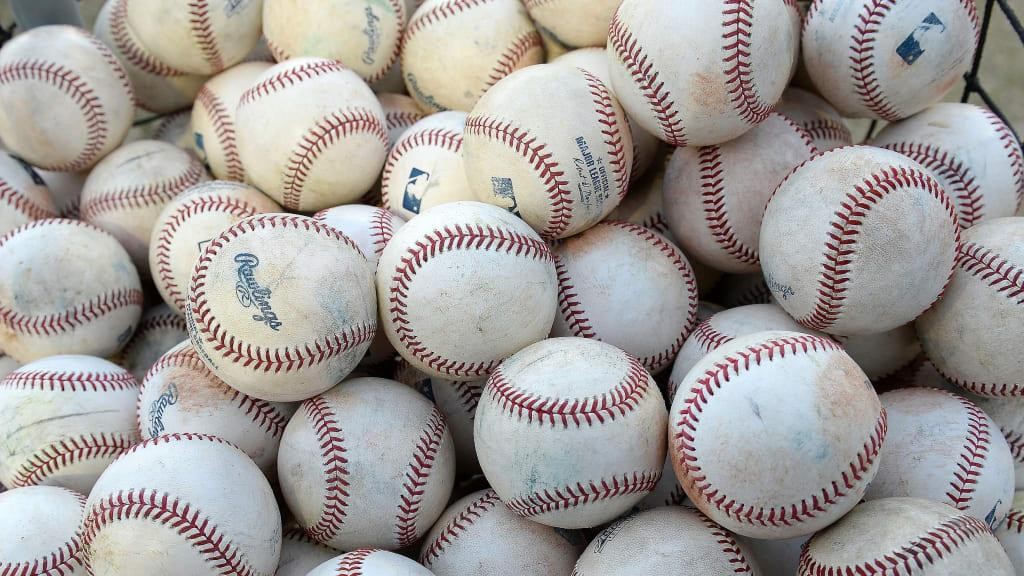
迴歸分析報告：常見的棒球進階數據對加權上壘率的影響

[壹、主題與動機介紹](#_bookmark0)

[貳、資料介紹](#_bookmark1)



[參、選擇模型與模型診斷](#_bookmark2)[肆、模型補救](#_bookmark3)

[伍、模型解釋與總結](#_bookmark4)

# 第 12 組

410978004 統計二 郭依璇

410378011 統計二 譚靖蓉

410978034 統計二 李海寧

410978044 統計二 謝瑋芸

410875019 統計三 林哲兆

410875024 統計三 李亮頤

# 壹、主題與動機介紹

棒球在台灣是一項盛行的運動，棒球比賽除了看場上球員的拚勁，從數據層面看這項運動也別有一番趣味。此外，隨著科技的進步，許多球隊也開始藉由數據調整對球員的訓練方式，因此我們想要研究一些擊球品質數據對加權上壘率的影響。

# 貳、資料介紹

資料來源為國外對於美國職棒大聯盟進階數據統計的網站 savant (<https://baseballsavant.mlb.com/>)。樣本為 2015~2022 年大聯盟中，所有打席數超過一定數量的打者之各項進階數據，包含打者的加權上壘率、年齡、平均擊球初速、平均擊球仰角、甜蜜點比例和出色擊球率。而由於資料中有部分為同一打者於不同年份的數據，因此可能存在互相不獨立的問題。

變數介紹：

1. woba (加權上壘率)：全名為 Weighted On-Base Average。它兼具上壘率和長打率對打者攻擊能力的見解，架構頗像把兩者的公式結合起來，能把打擊的各種情況考慮進去，並給予更合理的權重計算。
2. player\_age (年齡)：我們將年齡分為兩組，25~29 歲為 1，其餘的為 0，因為 25~29 歲是公認的美國職棒大聯盟運動員表現最好的時期。註 : 使用SAS時，將此變數名稱改為age0
3. exit\_velocity\_avg (MPH) (平均擊球初速)：代表的是被球棒打到的那顆球飛出去的初速。
4. launch\_angle\_avg (平均擊球仰角)：擊球仰角指的是打者把球擊出後，球離開球棒時的軌跡面與地面形成的角度。
5. sweet\_spot\_percent (%) (甜蜜點比例)：打者的擊球事件中，仰角落在甜蜜點的擊球所佔的比例。
6. sweet\_spot\_percent (%) (出色擊球率)：擊出出色擊球的比例，明確定義為「以擊球仰角和初速等條件來看，安打率不低於.500 且長打率至少在 1.500 之上」的擊球，其代表的意義是：打得非常紮實、擊球仰角和初速都在優良範圍的擊球事件。
7. 交互作用：x6 = launch\_angle\_avg \* barrel\_batted\_rate (擊球仰角\*出色擊球率)

x7 = sweet\_spot\_percent \* barrel\_batted\_rate(甜蜜點比例\*出色擊球率)

