R-web 資料分析應用：存活分析方法

陳逸萱 副統計分析師

生統 eNews 1-12 期之 R-web 資料分析應用專欄已向大家介紹了【雲端資料分析暨導引系統】（R-web, http://www.r-web.com.tw）中『初階使用者』的各類分析方法，分析的資料類型包含連續型變數與類別型變數。在『線性迴歸分析』中，我們可以評估應變數對於連續型結果變數的影響（如：血壓、BMI）；而『邏輯斯迴歸分析』則可用來評估應變數對於類別型結果變數之影響（主要針對二元變數，如：是否罹病、）。然而，在研究中，另一種常見的結果變數為『時間變數』，即研究開始到產生我們感興趣結果的時間長短（Time to event），如：治療後到發生死亡的時間（存活時間）、疾病復發的時間、燈泡使用壽命等，但實務上受限於研究時間的限制，以至於我們無法觀察到完整的存活時間，所以在『存活分析』的方法中，便會先定義"事件變數值"來表示觀察時間是否為存活時間的指標(即設限指標)。本期的生統 eNews 將跟大家介紹：R-web 裡面存活分析常見的方法，包含『Kaplan-Meier 存活函數估計』、『兩個(含)以上存活函數的比較』、『Cox 比例風險模式』。

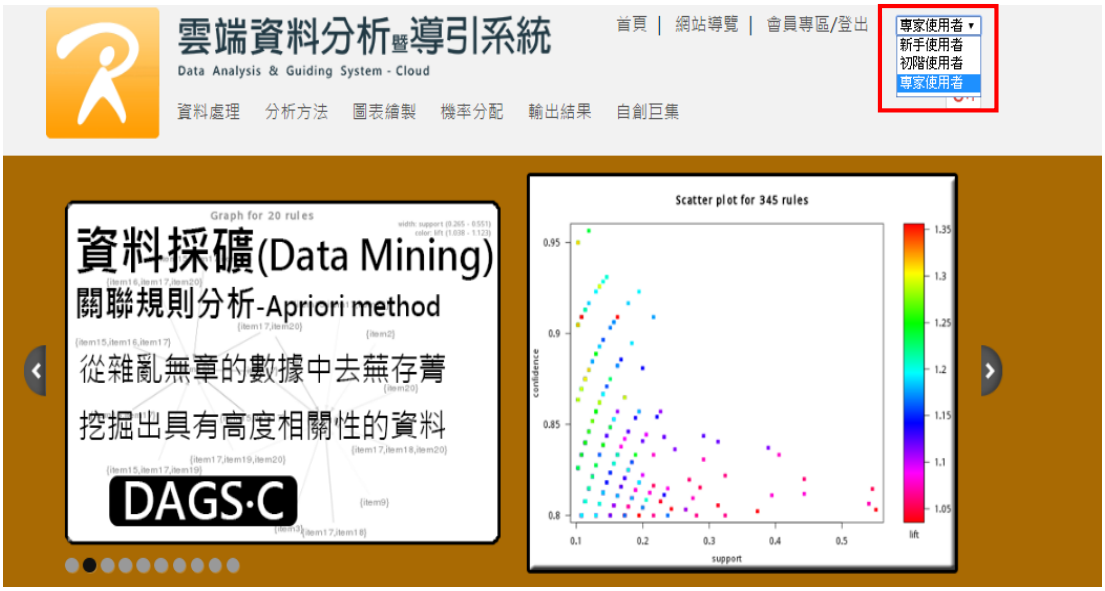
本系列分析將使用Survival Analysis: A Self-Learning Text [1] 書中的例子 “anderson.dat” [2] 來說明。此資料包含了42位血癌（leukemia）病人，其中一半接受標準治療（Rx=1），另一半則接受新治療（Rx=0），此研究欲觀察病人在治療後多久產生復發（Relapse=1），若在研究結束前未發生復發，則為設限資料（Relapse=0）；紀錄產生復發時間或最後觀察時間的時間變數為Surv（單位為週數）；同時，此研究亦記錄了病人的白血球數目（logWBC）與性別（Sex）變數，資料型態可見下頁【表一】。

【表一】: **anderson.dat 前五筆資料樣式**

變數名稱	Subj	Surv	Relapse	Sex	logWBC	Rx
變數型態	數值	數值	數值	數值	數值	數值
1.	1	35	0	1	1.45	0
2.	2	34	0	1	1.47	0
3.	3	32	0	1	2.2	0
4.	4	32	0	1	2.53	0
5.	5	25	0	1	1.78	0

