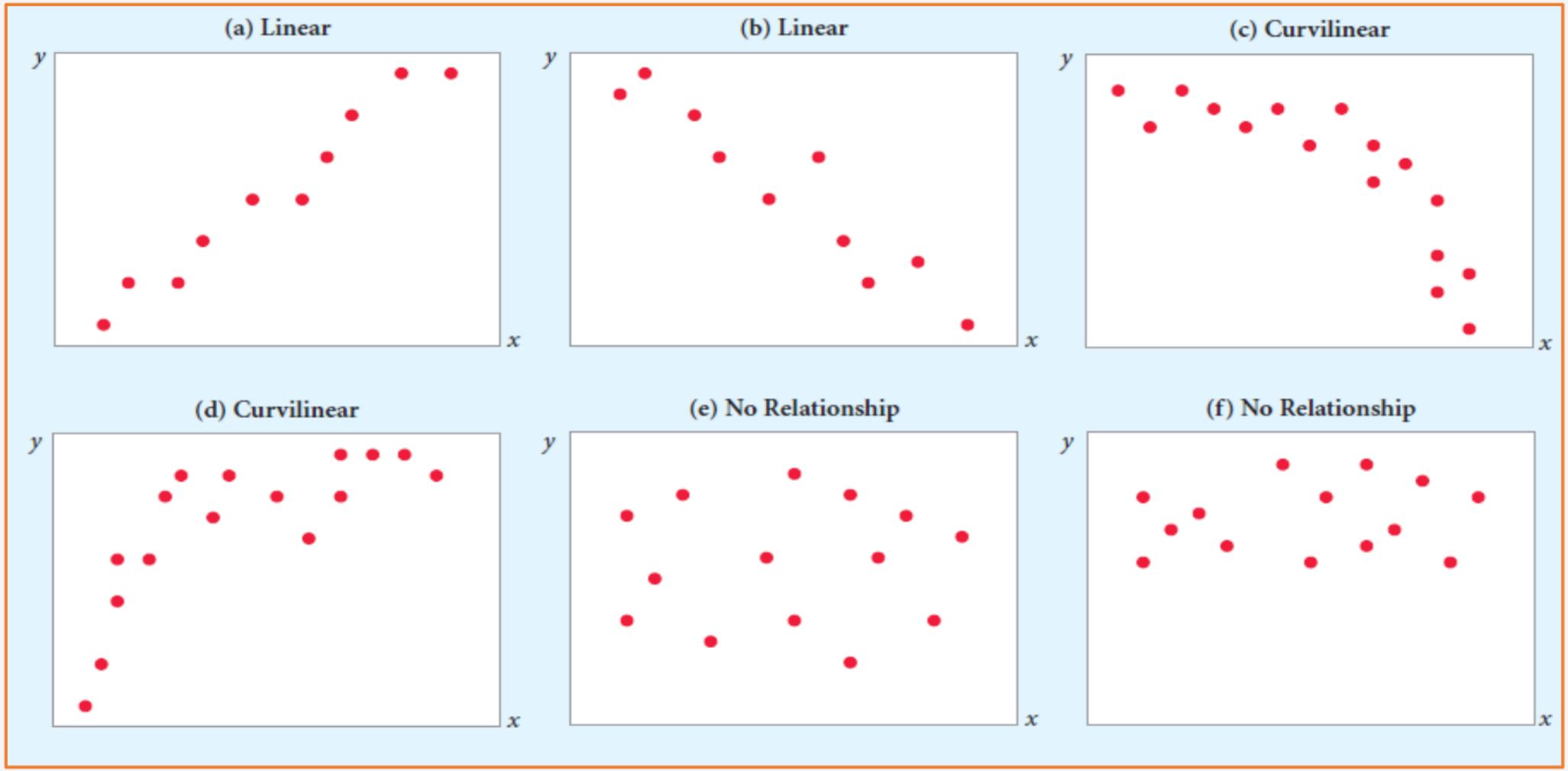
Linear Regression

Read Chapter 7 (Regression) of the Textbook

Regression Models

- To understand the application of regression analysis in data mining
    - Linear/nonlinear
    - Logistic (Logit)
- To understand the key statistical measures of fit

