R 軟體資料分析應用：列連表檢定與簡單線性迴歸分析

王博賢 副統計分析師

本期 eNews 將與各位討論使用 R 進行『列聯表檢定方法』，以及簡單線性回歸，本次分析同樣使用 CVD_ALL 這組資料作呈現，檔案位置可在 http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv 下載，資料詳述內容定義可至 http://biostat.tmu.edu.tw/attachment/25_help.docx 文件內觀看。

一、列連表檢定

首先介紹列連表檢定，若我們想觀察兩類別變數之間的關聯性，我們可以先將資料整理成『列聯表 (Contingency Table)』的形態。假設 A 類別變數有r個分組，B 類別變數有c個分組，計算資料中在此兩個變數產生的$r \times c$個類別組合的樣本次數，即可構成$r \times c$列聯表。列聯表檢定方法依據樣本的特性不同，可分為：卡方獨立性(或稱齊一性)檢定、費雪精確檢定、McNemar 檢定。

➢ 卡方獨立性檢定 (Chi-Squared Test)

當我們想評估資料中兩類別變數的關聯性，且資料樣本數較大時，即可使用『卡方獨立性檢定』。此方法的概念在比較列聯表中觀察次數和期望次數是否有差異，若兩變數獨立時，觀察次數和期望個數應很接近。以範例資料檔為例，我們想知道罹患心血管疾病與抽菸量是否存在相關性。

首先我們用 ?chisq.test，觀看一下 help 檔

【基本語法】

```
chisq.test(x, y = NULL, correct = TRUE,…)
```

【參數說明】

1. x : 一個變數或矩陣
2. y : 一個變數;x 為矩陣時忽略
3. correct : 是否要連續性校正

了解 chisq.test 如何使用後，我們就可以開始分析，程式碼如下:

```
#讀取檔案
cvd_all <- read.csv(
file = 'http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv '
)
#chi test
#排除抽菸量=0 的人
cvd_goal <-  cvd_all[cvd_all$抽菸量!=0,]
#利用 table 建立列連表
dt <- table(cvd_goal$心血管疾病,cvd_goal$抽菸量)
dt
#chisq test
chisq.test(dt, correct=FALSE)
#直接放兩個變數
c.t <- chisq.test(cvd_goal$心血管疾病,cvd_goal$抽菸量,
correct=FALSE)
#例用 summary 觀察分析結果包含甚麼東西
summary(c.t)
#觀察值
c.t$observed
#獨立時的期望值
c.t$expected
#殘差
c.t$residuals
```

output:

```
> #讀取檔案
> cvd_all <- read.csv(file = 'http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv '
> #chi test
> #排除抽菸量=0 的人
> cvd_goal <-  cvd_all[cvd_all$抽菸量!=0,]
> #利用 table 建立列連表
> dt <- table(cvd_goal$心血管疾病,cvd_goal$抽菸量)
> dt

      1     2     3
  0 13021  1420   144
  1  1383   176    24
> #chisq test
> chisq.test(dt, correct=FALSE)

        Pearson's Chi-squared test

data:  dt
X-squared = 7.1914, df = 2, p-value = 0.02744

> #直接放兩個變數
> c.t <- chisq.test(cvd_goal$心血管疾病,cvd_goal$抽菸量, correct=FALSE)
> #例用summary 觀察分析結果包含甚麼東西
> summary(c.t)
          Length Class  Mode
statistic 1      -none- numeric
parameter 1      -none- numeric
p.value   1      -none- numeric
method    1      -none- character
data.name 1      -none- character
observed  6      table  numeric
expected  6      -none- numeric
residuals 6      table  numeric
stdres    6      table  numeric
> #觀察值
> c.t$observed
                  cvd_goal$抽菸量
cvd_goal$心血管疾病     1     2     3
                  0 13021  1420   144
                  1  1383   176    24
> #獨立時的期望值
> c.t$expected
                  cvd_goal$抽菸量
cvd_goal$心血管疾病         1         2         3
                  0 12993.712 1439.7365 151.55121
                  1  1410.288  156.2635  16.44879
> #殘差
> c.t$residuals
                  cvd_goal$抽菸量
cvd_goal$心血管疾病          1          2          3
                  0  0.2393871 -0.5201505 -0.6133904
                  1 -0.7266303  1.5788532  1.8618716
```