。 。

欲使用R-web的『存活分析』模組，須先將使用者調為「專家使用者」（如下圖），再開始進行分析。



➢ Kaplan-Meier 存活函數估計（Kaplan-Meier survival function estimation）

當我們想觀察樣本的存活狀況時，即可使用『Kaplan-Meier 存活函數估計』。此方法的概念是在考量設限資料的狀況下，估計每個時間點的存活率，將各個時間的存活率同時於一張圖表示，將形成一階梯狀的函數圖形。以範例資料檔為例，我們想比較標準治療與新治療影響復發與否或復發時間的快慢，我們便可來繪製 Kaplan-Meier 存活函數圖形來比較兩種治療的復發時間差異。

在 R-web 主選單中依序點選【分析方法】→【存活分析】→【Kaplan-Meier 存活函數估計】即可進入分析頁面。

操作畫面如上圖所示。第一步，先選擇要進行分析的資料檔，點選”使用者個人資料檔”後選擇”Anderson_dat”的檔案（須先自行匯入此範例資料），系統將自動帶出參數設定畫面。在步驟二選擇要進行分析的變數，在此設定時間變數為” Surv” （Time to event）、事件變數為” Relapse” （是否發生復發）。最後，點選【進階選項】如下圖，選擇”分組變數”為” Rx” ，勾

選"顯示存活函數估計表"、與繪製"存活函數圖（y）"，接著點選【儲存設定】後即可【開始分析】。

下圖為分析結果，左上方可以看到設定的分析變數與相關設定，檢查沒問題即可往下看分析結果。第一個表格為兩組樣本之復發時間之百分位數估計值；第二個表格呈現兩組樣本之存活函數估計表，最後一部分即可看到兩組樣本之存活函數圖。由此我們可觀察血癌病人的新治療組(Rx=0)的療效比標準治療組好(Rx=1)，較不易復發。

存活函數估計 - 分析結果

- 分析方法：Kaplan-Meier 存活函數估計
- 資料名稱：Anderson_dat
- 時間變數：Surv
- 事件變數：Relapse (設限指標：0)
- 分組變數：Rx (0, 1)
- 顯著水準：0.05
- 信賴區間：Log-log
- 計算時間：0.342秒

- 百分位數估計值摘要：

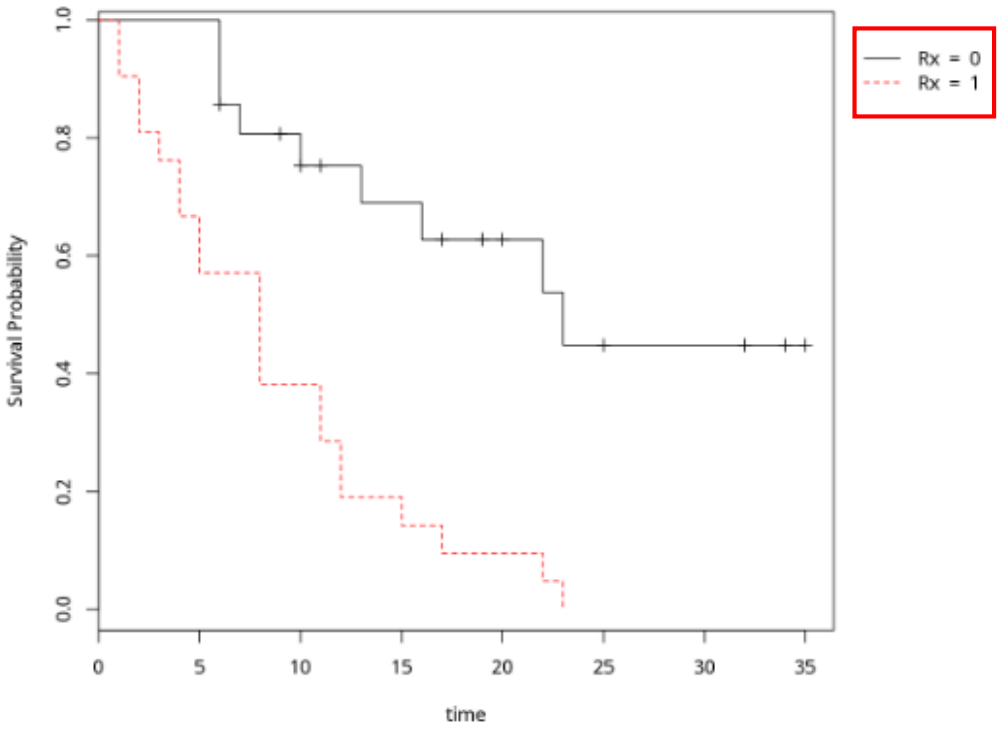
分組變數 grouping variable (Rx)	參數 parameters	估計 estimation	95% 信賴區間 95% C.I.	
			下界 lower	上界 upper
0	25 百分位數	NA	23	NA
	50 百分位數	23	13	NA
	75 百分位數	13	6	22
1	25 百分位數	12	8	22
	50 百分位數	8	4	11
	75 百分位數	4	1	5

