STATISTICAL LABORATORY

Applied Mathematics for Economics and Management
1st Year/1st Semester
2025/2026

CONTACT

Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

PROGRAM

1. Fundamental Concepts of Statistics

2. Exploratory Data Analysis

3. Organizing and Summarizing Data

4. Association and Relationships Between Variables

5. Index Numbers

6. Time Series Analysis

LECTURE 6: INTRODUCTION - CORRELATION AND REGRESSION

CORRELATION AND REGRESSION

Purpose	ScatterPlot	Correlation	Regression
• Analyse the relationship between **two quantitative variables** X and Y.	• **First step** in studying the relationship between variables. • Plots each pair (X, Y) as a point in a coordinate system. • Allows visual detection of: ➢ **Direction** (positive or negative) ➢ **Form** (linear or non-linear) ➢ **Strength** of the relationship	• Measures the **degree and direction** of a linear relationship. • Expressed by the **correlation coefficient (r)** → ranges from **–1 to +1**. ➢ $r > 0$: variables increase together. ➢ $r < 0$: one increases as the other decreases.	• Describes or predicts Y based on X. • **Simple Linear Regression Model:** $Y = \beta_0 + \beta_1 X + \varepsilon$ • Parameters estimated by the **least squares method**

Key Idea:

- **The scatterplot** shows the relationship.
- **Correlation** quantifies the strength of association.
- **Regression** explains and predicts the relationship.

LECTURE 6: COVARIANCE AND CORRELATION

MEASURES OF RELATIONSHIPS BETWEEN VARIABLES

Two measures of the relationship between variable are

- Covariance
  - a measure of the direction of a linear relationship between two variables
- Correlation Coefficient
  - a measure of both the direction and the strength of a linear relationship between two variables

- Only concerned with the strength of the relationship
- No causal effect is implied

Newbold et al (2013)

COVARIANCE

Definition:

Consider two variables X and Y, for which we have n paired observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

The sample covariance between X and Y is given by

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

or equivalently,

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

Silvestre (2007)

INTERPRETING COVARIANCE

- **Covariance** between two variables:

$\mathrm{Cov}(x, y) > 0 \rightarrow$ *x* and *y* tend to move in the same direction

$\mathrm{Cov}(x, y) < 0 \rightarrow$ *x* and *y* tend to move in opposite directions

$\mathrm{Cov}(x, y) = 0 \rightarrow$ *x* and *y* are independent

Newbold et al (2013)

Note:

- The sign of the covariance indicates the direction of the relationship between the variables.

COEFFICIENT OF CORRELATION

Definition:

The **sample correlation coefficient** measures the **degree and direction of the linear relationship** between two variables X and Y.

It is defined as

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y}$$

or equivalently,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Silvestre (2007)

FEATURES OF CORRELATION COEFFICIENT

- Unit free
- Ranges between −1 and 1
- The closer to −1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Newbold et al (2013)

Properties:

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} > 0$: positive linear relationship
- $r_{xy} < 0$: negative linear relationship
- $r_{xy} = 0$: no linear relationship

INTERPRETING CORRELATION COEFFICIENT (SILVESTRE, 2007)

Value of r_{xy}	Interpretation
$r_{xy} = -1$	Perfect negative linear correlation (exact linear relationship)
$-1 < r_{xy} < 0$	Negative linear relationship (not perfect)
$r_{xy} = 0$	No linear correlation (absence of linear relationship)
$0 < r_{xy} < 1$	Positive linear relationship (not perfect)
$r_{xy} = 1$	Perfect positive linear correlation (exact linear relationship)

Silvestre (2007)

