# 108 學年度專題報告競賽

## 題目:探討美國職棒與中華職棒的薪資與表現

The Case Study of Relationship between Salaries and Performance in Major League Baseball and Chinese Professional Baseball League


系所班別：國立臺北大學統計系

410578028 統計四 林偲晴

410578045 統計四 鄧揚耀

410578050 統計四 張芷庭

410578068 統計四 陳威傑

410474212 社學四 甘珮儒

指導教授: 蘇南誠 教授

## 摘要

本分析探討棒球場上薪資與表現，並且以美國職業棒球大聯盟與中華職業棒球大聯盟為例。利用1985-2016年美國職業棒球大聯盟的資料進行多元線性迴歸和主成分迴歸分析，並選取出重要變數，進行薪水預測模型。此薪水預測模型適用於美國職業棒球大聯盟與中華職業棒球大聯盟。

## 前言

棒球從日本殖民時代開始便是台灣的國球，就在去年 (2019)，代表台灣的中華隊在各層級的棒球國際賽取得佳績，更在爭取奧運資格的世界 12 強棒球賽取得第五名的優秀成績，這促使了我們想深入研究棒球這項運動。我們主要想了解球團薪資與球團和球員表現之間的關係，以及比較美國職棒和中華職棒可能存在的差異。

我們希望透過比較美國職棒和中華職棒的差異，以及分析薪資與球團表現間的影響，找出台灣棒球的缺陷與迷思並且將其改善。球員薪資與球員表現確實存在一定關係，球團訂定球員薪資合約時，顯然已將球員表現納入考量。然而，簽定薪資合約的時間點也是必須探討的重要因素之一。球團與球員洽談未來一年薪資的時間，通常為前一年球季結束後，球團並不能完全預測未來一年球員的表現，只能依照球員前一年的表現給予評估後的薪資；換句話說，球團簽定薪資合約是有風險的，故我們想透過研究試圖降低此風險。更希望透過這項研究使政府對台灣棒球政策有所改善，讓更多企業能夠支持及投資於這項運動，也能透過數據分析，讓球團簽訂薪資時更有依據，也更有保障。

本次選取最具代表性的美國職業棒球大聯盟與本國的中華職業棒球大聯盟為分析對象。由於中華職業棒球大聯盟球員薪資與團隊薪資為不開放資料，我們將利用 1985-2016 年美國職業棒球大聯盟的資料，透過多元線性迴歸(Multiple Linear Regression)以及主成分迴歸(Principal Component Regression)所得到的重要變數作為預測模型的基底，來預測職業棒球大聯盟球員薪資，並進行兩者比較。

# 文獻回顧

（一）薪資

於 Shorin (2017) 中提及，球團越高的薪資支出與球隊越高的勝率有關。四種運動（MLB、NBA、NHL、NFL）在例行賽中，由計量經濟分析顯示薪資花費越高的球隊在比賽上有著更好的表現。 而在季後賽中，僅有 NBA 和 NHL 達到統計上的顯著，然而若移除了團隊固定效果後，MBL 亦於球隊薪資和冠軍勝率具有顯著正向關係。

於 Wasserman (2013) 中，決定選手將獲得多少報酬的最大因素是他在運動場上的表現。一場比賽打得越好，他的薪水就越高。但是，還有許多其他因素通常會忽略影響工資收入的因素。除了球員的才華和生產水平，該項目還研究了許多其他因素，這些因素可能會影響他的薪水。研究使用線性回歸分析來隔離玩家薪水和可能與薪水有顯著關係的多種不同因素之間的關係。

於 Fields (2001) 中，研究了美國職棒大聯盟球員是否獲得邊際收益產品的報酬。採用 1990 年至 1999 年的數據進行模型配適。結果表明，職業棒球運動員的邊際收益被少付。 在 1994-95 年罷工後，球員的邊際產品報酬更高，而且年紀較大的球員似乎比年輕球員的邊際產品薪水更高。

（二）薪資不平等

於 Wiseman and Chatterjee (2003) 中提及，球隊內巨大的不平等對於運動的競爭平衡有所影響，且當球隊中個人的薪資越不平等時，球隊的勝場數越高。於 Hall, Szymanski et al. (2002) 中的數據顯示，在 1990 年代，球隊工資與績效之間的橫截面相關性顯著增加。

於 Jewell and Molina (2004) 中顯示，MLB 中的薪資隨著球隊內與球隊間的薪水不平等而升高。其中升高的不平等也可能影響單一球隊或整個聯盟的成功。並且發現，以一支球隊的獲勝百分比衡量，MLB 球隊內部的薪資分配確實對球隊的成功產生了顯著的負面影響。

又於 Mizak and Stair (2004) 中顯示於 MLB 內球隊間越大的薪資不平等，會導致越大的勝利不平等和越大的競爭不平等。亦發現自 1990 年代中期以來，尤其是在美國職棒的美國聯盟中，薪資差距越來越大，表明奢侈稅一直無效，並且在不久的將來可以預見更大的績效差距。

於 Cyrenne (2014) 中，研究了團隊薪水分配與其獲勝百分比之間的關係。發現相對薪資較高且薪資不平等程度較低的球隊獲勝率更高。還發現了超級巨星效應的證據，因為最高球員薪水更高的球隊獲勝率更高。但是，該結果對於所使用的工資不平等的具體度量以及工資分配的內生性很敏感。

（三）研究方法

於 Shorin (2017) 中提及不同的研究方法。經驗分析可平衡固定影響的迴歸模型以控制一些特權因素（例如：教練、管理、地點等）。而後續分析可以延伸不同的成功標準以了解薪資如何與其他結果相關聯，例如：排名、最終決戰勝率、團隊價值等等。然而，這個研究可能因專注於團隊層級的薪資支出，卻無法看到任何球員層級的淺在關係而產生限制。

（四）有效支出

於 Shorin (2017) 中指出，平均而言，增加美國職業棒球大聯盟，NBA 和 NHL 的球隊薪資支出是比較明智的，因為如果球隊取得更好的現場表現，可以抵消大部分支出，還可以為團隊如何最有效地花錢實現組織目標提供有價值的見解。

(五)模型準確性

於 Hoffman (2014) 中，對於 2010-2012 美國職業棒球大聯盟賽季，收集了幾種不同的生產統計數據。從每個賽季中隨機抽取一名球員樣本，並為位置球員和投手創建單獨的模型。為每個不同的模型選擇了有助於預測工資的重要生產統計數據。這些模型被認為是對每個不同模型的預測 r 平方值至少為 0.70 的良好模型。找到回歸模型後，通過預測 2013 MLB 賽季隨機樣本球員的薪水來測試模型的準確性。

而於 Magel and Hoffman (2015) 中，從 2010 年至 2012 年的每個賽季中，隨機抽取一名球員樣本。根據各種生產統計數據，開發了模型來預測薪水。為位置球員和投手創建了不同的模型。每個模型選擇了有助於預測工資的重要生產統計數據。為位置球員開發了兩個模型，為投手開發了兩個模型。每個組中的一個模型考慮了年度生產統計，另一個模型考慮了職業生產統計。考慮年度生產統計數據的模型可用於確定球員與其當年薪資相比是否表現不佳。這些模型不能用於預測年薪，因為提前知道年度生產統計信息。基於職業生產統計的兩個模型被認為是良好的預測模型，因為它們的預測 r 平方值至少為 0.68。通過預測 2013 MLB 賽季隨機樣本球員的薪水來測試開發的回歸模型的準確性。

