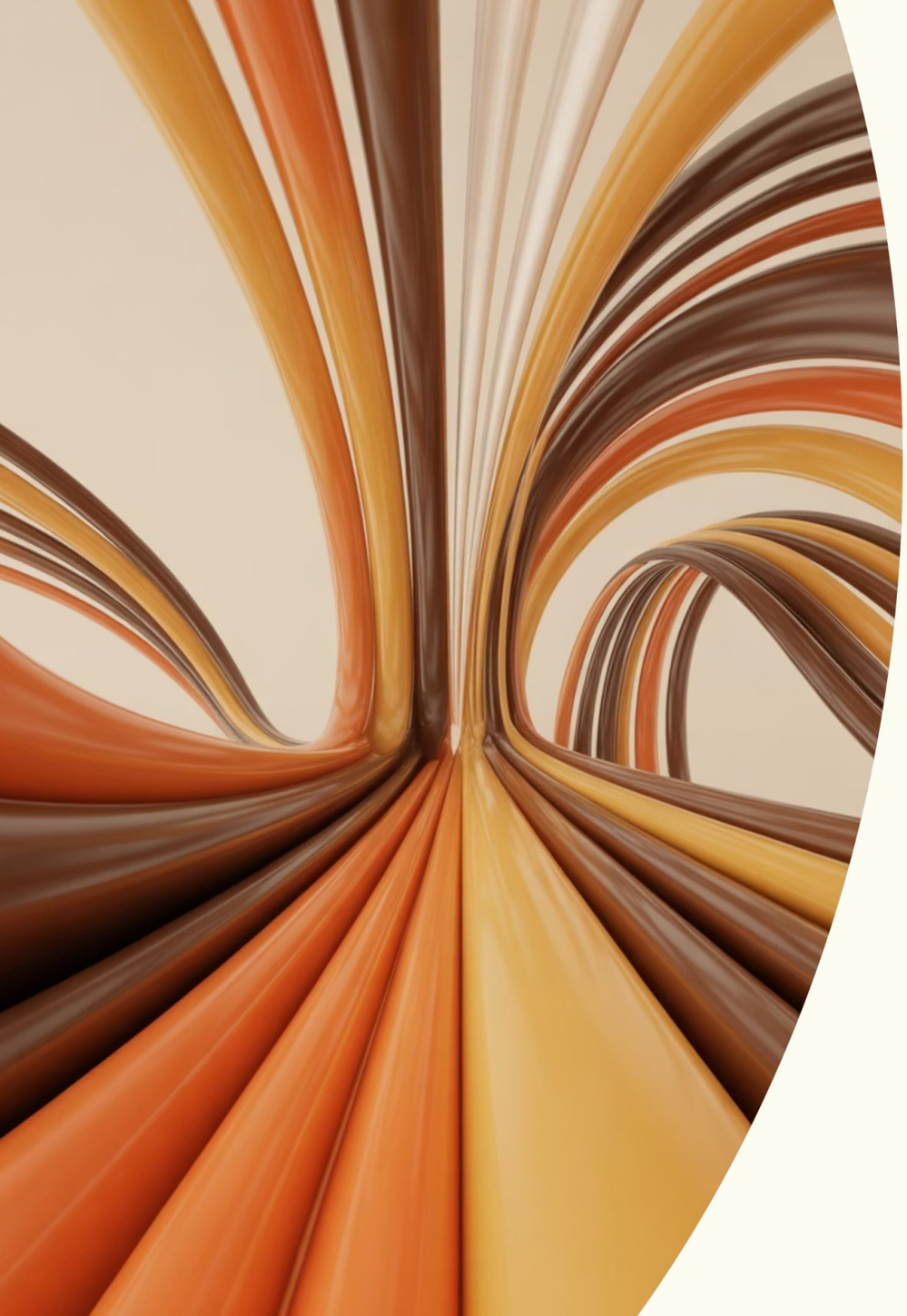


What is Regression and the Best Fit Line

A Beginner-Friendly Introduction to Predicting Continuous Values

Why Do We Need Regression?

