

The background of the slide is a green field with a white grid pattern, representing a baseball field. Various baseball-related items are scattered around the central text box: a pinstriped jersey and pants on the left, a baseball at the top center, a bat at the top right, a catcher's mask at the top right, a glove on the right, a catcher's chest protector at the bottom right, a baseball at the bottom center, and a cap at the bottom left.

Econometria: Determinantes de Victorias en MLB

Enrique Ulises Báez Gómez Tagle y Luis Alejandro Guillén Alvarez

Objetivo del trabajo

Explicar y pronosticar el número de victorias de un equipo de Grandes Ligas en una temporada a partir de múltiples variables 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.



***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





Datos (fuente y recorte)



Fuente

Lahman 1871-2024, Society for American Baseball Research (SABR)



Filtrar

Ligas Americana (AL) y Nacional (NL) periodo 2000-2019



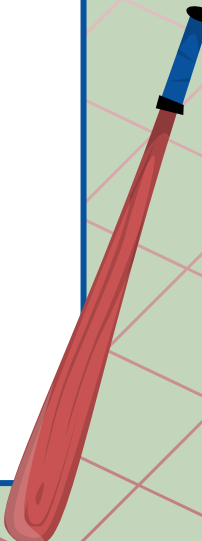
Variables principales

W, L, G, R, RA, ERA, HR y derivada (**RunDiff**).



Dataset Final

600 observaciones (30 equipos por temporada durante 20 años),



Selección de variables



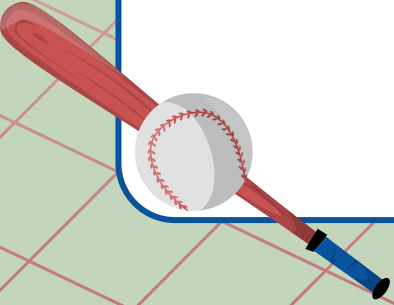
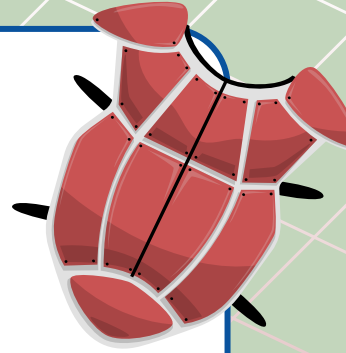
Dependientes

W (victorias por temporada).



Independientes

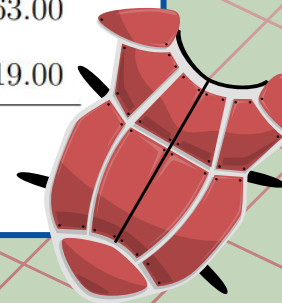
RunDiff, ERA, HR



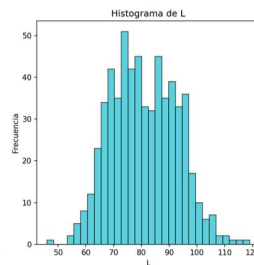
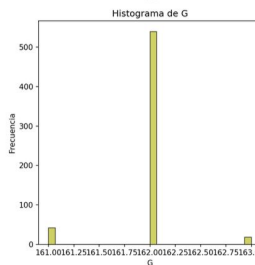
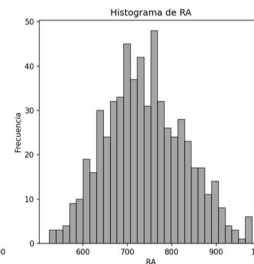
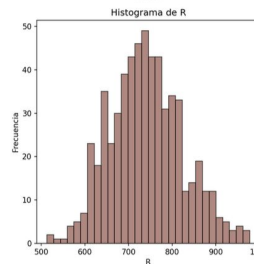
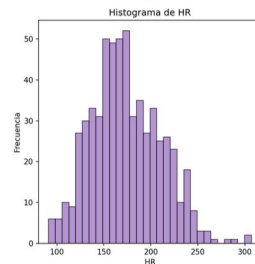
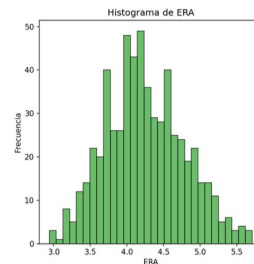
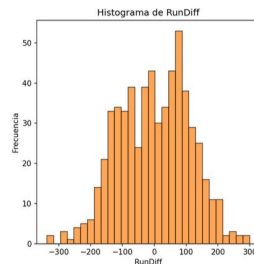
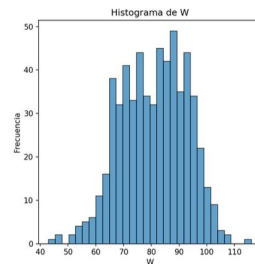
Estadísticas descriptivas



