Lecture No. 10

Introduction To Statistics, Statistics And Probability

Dr. Shabbir Ahmad

Assistant Professor,
Department of Mathematics,
COMSATS University
Islamabad, Wah Campus

Simple Linear Regression

Method of Lest Squares, Standard Error of Estimate, Coefficient of Determination

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

In this lecture

- Introduction to Regression Analysis
- Method Least Squares Method and Estimation Process
- Interpretation of the Slope and Intercept
- Predictions through Regression Model
- Interpolation vs. Extrapolation
- Standard Error of Estimate
- Measures of Variation under Regression Model
- Coefficient of Determination, r2

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Introduction to Regression Analysis

Dependent variable Y

The variable we wish to predict or explain.

Independent variable X

The variable used to explain the dependent variable.

- **Regression analysis** is used to:

  - Explain the impact of changes in an independent variable on the dependent variable.
  - Predict the value of a dependent variable based on the value of at least one independent variable.

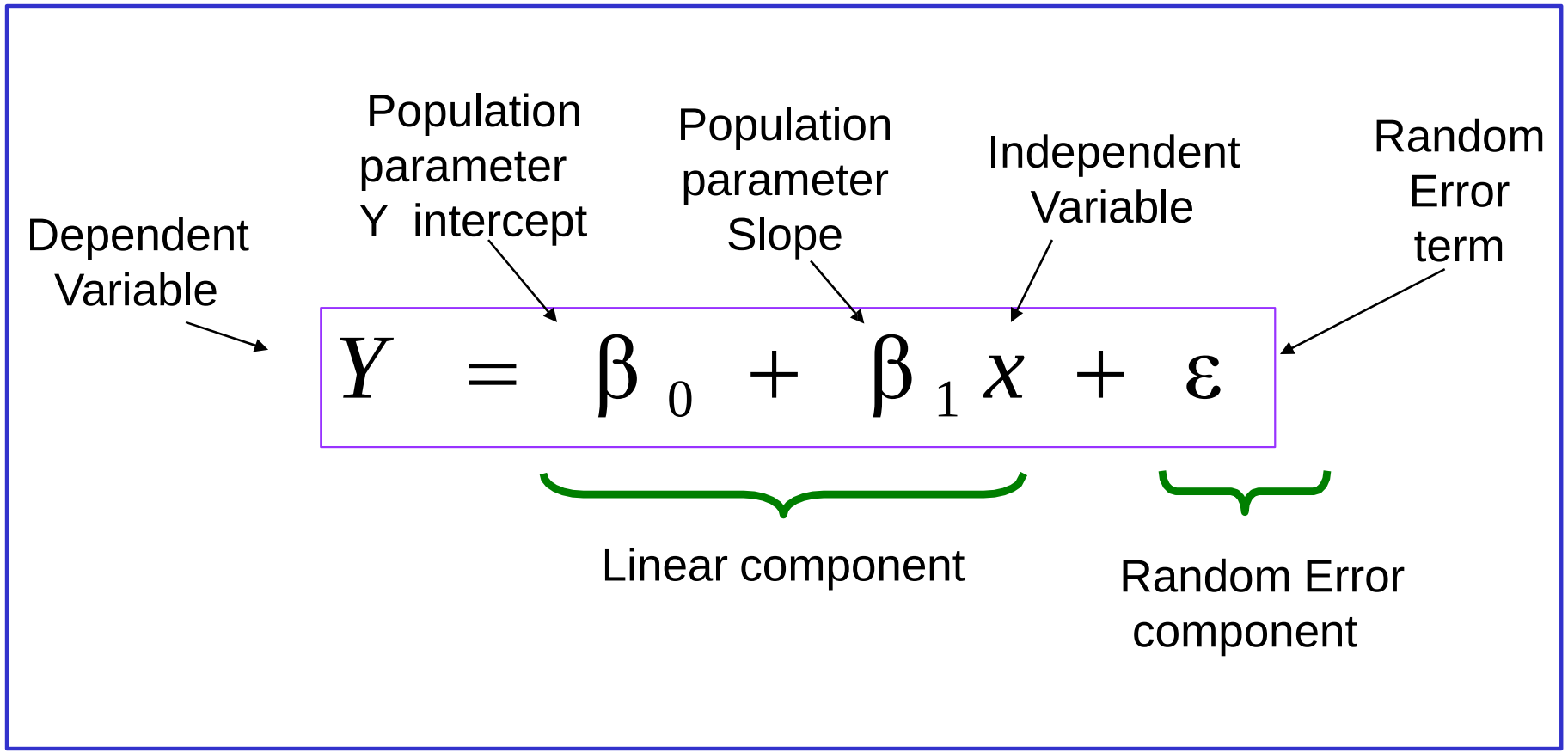
Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Simple Linear Regression Model

- Study the relationship between the dependent Y and one independent variable, X.
- The equation that describes how Y is related to X and an error term is called the **regression model**.
- Relationship between X and Y is described by a linear function.

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Simple Linear Regression Model



