

**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**

<b>Introducción</b>	<b>4</b>
Objetivo del trabajo: . . . . .	4
Justificación: . . . . .	4
Descripción de los datos: . . . . .	5
<b>Selección de Variables</b>	<b>6</b>
Variable dependiente: . . . . .	6
Variables independientes: . . . . .	6
<b>Análisis de Estadísticas Descriptivas</b>	<b>7</b>
Medidas de tendencia central y dispersión: . . . . .	7
Visualización de los datos: . . . . .	8
Identificación de valores atípicos: . . . . .	13
<b>Análisis de Correlación</b>	<b>14</b>
Correlación entre las variables: . . . . .	14
Interpretación de los resultados . . . . .	15
<b>Modelo de Regresión Múltiple</b>	<b>17</b>
Especificación y ajuste . . . . .	17
Resultados y significancia de coeficientes . . . . .	18
Bondad de ajuste y pruebas globales . . . . .	19
Comentarios sobre los resultados . . . . .	19

<b>Validación de los Supuestos de la Regresión Múltiple</b>	<b>20</b>
Linealidad . . . . .	20
Independencia de los errores . . . . .	20
Homocedasticidad . . . . .	20
Normalidad de los residuos . . . . .	20
Ausencia de multicolinealidad . . . . .	20
<b>Pronóstico</b>	<b>23</b>
Generación del pronóstico . . . . .	23
Intervalos de predicción . . . . .	23
Evaluación del pronóstico . . . . .	24
<b>Conclusiones</b>	<b>25</b>
Resumen de resultados . . . . .	25
Recomendaciones . . . . .	26
<b>Bibliografía</b>	<b>27</b>
<b>Anexo</b>	<b>27</b>
Link al repositorio con código fuente y salidas correspondientes . . . . .	27

## 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,

homocedasticidad, 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:  $\text{RunDiff} = R - RA$ ,  $\text{logHR1} = \ln(\text{HR}+1)$ .

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 dataset maestro con las variables 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
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:

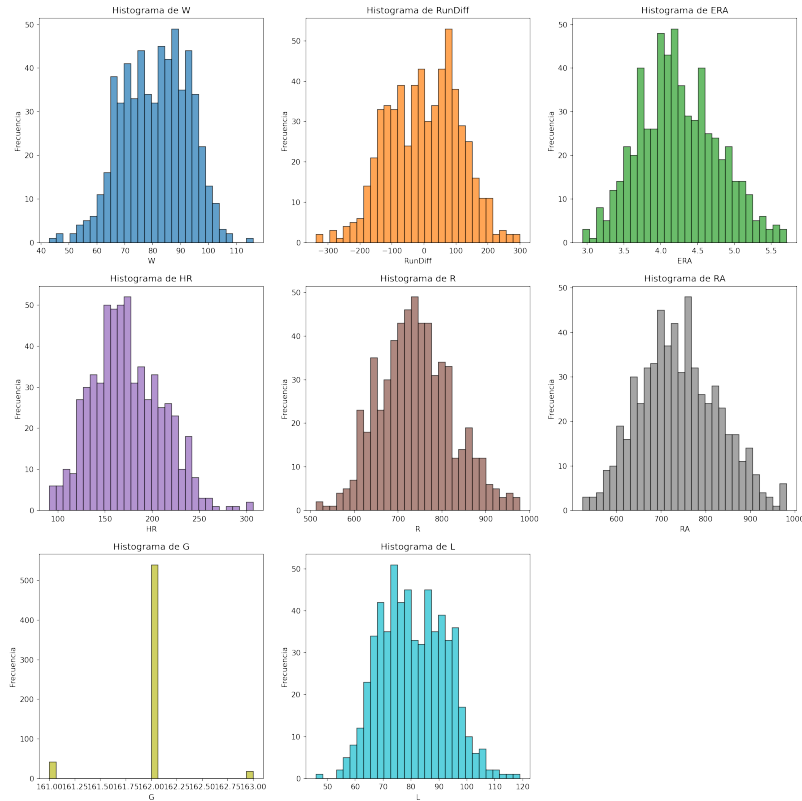
- **Victorias (W):** En promedio los equipos ganan 81 juegos por temporada , con una desviación estándar de 11.8. El rango va de 43 a 116 victorias, lo que refleja tanto equipos altamente competitivos como equipos en el extremo opuesto.
- **Diferencia de carreras (RunDiff):** Tiene media cercana a cero, 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.
  
- **Carreras anotadas (R) y recibidas (RA):** Ambas variables tienen media  $\approx 741$ , 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.

