STA 3180 Statistical Modelling: Multivariate Analysis

STA 3180 Statistical Modelling - Lecture Notes on Multivariate Analysis

Introduction

Multivariate analysis is a statistical technique used to analyze data from multiple sources. It is used to examine relationships between multiple variables and can be used to identify patterns and trends in data. Multivariate analysis can be used to identify the most important factors influencing a particular outcome, or to predict future outcomes.

Key Concepts

- **Correlation**: Correlation is a measure of the strength of the linear relationship between two variables. It is usually expressed as a number between -1 and 1, where a value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation.

- **Covariance**: Covariance is a measure of the degree to which two variables vary together. It is usually expressed as a number between -1 and 1, where a value of 1 indicates a perfect positive covariance, a value of -1 indicates a perfect negative covariance, and a value of 0 indicates no covariance.

- **Principal Component Analysis (PCA)**: Principal Component Analysis (PCA) is a multivariate technique used to reduce the dimensionality of a dataset by transforming it into a set of uncorrelated variables called principal components. PCA is used to identify patterns and trends in data, and can be used to reduce the complexity of a dataset.

- **Factor Analysis**: Factor analysis is a multivariate technique used to identify underlying factors or latent variables in a dataset. It is used to reduce the complexity of a dataset by identifying the most important variables that explain the variance in the data.

Definitions

- **Multivariate Analysis**: Multivariate analysis is a statistical technique used to analyze data from multiple sources. It is used to examine relationships between multiple variables and can be used to identify patterns and trends in data.

- **Correlation**: Correlation is a measure of the strength of the linear relationship between two variables. It is usually expressed as a number between -1 and 1, where a value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation.

- **Covariance**: Covariance is a measure of the degree to which two variables vary together. It is usually expressed as a number between -1 and 1, where a value of 1 indicates a perfect positive covariance, a value of -1 indicates a perfect negative covariance, and a value of 0 indicates no covariance.

- **Principal Component Analysis (PCA)**: Principal Component Analysis (PCA) is a multivariate technique used to reduce the dimensionality of a dataset by transforming it into a set of uncorrelated variables called principal components. PCA is used to identify patterns and trends in data, and can be used to reduce the complexity of a dataset.

- **Factor Analysis**: Factor analysis is a multivariate technique used to identify underlying factors or latent variables in a dataset. It is used to reduce the complexity of a dataset by identifying the most important variables that explain the variance in the data.

Coding Examples

Correlation
Start of Code

```
// Calculate the correlation between two variables

double x = 5;
double y = 10;

double correlation = (x - meanX) * (y - meanY) / (stdevX * stdevY);
```

End of Code

Principal Component Analysis (PCA)
Start of Code

```
// Perform PCA on a dataset

// Step 1: Calculate the covariance matrix
Matrix covMatrix = Matrix.cov(data);

// Step 2: Calculate the eigenvalues and eigenvectors of the covariance matrix
Matrix eigVals = covMatrix.eig();
Matrix eigVecs = covMatrix.eigVecs();

// Step 3: Select the principal components
Matrix pcMatrix = eigVecs.select(numComponents);

// Step 4: Transform the data using the principal components
Matrix transformedData = data.transform(pcMatrix);
```

End of Code

Practice Multiple Choice Questions

Q1. Which of the following is not a multivariate technique?

A) Correlation

B) Regression

C) Factor Analysis

D) Clustering

Answer: D) Clustering