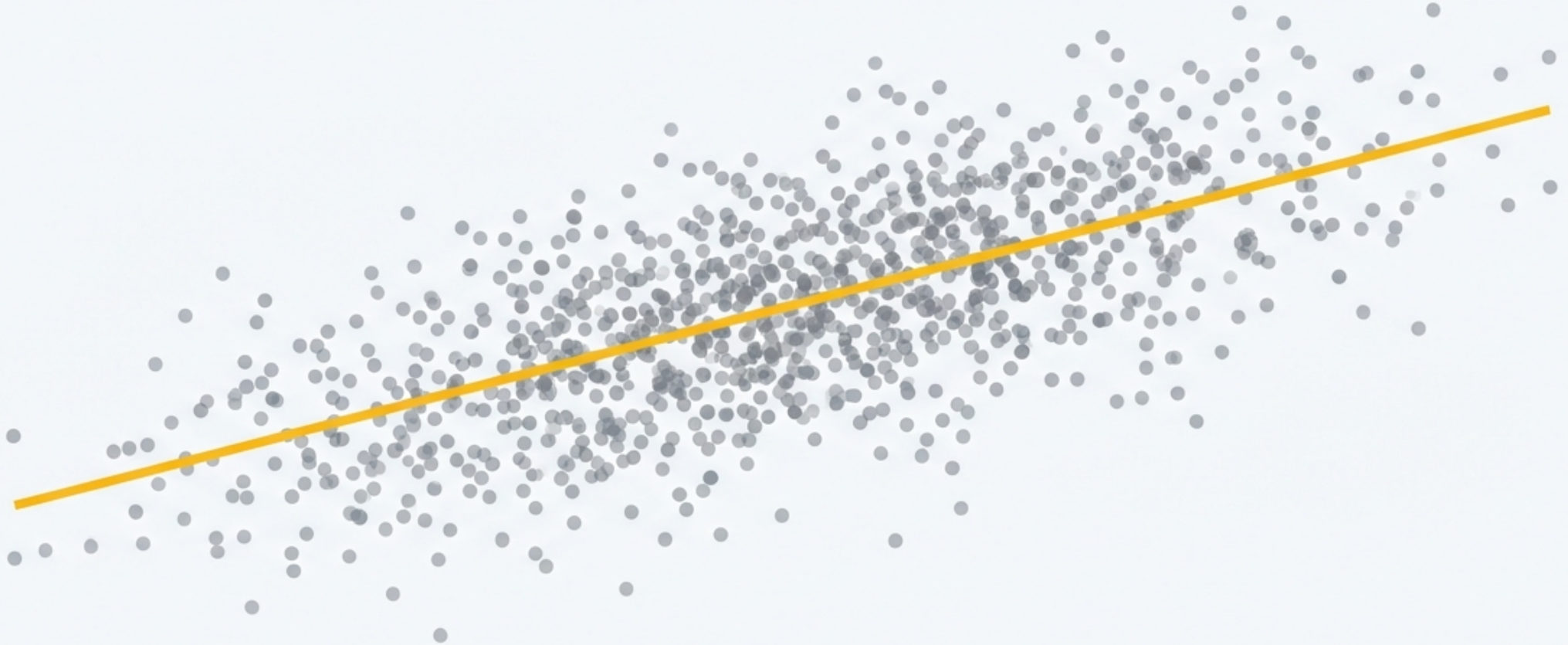
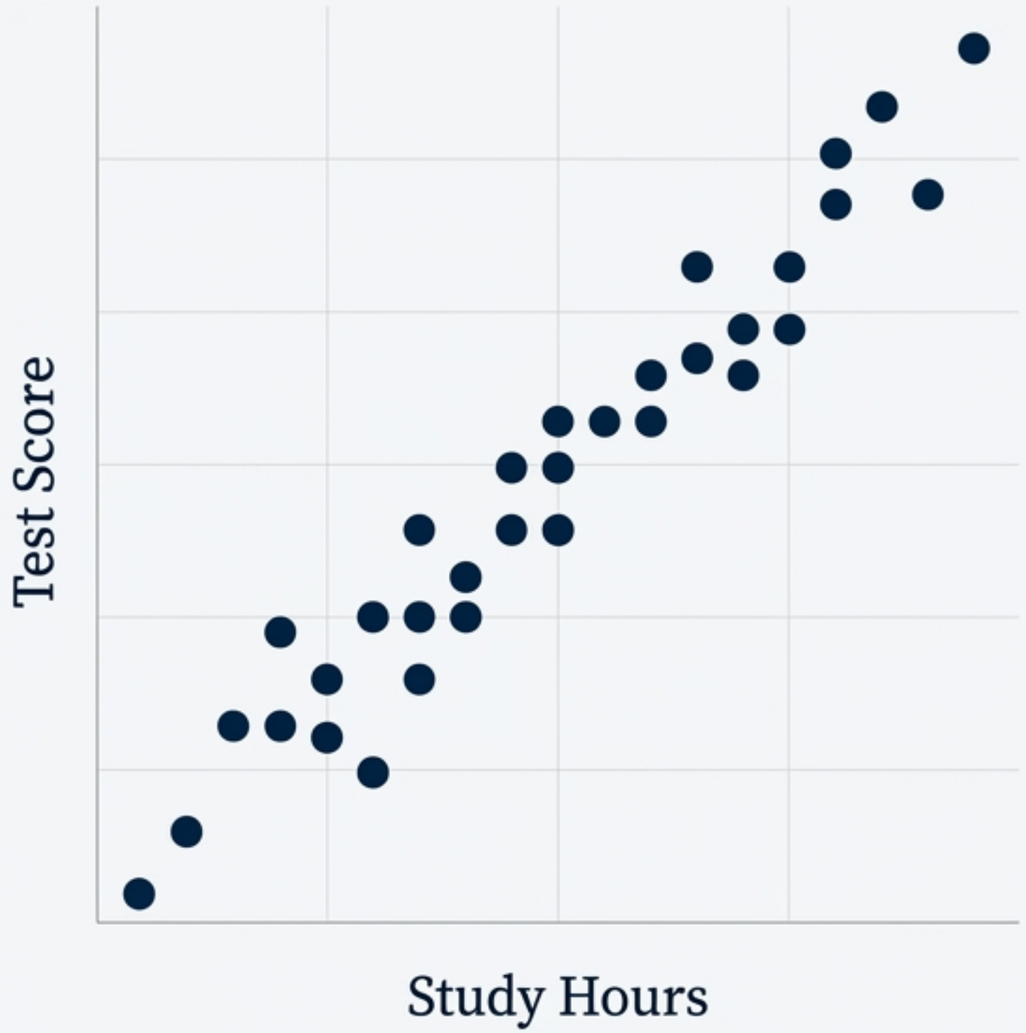
The Line of Best Fit: An Intuitive Guide to Least Squares Regression

