Universidad Nacional Autónoma de México

Facultad de Ciencias

Un proceso Poisson modificado y su aplicación a la predicción de marcadores en futbol soccer

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Matemático

PRESENTA:

Esteban Navarro Garaiz

DIRECTOR DE TESIS:
Dra. María Asunción Begoña Fernández Fernández

Ciudad de México, 2018

Índice general

Hoja de contacto

Esteban Navarro Garaiz

Un Proceso Poisson modificado y su aplicación a la predicción de marcadores en futbol soccer
Universidad Nacional Autónoma de México, Facultad de Ciencias

Carrera: Matemáticas
Número de cuenta: 309626680
Correo electrónico: enavarro@ciencias.unam.mx / EstebanNavarro-Garaiz@gmail.com

Grado y nombre de propietaria/tutora: Dra. María Asunción Begoña Fernández Fernández
Correo electrónico: bff@ciencias.unam.mx

Grado y nombre de propietaria: Dra. Ana Meda Guardiola
Correo electrónico: ana.meda@ciencias.unam.mx

Grado y nombre de propietario: Dr. Fernando Baltazar Larios
Correo electrónico: fernandobaltazar@ciencias.unam.mx

Grado y nombre de suplente: Dr. Yuri Salazar Flores
Correo electrónico: yurisf@ciencias.unam.mx

Grado y nombre de suplente: Dr. Arrigo Coen Coria
Correo electrónico: coen@ciencias.unam.mx

Agradecimientos

Esta tesis no habría sido posible sin el apoyo de mucha gente que puso su granito de arena, de una manera u otra. Estaré eternamente agradecido por sus invaluables contribuciones.

Primero, a mis padres, Diana y Alejandro, por su cariño y apoyo en todo el proceso.

Segundo, a mi tutora, Begoña, por su infinita paciencia y guía. Nunca olvidaré la sesión de ocho horas para arreglar el último problema fundamental en el trabajo.

Tercero, a mis sinodales, Ana, Fernando, Yuri, y Arrigo por sus atenciones y retroalimentación.

Además, me gustaría agradecer a toda persona que formó una parte importante de mi estancia en la UNAM, aportando a mi formación como matemático y persona en algún momento . No podría haberlo logrado sin ustedes. En ningún orden en particular: José Antonio Perusquia, Mario Delgadillo, Bruno Martínez, José de Jesús Arias, Antonio Soriano, Adrián Girard, Ruth Fuentes, Daniel Cervantes, Rafael Miranda, Alejandro García.

Por último, quiero agradecer a las personas que tengo el placer de llamar amigos, cuya influencia directa formó parte crucial de este trabajo: Ulises Hernández, Robert Pizzola, Blas Kolic, Ignacio Loaiza.

Introducción

La intención de esta tesis es desarrollar un modelo de predicción para partidos de futbol que sea capaz de generar una ganancia en el mercado de apuestas. Además de lo atractivo de ganar dinero apostando en futbol, que el modelo pueda ganar dinero - como en cualquier otro mercado - nos hace saber que está añadiendo información al mismo. En particular, el mercado de apuestas deportivas es altamente eficiente, por lo que sólo un modelo bien calibrado podría tener éxito a largo plazo.

En el capítulo 1, discutimos todas las matemáticas usadas en el trabajo. En Modelación histórica de futbol soccer, exploramos a detalle la literatura previa en modelación de futbol. Luego, en el capítulo 3, desglosamos el mercado de apuestas deportivas y su funcionamiento. En Estrategias de apuesta introducimos el Criterio de Kelly que, hasta donde sabe el autor, no ha sido usado nunca en el contexto de apuestas deportivas dentro de la literatura académica.

Posteriormente, en Análisis de marcadores y apuestas hacemos un análisis del mercado de apuestas y su relación con las probabilidades reales de ocurrencia. Discutimos la evidencia empírica para explorar si la naturaleza de los marcadores en futbol soccer está bien representada por una distribución Poisson. En Modelos que usan Poisson bivariada, exploramos minuciosamente el modelo Poisson bivariado, sus fortalezas, debilidades y uso histórico en la modelación de futbol.

En Poisson doble dinámico proponemos una modificación al modelo Poisson doble, presentado originalmente en Maher (1982) [1], para hacerlo dinámico en el tiempo.

Por último, el capítulo 8 discute el funcionamiento del modelo ∇-Poisson doble, propuesto por el autor, a detalle: cómo obtiene las apuestas a realizar y el tamaño de la apuesta. Exploramos además su funcionamiento en el mercado de apuesta para ver si puede generar una ganancia consistentemente.

Capítulo 1

Preliminares Matemáticos

La herramienta principal de esta tesis es la Distribución Poisson y algunas distribuciones derivadas de ella. Por ello, a continuación discutimos el origen histórico de la distribución, su definición y propiedades, además de las distribuciones relacionadas que se utilizan en el trabajo.

1.1. Origen de la distribución Poisson

La distribución Poisson recibe su nombre de Siméon Denis Poisson, un matemático francés nacido a finales del siglo XVIII, publicada en sus últimos años de vida en "Recherches sur la probabilité des jugements en matière criminelle et en matière civile"(Investigación sobre la probabilidad de juicios en materia criminal y civil). La publicación presentaba una teoría intentando medir el número de decisiones erroneas de jurados, dado el número de eventos (juicios) en un intervalo temporal de longitud dada. La distribución no obtuvo mucha atención en esta forma. Sin embargo, Abraham de Moivre había encontrado el mismo resultado más de cien años antes.

La derivación es obtenida por Poisson aproximando la distribución como el límite de una sucesión de variables aleatorias con distribución Binomial en dónde N (el número de ensayos) tiende a infinito, p tiende a cero y Np permanece finito e igual a λ.
Sin embargo, las distribuciones Binomial y Poisson tienen una diferencia fundamental: la distribución Binomial especifica explícitamente el número de fracasos, ya que sabemos el número de experi-

mentos y de éxitos. Por ejemplo, si tiramos una moneda diez veces y obtenemos un lado en tres ocasiones, sabemos que el otro lado ocurrió siete veces.
La distribución Poisson no determina el número de fracasos. Es decir, es una especie de extensión de la distribución binomial donde el número de volados no está determinado y es de alguna manera aleatorio también: aunque a largo plazo la proporción de éxitos está determinada por Np, en realidad a la aproximación no le importa cómo se llega a dicha proporción. Imagine dos distribuciones binomiales, una con $p = 0.1$ y $N = 20$ y otra con $p = 0.05$ y $N = 40$. A pesar de tener dos naturalezas distintas, particularmente en el dominio, la aproximación por una distribución Poisson será exactamente la misma en ambos casos, $X \sim Poisson(2)$.

Lo anterior responde a otra propiedad importante de la distribución: si $X \sim Poisson(\lambda)$, entonces $E[X] = Var[X] = \lambda$. Es decir, el primer momento es lo único que determina la distribución, no hay un segundo parámetro para la varianza, como en la distribución normal. Nuevamente, está ligado intrínsecamente a la naturaleza de la aproximación binomial: como sólo nos interesa el número de éxitos Np; λ es la única medida que cambia en la distribución. El número total de experimentos y la probabilidad de éxito no pueden ser estimados a nivel distribución.

Esta simplicidad representa la gran utilidad y, a su vez, el problema más grande de la distribución Poisson. Por un lado, dado que los primeros dos momentos son iguales, la estimación se vuelve muy simple. Como el cociente de la varianza y la media es uno, se dice que la distribución tiene equidispersión. No obstante, eso se encuentra en pocos datos en la vida real. Físicamente, surge cuando se cumplen dos condiciones: cuando los cambios son homogéneos en el tiempo y cuando los eventos futuros son independientes de lo sucedido en el pasado. Es una distribución discreta, entera, donde el número de eventos en un intervalo tiene una taza constante; es decir, el siguiente evento es independiente del último. Aunque ambos supuestos son en muchas ocasiones poco realistas, la distribución Poisson sigue haciendo un gran trabajo de aproximación dada su simpleza. Sin embargo, tendrá problemas ajustando datos que tengan sobredispersión (la varianza excede a la media) o infradispersión

(la media excede a la varianza).

1.2. Definición, usos y ejemplos de la distribución Poisson

Definición 1. Una variable aleatoria **X** que toma valores en $\mathbb{N}\cup\{0\}$ es una **varibale aleatoria Poisson** con parámetro $\lambda > 0$ si:

$$\mathbb{P}[X = x] = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in \mathbb{N} \cup \{0\}$$

Ésta es una variable aleatoria, pues:

$$\sum_{i=0}^{\infty} \mathbb{P}[\mathbf{X} = i] = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$$

La suma en el segundo término de la igualdad se convierte en e^{λ} pues ésta es la expresión en Serie de Taylor de e^{λ}.

Ejemplo 1. Suponga que el número de errores cometidos por una capturista de datos en cada diez páginas es una variable aleatoria Poisson con parámetro $\lambda = 15$.

a) ¿Cuál es la probabilidad de que cometa exactamente 15 errores en un documento de 10 páginas?

$$\mathbb{P}\left[15 \text{ errores}\right] = \frac{e^{-15}\,15^{15}}{15!} = 0.1024359$$

b) ¿Cuál es la probabilidad de que cometa al menos 10 errores?

$$\mathbb{P}\left[\text{Al menos 10 errores}\right] = 1 - \sum_{k=0}^{9} \mathbb{P}[\text{cometió k errores}] = 1 - \sum_{k=0}^{9} \frac{e^{-15}\,15^k}{k!} = .9301463$$

c) ¿Cuál es la probabilidad de que cometa más de 30 errores?

$$\mathbb{P}\left[\text{Más 30 errores}\right] = 1 - \sum_{k=0}^{30} \mathbb{P}[\text{cometió k errores}] = 1 - \sum_{k=0}^{30} \frac{e^{-15}\,15^k}{k!} = .0001973$$

1.2.1. Relación con Distribución exponencial

Aunque los supuestos para poder usar la distribución son fuertes - eventos independientes, taza constante y no co-ocurrencias (es decir, sólo puede ocurrir un evento en un instante cualquiera) - la distribución Poisson está intimamente ligada a muchas otras. Como ya hemos discutido antes, sirve para aproximar una Distribución binomial donde el número de experminetos n es grande y la probabilidad de éxito p es pequeña. Es decir, sirve muy bien para modelar eventos raros dónde tenemos muchas realizaciones del experimento. Además, está íntimamente relacionada con las variables aleatorias exponenciales.

Definición 2. Una variable aleatoria continua $\mathbf{X}$ tiene una **Distribución exponencial** con parámetro $\lambda > 0$ si su función de densidad está dada por:

$$f_{\mathbf{X}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Su función de acumulación está dada por:

$$F_{\mathbf{X}}(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

La función exponencial tiene propiedades fascinantes y muy útiles, que discutimos a continuación, tras algunas definiciones preliminares:

Definición 3. Una variable aleatoria tiene **pérdida de memoria** si:

$$\mathbb{P}[\mathbf{X} > s + t | \mathbf{X} > t] = \mathbb{P}[\mathbf{X} > s] \quad \forall s, t \geq 0$$

Equivalentemente, una variable tiene pérdida de memoria si, $\forall s, t \geq 0$:

$$\frac{\mathbb{P}[\mathbf{X} > s+t \cap \mathbf{X} > t]}{\mathbb{P}[\mathbf{X} > t]} = \mathbb{P}[\mathbf{X} > s]$$

$$\mathbb{P}[\mathbf{X} > s+t] = \mathbb{P}[\mathbf{X} > s]\,\mathbb{P}[\mathbf{X} > t]$$

Como $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$, las variables aleatorias exponenciales tienen pérdida de memoria; más aún, se puede probar que éstas son las únicas variables aleatorias continuas con ésta propiedad.

Definición 4. Sea $\mathbf{X}$ una variable aleatoria continua y positiva que representa la vida de un objeto. La **tasa de riesgo** $\lambda(t)$ de F es:

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)}$$

dónde $\bar{F} = 1 - F(t)$ representa a la función de supervivencia.

$\lambda(t)$ representa la distribución condicional de fallo para una unidad con t años de edad, es decir, cómo esperamos que se comporte la supervivencia de un objeto que ya ha sobrevivido t años.

Sin embargo, para una variable aleatoria $\mathbf{X}$ con distribución exponencial, la función de riesgo está dada por:

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

Es decir, la función de riesgo para una variable aleatoria exponencial es constante a través del tiempo e igual al parámetro de la distribución. Ésto es completamente consistente con la pérdida de memoria y quiere decir que, sin importar el tiempo que haya pasado ya, si un objeto ha sobrevivido hasta el tiempo t, su función de riesgo es exactamente igual a si no hubiera pasado tiempo y estuviéramos al principio de su vida.

En particular, el riesgo no aumenta con el tiempo (como con un electrónico, digamos, una computadora). Muchas funciones de

riesgo son de ésta forma; por ejemplo, objetos con tiempo de vida normal. La función de riesgo para una variable aleatoria normal es estríctamente creciente. Ésto tiene todo el sentido del mundo, pues la distribución normal (tomando sólo valores positivos) tiene una cola ligera que hace que una duración larga sea cada vez más rara. La función de riesgo podría disminuír con el tiempo también. Por ejemplo, la mortalidad infantil está concentrada en los primeros años de vida y mientras el infante sigue creciendo, tiene menor probabilidad de morir.

