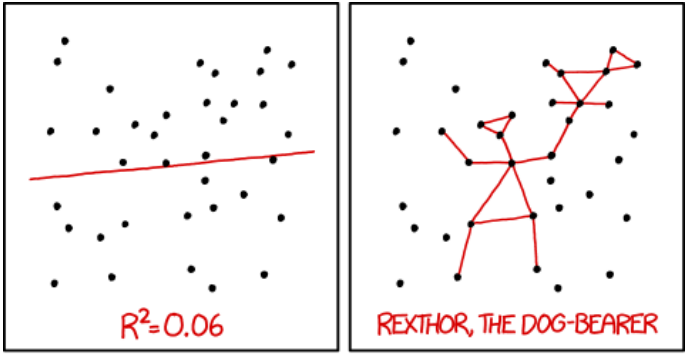
Linear Regression

Mini-Lecture: From Scatter Plots to Predictions

Methods & Algorithms

MSc Data Science – Spring 2026

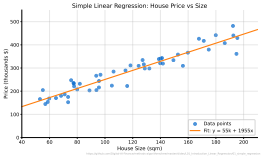
The Eternal Temptation of Fitting a Line



The Problem: Predicting Outcomes from Data

From Scatter Plot to Prediction

- A bank wants to predict house prices from square footage, location, and age
- Each data point is a past transaction – features on the x-axis, price on the y-axis
- Can we draw a line through the cloud that generalizes to new houses?



Key Question

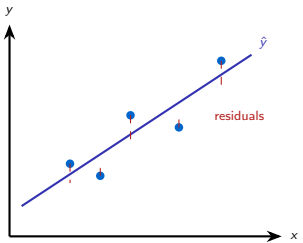
How do we find the "best" line, and what does "best" even mean?

Linear regression is the foundation of predictive modeling – nearly every ML method builds on or departs from it

The Idea: Fit a Line That Minimizes Error

The Linear Model

- Model: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- β_0 is the intercept (baseline prediction)
- Each β_j captures the effect of feature x_j on the outcome
- "Best" means minimizing the total distance between predictions and actual values



The residuals (red dashes) are the errors we want to make as small as possible

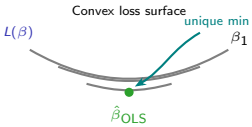
Ordinary Least Squares: The Closed-Form Solution

Minimize Squared Residuals

- Loss function: $L(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2$
- Take derivative, set to zero, solve
- Closed-form solution:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

This gives the unique global minimum – no iteration needed.



Insight

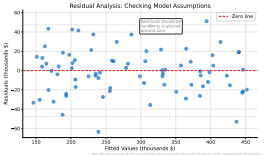
OLS is fast and exact, but requires $X^T X$ to be invertible. When features are correlated, this breaks down – motivating regularization.

OLS assumes: linearity, independence, homoscedasticity, normality of residuals (LINE assumptions)

How Good Is Our Model?

Evaluating Fit Quality

- R^2: fraction of variance explained, $R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$
- $R^2 = 1$ means perfect fit; $R^2 = 0$ means no better than the mean
- **Residual plots** reveal model violations: patterns mean the model is wrong



Insight

A high R^2 does not mean the model is correct – always check residual plots for non-linearity and heteroscedasticity.

Residuals should look like random noise. Any pattern signals a missing variable or wrong functional form.

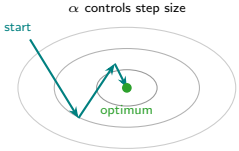
Gradient Descent: The Iterative Alternative

When OLS Is Too Expensive

- For millions of rows, inverting $X^T X$ is slow
- **Gradient descent**: update weights iteratively

$$\beta_{t+1} = \beta_t - \alpha \nabla L(\beta_t)$$

- α is the **learning rate** – too large overshoots, too small is slow
- Stochastic GD uses mini-batches for speed



Insight

Gradient descent scales to any dataset size and generalizes to non-linear models – it is the engine behind deep learning.

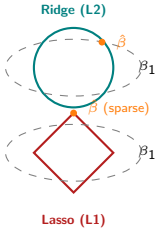
OLS: exact but $O(p^3)$. Gradient descent: approximate but $O(np)$ per step – choose based on data size.

Regularization: Preventing Overfitting

Ridge (L2) vs Lasso (L1)

- **Ridge**: adds $\lambda \sum \beta_j^2$ to loss – shrinks coefficients toward zero
- **Lasso**: adds $\lambda \sum |\beta_j|$ to loss – drives some coefficients exactly to zero (feature selection)
- λ controls the trade-off: more regularization = simpler model

Elastic Net combines both: $\lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2$



Insight

Use Ridge when all features matter; use Lasso when you suspect many features are irrelevant. Cross-validate λ.

Lasso's diamond constraint has corners on the axes – this is why it produces exact zeros (sparsity)

Finance Application: CAPM and Factor Models

Linear Regression in Finance

- **CAPM**: $R_i = \alpha + \beta R_m + \epsilon$
- β measures sensitivity to market risk; α is excess return
- **Factor models**: extend to multiple risk factors (Fama-French 3-factor, Carhart 4-factor)
- Portfolio risk decomposition via regression coefficients

Fama-French 3-Factor Model:

$$R_i - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 \cdot \text{SMB} + \beta_3 \cdot \text{HML} + \epsilon$$

SMB: Small Minus Big (size)
HML: High Minus Low (value)

Insight

Every hedge fund's risk report is built on linear regression. The β to market, size, value, and momentum factors determines portfolio exposure.

CAPM won the Nobel Prize (Sharpe, 1990). Linear regression is the language of asset pricing.

Summary: Linear Regression in 4 Takeaways

1. **The model**: $\hat{y} = X\beta$ – a linear combination of features, solved exactly via OLS or iteratively via gradient descent
2. **Evaluation**: R^2 measures fit quality, but always check residual plots – a good R^2 can hide a bad model
3. **Regularization**: Ridge (L2) shrinks all coefficients; Lasso (L1) sets irrelevant ones to zero. Cross-validate λ.
4. **Finance**: CAPM and factor models are linear regressions – β is the language of risk

Next Steps

Explore the deep dive slides for matrix derivations, statistical inference, and multicollinearity diagnostics. Try the Jupyter notebook to fit regressions on real housing data.

Linear regression is the starting point for all of machine learning – master it before moving on