R 軟體資料分析應用：相對風險、勝算比與邏輯斯迴歸分析

林怡諄 副統計分析師

本期 eNews 與各位討論如何使用 R 進行相對風險、勝算比與邏輯斯迴歸分析。相對風險(Relative Risk)與勝算比(Odds Ratio)是流行病學與醫療領域之中經常使用的風險測量方式，透過列聯表的方式，通常計算暴露於某種情況下，罹患特定疾病的風險；而透過邏輯斯迴歸分析，我們能夠估計風險因子與特定疾病間的風險關係。以下我們就逐一進行介紹並說明 R 程式步驟。

一、相對風險(Relative Risk)與勝算比(Odds Ratio)

(一) 相對風險(Relative Risk)

相對風險用於前瞻性研究(Prospective Study)，觀察目前暴露某種影響因子的情況下，追蹤未來罹患特定疾病的風險。我們可透過簡單的 2×2 列聯表來呈現上述風險情況，並計算暴露於(未暴露於)風險因子的疾病發生率(Incident Rate)，進一步推算其相對風險。

舉例而言，我們現在欲討論美國 4 個城市居民暴露於高血壓(HYPERTENSION)罹患心血管疾病(CVD)的風險，下表為示範資料格式，資料變數包含居民編碼(ID)、性別(GENDER)、年齡分組(AGE)、種族(RACE)、居住城市(CITY)、心血管疾病(CVD)、高血壓(HYPERTENSION)。

ID	GENDER	AGE	RACE	CITY	CVD	HYPERTENSION
1	male	41-50	White	Boston	no	no
2	female	41-50	Black	Detroit	no	no
3	male	41-50	Black	Boston	no	no
4	male	<=40	Black	Newyork	no	no
5	female	41-50	Black	Newyork	no	yes
6	male	41-50	Black	Detroit	no	no
7	male	>=60	White	LA	yes	yes

8	female	41-50	Black	Boston	no	no
9	female	41-50	Black	Boston	no	no
10	female	51-60	Black	Newyork	no	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮
960	female	51-60	Black	Newyork	no	no
961	male	<=40	Black	Boston	no	no
962	male	41-50	Black	Detroit	no	no
963	male	51-60	Black	Newyork	no	no
964	male	41-50	Black	Detroit	no	no
965	male	41-50	Black	Newyork	no	no
966	female	<=40	Black	Newyork	yes	yes

首先，我們先編制2×2 列聯表來呈現高血壓與心血管疾病的關係，透過統計上表找出「是否高血壓」以及「是否罹患 CVD」的次數，編表如下所示：

	CVD	
HYPERTENSION	YES	NO
YES	40 (a)	81 (b)
NO	11 (c)	834 (d)
	51 (N1)	915 (N2)

再者，暴露於高血壓的 CVD 發生率，為$\frac{a}{a+b} = \frac{40}{40+81} = 33.06\%$，而未暴露於高血壓的 CVD 發生率為$\frac{c}{c+d} = \frac{11}{11+834} = 1.30\%$，進一步計算相對風險(Relative Risk)為$\frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{33.06\%}{1.30\%} = 25.39$。

(二) 勝算比(Odds Ratio)

勝算比可用於前瞻性研究與回溯性研究，勝算比的定義是指兩個機率相除所得到的比值，在此，沿用上面的範例資料求算勝算比，罹患 CVD 組有暴露在高血壓(HYPERTENSION)的機率為 a/c，而未罹患 CVD 組有暴露在高血壓(HYPERTENSION)的機率為 b/d，在此，我們討論「"罹患 CVD 組"相對於"未罹患 CVD 組"，暴露於高血壓的機率比值」，即為勝算比。

因此，上述範例之勝算比(OR)，如下所示：

$$\mathrm{OR} = \frac{[(a/N1)/(c/N1)]}{[(b/N2)/(d/N2)]} = \frac{a/c}{b/d} = \frac{ad}{bc} = \frac{40 \times 834}{81 \times 11} = 37.44$$

此一勝算比的解釋與相對風險不同，在此，勝算比是指「罹患 CVD 組」有高血壓的勝算比是「未罹患 CVD 組」的 37.44 倍，是機率比值。我們可以從勝算比得知高血壓與心血管疾病具有高度相關性，換言之，暴露於高血壓者較易罹患心血管疾病，但是勝算比的數值並非多少倍的風險或是機率概念。

