STAT 215
Model Selection II

Colin Reimer Dawson

Oberlin College

November 7, 2017

Outline

So many models...

- How to decide among all these models?
  1. Understand the subject area! Build sensible models.
  2. Nested F-tests
  3. Model quality measures

What Makes a Good Model?

Fit

High R^2
Small SSE
Large F

Validity

Strong evidence for predictors
Simple (Parsimonious)
Generalizes outside sample

Why Does Parsimony Matter?

Don't we just care about good predictions?

Not exclusively...

- We also use models to *understand* the world (harder with more complexity)

And even so...

- We really care about making predictions for data we *haven't seen yet*.

Criteria to "score" models

1. high R^2/low SSE/low $\hat{\sigma}_\varepsilon^2$: always prefers more complex models
2. Adj. R^2: balances fit and complexity
3. Mallow's C_p / Akaike Information Criterion (AIC): estimates mean squared prediction error based on $\hat{\sigma}_\varepsilon^2$ from a "full" model
4. Out-of-sample predictive accuracy

Mallow's C_p / AIC

Two measures that reduce to the same thing in the case of MLR with independent, equal variance, Normal residuals. For a "reduced" model with $p_{reduced}$ total parameters (including the intercept) which is nested in a "full" model with p_{full} parameters, both fit using n observations:

$$C_p = \frac{SSE_{\text{reduced}}}{MSE_{\text{full}}} + 2p_{\text{reduced}} - n \tag{1}$$

$$= p_{\text{reduced}} + \frac{SSE_{diff}}{MSE_{full}} \tag{2}$$

where smaller values indicate a simpler model (smaller $p_{reduced}$) and/or a better fit (smaller SSE_{diff})

Cross-Validation

Validation is a technique whereby the full dataset is divided into training and validation (held-out) sets. The first is used for fitting parameters; the second for evaluating predictive power.
Cross-validation uses all the data but gives each piece a turn as the validation set.

Versions:

1. Two-fold: Divide data (randomly) in half. Fit two models, exchanging roles of training and validation.
2. k-fold: Divide data into k equal sized sets, fit k models letting each set as the validation set.
3. Leave-one-out (n-fold): Let each observation be its own validation set. Requires fitting n models.

Model Selection

Five predictor-selection methods:

1. Domain knowledge (+ a few F-tests)
2. Best subset
3. Forward selection
4. Backward selection
5. Stepwise selection

Automated exploration of predictor subsets

1. Best subset: consider all possible combinations (2^k)
2. Forward selection: start with null model, and consider adding one predictor at a time
3. Backward elimination: start with full model and consider removing one predictor at a time
4. Stepwise regression: alternate forward selection and backward elimination

Note: Choose best step based on adj-R^2 or C_p/AIC, *not* based on P-values

Model Selection

"Search"	"Scoring" $R^2_{adj.}$	"Scoring" C_p	"Scoring" CV Error
Domain Knowledge			
Best Subset			
Forward Selection			
Backward Selection			
Stepwise Selection			

Example: Baseball Win %

Demo