體育學報
中華民國體育學會
第 50 卷第 S 期，頁 79～90 (2017. 12.)
DOI:10.3966/102472972017125005006

以大數據分析球員技術面表現、對戰組合與中華職棒歷年票房之相關性

許懷中[1]、黃致豪[2]
[1]臺灣 臺中市 逢甲大學資訊工程學系
[2]臺灣 臺中市 國立臺灣體育運動大學運動資訊與傳播學系

摘 要

緒論：在棒球比賽中，賽場上有眾多元素能吸引現場觀眾，包括三振、盜壘、全壘打、投打對決等，而這些技術面表現，以及對戰組合，哪些眞正能影響比賽票房，爲本研究目的。研究團隊計算中職元年至 27 年間公開可取得的球員表現與票房之關係，以資料驅動 (data-driven) 的選模方法，由數據之間的關係以及資訊理論如關聯性分析、赤池信息量準則等歸納出最相關的表現參數。**方法：**本研究在不做任何假設的前提下，由網路爬蟲程式完整蒐集中華職棒大聯盟官網的 1990 到 2016 年每支球隊的各項數據，從 6,870 場比賽，714,480 筆資料中歸納出中職各隊各年度之 29 項攻守數據，並藉由關聯性分析、線性回歸模型進行解釋性分析，剖析球隊對戰、選手表現數據包括球隊的勝率、安打數、全壘打數、保送數、三振數、爲哪一隊等，對於各隊年度票房之影響，再以大數據自動建模方式將上述數據結合球隊因素，與中華職棒從創始到 2016 年的 27 年間的各隊票房做相關性分析。**結果：**研究結果發現全壘打、盜壘數、打者被三振頻率、失誤數、投手不被全壘打的能力、是否爲兄弟象及是否爲中信鯨爲對球迷進場票房影響最大的幾個因素，選出模型的決定係數 (R2) 爲 0.76，解釋力十分優秀，模型內變數之 p 值全數小於 .05。**結論：**球團可依此研究做爲挑選新秀之方針之一，而球員也可以依研究結果加強訓練重點，所有進場球迷最想看到的是拚勁與專注度，所以失誤對票房有最大的負面影響。而全壘打的激情、將球打進場內讓球迷看到精采攻防、投手壓制對手的能力則是對票房有最大的正面幫助。

關鍵詞：棒球、數據棒球、職業棒球票房、運動大數據

壹、緒 論

中華職棒大聯盟爲臺灣存在時間最悠久、觀眾數量也最多的職業聯賽，而在這近三十年的歷史中，各隊的戰績、票房隨著每年球員、教練組合的不同而有所起落。在比賽的現場，有眾多元素能讓現場觀眾充滿激情，包括三振、盜壘、全壘打、投打對決等，然而這些球員的技術面表現，是否眞正能影響比賽票房，亦或對戰組合才是影響票房的最大因素？本研究目的在於以大數據分析球員技術面表現及對戰組合對票房的影響。團隊藉由網路爬蟲程式完整蒐集中華職棒大聯盟官網的 1990 到 2016

通訊作者：許懷中 逢甲大學資訊工程學系 TEL: 04-22213108
407 臺中市西屯區文華路 100 號 E-mail: jhhuang@sportsnews.org.tw

年每支球隊的各項數據，從 6,870 場比賽，714,480 筆資料中歸納出中職各隊各年度之 29 項攻守數據，並藉由關聯性分析、多重線性回歸模型進行解釋性分析，剖析球隊對戰、選手表現數據包括球隊的勝率、安打數、全壘打數、保送數、三振數等，對於各隊年度票房之影響，可用做未來球團行銷以及球隊建構，甚至選手與球團談薪的參考。

職業運動的觀眾與球迷中永遠有爲數不少的戰績迷，而球星與戰績正是吸引更多球迷的不二法門。根據 Baseball Almanac (2017) 的資料，2010 年球季，堪薩斯皇家隊拿到美聯中區墊底的 67 勝 95 敗，主場進場觀眾人數平均爲 19,942 人；2015 年球季，他們以 95 勝 67 敗拿下美聯中區例行賽龍頭寶座，該年的主場觀眾人數提升到 33,439 人，最後也順利拿下了該年總冠軍。勝場數對進場觀眾人數及收視時數的影響，關係到球隊的管理階層願意付出多少代價在自由球員市場上，而每一勝的價值也不盡相同。Gennaro (2013) 調查，在美職大聯盟，80 勝時每一勝的價值從最低的海盜隊五十萬到最高的紐約大都會隊一百四十萬美金不等，接近分區冠軍的 90 勝時，每勝更飆漲到一百六十萬到四百三十萬美金。而除了勝場數外，球星與精采的球賽內容也是吸引球迷進場的因素，ESPN (2017) 的數據顯示，克利夫蘭騎士隊在 Lebron James 加入前一年的場均進場觀眾數是全聯盟最低的 11,496 人，他加入後馬上躍升至聯盟第 9 名，18,287 人，然而國內尚無針對中職每勝價值之研究。

