Regression

CS3300 Data Science

RJ Nowling

Readings

- Section 9.0-9.5

Common Forms of Machine Learning

- Supervised Learning
  - Regression – predicting a continuous output
  - Classification – predicting a categorical output
- Unsupervised Learning
  - Clustering – grouping similar records

Define a Problem

- We want to predict the sale price for real estate transactions.

  Regression

- We want to predict whether the animal in a picture is a cat or dog.

  Classification

Terminology

- Response – output variable we are trying to predict
- Predictor – input variable we are using to predict the response

(Multiple) Linear Regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where $\hat{y}$ is the predicted value, *p* is the number of features, x_i are the features, and β_i are the feature weights.

Simple Linear Regression

$$\hat{y} = \beta_0 + \beta_1 x_1$$

where $\hat{y}$ is the predicted value, x_1 is a single feature, and β_i are the feature weights.

Advertising Data Set

- Response: Sales
- Predictors:
  - Amount of TV advertisements
  - Amount of radio advertisements
  - Amount of newspaper advertisements

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Advertising Data

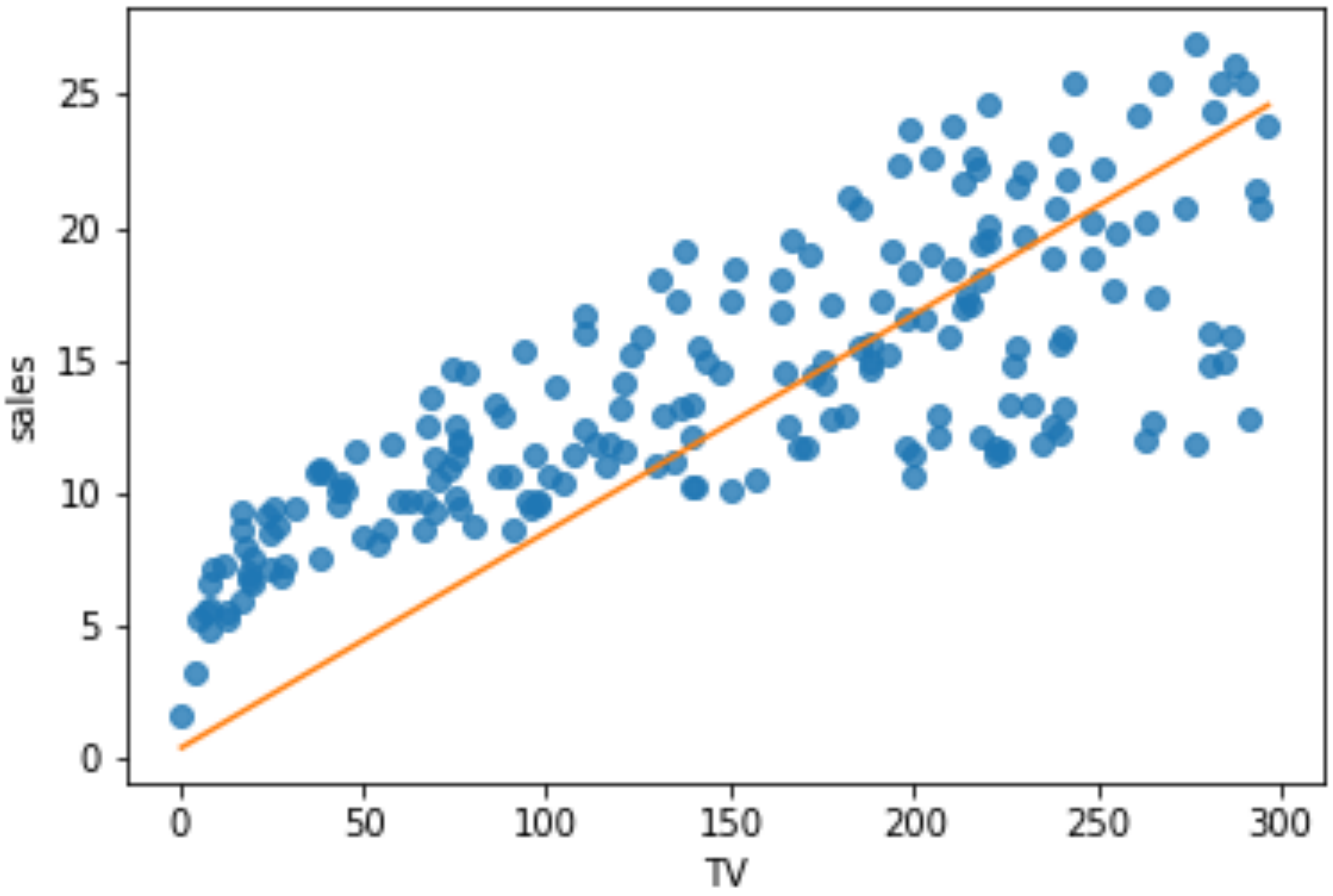
$$\widehat{sales} = \beta_0 + \beta_1 TV$$

