R-web 資料分析應用：邏輯斯迴歸分析

江 奕 副統計分析師

上期向大家介紹了簡單線性迴歸的使用方法，相信大家都了解線性迴歸分析中，反應變數需為連續型變數，解釋變數則接受連續與類別型變數。當反應變數為類別型的資料型態時，則可以使用邏輯斯迴歸來分析手邊的資料。

本期，將繼續使用基隆社區為基礎的整合篩檢計畫(Keelung Community-based Integrated Screen Program, KCIS)的心血管疾病資料作為範例資料檔搭配雲端資料分析暨導引系統】(R-web, http://www.r-web.com.tw) 介紹接下來的分析方法，此資料的變數定義可參考下表。詳細資料介紹請參閱首期生統 eNews。

變數名稱	變數定義	變數型態
性別(Gender)	女性(0)、男性(1)	類別
年齡(Age)	年齡	連續
腰圍(Waist)	公分(cm)	連續
心臟收縮壓(SysBP)	毫米汞柱(mmHg)	連續
心臟舒張壓(DiaBP)	毫米汞柱(mmHg)	連續
空腹葡萄糖(AC)	毫克/分升(mg/dl)	連續
高密度脂蛋白(HDL)	毫克/分升(mg/dl)	連續
三酸甘油酯(TG)	毫克/分升(mg/dl)	連續
嚼檳榔習慣(Betelnut)	無(0)、有(1)	類別
飲酒習慣(Alc_Drink)	無(0)、有(1)	類別
個人心血管疾病史(CVD)	無(0)、有(1)	類別
家族心血管疾病史(FamilyHx)	無(0)、有(1)	類別
抽菸習慣(Tobacco)	無(0)、有(1)	類別
菸草消費量(Tobacco_Consumption)	無(0)、每日一包(1)、每日兩包(2)、每日三包以上(3)	類別

➢ 邏輯斯迴歸分析(Logistic Regression Analysis)

邏輯斯迴歸是用來分析與解釋一個名義尺度的反應變數與一個以上的解釋變數間之關係，基本假設與線性迴歸類似。主要使用反應變數為二元型態之資料，例如死亡或未死亡。目的是為了要找出類別型態的反應變數和解釋變數之間的關係，因此和簡單線性迴歸分析中最大的差別在於反應變數型態的不同。

邏輯斯迴歸在運用上也需符合傳統迴歸分析的一般假設，也就是避免解釋變數之間共線性的問題，以及符合常態分配等的基本假設。因迴歸分析方法中，限制了反應變數為連續型變數，若欲分析的變數非為連續型時則無法使用，此時可選擇邏輯斯迴歸來處理，並同時針對類別型變項計算勝算比(odds ratio)指標來判斷其對於反應變數的影響強度。

➢ 勝算比(Odds Ratio)

首先，要先了解一個衡量風險的指標勝算（Odds），勝算定義為一件事情發生的機率與一件事情沒發生機率的比值。以拋硬幣為例，正面與反面的機率都為 0.5，所以 odds ratio 為 $\frac{0.5}{0.5}=1$。如果抽菸得肺癌發生的機率為 0.7 ，那勝算為 $\frac{0.7}{0.3}=2.33$ 。

此時我們可定義勝算比為暴露組勝算與非暴露組勝算之比值：

$$\text{odds ratio(OR)} = \frac{p_1/（1-p_1）}{p_2/（1-p_2）}$$ 。

OR > 1 表示曝露組(含危險因子)發生定義事件的風險比無曝露組較

大，OR < 1 則相反，OR =1 表示曝露與否和疾病的發生不相關。

針對範例資料，我們參考中央健康保險局，對於心血管疾病（CVD）所提出的危險因子有哪些，其中包含，1.高血壓，2.空腹葡萄糖，3.男性≧45 歲，4.有早發性冠心病家族史，5.女性≧55 歲或停經沒用雌激素療法者，6.有吸菸者。首先可做個簡單的敘述統計分析來了解資料(附錄一、二)。

我們利用菸草消費量，使用資料篩選功能(附錄三)篩選出其中有抽菸(危險因子)的人共 16168 筆作為後續分析用資料，初步使用『卡方獨立性檢定』(參閱第九期生統 eNews)來評估個人心血管疾病史與菸草消費量(1：每日一包、2：每日兩包、3：每日三包以上）兩類別變數的關聯性：