Mastering the technique that turns data points into powerful insights.

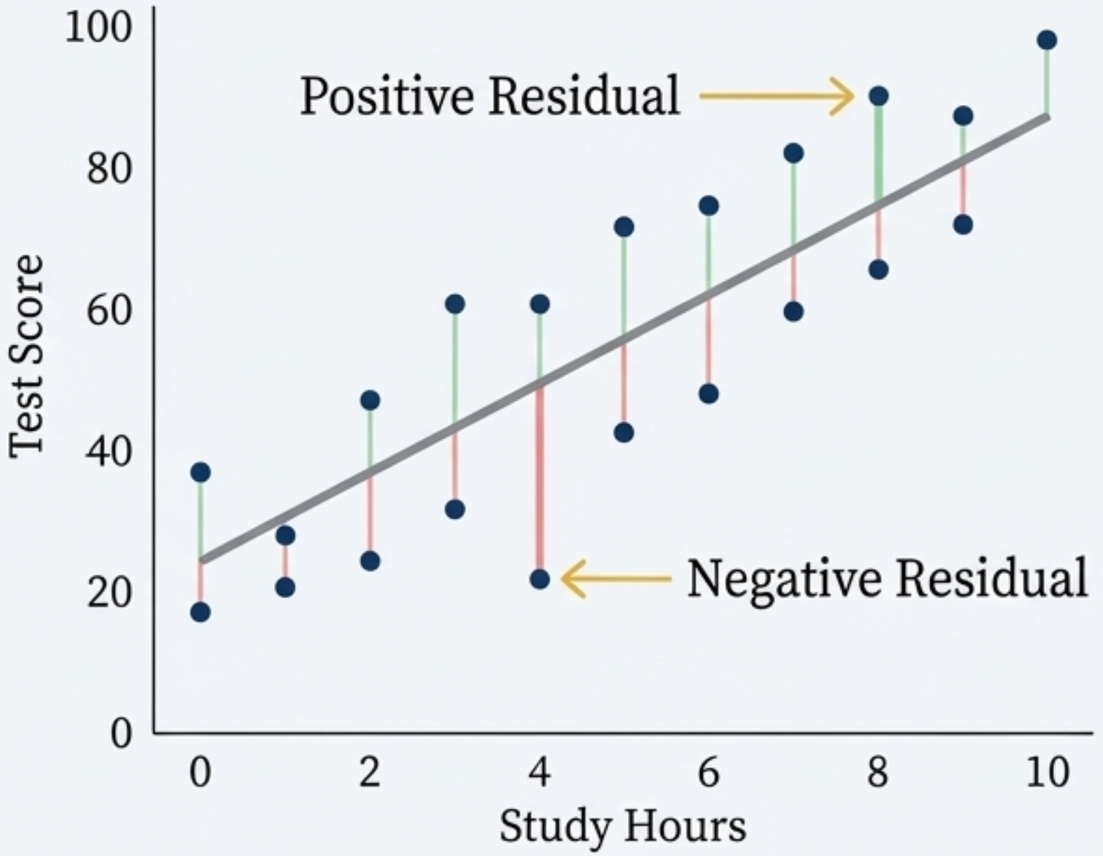




We see a pattern. How do we define it?

- How can we summarize this relationship with a single straight line?
- Any line drawn by hand is subjective. How do we find the one, mathematically “best” line?
- This “line of best fit” is our goal. It allows us to quantify the relationship, understand its strength, and make predictions.

The best line is the one with the smallest error.



Introduce Residuals (or Errors):

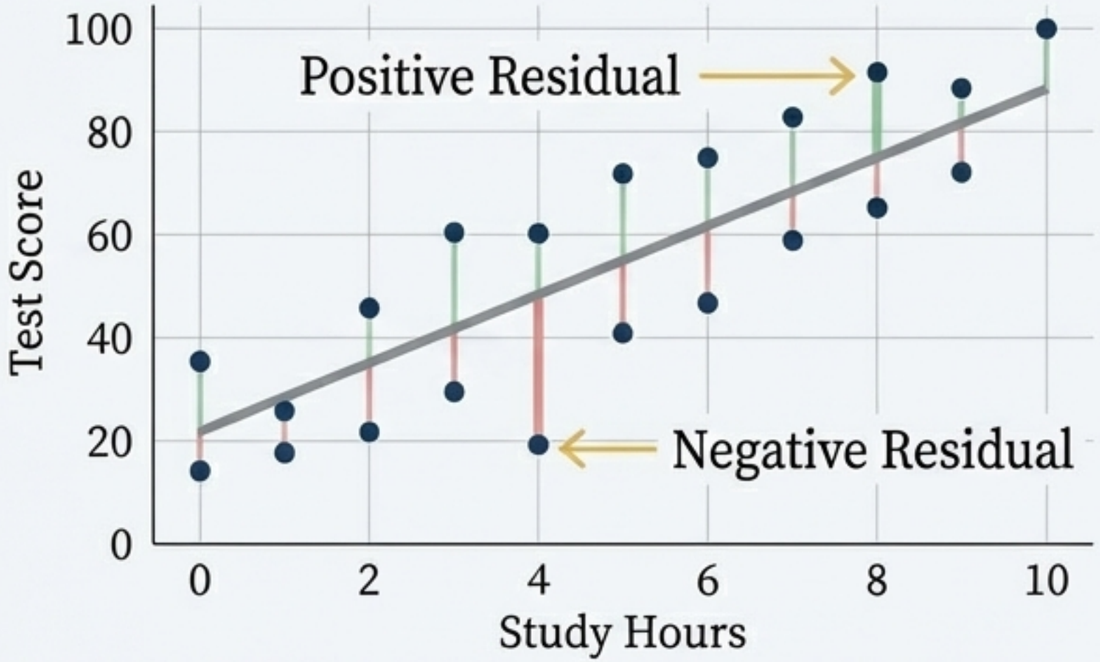
The residual is the vertical distance between an observed data point (y) and the line's predicted value ($\hat{y}$). It's a measure of the model's error for that specific point.