Regression helps us predict **numbers** based on patterns in data. It's everywhere in daily life, making predictions about real values we care about.

Home Prices

Predict how much a house will sell for based on size, location, and features

Weather Forecasting

Estimate tomorrow's temperature using historical patterns and current conditions

Student Performance

Predict exam scores based on study hours, attendance, and past grades

Sales Projections

Forecast next month's revenue using trends, seasons, and market data

What Does Regression Mean?

The Core Idea

Regression is a type of machine learning that predicts **continuous numeric outputs**. Unlike classification, which picks categories, regression gives you actual numbers.
Think of it as drawing a line through data points to capture the underlying pattern, then using that line to make predictions.

Task Type	Output	Example
Regression	Number	$325,000
Classification	Category	"Spam" or "Not Spam"

Our Running Example: Study Hours and Exam Marks

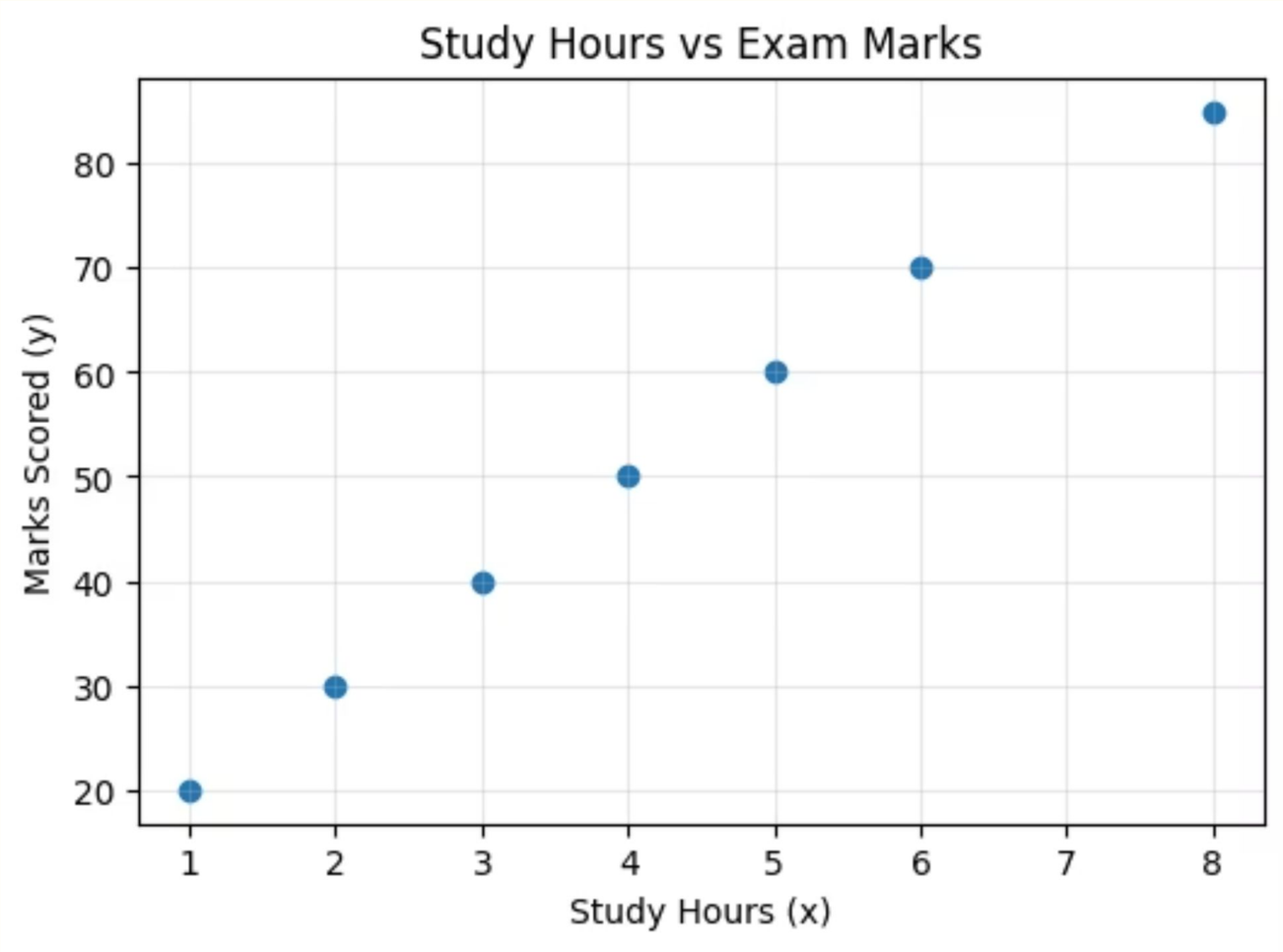
Let's learn regression using a simple, relatable dataset. We'll use this throughout the lesson to see how predictions work.