## 資料介紹及前置處理

### ·資料介紹

資料集來源自 Lahman’s Baseball Database 與 CPBLSTAT。資料集中的團隊薪水，投手薪水或擊球手薪水總計為所有球員，投手或擊球手的薪水，以百萬美元為單位。工資數據是營業日工資。在一個賽季中球員可能受傷，被交易或被召喚而無法上場的情況下，一個賽季中不參加任何一場比賽的球員的薪

水不包括在球隊總工資中。此外，為消除通貨膨脹的影響，所有薪資變量均除以 CPI，然後轉換為它們的實際價值。

美國職棒大聯盟（Major League Baseball，簡稱：MLB，或大聯盟），是世界水準最高的職業棒球比賽，由國家聯盟和美國聯盟在 1902 年成立，與職業籃球（NBA）、美式足球（NFL）、冰上曲棍球（NHL）並稱北美四大職業體育競賽。目前美國職棒大聯盟共有三十支球隊，分屬兩聯盟。其中國家聯盟十五隊，美國聯盟十五隊。兩聯盟各分為三區(東區，中區，西區)，各分區的冠軍球隊及兩聯盟的外卡球隊均可參加季後賽，爭奪世界大賽冠軍。

中華職業棒球大聯盟（CPBL，英文全稱為 Chinese Professional Baseball League，簡稱中華職棒大聯盟或中華職棒）是於 2003 年由「中華職棒聯盟」與「台灣大聯盟」兩職業聯盟合併而改制。1989 年中華職棒聯盟成立，1990 年 03 月 17 日中華職棒開幕戰開打，正式宣告台灣進入職棒元年。1996 年台灣職業棒球大聯盟成立，1997 年 02 月 28 日台灣大聯盟開幕戰揭幕，開啟我國職棒進入兩聯盟並存競爭的新時代。2003 年 01 月 13 日因兩聯盟長期惡鬥，包含爭奪球員、相互挖角、黑球或票房等問題，導致雙邊聯盟呈現「黑暗期」。在時任總統陳水扁指示體委會協調之下，兩聯盟合併（實質為台灣大聯盟合成兩隊併入中華職棒），並定名為「中華職業棒球大聯盟」；同年 03 月 01 日，合併後的中華職棒大聯盟開幕戰登場。自此，台灣走向單一聯盟化，而兩聯盟並立的時代正式走入歷史。

目前有富邦悍將隊、樂天桃猿隊、中信兄弟隊、統一 7-ELEVEn 獅隊、味全龍隊。

**·資料處理**

Lahman’s Baseball Database 為一美國職業棒球大聯盟的資料集，其包含 1871 年至 2019 年的完整擊球和投球統計數據，以及防守統計數據，排名，球隊統計數據，管理記錄，季後數據等。

首先，將 Lahman’s Baseball Database 的資料分成打者，投手，團隊討論，並取出其 1985-2016 年的資料，成立新的資料集。

關於打者的部分:我們將 1985-2016 年的資料集中的打者資料，將同一年度在多於 1 支球隊效力的球員進行年度數據合併，並且增加傳統以及進階打擊數據變數進入打者資料，最後，我們將打者資料篩選出完成規定打席之打者，清理掉不必要的資料，再合併薪水資料，完成打者資料集的前置處理。

投手的部分:跟打者一樣，我們將 1985-2016 年的資料集中的投手資料，將同一年度在多於 1 支球隊效力的球員進行年度數據合併，並且增加傳統以及進階投球數據變數進入投手資料，最後，將投手資料篩選出投球局數≥ 120 為先發投手，<100 為後援投手，清理掉不必要的資料。後援投手較為特殊，且上場機會較先發投手少，我們將與後援投手無關的變數剔除掉，並且清理掉無法呈現的資料。最後，再合併薪水資料，即完成投手資料集的前置處理。

團隊的部分: 我們將 1985-2016 年的資料集中的團隊資料新增 2 個變數最後，再合併團隊薪資之資料，完成團隊資料集的前置處理。

關於進階棒球數據，我們將分成打者與投手來介紹：

進階打者數據

**IsoP 純長打率(Isolated Power)**

計算公式： IsoP = S L G − A V G

純長打率=(長打率-打擊率)

**IsoD (Isolated Discipline) — 純選球率**

計算公式：IsoD = (OBP - AVG)

純選球率=(上壘率-打擊率)

說明：純選球率是判斷打者選球力的指標，比單純的上壘率好，因為可以有效降低上壘率被打擊率烘托的效應。一般來說，一位打者的純選球率要是大於0.100，即為絕佳的一棒人選；及格的一棒，最起碼的要求也是要大於0.060。

**SecA (Secondary AVG) — 第二打擊率**

計算公式：SecA = (BB + (TB – H) + (SB – CS)) / AB

第二打擊率=(四壞球+(總壘打數–安打)+(盜壘-盜壘刺))/打數

**EqA (Equivalent Average)**

計算公式: EqA = (H + TB + 1.5x(BB + HBP) + SB) / (AB + BB + HBP + CS + SB/3)

EqA = (安打+總壘打數+1.5x(四壞球+觸身球)+盜壘)/(打數+四壞球+觸身球+盜壘刺+盜壘/3)

**RC (Run Create) — 創造得分**

計算公式：

A = H + BB + HBP – CS – GIDP

B = TB + 0.24x(BB – IBB + HBP) + 0.62xSB + 0.5x(SH + SF) — 0.3xSO

C = AB + BB + HBP + SH + SF

RC(1) = A x B / C

RC(2) = [(2.4xC + A) x (3xC + B) / 9xC] — 0.9Xc

A = 安打+四壞球+觸身球-盜壘刺-雙殺打

B = 總壘打數+0.24x(四壞數-故意四壞+觸身球)+0.62x盜壘+0.5x(犧牲短打+犧牲高飛)-0.3x三振

C = 打數+四壞球+觸身球+犧牲短打+犧牲飛球

創造得分(1) = A x B / C

創造得分(2) = [(2.4xC + A) x (3xC + B) / 9xC] — 0.9Xc

說明：RC，創造得分，是一個比打點及得分更好的數據，可以看出一位打者整季替球隊貢獻幾分。

RC的公式共有非常多種版本，RC(1)的為2002年前的版本，是假定打線9位都是同一選手時，該位選手創造出來的得分。RC(2)所介紹的是Bill James於2002年提出的最新版本，此版本是假設該位球員與其他8位都是OBP.300及SLG.400的隊友 (約是聯盟平均成績) 時，替球隊創造出來的得分效益。提出此一版本的目地是在修正當某位選手有過於極端突出的表現時,RC(1)的版本會出現高估的效應。至於最原始版本「RC=(安打數+四壞球)x(總壘打數)/打席」則計算上比較簡單,目前也仍有人在使用!!

