中國統計學報
第28卷第1期
七十九年三月
93-103 頁

台北榮總病歷數據之邏輯迴歸分析

趙蓮菊　　王迺聖
清華大學統計所　　逢甲大學統計系

摘　要

我們利用邏輯迴歸的模式來分析各種屬性的病人一年內去台北榮總看診的機率多寡，結果顯示影響該機率的屬性包含病人的性別、住距、身份、年齡、是否初診、是否住院。利用我們選擇的模式可以探討各種屬性如何影響病人看診之機率。

關鍵詞：邏輯迴歸。
美國數學會分類索引：62P10。

1. 前言

本文所討論的眞實數據由清華工工所劉志明教授及東海大學翁榮銃先生經清華統計諮詢實驗室所提供。他們欲探討台北榮民總醫院(以下簡稱榮總)之病患看診與其屬性的關係，此即引發了一個有趣的統計問題：什麼樣的病人較常去榮總看病？數據係由翁先生在77年8月自榮總病歷中抽取的5135有效樣本爲依據(我們並未參與抽樣)。本文將說明如何用邏輯迴歸方法來分析此組數據並得到一些初步的結論。

邏輯迴歸早期源自 Berkson (1944, 1953, 1955) 的論文，自從 Truett, Cornfield 及 Kannel (1967) 三人首次應用以探討心臟病之相關危險因子後，邏輯迴歸便在各學科特別是生命相關科學上廣泛應用，一般的介紹與應用可參考 Hosmer 及 Lemeshow (1989) 所著有關邏輯迴歸的專書，最近類似的應用則可參考 Lemeshow, Teres, Avrunin 及 Pastides (1988)。

本文分析可能不盡完備，但因數據本身爲有趣的本土數據，使我們感覺值得介紹此組數據給大家參考及研究。

2. 本文：

根據我們的了解所知此組5135組數據係翁先生自榮總「流動病歷」(即三年內曾來看診的病人病歷)及「非流動病歷」(即三年內未來看診之病人病歷)分層抽出，但因流動病歷中許多被醫生調出、外帶或已抽出至非流動區，以致造成近三年看診的病人病例比例較母體比例爲少。這可能會因樣本的代表性問題產生分析上的一些偏差。但由於我們未參與抽樣，只能就得到的數據分析之。

2.1. 原始數據描述

對每一個抽到的病例，記錄該病人最近三次看診的時間(年，月)以及病人下列的屬性：

$V1$: 是否初診

$V1=0$ 代表初診病人(即只有一次看診時間記錄)，

$V1=1$ 代表複診病人(至少有二次看診記錄)。

$V2$: 身份

$V1=1$ 榮民，

$V2=2$ 榮眷，

$V2=3$ 榮總員工，

$V2=4$ 榮總員工之眷屬，

$V2=5$ 勞保病患，

$V2=6$ 農保病患，

$V2=7$ 公保病患，

$V2=8$ 眷公保病患，

$V2=9$ 一般人民。

$V3$: 性別

$V3=0$ 女性，

$V3=1$ 男性。

$V4$: 年齡

$V5$: 科別差異，即兩次看診之科別是否差異。

$V5{=}0$ 沒差異或初診者，

$V5{=}1$ 有差異。

$V6$: 住距

$V6{=}1.00$ 代表台北市內，其它城市則以1.00加上莒光號所需之時數代表之，如$V6{=}1.73$ 代表中壢，$V6{=}2.28$ 代表新竹等等。

$V7$: 住院次數

$V7{=}i$ 表示病例上記錄該病人共住院i次。

2.2. 將資料分組

由於原始資料的一些屬性過於細分，使得以分組資料來處理時，每組的資料太少。故我們先將某些屬性重組成新的分類變數如下：

ID ：身份

$$\begin{cases} ID=1 & \text{榮民、榮眷、員工、員眷}(V2{=}1\sim4)， \\ & \text{(由於員工、員眷人數很少，此類身份將簡稱榮民眷)} \\ ID=2 & \text{勞保、農保、公保、眷公保}(V2{=}5\sim8)， \\ & \text{(簡稱投保)} \\ ID=3 & \text{一般人民}(V2{=}9)。 \end{cases}$$