Study Hours (x)	Marks Scored (y)
1	20
2	30
3	40
4	50
5	60
6	70
8	85

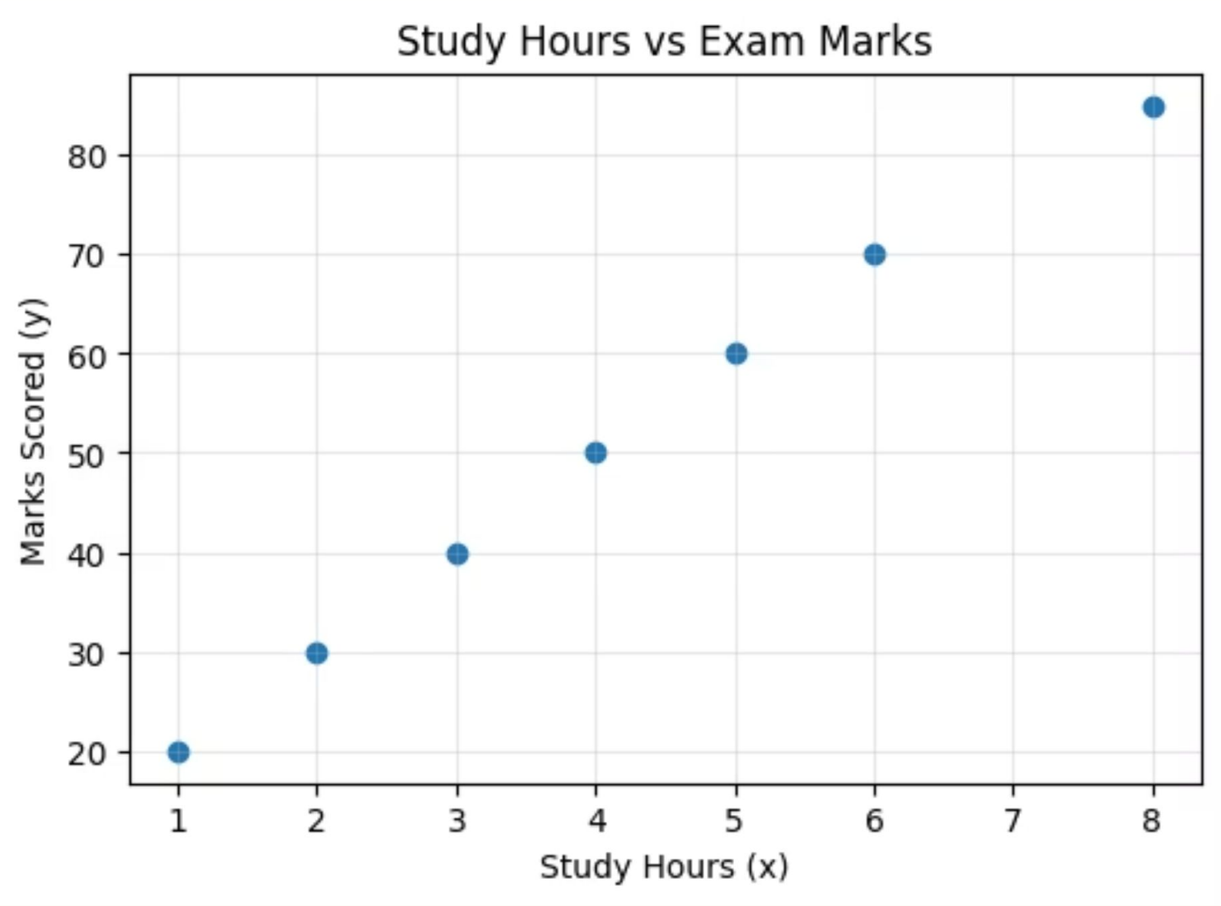
Note: Each row represents one student. The more hours studied, the higher the marks, but we need a model to predict scores for *any* number of study hours.

