Datachains

16 Sep 2025

Session #01

Regression Models – Quick Primer

What regression is, key types, and a focus on Linear & Multiple Linear Regression

What is a Regression ML Model?

- A supervised learning approach to predict a continuous value (e.g., price, demand, time).
- Learns a relationship between input features and a numeric target.
- Answers questions like: How much? How many? What value?

When to Use Regression

- You need numeric predictions (not labels).
- There is a measurable relationship between features and the target.
- You want interpretable coefficients (in simple models) or strong predictive power (tree/ensemble models).

Datachains

Common Regression Types

- Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Regularized Regression (Ridge/Lasso/Elastic Net)
- Tree-based Regression (Random Forest, Gradient Boosting)
- Time Series Regression (ARIMA, Prophet, RNN/LSTM)

Datachains

Handpicked for today ..

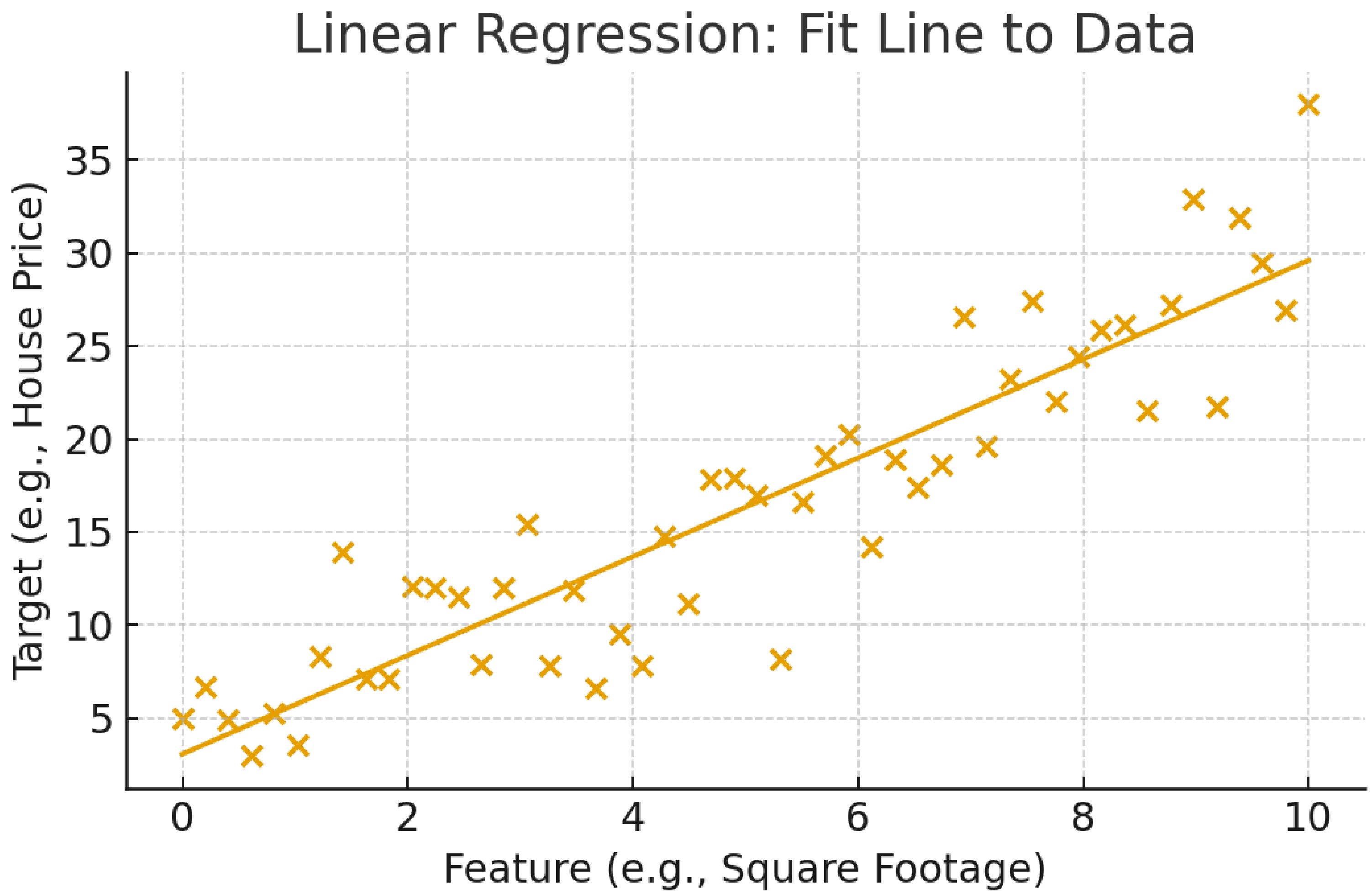
LINEAR REGRESSION — CORE INTUITION & ASSUMPTIONS