)

# 參、選擇模型與模型診斷

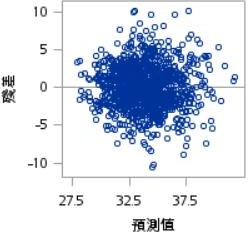
以下建立兩個模型來解釋 woba 跟其他變數之間的關係。

## 模型一：沒有交互作用

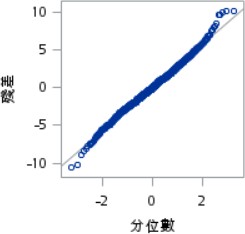
𝑤𝑜𝑏𝑎 = 𝛽0 + 𝛽1 𝑎𝑔𝑒0 + 𝛽2 𝑒𝑥𝑖𝑡\_𝑣𝑒𝑙𝑜𝑐𝑖𝑡𝑦\_𝑎𝑣𝑔 + 𝛽3 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔

+ 𝛽4 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡 + 𝛽5 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 + 𝜀𝑖

1. Linear Tendency：通過



殘差圖顯示，𝑦̂𝑖於紅色框起處，殘差的正值較多(其分布不是對於水平線 r = 0 上下對稱)，但我們認為該處資料數佔整體資料數之比例極低，因此仍判定此資料符合 linear tendency。

1. Error Normality：通過

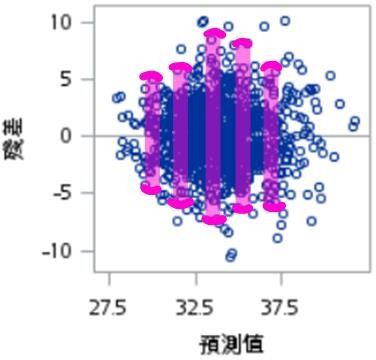
Normal Q-Q plot 呈現殘差頭尾偏離直線，但仍大致符合 45°線，判定殘差呈現常態分佈。

四個常態性檢定的結果中，除了 Kolmogorov-Smirnov Test，其餘均為 reject error normality。

綜合上述，我們認為是 Q-Q plot 頭尾偏離 45°線而導致常態性檢定未通過，但仍可看出絕大部分的資料是在 45°線上，所以最終判定殘差符合常態分佈。

1. Constant Variance：不通過





1. 殘差圖中呈現，當𝑦̂𝑖大約於 33 之處，殘差的分布似乎較 ŷi 在 30 及 37.5 之處來得廣，可知殘差的變異數會隨著 ŷi 的改變而改變。
2. White Test 的結果為 reject constant error variance。

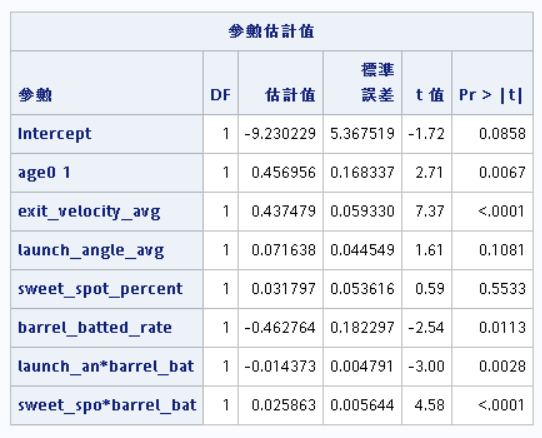
## 模型二：有交互作用

利用 Backward Elimination，保留 p-value < 0.05 的自變數以及交互作用的主作用項所建的模型。

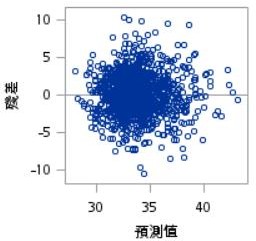
x6 = launch\_angle\_avg \* barrel\_batted\_rate x7 = sweet\_spot\_percent \* barrel\_batted\_rate

