前列腺癌資料分析

統計 3C　羅文傑

目錄：

研究目的----------3

研究流程----------3

文獻探討----------4

研究方法----------5

資料分析----------6~11

研究結果----------12.13

結論----------14

資料來源----------15

研究目的：

我們找到了某醫學中心泌尿科對血清前列腺特有抗原(PSA)與一些男性前列腺癌手術治療病患的資料，對其相關性感到興趣，欲了解：

1. **前列腺癌指標**
2. **前列腺重量**
3. **病人年齡**
4. **攝護腺面積(肥大量)**
5. **精囊滲入情形是否出現**
6. **包膜滲透程度**
7. **格里森分數和 PSA 等級是否有相關性。**

接著建立一回歸模型，對 PSA 等級做預測。

最後建立簡明的圖表對 PSA 等級做預測，以便臨床上之應用。

研究流程：

步驟 1-由某醫學中心泌尿科蒐集對於 PSA 和前列腺癌手術治療之資料

步驟 2-尋找個變數對患病率的發生因子

步驟 3-挑選最佳模型(格里森分數及其他相關變數之組合)

步驟 4-分析模型及匯整計算結果

步驟 5－計算變數勝算比

步驟 6－建立模型計算預測結果並對照實際結果

步驟 7－作結論

文獻探討

前列腺又稱做做攝護腺，屬於外分泌腺，是男性才具備的器官。

前列腺的主要功能是分泌和儲存前列腺液，它分泌的前列腺液，含有抗菌因子保護尿道，每天以 0.5-2ml 的量，經前列腺腺管，排到後尿道，隨尿液排出體外。前列腺腺管內存有一定量前列腺液，可與精子混合成精液。

前列腺位於骨盆腔的底部—膀胱下，尿道上，恥骨後，直腸前，是一個實心的腺體，這個腺體中有尿道和射精管穿過，其內有許多蜂巢狀的小腺體單位，最後匯流至前列腺部的尿道。

在之前曾有前列腺癌之相關報導，因惡性腫瘤為國人死大死因之冠，而其中前列腺癌為惡性腫瘤中男性不可忽視之癌症，男性中每 29 人之中就有 1 人罹患前列腺癌，而近年來前列腺癌的患病率及死亡率逐年上升，所以我們找到一組資料，想進一步分析探討這項疾病。

在攝護腺癌的分級方面，最常被使用者為格里森分級系統(Gleason grading system)，代表腫瘤細胞惡性的程度。

格里森分級系統是根據顯微鏡下腫瘤組織，依腺體細胞的排列形態來定出分級(grade)，其排列最接近正常者為第 2 級，排列最凌亂者為第 10 級。將腫瘤細胞中佔最多數與次多數者的分級相加就是格里森分數(Gleason score)。

而格里森分數也常用來表式治療後的惡化程度，由於此資料探討觀察值皆為嚴重前列腺患者，且已接受手術，欲想了解格里森分數和其變數的相關性。

研究方法

我們這組採用的研究方式為－

多元邏輯斯迴歸分析(Multiple logistic regression analysis)

符合此回歸分析的條件為：

條件 1：結果變數為類別型變數。

條件 2：主要針對二項式(binomial)的結果變數亦可推廣至依變數為多項用。

條件 3：迴歸模式中包含一個結果變數及多個自變數。

說明：

變數區分為結果變數與自變數，利用線性關係模式，了解自變數的改變對結果變數造成之影響；因此，迴歸模型亦可進行資料預測。

變數選取：

我們的變數選取方法為－Stepwise AIC 法

AIC 欄位中數值越小表示模式越佳，從數變數組中挑選出 AIC 植最小的那一組做為我們的最佳模型。

變數選取準則：min AIC(Akaike Information Criterion)

完整模式：Y= X1+ X2 + X3 + X4 + X5 + X6 + X7

準則選取之最佳模式：Y= **最佳模型變數+最佳模型變數**

(從 X1-X7 的變數中挑選出幾個最佳模型的變數相加)