國外有相當多對票房影響的相關研究，多數將重點放在整季勝率或是聯盟競爭平衡對當季票房的影響。Demmert (1973) 以及 Noll (1974) 於早年研究了票價與觀眾人數之間的關聯，Schmidt 與 Berri (2001) 用縱橫資料集 (panel data set) 研究，發現聯盟中的隊伍平衡競爭性對觀眾數量是有影響的。Davis (2008) 利用向量自回歸模型 (vector autoregression mode, VAR Model) 以及脈衝響應函數 (impulse response function)，分析大聯盟球隊戰績與票房數據，說明球隊戰績與票房之間的高相關性的因果關係爲戰績帶動票房，而非互有因果或者由於其他原因。除了勝率之外，Cebula (2013) 亦納入了可能影響票房的因素，包括主隊失誤次數及主隊得分，他採用固定效應縱橫最小平方模型 (fixed-effects panel least square) 研究上述數據與單場票房之間的關聯，發現均爲顯著相關，但因單場票房變因較多，其校正後判定係數僅有 0.43。Ahn 與 Lee (2014) 用 panel factor model 計算 1904 年到 2012 年百餘年間的大聯盟觀眾數，發現在 1904 年到 1957 年間，勝率是影響觀眾數量的惟一的顯著因子；然而從 1958 年到 2012 年間，除了隊伍的勝率之外，比賽結果的不確定性、球場品質、打擊能力，都成爲影響球迷入場意願的顯著因子，Lim (2017) 研究發現比賽結果的不確定性、季後賽的機率、票價、場館容量與票房有顯著性相關。以上的研究都是以整季的票房爲應變數。值得注意的是 1. 所有的研究都得到勝率對票房有正面顯著影響的結論，2. 幾乎所有研究聯盟球隊競爭平衡、比賽結果不確定性對票房影響的研究都得到其結果不確定票房有正面顯著影響的結論，而 Knowles 等 (1992) 研究發現最佳的票房落在主場球隊獲勝的機率爲 60% 時 (1992)。針對國內職業棒球的研究，王忠茂 (2005) 曾針對 2003 至 2004 的中職票房及可能原因進行分析。王志源 (2007) 研究中華職棒觀眾再購意願之影響因素，發現體驗行銷要素、體驗價值、涉入程度與中華職棒觀眾之再購意願有顯著相關。莊忠柱、陳天賜與姚爲守 (2004) 以表面似乎無關之迴歸模型計算中職 11 年至 13 年間的 540 場例行賽觀眾數資料，發現是否爲假日，及球隊本身與主場觀眾人數呈顯著相關，有否轉播與上一場得分差與進場人數無顯著相關，球場交通、當季勝率、上一季排名與降雨量則對各球隊的單場進場票房影響不一致。施致平、黃蕙娟與倪瑛蓮 (2010) 研究預測中職棒球賽勝負模式，計算 22 項攻擊指數、32 項投球指數、及 7 項防守指數，

與本研究使用的自變數最爲接近，不過其應變數爲球賽；其發現與勝負相關性最高的爲得分、打數、打點、失誤與強迫取分。廖主民、黃鈐雯與何鈺雯 (2008) 檢驗 295 位有固定單一支持球隊的球迷，分析其支持球隊的因素，發現對忠誠度預測力有顯著貢獻的只有團隊與形象因素，而與熟識與經營因素、親友與戰績因素、地利與主場因素無顯著相關。蔡銘仁 (2014) 在職棒觀眾涉入度與觀賽行爲之關聯性研究中則提到，觀眾對球賽涉入程度最高的兩個影響因素分別爲支持的球隊有好表現，及支持的球員有好表現時。陳成業 (2015) 探討中職運動賽會場館氣氛、個體情緒狀態與再購意願的結構路徑關係，發現中華職棒賽會場館氣氛由「球員表現」、「球迷熱情」、「硬體設施」、「現場活動」與「對戰之可看性」等因素所形成；中華職棒賽會場館氣氛正向預測現場觀眾情緒，而現場觀眾情緒正向預測其再購意願。陳成業 (2016) 亦檢視分析中職賽會服務品質的量表，發現賽前「球場內硬體條件」、「聲光效果」、「售票便利性」、「服務人員態度」、「開場活動」，球賽中「攻守交替的串場活動」、「場館清潔」、「球員表現」、「球賽氣氛」，球賽結束後「賽後活動」與「散場動線規劃」的品質影響最爲顯著。

在眾多相關研究中，Ahn 與 Lee (2014)，及 Regan (2012) 的研究除了勝率之外，加入了可能影響票房的選手表現因素，例如 Ahn 與 Lee 計算了長打率與三振數對票房影響，分析結果爲長打率對票房有正面顯著影響，三振數則無顯著影響。Regan 則採用多元線性迴歸分析計算了主隊失誤次數、盜壘、被保送，與場均全壘打數對票房的影響，發現隊失誤次數與被保送次數都對票房有負面影響，且達顯著水準，全壘打對票房有顯著正面影響，作者認爲觀眾並不喜歡刻意等待保送的打球方式，此偏好也反映在票房上；盜壘的影響雖接近但未達顯著水準。主隊被保送對票房有負面影響，與一般直覺認知並不相同，因被保送代表了上壘與接續得分的可能性，也因此本研究以這兩篇論文爲基礎，再發展到由所有能公開查詢到的球隊攻守數據來建模，希望能聚焦於攻守數據，瞭解對中華職棒的觀眾來說，哪一種打球風格是觀眾喜歡的；而因爲整季的票房較能表示出觀眾喜好的趨勢，因此與上述兩篇論文相同，我們以整季票房來做相關性研究。本研究的貢獻在於，過去的研究往往是由研究者以其主觀之領域知識挑選自變數，而本研究則完全以資料驅動 (data-driven) 之選模方法，藉由關聯性分析、多重回歸、自動選模技術挑選用已預測中職各隊年度票房之自變數，由此避免在挑選變數時代入研究者主觀之偏誤 (bias) 的狀況。

本研究試圖瞭解中職選手場上的表現對整季票房之影響，並且以聯盟的平均數做正常化，以進行跨年度、跨隊的比較性分析。研究數據包括了 1990 年到 2016 年中華職棒大聯盟的每一場比賽的票房，以及所有比賽的攻守相關數據，以大數據分析整隊的全壘打、失誤、打擊率對票房的影響。

貳、方　法

一、中華職棒歷年票房分析

圖 1 爲中華職棒歷年季賽單場平均票房的走勢變化，總和而言，中華職棒的票放經歷了初期成長、停滯、衰退、復甦、再次衰退、二次復甦等幾個階段；中職的票房在其創立的前三年快速地成長，而後由於新球隊加入以及新鮮感的流失等種種因素，場均票房陷入停滯，而 1995 年底以及 1997 年中所發生的黑虎以及黑鷹等職棒簽賭放水事件，更是重挫中職的票房，帶來了職棒票房的第一個黑暗期，爾後以陳致遠爲首的國手級球星率領兄弟象於 2001 至 2003 締造史上第一次的二度三連霸霸業，爲中職帶來了復甦的希望，然而從 2005 至 2009 期間連續發生的黑熊、黑鯨、黑米以及黑象事件，不僅再次重挫中職的票房、造成多支球隊解