𝑦(𝑤𝑜𝑏𝑎) = 𝛽0 + 𝛽1 𝑎𝑔𝑒0 + 𝛽2 𝑒𝑥𝑖𝑡\_𝑣𝑒𝑙𝑜𝑐𝑖𝑡𝑦\_𝑎𝑣𝑔 + 𝛽3 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔 +

𝛽4 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡 + 𝛽5 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 + 𝛾1 𝑥6 + 𝛾2 𝑥7 + 𝜀𝑖

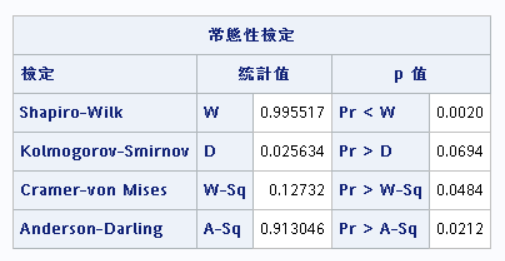
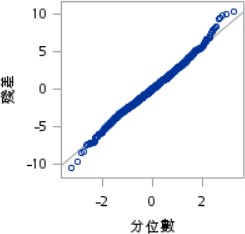


1. Linear Tendency：通過



殘差圖顯示，𝑦̂𝑖之分布對於水平線 r = 0 上下對稱，因此判定此資料符合 linear tendency。

1. Error Normality：通過

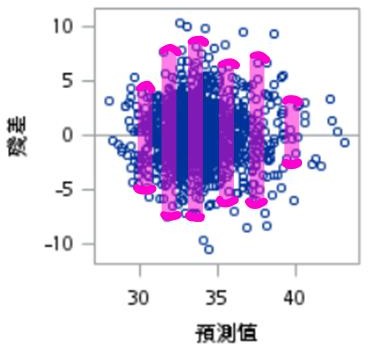


Normal Q-Q plot 呈現殘差頭尾偏離直線，但仍大致符合 45°線，判定殘差呈現常態分佈。

四個常態性檢定的結果中，除了 Kolmogorov-Smirnov Test，其餘均為 reject error normality。

綜合上述，我們認為是 Q-Q plot 頭尾偏離 45°線而導致常態性檢定未通過，但仍可看出絕大部分的資料是在 45°線上，所以最終判定殘差符合常態分佈。

1. Constant Variance：通過



1. 殘差圖中呈現當𝑦̂𝑖大約於 35 之處，殘差的分布似乎較𝑦̂𝑖在 30 及 40 之處來得廣，可知殘差的變異數會隨著𝑦̂𝑖的改變而改變。
2. White Test 的結果為 do not reject constant error variance。

由殘差圖可知，雖然殘差的變異數看似較不固定，但我們選擇相信 White Test

之結果。結論：

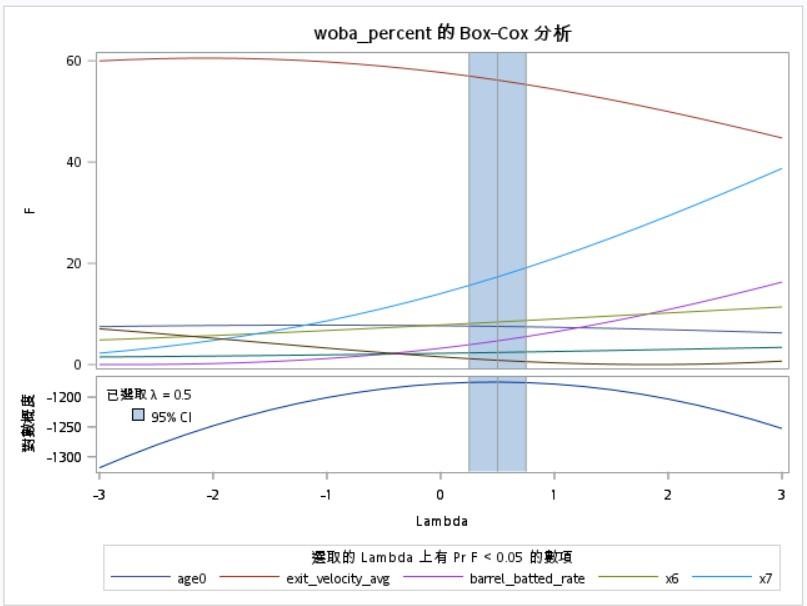
|  |  |  |
| --- | --- | --- |
|  | **模型一** | **模型二** |
| *Linear Tendency* | 通過 | 通過 |
| *Error Normality* | 通過 | 通過 |
| *Constant Variance* | 不通過 | 通過 |