INTERPRETING CORRELATION COEFFICIENT

$r = -1$

- Perfect negative linear correlation

$-1 < r \leq -0{,}8$

- Strong negative linear correlation

$-0{,}8 < r \leq -0{,}5$

- Moderate negative linear correlation

$-0{,}5 < r < 0$

- Weak negative linear correlation

$r = 0$

- No linear correlation

$0 < r < 0{,}5$

- Weak positive linear correlation

$0{,}5 \leq r < 0{,}8$

- Moderate positive linear correlation

$0{,}8 \leq r < 1$

- Strong positive linear correlation

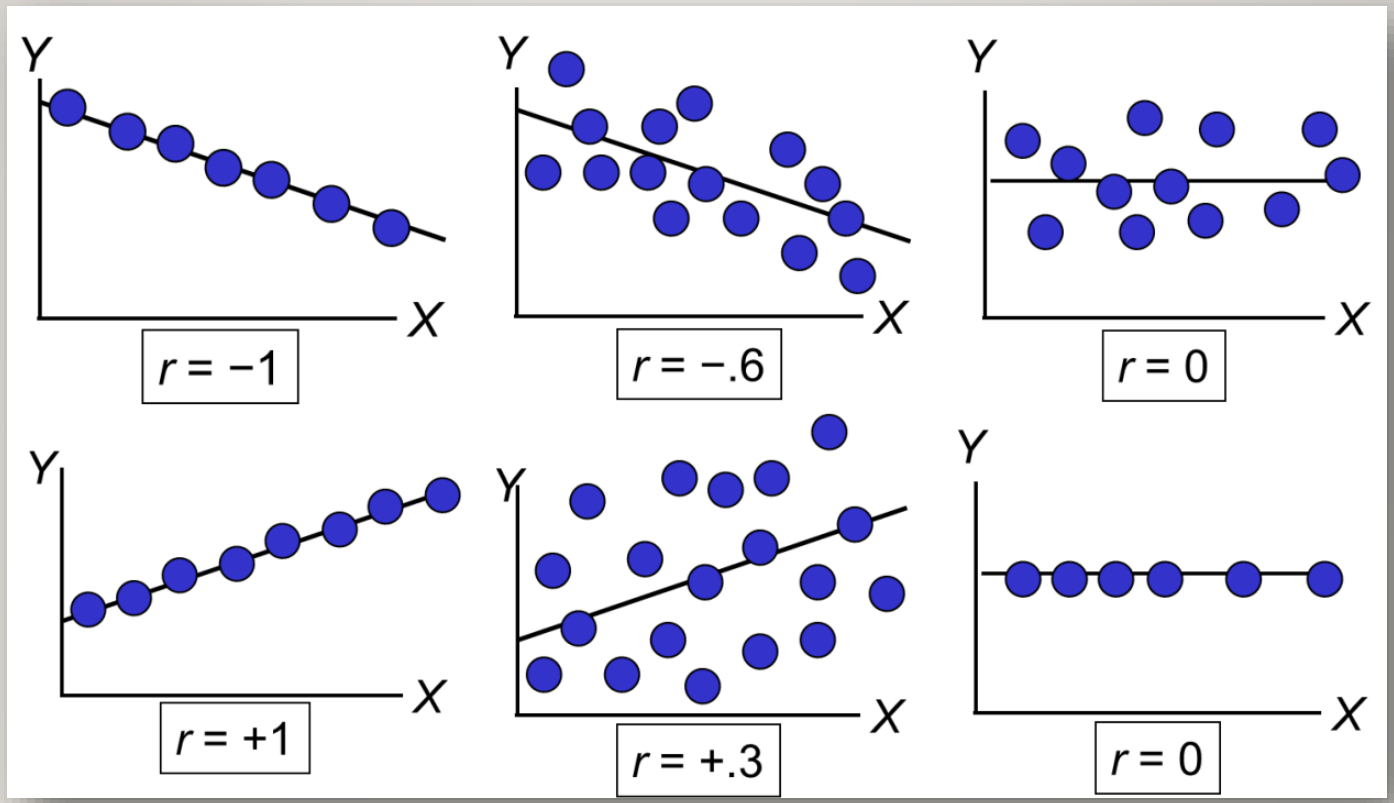
$r = 1$

- Perfect positive linear correlation

Note:

- The **sign** of r indicates the **direction** of the association.
- The **absolute value** |r| indicates the **strength** of the linear relationship.

SCATTERPLOTS OF DATA WITH VARIOUS CORRELATION COEFFICIENTS



Newbold et al (2013)

EXERCISE 2.55

2.55 A random sample for five exam scores produced the following (hours of study, grade) data values:

Hours Studied (x)	Test Grade (y)
3.5	88
2.4	76
4	92
5	85
1.1	60

a. Compute the covariance.
b. Compute the correlation coefficient

Newbold et al (2013)

EXERCISE 2.55: SOLUTION

Answer:

Data:

$$x = [3.5, 2.4, 4.0, 5.0, 1.1], \quad y = [88, 76, 92, 85, 60]$$

Step 1: Means

$$\bar{x} = 3.2, \quad \bar{y} = 80.2$$

Step 2: Covariance (Silvestre, 2007)

$$s_{xy} = \frac{1}{5}\sum x_i y_i - \bar{x}\bar{y} = 13.24$$

Step 3: Standard deviations

$$s_x = \sqrt{\frac{1}{5}\sum x_i^2 - \bar{x}^2} \approx 1.343, \quad s_y \approx 11.39$$

Step 4: Correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \approx 0.866$$

Summary:

- Covariance ≈ 13.24
- Correlation = r ≈ 0.866 (strong positive linear relationship, $0.8 \leq r < 1$)

STANDARDIZATION OF A VARIABLE

Let x be a variable with n observations $(x_1, x_2, \ldots, x_n)$.

