P9120 - hw 1

1. To prove the Bridge(β) function is convex, we need to use the first and second order characterization of convex function:

First we seperate the Bridge function into two parts:

Let $F(\beta) = (y - X\beta)^T(y - X\beta)$ be the first part,

$$G(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^q = \sum_{j=1}^{p} |\beta_j|_q^q$$

Expand $F(\beta) = y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta$

$$\Rightarrow \quad \frac{\partial F(\beta)}{\partial \beta} = -X^Ty - X^Ty + 2X^TX\beta = -2X^Ty + 2X^TX\beta$$

$$\Rightarrow \quad \frac{\partial F(\beta)}{\partial \beta^2} = 2X^TX \geq 0 \Rightarrow \text{It's a positive semi-definite matrix.} \Rightarrow F(\beta) \text{ is convex.}$$

By definition: $F(\beta)$ is convex since second-order $\nabla^2 F(\beta) \geq 0$ for all $\beta \in dom(f) \Leftrightarrow F(\beta)$ is convex.

Then we turn to $G(\beta)$, for any vector β_1, β_2.

$\|t\beta_1 + (1-t)\beta_2\|_q \leq t\|\beta_1\|_q + (1-t)\|\beta_2\|_q \quad (\forall t \in [0,1])$ [Minkovski Inequality]

By definition $|\beta|_q$ is convex if $q \geq 1$

So $G(\beta) = \lambda|\beta|_q^q$ is then rewritten as.

$\lambda\|t\beta_1 + (1-t)\beta_2\|_q^q \leq \lambda t\|\beta_1\|_q^q + \lambda(1-t)\|\beta_2\|_q^q$, thus, we derive that

$G(\beta)$ is also convex if $\lambda > 0$ and $q \geq 1$, since ($m^q \geq n^q$ for $m, n > 0$ and $q \geq 1$)

Thus, $Bridge_\lambda(\beta) = H(\beta) + G(\beta)$ is a convex function when $\lambda > 0$ and $q \geq 1$

Besides, for $q \neq 1$, the equal sign cannot be attained as below shown:

$$\lambda\|t\beta_1 + (1-t)\beta_2\|_q^q < \lambda t\|\beta_1\|_q^q + \lambda(1-t)\|\beta_2\|_q^q$$

So, we conclude that $Bridge_\lambda(\beta) = H(\beta) + G(\beta)$ is strictly convex when $\lambda > 0$, $q > 1$

To sum up: Bridge(β) is convex when $\lambda > 0$, $p \geq 1$.

Bridge(β) is strictly convex when $\lambda > 0$, $p > 1$.

(b) From Question (a), we know that:

① The bridge function is a convex function in β, and the second-order derivative proofs to be positive, So the minimum over the domain (f) can be attained. thus, the bridge function has at least one solution, there is minimizer for the function.

② However, to prove its uniqueness, we need to prove: if the is another β_m, Bridge(β_m) = Bridge(β_n) in which m, n allows $\arg\min$ Bridge(β), By definition of strictly convex:

For $t \in (0,1)$,

$$\text{Bridge}(t\beta_m + (1-t)\beta_n) < t\,\text{Bridge}(\beta_m) + (1-t)\,\text{Bridge}(\beta_n)$$
$$= t\,\text{Bridge}(\beta_m) + \text{Bridge}(\beta_n) - t\,\text{Bridge}(\beta_n)$$
$$= \text{Bridge}(\beta_n)$$

Since β_n is set to attain the min value, the above function can not hold, so the minimizer should be unique.

For $\sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q \le t_0$, we are assuming $\tilde{\beta}(\lambda)$ is the minimizer for $F(x) = (y - X\beta)^T(y - X\beta)$

So, for Bridge minimizer $\hat{\beta}(\lambda) \in \text{dom}(f)$:

$$(y - X\tilde{\beta}(\lambda))^T(y - X\tilde{\beta}(\lambda)) + \|\tilde{\beta}(\lambda)\|_q^q \le (y - X\hat{\beta}(\lambda))^T(y - X\hat{\beta}(\lambda)) + \|\hat{\beta}(\lambda)\|_q^q$$

(Assume $\|\tilde{\beta}(\lambda)\|_q^q \le \|\hat{\beta}(\lambda)\|_q^q$)

Then the equation suggests that $\hat{\beta}(\lambda)$ could not be the minimizer of Bridge(β) as there is another $\tilde{\beta}(\lambda)$ appears to have lower value.

So, the part of $\sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q \le t_0$.