散，同時也讓無數職棒球星失去舞台，陷入官司纏身的窘境。

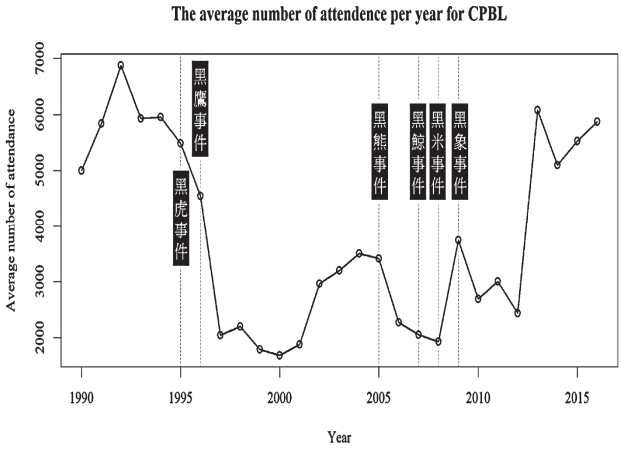

圖 1. **中華職棒歷年季賽單場平均票房變化**

由上述中職歷史大事紀與其平均票房走勢的對照，不難得知每次簽賭案對於中職年度票房都帶來巨大的負面影響。研究團隊對聯盟的票房層面進一步球隊票房的細部探討，如圖 2，進一步將曾經於中職活躍過的各球隊 (L：統一獅、D：味全龍、B：兄弟象|中信兄弟、T：三商虎、E：時報鷹、G：俊國熊|興農牛|義大犀牛|富邦悍將、W：和信鯨|中信鯨、M：第一金剛|La New 熊|Lamigo 桃猿、C：誠泰太陽|誠泰 Cobras|米迪亞暴龍) 之歷年單場平均票房做分析。

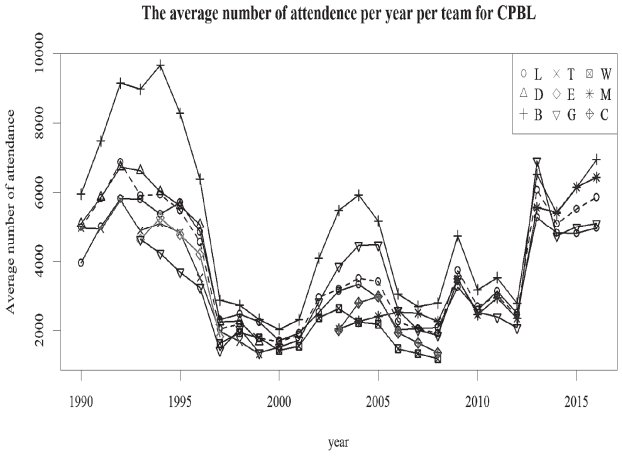

圖 2. **中職歷年各隊歷年季賽平均票房走勢**

由圖 2 可知，歷史事件如職棒簽賭放水案對於票房的影響是全體性的，既使是在我國職棒發展前二十年期間堪稱票房保證的兄弟象，其票房走勢也無法擺脫簽賭放水案的影響，而本研究試圖探討球隊因素、比賽以及球員技術對於各隊年度票房之影響，因此我們將各個隊伍的逐年票房以聯盟年均票房爲基礎進行正常化 (normalization)，正常化的方式爲將各隊各年度的單場平均票房與聯盟各年度的單場平均票房相除而得到一個調整後的年均票房指標 (adjusted average attendance)，其結果如圖 3 所示。

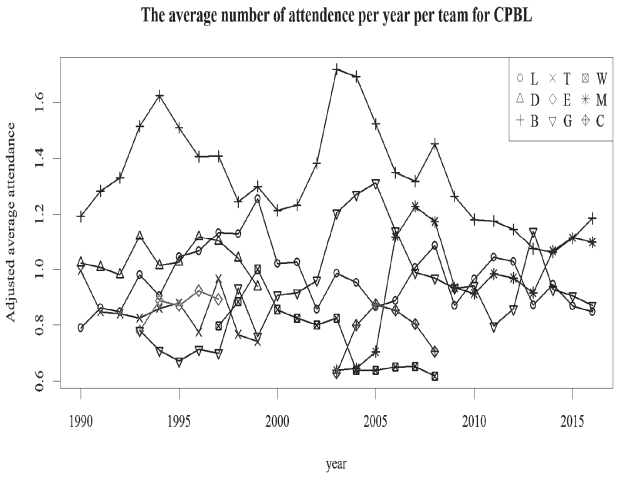

圖 3. **中職各隊歷年調整後之季賽單場平均票房數據**

圖 3 的 Y 軸單位爲倍數，也就是將聯盟票房正常化後，呈現出的是該隊該年度的票房是聯盟平均進場人數的倍數，例如兄弟象在 2003 年的票房約爲聯盟平均的 1.7 倍。從圖 3 可以發現，調整後的年均票房指標已經大致擺脫了簽賭放水案的因素，既使是在深受簽賭案影響的職棒黑暗期 (1995 至 2000)，各隊的票房仍然互有起落，例如 1995 兄弟象的票房下滑，而統一獅的票房上升，就明顯地是由於統一獅在戰績上壓倒兄弟象獲得總冠軍所致，由於職棒簽賭案所造成之票房全體性的下滑在正常化調整後已不復見。

因此，藉由正常化調整之後的票房指標，彌平了時間與大環境的所造成的影響，方便觀察其他影響票房的因素，有利於分析影響票房的隊伍以及賽事因素；進一步分析圖 3，可以看出兩個明顯影響票房的因素，其一是隊伍，如兄弟象的票房到無心經營轉賣之前，一直都是一枝獨秀；其二則是戰績對於票房的影響，如兄弟象在 1992 至 1994、2001 至 2003 兩次三連霸對於其票房帶來了巨大的提升，統一獅於 1995 開啓的首次連霸也對其票房有著顯著影響，此外興農牛於 2004 至 2005 的連霸以及 La New 熊於 2006 的隊史首冠，也都爲各自的票房帶來了成長；然而，並非每一次的總冠軍都能爲球隊帶來票房的效益，如味全龍隊於 1997 至 1999 的三連霸期間，其票房不斷地下滑，同一期間統一獅的票房卻始終維持高檔，此間緣由或許來自上述票房統計均來自季賽的票房數字，而味全龍在 1998、1999 這兩年都是以第三名之姿藉由挑戰賽制獲得總冠軍，統一獅這幾年的則始終處於冠軍候選的前二名球隊之列，而由於挑戰賽制，最終味全龍雖然由於挑戰賽制獲得冠軍，然而其球季中平凡的戰績，卻使得其票房與其最終獲得總冠軍的結果並不相稱。