where $\widehat{sales}$ is the predicted value, TV the feature, and β_i are the feature weights.

Fitted Model

- $\beta_0 = 0.0818$ base units of sales
- $\beta_1 = 0.3483$ units of sales per unit of TV advertisements over base

Plot of Linear Regression Model

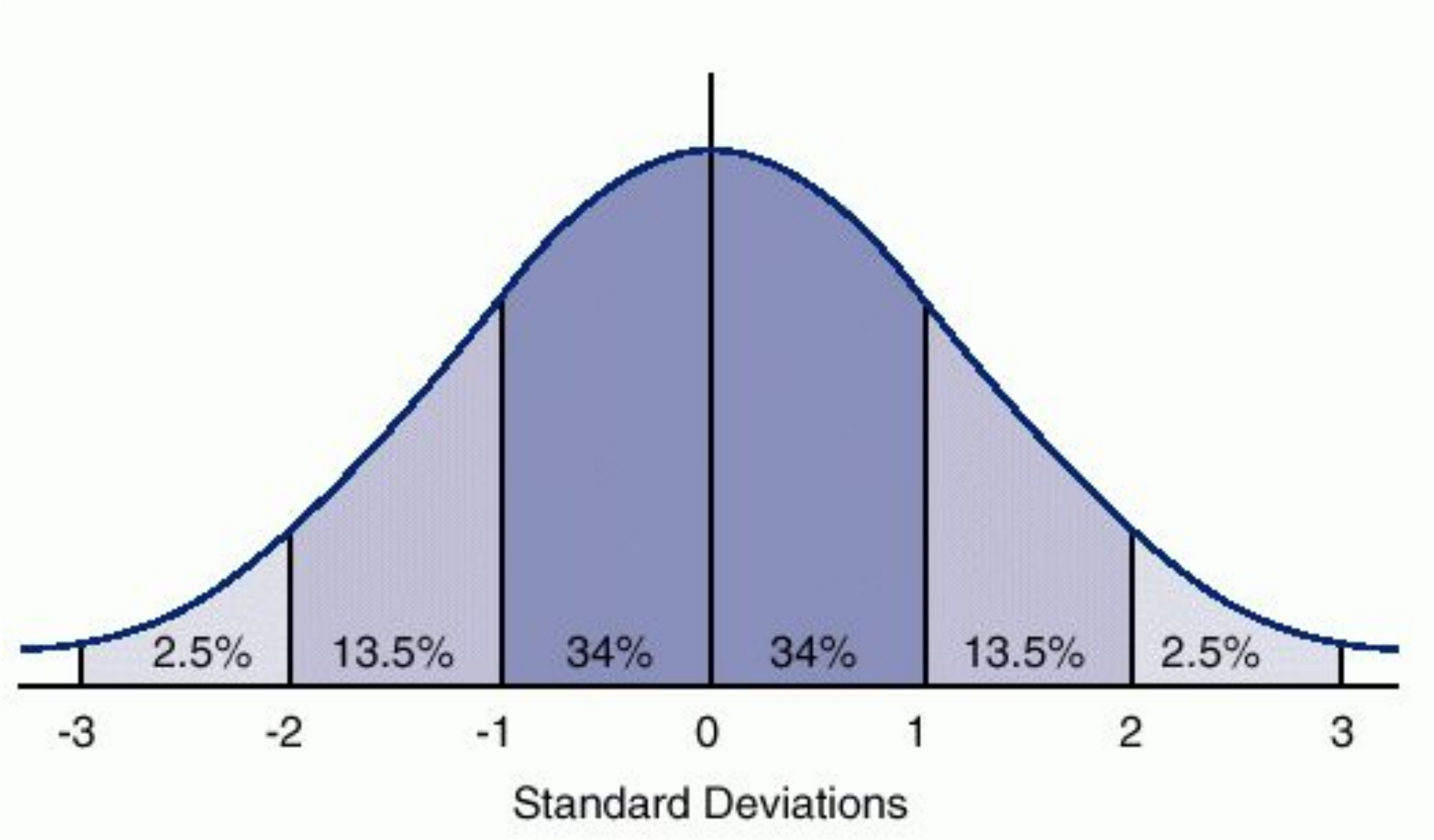


Advertising Data

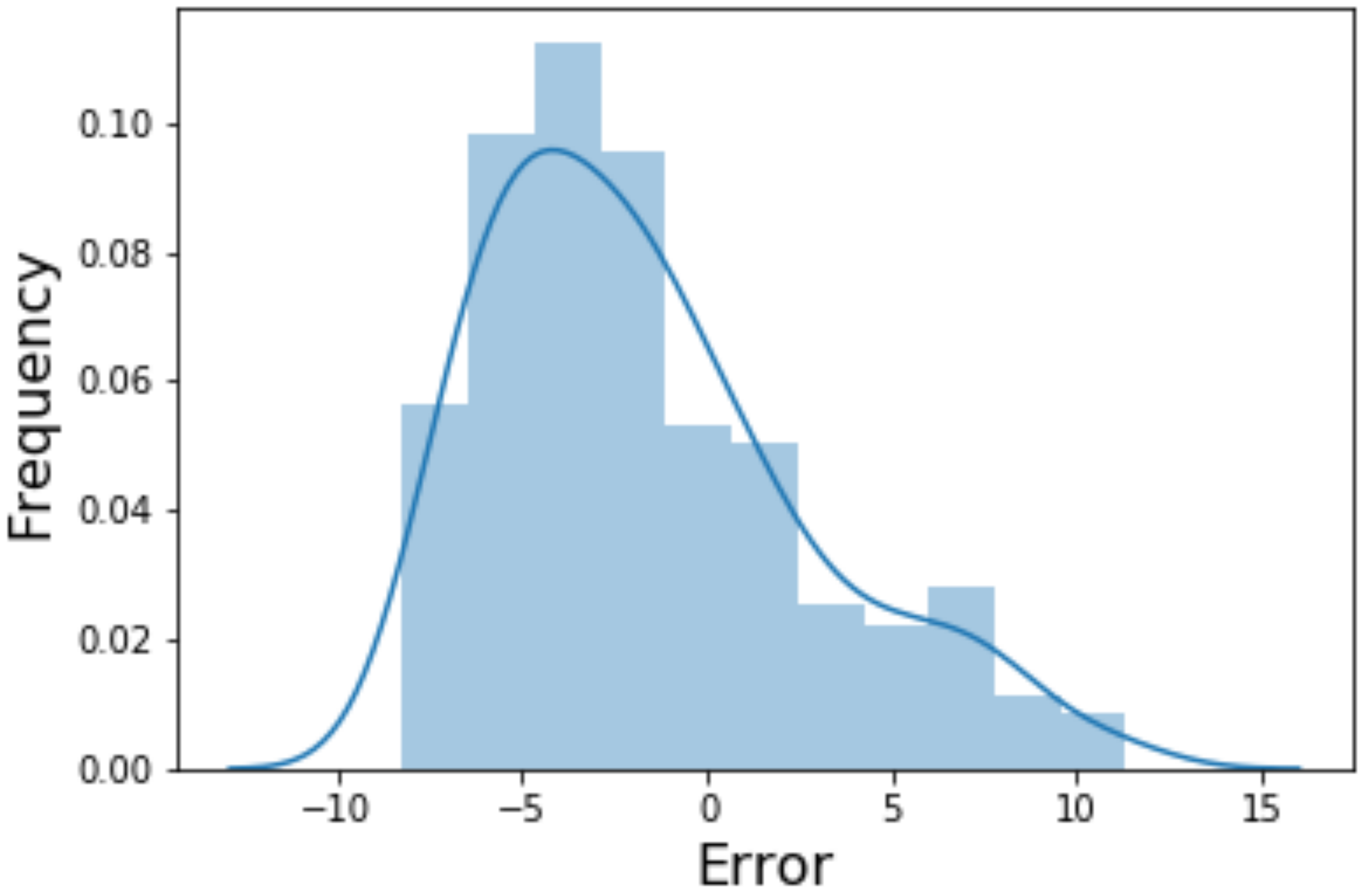
$$sales = \beta_0 + \beta_1 TV + \epsilon$$

