

Latent Class Model

研究目的

想要研究一些有關老年者頭部外傷的記憶情形，藉由作答情形，來探討並將其歸納為幾個類型，希望達到輔助判讀症狀等一些參考。

使用變數

我採用的變數為 SPMSO 認知功能量表，皆採用受傷後一年(period 為 3 的 datas)總共有 9 大問題，17 個小問題，分別編號，底下為此量表的編碼。

問題	編號	題目	回答記分方式
1	SPMSQ1	你今年幾歲? ___ 歲	答錯—0，答對—1
2-1	SPMSQ2	今天是幾年	答錯—0，答對—1
2-2	SPMSQ3	今天是幾月	答錯—0，答對—1
2-3	SPMSQ4	今天是幾日	答錯—0，答對—1
2-4	SPMSQ5	今天是星期幾	答錯—0，答對—1
3	SPMSQ6	請問您現在是在醫院(診所)、活動中心或家裡? 2=自然說出, 1=經提醒後說出	答錯—0，經提醒後說出—1 自然說出—2
4-1	SPMSQ7	20 減掉 3 等於多少?	答錯—0，答對—1
4-2	SPMSQ8	再減 3 等於多少?	答錯—0，答對—1
4-3	SPMSQ9	再減 3 等於多少?	答錯—0，答對—1
4-4	SPMSQ10	再減 3 等於多少?	答錯—0，答對—1
5	SPMSQ11	請告訴我你的地址。	答錯—0，答對—1
6	SPMSQ12	您母親姓什麼?	答錯—0，答對—1
7	SPMSQ13	現任總統是誰?	答錯—0，答對—1
8	SPMSQ14	上一任總統是誰?	答錯—0，答對—1
9-1	SPMSQ15	您是幾年出生的	答錯—0，答對—1
9-2	SPMSQ16	您是幾月出生的	答錯—0，答對—1
9-3	SPMSQ17	您是幾日出生的	答錯—0，答對—1

由於 1, 5, 7, 8, 11, 16, 17 題大家都答對，無鑑別作用，故這些題目就沒有放在此 model 的 observed variables 中。

Covariates:

原本我考慮有三個 covariates，分別為 old, sex, 與 disease.

研究方法

首先從受傷者的 data 中，挑出一些我感興趣的議題，並且其型態為 nominal

的形式，後來就挑選了 SPMSQ 認知功能量表，滿符合我的預期。

一開始我認為會影響此題作答的有可能有年齡(年齡越長，越難記憶)、性別、受傷之病的個數(可能越高的病數，記憶力會越差)，因此我就挑此三個為我的 covariates.

因此我將其改進的方式是先去看 data 型態，發現在 130 筆資料後，data 中出現 999 的數據，因此猜想是此人沒有做此項測驗，故把它刪掉，以取其有用的 data 來分析。發現觀察 data 果然很重要，不然無論我用何種模式，跑出來的結果都明顯太小，無法創造一個 model 來解釋它。

接著就採用 Latent Class Model 來試試看。我設定的 significance level $\alpha = 0.05$ ，發現在 1-cluster 時候，其 p-value 過小

Model1 - $L^2 = 184.6728$

1-Cluster Model		
Number of cases	129	
Number of parameters (Npar)	10	
Random Seed	131008	
Best Start Seed	131008	
Chi-squared Statistics		
Degrees of freedom (df)	119	p-value
L-squared (L?)	184.6728	0.00011
X-squared	13084210615357.3120	0.0e-2147483647
Cressie-Read	1027801643.1718	1.9e-223183861
BIC (based on L?)	-393.6449	
AIC (based on L?)	-53.3272	
AIC3 (based on L?)	-172.3272	
CAIC (based on L?)	-512.6449	
Dissimilarity Index	0.1701	

因此把它們全部歸類為只有一個 cluster 顯然不是很好。

接著我就嘗試了 2 個 clusters, 看起來就有不錯的改善

Model2 - $L^2 = 93.5765$

2-Cluster Model		
Number of cases	129	
Number of parameters (Npar)	24	
Random Seed	28856	
Best Start Seed	676575	
Chi-squared Statistics		
Degrees of freedom (df)	105	p-value
L-squared (L²)	93.5765	0.78
X-squared	599229.2612	1.0e-129906
Cressie-Read	13339.7597	1.6e-2767
BIC (based on L²)	-416.7038	
AIC (based on L²)	-116.4235	
AIC3 (based on L²)	-221.4235	
CAIC (based on L²)	-521.7038	
Dissimilarity Index	0.0924	

很顯然的，有明顯的改善，因此可以考慮採用此 model.

接著再嘗試比較使用其他 model

Model 3 採用 1-cluster, 並且將第 9 題與第 15 題的 locally independent 取消掉。

Model 4 採用 2-clusters, 並且將第 9 題與第 15 題的 locally independent 取消掉。

最後將這幾個模型做比較

		LL	BIC(LL)	Npar	L ² /b ²	df	p-value	Class.Err.
Model1	1-Cluster	-93.7227	236.0435	10	184.6728	119	0.00011	0.0000
Model2	2-Cluster	-48.1746	212.9846	24	93.5765	105	0.78	0.0002
Model3	1-Cluster	-88.2292	229.9164	11	173.6859	118	0.00065	0.0000
Model4	2-Cluster	-42.8252	212.0055	26	82.8778	103	0.93	0.0002

發現後面這3個models都fit的不錯，但由於Model 2的BIC與 L^2 顯然都比Model 3都要來的好，雖然比Model 4來的稍差，但由於其model較簡單，所以最後我採用Model 2這個model.