(三) 相對風險與勝算比之 R 程式語法

1. 完整程式列表

```
# 設定資料夾路徑與讀入資料 #
setwd("E:\\DEMO")
demo_data <- read.csv(file = "E:\\DEMO\\DEMO_CVD_DATA.csv", header=T)
attach(demo_data)

# 建立基本列聯表 (高血壓與心血管疾病) #
TAB1 <- table(HYPERTENSION,CVD)

# 建立完整列聯表 (高血壓與心血管疾病) #
d <- TAB1[1,1]
c <- TAB1[1,2]
b <- TAB1[2,1]
a <- TAB1[2,2]
EXP <- c("EXPOSES", "NON-EXPOSED")
outc <- c("DISEASE", "NON-DISEASE")
TAB2 <- matrix(c(a, b, c, d),2,2,byrow=TRUE)
dimnames(TAB2) <- list("HYPERTENSION" = EXP, "CVD" = outc)

# 計算基本的 RR 與 OR #

Relative_Risk <- (a/(a+b))/(c/(c+d))
Odds_Risk  <- (a/b)/(c/d)


# 使用epiR PACKAGE，計算 RR 與 OR ，並且計算95%信賴區間 #
install.packages('epiR')
library('epiR')
epi.2by2(TAB2, method = "cohort.count", conf.level = 0.95)
```

2. 程式與結果說明

[第一部分] 建立基本列聯表

<語法>

```
# 建立基本列聯表 (高血壓與心血管疾病) #
TAB1 <- table(HYPERTENSION,CVD)
```

使用 table 指令，將資料檔的 HYPERTENSION 與 CVD 等兩個變數資料，建立一個基本的列聯表。

<結果>

```
> TAB1
             CVD
HYPERTENSION  no yes
         no  834  11
         yes  81  40
```

[第二部分] 建立完整列聯表

<語法>

```
# 建立完整列聯表 (高血壓與心血管疾病) #
d <- TAB1[1,1]
c <- TAB1[1,2]
b <- TAB1[2,1]
a <- TAB1[2,2]
EXP <- c("EXPOSES", "NON-EXPOSED")
outc <- c("DISEASE", "NON-DISEASE")
TAB2 <- matrix(c(a, b, c, d),2,2,byrow=TRUE)
dimnames(TAB2) <- list("HYPERTENSION" = EXP, "CVD" = outc)
```

重新建構列聯表，將「HYPERTENSION = yes」與「CVD = yes」放在表格首欄與首列。首先，擷取 TAB1 表格資訊，分別建立 a,b,c,d 四個數值變數，而 EXP 變數為列名稱為(EXPOSES, NON-EXPOSED)，OUTC 為欄名稱為(DISEASE, NON-DISEASE)，並建立 TAB2 的 2 by 2 矩陣資料，再透過 dimnames 將正確欄名稱與列名稱放入列聯表。

<結果>

```
> TAB2
             CVD
HYPERTENSION  DISEASE NON-DISEASE
  EXPOSES           40          81
  NON-EXPOSED       11         834
```

[第三部分] 計算基本的 RR 與 OR

<語法>

```
# 計算基本的 RR 與 OR #

Relative_Risk <- (a/(a+b))/(c/(c+d))
Odds_Risk  <- (a/b)/(c/d)
```

根據相對風險與勝算比公式，可以計算出 Relative_Risk 變數，為相對風險值；而 Odds_Risk 變數，為勝算比值。

<結果>

```
> Relative_Risk
[1] 25.39444
> Odds_Risk
[1] 37.44108
```

[第四部分] 使用 epiR PACKAGE，計算 RR 與 OR ，並計算 95%信賴區間

<語法>

```
# 使用epiR PACKAGE，計算 RR 與 OR ，並且計算95%信賴區間 #
install.packages('epiR')
library('epiR')
epi.2by2(TAB2, method = "cohort.count", conf.level = 0.95)
```

