Universidad Panamericana

Maestría en Ciencia de Datos

Econometría

Proyecto Final: Determinantes de Victorias en MLB

Enrique Ulises Báez Gómez Tagle y Luis Alejandro Guillén Alvarez

Grupo 2

Universidad Panamericana

Universidad Panamericana

Maestría en Ciencia de Datos

Econometría

Proyecto Final: Determinantes de Victorias en MLB

Índice

Introducción

Objetivo del trabajo:

El propósito de este trabajo es aplicar un análisis de **regresión lineal múltiple** para explicar y pronosticar el número de **victorias (W)** de un equipo de Grandes Ligas en una temporada a partir de *múltiples* variables explicativas consideradas de forma *simultánea*. El modelo tendrá la forma general

$$W_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i,$$

donde X_1, X_2, X_3 representan predictores de desempeño ofensivo y defensivo (`RunDiff`, `ERA`, `HR`). Los objetivos específicos son: (i) cuantificar los *efectos marginales* de cada predictor sobre W; (ii) evaluar su significancia estadística y la bondad de ajuste del modelo; y (iii) generar pronósticos con intervalos de predicción para nuevas observaciones.

Justificación:

Se adopta una especificación **múltiple** porque permite estimar el efecto *parcial* de cada variable sobre W controlando por las demás, reduciendo el sesgo por omisión inherente a modelos univariados. En particular:

- **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.

Este enfoque permite comparar la *relevancia relativa* de los predictores (magnitud y significancia), mejorar el desempeño predictivo frente a especificaciones simples y verificar rigurosamente los supuestos del modelo (linealidad conjunta, independencia,

homoscedasticidad, normalidad de residuos), además de diagnosticar *multicolinealidad* mediante VIF.

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 panel (equipo–año), con una observación por equipo por temporada; con esto es posible ajustar y evaluar un modelo de regresión múltiple, además de realizar análisis descriptivos y de correlación.

Selección de Variables

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.

Variables independientes:

Para el análisis de regresión múltiple se consideran, de forma simultánea, tres variables explicativas:

- **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 ($\beta > 0$).

Análisis de Estadísticas Descriptivas

Medidas de tendencia central y dispersión:

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

Cuadro 1

Estadísticas Descriptivas del dataset

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 (≈ 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 $W + L \approx 162$.

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.

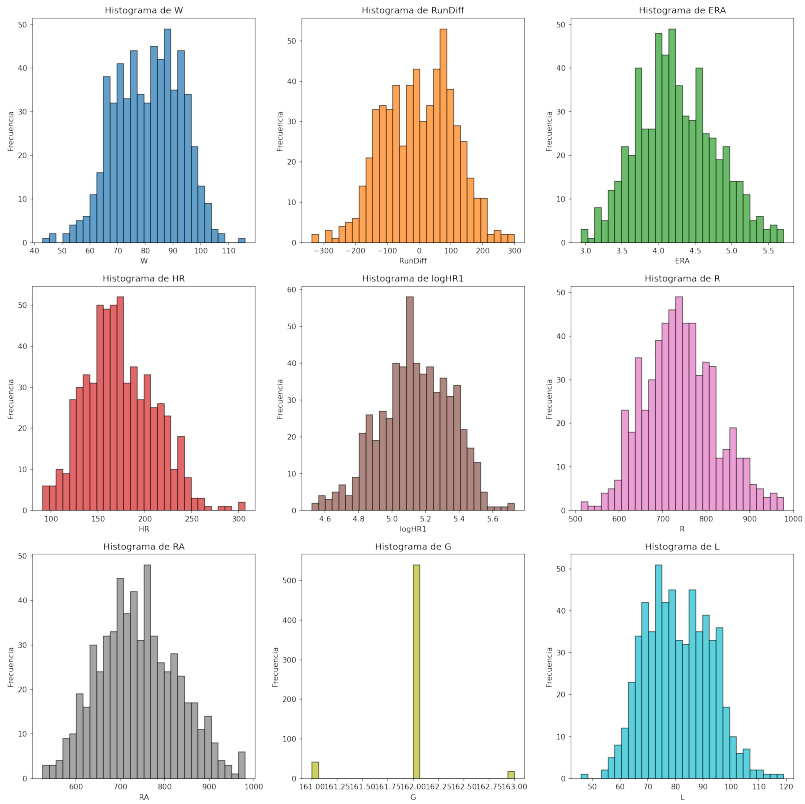

Figura 1

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.

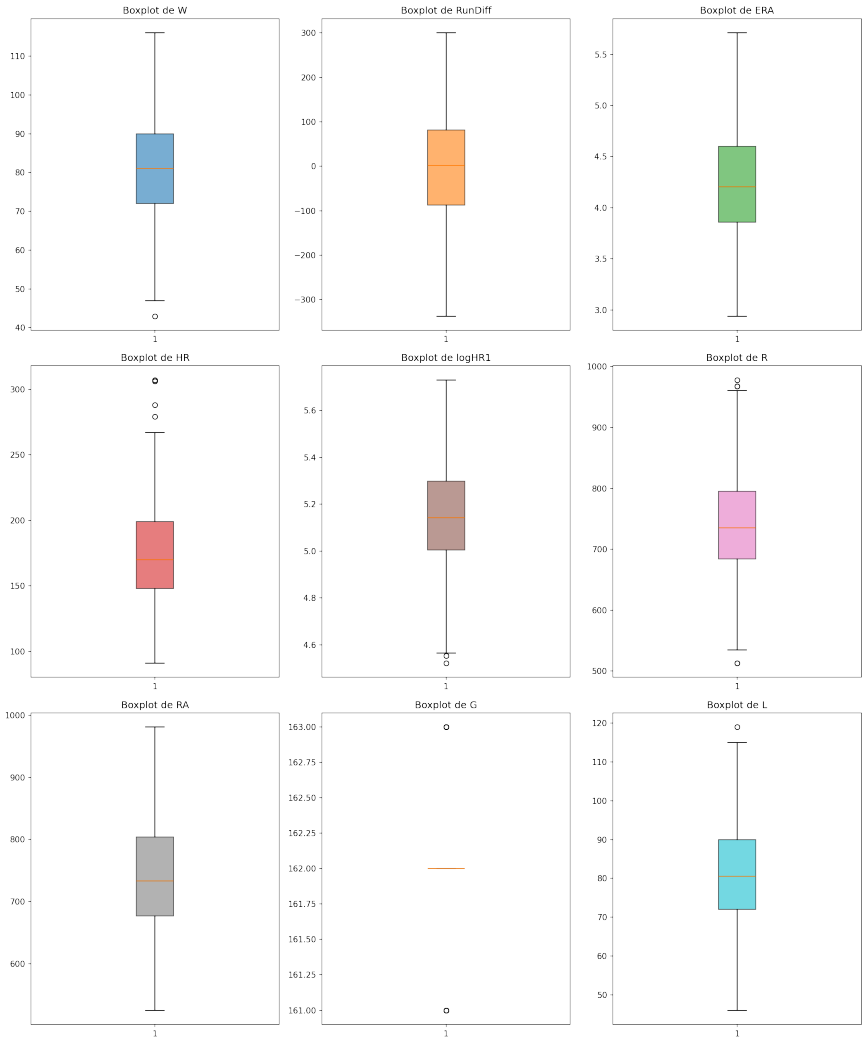

Figura 2

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.

