COMP809 – Logistic Regression

Lab 4

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 cardiovascular 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. The dataset framingham.csv is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. Each variable considered in this studied is a potential risk factor. There are both demographic, behavioral and medical risk factors. The variables are the following:

Demographic:

- sex: male or female.
- age: age of the patient.

Behavioural current

- Education: education level, being 0 the lowest level.
- Smoker: whether or not the patient is a current smoker.
- cigsPerDay: the number of cigarettes that the person smoked on average in one day.

Medical (history):

- BPMeds: whether or not the patient was on blood pressure medication.
- prevalentStroke: whether or not the patient had previously had a stroke.
- prevalentHyp: whether or not the patient was hypertensive.
- diabetes: whether or not the patient had diabetes.

Medical(current):

- totChol: total cholesterol level.
- sysBP: systolic blood pressure.
- diaBP: diastolic blood pressure.
- BMI: Body Mass Index.
- heartRate: heart rate.
- glucose: glucose level.

Predict variable (desired target):

- 10 year risk of coronary heart disease CHD ("1", means "Yes", "0" means "No").

Work in the following:

1. Is the response variable (TenYearCHD) unbalanced? If so,
    a. What are the implications of unbalanced data?
    b. How can you solve this problem?
    c. Implement the solution in 1.b and find a parsimonious logistic regression model. Check if the model provides an adequate fit for the data.
    d. Interpret the estimated coefficient associated to glucose.
2. Train the model with 70% of the data and with the remaining 30% calculate the accuracy, sensitivity, and specificity of the model. Comment on your findings.

3. Repeat question 2 using principal components as predictors. Compare the models.