- 存活函數估計表：

分組變數 grouping variable (Rx)	時間 time	涉險人數[I] no. at risk	事件人數 no. of event	K-M 存活率估計 K-M survival	標準差 std. err	95 % 信賴區間 95 % C.I.	
						下界 lower	上界 upper
0	6	21	3	0.8571	0.0764	0.6197	0.9516
	7	17	1	0.8067	0.0869	0.5631	0.9228
	10	15	1	0.7529	0.0963	0.5032	0.8894
	13	12	1	0.6902	0.1068	0.4316	0.8491
	16	11	1	0.6275	0.1141	0.3675	0.8049
	22	7	1	0.5378	0.1282	0.2678	0.7468
	23	6	1	0.4482	0.1346	0.1881	0.6801
1	1	21	2	0.9048	0.0641	0.67	0.9753
	2	19	2	0.8095	0.0857	0.5689	0.9239
	3	17	1	0.7619	0.0929	0.5194	0.8933
	4	16	2	0.6667	0.1029	0.4254	0.825
	5	14	2	0.5714	0.108	0.338	0.7492
	8	12	4	0.381	0.106	0.1831	0.5778
	11	8	2	0.2857	0.0986	0.1166	0.4818
	12	6	2	0.1905	0.0857	0.0595	0.3774
	15	4	1	0.1429	0.0764	0.0357	0.3212
	17	3	1	0.0952	0.0641	0.0163	0.2612
	22	2	1	0.0476	0.0465	0.0033	0.197
	23	1	1	0	NaN	NA	NA

I：該時間點存活人數

- 存活函數圖：



➢ 兩個(含)以上存活函數的比較（Comparison for two or more survival functions）

先前的『Kaplan-Meier 存活函數估計』可提供一組或多組的樣本存活函數估計，但並未進一步檢定各組存活函數有無差異。依照本次範例資料，若想我們檢定標準治療組與新治療組影響復發與否或復發時間的快慢的存活函數有無差異時，則可直接使用『兩個(含)以上存活函數的比較』的功能。

在 R-web 主選單中依序點選【分析方法】→【存活分析】→→【兩個(含)以上存活函數的比較】即可進入分析頁面。

操作畫面如上圖所示，首先，先選擇要進行分析的資料檔，點選”使用者個人資料檔”後選擇”Anderson_dat”的檔案（須先自行匯入此範例資料），系統將自動帶出參數設定畫面。在步驟二選擇要進行分析的變數，在此設定時間變數為”Surv”（Time to event）、事件變數為”Relapse”（是否發生復發），並且選擇欲檢定的分組變數”Rx”。完成後點選【進階選項】如下圖，在此可選擇欲使用的檢定方法，亦可依據自己需求設定是否顯示”存活函數估計表”、與繪製各類存活函數圖型，【儲存設定】後即可【開始分析】。

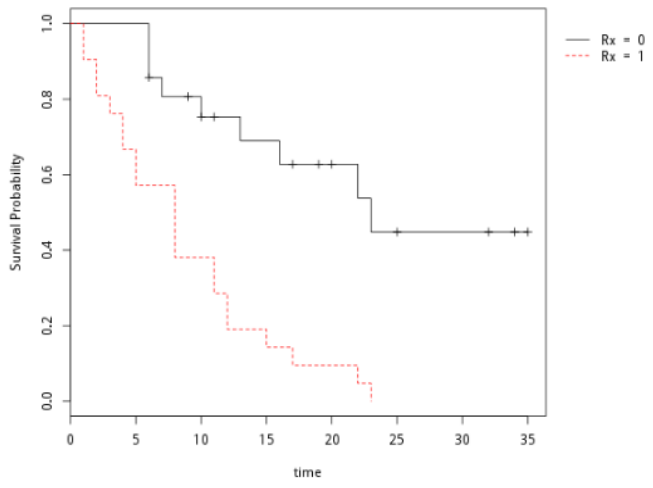
下圖為分析結果，左上方可以看到檢定的變數及相關設定，檢查沒問題後即可看分析結果。第一個表格顯示存活函數比較所用的檢定方法的與檢定結果的 p 值，本分析之虛無假設為兩組存活函數無差異，而 p-值＜0.0001 表顯著，拒絕虛無假設，我們可推論標準治療組與新治療組的復發狀況達顯著差異，且由存活函數圖我們可得知血癌病人的新治療組(Rx=0)的療效比標準治療組好(Rx=1)，較不易復發。

