Bases llenas con dos outs, Pitágoras al bate

Alejandro Crema

Noviembre de 2018
Happy Hour ASOVAC
Aula 35
Caracas, Venezuela

Con la asesoría de Fernando Crema

Contenido

Porqué hay tanta estadística en el beisbol?

• Secuencia de estados y eventos que producen la transición de un estado a otro. Se acumulan carreras anotadas.

Ejemplo:

$$s_1 = ((-,-,-),0),\ e_1 = \text{Out43},\ s_2 = ((-,-,-),1)$$

$$s_2 = ((-,-,-),1),\ e_2 = 2B,\ s_3 = ((-,*,-),1)$$

$$s_3 = ((-,*,-),1),\ e_3 = 1B + CA,\ s_4 = ((*,-,-),1)$$

$$s_4 = ((*,-,-),1),\ e_4 = \text{out64 y out43},\ ((X,X,X),3)$$

• Se anota TODO lo que ocurre en relación a estados y eventos (sin comentarios cualitativos).

• Es muy sencillo contar ocurrencias de eventos y estados e inventar métricas (variables) para jugadores y equipos.

Ejemplos de variables tradicionales

- Total de eventos por jugador, se suman, se calculan relaciones, promedios, promedios ponderados, ...

$$BB, BBI, GP, TS, SF, 1B, 2B, 3B, HR, BR, OR, CI, CA, AP, \cdots$$

$$H = 1B + 2B + 3B + HR,\ AB = AP - BB - GP - TS - SF, \cdots$$

$$AVE = \frac{H}{AB},\ SLG = \frac{1B + 2x2B + 3x3B + 4xHR}{AB}$$

$$OBP = \frac{H + BB + GP}{AP - TS}, \cdots$$

- Todas tienen un significado preciso en el contexto del beisbol .

Sabermetría

- SABR: Society for American Baseball Research (1971).

- METRICS: Métricas.

- SABR + METRICS = SAB**E**RMETRICS.

- Sabermetría: desarrollo EMPIRICO de variables sobre beisbol. Se va mucho más allá de conteo y promedios ponderados tradicionales.

- Algunos de los pioneros: Craig R. Wright (empleado de la MLB, el primer Sabermetrician, 1981), Davey Johnson (pelotero, manager y matemático, 1965), Bill James (escritor sobre beisbol, 1980), Billy Bean (gerente de los Atléticos de Oakland, 2001).

Para qué?

Por siempre:

- Para divertirnos, para discutir, para que hagamos click en págs WEB (FanGraphs, Baseball References, MLB, LVBP), para hacer un Happy Hour de Asovac...

- Para estimar el valor de un jugador.

En la era sabermétrica:

- Para pronosticar carreras anotadas, recibidas y victorias.

- Para estimar el valor de un jugador en función de su aporte a las carreras anotadas, a las recibidas y a las victorias.

- Para diseñar equipos y los lineups de los mismos.

Ejemplos de variables sabermétricas

• Variables sabermétricas: funciones de funciones de .. de variables tradicionales:

$$OPS = OBP + SLG$$

$$wOBA = \frac{1.95HR + 1.553B - \cdots + \cdots + 0.72(BB - BBI)}{AP - TS - BBI}$$

$$\text{pythagorean variable} = pyt = \frac{CA^{\gamma}}{CA^{\gamma} + CR^{\gamma}}$$

$$WAR, \ BABIP, \ RAR, \cdots$$

• Para entender lo que intentan medir debe examinarse la construcción de los parámetros que las definen. Pronostican carreras y victorias? Evalúan peloteros?

Construcción del *wOBA* (The Book: Tom M Tango, Mitchell G Lichtman, Andrew E Dolphin (2007) Ed. Potomac)

- Tango dice que The Book ES *el librito*.

- El libro tiene errores conceptuales asombrosos (de matemáticas, de estadística, de teoría de probabilidades).

- El libro es extraordinario (han podido obviar los basamentos matemáticos que en realidad no usan). Construyen paso a paso una de las variables sabermétricas mas reconocidas: el *wOBA*.

wOBA

Recordemos el SLG:

$$SLG = \frac{1B + 2x2B + 3x3B + 4xHR}{AB}$$

• En el SLG solo aparecen 4 eventos. Los eventos siempre valen lo mismo sin importar los estados antes y después del evento.