1.3. Procesos estocásticos

Definición 5. Sea $\mathcal{T} \subseteq [0, \infty)$. Una familia de variables aleatorias, $\{X_t\}_{t\in\mathcal{T}}$ se conoce como un **Proceso estocástico**. Si $\mathcal{T} = \mathbb{N}$ o a cualquier conjunto a lo más numerable, decimos que $\{X_t\}_{t\in\mathcal{T}}$ es un **Proceso estocástico discreto** Si $\mathcal{T} = [0, \infty)$ o a cualquier conjunto infinito, decimos que es un **Proceso estocástico continuo**

Si $\mathcal{T}$ es, por ejemplo, $\{1\}$, tenemos una sola variable aleatoria. Si $\mathcal{T}$ es un conjunto finito $\{1, 2, \ldots, N\}$, entonces tenemos un vector aleatorio. Los procesos estocásticos son una especie de generalización de un vector aleatorio, pero éste depende ahora del tiempo. Ofrecen una variedad de posibilidades que son más fáciles de entender con un ejemplo.

Ejemplo 2. Suponga que un restaurante tiene n mesas y sea $X_{i,t}$ el número de personas sentadas en la i-ésima mesa al tiempo t. Entonces podemos medir múltiples cosas.

Si fijamos i (es decir, una mesa), entonces tenemos una sola variable aleatoria: las personas sentadas en esa mesa durante el día a cualquier tiempo. De ésta podrían interesarnos el total de personas que se sentaron en esa mesa durante el día, el máximo/mínimo de personas a través de un día; la media, la varianza, cuantas horas estuvo vacía u ocupada. Además, podríamos comprar mesas entre sí. Podrían ser todas idénticas o no: por ejemplo, si una mesa está en la terraza, la ocupación de la misma podría depender de la hora y el clima. Además podría haber mesas de distintos tamaños (con

capacidad para 2, 4 o más personas) y tipos (por ejemplo, los gabinetes suelen usarse más las mesas acomodadas en el medio del restaurante).

Si ahora dejamos a i libre y fijamos t, lo que tenemos es la cantidad de personas sentadas en cada mesa a una hora determinada. Aquí podría interesarnos cuántas personas están en el restaurante en ese momento (la suma), comparar entre horas (hora de la comida/cena, contra el resto del día), o fijar la hora del día y comparar con el resto de los días (por ejemplo, viernes a las 3PM contra lunes a las 3PM).

Si ambos i y t corren libre, entonces tenemos un proceso estocástico: el número de ocupantes en cada una de las n mesas a través del tiempo.

1.3.1. El Proceso Poisson

Definición 6. Un Proceso estocástico $\{N(t); t \geq 0\}$ es un **Proceso de conteo** si $N(t)$ representa el número total de eventos hasta el tiempo t. Más formalmente, un Proceso de conteo debe satisfacer:
1.- $N(t) \in \mathbb{N} \cup \{0\}$
2.- Si $s < t$, $N(s) \geq N(t)$
3.- Si $s < t$; $N(t) - N(s)$ es el número de eventos ocurridos en el intervalo $(s, t]$

Definición 7. Un Proceso de conteo $\{N(t); t \geq 0\}$ tiene **incrementos independientes** si el número de eventos que ocurren en intervalos de tiempo disjuntos son independientes. Por ejemplo, sea $t < s$. Entonces el número de eventos ocurridos hasta t, $N(t)$, es independiente del número de eventos que ocurrirán en el intervalo $(t, t + s]$, $N(t + s) - N(t)$.

Definición 8. Un proceso de conteo $\{N(t); t \geq 0\}$ tiene **incrementos estacionarios** si la distribución del número de eventos que ocurren en un intervalo de tiempo, depende solamente de la longitud del intervalo. En otro términos, si el proceso tiene incrementos estacionarios, si el número de eventos en el intervalo $(0, t]$ es exactamente igual que el intervalo $(s, s + t]$ pues ambos son de longitud t.

Definición 9. Un proceso de conteo $\{N(t); t \geq 0\}$ es un **Proceso Poisson** con parámetro $\lambda > 0$ si:

1.-$N(0) = (0)$
2.- El proceso tiene incrementos independientes.
3.- El número de eventos en un intervalo de longitud t, tiene distribución Poisson con media λt, es decir:

$$\mathbb{P}[N(t+s) - N(s) = n] = e^{-\lambda t}\frac{(\lambda t)^t}{n!} \quad \forall s, t \geq 0 \quad n \in \{0, 1, \ldots\}$$

De 3, un Proceso Poisson tiene incrementos estacionarios y, además, $\mathbb{E}[N(t)] = \lambda t$. Sin embargo, 3 es muy difícil de determinar en la vida real, por lo que definimos a los Procesos Poisson equivalentemente usando una herramienta adicional: las funciones $o(h)$

Definición 10. La función f es $o(h)$ si:

$$\lim_{h\to 0}\frac{f(h)}{h} = 0$$

Es decir, f es $o(h)$ si para valores pequeños de h $f(h)$ es, relativamente, aún más pequeña.

Ejemplo 3. $f(x) = x^2$ es una función $o(h)$ pues:

$$\lim_{h\to 0}\frac{f(h)}{h} = \lim_{h\to 0}\frac{h^2}{h} = \lim_{h\to 0} h = 0$$

Teorema 1. *Si f y g son $o(h)$, $c \in \mathbb{R}^+$, entonces: $(f+g)(h)$, $c * f$ son $o(h)$, y, por lo tanto, cualquier combinación lineal de funciones $o(h)$ es $o(h)$ también.*

La prueba es sencilla y simplemente se hereda por propiedades de límites.

Estamos listos ahora para dar una nueva definición de Proceso Poisson que, en la práctica, es mucho más sencilla de utilizar:

Definición 11. Un Proceso de conteo $\{N(t); t \geq 0\}$ es un **Proceso Poisson** con parámetro $\lambda > 0$ si:
1.-$N(0) = (0)$

2.- El proceso tiene incrementos independientes y estacionarios.
3.- $\mathbb{P}[N(h) = 1] = \lambda h + o(h)$
4.- $\mathbb{P}[N(h) \geq 2] = o(h)$

Las condiciones 3 y 4 de esta nueva definición son fáciles de determinar pues, en general, conocemos la distribución del proceso de conteo $N(t)$. Eso hace que esta definición sea mucho más útil que la primera. Además, ambas son equivalentes. Una prueba se puede encontrar en Ross (2007).

Definimos ahora partes del Proceso Poisson que nos resultarán útiles.

Definición 12. Sea $\{N(t); t \geq 0\}$ un Proceso Poisson. Denotaremos por T_1 el **tiempo de ocurrencia para el primer evento**. Para $n > 1$, denotaremos por T_n el **tiempo de ocurrencia entre el evento** $n-1$ **y el** n**-ésimo evento**. La sucesión $\{T_n; n = 1, 2, \ldots\}$ es conocida como **tiempos interarribo**.

Ejemplo 4. Elsa tiene una papelería y mide el tiempo de llegada de los primeros cuatro clientes a partir de el momento en que abre, que denominaremos como $t = 0$. Supongamos que los clientes llegan en los siguientes tiempos: $c_1 = 20, c_2 = 47, c_3 = 49, c_4 = 95$. Entonces, los tiempos interarribo son: $T_1 = 20, T_2 = 27, T_3 = 2, T_4 = 46$.

El ejemplo deja entrever la relación entre los tiempos interarribo y los tiempos reales de llegada. Por ejemplo, el tercer cliente llega en el tiempo 49, pero éste es simplemente la suma de los primeros tres tiempos interarribo $20 + 27 + 2 = 49$. Más adelante trabajaremos con ésta relación y su distribución cuidadosamente.

Para poder determinar la distribución de T_n, trabajaremos primero con T_1. Sea $N(t)$ un Proceso Poisson de parámetro λ, i.e., $N(t) \sim Poisson(\lambda t)$. Entonces, para cualquier tiempo t el evento $[T_1 > t]$ ocurre si y sólo si $N(t) = 0$, pues si el tiempo de ocurrencia del primer evento es mayor a t, eso quiere decir que al tiempo t no ha ocurrido ningún evento. Entonces:

$$\mathbb{P}[T_1 > t] = \mathbb{P}[N(t) = 0] = e^{-\lambda t}$$

Por lo tanto: $T_1 \sim Exponencial(\lambda)$ i.e. media $\frac{1}{\lambda}$

Ahora, tras la primera ocurrencia, para T_2:

$$\mathbb{P}[T_2 > t | T_1 = s] = \mathbb{P}[0 \text{ eventos en}(s, s+t] \mid t_1 = s] = \mathbb{P}[0 \text{ eventos en}(s, s+t] \,] = e^{-\lambda t}$$

Éstas últimas dos igualdades se dan por tener incrementos independientes y estacionarios, respectivamente. Por lo tanto: $T_2 \sim Exponencial(\lambda)$ también.

Usando los incrementos independientes y estacionarios, podemos generalizar de la misma manera y probar que, $T_n \sim Exponencial(\lambda)$ para toda n.

Esto surge naturalmente pues, dadas las hipótesis sobre los incrementos, esperaríamos que el proceso no tuviera memoria y la Distribución exponencial es la única distribución continua con tal propiedad y, por tanto, la que tiene sentido.

1.4. Distribuciones relacionadas: Poisson doble y Poisson bivariada

Definición 13. Un vector aleatorio (X, Y) que toma valores en $\mathbb{Z}^+ \times \mathbb{Z}^+$ tiene una distribución Poisson doble con parámetros $\lambda, \mu > 0$ si su función de densidad está dada por:

$$\mathbb{P}[X = x, Y = y] = \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^y e^{-\mu}}{y!}$$

Alternativamente, (X, Y) tiene distribución Poisson doble si es el producto de dos variables aleatorias independientes X, Y tal que $X \sim Poisson(\lambda)$, $Y \sim Poisson(\mu)$

Definición 14. Sean X_1, X_2, X_3 variables aleatorias Poisson con parámetro λ_1, λ_2, λ_3 respectivamente. Definimos $X = X_1 + X_3$, $Y = X_2 + X_3$. Entonces, decimos que el vector aleatorio (X, Y) tiene distribución **Poisson bivariada**. Veamos cuál es su función de densidad:

$$\mathbb{P}[X = k_1, Y = k_2] = \mathbb{P}[X_1 + X_3 = k_1, X_2 + X_3 = k_2] =$$

$$\sum_{k=0}^{\infty} \mathbb{P}[X_1 + X_3 = k_1, X_2 + X_3 = k_2 | X_3 = k] \, \mathbb{P}[X_3 = k] = \psi$$

Como $X_3 = k$, podemos sustituir el condicional a la ecuación. Por otro lado, como X_1, X_2 tienen distribución Poisson, solo acumulan valor en los positivos. Así:

$$\psi = \sum_{k=0}^{min(k_1,k_2)} \mathbb{P}[X_1 = k_1 - k, X_2 = k_2 - k] \, \mathbb{P}[X_3 = k] = \sum_{k=0}^{min(k_1,k_2)} \frac{\lambda_1^{k_1-k} \, e^{-\lambda_1}}{(k_1-k)!} \frac{\lambda_2^{k_2-k} \, e^{-\lambda_2}}{(k_2-k)!} \frac{\lambda_3^{k} \, e^{-\lambda_3}}{k!}$$

$$= e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{k=0}^{min(k_1,k_2)} \frac{\lambda_1^{k_1-k} \lambda_2^{k_2-k} \lambda_3^{k}}{(k_1-k)!(k_2-k)!k!} = \Upsilon$$

Hay dos maneras alternativas de terminar las cuentas, que muestran la utilidad de la distribución. La primera es simplemente regresando la exponencial a la suma:

$$\Upsilon = \sum_{k=0}^{min(k_1,k_2)} \frac{e^{-\lambda_1} \lambda_1^{k_1-k} e^{-\lambda_2} \lambda_2^{k_2-k} e^{-\lambda} \lambda_3^{k}}{(k_1-k)!(k_2-k)!k!} = f_{Poiss(\lambda_1)}(k_1 - k) f_{Poiss(\lambda_2)}(k_2 - k) f_{Poiss(\lambda_3)}(k)$$

Es decir, con el dominio correcto, la distribución Poisson bivariada es simplemente el producto de tres Poisson univariadas independientes; las primeras dos evaluadas sustrayendo el valor evaluado en la tercera. La suma con límite en el mínimo deja entrever que hay muchas maneras de llegar al valor conjunto utilizando los valores marginales. Y con esa intuición, presentamos la segunda forma de la distribución, que es la más útil al hacer cálculos. Multiplicando por $1 = \frac{k!k_1!k_2!}{k!k_1!k_2!}$:

$$\Upsilon = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1}{k_1!} \frac{\lambda_2}{k_2!} \sum_{k=0}^{min(k_1,k_2)} \frac{k_1!}{(k_1-k)!k!} \frac{k_2!}{(k_2-k)!k!} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k$$

$$= e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1}{k_1!} \frac{\lambda_2}{k_2!} \sum_{k=0}^{min(k_1,k_2)} \binom{k_1}{k} \binom{k_2}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k$$

Entonces, decimos que el vector aleatorio (X, Y) se distribuye Poisson bivariado, $(X, Y) \sim BP(\lambda_1, \lambda_2, \lambda_3)$, con función de densidad:

$$\mathbb{P}[X = x, Y = y] = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1}{k_1!} \frac{\lambda_2}{k_2!} \sum_{k=0}^{min(k_1,k_2)} \binom{k_1}{k} \binom{k_2}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k$$

Las propiedades de la Poisson bivariada son muy nobles y útiles; por ello, es un modelo con muchas aplicaciones. Primero, las marginales son ambas Poisson: $X \sim Poisson(\lambda_1 + \lambda_3)$, $Y \sim Poisson(\lambda_2 + \lambda_3)$. Es en ese sentido, es una distribución muy similar a la Poisson doble, donde λ_3 es el parámetro que introduce la dependencia entre ambas distribuciones. En Kocherlakota (2006) [2], se prueba que $Cov(X, Y) = \lambda_3 > 0$, lo cual permite cuantificar exactamente la dependencia de las dos variables cuando ésta es positiva. Así, una condición necesaria y suficiente para que X,Y sean independientes, es que $\lambda_3 = 0$. Por ello, la distribución Poisson doble es un caso particular de la distribución Poisson bivariada. Dada la forma de la covarianza, sabemos entonces que la correlación está dada por: $\rho_{X,Y} = \frac{\lambda_3}{\sqrt{(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}}$

Como $\lambda_3 > 0$, entonces necesariamente $\rho_{X,Y} \geq 0$. Así, $\rho_{X,Y} = 0$ es también una condición necesaria y suficiente para la independencia de X, Y. Una discusión de todas las propiedades condicionales de la distribución se puede encontrar en Kocherlakota (2006) [2].

