Regression Analysis

Model Selection

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Introduction

Georgia Tech

About This Lesson

Georgia Tech

Objectives

- **High Dimensionality:** When we have a very large number of predicting variables to consider, it can be difficult to interpret and work with the fitted model.
- **Multicolliearity:** When the predicting variables are correlated, it is important to select variables in such a way that the impact of multicolliearity is minimized.
- **Prediction vs Explanatory Objective:** The variables selected for the two objectives will most often be different.

➔ **Variable Selection** addresses all these concerns.

Georgia Tech

Implications and Words of Caution

- **Confounding vs. Explanatory Variables**
  - Consider research hypothesis as well as potential confounding variables
- **Targeted Predicting Variables**
  - Include target variable in model if specified by research hypothesis
- **Over-Interpretation**
  - Selected variables are not necessarily special!
    - Highly influenced by correlations between variables
    - Interpretation of regression coefficients
    - Causality vs. Association

Georgia Tech

No Magic Bullet

- Variable selection for large number of predicting variables is an "unsolved" problem in statistics
- In some sense, model selection is "data mining"
- Data miners / machine learners often work with many predictors
- There are no magic procedures to get you the "best model"

"All models are wrong, but some are useful." —George Box

Georgia Tech

Notation

Given

$S \subset \{1, \ldots, p\}$ a subset of indices

and

(x_j for $j \in S$) the subset of predicting variables with indices in S:

- $\widehat{\boldsymbol{\beta}}(S)$ is the vector of estimated regression coefficients for the submodel with $\boldsymbol{X}_S = (x_j \text{ for } j \in S)$ predicting variables
- $\widehat{\boldsymbol{Y}}(S)$ is the vector of fitted values for the submodel with $\boldsymbol{X}_S = (x_j \text{ for } j \in S)$ predicting variables
  - E.g., for regression assuming normality, $\widehat{\boldsymbol{Y}}(S) = \boldsymbol{X}_S\widehat{\boldsymbol{\beta}}(S)$

➔ I will refer to this model as the **S submodel**.

Georgia Tech

Summary