廣義線性模式應用分析 **商研二 唐立人**

題目：冰與火之歌角色死亡預測模型

一、 簡介

冰與火之歌(A Song of Ice and Fire)於 1996 年開始由喬治馬丁所撰寫，主要描述虛幻大陸上七大王國的紛爭，其類似於中古世紀歐洲的故事背景，有王族、有權謀、有愛情也有背叛，不同的是，冰與火之歌融入了飛龍以及異鬼等神秘元素，使得劇情更有戲劇張力。其中，這部小說最讓人震撼的點是，它總能夠在你意想不到的時候將你所喜愛的角色賜死，可能某一主線人物上一章節才從戰場上殺出重圍，下一幕就馬上被親人背叛而死去，這樣的轉折總是讓讀者痛心不已，卻也因此讓它們看得目不轉睛。

有鑑於此，本研究想建立冰與火之歌角色死亡模型，除了找出潛在影響角色死亡的因素之外，也想作為預測模型來預測未來角色死亡的機率，讓讀者以及觀眾能提前做好心理準備！

二、 資料集介紹

本研究資料主要來自於 Github 上關於冰與火之歌的專案(附錄一)，由 pombredanne 網友所提供，其包含「章節資料集」、「角色資料集」、「事件資料集」共三個資料集，反映出第一本至第五本小說的人物劇情。以下逐一介紹各個資料集：

1. 章節資料

章節資料記錄每一本書的章節數目以及該章節下的人物視角，一共包含 5 本書、344 章節。

2. 角色資料

角色資料記錄 298 位角色的人口基本變項，包含姓名、所屬陣營、稱號、存活狀態等欄位。

3. 事件資料

事件資料記錄小說各章節發生的重大事件，共 4459 筆，其中事件類別一共 27 種，次數最多的前五種類別為人物登場(89%)、被殺害(3%)、換陣營(3%)、被取名(1%)、被捕捉(1%)，佔了將近 97%的資料量。

為取得更多關於角色的特徵，本研究也使用 Kaggle 上關於冰與火之歌的競賽資料(附錄二)，取其兩個「角色基本資料集」來使用，該資料集包含角色姓名、性別、是否為貴族、存活狀態等欄位。

三、 資料處理

為使資料具有一致性，本研究以 Github 專案上的角色為主、Kaggle 為輔進行分析。在排除掉角色編號重複與具有欄位遺漏值的角色後，本研究以 296 名角色為分析標的。整個資料處理步驟可分為三個部分，以下依序說明：

1. 人口背景變項

角色的人口背景變項包含存活狀態、性別、是否為貴族、是否為前五大陣營(狼家、鹿家、獅家、龍家、守夜人)共八個變數。針對角色陣營，由於資料集中共有 31 個陣營，且前五大陣營佔了 52%的人數，故以前五大陣營作為虛擬變數，其餘陣營作為參考變數，來探究不同陣營對角色死亡機率的影響。

2. 角色行為資料

角色的行為資料包含出現次數、擊殺數、是否換陣營共三個變數，皆由事件資料集所計算而成。

3. 社會網絡資料

本研究假設「只要角色在同一個章節被提到，則彼此間存在著某種程度的關聯」，並依照此假設去建構社會網絡，例如 A、B、C 三個角色都在某一章節被提到的話，則製造一個無方向性關係(Undirected relationship)資料集來呈現，如下表：

無方向性關係資料集

From	To	Weight
A	B	1
A	C	1
B	C	1

依照上述的處理方法，本研究為每一個章節製作關係資料集，並對 344 個資料集進行運算，將相同角色關係(From、To)的 Weight 進行加總，得到整個社會網絡的關係資料集。

為使用 Gephi 與 Ucinet 兩套開源軟體進行社會網絡分析，必須將資料集轉換成無方向性相鄰矩陣(Undirected adjacency matrix)來做輸入，該矩陣的對角線為 0 且存在相鄰關係則增加一個單位，該矩陣樣貌的範例如下表：

無方向性相鄰矩陣

	A	B	C
A	0	1	1
B	1	0	1
C	1	1	0

本研究使用 R 的 igraph 套件，將加總後的關係資料集轉換成矩陣型式，以使用上述兩套軟體(Gephi、Ucinet)產生社會網絡分析常用的指標，來描述每個角色在網絡中的特徵，包含 Degree、Betweeness centrality、Closeness centrality、Clustering coefficient、E-I index、Coreness 六個指標，各變數的定義如下表：