(c) The Bridge function is proved to be convex;

Suppose there're 2 minimizers $\hat{\beta}_1(\lambda)$ and $\hat{\beta}_2(\lambda)$.

for $\forall\, t \in (0,1)$ M denotes the min value obtained by $\hat{\beta}_1(\lambda)$ and $\hat{\beta}_2(\lambda)$ under the LASSO regression.

$$\|y - X[t\hat{\beta}_1(\lambda) + (1-t)\hat{\beta}_2(\lambda)]\|_2^2 + \lambda\|t\hat{\beta}_1(\lambda) + (1-t)\hat{\beta}_2(\lambda)\| < tM + (1-t)\cdot M = M.$$

where the equal sign cannot be achieved as the function is strict convex.

this is contradict to the Assumption M is the minimum value as a lower value was attained.

Thus, two minimizers should be equal and the panelty function should take the same value.

$$S(\lambda) \triangleq \sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q$$

While for proving $S(\lambda) \le t_0$, let $\hat{\beta}(\lambda)$ be the minimizer of Bridge function and $\tilde{\beta}(\lambda)$ be the minimizer of $F(\beta) = (y - X\beta)^T(y - X\beta)$

Suppose $\|\hat{\beta}(\lambda)\| > \|\tilde{\beta}(\lambda)\|$ holds,

$$\text{Bridge}_\lambda(\tilde{\beta}) = (y - X\tilde{\beta}(\lambda))^T(y - X\tilde{\beta}(\lambda)) + \|\tilde{\beta}(\lambda)\|$$
$$< (y - X\hat{\beta}(\lambda))^T(y - X\hat{\beta}(\lambda)) + \|\hat{\beta}(\lambda)\|$$

Cause $\hat{\beta}(\lambda)$ cannot be the unique minimizer, so there's some contradiction.

$$\Rightarrow S(\lambda) \triangleq \sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q \le t_0$$

(d) Since $G(\beta)$ is a convex function,

so, $G(\beta)$ has at least one subgradient at

every point in relint dom G, also G is differentiable $\nabla G(a)$ is subgradient of G at a.

we can define vector h as a subgradient at any β, and it satisfies that:

$$g(\beta') \geq g(\beta) + h^T(g(\beta') - g(\beta)) \quad \text{for } \forall\, \beta \in \text{dom}(\beta)$$

hence: $g(\beta') \leq g(\beta) \Rightarrow h^T(g(\beta') - g(\beta)) \leq 0$, $\partial g(\beta)$ denotes the subdifferential that includes all possible subgradients of g at β.

① To prove Condition in (d) $\Rightarrow$ Condidition in (C)

The lagrangian form of P_2:

$$L(\beta, r) = (y - X\beta)^T(y - X\beta) + r(\|\beta\|_q^q - s(\lambda))$$

$\Rightarrow$ the lagrangian condition: $0 \in -2X^Ty + 2X^TX\beta + r\partial G(\beta)$

② To prove Condition in (C) $\Rightarrow$ Condition in (d).

Due to the constraints specified, we checked:

for $(y - X\beta)^T(y - X\beta)$ subject to $\sum_{j=1}^{p} |\beta_j|^q \leq s(\lambda)$ ... (1)

a. $\|\beta\|_q^q - s(\lambda) \leq 0$ holds as β can be all zeros.

b. When $\lambda = r$, $\beta = \beta(r)$, the complementary condition is satisfied at solution point β.

To conclude, to minimize Bridge(β) in (c) is equvalent to that minimizing $(y - X\beta)^T(y - X\beta)$ with $\sum_{j=1}^{p} |\beta_j|^q \leq s(\lambda)$.

P9120 Homework #2, #3, #4
UNI: sj2921

2. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2, . . . , p predictors. Explain your answers:

   (a) Which of the three models with k predictors has the smallest training RSS?

   The smallest training RSS will be the model selected with the best subset approach, as this model have considered all models with predictor number ranging from 0 to P, so final model will be chosen with k parameters for best subset, the other two methods would not have such small RSS.

   (b) Which of the three models with k predictors has the smallest test RSS?

   The model with best subset approach will select a best model based on **training** error**, but it does not mean a necessary** smallest **test error**, so for any given data, the smallest test RSS can appear in either three models: best-subset model, forward stepwise and backward stepwise approach.

   (c) True or False:

   i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection.

      True, by definition, the forward stepwise adds variable once at a time and the K+1 variable should include all elements in K variable model, so the K variable model is a subset of (K+1) variable model.

   ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1) variable model identified by backward stepwise selection.

      True, also by definition, the backward algorithm reduces one variable at a time based on the original K+1 model, so the K-variable model should be a subset of the (K+1) variable model.

   iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1) variable model identified by forward stepwise selection.

