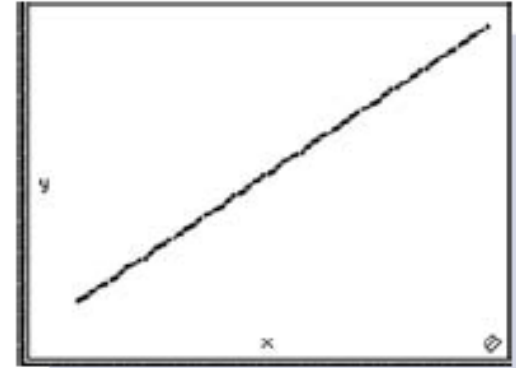
# Scatter plots



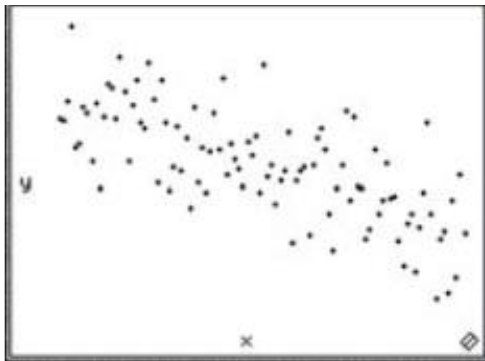
**Positive correlation:**  
 **$r = 0.851$**



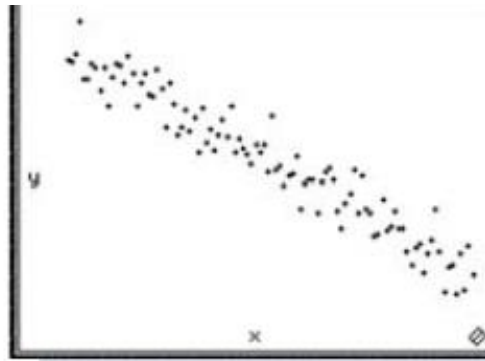
**Positive correlation:**  
 **$r = 0.991$**



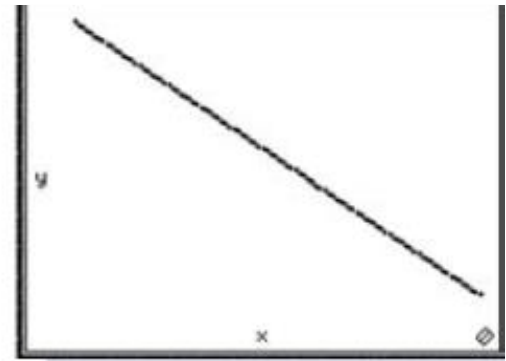
**Perfect positive correlation:**  
 **$r = 1$**



