Model Building
Part 1: Criteria for Model Selection

STAT 705: Regression and Analysis of Variance

Model Building

- How do we decide **how many** and **what** predictors to include in a statistical model?

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j Z_{ji} + \varepsilon_i$$

- Complex data sets can have dozens or more potential predictors
- Options:
  - Quantitative predictors: How many and which ones?
  - Transformations (squares, cubes, etc.): On what predictors?
  - Qualitative predictors
  - Interactions

Getting Started

- Ultimate goal
  - Find a set of predictors that fits the data well
  - We are NOT looking for the ‘best’ model
- Different strategies yield different subsets of “best” predictors
- Besides . . . “best” is really a relative term . . .



Model Misspecification

What is the danger?
Why care?

- Two few predictors
  - Biased point estimates
  - Consistently over- or under-estimates the magnitude of the relationship between Y and X's
- Too many predictors
  - Inflated variance of estimated parameters and predictions
  - Poor precision; reduced ability to find important differences

Considerations

- What is the nature of the study?
  - Controlled experiment
    - Blocking factors
    - Randomizations
  - Observational study
  - Combination: Part designed experiment, part observational

Design dictates the model!

- Prior knowledge and subject-matter expertise
  - Link predictors to the research objective
  - Include predictors that we want to make sure we adjust for

More Considerations

- Complexity of model
  - We want as few predictors as possible
  - But the model also needs to fit the data well
- Sample size
  - The complexity of the model (number of parameters) is limited by the available information
  - Rule of thumb: Need 6 to 10 observations for each predictor in the model
- *Always* check the model assumptions
  - Predictors may be needed, even if they do not contribute directly to the interpretation

Variable Selection

- It is not practical to look at every possible regression model
  - With 8 potential predictors there are 256 possible models
  - If we consider transformations (e.g. log(X), X^2), the possibilities are endless
- We need a systematic approach that will generate a few candidate models
  - Several different strategies can be used … resulting in several different candidate models
  - We take a closer look at the candidate models and use personal judgment to make a final choice
  - There is no guarantee that we will find the 'best' model

Criteria for Model Selection

- Summaries of how well the model fits the data (these are used to compare models)
  - Coefficient of determination R^2
  - Adjusted R^2
  - Residual Mean Square (ie. MSE)
  - Mallow's C_p
  - Akaike's Information Criterion (AIC)
  - Schwarz' Bayesian Criterion (SBC)

Goal: Estimating, explaining the data

  - PRESS (PREdiction Sum of Squares) Criterion

Goal: Predicting new observations

Basic Steps

1. Fit a series of competing models to the dataset
2. Compare the models using model selection criteria
3. Decide upon good candidate models
4. Among the candidates, check model assumptions
5. Make a final decision

Exercise common sense -- pay attention to

- Multicollinearity issues
- Outliers, influential points
- Subject matter expertise (may require some predictors to be in every candidate model)

Coefficient of Determination R^2

- Proportion of total variability in Y that is associated with the predictors fitted in the regression model

$$0 < R^2 = \frac{SSReg}{SSTot} = 1 - \frac{SSE}{SSTot} < 1$$

- We want R^2 to be large
- Adding more predictors *de*creases SSE, which *in*creases R^2
  - R^2 is not appropriate for choosing between models with different number of predictors
  - R^2 can be helpful to select between competing models that have the same number of predictors

Adjusted R^2

- When a new predictor enters the model,
  - SSE gets smaller -- this is good
  - We lose 1 df for Error -- this is not good
- Adjusted R^2 attempts to answer the question:

  Does the decrease in SSE offset the loss in error degrees of freedom?
- Adjusted R^2 can either increase or decrease as new predictors enter the model
  - It may be negative (for a really bad model)
  - It will never be larger than 1

Residual Mean Square

- This is another name for MSE
- Same criterion as adjusted R^2
  - These two criteria will always generate the same subset of predictors

Mallow's C_p

- Attempts to balance
  - Mistake of excluding important predictors
  - Including too many predictors
- Full model has all p predictors
- Calculate C_p on a reduced model

$$C_p = \frac{SSE(reduced)}{MSE(full)} - \left(n - 2 \cdot \#\,parameters\,in\,reduced\,model\right)$$

- Assumes that full model has no bias (C_p = # parameters)
- Desirable models have C_p close to or less than # parameters
- Poor models have C_p much larger than # parameters

AIC and SBC

$$AIC = n \cdot \log(SSE) - n \cdot \log(n) + 2p$$

$$SBC = n \cdot \log(SSE) - n \cdot \log(n) + p \cdot \log(n)$$

- Smaller values indicate better fit
- The '2p' and 'p·log(n)' are penalties associated with the number of parameters (p) in the model
- Penalty is heavier for SBC than AIC
  - SBC encourages smaller models
- General rule of thumb
  - A decrease of 2 or more points usually indicates a substantial improvement in model fit

PRESS

- PRESS is the preferred criterion if we want to use the fitted model to predict new observations
- Prediction Sum of Squares
  1. Delete the i^{th} observation
  2. Estimate the regression equation with the remaining (n-1) observations
  3. Predict the value of the i^{th} response
  4. The deleted residual is the difference between observed and predicted response
- PRESS is the sum of all the squared deleted residuals
- Small PRESS values are desirable (small prediction errors)

Which Criterion is “Best”?

- Criteria include R-square, adjusted R-square, Mallow’s Cp, RMSE, AIC, SBC, and PRESS
- We do not need to use all of these criteria to build a model
- Different researchers have different preferences – no one criterion is “best” for all situations

FOR DESIGNED EXPERIMENTS

IT IS THE SCIENTIFIC QUESTION, AND NOT THE DATA, THAT DRIVES THE MODELING APPROACH

AVOID DATA SNOOPING!

Snooping can generate spurious results ("flukes") that are not reproducible

What You Should Know

- Criteria for comparing models
  - Know what they are
  - How they are calculated
- Know that PRESS is different
  - prediction vs. estimation

In the next lesson, we work through an example that uses these criteria to choose a model. (We make SAS do the calculations.)