

## Homework 3 - ST 790b - Model Selection

Due March 25, 2009

1. I want you to get a little more familiar with the `lars` R package because it is very fast and useful. On my Variable Selection website under software, click on `Wrapper for LASSO` and then on `LASSO wrapper` to see one illustration of how to extract estimates using `mode="fraction"`. The adaptive lasso code given in class (see 2. below as well) shows how to use `mode="step"`. I want you to make a “wrapper” similar to my LASSO wrapper based on the `lars` R function that will

- a) Output the coefficients of the minimum  $C_p$  model. This is natural output from calling `lars`. So you just do something like `out=lars(x,y)` and `which.min(out$Cp)` to get the step of the minimum  $C_p$  and then call `predict.lars` to get the coefficients.
- b) Output the coefficients of the minimum BIC model, where

$$\text{BIC} = \frac{\text{RSS}_s}{\hat{\sigma}_{\text{Full}}^2} + \log(n)(\text{df}_s),$$

and here I purposely use  $\text{RSS}_s$  and  $\text{df}_s$  because those are natural outputs from the `lars` function for step  $s$ . Note I have called this version of BIC either “the  $C_p$  version” or the “known variance version.”

- c) Output the coefficients of the minimum AIC model, where

$$\text{BIC} = \frac{\text{RSS}_s}{\hat{\sigma}_{\text{Full}}^2} + 2(\text{df}_s).$$

You should be able to get  $\hat{\sigma}_{\text{Full}}^2$  from the last step using  $\text{RSS}/\text{df}$ , but you should check to make sure that it is correct. Finally, once the function is ready, illustrate it using the baseball data with the response on the square root scale and adding the interactions `x8*x13`, `x8*x15`, `x8*x14`, and `x10*x14` that you found important in HW 2. (For simplicity, you don’t need to center before creating the interactions.)

2. Go to the `Adaptive LASSO` section of my Variable Selection website and download and source the `lasso.adapt.bic2` as well as the LSA version. On the website I have illustrated that for linear regression, both programs give the same result. Now I want you to do the same thing for the baseball data set (constructed as in Problem 1 above), i.e., compare the coefficients from both

programs. In fact, use the variables as in Problem 1, and compare coefficient estimates from lasso, adaptive lasso, and the least squares fit with variables chosen by Hugh's program in HW 2 (You have to run the least squares again because Hugh's program gives estimates for the transformed data.)

3. On the website where this Homework 3 is posted, I have placed a small data set with 4 predictors and a response. I want you to read the data into R and fit the adaptive lasso with a 5th variable added,  $x_1 * x_2$ . Then fit the adaptive lasso a second time, but this time let the 5th variable be  $(x_1 - \text{mean}(x_1)) * (x_2 - \text{mean}(x_2))$ . Compare the fits and decide whether they are consistent as one might expect.