Linear Regression Final

Statistics and Big Data

Niccolò Salvini, PhD

UCSC

Academic Year 2025-2026

Course: Statistics and Big Data

2e

Overview

What is Linear Regression?

How can we predict one variable based on another?

What is Linear Regression?

How can we predict one variable based on another?

Definition

Linear regression provides a powerful framework for understanding relationships between variables.

What is Linear Regression?

How can we predict one variable based on another?

Definition

Linear regression provides a powerful framework for understanding relationships between variables.

In this lecture, we will explore the foundational concepts of linear regression, including fitting a line to data, calculating R^2, and determining statistical significance through p-values.

Concrete Example of Linear Regression

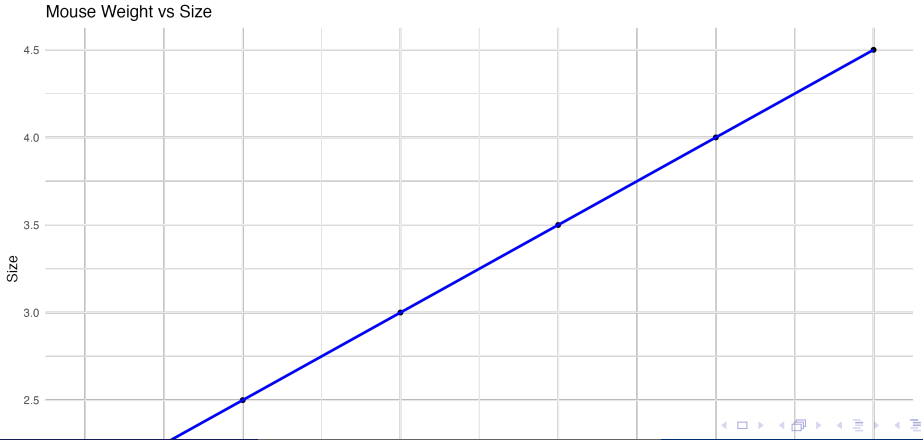
Consider a dataset where we measure the weight and size of mice.

Concrete Example of Linear Regression

Consider a dataset where we measure the weight and size of mice. Our goal is to predict mouse size based on mouse weight.

Concrete Example of Linear Regression

Consider a dataset where we measure the weight and size of mice. Our goal is to predict mouse size based on mouse weight. This scenario sets the stage for applying linear regression techniques.



Fitting a Line to Data

To fit a line to the data, we utilize the method of least squares.

Fitting a Line to Data

To fit a line to the data, we utilize the method of least squares. This involves the following steps:

1. Draw a line through the data.
2. Measure the distance from each data point to the line (these distances are called residuals).
3. Square each residual and sum them up.

Fitting a Line to Data

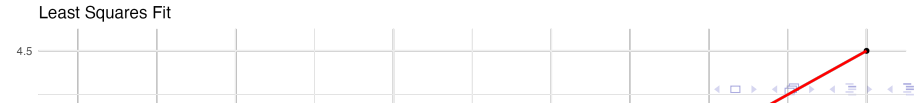
To fit a line to the data, we utilize the method of least squares. This involves the following steps:

1. Draw a line through the data.
2. Measure the distance from each data point to the line (these distances are called residuals).
3. Square each residual and sum them up.

The objective is to minimize the sum of squared residuals.

Mathematical Formulation

$$\text{Sum of Squared Residuals} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



Understanding Residuals

What do residuals tell us about our model?

Understanding Residuals

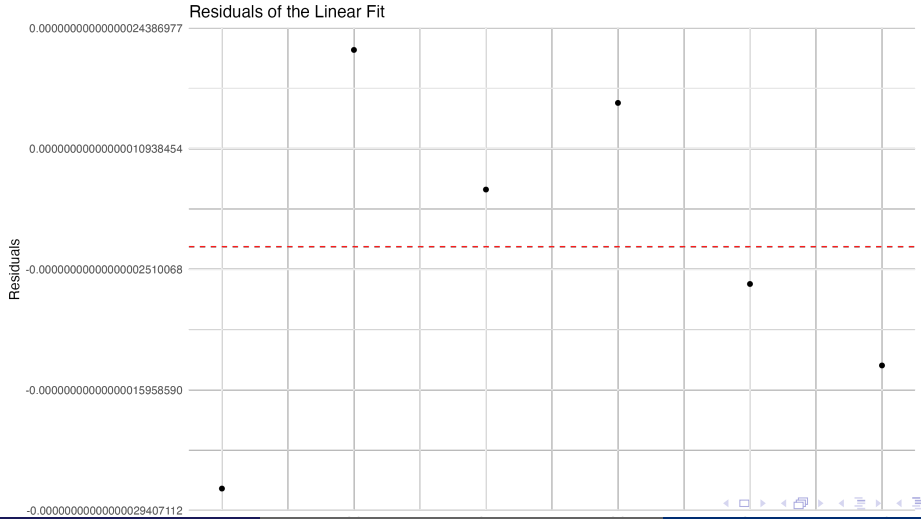
What do residuals tell us about our model? Residuals represent the difference between observed values and predicted values.

