**SOM 307: Data Analysis and Modeling for Business** **Due: March 23 (Mon), 2020**

# Homework 4: Logistic Regression

*Instructor: Dr. Akash Gupta* *Points: 30*

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This assignment intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The data is available on Canvas with the name *Heart Disease*.

Table 4.1: Meta Data

| Variable | Definition | Variable type |
|---|---|---|
| male | If sex is male (yes = 1) | Categorical |
| age | age of the patient | Continuous |
| education | education of the patient | Categorical |
| currentSmoker | whether or not the patient is a current smoker | Categorical |
| cigsPerDay | the number of cigarettes that the person smoked on average in one day | Continuous |
| BPMeds | whether or not the patient was on blood pressure medication | Categorical |
| prevalentStroke | whether or not the patient had previously had a stroke | Categorical |
| prevalentHyp | whether or not the patient was hypertensive | Categorical |
| diabetes | whether or not the patient had diabetes | Categorical |
| totChol: | total cholesterol level | Continuous |
| sysBP | systolic blood pressure | Continuous |
| diaBP | diastolic blood pressure | Continuous |
| BMI | Body Mass Index | Continuous |
| heartRate | Heart rate | Continuous |
| glucose | Glucose | Continuous |
| TenYearCHD | 10 year risk of coronary heart disease (CHD) | Categorical |

## Questions

1. Import dataset as *chdData*. Specify the number of variables and the number of observations.
2. Show first five observations.
3. Create a new dataset (*chdDataSelected*) by keeping following variables:
   - male
   - age
   - sysBP
   - heartRate
   - glucose

- TenYearCHD

4. Investigate the type of variables in *chdDataSelected*. Change the variable to the type shown in Table 4.1.
5. Produce a **table** tabulating the missing count of aforementioned six variables.
6. Impute missing values by the median for variables *glucose* and *heart rate*.
7. Make a density plot for variable *glucose*, similar to the one shown in Figure 4.1. (Hint: apply function *geom_density*).

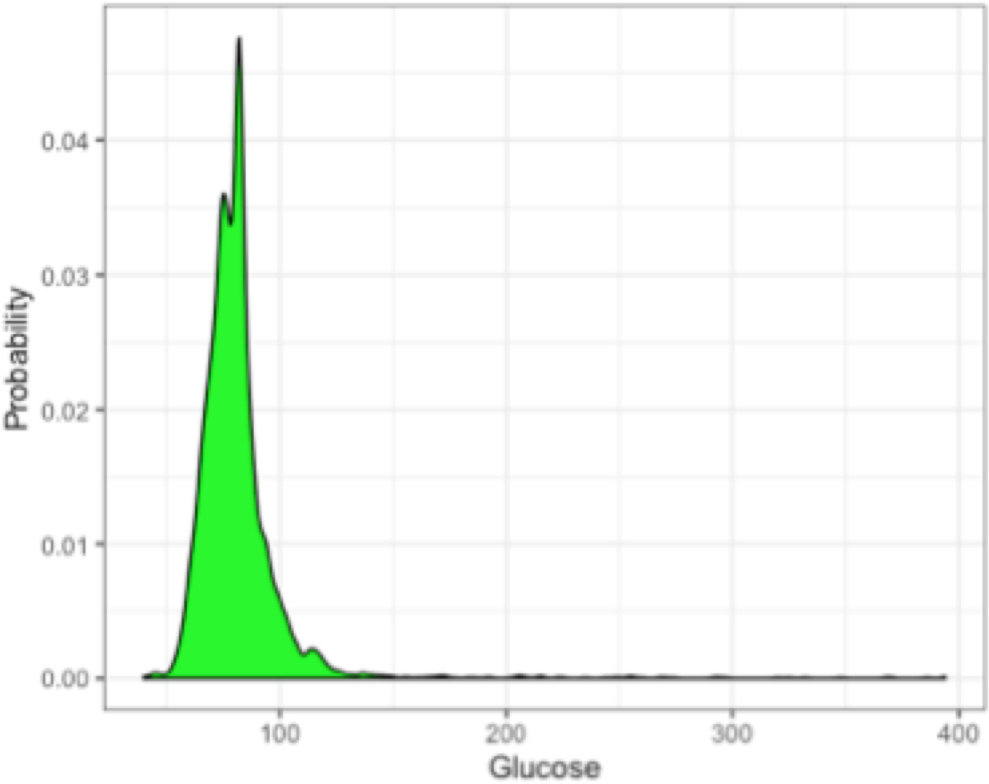

Figure 4.1: Density plot for Glucose

8. Remove observations that have missing values for other variables. Show your output after applying summary function. (Hint: you might want to use complete.cases function)
9. Make a bar plot showing ten year risk of heart disease for each gender
10. In this dataset (*chdDataSelected*), what is the proportion of the people with 10 year heart disease risk? (Hint:use function table)
11. In dataset (*chdDataSelected*), add another column *TenYearCHD_Encoded* representing *No* if TenYearCHD = 0, or *Yes* if TenYearCHD = 1.
12. Build three logistic regression models and report ROC. Your response variable is *TenYearCHD_Encoded*. (The arguements inside trainControl will be same as for the exercise we did in the class)

| Models | Variables |
|---|---|
| Model 1 | male + age |
| Model 2 | male + age + sysBP |
| Model 3 | male + age + sysBP + heartRate + glucose |

13. Write the logistic regression for the best model. You select the best model based on the model having the greatest value of ROC.

**Deliverables:**

- **Word File:** Your Word file should include segments of the code (NO SCREENSHOT) that address the specific answer. After each segment of the code should follow by the output. You can present the screenshot of the output from the **console** and the generated figure. If you present the screenshot of whole screen, I will NOT grade your submission. Your Word file is the most important document of the submission.
- **R Code:** Please also submit your R Code in a single file after adding the comments.