存活函數比較 - 分析結果

- 分析方法：存活函數比較
- 資料名稱：Anderson_dat
- 時間變數：Surv
- 事件變數：Relapse (設限指標：0)
- 分組變數：Rx (0, 1)
- 計算時間：0.251秒
- 存活函數比較：

虛無假設：各存活函數間無顯著差異			
檢定方法 method	檢定統計量 statistics	自由度 d.f.	p-值 p-value
Log-Rank檢定	16.7929	1	< 1e-04

- 存活函數圖：



➢ Cox 比例風險模式（Cox proportional hazards model）

前一方法『兩個(含)以上存活函數的比較』雖可檢定存活函數的差異，但卻無法控制其他干擾因子的影響。依照本次範例資料，若想我們知道在調整性別與白血球值的影響下，選擇標準治療組與新治療組是否仍會影響復發時間的快慢，則可使用『Cox 比例風險模式』的功能來建立模型。

在 R-web 主選單中依序點選【分析方法】→【存活分析】→→【Cox 比例風險模式】即可進入分析頁面。

操作畫面如上圖所示，首先，先選擇要進行分析的資料檔，點選”使用者個人資料檔”後選擇”Anderson_dat”的檔案（須先自行匯入此範例資料），系統將自動帶出參數設定畫面。在步驟二選擇要進行分析的變數，在此設定時間變數為”Surv”（Time to event）、事件變數為”Relapse”（是否發生復發），並且選擇欲放入模型的解釋變數”Rx”、”Sex”、”logWBC”。完成後點選【進階選項】如下圖，在此可選擇欲分層的分層變

數（若不進行分層分析則可忽略），亦可依據自己需求設定是否進行"變數選取"、是否顯示"存活函數估計表"等資訊，另外亦可選擇依照 Cox 比例風險模式繪製各類存活函數圖型，繪製的圖型可參照各變數的平均值或是使用者自行給定，在本次分析中，我們嘗試繪出新治療（Rx=0）的女性（Sex=0）患者，其 logWBC 值為平均值（2.93）的狀況下，所估計出的存活函數圖，設定好後，點選【儲存設定】後即可【開始分析】。

下圖為第一部份分析結果，左上方可以看到檢定的變數及相關設定，檢查沒問題後即可看分析結果。第一個表格顯示 Cox 比例風險模式之參數估計結果與檢定的 p 值，本次分析結果顯示 Rx 變項的 Hazard Ratio 與其 95%CI 為：4.01(1.64-9.83)，表示在控制性別與白血球值的狀況下，標準治

療組(Rx=1)相較於新治療組(Rx=0)復發的風險為 4.01 倍，p 值為 0.0023<0.05 達統計上的顯著，意即兩治療組復發的風險達顯著差異，標準治療組的復發風險較高。第二個表格則呈現模式訊息，可用來判斷本次分析的 Cox 比例風險模式的解釋能力。

Cox比例風險模式 - 分析結果

- 分析方法：Cox比例風險模式
- 資料名稱：Anderson_dat
- 時間變數：Surv
- 解釋變數：Rx, Sex, logWBC
- 事件變數：Relapse (設限指標：0)
- 顯著水準：0.05
- 計算時間：0.271秒
- 最終模式：

變數名稱 variable	係數估計值 coef. esti.	標準差 std. err.	z檢定統計量 z statistic	p值 p-value	估計值的指數 (風險比例) Exp(coef.) (Hazard Ratio)	Exp(coef.)的 95% 信賴區間 下界 lower	上界 upper
Rx	1.3909	0.4566	3.0459	0.0023	4.0184	1.6419	9.8344
Sex	0.2632	0.4494	0.5856	0.5582	1.301	0.5392	3.1394
logWBC	1.5936	0.33	4.8292	< 1e-04	4.9215	2.5775	9.3971

