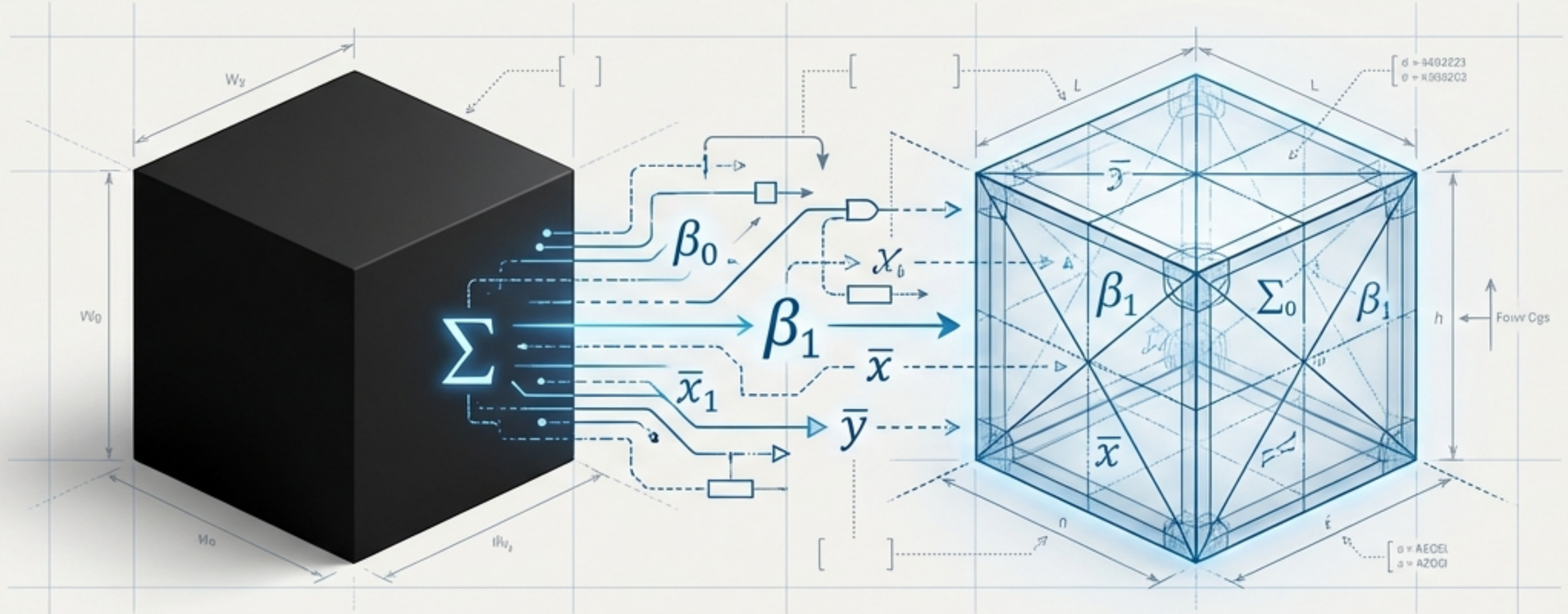
Beyond the Black Box: Building a Model from First Principles



In machine learning, we often use libraries that give us answers without revealing the process. They are 'black boxes.' This project is your opportunity to look inside.

Core Objective
Your mission is to build a simple linear regression model from the ground up. You will use fundamental mathematical formulas to compute every part of the model, transforming it from a mystery into a tool you truly understand.

Your Toolkit: The Data and The Model

The Data

Structure: One input variable (x) and one output variable (y).

Scale: A minimum of 30 observations.

Source: Must be real-world data. Cite your source with a link (e.g., Kaggle, World Bank, public datasets).

Examples:
Study Hours (x) → Exam Score (y)
Years of Experience (x) → Salary (y)
Temperature (x) → Electricity Usage (y)

The Model Blueprint

$$y = \beta_0 + \beta_1 x$$

β_1: The slope of the line.

β_0: The y-intercept.

