Lecture 16

Model Building: Automated Selection Procedures

STAT 512

Spring 2011

Background Reading

KNNL: Sections 9.4-9.6

Topic Overview

- Review: Selection Criteria
- CDI/Physicians Case Study
- “Best” Subsets Algorithms
- Stepwise Regression Methods
  - Forward Selection
  - Backward Elimination
  - Forward Stepwise
- Model Validation

Questions in Model Selection

- How many variables to use?
  - Smaller sets are more convenient, particularly for parameter *interpretation*.
  - Larger sets may explain more of the variation in the response and be better for prediction.
- Which variables to use?
  - F-tests, t-tests
  - Statistics (e.g., R^2, Mallows C_p)

Model Selection Criteria

- General Linear Tests
- Compare models of the same size using R^2 (maximize)
- Compare different sized models using adjusted R^2 (max) or AIC/SBC (min)
- Mallow's C_p Criterion for predictive ability of the model (minimize compared to p)
- PRESS statistic for predictive ability (measures prediction error, smaller is better)

Physicians Case Study

(cdi_modelselect1.sas)

What is the best model for predicting the # of active physicians in a county? Possible predictors include:

1. X1 = Total Population (dropped - multicollinearity)
2. X2 = Total Personal Income
3. X3 = Land Area
4. X4 = Percent of Pop. Age 65 or older
5. X5 = Number of hospital beds
6. X6 = Total Serious Crimes
7. X7 = Percent of population HS Grads
8. X8 = Percent Unemployment

Subset Models

- Number of possible models is:
  $2^{p-1}=2^7 = 128$
- Our previous code (from Lecture 15) gets the 3 “best” models for each possible number of variables
- Results in 19 models for us to compare (6*3+1=19)

Best Model?

- C_p suggests 5-6 variables
- Minimal AIC for 5-variable model excluding pop_eld and land_area.
- Maximum Adjusted R^2 for 6-variable model excluding land area.
- Follow up by looking at diagnostics, evaluating goals of study, and possibly confirming through model validation to choose final model.

Algorithmic Approaches

- Number of possible models is 2^{p-1}; this grows exponentially.
- If p is small (4 or 5), can compare all possible models.
- For larger p, do not want to have to look at thousands of models (even computers can struggle); hence need algorithmic procedure to decide on subset of variables that makes the "best" model.
- Often identify "good" subsets along the way

"Best" Subsets algorithms

- Get the best *k* subsets of each size according to a specific criterion.
- We used this in the example, looking at the best three subsets of each size, and using R^2 as our criterion.
- With computers, fairly easy to use for 5-40 variables; anything more can begin to require excessive processing time.

"Best" Subsets Algorithm (2)

- Usually results in a few models from which to choose; different statistics can point to different models.
- Should consider advantages/disadvantages of each (e.g. prediction vs. interpretation)
- Often there is no "right" answer – you have to use judgment.

Stepwise Regression

- Uses t-statistics to “search” for model.
- Based on t-statistic, choose one variable to add to (or delete from) the model.
- Repeat until no variables can be added or deleted (based on the alpha values you set initially).

Stepwise Regression (2)

- *Forward Selection* – From group of variables that "can" be added, add to the model the one with the largest "variable added-last" t-statistic.

- *Backward Elimination* – Start with full model and delete variables that "can" be deleted, one by one, starting with the smallest "variable-added-last" t-statistic.

Stepwise Regression (3)

- *Forward Stepwise Regression* – Combine forward selection with backward elimination, checking for entry, then removal, until no more variables can be added or removed.
- Each procedure requires ***only*** that we set significance levels (or critical values) for entry and/or removal. Once this is done, each has exactly one result.

Physicians Example

```
proc reg data=cdi outest=fits;
  model lphys = tot_pop tot_income land_area
    pop_elderly beds crimes hsgrad unemploy
    /selection=stepwise slentry=0.25
     slstay=0.1;
run;
```

- Checks to add variables with alpha=0.25, then delete variables with alpha=0.10.
- Note: If add/delete same variable, then procedure terminates.