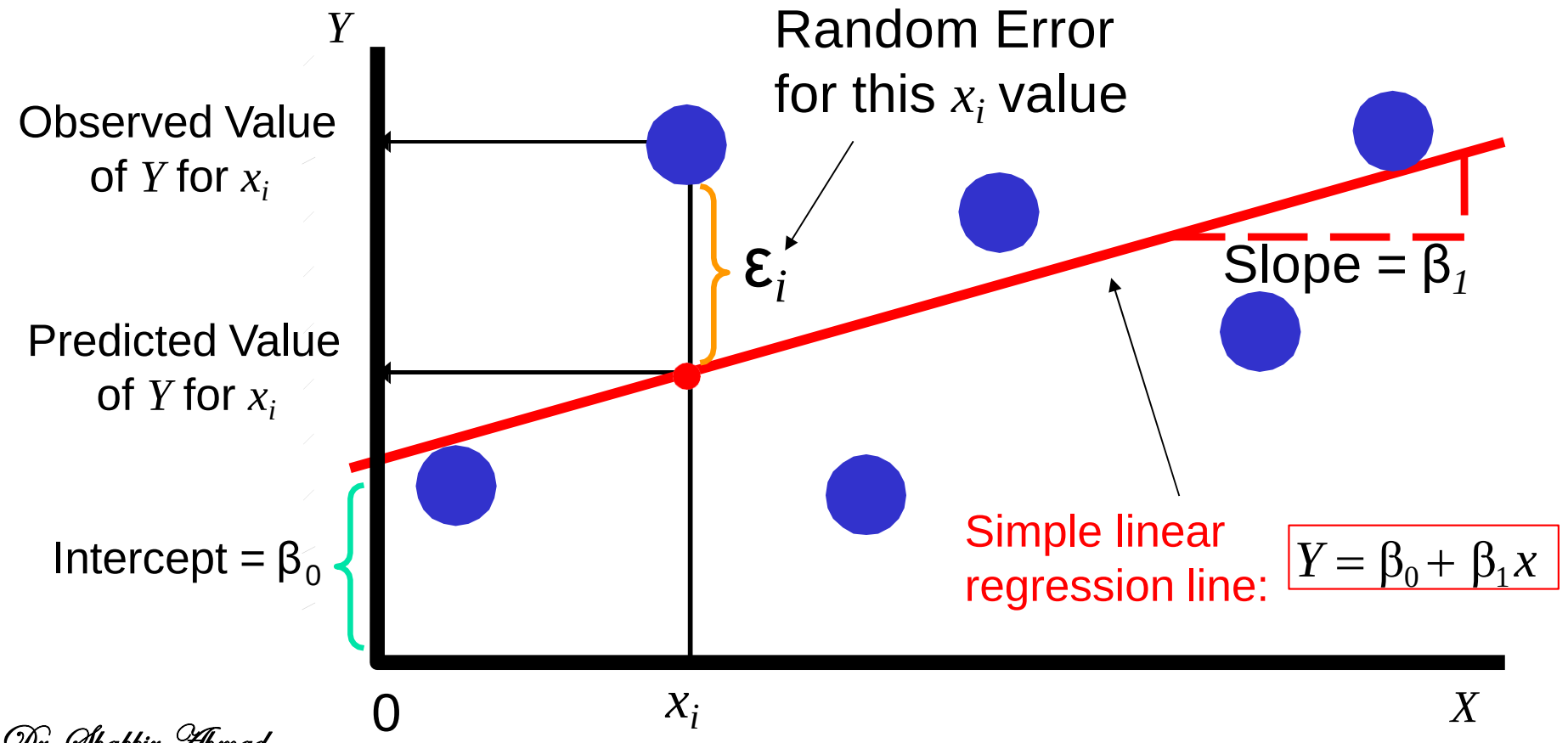
Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Simple Linear Regression Model

(continued)

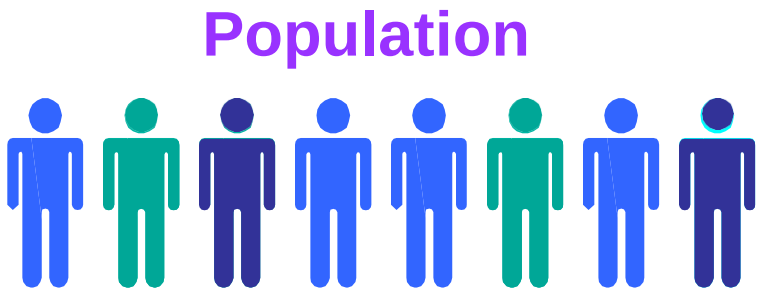
$$\text{Model}: Y = \beta_0 + \beta_1 x + \varepsilon$$

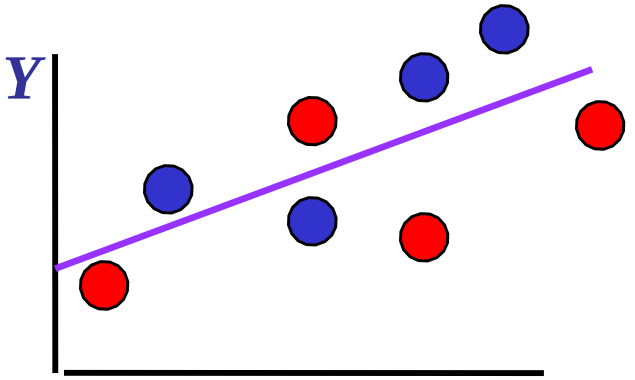
We consider all the data in a population.



Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

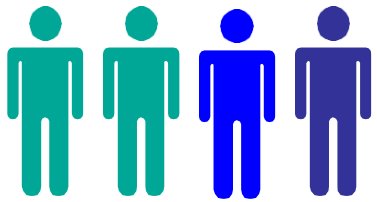
Estimation Process

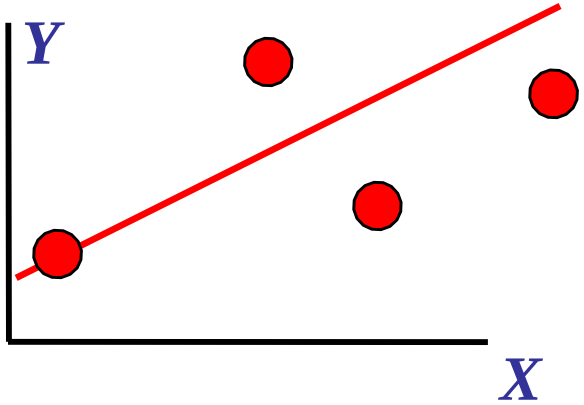




Simple linear regression equation: $y = \beta_0 + \beta_1 x$

Select a Random sample





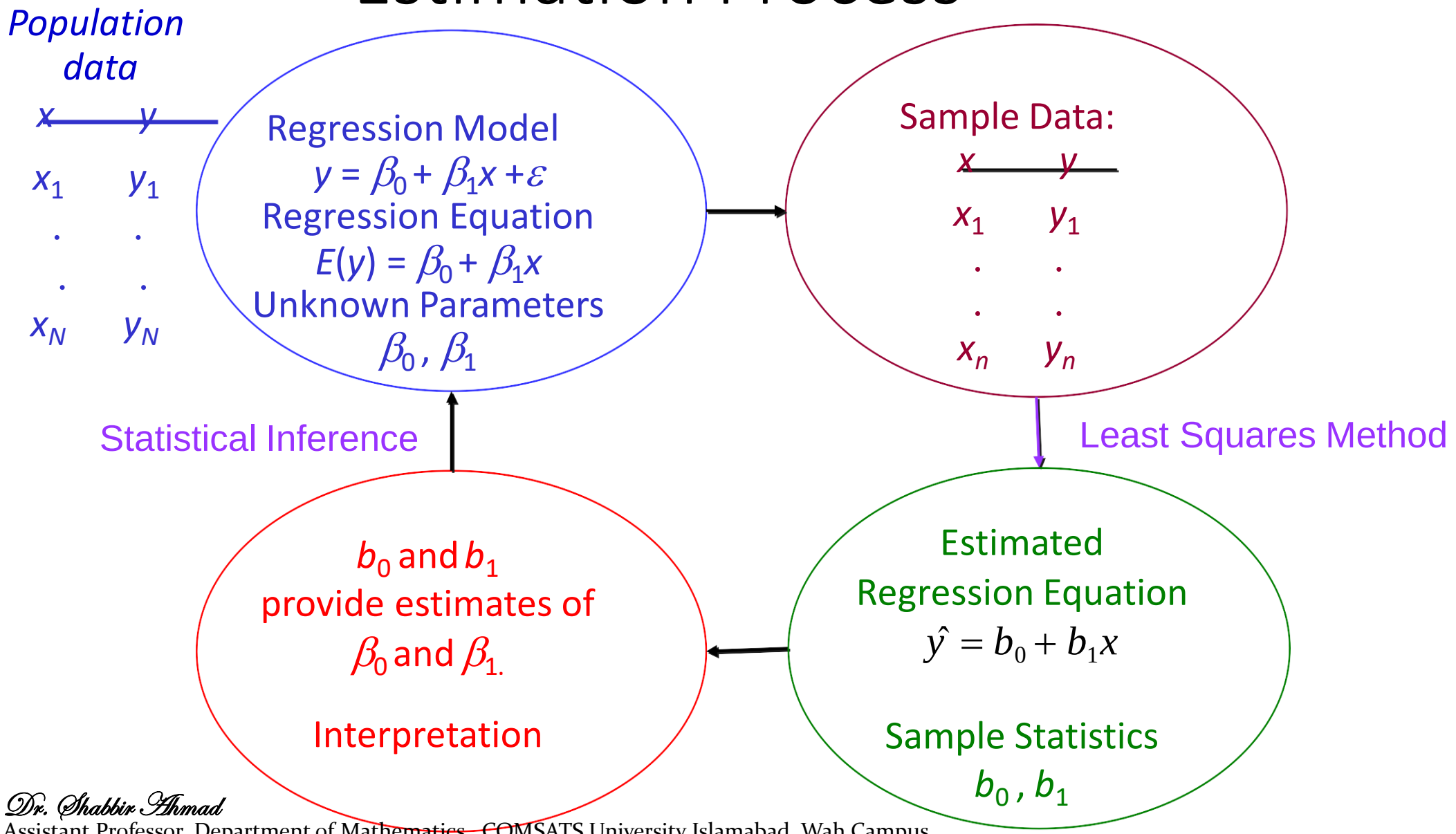
Estimated simple linear regression equation:

$\hat{y}_i = b_0 + b_1 x_i$

The estimated simple linear regression equation provides an estimate of the simple linear regression equation.

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Estimation Process



Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Estimated Simple Linear Regression Equation (Prediction Line)

The estimated simple linear regression equation provides an **estimate** of the simple linear regression equation.

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{y}_i = b_0 + b_1 x_i$$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Least Squares Method

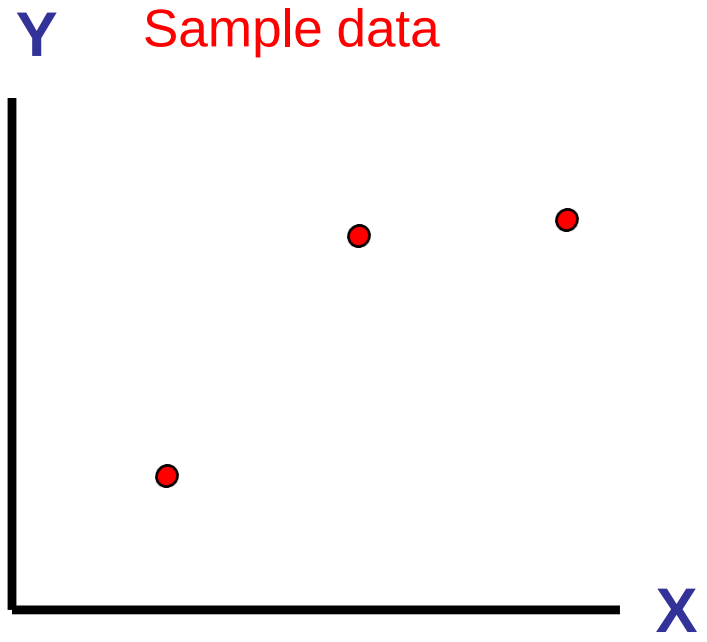
- y_i : observed value of the dependent variable for the ith observations.
- $\hat{y}_i$: estimated value of the dependent variable for the ith observations. (Fitted value)
- Find the values of b_0 and b_1 that minimize the sum of the squared differences between Y and $\hat{Y}$.

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Least Squares Method

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$



How can we find the estimated simple linear regression equation?

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Least Squares Method

Find b_0 and b_1 such that they min $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$

- Estimated slope: $b_1 = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$ Calculus
- Estimated intercept: $b_0 = \dfrac{\sum y}{n} - b_1 \dfrac{\sum x}{n}$ Results!
- The above results minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, so that it will be at it's minimum value.
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is not necessary = 0, but it will be at minimum.
- Note that : $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0.$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Simple Linear Regression
Example

Example: A real estate agent in US wishes to examine the relationship between the selling price of a home and its size (measured in square feet).

- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in US$1,000.
  - Independent variable (X) = house size