我們使用 R PACKAGE--「epiR」，建構高血壓與新血管疾病的列聯表，並且求算 RR 與 OR，與其信賴區間。

首先，使用 install.package 指令，安裝「epiR」package，再使用 library 指令呼叫「epiR」package。

再者，我們使用 epi.2by2 指令，將 TAB2 所內含的高血壓與新血管疾病之次數資料，建構列聯表與其 RR 以及 OR 數值。其中可透過 method 調整列聯表格式，epi.2by2 指令提供四種方式「cohort.count」、「case.control」、「cross.sectional」、「outcome = as.columns」，這裡為前瞻式研究，故選擇「cohort.count」。此外，也可以透過 conf.level 的設定來調整信賴區間的顯著水準。

<結果>

```
> epi.2by2(TAB2, method = "cohort.count", conf.level = 0.95)
             Outcome +    Outcome -      Total        Inc risk *        Odds
Exposed +           40           81        121             33.06      0.4938
Exposed -           11          834        845              1.30      0.0132
Total               51          915        966              5.28      0.0557

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 25.39 (13.40, 48.14)
Odds ratio                                     37.44 (18.50, 75.79)
Attrib risk *                                  31.76 (23.34, 40.17)
Attrib risk in population *                    3.98 (2.37, 5.58)
Attrib fraction in exposed (%)                 96.06 (92.54, 97.92)
Attrib fraction in population (%)              75.34 (58.65, 85.30)
-------------------------------------------------------------------
 Test that odds ratio = 1: chi2(1) = 213.443 Pr>chi2 = < 0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

由以上程式分析結果，相對風險為第一個紅框所示，在結果表標註為 Inc risk ratio 即為相對風險值，與上述直接計算值相同，RR 為 25.39，這裡提供了信賴區間，為(13.40, 48.14)。而勝算比為第二個紅框所示，標註為 Odds ratio，亦與直接計算值相同，為 37.44，其信賴區間為(18.50, 75.79)。

二、邏輯斯迴歸分析 (Logistic Regression)

若是我們所關心的解釋變數為二元類別變數，延續上面範例，「罹患心血管疾病與否」這個變數即為二元類別變數，而年齡、性別、是否高血壓、種族、居住城市等其他變數，則為可能的風險影響因子。在流行病學與醫學領域的模型分析方法，我們常用邏輯斯迴歸來討論心血管疾病與其他風險因子的關係。

為何不使用一般線性模型來估計心血管疾病與其他風險因子的關係，主要是因為線性模型在估計解釋變數為二元類別情況(Y= 0/1)時，估計完畢後進行預測時，在解釋變數的某些特定數值之下，可能會使得解釋變數的平均數不在(0,1)區間。而邏輯斯迴歸可以避免此一情況產生，使得預測值介於 0 與 1 之間，使得其估計結果較為準確。

假設 Y 為二元類別變數，代表罹患心血管疾病與否，而其他風險變數設為 X，如年齡、性別、高血壓...等，則罹患心血管疾病的條件機率為

$$\pi_{CVD} = \mathrm{P(Y=1|X)} = \frac{e^{f(x)}}{1+e^{f(x)}}\ ，其中\ f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

而沒有罹患心血管疾病的條件機率為

$$\pi_{nonCVD} = 1 - \mathrm{P(Y=1|X)} = \frac{1}{1+e^{f(x)}}$$

根據勝算比的定義為兩個機率的比值，如下所示

$$\frac{\pi_{CVD}}{\pi_{nonCVD}} = \frac{\frac{e^{f(x)}}{1+e^{f(x)}}}{\frac{1}{1+e^{f(x)}}} = e^{f(x)} = \mathrm{e}^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}$$

將上式取對數後，可得以下方程式

$$\log\left(\frac{\pi_{CVD}}{\pi_{nonCVD}}\right) = \log(\frac{\mathrm{P(Y=1|X)}}{1-\mathrm{P(Y=1|X)}}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

我們可藉由最大概似估計法進行係數估計，得到以上的係數值(β_0, β_1,…, β_k)。然而，以上為係數估計值，而非我們所常見的勝算比，因此，將係數估計值取對數後，即為勝算比。

(三) 邏輯斯迴歸之 R 程式語法

1. 完整程式列表

```
# 邏輯斯模型之估計#
logitic <- glm(CVD ~ HYPERTENSION+AGE+RACE+CITY, data = demo_data, family = "binomial")