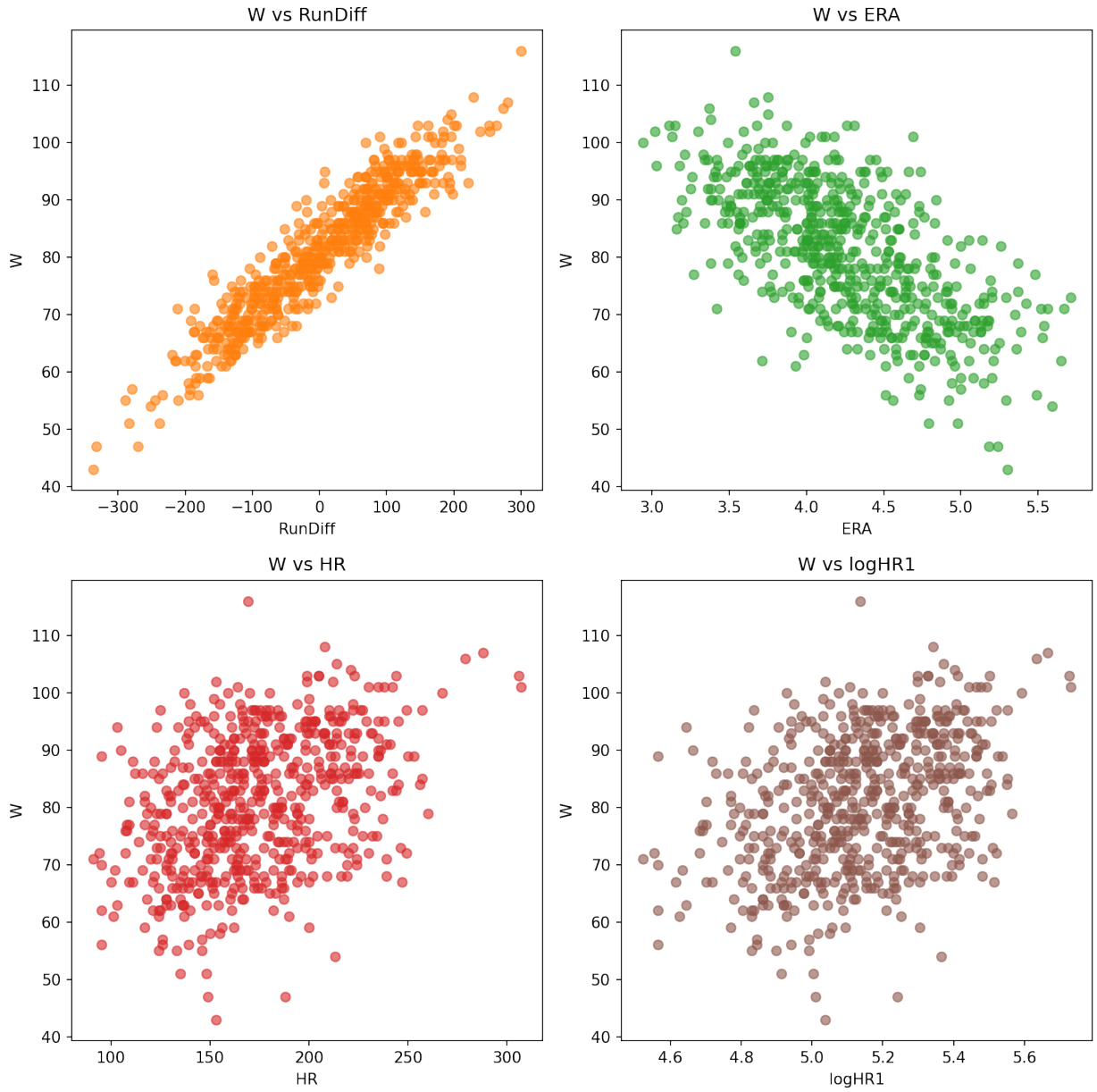

Figura 3

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.

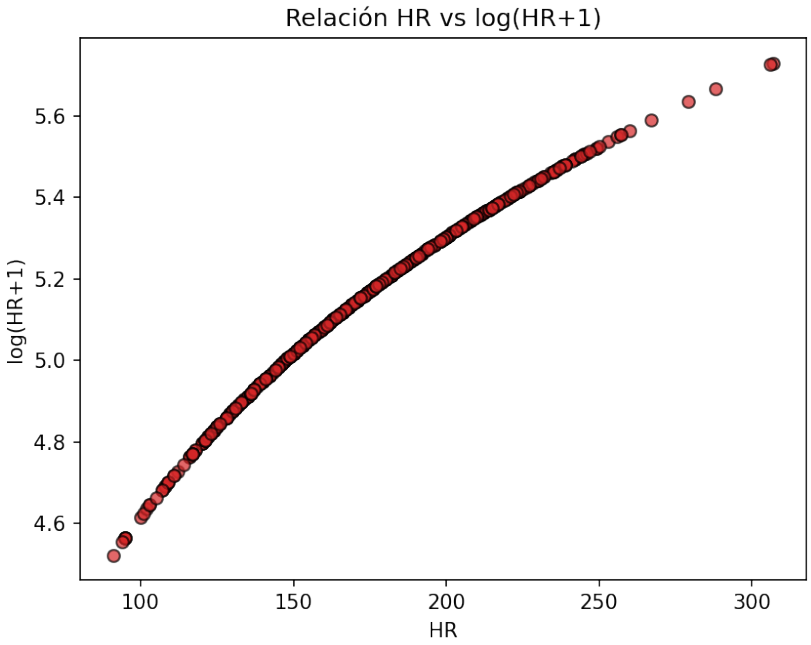

Figura 4

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.

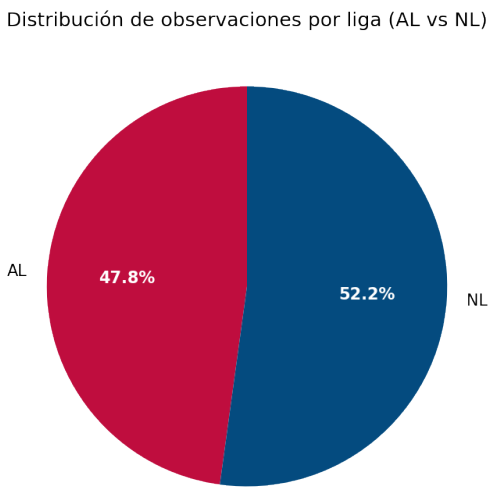

Figura 5

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.