**Negative correlation:**  
 **$r = -0.702$**

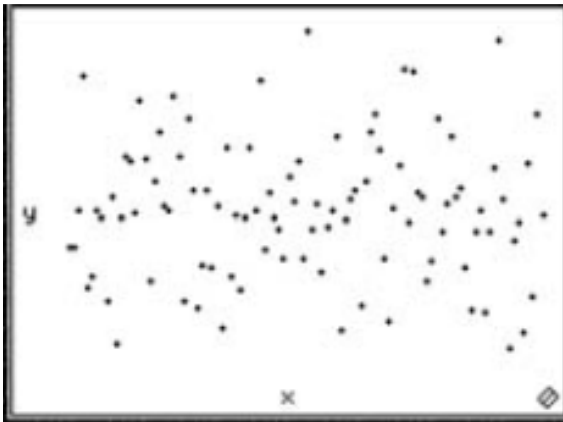


**Negative correlation:**  
 **$r = -0.965$**

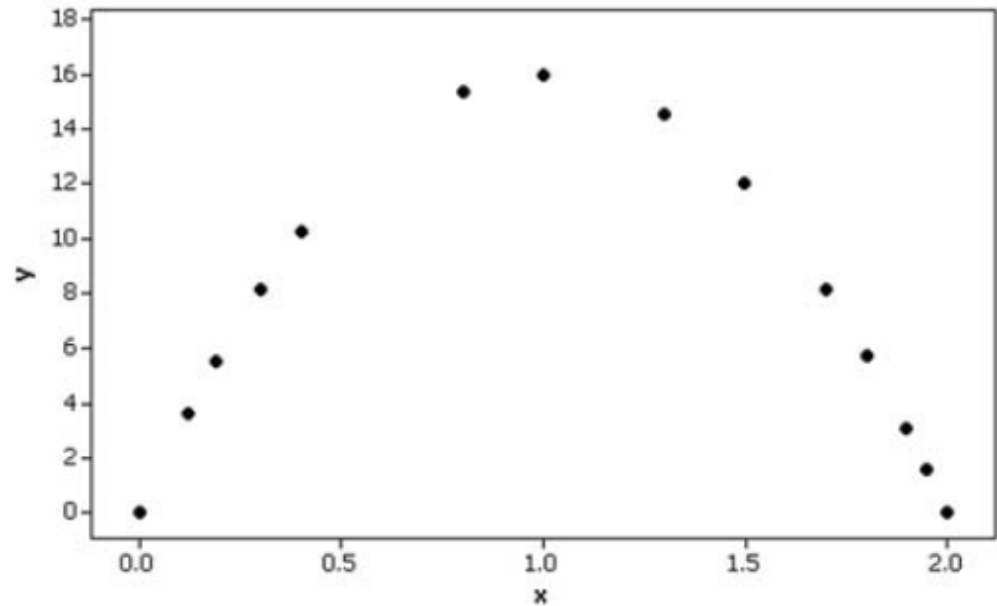


**Perfect negative correlation:**  
 **$r = -1$**

# Scatter plots



No correlation:  $r = 0$



Nonlinear relationship:  $r = -0.087$



# Palm reading



# Palm reading

- ❑ Some **people believe** that the **length of their palm's lifeline** can be used to **predict longevity**.
- ❑ In a letter published in the **Journal of the American Medical Association**, authors M. E. Wilson and L. E. Mather **refuted that belief** with a **study of cadavers**.
- ❑ **Ages at death** were recorded, along with the **lengths of palm lifelines**. The authors concluded that **there is no significant correlation between age at death and length of lifeline**. Palmistry lost, hands down.

# Requirements

Given any **collection of sample paired data**, the linear **correlation coefficient  $r$**  can always be computed, but the following requirements should be satisfied when **testing hypotheses** or making other **inferences about  $r$** .

1. The sample of **paired  $(x, y)$**  data is a ***random sample*** of **independent quantitative data**.
2. **Visual examination** of the **scatterplot** must confirm that the points approximate a **straight-line pattern**.
3. **Any outliers** must be removed if they are known to **be errors**. The effects of any other outliers should be considered by **calculating  $r$  with** and **without the outliers** included.

- ❑ **Note: Requirements 2 and 3** above are simplified attempts at checking this formal requirement:
- ❑ The **pairs of  $(x, y)$**  data must have **a bivariate normal distribution**. (This assumption basically requires that for any **fixed value of  $x$** , the **corresponding values of  $y$**  have a distribution that is **bell-shaped**, and for any fixed value of  $y$ , the values of  $x$  have a **distribution that is bell-shaped**.)
- ❑ This requirement is usually difficult to check, so for now, we will use Requirements 2 and 3 as listed above.

