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. Medidas de tendencia central y dispersión:

A partir del dateset maestro con las varibles y observaciones seleccionadas, se calculan las siguientes estadísticas descriptivas:

Variable	Count	Mean	Median	Mode	Std	Var	Min	Q1	Q3	IQR	Max
W	600	80.97	81.00	86.00	11.79	138.92	43.00	72.00	90.00	18.00	116.00
RunDiff	600	0.00	2.00	54.00	111.11	12344.79	-337.00	-87.00	81.25	168.25	300.00
ERA	600	4.25	4.21	4.01	0.53	0.29	2.94	3.86	4.60	0.74	5.71
HR	600	173.47	170.00	161.00	36.87	1359.12	91.00	148.00	199.00	51.00	307.00
logHR1	600	5.14	5.14	5.09	0.21	0.05	4.52	5.00	5.30	0.29	5.73
R	600	740.67	735.00	735.00	83.21	6924.21	513.00	684.00	795.25	111.25	978.00
RA	600	740.67	733.00	715.00	88.93	7909.25	525.00	676.75	804.00	127.25	981.00
G	600	161.96	162.00	162.00	0.31	0.10	161.00	162.00	162.00	0.00	163.00
L	600	80.97	80.50	76.00	11.76	138.34	46.00	72.00	90.00	18.00	119.00

A continuación se detallan las características principales:

- **Victorias (W):** En promedio los equipos ganan 81 juegos por temporada , con una desviación estándar de 11.8. El rango va de 43 a 116 victorias, lo que refleja tanto equipos altamente competitivos como equipos en el extremo opuesto.

- **Diferencia de carreras (RunDiff):** Tiene media cercana a cero, que sería esperado en un balance global de liga, pero una alta dispersión ($\approx$ 111, rango de -337 a +300). Esto muestra que algunos equipos dominan ampliamente a sus rivales mientras otros son ampliamente superados.
- **ERA (Efectividad del pitcheo):** Promedia 4.25, con valores típicos entre 3.9 y 4.6 (IQR = 0.74). La dispersión es moderada y refleja diferencias en la calidad del pitcheo entre equipos, con casos extremos desde 2.94 hasta 5.71.
- **Jonrones (HR):** Los equipos conectan en promedio 173 cuadrangulares por temporada, con un rango entre 91 y 307. Esta variabilidad se ve reflejada por las distintas filosofías ofensivas.
- **Transformación logarítmica (logHR1):** Reduce la dispersión (≈ 0.21) y comprime la escala, aunque en este rango de valores la distribución sigue un patrón casi lineal respecto a HR.
- **Carreras anotadas (R) y recibidas (RA):** Ambas variables tienen media ≈ 741, lo que es natural dado el equilibrio de la liga. Su dispersión ($\approx 83-89$) muestra diferencias en ofensiva y defensiva entre equipos.
- **Juegos (G):** Es prácticamente constante en 162, como dicta el calendario, con variaciones mínimas por suspensiones o ajustes.
- **Derrotas (L):** Presentan la misma estructura que las victorias, con media 81 y desviación de 11.7, dada la relación $W + L \approx 162$.

3.2. Visualización de los datos:

Con el fin de comprender mejor el comportamiento de las variables y su relación con las victorias, se generaron distintas visualizaciones: histogramas, boxplots, diagramas de dispersión, gráficas de pastel y una serie de tiempo de ejemplo.

