DATA ANALYTICS- UNIT 2

By

DEEPIKA KAMBOJ

Linear Regression Key Points

- Linear regression is a statistical method used to study the relationship between a dependent variable and one or more independent variables.
- The simplest form of linear regression involves fitting a straight line to a set of data points, with the goal of finding the line that best describes the relationship between the variables.
- Linear regression can be used to make predictions, estimate the strength and direction of the relationship between variables, and to identify patterns in data.
- The parameters of a linear regression model are estimated using a method called least squares, which involves minimizing the sum of the squared differences between the predicted values and the actual values.

Linear Regression Key Points

- Linear regression can be used for both simple regression (i.e. one independent variable) and multiple regression (i.e. more than one independent variable).
- Linear regression models can be evaluated using various metrics, such as R-squared (a measure of the proportion of the variation in the dependent variable that is explained by the independent variables), mean squared error (a measure of the average difference between the predicted and actual values), and residual plots.

Linear Regression Example

Hours studied	Exam score
2	65
3	72
4	80
5	84
6	89

Linear Regression Example

Step 1: Calculate mean of hours studied and exam score

Step 2: Calculate deviation of hours studied and exam score

Step 3: Calculate slope of linear regression

Slope = covariance (hours studied, exam score) / variance of hours studied

Step 4: Calculate Intercept

Intercept = mean of exam score - slope * mean of hours studied

Regression MCQs

Which of the following regression models is best suited for predicting a categorical outcome?

a. Linear regression

b. Logistic regression

c. Multiple regression

d. Polynomial regression

Regression MCQs

Which of the following regression models can handle multiple predictor variables simultaneously?

a. Linear regression

b. Logistic regression

c. Multiple regression

d. Polynomial regression

Regression MCQs

Which of the following regression models is best suited for capturing nonlinear relationships between predictor and outcome variables?

a. Linear regression

b. Logistic regression

c. Multiple regression

d. Polynomial regression

Regression MCQs

In multiple regression, what is the purpose of the coefficient of determination (R-squared)?

a. To measure the correlation between the predictor variables

b. To measure the correlation between the outcome variable and each predictor variable separately

c. To measure the overall fit of the model to the data

d. To measure the difference between the predicted and actual outcome values

Regression MCQs

Which of the following statements is true regarding linear regression?

a. It can only handle categorical predictor variables

b. It is only used for predicting binary outcomes

c. It assumes a linear relationship between the predictor and outcome variables

d. It is a non-parametric method

Regression MCQs

Which of the following statements is true regarding logistic regression?

a. It assumes a linear relationship between the predictor and outcome variables.

b. It can only handle continuous predictor variables.

c. It is used to predict continuous outcomes.

d. It is a non-parametric method.

Regression MCQs

Which of the following regression models can handle both categorical and continuous predictor variables?

a. Linear regression

b. Logistic regression

c. Multiple regression

d. Polynomial regression

Regression MCQs

Which of the following is a way to evaluate the fit of a linear regression model?

a. Mean squared error

b. R-squared

c. Mean absolute error

d. All of the above

Regression MCQs

Which of the following is a way to evaluate the fit of a logistic regression model?

a. Mean squared error

b. R-squared

c. Mean absolute error

d. None of the above

Regression MCQs

Which of the following is a potential issue with using polynomial regression?

a. It is computationally expensive to fit the model.

b. It can lead to overfitting if the degree of the polynomial is too high.

c. It can only handle categorical predictor variables.

d. It assumes a linear relationship between the predictor and outcome variables.

Multivariate Analysis

Multivariate analysis is a statistical technique used to analyse and understand the relationships between multiple variables