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

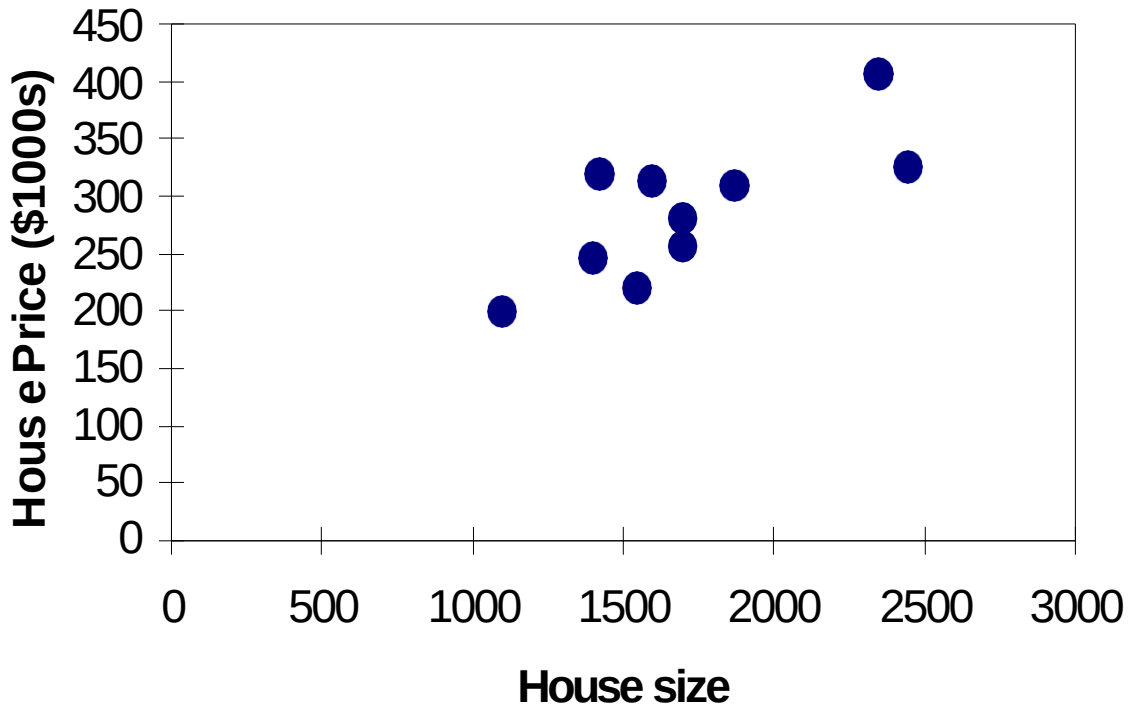
Sample Data for House Price (in US)

House Price (in US$1,000) ($y$)	**House size** (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Graphical Presentation

Construct a scatter plot of “house price” vs “house size”.



Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Estimated coefficient and regression equation.

Find the estimated regression equation.

$$b_1 = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{(10)(5{,}085{,}975) - (17{,}150)(2{,}865)}{(10)(30{,}983{,}750) - (17{,}150)^2} = 0.11$$

$$b_0 = \frac{\sum y}{n} - b_1\frac{\sum x}{n} = \frac{2{,}865}{10} - (0.11)\frac{17{,}150}{10} = 98.25$$

The regression equation is:

$$\widehat{\text{house price}} = 98.25 + 0.11(\text{house size})$$

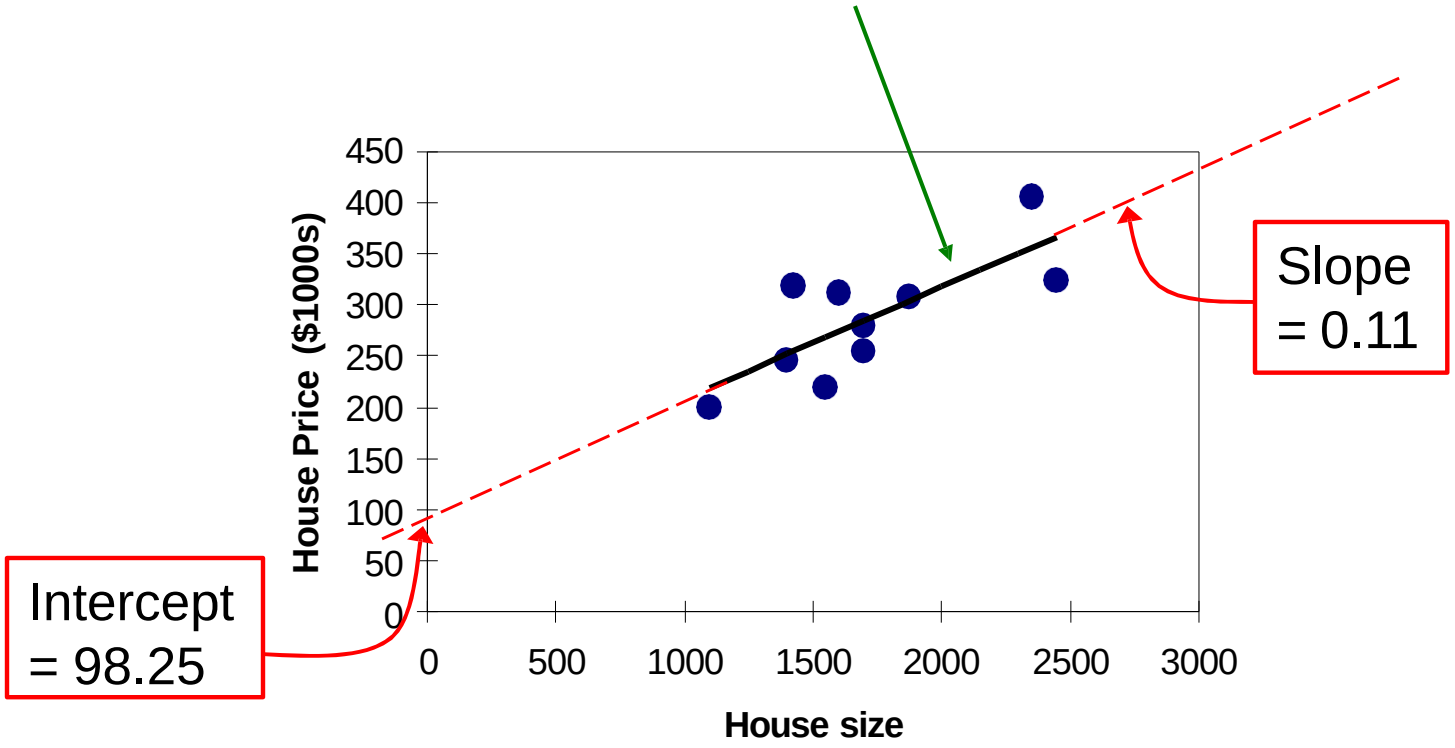
$$\hat{y} = 98.25 + 0.11x$$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Graphical Presentation

Put the data and the estimated regression line in a scatter plot.

$$\widehat{\text{house price}} = 98.25 + 0.11(\text{house size})$$



Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Interpretation of Intercept, b_0

$$\widehat{\text{house price}} = 98.25 + 0.11(\text{house size})$$

Interpret the estimated intercept.

- b_0 is the estimated average value of y when the value of x is zero (if x = 0 is in the range of observed x values)
  - Here, no house had 0 square feet, so b_0 = 98.25 just indicates that, for houses within the range of sizes observed, $98,250 is the portion of the house price not explained by house size.

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.25 + 0.11(\text{house size})$$

Interpret the estimated slope.