where sales is the true value, TV the feature, β_i are the feature weights, and ϵ is the error

Assumed Error Distribution



Error Distribution



Advertising Data

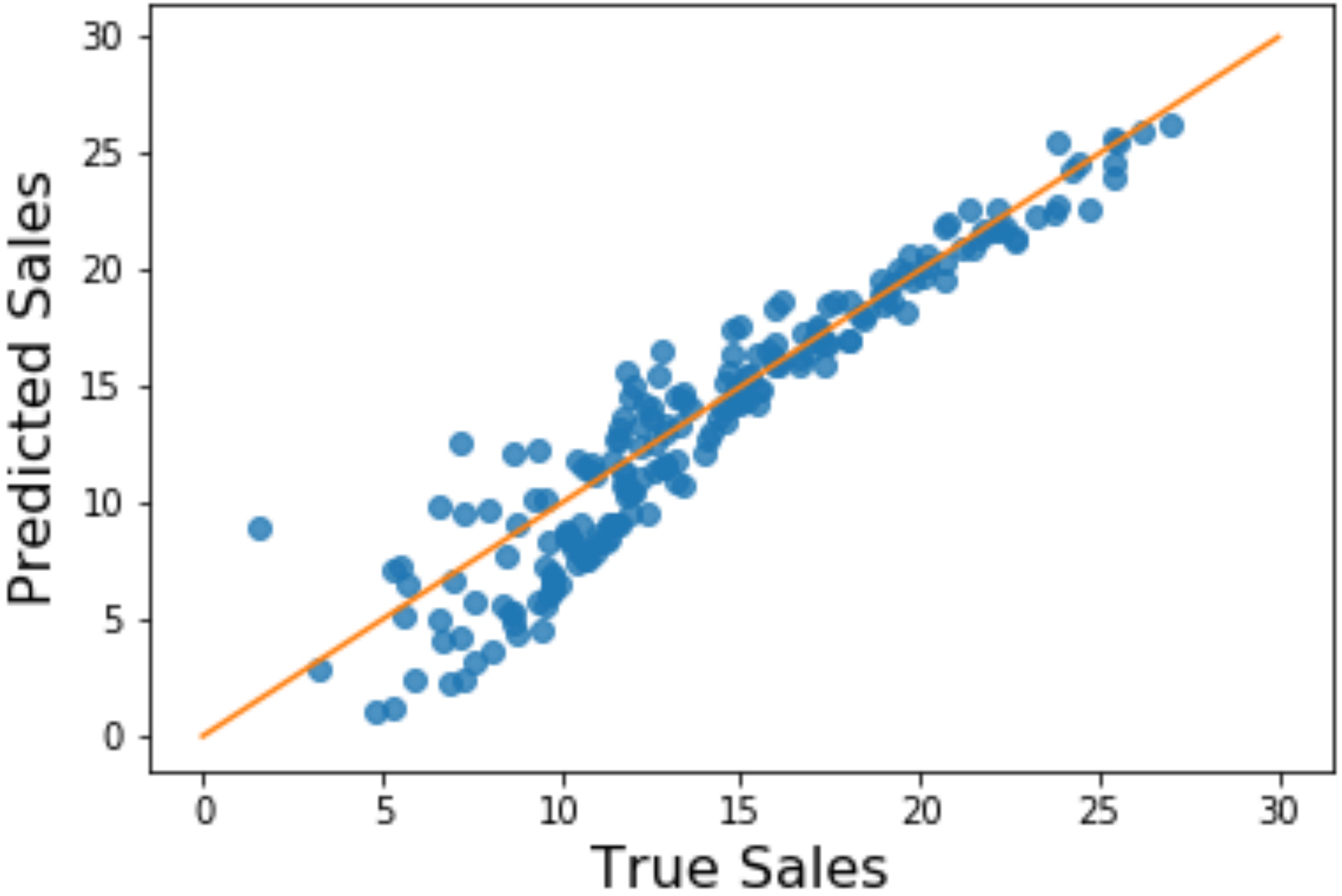
$$\widehat{sales} = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper$$

