STP598: Computational Statistics

Yiannis Kamarianakis

E-mail: yiannis76@asu.edu; Office: WXLR 531

August 30, 2017

Assignment 1 Part 1 (50%)

The dataset *HtVol.csv* which can be downloaded from Blackboard, contains data from 58 patients. The response variable is *HtVol*; note that some explanatory variables are categorical.

- Develop a predictive model for *HtVol* using *Male*, *Age*, *Ht* and *Wt*. Feel free to evaluate interaction terms. Report estimated coefficients using least squares (LS) and least absolute deviations (LAD).
- Develop a predictive model for *HtVol* using *Male*, *Age*, *BMI* and *BSA*. Feel free to evaluate interaction terms. Report estimated coefficients using least squares and least absolute deviations.
- Perform 10-fold cross-validation to examine the predictive power of your 4 predictive models. Report RMSE, MAE and Symmetric Median Absolute Percentage Error ($sMdAPE$) $= median(200\frac{|Y_t - F_t|}{Y_t + F_t})$ in 3 matrices (one per performance metric): alternative models should be in columns, folds in rows.
- Discuss your findings; use a few plots to support your arguments.

Assignment 1 Part 2 (50%)

- Develop predictive models for the probability of Death using logistic regression and KNN and selected predictors from *Death.csv*. Set a random seed before beginning your analysis.
- Estimate the test error for the 2 models using 10-fold cross-validation.
- Use the validation set approach to estimate the test error of your models. In order to do this, you must perform the following steps:
  - Split the sample set into a training set and a validation set.
  - Fit the logistic regression model and KNN using only the training observations.
  - Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
  - Repeat the process in above three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

Assignment 1

- Submit your responses in Blackboard in a **single** pdf file by Thursday Sept. 13, midnight.
- Use the following file name: LASTNAME_FIRSTNAME_ASUID_ASSIGNMENTNUMBER
- Submit the script containing the `R` commands you used; use the same file name as before.
- Prepare your pdfs carefully; each week some of you will present their work.