觀察中職歷年的場均票房變化已經可以獲得許多影響職棒票房因素的線索，然而此粗淺的分析仍有許多不確定的因素存在，下一節中，將進一步利用統計迴歸分析的方法，解析影響職棒票房的賽事因素。

二、賽事數據與中職票房

爲了進一步研究比賽因素是否對於中華職棒票房產生影響，研究團隊撰寫了網路爬蟲程式，從中華職棒官網 (中華職棒大聯盟全球資訊網，2017) 取得了中華職棒歷年各隊逐場的攻守數據，所蒐集的數據條列如下：

表 1

中華職棒賽事原始數據概觀

資料來源	中華職棒官方網站
資料時間	1990-2016
資料列數	6,870 場
資料項目	比賽時間 (年、月、日、星期)、比賽球場、隊戰隊伍名稱、是否為延長賽、主客隊得分數、安打數、失誤數、打數、打點、二壘打、三壘打、全壘打、雙殺打、打者被保送、打者被觸身、打者被三振、高飛犧牲打、犧牲短打、盜壘、盜壘被阻殺、投手投球局數、面對打者數、投球數、好球數、投手被安打、投手被全壘打、投手保送、投手觸身球、投手三振、暴投、投手犯規、投手失分、投手自責分、各局得分、主審與各壘審姓名、比賽時間、觀眾入場數等
總和資料數	714,480 筆資料 (資料列數*資料項目數)

依據表 1 中的原始資料彙整與計算出各隊各年度全年總合與平均的攻守表現得到包含勝率 (w.prn)、打數總和 (ab)、打者每打數被保送數 (bb)、打者每打數被三振數 (k)、場均得分 (run)、場均失誤數 (err)、場均對手失誤數 (err.opp)、打擊率 (ba)、HR/AB (hr2ab)、打者之四壞三振比 (bb2k.b)、場均滾地雙殺數 (gidp)、場均盜壘數 (sb)、場均盜壘被阻殺數 (cs)、場均犧牲打數 (sac)、每打數額外壘打數 (xbh)、長打率 (slg)、純長打率 (isop)、上壘率 (obp)、整體攻擊指數 (ops)、投手九局四壞數 (bb9)、投手九局奪三振數 (k9)、投手奪三振四壞比 (k2bb.p)、投手防禦率 (era)、投手被打擊率 (oba)、投手被長打率 (oslg)、投手被上壘率 (oobp)、投手九局被全壘打數 (ohr9)、聯盟校正後整理攻擊指數 (OPS+, ops.plus)、聯盟校正後投手防禦率 (ERA+, era.plus)、隊伍名稱 (team) 等，上述數據的定義與計算方式可以參考 (棒球統計，2013)，彙整後爲中職各隊各年度共 29 項攻守數據、135 列資料。

此外，攻守數據有其年度趨勢，如 2015 年球季是打擊年，其打擊數據爲十年之最，而職棒初期有投高打低的傾向，因此，我們亦將各隊各年度平均攻守數據正常化 (normalization) －除以聯盟年度平均數據，針對年度差異進行修正，以利不同年度間的比較。

初步計算正常化後的 29 項攻守數據與調整後年度平均票房指標之間的皮爾森相關係數 (Pearson correlation coefficient)，如下表：

表 2

攻守數據與調整後年度平均票房指標之相關係數

數據項目	相關係數	數據項目	相關係數	數據項目	相關係數
勝率 (w.prn)	0.485	場均盜壘數 (sb)	0.280	打者每打數被保送數 (bb)	-0.009
打者之四壞三振比 (bb2k.b)	0.471	投手奪三振四壞比 (k2bb.p)	0.262	投手九局被全壘打數 (ohr9)	-0.086
打擊率 (ba)	0.461	場均滾地雙殺數 (gidp)	0.202	投手九局四壞數 (bb9)	-0.180
上壘率 (obp)	0.423	投手九局奪三振數 (k9)	0.185	投手被打擊率 (oba)	-0.232
場均得分 (run)	0.369	場均犧牲打數 (sac)	0.155	投手被長打率 (oslg)	-0.240
聯盟校正後整理攻擊指數 (OPS+, ops.plus)	0.367	每打數額外壘打數 (xbh)	0.110	投手被上壘率 (oobp)	-0.265
整體攻擊指數 (ops)	0.364	打數總和 (ab)	0.103	投手防禦率 (era)	-0.296
場均對手失誤數 (err.opp)	0.323	場均盜壘被阻殺數 (cs)	0.096	場均失誤數 (err)	-0.469
長打率 (slg)	0.295	純長打率 (isop)	0.082	打者每打數被三振數 (k)	-0.606
聯盟校正後投手防禦率 (ERA+, era.plus)	0.283	HR/AB (hr2ab)	0.02		

由上表可知，29 項攻守數據各別而言，對於票房影響最大的爲相關係數 -0.606 的打者每打數被三振數 (負向相關)，其次爲勝率的 0.485、打者之四壞三振比的 0.471 等等，大部分攻守項目與票房的相關性大致在中度至低度相關不等，由此可知，單一數據並不足以準確地預測票房，而上述數據同時包含了基礎數據 (如打數、安打數等) 以及由基礎數據計算而來的進階數據 (如打擊率、上壘率、整體攻擊指數等)，變數彼此之間可能存在依賴關係，例如各項打擊數據對於勝率的影響，也需要進一步地研究與討論。

因此，我們進一步以各隊各年度校正後之票房爲應變數，以各隊各年度之平均攻守表現爲自變數，訓練多重回歸模型 (multiple regression model) 希望藉此了解球隊之攻守表現如何影響其票房。