# Notation for the Linear Correlation Coefficient

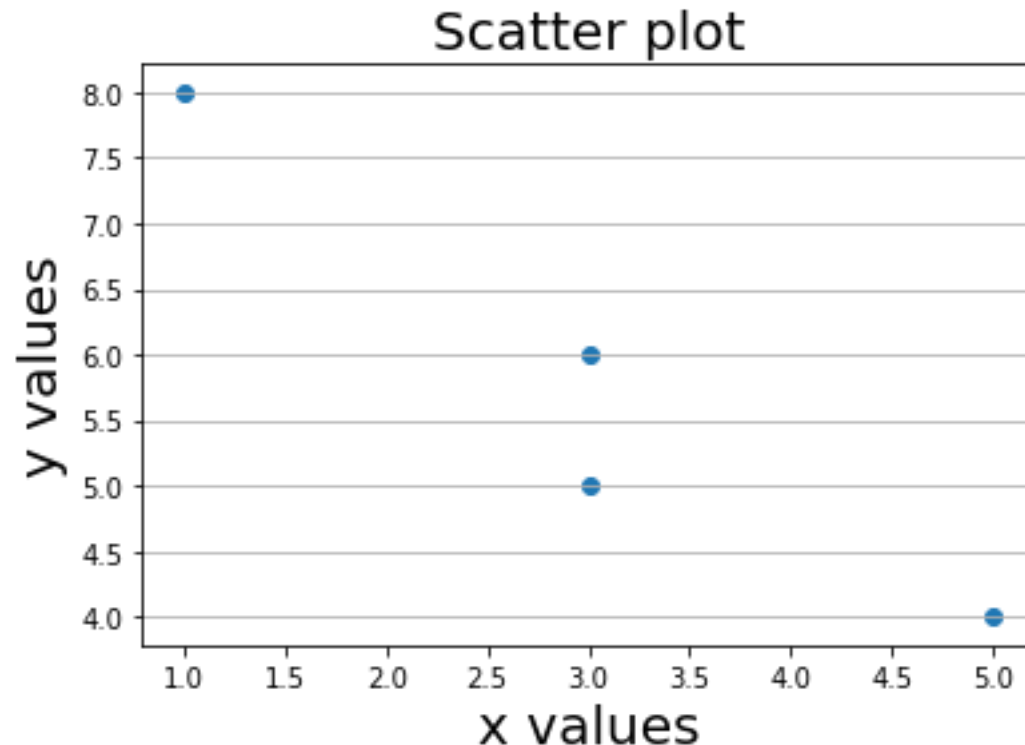
- **n**: represents the **number of pairs** of data present.
- **r**: represents the **linear correlation coefficient** for a **sample**.
- **$\rho$** : Greek letter **rho** used to represent the **linear correlation coefficient** for a **population**.

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

**Example Calculating  $r$**  Using the simple random sample of data given in the table, find the value of the **linear correlation coefficient  $r$** .

<b>x</b>	<b>3</b>	<b>1</b>	<b>3</b>	<b>5</b>
<b>y</b>	<b>5</b>	<b>8</b>	<b>6</b>	<b>4</b>

**REQUIREMENT** The data are a **simple random sample**. The accompanying **Python-generated scatterplot** shows **a pattern of points** that does appear to be a **straight-line pattern**. There are no outliers. We can proceed with the calculation of the linear correlation coefficient ***r***.





$x$	$y$	$xy$	$x^2$	$y^2$
3	5	15	9	25
1	8	8	1	64
3	6	18	9	36
5	4	20	25	16
$\sum x = 12$	$\sum y = 23$	$\sum xy = 61$	$\sum x^2 = 44$	$\sum y^2 = 141$

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

$$r = \frac{4(61) - (12)(23)}{\sqrt{4(44) - (12)^2} \sqrt{4(141) - (23)^2}}$$

$$r = \frac{-32}{\sqrt{32}\sqrt{35}} = -0.956$$

❑ These calculations get quite messy with larger data sets, so it's **fortunate** that the **linear correlation coefficient** can be **found automatically** with many different **calculators** and **computer programs**

# Interpreting the Linear Correlation Coefficient

- ❑ We need to interpret a calculated **value of  $r$** , such as the value of  **$-0.956$**  found in the preceding example.
- ❑ The value of  $r$  must always fall between  **$-1$  and  $+1$**  inclusive.
- ❑ If  **$r$  is close to  $0$** , we conclude that **there is no linear correlation** between  $x$  and  $y$ , but if  $r$  is close  **$-1$  to or  $+1$**  we conclude that there is a **linear correlation between  $x$  and  $y$** .

# Properties of the Linear Correlation Coefficient $r$

1. The value of  $r$  is always between **-1** and **+1** inclusive. That is,  **$-1 \leq r \leq +1$**
2. The value of  $r$  does not change if all values of either variable are **converted** to a **different scale**.
3. The value of  $r$  is **not affected** by the choice of  **$x$**  or  **$y$** . Interchange all  $x$ - and  $y$ -values and the value of  $r$  will not change.
4.  **$r$  measures** the strength of a **linear relationship**. It is **not designed** to measure the **strength of a relationship** that is **not linear**.

# Hypothesis Test for Correlation

Assume:  $r = 0.926$ ,  $n = 8$

1. **We state our hypothesis as:**

$H_0: \rho = 0$  (There is no linear correlation.)

$H_1: \rho \neq 0$  (There is a linear correlation.)