AGE ：年齡

$$\begin{cases} AGE=1 & \text{病人年齡小於或等於25 } (V4\le 25), \\ AGE=2 & 25<V4\le 45， \\ AGE=3 & 45<V4。 \end{cases}$$

DS ：住距

$$\begin{cases} DS=0 & \text{非台北附近地區}(V6\ge 2), \\ DS=1 & \text{台北附近地區}(V6<2)。 \end{cases}$$

HS ：住院否

$$\begin{cases} HS=0 & V7=0\text{，沒有住院過，} \\ HS=1 & V7\ge 1\text{，至少住院1次。} \end{cases}$$

另外保持不變，維持原分組的有：

SEX ：性別($V3$)，

FST ：是否初診($V1$)，

DIFF ：科別差異($V5$)。

雖然在原數據中記錄每一個病人最近三次來看診的時間，由於我們的問題是希望"預測"下一年(或兩年，三年，作法應相同)各種屬性病人來門診的機率，很自然的我們希望將最近一年內病人來看診的機率建立模式；因此對每一個病人而言，我們有興趣的反應變數爲是否病人最近一年內，即76年8月至77年8月曾來看診，因此定義反應變數y爲

$$Y=\begin{cases}1 & \text{若病人最近一年內曾來看診，}\\ 0 & \text{若病人最近一年內不曾來看診。}\end{cases}$$

因此我們實際上用到的只是病人最近一次來院的記錄，其它兩次的記錄並未用到。在5135人中，共有551人最近一年內曾來看診，其它4584人一年內未來看診。

如一般初步的分析，我們以Y和每一個屬性做成二個因子的列聯表，對所有選擇的屬性，卡方分配均強烈顯著。

2.3 邏輯迴歸分析

有關邏輯迴歸分析法背景知識請參考Breslow及Day (1987)之書，本文不再重復敘述。基本上我們的模式爲分組二項模式(grouped binomial model)，其二項變數之機率滿足

$$\log\frac{P(Y=1|\underset{\sim}{X}=\underset{\sim}{x})}{P(Y=0|\underset{\sim}{X}=\underset{\sim}{x})}\equiv \text{logit}P(Y=1|\underset{\sim}{X}=\underset{\sim}{x})=\beta_0+\sum_{i=1}^{P}\beta_iX_i\ ,$$

即

$$P(Y=1|\underset{\sim}{X}=\underset{\sim}{x})=e^{\beta_0+\sum_1^P\beta_iX_i}/[1+e^{\beta_0+\sum_1^P\beta_iX_i}]\ 。$$

($\underset{\sim}{X}$代表病人P個選定之解釋變數，可能爲屬性或其函數，或屬性間的交互作用)。

我們經由一連串模式選擇程序，選擇一個適當的模式。發現變數DIFF(科別差異)並非影響變數。事實上，因爲此變數DIFF亦包含初診的病人，所以直觀想起來它也不應是合理的解釋變數。我們選擇的解釋變數爲：

- 性別(SEX)
- 住距(DS)
- 身份*年齡(ID*AGE之交互作用，自然式模式中包含身份，年齡，各別之主因子效果)
- 是否住院*年齡(HS*AGE之交互作用)
- 初診或復診(FST)

由於身份、年齡各有三個分類層，其它性別、住距、是否住院、初診與否各有二個分類層，故總共將病人分成不同屬性的組別共有$3^2 \times 2^4 = 144$組。換言之，邏輯迴歸分析將此144組病人一年內會去榮總看病的機率找出一個模式，利用這個模式我們可以比較那些組的病人(或具那些屬性的病人)較常去榮總看病。事實上如此分組中有兩組因爲沒有病人，故總共比較的共有142組。