而爲了避免多重回歸時發生由於變數共線而影響回歸結果的情形，在進行多重回歸建模以前，我們先以統計教科書 The basic practice of statistics (Moore, 2013) 中所描述高度相關的標準 (皮爾森相關係數之絕對值大於 0.7)，針對 29 項攻守數據進行相關性分析，將彼此高度相關之變數歸納成群，在此過程中，我們使用 R 語言套件 caret 中的 findCorrelation 函式 (Kuhn, 2008)，在各群彼此高度相關之變數中，計算各變數與全部其他變數之相關係數，最終留下該群變數中，與其他所有變數相關係數最低者作爲代表數據進行建模，相關性分析的結果如下表：

表 3

高度相關變數與其代表數據之挑選

高度相關變數	代表數據
OPS+ (0.524)、整體攻擊指數 (0.523)、勝率 (0.512)、長打率 (0.484)、上壘率 (0.490)、平均得分 (0.482)、打擊率 (0.447)	*打擊率*
投手防禦率 (0.459)、投手被上壘率 (0.456)、ERA+ (0.452)、投手被長打率 (0.449)、投手被打擊率 (0.390)	*投手被打擊率*
每打數額外壘打數 (0.418)、純長打率 (0.403)、HR/AB (0.346)	*HR/AB*
投手奪三振四壞比 (0.359)、投手九局四壞數 (0.295)	*投手九局四壞數*

註：括號內為該變數與其他變數之相關係數絕對值平均

以表 3 第一列爲例，一支球隊單一年度其聯盟校正後整體攻擊指數 (OPS+)、整體攻擊指數 (OPS)、勝率、長打率、上壘率、平均得分、打擊率等數據彼此之間的相關係數超過 0.7 爲高度相關之數據群，而其中打擊率與全部其他的自變數之平均相關係數 (括號內之數據) 最小，因此被保留作爲這群彼此高度相關之變數的代表數據；依此類推，投手被打擊率、HR/AB、投手每九局四壞數，也分別被保留下來代表各自的一群彼此高度相關的數據。

經過相關性分析之後，原本的 29 項攻守數據，留下了 16 項彼此相關性較低之數據：打數總和 (ab)、打者每打數被保送數 (bb)、打者每打數被三振數 (k)、場均失誤數 (err)、場均對手失誤數 (err.opp)、打擊率 (ba)、HR/AB (hr2ab)、打者之四壞三振比 (bb2k.b)、場均滾地雙殺數 (gidp)、場均盜壘數 (sb)、場均盜壘被阻殺數 (cs)、場均犧牲打數 (sac)、投手九局四壞數 (bb9)、投手九局奪三振數 (k9)、投手被打擊率 (oba)、投手九局被全壘打數 (ohr9) 進行選模，此外，由於反應變數－調整後各隊的場均票房並未針對各個隊伍的場均入場數進行正常化，因此模型中也納入了由隊伍所構成之虛擬變項 (dummy variable) 以反映由於隊伍經營、或是號召力所造成票房的差異。

在進一步選模的方法上，研究團隊採用日本統計學家赤池弘次所發展的赤池信息量準則 (Akaike information criterion, AIC)，當模型的變數增加，雖然會提升模型的擬合度 (fitness)，但也同時可能增加過度擬合 (overfitting) 的風險；AIC 假設模型的誤差符合常態分布，以信息熵爲基礎，統計模型的複雜度與其擬合的優良性 (goodness of fit)，以求獲得在複雜度與過度擬合風險間具有良好平衡的模型 (王志源，2007；王忠茂，2005)。

在本研究中，以各隊各年度的 16 項攻守數據，加上代表資料所來自隊伍的虛擬變項 (dummy variable) 即是否爲兄弟象、是否爲統一獅等 9 項，共 24 項變數進行建模，在此我們使用 R 語言套件 MASS 中的 stepAIC 函式 (Ripley, et al., 2013)，以預設參數估算赤池信息量準則，進行選模，其算式如下：

$$\mathrm{AIC} = 2k + n\ln\left(\frac{\mathrm{RSS}}{n}\right)$$

其中，k 是模型的參數總數、RSS 是殘差平方和、n 爲觀察數，參數數量的增加會提高了擬合的優良性，然而也將提升過度擬合的風險，採用 AIC 的目的在於鼓勵數據良好擬合的同時，減少所使用的參數數量，盡量避免過度擬合的發生，也就是要建立能夠最好地解釋數據同時包含最少自由參數的模型 (王志源，2007；王忠茂，2005)，我們使用 R 語言套件 MASS 中的 stepAIC 函式 (B. Ripley, etc. 2013)，最終獲得一個具備最低 AIC，也就是符合赤池信息量準則的選模結果，細節將於下一章節中描述。

參、結　果

爲了讓最終模型的係數可以互相比較，研究團隊計算了 16 項攻守變數的標準分數 (standard score, Z-score) 進行標準化 (standardization) 後，進行 AIC 計算選模，選模結果如下表：

表 4

AIC 選模後模型

	估計係數值	p-value	顯著程度
截距	0.94696	< .001	***
打者每打數被保送數 (bb)	0.13328	.065	.
打者每打數被三振數 (k)	-0.16529	.013	*
場均失誤數 (err)	-0.03713	< .001	***
HR/AB (hr2ab)	0.03829	< .001	***
打者之四壞三振比 (bb2k.b)	-0.13246	.111	
盜壘數 (sb)	0.03401	.001	**
投手九局被全壘打數 (hr9)	-0.02413	.037	*
是否為兄弟象	0.34299	< .001	***
是否為中信鯨	-0.13517	< .001	***
是否為誠泰 Cobras	-0.08010	.097	.
Residual standard error: 0.1115 on 127 degrees of freedom			
Multiple R-squared: 0.7905, Adjusted R-squared: 0.7736			
F-statistic: 62.89 on 7 and 127 DF, p-value: < 0.001			