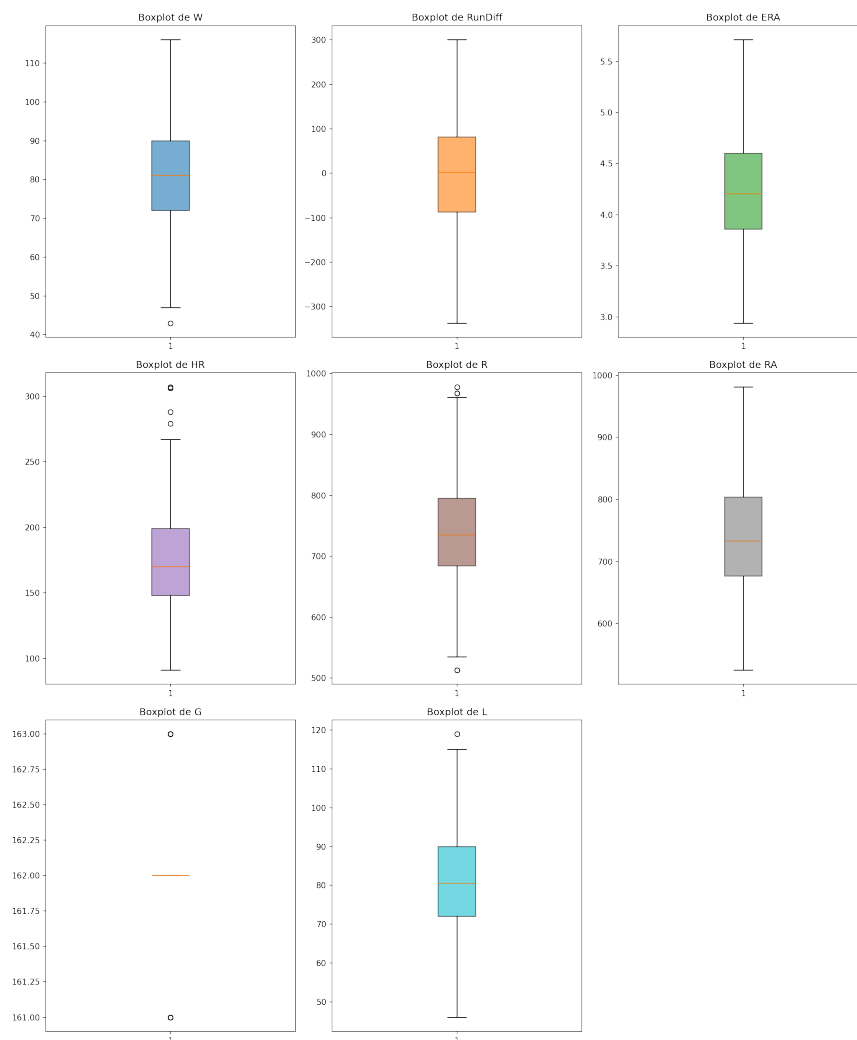




**Figura 1**

*Histogramas de  $W$ ,  $RunDiff$ ,  $ERA$ ,  $HR$ ,  $R$ ,  $RA$ ,  $G$  y  $L$  (2000–2019).*

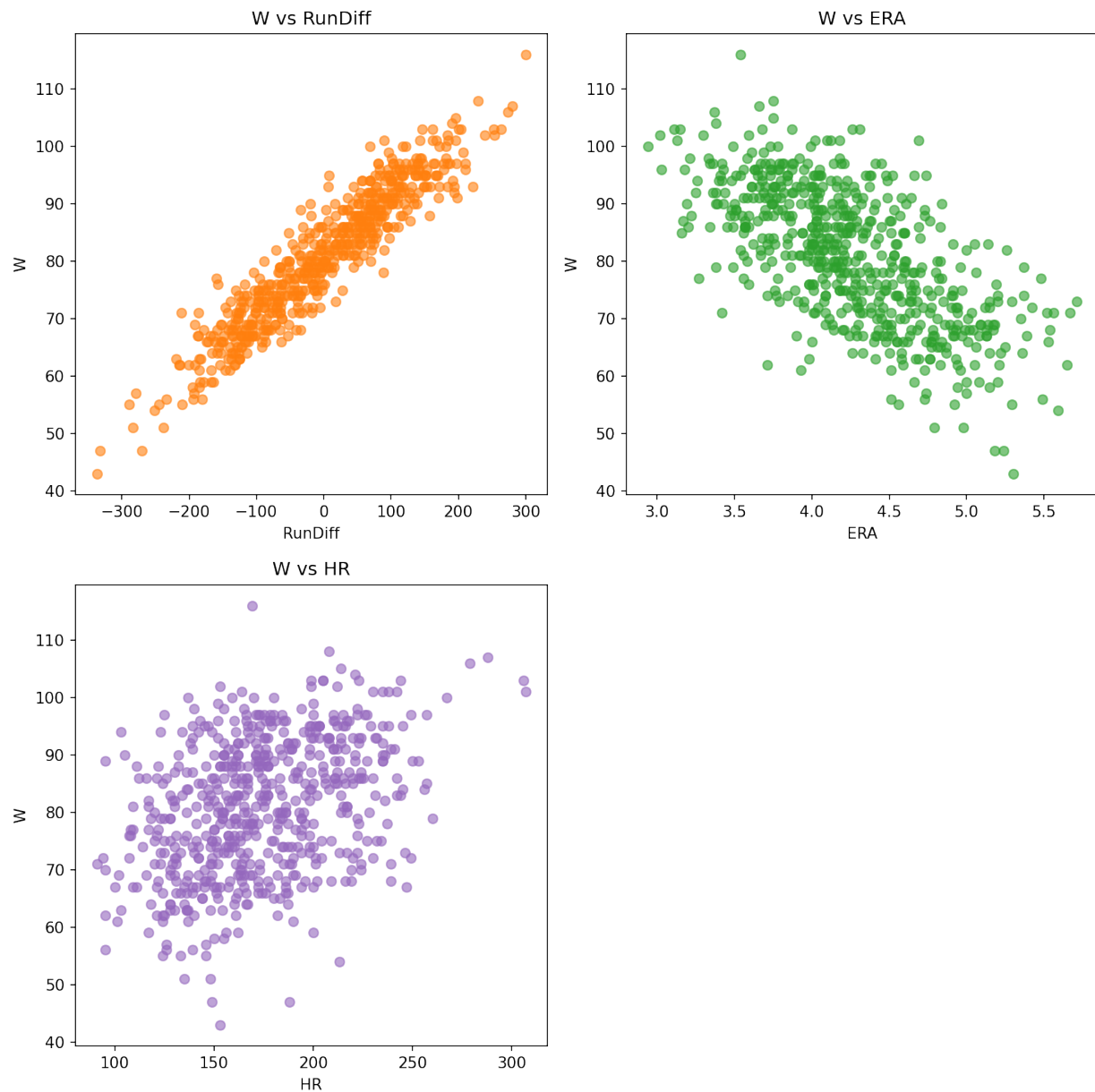
Los histogramas confirman distribuciones aproximadamente simétricas en **W** y **L**, con centro en 81 victorias/derrotas. **RunDiff** muestra gran dispersión, confirmando que algunos equipos superan a sus rivales por cientos de carreras, mientras otros son ampliamente superados. **ERA** se concentra en torno a 4, reflejando diferencias moderadas en pitcheo. **HR** se distribuye entre 100-300. **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.*

