**P9120 Homework #2, #3, #4**

**UNI: sj2921**

1. 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:

* 1. 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.

* 1. 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.

* 1. 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.

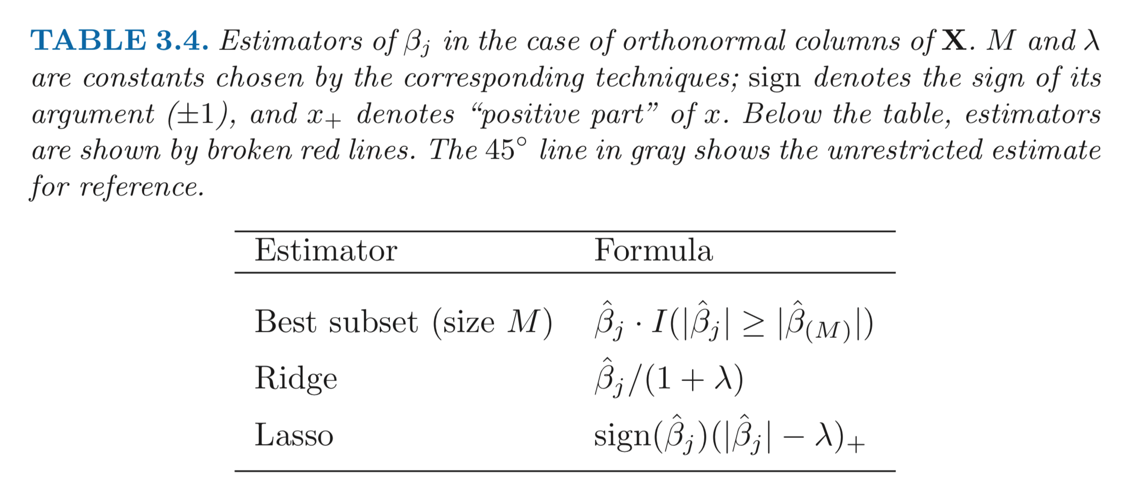
* 1. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

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.

* 1. 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.

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



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

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

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

1. For ridge regression estimates:

/ (1 +

1. For lasso regression, we get the following:

The **first order derivative** is:

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

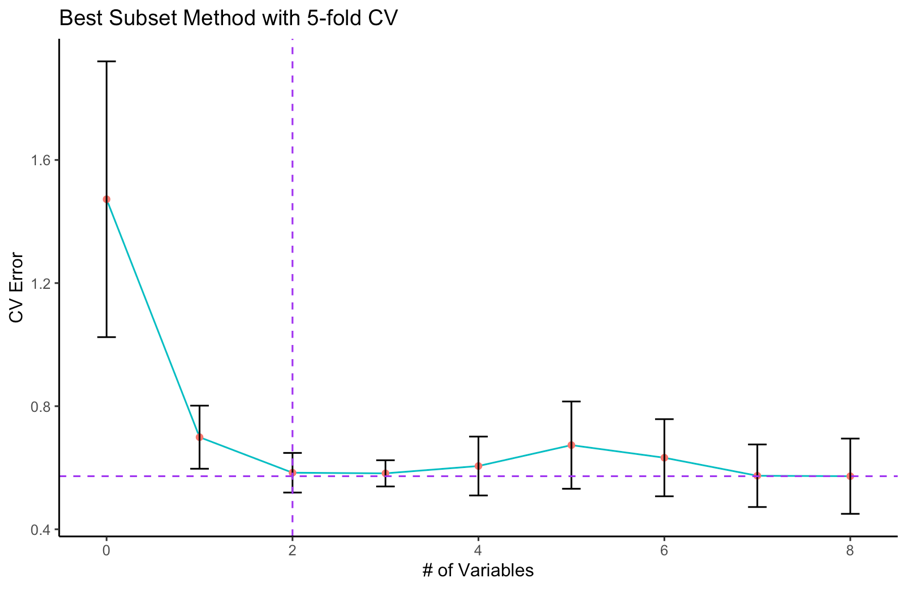
Q.E.D.

1. **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 |

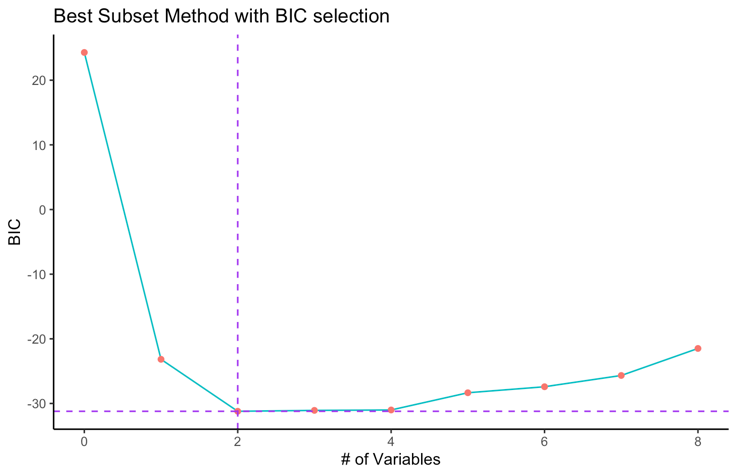
**Figure output:**

(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.



