Milestone 7

STA 2101: Statistics & Probability

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh (ULAB)

Prepared by: M.H. Ansha Hossain
Student ID: 242016008
Section: 04
Date: December 18, 2025

A. Introduction

The previous milestones explored single-variable distributions. This milestone introduces Simple Linear Regression (SLR), a fundamental method to model and quantify the linear relationship between two continuous variables. The core objective is to calculate the best-fit line parameters manually, adhering to the requirement of avoiding high-level machine learning libraries like scikit-learn for the main tasks.

B. Dataset

The healthcare dataset consists of $N = 55,500$ patient records. For this analysis, we investigate the relationship between the following two variables:

- **Independent Variable (X):** Age
- **Dependent Variable (Y):** Billing Amount

C. Task 1: Data Selection and Initial Visualization

We aim to determine if a patient's Age (X) is a predictor for their Billing Amount (Y).

1. Summary Statistics

Based on our preliminary analysis of the dataset:

- **Mean of X ($\bar{X}$):** 51.54
- **Mean of Y ($\bar{Y}$):** 25,539.32
- **Variance of X (s_x^2):** 384.26
- **Variance of Y (s_y^2):** 201,965,437.04

2. Scatter Plot

Below is the scatter plot visualizing the relationship between Age and Billing Amount.

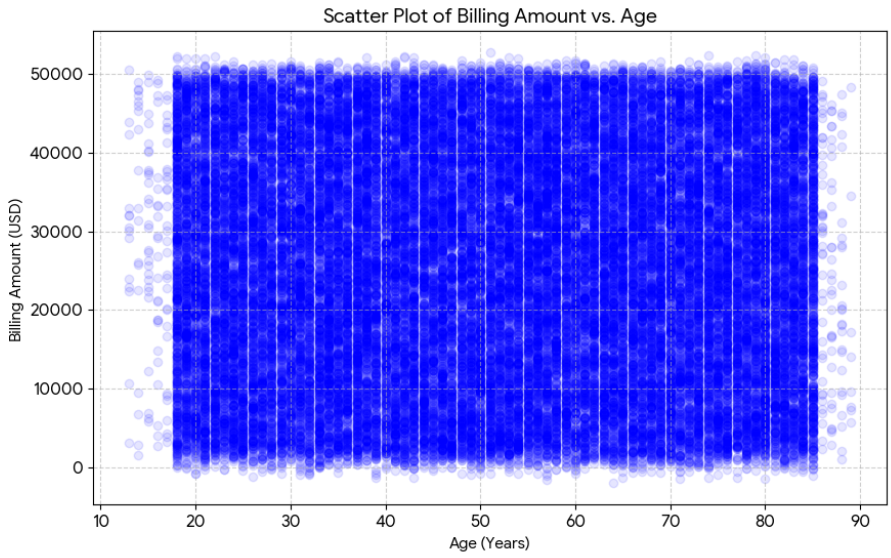

Figure 1: Scatter Plot of Billing Amount vs. Age.

D. Task 2: Manual Calculation of Regression Parameters

1. Slope (β_1)

The slope is calculated using the formula:

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

From our computed deviation scores:

$$\text{Numerator} = -59,245,185.60$$
$$\text{Denominator} = 21,325,834.58$$

$$\beta_1 = \frac{-59,245,185.60}{21,325,834.58} \approx -2.7781$$

2. Intercept (β_0)

The intercept is calculated as:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Substituting our values:

$$\beta_0 = 25,539.32 - (-2.7781 \times 51.54)$$

$$\beta_0 = 25,539.32 - (-143.18)$$

$$\beta_0 \approx 25,682.50$$

3. Estimated Regression Equation

The final linear model is:

$$\hat{Y} = 25,682.50 - 2.78X$$

E. Task 3: Visualization of the Fit

1. Regression Line Plot

The calculated regression line is overlaid on the data below.

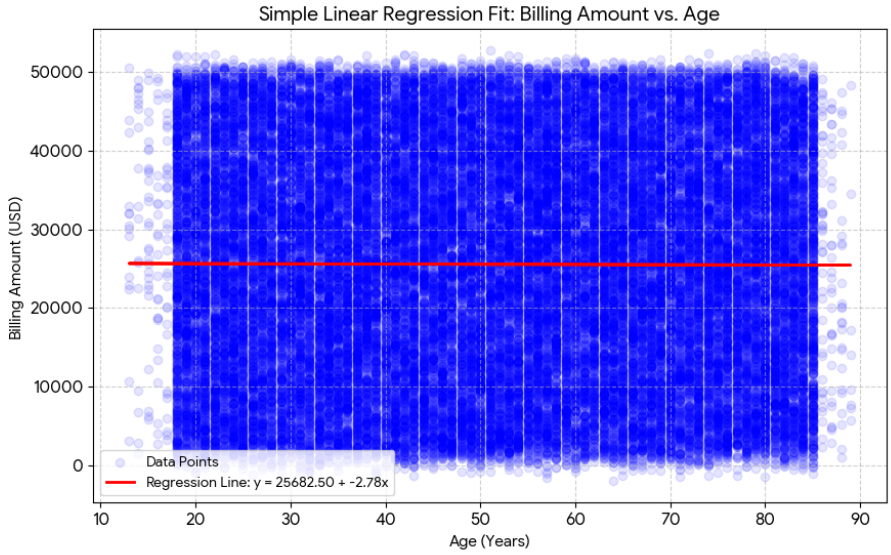

Figure 2: Linear Regression Fit.

2. Interpretation of Slope

The slope $\beta_1 = -2.78$ implies that for every one-year increase in a patient's age, the billing amount decreases by approximately \$2.78. Given the magnitude of the billing amounts, this change is negligible.

F. Task 4: Strength of Relationship

1. Pearson Correlation Coefficient (r)

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Using our calculated components:

$$r \approx -0.0038$$

2. Coefficient of Determination (R^2)

$$R^2 = r^2 = (-0.0038)^2 \approx 0.0000$$

3. Relationship Assessment

- **Direction & Strength:** The correlation $r = -0.0038$ is effectively zero, indicating no linear relationship.
- **Variance Explained:** An R^2 of 0.0000 means that 0% of the variation in Billing Amount is explained by Age.

G. Task 5: Reflection

Visual Fit Assessment: The regression line does not provide a good fit for the data points. The data is scattered uniformly, and the regression line is essentially horizontal.

R^2 **Support:** The R^2 value of 0.0000 quantitatively confirms the visual assessment, proving that Age has no predictive power for Billing Amount in this dataset.

Real-World Scenario: While Age was not a predictor here, this linear regression model could be useful in other healthcare contexts, such as predicting the cost of treatment based on the duration of a hospital stay (Length of Stay), where a stronger positive correlation would typically be expected.