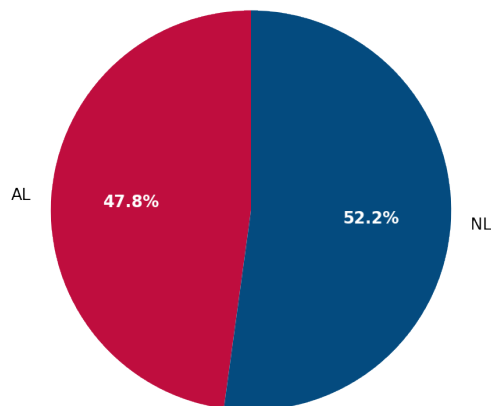
Los boxplots identifican *outliers* relevantes: (i) en **HR**, equipos con poder ofensivo distintivo (300+ HR), (ii) en **R** se ven reflejados esos mismos casos extremos de producción ofensiva. (iii) en **G**, ligeras desviaciones (161 o 163 partidos), explicadas por suspensiones o dobles juegos. En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

**Figura 3**

*Diagramas de dispersión:  $W$  vs  $RunDiff$ ,  $ERA$  y  $HR$ .*

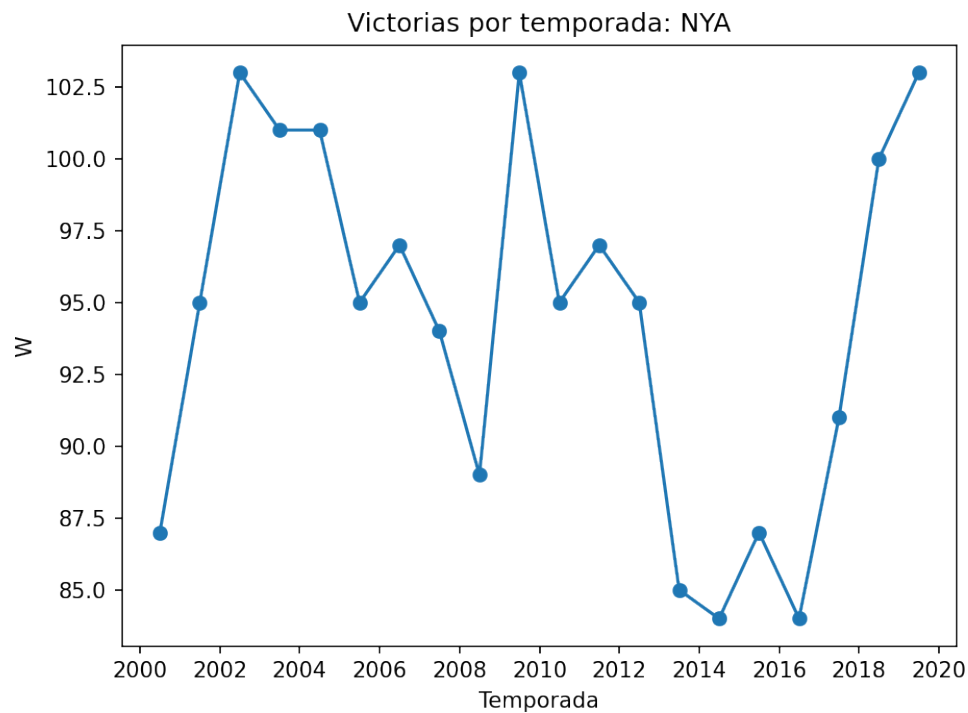
**W vs RunDiff** presenta la relación más fuerte y lineal: a mayor diferencial de carreras, más victorias, confirmando su validez como predictor central. **W vs ERA** muestra una relación negativa clara: equipos con menor efectividad del pitcheo (ERA baja) ganan más. **W vs HR** tiene asociación positiva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

Distribución de observaciones por liga (AL vs NL)

**Figura 4**

*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 5**

*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.
- **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.

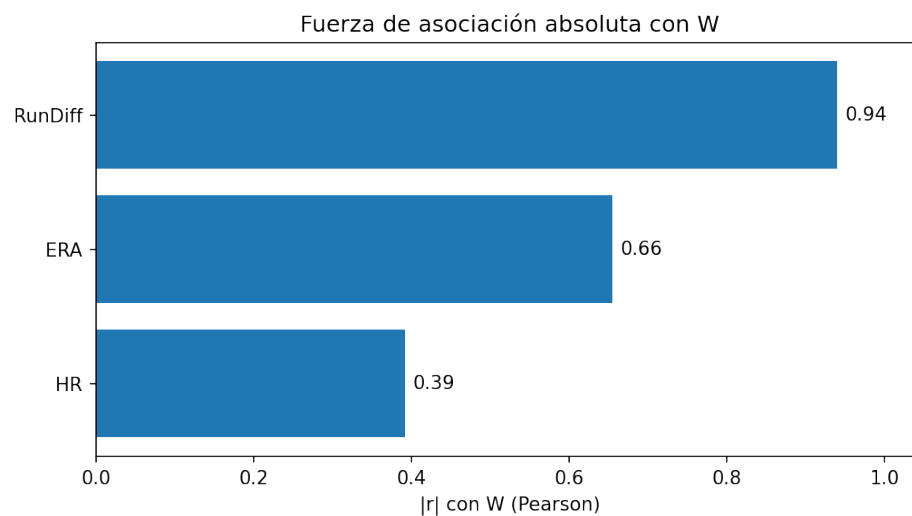
### Análisis de Correlación

#### 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 y HR). Los resultados se resumen en la Tabla 2 y en las Figuras 7–8.

