Homework 4

Stat 151A, Spring 2014

Due Thursday, April 17th

1. Question 22.1. Notice that part (f) asks you to do (a-e) several times; I take this to be at least 5 times. So write your code in such a way that you can easily repeat the process (e.g. write a function). In responding to questions (a-e) you should respond based on the results of all your repetitions of the process, not just a single one.

   A few clarifications:

   - For question c) you need to get 3 variables, so you should not run the default step since it will be difficult to get 3 variables. Instead you can do a 'forward' regression starting at a model with only 1 variable, where 'forward' limits step to only add. The syntax for this is a bit difficult because it's hard to get the 'scope' formula created from a large number of variables (you can't use $\sim .$ or $\sim X$). :

     ```
     allterms<-as.formula(paste("y~",paste(colnames(X),collapse="+")))
     colnames(X)<-paste("X",1:ncol(X),sep="")
     #notice in my null model I have to say "data=",
     #even though I'm not using it. Otherwise it can't update
     nullLm<-lm(y~1,data=data.frame(X))
     step(nullLm,scope=allterms,direction="forward",steps=3)
     ```

   - For question d) I interpret this to mean you do like in part c) only without the restriction of 3 variables (and thus also without the restriction of just forward stepwise).
   - e) asks for validation on a new data set. That would usually mean find the prediction error, but for this question we are focusing on variable selection and the problems with it, so for part e) validation mean refit the models from parts b)-d), i.e. use the same variables but get new coefficients and significance values.

2. In class we talked about the ridge regression estimator

$$\hat{\beta}_{RR}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y.$$

   For this problem work with a single X (mean-centered) and no intercept so that quantities can be expressed as scalars.

   (a) Show that the magnitude of $\hat{\beta}_{RR}$ is smaller than that of the OLS estimator (i.e., that ridge regression shrinks the LS estimator toward zero).

(b) Calculate the bias and variance of the ridge regression estimator and compare to those of the OLS estimator.

(c) In Stat135, you should have seen that the mean squared error of an estimator is bias squared plus variance.

Calculate the MSE of the two estimators and consider the conditions under which we would have

$$\text{MSE}_{\text{ridge}} < \text{MSE}_{\text{OLS}}$$

Specifically,

i. Isolate β on one side of the inequality and assess in relative terms whether smaller *true* values of β will favor ridge regression or OLS.

ii. Do the same for σ^2 to determine whether smaller values of σ^2 favor ridge regression or OLS

3. Consider choosing between two nested models, the larger model with RSS_1 and p_1 parameters and the smaller one with RSS_0 and p_0 parameters. Do not assume that either model is the full model.

Both the adjusted R^2,

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-(p+1)}\frac{\text{RSS}}{\text{TSS}}$$

and the AIC criterion,

$$AIC := n \log\left(\frac{RSS}{n}\right) + nlog(2\pi e) + 2(1+p)$$

penalize larger models (models with more parameters).

(a) Derive a formula for when the larger model will be favored under each criterion, in the form

$$\frac{\text{RSS}_1}{\text{RSS}_0} < f(n, p_1, p_0)$$

for some function f for each of the two criteria (one function for each criterion).

(b) Use these functions to define intervals of values for

$$\frac{\text{RSS}_1}{\text{RSS}_0}$$

under which (a) both criteria pick the larger model, (b) both pick the smaller model and (c) one picks the larger and the other the smaller. Which criterion penalizes complexity more strongly?

4. The US Census collects extensive individual data on individuals (on the so-called 'long form', now being replaced with the American Community Survey). These data are called the census microdata. I have extracted a number of variables

from the California subset of the microdata for the 2000 U.S. Census; some are continuous while others are categorical. Your task in this lab assignment is to predict income (the 'inctot' variable) for individuals from ages 21 to 65 using these other variables (I have chosen variables that are generally more easily collected than income).

The dataset comes from the Census as a relational database with two pieces, one being information on households and the other on individuals, that I have combined for you using the merge command based on the common serial number variable so that the household information for each individual is now part of the record for the individual. I have also created a second dataset to be used for testing your results; do NOT use this in fitting the model - you may only use it as the final step to calculate predictive performance. A copy of the dataset is on bspace as is a code book (data dictionary) that describes the variables. Note that I have extracted the variables from fixed-width text strings and excluded many variables, so not all the variables in the code book are in the cleaned dataset. Also note that given the size of the dataset and resulting computational demands, we are using random subsets of 10,000 individuals for each of the training and test sets, out of a total of about 1,000,000 individuals.

(a) Do some basic evaluation of the full model to determine if any transformations, outliers, or other corrections need to be done.

(b) Use model building techniques we have discussed to choose variables and build a final model to predict income:

   i. Stepwise
   ii. All subsets
   iii. Lasso
   iv. Ridge regression

   For each of the above, describe the process of how you chose the best model and compare the models you get.

(c) For each model found above, perform cross validation and estimate the error of each model (this should be separate from any cross-validation you might choose to do above).

(d) Report the performance of each model on the held-out test set, reporting the mean squared prediction error for income on the set and compare these with what your cross-validation predicted.

You might need to work on a random subset of observations for some of these techniques if they take too long to run otherwise. Be aware that the fitted models end up taking up a large amount of memory, so you may want to delete these as you go. You can check the size of objects in memory with: object.size(object)/1000000, which will give you the size in Mb in R.

Note that one difficulty for some model selection routines is that you can end up with linearly dependent columns of the X matrix because some of the factor variables place constraints on other factor variables. For example, using the table function, one can see that when LASTWRK is 2, CLWKR can only be 0

or 9 while when LASTWRK is 1, CLWKR can only take values 1-8. In these cases R drops columns and gives NAs for redundant factor levels. A similar issue arises with YR2US, which is 0 for all citizens; you only want to fit a linear term for YR2US for those who have a year of arrival. Be alert for other such issues.