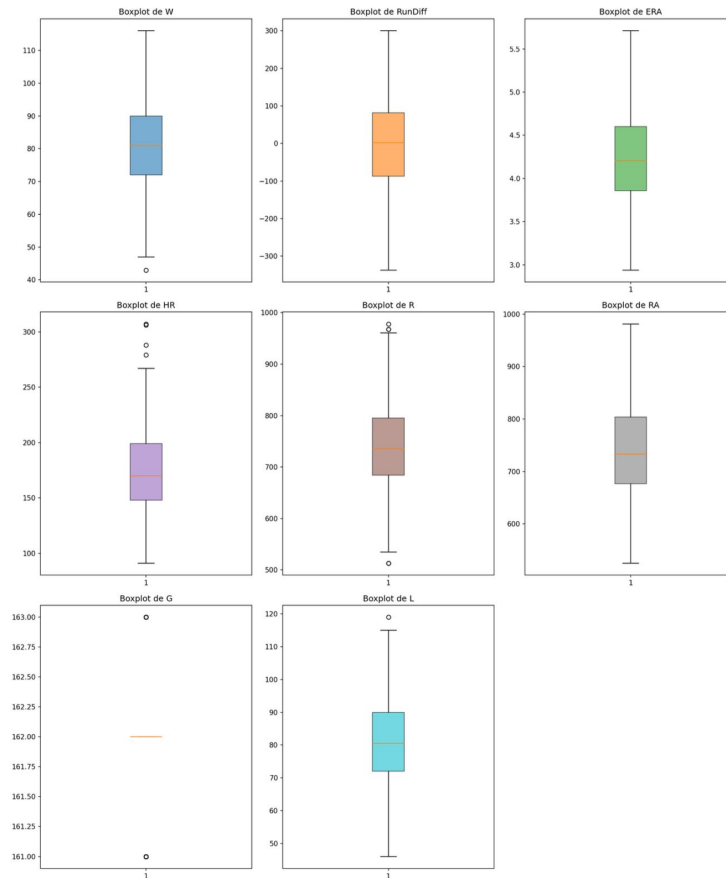
Variable	Count	Mean	Median	Mode	Std	Var	Min	Q1	Q3	IQR	Max
W	600	80.97	81.00	86.00	11.79	138.92	43.00	72.00	90.00	18.00	116.00
RunDiff	600	0.00	2.00	54.00	111.11	12344.79	-337.00	-87.00	81.25	168.25	300.00
ERA	600	4.25	4.21	4.01	0.53	0.29	2.94	3.86	4.60	0.74	5.71
HR	600	173.47	170.00	161.00	36.87	1359.12	91.00	148.00	199.00	51.00	307.00
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