**OPS+ (Adjusted Production) — 修正能力值**

計算公式：

OPS+ = (OBP/Lg OBP) + (SLG/Lg SLG) - 1

修正能力值=(上壘率/聯盟平均上壘率)+(長打率/聯盟平均長打率)-1

**ABR (Adjusted Batting Runs) — 修正每打數所創造的分數**

計算公式：

ABR = (0.47)x1B + (0.78)x2B + (1.09)x3B + (1.40)xHR + (0.33)x(BB+HBP) - (0.25)x(AB-H) - (0.50)x(H+BB+HBP-LOB-R-CS)

修正每打數所創造的分數=(0.47x一安)+(0.78x二安)+(1.09x三安)+(1.40x全壘打)+0.33x(四壞+觸身)-0.25X(打數-安打)-0.50x(安打+四壞+觸身-殘壘數-得分-盜壘刺)

進階投手數據

**ERA+ (Earned Runs Average Plus) — 修正ERA**

計算公式：ERA+ = LgERA / ERA

修正ERA=聯盟ERA平均/ERA

說明： ERA+是將原本的ERA去除掉「球場因素」，這個指標可以視為投手的PRO+(或OPS+)，可以測量出這個投手相較於聯盟所有其他投手的平均責失分比率。一個投手本身的ERA越低，當然ERA+就越高，也就代表此投手越好。如同OPS+一樣，當一個投手的ERA+是100%時，那就是聯盟平均水準；若是120%，就是高於聯盟平均水準20%。

**APR (Adjusted Preventing Runs) — 修正失分**

計算公式：APR = IP / 9 * (LgERA - ERA)

修正失分=投球局數/9*(聯盟平均ERA-ERA)

說明：當ERA已經經過球場修正後，這個指標可以測量出一個投手與一個聯盟平均等級的投手在一個「中立」的球場投球時，在同樣的投球局數下，可以少失多少(自責)分。

## 統計方法

(一)逐步多元線性迴歸模型(Multiple Linear Regression with Stepwise Selection)

多元線性迴歸可以用來解釋多個獨立的變數間的關係，或預測變數間的關係。其中依變項於模型中由多個自變項與相應係數與常數項的函數計算而成。$Y$為應變數，$x_1$ 、$x_2 \cdots x_k$為自變數，而應變數與自變數間具有線性關係時，則多元迴歸模型假設為

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

其中，$\beta_0$ 為截距， $\beta_1$ 、$\beta_2 \cdots \beta_k$ 為迴歸係數， $\varepsilon$ 為常態誤差項。

其建立模型時，為確保模型的解釋力與預測力，自變數對應變數應有顯著的影響，並呈現線性關係，且其線性相關須為真實存在而非形式上，又自變數間應有一定的互斥性，亦即變數間具有獨立性。R 值為觀察值與預測值間關係的測量方法，$R^2$為 R 的平方，可用於呈現模型解釋力的程度。在本研究我們將薪資設為 Y，在加入變數，進行逐步迴歸選取重要變數。

(二)主成份迴歸 (Principal Component Regression)

主成份迴歸為多元線性迴歸的另一種方法，以主成份分析法為基礎，從原始變數中找出少數幾個主成份來預測變量間的內部結構，可被用於估計未知的迴歸係數，並且可用於解決變數間共線性問題。

一般的迴歸模型等式為$Y = XB + e$。其中，Y 為依變數，X 為自變數，B 為預估計的迴歸模型係數，e 則為殘差。

在最小平方法中，迴歸係數是利用$\hat{B} = (X'X)^{-1}X'Y$ 估計。

而在主成份迴歸中，我們將反應變數設為 Y，解釋變數設為 X，這裡的 $X = (X_1, X_2, \cdots, X_p)^T$，即為主成分分析，我們選取 M 個主成分(new variables)為 $Z = (Z_1, Z_2, \cdots, Z_M)^T, M \ll p$，Y 對 Z 進行迴歸，使用普通最小二乘法(OLS, Ordinary Least Sqaure)，我們會得到近似 Y 對 X 進行迴歸的結果。

亦對迴歸係數具有限制式，若在三個變數的情形下，限制式為 $p_{13}b_1 + p_{23}b_2 + p_{33}b_3 = 0$ 。因此可以藉由限制解的區域以避免多重共線性

的問題。

## 研究結果

本次實例分析，我們是將 Lahman’s Baseball Database 的資料分成打者，投手，團隊討論，並取出其1985-2016年的資料，成立新的資料集。打者資料篩選出完成規定打席之打者(打席須≥3.1*162場比賽)，總共有3028筆資料，33個變數，其中1個類別變數，4個辨識碼，28個連續變數。先發投手資料篩選出完成將投手資料篩選出投球局數≥ 120，總共有3321筆資料，33個變數，其中1個類別變數，4個辨識碼，28個連續變數。後援投手資料篩選出完成將投手資料篩選出投球局數<100，清理掉無法呈現的資料，得到總共有3321筆資料，33個變數，其中1個類別變數，4個辨識碼，28個連續變數。團隊資料的部分將1985-2016年團隊資料合併團隊薪資之資料，得到總共有918筆資料，38個變數，其中1個類別變數，2個辨識碼，28個連續變數。

### 描述性統計

我們將打者、先發投手、後援投手、團隊資料進行描述性統計。在描述性統計中，我們發現投手資料中比較特別的是先發以及後援的ERA和ERA+的標準差有明顯差異，推測應該是投球局數所造成的，後援投手有時可能只出賽一局，雖然只失了一分，但ERA還是會偏高(見下 Table 1)。而打者與先發投手的平均薪資也有著將近一百萬美元的差異，後援投手平均薪資最低(見下 Table 2)。

Table 1 先發投手與後援投手的 ERA 和 ERA+

| | ERA 平均(標準差) | ERA+ 平均(標準差) |
|---|---|---|
| 先發投手 | 4.1 (0.9) | 135.3 (31.8) |
| 後援投手 | 4.7 (2.3) | 138.9 (71.4) |

Table 2 平均薪資 (單位:萬 美元)

| | 打者 | 先發投手 | 後援投手 |
|---|---|---|---|
| 平均薪資 (標準差) | 408 (483) | 317 (444) | 133 (221) |

團隊資料中，勝場數和敗場數的平均幾乎相同，說明比賽的勝負，在長期下來的機率會非常接近；另外，勝分差和勝率呈現高度線性正相關 (見下 Figure 1)。

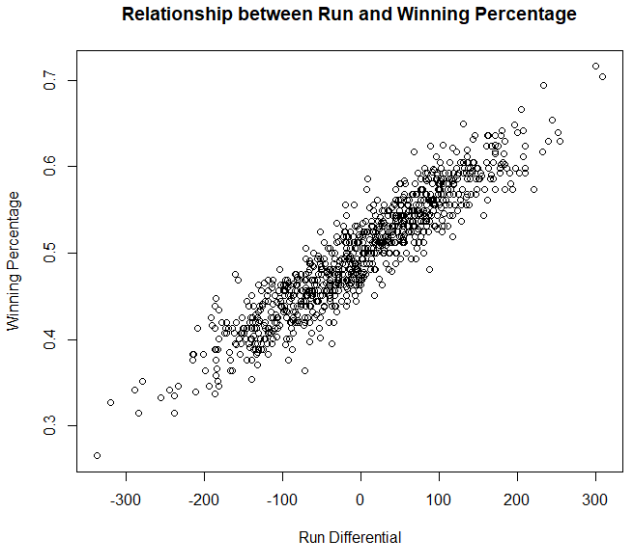


Figure 1 勝分差和勝率關係圖

## 迴歸分析

做完描述性統計後，進入迴歸分析，將薪資作為反映變數，進行多元線性迴歸以及主成分迴歸。我們分打者、先發投手、後援投手、團隊，四部分討論。最後，挑選最適合的模型進行模型預測。