The Execution, Part 1: Manually Forging the Parameters

This is where you build the model. You will compute the slope (β_1) and intercept (β_0) directly from the data using the following foundational formulas.

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Formula for Slope (β_1)

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Formula for Intercept (β_0)

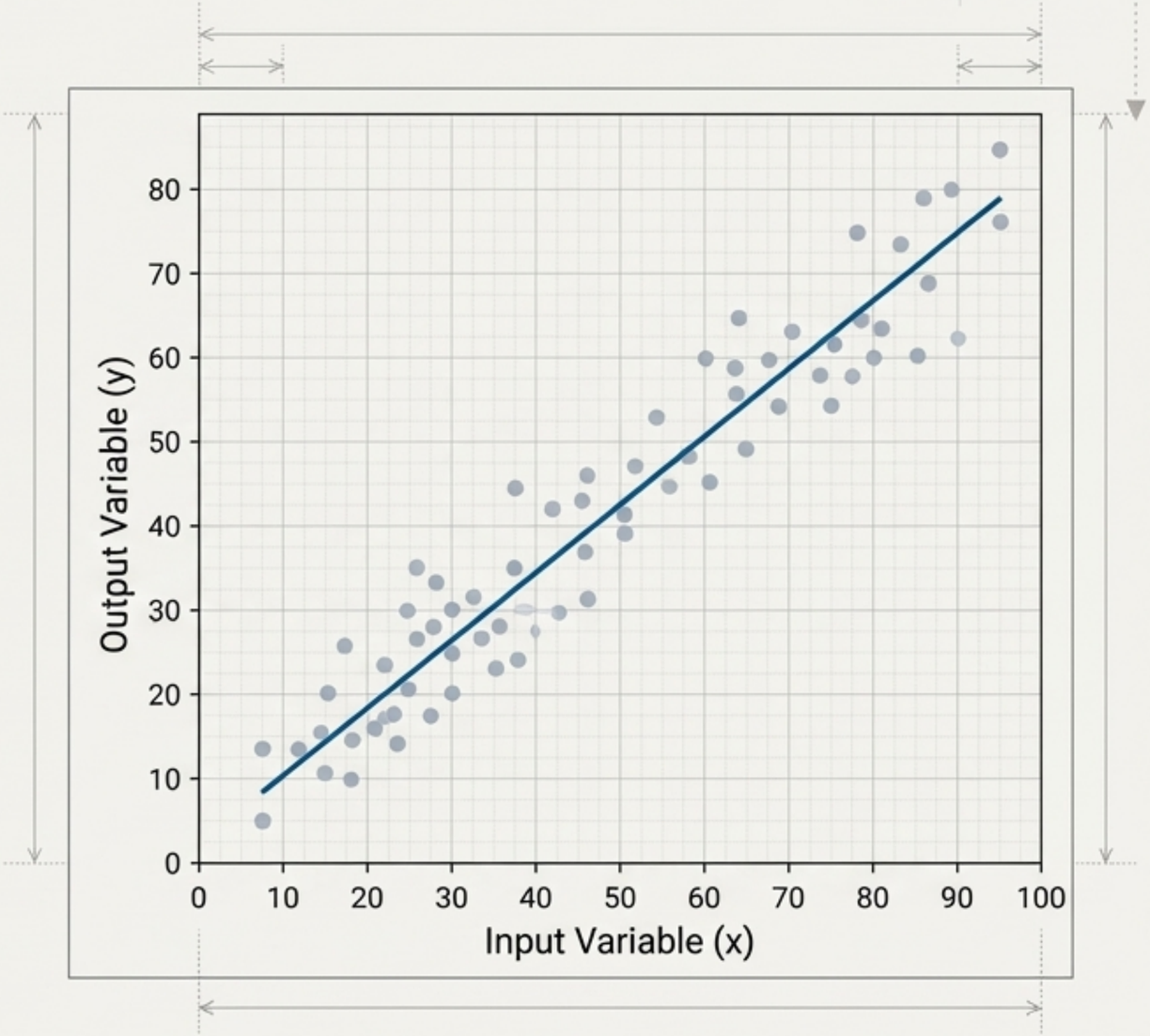
No Black Boxes Allowed

You may **NOT** use regression libraries like `sklearn`, `statsmodels`, or `np.polyfit` for these calculations. The goal is to compute them manually. Basic NumPy operations for sums and means are permitted.

