

2002 世界盃足球賽比賽結果預測之研究

許伯陽¹ 周靈山²

¹中華民國體育運動總會 ²吳鳳技術學院

摘 要

本研究的目的是以 2002 年世界盃足球賽 13 項攻防數據為預測變項，包括進球數、射門數、射在球門內次數、犯規、角球、任意球、點球、越位、烏龍球、黃牌、紅牌、控球百分比以及實際比賽時間等，由多元邏輯迴歸來建立足球賽比賽結果(勝、和、負)的預測模式。邏輯迴歸模式為：

$$\text{勝與不勝邏輯迴歸方程式：} \hat{p}_3 = \frac{1}{1 + e^{-(-2.89 + 1.75(\text{進球數}))}}$$

$$\text{負與不負邏輯迴歸方程式：} (\hat{p}_2 + \hat{p}_3) = \frac{1}{1 + e^{-(-0.95 + 1.75(\text{進球數}))}}$$

當進球數大於 1.4578 時，迴歸模式預測為勝的機率大於不勝的機率；進球數小於 0.7364 時，迴歸模式預測為負的機率大於不負的機率；若是在兩者之間，則迴歸模式預測為和的機率大於勝或負的機率。評估預測模式品質的一致性百分比為 71.2，不一致性百分比為 7.8，同值百分比為 21，其四種指標分別為 Somers'D：0.634，Gamma：0.802， τ -a：0.425，c：0.817。整體而言，2002 世界盃足球賽比賽結果的預測模式品質良好。

關鍵詞：世界盃足球賽，多元邏輯迴歸。

壹、緒論

一、研究動機

四年一度的世界盃足球賽象徵著足球賽的最高賽會，結合了如下的多項優勢，FIFA(世界足球總會)提供合理的競賽制度與比賽規則以及公平的賽程與公正的裁判；各參與國為籌組國家代表隊對足球所下的投資，包括足球硬體的建設、球員的培訓與教練的選用等；世界頂尖的球員為國家的榮譽、個人的價值全力爭取為國出賽；主辦國傾全力興建標準場地、舉辦無瑕疵的球賽。業餘賽比職業賽更受球團與球員的重視，是其他球類運動所無法比擬的。為此，世足賽的與賽各隊其實力之接近、素質之整齊，以及 FIFA 賽制賽程的規劃和各國重視的程度等等，皆能提供優質的攻防數據，提供其他盃賽無法呈現的有效性及邏輯性，世足賽的攻防數據確實是值得深入研究的資料。另外，準確預測足球比賽的勝負一直是球評專家致力想達成的目的之一，在合法博奕的國家中，博奕公司的賭盤也是專注於準確的預測球賽的勝負，為求開出合理的賭盤，賺取最多的利潤。因此，對於足球比賽勝負的預測是很受重視的主題，除了提供專家的論斷之外，數據引用方式的提升與適切性可增加數據對準確預測結果的影響，與專家的判斷相輔相成，更可提高預測的準確性以及合理性。

在數據分析的統計方法應用上，預測或分類的方法有線性區別函數(linear discriminate function)和邏輯迴歸(logistic regression)(陸偉明，民 85)，Hosmer、Hosmer 與 Fisher(1983)指出，當資料為混合變數時，線性區別函數對參數在連續變數上，有漸進不偏估計，但對於離散變數之參數估計有嚴重的偏差。Press 與 Wilson(1978)則指出，當資料是連續與離散混合型態時，邏輯迴歸的分類正確性高於線性區別函數。因此，由於本研究引用世足賽比賽的攻防數據不全然是連續變數，所以本研究採用邏輯迴歸分析建構預測世足賽球賽勝負的模式。

邏輯迴歸是指當依變項(Y)為二分名義變項(如男與女、成功與失敗、贊成與反對等)，或是多分類變項(如低收入、中收入與高收入；標準、過重與肥胖等)，這些依變項都是等距量尺以下，也就是名義(nominal)和順序(ordinal)量尺，無法和自變項直接構成一般的迴歸模式關係，因為在 X、Y 座標軸中，依變項(Y)永遠與自變項(X)平行，線性迴歸在依變項為等距量尺以下時，其模式是不適合的，因此要採用邏輯迴歸模式才能表達兩者間的關係(林清山，民 87)。本研究即是以順序量尺的「勝、和、負」為效標，由 2002 世足賽所提供的 13 項攻防數據為預測變項，由多元邏輯迴歸來建立預測模式。