社會網絡指標變數定義

變數	定義
Degree	總共與多少人聯繫
Betweeness centrality	網絡中位於中樞的程度
Closeness centrality	網絡中接近他人的程度
Clustering coefficient	朋友間互相認識的程度
E-I index (External - Internal)	與外人交流的程度(依自己所屬陣營)
Coreness	是否為核心群體

最後，本資料集除了角色名稱外共有 17 個變數，其涵蓋人口背景變項、行為資料以及社會網絡資料，可用於後續建立角色死亡預測模型。

四、 資料特性

為完整了解 296 位角色的存活狀態與特性，本研究以敘述統計與次數統計的方式呈現人口背景資料、行為資料、社會網絡資料三個層面下，各個欄位的統計值與分佈狀況。依序以類別型與量值型介紹如下：

1. 類別型變數

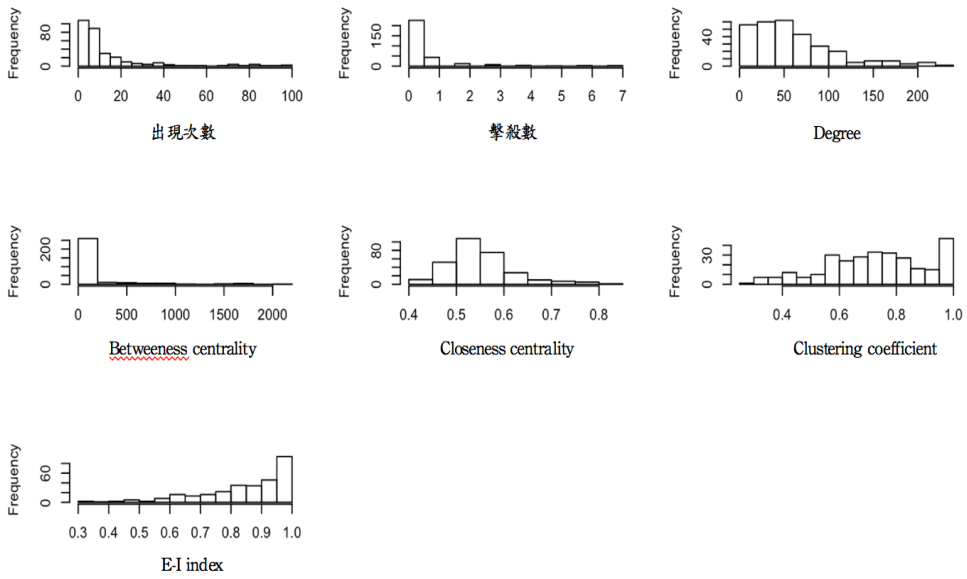
層面	變數		是否死亡			Odds ratio
			否	是	總計	
人口背景變項	是否男性	否	30	18	48	2.117
		%	0.63	0.37	1	
		是	109	139	248	
		%	0.44	0.56	1	
	是否貴族	否	48	81	129	0.50
		%	0.37	0.63	1	
		是	91	76	167	
		%	0.54	0.46	1	
	是否狼家	否	116	129	245	1.09
		%	0.47	0.53	1	
		是	23	28	51	
		%	0.45	0.55	1	

	是否鹿家	否	129	144	273	1.16
		%	0.47	0.53	1	
		是	10	13	23	
		%	0.43	0.57	1	
	是否獅家	否	124	142	266	0.87
		%	0.47	0.53	1	
		是	15	15	30	
		%	0.5	0.5	1	
	是否龍家	否	132	147	279	1.28
		%	0.47	0.53	1	
		是	7	10	17	
		%	0.41	0.59	1	
	是否守夜人	否	129	134	263	2.2
		%	0.49	0.51	1	
		是	10	23	33	
		%	0.3	0.7	1	
行為資料	是否換陣營	否	79	119	198	0.42
		%	0.4	0.6	1	
		是	60	38	98	
		%	0.61	0.39	1	
社會網絡資料	Coreness	否	126	149	275	0.52
		%	0.46	0.54	1	
		是	13	8	21	
		%	0.62	0.38	1	

