Session 2

Linear Regression with Lasso in R

Overview

Using a prepared R script, we will use the cross-validation (CV) option in the GLMNET package on simulated sample #1 in our training data (data file: "LMTr&Val.sav") as the analysis sample with N=50 to determine the magnitude of the penalty to use in the lasso model. The results from the CV show that B90 is the recommended model. This model has nonzero coefficients for 10 of the 14 valid predictors, 7 of the 28 irrelevant predictors, and 2 of the 14 extraneous predictors.

If you don't already have the R program, install the program (free) using the directions provided here:

http://cran.r-project.org/bin/windows/base/

Launch the R program

Install the required package "glmnet":

- In R, click Packages → Install Package(s)…
- Click on a mirror closest to your current location and click "OK"
- Scroll down and click on "glmnet"

You can download the GLMNET reference manual here:

http://cran.r-project.org/web/packages/glmnet/index.html

Change Directory:

- Click File→ Change dir…
- Navigate to the folder where you have saved the data file "LMTr&Val.sav" and the R script "PenalizedRegMethods.r"
- Click "OK"

Open Script:

- Click File→ Open script…
- Click on "PenalizedRegMethods.r"
- Click "OK"

You should now have 2 windows open:

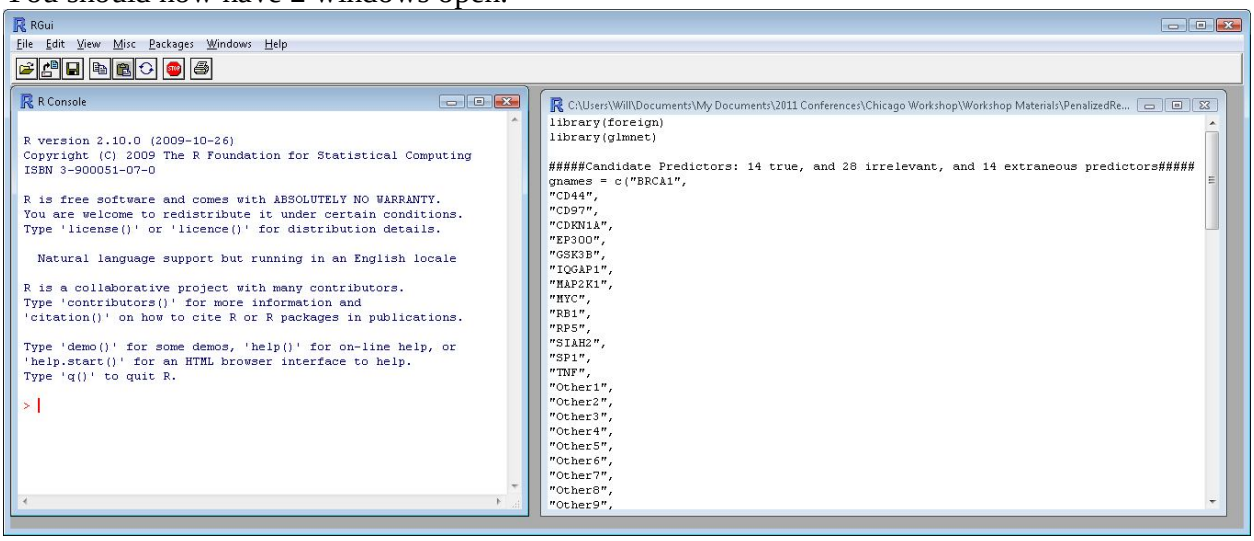

Fig. 1: RGui with R Console and R Script

R Script Details

Predictors

We have specified 14 true, and 28 irrelevant, and 14 extraneous candidate predictors ("gnames").

Specifying the Training Set

Using the following command, we have specified Validation=0 and simulation=1:

tbl0 = tbl[tbl$simulation == 1 & tbl$Validation == 0,]

Specifying the Dependent Variable

Using the following command, we have specified "Y" as the dependent variable:

dep0 = tbl0$Y

Specifying the Validation Set

Using the following command, we have specified Validation=1 and simulation~=1:

tbl1 = tbl[tbl$simulation != 1 & tbl$Validation == 1,]

Running Lasso (L1 penalty) in the GLMNET Package

lasso0 = glmnet(pred0, dep0, family="gaussian")

Run the Script:

- Click on the script window
- Type “CTRL + A” to select all of the script
- Type “CTRL + R” to run the selection

You can estimate Ridge regression (L2 penalty) models by setting “alpha = 0”. For example, in the cross-validation function:

cv.glmnet(pred0, dep0, family="gaussian", nfolds = 5, alpha = 0)

You can also estimate an Elastic Net (Elnet) model that equally weights the L1 and L2 penalties by setting “alpha = .5” in the cross-validation function:

cv.glmnet(pred0, dep0, family="gaussian", nfolds = 5, alpha = .5)