2. **The level of significance is set**  $\alpha = 0.05$ .

3. **Test statistic to be used is**  $t_{\text{cal}} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$

4. **Calculations:**

$$t_{\text{cal}} = \frac{0.926}{\sqrt{\frac{1 - (0.926)^2}{8 - 2}}} = 6.008$$

## 5. Critical region:

$|t_{\text{cal}}| > t_{\text{tab}}$ , where  $t_{\text{tab}} = t_{(\alpha/2, n-2)}$

$$t_{\text{tab}} = t_{(0.0250, 6)} = 2.447$$

$$6.008 > 2.447 \text{ (True)}$$

**6. Conclusion:** Since  $t_{\text{cal}}$  is greater than the  $t_{\text{tab}}$ , so we reject  $H_0$ .

□ There is **sufficient evidence** to support the claim of a linear correlation.

# Examples: Applications of correlation

Buying a TV Audience The *New York Post* published the **annual salaries (in millions)** and the **number of viewers (in millions)**, with results given below for **Oprah Winfrey, David Letterman, Jay Leno, Kelsey Grammer, Barbara Walters, Dan Rather, James Gandolfini, and Susan Lucci**, respectively. Is there **a correlation between salary and number of viewers**?

Salary	100	14	14	35.2	12	7	5	1
Viewers	7	4.4	5.9	1.6	10.4	9.6	8.9	4.2

# Examples: Applications of correlation

**Parent Child Heights Listed** below are **heights (in inches)** of **mothers** and **heights (in inches)** of their daughters (based on data from the National Health Examination Survey). Does there appear to be a **linear correlation between mother's heights** and the **heights of their daughters**?

Mother's height	63	67	64	60	65	67	59	60
Daughter's height	58.6	64.7	65.3	61.0	65.4	67.4	60.9	63.1



# Basic Concepts of Regression

- ❑ In some cases, **two variables** are related in a **deterministic way**, meaning that given **a value for one variable**, the value of the **other variable** is **automatically determined** without any **error**.
- ❑ For example, the **total cost  $y$**  of an item with a list price of  **$x$**  and a **sales tax of 5%** can be found by using the deterministic equation  **$y = 1.05x$** . If an item is priced at **\$100**, its total cost is **\$105**.

# Probabilistic Models

- ❑ In **probabilistic models**, meaning that one variable is **not determined** completely by the **other variable**.
- ❑ For example, a **child's height** is not determined completely by the **height of the father (or mother)**.
- ❑ **Sir Francis Galton (1822–1911)** studied the phenomenon of heredity and showed that when **tall or short couples have children**, the heights of those children **tend to regress, or revert** to the more typical **mean height** for people of the **same gender**.

# Notations

- ❑ The **regression equation** expresses a relationship between  **$x$**  (called the **explanatory variable**, or **predictor variable**, or **independent variable**) and (called the **response variable**, or **dependent variable**).
- ❑ The typical equation of a straight line  **$y = mx + b$**  is expressed in the form  **$\hat{y} = b_0 + b_1x$**  or  **$\hat{y} = a + bx$** , where  **$b_0$  or  $a$**  is the  **$y$ -intercept** and  **$b_1$  or  $b$**  is the **slope**.

- ❑ The given notation shows that  $b_0$  and  $b_1$  are **sample statistics** used to estimate the population parameters  $\beta_0$  and  $\beta_1$ .
- ❑ We will use **paired sample data** to **estimate the regression equation**. Using only sample data, we can't find the **exact values** of the population parameters  $\beta_0$  and  $\beta_1$ , but we can use the sample data to estimate them with  $b_0$  and  $b_1$ .

# Requirements

1. The sample of **paired  $(x, y)$  data** is a *random sample* of **quantitative data**.
2. **Visual examination** of the **scatterplot shows** that the **points** approximate **a straight-line pattern**.
3. Any **outliers** must be **removed** if they are known to be errors. Consider the effects of any outliers that are not known errors.

# Requirements

**Note:** Requirements 2 and 3 above are simplified attempts at checking these formal requirements for regression analysis:

- ☐ For each **fixed value of  $x$** , the **corresponding values of  $y$**  have a distribution that is **bell-shaped**.
- ☐ For the **different fixed values of  $x$** , the distributions of the corresponding  **$y$ -values all have the same variance**.
- ☐ For the different fixed values of  $x$ , the distributions of the corresponding  $y$ -values have **means that lie along the same straight line**.
- ☐ The  $y$  values are independent.