根據觀察，交互作用改善了 Constant Variance 不通過的問題，因此選擇模型二來做分析。

# 肆、模型補救

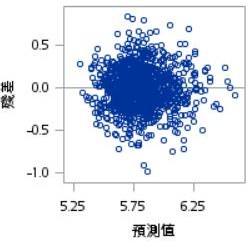
由於模型二之 White Test 的 p-value 為 0.0789，只略微大於 0.05，因此我們決定嘗試做模型補救，希望可以有所改善。

承上模型診斷可知 Constant Variance 不通過，所以選擇對反應變數(y)做轉換。參考 Box-Cox Transformation 的結果(如下圖)，選擇使用 𝑔(𝑦𝑖) = √𝑦𝑖。



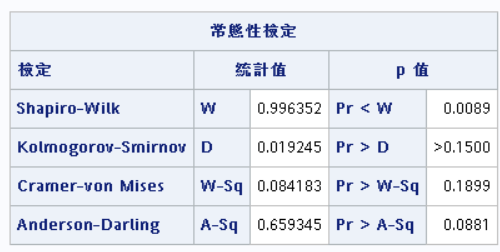
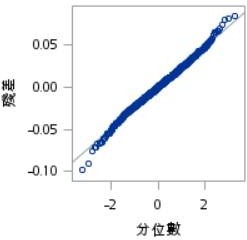
以下為補救後的診斷：

1. Linear Tendency：通過



殘差圖顯示，𝑦̂𝑖之分布對於水平線 r = 0 上下對稱，因此判定此資料符合 linear tendency。

1. Error Normality：通過

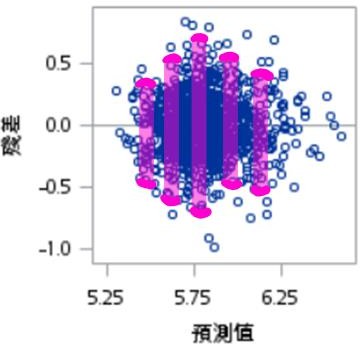


Normal Q-Q plot 呈現殘差頭尾偏離直線，但仍大致符合 45°線，判定殘差呈現常態分佈。

四個常態性檢定的結果中，除了 Shapiro-Wilk Test，其餘均為 do not reject error normality。

綜合上述，最終判定殘差符合常態分佈。

1. Constant Variance：通過



1. 殘差圖中呈現當√𝑦̂𝑖大約於 5.75 之處，殘差的分布似乎較𝑦̂𝑖在 5.3 及 6.2 之處來得廣，可知殘差的變異數會隨著√𝑦̂𝑖的改變而改變。
2. White Test 的結果為 do not reject constant error variance。

由殘差圖可知，雖然殘差的變異數看似較不固定，但我們選擇相信 White Test

之結果。

結論：

雖然補救後的檢定結果有改善，但改善幅度不大且轉換後的反應變數不利於解釋，因此仍採用補救前之模型二。

# 一張含有 桌 的圖片 自動產生的描述伍、模型解釋與總結

* 截距估計值的意義：

此處截距指的是模型中沒有任何變數的參數：𝛽0。其意義為：當年齡為 25~29 歲以外，且平均擊球初速為 0MPH、平均擊球仰角為 0°、出色擊球率和甜蜜點比例為 0%時的平均加權上壘率為-9.23023%，但在現實中不會出現類似數據。

* 年齡分組所代表的意義：

在其他自變數固定的情況下，我們有 95%的信心，平均而言 25~29 歲的選手加權上壘率比其他年齡層的選手高了 0.12667% ~ 0.78724%。

* 平均擊球初速估計值所代表的意義：

在其他自變數固定的情況下，我們有 95%的信心，平均而言平均擊球初速每增加 1MPH，加權上壘率會增加 0.32107% ~ 0.55389%。

* 平均擊球仰角所代表的意義：

在其他自變數固定，且出色擊球率為 0 的情況下，我們有 95%的信心，平均而言平均擊球仰角每增加 1°，加權上壘率會增加 -0.01577% ~ 0.15905%。由於檢定結果之 p-value > 0.05，表示此變數對 y (加權上壘率)的影響不大。