在邏輯迴歸的文獻上，陸偉明(民 85)由空軍官校飛訓生的動作技能測驗與多項心理測驗的成績，以初階飛行結果為效標，適配邏輯迴歸，建立預測模式，提供預測飛訓生初階飛行表現的機率值，其預測模式的分類正確率達 77.2%(正確錄取 66.2%+正確淘汰 11%)，錯誤淘汰率只有 5.6%，可以提升甄試的效用與節省施訓成本。姚漢禱(民 87)編製羽球效標參照測驗中，以全國甲組與乙組選手共 223 名為效標，以體能測驗與技術測驗為自變項，由邏輯迴歸建立的預測模式為： $\text{logit}(p) = -11.36 + 0.05 \times \text{網前平推球} + 0.03 \times \text{殺球} - 0.46 \times \text{五次前後跑} + 0.05 \times \text{立定跳遠}$ 。分類正確的一致性達 86.4%，不一致性為 13.5%，研究的結果可做為教練選材時的依據。對於預測的效標為二分類(或多分類)的名義(或順序)變項時，邏輯迴歸可做預測分類與建立模式的實用工具。

二、研究目的

本研究以 2002 年世界盃足球賽的 13 項攻防數據，包括進球數、射門數、射在球門內次數、

犯規、角球、任意球、點球、越位、烏龍球、黃牌、紅牌、控球百分比以及實際比賽時間等，建立世足賽比賽結果(勝、和、負)的預測模式。

三、研究限制

由於一場比賽中「勝、和、負」的出現並非完全獨立，如一隊得勝則另一隊必然為負，因此一場比賽的兩隊攻防數據在使用上不如每筆數據完全獨立般適合使用。另外，勝負之資料並非絕對性，而是相對性的資料，兩強之勝與兩弱之勝，其意義並不相同，惟因世足賽 32 支球隊皆為頂尖球隊，勝負間的差異極小，由此可提高資料的絕對性。

貳、研究方法

一、研究對象

本研究以 2002 世界盃足球賽的 32 支參賽球隊為研究對象，比賽賽制如下，32 隊分成八組，每組四隊打單循環賽，每隊各比三場，各組取前兩名晉級，共 16 隊參與第一階段的單敗淘汰賽制，勝隊晉級，共 8 隊參與第二階段單敗淘汰賽，勝隊晉級，共 4 隊參與準決賽賽程，勝的兩隊進行冠軍賽，敗的兩隊進行季軍賽。因此，2002 世界盃足球賽總共的賽程有 32 隊分組循環共計 48 場 ($C_2^4 \times 8$)，16 強單敗淘汰賽 8 場，8 強單敗淘汰賽 4 場，4 強準決賽 2 場，以及冠軍賽與季軍賽，合計 64 場，每一場比賽各可提供兩隊攻防數據，所以總共有 128 筆數據可供研究。

二、實驗設計

以 2002 世界盃足球賽所有比賽的攻防數據為本研究的預測變項，計有進球數(goal scored)(射進球門的次數)、射門數(shot)(起腳攻門的次數)、射在球門內次數(shot on goal)(射在球門範圍內的次數)、犯規(foul)(由裁判判定之違反規則情事)、角球(corner kick)(出底線的攻方球之發球方式)、任意球(free kick)(守方犯規的發球方式)、點球(penalty kick)(罰 12 碼球)、越位(offside)(攻方前方無防守者)、烏龍球(own goal)(守方誤將球踢進自己球門)、黃牌(caution)(警告)、紅牌(expulsion)(驅逐出場)、控球百分比(ball possession %)(控制球的時間比)、實際比賽時間(actual playing time)(有效掌握球的時間)等，共有 13 項(數據的內容為 FIFA 在賽後所統一提供的資料)，本研究為求數據統一起見，將有延長賽以及 PK 賽(penalty kick)的比賽視為和局，數據以 90 分鐘正規賽為主，合計有 16 強賽的西班牙 VS.愛爾蘭、韓國 VS.義大利、瑞典 VS.塞內加爾、8 強賽的西班牙 VS.韓國、土耳其 VS.塞內加爾等，共 5 場。由比賽的 13 項攻防數據為預測變項，以比賽結果「勝、和、負」等順序變項為效標，使用多元邏輯迴歸分析來預測比賽之結果，以及建立具有影響力的預測變項之迴歸模式。

