Universidad Panamericana
Maestría en Ciencia de Datos
Econometría

Proyecto Final: Determinantes de Victorias en MLB

Enrique Ulises Báez Gómez Tagle
Luis Alejandro Guillén Alvarez
Grupo2

31 de agosto de 2025

Índice

1. Introducción

1.1. Objetivo del trabajo:

El propósito de este trabajo es aplicar un análisis de regresión lineal simple para identificar y cuantificar la relación entre las victorias de un equipo de béisbol en una temporada y distintos indicadores de desempeño ofensivo y defensivo. Este análisis permite entender en qué medida factores como la diferencia de carreras, la efectividad del pitcheo (ERA) o el número de jonrones influyen en los triunfos obtenidos. De esta forma, se busca mostrar cómo las técnicas econométricas pueden emplearse para explicar y pronosticar resultados deportivos.

1.2. Justificación:

Las variables seleccionadas se eligieron por su relevancia directa en el rendimiento de un equipo de Grandes Ligas:

- **Victorias (W):** representa el desempeño global de un equipo en una temporada, es el objetivo principal a explicar.
- **Diferencia de carreras (RunDiff = R - RA):** refleja la solidez ofensiva y defensiva combinada; se espera una relación positiva con las victorias.
- **ERA (Earned Run Average):** mide la calidad del pitcheo, donde un menor valor debería asociarse con más victorias (relación negativa).
- **Jonrones (HR):** indicador clave del poder ofensivo; se espera una relación positiva con las victorias.
- **Transformación logarítmica de HR (log(HR+1)):** se incluye como forma funcional alternativa para evaluar si la relación no es estrictamente lineal.

Con este análisis se espera comprobar qué variable tiene mayor poder explicativo sobre las victorias, así como evaluar la utilidad de las transformaciones funcionales para mejorar la capacidad predictiva.

1.3. Descripción de los datos:

Se utiliza la Base de Datos de Béisbol de Lahman 1871-2024, publicada por la Society for American Baseball Research (SABR) con datos recopilados por Sean Lahman. La base está disponible en formato CSV, y específicamente se emplea el archivo `Teams.csv`, que contiene estadísticas anuales de desempeño de cada equipo de Grandes Ligas .

El archivo original incluye 48 columnas y 3075 observaciones, correspondientes a temporadas desde 1871 hasta 2024. Sin embargo, para este trabajo se decidió filtrar únicamente los equipos de las ligas Americana (AL) y Nacional (NL), ya que representan las ligas principales de las Grandes Ligas de Béisbol y permiten obtener datos más homogéneos en términos de reglas y estructura competitiva. Asimismo, se seleccionó el periodo 2000-2019 porque corresponde a una etapa reciente del béisbol moderno, con un calendario estable de 162 juegos por temporada y sin las distorsiones que generó la temporada 2020 por la pandemia de COVID-19. Con este filtro se obtuvieron 600 observaciones (30 equipos por temporada durante 20 años), lo cual asegura un tamaño de muestra suficiente para aplicar análisis con validez estadística.

El dataset maestro conserva las siguientes variables principales:

- Identificadores: `yearID`, `lgID`, `teamID`, `franchID`, `name`, `team_year`, `season_date`.
- Resultados: `W` (victorias), `L` (derrotas), `G` (juegos jugados).
- Estadísticas de desempeño: `R` (carreras anotadas), `RA` (carreras permitidas), `ERA` (efectividad), `HR` (jonrones).
- Variables derivadas: `RunDiff = R - RA`, `logHR1 = ln(HR+1)`.

El dataset es de tipo corte transversal en panel (equipo-año), con una observación por equipo por temporada, con esto es posible aplicar los modelos de regresión simple y realizar análisis descriptivos y de correlación.

2. Selección de Variables

2.1. Variable dependiente:

La variable dependiente seleccionada es el número de **victorias (W)** que obtiene cada equipo de las Grandes Ligas de Béisbol (MLB) en una temporada regular. Esta variable representa de manera directa el desempeño global de un equipo, ya que ganar más partidos es el objetivo principal dentro de una temporada. A partir de ella se busca explicar qué factores de rendimiento ofensivo y defensivo tienen mayor influencia en el éxito deportivo.