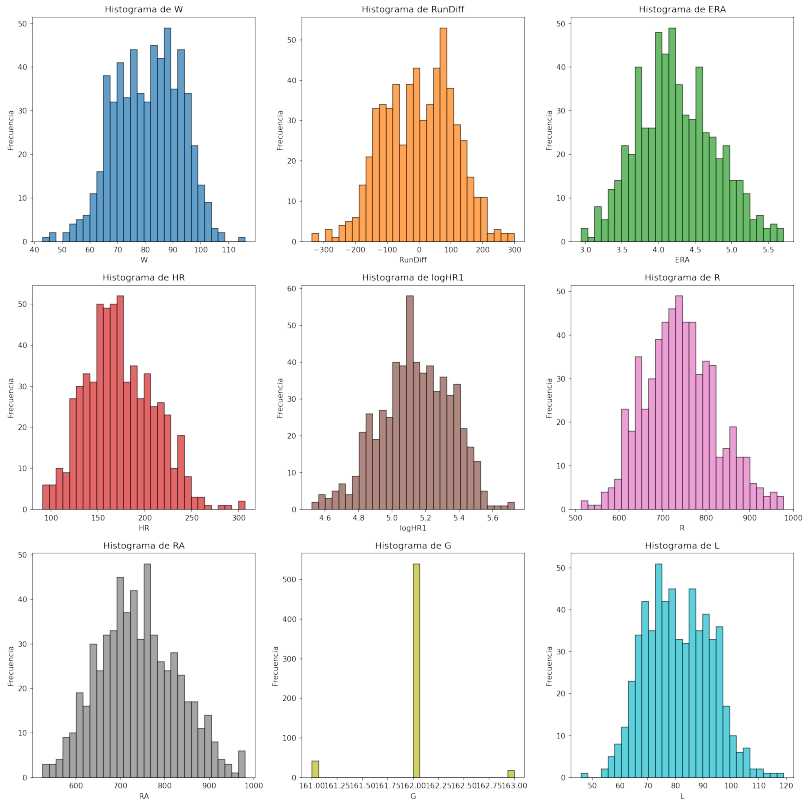

Figura 1: Histogramas de W, RunDiff, ERA, HR, logHR1, R, RA, G y L (2000–2019).

Los histogramas confirman distribuciones aproximadamente simétricas en **W** y **L**, con centro en 81 victorias/derrotas. **RunDiff** muestra gran dispersión, confirmarndo que algunos equipos superan a sus rivales por cientos de carreras, mientras otros son ampliamente superados. **ERA** se concentra en torno a 4, reflejando diferencias moderadas en pitcheo. **HR** se distribuye entre 100-300, y su transformación **logHR1** comprime los valores extremos, suavizando colas. **R** y **RA** tienen formas parecidas, centradas cerca de 740, lo que refleja equilibrio ofensivo-defensivo en la liga. **G** es casi una constante en 162, validando la homogeneidad del calendario.

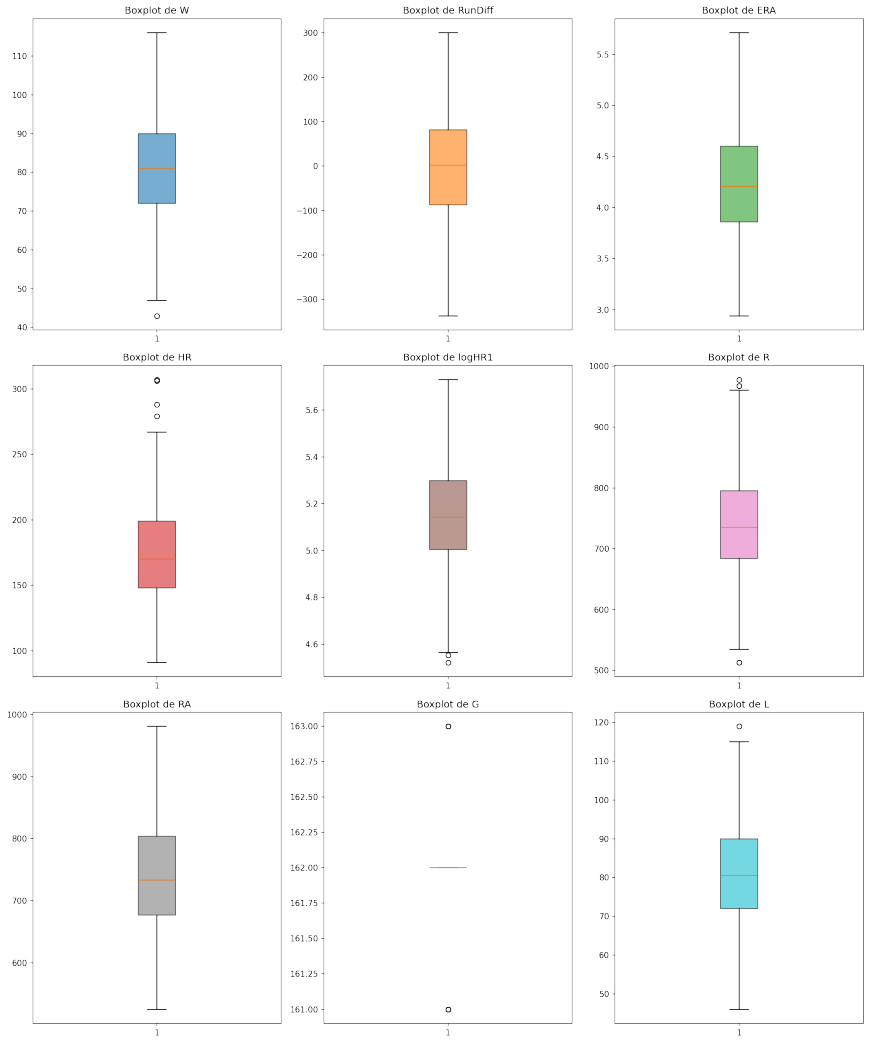

Figura 2: Boxplots por variable: dispersión, mediana y valores atípicos.