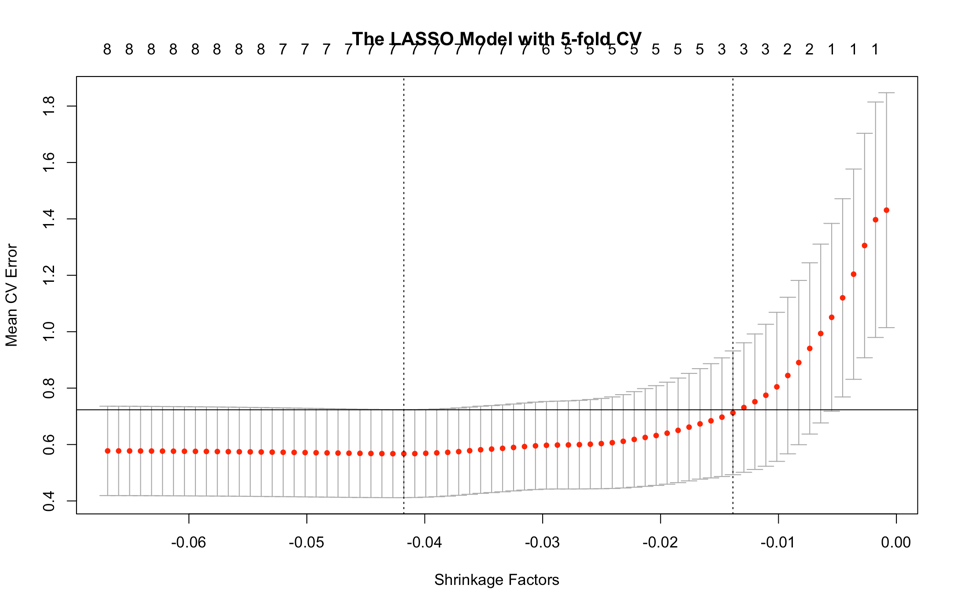
(b) **Best-subset with BIC**

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.



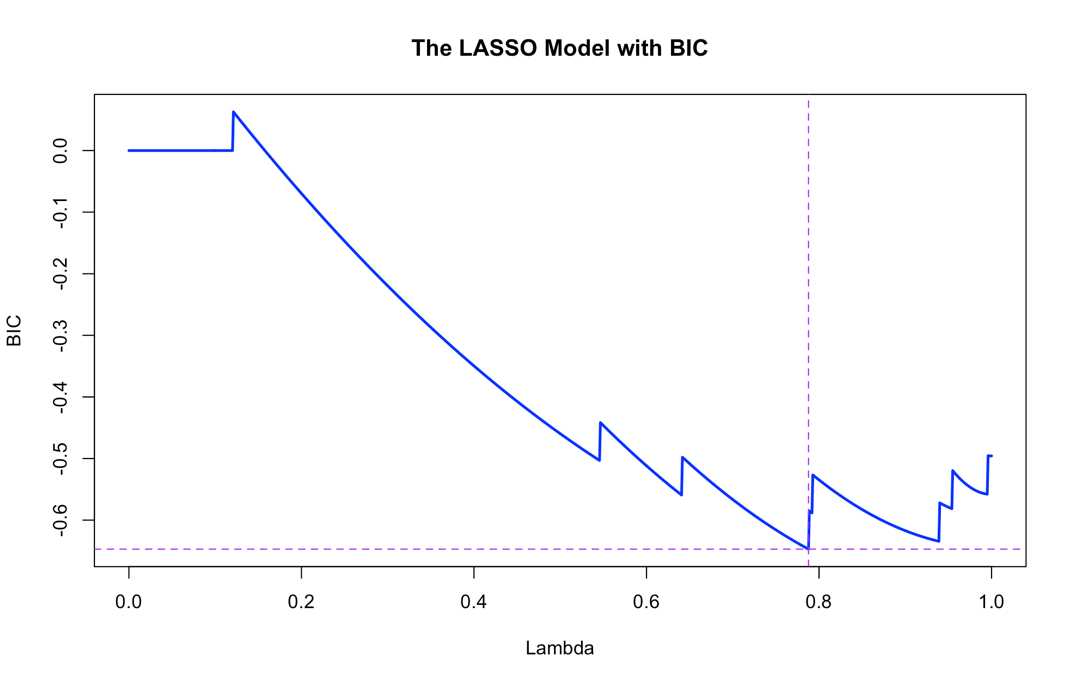
(c). **LASSO with 5-fold CV**

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 λ = 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.