三、資料分析

本研究先以 SPSS 10.0 版對 128 筆數據做敘述統計的分析以及「勝、和、負」三組分別在 13 項預測變項的獨立樣本單因子變異數分析，然後以 SAS 8e 版執行多元邏輯迴歸分析，分析與處理 128 筆數據。

參、結果與討論

一、勝、和、負的差異比較

2002 世界盃足球賽 128 筆的攻防數據中，勝的有 45 筆，和的有 38 筆，負的有 45 筆，以此

比賽結果在 13 項攻防數據的差異比較，進行獨立樣本單因子變異數分析，結果如表二所示(篇幅關係僅列有差異者)，13 項攻防數據中只有進球數、射門數、射在球門內次數和點球有差異，其餘的 9 項攻防數據無差異。事後比較(如表三)，發現在進球數變項上三組都有差異，勝組優於和組(平均差 1.22)，勝組優於負組(平均差 1.78)以及和組優於負組(平均差 0.56)；射門數變項上勝組優於和組(平均差 2.21)，勝組優於負組(平均差 3.84)；射在球門內次數變項上勝組優於和組(平均差 2.19)，勝組優於負組(平均差 3.04)；點球變項上勝組優於負組(平均差 0.20)，和組優於負組(平均差 0.24)。在四項有差異的預測變項中，點球該變項的差異由於負組中沒有罰進點球的資料，因此與勝組與和組有顯著差異，其餘三項變項的差異均以勝組為最佳，和組次之，負組最低。

表一 各組之敘述統計摘要表

預測變項	組別	次數	平均數	標準差
進球數	勝	45	2.22	1.31
	和	38	1.00	0.74
	負	45	0.44	0.69
射門數	勝	45	13.24	4.01
	和	38	11.03	3.51
	負	45	9.40	3.53
射在球門內次數	勝	45	7.11	2.21
	和	38	4.92	1.78
	負	45	4.07	1.97
點球	勝	45	0.20	0.40
	和	38	0.24	0.49
	負	45	0	0

表二 勝、和、負三組變異數分析摘要表

預測變項	變異來源	離均差平方和	自由度	均方	F 值	p 值
進球數	組間	74.08	2	37.04	39.61*	0.0001
	組內	116.89	125	0.94		
	全體	190.97	127			
射門數	組間	334.88	2	167.44	12.24*	0.0001
	組內	1710.09	125	13.68		
	全體	2044.97	127			
射在球門內次數	組間	220.46	2	110.23	27.45*	0.0001
	組內	502.01	125	4.02		
	全體	722.47	127			
點球	組間	1.40	2	0.70	5.45*	0.0050
	組內	16.07	125	0.13		
	全體	17.47	127			

*p<.05

表三 勝、和、負三組事後比較摘要表

變項	組別	勝	和	負	變項	組別	勝	和	負
進球數	勝($\bar{X}=2.22$)	--	1.22*	1.78*	射在球門內次數	勝($\bar{X}=7.11$)	--	2.19*	3.04*
	和($\bar{X}=1.00$)	--	--	0.56*		和($\bar{X}=4.92$)	--	--	0.85
	負($\bar{X}=0.44$)	--	--	--		負($\bar{X}=4.07$)	--	--	--
射門數	勝($\bar{X}=13.24$)	--	2.21*	3.84*	點球	勝($\bar{X}=0.20$)	--	-0.04	0.20*
	和($\bar{X}=11.03$)	--	--	1.63		和($\bar{X}=0.24$)	--	--	0.24*
	負($\bar{X}=9.40$)	--	--	--		負($\bar{X}=0$)	--	--	--

*p<.05

二、多元邏輯迴歸分析

本研究採用逐步法(stepwise)選取最佳的預測變項，投入迴歸公式設定 SLENTRY 為 0.01，邏輯迴歸按預測變項的重要性逐次增加投入模型中，直到沒有預測變項達顯著(SLENTRY<0.01)