      False, there is no direct connection between the forward selection algorithm and backward selection algorithm, so we have no rationale to infer from the K-variable model in backward method to K+1 model in forward stepwise method.

   iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

      False, there is no direct connection between the forward selection algorithm and backward selection algorithm, so we have no rationale to infer from the K-variable model in forward stepwise method to K+1 model in backward method.

   v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

No, since for each given N choose K, we can select more than 1 different models, and then pick up the best from the best subset method, so the K-variable model is not necessarily the subset of the (K+1)-variable model.

3. Derive the entries in Table 3.4, the explicit forms for estimators in the orthogonal case.

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	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

1. From the *Least Square method*, we can derive the formula:

$$\hat{\beta}^{ls} = (\mathbf{X}^{\mathrm{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathrm{y} = \mathbf{X}^{\mathrm{T}}\,\mathrm{y}$$

2. For *Best-subset*, we get the equation:

$$\hat{\beta}^{ls} = \hat{\beta}^{BS}(M = P) = \mathbf{X}^{\mathrm{T}}\,\mathrm{y}$$

where the coefficients are identical even if we take $M \leq p$ since the design matrix is orthogonal;

3. For *ridge regression* estimates:

$$\beta^{ridge} = (\mathbf{X}^{\mathrm{T}}\,\mathbf{X} + \lambda\,\boldsymbol{I})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathrm{y} = (\mathbf{I} + \lambda\,\mathbf{I})^{-1}\mathbf{X}^{T}\,y = \hat{\beta}^{ls}\,/\,(1+\lambda)$$

4. For *lasso regression*, we get the following:

$$L(\beta) = (\,y - \mathbf{X}\beta)^{\mathrm{T}}\,(y - \mathbf{X}\beta) + \lambda\,|\beta|$$

The **first order derivative** is:

$$\frac{\partial L(\beta)}{\partial(\beta)} = -\mathbf{X}^{\mathbf{T}}\mathrm{y} + \mathbf{X}^{\mathrm{T}}\,\mathbf{X}\,\beta + \lambda\,sign(\beta)$$

Setting the gradient as 0, we get that the solution with respect to β is:

$$\beta^{Lasso} = \mathbf{I}(\,\mathbf{X}^{\mathbf{T}}y - \lambda\,sign(\beta)) = sign(\beta)\,(\mathbf{X}^{\mathbf{T}}y - \lambda)$$ Q.E.D.

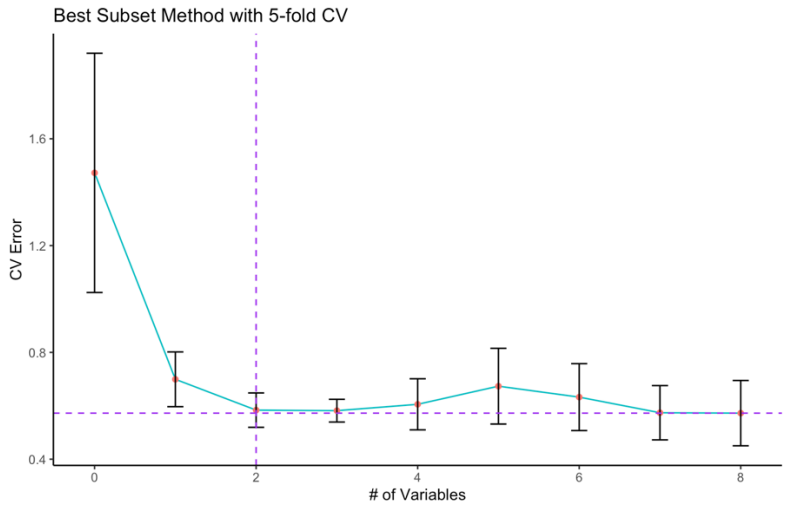
4. Table summary:

Table 1. Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data.

Term	Best Subset	LASSO with 5-fold CV	LASSO with BIC	PCR with 5-fold CV	LS
(Intercept)	2.4773	2.4686	2.4551	2.455	2.465
lcavol	0.7397	0.5259	0.0871	0.286	0.68
lweight	0.3163	0.1622		0.3391	0.263
age				0.0562	-0.141
lbph				0.1015	0.21
svi		0.0602		0.2614	0.305
lcp				0.2187	-0.288
gleason				-0.016	-0.021
pgg45				0.0617	0.267
Test Error	0.4713	0.5151	0.9472	0.5513	0.521
Std Error	0.0434	0.1762	0.3591	0.1328	0.179