Los boxplots identifican *outliers* relevantes: (i) en **HR**, equipos con poder ofensivo distintivo (300+ HR), (ii) en **R** se ven reflejados esos mismos casos extremos de producción ofensiva. (iii) en **G**, ligeras desviaciones (161 o 163 partidos), explicadas por suspensiones o dobles juegos. En **RunDiff** se observan extremos tanto positivos como negativos, reflejando temporadas históricas dominantes o muy pobres.

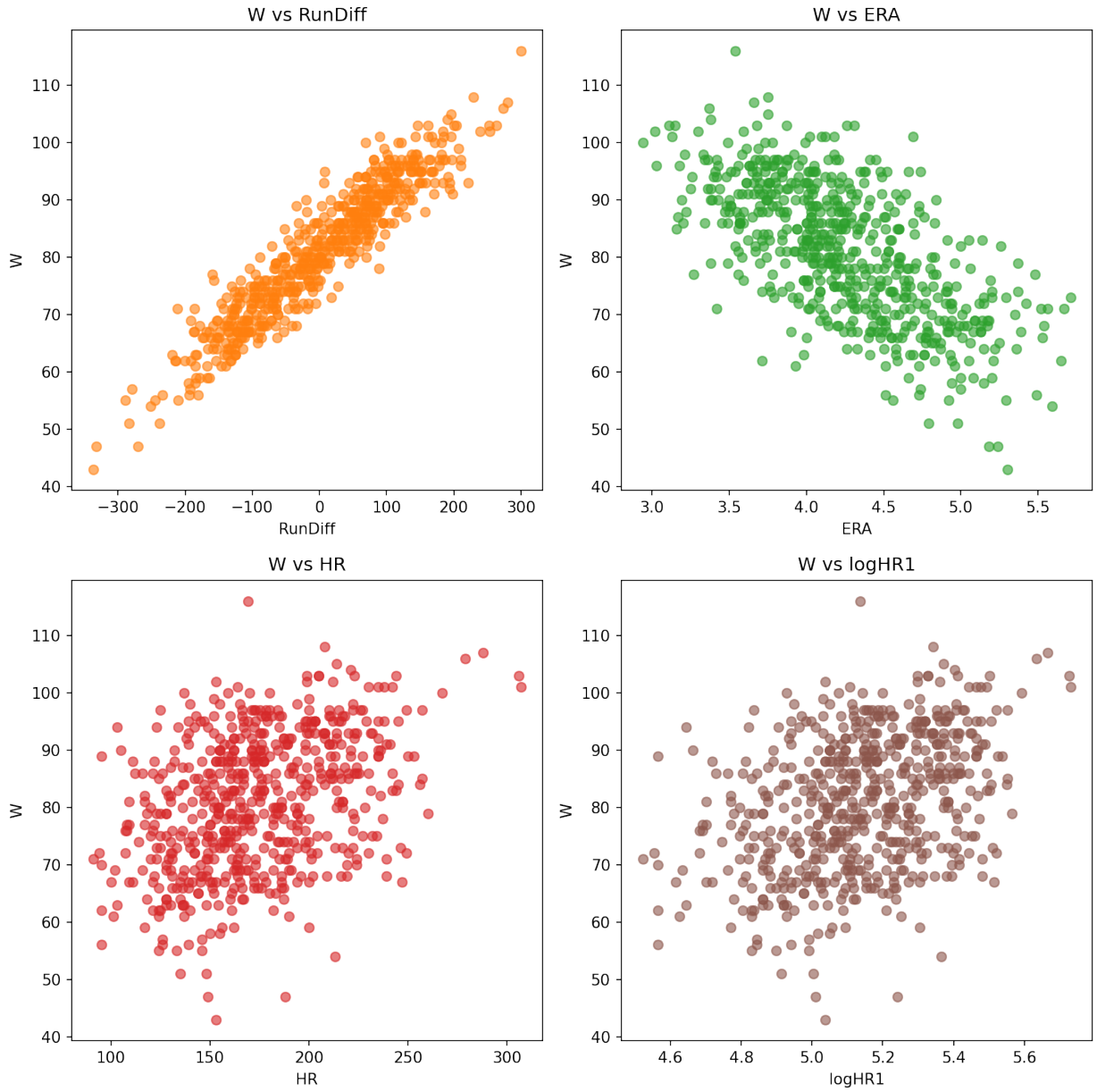

Figura 3: Diagramas de dispersión: W vs RunDiff, ERA, HR y log(HR+1).