where $\widehat{sales}$ is the predicted value, TV, radio, and newspaper are the features, and β_i are the feature weights.

Fitted Model

- $\beta_0 = 0.0874$ base units of sales
- $\beta_1 = 0.0530$ units of sales per unit of TV advertisements
- $\beta_2 = 0.2215$ units of sales per unit of radio advertisements
- $\beta_3 = 0.0162$ units of sales per unit of newspaper advertisements

Predictions vs True Sales



Metrics for Evaluating Regression Models

- Mean-Squared Error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2$$

- Root Mean-Squared Error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2}$$

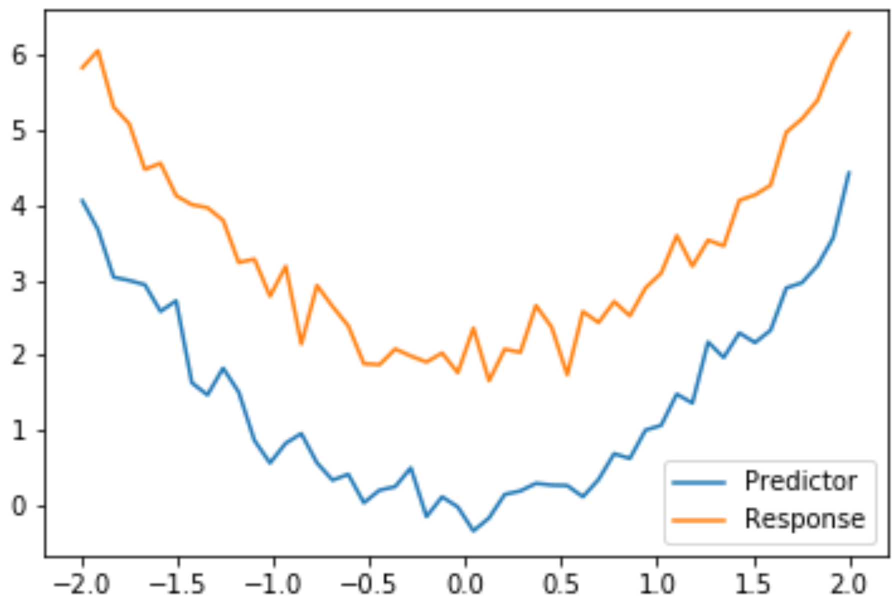
Advertising Error

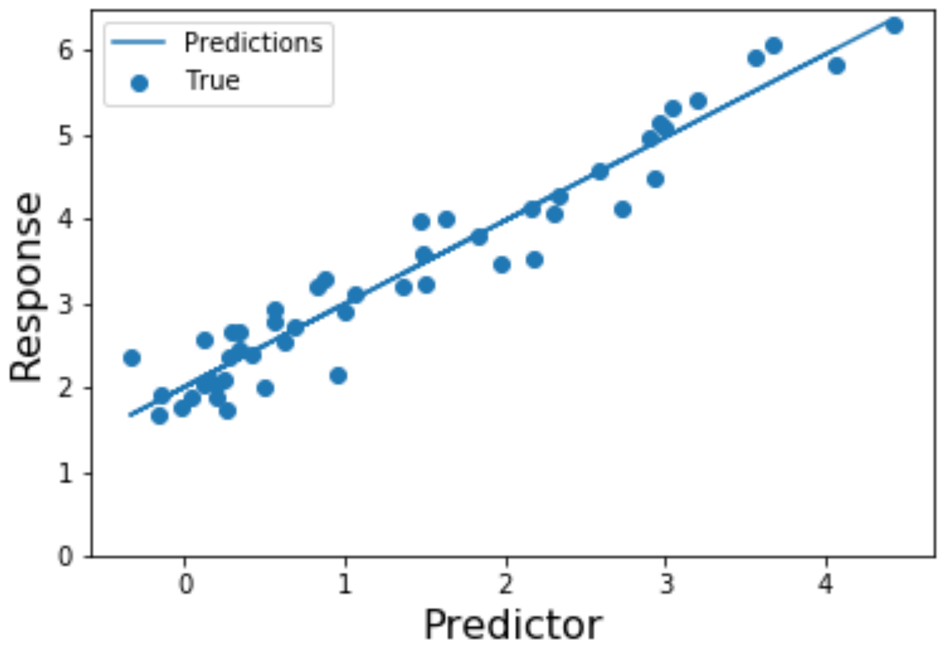
- Mean-Squared Error: 3.981
- Root Mean-Squared Error: 1.995

Linear Relationships

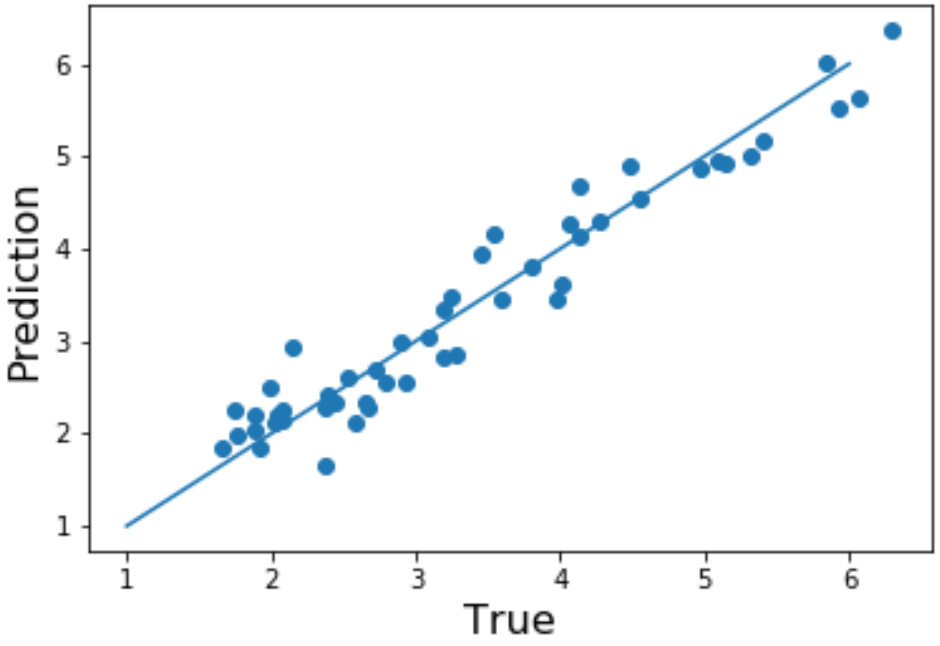
- The response variable does not need to be linear for linear regression to work
- The relationships between the response and predictors must be linear, however