由表 4 可知，以 AIC 選模後，模型中仍有部分不顯著之變數，進一步以後選取法 (backward elimination)，逐次移去模型中最不顯著的特徵，直到模型中所有特徵均爲顯著，此舉雖會降低擬合之良好度，然而最終模型中所有變數均爲顯著，可以確保模型之可解釋性，最終選出來的模型爲：

中職各隊該年票房/中職該年平均票房=

0.94367 - 0.05544*k - 0.03963*err + 0.04072*hr2ab + 0.03746*sb - 0.02871*hr9 + 0.34336*is.B - 0.13904*is.W

回歸模型的細節則如下表所示：

表 5

以隊伍年度攻守數據預測與中職年度調整後平均票房之多重回歸模型

	估計係數值	p-value	顯著程度
截距	0.94367	< .001	***
打者每打數被三振數 (k)	-0.05544	< .001	***
場均失誤數 (err)	-0.03963	< .001	***
HR/AB (hr2ab)	0.04072	< .001	***
盜壘數 (sb)	0.03746	< .001	***
投手九局被全壘打數 (hr9)	-0.02871	.0142	*
是否為兄弟象	0.34336	< .001	***
是否為中信鯨	-0.13904	< .001	***

Residual standard error: 0.1115 on 127 degrees of freedom
Multiple R-squared: 0.7761, Adjusted R-squared: 0.7638
F-statistic: 62.89 on 7 and 127 DF, p-value: < 0.001

此模型的輸入參數爲各職棒球隊各年度之攻守數據共有 135 筆，用以預測各球隊各年度調整後之場均票房指標，此模型的決定係數 (R^2) 到達 0.76、預測值與實際值之間的相關係數達 0.881，預測能力十分優秀，模型內變數之 p 值全數小於 .05，表示各變數對於模型的貢獻均有相當顯著性，圖 4 模型所預測之各隊校正後年度票房與眞實各隊校正後年度票房之比較圖 4 展示上述模型對於各隊年度校正後票房的預測與實際值的差異，其中灰實線代表預測與實際值完全相符，而灰色虛線則是預測值與實際值擬合的傾向，由此圖可知，此線性回歸模型做出了相當準確的預測。

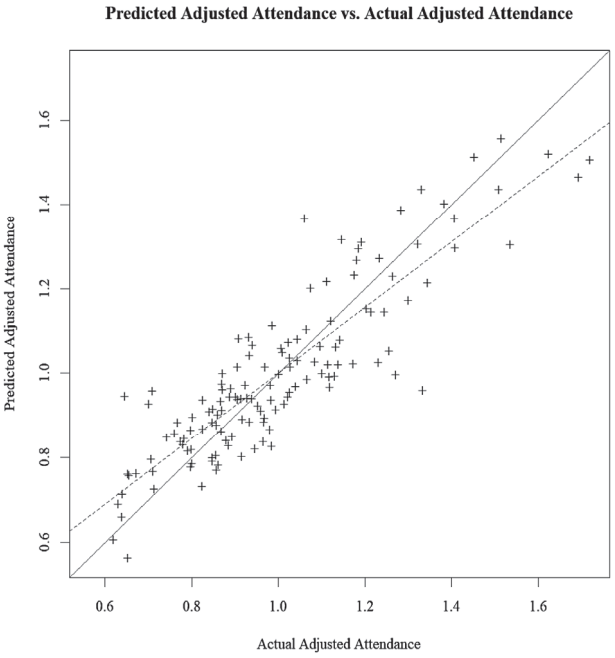

圖 4. **模型所預測之各隊校正後年度票房與真實各隊校正後年度票房之比較**

由於建模前已事先將變數標準化，我們可以藉由比較模型中各變數之係數瞭解變數對於模型的影響力，由此模型可知，兄弟象爲中華職棒長年的票房保證，對於票房的影響力 0.34336 超出其他各項變數甚多，假如不考慮其他變因，兄弟象的調整後年均票房指標約爲 1.286 (0.34336+0.94367)，表示其平均而言較聯盟平均票房多了接近三成;而中信鯨 (不包含中信兄弟) 則因經營上的失敗，與之對戰對票房都有負面的影響。

而所有賽事數據中影響觀眾進場意願最大爲打者被三振頻率，此處以每打數平均被三振的次數呈現 (K/AB)，此係數爲負值，表示球隊打者容易被三振會降低觀眾進場意願，顯見觀眾喜歡選球良好、不容易被三振的選手，因爲球打進場內就有希望，至少有看到精采攻防；其次的影響因素爲全壘打頻率 HR/AB，代表每平均每打數能打幾支全壘打，此係數對於票房的影響爲正，表示觀眾喜歡自己所支持的球隊擁有強大的打擊火力；而球隊場均的失誤數也會影響觀眾進場的意願，表示球迷們不喜歡看

到失誤頻頻的比賽，對球迷來說，這可能代表球員的不專注；此外，盜壘數高、投手抑制對手長打的能力 (九局被全壘打數 hr9 低) 等，對於票房都會有正面的影響。

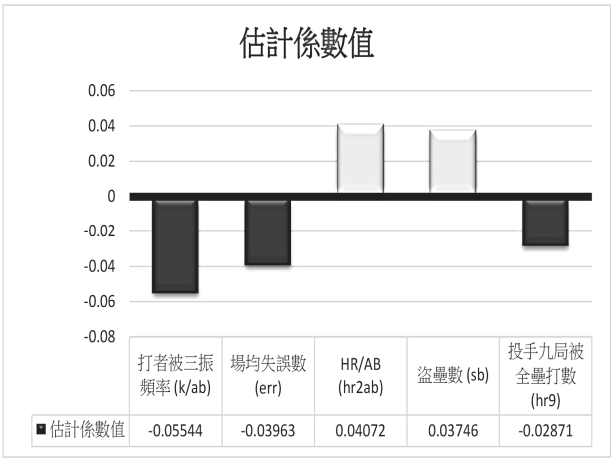

圖 5. **以圖表顯示的各項攻守數值數重要性**

肆、討　論

本論文藉由分析中華職棒歷年票房與球隊攻守數據之間的關係，剖析臺灣球迷進場觀賞職業賽事的偏好，由初步的分析可知，臺灣的棒球迷對於放水打假球深惡痛絕，每次假球案爆發都對票房產生巨大的傷害；而在放水打假球的原因以外，臺灣球迷對於兄弟象的支持十分熱烈，相較之下，中信鯨在其存在期間，一直都無法有效地號召球迷進場看球；由兄弟象與中信鯨的變數，我們可看出球員在球場上的表現固然對票房有不小的影響，但球隊的品牌與球迷經營更是票房保證，也是球團經營團隊應該重視的面相。在比賽數據方面，打者被三振的次數以及長打火力對於場均觀眾的影響極爲顯著，而野手優異的守備、投手抑制對手長打的能力以及有精彩盜壘表演，最能吸引觀眾進場看球。