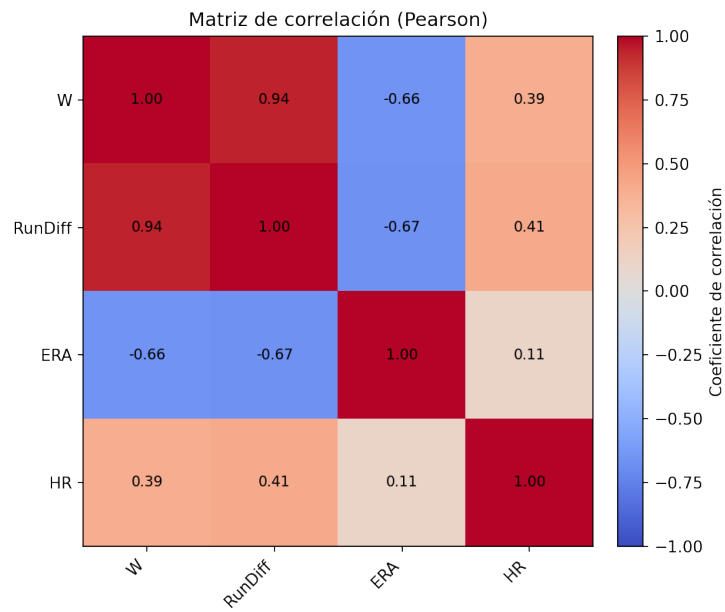
**Cuadro 2***Correlación de W con variables explicativas*

Variable	Pearson r	p (Pearson)	Spearman $\rho$	p (Spearman)	N
RunDiff	0.9395	0.0000	0.9398	0.0000	600
HR	0.3920	0.0000	0.3853	0.0000	600
ERA	-0.6554	0.0000	-0.6575	0.0000	600

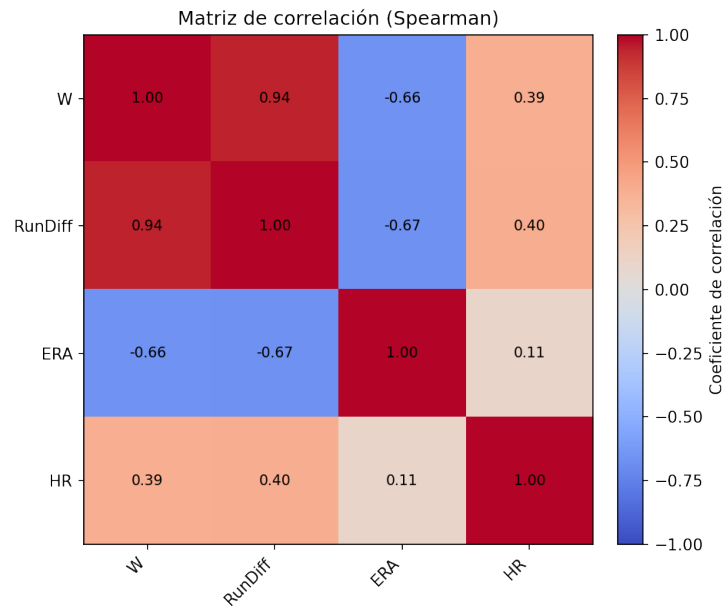
**Figura 6***Fuerza de asociación absoluta ( $|r|$ ) de cada predictor con W (Pearson).***Interpretación de los resultados**

- **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** exhibe una correlación positiva moderada ( $r \approx 0.39, p < 0.001$ ). Los jonrones

ayudan a ganar, aunque no explican tanto como RunDiff o ERA.



**Figura 7**  
*Matriz de correlación (Pearson) entre W y las variables explicativas.*



**Figura 8**  
*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.

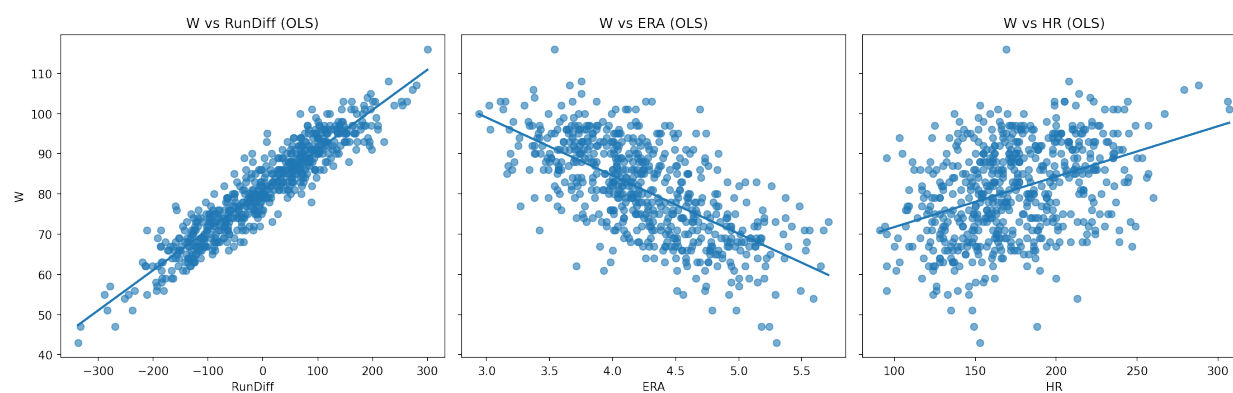
## Modelo de Regresión Múltiple

### Especificación y ajuste

El modelo estimado es:

