Logistic Regression 介紹

黃瓊瑤

Logistic Regression 是迴歸分析的一種，但與一般線性迴歸的依變項(Y)須為連續型變數不同，Logistic Regression 的依變項(Y)是類別變數，若是類別只有兩個，則為二元的邏輯式迴歸(Binary logistic regression)，若是類別超過三個以上則為 Polytomous logistic regression，model 相對複雜許多，本篇僅就 Binary logistic regression 做介紹。在常見的臨床醫學研究上，通常我們二元分類為有得病和沒有得病，有病設為 1，沒有設為 0。為了方便結果的解釋與理解，一般來說我們會將依變項為 0 設為參照組。自變項(X)可為類別變數或連續變數，用來討論對依變項(Y)的關係。通常使用「最大概似函數估計法(Maximum Likelihood Estimation)」對參數做估計。最常應用在流行病學的 Case-Control study，常見的指標是勝算比(Odds Ratio)。

在 SAS 中，Logistic Regression 的語法如下：

```
PROC LOGISTIC <options>;
   BY variables;
   CLASS variable <(options)> <variable <(options)> …> </ options>;
   CODE <options>;
   CONTRAST 'label' effect values<, effect values, …> </ options>;
   EFFECT name=effect-type(variables </ options>);
   EFFECTPLOT <plot-type <(plot-definition-options)>> </ options>;
   ESTIMATE <'label'> estimate-specification </ options>;
   EXACT <'label'> <INTERCEPT> <effects> </ options>;
   EXACTOPTIONS options;
   FREQ variable;
   ID variables;
   LSMEANS <model-effects> </ options>;
   LSMESTIMATE model-effect lsmestimate-specification </ options>;
   <label:> MODEL variable <(variable_options)> = <effects> </ options>;
   <label:> MODEL events/trials = <effects> </ options>;
   NLOPTIONS options;
   ODDSRATIO <'label'> variable </ options>;
   OUTPUT <OUT=SAS-data-set> <keyword=name <keyword=name …>> </ option>;
   ROC <'label'> <specification> </ options>;
   ROCCONTRAST <'label'> <contrast> </ options>;
   SCORE <options>;
   SLICE model-effect </ options>;
   STORE <OUT=>item-store-name </ LABEL='label'>;
```

STRATA *effects </ options>*;
<label:> TEST *equation1 <,equation2, …> </ option>*;
UNITS *<independent1=list1 <independent2=list2 …>> </ option>*;
WEIGHT *variable </ option>*;

其中 PROC LOGISTIC 和 MODEL 語法是在做 Logistic Regression 分析時一定必須要指定的，其他語法都是可以依需求選定。如果需要指定 CLASS 和 EFFECT 語法則必須放在 MODEL 語法之前，如果需要指定 CONTRAST，EXACT 和 ROC 語法則必須在 MODEL 語句之後。主要語法說明如下：

- CLASS：宣告自變項中哪些變項為類別變項。
- MODEL：有二種形式可以指定，第一種最常使用的是依變項是單次試驗且是二元的，第二種則是受試者有多次試驗的資料時，就要指定兩個變項，一個是二元的「事件」依變項，另一個是試驗次數，兩個變項之間要用「/」符號分隔，事件和試驗次數的值都不能是負數。
- STRATA：在資料有分層或是配對且是二元邏輯式迴歸時使用。這部份我們將會在之後的 Conditional Logistic Regression 介紹中說明。

以下藉由 SAS/STAT 15.1 User's Guide 範例中擷取資料做介紹，此資料欲研究不同治療方式及安慰劑對老年神經痛患者的鎮痛效果，依變項(Y)是患者是否疼痛。共收集了 60 名患者的年齡、性別及治療開始前的抱怨時間。首先將收集到的資料建立成一個名為 Neuralgia 的資料集。

```
DATA Neuralgia;
 INPUT Treatment $ Sex $ Age Duration Pain $ @@;
 DATALINES;
P F 68  1  No  B M 74 16 No  P F 67 30 No
P M 66 26 Yes B F 67 28 No  B F 77 16 No
A F 71 12 No  B F 72 50 No  B F 76  9 Yes
A M 71 17 Yes A F 63 27 No  A F 69 18 Yes
B F 66 12 No  A M 62 42 No  P F 64  1 Yes
A F 64 17 No  P M 74  4 No  A F 72 25 No
P M 70  1 Yes B M 66 19 No  B M 59 29 No
A F 64 30 No  A M 70 28 No  A M 69  1 No
B F 78  1 No  P M 83  1 Yes B F 69 42 No
B M 75 30 Yes P M 77 29 Yes P F 79 20 Yes
A M 70 12 No  A F 69 12 No  B F 65 14 No
```

```
B M 70 1 No B M 67 23 No A M 76 25 Yes
P M 78 12 Yes B M 77 1 Yes B F 69 24 No
P M 66 4 Yes P F 65 29 No P M 60 26 Yes
A M 78 15 Yes B M 75 21 Yes A F 67 11 No
P F 72 27 No P F 70 13 Yes A M 75 6 Yes
B F 65 7 No P F 68 27 Yes P M 68 11 Yes
P M 67 17 Yes B M 70 22 No A M 65 15 No
P F 67 1 Yes A M 67 10 No P F 72 11 Yes
A F 74 1 No B M 80 21 Yes A F 69 3 No
;
RUN;

DATA Neuralgia;
SET Neuralgia;
IF pain="No" THEN pain_new=0;
ELSE IF pain="Yes" THEN pain_new=1;
RUN;
```