【分析結果】

本分析之虛無假設為兩變數之間無關聯，而 p-值 0.02744 表顯著，拒絕虛無假設，我們可推論資料中是否罹患心血管疾病與抽菸量的高低分組有關。在分析結果中的殘差我們還可以觀察到菸抽菸量越高的分組（1：每日一包、2：每日兩包、3：每日三包以上），殘差越大，這也表示抽菸量越高者有心血管疾病的也越多，根據這個現象，研究者可以嘗試再做進一步的分析。

- **費雪精確檢定 (Fisher's exact test)**

當資料樣本數較小（以樣本筆數<30 為區分標準）時，卡方獨立性檢定的 p 值較不可靠，此時我們可改用『費雪精確檢定』來檢定兩類別變數的關聯性。費雪精確檢定是透過”超幾何分配”的公式來檢定兩變數的相關性，比起卡方獨立性檢定較精確，但是樣本數很大時會耗費較久的運算時間。比照前面的例子，我們可以嘗試用費雪精確檢定來檢定是否罹患心血管疾病與菸草消費量分組是否存在關聯性，雖然此範例的樣本數夠大，我們仍可大略比較兩方法的差異。

首先我們用 ？ fisher.test，觀看一下 help 檔

【基本語法】

fisher.test(x, y = NULL,alternative = "two.sided", conf.level = 0.95,…)

【參數說明】

1. x ： 一個變數或矩陣
2. y ： 一個變數;x 為矩陣時忽略
3. alternative ： 單尾，或雙尾檢定
4. conf.level ： 信賴區間範圍(只有在 2*2 列連表才有)

了解 fisher.test 如何使用後，我們就可以開始分析，程式碼如下:

```
dt
fisher.test(dt)
fisher.test(cvd_goal$心血管疾病,cvd_goal$抽菸量)
```

output:

```
> dt

        1     2     3
  0 13021  1420   144
  1  1383   176    24
> fisher.test(dt)

        Fisher's Exact Test for Count Data

data:  dt
p-value = 0.02829
alternative hypothesis: two.sided

> fisher.test(cvd_goal$心血管疾病,cvd_goal$抽菸量)

        Fisher's Exact Test for Count Data

data:  cvd_goal$心血管疾病 and cvd_goal$抽菸量
p-value = 0.02829
alternative hypothesis: two.sided
```

【分析結果】

本分析之虛無假設為兩變數之間無關聯，而 p-值 0.02829*表顯著，拒絕虛無假設，我們可推論資料中是否罹患心血管疾病與菸草消費量的高低分組有關。此分析結果與前面卡方獨立性檢定的趨勢相同，我們可知在大樣本的情況下，兩方法可得到相同的結論。

➢ McNemar 檢定 (McNemar's test)

當我們想比較類別為兩類的配對(matched pairs)資料，我們可以將資料轉換為成對資料的列聯表，並用『McNemar 檢定』進行分析。由於範例資料並非配對資料，在這邊我們改用生統教科書中的例子來說明：某一臨床試驗欲比較 A 和 B 兩種乳癌化療藥物的療效，納入了 621 對經過年齡配對的乳癌病人（共 1242 人），分別給予 A 藥和 B 藥的治療，而後觀察這些病人五年的存活狀況，觀察的結果整理成下表：有 90 對的病人無論進行 A 治療或 B 治療五年內皆死亡，而有 510 對的病人五年內皆存活；有 16 對的病人進行 A 治療者在五年內存活，但進行 B 治療者在五年內死亡；另有 5 對的病人進行 B 治療者在五年內存活，但進行 A 治療者在五年內死亡。