MULTIPLE LINEAR REGRESSION — ADDING MORE FEATURES

HOW TO SPOT ISSUES (E.G., MULTICOLLINEARITY) AND BEST PRACTICES

Datachains

Linear Regression – Intuition



Linear Regression – Key Assumptions

- Linearity: relationship between features and target is linear.
- Independence: observations are independent.
- Homoscedasticity: constant variance of errors.
- Normality of residuals: residuals roughly normal (for inference).
- No strong multicollinearity (applies more in MLR).

Linear Regression – Typical Workflow

- Understand the problem & collect data.
- Split into train/test sets; optionally cross-validate.
- Fit the model (least squares).
- Check residuals and goodness-of-fit (R^2, RMSE).
- Iterate: feature engineering, transformations, outlier handling.

Linear Regression – Pros & Cons

- ✅ Simple and fast; interpretable coefficients.
- ✅ Baseline model for many problems.
- ❌ Can underfit non-linear relationships.
- ❌ Sensitive to outliers and assumption violations.

Real-World Example: House Prices

- **Goal:** Predict sale price.
- **Features:** square footage, lot size, bedrooms, age, zip code.
- **Outcome:** A numeric price prediction used for appraisal & pricing tools.

Multiple Linear Regression (MLR) – What & Why

- Extends linear regression to multiple features ($y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$).
- Captures combined influence of many variables.
- Common in business/finance/science when outcomes depend on multiple drivers.

Datachains

MLR – Features & Multicollinearity

Feature scaling and encoding for categorical variables may be needed.

Multicollinearity (highly correlated features) inflates variance of coefficients.

Diagnostics: correlation matrix, VIF (Variance Inflation Factor).

Mitigation: drop redundant features, regularization (Ridge/Lasso), PCA.

MLR – Assumptions & Checks

- Same as Linear Regression plus careful attention to multicollinearity.
- Residual diagnostics: Q–Q plot, residuals vs. fitted, influence points (Cook's distance).
- Model selection: adjusted R^2, AIC/BIC, cross-validation.

Datachains

MLR – Pros & Cons

✅ USES MULTIPLE SIGNALS TO IMPROVE ACCURACY OVER A SINGLE-FEATURE MODEL.

✅ STILL INTERPRETABLE WITH CAUTION AROUND CORRELATED FEATURES.

❌ SENSITIVE TO MULTICOLLINEARITY AND OUTLIERS.

❌ LINEAR FORM MAY MISS COMPLEX NON-LINEAR EFFECTS.

Datachains

Marketing Mix Modeling

Goal: Predict weekly sales.

Features: spend on TV, search, social, promotions, seasonality, competitor price.

Outcome: Budget allocation recommendations based on coefficients and scenario planning.

Datachains

Wrap-up & Next Steps

Regression predicts numeric outcomes; start with Linear/MLR as baselines.

Validate assumptions and check diagnostics before trusting coefficients.

Next: Regularization (Ridge/Lasso), Polynomial terms, and Tree-based regressors.