Subset selection, Ridge, Lasso

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

January 22, 2020

Notation

- y - $n \times 1$
- x - $n \times p$
- β - $p \times 1$
- linear model: $y = x\beta$

Least squares

- estimates $\hat{\beta}$ by minimizing

$$\overline{\mathrm{err}}(\beta) = \sum_{i=1}^{n} L(y_i, x_i\beta))$$

where y_i and x_i are training examples and $x_i\beta$ is in matrix notation: $x_i\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

$$\overline{\mathrm{err}} = \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

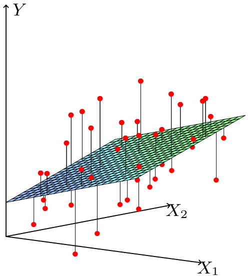


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

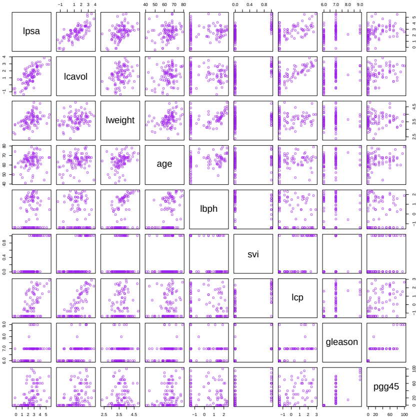


FIGURE 1.1. *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* `svi` *and* `gleason`*, are categorical.*

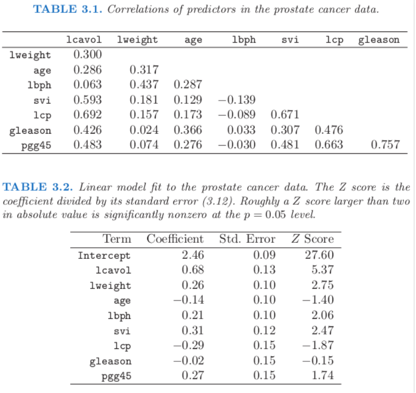


TABLE 3.1. *Correlations of predictors in the prostate cancer data.*

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

Test error

Once we have a predictor $\hat{Y} = f(X)$, test error is defined:

$$\mathrm{Err} = E_{X,Y}[L(Y, \hat{Y})]$$

Estimate Err using testing examples:

$$\overline{\mathrm{Err}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i^{\mathrm{test}}, \hat{y}_i^{\mathrm{test}})$$

Average loss when fitted model applied to testing examples.

Example: Prostate Cancer

- data is randomly split: training (2/3), testing (1/3)
- test error: 0.521:

$$\overline{\mathrm{Err}} = \frac{1}{n}\sum_{i=1}^{n}(y_i^{\mathrm{test}} - x_i^{\mathrm{test}}\hat{\beta})^2$$

- "base error" test error for intercept-only model : 1.057:

$$\overline{\mathrm{Err}}_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i^{\mathrm{test}} - \hat{\beta}_0)^2$$

Example: Prostate Cancer

- some predictors not important (e.g., `gleason`)
- using unnecessary predictors may cause overfitting
- reduce $\overline{\mathrm{Err}}$ by eliminating inputs or using penalty?

Sidebar on model selection

- modifying model after seeing data called model selection
- e.g., transforming inputs or outputs
- e.g., adding or eliminating inputs
- statistical inference is affected by model selection
- e.g., inflated type-I error
- model selection okay for prediction
- must use good estimate of Err

Best-subset selection

- suppose there are p predictors
- for each $k \in \{0, 1, \ldots, p\}$
    1. fit all possible combinations of k predictors among p total
    2. select combination that gives smallest training error $\overline{\mathrm{err}}$
- then choose k that minimizes test error $\overline{\mathrm{Err}}$

Best subset fitting

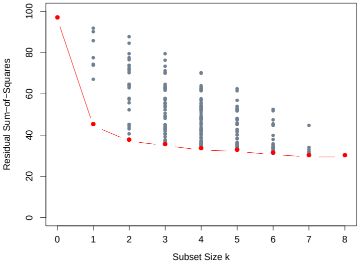


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Best subset tuning

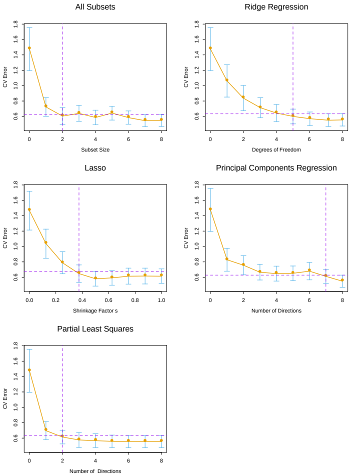


FIGURE 3.7. *Estimated prediction error curves and their standard errors for the various selection and shrinkage methods. Each curve is plotted as a function of the corresponding complexity parameter for that*