首先，打者的多元線性迴歸，我們發現 Variance 會隨著 Fitted Values 愈來愈大，根據此狀況，我們將殘差進行開平方根的處理，並發現狀況有所改善(見下 Figure 2 Figure 3)。

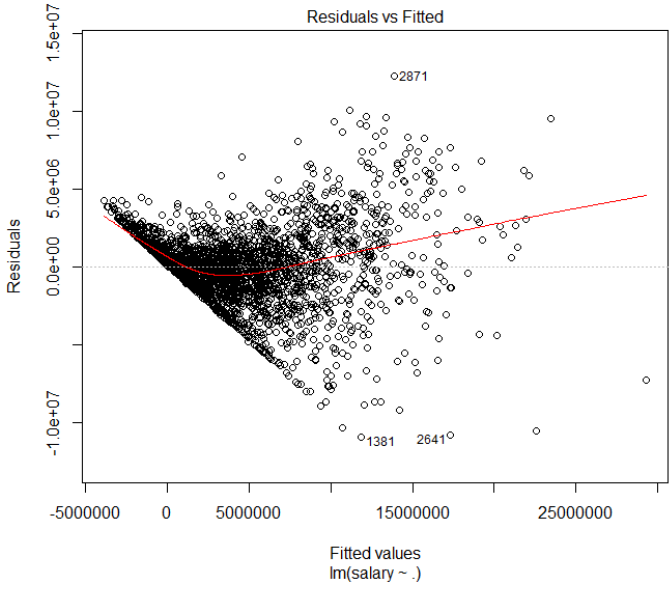


Figure 2 Residual vs Fitted

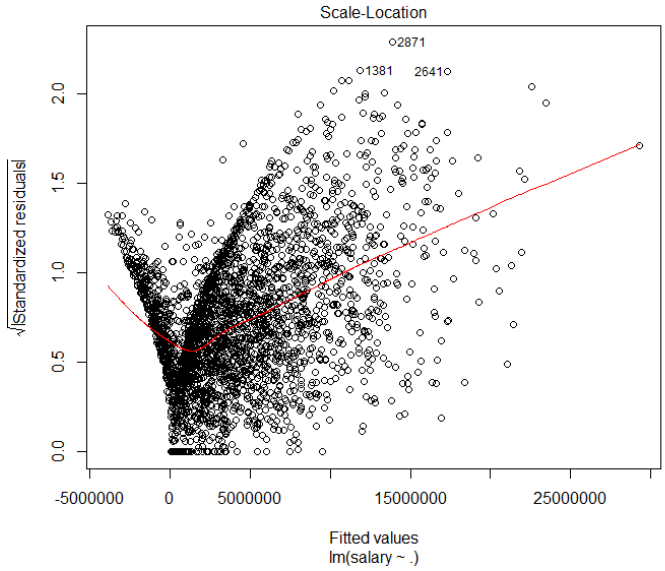


Figure 3 開平方根後的結果

接著我們進行殘差檢定:

(1) 常態性檢定

| Shapiro-Wilk Normality Test | |
|---|---|
| w = 0.94702 | p-value < 2.2e-16 |

$H_0$: 殘差服從常態分佈 versus $H_A$: 殘差不服從常態分佈

p-value <0.05，故拒絕虛無假設，殘差並不服從常態分佈

(2) 獨立性檢定

| Lag | Autocorrelation | D-W Statistic | p-value |
|---|---|---|---|
| 1 | 0.3643574 | 1.271072 | 0 |

$H_0$: 殘差互相獨立 versus $H_A$: 殘差互相獨立

p-value <0.05，故拒絕虛無假設，殘差並不互相獨立

(3) 變異數同質檢定

| Non-constant variance Score Test | | |
|---|---|---|
| Chi-Square=1613.352 | Df=1 | p-value < 2.22e-16 |

$H_0$: 殘差變異數具有同質性 versus $H_A$: 殘差變異數不具有同質性

p-value <0.05，故拒絕虛無假設，殘差並不具有同質性

從上，我們可知打者的多元線性迴歸，是很好的配適。

進行完多元線性迴歸，我們進行主成分迴歸，將薪資作為反映變數，呈現之

結果如下。前 3 組主成分，解釋力達到 63.3%，其殘差呈常態分佈，是好的配適模型(見下 Figure 4 Figure 5)。


Figure 4 MSEP 主成分數量

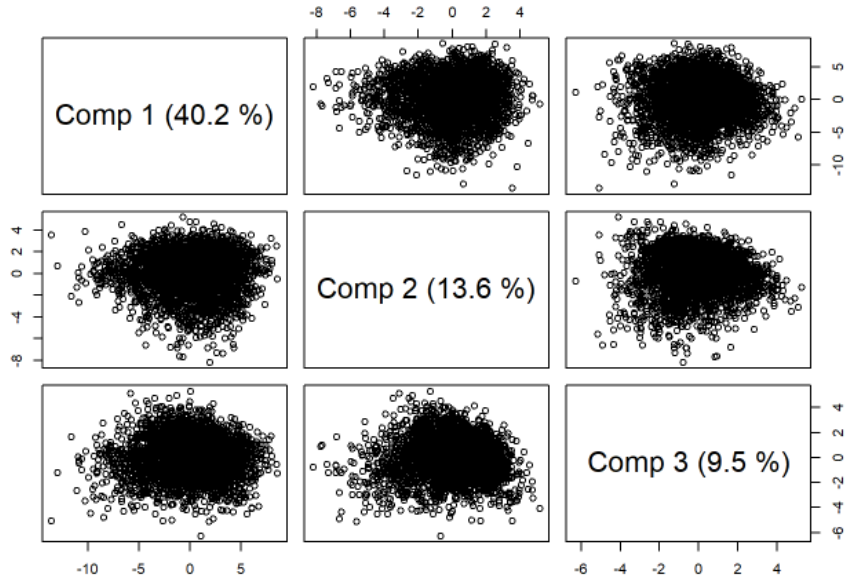


Figure 5 前三組主成分示意圖

最後，我們進行打者預測模型 RMSE、MAE 與 Rsqared 比較，主成分迴歸(PCR)的表現最佳(見下 Table 3 Table 4)。

Table 3 打者預測模型比較標準

| | Linear Regression+ Stepwise Selection | Linear Regression+ Forward Selection | Linear Regression+ Backward Selection |
|---|---|---|---|
| AIC | 92528.94 | 92542.17 | 92528.94 |
| RMSE (SD) | 4520505 (304966.4) | 4509252(269664.7) | 4502323 (201617.7) |
| Rsquared (SD) | 0.1256 (0.0473) | 0.1286 (0.0360) | 0.1323 (0.0387) |
| MAE (SD) | 3325595 (180515.3) | 3312727(158904.15) | 3301891(120273.01) |

Table 4 打者 PCR 預測模型比較標準

| | PCR |
|---|---|
| RMSE (SD) | 3808763 (319159.9) |
| Rsquared (SD) | 0.3936 (0.0405) |
| MAE (SD) | 2813234 (180903.9) |

接著，進行先發投手的多元線性迴歸，我們發現 Variance 會隨著 Fitted Values 愈來愈大，根據此狀況，我們將殘差進行開平方根的處理，並發現狀況有所改善(見下 Figure 6 Figure 7)。