		進行 B 治療的病人		
		是否五年內死亡		Total
進行 A 治療的病人		NO	YES	
是否五年內死亡	NO	510	16	526
	YES	5	90	95
Total		515	106	621

首先我們用 ? mcnemar.test，觀看一下 help 檔

【基本語法】

mcnemar.test(x, y = NULL, correct = TRUE)

【參數說明】

1. x : 一個變數或矩陣
2. y : 一個變數;x 為矩陣時忽略
3. correct : 是否要連續性校正

了解 fisher.test 如何使用後，我們就可以開始分析，程式碼如下:

```
dt <- matrix(c(510,5,16,90),2,2,byrow = F)
dt
mcnemar.test(dt)
```

output:

```
> dt <- matrix(c(510,5,16,90),2,2,byrow = F)
> dt
     [,1] [,2]
[1,]  510   16
[2,]    5   90
> mcnemar.test(dt)

        McNemar's Chi-squared test with continuity correction

data:  dt
McNemar's chi-squared = 4.7619, df = 1, p-value = 0.0291
```

【分析結果】

本分析之虛無假設為兩變數之間無關聯，而 p-值 0.0291*表顯著，拒絕虛無假設，我們可推論五年存活狀況與 A、B 治療種類有關。此資料中我們感興趣的為存活狀況不一致的配對，即[1,2] 、[2,1]的 21（16+5）對病人，其中進行 A 治療者在五年內存活、但進行 B 治療者在五年內死亡的 16 對病人占多數，我們可以推論 A 治療的療效比 B 治療好。

二、簡單線性迴歸

在日常生活中許多事物彼此間常常存在著線性關係，如要將變數與變數之間的關係以具體的式子表達，其中一個簡單且常用的方法就是利用簡單線性迴歸模型來分析，兩變項 X 與 Y 關係可表示成$y_i=\beta_0+\beta_1 x_i+\varepsilon_i$，其中 Y、X 分別稱為依變數(dependent variable)與自變數(independent variable)，ε 為隨機誤差項，由此式子模型可以很明確的從截距項 β_0 和係數 β_1 得知自變數改變時對依變數的影響，當自變數增加 1 單位，依變數則增加 β_1單位。

- **迴歸模型系數的估計-最小平方法**

截距項 β_0 和係數 β_1 要如何求得，最簡單的方法就是最小平方法，其精神在於讓迴歸模型的誤差項平方和能最小，即求$\min \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$，可利用微分的方式進而求得估計值$\hat{\beta}_1=\frac{\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}$、，$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$ 。且

MSE(Mean square error) = $\frac{\sum_{i=1}^n (y_i-\bar{y})^2}{n-2}$，而我們想知道 x 是否對 y 有顯著影響時，會檢定$\hat{\beta}_1$是否不等於 0，即虛無假設為H_0: $\beta_1=0$，檢定統計量為 t=$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$，其中$se(\hat{\beta}_1)=MSE/\sum_{i=1}^n (x_i-\bar{x})^2$，例用自由度為 n-2 的 t 檢定作檢定。

首先我們用?lm 看一下基本語法

【基本語法】
lm(formula, data,…)
【參數說明】
1. formula : 模型的樣式
2. data : 分析的資料集

了解 lm() 如何使用後，假設我們現在想利簡單迴歸分析建立一個模型，是利用年齡來預測收縮壓，則程式碼如下:

```
rm(list=ls())
cvd_all <- read.csv(
  file = 'http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv'
)
fit.1 <- lm(收縮壓 ~ 年齡,data=cvd_all)
fit.1

#利用 summary() 看更詳細的分析結果
summary(fit.1)
```