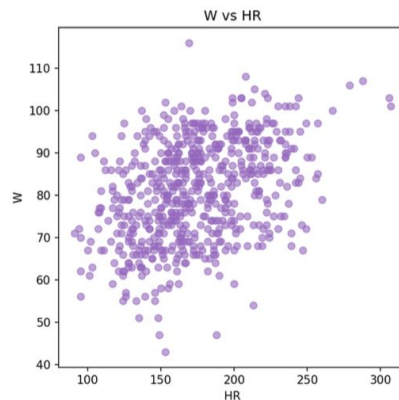
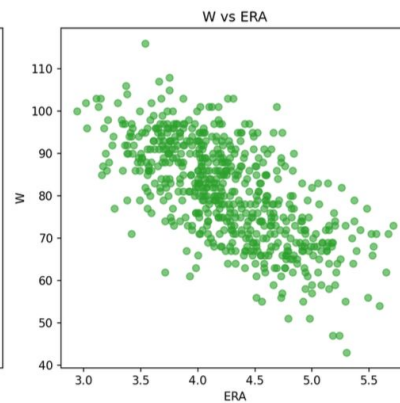
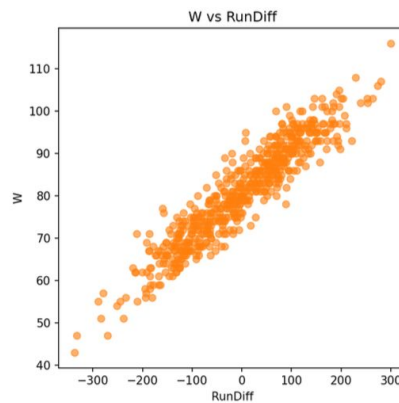
Histogramas



