Lab 5: Logistic Regression

Collaboration in groups is highly encouraged; however, each individual must submit their own work.

Background

Approximately 18 million deaths occur worldwide each year due to heart diseases, with half of these fatalities reported in the United States and other developed countries stemming from cardiovascular diseases. Early prognosis of cardiovascular diseases can significantly aid in decisions regarding lifestyle changes for high-risk patients, ultimately reducing complications.
In this lab, we will employ logistic regression to predict a patient's 10-year risk of future coronary heart disease (CHD). Logistic regression, a statistical method, predicts the outcome of a categorical dependent variable based on a set of predictor or independent variables. In logistic regression, the dependent variable is binary, and it is primarily utilized for prediction and calculating the probability of success.

Data Source and Dictionary

The lab dataset is publicly accessible on Kaggle, originating from a continuous cardiovascular study in Framingham, Massachusetts. The objective is to forecast a patient's 10-year risk of future CHD. With over 4,000 records and 15 attributes, the dataset contains patients' details, encompassing demographic, behavioral, and medical risk factors. See the data dictionary below for attribute details.

Variable	Definition
male	male or female(Nominal)
Age	Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
education	Categorical variable for education with 4 levels: 1: no high school degree/GED 2: High School Graduate/GED 3: College Graduate 4: Post-College education
currentSmoker	whether or not the patient is a current smoker (Nominal)
cigsPerDay	the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
BPMeds	whether or not the patient was on blood pressure medication (Nominal)
prevalentStroke	whether or not the patient had previously had a stroke (Nominal)
prevalentHyp	whether or not the patient was hypertensive (Nominal)
diabetes	whether or not the patient had diabetes (Nominal)
totChol	total cholesterol level (Continuous)
sysBP	systolic blood pressure (Continuous)
fiaBP	diastolic blood pressure (Continuous)
BMI	Body Mass Index (Continuous)
Heart Rate	heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
glucose	glucose level (Continuous)
TenYearCHD	10 year risk of coronary heart disease CHD (binary: "1", means "Yes" and "0" means "No")

Proceed through the following steps to construct your model, addressing the questions as you progress:

1. Use *.rename()* to rename the 'male' column to 'sex', so that 1 denotes Yes and 0 denotes No.

2. Check for missing values and report which column has the most. Replace missing values in 'cigsPerDay' with zeros and drop rows containing missing values in other columns.

3. For simplicity, let's drop the 'education' variable.

4. Create two count plots to determine the balance of the data regarding the 10-year risk of coronary heart disease (CHD) and sex.

5. Group the dependent and independent variables and then create a logistic regression model with a data test size of 30% and a random state of 82.

6. Evaluate the model and explain the meaning of each value in the confusion matrix within the context of the data. Based on the balance of the 10-year risk of CHD, select and explain the values of at least two relevant metrics.

Submission

Upload only your **.ipynb** file on course site. Please save the file using the usual file naming format.