Teicher (1954) **??** da fórmulas de recurrencia para obtener todos los valores de la función de probabilidad a partir de los valores de la función de probabilidad marginal, que sabemos es Poisson univariada y por tanto fácil de calcular.

Una de las mayores ventajas del modelo radica precisamente en que se puede tomar en cuenta y reflejar la covarianza en los datos. Además, es fácil de interpretar y manejar, dadas las marginales

Poisson. En su simplicidad, sin embargo, se encuentran sus mayores problemas. El modelo sólo puede modelar covarianza positiva entre las variables, lo cual lo limita. Además, ¿cómo podemos modelar λ_3? ¿es constante en el tiempo? ¿depende de las variables en cuestión? ¿cómo podemos hacerla variar correctamente?

En Karlis, Ntzoufras (2003) [3], los autores proponen un modelo inflado por la diagonal, por lo que exploramos brevemente los modelos inflados.

En cuanto a la distribución Poisson, el más común es un modelo inflado en cero - en inglés ZIP: Zero-Inflated Poisson - los cuales son muy populares en la práctica, particularmente en medicina y seguros. Por ejemplo, el número de reclamaciones a un seguro dentro de una población estaría inflado en cero por la gente que no se ha asegurado contra el riesgo en cuestión y, por lo tanto, son incapaces de generar una reclamación.

La idea del modelo es la siguiente:

Para cada observación, hay dos casos posibles. En el primer caso, la observación es cero. En el segundo caso, la observación se genera a través de un modelo Poisson(λ) - la cual podría generar un cero inherente a la distribución también (con probabilidad $e^{-\lambda}$). Supongamos que la probabilidad del caso 1 es π y la probabilidad del caso 2 es $1-\pi$. Entonces, la función de acumulación de un modelo inflado por cero está dada por:

$$\mathbb{P}[X = k] = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{si } k = 0 \\ (1-\pi)\frac{\lambda^k e^{-\lambda}}{k!} & \text{si } k > 0 \end{cases} \tag{1.1}$$

Los casos en la literatura bivariada son escasos. Karlis, Ntzoufras (2003)[3] proponen un modelo inflado por la diagonal para modelar marcadores de futbol, es decir, las probabilidades donde $X = Y$. En futbol, estos son empates. La modificación propuesta al modelo Poisson bivariado, que busca tener una mejor estimación de los empates, tiene la siguiente forma:

$$f_{IBP}(x,y) = \begin{cases} (1-p) f_{BP}(x,y|\lambda_1,\lambda_2,\lambda_3) & \text{si } x \neq y \\ (1-p) f_{BP}(x,y|\lambda_1,\lambda_2,\lambda_3) + p f_d(X|\theta) & \text{si } x = y \end{cases}$$

dónde $f_d(X|\theta)$ es la función de probabilidad de alguna distribución discreta (otra Poisson, una geométrica, o una distribución discreta general con todas las probabilidades especificadas). $D(X;\theta)$ es dicha distribución discreta con dominio en $\mathbb{Z}^+$ y vector de parámetros θ.

A continuación, exploramos los antecedentes a este trabajo en la literatura, sobre modelación de marcadores en futbol soccer.

Capítulo 2

Modelación histórica de futbol soccer

La literatura en modelación de futbol soccer es mucho más reducida que en otros deportes como, por ejemplo, béisbol, donde las estadísticas avanzadas tienen décadas de desarollo.

Las primeras pruebas para intentar modelarlo, atacan la información en su totalidad, sin prestar atención a la influencia de los distintos equipos a cada uno de los resultados. Es decir, ¿cómo se comportan los marcadores de futbol en general, sin importar cuáles dos equipos jueguen el partido?

El porqué estos primeros intentos no parecen lidiar con los datos individualmente no es muy claro. La dificultad de lidiar con ellos granularmente es mucho mayor y al momento de los inicios el poder computacional era muy limitado, por lo cual quizá no era siquiera una posibilidad realista.

A continuación, discutimos estos primeros intentos. Después, hacemos una exposición de los modelos principales en la literatura, en los que está sustentado el de esta tesis. Por último, discutimos algunas otras investigaciones relevantes.

2.1. Primeros intentos

Moroney (1953) [4] es la primer investigación que utiliza datos de futbol en un intento de modelación. Aunque no sabemos nada del origen de los datos utilizados, Moroney tiene una muestra de 240 partidos (i.e. 480 marcadores registrados) en donde los goles promedio anotados por equipo en cada partido partido son 1.7. Moroney utiliza estos datos como contraejemplo - además de datos sobre suicidios - para mostrar que la distribución Poisson no modela perfectamente el fenónemno: "(...) no deberíamos esperar que la distribución Poisson de una descripción perfecta del número de goles anotados por equipo en un partido de futbol, dado que el numero esperado de goles depende, entre otras cosas, de los equipos que se enfrentan, de las condiciones climáticas, etcétera". Moroney encuentra que la varianza (1.9) es un poco más grande que la media (1.7), por lo que propone "una modificación de la Poisson"para amoldarse a la sobredispersión. A pesar de haber dicho que los equipos en el partido deben influir en el marcador, Moroney no trata los datos individualmente en ningún momento; modela los datos en conjunto, sin intentar descifrar la influencia de equipos y si hay dierencias significativas entre uno y otro. Su conclusión es que: "los factores de clima y equipos enfrentándose no ejercen un efecto tan grande como es usualmente supuesto".

Contradiciendo a Moroney; Ugarte, Militino, Arnholt (2016)[5], usando datos de la Copa Mundial desde 1990 y hasta 2002, encuentran que la distribución Poisson modela muy bien el total de goles anotado en un partido - por ambos equipos, a diferencia de Moroney que los intentaba modelar para cada equipo - con una media muestral de 2.478 y una varianza muestral de 2.458. Los valores observados y esperados son casi iguales para la distribución completa.

Tras la investigación de Moroney; Reep, Pollard, Benjamin (1971) [6] proponen una binomial negativa - que argumentan es lo que había propuesto Moroney sin nombrarla - puesto que:"la probabilidad de que se anote un gol a través del partido no es constante y es afectada por goles previos y otros factores". Los autores tienen datos de la Primera División de Inglaterra de 1965-1969 (cuatro temporadas). Nuevamente, los datos presentan varianza un poco más grande que

la esperanza (entre 15 % y 25 % mayor en las distintas temporadas). Encuentran además, que el ajuste puede ser usado con resultados similares en Hockey y las carreras anotadas por inning en la Serie Mundial de beisbol. Los autores concluyen que la modelación por binomial negativa funciona si: "hay varios jugadores involucrados y la probabilidad de éxito es relativamente invariante para todo momento y entre los mismos jugadores". Nuevamente, los autores no tratan los datos por equipo, sino el cúmulo de todos los equipos en la liga.
Sin embargo, en otra investigación, al intentar extender su estudio sobre cadenas de pases consecutivos en futbol a secuencias de golpes en tenis (Reep, Benjamin [1968] [7]), encuentran que el ajuste es muy pobre e hipotetizan que la habilidad de los jugadores en tenis es más importante y varía más entre jugadores y partidos que en futbol. Esto parece ser un buen precusrsor sobre la influencia de distintos jugadores en cualquier deporte.

Hill(1974) [8] retoma esta idea final y cree que es obvio simplemente de ver un partido de futbol que: "ambas habilidad y suerte están involucradas" y tienen incidencia en el resultado y que, aunque: "es bien sabido que al pronosticar el resultado de un solo partido hay mucha suerte involucrada", en el largo plazo la habilidad del equipo dominará los resultados. Para probar su hipótesis, Hill utliza los pronósticos de la revista Goal para la temporada 1971-1972 sobre las cuatro divisiones del futbol inglés y las dos temporadas del futbol escocés y los compara con la tabla final. Encuentra que el coeficiente de correlación de Kendall es positivo en los seis casos. Más aún, para ambas primeras divisiones - que es donde, naturalmente, hay más información para hacer pronósticos - ambas correlaciones son mayores a 0.5.

2.2. Modelación individual de equipos

Un año después, Thompson (1975) [9] es la primera investigación en modelación individual de equipos. Thompson plantea el problema de obtener "comparaciones justas y precisas entre competidores, cuando muchos de los pares posibles no se han enfrentado y la calidad promedio de los oponentes enfrentados varía mucho" dentro de

la temporada 1973 de la NFL, a través de una máxima verosimilitud. Es decir, ¿cómo podemos determinar fuerza relativa de equipos, si muchos de estos no se han enfrentado ente si?

Esencialmente, la meta de Thompson es modificar el record de ganados y perdidos en una temporada, ajustándolo por la dificultad de los oponentes enfrentados, para maximizar la verosimilitud sobre los resultados observados en la temporada. Esto es algo que se hace muy comunmente en analítica del deporte moderna.
Thompson prueba con distintas fórmulas en la verosimilitud para obtener la probabilidad de que un equipo en particular gane un partido, dadas las posiciones en el ranking de los dos equipos enfrentándose. Encuentra que una función cuadrática de la diferencia de rankings entre los dos equipos devuelve la mejor predicción. Thompson comenta que este método tiene un problema, dado que las diferencias no son uniformes (por ejemplo,la diferencia entre el 2do y 3er equipo en el ordenamiento, puede ser mucho mayor a la diferencia del 3ro al 4to equipo). Esto ocurre con frecuencia en el futbol soccer moderno, donde los mejores equipos suelen tener una distancia considerable a los siguientes en el ranking. No obstante, Thompson no encuentra un cambio significativo de la verosimilitud al probar un método que permite empates en el ranking.
Por último, al comprar sus rankings con los records reales, encuentra que: "las diferencias que resultan por la discrepancia en la dificultad de oponentes no se emparejan tras una temporada". Es decir, que en una muestra pequeña - la temprada de la NFL tiene 16 juegos por equipo - la suerte es un factor importante en determinar los equipos que van a postemporada, puesto que hay errores de 2 partidos (12.5 %) consistentemente.

Modelos en futbol soccer

Maher (1982) [1] es el primer trabajo sobre futbol soccer en modelar la fuerza individual de cada equipo y su influencia en los partidos. Su investigación desata una importante línea de investigación de modelos Poisson para la predicción de partidos y es la mayor influiencia en esta tesis.

Maher argumenta que hay "buenas razones para pensar que el

número de goles anotado por un equipo en un partido es una variable Poisson"pues las posesiones - los ataques, las veces que tiene la pelota un equipo - son muchas y la probabilidad de anotar en una posesión cualquiera es muy pequeña. Maher nota que asumir a todos los equipos iguales daría efectivamente una Binomial Negativa, como la encontrada por Moroney (1953) [4] y Reep, Pollard, Benjamin (1971) [6].

De esta forma, propone un modelo Poisson para los marcadores. Sea $(X_{i,j}, Y_{i,j})$ el marcador observado cuando se enfrentan el equipo local i, contra el equipo visitante j:

$$X_{i,j} \sim \text{Poisson}(\alpha_i \beta_j) \quad Y_{i,j} \sim \text{Poisson}(\alpha_j \beta_i k^2)$$

donde $X_{i,j}$ y $Y_{i,j}$ son independientes. Podemos pensar a α_i como la habilidad de ataque del equipo i y a β_j como la habilidad de la defensa del equipo j. Similarmente, α_j es la habilidad de ataque del equipo visitante j y β_i la habilidad para defender del equipo i. k es un parámetro que modela la ventaja de localía, definido a detalle más adelante.