爲了討論方便，對身份及年齡，我們進一步再定義兩個虛構變數(一般而言，統計軟體如SAS或GLIM等對分類變數會自動設定虛構變數。)此地我們對身份及年齡分別設定兩個虛構變數如下：

ID：

$$ID1 = \begin{cases} 1 & \text{若ID=榮民眷，} \\ 0 & \text{其它。} \end{cases}$$

$$ID2 = \begin{cases} 1 & \text{若ID=投保，} \\ 0 & \text{其它。} \end{cases}$$

AGE：

$$AGE2 = \begin{cases} 1 & \text{若} AGE \leq 25\text{，} \\ 0 & \text{其它。} \end{cases}$$

$$AGE1 = \begin{cases} 1 & \text{若} 25 < AGE < 45\text{，} \\ 0 & \text{其它。} \end{cases}$$

我們採用最大概似估計法來估計邏輯迴歸中的各項係數，則模式之配合值(fitted value)經SAS軟體處理所得如下：

模式一 ：P(一年內看診各種屬性)

$$
\begin{aligned}
&= -4.624 + 1.495(ID1) + 0.425(ID2) \\
&-0.318(AGE1) - 0.706(AGE2) \\
&-0.717(ID1 * AGE1) - 0.383(ID1 * AGE2) \\
&-0.376(ID2 * AGE1) - 1.166(ID2 * AGE2) \\
&+0.767(DS) + 0.645(HS) + 2.048(FST) - 0.319(SEX) \\
&+0.800(HS * AGE1) + 0.212(HS * AGE2)
\end{aligned}
$$

對上面模式是否適合的問題，我們首先看檢定適合度之可能比檢定：離差檢定 (deviance) (參考 Fienberg (1982), p.40) 。此時卡方值爲 136.17 ，自由度爲 127 (142-15 ，模式中有 15 個參數) ，P- 值爲 0.27 。另外我們再檢查每一組之標準化殘差值 (standardized residual) 而確定模式確爲合適的模式。

由我們得到的模式，我們便可推算各組病人一年內會去看病之機率以及估計各屬性的影響，我們先討論沒出現交互作用的屬性影響：

(1) 住距的影響：
由住距的係數 0.767 我們很容易得到台北附近的人較非台地區的人去榮總看病的機率大，且其對比爲 $exp(0.767) = 2.15$ 倍。(信賴區間亦可容易求出，本文均忽略之) 。

(2) 性別的影響：
男性的看病機率較小，男對女之對爲 $exp(-0.319) = 0.73$ 倍，亦即女對男之對比爲 1.37 倍。而女性爲何機率較高的可能解釋爲投保者及一般人民中產婦人數不少 (請參考後面對年齡的分析) 。

(3) 初診或複診的影響：
複診的病人一年內看病的機率較大，複診對初診的對比爲 $exp(2.048)$ =7.75 倍。

其次我們再來討論出現交互作用之各屬性的影響：

(4) 身份的影響：
由於身份與年齡交互作用出現，故討論身份的影響必須看年齡層而定：

I. 年齡 ≤ 25 ：
我們由模式之迴歸係數很容易得知榮民眷對一般人民的對比爲

$exp(1.495 - 0.383) = 3.04$ ，同理可計算其它對比。可知在此年齡層榮民眷之機率高於一般人民，而一般人民又高於投保者(對比2.08)。我們按機率從大排到小，下面之係數則表示對比值：