Residual = Observed value – Predicted value

$$= y - \hat{y}$$

Why Squaring the Errors is the Key

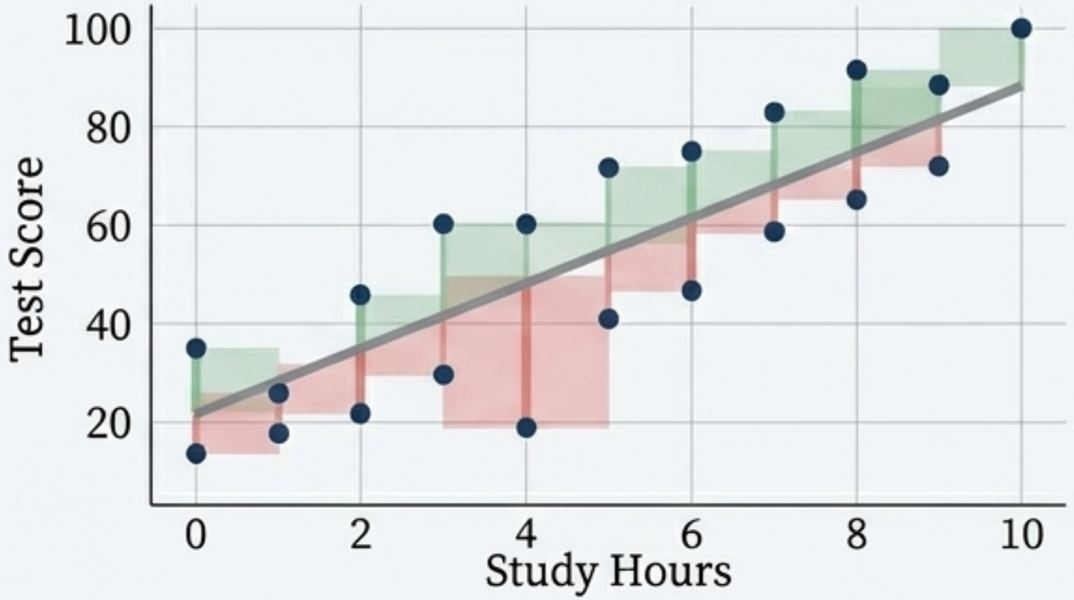
The Problem



If we simply add up the residuals, positive and negative errors cancel each other out. A line can have a total error of zero but be a terrible fit.