Propuesta de Tango:

$$E = \{BB, BBI, GP, 1B, \cdots, SO, out, \cdots\}, \ \ |E| = 20$$

$$S = \{((-,-,-),0), \cdots, ((*,*,*),2), \ |S| = 24$$

$v(e, sI, sF)$: valor del evento; e al pasar del estado sI al estado sF

$ce(s)$: carreras anotadas en promedio partir de s

wOBA

Ejemplos:

- $sI = ((-,*,-),0),\ e = 1B + CA,\ \ sF = ((*,-,-),0)$

$$v(e,sI,sF) = 1 + ce(sF) - ce(sI) = 1 + 0.953 - 1.189 = 0.764$$

- $sI = ((-,*,-),0),\ e = 1B + CA + out4,\ sF = ((-,-,-),1)$

$$v(e,sI,sF) = 1 + ce(sF) - ce(sI) = 1 + 0.297 - 1.189 = 0.108$$

- $sI = ((*,*,*),2),\ e = HR + 4 * CA,\ sF = ((-,-,-),0)$

$$v(e,sI,sF) = 4 + ce(sF) - ce(sI) = 4 + 0.117 - 0.815 = 3.302$$

Paréntesis

Para los defensores de los toques de sacrificio:

$$ce((*, -, -), 0) = 0.953, \quad ce((-, *, -), 1) = 0.725$$

$$ce((*, *, 0)) = 1.573, \quad ce((-, *, *)) = 1.467$$

wOBA

$$wOBAjug = \frac{\sum_{e \in E} \sum_{sI \in S} \sum_{sF \in S} v(e, sI, sF) njug(e, sI, sF)}{APjug - BBIjug - TSjug}$$

$$= \cdots = \sum_{e \in E} vjug(e) propjug(e)$$

Tango (2007): No se conoce $njug(e, sI, sF)$ pero sí conoce $nliga(e, sI, sF)$ y por supuesto $propliga(e, sI, sF)$. Eso es falso hoy en día. Pueden calcular $wOBAjug$ correctamente. El $wOBA$ original es:

$$wOBA = \sum_{e \in E} vliga(e) propjug(e)$$

$wOBA$(2018) según FanGraphs

$$wOBA =$$

$$\frac{0.69BBnI + 0.72GP + 0.89(1B) + 1.27(2B) + 1.62(3B) + 2.1HR}{AP - BBI - TS}$$

- Tango: para los defensores del OPS

$$wOBA \approx \frac{2 * OBP + SLG}{3} \;\; \text{vs} \;\; OPS = OBP + SLG$$

- $wOBA$: weighted On Base Average.

- Hay factores de corrección por stadiums (no se batea igual en todas partes, usan Estadística y Física para corregir)!

Cómo usar y mejorar el wOBA y otras variables sabermétricas

- A partir del *wOBA* se definen otras variables.

- Usar $njug(e, sI, sF)$ y no $nliga(e, sI, sF)$. Lo usan?.

- El *wOBA* claramente evalua jugadores. Mientras más alto mejor.

- No hay que exagerar: *wOBA is a statistic that uses linear weights (read: math) to determine* **exactly** *how valuable each offensive outcome truly is)*

- Bill James: calcular $v(e, sI, sF)$ según el contexto. Puede definirse el *wOBA* y todas las variables sabermétricas para situaciones críticas.

- Sirve para pronosticar carreras y victorias? Conexión con Programación Matemática.

Programación matemática para la selección de variables

- Modelo para pronosticar Y (CA, CR, JG) usando las variables X_i ($H, 1B, \cdots, BABIP, WAR$).

- Datos: Para cada variable X_i y cada equipo j en la muestra histórica. Valores de las variables $X_i^{(j)}$ y Y_j.

- El Modelo lineal en X_i:

$$Y = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i + \epsilon$$

Selección de variables

Cómo hallar el mejor modelo que use solamente k variables?

$$P(k) \;\; min \;\; \sum_{j=1}^{n} |\epsilon_j| \;\; s.a:$$

$$\epsilon_j = Y_j - (\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i^j)$$

$$z_i = 0 \implies \alpha_i = 0$$

$$\sum_{i=1}^{n} z_i = k$$

$$\alpha_0, \alpha_i \in \Re^n, \epsilon_j \in \Re, \;\; z_i \in \{0, 1\}$$

Selección de variables. A.Crema y F.Crema (2016)

• $P(k)$ se reescribe adecuadamente para que se pueda, en teoría, resolver con algún software de Programación Entera apropiado (por ejemplo CPLEX). Pero, si n es grande y k está cerca de $\frac{n}{2}$ el tiempo de CPU puede ser inadmisible.

• Un buen modelo usando k variables puede construirse con un buen modelo usando $k-1$ variables añadiendo y eliminando un número controlado de variables.

• Se resuelve $P(1), P(2), \cdots, P(k_1)$. Luego se usan Búsquedas Locales para hallar buenos modelos para $k_1+1, k_1+2, \cdots, k_2$ para luego resolver $P(k_2+1), \cdots, P(n)$

Selección de variables

Table: Modelos de pronóstico de Carreras. Busquedas locales sucesivas 1389 equipos de MLB (1967-2914) (kmax=7)

Modelo	k	error promedio	maxerror	Tiempo (seg.)
Atomos	13	26.68	296	20
At+Ave	19	26.31	282	64
Sabr	4	16.99	75	7
	8	16.92	77	
Ave+Sabr	6	16.82	74	23
	14	16.76	75	
Todas	27	15.69	88	514

Selección de variables

Table: Modelos de pronóstico de Porcentaje de Victorias. Busquedas locales sucesivas 1389 equipos de MLB (1967-2914) (kmax=5)

Modelo	k	error promedio	maxerror	Tiempo (seg.)
Atomos	11	2.5	10.74	78
	26	2.4	10.08	
At+Ave	26	1.9	8.70	437
	33	1.89	8.71	
Sabr	24	1.85	8.32	1610 (12600)
	40	1.83	8.38	

Construcción de la variable Pitagórica: *pyt* (Baseball abstracts: Bill James (1983) Ed. Ballantine). La idea primitiva: un modelo elemental válido.

