STAT 602: Modern Multivariate Statistical Learning

Homework Assignment 2

Spring 2015, Dr. Stephen Vardeman

Name:	Ian Mouzon
email:	imouzon [@iastate]

Assignment:	Problem 15, 16 Vardeman's HW Set (link)
Due Date:	February 10, 2015

The following packages are used in these solutions. That said, there are many ways to arrive at satisfiable results.

Problem 16

This question concerns analysis of a set of sale price data obtained from the Ames City Assessor's Office. There is an `Ames Home Price Data` file on the course web page containing information on the sales May 2002 through June 2003 of $1\frac{1}{2}$ and 2 story homes built 1945 and before, with (above grade) size of 2500 sq ft or less and lot size 20,000 sq ft or less, located in Low- and Medium-Density Residential zoning areas. $n = 88$ different homes fitting this description were sold in Ames during this period. (2 were actually sold twice, but only second sales prices of these were included in our data set.) (The rows of the file have been shuffled randomly, so that you may use 8 successive sets of 11 rows as folds for cross-validation purposes if you end up programming your own cross-validations.)

For each home, the value of the response variable, *Price*, and the values of 14 potential explanatory variables were obtained.

Size	the floor area of the home above grade in sq ft,
Land	the area of the lot the home occupies in sq ft,
Bedrooms	a count of the number of bedrooms in the home
Central Air	a dummy variable that is 1 if the home has central air conditioning and is 0 if it does not,
Fireplace	a count of the number of fireplaces in the home,
Full Bath	a count of the number of full bathrooms above grade,
Half Bath	a count of the number of half bathrooms above grade,
Basement	the floor area of the home's basement (including both finished and unfinished parts) in sq ft,
Finished Bsmnt	the area of any finished part of the home's basement in sq ft,
Bsmnt Bath	a dummy variable that is 1 if there is a bathroom of any sort (full or half) in the home's basement and is 0 otherwise,
Garage	a dummy variable that is 1 if the home has a garage of any sort and is 0 otherwise,
Multiple Car	a dummy variable that is 1 if the home has a garage that holds more than one vehicle and is 0 otherwise,
Style (2 Story)	a dummy variable that is 1 if the home is a 2 story (or a 2 12 story) home and is 0 otherwise, and
Zone(Town Center)	a dummy variable that is 1 if the home is in an area zoned as "Urban Core Medium Density" and 0 otherwise.

a) In preparation for analysis, standardize all explanatory variables that are not dummy variables (those we'll leave in raw form), and center the price variable making a data frame with 15 columns. Say clearly how one goes from a particular new set of home characteristics to a corresponding set of predictors. Then say clearly how a prediction for the centered price to a prediction for the actual dollar price.

b) Find linear predictors for centered price of all the following forms:

- OLS
- Lasso (choose λ by 8-fold cross-validation)
- Ridge (choose λ by 8-fold cross-validation)
- Elastic Net with $\alpha = 0.5$ (choose λ by 8-fold cross-validation)
- PCR (choose the number of components by 8-fold cross-validation)
- PLS (choose the number of components by 8-fold cross-validation)

For each predictor that you have a way to do so, evaluate the effective degrees of freedom.

c) Plot on the same set of axes as a function of index (1 through 15) the values of co-ordinates $\hat{\beta}_j$ of $\hat{\beta}$ for each predictor in b). (Connect successive coordiantes of a give $\hat{\beta}$ with the lines segments so that you can track the different methods across the plot. Use different symbols and colors for the 6 different methods.) If you see anything interesting in the plot, comment on it.

d) Plot on the same set of axes as a function of index (1 through 88) the values of co-ordinates $\hat{y}_i$ of $\hat{\mathbf{Y}}$ for each predictor in b). (Connect successive coordiantes of a give $\hat{\beta}$ with the lines segments so that you can track the different methods across the plot. Use different symbols and colors for the 6 different methods.) If you see anything interesting in the plot, comment on it.

Solution

a) The data can read into R directly from the excel spreadsheet found on the course page using the `gdata` package:

```
h <- read.csv("~/Downloads/ames_data.csv")
```

Centering and standardizing the columns of $\mathbf{X}$ can be accomplished by subtracting the column mean from each observation in the column and dividing each column by its variance:

```
# center the predictors
h.c <- h
cols <- c(2, 5:9, 13)
names(h)[cols]

## [1] "Size"        "Bed.Rooms"   "Central.Air"
## [4] "Fireplace"   "Full.Bath"   "Half.Bath"
## [7] "Land"

for (i in 1:length(cols)) h.c[, cols[i]] <- (h.c[,
    cols[i]] - mean(h.c[, cols[i]]))/sd(h.c[, cols[i]])
X <- h.c[, 2:ncol(h.c)]

# center the response as well
h.c[, 1] <- h.c[, 1] - mean(h.c[, 1])
y <- h.c[, 1]
```

Effective degrees of freedom:

OLS

```
ols.mod <- lm(Price ~ ., data = h.c)
beta.ols <- ols.mod$coefficients
fitted.ols <- ols.mod$fitted.values
```

Lasso

```
require(glmnet)
K.folds <- 8
cv.glmnet(X, y, alpha = 1, nfolds = K.folds)$lambda.lse

## Error in elnet(x, is.sparse, ix, jx, y, weights, offset, type.gaussian, :  (list) object
cannot be coerced to type 'double'

lasso.mod <- lm(Price ~ ., data = h.c)
beta.lasso <- lasso.mod$coefficients
fitted.lasso <- lasso.mod$fitted.values
```