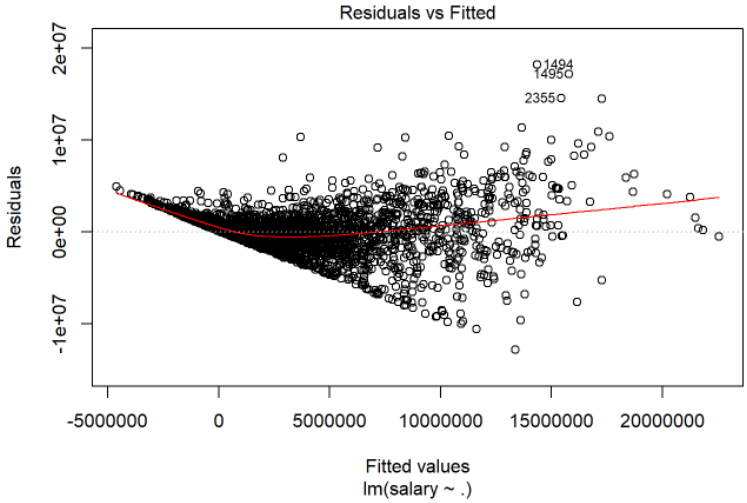


Figure 6 Residual vs Fitted

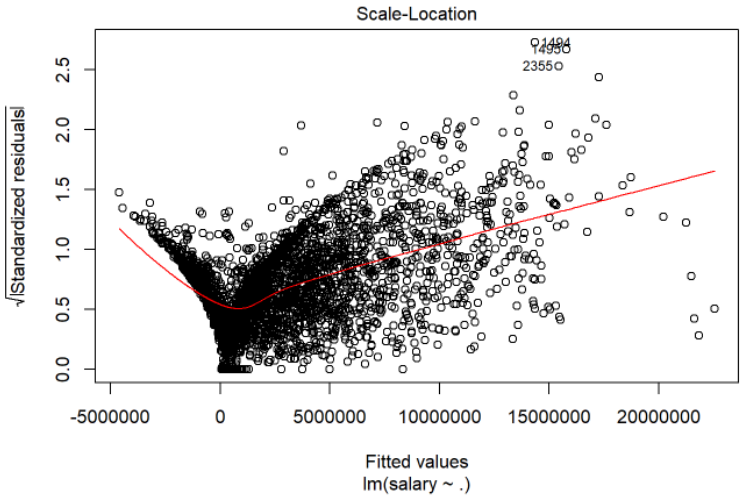


Figure 7 開平方根後的結果

接著我們進行殘差檢定:

(1) 常態性檢定

| Shapiro-Wilk Normality Test | |
|---|---|
| w = 0.90469 | p-value < 2.2e-16 |

$H_0$: 殘差服從常態分佈 versus $H_A$: 殘差不服從常態分佈
p-value <0.05，故拒絕虛無假設，殘差並不服從常態分佈

(2) 獨立性檢定

| Lag | Autocorrelation | D-W Statistic | p-value |
|---|---|---|---|
| 1 | 0.4060344 | 1.187153 | 0 |

$H_0$: 殘差互相獨立 versus $H_A$: 殘差互相獨立
p-value <0.05，故拒絕虛無假設，殘差並不互相獨立

(3) 變異數同質檢定

| Non-constant variance Score Test | | |
|---|---|---|
| Chi-Square=2859.752 | Df=1 | p-value < 2.22e-16 |

$H_0$: 殘差變異數具有同質性 versus $H_A$: 殘差變異數不具有同質性
p-value <0.05，故拒絕虛無假設，殘差並不具有同質性

從上，我們可知先發投手的多元線性迴歸，是很好的配適。

進行完多元線性迴歸，我們進行主成分迴歸，將薪資作為反映變數，呈現之結果如下。前 3 組主成分，解釋力達到 55.9%，其殘差呈常態分佈，是好的配適模型(見下 Figure 8 Figure 9)。

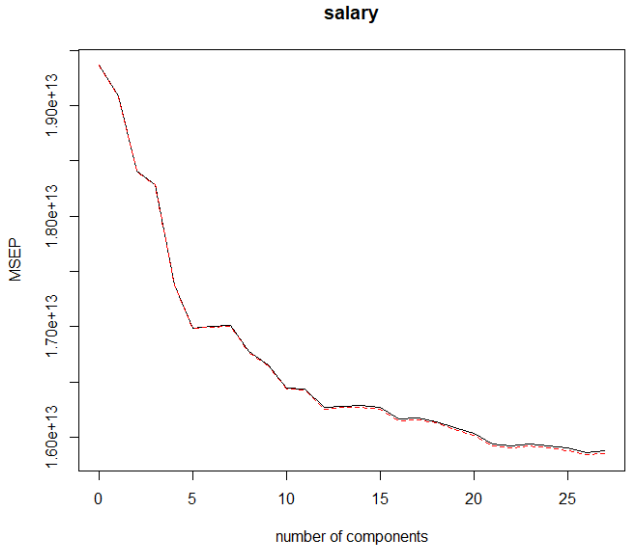


Figure 8 MSEP 主成分數量

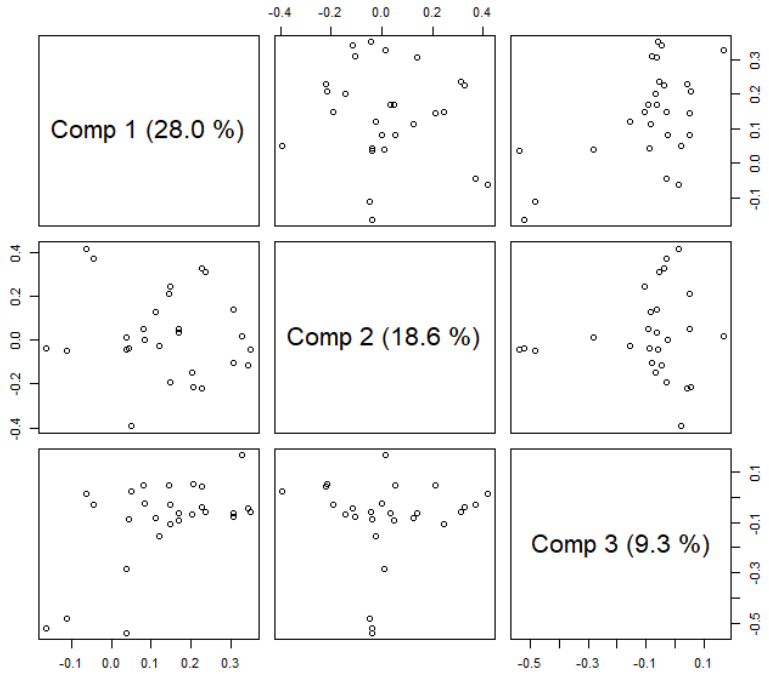


Figure 9 前三組主成分示意圖

最後，我們進行先發投手預測模型 RMSE、MAE 與 Rsqared 比較，主成分迴歸(PCR)的表現最佳(見下 Table 5 Table 6)。

Table 5 先發投手預測模型比較標準

| | Linear Regression+ Stepwise Selection | Linear Regression+ Forward Selection | Linear Regression+ Backward Selection |
|---|---|---|---|
| AIC | 100943.6 | 100955.9 | 100943.6 |
| RMSE (SD) | 4056529 (401964.7) | 4047720(266040.9) | 4044236(308680.9) |
| Rsquared (SD) | 0.1597 (0.0519) | 0.1662 (0.0248) | 0.1672 (0.0308) |
| MAE (SD) | 2803397 (175527.3) | 2792467(164092.6) | 2788712(141542.4) |

