logistic 回归案例：健康信息搜寻行为研究

吴翔

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

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

# set seeds
set.seed(123)

# read dataset
suppressMessages(library(tidyverse))
suppressMessages(library(pander))
panderOptions('round',2)
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？**

当观测样本 i 使用互联网作为健康信息来源时，记作 $y_i = 1$；否则，记作 $y_i = 0$。将所有其它变量纳入模型作为自变量，用以解释民众使用互联网作为健康信息来源的概率 p。因此，二项 logistic 回归模型如下：

$$\text{logit}(p_i) = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 Race_i + \beta_4 Educ_i + \beta_5 Inc_i + \epsilon_i.$$

进一步，考虑到二水平和多水平的分类自变量（categorical independent variable），我们将其虚拟变量化，用 $k-1$ 个虚拟变量来表示 k 个水平的分类自变量。因此，二项 logistic 回归模型重新表示为：

$$\text{logit}(p_i) = \beta_0 + \beta_1 Age_i + \beta_2 GenderM_i + \beta_3 RaceW_i + \beta_4 EducH_i + \beta_5 IncM_i + \beta_6 IncH_i + \epsilon_i.$$

注意，收入变量有三个水平，我们以低收入水平（年收入 19,999 美元以内）作为参照水平（reference level），而将其它中等收入和高等收入水平作为虚拟变量纳入模型。只使用 $k-1$ 个虚拟变量的原因在于，避免出现完全多重共线性。

我们采用 `glm()` 函数估计二项 logistic 回归模型，得到如下结果：

```r
# 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
```

考虑到 age 不可能为 0，为了使截距项有实际意义，我们将年龄变量做对中（即减去其均值）处理。

```
# centering age variable
hisb.bl$age <- scale(hisb.bl$age, center = TRUE, scale = FALSE)
# 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)                 -0.43365    0.17298   -2.51    0.0122 *
## 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,
pandoc.table(orstat.bl, digits = 2)
```

	Estimate	OR	2.5 %	97.5 %	Pr(>|z|)
intercept	-0.43	0.65	0.46	0.91	0.01
age	-0.05	0.95	0.94	0.96	0
male	-0.04	0.96	0.76	1.2	0.75
white	0.65	1.9	1.4	2.5	0
college and above	0.37	1.4	1.1	1.8	0
$20,000 to 74,999	0.88	2.4	1.7	3.3	0
$75,000 or more	1.3	3.5	2.5	5.1	0

类似的，我们返回最大对数似然值。

```r
# LL
logLik(bl.fit)
```

```
## 'log Lik.' -907 (df=7)
```

最后，估计空模型。

```r
# fit the null logistic regression model
bl.fit.null <- glm(y ~ 0, family = binomial(), data = hisb.bl)
summary(bl.fit.null)
```

```
##
## Call:
## glm(formula = y ~ 0, family = binomial(), data = hisb.bl)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.18   -1.18    1.18    1.18    1.18
##
## No Coefficients
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2514.7  on 1814  degrees of freedom
## Residual deviance: 2514.7  on 1814  degrees of freedom
## AIC: 2515
##
## Number of Fisher Scoring iterations: 0
```

返回空模型的 LL，并由此可以计算伪 R^2。

```r
# calculate R square
logLik(bl.fit.null)
```

```
## 'log Lik.' -1257 (df=0)
```

```r
rsq <- (logLik(bl.fit.null) - logLik(bl.fit)) / logLik(bl.fit.null)
rsq
```

```
## 'log Lik.' 0.28 (df=0)
```

多项 logistic 回归

类似地，我们采用 `nnet` 包中的 `multinom()` 函数估计多项 logistic 模型。

```r
rm(list = ls())
load("hisb.RData")

# create a binary response variable
hisb.ml <- hisb
hisb.ml$age <- scale(hisb.ml$age, center = TRUE, scale = FALSE)

# fit the multinomial logistic regression model
suppressMessages(library(nnet))
ml.fit <- multinom(y ~ ., data = hisb.ml)
```

```
## # weights:  24 (14 variable)
## initial  value 1992.882692
## iter  10 value 1253.592652
## iter  20 value 1239.182484
## final  value 1239.182108
## converged
```

```r
summary(ml.fit)
```

```
## Call:
## multinom(formula = y ~ ., data = hisb.ml)
##
## Coefficients:
##          (Intercept)     age genderMale raceWhite educationCollege and above
## Internet       0.244 -0.0528    -0.0405      0.53                      0.397
## Others        -0.026 -0.0057    -0.0086     -0.26                      0.065
##          income$20,000 to $74,999 income$75,000 or more
## Internet                    0.839                  1.13
## Others                     -0.081                 -0.31
##
## Std. Errors:
##          (Intercept)    age genderMale raceWhite educationCollege and above
## Internet        0.20 0.0052       0.14      0.17                       0.15
## Others          0.23 0.0068       0.19      0.21                       0.20
##          income$20,000 to $74,999 income$75,000 or more
## Internet                     0.19                  0.22
## Others                       0.23                  0.28
##
## Residual Deviance: 2478
## AIC: 2506
```

类似地，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.ml <- function(fit, j = 1){
    # calculate OR and CI
    y <- exp(cbind(coef(fit)[j, ], confint(fit)[,,j]))
    # calculate z values
    zvalues <- summary(fit)$coefficients / summary(fit)$standard.errors
    # calculate p values
    pvalues <- pnorm(abs(zvalues[j, ]), lower.tail = F) * 2
    # column bind with estimate and p-value
    y <- cbind(coef(fit)[j, ], y, pvalues)
    # rename column names
    colnames(y)[c(1, 2, 5)] <- c("Estimates", "OR", "Pr(>|z|)")
    # return the matrix
    return(y)
}

# calculate model statistics
internet.or <- orsummary.ml(ml.fit, j = 1)
other.or <- orsummary.ml(ml.fit, j = 2)
```

最后，展示最终结果。

```
# display the ORs
rn <- c("intercept", "age", "male", "white", "college and above", "$20,000 to 74,999", "$75,000 or
rownames(internet.or) <- rn
rownames(other.or) <- rn
pandoc.table(internet.or)
```

	Estimates	OR	2.5 %	97.5 %	Pr(>|z|)
intercept	0.24	1.28	0.86	1.9	0.23
age	-0.05	0.95	0.94	0.96	0
male	-0.04	0.96	0.73	1.27	0.78
white	0.53	1.7	1.22	2.38	0
college and above	0.4	1.49	1.11	1.99	0.01
$20,000 to 74,999	0.84	2.31	1.58	3.38	0
$75,000 or more	1.13	3.11	2.03	4.77	0

```
pandoc.table(other.or)
```

	Estimates	OR	2.5 %	97.5 %	Pr(>|z|)
intercept	-0.03	0.97	0.62	1.53	0.91
age	-0.01	0.99	0.98	1.01	0.4
male	-0.01	0.99	0.68	1.44	0.96
white	-0.26	0.77	0.51	1.16	0.21
college and above	0.06	1.07	0.72	1.58	0.75
$20,000 to 74,999	-0.08	0.92	0.59	1.44	0.72
$75,000 or more	-0.31	0.73	0.42	1.27	0.27