Example 3 Explained

Yu-Chen Xue

Apr 25, 2018

Example 3 – UCBAdmission Explained

1. UCBAdmissions 資料集的 Overview

UCBAdmissions 是儲存為 3D array 格式的資料集，這裡先將 UCBAdmissions 轉換為 dataframe，並命名為 ucb_df，以直觀地展示這個資料集

```
ucb_df <- data.frame(UCBAdmissions)
ucb_df
```

```
##       Admit Gender Dept Freq
## 1  Admitted   Male    A  512
## 2  Rejected   Male    A  313
## 3  Admitted Female    A   89
## 4  Rejected Female    A   19
## 5  Admitted   Male    B  353
## 6  Rejected   Male    B  207
## 7  Admitted Female    B   17
## 8  Rejected Female    B    8
## 9  Admitted   Male    C  120
## 10 Rejected   Male    C  205
## 11 Admitted Female    C  202
## 12 Rejected Female    C  391
## 13 Admitted   Male    D  138
## 14 Rejected   Male    D  279
## 15 Admitted Female    D  131
## 16 Rejected Female    D  244
## 17 Admitted   Male    E   53
## 18 Rejected   Male    E  138
## 19 Admitted Female    E   94
## 20 Rejected Female    E  299
## 21 Admitted   Male    F   22
## 22 Rejected   Male    F  351
## 23 Admitted Female    F   24
## 24 Rejected Female    F  317
```

製作 UCBAdmissions 資料集的 ‘flat’ contingency tables，並指定欄 “Admit” 為比較的目標欄

```
ftable(UCBAdmissions, col.vars="Admit")
```

```
##             Admit Admitted Rejected
## Gender Dept
## Male   A               512      313
##        B               353      207
##        C               120      205
##        D               138      279
##        E                53      138
##        F                22      351
## Female A                89       19
##        B                17        8
##        C               202      391
##        D               131      244
##        E                94      299
##        F                24      317
```

顯示 UCBAdmissions 資料集中解釋變數的組成

```
dimnames(UCBAdmissions)
```

```
## $Admit
## [1] "Admitted" "Rejected"
##
## $Gender
## [1] "Male"   "Female"
##
## $Dept
## [1] "A" "B" "C" "D" "E" "F"
```

顯示 UCBAdmissions 資料集的 margin table（一個顯示某個欄位的各個數值的個數的表格），這裡選取 UCBAdmissions 的第 1 個解釋變數 (Admit) 相對第 2 個解釋變數 (Gender) 的個數

```
margin.table(UCBAdmissions, c(2,1))
```

```
##         Admit
## Gender   Admitted Rejected
##   Male       1198     1493
##   Female      557     1278
```

顯示 **UCBAdmissions** 資料集的 **margin table**，這裡選取 **UCBAdmissions** 的第 **1** 個解釋變數 **(Admit)** 相對第 **3** 個解釋變數 **(Dept)** 的個數

```
margin.table(UCBAdmissions, c(3,1))
```

```
##     Admit
## Dept Admitted Rejected
##    A      601      332
##    B      370      215
##    C      322      596
##    D      269      523
##    E      147      437
##    F       46      668
```

顯示 **UCBAdmissions** 資料集的 **margin table**，這裡選取 **UCBAdmissions** 的第 **3** 個解釋變數 **(Dept)** 相對第 **2** 個解釋變數 **(Gender)** 的個數

```
margin.table(UCBAdmissions, c(2,3))
```

```
##         Dept
## Gender     A   B   C   D   E   F
##   Male   825 560 325 417 191 373
##   Female 108  25 593 375 393 341
```

2. 對 UCBAdmissions 資料集進行建模、分析

以另一種方法展現這個資料集，將 **Admit** 拆分為 **yes** 和 **no** 兩種情況

```
### begin copying here
ucb.df = data.frame(gender=rep(c("Male","Female"),c(6,6)),
                    dept=rep(LETTERS[1:6],2),
                    yes=c(512,353,120,138,53,22,89,17,202,131,94,24),
                    no=c(313,207,205,279,138,351,19,8,391,244,299,317))
### end copying here and paste into the R Console

ucb.df
```

```
##    gender dept yes  no
## 1    Male    A 512 313
## 2    Male    B 353 207
## 3    Male    C 120 205
## 4    Male    D 138 279
## 5    Male    E  53 138
## 6    Male    F  22 351
## 7  Female    A  89  19
## 8  Female    B  17   8
## 9  Female    C 202 391
## 10 Female    D 131 244
## 11 Female    E  94 299
## 12 Female    F  24 317
```