Visualizing the Data: Scatter Plot

Each dot represents one student. The x-axis shows study hours, and the y-axis shows marks scored. This visualization helps us see patterns in the data at a glance.



Understanding the Trend



What We Observe

The points move **upward** as we go from left to right. This tells us there's a positive relationship between study hours and marks.

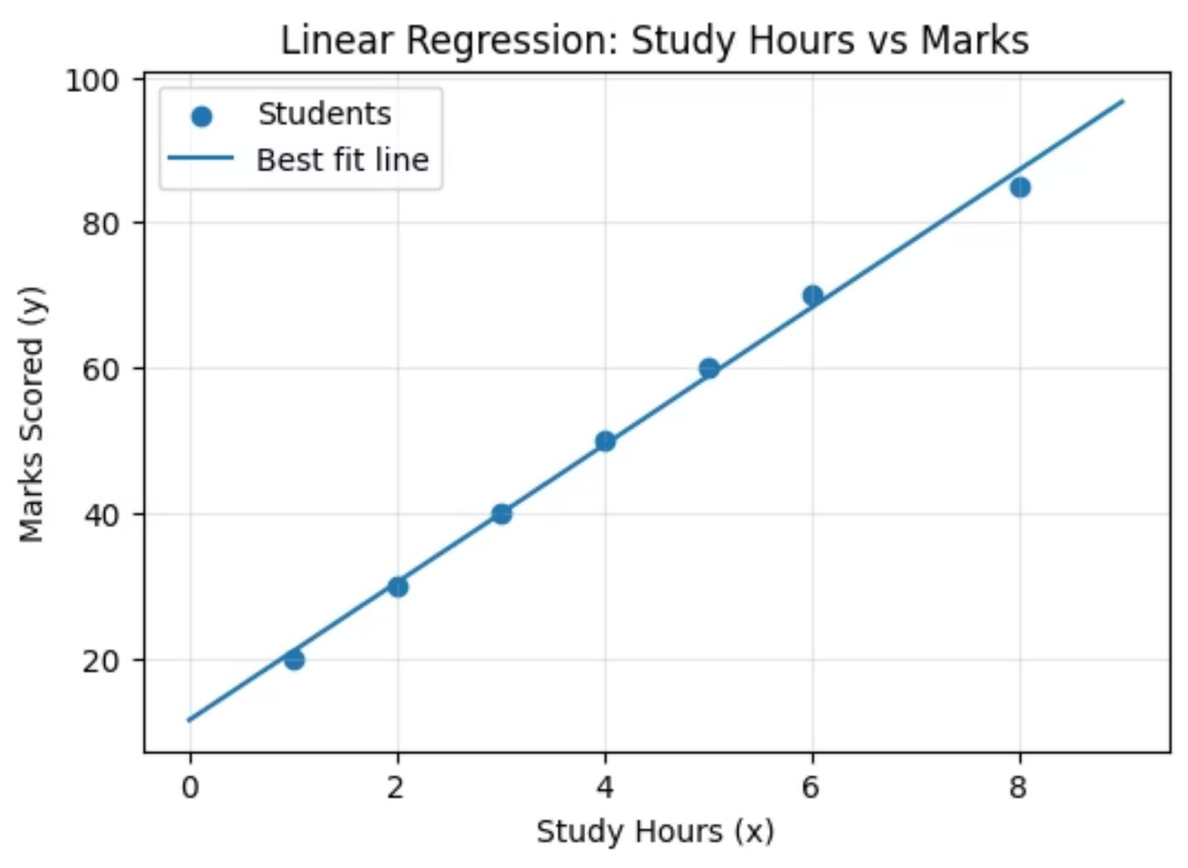
More study time generally leads to higher scores.

