Introduction & Linear Regression

Overview

Methods and Algorithms

MSc Data Science

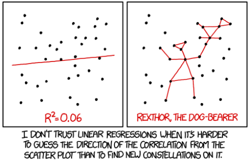
Spring 2026

Outline

Situation: Banks approve millions of mortgage applications each year. Every one requires an accurate property valuation.

Complication: Manual appraisals are slow, expensive, and inconsistent. Two appraisers can disagree by 20% on the same property.

Question: Can we build a model that predicts property value from its characteristics—automatically and consistently?



XKCD #1725 by Randall Munroe (CC BY-NC 2.5) – This is the problem we solve today

What Are the Two Types of Learning?

Supervised Learning

- We have the answers (labels) for training data
- Goal: predict the answer for new data
- Examples: predicting house prices, classifying emails

Unsupervised Learning

- No answers – just data
- Goal: find hidden structure or groups
- Examples: customer segments, anomaly detection

This course covers both: L01–L04 supervised, L05 unsupervised

What is Regression?

Predicting a Number (Not a Category)

- Predicting a number (not a category) – for example, a house price
- Example: predicting house price from its size in square meters
- We draw the best line through the data points
- The line captures the relationship between input and output

---

Regression = predicting a continuous value. Classification = predicting a category.

How Do We Model a Relationship?

The Idea
Price goes up as size increases. We write this as:

$$y = a + b \times x$$

- y = price (what we predict)
- x = size (what we know)
- a = starting price, b = price per sqm

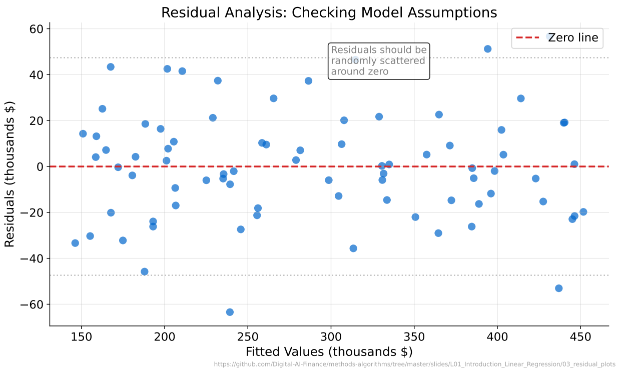
Example

$$\text{Price} = 50{,}000 + 200 \times \text{Size}$$

- A 100 sqm apartment: $50{,}000 + 200 \times 100 = 70{,}000$
- A 150 sqm apartment: $50{,}000 + 200 \times 150 = 80{,}000$
- Each extra square meter adds 200 to the price

This is the entire idea of linear regression – the rest is details

What Makes a Good Prediction?



- The red lines show prediction errors (residuals)
- A good model has small, random errors
- Patterns in errors = something the model missed

We want errors to be small and random – not systematic

How Does the Computer Learn?

The Learning Process

Step 1: Start with a random line

Step 2: Measure how wrong it is (errors)

Step 3: Adjust the line to reduce errors

Step 4: Repeat until errors are small

Step 5: Use the final line to predict

This process is called **optimization** – the computer tries thousands of lines and keeps the best one

The computer tries thousands of lines and keeps the best one

What Terms Do You Need to Know?

Data

- Features (inputs): size, bedrooms, age
- Target (output): price
- Observation: one data point (one house)

Learning

- Training: fitting the model to data
- Coefficients: the numbers the model learns
- Loss: how wrong the model is

Evaluation

- Prediction: model output for new data
- Error: difference between prediction and truth
- Overfitting: memorizing instead of learning

Master these terms – they appear in every ML method we study

Road Map for Today

Now that you know the basics, we go deeper:

1. A real business problem (house prices for banks)
2. The math behind finding the best line
3. How to tell if our model is good
4. How to prevent our model from being too complex

Everything ahead builds on the intuition from this intro

Why Do Banks Need Predictions?

Finance Use Case

- Banks need accurate property valuations for mortgages
- Insurance companies assess property risk
- Investors evaluate real estate portfolios

Why Linear Regression?

- Interpretable coefficients (price per square meter)
- Fast, well-understood method
- Strong baseline for comparison

Linear regression: the "hello world" of machine learning

What Does the Model Look Like?

The Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

- y: target variable (house price)
- $x_1, \ldots, x_p$: features (size, bedrooms, age)
- $\beta_0, \ldots, \beta_p$: coefficients to learn
- ε: random error term

Worked Example

$$\text{Price} = 50{,}000 + 200 \times \text{Size} + 15{,}000 \times \text{Bed} - 1{,}000 \times \text{Age}$$

- Base price: \$50,000
- Each sqm adds \$200
- Each bedroom adds \$15,000
- Each year of age subtracts \$1,000

Goal: Find coefficients that minimize prediction error

How Do We Find the Best Coefficients?