$$\sum(y - \hat{y}) \approx 0$$

The Solution

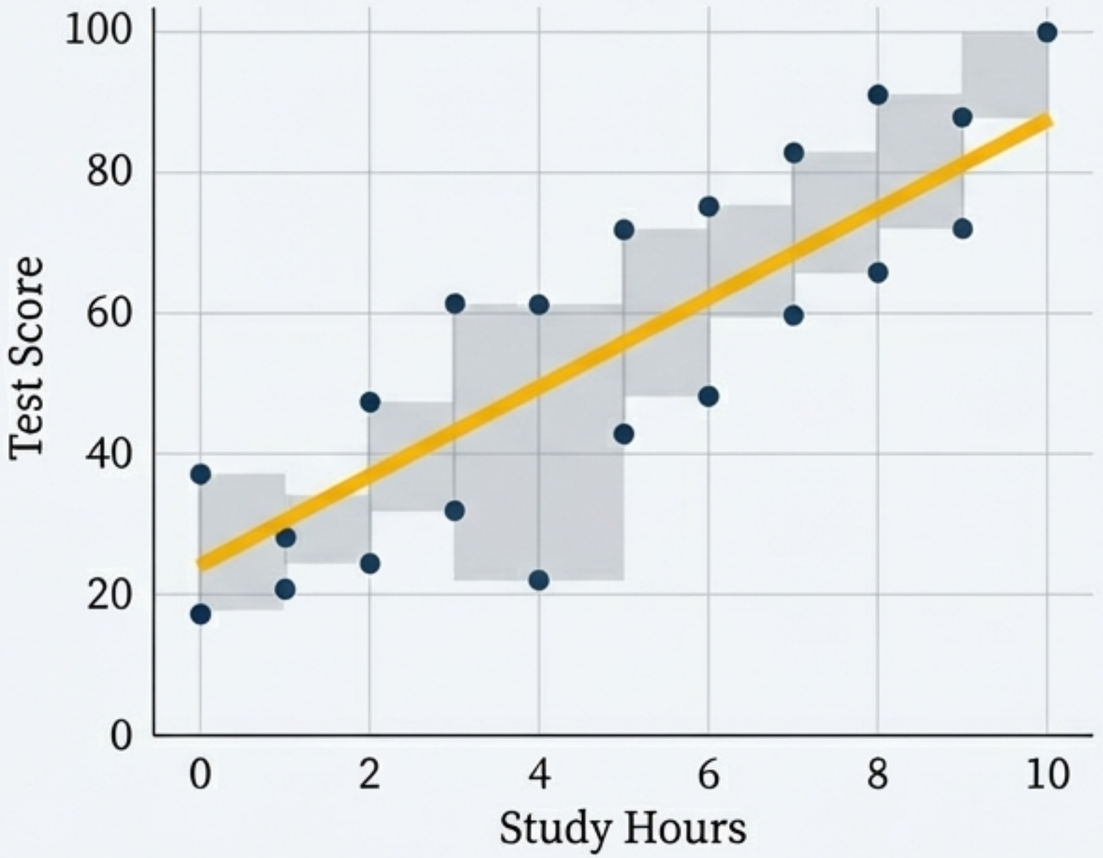


The Insight: By squaring each residual, we achieve two things: 1) every error becomes positive, eliminating cancellation, and 2) larger errors are penalized more heavily.

The Goal: The best line is the one that minimizes the **Sum of Squared Errors (SSE)**.

$$\text{SSE} = \sum(y - \hat{y})^2$$

The unique line that minimizes the sum of squared errors.



Definition: The **Least Squares Regression Line** is the one straight line that fits the data better than any other by minimizing the sum of the squared errors.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{\beta}_1$ **(Slope):** The average change in y for a one-unit increase in x.
$\hat{\beta}_0$ **(Y-Intercept):** The predicted value of y when x is zero.

The Mechanism: Calculus provides the method to find the exact values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that guarantee the minimum possible SSE for any given dataset.

The Mechanics: Calculating the Slope ($\hat{\beta}_1$)

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

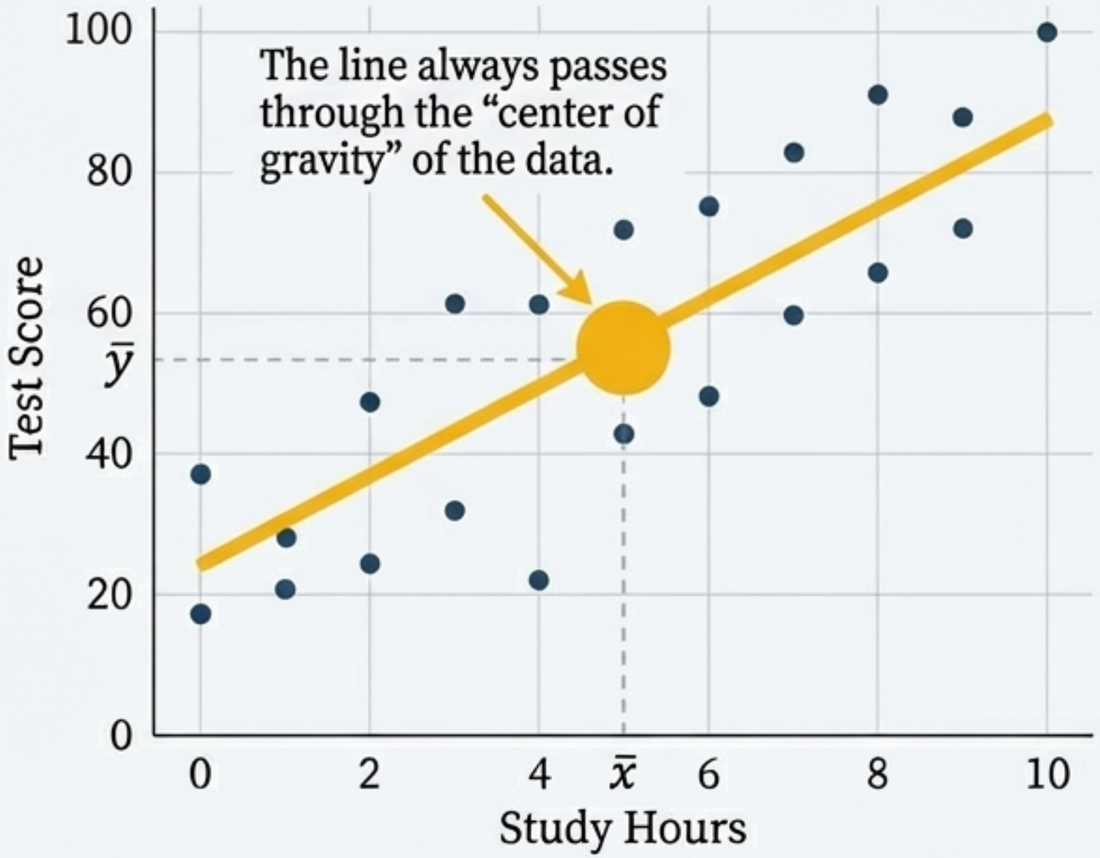
Sum of cross-products of deviations. This measures how `x` and `y` vary together (their co-movement).

$$SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

Sum of squared deviations for `x`. This measures the total variability in `x`.

Conceptual Explanation: The slope is essentially a ratio: it compares how much `x` and `y` move together against how much `x` moves on its own.

The Mechanics: Calculating the Intercept ($\hat{\beta}_0$)



$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Intuition: This formula guarantees that the regression line passes through the the regression line passes through the central point of the data: the point of averages, $(\bar{x}, \bar{y})$.

