統計諮詢期末報告計畫書

第三組　洪毅荃 余振瑋 林宇軒 郭懿葶 陳志恆

一、 資料來源 ： 野球革命、中華職棒大聯盟全球資訊網

二、 資料介紹 ：

本次分析所使用之資料涵蓋 2025 年上半季中華職棒比賽數據（預計截至 2025 年 5 月 31 日），共包含兩筆資料集：

1. **隊伍逐日表現資料**：紀錄各球隊每日的比賽表現，共計 19 個變數。
   - **一般變數**：比賽日期、球隊名稱、勝負結果。
   - **打擊相關變數（共 10 項）**：包含 AVG、OPS、wOBA 等。
   - **投球相關變數（共 3 項）**：包含先發投手 ERA、非先發投手 FIP 及 WHIP。
   - **近期表現變數（共 3 項）**：包含連勝/敗紀錄、以近 10 場勝負反映隊伍的近期狀態。
2. **隊伍對戰資料**：紀錄每場比賽雙方對戰結果，共計 8 個變數。
   - **一般變數**：比賽日期與主審姓名。
   - **主場與客場隊伍資料（各 3 項）**：球隊名稱、勝負結果與該隊得分。

三、 研究目標 ：

探討影響棒球比賽勝率的關鍵因素，評估各項打擊、投球及近期表現等變數對勝率的影響力。目標為預測任意兩支球隊於未來對戰中的勝率，藉由建立中華職棒勝率模型，提供具參考價值的賽前分析依據。

四、分析方法 ：

首先，我們將使用廣義線性模型（Generalized Linear Model），以比賽勝負結果作為應變數，並加入各項打擊與投球指標作為自變數進行建模。然而，由於棒球數據中的變數（特別是打擊與投球相關指標）常具有高度共線性，因此我們將使用 Elastic Net 進行變數選擇，以兼顧 Ridge 與 Lasso 的優勢，提升模型穩定性與預測準確性。

然而，傳統以戰績或排名衡量球隊實力的方式，往往無法反映球隊間的相對實力。例如某些球隊可能出現「贏強隊、輸弱隊」的情況，使整體勝率或排名與實際實力不符。此外，由於職棒球員在各賽季的流動性高，我們僅使用當前賽季數據的資料，預測未來比賽的結果。

在考量球隊對戰配對性質與賽程不平衡（部分配對缺失）等問題下，我們進一步使用 Bradley-Terry Model（BTM）去預測棒球比賽的勝率。BTM 是設計用於估計兩兩對戰實力差異的模型，透過比賽勝負結果推估各球隊的潛在實力值，進而用以預測比賽勝率。傳統的 BTM 僅考慮勝負資料，為提升模型的預測準確度，我們將在 BTM 架構下加入更多比賽層級與球隊層級的變數，包括主客場因素、打擊與投球表現指標等。

最後，為評估各模型的預測效能，我們將透過 AUC（Area Under the ROC Curve）與交叉驗證進行模型比較，評估 GLM、基本 BTM 以及擴充變數版本 BTM 模型在預測準確性上的表現差異。