Objective: Minimize the sum of squared errors (SSE)

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

Why Squared Errors?

- Penalizes large errors more than small errors
- Differentiable (enables gradient-based optimization)
- Leads to closed-form solution

This defines "ordinary least squares" (OLS)

Two Ways to Solve: Which is Better?

Normal Equation

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \tag{3}$$

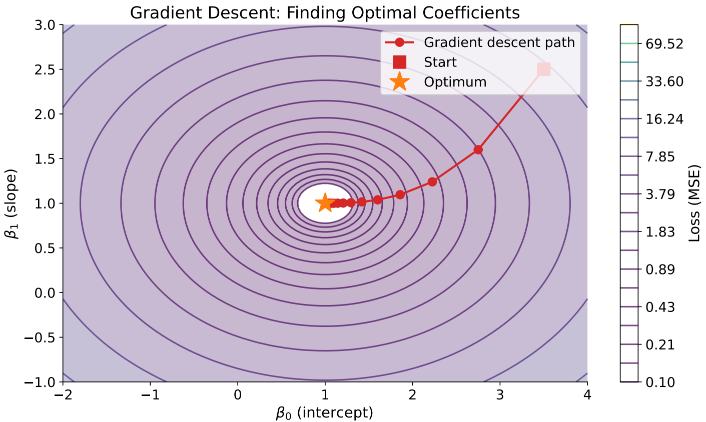
- Exact, closed-form solution
- One computation – no iterations
- Slow for very large datasets

Gradient Descent

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \alpha \nabla L(\boldsymbol{\beta}_t) \tag{4}$$

- Scales to big data
- Memory efficient
- Requires tuning learning rate α

Both methods yield the same solution for OLS



https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/L01_Introduction_Linear_Regression/04_gradient_descent

Iteratively update parameters in the direction of steepest descent

What Do the Numbers Mean?

Example: House price model

$$\text{Price} = 50{,}000 + 200 \times \text{Size} + 15{,}000 \times \text{Bedrooms} - 1{,}000 \times \text{Age} \tag{5}$$

Interpretation:

- $\beta_0 = 50{,}000$: Base price (all features = 0)
- $\beta_1 = 200$: Each extra sqm adds \$200
- $\beta_2 = 15{,}000$: Each bedroom adds \$15,000
- $\beta_3 = -1{,}000$: Each year of age subtracts \$1,000

Coefficients show marginal effect, holding others constant

How Good Is Our Model?

Key Metrics:

- R^2 **(Coefficient of Determination):** Variance explained

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{6}$$

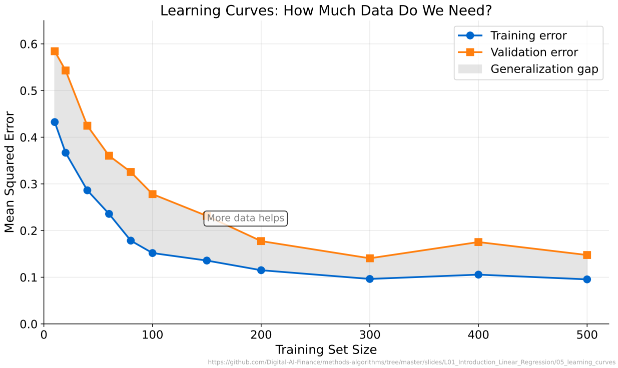
- **RMSE (Root Mean Squared Error):** Prediction accuracy

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{7}$$

In Practice: $R^2 = 0.75$ means model explains 75% of price variance

Always evaluate on held-out test data

Is Our Model Learning or Memorizing?



- Gap between train and test error = overfitting
- Curves converging = more data won't help

Learning curves diagnose underfitting vs overfitting

When to Use Linear Regression

Advantages

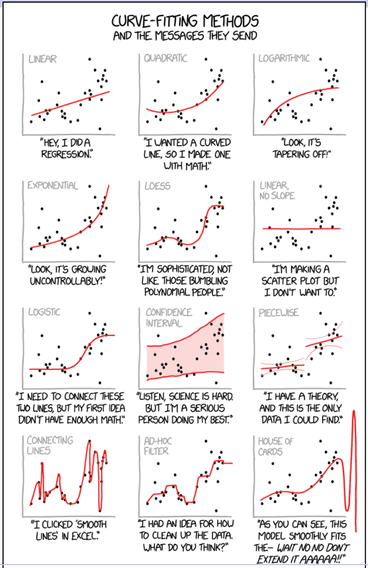
+ Interpretable coefficients
+ Fast training and prediction
+ Well-understood theory
+ Works as strong baseline