爲止；同時也進行剔除預測變項，設定 SLSTAY 爲 0.01，也就是投入迴歸模型中的預測變項未達設定值顯著者(SLSTAY<0.01)，予以剔除。由邏輯迴歸所建立的預測模式首先選取的預測變項爲進球數(如表五)，但之後沒有其他預測變項達到選入(SLENTY<0.01)與剔除(SLSTAY<0.01)的標準，所以邏輯迴歸的預測變項只有進球數。

(一)截距與預測變項的估計

因效標變項爲「勝、和、負」三組，在多元邏輯迴歸中會構成兩條迴歸線，勝與不勝(即爲和與負)構成第一條迴歸線，負與不負(即爲勝與和)構成第二條迴歸線；由最大可能率估計法(maximum likelihood estimation)得到截距 1 的估計值爲-0.61， χ^2 值爲 10.94，達顯著水準；截距 2 的估計值爲 0.61， χ^2 值爲 10.94，達顯著水準，表示截距 1 與截距 2 的存在有統計上的意義。

另外，預測變項的選取與剔除是根據 SCORE 統計量來決定，本研究 13 項預測變項的 SCORE 統計量(如表五)；預測變項達顯著水準的大小順序依序爲：進球數、射在球門內次數、射門數與點球，其 SCORE χ^2 值分別爲 47.66、32.76、15.86 及 6.59。首先選取 SCORE χ^2 值最大者投入模式中，也就是進球數。

表四 截距投入模式之檢定摘要表

參數	自由度	估計值	標準誤	χ^2 值	p 值
截距 1	1	-0.61	0.19	10.94*	0.0009
截距 2	1	0.61	0.19	10.94*	0.0009

*p<.05

表五 預測變項的 SCORE 統計量摘要表

預測變項	自由度	SCORE χ^2 值	p 值
進球數	1	47.66	0.0001*
射門數	1	15.86	0.0001*
射在球門內次數	1	32.76	0.0001*
犯規	1	0.64	0.42
角球	1	0.48	0.49
任意球	1	0.19	0.66
點球	1	6.59	0.01*
越位	1	0.67	0.41
烏龍球	1	0.09	0.76
黃牌	1	1.39	0.24
紅牌	1	2.31	0.13
控球百分比	1	0.07	0.79
實際控球時間	1	0.01	0.91

*p<.05

(二)投入進球數

1.參數的模式適合度檢定

在投入進球數爲唯一的預測變項時，首先要檢驗模式的適合度(如表六)，檢定預測變項與效標變項間的關係，AIC(Akaike information criterion)、SC(Schwartz criterion)和-2log L(-2log likelihood statistic)三指標都是概似率的函數(彭昭英，民 90)，其值達顯著，表預測變項與效標變項有顯著相關，迴歸模式不是隨機因素所造成的，所建立的模式是有意義的。

表六 參數的模式適合度檢定摘要表

適合度檢定指標	截距	截距加共變項	p 值
AIC	284.46	213.95	<0.0001*
SC	290.17	222.51	<0.0001*
-2log L	280.46	207.95	<0.0001*

*p<.05

2. 迴歸係數(β)的檢定

接著檢定迴歸係數(β)(如表七)，同樣有三種指標可供觀察，概似率 (likelihood ratio)、SCORE 統計量和 Wald 統計量，三指標的 χ^2 值都達顯著水準，表示拒絕虛無假設($\beta=0$)，迴歸係數是存在的。

表七 迴歸係數檢定摘要表

檢定指標	自由度	χ^2 值	p 值
概似率	1	72.52	0.0001*
SCORE	1	47.66	0.0001*
Wald	1	45.76	0.0001*

*p<.05

3. 迴歸分析

逐步法首先選取的預測變項為進球數，經由最大可能率估計分析得知參數估計值為(如表八)：截距 1 為-2.89，截距 2 為-0.95，進球數為 1.75，轉換為 χ^2 值截距 1 為 51.26，截距 2 為 9.82，進球數為 45.76，三者都有達到顯著水準(p<0.05)，亦即接受截距和進球數不為零的假設，以及有存在的意義。另外，進球數的勝數比(odds Ratio)點估計為 5.78，意思為每增加一個進球，勝利發生率增加 5.78 倍，也可說進球數對勝利有正向的影響。另外，由圖一可知「勝、和、負」三組由兩條平行的迴歸線構成，斜率為 1.75，例如當進球數為 0 時， $\logit(\hat{p}_3) = -2.89$ ， $\logit(\hat{p}_2 + \hat{p}_3) = -0.95$ ，斜率為正的意義為當進球數增加時預估的邏輯值也增加，呈正向的反應。