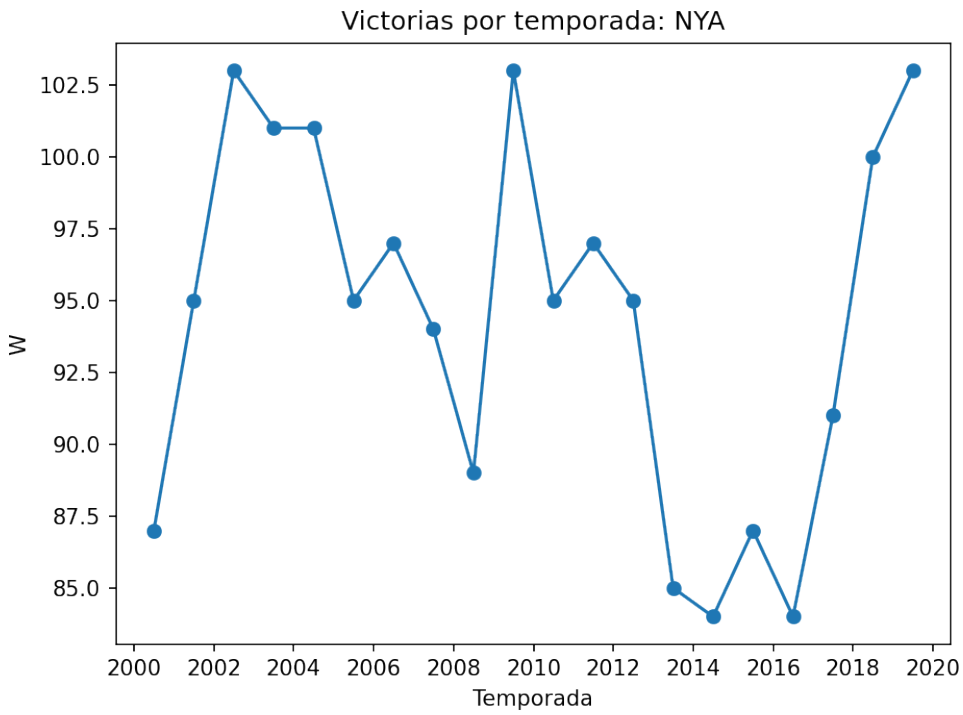

Figura 6

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.

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.

Análisis de Correlación

Correlación entre las variables:

Se calculó el coeficiente de correlación de Pearson y Spearman entre el número de victorias (W) y las variables explicativas (RunDiff, ERA, HR y logHR1). Los resultados se resumen en la Tabla 2 y en las Figuras 8–9.

Cuadro 2

Correlación de W con variables explicativas

Variable	Pearson r	p (Pearson)	Spearman ρ	p (Spearman)	N
RunDiff	0.9395	0.0000	0.9398	0.0000	600
HR	0.3920	0.0000	0.3853	0.0000	600
logHR1	0.3901	0.0000	0.3853	0.0000	600
ERA	-0.6554	0.0000	-0.6575	0.0000	600

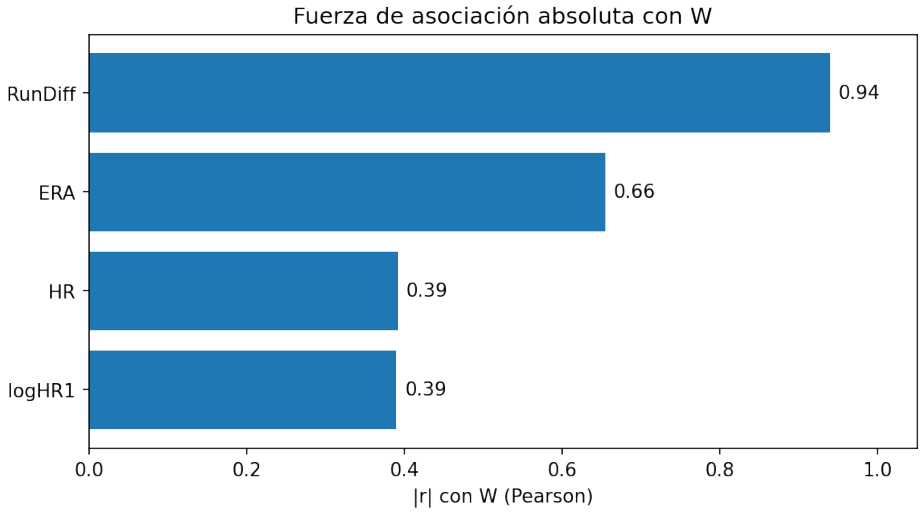

Figura 7

Fuerza de asociación absoluta (|r|) de cada predictor con W (Pearson).

Interpretación de los resultados

- **RunDiff** presenta la correlación más fuerte con $W (r = 0.94, p < 0.001)$, lo que confirma que la diferencia de carreras es un predictor casi determinístico del número de victorias.

- **ERA** muestra una correlación negativa alta $(r = -0.66, p < 0.001)$. Esto indica que un menor promedio de carreras limpias permitidas (mejor pitcheo) está fuertemente asociado con más victorias.

- **HR** y su transformación **logHR1** exhiben correlaciones positivas moderadas $(r \approx 0.39, p < 0.001)$. Los jonrones ayudan a ganar partidos, aunque no explican tanto como RunDiff o ERA. La similitud entre HR y logHR1 confirma que la transformación logarítmica apenas cambia la relación en el rango observado (90–300 jonrones).

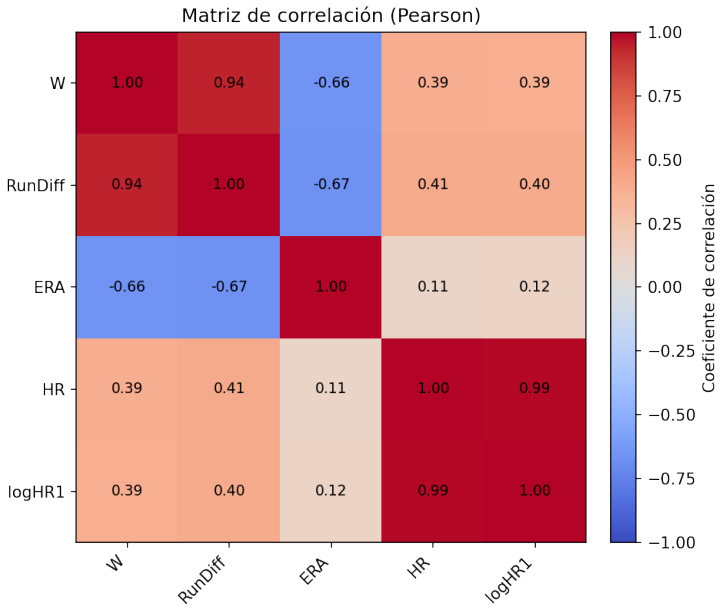

Figura 8

Matriz de correlación (Pearson) entre W y las variables explicativas.