296 位角色中，死亡事件將近各佔一半，故無不均衡資料的問題。根據單變量的 Odds ratio，如該角色為男性或守夜人，則會大幅提高其死亡機率；如該角色為貴族、換過陣營或身為核心群體，則會大幅降低其死亡機率。

2. 量值型變數

層面	變數	最小值	Q1	中位數	平均數	Q3	最大值
行為資料	出現次數	1	3	7	13.4	15	98
	擊殺數	0	0	0	0.44	0	7
社會網絡資料	Degree	4	28.75	50	60.93	81.5	231
	Betweeness centrality	0	2.95	18.879	124.828	85.145	2159.783
	Closeness centrality	0.40	0.51	0.54	0.55	0.58	0.82
	Clustering coefficient	0.29	0.60	0.74	0.73	0.86	1
	E-I index	0.33	0.77	0.88	0.85	0.97	1



根據量值型變數散佈圖，可發現出現次數、擊殺數、Degree、Betweeness centrality、Closeness centrality 為右偏分佈，表示僅有少部分的角色會大量的出現、殺人或是佔據網絡中重要的位置；Clustering coefficient 與 E-I index 則為左偏分佈，表示角色多半都會與外人做接觸。

五、 分析方法與結果

1. 共線性檢測

針對自變數間的共線性問題，本研究以變異數膨脹係數(VIF)作為判別指標，當 VIF 大於 5 則認為有共線性存在，應刪除該變數或進行維度縮減。使用 R 的 car 套件來計算 VIF，發現有五個變數的 VIF 值大於 5，詳細列於下表：

VIF 異常之變數

變數	Degree	Closeness centrality	出現次數	Clustering Coefficient	Betweeness centrality
VIF	39.44	32.57	16.97	6.43	8.41

有鑒於共線性問題存在且想保留各項變數的特性，本研究使用主成份分析法來進行維度縮減。由於各變數的數值範圍有極大的差異，故使用相關係數矩陣來做標準化，其主成份分析結果如下表。

主成份分析法萃取出一個主成份，其解釋了 89%的變異量，其中每一個變數皆與該成份有關聯(每個主成份負荷量皆 0.4 以上)，故不需要更動變數。該主成份所代表的意涵為「該角色經常出現且交友廣闊，重要的是，其位於社會網絡的中樞」，故將該主成份命名為「網絡重要性」，並以主成份計分作為網絡重要性的分數，分數越高表示該角色在網絡的重要性越高。

主成份分析結果

	網絡重要性
Degree	0.47
出現次數	0.45
Clustering Coefficient	-0.43
Betweeness centrality	0.42
Closeness centrality	0.46
特徵值	4.44
解釋變異量	0.89

經由主成份法處理後，所有自變數皆通過共線性檢測，故可進行模型建立的步驟。

2. 建立死亡預測模型

本研究使用羅吉斯迴歸(Logistic regression)來作為預測模型，依變數為死亡機率，自變數為人口背景變項、行為資料、社會網絡資料三個層面下的變數，共 12 個變數，其統計模式如下列公式。

$g(\mu i) = \beta 0 + \beta 1$*性別$+ \beta 2$*是否貴族$+ \beta 3$*所屬團隊…$+ \beta 7$*所屬團隊

$+ \beta 8$*擊殺數$+ \beta 9$*換陣營$+ \beta 10$*網絡重要性$+ \beta 11$*E-I index$+ \beta 12$*coreness

$\mu i = P$(是否死亡 $= 1 | Xi$) 為死亡機率

使用逐步分析法(Stepwise)來挑選適當的變數，並以最小化的 AIC 指標(Akaike information criterion)作為模型選擇的目標。其統計結果如下表。

經由逐步分析法，該羅吉斯迴歸模型總共採納了十個變數，在顯著水準為 0.05 的情況下，有五個變數達到顯著水準。是否為守夜人、Coreness 皆對死亡機率有顯著地正面影響，其中又以 Coreness 變數影響的程度最大，表示該角色屬於守夜人、該角色屬於核心團體皆有較高的機率死亡，前者反映出守夜人必須與牆外異鬼、野人對戰，暴露在死亡風險下的機會較高，後者反映作者一定程度上喜愛將推動故事進行的主要角色賜死，以製造轉捩點。E-I index、擊殺數、網絡重要性皆對死亡機率有顯著地負面影響，表示該角色越常與外人聯繫、該角色殺人數目越高、該角色在網絡的重要性越高皆使得其死亡機率降低，E-I index 與網絡重要性反映了故事中到處遊說、離間各國的權臣通常不易死亡，殺人數目則反映了作者為了累積反派在觀眾心中的仇恨值，通常會容忍他們持續作惡，直至劇情的高潮再將它們賜死。

