Logistic Regression in R

By Kikonyogo Steven

Definition

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, in health, Logistic regression can also be used in the following areas:

- To identify risk factors and plan preventive measures;
- In drug research to tease apart the effectiveness of medicines on health outcomes across age, gender and ethnicity;

About the project

I used data of 318 individuals with and without diabetes type 2.The aim is to examine the relationship between age, gender, BMI, diet type, smoking status and family history of disease and build a model for predicting diabetes risk.

(i) Importing data into R and loading packages

```
Diabetes <- rio::import(here::here("diabetes.csv"))
 if(!require(pacman)) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load( tidyverse, janitor, plotrix, gtsummary, survival )
```

(ii) Data manipulation and exploration

```
Diabetes$gender <- factor(Diabetes$gender, levels = c("1","2"),
                          labels = c("male","female"))

Diabetes$smoking <- factor(Diabetes$smoking, levels = c("1","2"),
                          labels = c("smokers","Non-smokers"))
Diabetes$diabetes <- factor(Diabetes$diabetes, levels = c("0","1"),
                           labels = c("Non-diabetic","Diabetic"))

Diabetes$veg <- factor(Diabetes$veg, levels = c("1","2"),
                           labels = c("vegeterian","Non-vegeterian"))

Diabetes$familiy_history <- factor(Diabetes$familiy_history, levels = c("1","2"),
```

```
                      labels = c("Yes","No"))



Diabetes <- Diabetes %>%
  mutate(BMI = weight/height^2)
Diabetes <- Diabetes %>%
  select(age, gender, smoking,weight, diabetes,veg, familiy_history, BMI) %>%
  mutate(Agegroup=
           ifelse(age < 65, "< 65", "> =65"),
    BMIclass = case_when(
    BMI < 18.5~ "Underweight",
    BMI  >=18.5 & BMI < 24.9 ~ "Normal weight",
   TRUE ~ "Obese"
  ))
dim(Diabetes)
```

```
## [1] 318  10
```

```
summary(Diabetes)
```

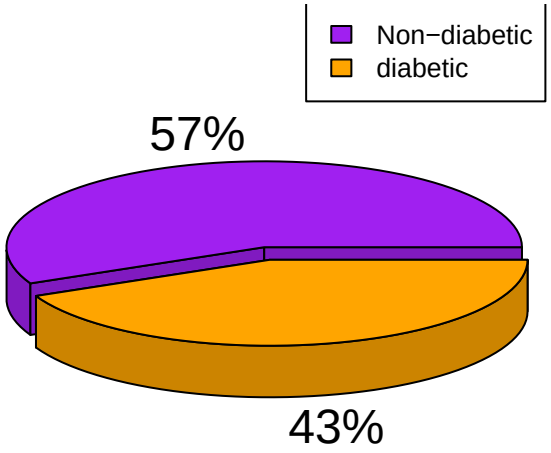
```
##       age           gender           smoking        weight
##  Min.   :17.00   male  :127   smokers    :154   Min.   : 39.30
##  1st Qu.:54.00   female:191   Non-smokers:164   1st Qu.: 61.00
##  Median :65.00                                  Median : 67.00
##  Mean   :63.29                                  Mean   : 72.01
##  3rd Qu.:75.00                                  3rd Qu.: 78.00
##  Max.   :89.00                                  Max.   :186.00
##          diabetes               veg      familiy_history      BMI
##  Non-diabetic:181   vegeterian    :165   Yes: 81         Min.   :13.60
##  Diabetic    :137   Non-vegeterian:153   No :237         1st Qu.:21.08
##                                                         Median :23.88
##                                                         Mean   :25.01
##                                                         3rd Qu.:27.40
##                                                         Max.   :57.52
##    Agegroup           BMIclass
##  Length:318         Length:318
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
table1 <- table(Diabetes$diabetes)

piepercent <-paste0(round(100 * table1/sum(table1)), "%")

plotrix::pie3D(table1,radius = 1.0,
               explode = 0.05,
               labels = piepercent,
               col = c("purple","orange"),
               main = "Diabetic Vs Non-diabetic compositions",
```

```
               col.main="blue")
legend("topright",
       c("Non-diabetic","diabetic"),
       cex = 0.8,
       fill =c("purple","orange") )
```



(iii) **Descriptive statistics**

```
 descriptives <- Diabetes %>%
  select(age,BMI,diabetes,Agegroup,gender,BMIclass,smoking,veg,familiy_history)
tbl_summary(descriptives,
            type = list(familiy_history ~ "categorical") ) %>%
  modify_caption("Table1: Descriptive statistics") %>%
  bold_labels()
```

Table 1: Table1: Descriptive statistics

Characteristic	**N = 318**
age	65 (54, 75)
BMI	23.9 (21.1, 27.4)
diabetes	
Non-diabetic	181 (57%)
Diabetic	137 (43%)

Characteristic	N = 318
Agegroup	
< 65	152 (48%)
> =65	166 (52%)
gender	
male	127 (40%)
female	191 (60%)
BMIclass	
Normal weight	164 (52%)
Obese	131 (41%)
Underweight	23 (7.2%)
smoking	
smokers	154 (48%)
Non-smokers	164 (52%)
veg	
vegeterian	165 (52%)
Non-vegeterian	153 (48%)
familiy_history	
Yes	81 (25%)
No	237 (75%)

The minimum age was 17yrs and maximum was 89yrs.The median age was 65 yrs and mean age was 63 yrs. 127 were males and 191 were females. 137 were diabetic and 181 were not.