The standardized values z_i are defined as:

$$z_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, 2, \ldots, n$$

where:

- $\bar{x}$ is the **mean of** x
- s_x is the **standard deviation of** x
- z_i is a **relative number**
- The standardized variable z has **mean 0** and **variance 1**

Silvestre (2007)

LECTURE 6: LEAST SQUARES REGRESSION LINE

LEAST SQUARES REGRESSION LINE

Let x and y be variables with n observations $(x_i, y_i), i = 1, 2, \ldots, n$, which appear to be correlated.

The regression line is estimated as:

$$\hat{y}_i = a + bx_i$$

The **Least Squares Method** is used to determine the coefficients a and b by **minimizing the estimation errors**:

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i), \quad i = 1, 2, \ldots, n$$

- e_i represents the **difference between the observed and predicted values**

Silvestre (2007)

LEAST SQUARES METHOD AND ESTIMATORS

The **Least Squares Method** minimizes the sum of squared errors:

$$\min_{a,b} Q(a,b) = \min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- The **slope estimator** b is:

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{s_{yx}}{s_x^2}$$

$$\Leftrightarrow \quad b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- The **intercept estimator** a is:

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

- The **regression line** is:

$$\hat{y} = a + bx$$

Silvestre (2007)

PROPERTIES OF THE LEAST SQUARES REGRESSION

1. The **mean of the estimation errors** is zero:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

2. The **regression line passes through the point of means**:

$$(\bar{x}, \bar{y})$$

3. The **mean of the predicted values** equals the **mean of the observed values**:

$$\overline{\hat{y}} = \bar{y}$$

4. The **estimation errors** are **uncorrelated with the variable** x:

$$\sum_{i=1}^{n} x_i e_i = 0$$

Silvestre (2007)

EXERCISE 11.19

11.19 A company sets different prices for a particular DVD system in eight different regions of the country. The accompanying table shows the numbers of units sold and the corresponding prices (in dollars).

Sales	420	380	350	400	440	380	450	420
Price	104	195	148	204	96	256	141	109

a. Graph these data, and estimate the linear regression of sales on price.
b. What effect would you expect a $50 increase in price to have on sales?

Newbold et al (2013)

EXERCISE 11.19 A): SOLUTION

✅ Answer:

Step 1: Means

$$\bar{x} \approx 156.63, \quad \bar{y} = 405$$

Step 2: Sums

$$\sum x_i y_i = 496{,}930, \quad \sum x_i^2 = 218{,}875$$

Step 3: Least Squares Estimators

$$b = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} \approx -0.454$$

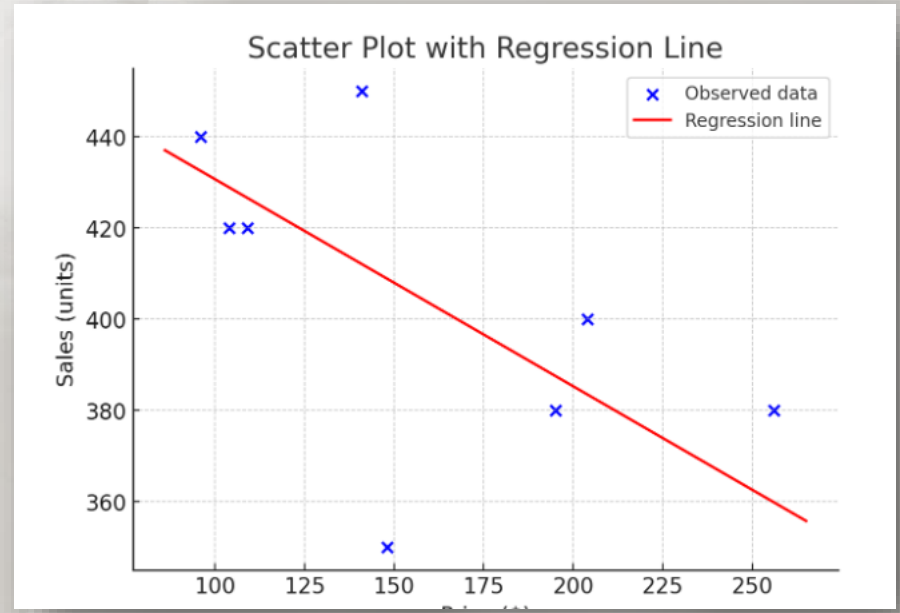
$$a = \bar{y} - b\bar{x} \approx 476.1$$

