Model Building
Part 2: Procedures for Model Selection

STAT 705: Regression and Analysis of Variance

Variable Selection Procedures

- All possible regressions
  - For p predictors, there are 2^p possible models
    - » p = 3 ⇨ 2^3 = 8 possible models
    - » p = 4 ⇨ 2^4 = 16 possible models
    - » p = 8 ⇨ 2^8 = 256 possible models
    - » p = 10 ⇨ 2^{10} = 1024 possible models !!!
  - Generally, too many models to consider them all
  - Computer search algorithms can make this more efficient
  - Not something we want to do 'by hand'

Automated Search Methods

- Forward Selection
  - Starts with an intercept-only model, and adds predictors one at a time
- Backward Elimination
  - Starts with all predictors in the model, and removes them one at a time
- Stepwise
  - Combination of Forward Selection and Backward Elimination

Forward Selection

- Start with only the intercept (no predictors)
- Add the predictor that contributes the most to the fit of the model
- Continue adding “important” predictors until none of the remaining predictors contribute significantly to the fit of the model
- We specify
  - Level of significance for entering the model
  - How to measure “importance”
- Once a predictor enters the model, it never leaves

Backward Elimination

- Start with all predictors in the model
- Remove the least important predictor
- Continue removing “unimportant” predictors until all predictors remaining in the model are considered relevant to explain the behavior of Y
- We specify
  - Level of significance for exiting the model
  - How to measure “unimportant”
- Once a predictor is removed from the model, it never returns

Stepwise

- Combination of Forward and Backward
  - Start with no predictors
  - Perform one step of Forward Selection (i.e., add one "important" predictor)
  - Perform one step of Backward Elimination (i.e., remove one "unimportant" predictor)
  - Repeat one step of Forward and one step of Backward until no predictors are added or removed
- Predictors can move in and out of the model many times

Comparison of Methods

- Each method generates a set of predictors that are considered “important” for explaining the behavior of Y
- The methods can produce different sets of predictors
- We specify
  - Original collection of predictors that we want to consider, including any transformations, interactions, etc.
  - Criteria for comparing models (adjusted R^2, AIC, SBC, etc.)

Limitations

- The methods are automatic
  - Require no additional input from the user
  - Do not consider subjective criteria
  - Results must still be examined for relevancy
  - May generate models that
    - » are not defensible from a practical standpoint
    - » do not satisfy model assumptions
- Remember to exercise common sense in your final selection of a model

Hybrid Approach

When the number of potential predictors is very large

- Stage 1: Screening
  - Automatic search methods
  - Dismiss predictors with negligible effects
- Stage 2: Fine tune
  - Use all possible regressions to the reduced set of potential predictors

SAS Implementation

```
proc reg data = datasetname;
model Y = Z1 Z2 Z3 Z4 Z5 / selection = forward;
```

Use the 'selection' option on the model statement

- Automated methods
  - `selection = forward`
  - `selection = backward`
  - `selection = stepwise`
- All possible regressions
  - `selection = adjrsq`
  - `selection = cp`

A complete example is given in the file 'Example.SENICdata.pdf' on the course website. Your should review this example as part of today's lesson.

What You Should Know

- How to implement the model selection procedures in SAS
- How to interpret the SAS output for these procedures
- Use the results of these procedures to develop a linear model
- Justify (in statistical and/or logical terms) the choice of predictors in the final model
- Be able to write a report of the process and final results