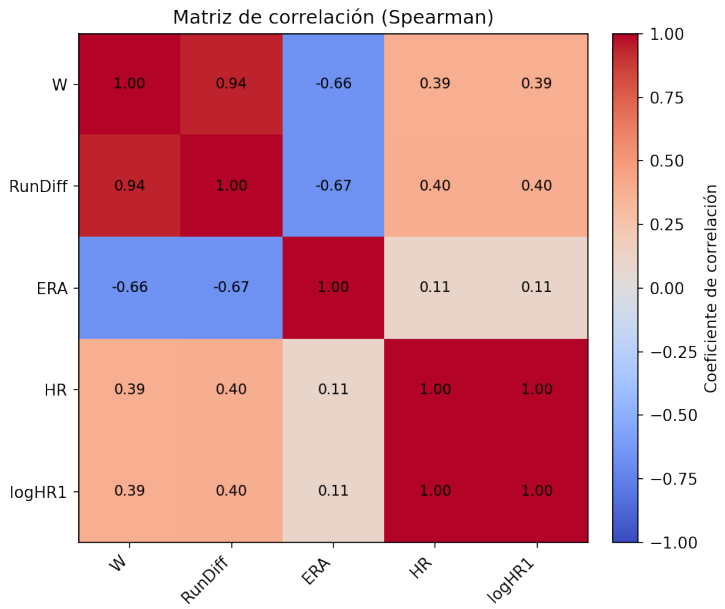

Figura 9

Matriz de correlación (Spearman) entre W y las variables explicativas.

Tanto Pearson como Spearman producen resultados consistentes: RunDiff y ERA son los predictores más fuertemente asociados con las victorias, mientras que los jonrones tienen un efecto más limitado. Esto anticipa que los modelos de regresión simple con RunDiff y ERA tendrán mayor poder explicativo que aquellos con HR.

Modelo de Regresión Simple (3 modelos)

Resultados y resumen de estimaciones

Se ajustaron cuatro modelos de regresión lineal simple con **W** (victorias por temporada) como variable dependiente y, por separado, como variables explicativas: **RunDiff**, **ERA**, **HR** y $\log(\mathbf{HR}+1)$. Los resultados completos se muestran en la Tabla 3:

Cuadro 3

Modelos de regresión lineal simple: coeficientes y bondad de ajuste

Modelo	β_0	β_1	$p(\beta_1)$	CI95 % Inf	CI95 % Sup	R^2	R^2_{adj}	F	$p(F)$	AIC	BIC	RMSE	MAE	N
W RunDiff	80.9700	0.0997	0.0000	0.0967	0.1026	0.8827	0.8825	4498.4794	0.0000	3380.4384	3389.2322	4.0340	3.2120	600
W ERA	142.3221	-14.4426	0.0000	-15.7791	-13.1060	0.4296	0.4286	450.3684	0.0000	4329.2274	4338.0212	8.8944	7.1557	600
W HR	59.2303	0.1253	0.0000	0.1017	0.1489	0.1537	0.1522	108.5754	0.0000	4565.9635	4574.7574	10.8341	8.9117	600
W logHR1	-29.3238	21.4615	0.0000	17.3937	25.5292	0.1522	0.1508	107.3631	0.0000	4566.9939	4575.7878	10.8434	8.9306	600

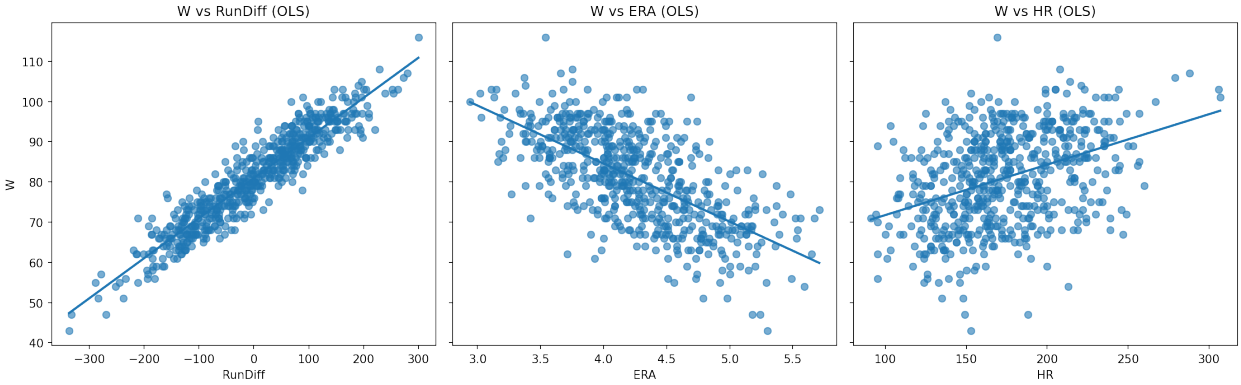

Figura 10

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

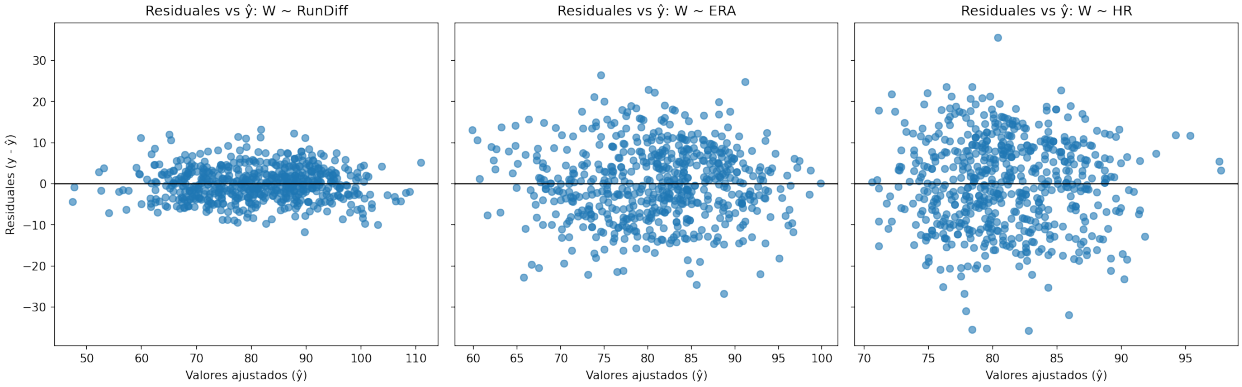

Figura 11

Residuales vs valores ajustados: W vs RunDiff, ERA, HR y log(HR+1).

Interpretación de los resultados

W ~ RunDiff. Este modelo presenta un ajuste sobresaliente: $R^2 = 0.883$, con un error medio de apenas 4 juegos (RMSE = 4.03, MAE = 3.21). El coeficiente estimado de

0.0997 implica que por cada 10 carreras de diferencia anotadas sobre el rival, un equipo gana en promedio una victoria adicional. El intercepto de 80.97 refleja que un equipo "neutral" (RunDiff=0) tiende a terminar con 81 victorias, lo cual coincide con el balance teórico de .500. Este es, con diferencia, el modelo más predictivo.

W ∼ ERA.. La relación entre la efectividad de los lanzadores y las victorias también es significativa ($p < 0.001$), con una pendiente de -14.44: reducir en 1.00 la ERA equivale aproximadamente a 14 victorias adicionales. Aunque el ajuste es más modesto que con RunDiff ($R^2 = 0.43$), este modelo ofrece una visión valiosa del rol del pitcheo en el éxito global del equipo.