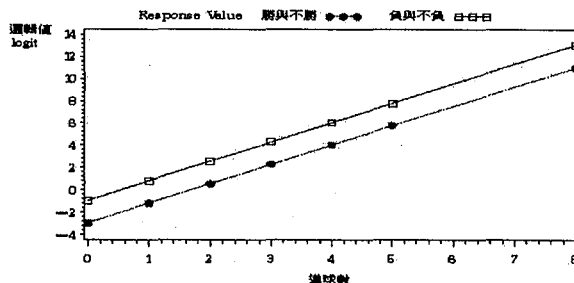
表八 多元邏輯迴歸摘要表

參數	自由度	估計值	標準誤	χ^2 值	p 值	勝數比 點估計
截距 1	1	-2.89	0.40	51.26	<0.0001*	
截距 2	1	-0.95	0.30	9.82	0.0017*	
進球數	1	1.75	0.26	45.76	<0.0001*	5.78

*p<.05

所以 2002 世足賽比賽結果的預測模式為：

(1) 邏輯反應函數方程式



圖一 進球數邏輯反應函數圖

兩迴歸線的邏輯反應函數方程式如下(\hat{p}_3 、 \hat{p}_2 、 \hat{p}_1 分別表示勝、和、負的機率值)：

勝與不勝的邏輯反應函數方程式： $\logit(\hat{p}_3) = -2.89 + 1.75(\text{進球數})$

負與不負的邏輯反應函數方程式： $\logit(\hat{p}_2 + \hat{p}_3) = -0.95 + 1.75(\text{進球數})$

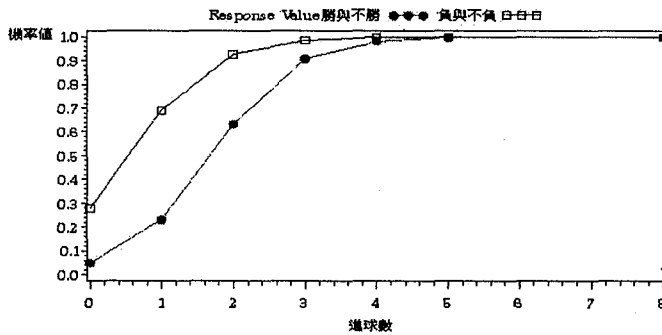
(2) 邏輯迴歸方程式

以下為「勝、和、負」三組的邏輯迴歸方程式，由圖二中可看出，轉換後的邏輯

值會成爲 0 到 1 之間的機率值，當進球數爲 0 時，分出勝與不勝的機率爲 0.0527，低於分出負與不負的機率 0.2789；當進球數爲 8 時，分出勝與不勝以及分出負與不負的機率都接近 1，可說當進球數增加，分出「勝、和、負」的機率也跟著增加。其邏輯迴歸方程式如下：

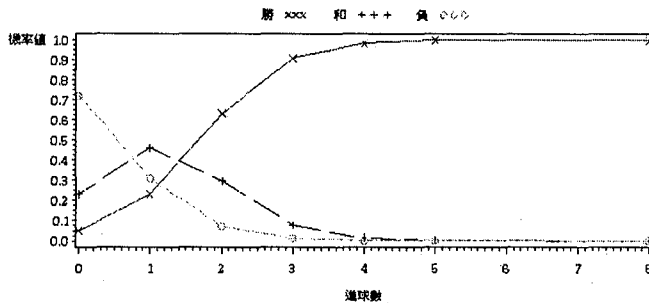
$$\text{勝與不勝邏輯迴歸方程式： } \hat{p}_3 = \frac{1}{1 + e^{-(-2.89 + 1.75(\text{進球數}))}}$$

$$\text{負與不負邏輯迴歸方程式： } (\hat{p}_2 + \hat{p}_3) = \frac{1}{1 + e^{-(-0.95 + 1.75(\text{進球數}))}}$$



