Lab 8: Penalized regression

Data: snps.csv (see the snps.pdf file for data dictionary/useful information).

Practical objective: To practice implementation and interpretation of penalized regression models.

Scientific objective: To develop a polygenic risk score to predict cardiovascular disease risk.

Noteworthy commands: Below is a list of Stata commands and options that will be helpful for this lab.

- `logit`
- `splitsample`
- `lasso logit`
- `elasticnet logit`

Exercises: Below is a set of exercises that we will go through individually, in small groups, and/or together as appropriate and as time permits.

Exercise 1: Load the snps.csv data set into Stata. Summarize the distribution of SNP 96 and SNP 329. What do you think will happen if you include all SNPs in a logistic regression model? Verify your response and show that logistic regression fails spectacularly without further modification. You will find the shortcut 'snp1-snp500' helpful to include all five-hundred SNPs in the model. Does it seem that SNPs 96 and 329 are the *only* ones causing a problem?

Exercise 2: Split the data set into a training set and a test set. For reproducibility purposes, set the seed as 6312 this problem (and for all problems in this lab).

Exercise 3: On the training set (sample = 1), try fitting a logistic regression model with SNPs 50 through 75 as predictors. Determine the training and test prediction error based on a binary classifier (specifically, try a cut-off of 0.5 for the predicted probability). Comment on your findings.

Exercise 4: Build a model based on the previous model, only including SNPs that produced a p-value smaller than 0.1. This should be SNPs 52, 54, 59, 64, 65, 69, and 73. Repeat Exercise 3 based on this updated model. Does it do any better? What does this suggest?

Exercise 5: On the training set only, fit an elastic net model that lives "halfway" between LASSO and ridge regression (i.e., with α=0.50). How many SNPs are selected into the model? Determine the training and test error and compare to Exercises 4 and 5.

Exercise 6: Before running an "almost-LASSO" model (i.e., with α=0.95), do you expect it to select *more*, *fewer*, or *about as many* variables into the model as compared to the model of Exercise 5? Verify your answer by running this model; determine the training and test error and compare to those of Exercise 5.

Exercise 7: Repeat Exercise 6 with an "almost-ridge" model (i.e., with α=0.05).