

Advanced Statistics

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

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

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**) **\hat{y}** 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 β_0 and β_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 β_0 and β_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	β_0	b_0
Slope of regression equation	β_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 $\hat{y} = b_0 + b_1 x$

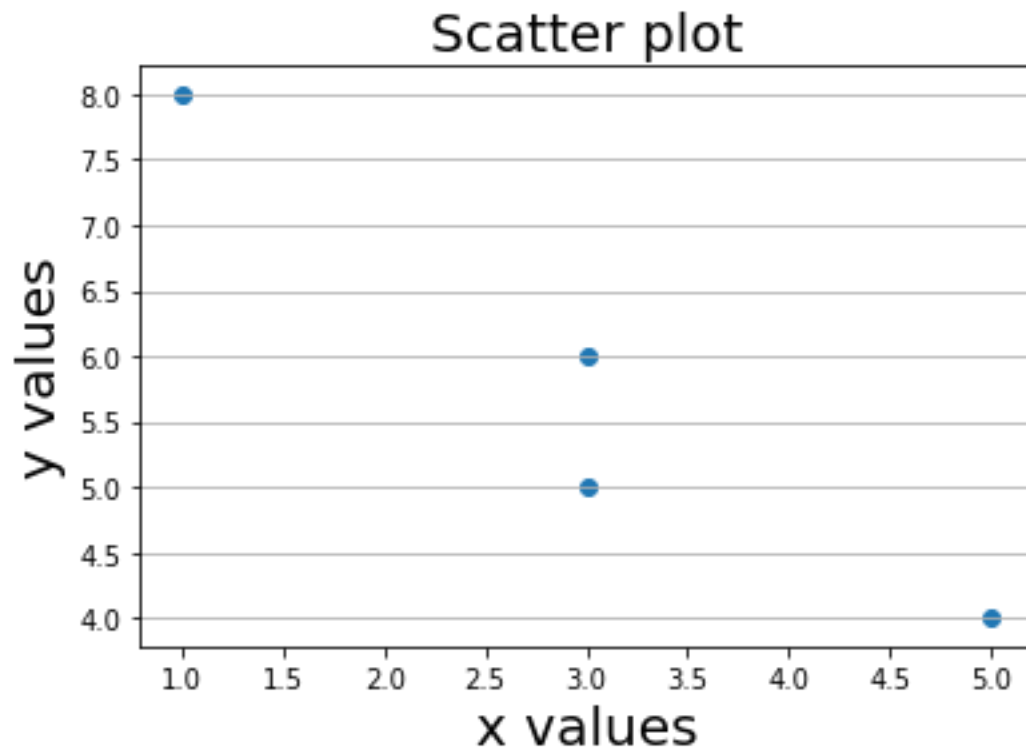
Slope	$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$
y-intercept:	$b_0 = \bar{y} - b_1\bar{x}$ <p>or</p> $b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$

Example Finding the Regression Equation

Use the given sample data to find the regression equation.

x	3	1	3	5
y	5	8	6	4

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
$\sum x = 12$	$\sum y = 23$	$\sum xy = 61$	$\sum x^2 = 44$	$\sum y^2 = 141$

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

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

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

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

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

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

$$b_0 = 8.75$$

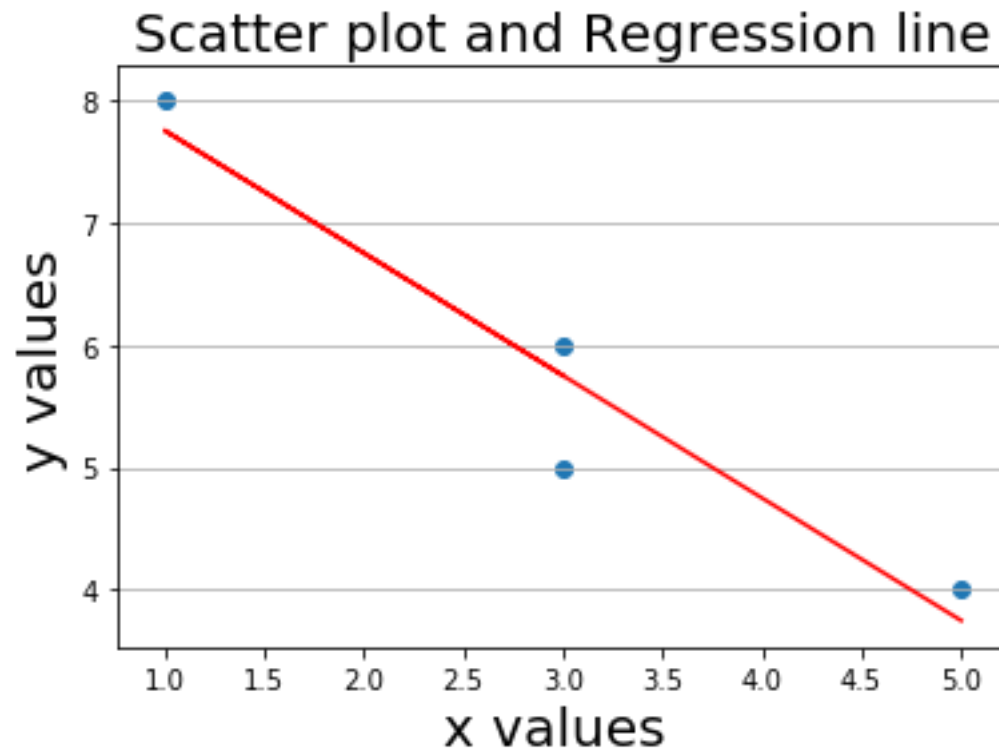
- Knowing the slope b_1 and y -intercept b_0 , we can now express the **estimated equation** of the **regression line** as