- b_1 measures the estimated change in the average value of y as a result of a one-unit change in x.
  - Here, $b_1 = 0.11$ tells us that the average value of a house increases by 0.11($1000) = $110, on average, for each additional one square foot of size.

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Predictions through Regression Analysis

Predict the price for a house with 1,400 square feet.

$$\hat{y} = 98.25 + 0.11(\text{housesize})$$
$$= 98.25 + 0.11(1{,}400)$$
$$= 252.25 \ (\text{in US\$1,000})$$

Find the prediction error.

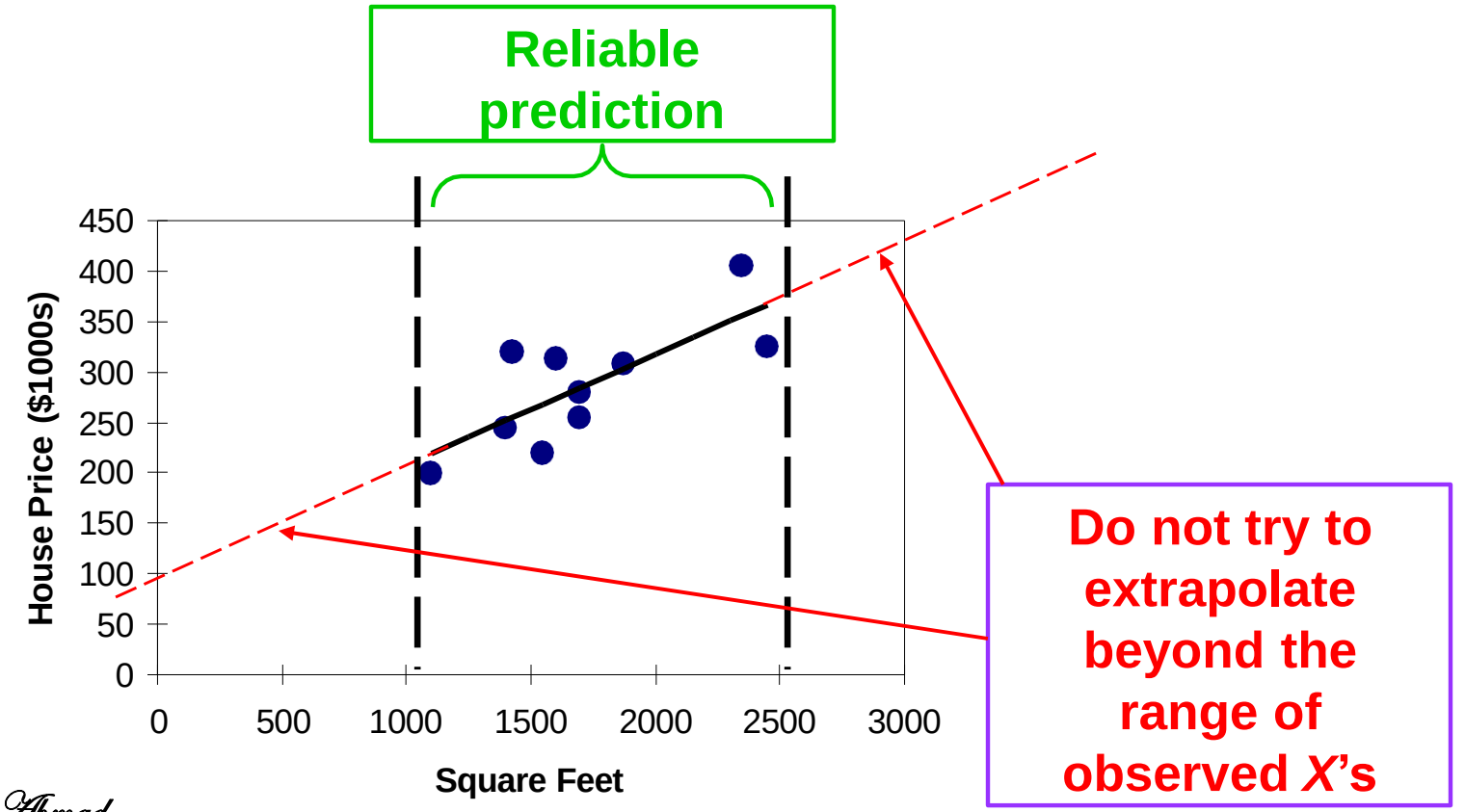
$$y_i - \hat{y}_i = 245 - 252.25 = -7.25 \ (\text{in US\$1,000})$$

Predict the average price for houses with 1,400 square feet.

$$\hat{y} = 98.25 + 0.11(1{,}400) = 252.25 \ (\text{in US\$1,00 0})$$

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Standard Error of Estimate

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{13{,}667.3}{10-2}} = 41.3$$

In our example, we have $S_{YX} = 41.3$ (in US$1,000)

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 13{,}667.3$$

House Price (in US$1,000) ($y$)	House Size (x)	$\hat{y} = 98.25 + 0.11x$	$(y_i - \hat{y}_i)^2$
245	1400	252.3	52.6
312	1600	274.3	1425.1
279	1700	285.3	39.1
308	1875	304.5	12.3
199	1100	219.3	410.1
219	1550	268.8	2475.1
405	2350	356.8	2328.1
324	2450	367.8	1914.1
319	1425	255.0	4096.0
255	1700	285.3	915.1

Computing Standard Error of Estimate

Method A

$$S_{YX} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}} = \sqrt{\frac{853423 - (98.25)(2865) - (0.11)(5085975)}{10-2}} \approx 41.3$$

