University of Essex

Regression Analysis

Akitaka Matsuo

Content

- Regression problem
- Objective function
- Ordinary Least Square (OLS) model
- Data in the demo

University of Essex

Regression problem

University of Essex

- Outcome Y is a continuous variable
  - Example:
    - Income
    - Number of electoral votes a candidate won
- Inputs X can be anything
  - continuous
  - categorical

The model

$$Y = f(X) + \epsilon$$

- The output is a product of some function of X and an error.
- We want to find a good $f(X)$

Objective function

University of Essex

- Objective function is a function you want to optimize (i.e. minimize or maximize) and evaluate the model performance. $\hat{Y}$ is the prediction from $f(X)$.
- Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum|Y - \hat{Y}|$$

- Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum(Y - \hat{Y})^2$$

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(Y - \hat{Y})^2}$$

Linear model

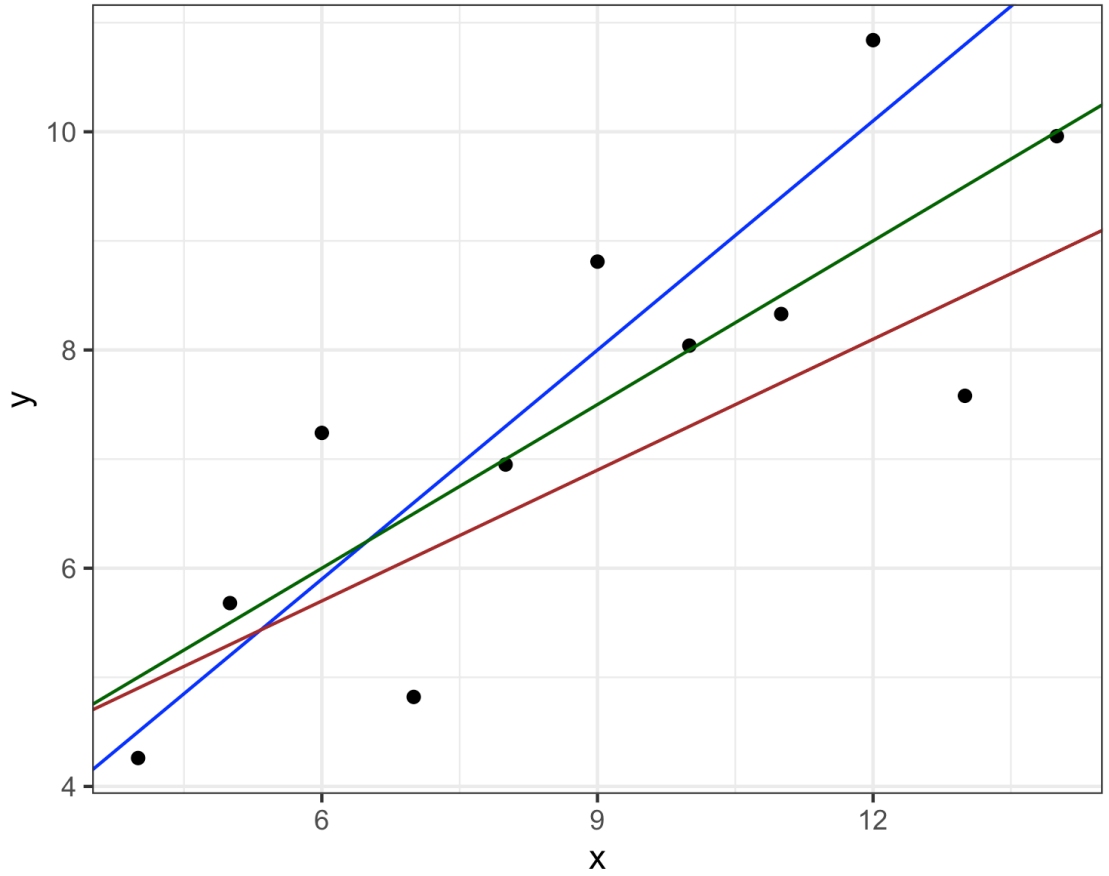
University of Essex

- Suppose that we assume the linearity for $f(X)$:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

$$= \beta_0 + \sum_j \beta_j X_j$$

- In this setup, the relation between x and y is a straight line. Which line looks the best?
- Linear regression provide an answer.



Liner regression

University of Essex

- Linear regression is a method that minimize MSE/RMSE among linear models.
- The minimization problem here is:

$$\text{argmin}_{\beta} \sum_{i} (Y_i - (\beta_0 + \sum_{j} \beta_j X_{ij}))^2$$

- OLS regression is BLUE (= best linear unbiased estimator, or the best solution for minimizing train RMSE, with linear $f(X)$)

$$\hat{Y} = \hat{f}(X) = \widehat{\beta_0} + \sum_{i} \hat{\beta}_i X_i$$

- There is no need for tuning parameter (except for variable selection)
- This solution is BLUE, but may be too ignorant for the model variance (so we need regularized regressions)

Boston data

University of Essex

- We use Boston data which is a part of **MASS** package in R
- “Housing data for 506 census tracts of Boston from the 1970 census.”
- Variables:
  - **crim**: per capita crime rate by town
  - **zn**: proportion of residential land zoned for lots over 25,000 sq.ft
  - **indus**: proportion of non-retail business acres per town
  - **chas**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  - **nox**: nitric oxides concentration (parts per 10 million)
  - **rm**: average number of rooms per dwelling
  - **age**: proportion of owner-occupied units built prior to 1940
  - **dis**: weighted distances to five Boston employment centres
  - **rad**: index of accessibility to radial highways
  - **tax**: full-value property-tax rate per USD 10,000
  - **ptratio**: pupil-teacher ratio by town
  - **b**: 1000(B - 0.63)^2 where B is the proportion of blacks by town
  - **lstat**: percentage of lower status of the population
  - **medv**: median value of owner-occupied homes in USD 1000’s