Non-Linear Data





Non-Linear Data



Error

- Mean-Squared Error: 0.114
- Root Mean-Squared Error: 0.337

Credit Data Set

- Response: Balance
- Numerical Predictors:
  - Income, Limit, Ratings, Cards, Age, Education
- Categorical Predictors:
  - Gender: Male, Female
  - Student: Yes, No
  - Married: Yes, No
  - Ethnicity: African-American, Caucasian, Asian

Categorical Variables

- How do we interpret categories numerically?
  - We can't
- We can use one-hot encoding to create a separate numerical variable for each category in a categorical variable
- If there are N categories (e.g., is a student, is not a student), then we create N new "dummy" variables
- We set one of the N dummy variables to 1, the rest to 0

Credit Student Example

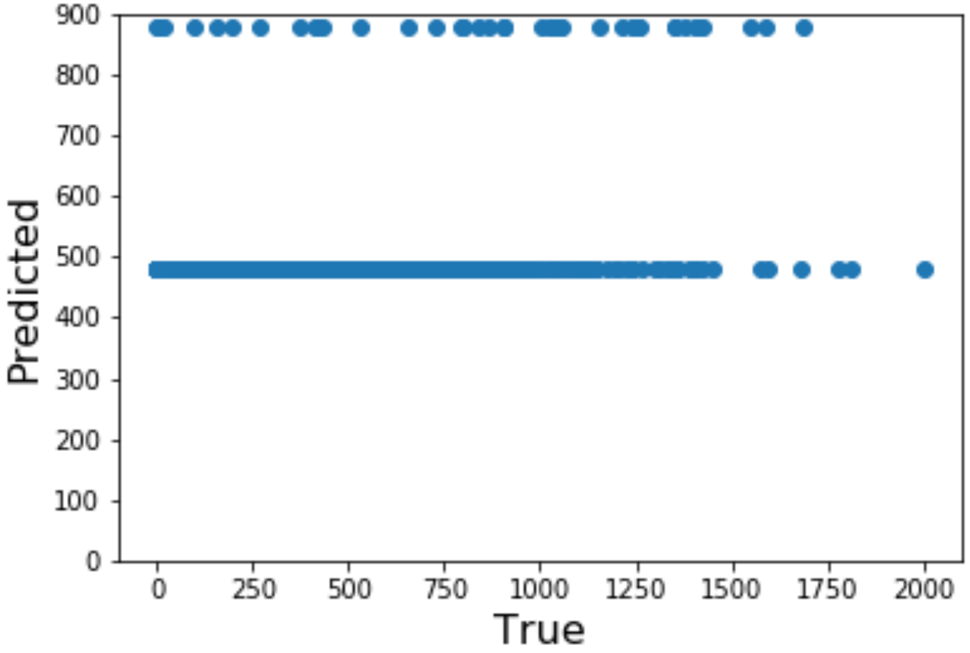
	Student_No	Student_Yes
0	1	0
1	0	1
2	1	0
3	1	0
4	1	0

Credit Data

$$\widehat{balance} = \beta_0 + \beta_1 StudentYes + \beta_2 StudentNo$$

where $\widehat{sales}$ is the predicted value, StudentYes and StudentNo are the features, and β_i are the feature weights.

Credit Predictions



- MSE: 196704.1
- RMSE: 443.5

Regularization

- Regularization involves adding a penalty term when fitting a model
- The penalty term is based on the weights of the model

Lasso

- The regularization term is based on the L2 norm of the weights
- L2 regularization improves the model stability
- E.g., helps the model handle collinear predictors
- Default in Scikit Learn's SGDRegressor class

Ridge

- The regularization term is based on the L1 norm of the weights
- Enables the model to pick a subset of the predictors by setting the weights for unused features to 0

Other Regression Techniques

- Polynomial regression
- Multivariate adaptive regression splines (MARS) regression
- Random Forest Regression
- k-Nearest Neighbor (kNN) Regression