$$\text{榮民眷} \underset{3.04}{>} \text{一般人民} \underset{2.08}{>} \text{投保者。}$$

II. $25 <$ 年齡 ≤ 45 ：

$$\text{榮民眷} \underset{2.08}{>} \text{投保者} \underset{1.05}{\cong} \text{一般人民。}$$

III. 年齡 > 45 ：

$$\text{榮民眷} \underset{2.92}{>} \text{投保者} \underset{1.53}{>} \text{一般人民。}$$

由此可知在各年齡層，榮民眷之機率均爲最高者，而投保者在第二、三層爲次高在第一年齡層爲最低；一般人民在第二、三層均爲最低，在第一年層高於投保者。

(5) 住院與否的影響：

由於住院與否(HS)與年齡層有交互作用，故我在每一個年齡層來討論住院與否的影響：

I. 年齡 ≤ 25 ：

$$\text{住院} \underset{2.36}{>} \text{未住院。}$$

II. $25 <$ 年齡 ≤ 45 ：

$$\text{住院} \underset{4.24}{>} \text{未住院。}$$

III. 年齡 > 45

$$\text{住院} \underset{1.9}{>} \text{未住院。}$$

可知在各年齡層，曾住院者較未住院者來看診的機會均高，而之所以有交互作用，是因爲各年齡層之對比相差不小。

(6) 年齡的影響：

由於年齡與身份、住院與否均有交互作用，故必須分爲六層來討論：

I. 一般人民，未曾住院者：

$$(AGE > 45) \underset{1.37}{>} (25 < AGE \leq 45) \underset{1.47}{>} (AGE \leq 25) \circ$$

表示年齡大於45歲有最大的機率，中年者次之，年輕的人最低。其對比值列於 > 號之下。

II. 一般人民，曾住院者：

$$(25 < AGE \leq 45) \underset{1.62}{>} (AGE > 45) \underset{1.65}{>} (AGE \leq 25) \circ$$

III. 投保者，未曾住院：

$$(AGE > 45) \underset{2.0}{>} (25 < AGE \leq 45) \underset{3.23}{>} (AGE \leq 25) \circ$$

IV. 投保者，曾住院：

$$(25 < AGE \leq 45) \underset{1.1}{\simeq} (AGE > 45) \underset{5.26}{>} (AGE \leq 25) \circ$$

V. 榮民眷，未曾住院者：

$$(AGE > 45) \underset{2.8}{>} (25 < AGE \leq 45) \underset{1.05}{\simeq} (AGE < 25) \circ$$

VI. 榮民眷，曾住院者：

$$(AGE > 45) \underset{1.26}{>} (25 < AGE \leq 45) \underset{1.89}{>} (AGE < 25) \circ$$

由此，在所有情況下，年輕者看病的機會都最少。而對一般人民及投保者，中年者看診機會最大(猜測可能係產婦)，其它各情況，年齡者機會最大。

我們對142組不同屬性病人根據模式一可分別計算其看診機率，將其機率由大到小可排出大致前後順序，表一我們列出前八名者，由於各組人數相差很大，標準差相差也大，故順序無絕對的前後意義，但可看出一般傾向。最有可能的八種人列在下表：

表一 一年內最有可能看診的一羣人

模式一名次	身份	性別	年齡	住址距離	住院否	初診否	模式二名次
1	榮眷	女	>45	台北	住院	複診	1
2	榮眷	女	25-45	台北	住院	複診	3
3	榮民	男	>45	台北	住院	複診	2
4	榮民	男	25-45	台北	住院	複診	6
5	榮眷	女	>45	台北	未住院	複診	7
6	榮眷	女	>45	非台北	住院	複診	4
7	榮眷	女	≤25	台北	住院	複診	5
8	投保	女	25-45	台北	住院	複診	

注意在所列前八名中之前七名均爲榮民眷屬且均爲複診病人，而第八名很顯然最可能爲投保之女性產婦。總結表一之結論得知一年內最可能看診的一群人爲：

(1) 女性而言：包含台北地區住院之複診，榮眷，非台北地區之老年住院複診榮眷及台北地區未住院之老年複診榮眷。

(2) 男性而言：超過25歲台北地區曾住院之複診榮民。

在上述的模式一中，我們並未考慮病人距抽樣時已多久沒來看病的"時間"因素。直觀上我們會認爲一個病人若已很久沒來看病則較一個最近曾來看診的病人有較小的機會在最近一年內會來看診。如果考慮此種時間因素，我們可定義。

