جامـــــعة المـــــستـقبـل
AL MUSTAQBAL UNIVERSITY

قســـم الامـــــن الـــــــسيبرانـــــــي

Department of Cyber Security

Subject:

Regression and Correlation

Class:

first

Lecturer:

Asst. lect. Mustafa Ameer Awadh

Lecture: (11)

Introduction

Regression and correlation are statistical methods used to analyze the relationship between two or more variables. Regression helps predict values, while correlation measures the strength and direction of the relationship between variables.

---

1. Correlation Analysis

Definition:

Correlation measures the degree to which two variables move in relation to each other. It is quantified by the correlation coefficient (r).

Types of Correlation:

- **Positive Correlation:** As one variable increases, the other also increases.
- **Negative Correlation:** As one variable increases, the other decreases.
- **No Correlation:** No relationship between the variables.

Correlation Coefficient (r):

The Pearson correlation coefficient is given by:

where:

- are individual data points
- are mean values of X and Y

Interpretation of r:

r Value	Strength of Relationship
-1.0 to -0.7	Strong Negative Correlation
-0.7 to -0.3	Moderate Negative Correlation
-0.3 to 0.3	Weak or No Correlation
0.3 to 0.7	Moderate Positive Correlation
0.7 to 1.0	Strong Positive Correlation

Limitations:

- Correlation does not imply causation.
- Outliers can distort the correlation coefficient.

---

2. Regression Analysis

Definition:

Regression analysis is used to model the relationship between a dependent variable (Y) and one or more independent variables (X).

Types of Regression:

- **Simple Linear Regression:** Relationship between one independent variable (X) and one dependent variable (Y).
- **Multiple Linear Regression:** Relationship between multiple independent variables and a dependent variable.

Simple Linear Regression Equation:

where:

- = dependent variable
- = independent variable
- = intercept
- = slope (rate of change of Y with respect to X)
- = error term

Finding the Best Fit Line (Least Squares Method):

The slope (b) and intercept (a) are calculated as:

Interpreting the Slope and Intercept:

- **Slope (b):** The amount by which Y changes for a one-unit increase in X.
- **Intercept (a):** The expected value of Y when X = 0.

---

3. Coefficient of Determination (R^2)

Definition:

The coefficient of determination () measures how well the regression line fits the data.

where:

- = sum of squared residuals (error)
- = total sum of squares

Value	Explanation
0 to 0.3	Weak fit
0.3 to 0.6	Moderate fit
0.6 to 1.0	Strong fit

- High does not always indicate a good model.
- Outliers can influence the regression results.

---

4. Hypothesis Testing in Regression

Testing the Significance of Regression Coefficients

- **Null Hypothesis ()**: The coefficient is not significantly different from zero ().
- **Alternative Hypothesis ()**: The coefficient is significantly different from zero ().
- **t-test:**

where is the standard error of the slope.

- **Decision Rule:** Compare the calculated t-value with the critical t-value from the t-distribution table.

---

5. Practical Example

Problem: A company wants to predict sales (Y) based on advertising expenditure (X). Given the dataset:

Advertising (X)	Sales (Y)
10	200
15	250
20	300
25	330
30	390

1. Compute the correlation coefficient (r).
2. Fit a linear regression model.
3. Interpret the slope and intercept.
4. Determine .

Conclusion

- **Correlation** measures the strength of a relationship between two variables.
- **Regression** helps predict outcomes based on independent variables.
- assesses how well the model fits the data.
- **Hypothesis testing** determines if relationships are statistically significant.

Practice Questions

1. Given the dataset below, calculate the Pearson correlation coefficient:

X	Y
1	3
2	6
3	9
4	12
5	15

2. If a regression equation is , interpret the slope and intercept.
3. A dataset has . What does this imply about the model's fit?