Maher prueba varios modelos, principalmente, uno donde en vez de dos parámetros por equipo (α, β), cada equipo tiene cuatro parámetros $(\alpha, \beta, \delta, \gamma)$: las habilidades de el ataque (α, δ) y la defensa (β, γ) son divididas en partidos como locales y visitantes, como si fueran dos cosas distintas. Sin embargo, encuentra que el primero modelo con dos parámetros por equipo es más adecuado y que no es necesario distinguir entre habilidades como local y visitante.

Además, experimenta dándole un parámetro de ventaja de localía a cada equipo y encuentra que, aunque la influencia de jugar como local es muy importante para el modelo, se puede utilizar un sólo parámetro global para todos los equipos. Es decir, que la ventaja no varía significativamente de equipo a equipo. Maher elige utilizar el parámetro de localía para reducir la fuerza de ataque del visitante en vez de aumentar la del local; aunque ambas parametrizaciones funcionan.

De esta manera, Maher puede modelar los goles anotados por

cada equipo utilizando variables aleatorias Poisson independientes, con medias equivalentes a la fuerza relativa entre los equipos como en la ecuación (2.2). Así, el modelo propuesto puede obtener la probabilidad de cualquier marcador posible, $(X_{i,j}, Y_{i,j}) = (x, y)$, dados los parámetros de los equipos enfrentándose y el parámetro de localía. Puesto que $X_{i,j}, Y_{i,j}$ son Poisson independientes, los parámetros máximo verosímiles están dados por:

$$\hat{\alpha}_i = \frac{\sum\limits_{j \neq i} x_{i,j} + y_{j,i}}{(1+\hat{k}^2)\sum\limits_{j \neq i} \beta_j} \qquad \hat{\beta}_j = \frac{\sum\limits_{i \neq j} x_{i,j} + y_{j,i}}{(1+\hat{k}^2)\sum\limits_{i \neq j} \alpha_i} \qquad \forall i, j \tag{2.1}$$

El parámetro de localía está dado por:

$$\hat{k}^2 = \frac{\sum\limits_i \sum\limits_{j \neq i} y_{i,j}}{\sum\limits_i \sum\limits_{j \neq i} x_{i,j}} \tag{2.2}$$

De la forma de los estimadores, se sigue que:

$$\sum_i \sum_{j \neq i} \hat{\alpha}_i \hat{\beta}_j = \sum_i \sum_{j \neq i} x_{i,j} \qquad \sum_i \sum_{j \neq i} \hat{k}^2 \hat{\alpha}_j \hat{\beta}_i = \sum_i \sum_{j \neq i} y_{i,j} \tag{2.3}$$

Eso significa que la suma de las media del modelo Poisson ajustado es igual al número de goles anotados observados.

Maher se da cuenta que su modelo subestima la probabilidad de que los equipos anoten 1 o 2 goles, mientras que sobrestima la probabilidad de anotar 0 o 4+ goles. Por ello, cree necesario usar una distribución más "angosta" (refiriéndose al cociente de varianza y media). Para analizar los marcadores más a detalle, define una variable adicional $Z_{i,j} = X_{i,j} - Y_{i,j}$ para explorar la diferencia entre los goles locales y visitantes. Cuando $Z_{i,j} > 0$, ha gando el equipo local pues $X_{i,j} > Y_{i,j}$; si $Z_{i,j} < 0$, ha gando el equipo visitante dado que $X_{i,j} < Y_{i,j}$.
Maher nota que el modelo subestima sistemáticamente $Z_{i,j} = 0$ - la probabilidad de empate - al modelar partidos individualmente en

los datos disponibles.

Tras una exploración, concluye que hay una pequeña correlación entre los dos marcadores $X_{i,j}$,$Y_{i,j}$ y sugiere un modelo Poisson bivariado en vez de modelar cada uno de los marcadores como variables Poisson independientes.

Sean X_1, X_2, X_3 variables aleatorias Poisson con parámetro λ_1, λ_2, λ_3 respectivamente. Definimos $X = X_1 + X_3$, $Y = X_2 + X_3$. Entonces, decimos que el vector aleatorio (X, Y) tiene distribución Poisson bivariada, $\mathrm{BP}(\lambda_1, \lambda_2, \lambda_3)$, con función de densidad:

$$\mathbb{P}[X = x, Y = y] = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1}{k_1!} \frac{\lambda_2}{k_2!} \sum_{k=0}^{min(k_1,k_2)} \binom{k_1}{k} \binom{k_2}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k \tag{2.4}$$

El modelo específico para Maher mantiene las marginales Poisson. Los goles locales $X_{i,j}$ y visitantes $Y_{i,j}$ para un partido donde se enfrentan el equipo local i y el equipo visitante j están definidos como:

$$X_{i,j} = X_1 + X_3 \qquad Y_{i,j} = X_2 + X_3 \tag{2.5}$$

$$X_1 \sim \mathrm{Poisson}(\lambda_1 - \lambda_3) \quad X_2 \sim \mathrm{Poisson}(\lambda_2 - \lambda_3) \quad X_3 \sim \mathrm{Poisson}(\lambda_3) \tag{2.6}$$

$$\lambda_1 = \alpha_i\, \beta_j \quad \lambda_2 = \alpha_j\, \beta_i\, \hat{k}^2 \quad \lambda_3 = Cov(X_{i,j}, Y_{i,j}) = \rho \sqrt{\lambda_1\, \lambda_2} \tag{2.7}$$

Maher prueba varios valores de ρ - la correlación entre los marcadores local y visitante - y concluye que el más apropiado se encuentra alrededor de $\rho = 0.2$

La gran ventaja del modelo de Maher es la simplicidad de los estimadores y su obtención, en ambos la versión Poisson doble y Poisson bivariada. Sin embargo, es un modelo estático, y en ello lleva su mayor desventaja. La restricción planteada por la ecuación 2.3 significa además que los parámetros no se amoldan a un partido real con información parcial: aunque los mismos reflejan la diferencia relativa entre un equipo y otro en los datos disponibles, los parámetros serán por naturaleza pequeños, puesto que el número total de goles en las primeras semanas es muy pequeño también. Esto quiere

decir que los parámetros no reflejarán el número promedio de goles anotados por partido hasta el final de la temporada, haciendo que no se pueda implementar este mismo modelo dinámicamente. Más adelante discutimos los detalles con detenimiento.

Dixon, Coles(1997) [10], al revisar la literatura, encuentran que "en el largo plazo, no es difícil predecir satisfactoriamente qué equipos tienen alta probabilidad de ser exitosos, pero el desarrollo de modelos que tengan suficiente alcance para explotar esta predictibildad del largo plazo en partidos individuales es sustancialmente más difícil". Esto coincide con lo que encuentran los primeros intentos de modelación. Por ello, trabajan en mejorar el modelo de Maher (1982) [1], proponiendo que en vez de que la distribución conjunta del marcador en un partido sea una Poisson bivariada, esté dada por:

$$\mathbb{P}[X_{i,j} = x, Y_{i,j} = y] \quad = \quad \tau_{\lambda,\mu}(x,y)\frac{\lambda^x e^{-\lambda}}{x!}\frac{\mu^y e^{-\mu}}{y!} \tag{2.8}$$

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1-\lambda\mu\rho & \text{si } x=y=0 \\ 1+\lambda\rho & \text{si } x=0, y=1 \\ 1+\mu\rho & \text{si } x=1, y=0 \\ 1-\rho & \text{si } x=1, y=1 \\ 1 & \text{en otro caso} \end{cases} \tag{2.9}$$

$$X_{i,j} \sim Poisson(\lambda = \alpha_i\,\beta_j\,\gamma) \quad Y_{i,j} \sim Poisson(\mu = \alpha_j\,\beta_i) \tag{2.10}$$

$$\text{máx}(\frac{-1}{\lambda}, \frac{-1}{\mu}) \;\leq\; \rho \;\leq\; \text{mín}(\frac{1}{\lambda\,\mu}, 1) \tag{2.11}$$

donde $X_{i,j}, Y_{i,j}$ son independientes. La intepretación de α, β es la misma que en el modelo de Maher. $\gamma > 0$ es el parámetro de localía que, a diferencia de Maher, ahora aumenta la capacidad de ataque del equipo local.

ρ actúa como un parámetro de dependencia: si fuera cero, habría independencia total y, por lo tanto, un modelo Poisson doble - de dos Poisson independientes, una por marcador. Las distribuciones

marginales siguen siendo Poisson con parámetros λ y μ respectivamente. La modificación al modelo de Maher sólo ocurre en los marcadores 0-0, 1-0, 1-1 y 0-1, dada por τ, puesto que en éstos los autores encuentran una diferencia significativa - respecto a errores estándar Bootstrap - entre las distribuciones empíricas marginales y la distribución empírica conjunta.

Por ejemplo, en los datos utilizados por los autores, un equipo local anota 0 goles 22.1 % de los partidos y un equipo visitante anota 0 goles 33.4 %. Entonces esperaríamos ver un empate 0-0 bajo total independencia en 7.38 % de los partidos. Sin embargo, en los datos se presenta un empate 0-0 8.2 %, que dada la frecuencia de ocurrencia de un empate sin goles - los marcadores con pocos goles son los más comunes - es una diferencia significativa.

Dado el modelo antes propuesto, en una liga con N equipos, se tienen que estimar los parámetros de ataque $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, los parámteros de defensa $\{\beta_1, \beta_2, \ldots, \beta_n\}$, el parámetro de dependencia ρ y el parámetro de localía γ. Los autores establecen estimadores para las cuatro divisiones de Inglaterra simultáneamente, por lo que $N = 92$. De esta forma, los autores pueden modelar partidos de copa, donde juegan equipos de distintas divisiones y, además, tienen información sobre los equipos cuando estos cambian de división al final de una temporada por ascenso y descenso. Esto, por supuesto, tiene un costo computacional importante; ambos en los datos utilizados y en los parámetros a estimar, que son 185.

Para los partidos $k = 1, 2, \ldots, N$ y sus correspondientes marcadores observados (x_k, y_k), la función de verosimilitud (hasta una constante de proporcionalidad) está dada por:

$$L(\alpha_i, \beta_i, \rho, \gamma; i = 1, 2, \ldots, n) = \prod_{k=1}^{N} \tau_{\lambda_k, \mu_k}(x_k, y_k)\, e^{-\lambda_k}\, \lambda_k^{x_k}\, e^{-\mu_k}\, \mu_k^{y_k} \tag{2.12}$$

donde

$$\lambda_k = \alpha_{i(k)}\beta_{j(k)}\gamma \qquad \mu_k = \alpha_{j(k)}\beta_{i(k)} \tag{2.13}$$

$i(k), j(k)$ denotan, respectivamente, los índices del equipo local y visitante jugando en el k-ésimo partido. El modelo está restringido

a maximizar numéricamente la ecuación (2.12) para poder obtener los parámetros.
Incluír todas las divisiones en la verosimilitud tiene una ventaja fundamental: los parámetros mismos deberían reflejar la calidad relativa de las distintas divisiones y así poder cuantificar fácilmente cuando dos equipos de distintas divisiones se enfrentan.

Además, los autores proponen que el modelo debe ser dinámico y no estático; es decir, que los parámetros de los equipos varíen a través del tiempo, intentando así arreglar una de las mayores fallas en el modelo de Maher. Esto pues las actuaciones de cada equipo fluctúan en el tiempo, por múltiples razones: táctica, lesiones, nuevos jugadores contratados, cambio de técnico, etcétera. Los autores creen que el nivel presente de un equipo se asemeja mucho más al mostrado en los últimos partidos que en partidos anteriores. Así, proponen que la importancia de cada juego sea determinada por una función de pesos ϕ. De esta manera, modifican la verosimilitud de la ecuación (2.12) y proponen una pseudo-verosimilitud para cada punto en el tiempo t, dada por:

$$L(\alpha_i, \beta_i, \rho, \gamma; i = 1, 2, \ldots, n) = \prod_{k \in A_t} \{\tau_{\lambda_k, \mu_k}(x_k, y_k)\, e^{-\lambda_k}\, \lambda_k^{x_k}\, e^{-\mu_k}\, \mu_k^{y_k}\}^{\phi(t-t_k)} \tag{2.14}$$

donde t_k es el momento en el que se juega el k-'esimo partido, $A_t = \{k | t_k < t\}$ el el conjunto de partidos que se han jugado hasta el momento t; λ_k, μ_k son como en la ecuación (2.13) y ϕ es una función no creciente que depende del tiempo. Maximizar la ecuación (2.14) al tiempo t devuelve estimaciones de los parámetros que sólo utilizan información antes de los partidos que se juegan en el tiempo t.

Dixon y Coles proponen usar una función con decaimiento exponencial para la función de pesos, es decir, $\phi = e^{-\xi t}$; donde un modelo estático estaría determinado por $\xi = 0$.

Al maximizar la probabilidad de predecir el resultado de los partidos en sus datos, encuentran que el mejor valor está dado por $\xi = 0.0065$.