		Tobacco_Consumption			**合計**
		1	2	3	**Total**
CVD	0	13021 80.54 89.28 90.40	1420 8.78 9.74 88.97	144 0.89 0.99 85.71	14585
	1	1383 8.55 87.37 9.60	176 1.09 11.12 11.03	24 0.15 1.52 14.29	1583
合計 Total		14404	1596	168	16168
列聯表內容為觀察值個數 / 百分比 / 列百分比 / 行百分比					
卡方獨立性檢定 虛無假設：兩變數之間無關聯					
卡方檢定統計量			自由度		p-value
7.1914			2		**0.027441**

上表，p-value 小於我們設定的顯著水準 α =0.05，由此推論，是否罹患心

血管疾病與菸草消費量的分組有顯著的相關性，亦可從列聯表觀察到菸草消費量越高，罹患心血管疾病的比例越高。以下將應用此組資料進行邏輯斯迴歸分析。

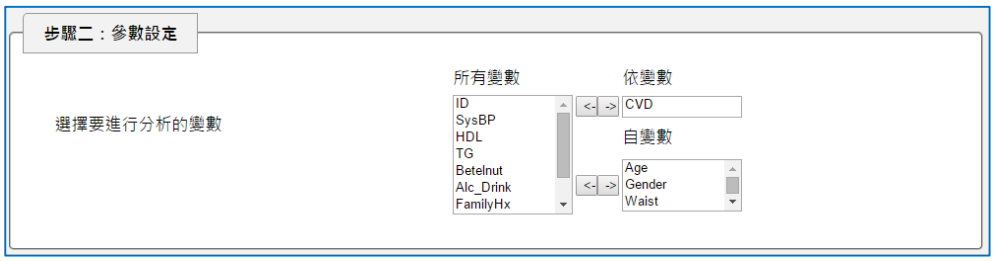
➢ R-web 操作分析步驟

我們將從 16168 筆抽菸的人中，挑選出心血管疾病的危險因子以及部分感興趣的變項來作為邏輯斯迴歸分析中的解釋變數：年齡、性別、腰圍、舒張壓、空腹葡萄糖。

依序點選主選單中【分析方法】→【迴歸模式】→【邏輯斯迴歸分析】

步驟一:資料匯入，從個人資料檔中選取欲分析的資料名稱

步驟二:參數設定，選取依變數以及自變數名稱。此處依變數:CVD，自變數: Age, Gender, Waist, DiaBP, AC。



進階選項中，可設定是否使用交互作用項、顯示樣本敘述統計量、是否使用 AIC 法進行變數選取以及是否顯示分類表等設定，確認無誤儲存後，點選【開始分析】進行分析。

➢ R-web 輸出結果:

◆ 樣本敘述統計量[I]:

數值變數(numerical)

變數名稱 Variable	樣本數 Count	平均數 Mean	中位數 Median	最小值 Minimum	最大值 Maximum	標準差 Std. dev.
Age	14964	46.8197	45	19	80	13.8133
Waist	14964	82.4255	83	40	175	10.4177
DiaBP	14964	79.7716	79	40	139	12.3217
AC	14964	93.5293	87	51	476	29.454

類別變數(categorical)

變數名稱 Variable	變數值 Value	編碼 Coded	個數 Count
CVD	0	0	13617
	1	1	1347
Gender	0	0	2873
	1	1	12091

I：變數訊息皆不包含遺失值

◆ 模式係數估計[I]：

係數 coefficient	估計值 estimation	p 值[II] p-value	參數 95% 信賴區間		估計值的勝算比 exp(coef.)
			下界 lower	上界 upper	
(截距項)	-7.9439	<1e-04 ***	-8.5445	-7.3506	---
Age	0.0666	< 1e-04 ***	0.0621	0.0711	---
Gender(1)	-0.2602	0.0021 **	-0.4243	-0.0929	0.7709
Waist	0.016	< 1e-04 ***	0.0097	0.0223	---
DiaBP	0.0093	2e-04 ***	0.0044	0.0142	---
AC	0.0027	7e-04 ***	0.0011	0.0042	---