勝率在相關性分析中看來是影響票房的首要因素之一，在許多其他研究中勝率也經常被選爲影響球迷進場的當然因素，然而由於有許多攻守數據都會影響球隊年度勝率的高低，因此在分析變數之間彼此相依性的分析過程中，勝率與攻擊數據之間的高度相依性被凸顯，而在自動選模過程中由打擊率代表，而打擊率則在 AIC 的選模過程中，由於有其他變數可以對預測的準確性做出更大貢獻而被割捨，而最後所留下的攻擊數據則是在相關性分析中看起來與票房較無直接關聯性的每打數全壘打數 (HR/AB)，大數據研究透過統計以客觀的自動選模方式找出吸引球球進場的相關數據，由數據解讀，失誤對票房帶來的負面影響，幾乎等同全壘打頻率的正面影響。對戰組合包含了哪一隊，更是對票房高低決定性的因素。球團的教練、球探及經營團隊在選材及訓練時可將這些因素列入考慮，選擇手眼協調性佳、三振率低、長打率高的打者，投手則以三振型投手對票房較有助益。失誤則是需由練習來降低發生機率，可讓選手瞭解失誤對票房的影響，再依失誤數訂定罰則，以期降低失誤數。盜壘除了是推進壘包的手段，也是拉高票房的因素，各球團可依本研究爲據，多發動盜壘戰術。

本研究以年度爲單位解析了中華職棒球員技術面影響球迷進場的影響因素，同時使用純資料驅動之方法，在過程中，完全依據資料之間的關係以及資訊理論如關聯性分析、赤池信息量準則等來揀選模型，主要是希望在球界專家、球迷、球團經營者的主觀想法以外，提供一個純然客觀的研究方法來探討在大環境以外影響職業球賽票房的因素，也希望球團體認到場上表現對票房影響的重要性。研究以資料驅動方式，歸納出以七項包含攻守數據、球隊品牌因素之數據預測中職球隊年度修正後票房的模型，此模型之決定係數達 0.76，表示這些變數確實相當程度的解釋了中職票房的影響要素；然而，決定係數 0.76 也表示仍有部分因素，受到資料收集的限制，未能納入模型中，例如球隊的經營手法、行銷活動以及電視轉播等；Ahn 與 Lee (2014) 的研究中，長打率對票房有正面顯著影響，三振數則無顯著影響。Regan (2012) 的研究結論是失誤次數、被保送次數及

全壘打對票房有顯著影響。針對中職觀眾，我們的研究結果顯示被三振數、失誤、全壘打、盜壘、投手不被全壘打能力對票房有顯著影響，與上述兩個研究不盡相同，其中原因除了研究方法的不同外，研究對象觀眾群也並不相同，因此導致不同的結果。

除了以年度平均票房的觀點來探討影響票房之因素，未來我們也期許可以進一步縮小研究的粒度 (granularity)，從單場的觀點來探討影響票房的因素，如此，一些在年度研究中屬於週期因素而未納入的資料，如比賽時間與天氣勢必也將在模型中扮演重要的角色。與單場票房相關的因素更多、計算更複雜，因此未來我們將針對逐場票房進行預測，探討逐場票房與球員逐場攻守表現、隊伍戰績與首位勝差之間的關聯性，同時更進一步研究與非競賽因素如比賽時間 (如是否爲周末)、天氣 (如是否下雨)、比賽地點 (是否偏僻)、球風 (防守/進攻野球)，以及近年來對票房大有助益的主題日等之間的關聯性，而臺灣的環境與美國不同的是，夏天晚上氣溫、溼度仍高，有時看球並不舒適，我們也會試著計算溫度對票房的影響；此外，近年來，Lamigo 隊的全猿主場帶起一陣新的行銷與應援風潮，然而礙於以可取得公開數據有限，難以將其行銷的投入與票房的關係關聯量化。未來希望能找到好的方法，收集中職行銷投入大數據資料，以計算行銷、球員表現對球迷進場意願的綜合效果。

引用文獻

中華職棒大聯盟全球資訊網 (2017)。*中華職棒歷年各隊逐場攻守數據*，取自 http://www.cpbl.com.tw/

[Official website of Chinese professional baseball league (2017). *Statistics of each team and player of all years*. Retrieved from http://www.cpbl.com.tw/]

王志源 (2007)。體驗行銷要素：體驗價值與涉入程度對中華職棒觀眾再購意願之影響 (未出版碩士論文)。*國立臺灣師範大學*，臺北市。

[Wang, C. Y. (2007). *The effect of experiential marketing, experiential value, and involvement to the repurchase intention of audiences of the CPBL* (Unpublished master thesis). N*ational Taiwan Normal University, Taipei.*]

王忠茂 (2005)。2003-2004 年中華職棒大聯盟觀眾人數之分析。*中華體育季刊，19*(3)，53-60。

[Wong, J. M. (2005). The analysis of attendance of Chinese professional baseball league during 2003 to 2004. *Quarterly of Chinese Physical Education, 19*(3), 53-60.]

施致平、黃蕙娟、倪瑛蓮 (2010)。中華職棒比賽勝負預測模式之建構。*體育學報，43*(2)，115-130。

[Shih, C. P., Huang, H. C., & Ni, Y. L. (2010). Establishing models to predict the outcomes of baseball games in CPBL. *Physical Education Journal, 43*(2), 115-130.]