Utilizan 60 de las 174 “medias semanas” - usualmente hay dos

conjuntos de partidos por semana en Inglaterra, uno a media semana y otro el fin de semana, por lo que los autores dividen sus tiempos para coincidir con el calendario - en sus datos para calibrar el modelo. Tras la calibración, el modelo es probado contra líneas de apuesta, con una estrategia de apuesta fija; es decir, apostar una unidad a cualquier juego donde tengan una ventaja porcentual sobre un umbral predeterminado.
Sea $\hat{p}$ la probabilidad estimada por el modelo y b_k la probabilidad implícita en el momio de apuesta disponible. Los autores proponen que se apueste una unidad si: $\frac{\hat{p}}{b_k} > r$ para algún valor de $r > 1$.

La elección de r es clave para determinar la cantidad de apuestas que habrá en una temporada: aumentar r quiere decir que hay menos apuestas, pero con una ventaja mayor en cada una de ellas; disminuír r quiere decir que habrá más apuestas, pero con una ventaja menor en cada una. Los autores encuentran que el retorno esperado es positivo para cualquier estrategia donde $r > 1.1$.

Dixon, Pope (2004) [11] hace una exploración más exhaustiva del modelo Dixon, Coles (1997) [10] en el mercado y su comportamiento. De interés, los autores encuentran que los momios de empate ofrecidos por las casas de apuesta varían significativamente menos que los de victoria local o visitante. Además, encuentran que si se fuera a seguir una estrategia ciega, apostar a todos los empates es la que da mejores resultados (-4.9 %, que es menor a la comisión de la casa que tienen su líneas de aproximadamente -11 %). Por otro lado, encuentran que el modelo se puede utilizar y explotar en mercados de marcadores exactos, aunque la varianza es mucho más grande a la de apostar al resultado del partido.

Karlis, Ntzoufras (2003 y 2005) [3],[12] presentan distintas alternativas para datos bivariados que se parecen mucho a una distribución Poisson doble pero tienen algún problema en la especificación. Primero trabajan con una de las parametrizaciones de la distribución Poisson bivariada, cuyo caso especial $\lambda_3 = Cov(X, Y) = 0$ es una Poisson doble. Sin embargo, el modelo Poisson bivariado sólo puede modelar datos con covarianza positiva. Además, como ambas marginales son Poisson, no hay manera de modelar una dispersión distinta a la dada: un modelo Poisson tiene media y varianza iguales.

Por ello, los autores proponen un modelo inflado por la diagonal. En el caso univariado, los modelos inflados en cero son muy populares en la práctica, particularmente en medicina. Sin embargo, los casos en la literatura bivariada son escasos. Por inflado por la diagonal, los autores se refieren a las probabilidades donde $X = Y$. En el caso del futbol, estos son los empates. La modificación propuesta al modelo poisson bivariado tiene la siguiente forma:

$$f_{IBP}(x,y) = \begin{cases} (1-p) f_{BP}(x,y|\lambda_1,\lambda_2,\lambda_3) & \text{si } x \neq y \\ (1-p) f_{BP}(x,y|\lambda_1,\lambda_2,\lambda_3) + p f_d(X|\theta) & \text{si } x = y \end{cases}$$

dónde f_{BP} es la función de probabilidad de una distribución Poisson Bivariada, $f_d(X|\theta)$ es la función de probabilidad de alguna distribución discreta (otra Poisson, una geométrica, una Bernoulli, o una distribución discreta general con todas las probabilidades especificadas). $D(X;\theta)$ es dicha distribución discreta con dominio en $\mathbb{Z}^+$ y vector de parámetros θ. Estos modelos pierden la propiedad de tener marginales Poisson; éstas serán ahora una mezcla con al menos un componente Poisson. Aunque será computacionalmente mucho más complicado lidiar con ellas, permiten cambiar la dispersión y no estar completamente atados a una Poisson.

Los autores desarrollan un paquete en el lenguaje de programación R para implementar dichos modelos mediante un Algoritmo esperanza-maximización. Entre los ejemplos desarrollados en la publicación, utilizan los datos de la temproada 1991-1992 de la Serie A italiana - que habían modelado previamente en Karlis, Ntzoufras (2003) - para mostrar el uso del paquete en R y sus aplicaciones. Los autores habían notado en sus dos publicaciones previas que parece haber un exceso de empates y una pequeña sobredispersión respecto al modelo Poisson doble y reproducen los resultados con el paquete que han creado en R. Encuentran que el mejor modelo es un Poisson bivariado con $\lambda_3 = Cov(X,Y) = 0.21$ y la distribución para inflar la diagonal una degenerada con toda su masa en 1 y parámetro de mezcla $p = 0.09$, como en la ecuación (2.2). Esto pues 1-1 fue un marcador muy popular en Italia ese año. Es importante recordar aquí que entonces Italia tenía un sistema de puntos 2-1-0, es decir, las victorias valían solamente 2 puntos en vez de los 3 que otorga

el sistema moderno. Esto podría explicar por qué la alta correlación y empates, dado que los equipos no tenían un incentivo tan grande para ir por la victoria. Para comparar, en los 306 partidos de la temporada 1991-1992 en Italia, 111 terminaron en empate o 36.27 % de los partidos. El equipo que más empates tuvo, Internazionale Milano, empató 17 juegos o 50 % de sus partidos. En los 380 partidos de la temporada 2016-2017 en Italia, solo 80 terminaron en empate o 21.05 %. El equipo que más empates tuvo, Torino FC, empató 14 juegos o 36.84 % de sus partidos. Una investigación exhaustiva del cambio al sistema de competencia y los incentivos para ir por la victoria se puede encontrar en Sumpter (2016) [13]

Koopman, Lit (2015) [14] creen que hay un hueco en la literatura para modelos que usen la Poisson bivariada y además sean estocásticos. Por ello, proponen que la distribución de los goles en un partido sea Poisson bivariada y permiten que los parámetros varíen ligeramente con el tiempo al hacerlos autoregresivos de primer orden con innovaciones normales.

Es decir, sea $(X, Y) = (X_{i,t}, Y_{j,t})$ un partido jugado entre el equipo local i y el visitante j a tiempo t, donde $t = 1, \ldots, n$ es el número de semanas disponibles en los datos, $i, j = 1, \ldots, J$ es el número de equipos en la liga, $i \neq j$. Entonces:

$$(X, Y) \sim BP(\alpha_x, \alpha_y, \gamma) \quad \mathbb{E}(X) = \lambda_x + \gamma \quad \mathbb{E}(Y) = \lambda_y + \gamma \tag{2.15}$$

$$Cov(X, Y) = \gamma \qquad \rho = \frac{\gamma}{\sqrt{(\lambda_x + \gamma)(\lambda_y + \gamma)}} \tag{2.16}$$

$$\lambda_{x;i,j,t} = exp(\alpha_{i,t} - \beta_{j,t} + \delta) \qquad \lambda_{y;i,j,t} = exp(\alpha_{j,t} - \beta_{i,t}) \tag{2.17}$$

donde $\alpha_{i,t}$ es la fuerza del ataque para el equipo i en la semana t; $\beta_{i,t}$ es la fuerza de la defensa para el equipo i en la semana t. δ denota la ventaja de localía, la cual es igual para todos los equipos. Los autores asumen que la dependencia entre los dos marcadores, denotada γ, es igual para todos los partidos jugados. $\alpha_{i,t}, \beta_{i,t}$ se asumen independientes y cambiantes en el tiempo bajo un proceso autoregresivo de orden uno dado por:

$$\alpha_{i,t} = \mu_{\alpha,i} + \phi_{\alpha,i}\,\alpha_{i,t-1} + \eta_{\alpha,i,t} \qquad \beta_{i,t} = \mu_{\beta,i} + \phi_{\beta,i}\,\beta_{i,t-1} + \eta_{\beta,i,t} \quad (2.18)$$

donde $\mu_{\alpha,i}, \mu_{\beta,i}$ son constantes desconocidas, $\phi_{\alpha,i}, \phi_{\beta,i}$ son coeficientes autoregresivos y $\eta_{\alpha,i,t}, \eta_{\beta,i,t}$ son innovaciones distribuidas como una normal, las cuales son independientes entre ellas para todo equipo y a todo tiempo. Es decir:

$$\eta_{k,i,t} \sim NID(0, \sigma^2_{k,i}) \qquad k = \alpha, \beta \quad (2.19)$$

Los autores asumen que los procesos son independientes y estacionarios para ayudar a la estimación de los parámetros, lo cual requiere que $|\phi_{k,i}| < 1$ donde $k = \alpha, \beta$ y $i = 1, \ldots, j$.

Para estimar los parámetros necesarios, utilizan máxima verosimilitud, pero necesitan estimaciones numéricas complicadas para poder obtenerlos, por lo que optan por un enfoque bayesiano y estimaciones por Monte Carlo.

Utilizan datos de nueve temporadas, desde la temporada 2003/2004 hasta la temporada 2011/2012. Dado que tres equipos cambian (ascienden y descienden) cada temporada, los autores tienen en total 36 equipos en los nueve años de datos que necesitan estimar simultáneamente; es decir, 72 estimadores (α's y β's) para todo tiempo, con muchas de las observaciones como faltantes - no habrá datos para los equipos que no estén jugando en la primera división en esa temporada en específico. Utilizan las primeras siete temporadas para la estimación de los parámetros y las últimas dos para probar el modelo bajo una ventana fija; es decir, cuando agregan una semana de información, quitan la primera semana de siete años antes.

Dado el problema de dimensionalidad, los autores deciden que los coeficientes autoregresivos y las varianzas de las innovaciones sean las mismas para todos los equipos, por lo que sólo tienen que estimar $\phi_\alpha, \phi_\beta, \sigma_\alpha, \sigma_\beta$, además de los estimadores originales del Poisson bivariado (un par α, β por equipo, la ventaja de local para todos δ y el coeficiente de correlacion en la poisson bivariada γ).

Ambos coeficientes para el proceso autoregresivo son muy cercanos a uno (phi_α =0.9985, ϕ_β = .9992), lo cual sugiere una alta peristencia en las habilidades de los equipos a través del tiempo. De

interés, los autores encuentran que $\gamma = .0966$, es decir, mucho menor que la encontrada por Maher (1982) [1] y Karlis, Ntzoufras (2003) [3], quizá porque todos los datos utilizados son bajo el sistema de puntuación 3-1-0. Por último, la estimación máximo verosimil del parámetro de localía es $\delta = .3641$. Dado que ésta entra al modelo como e^{δ}, la estimación del efecto es $e^{.3641} = 1.44$. Esto es parecido en valor al parámetro de localía utlizado en Maher (1982), aunque éste utiliza el inverso multiplicativo para reducir el ataque del visitante.

Los autores comparan seis modelos distintos:
1.- El propuesto por ellos, que ya ha sido descrito.
2.- El modelo 1, pero con $\gamma = 0$. Es decir, un Poisson doble (que son dos Poisson independientes).
3.- El modelo 1, pero con parámetro de dependencia γ variable entre los equipos y el momento del partido, en vez de ser constante e invariante en el tiempo ($\gamma_{i,j,t}$).
4.- El modelo 1, pero con un único parámetro por equipo, que depende del tiempo. Específicamente, $\lambda_{i,t} = exp(\theta_{i,t})$ con las mismas especificaciones autoregresivas de la ecuación (2.18), donde $\alpha_{i,t}$ es remplazado por $\theta_{i,t}$ y $\beta_{i,t}$ desaparece.
5.- El modelo 1, pero donde $\alpha_{i,t}, \beta_{i,t}$ ahora son constantes en el tiempo.
6.- El modelo 4, pero constante en el tiempo.

Encuentran que los últimos tres modelos son rechazados fácilmente, es decir, que la fuerza de ataque y defensa deben ser cuantificadas por separado y, además, que no son constantes en el tiempo.

Aunque el modelo propuesto por los autores tiene el menor valor en la función de pérdidas, los modelos con $\gamma = 0$ y $\gamma_{i,j,t}$ están muy cercanos. Parece entonces, sugieren los autores, que el parámetro de dependencia γ no tiene un impacto grande al predecir valores fuera de la muestra, a pesar de que tiene una fuerte significancia dentro de la muestra: en la especificación dentro de la muestra, todos los modelos son rechazados y sólo el propuesto por los autores es aceptado. De hecho, bajo la prueba estadística que realizan los autores para la predicción fuera de la muestra, no pueden rechazar la hipótesis nula de que los modelos 2 y 3 son tan precisos como el modelo 1.

Por último, los autores prueban su modelo contra los momios promedio dados por Football Data. La estrategia utlizada es apostar una unidad para cualquier apuesta con valor esperado EV(A) $> \tau$, para $\tau > 0$, sin importar el tamaño de la ventaja. Además, consideran apostar en todos los eventos que tienen poca probabilidad (que ellos definen como un momio > 7, es decir, de probabilidad de ocurrencia menor a 15 %) y deciden reducir el tamaño de sus apuestas a 0.3 unidades en estos casos. Las razones para éste segundo criterio no son explicadas. Prueban distintos valores de τ: para $\tau = 0$ el modelo tiene una apuesta en la mayoría de los partidos. Para $0 < \tau < .12$, el rendimiento promedio es alrededor de cero, que los autores atribuyen a incertidumbre en los parámetros. Los rendimientos positivos comienzan alrededor de $\tau > .12$. Sin embargo, al aumentar τ disminuye naturalmente la cantidad de apuestas que el modelo puede realizar. Los autores encuentran que el modelo hace menos de media apuesta por semana en promedio cuando $\tau > 0.45$, lo cual hace crecer mucho la varianza de los rendimientos. La figura 6.2 presenta el rendimiento promedio con intervalo de confianza y el número de apuestas que realiza el modelo de los autores, en función de τ

2.3. Otras investigaciones relevantes

Lee (1997) [15] utiliza el método planteado por Maher (1982) para explorar la temporada completa 1995-96 de English Premier League y simularla muchas veces, buscando entender si el mejor equipo ha ganado y poder generar una tabla final esperada, simulando a partir de los parámetros del modelo. Este método es común en la analítica actual para encontrar qué equipos han tenido suerte en una temporada, con respecto a las habilidades subyacentes del equipo. Además, su uso es común en otros deportes, como en béisbol donde, por ejemplo, Baseball Prospectus utiliza tres fórmulas distintas para ajustar la suerte de algunos equipos.