Neuralgia資料集中(圖一)，Pain是依變項(Y)，YES代表疼痛存在，No則表示沒有疼痛。Treatment是分成三類的類別變項，A和B代表兩種治療，P代表安慰劑治療。Sex代表患者性別。Age為患者開始治療的年齡(以年為單位)。Duration則為開始治療前的抱怨時間(以月為單位)。

圖一 截取前 20 筆資料

	Treatment	Sex	Age	Duration	Pain	pain_new
1	P	F	68	1	No	0
2	B	M	74	16	No	0
3	P	F	67	30	No	0
4	P	M	66	26	Yes	1
5	B	F	67	28	No	0
6	B	F	77	16	No	0
7	A	F	71	12	No	0
8	B	F	72	50	No	0
9	B	F	76	9	Yes	1
10	A	M	71	17	Yes	1
11	A	F	63	27	No	0
12	A	F	69	18	Yes	1
13	B	F	66	12	No	0
14	A	M	62	42	No	0
15	P	F	64	1	Yes	1
16	A	F	64	17	No	0
17	P	M	74	4	No	0
18	A	F	72	25	No	0
19	P	M	70	1	Yes	1
20	B	M	66	19	No	0

在 SAS 軟體中，是以 PROC LOGISTIC 指令來執行 Logistic Regression 分析，並**預設**是對依變項(Y)資料升序**排列後取排序值最小的進行建模**，在上述的範例資料中即是以 Pain=No 來建模(圖二)，也是就將 Pain=Yes 設定為參考組。但一般為了讓結果解釋方便好理解，會將依變項(Y)資料以 0(沒有疼痛)、1(疼痛)表示，且會將 Y=0 設定為參考組。所以我們先將 Pain 資料重新編碼，Pain=Yes 轉換成 pain_new=1，Pain=No 轉換成 pain_new=0。要將 pain_new=0 設定為參考組的方法就是在指令 PROC LOGISTIC 的最後面加上 DESCENDING 讓資料降序排列後以 pain_new=1 建模，就會是我們要的將沒有疼痛設定為參考組(圖三)。

圖二

回應概況

已排序值	Pain	總次數
1	No	35
2	Yes	25

建立模型的機率是 Pain='No'。

圖三

回應概況

已排序值	pain_new	總次數
1	1	25
2	0	35

Probability modeled is pain_new=1.

CLASS 指令則是宣告 Treatment 和 Sex 為類別變項，且將 Treatment 另創建了兩個設計變項來代表三種治療方式，而 Sex 則是創建了一個設計變項來代表(圖四)。/PARAM= *<keyword>*則是用來定義設計變項呈現的方法。SAS 軟體會對變項數據資料升序排列後，取最後順位的當基準組，並將其設計變項的值預設為-1，指令為/PARAM=EFFECT REF=LAST。但為了結果解釋方便，我們通常習慣將基準組設定成 0，所以指令則為/PARAM=REFERENCE 或精簡為 REF(圖五)。若想要自己定義基準組則可在類別變項後面加上 REF= *'level'| keyword*。例如 CLASS Treatment(ref='A') Sex(ref='F');(圖六)。

