

Correlation & Regression

Part I

Prepared by Dr. Hesham Mousli

MHA 613
Statistics in Healthcare

Lecture 9



Intended Objectives

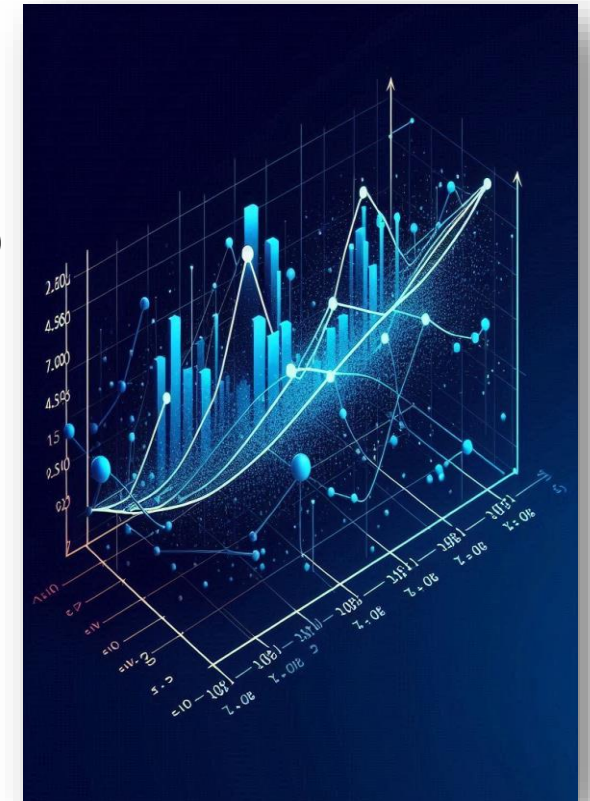
1. Distinguish between correlation and regression, their purposes, and their differences in analyzing relationships between variables.
2. Learn how to interpret correlation coefficients (e.g., Spearman's r).
3. Understand the steps and best practices for performing correlation and regression analyses.

Correlation vs Regression

- **Correlation:** Measures the strength of the relationship between variables.
- **Regression:** Describes the relationship and provides predictive equations.

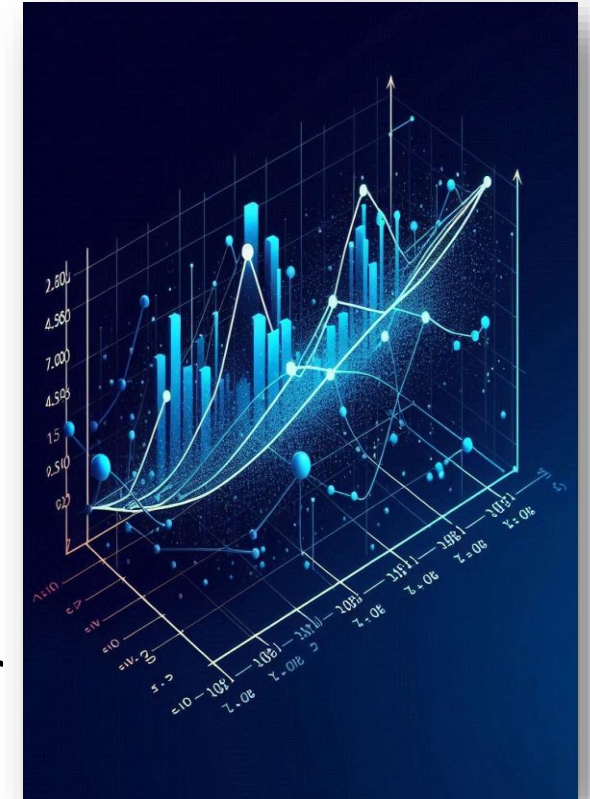
Definition of Correlation

- **Correlation** measures **how strongly** two variables are **associated**.
- It does not imply causation.



Correlation Coefficient

- **The Correlation Coefficient (r):**
- Measures linear relationships.
 - Ranges from -1 to +1.
 - **Positive:** Variables increase together.
 - **Negative:** One variable increases as the other decreases.



Interpreting r Values

- **$r = 1$ or -1 :** Perfect linear relationship.
- **$r = 0$:** No linear relationship.
- **Intermediate values:** Degree of association.