莊忠柱、陳天賜、姚爲守(2004)。職業棒球主場觀眾人數的影響因素之探討—以中華職棒聯盟爲例。*體育學報，37*，163-175。

[Chung, C. C., Chen,T. T., & Yao, W. S. (2004). Influence factors of the attendance numbers in home-field for professional baseball-case of CPBL. *Physical Education Journal, 37*, 163-175.]

蔡銘仁 (2014)。職棒觀眾涉入度與觀賽行爲之關聯性研究 (未出版碩士論文)。*中國文化大學*，臺北市。

[Tsai. M. J. (2014). *A study on the involvement and behavior of CPBL spectators* (Unpublished master thesis). *Master thesis, Chinese culture university, Taipei.]*

陳成業 (2015)。環境心理學觀點探討運動賽會場館氣氛。*體育學報，48*(4)，417-430。

[Chen. C. Y. (2015). A study on sport stadium atmosphere of chinese Professional Baseball League - The perspective of environmental psychology. *Physical Education Journal, 48*(4), 417-430.]

陳成業 (2016)。觀賞性運動賽會服務品質：量表發展與驗證。*體育學報，49*(2)，195-207。

[Chen. C. Y. (2016). Service quality in spectator sporting events: Scale development and validation. *Physical Education Journal, 49*(2), 195-207.]

棒球統計 (2013)。*各項棒球數據計算方法*。取自 http://twbsball.dils.tku.edu.tw/wiki/index.php/棒球統計

[Baseball statistics. (2013). *Methods for calculating different baseball statistics*. Retrieved from http://twbsball.dils.tku.edu.tw/wiki/index.php/%E6%A3%92%E7%90%83%E7%B5%B1%E8%A8%88]

廖主民、黃鈐雯、何鈺雯 (2008)。職棒球迷支持因素：戰績歸因對球迷忠誠度之預測。*體育學報，41*(3)，43-55。

[Liao, C. M., Huang, L. W., & Ho, Y. W. (2008). The prediction of baseball fans' reasons for supporting a team and attributions to team performance on fan loyalty. *Physical Education Journal, 41*(3), 43-55.]

Ahn, S. C., & Lee, Y. H. (2014). Major League Baseball attendance. *Journal of Sports Economics, 15*(5), 451-477. doi:10.1177/1527002514535171

Baseball Almanac, Inc. (2017). *Kansas city royals fan attendance data.* from http://www.baseball-almanac.com/teams/kcratte.shtml

Cebula, R. J. (2013). A panel data analysis of the impacts of regional economic factors, marketing and promotions, and team performance on minor league baseball attendance. *The Annals of Regional Science, 51*(3), 695-710.

Davis, M. C. (2008). The interaction between baseball attendance and winning percentage: A VAR analysis. *International Journal of Sport Finance, 3*(1), 58.

Demmert, H. G. (1973). *The economics of professional team sports*: Lexington, Mass: Lexington Books.

ESPN. (2017). *NBA games attendance*. from http://www.espn.com/nba/attendance

Gennaro, V. (2013). *Diamond dollars: The economics of winning in baseball*: Diamond Analytics.

Knowles, G., Sherony, K., & Haupert, M. (1992). The demand for major league baseball: A test of the uncertainty of outcome hypothesis. *The American Economist, 36*(2), 72-80.

Kuhn, M. (2008). Care package. *Journal of Statistical Software, 28*(5), 1-26.

Lim, N. (2017). *Examining professional baseball attendance determinants: A multilevel analysis of Major League Baseball (MLB) Seasons*. Indiana University.

Moore, D. S., Notz, W., & Fligner, M. A. (2013). The basic practice of statistics: WHFreeman.

Noll, R. (1974). *Attendance and price setting. government and the sports business.* RG Noll. Washington, DC. *The Brookings Institution*.

Regan, C. S. (2012). The price of efficiency: Examining the effects of payroll efficiency on Major League Baseball attendance. *Applied Economics Letters, 19*(11), 1007-1015.

Ripley, B., Venables, B., Bates, P. M., Hornik, K., Gebhard, A., Firth, D., & Ripley, M. B. (2013). Package'MASS'. CRAN Repos. Httpcran R-Proj. Orgwebpackage. MASSMASS Pdf.

Schmidt, M. B., & Berri, D. J. (2001). Competitive balance and attendance: The case of Major League Baseball. *Journal of Sports Economics, 2*(2), 145-167.

投稿日期：106 年 06 月

通過日期：106 年 12 月

Physical Education Journal
Society of Physical Education, Taipei, Taiwan
Vol. 50, No. S, pp. 79～90 (2017. 12.)

Using big data to analyze the correlation between baseball players' performance, the matching teams, and fans attendance of Chinese professional baseball league

Huai-Chung Hsu[1] and Chih-Hao Huang[2]
[1]Department of Information Engineering, Feng Chia University, Taichung City, Taiwan
[2]Department of Sports Information and Communication, National Taiwan University of Sport, Taichung City, Taiwan

Abstract

Introduction: The Chinese Professional Baseball League (CPBL) has been in operation for 28 years and, among all factors of player performance, including stealing bases, homeruns, strike outs, we want to find out what are the most important factors affecting fan attendance. **Method:** In this research, we use big data, auto model selection method to look for significant factors from all performance associated numbers, including winning percentage, hits, home runs, walks, strikeouts, and 23 other team statistics. We use a crawler program to collect data from the Chinese Professional Baseball League (CPBL) website and exam the teams' performance in more than six thousand games versus the fans' attendance of the games from 1990 to 2016, 27 years in total. We use auto model selection to find the best model and coefficients, and the most relevant factors to the fan attendance. **Result:** We found that errors, home runs, K/AB, steals, and pitchers' HR/AB are the most important factors that affect the attendance of CPBL. The R^2 of the selected model is 0.76, which is high correlated. All the p values in the model are less than .05. **Conclusion:** While errors is an indication to the fans that a player is not focused and thus drive the fans away, home runs, stealing bases, strike outs and the pitchers ability to prevent homeruns are all important factors to attract fans. The clubs can have these ideas in mind when they draft and trade players. And coaches can adjust the team's style of playing. Fans rather see the players they support play actively, stealing some bases, than wait for a walk passively.

Key words: baseball, sabermetrics, professional baseball attendance, sport big data