Regression Analysis

Model Selection

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Regularized Regression: Approaches

About This Lesson

Variable Standardization & Notation

For regularized regression, center each column's values at zero and rescale so that the sum of squares of each column's values is 1. That is,

- Rescale the values for each j-th predicting variable, x_j, $j=1,.., p$, as follows:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$$

- It is also recommended to rescale the response variable in the same way:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = 0 \text{ and } \frac{1}{n}\sum_{i=1}^{n} y_{ij}^2 = 1$$

➔ **Use the original scale when fitting the selected model for interpretation of the regression coefficients.**

Ridge Regression

- Minimizes SSE plus the penalty the penalty term

$$SSE_{\lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

- Provides closed-form estimate of regression coefficients ($\hat{\boldsymbol{\beta}}$)

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

  - $\mathbf{I}$ is the identity matrix
- $\lambda = 0$ gives least squares estimate (low bias, high variance)
- $\lambda \rightarrow \infty$ gives $\hat{\boldsymbol{\beta}} \rightarrow 0$ (high bias, low variance)
- Commonly used under multicollinearity
- Not used for model selection
  - Shrinks but does not “force” any $\hat{\beta}_j$ to equal 0

LASSO Regression

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator
- Normal Linear Regression minimizes

$$SSE_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Generalized Linear Model minimizes

$$SSE_\lambda(\boldsymbol{\beta}) = -\ell(\beta_0, \cdots, \beta_p) + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - $\ell(\boldsymbol{\beta})$ is the log-likelihood function
- Estimated regression coefficients
  - Must use numerical algorithms
  - No closed-form expression
- Used for model selection
  - Does “force” some $\hat{\beta}_j$ to equal 0

LASSO Regression

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator
- Normal Linear Regression minimizes

$$SSE_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Generalized Linear Model minimizes

$$SSE_\lambda(\boldsymbol{\beta}) = -\ell(\beta_0, \cdots, \beta_p) + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - $\ell(\boldsymbol{\beta})$ is the log-likelihood function
- Estimated regression coefficients
  - Must use numerical algorithms
  - No closed-form expression
- Used for model selection
  - Does “force” some $\hat{\beta}_j$ to equal 0

- LASSO performs estimation of regression coefficients and variable selection simultaneously.
- The regression coefficients obtained from LASSO are less efficient than those obtained from Ordinary Least Squares (OLS).

➔ **After using LASSO to select the model, use OLS to estimate the (final) regression coefficients.**

Georgia Tech

Choosing λ: Cross-Validation

Split the data $\{(x_{11}, \cdots, x_{1p}), y_1\}, \cdots, \{(x_{n1}, \cdots, x_{np}), y_n\}$ into two sets.

- **Training set**
  - Use to fit the penalized model
    - Given l, estimate $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$
- **Testing/Validation set**
  - Use to evaluate performance of model obtained with training set
    - Estimate mean squared error (MSE) for normal regression
    - Estimate classification error rate for logistic regression
    - Estimate sum of squared deviances for Poisson regression
    - Generally, estimate a scoring rule depending on the regression problem

The process can be repeated for multiple λs.

Cross Validation: How to Split Data?

K-fold cross-validation (KCV)

- Divide data into K chunks of approximately equal size
- For a range of λ penalty values, e.g., $\lambda_1, \cdots, \lambda_B$, and for $k = 1 \text{ to } K$
  - The training set consists of data without the $k\text{-th}$ fold of data, and the testing set consists of the $k\text{-th}$ fold
  - Given λ, fit a model on the training data and predict responses
  - Given λ, compute mean squared error or classification error rate for the $k\text{-th}$ fold testing data
  - Given λ, after K folds have been processed, compute overall error (e.g., MSE or classification error) for that λ for all folds

➔ Select λ penalty providing minimum overall error

Ridge vs. LASSO Regression

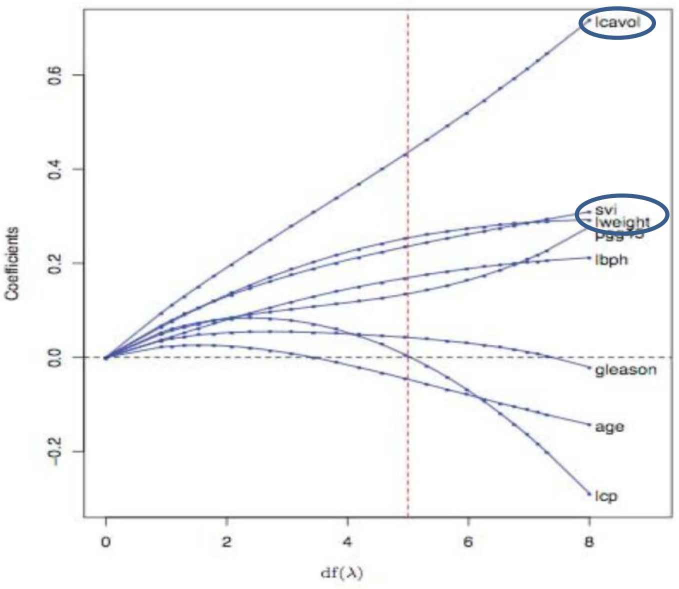

FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

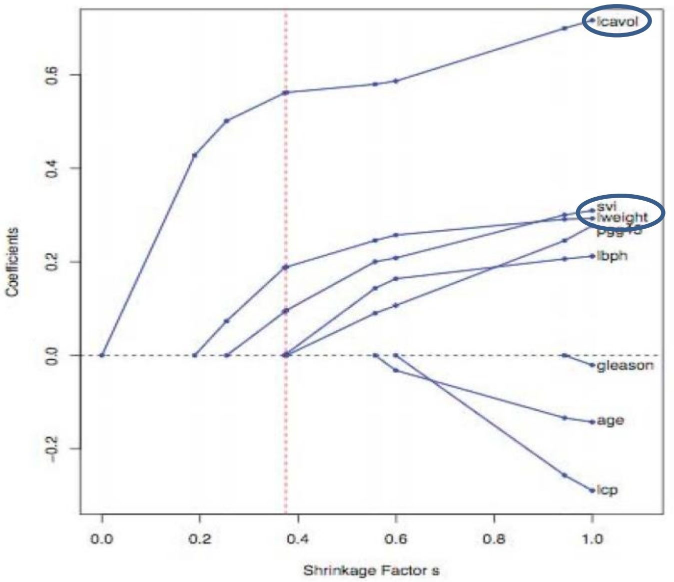

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$*. A vertical line is drawn at* $s = 0.36$*, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

<u>Acknowledgement</u>: From Hastie, T., Tibshirani, R., Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics.

LASSO: Limitations

- LASSO selects only up to n variables
  - n is the number of observations
  - If the number of potential predictors is greater than the number of observations, LASSO will select at most n of them
  - Since, normally, $n > p$, not a significant limitation
- If there are high correlations among predictors
  - LASSO is dominated by ridge regression
- If there is a group of variables with high correlation
  - LASSO tends to select only one variable from the group
    - LASSO doesn't care which one

Elastic Net

Elastic Net minimizes

$$\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j{}^2$$

- $\boldsymbol{L_1}$ penalty generates a sparse model
- $\boldsymbol{L_2}$ penalty
  - Removes the limitation on the number of selected variables
  - Encourages group effect
  - Stabilizes the $\boldsymbol{L_1}$ regularization path

Reference: Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B* 67.2 (2005): 301-320.

Georgia Tech

Summary