Techniques for Choosing Predictors

READING:	4.2
EXERCISES:	CH 4. 4-6
ASSIGNED:	HW 8
PRODUCER:	DR. MARIO

IMG CREDIT: ALEX RIEGERT-WATERS

Family of Models

- Consider the Linear Regression Model

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Full Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Single Variable Model:
  - $\hat{Y} = \beta_0 + \beta_1 X_1 + \epsilon$
  - $\hat{Y} = \beta_0 + \beta_2 X_2 + \epsilon$
- Empty Model: $\hat{Y} = \beta_0 + \epsilon$

Which Model is Best?

Family of Models

- When There are a Few Predictors We Can Fit All Possible Models
- As k Increases, the Number of Models in the Family Increases
- There are 2^k Different Linear Regression Models in the Family
- We Desire to Know What Predictors Need to Be Included
- Bad Option: Fit Full Model and **Remove** Insignificant Predictors

Family of Models

- Problem with Selecting Predictors
    - R-Squared **Always Increases** as I Add Predictors
    - Standard Error of Regression Always Decreases as I Add Predictors
    - Question: *Is the Relationship Between a Predictor and the Response Variable Due to Chance or Is There an Actual Relationship?*

- Typically, There are Multiple Competing Models that Are All **Similar**

- Ideally, We Want the **Simplest** Model that Explains the **Variation in Y**

Family of Models

- Hypothetical Situation From Dataset with 100 Potential Predictors
  - Model 1: *k=100, R-Squared = 0.90*
  - Model 2: *k=1, R-Squared = 0.89*
- Question: *Which of These Two Models is Better?*
- Goal: Find the "Best" Model, but Combat "Overfitting"

Best Subsets

- Computers are Fast = We Can Fit All Possible Models Most of the Time
- Good to Start By Removing **Unimportant** or **Correlated** Predictors
- Adjusted R-Squared
  - Accounts for the Complexity of the Model
  - Won’t Always Increase Just by Adding More Predictors
  - Find the Best Subset of Predictors that **Maximizes** *adj R-Squared*

Best Subsets

- Mallow's C_p
  - Most of the Criteria Depend Only On What is **in** the Model
  - Formula:

$$C_p = \frac{SSE_m}{MSE_k} + 2(m+1) - n$$

  - $m = \#\ of\ Predictors\ in\ Subset\ Model$
  - $SSE_m = Sum\ of\ Squared\ Residuals\ of\ Subset\ Model$
  - $MSE_k = Mean\ Squared\ Error\ from\ Full\ Model$
  - Find the Best Subset of Predictors that **Minimizes** Mallow's C_p

Best Subsets

- Akaike's Information Criterion (AIC)
  - Penalizes Based Off Complexity and Smaller is Better
  - Formula:

$$AIC = -2\log(Likelihood) + 2k$$

- Bayesian Information Criterion (BIC)
  - Penalizes Based Off Complexity and Smaller is Better
  - Formula:

$$BIC = -2\log(Likelihood) + \log(n)\,k$$

- Penalty in BIC is Larger than Penalty in AIC when $n>7$.

Model Selection Algorithms

- Backwards Elimination
  - Start by Fitting Full Model
  - Remove Variable with the **Largest P-value** and **>0.05**
  - Refit Model Without Previous Variable
  - Remove Variable with the **Largest P-value** and **>0.05**
  - Continue…
  - Stop Removing Variables Once All Variables Have a P-value <0.05

Model Selection Algorithms

- Positives and Negatives of Backwards Elimination
  - (+) We are Not Fitting All Subsets
  - (+) We Start By Seeing What the Full Model Looks Like
  - (+) We Make Use of All Predictor Variables
  - (-) Initial Models are Very Complex and Overfitted
  - (-) Still Likely to Lead to an Overfitted Model
  - (-) Multicollinearity Could Lead Us to Making a Mistake
  - (-) Once a Predictor is Gone It Will Never Return

Model Selection Algorithms

- Forwards Selection
  - Fit Model With **Only** the **Single** Predictor with Strongest ***r***
  - If **P-value > 0.05**, then **Stop** and **Remove**
  - Otherwise, Find the **Next** Predictor that Maximizes R-squared When Added to the Model
  - If **P-value > 0.05**, then **Stop** and **Remove**
  - Otherwise, Keep Repeating Until **P-value > 0.05**

Model Selection Algorithms

- Positives and Negatives of Forwards Selection
  - (+) Combats Against Multicollinearity Better
  - (+) Favors Smaller Models
  - (-) Typically Requires Fitting More Models
  - (-) Very Unlikely that the Full Model Ever Gets Fit
  - (-) May Mislead to Thinking Important Variables are Not Important

Model Selection Algorithms

- Stepwise Regression
  - Uses **Forward Selection** to **Add** Variables
  - Uses **Backward Elimination** to **Remove** Redundant Information
- Advice: Run All Three Algorithms to Identify Competing Models
- You Can Use **Better** Criteria to Determine When to **Add** or **Remove**
  - Adjusted R-Squared
  - Mallow's Cp
  - AIC or BIC

Example: Predicting Body Fat

- Question: *Can we predict the body fat percentage of an individual using the age of the individual and other body measurements?*

- List of Potential Predictor Variables in Dataset
  - Age (yrs)
  - Weight (lbs)
  - Height (in)
  - Circumference of Neck, Chest, Abdomen, Ankle, Biceps, Wrist (cm)

Supplement for Lecture 16

- Examine Multicollinearity and Look at VIF
- Use **regsubsets()** to Fit All Possible Models
  - Look at R-Squared, Adjusted R-Squared, Mallow’s Cp, and BIC
- Backwards, Forwards, and Stepwise Algorithms for Selecting Variables

Thank You

Make Reasonable Decisions