UNIVERSITY OF SAN FRANCISCO

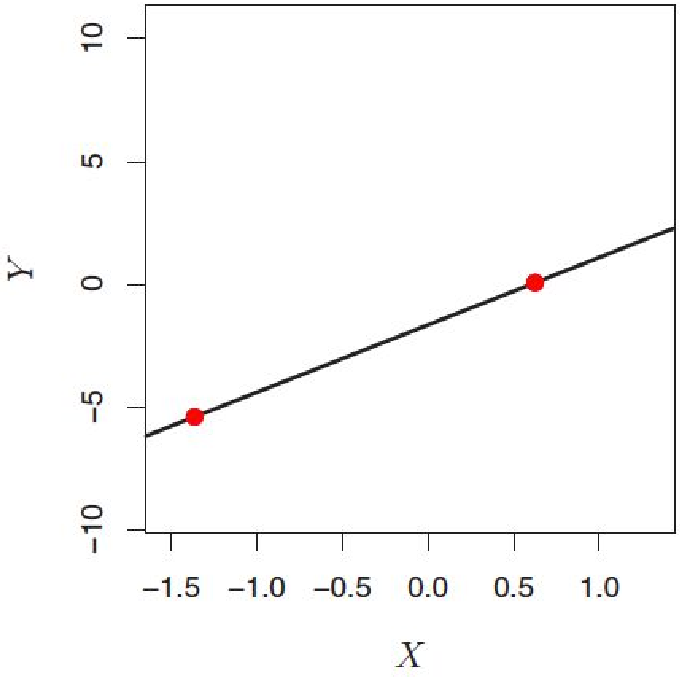
Master of Science in Analytics

Machine Learning in High Dimensions

Machine Learning 1

High-dimensions; high problems

- How high is high?
  - Any time n is close to or smaller than count of features (p)
  - More common as time moves forward
- Simple Linear Regression fails
  - Example case: n = 2 & p = 1
  - MSE_{Train} = 0; MSE_{Test} = ??
  - All σ^2-based techniques fail
    - C_p AIC, BIC
    - *Adjusted R^2*



What could work

- Generally: fitting less flexible models
  - Any technique which (aggressively) avoids overfitting
  - Forward stepwise selection, ridge regression, the lasso, PCA
- Estimating test error
  - Directly: use validation set / cross validation
  - Indirectly: make adjustment to training error, account for overfitting

Conceptual Example

- Blood pressure
  - Assume model is predicting blood pressure
  - Features: height, weight, single-nucleotide polymorphism (SNP)
  - SNP is a variation in a single nucleotide in a specific genomic position
  - There could be ~ 500,000 SNPs
- Task:
  - Find the feature sets which predict high blood pressure
  - May use (for example) forward stepwise selection to create model
- If you find a solution, you may say
  - Your model is *one of possibly several* to predict the outcome
  - Your model should be validated on independent data sets