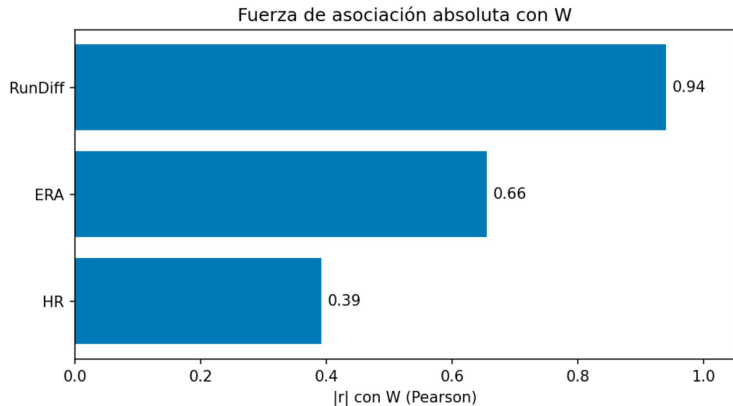
Boxplots & Outliers



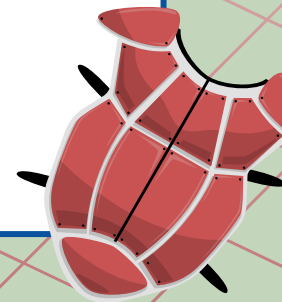
Dispersión: W vs. predictores



Correlación de W con variables explicativas



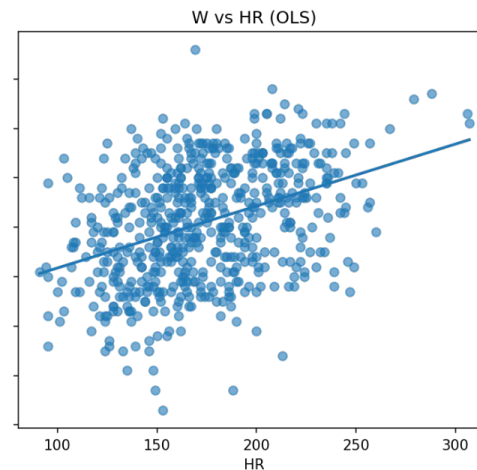
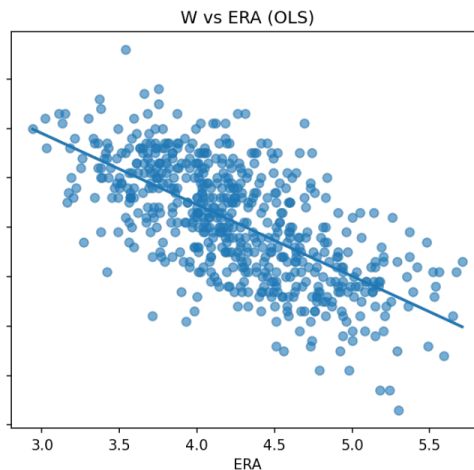
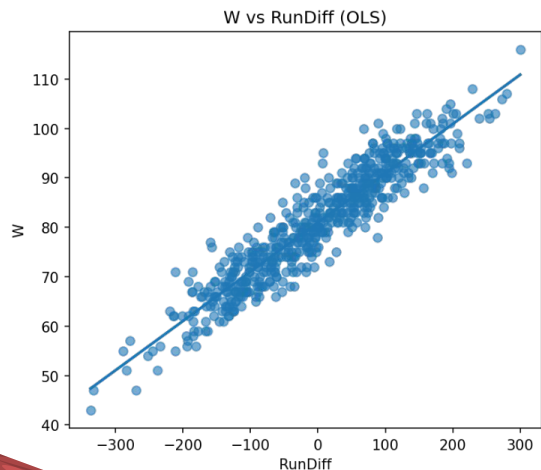
Variable	Pearson r	p (Pearson)	Spearman ρ	p (Spearman)	N
RunDiff	0.9395	0.0000	0.9398	0.0000	600
HR	0.3920	0.0000	0.3853	0.0000	600
ERA	-0.6554	0.0000	-0.6575	0.0000	600





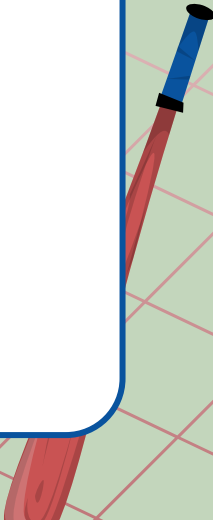
Modelo de regresión múltiple

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





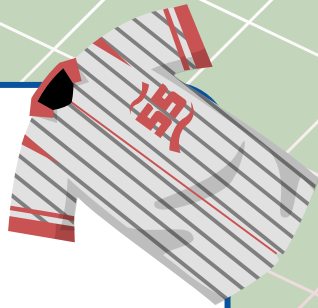
Resultados



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



Bondad de ajuste y desempeño



Medida	Valor	<i>p</i> -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



Supuestos del modelo



Linealidad

Rainbow $p=0.334$; RESET $p=0.053$



Homoscedasticidad

Breusch–Pagan $p=0.26$, White $p=0.38$.



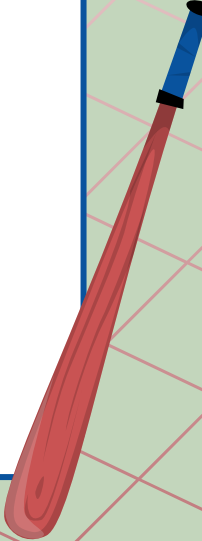
Multicolinealidad

VIF RunDiff ≈ 3.17 , ERA ≈ 2.68 ,
HR ≈ 1.78



Normalidad residuos

Shapiro–Wilk $p=0.20$, Jarque–Bera $p=0.15$.





Pronóstico: split, desempeño y paridad


Train **2000–2016 (510)**;

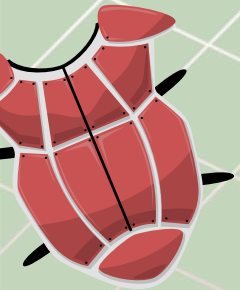
Test **2017–2019 (90)**

Mejor modelo: el múltiple

(RunDiff+ERA+HR).

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





Conclusiones ***&*** ***Recomendaciones***





Gracias
Por su
Atención

Referencias



Lahman Baseball
Database -Society for
American Baseball
Research. (2025)

Github Repository:
Código fuente y salidas
correspondientes