圖四

類別層級資訊

類別	值	設計變數	
Treatment	A	1	0
	B	0	1
	P	-1	-1
Sex	F	1	
	M	-1	

圖五

類別層級資訊

類別	值	設計變數	
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

圖六

類別層級資訊

類別	值	設計變數	
Treatment	A	0	0
	B	1	0
	P	0	1
Sex	F	0	
	M	1	

接著我們就以下列指令來執行 Logistic Regression 分析。結果如圖七。

```
PROC LOGISTIC DATA= Neuralgia DESCENDING;
CLASS Treatment Sex / PARAM=REF REF=LAST;
MODEL pain_new= Treatment Sex Age Duration;
RUN;
```

圖七

最大概度估計值分析

參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	-15.5744	6.5915	5.5828	0.0181
Treatment	A	1	-3.1817	1.0161	9.8049	0.0017
Treatment	B	1	-3.7085	1.1407	10.5700	0.0011
Sex	F	1	-1.8322	0.7963	5.2946	0.0214
Age		1	0.2621	0.0970	7.2977	0.0069
Duration		1	-0.00586	0.0330	0.0315	0.8591

勝算比估計值

效果	點估計值	95% Wald 信賴界限	
Treatment A 與 P	0.042	0.006	0.304
Treatment B 與 P	0.025	0.003	0.229
Sex F 與 M	0.160	0.034	0.762
Age	1.300	1.075	1.572
Duration	0.994	0.932	1.061

上列結果為在控制年齡、性別及治療開始前的抱怨時間的影響後，使用 A 藥還會造成疼痛的的勝算比僅為 P 安慰劑的 0.042 倍，且有統計上的顯著差異(p=0.0017)。使用 B 藥治療還會造成疼痛的勝算比(Odds Ratio: OR)是 P 安慰劑的 0.025 倍，並有統計上的顯著差異(p=0.0011)。表示無論 A 或 B 治療都比 P 有鎮痛效果。在控制治療方式、年齡及治療開始前的抱怨時間的影響後，男性的勝算是女性的 0.160 倍，且有統計上的顯著差異(p=0.0214)。即男性相較於女性而言，較不易神經痛。而對於連續型變項 Age 的結果解釋為控制治療方式、性別及治療開始前的抱怨時間的影響後，年齡每增加 1 歲疼痛的風險會上升 1.30 倍，且有統計上的顯著差異(p=0.0069)。

在此提供另一可將 pain_new=0 設定為參考組的指令，不用下 DESCENDING 指令，直接在 MODEL pain_new(event='1')= 定義 pain_new 的事件是 1，就等於是以 pain_new=1 建模。雖圖八風險估算結果和圖七一致，一樣也是 pain_new=1 建模，但對 pain_new 資料仍是預設的升序排列，與圖三不同。指令如下：

```
PROC LOGISTIC DATA= Neuralgia;
CLASS Treatment Sex / PARAM=REF REF=LAST;
MODEL pain_new(event='1')= Treatment Sex Age Duration;
RUN;
```

圖八

回應概況

已排序值	pain_new	總次數
1	0	35
2	1	25

Probability modeled is pain_new=1.

類別層級資訊

類別	值	設計變數	
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

最大概度估計值分析

參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	-15.5744	6.5915	5.5828	0.0181
Treatment	A	1	-3.1817	1.0161	9.8049	0.0017
Treatment	B	1	-3.7085	1.1407	10.5700	0.0011
Sex	F	1	-1.8322	0.7963	5.2946	0.0214
Age		1	0.2621	0.0970	7.2977	0.0069
Duration		1	-0.00586	0.0330	0.0315	0.8591

勝算比估計值

效果	點估計值	95% Wald 信賴界限	
Treatment A 與 P	0.042	0.006	0.304
Treatment B 與 P	0.025	0.003	0.229
Sex F 與 M	0.160	0.034	0.762
Age	1.300	1.075	1.572
Duration	0.994	0.932	1.061

Reference

1. SAS/STAT 15.1 User's Guide,The LOGISTIC Procedure.
2. SAS/STAT 15.1 User's Guide,The LOGISTIC Procedure：Example 76.2 Logistic Modeling with Categorical Predictors.
3. Institute for Digital Research & Education, University of California Los Angeles. LOGIT REGRESSION | SAS DATA ANALYSIS EXAMPLES, Available from National Technical Information Service Web site, https://stats.idre.ucla.edu/sas/dae/logit-regression/
4. Institute for Digital Research & Education, University of California Los Angeles. IN PROC LOGISTIC WHY AREN' T THE COEFFICIENTS CONSISTENT WITH THE ODDS RATIOS? | SAS FAQ, Available from National Technical Information Service Web site, https://stats.idre.ucla.edu/sas/faq/in-proc-logistic-why-arent-the-coefficients-consistent-with-the-odds-ratios/