Regression line:

$$\hat{y} = 476.1 - 0.454x$$

EXERCISE 11.19 A): SOLUTION

✅ Answer:



Here's the **scatter plot of Sales vs. Price** with the **regression line** overlaid.

- **Blue points**: observed data
- **Red line**: regression line $\hat{y} = 476.1 - 0.454x$

EXERCISE 11.19 B): SOLUTION

✅ Answer:

b. Effect of $50 increase in price

$$\Delta y = b \cdot 50 \approx -23$$

Interpretation: Sales are expected to **decrease by about 23 units**.

ASSESSMENT OF FIT QUALITY: MSE AND RESIDUAL VARIANCE

Mean Squared Error (MSE):

Average of the squared prediction errors

$$MSE = \frac{1}{m}\sum_{k=1}^{m}(\hat{y}_{n+k} - y_{n+k})^2$$

Assuming m forecasts and corresponding observed values

$(\hat{y}_{n+k}, y_{n+k}),\ k = 1, 2, \ldots, m$

- m = number of **predictions** you are evaluating.
- y_{n+k} = **observed value** at step $n + k$.
- $\hat{y}_{n+k}$ = **predicted value** for step $n + k$.
- $k = 1, 2, \ldots, m$ is just an **index for each prediction** after the initial sample of size n.

MSE is used to assess the quality of out-of-sample or future predictions. The MSE is computed using data points outside the original dataset

Variance of Estimation Errors:

Measures the dispersion of the estimation (or residual) errors

$$s_e^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Silvestre (2007)

Interpretation:

- **Smaller values** of both **MSE** and **residual variance** indicate **better model fit**.
- Both measure the magnitude of prediction errors: the closer the predicted values are to the observed values, the smaller these quantities will be. Thus, large values suggest poor predictive performance or high variability in the errors.

ASSESSMENT OF FIT QUALITY: COEFFICIENT OF DETERMINATION

Coefficient of Determination (R^2)

Starts from the following decomposition:

$$\text{Total Variation} = \text{Residual Variation} + \text{Explained Variation}$$

By dividing by n, we obtain:

$$\frac{1}{n}\sum(y_i - \bar{y})^2 = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

Using variance notation:

$$s_y^2 = s_e^2 + \hat{s}_y^2$$

where:

- s_y^2 = total variance of y
- s_e^2 = variance of the residuals (errors)
- $\hat{s}_y^2$ = variance of the fitted values (explained by the model)

Dividing both sides by s_y^2 gives:

$$1 = \frac{s_e^2}{s_y^2} + \frac{\hat{s}_y^2}{s_y^2}$$

From this, the **Coefficient of Determination** is defined as:

$$R^2 = \frac{\hat{s}_y^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Interpretation:

- $0 \leq R^2 \leq 1$
- The closer R^2 is to 1, the better the model fits the data.

Silvestre (2007)

GENERAL DECOMPOSITION IN REGRESSION

In simple linear regression, the total variation in the dependent variable y can be decomposed as:

$$\text{Total Variation (SST)} = \text{Explained Variation (SSR)} + \text{Residual Variation (SSE)}$$

1. **Total Sum of Squares (SST):**

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Measures the total variability of the observed values around their mean.

2. **Regression Sum of Squares (SSR):**

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- Measures the variability explained by the regression model.

3. **Residual Sum of Squares (SSE):**

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

- Measures the variability that the model **fails to explain** (errors).

Relationship:

$$SST = SSR + SSE$$

Coefficient of Determination (R^2):

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Represents the proportion of the total variation in y explained by the model.
- Higher R^2 → better fit.

$$R^2 = \frac{\hat{s}_y^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Silvestre (2007)

ASSESSMENT OF FIT QUALITY

There is **no fixed cutoff** for MSE or residual variance (s_e^2) to define a "bad" fit, because their values depend on the scale of the data and context.