圖二 進球數預測機率曲線圖

由圖三可知，在機率爲 0 到 1 的範圍內，當進球數增加時，負的機率密度迅速減小，和的機率密度先增加之後才減小，而勝的機率是隨著進球數的增加而增加。在圖三中可以觀察負的機率函數先與和的機率函數產生一個截點(cutpoint)，然後和的機率函數在與勝的機率函數產生第二個截點，這表示「勝、和、負」雖不等距，但還是有一定的順序關係。這兩個截點爲分組的臨界點，由 $\hat{p}_1 = \hat{p}_2$ 可算出負與和的截點，爲 0.7364，由 $\hat{p}_2 = \hat{p}_3$ 可算出和與勝的截點，爲 1.4578，也就是說當進球數大於 1.4578 時，迴歸模式預測爲勝的機率大於不勝的機率；進球數小於 0.7364 時，迴歸模式預測爲負的機率大於不負的機率；若是在兩者之間，則迴歸模式預測爲和的機率大於勝或負的機率。



圖三 各組進球數機率密度圖

個別的機率密度函數分別為：勝的機率密度函數： $\hat{p}_3 = \frac{1}{1 + e^{-(-2.89+1.75(\text{進球數}))}}$

和的機率密度函數： $\hat{p}_2 = \frac{1}{1 + e^{-(-0.95+1.75(\text{進球數}))}} - \hat{p}_3$

負的機率密度函數： $\hat{p}_1 = 1 - \hat{p}_3 - \hat{p}_2$

(三)邏輯迴歸模式品質的評估

由觀察反應值和預測機率值之間的相關來評估模式品質，一致性(concordant)百分比為 71.2，不一致性(discordant)百分比為 7.8，同值(tied)百分比為 21，代表模式將比賽結果正確分類的比率，一致性百分比愈高愈能說明模式的品質優良。另外，四個等級相關的指標，用以說明模式品質。Somers'D 為 0.634，Gamma 為 0.802， τ -a 為 0.425，c 為 0.817，指標愈高表示分類的愈正確。

肆、結論與建議

一、結論

本研究以 2002 年世界盃足球賽 128 筆攻防數據為研究對象，由攻防數據建立足球賽比賽結果(勝、和、負)的預測模式，研究的結論如下：

(一)13 項攻防數據的重要性依序如下(如表五)：進球數、射在球門內次數、射門數、點球、紅牌、黃牌、越位、犯規、角球、任意球、烏龍球、控球百分比及實際控球時間等，投入迴歸模式的預測變項為進球數，顯示最能影響比賽結果的攻防數據為進球數。

(二)建立的邏輯迴歸模式為：

勝與不勝邏輯迴歸方程式： $\hat{p}_3 = \frac{1}{1 + e^{-(-2.89+1.75(\text{進球數}))}}$

負與不負邏輯迴歸方程式： $(\hat{p}_2 + \hat{p}_3) = \frac{1}{1 + e^{-(-0.95+1.75(\text{進球數}))}}$

(三)進球數的勝數比(odds ratio)點估計為 5.78，意思為每增加一個進球，勝利發生率增加 5.78 倍，可說進球數對勝利的影響為正，進的球數愈多，愈能獲得勝利。

(四)由圖三發現：勝、和、負產生了兩個截點(cut point)，負與和的截點為 0.7364，和與勝的截點為 1.4578，當進球數大於 1.4578 時，迴歸模式預測為勝的機率大於不勝的機率；進球數小於 0.7364 時，迴歸模式預測為負的機率大於不負的機率，若是在兩者之間，則迴歸模式預測為和的機率大於勝或負的機率。

(五)評估預測模式品質的一致性(concordant)百分比為 71.2，不一致性(discordant)百分比為 7.8，同值(tied)百分比為 21。四個等級相關的指標用以說明模式品質，Somers'D 為 0.634，Gamma 為 0.802， τ -a 為 0.425，c 為 0.817。整體而言，2002 世足賽比賽結果的預測模式品質良好。

二、建議

以數據為預測的依據，在本研究預測模式的應用上，可以同一賽事的進球數平均為依據，

例如：預測本屆世界盃冠軍戰巴西與德國的勝負，巴西在前六場比賽中總共射進 16 球，平均一場比賽可進 2.67 球，德國隊在前六場比賽中總共射進 14 球，平均一場比賽可進 2.33 球，以此數據帶入預測模式中預估冠軍戰的勝負(本預測模式包含冠軍賽的攻防數據)，可知巴西 90 分鐘正規賽獲勝的機率為 0.8560，不敗的機率為 0.9764；德國隊 90 分鐘正規賽獲勝的機率為 0.7684，不敗的機率為 0.9580。兩對取得勝利的機率都很高，但兩相比較之下，還是巴西隊獲得勝利的機率高一些。