Table 6 先發投手 PCR 預測模型比較標準

| | PCR |
|---|---|
| RMSE (SD) | 3677556 (255790.4) |
| Rsquared (SD) | 0.2983 (0.0428) |
| MAE (SD) | 2590098 (210773.6) |

接著，進行後援投手的多元線性迴歸，我們發現 Variance 會隨著 Fitted Values 愈來愈大，根據此狀況，我們將殘差進行開平方根的處理，並發現狀況有所改善(見下 Figure 10 Figure 11)。

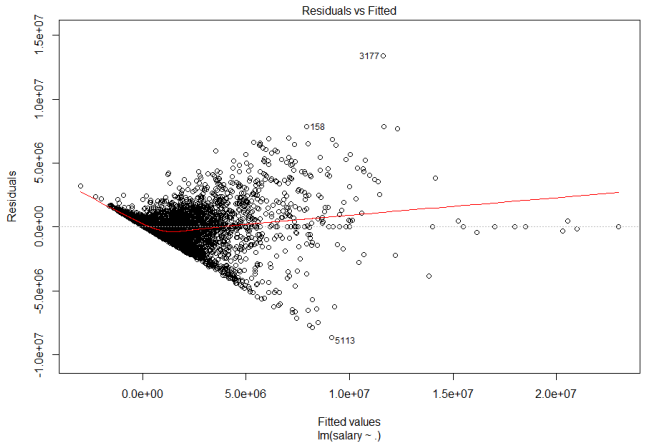


Figure 10 Residual vs Fitted

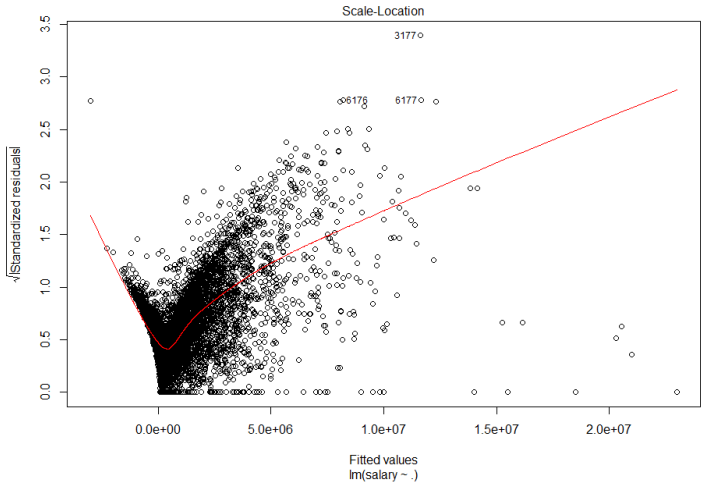


Figure 11 開平方根後的結果

接著我們進行殘差檢定:

(1) 常態性檢定，因為 Shapiro-Wilk Normality Test 的樣本數介在 3 和 5000 之間。因為後援投手的樣本數>5000，會拒絕服從常態。

(2)獨立性檢定

| Lag | Autocorrelation | D-W Statistic | p-value |
|---|---|---|---|
| 1 | 0. 1525284 | 1.694917 | 0 |

$H_0$: 殘差互相獨立 versus $H_A$: 殘差互相獨立

p-value <0.05，故拒絕虛無假設，殘差並不互相獨立

(3)變異數同質檢定

| Non-constant variance Score Test | | |
|---|---|---|
| Chi-Square=10556.15 | Df=1 | p-value < 2.22e-16 |

$H_0$: 殘差變異數具有同質性 versus $H_A$: 殘差變異數不具有同質性

p-value <0.05，故拒絕虛無假設，殘差並不具有同質性

從上，我們可知後援投手的多元線性迴歸，是很好的配適。

進行完多元線性迴歸，我們進行主成分迴歸，將薪資作為反映變數，呈現之結果如下。前 3 組主成分，解釋力達到 97%，其殘差呈常態分佈，是好的配適模型(見下 Figure 12 Figure 13)。

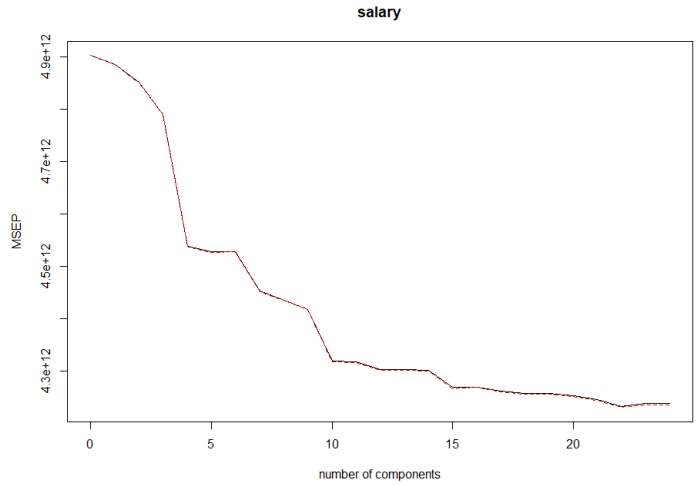


Figure 12 MSEP 主成分數量

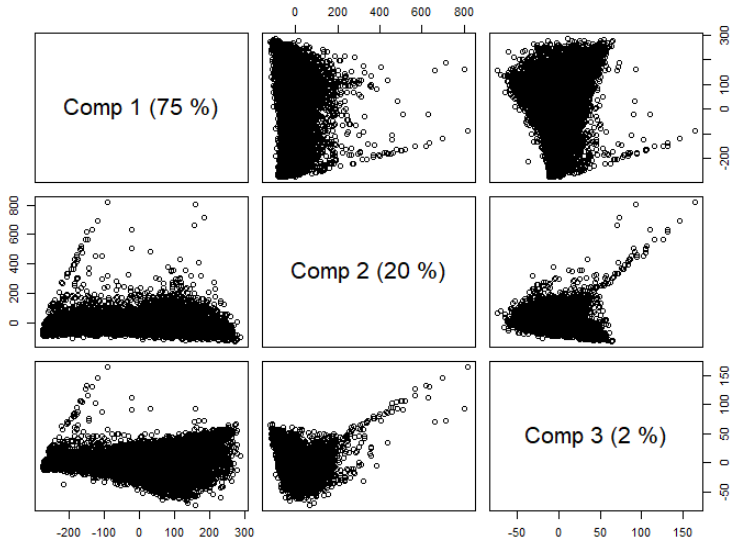


Figure 13 前三組主成分示意圖

最後，我們進行先發投手預測模型 RMSE、MAE 與 Rsqared 比較，主成分迴歸(PCR)的表現最佳(見下 Table 7 Table 8)。

Table 7 後援投手預測模型比較標準

| | Linear Regression+ Stepwise Selection | Linear Regression+ Forward Selection | Linear Regression+ Backward Selection |
|---|---|---|---|
| AIC | 192170.1 | 192176.8 | 192170.1 |
| RMSE (SD) | 2059412 (164786) | 2059012(173927.7) | 2054572 (223427.2) |
| Rsquared (SD) | 0.1312 (0.0202) | 0.1322 (0.0264) | 0.1341 (0.0362) |
| MAE (SD) | 1232387 (59968.71) | 1231124 (49821) | 1231038 (68964.5) |