I：依變數為 CVD

II：顯著性代碼：‘***’ ：< 0.001, ‘**’ ：< 0.01, ‘*’ ：< 0.05, ‘#’ ：< 0.1

由上方模式係數估計表可以看出，年齡(Age)，性別(Gender)，腰圍(Waist)，舒張壓(DiaBP)，空腹葡萄糖(AC)等，其檢定結果 P 值皆小於α = 0.05，因此推論模型中整體自變數對依變數的解釋能力具有統計上之顯著性。

邏輯斯迴歸模型的估計結果顯示年齡，性別，腰圍，舒張壓，空腹葡萄糖對於心血管疾病具有顯著影響。以年齡為例，表示隨著年齡每增加 1 歲，其患有心血管疾病的風險是未增加前的 exp(0.0666) = 1.07 倍(增加 7%)。而對於性別來說，男性抽菸比女性抽菸患有心血管疾病的勝算率為 exp(-0.2602) = 0.77 倍(減少 23%)，此即意謂：與男性相較之下，抽菸對女性得到心血管疾病的影響要比抽菸對男性的影響來得大。

分析結果中，亦有提供此模型之預測/觀察分類表，如下表所示：敏感度:1344/1347 = 99.8%，精確度:3/1347 = 22%，正確度:13617/14964=91%。

CVD		預測次數		
		0	1	總和
觀察次數	0	13614	3	13617
	1	1344	3	1347
	總和	14958	6	14964
敏感度(sensitivity)： 99.98 % 精確度(specificity)： 0.22 % 正確度(accuracy)： 91 % 偽陽性(false positive)： 8.99 % 偽陰性(false negative)： 50 %				

- 分類表解釋：

類別依變數		預測次數		
		0	1	總和
觀察次數	0	a	b	r_0
	1	c	d	r_1
	總和	c_0	c_1	n

敏感度(sensitivity)：a/r_0
精確度(specificity)：d/r_1
正確度(accuracy)：$(a+d)/n$
偽陽性(false positive)：c/c_0
偽陰性(false negative)：b/c_1

註解：邏輯斯迴歸分析中，類別變數通常會轉換為虛擬變數，若該類別共有 K 類，則需要 K-1 個虛擬變數。此外為了使輸出報表較好解讀，建議於輸入資料前，將類別資料轉換為 0，1 或 0，1，2 等組別，R-web 預設使用最小的數字做為參照組來呈現結果。比如說 0=沒抽菸、1=有抽菸，即表示在報表解讀時，應以沒抽菸的組別作為比較基準。

本期的生統 eNews 就介紹到這裡，這次介紹了如何將邏輯斯迴歸分析方法運用在依變數為類別型態的資料，並搭配前幾期 eNews 所介紹的列聯表檢定方法：卡方獨立性檢定，來檢視變數間的相關性。希望大家能搭配多種檢定方法來看不同的結果，並能了解使用時機與軟體操作。下一期的生統 eNews 將為大家介紹更進階的分析方法—『無母數方法』，敬請期待！

附錄一

變數名稱	年齡	腰圍	收縮壓	舒張壓	空腹葡萄糖	高密度脂蛋白	三酸甘油酯
樣本數	64484	62852	63256	63245	60978	60084	60891
平均數	46.82	78.34	123.27	78.07	93.16	57.31	121.07
中位數	45	78	120.5	77	87	57	92
標準差	13.9	10.68	20.82	11.98	28.93	12.19	111.08
最小值	19	37	70	40	49	10	11
最大值	80	179	276	140	606	154	4137

附錄二：

變數名稱 以性別分組		年齡	腰圍	收縮壓	舒張壓	空腹葡萄糖	高密度脂蛋白	三酸甘油酯
平均數	0	45.9	74.8	119.4	76	92.5	60.7	105.5
	1	48.3	84.4	129.8	81.7	94.3	51.6	147.3
中位數	0	45	74	116	74.5	87	61	82
	1	47	84	127	80.5	88	51	113
眾數	0	38	70	110	70	87	63	58
	1	44	84	120	80	87	52	77

0:女性 1:男性

附錄三：

依序點選主選單中【資料處理】→【資料篩選】。如下圖，步驟一，選取資料後，於步驟二的參數設定中，將篩選條件設定為(Tobacco_Consumption ≠ 0) 。(無(0)、每日一包(1)、每日兩包(2)、每日三包以上(3))，並且保留欲保留的變數。點選【開始處理】。