

New Methods and Software for Variable Selection in Regression

Chris Fraley & Tim Hesterberg
Insightful Corporation

JSM 2006, Seattle



Supported by NIH SBIR Phase I 1 R43 GM074313-01

- ▶ Significance
- ▶ Estimation methods and their relationship
- ▶ Available software
- ▶ S+GLARS package
- ▶ Plans for future development

Predicting outcomes based on covariates, and determining which covariates most affect outcomes, are among the most important problems in statistics.

$$y \sim \text{Model}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots)$$

response y ; predictors/variables x_i ; coefficients β_i

Goals in model (predictor) selection include:

- ▶ accuracy
- ▶ parsimony
- ▶ interpretability
- ▶ stability
- ▶ avoiding bias in inference

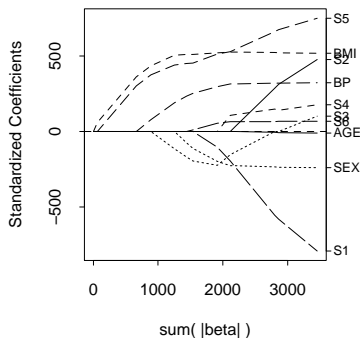
Estimation Methods

- ▶ methods that successively add or delete variables:
stepwise, forward stagewise, all subsets (leaps and bounds), ...
- ▶ penalized approaches:

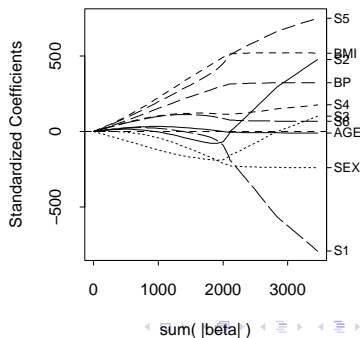
l_1 penalty: $y \sim \text{Model}(X\beta) + \lambda \sum |\beta_i|$ (lasso)

l_2 penalty: $y \sim \text{Model}(X\beta) + \lambda \sum \beta_i^2$ (ridge regression)

LASSO



Ridge Regression



Least Angle Regression

Efron, Hastie, Johnstone, Tibshirani *The Annals of Statistics*, 2004
(with discussion)

- ▶ new method “Least Angle Regression” that generates models by adding one variable at a time
- ▶ close relationship to forward stagewise (Hastie, Tibshirani, Friedman, 2001) and the lasso (Tibshirani, 1996)
- ▶ simple data-based rule (C_p statistic) for model-selection

Stepwise / Least Angle / Forward Stagewise / Lasso

All of these methods

- ▶ start with only the intercept in the model
- ▶ increment coefficients in the least-squares direction.

Stepwise Regression:

- ▶ Full least-squares fit incorporating new predictor that is most correlated with the response.

Least Angle regression:

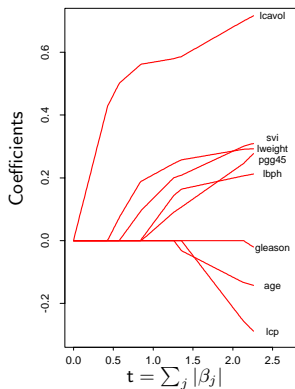
- ▶ Incremented only until another variable becomes as correlated as the current set of predictors.

Forward Stagewise / Lasso:

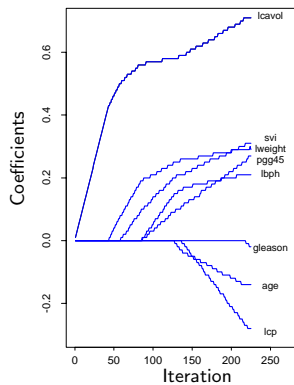
- ▶ Increment usually by a smaller amount.
- ▶ Predictors may leave / re-enter the model.

Prostate Cancer Data

Lasso



Forward Stagewise



Least-Squares Direction

$$\min \left\| y - X^{(k)}(\beta^{(k)} + \Delta\beta^{(k)}) \right\|_2^2$$

$\beta^{(0)} = 0$; $X^{(0)}$: predictor(s) most correlated with y

least-squares direction: $\Delta\beta^{(k)} \equiv (X^{(k)T} X^{(k)})^{-1} X^{(k)T} r^{(k)}$

$$r^{(k)} \equiv y - X^{(k)}\beta^{(k)}$$

LAR, lasso, forward stagewise, . . . , move along $\Delta\beta^{(k)}$

LAR adds new predictors based on correlation with $r^{(k)}$

Some Available Software in R and/or S-PLUS

lasso : Tibshirani (1995)
linear models with ℓ_1 constraint

lasso2 : Lokhorst, Turlach, Venables (1999)
linear and generalized linear models with ℓ_1 constraint

brdgrun : Fu (2000)
shrinkage estimator for LMs with bridge penalty $\lambda \sum |\beta_j|^\gamma$

lars : Efron and Hastie (2004)
linear regression: least angle, lasso, forward stagewise

elasticnet : Zou and Hastie (2005)
LM with penalty $\lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$

glasso : Kim, Kim, Kim (2005) generalized lasso:
multiple ℓ_1 constraints; multiple linear predictors

glmpath : Park and Hastie (2006)
GLM and Cox proportional hazards

- ▶ S-PLUS and R, open source
 - ▶ User-friendly
 - ▶ Consistent interface (GUI and non-GUI)
 - ▶ Visualization Tools
 - ▶ Documentation
 - ▶ Incorporate some existing packages: `lars`, `glm`, `path`
 - ▶ Methods: `plot`, `print`, `predict`, `cv`, `coef`
 - ▶ Framework for future work by Insightful and others
- ▶ Possible Extensions
 - ▶ Improved efficiency and accuracy of computations
 - ▶ Variables: factors, splines, polynomials, interactions, ...
 - ▶ Models: robust regression, survival analysis, ...
 - ▶ Tools for diagnostics
(including variable, model, and tuning parameter selection)
 - ▶ Missing data
 - ▶ Big data sets

S+GLARS (continued)

- ▶ NIH SBIR funding
 - ▶ research / risk
 - ▶ commercial potential
 - ▶ Phase I completed, software available for alpha testing
 - ▶ Phase II pending ...
- ▶ Outside contributors and collaborators
- ▶ Indirect benefit for Insightful Corporation
 - ability to ship with S-PLUS, other products

- ▶ Goals:
 - ▶ Turn research into software for use by a much larger community
 - ▶ Establish a high standard in terms of ease of use and robustness
- ▶ Projects: resampling, missing data, group sequential designs, simulation-based econometric software, functional data, stable distributions, proteomics, frailty models, ...
- ▶ External funding — SBIR / STTR grants (NIH, NSF, DOD, ...)
- ▶ Collaboration with academic and industrial researchers (we're always looking for good projects and collaborators)