An Introduction to Linear Regression

The "Hello, World!" of AI Models

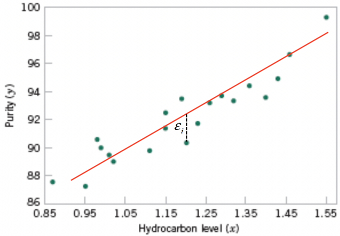
Jie Wang & Zhaoliang Yuan

The Chinese University of Hong Kong, Shenzhen

November 11, 2025

What is Regression?

- A fundamental **supervised learning** task.
- Goal: Predict a **continuous** (numerical) output value based on input data.
- Examples:
  - Predicting house prices based on size, location, etc.
  - Forecasting sales based on advertising budget.
  - Estimating a student's final exam score based on hours studied.



Response   Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$

Intercept   Slope   Random error

The Simplest Model: One Input, One Output

We start with one input feature (or variable) x to predict one output y.

Example

Input (x): Hours Studied
Output (y): Exam Score

The Model

We assume a **linear** (straight-line) relationship:

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

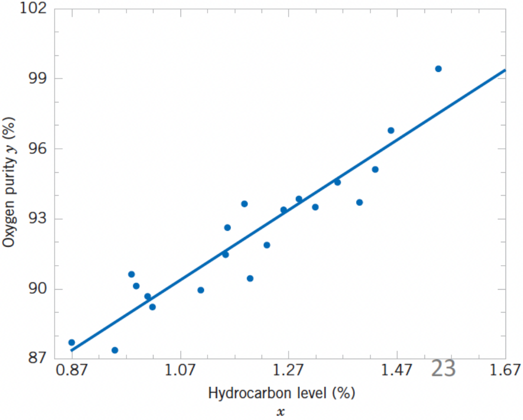
Breaking Down the Simple Model

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

- y: The **predicted** output (e.g., predicted exam score).
- x: The **input** feature (e.g., hours studied).
- β_1 (Slope): How much y changes for a one-unit change in x.
  - "For each additional hour studied, your score increases by β_1 points."
- β_0 (Intercept): The predicted value of y when x is 0.
  - "The expected score if you didn't study at all." (Often less meaningful)

Finding the Best-Fit Line

- In real data, points don't fall perfectly on a straight line.
- We need to find the line that **best fits** the data.

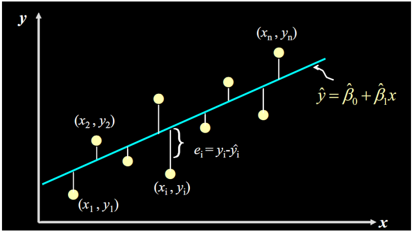


How Do We Measure "Best"? The Cost Function

We use **Ordinary Least Squares (OLS)**.

The Idea

Find the line that minimizes the sum of the squared **errors** (the vertical distances between the data points and the line).



Each white line is an error (or residual):
$Error = (True\ \ Value) - (Predicted\ \ Value)$

Find "optimal" coefficient of simple regression

Model and Objective

Linear model: $y = \beta_0 + \beta_1 \cdot x$

Minimize Sum of Squared Errors (SSE):

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Unconstrained Optimization

- Decision variables: β_0, β_1
- No constraints
- Objective: $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Find "optimal" coefficient of simple regression

Optimality Conditions

Set derivatives to zero:

$$\frac{\partial f}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial f}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Solution(do it by yourself!)

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

$\beta_1^* = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$

The Real World Has Many Factors

What if the exam score depends on more than just study hours?

- Hours of sleep?
- Attendance?
- Previous GPA?

Extending the Model

We can include **multiple input features**:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Understanding the Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Example

- y: Exam Score
- x_1: Hours Studied
- x_2: Hours of Sleep
- x_3: Attendance (%)

- β_1: The effect of *one more hour of study* on the score, **while holding sleep and attendance constant**.
- Each coefficient (β) shows the **individual contribution** of that feature.

Hands-on Lab 1

- Download the MOSEK solver to your local computer. https://www.mosek.com/downloads/
- receive the email from MOSEK and follow the instruction to create a file name "mosek" and put the license into that file
- Use cvxpy to construct linear regression models
- Test the models on Housing-price prediction

A Common Problem: Too Many Features?

What if we have 100 possible features to predict house price?

- Size, bedrooms, bathrooms, zip code, proximity to school, year built, roof color, ...

The Challenge

- **Overfitting:** A model with too many features becomes overly complex. It memorizes the training data (including noise) but fails to predict new data well.
- **Interpretability:** A simpler model is easier to understand and explain.

How to Choose the Right Features?

This is called **Variable Selection** or **Feature Selection**.

Common Methods

1. **Expert Knowledge:** Use what you know about the problem.
2. **Exploratory Data Analysis:** Look for relationships visually.
3. **Automated Algorithms:**
   - **Forward Selection:** Start with no variables, add one at a time.
   - **Backward Elimination:** Start with all variables, remove the least useful one at a time.

Goal: Find a model that is accurate but also simple and robust.

Sparse Regression Problem

Original Goal

Find at most k non-zero coefficients:

$$\min_{\beta \in \mathbb{R}^n, \|\beta\|_0 \leq k} \|y - X\beta\|_2^2$$

Mixed-Integer Reformulation

Introduce binary variables $q_i \in \{0, 1\}$:

$$\begin{aligned} \min_{\beta, q} \quad & \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} q_i \leq k \quad -M \cdot q_i \leq \beta_i \leq M \cdot q_i, \quad i = \{1, \ldots, n\} \end{aligned}$$

Key Concepts and Parameters

The Big-M Method

- **M**: Upper bound on coefficient size
- If $q_i = 0$: $\beta_i = 0$ (feature excluded)
- If $q_i = 1$: $-M \leq \beta_i \leq M$ (feature included)

Regularization

- λ: Ridge regularization parameter
- Stabilizes optimization
- Prevents overfitting

Hands-on lab 2

- Using CVXPY to select variables for house pricing problem
- Use MOSEK to solve the problem
- Display the optimal features that satisfy the sparsity constraint

Summary: The Linear Regression Toolkit

- **Simple Regression:** Models the relationship between one input and one output. $y = \beta_0 + \beta_1 x$
- **Multiple Regression:** A powerful extension that uses many inputs to make a better prediction. $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$
- **Variable Selection:** The art and science of choosing the right features to build a model that generalizes well and is easy to interpret.

Why is this in an AI course?

Linear regression is a foundational **predictive model**. Understanding its concepts (features, coefficients, training, prediction) is the first step toward more complex AI like neural networks!

Questions?