# Requirements

- ❑ Results are **not seriously affected** if departures from **normal distributions** and equal variances are not too extreme.

# Definitions

Given a collection of paired sample data, the **regression equation**

$$\hat{y} = b_0 + b_1x$$

algebraically describes the relationship **between the two variables**. The graph of the **regression equation** is called the **regression line** (or *line of best fit*, or *least-squares line*).



# Notation for Regression Equation

	Population Parameter	Sample Statistic
y-intercept of regression equation	$\beta_0$	$b_0$
Slope of regression equation	$\beta_1$	$b_1$
Equation of the regression line	$Y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Finding the slope  $b_1$  and y-intercept  $b_0$  in the regression equation  $\hat{y} = b_0 + b_1 x$

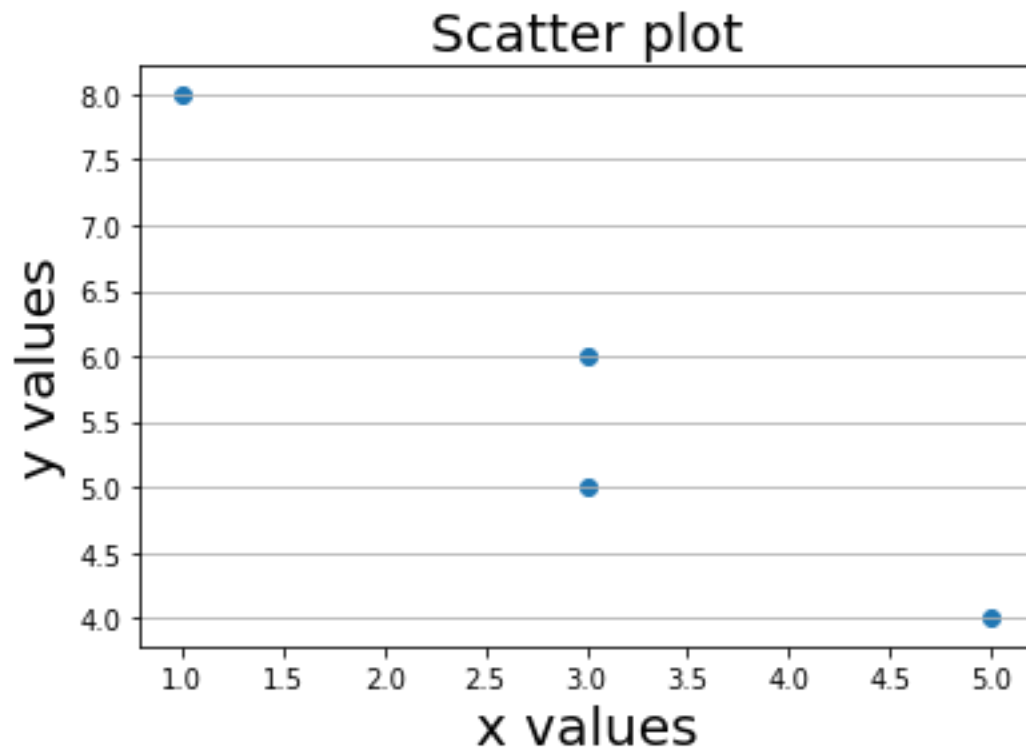
<b>Slope</b>	$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$
<b>y-intercept:</b>	$b_0 = \bar{y} - b_1\bar{x}$ <p>or</p> $b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$

## Example Finding the Regression Equation

Use the given sample data to find the regression equation.

<b>x</b>	<b>3</b>	<b>1</b>	<b>3</b>	<b>5</b>
<b>y</b>	<b>5</b>	<b>8</b>	<b>6</b>	<b>4</b>

**REQUIREMENT** The data are a simple random sample. The accompanying Python-generated scatterplot shows a pattern of points that does appear to be a straight-line pattern. There are no outliers. We can proceed to find the slope and intercept of the regression line.



