Multiple linear regression model

Statistics and Big Data

Niccolò Salvini, PhD

UCSC

Academic Year 2025-2026

Course: Statistics and Big Data

Overview

What is Multiple Regression?

Definition

Multiple regression is a statistical technique used to predict an outcome using multiple variables instead of just one.

What is Multiple Regression?

Definition

Multiple regression is a statistical technique used to predict an outcome using multiple variables instead of just one.

Importance

This approach differs from simple linear regression and is crucial for understanding complex relationships in data.

Understanding the Basics of Multiple Regression

Example

Consider predicting the body length of mice. In simple linear regression, we might use just one variable, such as mouse weight.

Understanding the Basics of Multiple Regression

Example

Consider predicting the body length of mice. In simple linear regression, we might use just one variable, such as mouse weight.

Multiple Variables

In multiple regression, we can include additional variables, such as tail length, amount of food eaten, and time spent running on a wheel.

Core Concept of Multiple Regression

Key Insight

Multiple regression allows us to fit a plane (or higher-dimensional object) to our data, accommodating multiple predictors.

Core Concept of Multiple Regression

Key Insight

Multiple regression allows us to fit a plane (or higher-dimensional object) to our data, accommodating multiple predictors.

Mathematical Formulation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where Y is the dependent variable, X_i are the independent variables, β_i are the coefficients, and ϵ is the error term.

Visualizing Multiple Regression

Three-Dimensional Plot

Consider a three-dimensional plot where body length is predicted by mouse weight and tail length.

Calculating R-squared in Multiple Regression

Definition

R-squared measures the proportion of variance in the dependent variable that can be explained by the independent variables.

Calculating R-squared in Multiple Regression

Definition

R-squared measures the proportion of variance in the dependent variable that can be explained by the independent variables.

Mathematical Formulation

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

Calculating R-squared in Multiple Regression

Definition

R-squared measures the proportion of variance in the dependent variable that can be explained by the independent variables.

Mathematical Formulation

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

Interpretation

A higher R^2 indicates a better fit of the model to the data.

Adjusted R-squared for Multiple Regression

Adjustment

In multiple regression, we adjust R-squared to account for the number of predictors.

Mathematical Formulation

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1} \right)$$

where n is the number of observations and p is the number of predictors.

Adjusted R-squared for Multiple Regression

Adjustment

In multiple regression, we adjust R-squared to account for the number of predictors.

Mathematical Formulation

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1} \right)$$

where n is the number of observations and p is the number of predictors.

Purpose

This adjustment prevents overfitting by penalizing the addition of unnecessary predictors.

Hypothesis Testing in Multiple Regression

F-statistic

To evaluate the significance of our model, we calculate the F-statistic:

Mathematical Formulation

$$F = \frac{MS_{reg}}{MS_{res}} = \frac{SS_{reg}/p}{SS_{res}/(n-p-1)}$$

where MS_{reg} is the mean square regression and MS_{res} is the mean square residual.

Hypothesis Testing in Multiple Regression

F-statistic

To evaluate the significance of our model, we calculate the F-statistic:

Mathematical Formulation

$$F = \frac{MS_{reg}}{MS_{res}} = \frac{SS_{reg}/p}{SS_{res}/(n-p-1)}$$

where MS_{reg} is the mean square regression and MS_{res} is the mean square residual.

Interpretation

A significant F-statistic indicates that at least one predictor variable has a non-zero coefficient.

Comparing Simple and Multiple Regression

Model Improvement

To determine if adding a variable (like tail length) improves our model, we compare the R-squared values of the simple regression model to the multiple regression model.

Comparing Simple and Multiple Regression

Model Improvement

To determine if adding a variable (like tail length) improves our model, we compare the R-squared values of the simple regression model to the multiple regression model.

Justification

If the increase in R^2 is substantial and the p-value is small, adding the variable is justified.

Summary of Key Concepts

- Multiple regression extends simple regression by incorporating multiple predictors.
- R-squared and adjusted R-squared are crucial for evaluating model fit.
- The F-statistic helps assess the significance of the overall model.

Exercises

Exercise 1

Explain the difference between simple and multiple regression in your own words.

Exercises

Exercise 1

Explain the difference between simple and multiple regression in your own words.

Exercise 2

Given a dataset with body length, weight, and tail length, calculate the R-squared value for a multiple regression model.

Exercises

Exercise 1

Explain the difference between simple and multiple regression in your own words.

Exercise 2

Given a dataset with body length, weight, and tail length, calculate the R-squared value for a multiple regression model.

Exercise 3

Discuss a real-world scenario where multiple regression could provide better insights than simple regression.

Exercises

Exercise 1

Explain the difference between simple and multiple regression in your own words.

Exercise 2

Given a dataset with body length, weight, and tail length, calculate the R-squared value for a multiple regression model.

Exercise 3

Discuss a real-world scenario where multiple regression could provide better insights than simple regression.

Exercise 4

If you were to add another predictor to your model, how would you assess its impact on the overall model fit?