Walkthrough: Car Age vs. Value

Sample Data (n=10)

x (Age in years)	y (Value in $1000s)
2	28.7
3	24.8
3	26.0
3	30.5
4	23.8
4	24.6
5	23.8
5	20.4
5	21.6
6	22.1

1. Summary Statistics

$\sum x = 40 \qquad \sum x^2 = 174 \qquad \bar{x} = 4$

$\sum y = 246.3 \qquad \sum xy = 956.5 \qquad \bar{y} = 24.63$

2. Calculate Sums of Squares

$SS_{xx} = \sum x^2 - (\sum x)^2/n = 174 - (40)^2/10 = 14$

$SS_{xy} = \sum xy - (\sum x)(\sum y)/n = 956.5 - (40)(246.3)/10 = -28.7$

3. Calculate Slope ($\hat{\beta}_1$)

$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}/SS_{xx} = \frac{-28.7}{14} = -2.05$

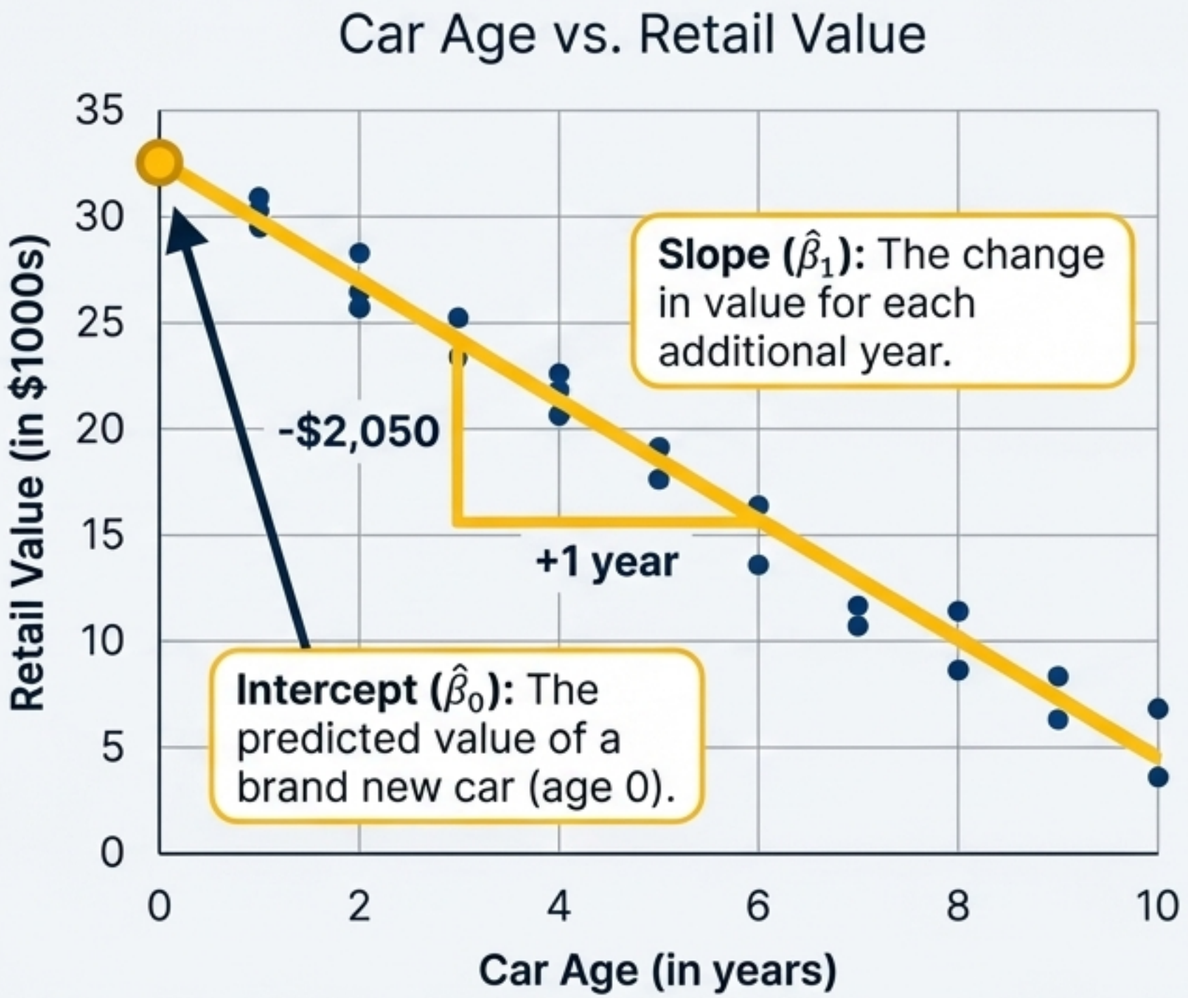
4. Calculate Intercept ($\hat{\beta}_0$)

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 24.63 - (-2.05)(4) = 32.83$