$x$	$y$	$xy$	$x^2$	$y^2$
3	5	15	9	25
1	8	8	1	64
3	6	18	9	36
5	4	20	25	16
$\sum x = 12$	$\sum y = 23$	$\sum xy = 61$	$\sum x^2 = 44$	$\sum y^2 = 141$

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{4(61) - (12)(23)}{4(44) - (12)^2} = \frac{-32}{32} = -1$$

$$\bar{x} = \frac{12}{4} = 3$$

$$\bar{y} = \frac{23}{4} = 5.75$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = 5.75 - (-1)(3)$$

$$b_0 = 8.75$$

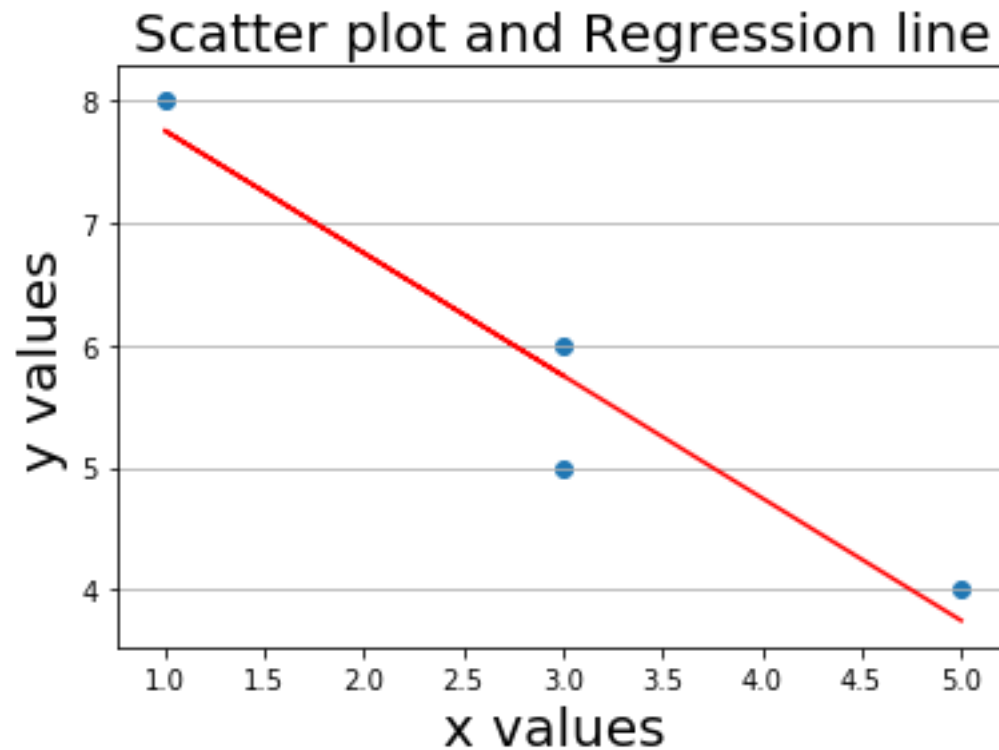
- Knowing the slope  $b_1$  and  $y$ -intercept  $b_0$ , we can now express the **estimated equation** of the **regression line** as

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 8.75 - 1x$$

- We should realize that this equation is an *estimate* of the **true regression equation**  $Y = \beta_0 + \beta_1 x$ . This estimate is based on one **particular set of sample data**, but another sample drawn from the same population would probably lead to a slightly different equation.