Relationships between variables



When the data shows linear relationship

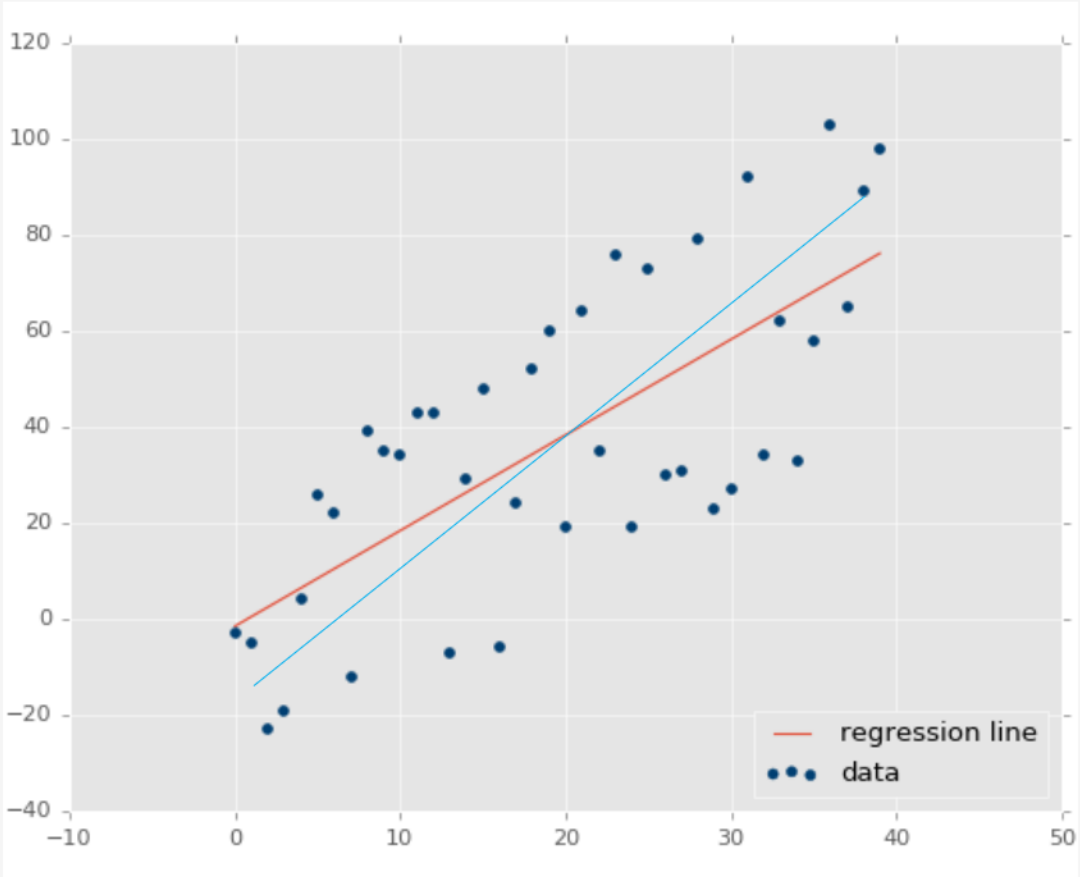
Correlation is high (positive or negative) and Scatter plots display a linear relationship

First model come to mind is

$$Y = m X + b$$

But still, there can be many lines that can “kind of” fit the data as well

Question: How to pick the “best-fit” line?

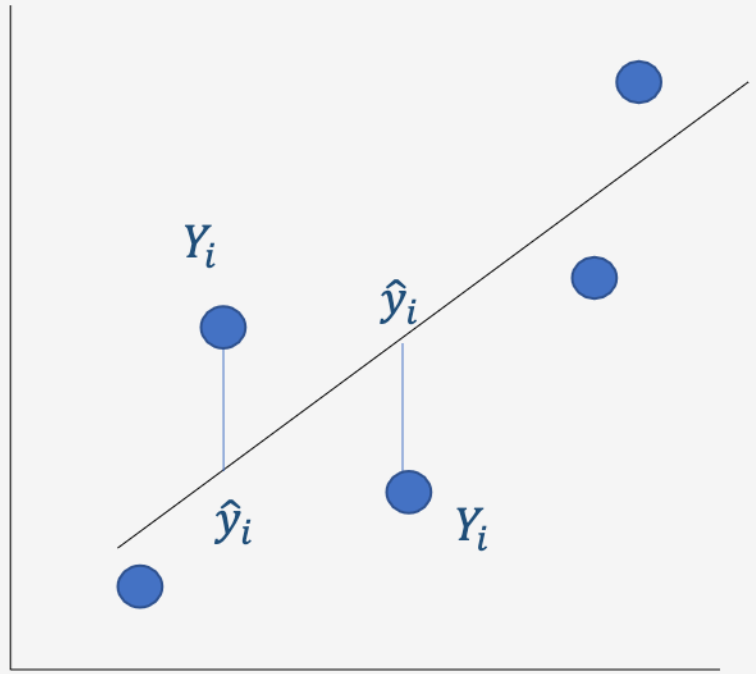


How to find the best fitting line?

Define Mean Squared Error (MSE)
To be the square of the distance between actual and predict Y values

$\hat{y}_i$ = prediction, $Y_i = actual\ value$

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



Best fitted line is the line that minimize the MSE =>

Least Square Methods

R-square as metrics for determining "goodness" of the fit

- Determining the relationship between predictor & outcome
- Relationship Among SST, SSR, SSE

$$r^2 = \text{SSR}/\text{SST}$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

Higher R-square => Lower SSE => Better Model

R-square is 0% to 100%, anything > 70% is great

Common Theme, Toolbox and Research workflow in Data Science

Apply different algorithms to solve different problems based on the same <Theme> and <Research Workflow>

Algorithms

- SVM
- KNN
- Naïve Bayes
- Neural Network
- Logistics Regression
- NLP

Problems

- Regression
- Classification
- Recommendation System
- Clustering
- Association

Common Theme, Toolbox and Research workflow in Data Science

Will use Linear Regression for many of the general practices in building models, some of them are

- Split the dataset into training set and a testing set
- Use standard metrics to judge model performance
- K-fold cross validation

Learning by doing