$MONTH$ = (77年8月) − (最近一次看診時間)〔化成月單位〕，

即爲此一病人有多月已沒有來看診(注意此地我們仍只用到最近一次看診時間，其它兩次看診時間均未用到。)如果我們加入此種"時間"因素，則模式一中的住院與年齡之交互作用項變成不顯著，而得下列模式二。

模式二：P(一年內看診 | 各種屬性)

$$= -4.356 + 1.258(ID1) + (ID2) - 0.197(AGE1) - 0.644(AGE2)$$
$$-0.401(ID1 * AGE1) - 0.193(ID2 * AGE2)$$
$$-0.032(ID2 * AGE2) - 1.100(ID2 * AGE2)$$
$$+0.812(DS) + 0.934(HS) + 1.945(FST) - 0.292(SEX)$$
$$-0.00915(MONTH)\ 。$$

將此模式二與模式一的係數相比，除了交互項係數有些差別外，其它的差別都非常的少。因此各種屬性影響之對比結果與模式一大致相同。我們僅敘述"時間"的影響。由$MONTH$的係數0.00915(標準誤差0.00258)可知一個月沒來的病人較二個月沒來之病人看診機會較高，其對比爲$exp(0.00915) = 1.0092$；一個月沒來與三年沒來的對比爲1.38；一個月沒來與十年沒來的對比爲2.97，可依此類推其它對比情況。

如果按照模式二，我們將病人分成142組後，每一組中病人的"時間"長短均不同，若按其平均機率來比較，也將最高機率的列出，我們很有興趣的發現模式一中最有可能看診的七種病人與模式二中的二中的最有可能的七種病人完全相同，只是順序成爲1,3,2,6,7,4,5(參考表一中最後一行)。因此基於模式一的結論亦適用於此。

3. 討論

事實上我們覺得，有一個重要的屬性可能在抽樣時，應該記錄的，即就是應記錄病人是否爲慢性病人，因爲這種人可能是最需經常看診的人。同時若能記錄女性中是否爲產婦或每病人看診科別將更有助分析此組數據。

致謝詞：我們感謝翁榮鈁先生同意讓我們使用此組數據。

參考文獻

Berkson, J. (1944). Application of the logistic function to bioassay. Journal of American Statistical Association, 39, 357 –65.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. Journal of Amrerican Statistical Association, 50, 130–62.

Breslow, N. E. and Day, N. E. (1980). Statistical Methods in Cancer Research. International Agency for Research on Cancer Publication, No. 32.

Fienberg, S. E. (1981). The Analysis of Cross-Classified Categorical Data. Second Ed. The MIT Press.

Hosmer, D.W. and Lemeshow, S. (1989). Applied Logisitic Regression. John Wiley & Sons.

Lemeshow, S., Teres, D., Avrnuin, J.S. and Pastides, H. (1988). Predicting the outcome of intensive care unit patients. Journal of American Statistical Association, 83, 348–56.

Truett, J., Cornfield, J. and Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. Journal of Chronic Diseases 20, 511–24.

[民國78年12月11日收稿，1月10日修訂]

Logistic regression analysis for hospital data

Anne Chao

Institute of Statistics
National Tsing–Hwa University

N. S. Wang

Department of Statistics
Feng–Chia University

ABSTRACT

Using logistic regression, we model the conditional probability of seeing a doctor within one year for patients with various covariates in Veteran General Hospital of Taipei. The chosen model indicates that significant covariates include patient's sex, distance of residence, identity, age, whether the patient is first or not, and whether the patient was hospitalized before. The effect (odds ratio) of each covariate is also calcuated.

Key words and phrases. logistic regression.

AMS 1980 subject classifications. 62P10.