W vs RunDiff presenta la relación más fuerte y lineal: a mayor diferencial de carreras, más victorias, confirmando su validez como predictor central. **W vs ERA** muestra una relación negativa clara: equipos con menor efectividad del pitcheo (ERA baja) ganan más. **W vs HR** y **W vs logHR1** tienen asociación positiva pero más difusa; los cuadrangulares ayudan a ganar, aunque con variabilidad, lo cual lleva a explorar transformaciones.

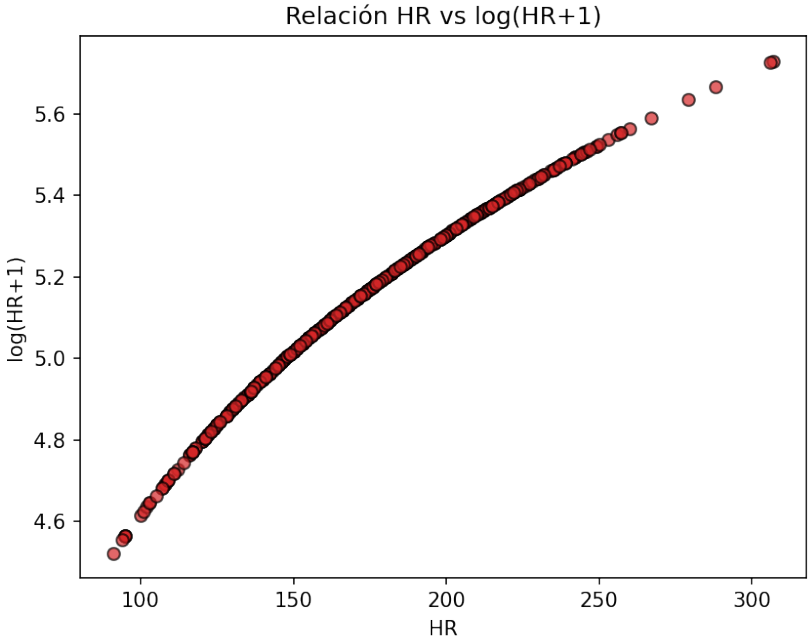

Figura 4: Relación funcional entre HR y log(HR+1).

La curva muestra que $\log(HR+1)$ suaviza el crecimiento de los HR. Aunque en el rango 90-300 se mantiene casi lineal, la transformación previene que valores extremos dominen el ajuste del modelo, haciendo el análisis más robusto.

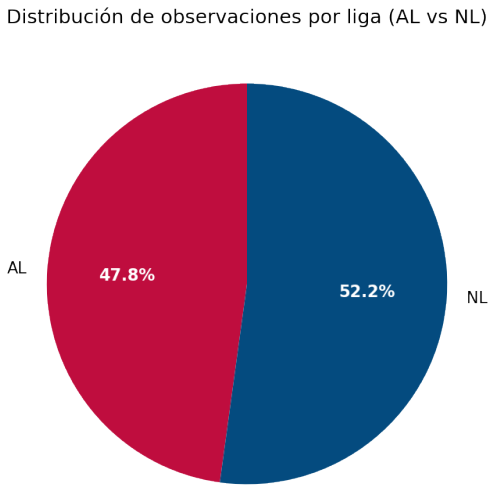

Figura 5: Distribución de observaciones por liga (AL vs NL), 2000–2019.

El dataset está balanceado entre **NL** (52.2 %) y **AL** (47.8 %),y con esto se garantiza representatividad de ambas ligas, sin sesgos por desbalance en la muestra.