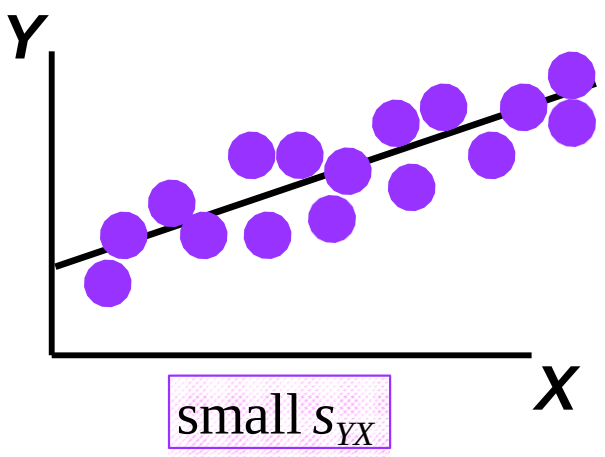
Method B

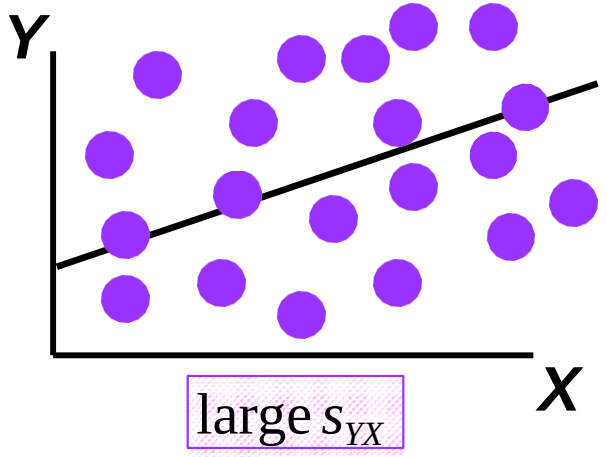
1. Enter all the data into your calculator.
2. Compute sample variance, s^2.
3. SST = (n – 1)(sample variance)
   = (10 – 1)(3,622.278)
   = 32,600.5
4. SSE = SST(1 – r^2) = (32,600.5)(1 – 0.5808) = 13,666.13
5. $S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{13{,}666.13}{10-2}} = 41.3$ (in US$1,000)

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line.





The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data.

i.e., S_{YX} = \$41,300 is moderately small relative to house prices in the \$200,000 - \$400,000 range

Measures of Variation under Regression Model

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (y - \bar{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Predicted value of y for the given x_i value

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:29:57 AM

Measures of Variation

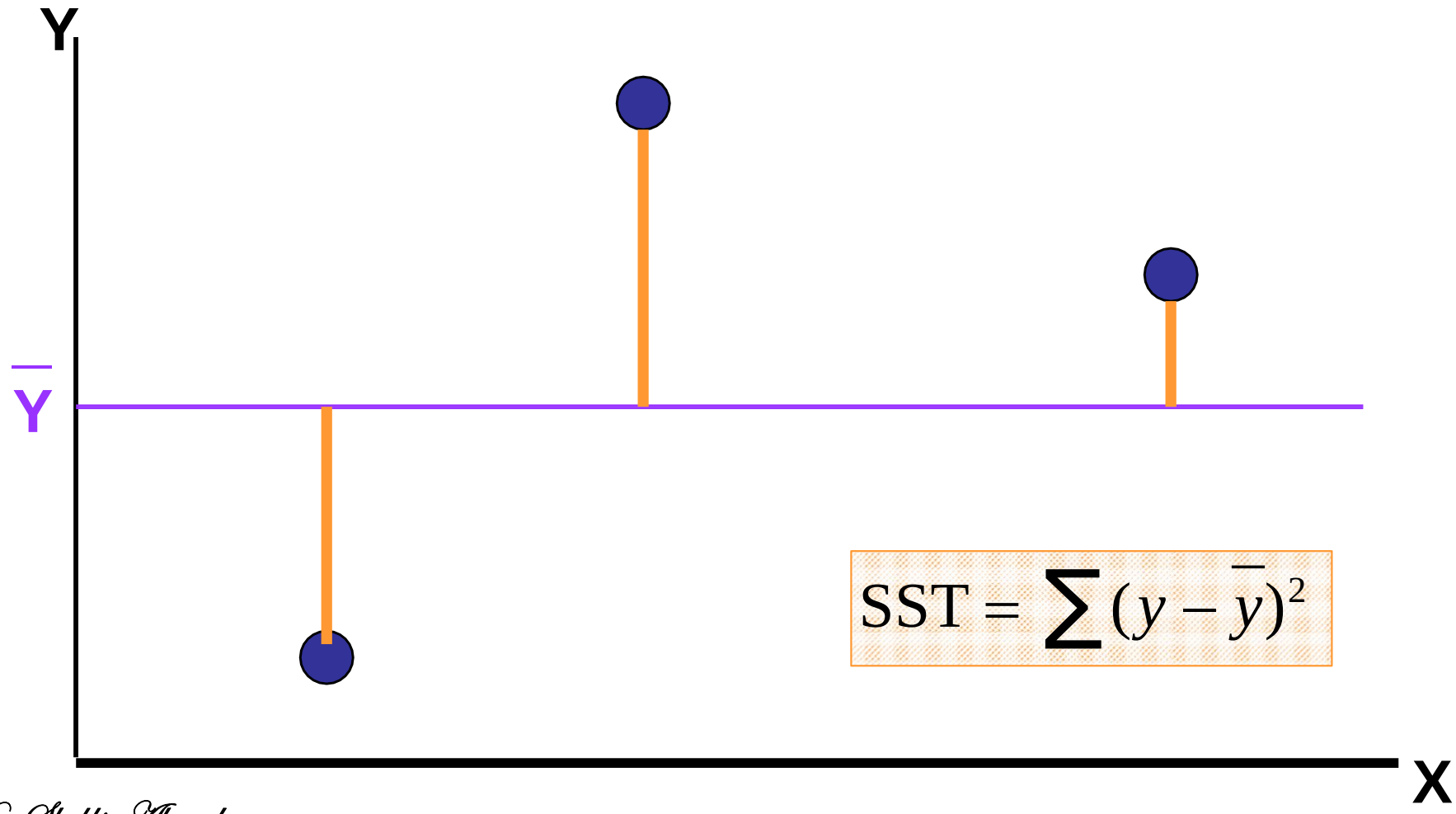
(continued)

- SST = total sum of squares
  - Measures the variation of the y values around their mean y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between x and y

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

SST = total sum of squares

(continued)

