15.072: The Analytics Edge **Fall 2021**
Short Homework 2: Logistic Regression

Out: September 27; Due: October 07, 11:59 pm.

Framingham Heart Study

Heart disease is the leading cause of death worldwide. About 17.9 million people died from coronary heart disease (CHD) in 2016—over 25% of all deaths that year across the globe.

In 1928, the U.S. government started to track a cohort of people in Framingham, MA. The study comprised initially 5,209 participants, who were given a questionnaire and a medical exam every two years. Data were also collected on their physical and behavioral characteristics. Over the years, the study has grown to include multiple generations and more variables. This dataset is known as the Framingham Heart Study.

The data is contained in the file `framingham.csv`. There are 3,658 observations, each corresponding to a participant. The 16 variables are described in Table 1.

Table 1: Variables in the dataset `framingham.csv`.

Variable	Description
`male`	Gender of patient (1 if male, 0 if female)
`age`	Age (in years) at first examination
`education`	Some high school, high school/GED, some college/vocational school, college
`currentSmoker`	1 if patient is a current smoker, 0 otherwise
`cigsPerDay`	Number of cigarettes per day
`BPMeds`	1 if patient is on blood pressure medication at time of first examination, 0 otherwise
`prevalentStroke`	1 if patient previously had a stroke, 0 otherwise
`prevalentHyp`	1 if patient is currently hypertensive, 0 otherwise
`diabetes`	1 if patient currently has diabetes, 0 otherwise
`totChol`	Total cholesterol (mg/dL)
`sysBP`	Systolic blood pressure
`diaBP`	Diastolic blood pressure
`BMI`	Body Mass Index: weight (kg)/height $(m)^2$
`heartRate`	Heart rate (beats/minute)
`glucose`	Blood glucose level (mg/dL)
`TenYearCHD`	1 if patient is experienced coronary heart disease within 10 years of first examination, 0 otherwise

To lower the risk of CHD, physicians can prescribe preventive medication that lowers blood pressure or cholesterol. Given the cost and possible side effects of preventive medications, these prescriptions require hard evidence on CHD risks. To support these decisions, you will predict whether a patient will experience CHD within 10 years of his/her first examination (`tenYearCHD`), and analyze risk factors.

A colleague of yours has just completed a health economics study to assess a recently approved medication. The study has estimated that patients who experience CHD within the next 10 years are expected to incur a lifetime cost of $165,000 associated with the disease—including the costs of treatment ($80,000) as well as lower quality of life and life expectancy ($85,000). The study has determined that the medicine will lower patients' risk of developing CHD within the next 10 years by a factor of 2.3. Regardless of whether a patient eventually develops CHD, the preventive medication costs $7,500.

Import the data into `R` and split them randomly into a training set (containing 75% of the data) and a test set (containing the remaining 25% of the data). Use the following seed and commands:

```
library(caTools)
data = read.csv("framingham.csv")
data$TenYearCHD <- factor(data$TenYearCHD)
data$male <- factor(data$male)
data$currentSmoker <- factor(data$currentSmoker)
data$BPMeds <- factor(data$BPMeds)
data$prevalentStroke <- factor(data$prevalentStroke)
data$prevalentHyp <- factor(data$prevalentHyp)
data$diabetes <- factor(data$diabetes)
set.seed(38)
N <- nrow(data)
idx = sample.split(data$TenYearCHD, 0.75)
train <- data[idx,]
test = data[!idx,]
```

a. Perform a brief exploratory data analysis. **[20 pt]**

   (i) Using two variables of your choice, construct conditional density plots to visualize the impact of the two factors on the patients' propensity to develop CHD.

   (ii) Comment briefly.

b. Using all the independent variables in the dataset, construct a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. **[20 pt]**

   (i) Describe, interpret and explain the coefficients of the model.

c. You aim to improve your model using Lasso. Perform 5-fold cross-validation, 10-fold cross-validation, and leave-one-out cross-validation, using the deviance as your performance metric. **[30 pts]**

   (i) For each cross-validation instance, report the computational time.

   (ii) For each cross-validation instance, plot the outputs.

   (iii) For each cross-validation instance, report the selected value of λ.

   (iv) Report the model's coefficients obtained with each selected value of λ.

   Hint: The computational time of a procedure in `R` can be obtained by applying the `Sys.time()` command before and after the procedure and subtracting the two values. If the procedure is one line, you can simply wrap it in the `system.time()` command.

   FYI: The deviance is equal to negative two times the maximum log-likelihood of your model, i.e.:

$$\text{Deviance} = -2\sum_{i=1}^{n}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

d. Build predictions on the test set, using the "full model" from Question b. and the one obtained with 10-fold cross-validation in Question c. For each one, denote by $\hat{p}_i$ the estimated probability that patient i will develop CHD. Let α be the threshold such that medication is prescribed to all patients i satisfying $\hat{p}_i \geq \alpha$. To assess your models, define two benchmarks: "baseline" practice (under which no medication is prescribed) and "ideal" practice (under which medication is only prescribed to patients that would otherwise develop CHD, assuming perfect *ex post* information on the test set). **[30 pt]**

   (i) Provide a plot showing the number of patients treated as a function of α, for the four models.

   (ii) Provide a plot showing the expected cost for all patients in the test set as a function of α, for the four models.

   (iii) Provide a plot showing the Receiver Operating Characteristics curve, for the four models.

   (iv) Report the Area under the Curve of the four models.

   (v) Comment briefly on your results.