Final Result: Regression Equation

Predicted Value = 32.83 – 2.05 × Age

Turning Numbers into Narrative



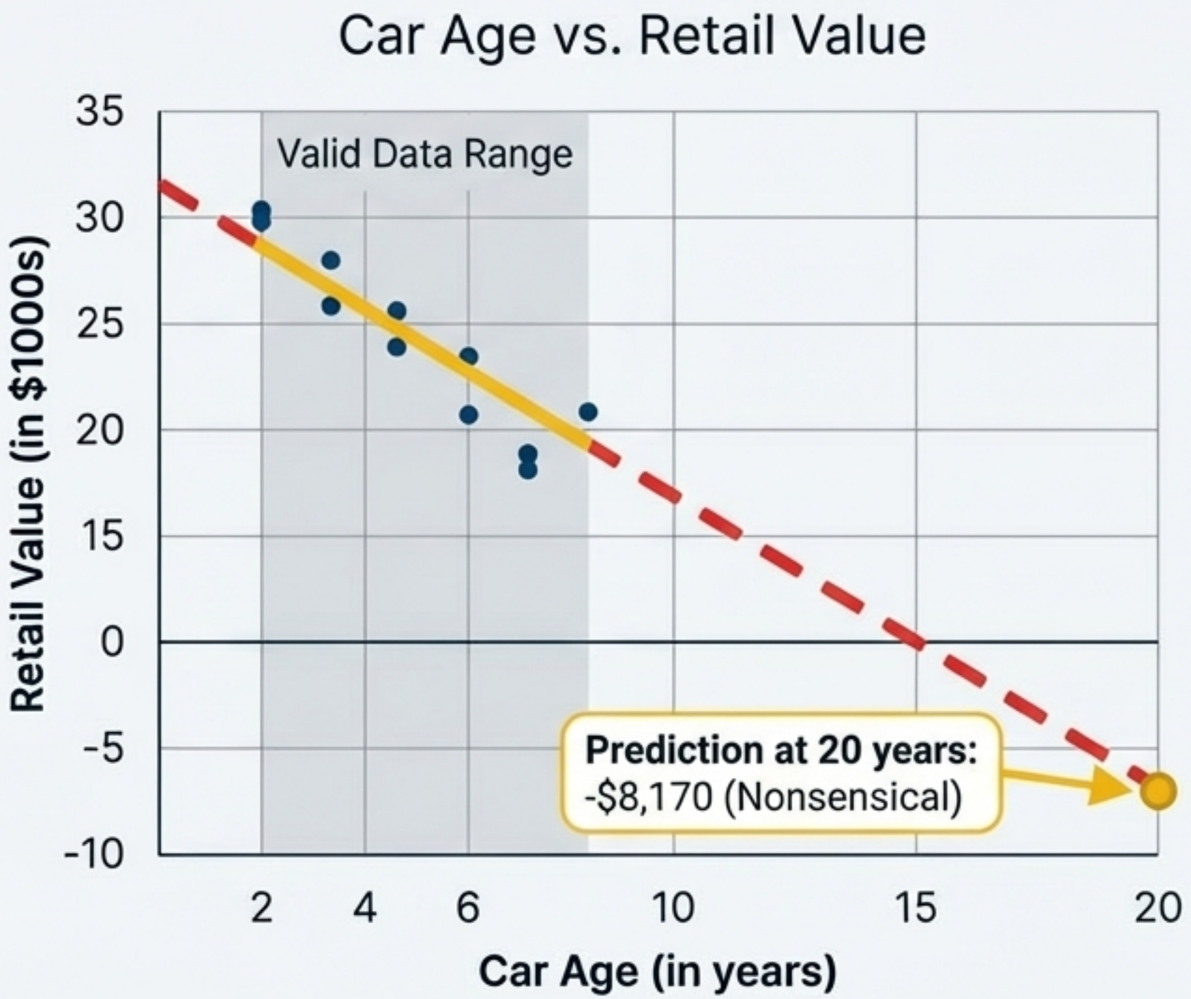
Slope ($\hat{\beta}_1 = -2.05$):

For each additional year of age, the average value of this make and model vehicle is predicted to decrease by about $2,050.

Intercept ($\hat{\beta}_0 = 32.83$):

A brand new (0-year-old) car of this model is predicted to have a retail value of $32,830. This is the value of $\hat{y}$ when x=0.

A Model is Only Valid Within the Range of Your Data



Definition

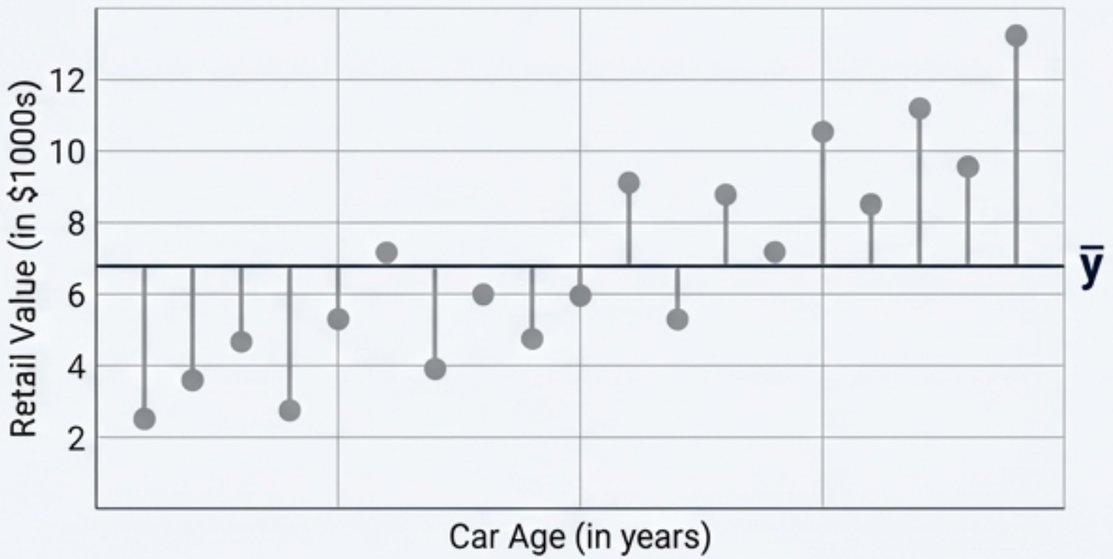
Applying the regression equation to a value of x outside the range of x-values in the original data is called **extrapolation**. It is an invalid use of the regression equation and should be avoided.

The Danger Illustrated

Using the equation to predict the value of a 20-year-old car yields $\hat{y} = -2.05(20) + 32.83 = -8.17$, or -$8,170. This makes no sense and shows the error of applying the model to a value of x not in the original data range.