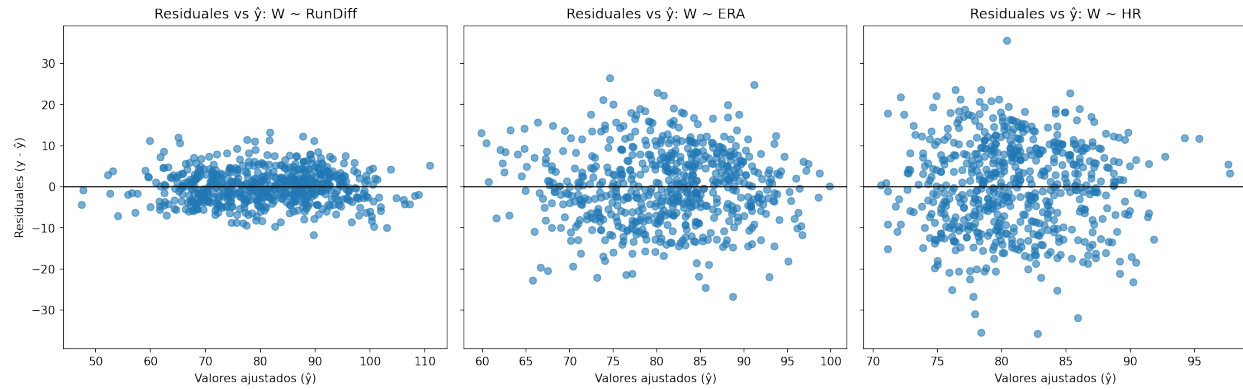
$$W_i = \beta_0 + \beta_1 \text{RunDiff}_i + \beta_2 \text{ERA}_i + \beta_3 \text{HR}_i + \varepsilon_i,$$

que se ajustó por Mínimos Cuadrados Ordinarios . Las siguientes figuras muestran, a modo exploratorio, las relaciones con  $W$  y los residuales del ajuste (para evaluar linealidad y patrones):



**Figura 9**

*Diagramas de dispersión: W vs RunDiff, ERA y HR .*

**Figura 10**

*Residuales del modelo múltiple vs valores ajustados por predictor.*

## Resultados y significancia de coeficientes

**Cuadro 3**

*Modelo múltiple:  $W \sim RunDiff + ERA + HR$  (coeficientes,  $p$  e IC95 %)*

Variable	$\hat{\beta}$	$p$ -valor	IC95 % inf	IC95 % sup
Constante	86.015950	0.000000	82.550769	89.481130
RunDiff	0.091675	0.000000	0.086527	0.096823
ERA	-1.824850	0.000288	-2.807436	-0.842265
HR	0.015600	0.008514	0.003994	0.027206

**Interpretación de signos y magnitudes.** Manteniendo constantes las demás variables: (i) un incremento de 10 unidades en RunDiff se asocia con  $\approx 0.92$  victorias adicionales; (ii) reducir la ERA en 1.00 se asocia con  $\approx 1.82$  victorias más; (iii) sumar 10 jonrones (HR) se asocia con  $\approx 0.156$  victorias adicionales. Los tres coeficientes son estadísticamente significativos.

## Bondad de ajuste y pruebas globales

### Cuadro 4

*Estadísticos de ajuste y desempeño del modelo múltiple*

Medida	Valor	$p$ -valor
$R^2$	0.885314	
$R^2$ ajustado	0.884737	
Estadístico $F$	1533.603	0.000000
AIC	3370.731	
BIC	3388.319	
Observaciones ( $N$ )	600	
RMSE	3.988	
MAE	3.177	
Durbin–Watson	1.789	
Omnibus (normalidad)	4.013	0.134
Jarque–Bera	3.840	0.147

El modelo explica **88.5 %** de la variación en  $W$  ( $R^2 = 0.885$ ) con  $R^2_{adj} = 0.885$ . La prueba  $F(3, 596) = 1533.60$  es altamente significativa ( $p < 10^{-270}$ ), confirmando relevancia conjunta. Los errores medios de pronóstico in-sample son bajos ( $RMSE \approx 3.99$ ,  $MAE \approx 3.18$ ). Los diagnósticos básicos no rechazan normalidad de residuos (Omnibus/JB,  $p > 0.10$ ) y el estadístico Durbin–Watson  $\approx 1.79$  no sugiere autocorrelación preocupante en este contexto. El número de condición reportado ( $\approx 1.98 \times 10^3$ ) advierte posible *multicolinealidad moderada*; por lo que se complementará con VIF en la sección de supuestos.

### Comentarios sobre los resultados

- **RunDiff** es el predictor dominante en magnitud y precisión, consistente con la teoría sabermétrica: resume balance ofensivo–defensivo.

- **ERA** conserva un efecto negativo y significativo una vez controlado RunDiff y HR: mejor pitcheo, más victorias.
- **HR** aporta señal positiva pero de menor tamaño relativo; su significancia indica que, manteniendo constante RunDiff y ERA, el poder de bateo agrega información marginal.

## Validación de los Supuestos de la Regresión Múltiple

### Linealidad

La linealidad del modelo se evaluó con gráficos exploratorios y de regresión parcial (Fig. 11), así como con pruebas formales. El test **Rainbow** no rechaza la especificación lineal ( $p=0.334$ ). El **RESET de Ramsey** arroja  $p=0.053$ , valor cercano al umbral del 5 %; lo consideramos una señal de cautela, pero no evidencia concluyente para modificar la forma funcional. En conjunto, la especificación lineal resulta adecuada para los fines del estudio.

### Independencia de los errores

La independencia se contrastó con el estadístico **Durbin–Watson**, que resultó en  $DW \approx 1.79$ . Este valor es cercano a 2 y no sugiere autocorrelación de primer orden preocupante en los residuos.

### Homocedasticidad

Se revisaron los *residuales vs. valores ajustados* (Fig. 12) sin patrones de abanico. Las pruebas **Breusch–Pagan** ( $p=0.26$ ) y **White** ( $p=0.38$ ) no rechazan la hipótesis nula de varianza constante. Concluimos que no hay evidencia estadística de heterocedasticidad.