Shrinkage: Ridge

- synonyms: Penalization, Regularization, Shrinkage
- minimize penalized training error:

$$\overline{\mathrm{err}} = \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- $\hat{\beta}$ has a “closed form” solution
- shrinkage applies to $\hat{\beta}$, no subsetting of inputs X
- thus, number of β’s stays the same for all λ
- can use concept of effective degrees of freedom $df(\lambda)$
- one-to-one relationship between $df(\lambda)$ and λ
- $df(\lambda) = p$ when $\lambda = 0$
- $df(\lambda) \to 0$ as $\lambda \to \infty$

Shrinkage: Ridge

- $\hat{\beta}$ always unique, even when inputs perfectly correlated
- usually the intercept β_0 is not penalized
- can do this by centering outcome and inputs: $y_i = y_i - \bar{y}$ and $x_{ij} = x_{ij} - \bar{x}_j$; forces intercept to be zero; only remaining β estimated using ridge penalty
- λ parameterizes the "path" of estimates $\hat{\beta}$
- ridge and lasso are two of many such "path algorithms"
- graph of $\hat{\beta}$ as function of λ called a "path diagram"

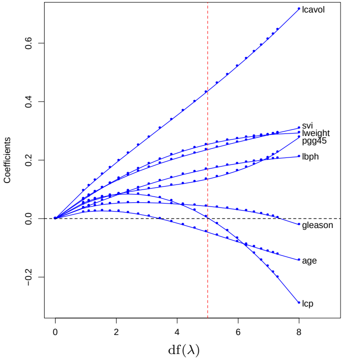


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

Shrinkage: Lasso

- minimize penalized training error:

$$\overline{\mathrm{err}} = \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- no closed form solution for $\hat{\beta}$
- making λ large causes some $\hat{\beta}$ to be exactly zero
- thus, lasso has a predictor selection effect

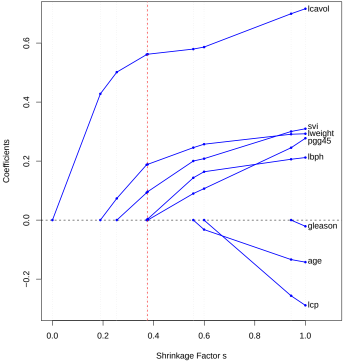


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for 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.*

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

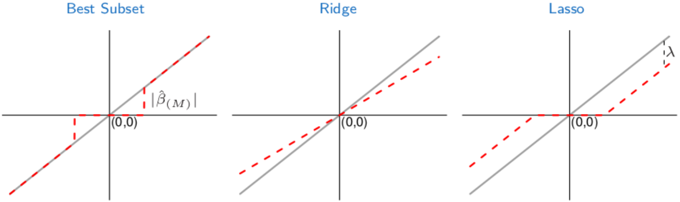
Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Consider independent inputs

- columns of x are uncorrelated, “orthogonal”
- $\hat{\beta}$ are independent; can be estimated separately
- can think about effect of selection/shrinkage on each coefficient separately

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat\beta_j \cdot I(\lvert\hat\beta_j\rvert \geq \lvert\hat\beta_{(M)}\rvert)$
Ridge	$\hat\beta_j/(1+\lambda)$
Lasso	$\text{sign}(\hat\beta_j)(\lvert\hat\beta_j\rvert - \lambda)_+$


Best Subset
Ridge
Lasso

Ridge and lasso penalties as constraints

- Ridge and lasso estimation criteria can be rewritten as constrained estimation problems:
- minimize $\overline{\mathrm{err}}$ subject to constraint:
- ridge: $\beta_1^2 + \cdots + \beta_p^2 \leqslant t^2$
- lasso: $|\beta_1| + \cdots + |\beta_p| \leqslant t$
- one-to-one relationship between t and λ

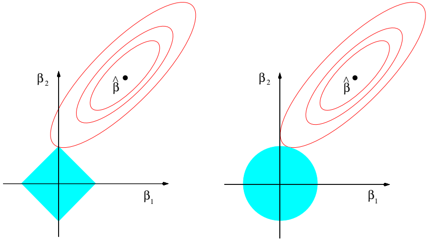


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

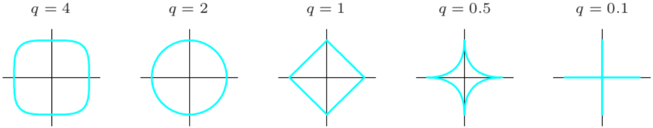


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

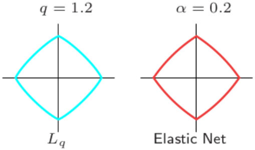


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j(\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

Code example

`lasso-examples.R`