Se gana si $CA > CR$. Se define la Calidad de los equipos:

$$Cal_1 = \frac{CA_1}{CR_1}, \quad Cal_2 = \frac{CA_2}{CR_2} = \frac{CR_1}{CA_1}$$

Modelo probabilístico elemental:

$$ProbGana_1 = \frac{Cal_1}{Cal_1 + Cal_2}, \quad ProbGana_2 = \frac{Cal_2}{Cal_1 + Cal_2}$$

con lo cual:

$$ProbGana_1 = \frac{\frac{CA_1}{CR_1}}{\frac{CA_1}{CR_1} + \frac{CR_1}{CA_1}} = \frac{CA_1^2}{CA_1^2 + CR_1^2}$$

Construcción de la variable Pitagórica. La práctica

$$pyt = \frac{CA^2}{CA^2 + CR^2}, \; pyt = \frac{CA^\gamma}{CA^\gamma + CR^\gamma}$$

γ se ajusta con los datos de una liga usando métodos estadísticos

$$min \sum_{i=1}^{n} (Prop_i - pyt_i)^2$$

- Asumen el mismo γ para toda la liga.

- No toman en cuenta el calendario restante. Las proyecciones podrían ser incompatibles. Eso puede arreglarse añadiendo restricciones al problema.

Validación experimental-analítica de la variable pitagórica (Steven J. Miller: A Derivation of the Pytahgorean won-loss formula in Baseball. Chance 20 (1), 40-48 (2007))

• Miller valida experimentalmente que CA y CR pueden ajustarse con Variables Weibull independientes con parámetros α_A, γ y α_R, γ respectivamente.

• Lo que sigue es analítico (luego de algunas cuentas muy complicadas):

$$\text{Prob de ganar} = \text{Prob}\{CA > CR\} = \frac{\mu_A^\gamma}{\mu_A^\gamma + \mu_R^\gamma}$$

con μ_A y μ_R los valores esperados (promedios) de CA y CR
• γ debería depender del equipo y de si se trata de CA o de CR. En la práctica se busca un solo γ.

Validación experimental-analítica de la fórmula pitagórica

$$Prob\{CA > CR\} = \frac{\mu_A^\gamma}{\mu_A^\gamma + \mu_R^\gamma}$$

con lo cual en la práctica:

$$Prob(ganar) = pyt = \frac{CA^\gamma}{CA^\gamma + CR^\gamma}$$

Para qué sirve la fórmula pitagórica

A mitad de temporada tenemos $prop, CA$ y CR para un equipo

Calculamos

$$pyt = \frac{CA^\gamma}{CA^\gamma + CR^\gamma}$$

y comparamos *prop* con *pyt*

Ya que el ajuste es bueno podemos concentrarnos en desarrollar modelos para proyectar CA (con un modelo de la ofensiva) y CR (con un modelo para la defensiva y el pitcheo) para luego proyectar Prob de ganar usando la fórmula pitagórica.

pyt para MLB y LVBP

- Se ajustó el modelo para la MLB con 1389 equipos desde 1967 hasta 2014.

$$\gamma = 1.85, \;\; \text{error promedio} = 0.0198, \; \text{maxerror} = 0.0895$$

- idem para la LVBP con 32 equipos desde 2015 hasta el domingo 18 de Noviembre de 2018.

$$\gamma = 1.29, \;\; \text{error promedio} = 0.0372, \; \text{maxerror} = 0.1400$$

pyt para la temporada 2018

- Se ajustó el modelo para la LVBP con 8 equipos (temporada 2018 hasta el domingo 18 de noviembre)

$$\gamma = 1.4784, \;\; \text{error promedio} = 0.0323, \; \text{maxerror} = 0.0564$$

Proyección temporada 2018

Table: Proyección temporada 2018

Equipo	G	P	prop	pyt	G	G
Cardenales	25	12	.636	.597	38.9	39
Navegantes	18	14	.563	.526	34.3	34
Leones	16	15	.516	.510	32.3	**32**
Bravos	17	16	.515	.548	33.4	**33-34**
Aguilas	15	17	.469	.514	30.9	31
Tigres	14	16	.467	.453	28.9	29
Tiburones	13	18	.419	.383	25.2	**25**
Caribes	13	19	.406	.456	27.1	**27**

- Con Bravos 34: 251 Ganados (Temporada tiene 252 juegos).