### Normalidad de los residuos

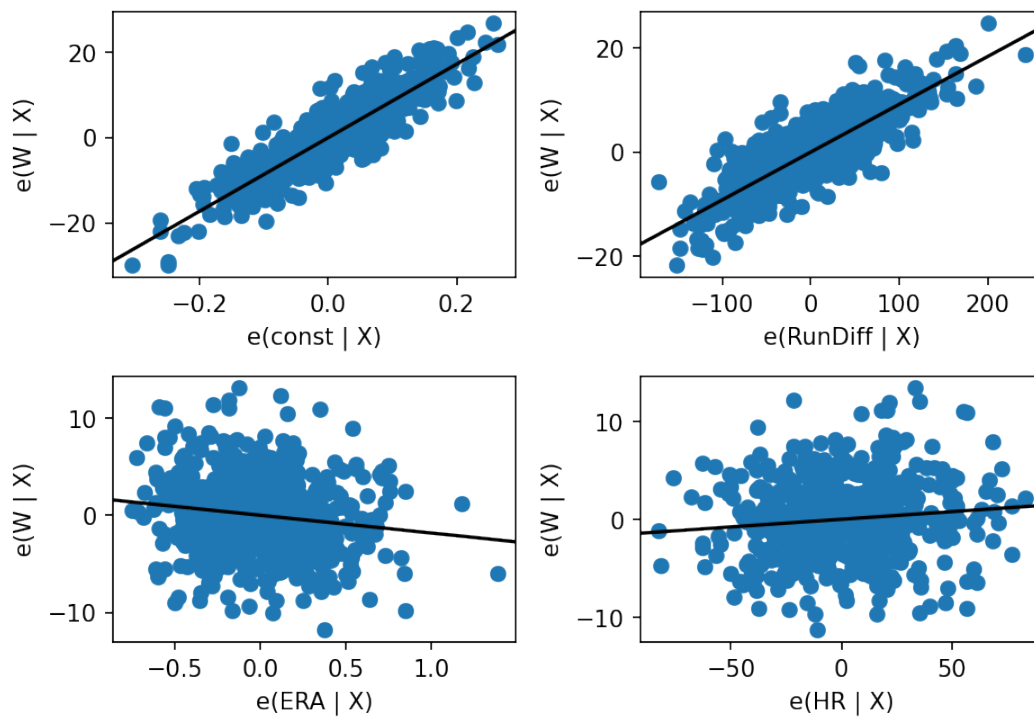
La normalidad se verificó mediante **Shapiro–Wilk** ( $p=0.20$ ) y **Jarque–Bera** ( $p=0.15$ ), que no rechazan normalidad al 5 %. El *QQ-plot* (Fig. 13) muestra alineación razonable con la diagonal, con desvíos menores en colas.

### Ausencia de multicolinealidad

Se calcularon los **VIF**:  $\text{RunDiff} \approx 3.17$ ,  $\text{ERA} \approx 2.68$  y  $\text{HR} \approx 1.78$ , todos por debajo de 5, por lo que no hay multicolinealidad grave. El **número de condición** del diseño es

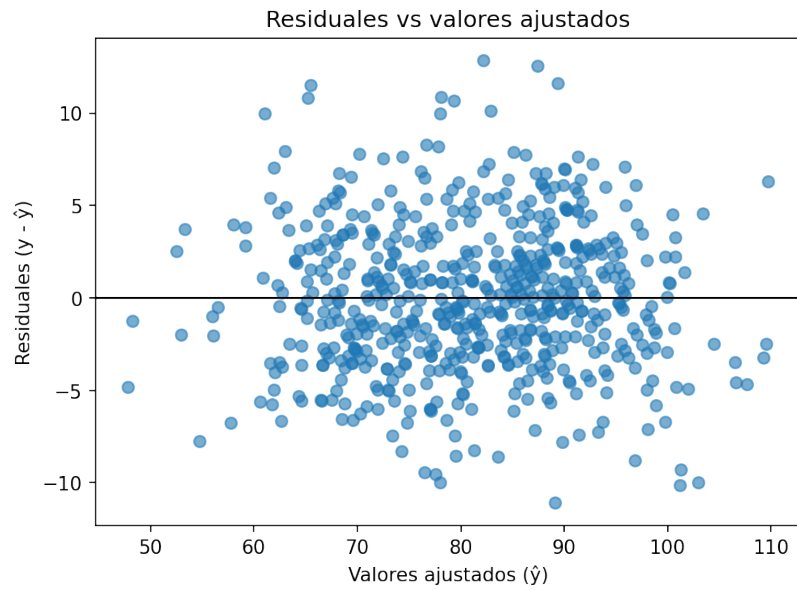
$\kappa \approx 1,980$ , lo que sugiere colinealidad *moderada*. Esto es esperable porque las métricas de rendimiento de un equipo están naturalmente relacionadas entre sí.

Partial regression (added-variable) –  $W \sim \text{RunDiff} + \text{ERA} + \text{HR}$