將 yes/no 作為響應變數，Gender * Dept 作為解釋變數，對 UCBAdmission 資料集構建廣義線性模型，其中 family=binomial(logit) 表示指定使用邏輯回歸。

```
mod.form = "cbind(yes,no) ~ gender * dept"     # mind the quotes here!
glm.out = glm(mod.form, family=binomial(logit), data=ucb.df)
```

對模型作 anova 表格，test=“Chisq” 表示在輸出的表格中加上 Pr(>Chi) 一欄

可以看到 “gender” 和 “detp” 對學生是否被錄取都很有關係

```
anova(glm.out, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(yes, no)
##
## Terms added sequentially (first to last)
##
```

```
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           11     877.06
## gender       1    93.45        10     783.61 < 2.2e-16 ***
## dept         5   763.40         5      20.20 < 2.2e-16 ***
## gender:dept  5    20.20         0       0.00  0.001144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

對模型作 summary 表格

其中 z-value 為 Critical value，“genderMale:deptB” 的期望值表示若同時為 gemderMale 和 deptB，則計算這種情況的 ln(odds) 的式子需要額外考慮這一項，即 1.5442-1.0521-0.7904+0.8321 “Residual deviance: 1.0791e-13 on 0 degrees of freedom” 說明該模型很完善地解釋了這個資料集

```r
summary(glm.out)
```

```
##
## Call:
## glm(formula = mod.form, family = binomial(logit), data = ucb.df)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.5442     0.2527   6.110 9.94e-10 ***
## genderMale          -1.0521     0.2627  -4.005 6.21e-05 ***
## deptB               -0.7904     0.4977  -1.588  0.11224
## deptC               -2.2046     0.2672  -8.252  < 2e-16 ***
## deptD               -2.1662     0.2750  -7.878 3.32e-15 ***
## deptE               -2.7013     0.2790  -9.682  < 2e-16 ***
## deptF               -4.1250     0.3297 -12.512  < 2e-16 ***
## genderMale:deptB     0.8321     0.5104   1.630  0.10306
## genderMale:deptC     1.1770     0.2996   3.929 8.53e-05 ***
## genderMale:deptD     0.9701     0.3026   3.206  0.00135 **
## genderMale:deptE     1.2523     0.3303   3.791  0.00015 ***
## genderMale:deptF     0.8632     0.4027   2.144  0.03206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.7706e+02  on 11  degrees of freedom
## Residual deviance: 1.0791e-13  on  0  degrees of freedom
## AIC: 92.94
##
## Number of Fisher Scoring iterations: 3
```

計算 genderMale 的期望值的 exponential

```r
exp(-1.0521)
```

```
## [1] 0.3492037
```

計算 genderMale 的期望值的 exponential 的相反數

```r
1/exp(-1.0521)
```

```
## [1] 2.863658
```

計算 deptC 的期望值的 exponential

```r
exp(-2.2046)
```

```
## [1] 0.1102946
```

將上面的兩個 exponential 數值相除，得到 genderMale:deptD 的近似值

```r
exp(-2.2046) / exp(-2.1662)          # C:A / D:A leaves C:D
```

```
## [1] 0.9623279
```

將 yes/no 作為響應變數，Gender * Dept 作為解釋變數，對 UCBAdmission 資料集構建邏輯回歸模型，並對該模型作 anova 表格進行分析

在這個模型下，“gender” 對結果沒有什麼貢獻

```
mod.form="cbind(yes,no) ~ dept + gender"
glm.out=glm(mod.form, family=binomial(logit), data=ucb.df)
anova(glm.out, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(yes, no)
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                      11     877.06
## dept    5   855.32         6      21.74   <2e-16 ***
## gender  1     1.53         5      20.20   0.2159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

對模型作 summary 表格進行分析

“genderMale” 對結果沒什麼貢獻 “Residual deviance: 20.204 on 5 degrees of freedom” 說明這個模型的解釋能力不如前面的模型

```
summary(glm.out)
```

```
##
## Call:
## glm(formula = mod.form, family = binomial(logit), data = ucb.df)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -1.2487  -0.0560   1.2533   0.0826   1.2205  -0.2076   3.7189   0.2706
##       9       10       11       12
## -0.9243  -0.0858  -0.8509   0.2052
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68192    0.09911   6.880 5.97e-12 ***
## deptB       -0.04340    0.10984  -0.395    0.693
## deptC       -1.26260    0.10663 -11.841  < 2e-16 ***
## deptD       -1.29461    0.10582 -12.234  < 2e-16 ***
## deptE       -1.73931    0.12611 -13.792  < 2e-16 ***
## deptF       -3.30648    0.16998 -19.452  < 2e-16 ***
## genderMale  -0.09987    0.08085  -1.235    0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.056  on 11  degrees of freedom
## Residual deviance:  20.204  on  5  degrees of freedom
## AIC: 103.14
##
## Number of Fisher Scoring iterations: 4
```