W ∼ HR.. Cada 10 jonrones se asocian con 1.25 victorias adicionales. Sin embargo, el poder explicativo es limitado ($R^2 = 0.154$, RMSE=10.83). Esto refleja que, aunque los cuadrangulares ayudan, existen múltiples caminos ofensivos para anotar que no se capturan solo con HR.

W ∼ log(HR + 1). La transformación logarítmica produce resultados prácticamente idénticos a HR lineal ($R^2 = 0.152$, RMSE=10.84). Esto confirma lo observado en el análisis exploratorio: en el rango de 90–300 HR, la relación con victorias es casi lineal y la transformación no mejora el ajuste.

Evaluación de la bondad de ajuste

- **Mejor modelo:** RunDiff domina en todos los indicadores (R^2, AIC, BIC, RMSE, MAE).
- **Modelo intermedio:** ERA es una métrica útil, con un ajuste razonable, que complementa la interpretación sabermétrica con una perspectiva puramente de pitcheo.
- **Modelos débiles:** HR y $\log(HR + 1)$ muestran asociaciones significativas pero débiles, explicando solo un 15 % de la variabilidad de W.
- **Pruebas de significancia:** En todos los casos, la prueba F global confirma que los modelos son estadísticamente significativos ($p < 0.001$).

- **Error de predicción:** Con RunDiff, el error típico es de ≈ 3–4 victorias por temporada; con ERA, sube a ≈ 7–9; con HR, a casi ≈ 9–11.

Bajo este análisis, se puede concluir que el diferencial de carreras es el mejor predictor de victorias. Sin embargo, ERA aporta información clave sobre la importancia del pitcheo, mientras que los jonrones reflejan solo una parte limitada de la ofensiva.

Formas Funcionales

Para cada predictor probamos una forma lineal y una alternativa no lineal:

1. **HR** vs. $\log(\text{HR} + 1)$;

2. **ERA** lineal vs. cuadrática (ERA^2);

3. **RunDiff** lineal vs. cuadrática (RunDiff^2).

La comparación se basó en R^2, R^2_{adj}, AIC/BIC, RMSE in–sample y RMSE de validación cruzada (10-fold), además de diagnósticos Breusch-Pagan (heterocedasticidad) y RESET de Ramsey (no linealidades u omisiones).

Cuadro 4

Comparación de formas funcionales por predictor

Modelo	k	R^2	R^2_{adj}	AIC	BIC	RMSE	MAE	RMSE_{CV} (media)	RMSE_{CV} (sd)	BP F	p(BP)	RESET F	p(RESET)	N
W HR	1	0.1537	0.1522	4565.9635	4574.7574	10.8341	8.9117	10.8372	0.6868	0.6344	0.4261	0.3085	0.7346	600
W log(HR+1)	1	0.1522	0.1508	4566.9939	4575.7878	10.8434	8.9306	10.8479	0.6725	0.3728	0.5417	0.5937	0.5526	600
W ERA	1	0.4296	0.4286	4329.2274	4338.0212	8.8944	7.1557	8.8994	0.4535	2.6423	0.1046	4.1149	0.0168	600
$W \sim ERA + ERA^2$	2	0.4296	0.4277	4331.2180	4344.4088	8.8943	7.1572	8.9189	0.4407	2.4197	0.0898	5.4130	0.0047	600
W RunDiff	1	0.8827	0.8825	3380.4384	3389.2322	4.0340	3.2120	4.0259	0.4107	0.0072	0.9322	3.5185	0.0303	600
$W \sim \text{RunDiff} + \text{RunDiff}^2$	2	0.8835	0.8831	3378.1195	3391.3103	4.0195	3.1946	4.0256	0.3990	0.0806	0.9226	1.6484	0.1932	600

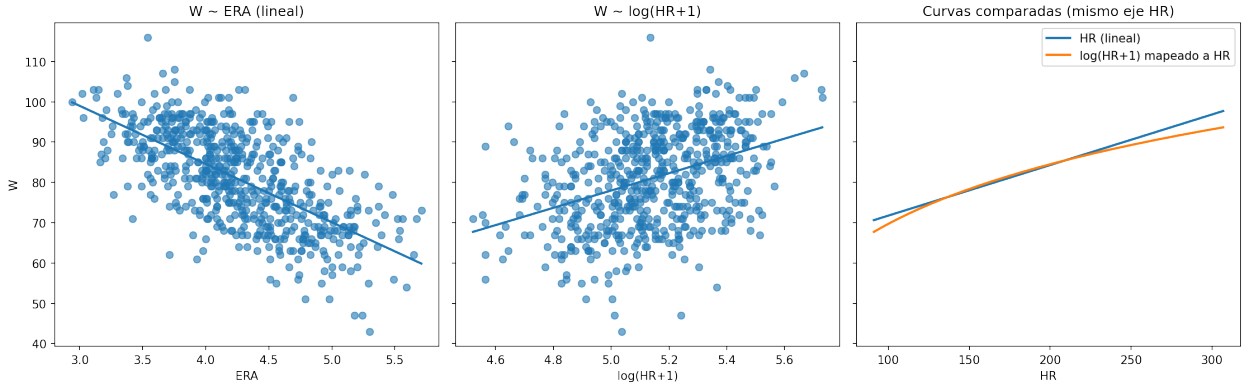

Figura 12

HR: comparación lineal vs. $\log(HR+1)$.

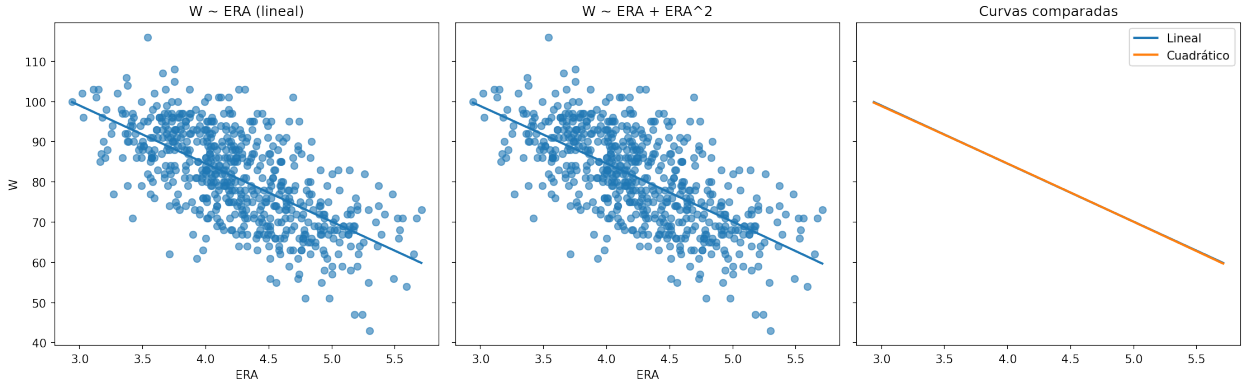

Figura 13

ERA: comparación lineal vs. cuadrática.

