Course name: Engineering statistics
Course Code: STAT- 2201

Presentation about:

Test for Correlation and Regression

Understanding the Significance of Relationships in Data

Presented by: MD Rahat Mahbub
Roll: 220641
Reg: 1065488
2nd Year
2nd Semester,
Pabna University
Of Science And Technology

Presented to: Dr. Md. Sarwar Hosain
Associate Professor
Department of
Information & Communication engineering,
Pabna University
Of Science And Technology

Contents

01

Introduction to Correlation and Regression

Understanding the concept

Definition of Correlation

Correlation measures the strength and direction of a relationship between two variables. It indicates how changes in one variable may be associated with changes in another, ranging from -1 (perfect negative) to +1 (perfect positive), with 0 indicating no correlation.

Definition of Regression

Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It helps to predict the value of the dependent variable based on the values of the independent variables, illustrating how changes in one variable can affect another.

Importance in Statistics

Correlation and regression are essential statistical tools that measure relationships between variables. Understanding their significance aids in data analysis, guiding predictions, and making informed decisions, thereby enhancing research, finance, and social sciences applications.

Types of Correlation

Positive vs Negative Correlation

Positive correlation indicates that as one variable increases, the other also increases, while negative correlation signifies that as one variable increases, the other decreases. Understanding these relationships is essential in regression analysis, enabling predictions and insights based on the nature of variable interactions.

Strong vs Weak Correlation

Correlation measures the strength and direction of a relationship between two variables. A strong correlation indicates that changes in one variable significantly affect the other, while a weak correlation suggests a minor or negligible relationship, making predictions less reliable.

Pearson vs Spearman Correlation

Pearson correlation measures linear relationships between continuous variables, assuming normality, while Spearman correlation assesses monotonic relationships, suitable for ordinal data or non-normally distributed variables. Understanding the differences is crucial for selecting the appropriate test based on data characteristics.

02

Correlation Coefficients

Pearson Correlation Coefficient

Formula and Calculation

The Pearson Correlation Coefficient has limitations, including sensitivity to outliers which can skew results, assumption of linearity between variables, and inability to establish causation. It also requires both variables to be measured on an interval or ratio scale for accurate analysis.

The Pearson Correlation Coefficient (r) is calculated using the formula (r=[nΣx2−(Σx)2][nΣy2−(Σy)2]n (Σxy)−(Σx)(Σy)). This measures the strength and direction of linear relationships between two variables.

limitations

Interpretation of Results

The Pearson Correlation Coefficient quantifies the strength and direction of a linear relationship between two variables. A value close to +1 indicates a strong positive correlation, while a value near -1 signifies a strong negative correlation. A value around 0 suggests no correlation.

Spearman Rank Correlation Coefficient

Formula and Calculation

The Spearman Rank Correlation Coefficient (ϱ) is calculated using the formula $\varrho = 1 - (6\Sigma d^2) / (n(n^2 - 1))$, where d is the difference between ranks and n is the number of pairs. This non-parametric measure assesses the strength and direction of monotonic relationships.

When to Use Spearman

Use the Spearman Rank Correlation Coefficient when data is non-parametric, ordinal, or not normally distributed. It is ideal for small sample sizes and when relationships between variables are monotonic rather than strictly linear, making it a versatile tool in statistical analysis.

Advantages and Disadvantages

Advantages of the Spearman Rank Correlation Coefficient include its robustness to non-normal distributions and its effectiveness with ordinal data. However, its disadvantages include less sensitivity to linear relationships compared to Pearson's correlation and potential loss of information due to ranking.

03

Regression Analysis

Simple Linear Regression

Definition and Formula

- Regression analysis predicts the relationship between variables using a linear equation. The formula for a simple linear regression is ($Y = a + bX$), where (Y) is the dependent variable, (a) is the intercept, (b) is the slope, and (X) is the independent variable.

Assumptions of Simple Linear Regression

- Assumptions of Simple Linear Regression include linearity, independence of errors, homoscedasticity (equal variance), normality of error terms, and no multicollinearity. Ensuring these assumptions are met is crucial for valid and reliable regression results and insights.

Interpreting the Regression Coefficients

- In simple linear regression, the regression coefficient indicates the change in the dependent variable for each one-unit change in the independent variable. A positive coefficient suggests a direct relationship, while a negative coefficient indicates an inverse relationship. The magnitude reflects the strength of the association.

Simple vs Multiple Regression

Simple regression deals with the relationship between two variables, involving one dependent and one independent variable. Multiple regression, on the other hand, includes multiple independent variables to predict the dependent variable, allowing for more complex interrelationships.

04

Hypothesis Testing for Correlation

Hypothesis Testing for Correlation

Setting up the Hypotheses

- In hypothesis testing for correlation, the null hypothesis asserts no correlation (H0: $\rho = 0$), while the alternative hypothesis suggests a significant correlation exists (H1: $\rho \neq 0$). This framework guides statistical analysis and interpretation of correlation coefficients.

Types of Errors (Type I & II)

- In hypothesis testing for correlation, the null hypothesis posits no relationship between variables, while the alternative suggests a significant correlation exists. Type I error occurs when rejecting a true null hypothesis, whereas Type II error happens when failing to reject a false null hypothesis. Understanding these errors is crucial for accurate interpretation.

Testing Procedure

- In hypothesis testing for correlation, the null hypothesis states that there is no correlation between variables ($\rho = 0$), while the alternative hypothesis suggests that a correlation exists ($\rho \neq 0$). The testing procedure involves calculating the correlation coefficient and determining its significance using a p-value against a chosen significance level.

05

Assumptions and Diagnostics

Assumptions in Regression Analysis

Linearity

Linearity assumes a straight-line relationship between independent and dependent variables. This can be assessed through scatter plots, residual plots, and correlation coefficients. Non-linear patterns indicate that a different model may be needed for accurate predictions.

Independence of Errors

Independence of errors assumes that the residuals (errors) from the regression model are uncorrelated. This is crucial as correlated errors can indicate model misspecification, leading to biased estimates and unreliable statistical inferences. Autocorrelation tests, such as the Durbin-Watson test, can help assess this assumption.

Homoscedasticity

Homoscedasticity refers to the consistency of variance in the residuals throughout the range of predicted values. If the spread of the residuals increases or decreases, it indicates a violation of this assumption, potentially leading to inefficient estimates and unreliable hypothesis tests.

06

Applications of Correlation and Regression

Case Studies in Various Fields

Business and Economics

Correlation and regression analysis are vital in business and economics for forecasting sales, analyzing market trends, and evaluating the impact of marketing strategies. They enable decision-makers to understand relationships between variables, optimize resource allocation, and enhance overall operational efficiency.

Health Sciences

n health sciences, correlation and regression analysis are used to examine relationships between variables, such as the impact of lifestyle factors on disease risk. These methods help identify trends and inform preventive measures, enhancing patient outcomes and guiding public health policies.

Social Sciences

In social sciences, correlation and regression analysis are used to explore relationships between variables, such as the impact of education on income levels or the influence of social media usage on mental health. These analyses help researchers predict trends and inform policy decisions.

07

Conclusion

Key TakeawaysContent:

Correlation and regression are powerful tools that allow us to examine and predict relationships between variables . Statistical hypothesis testing enables us to validate the significance of these relationships . A non-significant test result implies that the relationship could be due to random variation, whereas a significant result supports a meaningful connection . These techniques are widely used in research, business, economics, psychology, medicine, and more . Always ensure assumptions (e.g., linearity, normality) are checked before performing these tests to avoid misleading conclusions.

Thank you

The End