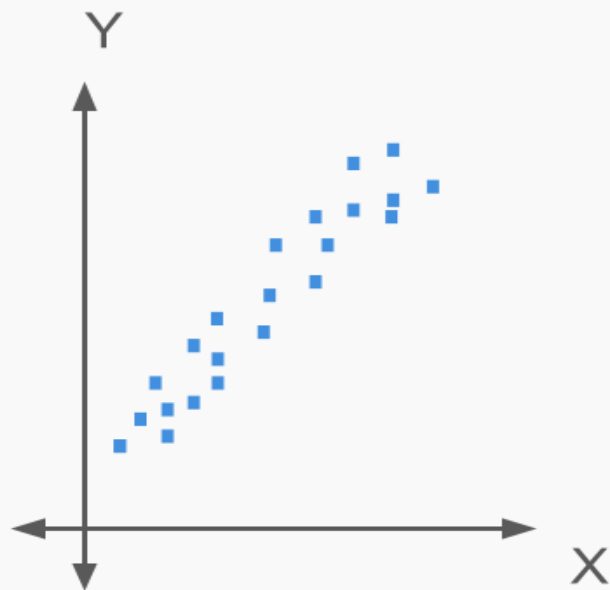
Testing Significance of r

- Even independent variables may show small r values due to random fluctuations.
- Significance testing (**p-value**) determines if r differs from zero.

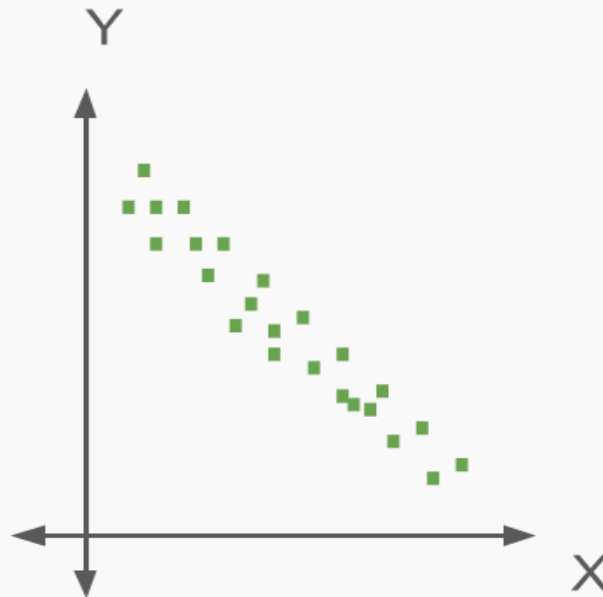
Visualization of Correlation

- **Scatter plots** to observe relationships.
- **Linearity and clustering** of points provide insights into **strength and direction**.