Goddard (2005) compara los dos enfoques grandes de modelación: centrarse en el proceso de goles para ambos equipos o intentar modelar directamente las probabilidades de victoria local, visitante o empate. Goddard piensa que, como un modelo que lidia con

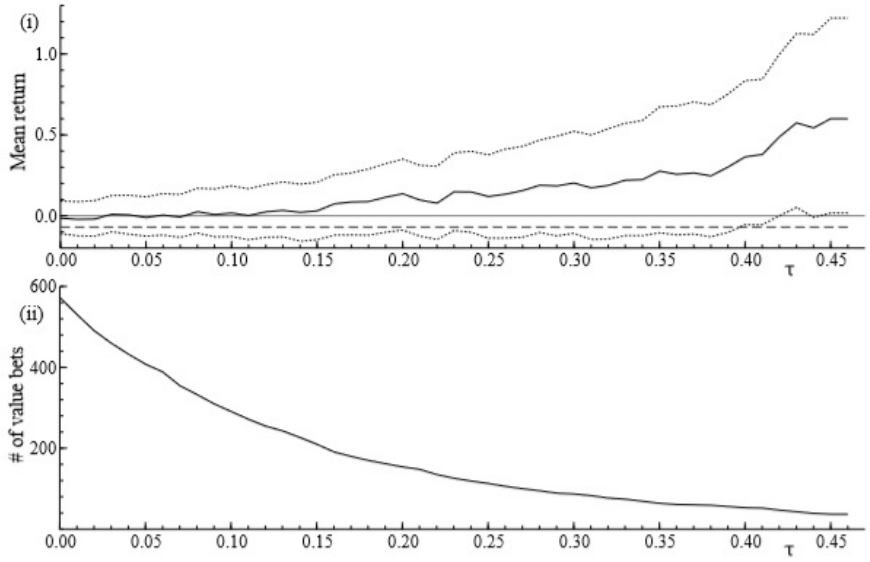

Figura 2.1: Rendimiento promedio, con intervalo de confianza y número de apuestas en función de τ que realiza el modelo específicado por Koopman, Lit (2012)

los reultados está intrínsecamente anidado en uno de modelación de goles - si se calculan las probabilidades para todos los marcadores, sólo basta sumar las de interés - el segundo debe tener mucho más ruido, pues en uno de resultados el número de goles es incidental; lo único que importa es el resultado. Y encuentra que el mejor modelo es un híbrido de ambos, pero que la diferencia en resultados es muy pequeña. Esto apoya la noción de que los dos enfoques son válidos.

Deschamps, Gergaud (2007) [16] hacen una investigación de mercado buscando si hay alguna estrategia totalmente basada en los momios que pudiera generar una inversión positiva, dado que la literatura previa ha encontrado resultados contradictorios: algunos inverstigadores - como Dixon, Pope (2004) [11] - encontraron que apostar a equipos de momios muy bajos (i.e. a eventos de probabilidad alta) generaba un rendimiento menor a eventos con momios muy altos; otros autores - como en Cain, Law, Peel (2000) [17] - encontran lo opuesto, que los eventos de poca probabilidad (i.e. momios muy altos) están infraestimados por los momios. Deschamps

y Gergaud utilizan casi 8400 partidos y encuentran resultados mixtos. Al apostar por el local o el visitante, encuentran que es mucho mejor apostar ciégamente a favoritos y que ésto genera mejores rendimientos que apostar a momios altos. Para las apuestas de empate, sin embargo, encuentran lo contario: apostar a momios altos devuelve un mejor rendimiento que a momios bajos. Sin embargo, a pesar del sesgo, muestran que no existe ninguna estrategia basada sólamente en las líneas que tenga un rendimiento positivo. Por último, encuentran otro resultado que es consistente con la literatura previa: si se apostara ciégamente a cualquiera de los tres resultados todo el tiempo, el empate devuelve los mejores rendimientos (-7 %).

Capítulo 3

Apuestas deportivas

3.1. Introducción

En esta sección, discutimos qué son y cómo funcionan las apuestas deportivas. Primero, cómo funciona un momio de apuesta y el cálculo de un momio justo. Posteriormente, discutimos las apuestas en el contexto deportivo: cómo calcular probabilidades implícitas en la línea, cómo gana dinero la apuesta, la forma de determinar apuestar favorables y las apuestas en futbol soccer. Finalmente, discutimos el comportamiento del mercado de líneas de apuestas: si es eficiente, cómo se crean las líneas y el tamaño del mercado global.

Como al empezar un curso de Probabilidad, la mejor manera de introducir el tema es con algunos ejemplos sencillos usando monedas y dados.

Ejemplo 5. Imagine que vamos a lanzar una moneda justa y apostar en el resultado, pero usted va a pagar $15 pesos por entrar al juego y yo pagaré $10. El que elija el resultado correctamente se lleva los $25 que suman las entradas, o el pozo. ¿Estaría usted dispuesto a entrar a dicho juego?

Probablemente no, y la razón es que usted entiende que la moneda tiene 50 % de probabilidad de caer cara y 50 % de probabilidad de caer cruz y, por lo tanto, es injsuto que usted ponga menos dinero al pozo que yo. Como el juego tiene un resultado 50 % − 50 %, entonces parece justo que la proporción de dinero en el pozo se divida 50 % − 50 % también.

Ejemplo 6. Imagine ahora que en vez de una moneda, vamos a lanzar un dado justo. Usted elige un número del dado y, si ese número cae, se lleva el pozo; en caso contrario, yo gano el pozo. Le propongo que usted ponga $10 al pozo y yo ponga $10; una división justa, como en el caso de la moneda. ¿Aceptaría jugar?

La respuesta es no otra vez, la probabilidad de elegir el número correcto del dado es baja; intuitivamente, más baja que la de una moneda, por lo que parece injusto que la proporción de dinero apostada sea la misma para ambas partes. Es decir, para entrar a un juego donde la probabilidad de ganar es más baja, el premio necesita ser más alto.

En cambio, si le propongo que usted ponga $10 al pozo y yo ponga $100, ahora el juego se vuelve deseable para usted: aunque la probabilidad de adivinar correctamente sigue siendo menor a una moneda (exactamente, $\frac{1}{6}$), el premio correspondiente ha hecho que esta sea una apuesta favorable a usted de alguna manera.

3.1.1. Momios de apuesta

La idea detrás de un momio de apuesta es exactamente esa: cuando el juego tiene probabilidades desiguales, el pago correspondiente disminuye/aumenta para compensar la diferencia en probabilidad. Un momio justo es, como en el caso de la moneda, el que hace que el juego sea equilibrado.

Formalmente, un momio justo es aquel pago que, bajo la probabilidad del juego, hace que la ganancia esperada para los jugadores involucrados sea 0.

Sea p la probabilidad de ganar el juego y $q = 1 - p$ la probabilidad de perderlo. Definimos la entrada como lo que el jugador debe pagar para jugar; llamémoslo por ahora e. Esto es lo que el jugador perderá si el resultado no es favorable; lo apostado.
Llamemos r lo que el juego regresa si el jugador gana el juego; el rendimiento. Entonces, la posible ganancia del jugador está dada por $g = p - e$; es decir, lo que en caso de ganar, el juego regresa,

menos lo que ya era del jugador.
Usemos el ejemplo de la moneda. La probabilidad de ganar es $p = \frac{1}{2}$ y queremos calcular cual sería el pago justo para que el jugador accediera a jugar; es decir, el de esperanza cero. La esperanza del juego para el jugador está dada como lo que éste puede ganar, por la probabilidad de ganar; menos lo que este puede perder, por la probabilidad de perder.

Calculamos el momio justo con una ecuación de esperanza:

$$\mathbb{E} = p * g + q * (-e) = 0 \iff \frac{1}{2}\, g - (1 - \frac{1}{2})\, e = 0 \iff g = e$$

La ecuación confirma lo que nuestra intuición ya sabía: para que un juego con una moneda sea equilibrado, la entrada al juego debe ser igual a la posible ganancia.

De igual manera, podemos calcular la entrada justa apostando con un dado. La probabilidad de ganar eligiendo un lado es $p = \frac{1}{6}$. Entonces:

$$\mathbb{E} = \frac{1}{6} * g - \frac{5}{6} e = 0 \iff g = 5e$$

Esto quiere decir que el juego sería justo para ambas partes si su potencial ganancia fuera 5 veces mayor que la entrada. Es decir, si yo pusiera \$50 al pozo y usted pusiera \$10 al pozo y repitiéramos el mismo juego muchas veces, a largo plazo, los dos esperaríamos tener ganancia cero. La idea detrás de un momio para apuestas deportivas es muy similar. A continuación, exploramos esto a detalle.

3.2. Las apuestas en deportes

Las apuestas deportivas involucran intentar predecir el resultado de algún evento futuro. El evento podría ser cualquier cosa, desde quién gana un partido, hasta el número de tarjetas amarillas que habrá entre los minutos 60:00 y 75:00.
La apuesta más popular en futbol soccer es el resultado del partido, que tiene tres posibilidades: gana el equipo local, gana el equipo visitante o el juego termina en empate. Las casas de apuesta publican

sus líneas típicamente algunos días antes del partido - aunque para eventos grandes, como la Copa Mundial, con meses de anticipación - y reciben dinero hasta el momento en el que éste inicia. Las apuestas son realizadas contra las líneas fr apuesta publicadas o momios, por lo cual son conocidas como *fixed-odds betting*; el *fixed* se refiere a que la línea al momento de apostar es la línea contra la cual se paga la apuesta, a diferencia de otros sistemas, como el de apuestas mutuas (*parimutuel betting*) que se utiliza en las carreras de caballos.

La palabra predecir en el párrafo anterior tiene una connotación distinta al uso de todos los días: a menos de que el partido esté de alguna manera arreglado yo, como apostador, no puedo saber el resultado de un partido de futbol. Pero, como en el ejemplo del dado y la moneda, hay situaciones donde el juego es favorable a mí, dadas la probabilidades de ocurrencia en el partido. Ganar dinero en apuestas deportivas, se trata fundamentalmente de eso: encontrar eventos donde la probabilidad de que un evento ocurra sea mayor a la probabilidad implícita en la línea de apuesta para dicho evento, dándole una ventaja al jugador.

Las líneas de apuesta o momios representan la creencia de la casa sobre el evento en cuestión. Intrínsecamente, los pagos ofrecidos en cada apuesta reflejan la opinión de la casa sobre la probabilidad de que dicho evento ocurra. Si el pago es alto, la casa cree que el evento tiene poca probabilidad de ocurrir. Si el pago es bajo, la casa cree que el evento tiene mucha probabilidad de ocurrrir.

A continuación, un ejemplo de cómo funcionan las líneas de apuesta:

Ejemplo 7. En la figura 3.1 se observan las líneas de apuesta disponibles para el juego Tigres - América del 10/febrero/2018, publicadas por la casa en línea Pinnacle. Los momios están en forma decimal; de esto más en un segundo. Tigres es el local y favorito, con el momio más pequeño (1.877), el empate paga 3.58 y la victoria visitante 4.7. Esto quiere decir, por ejemplo, que si apostamos $100 al América y el resultado es ese, la casa nos pagaría $470 pesos. Es decir, nuestra ganancia sería $370; el rendimiento de la apuesta es simplemente lo apostado multiplicado por el momio y la ganancia se obtiene restando la apuesta original.

Figura 3.1: Líneas disponibles para Tigres - América en Pinnacle.

3.2.1. Apuestas favorables

La pregunta entonces es: ¿cómo determinar si una apuesta nos conviene?

Como todo evento aleatorio y con incertidumbre, nos gustaría poder cuantificar sus propiedades estadísticas. La más importante, como suele suceder en modelación, es el valor esperado. Por ello, es crucial determinar cuándo el valor esperado de una apuesta es positivo. Al punto en el que el valor esperado de una apuesta es cero se le llama punto de quiebre o *break-even point*. Éste es el momio justo que hemos determinado en el ejemplo del dado. Así, cualquier apuesta cuya probabilidad de ganar sea mayor al punto de quiebre implicado en la línea, tendrá un valor esperado positivo para el jugador. Usando nuestro ejemplo, queremos determinar si vale la pena apostar al empate en el partido Tigres - América. Esto puede hacerse con una ecuación de valor esperado sencilla usando la posible ganancia/pérdida. Supongamos que apuesto \$100 al empate. Esto quiere decir que si el juego termina en empate, la casa me pagará $\$100 * 3.58 = \$358$, con ganancia $\$358 - \$100 = \$258$. Si el juego no termina en empate, habré perdido \$100. Con ello, calculamos el punto de quiebre:

$$258X + (-100)(1 - X) = 0 \iff 358X = 100 \iff X = .2793$$

Entonces el punto de quiebre es 27.93 %. Si yo estimo que la probabilidad de empate para el partido es mayor a 27.93 %, entonces dicha apuesta tiene un valor esperado positivo; en caso contrario, negativo.