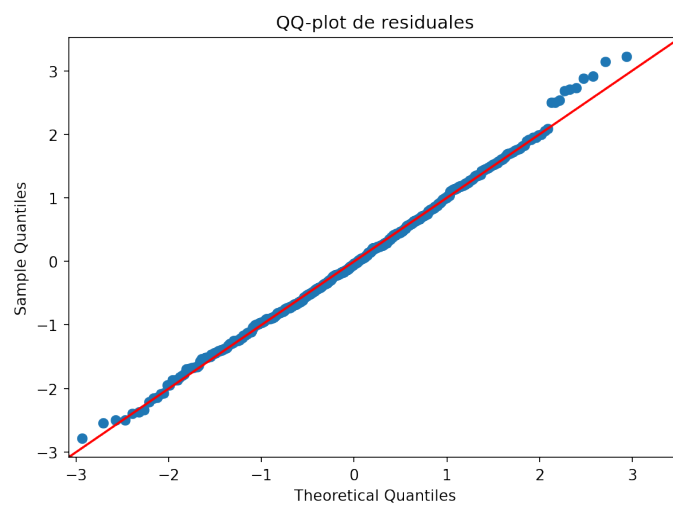


**Figura 11**

*Regresión parcial (added-variable): grilla de gráficos.*

**Figura 12**

*Residuos vs. valores ajustados del modelo múltiple.*

**Figura 13**

*QQ-plot de los residuos del modelo.*

## Pronóstico

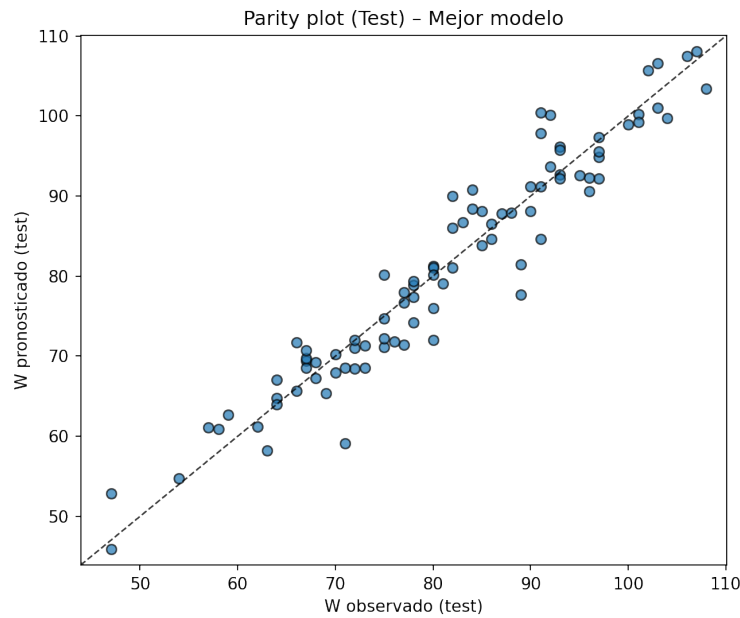
### Generación del pronóstico

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 de ajuste y validación, el modelo usado para pronosticar fue el de mejor desempeño:

$$W = \beta_0 + \beta_1 \text{RunDiff} + \beta_2 \text{ERA} + \beta_3 \text{HR},$$

ajustado únicamente con el conjunto de entrenamiento y evaluado luego en prueba.

La Figura 14 muestra el *parity plot* (observado vs. pronosticado) en prueba: los puntos se alinean alrededor de la diagonal, lo que refleja buen desempeño predictivo sin sesgos sistemáticos.



**Figura 14**

*Parity plot (set de prueba) del mejor modelo:  $W \sim \text{RunDiff} + \text{ERA} + \text{HR}$ .*

### Intervalos de predicción

Además del pronóstico puntual  $\hat{W}$ , reportamos (i) el intervalo de confianza al 95 % para la media condicional y (ii) el intervalo de *predicción* al 95 % para observaciones futuras.

**Cuadro 5**

*Pronósticos puntuales e intervalos (IC 95 % de la media; predicción 95 %).*

RunDiff	ERA	HR	$\hat{W}$	IC 95 % inf	IC 95 % sup	Pred. 95 % inf	Pred. 95 % sup
-150	5.0	140	65.11	64.39	65.83	57.19	73.03
-50	4.3	170	76.26	75.83	76.70	68.37	84.16
0	4.0	200	82.02	81.31	82.74	74.11	89.94
50	3.8	210	87.21	86.41	88.02	79.29	95.14
150	3.5	230	97.40	96.41	98.39	89.45	105.34
250	3.2	250	107.58	106.33	108.82	99.60	115.56

**Evaluación del pronóstico**

La Tabla 6 compara el desempeño en entrenamiento y prueba entre el mejor modelo múltiple y tres modelos lineales simples.

**Cuadro 6**

*Desempeño predictivo (MSE, RMSE, MAE) por modelo y partición.*

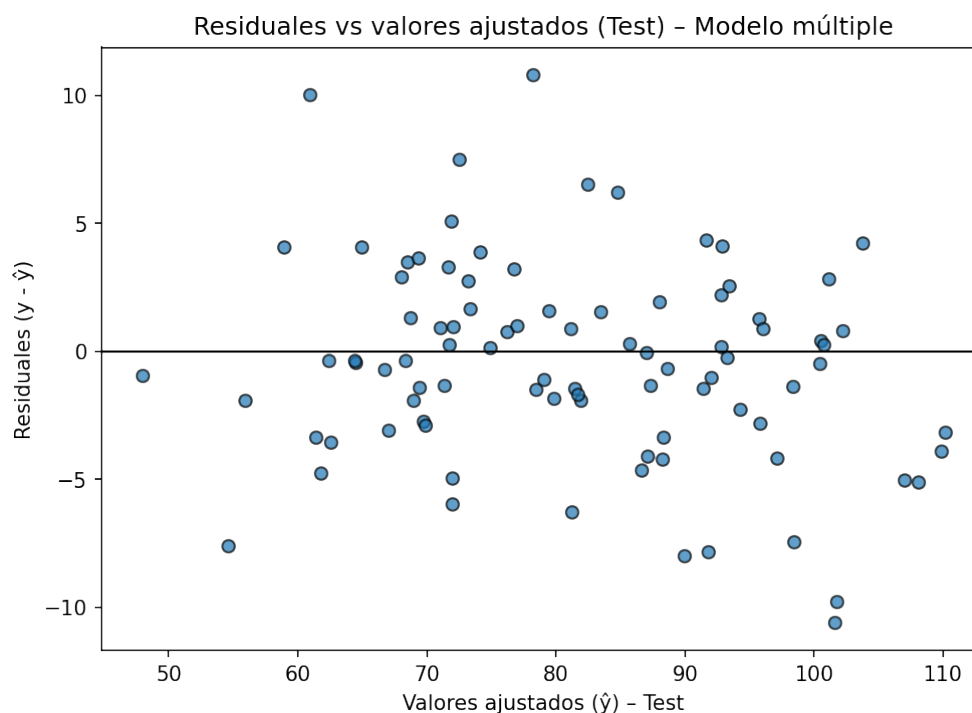
Modelo	Split	MSE	RMSE	MAE	N
Best: $W \sim \text{RunDiff} + \text{ERA} + \text{HR}$	Train (2000–2016)	15.98	3.998	3.205	510
Best: $W \sim \text{RunDiff} + \text{ERA} + \text{HR}$	Test (2017–2019)	15.72	3.965	3.027	90
RunDiff (lineal)	Train (2000–2016)	16.42	4.052	3.252	510
RunDiff (lineal)	Test (2017–2019)	15.57	3.946	2.978	90
ERA (lineal)	Train (2000–2016)	78.79	8.877	7.177	510
ERA (lineal)	Test (2017–2019)	83.49	9.137	7.209	90
HR (lineal)	Train (2000–2016)	109.49	10.464	8.630	510
HR (lineal)	Test (2017–2019)	165.82	12.877	10.450	90