Regression helps us capture this trend mathematically.

Introducing the Best Fit Line

> "This line summarizes the trend and helps us make predictions for new data points"

The best fit line (also called the regression line) runs through the middle of the data. It captures the general pattern while staying as close as possible to all points.



The Line Equation: y = mx + c

$$y = mx + c$$

Every straight line can be described by this simple equation. Let's break down what each part means:

m (Slope)

How steep the line is—how much y changes for each unit increase in x

c (Intercept)

Where the line crosses the y-axis—the value of y when x = 0

Example Calculation

If our line is **y = 10x + 10**, and a student studies for 7 hours:

y = 10(7) + 10 = 80 marks

Making a Prediction

1

Step 1

Find x = 7 on the horizontal axis

2

Step 2

Move up to the best fit line

3

Step 3

Read the y-value (predicted marks)

The line gives us our prediction: approximately **80 marks** for 7 hours of study. This is how regression makes predictions for new, unseen data.

Real-Life Example: Car Price Prediction

Let's look at another practical application of regression analysis: predicting the price of a used car based on its mileage.

Used Car Data

Mileage (km)	Price ($)
10,000	28000
30,000	24000
50,000	20000
70,000	16000
90,000	12000

Notice the negative relationship: as mileage increases, price decreases. The best fit line helps dealers and buyers estimate fair prices.

Using regression, we can find a formula to predict the price:

Price = -0.18 × Mileage + 29,800

For a car with 60,000 km:

Price = -0.18(60,000) + 29,800 = $19,000

How Did We Find This Formula?

Let's delve into the process by which regression analysis determines the formula for the best fit line.

1

Analyzing Data Points

Regression algorithms analyze all the data points to find the line that best fits the observed pattern. This involves identifying the trend between mileage and price.