資料分析

變數定義

我們找到的資料裡包括編號共有 9 筆變數，可用來了解血清前列腺特有抗原(PSA)與一些男性前列腺手術治療病患之間的關係。

變數說明如下：

變數名稱	說明
編號	1-97
SPA 等級	血清前列腺特有抗原等級(mg/ml)
癌症指數	前列腺癌指標估計量(cc)
重量	前列腺重量(gm)
年齡	病人年齡(歲)
肥大量	攝護腺面積(平方公分)
精囊滲入	是否出現(1：有，0：沒有)
胞膜滲透	胞膜滲透程度(cm)
格里森分數	使用兩種型態指標總和來表示病理上的嚴重程度(>6 為嚴重)

我們採用的 Y= **格里斯分數**

X1 = PSA **等級**　X2 = **癌症指標**　X3 = **前列腺重量**

X4 = **年齡**　X5 = **肥大量**　X6 = **精囊滲入**　X7 = **孢膜滲透**

變數解釋：

解釋 PSA 等級.精囊滲入.胞膜滲透.格里森分數等一些較為特別的專有名詞

PSA 等級

PSA 是前列腺上皮細胞中所特有的一種組織特異性蛋白，即前列腺細胞裡的蛋白成分。

當前列腺出現癌細胞，這些蛋白便會流入血液裡。

一般用來檢測前列腺癌治療後癌症是否有被抑制的程度。

精囊滲入

滴蟲、淋病癌症病菌經尿道逆行感染進入精囊。

(因 PSA 等級高有可能為精囊炎)

胞膜滲透

藥物有效滲透進入前列腺胞膜的程度。

格里森分數

分數由 2 至 10 分。

數字越小，等級越小，亦即癌症擴散的速度越慢；

反之，數字越大，等級越高，癌腫瘤的生長速度亦較快。

挑選最佳模型

首先，我們要先挑選出最佳的一組模型，我們分配出了五個變數組，並對這五個變數組做出挑選的動作，AIC 越小的那組則越適合做為我們的模型

以下是我們分析出來的結果：

```
Start:  AIC=176.92
格里森分數 ~ PSA + 癌症指標 + 前列腺重量 + 年齡 + 肥大量 + 精囊滲入 +
    胞膜滲透

             Df    AIC
- 肥大量      1 174.93
- 精囊滲入    1 175.13
- 癌症指標    1 175.24
- 前列腺重量  1 175.95
<none>          176.92
- 胞膜滲透    1 177.63
- 年齡        1 179.43
- PSA         1 182.72

Step:  AIC=174.93
格里森分數 ~ PSA + 癌症指標 + 前列腺重量 + 年齡 + 精囊滲入 +
    胞膜滲透

             Df    AIC
- 精囊滲入    1 173.16
- 癌症指標    1 173.25
- 前列腺重量  1 174.03
<none>          174.93
- 胞膜滲透    1 175.67
- 年齡        1 178.13
- PSA         1 180.98

Step:  AIC=173.15
格里森分數 ~ PSA + 癌症指標 + 前列腺重量 + 年齡 + 胞膜滲透

             Df    AIC
- 癌症指標    1 171.50
- 前列腺重量  1 172.21
<none>          173.16
- 胞膜滲透    1 173.76
- 年齡        1 176.22
- PSA         1 179.61
```

```
Step:  AIC=171.5
格里森分數 ~ PSA + 前列腺重量 + 年齡 + 胞膜滲透

             Df    AIC
- 前列腺重量  1 170.62
<none>          171.50
- 年齡        1 174.87
- 胞膜滲透    1 175.55
- PSA         1 185.84

Step:  AIC=170.62
格里森分數 ~ PSA + 年齡 + 胞膜滲透

           Df    AIC
<none>        170.62
- 年齡      1 173.29
- 胞膜滲透  1 174.83
- PSA       1 184.42
Call:
polr(formula = 格里森分數 ~ PSA + 年齡 + 胞膜滲透, data = data,
    Hess = T, method = "logistic")
```

