Assignment2_SG

Siliang Gong

November 13, 2021

Description of the data

The data used in this assignment is a Breast Cancer Coimbra data set, which is publicly available at UCI machine learning repository (link). The data set includes clinical features measurement for 64 patients with breast cancer and 52 healthy controls. There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models could be build to investigate the association between the clinical features and the disease status, and potentially benefit diagnosis. The data are stored in a csv format excel file, which is comma delimited.

Read the data

```
# read the data into a dataframe

breast.data <- read.csv("dataR2.csv")
```

Clean the data

```
# rename Classification to group and change group = 1 to "Healthy" and group =2 to "Patients"

breast.data = rename(breast.data, group = Classification)
breast.data$group[breast.data$group == 1] = "Healthy"
breast.data$group[breast.data$group == 2] = "Patients"

# select a subset of features
breast.data = breast.data%>%select(Age, BMI, Glucose, Insulin, group)
```

Data characteristics

This dataframe has 116 rows and 5 columns. The names of the columns and a brief description of each are in the table below:

```
Variable = colnames(breast.data)
Description = c("Age for each sample in years",
                "Body mass index (BMI), defined as the body mass divided
                by the square of the body height, in units of kg/m2",
                "Glucose level, measured in mg/dL",
                "Insulin level, measured in µU/mL",
                "Group indicator for each sample, either healthy control or patients")
vars = data.frame(Variable, Description)
kable(vars, caption =  "Summary of variables")
```

Table 1: Summary of variables

Variable	Description
Age	Age for each sample in years
BMI	Body mass index (BMI), defined as the body mass divided by the square of the body height, in units of kg/m2
Glucose	Glucose level, measured in mg/dL
Insulin	Insulin level, measured in µU/mL
group	Group indicator for each sample, either healthy control or patients

Summary statistics

```
# select Age, BMI and Glucose from the original data
sub.data = breast.data%>%select(Age, BMI, Glucose)

# obtain summary statistics

sub.summary = summary(sub.data)
sub.summary
```

```
##       Age             BMI           Glucose
##  Min.   :24.0   Min.   :18.37   Min.   : 60.00
##  1st Qu.:45.0   1st Qu.:22.97   1st Qu.: 85.75
##  Median :56.0   Median :27.66   Median : 92.00
##  Mean   :57.3   Mean   :27.58   Mean   : 97.79
##  3rd Qu.:71.0   3rd Qu.:31.24   3rd Qu.:102.00
##  Max.   :89.0   Max.   :38.58   Max.   :201.00
```

There are no missing values in any of the three variables.