- 模式訊息：

模式適合度(model fitness)			
檢定方法 method	統計量 statistic	自由度 d.f.	p值 p-value
概似比檢定(Likelihood ratio test)	43.752	3	< 1e-04
華德檢定(Wald test)	44.5708	3	< 1e-04
分數(對數-秩)檢定(Log-rank test)	31.74	3	< 1e-04
模式一致性(Concordance)：0.8513			
模式判定係數(R-square)：64.72 %			

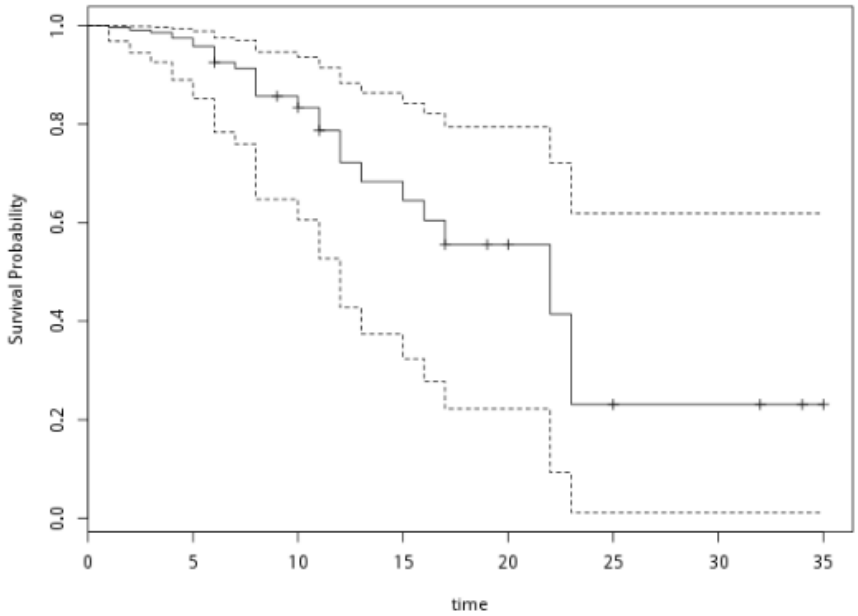
第二部份分析結果可見下圖，由於在我們有在進階選項中，勾選”繪製存活函數圖（共變數=給定值）”，在此便會顯示在給定新治療（Rx=0）、女性（Sex=0）、logWBC 值為平均值（2.93）的狀況下，所估計的存活函數估計表以及存活函數圖。

- 存活函數估計表(共變數值=給定值)：

時間 time	涉險人數[I] no. at risk	事件人數 no. of event	存活率估計 survival	標準差 std. err	95 % 信賴區間 95 % C.I.	
					下界 lower	上界 upper
1	42	2	0.9962	0.0041	0.9685	0.9996
2	40	2	0.9911	0.0084	0.9447	0.9986
3	38	1	0.9862	0.012	0.9259	0.9975
4	37	2	0.9748	0.0193	0.8899	0.9945
5	35	2	0.9586	0.0276	0.8513	0.989
6	33	3	0.9255	0.0418	0.7841	0.9757
7	29	1	0.9131	0.0469	0.7594	0.9704
8	28	4	0.8569	0.0699	0.6475	0.9466
10	23	1	0.8338	0.0784	0.6058	0.9362
11	21	2	0.7873	0.0946	0.5271	0.9145
12	18	2	0.7227	0.1151	0.4276	0.8832
13	16	1	0.6836	0.1261	0.3737	0.8633
15	15	1	0.6447	0.1363	0.3237	0.843
16	14	1	0.6047	0.1454	0.277	0.8211
17	13	1	0.5556	0.1567	0.222	0.795
22	9	2	0.4146	0.1851	0.0927	0.7219
23	7	2	0.2309	0.1931	0.0113	0.6193

I：該時間點存活人數

- 存活函數圖(共變數值=給定值)：
  - 存活函數圖：



本期生統 eNews 的介紹到此告一段落，這次介紹了 R-web 存活分析的三種功能：Kaplan-Meier 存活函數估計、兩個(含)以上存活函數的比較、Cox 比例風險模式，希望大家能更加熟悉這些方法的使用時機與操作方式。存活分析的方法尤其常用於醫學相關領域，若讀者對於這些方法的概念尚不熟悉，建議先閱讀相關參考書，再實際進行分析。

參考資料

1. David G. Kleinbaum, Mitchel Klein (2006), *Survival Analysis A Self-Learning Text*, 2nd Edition.
2. Freireich et al.(1996), *The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia*, Blook 21, 699-716