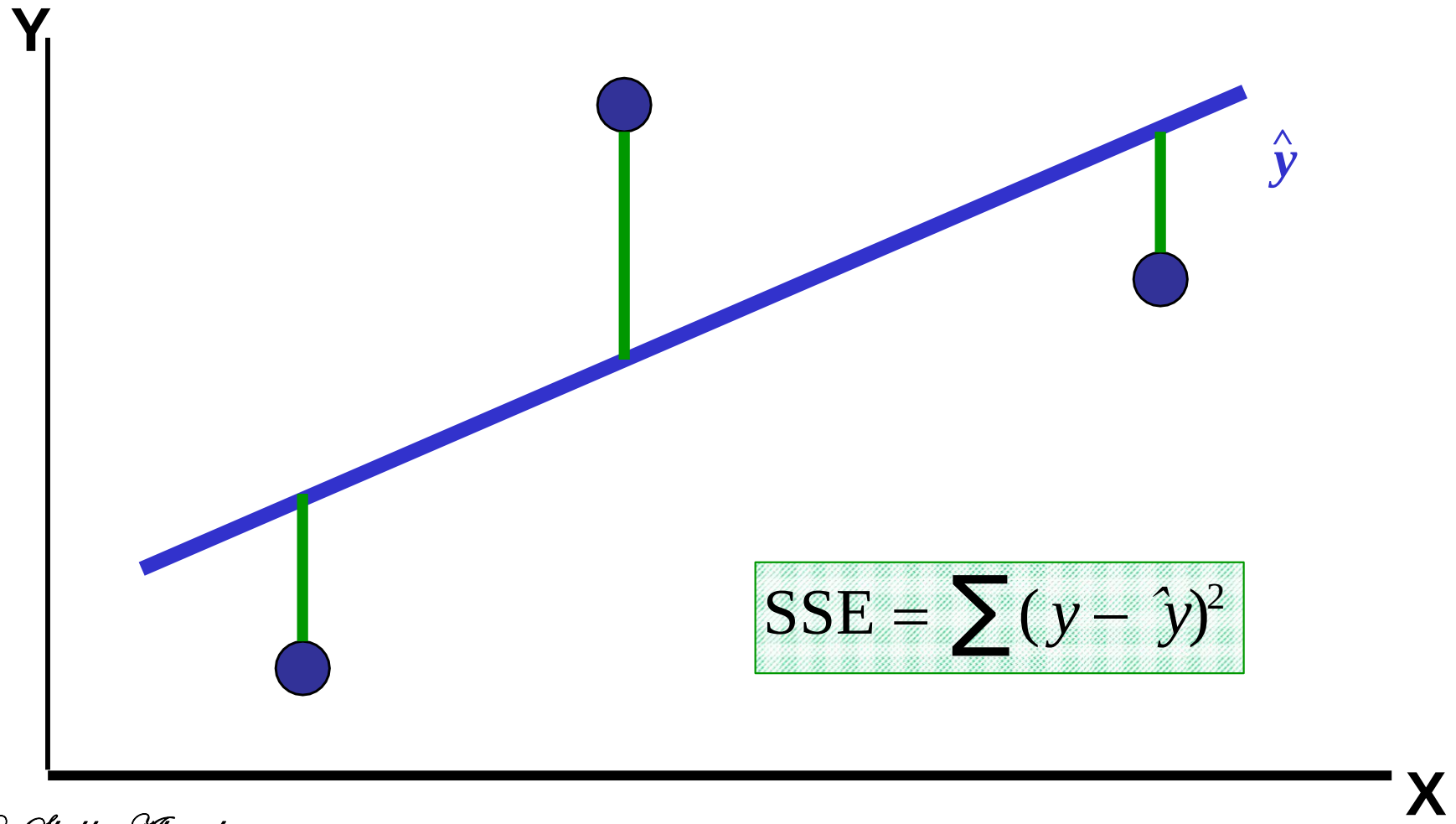


$$SST = \sum (y - \bar{y})^2$$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

SSE = error sum of squares

(continued)

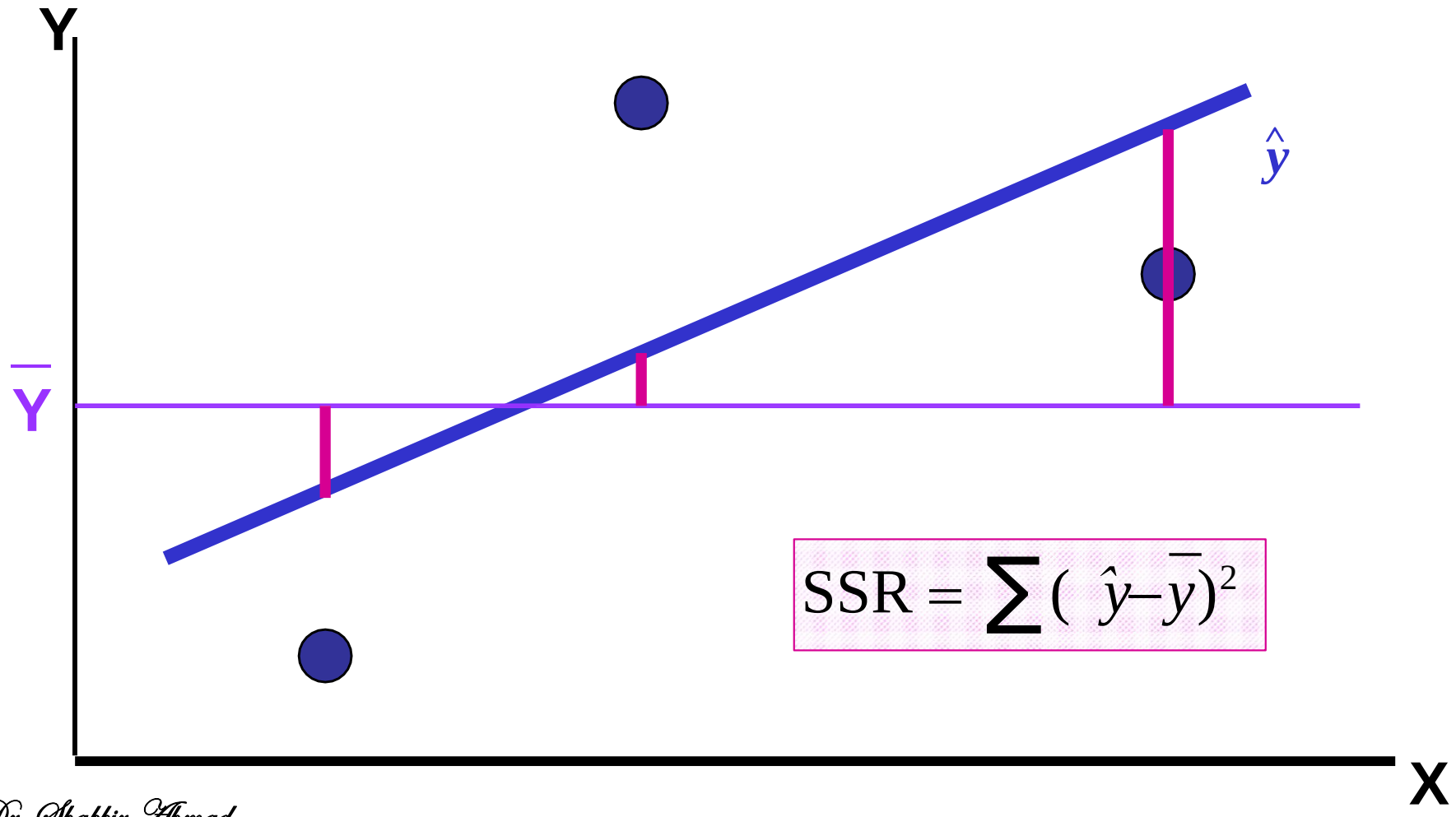


$$SSE = \sum (y - \hat{y})^2$$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

SSR = regression sum of squares

(continued)



$$SSR = \sum (\hat{y} - \bar{y})^2$$

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Coefficient of Determination, r^2

- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.
- The coefficient of determination is denoted as r^2.

$$r^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

- Note: $0 \le r^2 \le 1$
-
- Note: $r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$

$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$

where: b_1 = the slope of the estimated regression equation.