# Scatter plot and Regression line





# Using the Regression Equation for Predictions

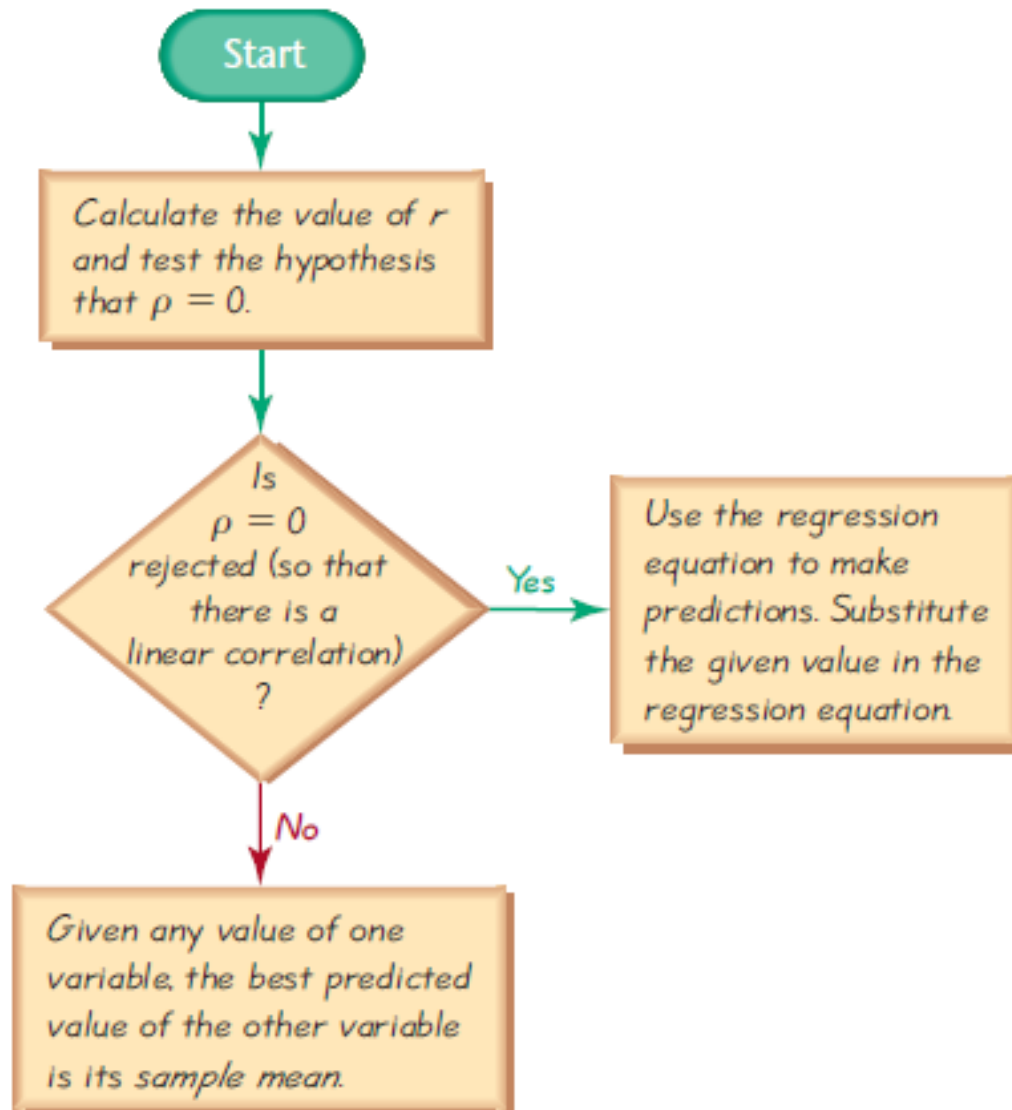
- ❑ Regression equations are often useful for *predicting* the **value of one variable**, given **some particular value of the other variable**.
- ❑ If the **regression line fits** the data quite well, then it makes sense to use its **equation for predictions**, provided that we don't go beyond the scope of the available values.

# Using the Regression Equation for Predictions

In **predicting a value of  $y$  based on some given value of  $x$**  . . . .

1. If there is ***not* a linear correlation**, the best predicted  **$y$ -value** is  $\bar{y}$ .
2. If there is a **linear correlation**, the **best predicted  $y$ -value** is found by **substituting the  $x$ -value** into the **regression equation**.

# Procedure for Predicting



# Guidelines for Using the Regression Equation

1. If there is **no linear correlation**, don't use the **regression equation** to make predictions.
2. When using the **regression equation for predictions**, stay within the **scope of the available sample data**. If you find a regression equation that **relates women's heights** and **shoe sizes**, it's absurd to predict the shoe size of a woman who is **10 ft tall**.

# Guidelines for Using the Regression Equation

3. A regression equation based on old data is not necessarily valid now. The regression equation relating used-car prices and ages of cars is no longer usable if it's based on data from the 1990s.

4. Don't make predictions about a population that is different from the population from which the sample data were drawn. If we collect sample data from men and develop a regression equation relating age and TV remote-control usage, the results don't necessarily apply to women. If we use state averages to develop a regression equation relating SAT math scores and SAT verbal scores, the results don't necessarily apply to individuals.