However, for the **coefficient of determination** (R^2), general guidelines can be used:

- $R^2 > 0.9$ → excellent fit
- $0.7 < R^2 \leq 0.9$ → good fit
- $0.5 < R^2 \leq 0.7$ → moderate fit
- $R^2 < 0.5$ → weak fit, the model explains little of the variance

EXERCISE 11.75

11.75 The following table shows, for eight vintages of select wine, purchases per buyer (y) and the wine buyer's rating in a year (x):

x	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
y	24	21	22	22	18	13	9	6

a. Estimate the regression of purchases per buyer on the buyer's rating.
b. Interpret the slope of the estimated regression line.
c. Find and interpret the coefficient of determination.

Newbold et al (2013)

EXERCISE 11.75 A): SOLUTION

✅ Answer:

The simple linear regression model is:

$$y = a + bx + \epsilon$$

where:

- b = slope (declive)
- a = intercept (ordenada na origem)

1. Compute the means:

$$\bar{x} = 2.6875, \quad \bar{y} = 16.875$$

2. Compute slope b:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \approx 8.39$$

3. Compute intercept a:

$$a = \bar{y} - b\bar{x} = 16.875 - 8.39 \cdot 2.6875 \approx -5.54$$

✅ **Estimated regression line:**

$$\hat{y} = -5.54 + 8.39x$$

EXERCISE 11.75 B): SOLUTION

✅ Answer:

Interpret the slope (b)

The slope $b \approx 8.39$ means:

For each **1-unit increase in the buyer's rating**, purchases per buyer **increase on average by about 8.39 bottles**.

Interpretation of the Intercept (a)

- The intercept $a \approx -5.54$ represents the **expected number of purchases per buyer** when the buyer's rating $x = 0$.
- In practice, since a rating of 0 may not exist in this context, the intercept is mainly a **mathematical reference point** for the regression line.

EXERCISE 11.75 C): SOLUTION

✅ Answer:

. Find and interpret the coefficient of determination (R^2)

1. Compute $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

- Total sum of squares:

$$SST = \sum (y_i - \bar{y})^2 = 50.766$$

- Regression sum of squares:

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = 42.689$$

- Residual sum of squares:

$$SSE = SST - SSR = 50.766 - 42.689 = 8.077$$

$$R^2 = \frac{SSR}{SST} = \frac{42.689}{50.766} \approx 0.841$$

Interpretation:

- About **84.1% of the variation in purchases** is explained by the buyer's rating.
- The model fits the data **quite well**.

RELATIONSHIP BETWEEN CORRELATION COEFFICIENT AND REGRESSION COEFFICIENT

Given:

Simple linear regression:

$$y = a + bx$$

Slope:

$$b = \frac{s_{xy}}{s_x^2} \quad \text{where } s_{xy} = \text{cov}(X, Y)$$

Pearson correlation:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Step 1: Express r_{xy} in terms of b

Substitute $s_{xy} = b s_x^2$ into $r_{xy} = s_{xy}/(s_x s_y)$:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{b s_x^2}{s_x s_y} = b\frac{s_x}{s_y}$$

✅ This proves:

$$\boxed{r_{xy} = b\frac{s_x}{s_y}}$$

Step 2: Express r_{xy}^2

$$r_{xy}^2 = \left(b\frac{s_x}{s_y}\right)^2 = b^2\frac{s_x^2}{s_y^2}$$

✅ This proves:

$$\boxed{r_{xy}^2 = \frac{b^2 s_x^2}{s_y^2}}$$

Conclusion / Interpretation

- The square of the correlation coefficient is directly related to the **squared slope** of the regression line, scaled by the ratio of the variances of X and Y.
- This shows mathematically that **strong correlation → steep slope (if S_x, S_y fixed)**.

RELATIONSHIP BETWEEN *R* AND R^2

Simple Linear Regression Model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

Definitions:

$$r = \frac{s_{xy}}{s_x s_y} \quad \text{and} \quad R^2 = \frac{SSR}{SST}$$

Derivation:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$SSR = \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 = (n-1)\frac{s_{xy}^2}{s_x^2}$$

$$SST = (n-1)s_y^2$$

$$\Rightarrow R^2 = \frac{SSR}{SST} = \frac{s_{xy}^2}{s_x^2 s_y^2} = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = r^2$$

Therefore:

Model	Relationship
Simple Linear Regression	$R^2 = r^2$
Multiple Linear Regression	$R^2 \neq r^2$

Interpretation:
In simple linear regression, the coefficient of determination (R^2) equals the square of the correlation coefficient (r), meaning the proportion of variance in Y explained by X is directly related to their linear association.

THANKS!

Questions?