Disadvantages

- Assumes linear relationships
- Sensitive to outliers
- Limited flexibility
- Needs feature engineering for non-linear patterns

When in doubt, start with linear regression as your baseline

The Danger of Overfitting



XKCD #2048 – Simpler models often generalize better (CC BY-NC 2.5)

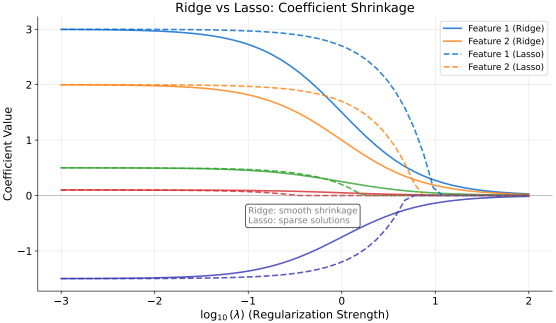
How Do We Prevent Overfitting?

Ridge (L2)

- Adds penalty on squared coefficients
- Shrinks all coefficients toward zero
- Never removes features entirely

Lasso (L1)

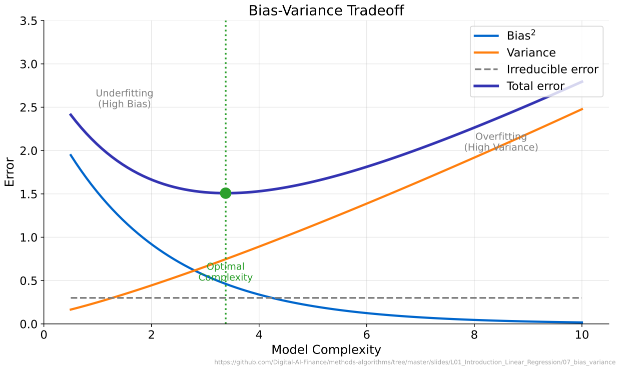
- Adds penalty on absolute coefficients
- Can set coefficients exactly to zero
- Automatic feature selection



https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/L01_Introduction_Linear_Regression/06_regularization_comparison

Regularization prevents overfitting by penalizing large coefficients

Why Can't We Minimize Both Errors?



- Bias: error from wrong assumptions (too simple)
- Variance: error from sensitivity to training data (too complex)
- Sweet spot: minimize total error

Model complexity controls the tradeoff between bias and variance

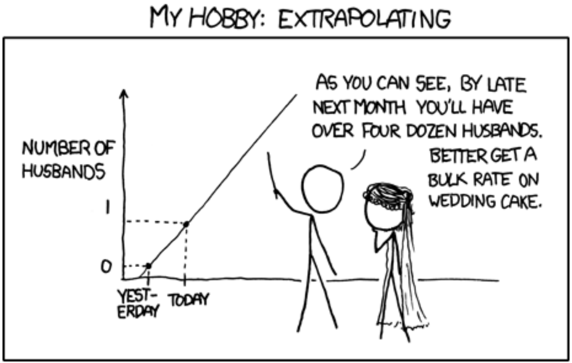
Hands-on Exercise

Open the Colab Notebook

- Exercise 1: Implement OLS from scratch
- Exercise 2: Use scikit-learn LinearRegression
- Exercise 3: Compare with gradient descent

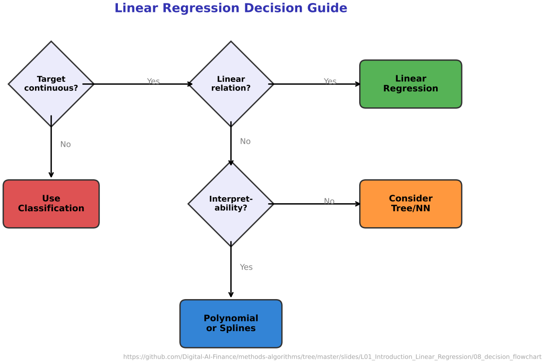
Link: See course materials on GitHub

Start with Exercise 2 if short on time



XKCD #605 by Randall Munroe (CC BY-NC 2.5) – Extrapolation is dangerous!

Which Method Should You Choose?



Use this flowchart to decide when linear regression is appropriate

Key Takeaways

Key Concepts

- Linear regression predicts continuous outcomes
- OLS minimizes sum of squared errors
- Solve via normal equation or gradient descent

Next: Deep dive into mathematics and implementation

References:

- ISLR Chapter 3: Linear Regression
- ESL Chapter 3: Linear Methods for Regression

Practical Guidance

- Coefficients are directly interpretable
- Evaluate with R^2 and RMSE on test data
- Use regularization to prevent overfitting

Foundation for all supervised learning methods