2.2. Variables independientes:

Para el análisis de regresión simple se seleccionaron tres variables distintas, cada una analizada en un modelo separado:

- **Diferencia de carreras (RunDiff = R - RA):** mide la diferencia entre las carreras anotadas (`R`) y las carreras permitidas (`RA`). Es un indicador directo del dominio de un equipo sobre sus rivales; se espera que un mayor diferencial de carreras se traduzca en un mayor número de victorias ($\beta > 0$).
- **ERA (Earned Run Average):** representa el promedio de carreras limpias permitidas por cada nueve entradas lanzadas. Es una métrica clave de la calidad del pitcheo: un valor más bajo de ERA refleja un mejor desempeño de los lanzadores y, por lo tanto, debería estar negativamente correlacionado con las derrotas y positivamente con las victorias ($\beta < 0$).
- **Jonrones (HR):** corresponde al número total de cuadrangulares conectados por un equipo en una temporada. Dado que los jonrones aportan carreras directas, se espera que tengan una relación positiva con las victorias ($\beta > 0$). Además, se incluirá una transformación funcional $\log(HR + 1)$ para evaluar si la relación entre jonrones y victorias presenta un comportamiento no lineal, suavizando el efecto de valores extremos.

3. Análisis de Estadísticas Descriptivas

3.1. Objetivo:

Resumir y describir las características principales de los datos.

3.2. Pasos:

Calcular medidas de tendencia central (media, mediana, moda).
Calcular medidas de dispersión (desviación estándar, varianza, rango intercuartílico).
Visualización de los datos (histogramas, diagramas de dispersión, boxplots) para cada variable.
Identificar la presencia de valores atípicos o extremos.

4. Análisis de Correlación

4.1. Correlación entre las variables:

Realizar un análisis de correlación para examinar las relaciones entre la variable dependiente y las variables independientes, utilizando el coeficiente de correlación de Pearson o Spearman según corresponda.

4.2. Interpretación:

Identificar si existe una relación significativa entre las variables.

5. Modelo de Regresión Simple (3 modelos)

5.1. Ajustar el modelo de regresión:

Utilizar una herramienta estadística (como Excel, R, Python) para obtener los coeficientes de la regresión.

5.2. Interpretación de los resultados:

Analizar los coeficientes obtenidos, el valor de R^2, el valor de p para cada variable independiente (para determinar si son significativas de manera individual).

5.3. Evaluación de la bondad de ajuste:

Evaluar si el modelo explica adecuadamente la variabilidad de la variable dependiente.

6. Formas Funcionales

6.1. Identificación de la forma funcional adecuada:

Dependiendo de la naturaleza de las variables, se puede considerar aplicar formas funcionales no lineales (por ejemplo, logarítmica, cuadrática, exponencial).

6.2. Transformaciones de las variables:

Si es necesario, aplicar transformaciones a las variables para mejorar el ajuste del modelo (por ejemplo, $\ln(X)$, X^2, etc.).

6.3. Validación del modelo transformado:

Volver a ajustar el modelo con las variables transformadas y comparar el ajuste con el modelo lineal original.

7. Evaluación del Modelo de Regresión

7.1. Pruebas de significancia:

Realizar las pruebas estadísticas necesarias para evaluar la significancia global del modelo (prueba F) y la significancia de cada coeficiente individual (pruebas t).

8. Pronóstico

8.1. Generación del pronóstico:

Usar el modelo ajustado para realizar predicciones de la variable dependiente.

8.2. Intervalos de predicción:

Obtener intervalos de confianza o predicción para las futuras observaciones de la variable dependiente.

8.3. Evaluación del pronóstico:

Comparar las predicciones con los valores reales (si están disponibles) utilizando medidas de error como el MSE (Error Cuadrático Medio), RMSE (Raíz del MSE), y el MAE (Error Absoluto Medio).

9. Conclusiones

9.1. Resumen de los hallazgos:

Resumir los resultados obtenidos del análisis de regresión, incluyendo la relación entre las variables y la efectividad del modelo.

9.2. Recomendaciones:

Si es aplicable, proporcionar recomendaciones basadas en los resultados del análisis de regresión.

10. Bibliografía

Incluir todas las referencias bibliográficas utilizadas durante la investigación y el análisis, como libros, artículos de investigación, y manuales de herramientas estadísticas.

11. Anexo

11.1. Link al repositorio con código fuente y salidas correspondientes

https://github.com/enriquegomeztagle/MCD-ProyectoFinalEconometria-DeterminantesVictoriasMLB