Model2 - L²= 93.5765

2-Cluster Model			
Number of cases	129		
Number of parameters (Npar)	24		
Random Seed	28856		
Best Start Seed	676575		
Chi-squared Statistics			
Degrees of freedom (df)	105	p-value	
L-squared (L²)	93.5765	0.78	
X-squared	599229.2612	1.0e-129906	
Cressie-Read	13339.7597	1.6e-2767	
BIC (based on L²)	-416.7038		
AIC (based on L²)	-116.4235		
AIC3 (based on L²)	-221.4235		
CAIC (based on L²)	-521.7038		
Dissimilarity Index	0.0924		

Log-likelihood Statistics			
Log-likelihood (LL)	-48.1746		
Log-prior	-5.5301		
Log-posterior	-53.7047		
BIC (based on LL)	212.9846		
AIC (based on LL)	144.3491		
AIC3 (based on LL)	168.3491		
CAIC (based on LL)	236.9846		
Classification Statistics	Clusters		
Classification errors	0.0002		
Reduction of errors (Lambda)	0.9941		
Entropy R-squared	0.9899		
Standard R-squared	0.9940		
Classification log-likelihood	-48.3373		
AWE	401.9457		
Classification Table	Modal		
Probabilistic	Cluster1	Cluster2	Total
Cluster1	125.9795	0.0000	125.9795
Cluster2	0.0205	3.0000	3.0205
Total	126.0000	3.0000	129.0000
Covariate Classification Statistics	Clusters		
Classification errors	0.0271		
Reduction of errors (Lambda)	0.0000		

Entropy R-squared	0.0958		
Standard R-squared	0.0297		

Parameters

Models for Indicators						
	Cluster1	Cluster2	Wald	p-value	R2	
spmsq2						
0	-2.5059	2.5059	0.7411	0.39	0.5613	
1	2.5059	-2.5059				
spmsq3						
0	-2.5059	2.5059	0.7411	0.39	0.5613	
1	2.5059	-2.5059				
spmsq4						
0	-2.5059	2.5059	0.7411	0.39	0.5613	
1	2.5059	-2.5059				
spmsq6						
1	-0.8054	0.8054	5.4621	0.019	0.0859	
2	0.8054	-0.8054				
spmsq9						
0	-2.3805	2.3805	0.3332	0.56	0.2785	
1	2.3805	-2.3805				
spmsq10						
0	-2.3805	2.3805	0.3332	0.56	0.2785	
1	2.3805	-2.3805				
spmsq12						
0	-0.9828	0.9828	6.3603	0.012	0.1360	
1	0.9828	-0.9828				

spmsq13						
	0	-2.5057	2.5057	0.7411	0.39	0.5613
	1	2.5057	-2.5057			
spmsq14						
	0	-2.5057	2.5057	0.7411	0.39	0.5613
	1	2.5057	-2.5057			
spmsq15						
	0	-2.3805	2.3805	0.3332	0.56	0.2785
	1	2.3805	-2.3805			
Intercepts						
		Overall	Wald	p-value		
spmsq2						
	0	-2.3645	0.6599	0.42		
	1	2.3645				
spmsq3						
	0	-2.3645	0.6599	0.42		
	1	2.3645				
spmsq4						
	0	-2.3645	0.6599	0.42		
	1	2.3645				
spmsq6						
	1	-1.2579	13.3636	0.00026		
	2	1.2579				
spmsq9						
	0	-2.8401	0.4743	0.49		
	1	2.8401				
spmsq10						

	0	-2.8401	0.4743	0.49		
	1	2.8401				
spmsq12						
	0	-1.4339	13.7378	0.00021		
	1	1.4339				
spmsq13						
	0	-2.3644	0.6598	0.42		
	1	2.3644				
spmsq14						
	0	-2.3644	0.6598	0.42		
	1	2.3644				
spmsq15						
	0	-2.8401	0.4743	0.49		
	1	2.8401				
Model for Clusters						
Intercept	Cluster1	Cluster2	Wald	p-value		
	3.6674	-3.6674	6.9088	0.0086		
Covariates						
old						
	-0.0180	0.0180	0.9808	0.32		
sex						
	-0.3737	0.3737	0.4293	0.51		
disease						
	-0.1604	0.1604	0.1784	0.67		

ProbMeans

	Cluster1	Cluster2
Overall	0.9729	0.0271
Indicators		
spmsq2		
0	0.0000	1.0000
1	0.9907	0.0093
spmsq3		
0	0.0000	1.0000
1	0.9907	0.0093
spmsq4		
0	0.0000	1.0000
1	0.9907	0.0093
spmsq6		
1	0.6320	0.3680
2	0.9815	0.0185
spmsq9		
0	0.0000	1.0000
1	0.9817	0.0183
spmsq10		
0	0.0000	1.0000
1	0.9817	0.0183
spmsq12		
0	0.4587	0.5413
1	0.9817	0.0183
spmsq13		
0	0.0000	1.0000

1	0.9907	0.0093
spmsq14		
0	0.0000	1.0000
1	0.9907	0.0093
spmsq15		
0	0.0000	1.0000
1	0.9817	0.0183
Covariates		
old		
1 - 10	1.0000	0.0000
11 - 22	0.9587	0.0413
23 - 34	0.9999	0.0001
35 - 45	0.9998	0.0002
46 - 63	0.9154	0.0846
sex		
1	0.9837	0.0163
2	0.9594	0.0406
disease		
1	0.9839	0.0161
2	0.9994	0.0006
3	0.9258	0.0742

因此我的 model 為

$$P(\text{spmsq}_i | \text{old}, \text{sex}, \text{disease}) = \sum_{k=1}^2 P(\text{cluster}_k | \text{old}, \text{sex}, \text{disease}) \prod_{t=1}^T P(\text{spmsq}_{it} | \text{cluster}_k)$$