3.2.2. ¿Cómo gana dinero la casa de apuestas?

Si calculamos los puntos de quiebre restantes de la misma manera, estos son 53.28 % para Tigres y 21.28 % para el América. Al sumar los tres, obtenemos $53.28\,\% + 27.93\,\% + 21.28\,\% = 102.49\,\%$. La diferencia entre el 100 % que deberían sumar las probabilidades y lo que suman en verdad es conocido como el *vigorish*, *vig* o *take*, y es la comisión que la casa cobra para aceptar las apuestas; la forma en que gana dinero. En el ejemplo, teóricamente, la casa se quedaría con 2.49 % de todas las apuestas realizadas. Esto, claro, si lograra balancear las apuestas en los tres resultados. Más sobre cómo establece sus líneas una casa un poco después.

La cantidad de *vig* en un mercado cualquiera depende totalmente del tamaño y la eficiencia del mismo. Por ejemplo, un mercado grande como la English Premier League recibe muchas apuestas, además de información sobre todos los jugadores y lesiones, por lo que la casa puede ganar dinero con un vig bajo, atrayendo un volumen de apuestas considerable. Pero en una liga mucho menos conocida, como la K League (la primera división en Corea del Sur), la casa no puede atraer tantas apuestas, además de que la información sobre la misma será mucho más limitada, por lo que las líneas publicadas no son tan exactas; tienen mucho mayor margen de error. Por ello, la casa se protege poniendo un *vig* mayor en las líneas.

La mayoría de las casas de apuesta no buscan atraer apostadores profesionales, sino al público en general. Usualmente, se encuentran en casinos y tienen una sala muy atractiva donde se reciben las apuestas: la comida y las bebidas alcohólicas suelen ser baratas; hay muchos televisores para la mayoría de los partidos en vivo y carreras de caballos. Su negocio no radica en hacer líneas muy eficientes, sino en ser un servicio complementario del casino. Por lo mismo, el *vig* en las líneas es significativo y poco atractivo para los apostadores. Muchas de las casas por internet han adoptado este mismo modelo: la casa deportiva es simplemente un complemento para las demás partes de su negocio que conllevan menor riesgo y siempre son favorables hacia la casa: las máquinas tragamoneda, las ruletas, el poker entre usuarios (en el que la casa cobra una comisión por mano) y las loterías.

Como contexto, las ganancias netas para casinos en Las Vegas por deportes fueron 248.7 millones de dólares [18], mientras que para las máquinas tragamonedas fueron 3,100 millones de dólares y 1,200 millones para las mesas de Blackjack [19]. Sin embargo, los eventos grandes atraen una cantidad fuerte de dinero en apuestas y hacen que el resto de los negocios sean atractivos para el casino. Las casas de deportes en Nevada recibieron un total de 158.58 millones de dólares en apuestas en el reciente Super Bowl LII jugado entre Patriots - Eagles, aunque la ganancia neta fue solamente 1.17 millones o 0.7 % de las apuestas [20]. Esto pues la mayoría de las apuestas en el juego fueron ganadoras (Eagles +4.5 y *over* 48.5). Las casas han salido adelante en 26 de los 28 Super Bowls celebrados desde 1991, cuando la Comisión de Control de Juego de Nevada comenzó a llevar un registro de las apuestas.

La comisión de juegos de Nevada tiene números públicos detallados para toda la industria. En 2017 las casas de apuestas en Nevada ganaron un record $248.7 millones de dolares, de un total de $4,800 millones de dólares apostados. El futbol americano fue el deporte más apostado, con $1,700 millones de dólares entre la NFL y el colegial; seguido por el basketball, $1,400 millones y el béisbol, $1,100 millones. Al cierre de 2017, la casas habían generado ganancias en 53 meses consecutivos, una racha que empezó en julio de 2013 [**?**].

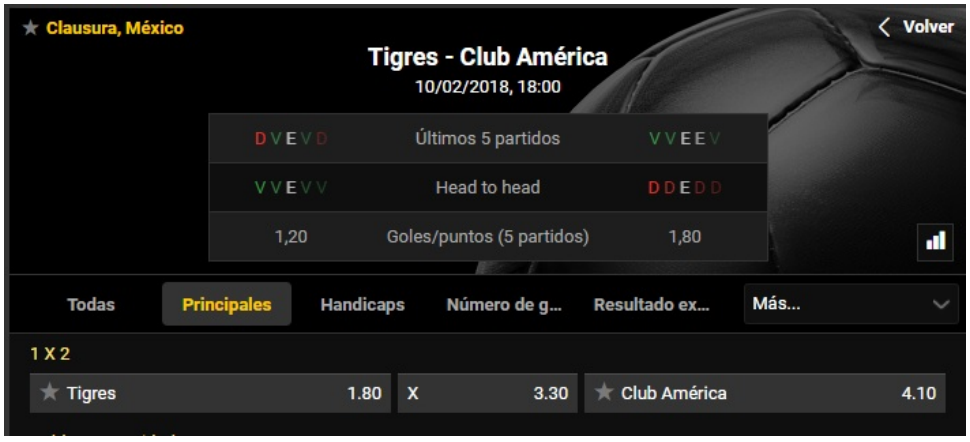

Figura 3.2: Líneas disponibles para Tigres - Ameríca en Bwin.

Estos porcentajes de *vig* tan altos hacen que la casa tenga un margen de error muy grande. En la Figura 3.2 se observa la línea publicada por la casa de apuesta Bwin para el mismo partido Tigres - América. Los puntos de quiebre son, respectivamente, $55.56\,\% + 30.30\,\% + 24.39\,\% = 110.25\,\%$. Los takes cercanos a $10\,\%$ son comunes para el futbol. Obviamente, el jugador quisiera que el número fuera tan cercano a 0 como le fuere posible y la manera más fácil de lograrlo es tener cuentas en muchas casas distintas. Esto le permite al jugador apostar conta la línea más favorable entre sus opciones y reducir la ventaja que tiene la casa.

En la Figura 3.3 se observa el partido del ejemplo en el sitio OddsPortal, que compila los momios de más de 60 casas por partido. En el renglón rotulado *Highest*, se encuentran los mejores momios para cada una de las apuestas. Estos máximos y sus respectivos puntos de quiebre son $1.93\ (51.81\,\%) + 3.60\ (27.78\,\%) + 4.65\ (21.51\,\%) = 101.10\,\%$. Es decir, si el jugador tuviera estas tres líneas disponibles, reduciría la ventaja de la casa a solamente $1.1\,\%$. Por supuesto, tener cuentas en 60 casas distintas es irreal, dada la cantidad de dinero necesaria. Sin embargo, es conocido en la industria cuales son las casas que, en general, ofrecen márgenes favorables al jugador, por lo cual cuatro o cinco cuentas es factible y muy importante para el jugador. En la Figura 3.3 se observa que de las casas disponibles en pantalla, 5Dimes, Pinnacle y AssianBet son las que tienen menor comisión. En el mercado hay pocas casas con comisiones pequeñas dispuestas a recibir apuestas grandes. Éstas son en las que el jugador debe centrar la mayoría de sus apuestas.

3.2.3. Notaciones alternativas para los momios

En el Reino Unido y las carreras de caballos, se utiliza una notación distinta para los momios, conocida como fraccional. Como su nombre lo indica, siempre son una fracción entre dos números enteros. Y son muy intuitivas, puesto que indician cuánto ganará el apostador relativo a lo que apostado. Por ejemplo, un momio de $\frac{5}{1}$ indica que el jugador ganaría \$500 al hacer una apuesta de \$100; un momio de $\frac{1}{5}$ indica que el jugador ganaría \$20 en caso de que apueste \$100. Además de la ganancia, el jugador recibe lo que apostado ori-

U.A.N.L.- Tigres - Club America

Tomorrow, 11 Feb 2018, 01:00

1X2 | AH | O/U | DNB | EH | DC | CS | More bets

Full Time | 1st Half | 2nd Half

Bookmakers	1	X	2	Payout
10Bet	1.87	3.40	4.15	93.5%
18bet	1.88	3.45	3.70	91.6%
1xBet	1.92	3.54	4.56	97.8%
5Dimes	1.93	3.49	4.60	97.8%
Asianodds	1.93	3.52	4.50	97.6%
bet-at-home	1.81	3.19	4.12	90.2%
bet365	1.80	3.60	4.33	94.0%
Bethard	1.87	3.40	4.15	93.5%
Betrally	1.83	3.35	4.10	91.8%
bwin	1.80	3.30	4.10	90.7%
Jetbull	1.82	3.40	4.00	91.4%
Marathonbet	1.90	3.55	4.60	97.5%
Pinnacle	1.93	3.52	4.50	97.6%
TonyBet	1.85	3.30	4.30	92.9%
Unibet	1.79	3.45	3.95	90.8%
Click to show 60 more bookmakers!				
= Average	1.86	3.43	4.24	94.0%
Highest	1.93	3.60	4.60	98.7%

Figura 3.3: Líneas disponibles para Tigres - Ameríca en múltiples casas, tomado de OddsPortal.

Sat 2/10	6266	Tigres	-0.5 -114	-114
05:00 PM	6267	CF America	+0.5 +104	+370
Live !	6268	- Draw		+258

Figura 3.4: Líneas disponibles para Tigres - Ameríca en notación americanda, tomado de Pinnacle

ginalmente también: si el momio fuera $\frac{5}{1}$, el jugador recibiría \$600 al haber apostado \$100 (los \$500 de ganancia, más los \$100 apostados originalmente).

A diferencia de los momios decimales con los que hemos tratado hasta ahora, los momios británicos muestran solamente la ganancia potencial; los momios decimales/europeos muestran el pago neto que tendrá la apuesta, incluida la apuesta original. De esta manera, un momio británico $\frac{5}{1}$ es equivalente a un momio decimal 6.

Hay otra notación común en la que se publican los momios. Es la más utlizada en Estados Unidos y, por tanto, es la notación usada en muchas de las casas de apuesta mexicanas. En la Figura 3.4 se observa la misma línea para Tigres - América, tomada de Pinnacle, en la notación Americana. La diferencia fundamental es que ahora los momios tienen dos símbolos (+ o −) junto a la línea. El favorito está usualmente marcado con −(en deportes de dos resultados en vez de tres, como el béisbol, podría haber dos equipos con un momio −. En ese caso, el favorito es aquel cuyo número sea mayor. Más adelante, ilustramos esto con un ejemplo de la NFL). La función del − es indicarle al apostador cuánto tiene que apostar para ganar una unidad de apuesta. En el ejemplo, Tigres es favorito con un momio de −114. Esto quiere decir que el apostador tendría que apostar \$114 para poder ganar \$100. En el fondo, es exactamente equivalente al mimio decimal: $\$114 * 1.877 \approx \$214$ y la ganancia es \$100. El momio con signo + le indica al jugador cuanto ganará si apuesta una unidad. En el ejemplo, América tiene un momio de +370. Es decir, si el jugador apostará \$100 y éste ganare, entonces la casa le pagaría al jugador \$470 y su ganancia sería \$370. En el futbol es común que los tres momios tengan un signo +, en cuyo caso, el momio más pequeño es el que indica al favorito.

La probabilidad implícita (punto de quiebre) en el que un momio cambia de + a − es %50. Es decir, si un evento tiene probailidad implícita mayor a %50, tendrá un signo − (en notación decimal esto ocurre si el momio es menor a 2); si tiene probabilidad implícita menor a 50 %, tendrá un signo más (en notacion decimal, el momio es mayor a 2).

En la figura 3.5 se muestran algunos momios en las tres notaciones, además de sus respectivas probabilidades implícitas.

Probabilidad Implícita	Decimal / Europeo	Moneyline / Americano	Fraccional / Británico
90%	1.11	-900	1/9
80%	1.25	-400	1/4
66 .70 %	1.5	-200	1/2
60%	1.667	-150	2/3
55 . 55%	1.8	-125	4/5
52 . 38%	1.9091	-110	10 / 11
50%	2	+100 / -100	1/1
40%	2.5	+150	3/2
33%	3	+200	2/1
25%	4	+300	3/1
20%	5	+400	4/1
10%	10	+900	9/1

Figura 3.5: Probabilidades implícitas y sus equivalencias en las tres notaciones.

3.2.4. Total de goles en un partido

Si el mercado involucra predecir algún total en el partido, la línea disponible se llama *over/under* o altas/bajas. Por ejemplo, el jugador podría apostar si habrá más o menos de 7.5 tiros de esquina en el partido. Si apuesta las altas y hay 8 o más tiros de esquina, el jugador gana su apuesta. Si apostara las bajas, necesitaría que hubiera 7 o menos tiros de esquina en el partido para ganar.