Understanding Residuals

What do residuals tell us about our model? Residuals represent the difference between observed values and predicted values. Smaller residuals indicate a better fit of the model to the data.

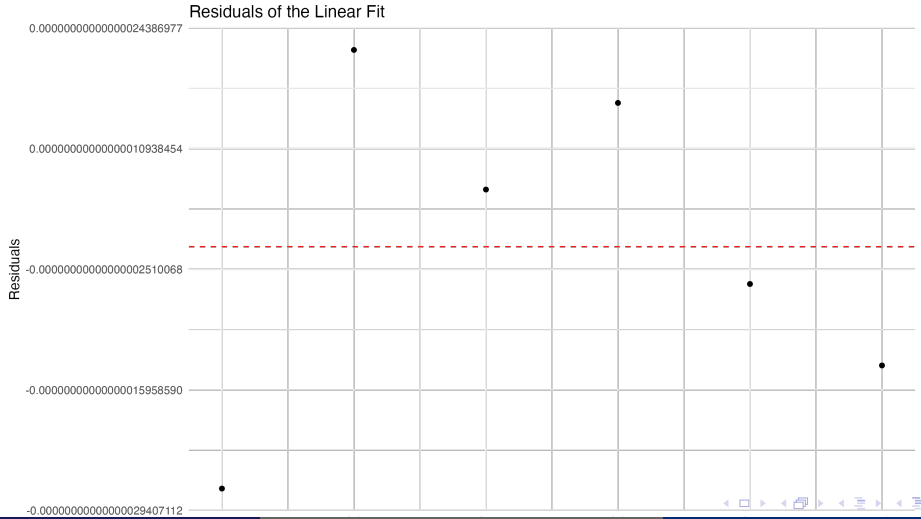
Visualizing the Fitting Process

Here we visualize the process of fitting a line to our dataset.



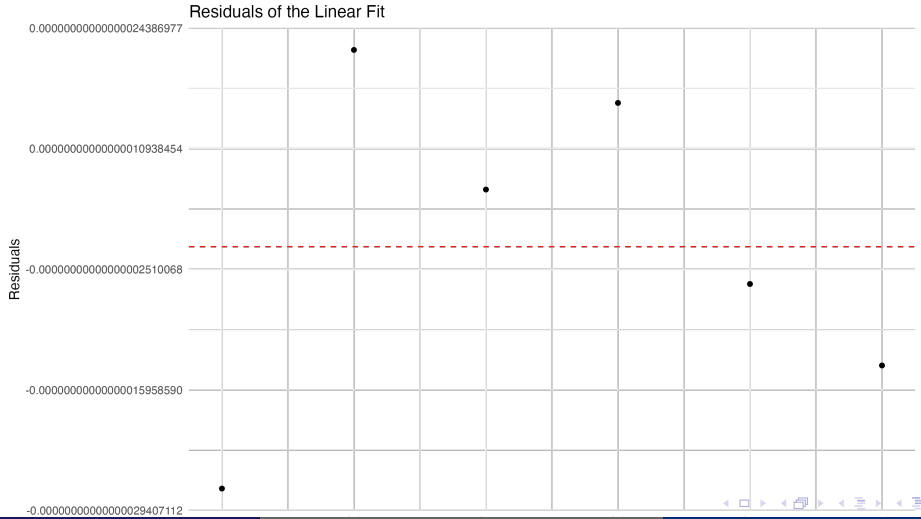
Visualizing the Fitting Process

Here we visualize the process of fitting a line to our dataset.



Visualizing the Fitting Process

Here we visualize the process of fitting a line to our dataset.



The Equation of the Fitted Line

Once we have our fitted line, we can express it mathematically:

Equation

$$y = a + b \cdot x$$

Where:

- y is the predicted size,
- a is the y-intercept,
- b is the slope,
- x is the weight.

The Equation of the Fitted Line

Once we have our fitted line, we can express it mathematically:

Equation

$$y = a + b \cdot x$$

Where:

- y is the predicted size,
- a is the y-intercept,
- b is the slope,
- x is the weight.

This equation allows us to make predictions about mouse size based on weight.

Evaluating the Fit with R^2

How do we assess the quality of our predictions?