output:

```
> rm(list=ls())
> cvd_all <- read.csv(
+   file = 'http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv '
+ )
>
>
> fit.1 <- lm(收縮壓 ~ 年齡,data=cvd_all)
> fit.1

Call:
lm(formula = 收縮壓 ~ 年齡, data = cvd_all)

Coefficients:
(Intercept)         年齡
    93.7881       0.6298

>
> #利用 summary() 看更詳細的分析結果
> summary(fit.1)

Call:
lm(formula = 收縮壓 ~ 年齡, data = cvd_all)

Residuals:
    Min      1Q  Median      3Q     Max
-67.916 -13.020  -1.689  10.947 144.421

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.788102   0.263950   355.3   <2e-16 ***
年齡         0.629841   0.005407   116.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.89 on 63249 degrees of freedom
  (1238 observations deleted due to missingness)
Multiple R-squared:  0.1767,    Adjusted R-squared:  0.1767
F-statistic: 1.357e+04 on 1 and 63249 DF,  p-value: < 2.2e-16
```

【分析結果】

從結果來看，假設在顯著水準為 0.05 時，年齡是顯著的，且年齡每增加一歲，收縮壓會增加 0.6298，而截距項為 93.7881，故我們得到的迴歸模型如下

$$\hat{y}_i = 93.7881 + 0.6298x_i$$

有了此預估模型則可以用來預測依變數，例如有一人的年齡為 27 歲，則套入此預估模型可估計此人心臟收縮壓平均測量值為 110.7927。而判斷此模型時否建立的好時，最常被使用的是判定係數(coefficient of determination，R^2)，從定義上來說，R^2 可以表示自變數能解釋多少比例的依變數變異，數值會介於 0~1 之間，愈接近 1 代表此模型愈能解釋依變數的變化，其等式為