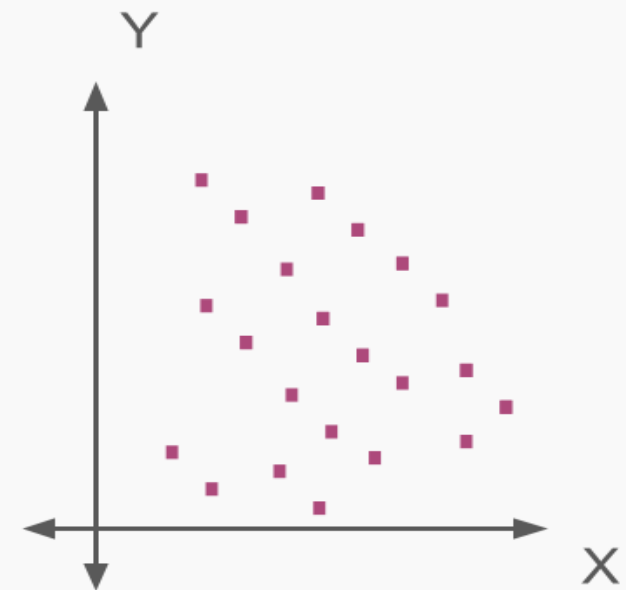
Scatter Plot Correlation Examples



**Positive
Correlation**



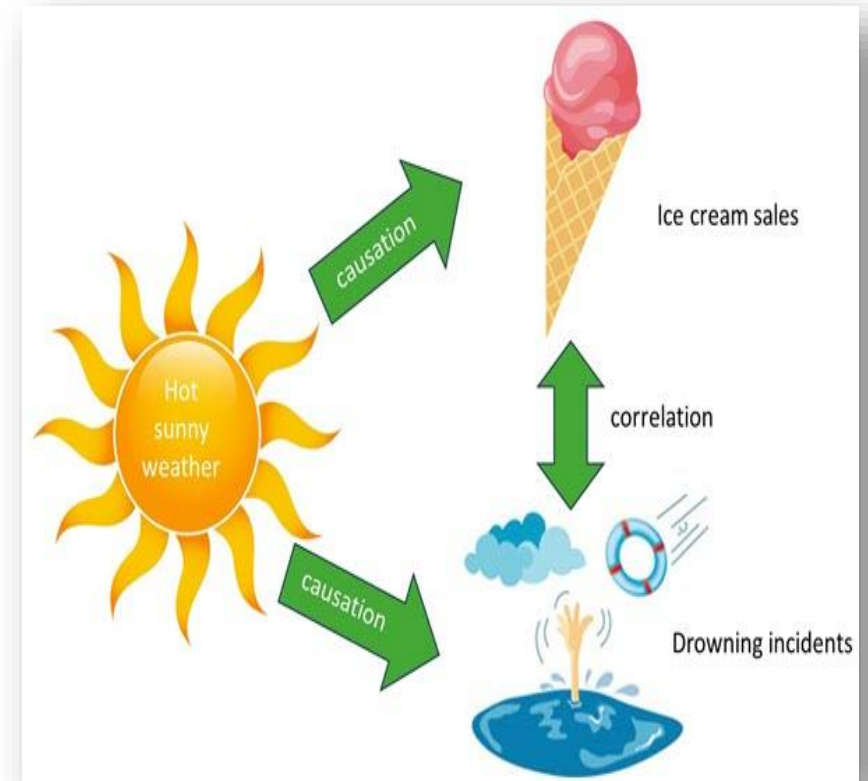
**Negative
Correlation**



**No
Correlation**

Limitations of Correlation

- Does not imply causation.
- Sensitive to outliers.
- Limited to linear relationships.



Applications of Correlation in Healthcare

- **Hospital Bed Occupancy and Infection Rates:**
- **Correlation:** There is often a positive correlation between high hospital bed occupancy rates and the incidence of hospital-acquired infections (HAIs). Overcrowding can strain hygiene protocols and increase the risk of cross-infections among patients.
- **Implication:** Monitoring and managing bed occupancy can help reduce the risk of HAIs and improve patient safety.

Case Scenario (1)

- A healthcare administrator wants to examine whether the **frequency of staff training sessions** (measured in hours per month) is correlated with the **rate of medication errors** in a hospital ward.
- **Objective:**
 - To identify whether increasing training hours is associated with a reduction in medication errors.

 Training sessions (hours)	 Medication errors rate (%)
4	2.3
3	1.4
6	2.1
6	1.3
7	2.1
5	3.1
6	1.7
8	2.1
4	1.3

Statistical Analysis Menu:

- Edit Data
- Descriptives
- T-Tests
- ANOVA
- Mixed Models
- Regression
- Frequencies

Training sessions (hours)	Medication errors rate (%)
4	2.3
3	1.4
6	2.1
6	1.3
7	2.1
5	3.1
6	1.7
8	2.1

Regression Analysis Options:

- Classical**
 - Correlation
 - Linear Regression
 - Logistic Regression
 - Generalized Linear Model
- Bayesian**
 - Correlation
 - Linear Regression
 - Logistic Regression

Correlation

Training sessions (hours)

Medication errors rate (%)

↓^A₂

▶

▶

Variables

Partial out

Sample Correlation Coefficient

☐ Pearson's r

☒ Spearman's rho

☐ Kendall's tau-b

Additional Options

☐ Display pairwise

☒ Report significance

☐ Flag significant correlations

☐ Confidence intervals

Results

Correlation

Spearman's Correlations

Variable		Medication errors rate (%)
1. Medication errors rate (%)	Spearman's rho	—
	p-value	—
2. Training sessions (hours)	Spearman's rho	-0.028
	p-value	0.902

Case Scenario (1)

- **Correlation coefficient (r) = -0.028:** This indicates a very weak negative correlation between the frequency of staff training sessions and the rate of medication errors. The magnitude of the correlation is so small that it is practically negligible.
- **p-value = 0.902:** This is far greater than the common significance level (e.g., 0.05), meaning the result is not statistically significant.

