2025

4Geeks Academy: data science cohort 12

DAY 19: LINEAR REGRESSION

TODO

LINEAR REGRESSION

Model details, evaluation, overfitting and regularization

LINEAR REGRESSION PROJECTS

- **Health insurance cost project** - ‘basic’ linear regression
- **Health demographics project** - regularized linear regression

EDA PROJECT

Submit Data Preprocessing Project Tutorial (Exploratory data analysis project) if you haven’t done so already.

TOPICS

01 LINEAR REGRESSION MODELS

02 REGRESSION METRICS

03 OVERFITTING

LINEAR REGRESSION

WHAT

Predicts a continuous label from input feature or features

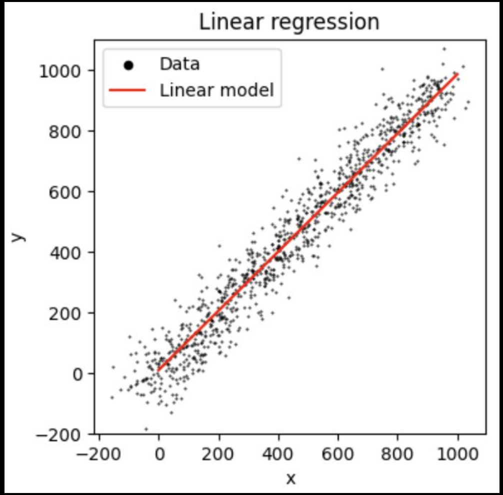
WHY

- Simple
- Easy to train - does not require large amounts of data, time or compute resources
- Interpretable - can tell you about the relative importance of features

HOW

$Y = mX + b$

Minimizes error by finding ‘best’ values for m and b.



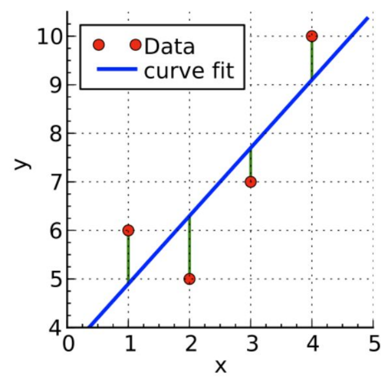
- **Simple linear regression**: one input variable
- **Multiple linear regression**: two or more input variables

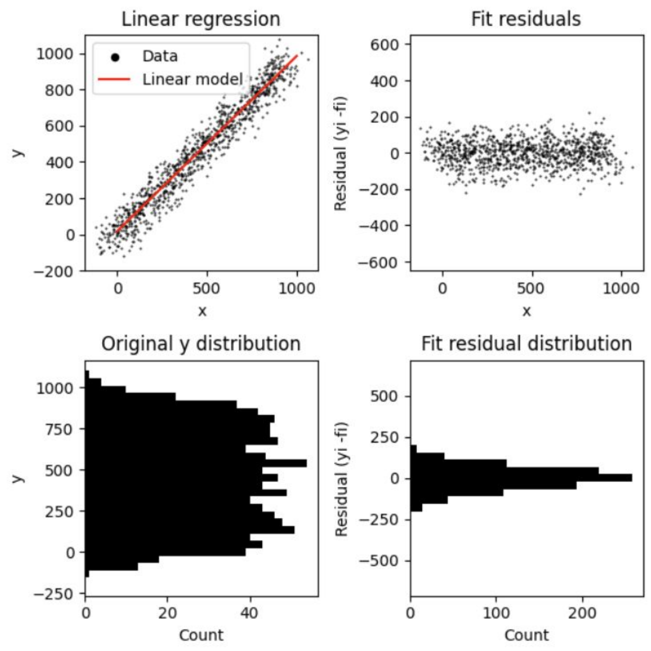
Available via Scikit-learn `LinearRegression()`

REGRESSION METRICS

RMSE

- Quantifies on average, how far predicted value is from the true value
- Higher is better
  - Zero -> model has no explanatory power
  - One -> model makes perfect predictions





R-squared

- Quantifies fraction of variance in the label explained by the model
- Higher is better
  - Zero -> model has no explanatory power
  - One -> model makes perfect predictions

OVERFITTING

BIAS-VARIANCE TRADEOFF

Very good fit on training data often leads to bad generalization

OVERFITTING

The model is too tuned to the training data, it can’t predict the test data

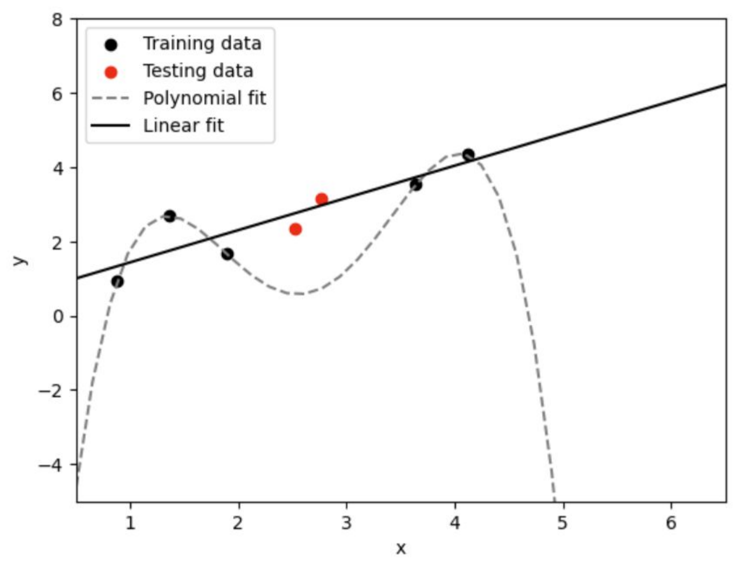
REGULARIZATION

Set of techniques to penalize/mitigate overfitting

- Don’t use a over-powered model
- Don’t use too many features

Linear regression:

- **Ridge regression** - adds a penalty to constrain size of coefficients
- **Lasso regression** - adds a penalty to force coefficients to zero


● Training data
● Testing data
Polynomial fit
Linear fit
y
x