HOMEWORK 3
BIOSTATISTICS 825
DUE OCTOBER 10TH, 2023

1. For this question, we'll use the dataset "parkinsons_updrs.txt" used in the previous homework. Recall, the data is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring.

   For all of the following, ignore that multiple observations come from the same person. This should not be done in practice, especially when we're doing any type of inference.

   (a) Expand the first 16 voice measurements using B-splines with 4 degrees of freedom and degree 3, i.e., use `bs( , df=4, degree = 3)`. Use this data for the rest of the problem.

   (b) Analyze the data using the grouped lasso model, similar to the last homework. However, this time use data splitting to get p-values and confidence intervals of all coefficients retained in the grouped lasso model. Note the "inference set" analyses will be based on standard least squares.

   (c) Analyze the data using a lasso model. Use *Selective Inference Tools* to get p-values and confidence intervals on the variables retained by the lasso model.

   (d) Compare the results from the previous two models. Which do you think is a better approach and why?

   (e) Using conformal inference, get prediction intervals for 100 observations left out of the analysis. That is, extract 100 observations then use the rest to train the model and get the conformal scores. Use a lasso regression with the same λ you used above.

   (f) How many of the prediction intervals contain the true Y value?

2. The dataset "pima.indians.diabetes3.xlsx" contains data from 532 females whom are at least 21 years of age and of Pima Indian heritage. The dataset consists of:

```
npregnant: Number of times pregnant
glucose: Plasma glucose concentration a 2 hours in an oral
glucose tolerance test
diastolic.bp: Diastolic blood pressure (mm Hg)
skinfold.thickness: Triceps skin fold thickness (mm)
bmi: Body mass index (weight in kg/(height in m)^2)
pedigree: Diabetes pedigree function
age: Age (years)
classdigit: Numerical class variable
class: Alphanumeric class variable
```

The point of this study was to indicate whether the patient shows signs of diabetes mellitus according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine care). The two class variables indicate diabetes status using a blood tests of 2-hour serum insulin. The goal is to use the other variables to predict diabetes status since getting information on them is easier for the patients, less invasive and cheaper than a blood-test. Use all other variables in each analysis.

(a) Compute a LDA, draw histograms of the first LDF coordinate by diabetes status.

(b) Compute a QDA of the data, draw histograms of the first LDF coordinate by diabetes status.

(c) Perform a logistic regression. Draw histograms of the estimated logit function by diabetes status.

(d) Comment on the figures. Which do you think will perform better classification?

(e) Using the pros and cons of LDA versus QDA versus logistic regression: which procedure do you think is the most appropriate for this data?

(f) Perform a leave one out cross-validation to see which method (LDA vs QDA vs Logistic) more accurately predicts the response.

(g) Saying a subject doesn't have diabetes when they actually do is dangerous. Ignored diabetes can lead to complications that include low blood sugar, cardiovascular disease along with damage to the eyes, kidneys and nerves. An expert panel suggested that saying someone doesn't have diabetes when they actually do is 4 times worse than saying they do when they actually don't. Report the confusion matrix (i.e., a 2×2 table of the true by the predicted classification) using the cost matrix.

(h) Perform a leave one out cross-validation to see which method (LDA vs QDA vs Logistic) leads to a lower expected cost due to misclassification. Use $1 and $4 (as appropriate) as the misclassification costs.

3. For this problem we'll use the "Communities and Crime Data Set" we used on the previous homework, however, you must re-download the data (see `Communities.xlsx`) as the variables have been streamlined to minimize missing data. The goal is to predict the variable "ViolentCrimesPerPop" (Y in the data) using the first 100 predictors (V1–V100).

   (a) Assuming the predictor variables for this data are economic and socio-demographic indicators, discuss whether 'local' or 'global' structure should be the focus of dimension reduction efforts.

   (b) Apply 3 of the dimension reduction techniques to the predictor variables.

   (c) Plot the first two coordinates of the methods applied in the previous question on separate plots (3 plots total). Make the color of the points equal to red if $Y > 0.5$ and black if $Y \leq 0.5$.

(d) Which of the dimension reduction techniques appears to separate $Y > 0.5$ vs $Y \leq 0.5$ most effectively?

(e) Which dimension reduction technique would you use to classify $Y > 0.5$ vs $Y \leq 0.5$? Use the pros and cons discussed in class along with your answer to the previous question.