STAT S4240 002, Homework 4

Brian Weinstein (bmw2148)

August 13, 2015

Problem 1: James 6.8.1

(a) The model obtained through *best subset selection* has the smallest training RSS for $k \leq p$ predictors. This is because best subset selection tests all possible subsets of size k and chooses the one that minimizes the training RSS.

There is a chance that by using *forward stepwise selection* or *backward stepwise selection* we'd end up with the same optimal subset of k predictors (i.e., the one we obtain via best subset selection), but since the k-sized subset is dependent on the covariates included in the prior steps in each of these methods, there's no guarantee that this will be the case.

(b) Without either valuating each model on a testing set or estimating test error, none among the three models is guaranteed to have the lowest test RSS. These methods only choose the subset that results in the lowest training RSS, which in no way indicates if the model will also result in a low test RSS.

(c)   i. True. In a model identified by forward stepwise selection, the subset of predictors in a given step (with k predictors) is the base upon which the next predictor is added in the subsequent step (with $k + 1$ predictors).

  ii. True. In a model identified by backward stepwise selection, the subset of predictors in a given step (with k predictors) is itself a subset of the predictors used in the previous step (with $k + 1$ predictors).

 iii. False. The forward stepwise and backward stepwise models using k predictors are not guaranteed to use the same subset of predictors. As such, once an additional variable is added to the forward stepwise model (or equivalently, once a variable is removed from the backward stepwise model), there's no guarantee that the variables in the k-variable backward stepwise model are a subset of the $k + 1$-variable forward stepwise model.

 iv. False, with similar reasoning to Part (iii). The k predictors in a forward stepwise model are the predictors that, at each incremental step, kept the error as low as possible. Similarly, the $k + 1$ predictors in a backward stepwise model are the predictors at each incremental step that kept the error as low as possible. Since each subset of predictors depends on those used in the previous step, there's no guarantee (although it may happen to be the case) that a set of predictors obtained via forward selection is a subset of those obtained via backward selection.

  v. False. The set of k predictors chosen in best subset selection are those that keep the error in the k-variable model as low as possible. Independently, the set of $k + 1$ predictors chosen in best subset are those that keep the error in the $k + 1$-variable model as low as possible. There's no guarantee that k of the variables included in the $(k + 1)$-variable model are those that minimize the error in the k-variable model.

Problem 2: James 6.8.3

Estimating the regression coefficients in a linear regression model via

$$\underset{\hat{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

for a given value of s.

As we increase s from 0:

(a) The training RSS will (**iv.**) steadily decrease. As we make the constraint on $\sum_{j=1}^{p} |\beta_j|$ less strict, the limitation on the size of $\hat{\beta}$ disappears and we eventually get the standard "least residuals" solution. At $s = 0$ we start with only the intercept β_0 non-zero. As we increase s we allow for additional non-zero coefficients, eventually including all coefficients in the model. With all coefficients included, we minimize the training error (while potentially overfitting).

(b) The testing RSS will (**ii.**) decrease initially, and then eventually start increasing in a U shape. At $s = 0$ we start with a highly-biased model with only the intercept β_0 non-zero. As we increase s we allow for *some* of the $\beta_{j \neq 0}$ to take non-zero values (since the constraint is the same as that used in the lasso), which reduces the bias and increases the variance, for an overall drop in RSS. Once we increase s enough, *all* of the $\beta_{j \neq 0}$ will take non-zero values and we're left with a model with very low bias, but very high variance, increasing the RSS.

(c) The variance will (**iii.**) steadily increase. The reasoning here is similar to that used in Part (ii): At $s = 0$ we start with a very inflexible model, and as we increase s to allow more non-zero coefficients, the model flexibility, and thus our variance, increases.

(d) The (squared) bias will (**iv.**) steadily decrease. The reasoning here is similar to that used in Part (ii): At $s = 0$ we start with a highly-biased, very inflexible model with only one coefficient, β_0. As we increase s to allow more non-zero coefficients, the model flexibility, and thus our squared bias, decreases.

(e) The irreducible error will (**v.**) remain constant. The irreducible error is a function of the data we're given, not the model we choose, so our choice of s has no effect on the size of the irreducible error.

Problem 3: James 6.8.5

(a) _____ Problem 3

Problem 4: James 8.4.5

(a) _____ Problem 4

Todo list