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

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

- 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

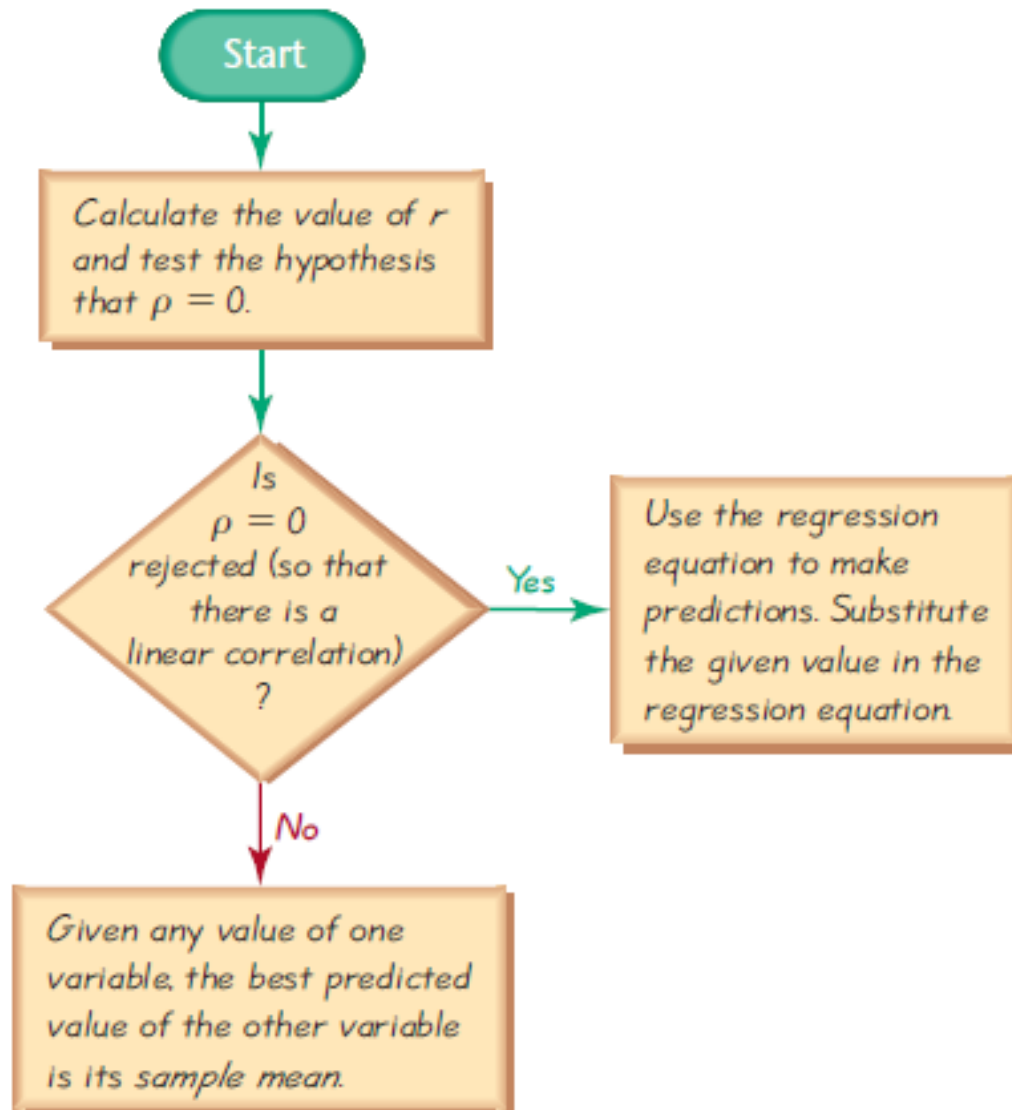
- ❑ Regression equations are often useful for *predicting* the **value of one variable**, given **some particular value of the other variable**.
- ❑ If the **regression line fits** the data quite well, then it makes sense to use its **equation for predictions**, provided that we don't go beyond the scope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