2

Understanding the Slope

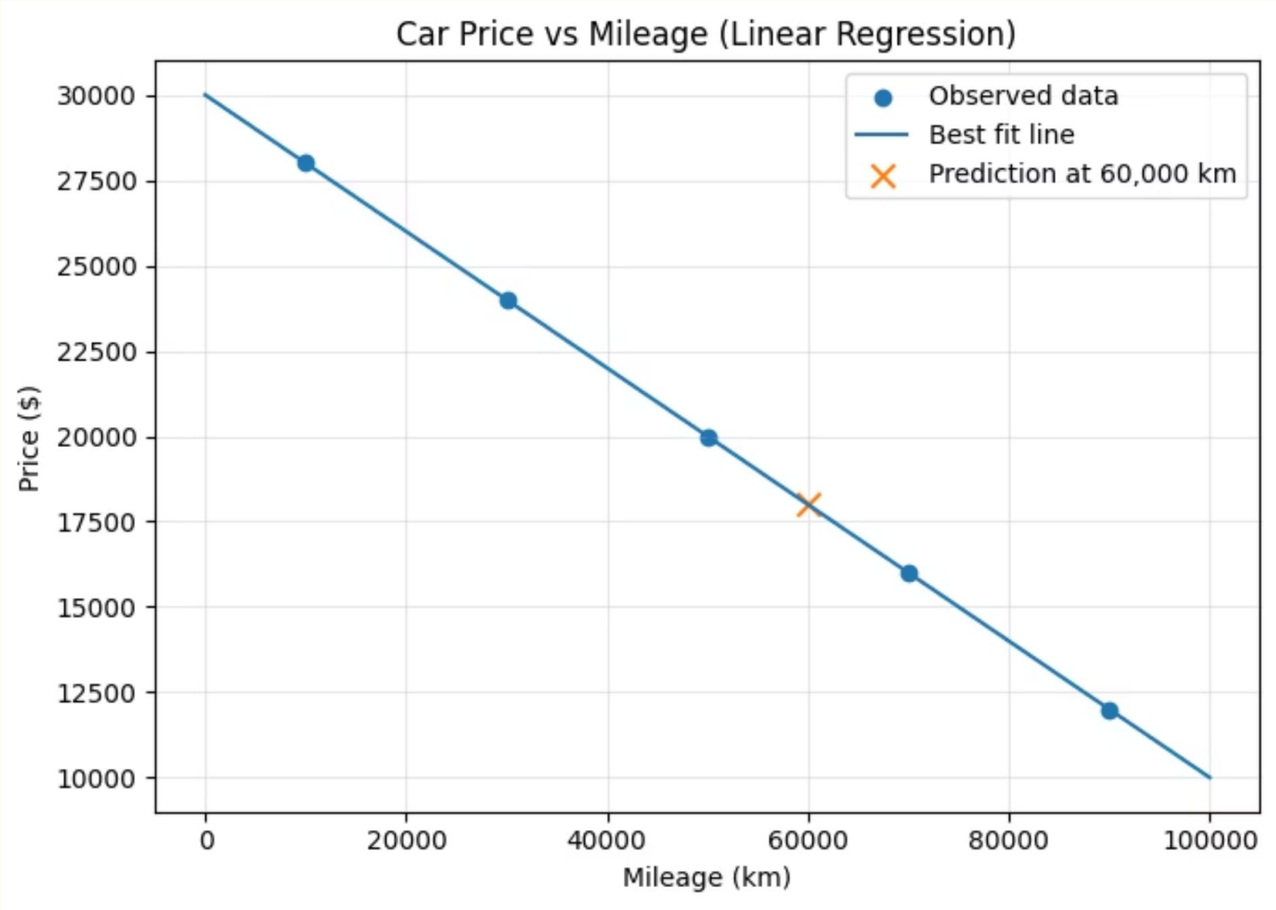
The algorithm calculates the slope (m = -0.18). This tells us that for every 1 km increase in mileage, the car's price is predicted to drop by $0.18. It quantifies the negative relationship.

3

Interpreting the Intercept

The intercept (c = 29,800) represents the theoretical price of a brand new car with 0 km. The algorithm minimizes the distance between this line and all actual data points to determine the "best fit" line.

Finding the Best Fit: Minimizing Error



The "best fit" line isn't just any line. It's the one that minimizes the total distance between itself and every data point. These distances are often called "residuals" or "errors." The regression algorithm systematically tests various lines, ultimately selecting the one that results in the smallest overall error. This rigorous process guarantees that our predictions are as accurate and reliable as possible.

Regression vs Classification

Regression

Predicts Numbers

- House price: $425,000
- Temperature: 72°F
- Exam score: 85 marks
- Sales revenue: $12,500

Output is *continuous* and can be any value within a range

Classification

Predicts Categories

- Email: Spam or Not Spam
- Image: Cat or Dog
- Grade: A, B, C, D, or F
- Diagnosis: Positive or Negative

Output is *discrete* and belongs to a fixed set of classes

Key Takeaways

1

Regression predicts continuous numbers

Use it when your output is a measurable value, not a category

2

The best fit line captures the trend

It summarizes the relationship between input and output variables

3

The equation y = mx + c powers predictions

Once we know m and c, we can predict y for any value of x

4

Visualization helps understanding

Scatter plots and regression lines make patterns visible and intuitive

Next up: We will explore the cost function and gradient descent.