# 邏輯斯模型之估計結果呈現#
summary(logitic)

# 邏輯斯模型之勝算比與信賴區間#
OR <- exp(cbind(OR = coef(logitic), confint(logitic)))
```

2. 程式與結果說明

[第一部分] 邏輯斯模型之估計

<語法>

```
# 邏輯斯模型之估計#
logitic <- glm(CVD ~ HYPERTENSION+AGE+RACE+CITY, data = demo_data, family = "binomial")

# 邏輯斯模型之估計結果呈現#
summary(logitic)
```

首先，我們使用 glm 指令來進行 logistic regression 估計，解釋變數為 CVD，被解釋變數為 HYPERTENSION+AGE+RACE+CITY，並且設定分配為 binomial (family="binomial")。

再者，使用 summary 指令，將估計結果彙整顯示。

<結果>

```
glm(formula = CVD ~ HYPERTENSION + AGE + RACE + CITY, family = "binomial",
    data = demo_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.56508  -0.17406  -0.12678  -0.09494   3.10733

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.3169     0.7051  -7.541 4.66e-14 ***
HYPERTENSIONyes   3.5443     0.3922   9.038  < 2e-16 ***
AGE>=60           0.1353     0.6096   0.222    0.824
AGE41-50         -0.7205     0.6511  -1.107    0.268
AGE51-60          0.4972     0.5901   0.843    0.399
RACEWhite         0.1404     0.4203   0.334    0.738
CITYDetroit       0.1325     0.7074   0.187    0.851
CITYLA            2.1522     0.5467   3.937 8.25e-05 ***
CITYNewyork       0.6374     0.5657   1.127    0.260
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 399.28  on 965  degrees of freedom
Residual deviance: 236.14  on 957  degrees of freedom
AIC: 254.14

Number of Fisher Scoring iterations: 7
```

紅框內為估計係數的結果，由於 HYPERTENSION 與 RACE 為二元變數，在此，分別取 HYPERTENSION = no 與 RACE =Black 作為參考組，故結果只有顯示 HYPERTENSION = yes (HYPERTENSIONyes) 與 RACE =White (RACE White)，係數分別為 3.5443 以及 0.1404，觀看 P 值(Pr(<|Z|))，HYPERTENSIONyes 的 P 值為 2e-16 < 0.05，其係數顯著。

而 AGE 與 CITY 亦為類別變數，且均為 4 組，在此，R 程式選擇 AGE<=40 作為參考組，CITY=Boston 為參考組，分別對應其餘 3 組。觀看結果，LA 相對於 Boston 為 CVD 的罹患風險顯著較高。

[第二部分] 邏輯斯模型之勝算比與信賴區間

<語法>

```
# 邏輯斯模型之勝算比與信賴區間#
OR <- exp(cbind(OR = coef(logitic), confint(logitic)))
```

透過 coef 與 confint 指令獲得邏輯斯迴歸係數與信賴區間，並且使用 exp 指令將迴歸係數與信賴區間取指數(exponential)，再使用 cbind 指令將勝算比與其信賴區間合併顯示。

<結果>

```
> OR
                          OR        2.5 %      97.5 %
(Intercept)       0.004907751  0.001075804  0.01735451
HYPERTENSIONyes  34.616624547 16.616205110 78.17591552
AGE>=60           1.144870726  0.359482585  4.08363280
AGE41-50          0.486528928  0.137144332  1.84317236
AGE51-60          1.644131297  0.542585986  5.71042874
RACEWhite         1.150738047  0.496704412  2.60120703
CITYDetroit       1.141678805  0.263841715  4.53284564
CITYLA            8.603964761  3.121858524 27.26052226
CITYNewyork       1.891603468  0.644791905  6.10987426
```

以上 R 執行結果，OR 為勝算比，2.5%為信賴區間下界，97.5%為信賴區間上界，整體為 95% 信賴區間。紅框為暴露於高血壓相對於未有高血壓者罹患心血管疾病的勝算比(OR)為 34.6166，其信賴區間為(16.6162-78.1759)，表示其勝算比顯著高於 1。此一勝算比有經過其他風險因子調整過，是為考量其他風險因子情況之下，所獲得之勝算比，與之前單純計算高血壓與心血管疾病的列聯表而得的勝算比有所差異。

以上為本期生統 eNEWs 的內容，介紹如何使用 R 計算相對風險與勝算比，並使用邏輯斯迴歸分析，得到考量其他因子的調整勝算比，希望本期 eNEWs 能夠提供大家對於 R 分析操作之參考。