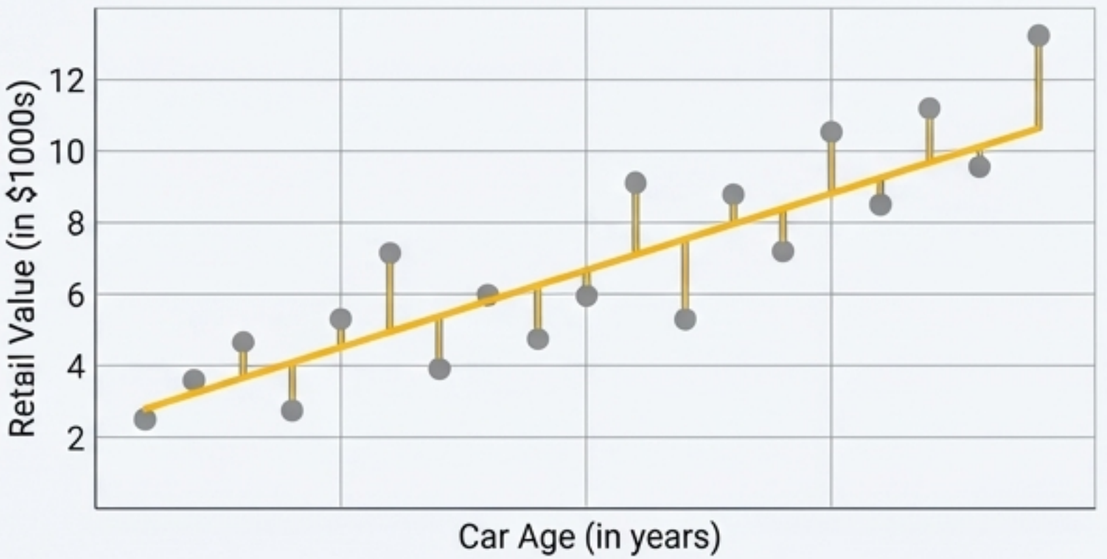
How Good is the Fit? Measuring Explanatory Power with R^2

Total Variation (SST)



$$SST = \sum(y - \bar{y})^2$$

Unexplained Variation / Residuals (SSE)



$$SSE = \sum(y - \hat{y})^2$$

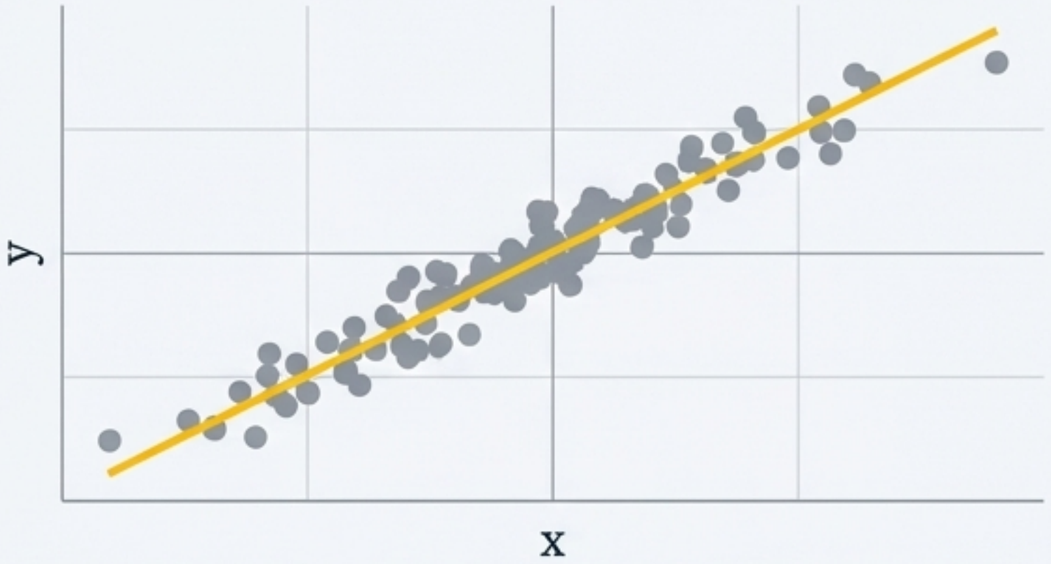
The Core Question:
How much of the variation in our `y` variable (car value) is actually explained by our `x` variable (car age)?

The Answer: R-squared (R^2)
R-squared is the proportion of the total variance in the dependent variable that is predictable from the independent variable. It ranges from 0 (the model explains no variance) to 1 (the model explains all variance).

$$R^2 = 1 - (SSE / SST)$$

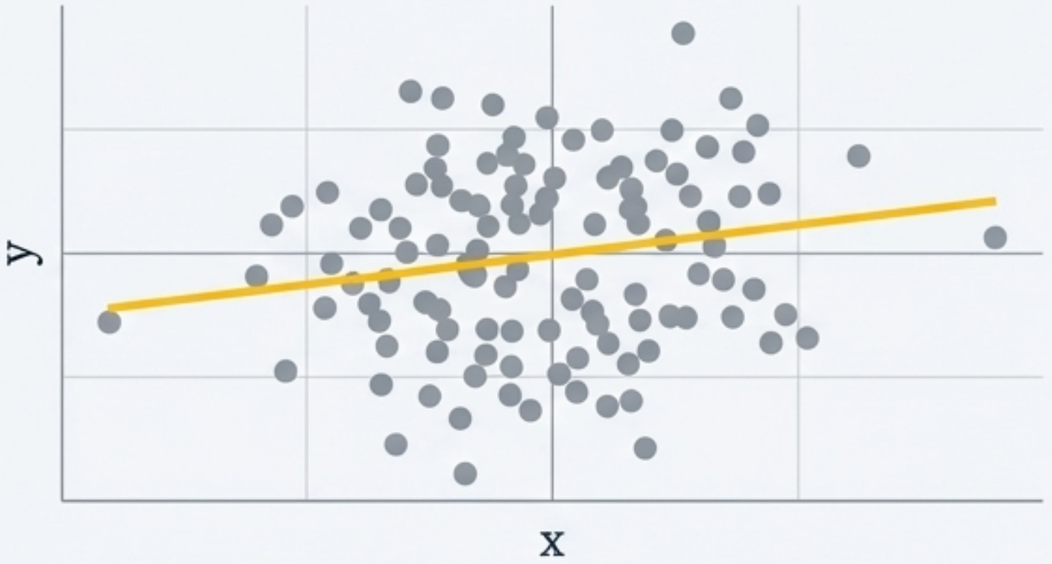
Visualizing R-squared

High R^2 (~0.90)



The data points are tightly clustered around the line. The model is a good fit and explains most of the variability in `$y$`.