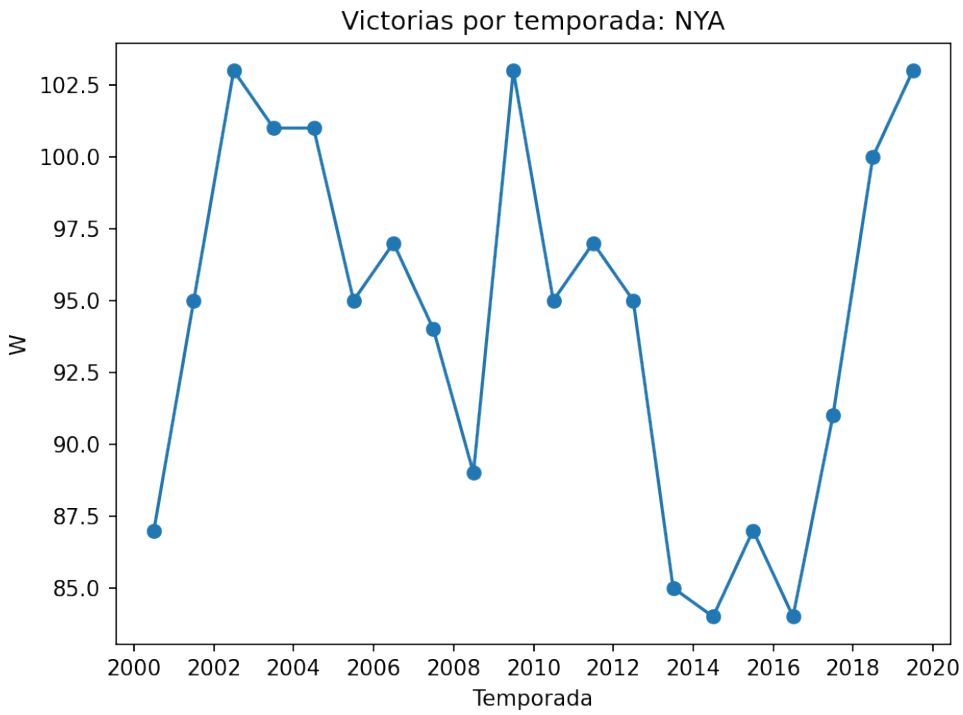

Figura 6: Serie de tiempo de victorias (ejemplo: NYA), 2000–2019.

Los Yankees de Nueva York (NYA) ilustran la variabilidad interanual en victorias. Se observan picos de más de 100 triunfos en varias temporadas y caídas a la franja de 85-90 victorias en otras. Este patrón muestra que incluso equipos consistentemente competitivos presentan fluctuaciones naturales, útiles para entender la estabilidad del modelo a lo largo del tiempo.

3.3. Identificación de valores atípicos:

El análisis mediante el método del rango intercuartílico (IQR) permitió identificar observaciones atípicas en varias variables:

- **Victorias (W):** El caso más extremo corresponde a los Detroit Tigers en 2003, con solo 43 victorias, claramente fuera del rango intercuartílico (45–117). Esto refleja una de las peores campañas en la historia reciente de MLB.
- **Jonrones (HR):** En 2019 se detectaron valores extraordinariamente altos en equipos como Minnesota Twins (307), New York Yankees (306), Houston Astros (288) y Los Angeles Dodgers (279), todos por encima del umbral superior (275.5). Esto coincide con el "Año del jonrón". en 2019, cuando se registró un récord colectivo histórico de cuadrangulares.
- **Transformación log(HR+1):** Aunque la mayoría de observaciones están dentro del rango, aparecen valores bajos en equipos con ofensivas débiles como los San Diego Padres (2011) y San Francisco Giants (2008), lo que confirma que esta transformación ayuda a suavizar pero no elimina del todo los outliers.
- **Carreras anotadas (R):** Se identifican equipos con valores extremos, por ejemplo, los Yankees (2007) y Rockies (2000) con más de 968 carreras, y los Marlins (2013) o Mariners (2010) con apenas 513, fuera del rango esperado (517–962).
- **Juegos disputados (G):** Aunque la liga establece un calendario de 162 juegos, se detectan temporadas con 161 o 163 partidos, resultado de suspensiones o reprogramaciones (e.g., Cubs 2009, Rockies 2007, Rangers 2013).

- **Derrotas (L):** Nuevamente destacan los Tigers de 2003, con 119 derrotas, simétrico al outlier en victorias.

Estos valores atípicos no necesariamente representan errores de medición, sino hechos históricos del béisbol (equipos en un muy bajo nivel, ofensivas históricas, o particularidades del calendario). Sin embargo, es importante tenerlos en cuenta porque pueden influir en el ajuste de los modelos de regresión, afectando los coeficientes e incrementando la dispersión residual.

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

Lahman Baseball Database -Society for American Baseball Research. (2025). Sabr.org. https://sabr.org/lahman-database/

11. Anexo

11.1. Link al repositorio con código fuente y salidas correspondientes

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