The Execution, Part 2: Visualizing Your Model

Once calculated, a model must be seen. Create a single, clear plot that tells the story of your data and your model's fit.

1. **Scatter Plot**: Plot all your raw data points (x, y).
2. **Regression Line**: Overlay the regression line ($y = \beta_0 + \beta_1 x$) you calculated on the same plot.
3. **Clarity**: Both the X and Y axes must be clearly and accurately labeled.

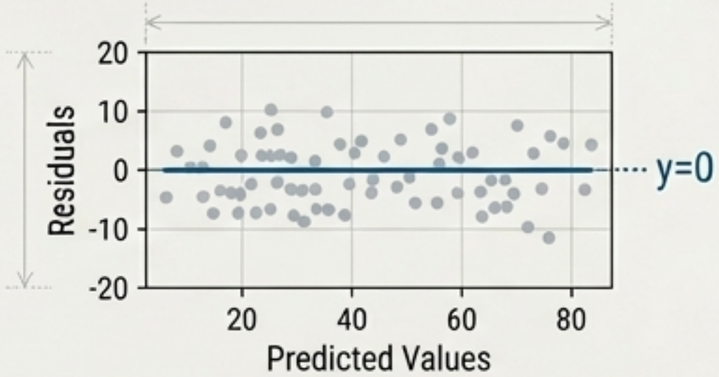


The Execution, Part 3: Analyzing the Error

A model is only as good as its predictions. Now, you will quantify your model's accuracy and diagnose its fit.

Step 1: Compute Mean Squared Error (MSE)

Calculate the MSE to get a single metric for the average squared difference between the observed and predicted values.



Residual Plot

Step 2: Create a Residual Plot

Plot the residuals (the errors) to visually inspect for patterns. This helps determine if a linear model is appropriate.

Step 3: Provide Commentary

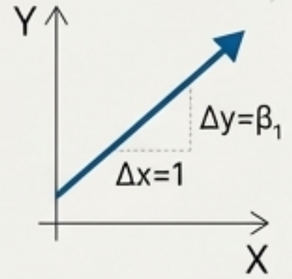
Based on your MSE and residual plot, briefly comment on whether you believe a linear model is a reasonable choice for your dataset.

The Insight: What Does Your Model Actually Mean?

The final step is to translate the numbers into a narrative. Your report must explain the practical meaning of your findings. Address these three points:

1.

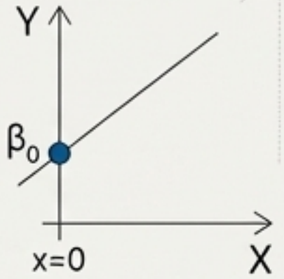
The Slope (β_1)



What is the real-world meaning of your calculated slope? For every one-unit increase in x, what happens to y?

2.

The Intercept (β_0)



What does the intercept represent in the context of your data? Is this value meaningful or is it an artifact of the model? (e.g., Does a house with zero size have a price?)

3.

Model Limitations

What external factors or underlying assumptions could limit or break your model's predictive power?

The Submission: Your Final Report

Consolidate your work into a single, professional document.

Format: One PDF file, maximum 3 pages.