Low R^2 (~0.20)



The data points are widely scattered. The model is a poor fit and explains very little of the variability in `$y$`.

Our Example's Result:

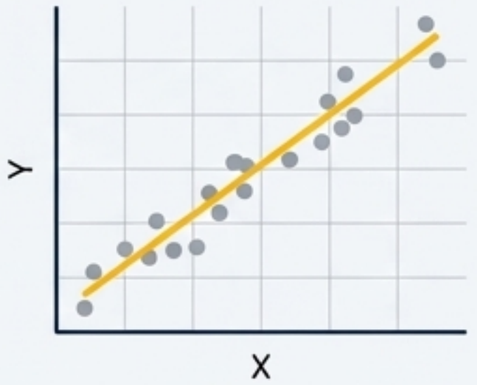
For the car data, the correlation $r = -0.819$, so $R^2 = (-0.819)^2 \approx 0.67$.

Interpretation: 67% of the variation in the used car prices in our sample can be explained by the car's age.

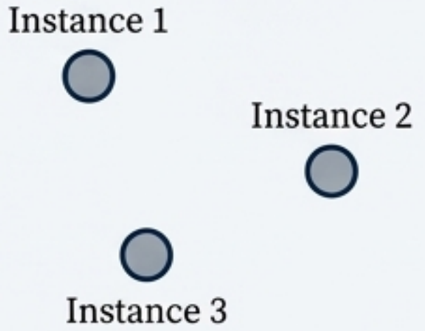
The Rules of the Road: Key Model Assumptions

For the interpretations of the coefficients and the confidence intervals to be valid, the linear regression model relies on several key assumptions about the data.

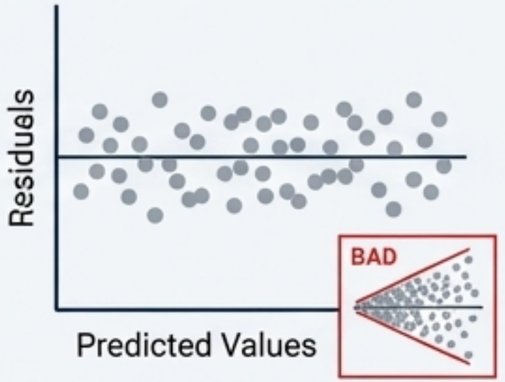
Linearity: The underlying relationship between X and Y is linear.



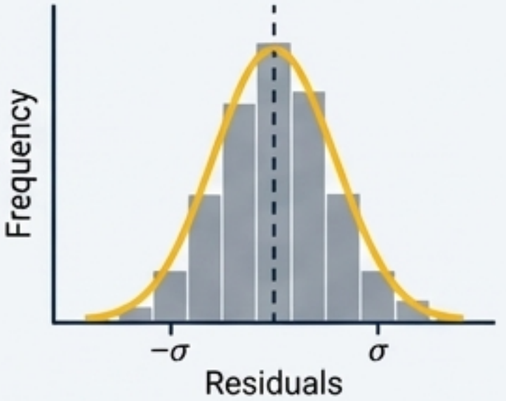
Independence: Each instance (and its error) is independent of any other instance.



Homoscedasticity: The variance of the error terms is constant across all values of `x`.



Normality: The error terms are normally distributed.



From Data Chaos to Predictive Clarity

The Principle

Least squares regression finds the unique line that **minimizes the sum of squared residuals**, providing the single best mathematical fit to the data.

The Interpretation

The **slope** ($\hat{\beta}_1$) quantifies the average change in `y` for a one-unit change in `x`, while the **intercept** ($\hat{\beta}_0$) provides a baseline value.

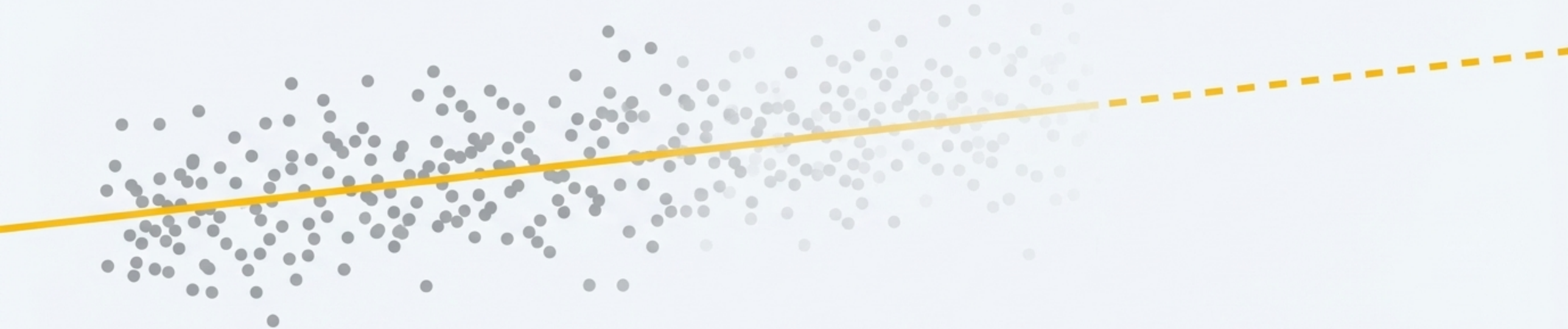
The Evaluation

R-squared (R^2) measures the proportion of the variance in `y` that is explained by the model, indicating the goodness-of-fit.

The Responsibility

Always interpret results **in context**, be wary of **extrapolation** beyond your data's range, and verify model **assumptions** for valid conclusions.

A Foundation for Insight



Least Squares Regression is more than a formula; it's a foundational tool for moving from simple observation to quantifiable understanding. By modeling relationships, we can describe the world more accurately, test hypotheses rigorously, and make data-driven predictions.