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

從分析結果我們可以知道這個模型的$R^2 = 0.176$，表示使用此預估模型自變數對於解釋依變數變異的能力不是很好。

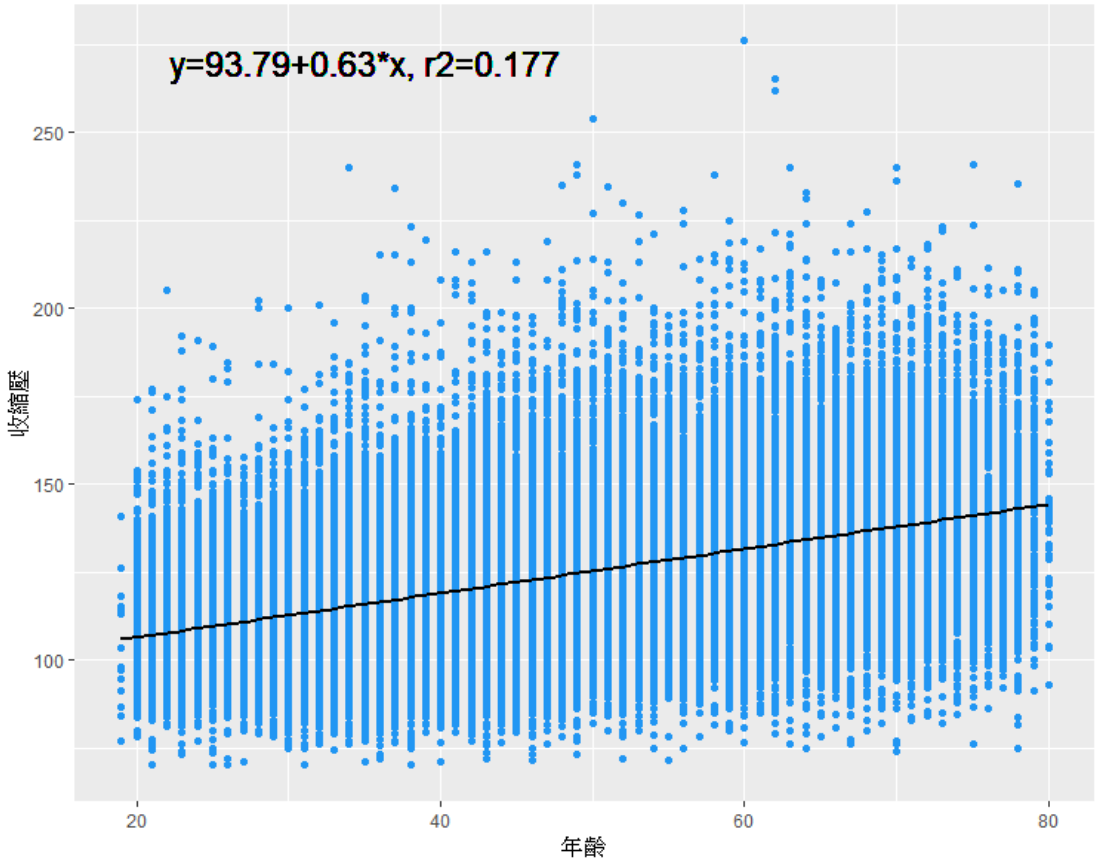
另外我們也可以配合圖形來看兩變數之間的關係。

程式碼如下:

```
#自訂 function 輸出模型形式
lm_eqn <- function(x,y,df){
  m <- lm(get(y) ~ get(x), get(df))
```

```
      #抓取截距項
      intercept <- round(m$coefficients[1],2)
      #抓取 x 係數
      beta <- round(m$coefficients[2],2)
      #抓取 r.squared
      r2 <- round(summary(m)$r.squared,3)
      #合併資訊為一字串
      eq <- paste0("y=",intercept,"+",beta,"*x, r2=",r2)
      return(eq)
    }

    ##利用 ggplot2 畫圖
    library(ggplot2)
    ggplot(cvd_all,aes(x=年齡,y=收縮壓))+geom_point(color="#2196F3")+
      geom_smooth(method = "lm", se=FALSE, color="black")+
      geom_text(x = 35, y = 270, label = lm_eqn(x="年齡",y="收縮壓
",df="cvd_all"),size=6)
```

output:



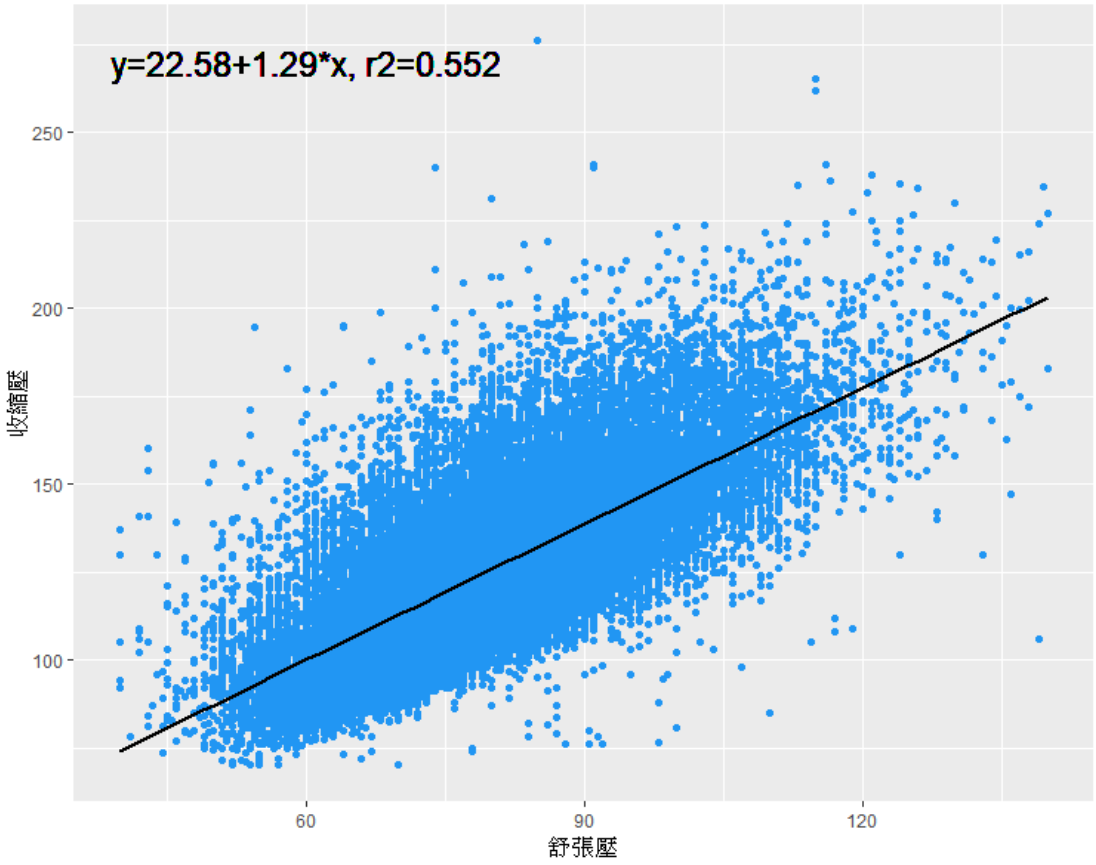
上圖為 “年齡” 與 “收縮壓” 的散佈圖，藍色點代表各個實際資料點，而黑色線為依照迴歸預估模型 $\hat{y}_i = 93.7881 + 0.6298x_i$ 所得的迴歸線，可看出年齡與收縮有線性關係但並不非常的明顯，且資料分佈的位置並沒有明顯向迴歸線集中，與 R^2 值 0.1767 相符合。