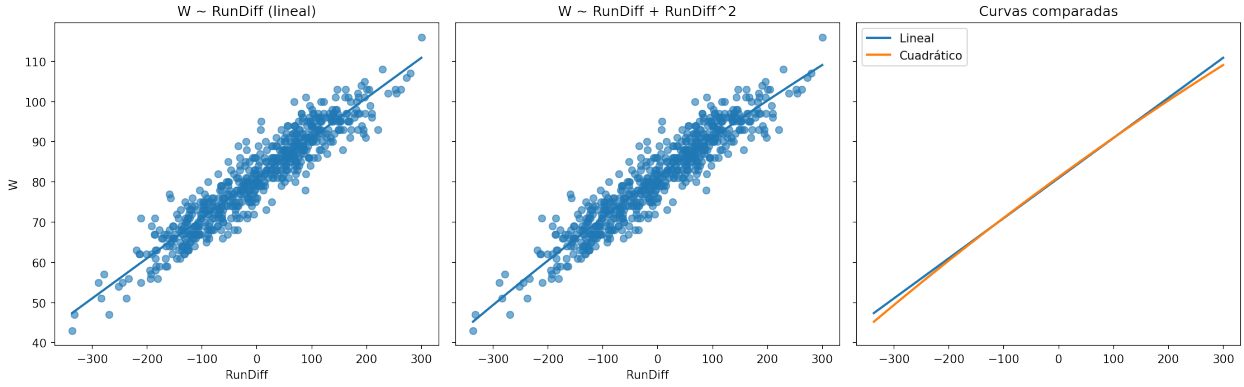

Figura 14

RunDiff: comparación lineal vs. cuadrática.

Resultados e interpretación

HR ($W \sim HR$ vs. $W \sim \log(\mathbf{HR}+1)$). Las métricas son prácticamente idénticas entre la forma lineal y la logarítmica: $R^2 \approx 0.153$, RMSE ≈ 10.84 y AIC/BIC casi iguales. Ni el test RESET (p=0.59 para log y p=0.31 para lineal) ni Breusch–Pagan (p$\approx 0.37-0.54$) sugieren problemas de especificación u heterocedasticidad. *Conclusión:* la transformación $\log(\mathrm{HR}+1)$ **no aporta mejora** sustantiva; mantenemos **HR lineal** como especificación base y usamos la versión logarítmica sólo como contraste funcional.

ERA ($W \sim \mathbf{ERA}$ vs. $W \sim \mathbf{ERA} + \mathbf{ERA}^2$). El término cuadrático no mejora el ajuste: R^2 se mantiene (0.4296 vs. 0.4296), AIC/BIC *empeoran* levemente con el cuadrático, y el RMSE-CV también es ligeramente mayor (8.92 vs. 8.90). RESET es significativo en ambas (p=0.0168 lineal; p=0.0047 cuadrática), por lo que la curvatura de segundo grado **no** corrige del todo la posible omisión de forma o variables. BP no detecta heterocedasticidad (p$\approx 0.09-0.10$). *Conclusión:* **era suficiente la forma lineal** para ERA; incorporar ERA^2 no justifica la mayor complejidad.

RunDiff ($W \sim \mathbf{RunDiff}$ vs. $W \sim \mathbf{RunDiff} + \mathbf{RunDiff}^2$). Aquí sí aparece una *mejora pequeña pero consistente*: el R^2 sube de 0.8827 a 0.8835, AIC baja (3380.44 $\rightarrow$ 3378.12), RMSE disminuye (4.034 $\rightarrow$ 4.020) y el RESET deja de ser significativo (p=0.0303 $\rightarrow$ 0.193), lo que sugiere que el término cuadrático captura una leve curvatura. BP no señala heterocedasticidad. *Conclusión:* **RunDiff cuadrático** ofrece el mejor compromiso (ligera mejora de ajuste y especificación más estable), aunque el *gain* es modesto.

Validación del modelo transformado

Con base en los resultados anteriores y en las Figuras 1213–14, seleccionamos como especificaciones finales por predictor:

- $W \sim \mathrm{RunDiff} + \mathrm{RunDiff}^2$ (mejor AIC y RESET no significativo).
- $W \sim \mathrm{ERA}$ (lineal; el término cuadrático no mejora).
- $W \sim \mathrm{HR}$ (lineal; la versión logarítmica es equivalente en ajuste).

Evaluación del Modelo de Regresión

Pruebas de significancia

Tablas de resultados. A continuación se muestran (i) la significancia *global* por modelo mediante la prueba F, y (ii) la significancia *individual* de los coeficientes mediante pruebas t con intervalos de confianza al 95 %.

Cuadro 5

Significancia global por modelo (prueba F)

Modelo	Fórmula	gl (modelo,resid)	F	$p(F)$	R^2	R^2_{adj}	N
Cuadrático: W RunDiff + RunDiff²	W RunDiff + I(RunDiff**2)	(2,597)	2263.8564	0.0000	0.8835	0.8831	600
Lineal: W RunDiff	W RunDiff	(1,598)	4498.4794	0.0000	0.8827	0.8825	600
Lineal: W ERA	W ERA	(1,598)	450.3684	0.0000	0.4296	0.4286	600
Lineal: W HR	W HR	(1,598)	108.5754	0.0000	0.1537	0.1522	600
Lineal: W log(HR+1)	W logHR1	(1,598)	107.3631	0.0000	0.1522	0.1508	600

Cuadro 6

Pruebas de significancia: F global y t por coeficiente

Modelo	Fórmula	gl (modelo,resid)	F	$p(F)$	Término	β	EE(β)	t	$p(t)$	CI95 % inf	CI95 % sup	Signif. 5 %	R^2	R^2_{adj}	N
Lineal: W RunDiff	W RunDiff	(1,598)	4498.4794	0.0000	Intercept	80.9700	0.1650	490.8379	0.0000	80.6460	81.2940	Sí	0.8827	0.8825	600
Lineal: W RunDiff	W RunDiff	(1,598)	4498.4794	0.0000	RunDiff	0.0997	0.0015	67.0707	0.0000	0.0967	0.1026	Sí	0.8827	0.8825	600
Lineal: W ERA	W ERA	(1,598)	450.3684	0.0000	Intercept	142.3221	2.9138	48.8446	0.0000	136.5996	148.0446	Sí	0.4296	0.4286	600
Lineal: W ERA	W ERA	(1,598)	450.3684	0.0000	ERA	-14.4426	0.6806	-21.2219	0.0000	-15.7791	-13.1060	Sí	0.4296	0.4286	600
Lineal: W HR	W HR	(1,598)	108.5754	0.0000	Intercept	59.2303	2.1329	27.7702	0.0000	55.0415	63.4191	Sí	0.1537	0.1522	600
Lineal: W HR	W HR	(1,598)	108.5754	0.0000	HR	0.1253	0.0120	10.4200	0.0000	0.1017	0.1489	Sí	0.1537	0.1522	600
Lineal: W log(HR+1)	W logHR1	(1,598)	107.3631	0.0000	Intercept	-29.3238	10.6537	-2.7525	0.0061	-50.2470	-8.4006	Sí	0.1522	0.1508	600
Lineal: W log(HR+1)	W logHR1	(1,598)	107.3631	0.0000	logHR1	21.4615	2.0712	10.3616	0.0000	17.3937	25.5292	Sí	0.1522	0.1508	600
Cuadrático: W RunDiff + RunDiff²	W RunDiff + I(RunDiff**2)	(2,597)	2263.8564	0.0000	Intercept	81.2425	0.2104	386.1012	0.0000	80.8292	81.6557	Sí	0.8835	0.8831	600
Cuadrático: W RunDiff + RunDiff²	W RunDiff + I(RunDiff**2)	(2,597)	2263.8564	0.0000	RunDiff	0.0994	0.0015	66.8554	0.0000	0.0965	0.1023	Sí	0.8835	0.8831	600
Cuadrático: W RunDiff + RunDiff²	W RunDiff + I(RunDiff**2)	(2,597)	2263.8564	0.0000	I(RunDiff ** 2)	-0.0000	0.0000	-2.0767	0.0383	-0.0000	-0.0000	Sí	0.8835	0.8831	600