Case Scenario (2)

- A public health department wants to study the relationship between **healthcare spending per capita** and **life expectancy** in different regions.
- **Objective:**
 - To determine if higher healthcare expenditure is correlated with improved life expectancy.

	healthcare spending per capita (\$)	life expectancy (years)
1	750	15
2	430	12
3	250	10
4	360	5
5	470	13
6	650	14
7	632	13.5
8	544	12.5
9	855	18

The screenshot displays a statistical software interface. At the top, a toolbar contains icons for various statistical functions: Edit Data, Descriptives, T-Tests, ANOVA, Mixed Models, Regression, and Frequencies. Below the toolbar is a data table with two columns: 'healthcare spending per capita (\$)' and 'life expectancy (years)'. The table contains 8 rows of data. To the right of the table, a 'Regression' menu is open, showing options for 'Classical' and 'Bayesian' regression. The 'Classical' section includes 'Correlation', 'Linear Regression', 'Logistic Regression', and 'Generalized Linear Model'. The 'Bayesian' section includes 'Correlation', 'Linear Regression', and 'Logistic Regression'. Two orange arrows point to the 'Classical' and 'Correlation' options in the menu.

	healthcare spending per capita (\$)	life expectancy (years)
1	750	15
2	430	12
3	250	10
4	360	5
5	470	13
6	650	14
7	632	13.5
8	544	12.5

Regression Menu Options:

- Classical**
 - Correlation
 - Linear Regression
 - Logistic Regression
 - Generalized Linear Model
- Bayesian**
 - Correlation
 - Linear Regression
 - Logistic Regression

Correlation

Variables

- healthcare spending per capita (\$)
- life expectancy (years)

Partial out

Sample Correlation Coefficient

- ☐ Pearson's r
- ☒ Spearman's rho
- ☐ Kendall's tau-b

Additional Options

- ☐ Display pairwise
- ☒ Report significance
- ☐ Flag significant correlations

Correlation ▼

Spearman's Correlations ▼

Variable	healthcare spending per capita (\$)	
1. healthcare spending per capita (\$)	Spearman's rho	—
	p-value	—
2. life expectancy (years)	Spearman's rho	0.793
	p-value	< .001

Case Scenario (2)

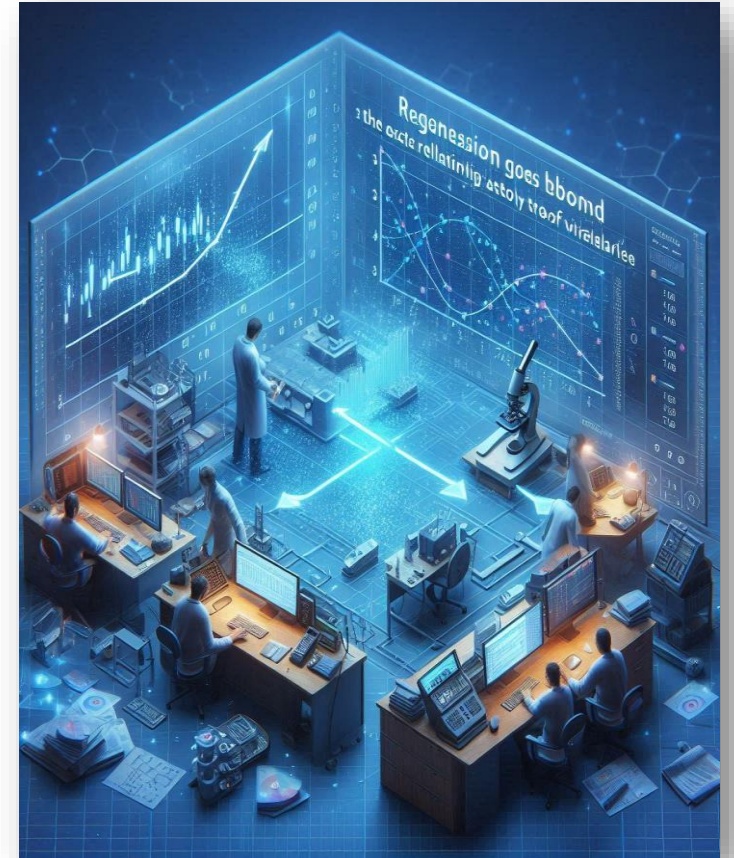
- The **p-value (< 0.001)** indicates that the observed correlation between healthcare spending per capita and life expectancy is **statistically significant**.
- This means **there is strong evidence to reject the null hypothesis**, which assumes no correlation between these variables.

