logistic回归案例：健康信息搜寻行为研究

吴翔

2019-03-22

概述

我们通过案例来阐述如何使用logistic回归模型。

- 二项logistic回归
- 多项logistic回归
- 次序logistic回归

```
# clean the work directory
rm(list = ls())

# set seeds
set.seed(123)

# read dataset
suppressMessages(library(tidyverse))
suppressMessages(library(pander))
suppressMessages(library(stargazer))
load("hisb.RData")
```

可以看到，数据集包含1814个样本和6个变量。

```
# display variables
str(hisb)
```

```
## 'data.frame':    1814 obs. of  6 variables:
##  $ age      : num  49 72 38 55 67 40 86 40 73 52 ...
##  $ gender   : Factor w/ 2 levels "Female","Male": 1 2 1 2 2 1 2 1 2 2 ...
##  $ race     : Factor w/ 2 levels "Others","White": 2 2 2 2 2 1 2 2 2 2 ...
##  $ education: Factor w/ 2 levels "Under College",..: 1 1 1 2 2 2 2 2 1 1 ...
##  $ income   : Factor w/ 3 levels "$0 to $19,999",..: 3 2 2 3 2 3 3 3 2 3 ...
##  $ y        : Factor w/ 3 levels "Doctor","Internet",..: 2 3 2 2 2 2 2 2 3 2 ...
```

各变量含义如下：

- 健康信息来源 `y`：包括互联网、医生和其它来源。
- 年龄 `age`
- 性别 `gender`
- 种族 `race`
- 教育水平 `education`
- 收入 `income`

各个变量分布情况如下：

```
# age
summary(hisb$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      19      43      57      55      66     101
```

```
# gender
table(hisb$gender)
```

```
## 
## Female   Male 
##   1050    764
```

```
# race
table(hisb$race)
```

```
## 
## Others  White 
##    355   1459
```

```
# education
table(hisb$education)
```

```
## 
##     Under College College and above 
##               838               976
```

```
# income
table(hisb$income)
```

```
## 
##      $0 to $19,999 $20,000 to $74,999    $75,000 or more 
##                237                808                769
```

```
# hisb
table(hisb$y)
```

```
## 
##   Doctor Internet   Others 
##      291     1320      203
```

二项logistic回归

考虑如下问题：哪些民众更倾向使用互联网作为健康信息来源？我们采用 `glm()` 函数估计二项logistic回归模型。

```
# create a binary response variable
hisb.bl <- hisb
hisb.bl$y <- ifelse(hisb.bl$y == "Internet", 1, 0)

# fit the logistic regression model
bl.fit <- glm(y ~ ., family = binomial(), data = hisb.bl)
summary(bl.fit)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(), data = hisb.bl)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.566  -0.862   0.510   0.780   1.817
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.35259    0.29586    7.95  1.8e-15 ***
## age                       -0.05043    0.00431  -11.69  < 2e-16 ***
## genderMale                -0.03720    0.11918   -0.31   0.7550
## raceWhite                  0.64694    0.14190    4.56  5.1e-06 ***
## educationCollege and above  0.37010    0.12278    3.01   0.0026 **
## income$20,000 to $74,999    0.87564    0.16555    5.29  1.2e-07 ***
## income$75,000 or more       1.26223    0.18502    6.82  9.0e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2124.4  on 1813  degrees of freedom
## Residual deviance: 1813.9  on 1807  degrees of freedom
## AIC: 1828
##
## Number of Fisher Scoring iterations: 4
```

由于原始参数$\hat{\beta}$不易解释，我们撰写函数计算相应的OR值和置信区间。

```
# write a function to calculate the OR and CI
orsummary.bl <- function(fit){
    # calculate OR and CI
    y <- exp(cbind(coef(fit), confint(fit)))
    # rename the matrix y
    colnames(y)[1] <- "OR"
    # column bind with estimate and p-value
    y <- cbind(summary(fit)$coef[, c(1, 4)], y)
    # adjust column order
    y <- y[, c(1, 3:5, 2)]
    # return the matrix
    return(y)
}
# calculate OR and CI
orstat.bl <- orsummary.bl(bl.fit)
# display the ORs
rownames(orstat.bl) <- c("intercept", "age", "male", "white", "college and above", "$20,000 to 74,999",
"$75,000 or more")
pandoc.table(orstat.bl)
```

	Estimate	OR	2.5 %	97.5 %	Pr(>|z|)
intercept	2.353	10.51	5.924	18.91	1.838e-15
age	-0.05043	0.9508	0.9427	0.9588	1.373e-31
male	-0.0372	0.9635	0.7629	1.217	0.755

	Estimate	OR	2.5 %	97.5 %	Pr(>|z|)
white	0.6469	1.91	1.445	2.521	5.135e-06
college and above	0.3701	1.448	1.138	1.842	0.002575
$20,000 to 74,999	0.8756	2.4	1.737	3.325	1.228e-07
$75,000 or more	1.262	3.533	2.461	5.085	8.967e-12

```
# output as a table
stargazer(bl.fit, type = "html")
```

	Dependent variable:
	y
age	-0.050***
	(0.004)
genderMale	-0.037
	(0.120)
raceWhite	0.650***
	(0.140)
educationCollege and above	0.370***
	(0.120)
74,999	0.880***
	(0.170)
75,000 or more	1.300***
	(0.180)
Constant	2.400***
	(0.300)
Observations	1,814
Log Likelihood	-907.000
Akaike Inf. Crit.	1,828.000
Note:	*p<0.1;* ***p<0.05;*** p<0.01

多项logistic回归

次序logistic回归