Evaluating the Fit with R^2

How do we assess the quality of our predictions? The coefficient of determination, R^2, quantifies how much of the variation in the dependent variable can be explained by the independent variable.

Mathematical Formulation

$$R^2 = \frac{SS_{mean} - SS_{fit}}{SS_{mean}}$$

Where:

- SS_{mean} is the total sum of squares around the mean,
- SS_{fit} is the sum of squares around the fitted line.

Practical Example of R^2

Let's calculate R^2 using specific values:

- $SS_{mean} = 11.1$
- $SS_{fit} = 4.4$

Practical Example of R^2

Let's calculate R^2 using specific values:

- $SS_{mean} = 11.1$
- $SS_{fit} = 4.4$

Plugging these into our formula gives:

Calculation

$$R^2 = \frac{11.1 - 4.4}{11.1} = 0.6$$

Practical Example of R^2

Let's calculate R^2 using specific values:

- $SS_{mean} = 11.1$
- $SS_{fit} = 4.4$

Plugging these into our formula gives:

Calculation

$$R^2 = \frac{11.1 - 4.4}{11.1} = 0.6$$

This indicates that 60% of the variation in mouse size can be explained by mouse weight.

Exploring Different Scenarios for R^2

What happens when the relationship changes?

1. **Perfect Prediction:** If knowing mouse weight perfectly predicts size, $R^2 = 1$.
2. **No Prediction:** If mouse weight does not help at all, $R^2 = 0$.

Exploring Different Scenarios for R^2

What happens when the relationship changes?

1. **Perfect Prediction:** If knowing mouse weight perfectly predicts size, $R^2 = 1$.
2. **No Prediction:** If mouse weight does not help at all, $R^2 = 0$.

These scenarios illustrate the extremes of model performance.

Moving to Multiple Variables

What if we want to include more predictors, such as tail length?

Moving to Multiple Variables

What if we want to include more predictors, such as tail length? In this case, we fit a plane instead of a line, using multiple dimensions.

Equation for the Plane

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2$$

Where x_1 is weight and x_2 is tail length.

The Role of Additional Parameters

How do additional parameters affect our model?

The Role of Additional Parameters

How do additional parameters affect our model? Adding parameters can improve the fit, but it also requires more data to estimate them accurately.

The Role of Additional Parameters

How do additional parameters affect our model? Adding parameters can improve the fit, but it also requires more data to estimate them accurately. This leads to the concept of adjusted R^2, which accounts for the number of predictors in the model.

Understanding p-values in Regression

Why do we need p-values?

Understanding p-values in Regression

Why do we need p-values? While R^2 tells us how well our model fits the data, p-values help us determine if the relationship is statistically significant.

Understanding p-values in Regression

Why do we need p-values? While R^2 tells us how well our model fits the data, p-values help us determine if the relationship is statistically significant. The p-value is derived from the F-statistic, which compares the variance explained by the model to the variance not explained.

Mathematical Formulation

$$F = \frac{\text{Variance explained}}{\text{Variance not explained}}$$

Calculating the p-value

How do we compute the p-value?

Calculating the p-value

How do we compute the p-value?

1. Generate random datasets and calculate their F-statistics.
2. Compare the F-statistic from our model to the distribution of F-statistics from random datasets.

Calculating the p-value

How do we compute the p-value?

1. Generate random datasets and calculate their F-statistics.
2. Compare the F-statistic from our model to the distribution of F-statistics from random datasets.

This process helps us determine the significance of our findings.

Summary of Key Concepts

- **Linear Regression**: A method to predict one variable based on another.
- **Least Squares**: A technique to minimize the sum of squared residuals.
- R^2: A measure of how much variance is explained by the model.
- **p-value**: Indicates the statistical significance of the relationship.

Exercises

Exercise 1

Explain the significance of residuals in linear regression.

Exercises

Exercise 1

Explain the significance of residuals in linear regression.

Exercise 2

Given $\text{SS}_{\text{mean}} = 20$ and $\text{SS}_{\text{fit}} = 5$, calculate R^2.

Exercises

Exercise 1

Explain the significance of residuals in linear regression.

Exercise 2

Given $SS_{mean} = 20$ and $SS_{fit} = 5$, calculate R^2.

Exercise 3

Discuss how adding more predictors can affect the R^2 value.

Exercises

Exercise 1

Explain the significance of residuals in linear regression.

Exercise 2

Given $SS_{mean} = 20$ and $SS_{fit} = 5$, calculate R^2.

Exercise 3

Discuss how adding more predictors can affect the R^2 value.

Exercise 4

Describe the process of calculating a p-value in the context of linear regression.