Case Scenario (2)

- The **correlation coefficient ($r = 0.793$)** suggests a **strong positive relationship** between healthcare spending per capita and life expectancy.
- This means that regions with higher healthcare spending per capita tend to have higher life expectancy, and vice versa.

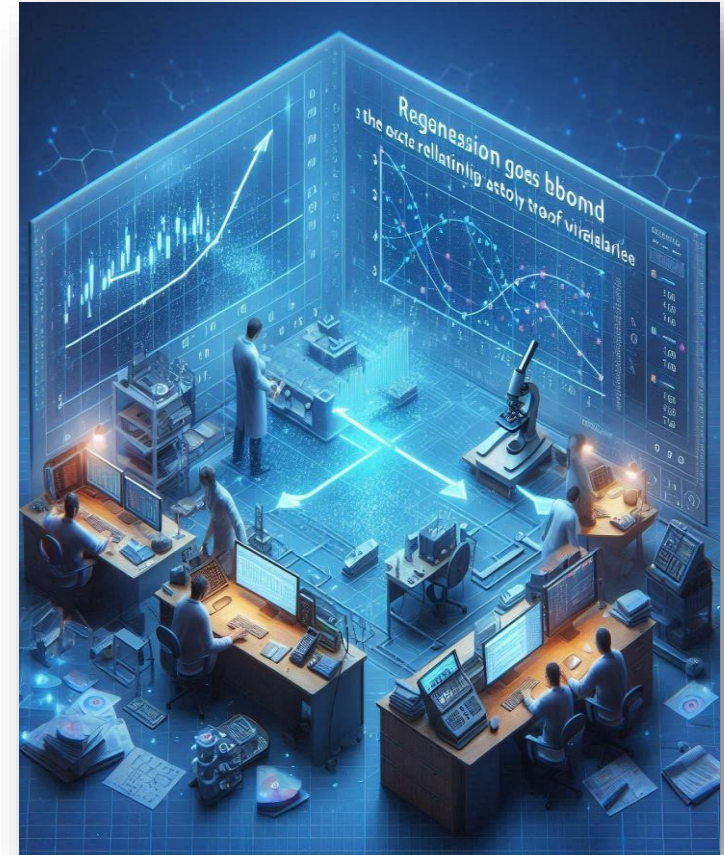
Introduction to Regression

- Regression goes **beyond correlation** by determining the **exact relationship** between variables.