Significance Levels

- Often we specify alpha levels that are somewhat liberal in terms of allowing variables in the model.
- The reason to do this is to keep the model selection procedure from getting "stuck" too early in situations where there is an abundance of intercorrelation.
- In SAS, specify slentry= and slstay= (defaults are 0.15 for stepwise)

Results

- For each variable added/removed, SAS provides ANOVA table, and usual statistics.

- Summary follows at the end of the output

Results (2)

Step	Var Entered	Var Removed	Num Vars	Part R-Sq	Model R-Sq	C(p)	F Value	Pr > F
1	beds		1	0.6265	0.6265	162	731.21	<.0001
2	tot_income		2	0.0651	0.6916	59.9	91.86	<.0001
3	hsgrad		3	0.0285	0.7201	16.5	44.13	<.0001
4	tot_pop		4	0.0062	0.7263	8.6	9.84	0.0018
5	crimes		5	0.0018	0.7281	7.7	2.85	0.0923
6	unemploy		6	0.0016	0.7297	7.2	2.54	0.1120
7		tot_pop	5	0.0015	0.7281	7.6	2.42	0.1203

- Note that total population is added at step 4, then removed at step 7 (often happens when variables are intercorrelated).

Forward Selection

- Some disadvantages; generally not the best choice of selection procedure
- For initial phases, MSE can be biased (too large) since important variables aren't yet in the model.
- Will not adjust in any way for intercorrelation as we can do if we allow both addition and removal.

Backward Elimination

- Backward Elimination has some advantage in that MSE will start out reasonable

- If start with 1-variable models, MSE may be inflated because other important predictors are removed.

Physicians Example

- Backward and stepwise both lead to the same model in this case.
- Forward does not work well; gets stuck with the total income and total population in the model together; which is not the best predictive model, nor is it good for interpretation because of high inter-correlation between these variables.

Selection Options in PROC REG

- STEPWISE, FORWARD, or BACKWARD

- RSQUARE, ADJRSQ, or CP (used in conjunction with BEST=*n*)
  - Note: Cannot obtain PRESS statistic when these are specified.

Other Options in PROC REG

- BEST = <*n*> used to get the best *n* subset models.
- INCLUDE = <n> used to force the first *n* variables listed into the model
- DETAILS = <option> indicates what summary statistics you want (ALL, STEPS, or SUMMARY)

BEST= option

```
proc reg data=cdi outest=fits;
  model lphys = tot_income land_area pop_elderly beds
                crimes hsgrad unemploy
    /selection=adjrsq best=4;
run;
```

Number in Model	Adjusted R-Square	R-Square	Variables in Model
6	0.7254	0.7291	tot_income pop_elderly beds crimes hsgrad unemploy
5	0.7250	0.7281	tot_income beds crimes hsgrad unemploy
7	0.7248	0.7292	tot_income land_area pop_elderly beds crimes hsgrad unemploy
6	0.7244	0.7282	tot_income land_area beds crimes hsgrad unemploy

Things to Note

- Be cautious when employing an automated procedure such as backward/forward/ stepwise regression
- Only one model is selected
  - Not guaranteed to be the "best"
  - There may be other, more parsimonious or reasonable models, but which the particular heuristic method employed does not find.

Things to Note (2)

- PROC GLMSELECT is a newer procedure (added in v.9.1) that provides a some additional options, including a couple of new selection methods:
  - Least Angle Regression (LAR)
  - LASSO

- We won't use these methods or procedures in this class, just know that it is available if you need to do model selection in your research.

Model Validation

- See how well model fits with:
  - Newly collected data
  - Results from theoretical expectations
  - A holdout (validation) sample

- Collecting new data preferred, but often not practical or feasible.

Model Validation (2)

- Mean Square Prediction Error

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*}$$

  - Y_i=value of response in i^{th} validation case
  - $\hat{Y}_i$=predicted value for i^{th} validation case based on model selected in model-building data set.
  - n^*=number of obs. in validation dataset

Model Validation (3)

- Use old model to predict for new data, then compute MSPR.
- If MSPR is fairly close to MSE suggests model is reasonable.
- If MSPR is much larger than MSE, suggests that one should use MSPR rather than MSE as an indicator of who well the model will predict in the future.
- See Section 9.6 for more information.

Upcoming in Lecture 17...

- Diagnostics to Identify Outliers & Influential Observations