* 甜蜜點比例所代表的意義：

在其他自變數固定，且出色擊球率為 0 的情況下，我們有 95%的信心，平均而言出色擊球率每增加 1%，加權上壘率會增加 -0.07340% ~ 0.13700%。由於檢定結果之 p-value > 0.05，表示此變數對 y (加權上壘率)的影響不大。

* 出色擊球率所代表的意義：

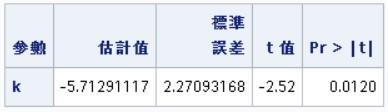
在其他自變數固定，且平均擊球仰角與甜蜜點比例皆為 0 的情況下，我們有

95%的信心，平均而言出色擊球率每增加 1%，加權上壘率會減少 0.10509% ~

0.82044%。

* 平均擊球仰角與出色擊球率的交互作用：

在其他自變數固定的情況下，平均而言平均擊球仰角每增加 1°，加權上壘率會增加

𝛽3 + 𝛾1 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 = (0.07164 − 0.01437 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒)%。平均擊球仰角的邊際效果會隨著出色擊球率的值改變而改變。

因為資料中所有打者的出色擊球率皆落在 0% ~ 25.7%之間，所以我們將出色擊球率設為 12.5%，代入𝛽3 + 𝛾1 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 中。我們有 95%的信心，平均而言平均擊球仰角每增加 1°，加權上壘率會減少 1.24545% ~ 10.1804%。

而在其他自變數固定的情況下，平均而言出色擊球率每增加 1%，加權上壘率會增加𝛽5 + 𝛾1 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔 = (−0.46276 − 0.01437 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔)% 。出色擊球率的邊際效果會隨著平均擊球仰角的值改變而改變。

* 甜蜜點比例與出色擊球率的交互作用：

在其他自變數固定的情況下，平均而言甜蜜點比例每增加 1%，加權上壘率會增加𝛽4 + 𝛾2 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 = (0.03180 + 0.02586 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒)%。甜蜜點比例的邊際效果會隨著出色擊球率的值改變而改變。

而在其他自變數固定的情況下，我們有 95%的信心，平均而言出色擊球率每增加 1%，加權上壘率會增加

𝛽5 + 𝛾2 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡 = (−0.46276 + 0.02586 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡)% 。出色擊球率的邊際效果會隨著甜蜜點比例的值改變而改變。

總結：

從以上模型檢定之 p-value 來看，對加權上壘率影響較顯著的自變數為平均擊球初速、甜蜜點比例與出色擊球率的交互作用、年齡、平均擊球仰角與出色擊球率的交互作用。我們推測可能原因如下：

* 平均擊球初速：擊球初速越高，防守球員反應時間越短，因此形成安打上壘機率較高。
* 甜蜜點比例與出色擊球率的交互作用：從結果來看，甜蜜點比例與出色擊球率的交互作用為正的，推測的原因是因為兩項數據的定義都有將擊球仰角考慮進去，且理論上出色擊球率越高的球員甜蜜點比例也會越高，因此兩者有顯著交互作用且為正的。
* 年齡：因為 25~29 歲是公認美國職棒大聯盟運動員表現最好的時期，所以

25~29 歲對加權上壘率的影響是顯著的。

* 平均擊球仰角與出色擊球率的交互作用：從結果來看，平均擊球仰角與出色擊球率的交互作用為負的，推測的原因是因為擊球仰角過高的話，就並非出色擊球，此外出色擊球的定義當中就包括擊球仰角，因此兩者有顯著交互作用且為負的。

從以上模型檢定之 p-value 來看，對加權上壘率影響較不顯著的自變數為平均擊球仰角、甜蜜點比例。我們推測可能原因如下：

* 平均擊球仰角：推測原因為擊球仰角太高或太低都不易形成安打上壘。擊球仰角要在某特定區間，才容易形成扎實的擊球形成安打上壘，因此平均擊球仰角對加權上壘率不會特別顯著。
* 甜蜜點比例：因為甜蜜點比例是針對擊球的品質去推估，但有時打得扎實的飛球會被外野手接殺出局，才會造成甜蜜點比例對平均加權上壘率的影響不顯著。