Table 8 後援投手 PCR 預測模型比較標準

| | PCR |
|---|---|
| RMSE (SD) | 1871918 (175472.7) |
| Rsquared (SD) | 0.2478 (0.0235) |
| MAE (SD) | 1149841 (52710.84) |

接著，進行團隊的多元線性迴歸，我們發現 Variance 會隨著 Fitted Values 愈來愈大，根據此狀況，我們將殘差進行開平方根的處理，並發現狀況有所改善(見下 Figure 14 Figure 15)。

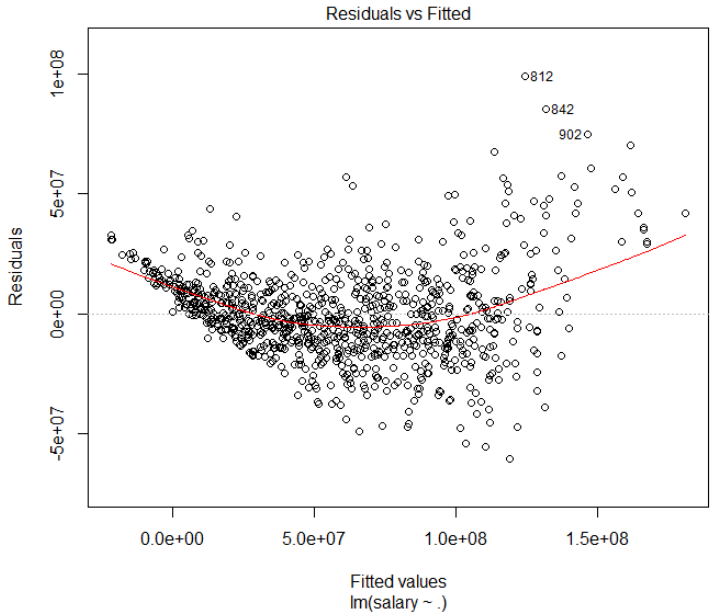


Figure 14 Residual vs Fitted

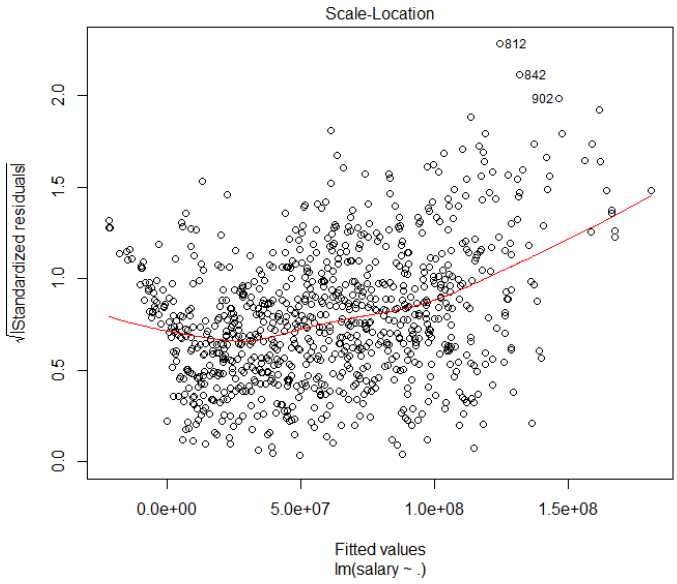


Figure 15 開平方根後的結果

接著我們進行殘差檢定:

(1)常態性檢定

| Shapiro-Wilk Normality Test | |
|---|---|
| w = 0.97298 | p-value < 5.003e-12 |

$H_0$: 殘差服從常態分佈 versus $H_A$: 殘差不服從常態分佈

p-value <0.05，故拒絕虛無假設，殘差並不服從常態分佈

(2)獨立性檢定

| Lag | Autocorrelation | D-W Statistic | p-value |
|---|---|---|---|
| 1 | 0.01119571 | 1.976132 | 0.66 |

$H_0$: 殘差互相獨立 versus $H_A$: 殘差互相獨立

p-value > 0.05，故無法拒絕虛無假設，殘差間互相獨立

(3)變異數同質檢定

| Non-constant variance Score Test | | |
|---|---|---|
| Chi-Square=228.007 | Df=1 | p-value < 2.22e-16 |

$H_0$: 殘差變異數具有同質性 versus $H_A$: 殘差變異數不具有同質性

p-value <0.05，故拒絕虛無假設，殘差並不具有同質性

從上，我們可知團隊的多元線性迴歸，是很好的配適。

進行完多元線性迴歸，我們進行主成分迴歸，將薪資作為反映變數，呈現之結果如下。前 5 組主成分，解釋力達到 72%，其殘差呈常態分佈，是好的配適模型(見下 Figure 16 Figure 17)。

