Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use

Peter L. Flom[1,2,3] David L. Cassell[4]

National Development and Research Institutes, Inc.

BrainScope, Inc.

Peter Flom Consulting

Design Pathways, Inc.

NESUG, November, 2007

Outline

Outline

Introduction

Garbage in, fake pearls out

Sorting pearls from garbage

Garbage in, total garbage out

Garbage disposal: Some solutions

Summary and further reading

Outline

Outline

Outline

Outline

Outline

The problem

- Too many IVs in regression
- Not sure which are good
- Need some help in speeding the selection process
- Problem exists in many types of regression

Extent of this talk

- Brief theory on stepwise
- Various example data sets
- No bootstrapping, etc.
- PROC GLMSELECT, lasso and lars
- Only OLS regression
- 'Stepwise' used for forward, backward, stepwise etc.

Some theory on why stepwise is bad

- The basic problem - one test vs. many
- The result:
  - Standard errors too small
  - p-values too small
  - Parameter estimates biased away from 0
  - Models too complex

Outline

Introduction

- You hope that you never do a regression with all noise
- If you do, you hope that the output says it's all noise

100 cases, 50 variables

- For the first test, we ran a regression with 100 subjects and 50 independent variables — all noise
- The defaults in stepwise are SLE = .15, SLS = .15
- The final model included 15 IVs, 5 sig at $p < .05$
- Forward: default SLE = .50, 29 IVs, 5 sig at $p < .05$
- Backward: default SLS = .10, 10 IVs, 8 sig at $p < .05$

1000 cases, 50 variables

- That's a lot of IVs per subject, but with N = 1000
- The final stepwise model had 10 IVs, again, 5 sig. at $p < .05$
- Forward: 28 IVs, 5 sig. at $p < .05$
- Backward: 8 IVs, 5 sig. at $p < .05$

Outline

Introduction

- More often, some of your IVs are real and some are noise
- Here, you want to sort the pearls from the garbage

100 cases, 50 + 1 variables

- If we add one IV that is linearly related to the DV, r = .32
- Stepwise with the default settings has 4 IVs, including the real one
- Backward: Real one plus 1
- Forward: Real one plus 23

1000 cases, 50 + 1 variables

- With 1000 cases, 51 IVs, one real, same r
- Stepwise: Real variable is included, but so are 9 others, 5 at .05
- Forward: Real and 27 others, 5 at .05
- Backward: Real and 6 others, all at .05

Outline

A single outlier in a perfect world

- Sometimes, you violate assumptions
- Outliers and leverage points happen
- N = 100, 5 noise IVs, 1 real, 1 outliers
- Forward: real variable included, and 2 others, param est on real was .76 (not 1)
- Stepwise and backward: Only real variable included, but param est now .72 (not 1)

Multiple outliers in that perfect world

- N = 100, 5 noise IVs, 1 real, 2 outliers
- Stepwise and backward: Only real variable included, but param est now .44 (not 1)
- Forward: Real and 2 others, parameter est = .44

Outline

Introduction

- If you can, the best solution is expert knowledge
- If the key is not the particular IVs and explanation: PLS, multimodel averaging
- If there are a small number of sensible models: AIC or BIC
- Otherwise.....LASSO or LAR

PROC GLMSELECT - introduction

- Experimental PROC in v9
- Download from SAS website
- Implements a variety of model selection schemes
- Has a variety of cross-validation methods
- Not intended to replace PROC GLM or REG, too few options

PROC GLMSELECT — key syntax

```
PROC GLMSELECT <options>;
CLASS variable;
MODEL variable = <effects></options>;
SCORE <DATA = dataset> <OUT = dataset>;
```

PROC GLMSELECT statement — key options

- DATA =
- TESTDATA =
- VALDATA =
- PLOTS =

MODEL statement - selection options

- Forward
- Backward
- Stepwise
- Lasso
- LAR

MODEL statement - choose options

- The CHOOSE = *criterion option chooses from a list of models based on a criterion*
- *Available criteria are: adjrsq, aic, aicc, bic, cp ,cv, press, sbc, validate*
- *CV is residual sum squares based on k-fold CV*
- *VALIDATE is avg. sq. error for validation data*

MODEL statement - stop options

- The STOP = *criterion option stops the selection process.*
- *Available criteria are: adjrsq, aic aicc, bic, cp cv, press, sbc, sl, validate*

MODEL statement - some other options

- HIERARCHY =
- CVDETAILS= AND CVMETHOD=
- STATS =
- STB

Uses of GLMSELECT

- You can combine the options in lots of ways. e.g.
  `selection = forward(stop = AIC sle = .2)`
  `selection = forward(stop = 20 choose = AICC)`

Brief theory of LASSO

- Least Absolute Shrinkage Selection Operator — Developed by Tibshirani (1994)
- Shrinkage method
- Constrains the sum of the absolute regression coefficients
- Center and scale all variables then minimize

$$||y - X\beta||^2 \quad \text{subject to} \sum_{j=1}^{m} |\beta_j| \leq t$$

LASSO with defaults applied to the above problems

- N = 100, 50 IVs, all noise . . . none selected
- N = 1000, 50 IVs, all noise . . . none selected
- N = 100, 50 noise variables, 1 real . . . none selected
- N = 1000, 50 noise variables, 1 real . . . only real selected
- N = 100, 5 noise variables, 1 real, 1 outlier . . . . param est now .99
- N = 100, 5 noise variables, 1 real, 2 outliers . . . .no variables included

Brief theory of LAR

- Least Angle Regression - developed by Efron, Hastie, Johnstone & Tibshirani (2004)
- All variables are centered, covariates are scaled
- Starts with all parameters = 0
- Adds parameters based on correlations with current residual
- Results on above problems essentially identical with LASSO

Outline

Summary

- In any statistical problem, the key is substantive knowledge
- If that is not available, then methods such as LASSO and LAR are better than standard methods

Further reading

- On the general problem:
  1. Harrell: Regression modeling strategies
  2. Burnham and Anderson: Model selection and multmodel averaging
- On LASSO and LARS
  1. Efron et al. (2004). Least Angle Regression (with discussion). *Annals of Statistics, 32, 407-499.*
  2. *Tibshirani (1996). Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society, series B, 58, 267-288.

Contact information

Peter L. Flom
peterflomconsulting@mindspring.com
(917) 488 7176

David L. Cassell
mathematical statistician
Design Pathways
3115 NW Norwood Pl.
Corvallis OR 97330