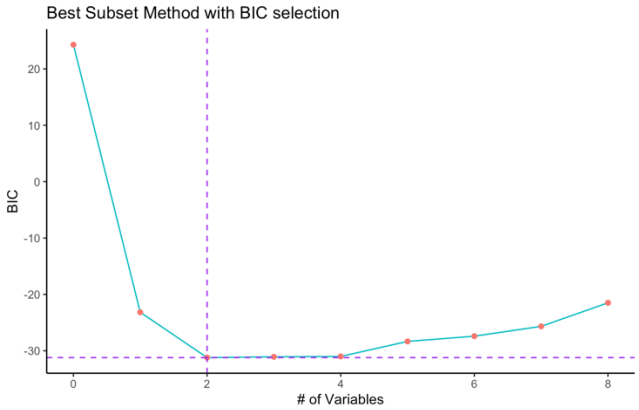
Figure output:

(a). The best-subset method produced with 5-fold cross-validation shown below just chose two variables in the model, and has a relatively small test error with small std.dev.



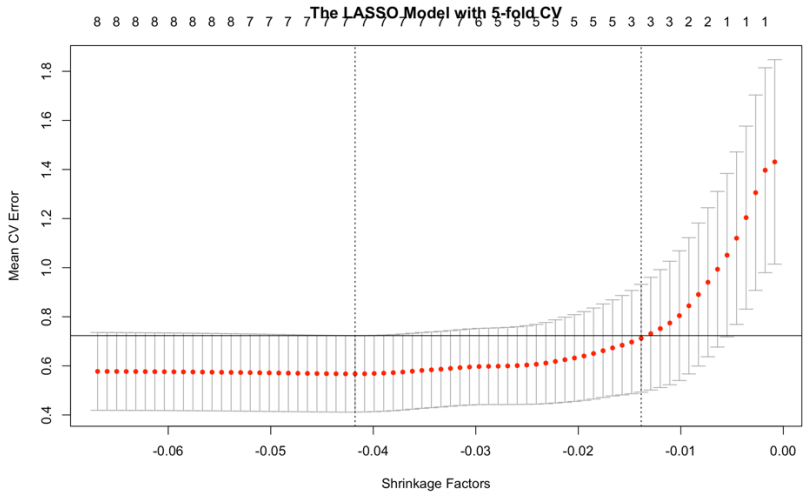
(b) **Best-subset with BIC**

This method selects variables “lcavol” and “lweight” into the model, obtained the lowest BIC value, compared with Best-subset with 5-fold CV, these two have the same model selected and the test error and the standard error of test error are the lowest.



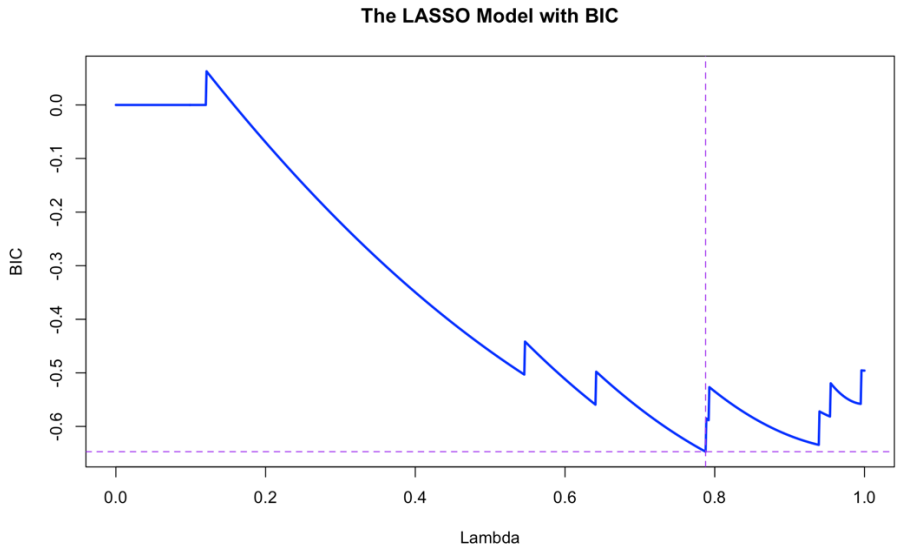
(c). **LASSO with 5-fold CV**

In the LASSO with 5-fold CV, we select from multiple lambdas, and found the best one is λ = 0.250016 provides the best CV error under the one-standard deviation rule. This method selects three variables “lcavol”, “lweight” and “svi”.



d. LASSO with BIC

We find $\lambda = 0.7877$ provides the lowest BIC value. This method selects only variable "lcavol" into the model. LASSO with BIC criteria for model fitting has the largest test error among all the models.



e. **PCR** with 5-fold cross-validation

Principle component regression presents three factors provide the best CV error under the one-standard deviation rule. The fitted coefficients in model PCR shown below has selected model with all 8 variables.