羅吉斯迴歸結果

變數	迴歸係數	風險值	係數標準差	P 值	人數占比 (是/總人數)
截距項	1.12	3.06	0.90	0.21	
是否男性	0.59	1.80	0.36	0.1	248/296
是否狼家	0.69	1.99	0.37	0.06	51/296
是否鹿家	1.02	2.77	0.53	0.05	23/296
是否龍家	0.93	2.53	0.61	0.12	17/296
是否守夜人	0.89	2.44	0.45	0.04*	33/296
是否換陣營	-0.55	0.58	0.30	0.07	98/296
Coreness	3.41	30.27	0.91	0***	21/296
E-I index	-2.04	0.13	0.96	0.03*	
擊殺數	-0.45	0.64	0.19	0.02*	
網絡重要性	-0.57	0.57	0.12	0***	

接著進行模型配適度檢定(Goodness of fit)，其 Residual Deviance 為 341.23、自由度為 285，在顯著水準為 0.05 的情況下，其卡方統計量的 P 值為 0.01(1-pchisq(341.23,285))故拒絕虛無假設，表示該模型與飽和模型仍有落差，仍有重要因子尚未被納入，需考慮增加新變數、交互作用項或非線性項。

3. 交叉驗證

建立完死亡預測模型後，接著進行交叉驗證，除了驗證該模型沒有過度配適(Over-fitting)的情況外，也同時取得該模型的準確度(Accuracy)。本研究以 10-fold cross-validation 作為交叉驗證的方法，隨機將樣本分成十等份，其中九等份用於建模、一等份用來驗證，重複驗證十次，讓每個子樣本皆被驗證一次，最後平均十次的結果以得到平均準確率。本研究交叉驗證的結果如下表。

根據下表，該模型的平均準確率為 0.67，從十次驗證結果的準確率來看，其值介於 0.5 到 0.83 之間，表示模型準確率一定程度上受到資料所影響。

10 fold cross-validation 準確率

10-fold	準確率
Fold1	0.60
Fold2	0.76
Fold3	0.68
Fold4	0.70
Fold5	0.53
Fold6	0.69
Fold7	0.77
Fold8	0.83
Fold9	0.63
Fold10	0.5
平均	0.67

六、 結論與建議

1. 結論

本研究以 296 位冰與火之歌的角色為研究對象，使用人口背景變項、行為資料、社會網絡資料三個層面下的變數，來建立角色死亡預測模型。根據本研究模型得到以下三個發現：

(1)角色所屬的陣營決定其生存環境，使得死亡機率受到影響。

(2)為了累積壞人在觀眾心中的仇恨值，作者會讓他持續作惡，直至劇情高潮。

(3)透過殺害觀眾心愛的核心角色，來製造劇情轉折點。

除了探討不同變數對角色死亡機率的影響，本研究也透過 10 fold cross-validation 來檢視該模型的配適情形以及準確率，發現其平均準確率為 67%，優於隨機猜測的 50%，表示該模型確實能提高正確預測角色死亡的機率。

2. 建議

本研究根據結論提出三點建議，作為未來研究可以改善的方向：

(1)由於小說故事採人物視角進行，故可能會有時間軸重疊的問題，未來研究可考慮以單一主角的視角來進行分析，以確保時間軸的一致性。

(2) 有鑒於該模型的 Goodness of fit 並沒有通過（顯著，拒絕虛無假設），表示仍有重要變數未被納入模型，故未來研究可考慮加入新變數、交互作用項或非線性項。

(3)從 10 fold cross-validation 發現，不同驗證模型的準確度存在差異，表示資料集內不同角色的死亡特性存在著極大的差異，故未來可考慮更近一步篩選樣本，例如剔除出現次數過少或無關緊要的人物。

附錄

1. Github 冰與火之歌專案
   https://github.com/pombredanne/GoT-SNA/tree/master/data
2. Kaggle 冰與火之歌競賽
   https://www.kaggle.com/mylesoneill/game-of-thrones