Homework 03

Brandon Hosley

University of Illinois - Springfield
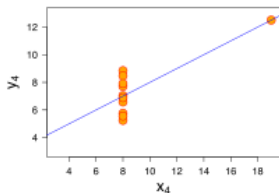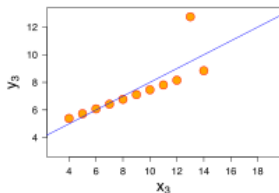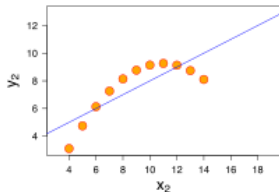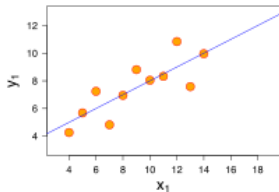
September 16, 2020

Overview

Linear Regression Problem 1: Simplicity

+ Computationally easy
- Can only accurately represent simple relationships
- Can only accurately represent linear relationships

Linear Regression Problem 2: Selection Bias

Linear regression is susceptible to selection bias

- A type of overfitting
- Occurs when a type of data is over-represented in test set

Average Annual Rainfall for Validation Period
(1976-1999)
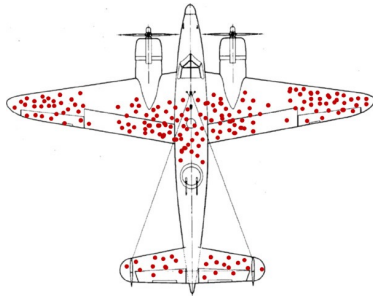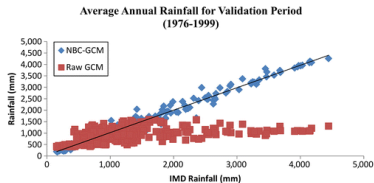
Linear Regression Problem 3: Multicolinearity

- When multiple predictors share a linear relationship
- Small changes are magnified in the model
- Heavily correlated predictors may cause redundancy in model
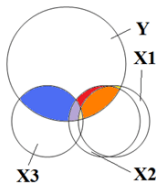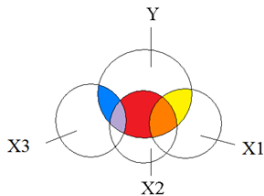
Linear Regression Problem 4: Heteroscedasticity

- Divergent data
- May be easily bound between two curves
- Difficult to accurately predict

Tibshirani Lecture: Linear Regression

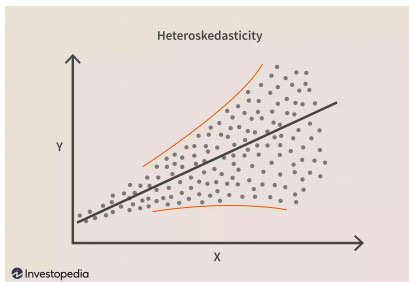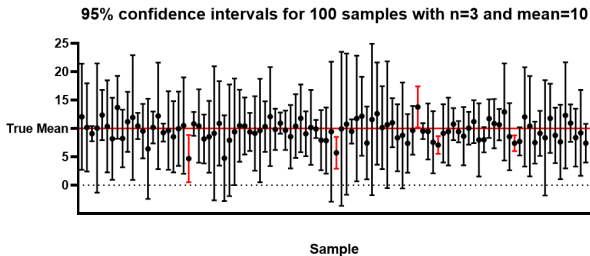- Linear Regression
  - Simple approximation method
  - Great for estimating slope of data
  - From this one may generate confidence interval

95% confidence intervals for 100 samples with n=3 and mean=10
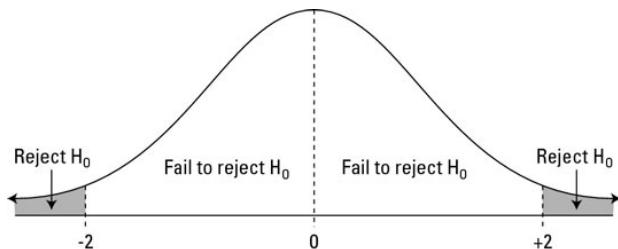
Sample

Tibshirani Lecture: Linear Regression

- Linear Regression
  - Simple approximation method
  - Great for estimating slope of data
  - From this one may generate confidence interval
- Hypothesis Testing
  - Testing against a null hypothesis
    H_0 : No relationship between X and Y.
  - Testing for the probability of independent variable distribution

Tibshirani Lectures: Multiple Linear Regression

- Multiple Linear Regression
  - Fitting data to a hyperplane instead of line
  - Works best when variables are independent
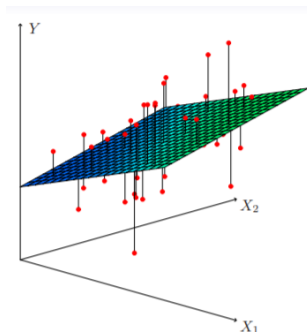
Tibshirani Lectures: Choosing Variables

- Selection from all subsets
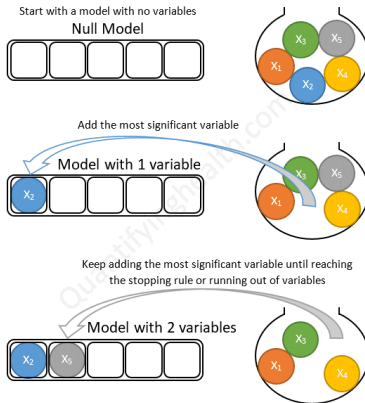
Tibshirani Lectures: Choosing Variables

- Selection from all subsets
- Forward Selection - adding variables with highest significance

Forward stepwise selection example with 5 variables:

Start with a model with no variables
Null Model

Add the most significant variable
Model with 1 variable

Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables

Tibshirani Lectures: Choosing Variables

- Selection from all subsets
- Forward Selection - adding variables with highest significance
- Backward Selection - removing variables with least significance

Backward stepwise selection example with 5 variables:

Start with a model that contains all the variables
Full Model

Remove the least significant variable

Model with 4 variables

Keep removing the least significant variable until reaching the stopping rule or running out of variables

Model with 3 variables