De las tablas anteriores, se puede concluir que:

- **Todos los modelos son globalmente significativos** (Tabla 5):
  - $W \sim \text{RunDiff} + \text{RunDiff}^2$: $F(2, 597) = 2263.86$, $p < 0.001$, $R^2 = 0.8835$, $R^2_{adj} = 0.8831$.
  - $W \sim \text{RunDiff}$: $F(1, 598) = 4498.48$, $p < 0.001$, $R^2 = 0.8827$.

  - $W \sim$ ERA: $F(1, 598) = 450.37$, $p < 0.001$, $R^2 = 0.4296$.
  - $W \sim$ HR y $W \sim \log(\text{HR} + 1)$: $F \approx 108$, $p < 0.001$, $R^2 \approx 0.153$.

- **Significancia individual de coeficientes** (Tabla 6):
  - RunDiff en el modelo lineal: $\hat{\beta}_1 = 0.0997$ (EE $= 0.0015$), $t = 67.07$, $p < 0.001$, IC95 % [0.0967, 0.1026].
  - ERA: $\hat{\beta}_1 = -14.44$ (EE $= 0.68$), $t = -21.22$, $p < 0.001$, IC95 % $[-15.78, -13.11]$.
  - HR: $\hat{\beta}_1 = 0.1253$ (EE $= 0.0120$), $t = 10.42$, $p < 0.001$, IC95 % [0.1017, 0.1489].
  - $\log(\text{HR} + 1)$: $\hat{\beta}_1 = 21.46$ (EE $= 2.07$), $t = 10.36$, $p < 0.001$, IC95 % [17.39, 25.53].
  - **Cuadrático** RunDiff^2: $\hat{\beta}_2 = -2.2 \times 10^{-5}$ (EE $= 1.1 \times 10^{-5}$), $t = -2.08$, $p = 0.038$, IC95 % $[-4.3 \times 10^{-5}, -1.0 \times 10^{-6}]$. Aunque el efecto marginal es muy pequeño, es estadísticamente distinto de cero al 5 %.

El mejor ajuste lo ofrece RunDiff (y su versión cuadrática con ligera mejora), seguido por ERA. Los modelos con HR o $\log(\text{HR} + 1)$ son significativos pero con R^2 notablemente menor.

Pronóstico

Generación del pronóstico

Para evaluar la capacidad predictiva, dividimos el panel equipo-año en un *split* temporal: *entrenamiento* 2000–2016 (510 obs.) y *prueba* 2017-2019 (90 obs.). Con base en la evidencia previa (mayor R^2, AIC/BIC y CV), el modelo usado para pronosticar fue el de mejor desempeño dentro de las formas funcionales consideradas:

$$\text{W} \;=\; \beta_0 \;+\; \beta_1\,\text{RunDiff} \;+\; \beta_2\,\text{RunDiff}^2,$$

ajustado sólo con los datos de entrenamiento. Sobre el conjunto de prueba, el modelo alcanza RMSE = 3.89 y MAE = 2.94 (frente a RMSE = 3.95 y MAE = 2.98 del modelo lineal en `RunDiff`). La Figure 15 muestra el *parity plot* (observado vs. pronosticado) en prueba: los

puntos se alinean alrededor de la diagonal, lo que refleja buen ajuste predictivo sin sesgos evidentes.

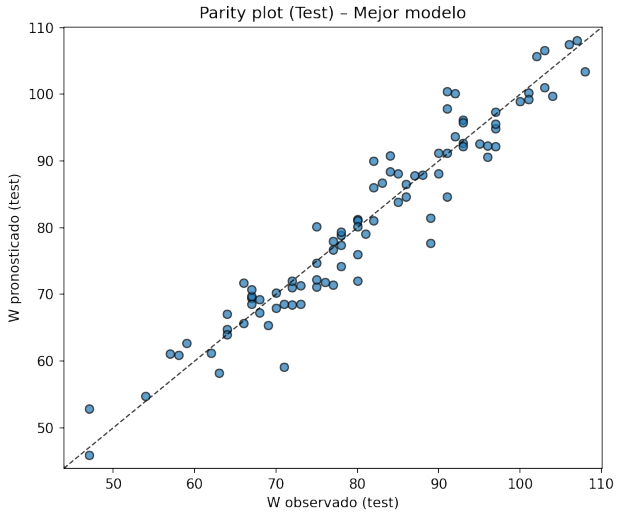

Figura 15

Parity plot (set de prueba) del mejor modelo: $W \sim RunDiff + RunDiff^2$.

Intervalos de predicción

Además del pronóstico puntual $\hat{W}$, se calcularon: (i) el intervalo de confianza al 95 % para la media condicional $E[W \mid \text{RunDiff}]$, y (ii) el intervalo de *predicción* al 95 % para observacionesau futuras W^*. La Figure 16 ilustra ambas bandas sobre el set de prueba: la banda verde (IC de la media) es más estrecha, mientras que la banda naranja (intervalo de predicción) refleja la variabilidad individual.