選出 AIC 值最小的做為最是最佳模型

(PSA+前列腺重量+年齡+包膜滲透)這組的模型之 AIC 最小，固為最佳模型

分析模型

對我們挑選出的最是最佳模型進行分析

```
polr(formula = 格里森分數 ~ PSA + 年齡 + 胞膜滲透, data = data,
    Hess = T, method = "logistic")

Coefficients:
       PSA       年齡   胞膜滲透
0.05155470 0.06563796 0.18476756

Intercepts:
     6|7      7|8
4.516666 7.290247

Residual Deviance: 160.6189
AIC: 170.6189
```

以下為 Zeta 元素傳回截距項估計值以及

Coef 元素傳回解是變數前的參數估計值

```
> result.polr$zeta
     6|7      7|8
4.516666 7.290247
>
> result.polr$coef
       PSA       年齡   胞膜滲透
0.05155470 0.06563796 0.18476756
```

最佳模型的分析結果：

因我們的 Y(格里森分數)分為 6.7.8 三種等級，上圖 result.polr$zeta 下面的數字 4.516666 指的是第 6 與第 7 等級之間的截距，而 7.290247 指的是第 7 與第 8 等機之間的截距。

而下面 result.polr$coef 下面的三個小數分別為 PSA.年齡.包膜滲透的 B 值

模型的設計矩陣

如下圖：

	(Intercept)	PSA	年齡	胞膜滲透
1	1	0.651	50	0.0000
2	1	0.852	58	0.0000
3	1	0.852	74	0.0000
4	1	0.852	58	0.0000
5	1	1.448	62	0.0000
6	1	2.160	50	0.0000
7	1	2.160	64	0.0000
8	1	2.340	58	0.0000
9	1	2.858	47	0.0000
10	1	2.858	63	0.0000
11	1	3.561	65	0.0000
12	1	3.561	63	0.0000
13	1	3.561	63	0.5488
14	1	3.857	67	0.0000
15	1	4.055	57	0.6505
16	1	4.263	66	0.0000
17	1	4.349	70	0.5488
18	1	4.437	66	1.4477

彙整統計計算結果：

算出模型中三個變數之 95%信賴區間以及 P-VALUE 來檢驗顯著與否。

計算結果如下表：

	Ordinal OR	lower95ci	upper95ci	P value
PSA	1.053	1.022	1.091	0.00123
年齡	1.068	1.006	1.138	0.01923
胞膜滲透	1.203	1.039	1.421	0.01051

三個解釋變數前面的回歸係數之 95%信賴區間皆不包含 0，

T 檢定的 P-VALUE 也都小於 α=0.05，因此三項解釋變數皆為顯著

正確預測比例：

我們計算估計模型的混淆矩陣和正確預測比例

計算結果如下表：

```
> tab=table(Real=data$格里森分數,Prediction=Ypred)
> tab
    Prediction
Real  6  7  8
   6 17 14  2
   7  8 31  4
   8  2  9 10
>
> cat("正確預測比例 =",100*sum(diag(tab))/sum(tab), "%\n")
正確預測比例 = 59.79381 %
>
```

由表中得知，最後的正確預測比例為 59.79381%

研究結果

勝算比：

計算出三個變數之勝算比，如下表

```
> (zeta=exp(result.polr$zeta))
      6|7        7|8
  91.52988 1465.93310
>
> (beta=result.polr$coef)
      PSA       年齡    胞膜滲透
0.05155470 0.06563796 0.18476756
> exp(beta)
    PSA     年齡 胞膜滲透
1.052907 1.067840 1.202939
>
```

