Least Squares Regression Line

Introduction

In this assignment, you will derive the coefficients of the best-fit line for a set of data points using the **least squares method**. This method finds the linear function $y = ax + b$ that minimizes the sum of squared vertical distances from the data points to the line.

Objective

Given data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, derive formulas for:

$$\text{slope } a \quad \text{and} \quad \text{intercept } b$$

such that the total squared error $D = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$ is minimized.

1. Define the Error Function

Start by expressing the total squared error D as a function of a and b.

Task 1

Write the expression for $D = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$. What does this represent geometrically?

2. Fix a, Minimize with Respect to b

Now treat a as fixed. Let $u_i = y_i - ax_i$. This simplifies the expression:

Task 2

Rewrite D in terms of u_i and b. Expand the square and simplify to express D as a quadratic in b.

Task 3

Minimize this quadratic expression by taking its derivative with respect to b, setting it to zero, and solving for b.

Task 4

Substitute $u_i = y_i - ax_i$ back into your expression for b. Then simplify the result using mean notation (e.g., $\bar{x}$, $\bar{y}$).

3. Substitute b Back into the Error Function

You now have b as a function of a. Plug it back into the error function D and simplify.

Task 5

Substitute $b = \bar{y} - a\bar{x}$ into the expression for D, and show that it simplifies to:

$$D = \sum_{i=1}^{n} [(y_i - \bar{y}) - a(x_i - \bar{x})]^2$$

Why is centering around the means useful?

4. Minimize with Respect to a

Now treat D as a function of a. Expand the square and identify this as a quadratic in a.

Task 6

Expand the squared expression and write D as a quadratic in a. Identify the constant, linear, and quadratic terms.

Task 7

Take the derivative of D with respect to a, set it to zero, and solve for a. Express your result using summation notation.

5. Final Expressions

Now that you've derived both expressions, summarize them.

Task 8

Write down the final expressions you derived for a and b. Interpret each term. What do the numerator and denominator of a represent?