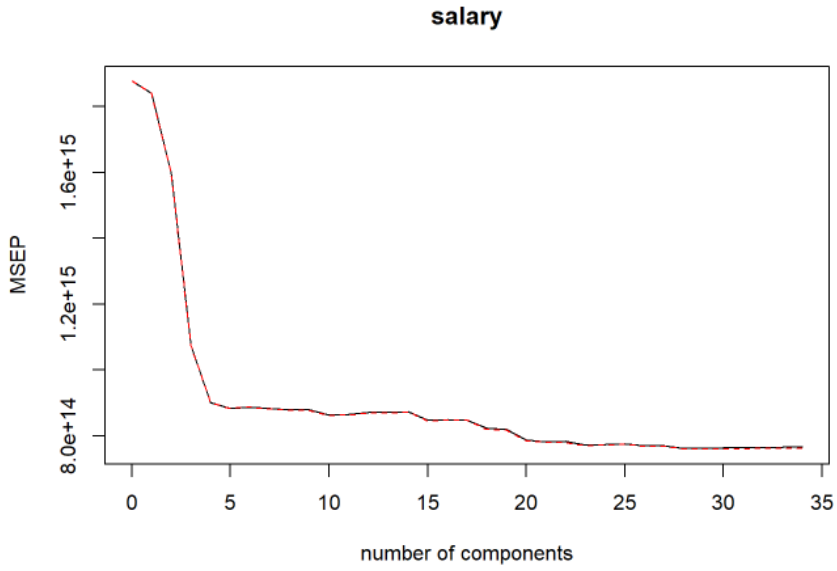


Figure 16 MSEP 主成分數量

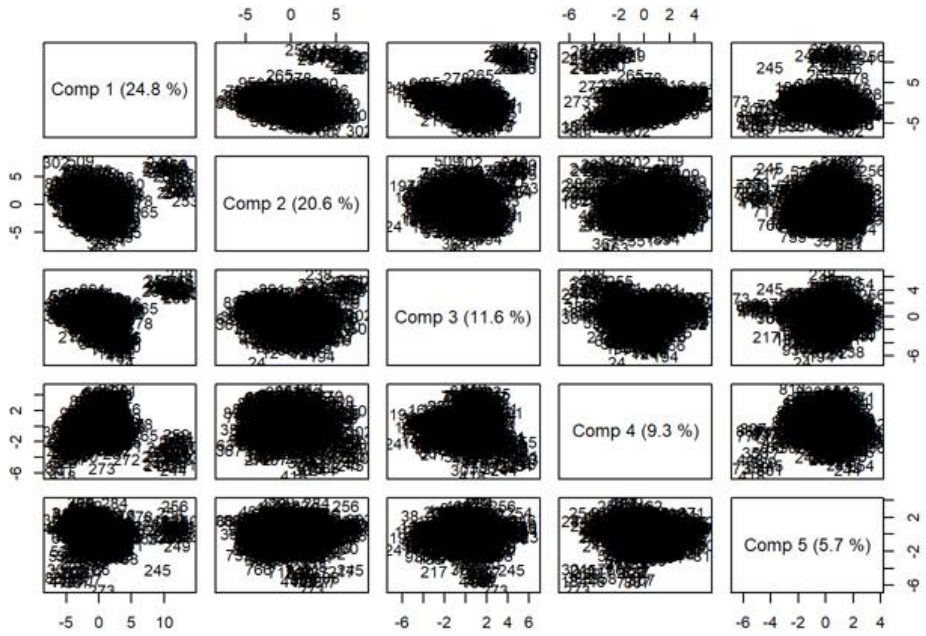


Figure 17 前五組主成分示意圖

最後，我們進行團隊預測模型 RMSE、MAE 與 Rsqared 比較，主成分迴歸(PCR)的表現最佳(見下 Table 9 團隊預測模型比較標準 Table 9 Table 10)。

Table 9 團隊預測模型比較標準

| | Linear Regression+ Stepwise Selection | Linear Regression+ Forward Selection | Linear Regression+ Backward Selection |
|---|---|---|---|
| AIC | 31438 | 31455.08 | 31438 |
| RMSE (SD) | 28389645(3813656) | 27823997(1921978) | 27764338 (2042503) |
| Rsquared (SD) | 0.5702 (0.1131) | 0.5929 (0.0351) | 0.5941 (0.0402) |
| MAE (SD) | 20816396(2631816) | 20344321(158904.15) | 20330898 (1205981) |

Table 10 團隊 PCR 預測模型比較標準

| | PCR |
|---|---|
| RMSE (SD) | 21840374 (3192478) |
| Rsquared (SD) | 0.7503 (0.0419) |
| MAE (SD) | 16279514 (1968516) |

**預測模型**

完成迴歸分析後，接下來我們要建立薪水預測模型。從上描述性統計，我們可以發現，美國職業棒球大聯盟的生態為打者薪水遠高於投手薪水；而在台灣的中華職業棒球大聯盟，近年來的生態為打高投低，被專家譽為「打擊聯盟」。我們可知兩聯盟都相當重視打擊。我們將針對美國職業棒球大聯盟 1985-2016 年的打者資料進行薪水預測模型。

首先，透過模型變數重要性，我們可以透過圖形來探討那些變數是在模型中相對重要的變數，呈現如下 Figure 18。可以觀察到，變數中占重要性比較重的變數與長打與打點能力。換句話說，美國職業棒球大聯盟可獲高薪的打者，其為球隊得分的能力強；與中華職業棒球大聯盟的生態類似，雖然中華職業棒球大聯盟的薪水資料是不公開的，我們可以透過球團與媒體官方的報導得知球員薪水，例如: 樂天桃猿的選手-林泓育，透過CPBL STATS找出其2019年的數據為打擊率:0.350，全壘打:26隻，打點:95分，SLG:0.585，OPS:1.001，OPS+為151.7。我們建立一個以打擊率，全壘打，打點，SLG，OPS，OPS的模型進行預測，透過我們的模型我們發現林泓育的年薪高達2億，雖然數值稍有誇飾，但是，透過樂天桃猿的官方消息與媒體的官方報導，林泓育2020賽季，月薪為75萬新台幣，為本賽季中華職業棒球大聯盟的最高薪選手，與模型的高薪不謀而合。但是，此模型還需要透過匯率以及各國CPI指數作調整，才能更加準確。

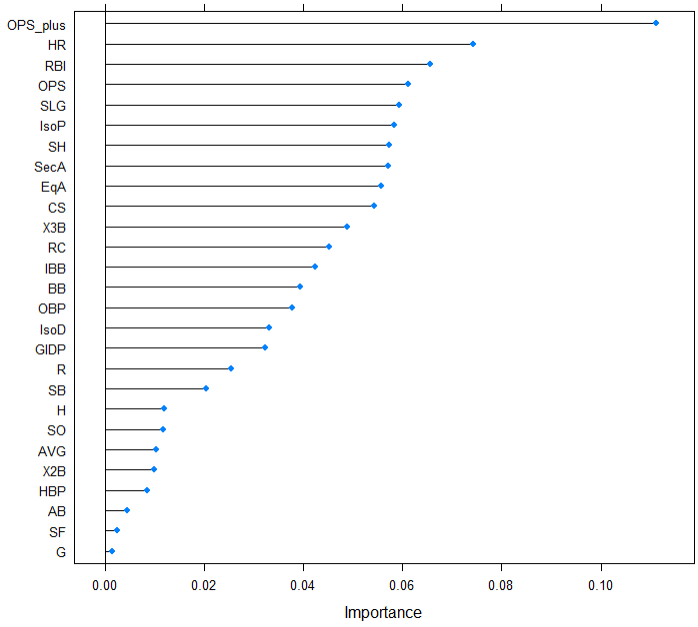


Figure 18 模型變數重要性

## 結論

中華職業棒球大聯盟常常出現球團無法負擔虧損，導致球隊破產，造成球團解散或是轉賣，透過這一次利用美國職業棒球大聯盟 1985-2016 年的打者資料進行薪水預測模型，我們發現是可以套用到中華職業棒球大聯盟上的。透過這個系統，球團可以降低風險控管，不至於造成虧損，找到合適的球員；球員也可以透過此模型，爭取加薪。

# 参考文獻

Cyrenne, P. (2014). Salary Inequality, Team Success and the Superstar Effect.

Fields, B. J. U. m. s. t., East Carolina University, Greenville, NC, USA (2001). "Estimating the value of Major League Baseball players."

Hall, S., et al. (2002). "Testing causality between team performance and payroll: the cases of Major League Baseball and English soccer." **3**(2): 149-168.

Hoffman, M. G. (2014). Analysis of salary for Major League Baseball players, North Dakota State University.

Jewell, R. T. and D. J. J. E. I. Molina (2004). "The effect of salary distribution on production: An analysis of major league baseball." **42**(3): 469-482.

Magel, R. and M. J. I. J. o. S. S. Hoffman (2015). "Predicting salaries of major league baseball players." **5**(2): 51-58.

Mizak, D. and A. J. E. B. Stair (2004). "The relationship between payroll and performance disparity in major league baseball: An alternative measure." **12**(9): 1-14.

Shorin, G. (2017). Team Payroll Versus Performance in Professional Sports: Is Increased Spending Associated with Greater Success?, Duke University Durham.

Wasserman, T. (2013). "Determinants of Major League Baseball Player Salaries."

Wiseman, F. and S. J. E. B. Chatterjee (2003). "Team payroll and team performance in major league baseball: 1985–2002." **1**(2): 1-10.