Stat 897 Fall 2017 Project 1

Penn State

Due October 1, 2017

Linear Regression, Variable Selection, Ridge Regression, Lasso

This project is to be completed individually. You may submit pdf only (Rmd is not needed and you can use another word processing tool if you like).

The diabetes data in Efron et al. (2003) will be used: ten baseline variables: age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. The data is available in R package `lars`. Load the data:

```
library(lars)
```

```
## Loaded lars 1.2
```

```
data(diabetes)
data.all = data.frame(cbind(diabetes$x, y = diabetes$y)) # change to normal formatting
```

Partition the patients into two groups: training (~75%) and test (~25%). Please use use the random number generator seed specified below before randomly splitting the data. Since you will not submit a markdown file, please use the set.seed at each noted place.

```
set.seed(38723) # set random number generator seed to enable reproducibility of results
```

Project Requirements

Write up your results in a professional report, like you would present to a client or internal customer for your analysis. The report should be no more than 4 single-spaced pages long and submitted in PDF format.

It should include coefficient estimates for each model and test data mean prediction errors.

Include any other details from your analysis that you feel are worthy of mention.

The report should have sections (e.g., Introduction, Analysis, Results, Conclusion) and provide sufficient details that anyone with a reasonable statistics background could understand exactly what you've done and what you concluded.

Consider using tables and figures to enhance your report. You might use the package "pander" if you are using Rmarkdown for nicely formatted tables.

Do not embed R code in the body of your report (if you are using rmarkdown, use {r echo=FALSE} to supress the printing of the r code), but instead attach the code (code only, not output) in an appendix. The appendix does not count towards the page limit.

Grading criteria (out of 25)

15 points: fulfilling the project requirements and matching results exactly (this is why you should use the specific random number generator seeds).

10 points: the quality of your report (including: clarity of writing, organization, and layout; appropriate use of tables and figures; careful proof-reading; adherence to report guidelines

Fit the following models to the training set. For each model extract the model coefficient estimates and calculate the "mean prediction error" in the test set.

1. Least squares regression model using all ten predictors.
2. Apply best subset selection using BIC to select the number of predictors.
3. Apply best subset selection using 10-fold cross-validation to select the number of predictors. Please use a random number seed of 38723 immediately before entering the command.
4. Ridge regression model using 10-fold cross-validation to select the largest value of λ such that the cross-validation error is within 1 standard error of the minimum (R functions glmnet and cv.glmnet in package glmnet). Please use a random number seed of 38723 immediately before entering the command.
5. Lasso model using 10-fold cross-validation to select the largest value of λ such that the cross-validation error is within 1 standard error of the minimum (R functions glmnet and cv.glmnet in package glmnet). Please use a random number seed of 38723 immediately before entering the command.