En prueba, el modelo múltiple alcanza  $\text{RMSE} \approx 3.97$  y  $\text{MAE} \approx 3.03$ , superando



claramente a los modelos que usan *una sola* variable y quedando a la par del lineal en RunDiff. Dado que el múltiple integra simultáneamente información de bateo (HR), diferencial agregado (RunDiff) y pitcheo (ERA), ofrece una *mejor base explicativa* con errores similares o menores.

Finalmente, verificamos los residuales del modelo en el set de prueba (Fig. 15): la nube se distribuye alrededor de cero sin patrón de abanico, lo cual es consistente con la homocedasticidad indicada por las pruebas.



**Figura 15**

*Residuos vs. valores ajustados (Test) – Modelo múltiple.*

## Conclusiones

### Resumen de resultados

El modelo de **regresión lineal múltiple**  $W \sim \text{RunDiff} + \text{ERA} + \text{HR}$  explica el **88.5 %** de la variación en las victorias por temporada ( $R^2 = 0.885$ ;  $R_{aj}^2 = 0.885$ ). La prueba global  $F(3, 596) = 1533.60$  ( $p < 10^{-270}$ ) confirma la *relevancia conjunta* de los predictores.

En términos de *efectos parciales* (manteniendo constantes las demás variables):

- **RunDiff**: coeficiente  $\hat{\beta} \approx 0.0917$ . Un aumento de 10 carreras en el diferencial se asocia con  $\approx 0.92$  victorias adicionales. Es el predictor con mayor precisión y magnitud relativa.
- **ERA**: coeficiente  $\hat{\beta} \approx -1.825$ . Reducir la efectividad en 1.00 se asocia con  $\approx 1.82$  victorias más, consistente con el papel del pitcheo en el desempeño global.
- **HR**: coeficiente  $\hat{\beta} \approx 0.0156$ . Sumar 10 jonrones se asocia con  $\approx 0.156$  victorias adicionales; el efecto es positivo pero de menor tamaño que RunDiff y ERA.

Los diagnósticos apoyan la validez del modelo para los fines del estudio: **DW**  $\approx 1.79$  no sugiere autocorrelación de primer orden; **Omnibus/JB** no rechazan normalidad ( $p > 0.10$ ); **BP** y **White** no rechazan homocedasticidad ( $p = 0.26$  y  $p = 0.38$ ). Los **VIF** ( $\approx 3.17, 2.68, 1.78$ ) y el número de condición ( $\kappa \approx 1,980$ ) indican colinealidad *moderada*, esperable en métricas de rendimiento de un mismo equipo.

En evaluación temporal, el modelo múltiple alcanzó en prueba (2017–2019)  $\text{RMSE} \approx 3.97$  y  $\text{MAE} \approx 3.03$ , con buen alineamiento observado–pronosticado y residuales sin patrones sistémicos, lo que respalda su uso para pronóstico de corto plazo.

### Recomendaciones

1. **Uso práctico del modelo.** Emplear  $W \sim \text{RunDiff} + \text{ERA} + \text{HR}$  para pronósticos explicables y consistentes. Reportar siempre el error típico (RMSE/MAE) como referencia operativa.
2. **Comunicación de incertidumbre.** Acompañar cada pronóstico puntual  $\hat{W}$  con su **intervalo de predicción al 95 %**; evitar interpretar el IC de la media como incertidumbre individual.
3. **Monitoreo de supuestos.** Mantener verificación periódica de linealidad, homocedasticidad y normalidad; ante señales de heteroscedasticidad, complementar

con errores estándar robustos (HC3) como sensibilidad.

4. **Colinealidad.** Vigilar VIF y el número de condición al incorporar nuevas variables; priorizar parsimonia para preservar estabilidad de coeficientes e interpretabilidad.
5. **Mantenimiento del modelo.** Reentrenar con datos recientes (p. ej., anual) y validar con *split* temporal; considerar *rolling-origin* cuando el objetivo principal sea pronóstico.
6. **Extensiones posibles.** Explorar especificaciones de panel con efectos fijos por franquicia/año y la inclusión de métricas complementarias (BB, SO, OBP, SLG) para descomponer aportes ofensivos y de pitcheo, evaluando siempre el impacto en VIF y desempeño predictivo.

### 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/enriuegomeztagle/  
MCD-ProyectoFinalEconometria-DeterminantesVictoriasMLB](https://github.com/enriuegomeztagle/MCD-ProyectoFinalEconometria-DeterminantesVictoriasMLB)