Prediction of Type 2 Diabetes Using Logistic Model

MAT325 Final Project

Due: Friday, 5/12/2023

Contents

This project is an extension of the case study on predicting type II diabetes.

The data is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of collecting the data is to diagnostically **predict whether or not a patient has diabetes**, based on certain diagnostic measurements included in the data. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of *Pima Indian heritage*.

A portion of the above data with missing components removed is stored in the GitHub repository of this course: https://raw.githubusercontent.com/pengdsci/MAT325/main/w12/diabetes.csv

The following example code shows in the lecture note shows how to load data to R and extract variables for modeling through Newton method.

```
diabetes = read.csv("https://raw.githubusercontent.com/pengdsci/MAT325/main/w12/diabetes.csv")
diabeticStatus = diabetes$Outcome # y-variable: 1 = diabetes, 0 = no-diabetes
BMI = diabetes$BMI                # x-variable (risk factor)
```

Specific requirements:

1. The dependent variable is the **outcome** of diabetes (same as the one in the case study).
2. Choose **at least two** independent variables to build the *logistic predictive model*. (MBI was used in the case study, you can choose any two or more variables that you believe relevant to diabetes).
3. The general form of probability model and the objective function to maximize are

$$P[\text{d} = 1] = \frac{e^{\beta_0+\beta_1 x+\beta_2 y+\beta_3 z+\cdots}}{1+e^{\beta_0+\beta_1 x+\beta_2 y+\beta_3 z+\cdots}}$$

where d is the diabetes status (1 = diabetes, 0 = diabetes-free), $x, y, z, \cdots$ are variables (risk factor of diabetes) you choose to build your model. For a set of given points (tuples) based on historical records $\{(d_i, x_i, y_i, z_i, \cdots)\}_{i=1}^{n}$. The objective function of $\beta_0, \beta_1, \beta_2, \beta_3, \cdots$ to be maximized is

$$\mathbf{LL}(\beta_0, \beta_1, \beta_3, \cdots) = -\sum_{i=1}^{n} \ln[1 + e^{\beta_0+\beta_1 x_i+\beta_2 y_i+\beta_3 z_i+\cdots}] + \sum_{i=1}^{n} d_i(\beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 z_i + \cdots)$$

3.1. Derive the gradient vector $(\mathrm{LL}_{\beta_0}, \mathrm{LL}_{\beta_1}, \mathrm{LL}_{\beta_2}, \mathrm{LL}_{\beta_3}, \cdots)$.

3.2. Derive the Hessian matrix

$$H(\beta_0, \beta_1, \beta_3, \cdots) = \begin{bmatrix} \mathrm{LL}_{\beta_0\beta_0} & \mathrm{LL}_{\beta_0\beta_1} & \mathrm{LL}_{\beta_0\beta_2} & \mathrm{LL}_{\beta_0\beta_3} & \cdots \\ \mathrm{LL}_{\beta_1\beta_0} & \mathrm{LL}_{\beta_1\beta_1} & \mathrm{LL}_{\beta_1\beta_2} & \mathrm{LL}_{\beta_1\beta_3} & \cdots \\ \mathrm{LL}_{\beta_2\beta_0} & \mathrm{LL}_{\beta_2\beta_1} & \mathrm{LL}_{\beta_2\beta_2} & \mathrm{LL}_{\beta_2\beta_3} & \cdots \\ \mathrm{LL}_{\beta_3\beta_0} & \mathrm{LL}_{\beta_3\beta_1} & \mathrm{LL}_{\beta_3\beta_2} & \mathrm{LL}_{\beta_3\beta_3} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots \end{bmatrix}$$

4. Write code to implement the Newton method to find the value of $(\beta_0, \beta_1, \beta_2, \beta_3, \cdots)$ that maximize $\mathbf{LL}(\beta_0, \beta_1, \beta_2, \beta_3, \cdots)$.

5. Write the explicit prediction model with the approximated values of $(\beta_0, \beta_1, \beta_2, \beta_3, \cdots)$ and interpret the model (the positive or negative relationship between the risk variables and the disease status).

6. Since this logistic predictive model is a popular in both statistics and machine, there are well-developed functions to find the solution to the optimization problem. However, the goal of this project is to implement the Newton-Raphson method for the optimization. You are not expected to use any other existing 'black-box' functions to solve the optimization problem in this project.

7. Format of the project report.

   7.1 Title page: Title, your name, table of contents, etc

   7.2 Introduction: explain what is the goal and data source(s).

   7.3 The description of the model and the objective function to optimize.

   7.4 Description of the components in implementing the Newton methods: gradient vector, Hessian matrix, etc, and their derivations

   7.5 Numerical experiment: algorithm, visuals (figures and charts)

   7.6 Results and model interpretation and implications.

   References: data sources and other published papers you cited and/or summarized.

   Appendix: code and other supplementary figures and tables (if any).

Remarks

1. You can use code in any of the lecture notes.

2. Try to find the initial values by yourself first (hope you will find one). If you have difficult to find an initial value, please let me know, I will find one for you. Please keep in mind the you need to tell me what variables are used in your predictive model.