看到上表中最下面 exp(beta)那個欄位，三變數中，年齡的值為 1.067840，包膜滲透的值為 1.202939，PSA 的值為 1.052907。

意即若病患 A 與病患 B 其他條件完全相同，只有年齡差異的話，A 比 B 大 1 歲之勝算比為 exp(beta[“年齡”]*1)=1.06784。

而若病患 A 與病患 B 其他條件完全相同，只有包膜滲透程度有差異的話，A 比 B 多 1 單位之勝算比為 exp(beta[“胞膜滲透”]*1)=1.202939

而要是病患 A 與病患 B 其他條件完全相同，只有 PSA 程度有差異，A 比 B 多 10 單位之勝算比為 exp(beta[“PSA”]*10)=1.674554。

原始資料與預測模型結果之比較

原始資料第 10 筆與預測模型結果比較，如下表

```
> #原始資料10 與 預測模型結果比較
> (Y10=data[10,])
   id    PSA 癌症指標 前列腺重量 年齡 肥大量 精囊滲入 胞膜滲透 格里森分數
10 10 2.858   1.2461     25.534   63      0         0         0          6
> predict(result.polr,newdata=Y10,type="probs")
         6          7          8
0.55826674 0.39465460 0.04707867
>
>
>
```

原始資料第 20 筆與預測模型結果比較，如下表

```
> #原始資料20 與 預測模型結果比較
> (Y20=data[20,])
   id    PSA 癌症指標 前列腺重量 年齡 肥大量 精囊滲入 胞膜滲透 格里森分數
20 20 4.953   1.1972     46.063   70 5.2593         0         0          7
> predict(result.polr,newdata=Y20,type="probs")
         6          7          8
0.41742764 0.50241711 0.08015525
>
>
```

原始資料第 70 筆與預測模型結果比較，如下表

```
> #原始資料70 與 預測模型結果比較
> (Y70=data[70,])
   id     PSA 癌症指標 前列腺重量 年齡   肥大量 精囊滲入 胞膜滲透
70 70 19.492   3.2871    119.104   72 10.2779         0   0.4493
   格里森分數
70          7
> predict(result.polr,newdata=Y70,type="probs")
         6          7          8
0.2146387 0.5993886 0.1859727
>
```

原始資料第 10 筆.第 20 筆以及第 70 筆預測模型之最大機率結果皆相符

第 10 筆的預測模型最大機率 0.55826674 對到格里森分數 6，與原始資料相同

第 20 筆及第 70 筆的最大機率分別為 0.50241711 及 0.5993886，對到的格里森分數皆為 7，與原始資料相同。

結論

普遍來說，50 歲以上的男性較易患上前列腺癌。 前列腺癌與發生在其他身體部位的癌症略有不同。

若癌症只發生在小部分的前列腺上，癌細胞可以維持一段時間不變，然後才開始擴散、生長。

50 歲或以上的男性---約有三分之一的人能在其前列腺發現若干癌細胞

80 歲或以上的男性---幾乎全部的人也能在其前列腺發現癌細胞。

於前列腺內的癌細胞，通常以很慢的速度生長，尤其是在年長者裡，可能永遠也不會發生問題。

某些特殊案例裡，癌細胞可以生長得很快，並且擴散至身體其他部位，例如**骨髓**。

前列腺癌的症狀：

年紀大的男性常有前列腺腫大的情況，通常是屬於良性的前列腺增生。

良性和惡性腫瘤的徵狀有很多相似之處，可能出現的徵狀如下：

1. 排尿困難

2. 排尿時刺痛

3. 排尿時帶血

4. 排尿次數頻密

5. 排尿不順或微弱

6. 時有沮滯腰背、盆骨及大腿骨位置持續疼痛

資料來源

http://website.tpech.gov.tw/cancer/.htm（台北市立聯合醫院）

http://www.cancerfund.org/prostate/chi/detection/detection.html

（癌症基金會 cancerfoud）

Applied linear statistical models : regression, analysis of variance, and experimental designs /