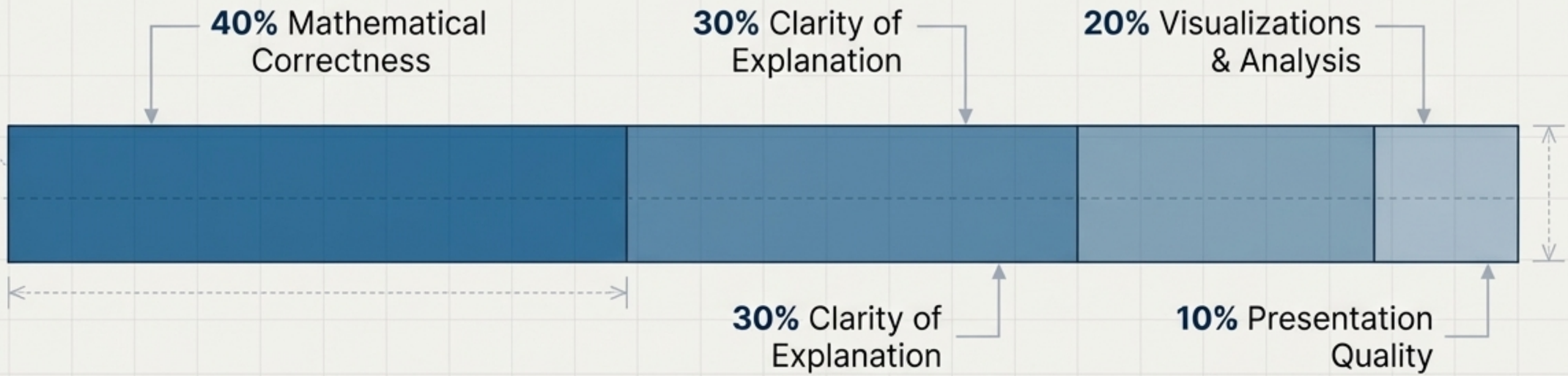
Submission Checklist

Your PDF must include the following sections in a clear, logical order.

- [ ] Dataset Description & Source Link
- [ ] The Mathematical Formulas Used
- [ ] Your Final Calculated Values for β_1 and β_0
- [ ] The Scatter Plot with Regression Line
- [ ] The Residual Plot and MSE Calculation
- [ ] Your Full Interpretation (Slope, Intercept, Limitations)

How Your Work Will Be Evaluated

Your project will be assessed on both your technical execution and your ability to communicate your findings. The grading is weighted as follows:

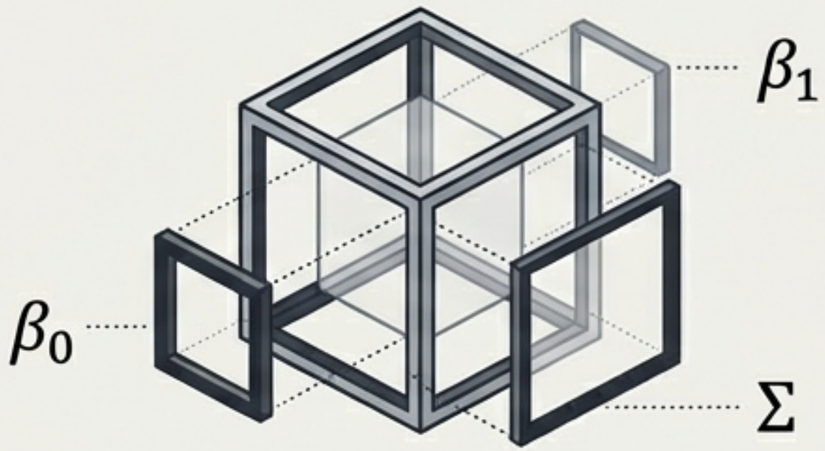


40% - Mathematical Correctness: Accurate calculation of β_0, β_1, and MSE.
30% - Clarity of Explanation: Your interpretation of the model and its components is clear, insightful, and well-written.
20% - Visualizations & Analysis: Plots are clear, well-labeled, and your error analysis is sound.
10% - Presentation Quality: The final PDF is professional, well-organized, and easy to read.

Bonus Challenge: Verify with the ‘Black Box’

Now that you’ve built the engine by hand, let’s see how it compares to the automated tool.

Your Manual Model



Library Model

The Task (+10%)

1. Use a library like `sklearn` to build a linear regression model on the same data.
2. Compare the library's results for β_1 and β_0 to your manually calculated values.
3. Include a brief explanation in your report for why the results match. This closes the loop on understanding the model's inner workings.