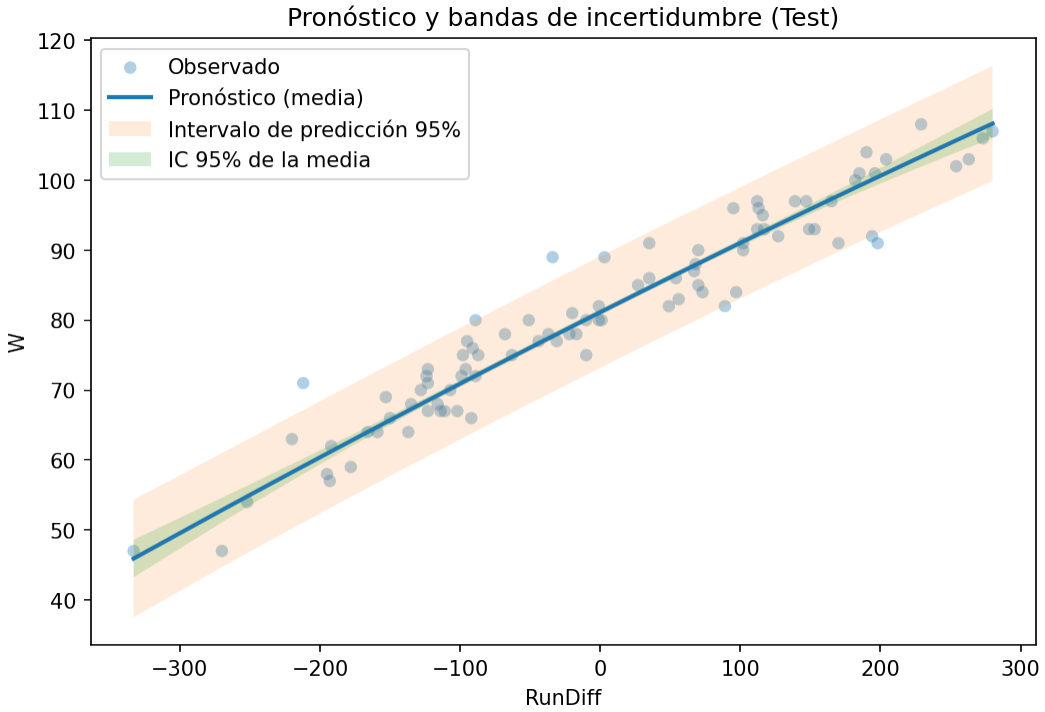

Figura 16

Pronóstico en prueba con bandas de incertidumbre (IC 95 % de la media y predicción 95 %).

Como referencia, para valores representativos de `RunDiff` se obtuvieron los siguientes intervalos (en el set de prueba):

Cuadro 7

Intervalos de predicción

RunDiff	$\hat{W}$	IC 95 % inf	IC 95 % sup	Pred. 95 % inf	Pred. 95 % sup
-150	65.69	65.04	66.35	57.69	73.69
-50	76.06	75.60	76.52	68.08	84.05
0	81.13	80.68	81.59	73.15	89.12
50	86.13	85.69	86.57	78.14	94.11
150	95.88	95.17	96.60	87.88	103.89
250	105.33	103.63	107.03	97.18	113.49

Obsérvese que la amplitud típica del intervalo de predicción ronda 15-18 victorias, mientras que el IC de la media es mucho más estrecho ($\approx 0.9 - 1.9$), como se espera teóricamente.

Evaluación del pronóstico

El desempeño predictivo se comparó entre modelos usando métricas estándar sobre *train* y *test*. En prueba (2017–2019), los resultados fueron:

- **Mejor modelo** $W \sim \text{RunDiff} + \text{RunDiff}^2$: MSE = 15.10, RMSE = 3.89, MAE = 2.94.
- **Lineal en RunDiff**: MSE = 15.57, RMSE = 3.95, MAE = 2.98.
- **Lineal en ERA**: RMSE = 9.14, MAE = 7.21.
- **Lineal en HR**: RMSE = 12.88, MAE = 10.45.

La mejora del término cuadrático frente al lineal en `RunDiff` es pequeña pero consistente y se acompaña de mejores AIC/BIC y validación cruzada. Además, `RunDiff` explica de forma sustantiva las victorias, y el modelo cuadrático aporta un refinamiento marginal útil para pronóstico operativo.

Conclusiones

Resumen de los hallazgos

El análisis confirma que la **diferencia de carreras** (`RunDiff`) es, con amplio margen, el mejor determinante de las victorias (`W`) a nivel equipo–año (2000–2019). En los modelos lineales simples, `RunDiff` explica cerca del 88 % de la variación de `W` ($R^2 \approx 0.883$), con errores típicos de pronóstico en el orden de 3–4 victorias por temporada (RMSE ≈ 4). La interpretación es directa: $\hat{\beta}_1 \approx 0.10$ implica que $\Delta 10$ carreras en el diferencial se traducen, en promedio, en ≈ 1 victoria adicional, y el intercepto cercano a 81 victorias es coherente con un equipo "neutral" (`RunDiff`=0).

La **efectividad del pitcheo** (ERA) exhibe una relación negativa y sustantiva con `W` ($r \approx -0.66$; $R^2 \approx 0.43$): reducir en una unidad la ERA se asocia con ≈ 14 victorias más.

Aunque su poder explicativo es menor que el de `RunDiff`, ofrece evidencia clara del rol del pitcheo en el desempeño global.

Los **jonrones** (HR) y su transformación $\log(\text{HR}+1)$ muestran asociaciones positivas pero *moderadas* ($R^2 \approx 0.15$). En el rango observado (90–300 HR) la versión logarítmica no mejora de manera apreciable al modelo lineal, lo que sugiere que el conteo de HR, por sí solo, captura sólo un fragmento del aporte ofensivo a las victorias.

En **formas funcionales**, un término cuadrático en `RunDiff` aporta una *mejora pequeña pero consistente*: R^2 y AIC/BIC mejoran levemente, el RMSE baja ($\approx 4.03 \to 4.02$) y el test RESET deja de ser significativo, lo que indica mejor especificación. En `ERA`, el término cuadrático no agrega valor y empeora marginalmente los criterios de información; para `HR`, la transformación logarítmica es prácticamente equivalente a la lineal.

Para **pronóstico** con *split* temporal (train 2000–2016, test 2017–2019), el modelo $W \sim \text{RunDiff} + \text{RunDiff}^2$ logra $\text{RMSE}_{\text{test}} \approx 3.89$ y $\text{MAE}_{\text{test}} \approx 2.94$, superando ligeramente al modelo lineal en `RunDiff`. Las bandas de incertidumbre muestran que, aunque el IC de la media es estrecho, los **intervalos de predicción** son naturalmente más amplios (del orden de 15–18 victorias), reflejando la variabilidad individual de equipos y temporadas.

Recomendaciones

1. **Uso operativo del modelo.** Para pronósticos rápidos y explicables, emplear $W \sim \text{RunDiff}$ como base; cuando se busque el mejor desempeño posible con mínima complejidad adicional, preferir $W \sim \text{RunDiff} + \text{RunDiff}^2$ (mejora marginal pero consistente y mejor especificación).

2. **Comunicación de incertidumbre.** Reportar siempre los *intervalos de predicción* junto con el pronóstico puntual ($\hat{W}$); el IC de la media no sustituye la amplitud de la incertidumbre a nivel de observación futura.

3. **Interpretación de ERA y HR.** Utilizar `ERA` como indicador complementario de diagnóstico (impacto del pitcheo) y evitar usar `HR` como único predictor de victorias; el

poder explicativo de HR aislado es limitado.

4. **Tratamiento de atípicos.** No eliminar outliers históricos por defecto (p. ej., 2019 en HR o 2003 DET en W) y, si se requiere robustez adicional, contrastar con estimadores/intervalos robustos o winsorización como análisis de sensibilidad.

5. **Extensiones futuras.**

   - Pasar a **regresión múltiple** (p. ej., `RunDiff` + métricas ofensivas y de pitcheo adicionales) para descomponer mejor contribuciones y reducir sesgo por omisión.
   - Explorar **modelos de panel** con efectos fijos por franquicia/año para capturar heterogeneidades persistentes (manager, estadio, presupuesto).
   - Evaluar especificaciones con **heterocedasticidad robusta** y **validación temporal** (rolling-origin) para escenarios de predicción real.

Bibliografía

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

Anexo

Link al repositorio con código fuente y salidas correspondientes

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