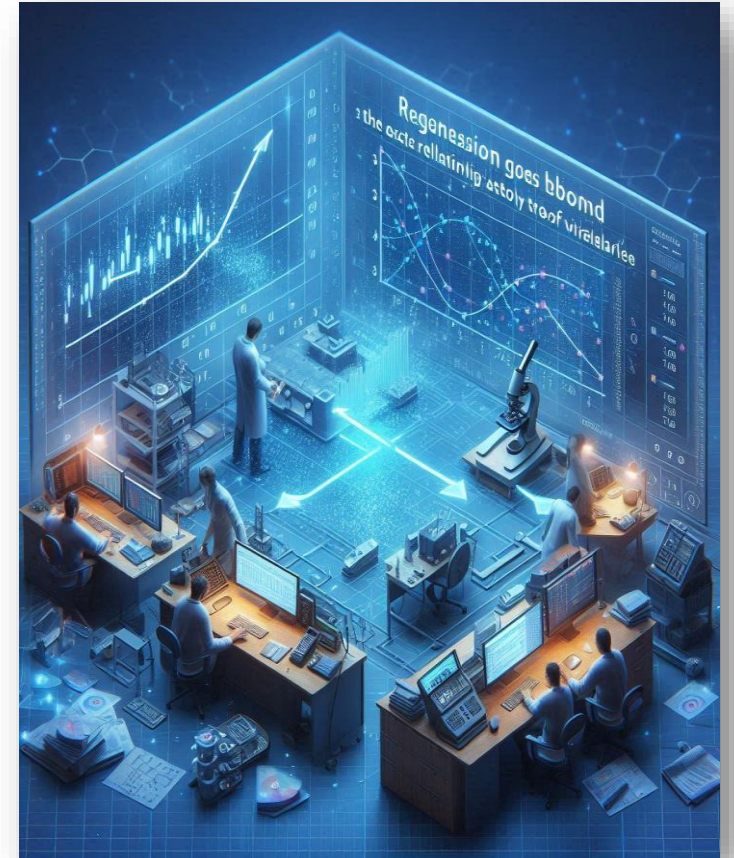
Purpose of Regression Analysis

- Test for **significant associations**.
- Provide a **compact representation of data**.
- Make **predictions and interpolations**.



Key Terms in Regression

- Independent variables (**predictors**).
- Dependent variable (**outcome**).
- Parameters (**e.g., slope, intercept**).



1. Simple Linear Regression

- It models the relationship between one predictor and one outcome.
- Formula: $Y = a + bX$.

- A healthcare administrator wants to determine whether the number of available hospital beds in a facility can predict the patient occupancy rate.
- **Dataset Variables:**
 - **Independent Variable (X):** Number of hospital beds (e.g., 20, 30, 40, etc.).
 - **Dependent Variable (Y):** Patient occupancy rate (percentage of beds occupied).

➤ **Hypothesis:**

- **Null Hypothesis (H_0):** The number of hospital beds does not significantly predict the patient occupancy rate.
- **Alternative Hypothesis (H_1):** The number of hospital beds significantly predicts the patient occupancy rate.

2. Multiple Linear Regression

- It **models** the relationship with **multiple predictors**.
- **Formula:** $Y = b_0 + b_1X_1 + b_2X_2 + \dots$

➤ **Title:** Predicting Hospital Length of Stay (LOS):

➤ **Objective:**

To understand how various factors influence the length of stay (LOS) in a hospital and develop a model to predict it.

➤ **Dependent Variable (Y):**

- Length of Stay (in days).

➤ **Independent Variables (X1, X2, X3, etc.):**

- **Age of the patient (X1):**
Older patients may have longer stays due to complications or comorbidities.
- **Severity of the condition (X2):**
Categorized by diagnostic scores or emergency admission status.
- **Type of procedure or treatment (X3):**
Major surgeries may require extended recovery periods.
- **Insurance coverage status (X4):**
Patients with insurance may receive more comprehensive care.
- **Hospital capacity and staffing levels (X5):**
Understaffed facilities may influence LOS due to delayed treatments.

3. Logistic Regression

- Used when **the outcome is categorical** (e.g., disease/no disease).
- Outputs **probabilities**.

- A hospital's administration wants to predict whether patients admitted to the emergency department (ED) are **likely to be readmitted within 30 days** after discharge.
- This helps the hospital **allocate resources effectively** and **improve patient outcomes**.

➤ **Data Used:**

➤ **Dependent Variable (Outcome):** Readmission status (**1 = Readmitted within 30 days, 0 = Not readmitted**).

➤ **Independent Variables (Predictors):**

- Patient age, Length of stay, Number of comorbidities, Previous hospitalizations, Insurance type, Discharge condition, Medication adherence, Socioeconomic status.

Steps in Regression Analysis

1. Define the research question.
2. Collect and clean data.
3. Select the appropriate regression model.
4. Evaluate model performance.

Wrap Up

1. **Correlation** measures the strength and direction of variable relationships without implying causation.
2. **Regression** explains relationships and predicts outcomes between dependent and independent variables.
3. **Types of Regression:** Includes simple, multiple, and logistic regression for continuous and categorical outcomes.
4. **Applications:** Useful in healthcare for predicting outcomes, analyzing risk factors, and resource allocation.



Thank You