homework1

Apoorva Srinivasan

10/17/2019

Problem 3

The estimator of $\hat{\beta}$ for leat squares are:

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

In the case of orthonormal designs, the esimator would simplified as

$$\hat{\beta}^{ls} = X^T y$$

Best subset

We know that the estimator of $\hat{\beta}$ for best subset is:

$$\hat{\beta}^{ls} = \hat{\beta}^{bss}(M = p) = X^T y$$

In the case of orthorgonal designs, the coefficients are identical even if $M \leq p$ So, we get

$$\hat{\beta}^{bss} = \hat{\beta}^{ls} I(|\hat{\beta}^{ls}| \geq |\hat{\beta}^{M}|)$$

Ridge Regression

We know that the estimator of $\hat{\beta}$ for best subset is:

$$\hat{\beta}^{Ridge} = (X^T X + I\lambda)^{-1} X^T y = (I + I\lambda)^{-1} X^T y = \frac{\hat{\beta}^{ls}}{1 + \lambda}$$

Lasso Regression

For lasso regression, we get the following be definition:

$$L(\beta) = (y - X\beta)^T (y - X\beta) + \lambda|\beta|$$

Take the derivative at both side we get and set $\hat{\beta}$ as 0 we get:

$$\hat{\beta}^{Lasso} = X^T y - \lambda sign(\beta) = sign(\beta)(|X^T y| - \lambda)_+ = sign(\beta)(|\hat{\beta}^{ls}| - \lambda)_+$$

Problem 2

(a)

Best subset has the smallest training RSS. Other two models may also yield the smallest training RSS, but best subset would yield in certain. Because all possible combinations of parameters are considered in given k in the best subset model and it finds the subset that minimizes the training error.

(b)

Best subset has the smallest testing RSS for sure. But other two models might also find the smallest model by luck.

(c)

- 1: True: Forward selection is a type of stepwise regression which begins with an empty model and adds in significant variables one by one. So the k+1 - variable model already contains the k - variable model.
- 2: True: Backward selection is a type of stepwise regression which begins with an full model and deletes in non-significant variables one by one. So the k+1 - variable model already contains the k - variable model.
- 3: False: This is not always true. This is because forward selection starts with an empty model and keeps adding a single predictor each time to check for lowest RSS, however, the backward model depends on the combined effects of predictors in resulting the lowest RSS and hence the k-variable model identified by backward stepise won't always be a subset of the predictors in this scenario.
- 4: False: The reason's similar to the above since they start differently looking for different ways of lowest training RSS.
- 5: True: The k-variable model identified by best subset is defined as the model that has the smallest RSS. Same as the k+1 variable. So, the k+1 variable model does not have to include the varibales that selected in k variable model.