(iv) **Testing for significance between diabetic and non diabetic**

```
Diabetes %>%
 select(diabetes,age,BMI,Agegroup,gender,BMIclass,smoking,familiy_history,veg) %>%
 tbl_summary(by=diabetes,
             label = list(age ~ "mean age,yrs (sd)",
                          BMI~ "median BMI (IQR)",
                         smoking ~ " smoking status",
                         veg ~ " Diet type",
                         familiy_history ~ " Family history"),
         statistic = list(age ~ "{mean}({sd})"),
type = list(familiy_history ~ "categorical")) %>%
     bold_labels() %>%
  add_p(test = list(
    all_continuous() ~ "t.test",
    all_categorical() ~ "fisher.test")) %>%
 modify_caption("Table2: Statistical significance difference by diabetes status")
```

Table 2: Table2: Statistical significance difference by diabetes status

Characteristic	**Non-diabetic**, N = 181	**Diabetic**, N = 137	**p-value**
mean age,yrs (sd)	57(16)	71(12)	<0.001
median BMI (IQR)	23.3 (20.7, 26.2)	24.4 (21.3, 28.5)	0.001
Agegroup			<0.001
< 65	115 (64%)	37 (27%)	
> =65	66 (36%)	100 (73%)	

Characteristic	**Non-diabetic**, N = 181	**Diabetic**, N = 137	**p-value**
gender			0.4
male	68 (38%)	59 (43%)	
female	113 (62%)	78 (57%)	
BMIclass			0.009
Normal weight	100 (55%)	64 (47%)	
Obese	63 (35%)	68 (50%)	
Underweight	18 (9.9%)	5 (3.6%)	
___ smoking status___			0.055
smokers	79 (44%)	75 (55%)	
Non-smokers	102 (56%)	62 (45%)	
___ Family history___			0.020
Yes	37 (20%)	44 (32%)	
No	144 (80%)	93 (68%)	
___ Diet type___			0.002
vegeterian	108 (60%)	57 (42%)	
Non-vegeterian	73 (40%)	80 (58%)	

There was significant difference in mean age, mean BMI, smoking status, diet type and family history of the disease between the diabetic and non diabetic individuals. (All $p < 0.05$)

(vi) **Univariate analysis of diabetes risk factors**

```
Diabetes %>%
  select(diabetes,Agegroup,gender,BMIclass,smoking,familiy_history,veg) %>%
  tbl_uvregression(
    method = glm,
    y = diabetes,
    method.args = list(family = binomial),
    exponentiate = TRUE) %>%
  modify_caption("Table3: Univariate analysis of diabetes risk factors") %>%
  bold_p() %>%
  italicize_levels() %>%
  bold_labels()
```

Table 3: Table3: Univariate analysis of diabetes risk factors

Characteristic	**N**	**OR**	**95% CI**	**p-value**
Agegroup	318			
< 65		—	—	
> =65		4.71	2.93, 7.71	**<0.001**
gender	318			
male		—	—	
female		0.80	0.51, 1.25	0.3
BMIclass	318			
Normal weight		—	—	
Obese		1.69	1.06, 2.69	**0.027**
Underweight		0.43	0.14, 1.15	0.12
smoking	318			
smokers		—	—	

Characteristic	N	OR	95% CI	p-value
Non-smokers		0.64	0.41, 1.00	0.050
familiy_history	318			
Yes		—	—	
No		0.54	0.33, 0.90	**0.019**
veg	318			
vegeterian		—	—	
Non-vegeterian		2.08	1.33, 3.27	**0.002**

(vi) **Multivariate analysis of diabetes risk factors**

```
multivariate <- glm(
  diabetes~ Agegroup+BMIclass+smoking+familiy_history+veg,
  data = Diabetes, family=binomial)


tbl_regression(multivariate, exponentiate = TRUE) %>%
  bold_p() %>%
  bold_labels() %>%
  italicize_levels() %>%
  modify_caption("Table4: Multivariate analysis of diabetes risk factors")
```

Table 4: Table4: Multivariate analysis of diabetes risk factors

Characteristic	OR	95% CI	p-value
Agegroup			
< 65	—	—	
> =65	5.40	3.24, 9.22	**<0.001**
BMIclass			
Normal weight	—	—	
Obese	1.85	1.10, 3.13	**0.020**
Underweight	0.36	0.10, 1.05	0.076
smoking			
smokers	—	—	
Non-smokers	0.55	0.33, 0.92	**0.024**
familiy_history			
Yes	—	—	
No	0.55	0.30, 0.97	**0.039**
veg			
vegeterian	—	—	
Non-vegeterian	1.96	1.19, 3.25	**0.008**

All the risk factors that were significant risk factors for diabetes in univariate analysis were also significant in multivariate analysis. The factors were: (i) age(>=65yrs)[OR = 5.4(3.2 - 9.2; p < 0.001)] (ii) Obesity[OR = 1.85(1.1 - 3.1:; p = 0.02)] (iii) Non smoking[OR = 0.55(0.33 - 0.92); P = 0.02] (iv) Non vegetarian[OR = 1.96(1.19 - 3.25); P = 0.008] (v) No family history of disease[OR = 0.55(0.3 -30.97);p=0.04]