因為在此預估迴歸模型下，自變數“年齡”並不能充分解釋依變數“收縮壓”的變異，且並無非常明顯的線性關係，建議可以換個變數試試，以下我們選擇 “舒張壓” 為自變數且重複與之前同樣的步驟，程式碼如下:

```
fit.1 <- lm(收縮壓 ~ 舒張壓,data=cvd_all)
summary(fit.1)
##利用 ggplot2 畫圖
library(ggplot2)
ggplot(cvd_all,aes(x=舒張壓,y=收縮
壓))+geom_point(color="#2196F3")+
  geom_smooth(method = "lm", se=FALSE, color="black")+
  geom_text(x = 60, y = 270, label = lm_eqn(x="舒張壓",y="收縮壓
",df="cvd_all"),size=6)
```

output:

```
> fit.1 <- lm(收縮壓 ~ 舒張壓,data=cvd_all)
> summary(fit.1)

Call:
lm(formula = 收縮壓 ~ 舒張壓, data = cvd_all)

Residuals:
    Min      1Q  Median      3Q     Max
-95.776  -9.408  -2.022   7.261 143.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.58188    0.36490   61.88   <2e-16 ***
舒張壓       1.28916    0.00462  279.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.9 on 63203 degrees of freedom
  (1284 observations deleted due to missingness)
Multiple R-squared:  0.552,     Adjusted R-squared:  0.552
F-statistic: 7.788e+04 on 1 and 63203 DF,  p-value: < 2.2e-16

> ##利用ggplot2 畫圖
> library(ggplot2)
> ggplot(cvd_all,aes(x=舒張壓,y=收縮壓))+geom_point(color="#2196F3")+
+   geom_smooth(method = "lm", se=FALSE, color="black")+
+   geom_text(x = 60, y = 270, label = lm_eqn(x="舒張壓",y="收縮壓",df="cvd_all"),size=6)
Warning messages:
1: Removed 1284 rows containing non-finite values (stat_smooth).
2: Removed 1284 rows containing missing values (geom_point).
>
```



【分析結果】

從結果來看，假設在顯著水準為 0.05 時，舒張壓是顯著的，且$R^2 = 0.552$，顯然我們利用舒張壓建立的模型，比利用年齡建立的模型來說結果更好，從圖形也可發現兩者有較為明顯的線性關係，且資料也較向迴歸線集中，因此可判斷依變數“收縮壓”與自變數“舒張壓”有更高度的線性相關。

本期生統 eNews 的介紹到此告一段落，此次介紹了如何利用 R 軟體進行列聯表檢定以及簡單線性迴歸，希望本期生統 eNews 能幫助大家更加熟悉 R 中這些方法的操作方式。