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

Coefficient of Determination, r^2

$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

From our house data, find the coefficient of determination.

- **Method A**: By formula

$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{18934.9348}{32600.5000} = 0.581$$

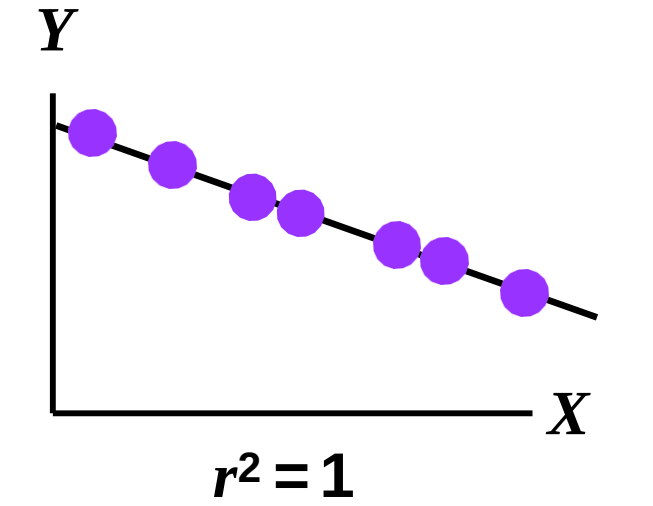
- **Method B**: Squaring the coefficient of correlation.

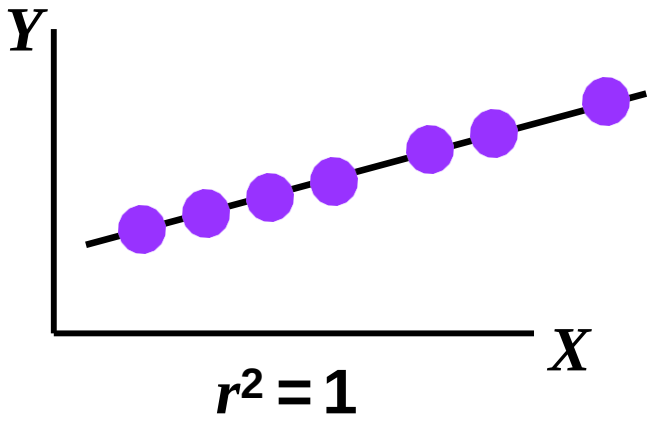
$$r^2 = (0.7621)^2 = 0.581$$

Interpretation: 58.1% of the variation in house prices is explained by variation in house size.

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

r^2 Values





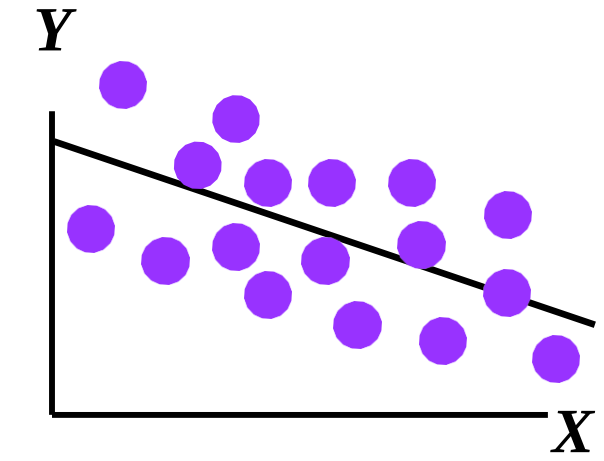
$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

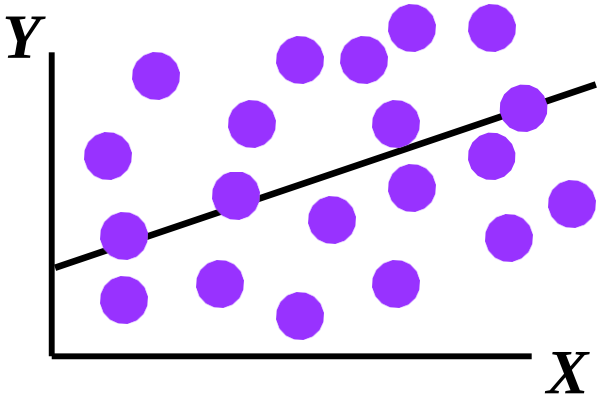
Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

r^2 Values



$0 < r^2 < 1$

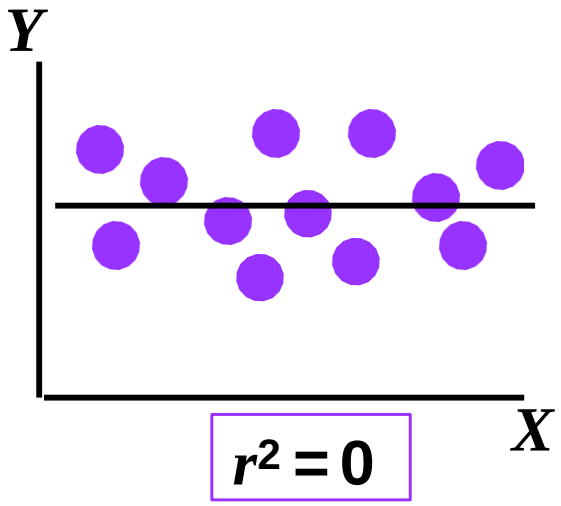
Some but not all of the variation in Y is explained by variation in X



Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

r^2 Values

Examples of Approximate



$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Dr. Shabbir Ahmad
Assistant Professor, Department of Mathematics, COMSATS University Islamabad, Wah Campus
Cell # 0323-5332733, 0332-5332733. Date: 10/14/2021 9:23:55 AM

ANY QUESTION