適切的引用數據與適當的分析方法可增加數據對準確預測結果的影響，但數據依然是已經發生過的結果，對於尚未進行的比賽確實無法做到正確無誤的預測，只能儘量減低發生錯誤的機率，提供合宜的分析結果給專家去做更專業的判斷，例如：可以上例於賽前提供給足球專業人士，做為預測結果的依據之一，其他依據如球員的健康狀況、犯規狀況如：德國隊中場重要球員巴拉克(Ballack, M.)禁賽、教練的戰術、球隊的氣勢運氣等，將諸多因素匯集成最後的結論，結合專家的質與數據的量才能做出最合理的判斷。

參考文獻

- 沈明來(民 86)：實用無母數統計學與計數資料分析。台北：九州圖書文物。
- 林清山(民 87)：多元羅吉式迴歸係數的最大可能性估計、顯著性考驗以及多元羅吉式迴歸分析模式的適合度考驗。中國測驗學會測驗年刊，45 輯 1 期，181-200 頁。
- 彭昭英(民 90)：SAS 與統計分析。台北：儒林圖書有限公司。
- 姚漢禱(民 87)：編製羽球效標參照組合測驗。國科會專案計畫。計畫編號：NSC 88-2413-H-179-007。
- 陸偉明等(民 85)：一個預測初階飛行表現之羅吉斯模式。中國測驗學會測驗年刊，43 輯 1 期，385-394 頁。
- Agresti, A.(1996). An introduction to categorical data analysis. New York: John Wiley & Sons Inc.
- Hosmer, D. W. & Lemeshow, S.(1989). Applied logistic regression. New York: John Wiley & Sons Inc.
- Hosmer, T. A., Hosmer, D. W., & Fisher, L.(1983). A comparison of three methods of estimating the logistic regression coefficient, Communication Statistics Computation, 12, 727-751.
- Press, S. J. & Wilson, S. (1978). Choosing between logistic regression and discriminate analysis. Journal of the American Statistical Association, 73, 699-705.
- SAS Institute Inc. (1995). Logistic regression examples using the SAS system, version 6, first edition. North Carolina: SAS Institute Inc.

投稿日期：91 年 08 月

審查日期：91 年 09 月

接受日期：91 年 12 月

A STUDY OF PREDICTION BY THE RESULTS OF LOGISTIC REGRESSION ON THE GAME IN 2002 FIFA WORLD CUP

Po-yang Hsu¹ & Lin-sang Chou²

¹Republic of China Sports Federation & ²Wu-Feng Institute of Technology

ABSTRACT

The purpose of the study was to construct a prediction model by the results of the game in 2002 world cup. The dependence variables were the group of win, draw and losing. The independence variables were including 13 offense-defense indexes (goal scored, shot, shot on goal, foul, corner kick, free kick, penalty kick, offside, own goal, caution, expulsion, ball possession %, actual playing time). Model construction could be analyzed by multiple logistic regression. The results of logistic regression equation were

$$\text{Win and draw or losing: } \hat{p}_3 = \frac{1}{1 + e^{-(-2.89 + 1.75(\text{goalscored}))}}$$

$$\text{Losing and win or draw: } (\hat{p}_2 + \hat{p}_3) = \frac{1}{1 + e^{-(-0.95 + 1.75(\text{goalscored}))}}$$

If the goals scored were more than 1.4578, the probability of win was more than draw or losing in regression prediction model. If the goals scored were less than 0.7364, the probability of losing was more than win or draw in regression prediction model. If the goals scored were between 1.4578 and 0.7364, the probability of draw was more than win or losing in regression prediction model. The quality of prediction model were percent concordant 71.2, percent discordant 7.8 and percent tied 21. The indexes of prediction model were Somers'D 0.634, Gamma 0.802 τ -a 0.425, c 0.817. On the whole, the results of the game had good characteristics in 2002 world cup.

Key words: world cup, multiple logistic regression.