

Introduction to Linear Regression

Supervised learning

assumes linearity

There is a target variable to be modeled

1. Introduction to Linear Regression

Linear Regression is a fundamental algorithm in supervised machine learning used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the inputs and the output.

2. Mathematical Foundation

The model predicts the output y from input x using the following formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where:

- \hat{y} : Predicted value
- β_0 : Intercept term
- $\beta_1, \beta_2, \dots, \beta_n$: Coefficients for input features

The coefficients are learned by minimizing the Mean Squared Error (MSE) between predicted and actual values:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

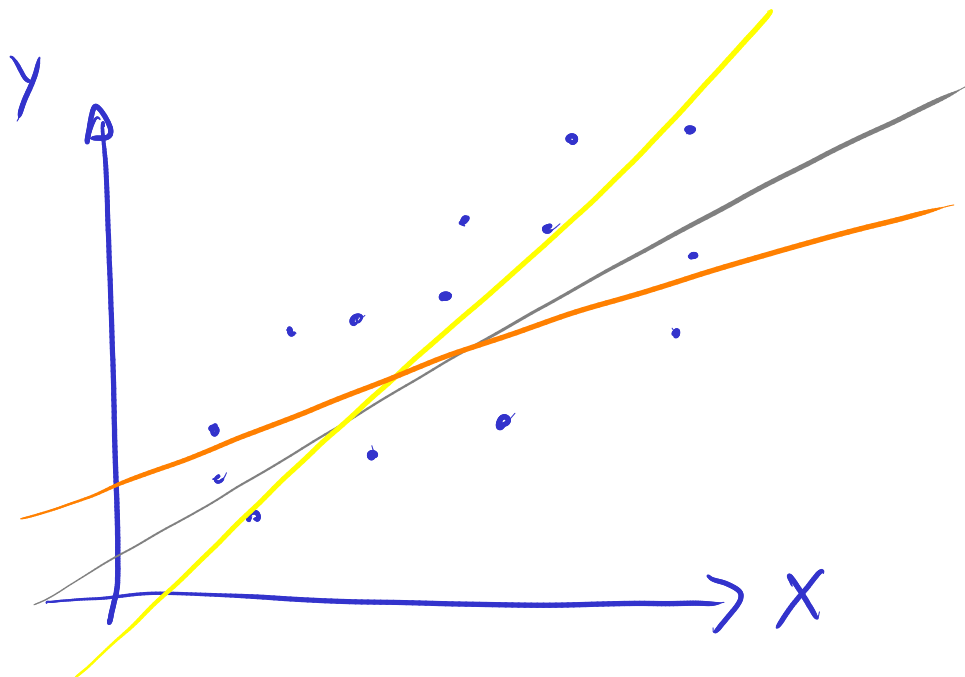
model prediction

True value

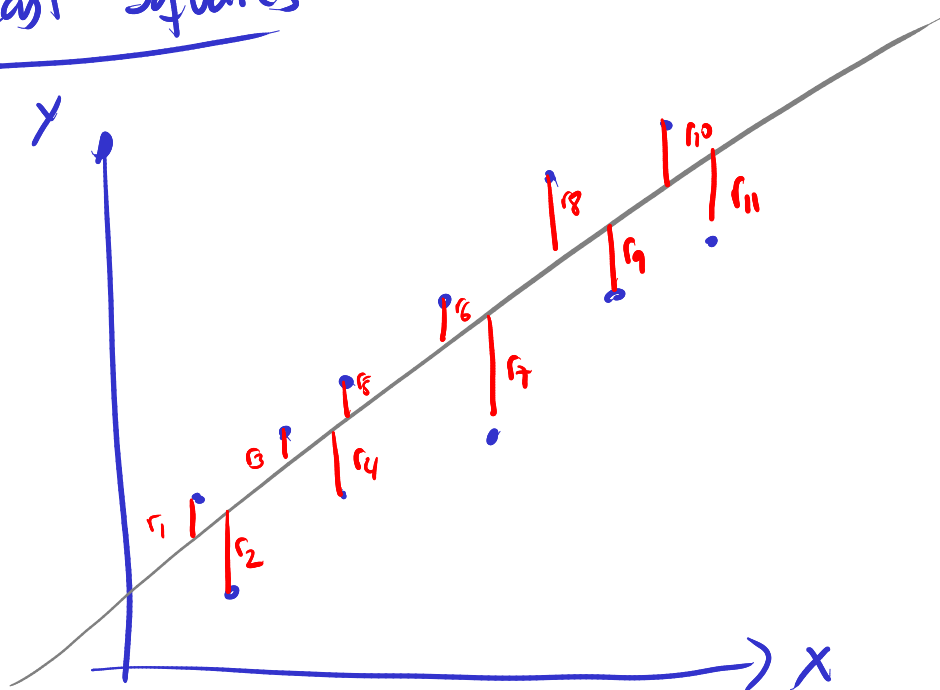
3. Applications in Data Science

Linear regression is widely used in various industries for predictive analytics:

- **Finance:** Predicting stock prices or returns.
- **Marketing:** Estimating advertising effectiveness on sales.
- **Real Estate:** Estimating house prices from square footage and location.
- **Operations:** Forecasting demand or resource usage.



Ordinary least Squares



The goal is to minimize

$$r_1^2 + r_2^2 + r_3^2 + \dots + r_{10}^2 + r_{11}^2$$

4. Pros of Linear Regression

- **Simple and interpretable:** Easy to understand model coefficients.
- **Fast to train:** Scales well with large datasets.
- **Good baseline:** Often used as a starting point for regression problems.
- **Analytical solution:** Closed-form solution using Ordinary Least Squares (OLS).

5. Cons of Linear Regression

- **Assumes linearity:** Fails to capture nonlinear relationships.
- **Sensitive to outliers:** Outliers can heavily influence predictions.
- **Multicollinearity issues:** Highly correlated inputs affect coefficient stability.
- **Overfitting/underfitting:** Without regularization, the model can behave poorly.

6. When to Use Linear Regression

- When the relationship between inputs and output is approximately linear.
- When interpretability of coefficients is important.
- As a quick baseline for regression tasks.

7. Summary

Linear Regression is one of the most widely used and well-understood models in data science. Though simple, it provides a solid foundation for understanding more complex regression and predictive modeling techniques.