En el futbol soccer, el segundo mercado más popular es el total de goles anotado en un partido por ambos equipos, es decir, la suma de los marcadores. En la figura 3.6, se muestra que el total para nuestro juego ejemplo Tigres - América es 2.5. Si se apuestan las bajas (*under*), la apuesta gana si el partido tiene un total de goles anotados de 2 o menos. Si se apuestan las altas (*over*), la apuesta gana si el partido tiene un total de goles anotados de 3 o más. El total de 2.5 es el más popular en el mercado de apuestas, puesto que suele ser el más cercano a la realidad del partido y, por lo tanto, suele tener una línea cercana a $50\,\% - 50\,\%$. En el ejemplo, los momios de altas y bajas son, respectivamente, 2.170 y 1.751, con puntos de quiebre $46.08\,\%$ y $57.11\,\%$.

Sat 2/10	6266	Tigres	-0.5 1.877	1.877	Over 2.5 2.170
05:00 PM	6267	CF America	+0.5 2.040	4.700	Under 2.5 1.751
Live !	6268	- Draw		3.580	

Figura 3.6: Línea para Tigres - América, incluyendo el total de goles.

3.2.5. Otros tipos de apuesta

Crystal Palace	7.300	+1	2.060
Tottenham Hotspur	1.490	-1	1.840
Draw	4.500		

Figura 3.7: Líneas disponibles para Crystal Palace - Tottenham, en la cása en línea Pinnacle.

Muchos de los partidos en una temporada tienen a un favorito muy claro, lo cual limita qué tan atractiva es la línea en ambas direcciones.

La mayoría de las ligas deportivas en Estados Unidos no son tan parejas como un partido de soccer promedio. Además, en las cuatro ligas grandes de EUA (NFL, MLB, NBA, NHL) los juegos no pueden terminar en empate: se juega el partido hasta que se tiene un ganador. Es por ello que en la NFL, por ejemplo, las apuestas más populares son aquellas con *handicap*. Los *handicaps* son una manera de equilibrar el partido, de tal manera que la línea de apuesta tenga una ocurrencia cercana a $50\,\% - 50\,\%$. Esto se logra quitándole puntos (en el caso del futbol, goles) al favorito para emparejar el juego. A continuación, un ejemplo para entenderlos:

En la figura 3.7, se muesta una línea para un juego en el que Crystal Palace recibe a Tottenham. Los respectivos puntos de quiebre para la victoria local, empate y visitante son $13.70\,\%$, $22.22\,\%$ y $67.11\,\%$. Sin embargo, la casa ofrece una línea con handicap 1 hacia el favorito: Tottenham -1 (1.84 - $54.35\,\%$) / Crystal Palace +1 (2.06 - $48.54\,\%$). Supongamos que apostamos \$100 a Tottenham -1. Nuestra apuesta tiene ahora tres posibles resultados, en vez de los

dos que tendría si el handicap fuera uno con medio gol:

1.- Si el Tottenham gana por dos o más goles, nuestra apuesta gana y la ganancia será $84.
2.- Si el Tottenham pierde o el juego termina en empate, nuestra apuesta pierde y lq ganancia es -$100.
3.- Si el Tottenham gana por un gol, como el juego cayó exactamente en el *handicap*, nuestra apuesta será rembolsada, en lo que en inglés se conoce como un *push*. Es importante notar aquí que, como la ganancia de un *push* es 0, éste no entra en ningún momento en nuestro cálculo de el valor esperado de la apuesta; sólo los escenarios en los que la apuesta sea decidida influyen en el cálculo.

Es decir, la línea le está quitando un gol al Tottenham para hacer más parejo el encuentro. Los handicaps no necesariamente son enteros, pueden involucrar .5 goles $(-1.5, -2.5, \ldots)$ y hasta .25 de gol en los que son conocidos como *Asian Handicaps*, que no explicamos aquí. Para los *handicaps* con medios goles, los casos son los mismos que en el ejemplo, excepto que ahora el marcador no puede caer exactamente en el handicap, por lo que la apuesta siempre tendrá un resultado y no es devuelta en ninguna situación.

Otros mercados de apuesta populares son si un jugador anotará o no un gol en un partido y si un equipo ganará o no un torneo. En el primero, parece muy complicado poder predecir el comportamiento individual de un jugador, aunque sería interesante ajustar algún modelo e intentarlo.

Para el segundo, estos mercados usualmente están abiertos con límites más grandes antes de empezar la temporada y no durante. Eso quiere decir que el *vig* que las casas usan para las líneas es enorme y que encontrar una apuesta con valor esperado positivo es muy complicado. En ligas con postemporada, como la NFL, MLB o la liguilla en el futbol mexicano, usualmente los momios ofrecidos al apostar pot un equipo al principio de la postemporada no son mucho peores que los ofrecidos al inicio de la temporada, por lo que la casa se tiene que haber equivocado en modelar un equipo fuertemente al inicio del año para que apostar por éste tenga valor. Reciéntemente, las casas de apuesta en Inglaterra aparecieron en las noticias cuando

Leicester City, ganó la English Premier League 2015-2016. El momio al inicio de temporada para el campeonato de Leicester City era 5000-1 (con probabilidad implícita de 0.02 %)
Muchos de los aficionados que apostaron a su equipo en este momio se hicieron famosos, sin embargo, la gran pérdida de la casa vino en momios menores, entre 100 y 500, cuando las casas no ajustaron suficientemente rápido tras el inicio de la temporada [21]. El caso de Leicester es fascinante y un evento de quizá una vez en una generación, donde muchos factores tuvieron que conjuntarse. Leicester se salvó milagrosamente del descenso la temprada anterior a ganar el título: eran últimos hasta la jornada 32, pero ganaron 7 de sus últimos 8 partidos - y empataron el octavo, para acumular 22 de 24 puntos - por lo que lograron salvarse, acabando 6 puntos arriba de la posición de descenso. Para poner la racha en perspectiva, en los primeros 30 partidos, el equipo había ganado solamente 17 puntos. Para la temporada 2015-2016, el factor principal fue que todos los equipos grandes, los favoritos para ganar el título, tuvieron una temporada mala al mismo tiempo. Por ejemplo, Chelsea, el campeón de la temporada anterior, terminó 10°, consiguiendo 37 puntos menos que en la 2014-2015. Sin embargo, Alex Donhoue, que maneja las Relaciones Públicas de futbol para Ladbrookes, una de las casas más grandes en el Reino Unido, declaró que "es una falsedad si cualquier casa dice que ha perdido dinero en total con Leicester. Aunque £3 millones es un record en un pago individual para un ganador de título, a la casa le fue bien con Leicester sorprendiendo semana a semana para llegar ahí. Ninguna queja en absoluto." [21]

Otra diferencia fundamental del futbol soccer y un deporte como el futbol americano es que, como el número de goles por partido es relativamente bajo, las permutaciones de posibles marcadores lo son también. Esto permite un mercado de marcadores exactos en el futbol soccer. Es decir, yo puedo apostar que el partido quedará exactamente 3-1 hacia el equipo local, por ejemplo. De nuevo, dado que los pagos para los marcadores exactos son altos y la modelación complicada, la casa se protege cobrando una comisión muy alta para estas líneas.

3.3. El mercado de apuestas

3.3.1. La eficiencia del mercado

Es natural preguntarse, ¿qué tan difícil es vencer las líneas que las casas publican? O, planteado de otra manera, ¿qué tan buen predictor son las líneas de apuestas disponibles?

Lopez, Matthews, Baumer (2017) [22] analizan esto, utilizando datos de las cuatro ligas mayores de EUA (MLB, NBA, NHL y NFL) desde 2006 hasta 2016, con líneas de apuesta obtenidas de Sports Insights. Los autores tienen una línea de apuesta disponible para $99 + \%$ de los juegos en sus datos. El *vig* promedio en los datos es $1.93\,\%$, indicando mercados muy líquidos. Los autores utilizan la línea disponible más cercana al inicio del partido, es decir, la línea de cierre. Esta consideración - que se utilicen las líneas de cierre - es sumamente importante; más adelante se explica por qué.

Un resumen de lo encontrado por los autores se puede ver en la Figura 3.8. Como vemos, la victoria local en las cuatro ligas se da entre $54\,\%$ y $59\,\%$ de los partidos. p_{games} es la media de la probabilidad observada para una victoria local, p_{bets} es la media de la probabilidad implícita en los momios para una victoria local. Para las cuatro ligas, las dos son muy cercanas.

Para comparar estas ligas con futbol, en la English Premier League, temporada 2016-2017, la victoria local se dio en sólo $49.2\,\%$ de los partidos. En ambos torneos de la Liga MX celebrados en 2016-2017, la victoria local se dio solamente en $44.6\,\%$ de los partidos. Esto responde además, a los tres posibles resultados de un partido: ninguno de los partidos en las cuatro ligas del estudio puede terminar en empate.

Los autores comparan la probabilidad promedio de una victoria local observada en los datos con la probabilidad promedio implícita en las líneas de apuesta granularmente. El método es muy intuitivo: se agrupan todos los partidos cuya línea predice tendrán victoria local con alguna probabilidad - redondeado hasta la centena más cercana - digamos, $50\,\%$, y se observa cuantas veces ganó el equipo local estos partidos. La figura 3.9 muestra un resumen gráfico de

Sport (q)	t_q	n_{games}	$\bar{p}_{games}$	n_{bets}	$\bar{p}_{bets}$	Coverage
MLB	30	26728	0.541	26710	0.548	0.999
NBA	30	13290	0.595	13245	0.615	0.997
NFL	32	2560	0.563	2542	0.589	0.993
NHL	30	13020	0.548	12990	0.565	0.998

Figura 3.8: Resumen de la información para los cuatro deportes. t_q es el número de equipos, n_{games} el número de juegos, n_{bets} el número de juegos con línea de apuesta disponible, p_{games} es la media de la probabilidad observada para una victoria local, p_{bets} es la media de la probabilidad implícita en los momios para una victoria local.

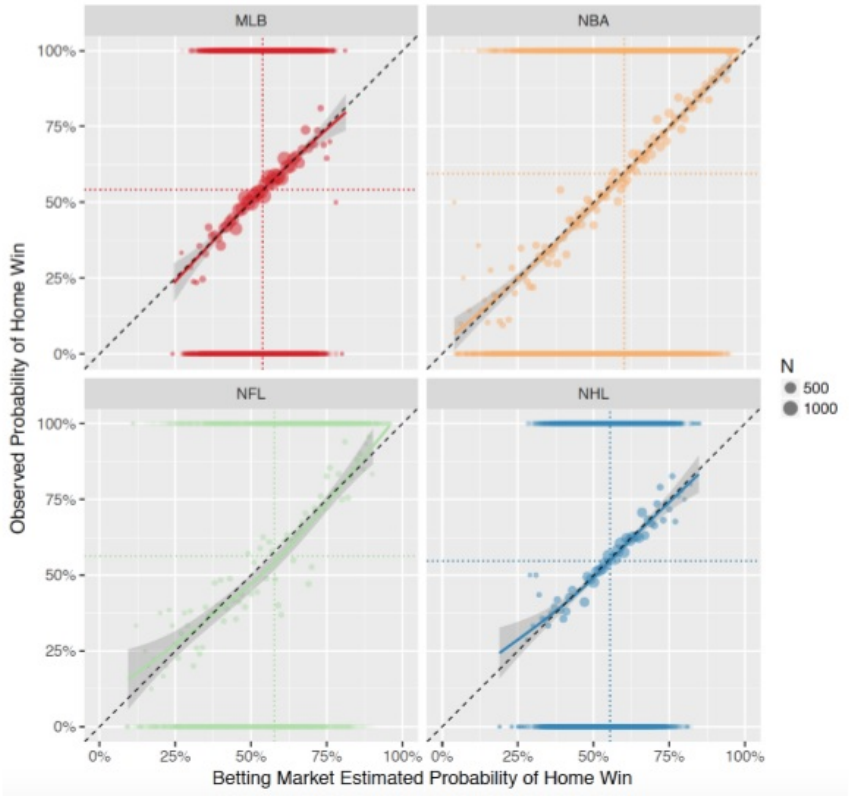

Figura 3.9: Exactitud de la probabilidad implícita de los mercados de apuestas. Cada punto representa un cúmulo de probabilidades, redoneadas a la centena más cercana. El tamaño del punto (N) es proporcional al número de juegos en el cúmulo. La línea diagonal implica un mercado exacto donde las probabilidad implícitas y observadas coinciden perfectamente.

la información, donde se compara la probabilidad de victoria local observada en los datos, contra la probabilidad de victoria implícita en los momios. El tamaño de los puntos está relacionado a cuántos partidos hay en ese cúmulo. En ninguno de los cuatro deportes se puede rechazar la hipótesis de un mercado eficiente, por lo cual los autores concluyen que no hay evidencia para sugerir que las probabilidades de cierre del mercado de apuestas son inexactas o tienen un sesgo.

3.3.2. ¿Cómo se crean las líneas de apuesta?

El hecho de que las líneas de cierre sean eficientes colectivamente, no quiere decir que no se puedan encontrar oportunidades de apuesta redituables o que la casa modele perfectamente todos los eventos. Para ello, se debe entender cómo funciona el mercado de momios. El director de Trading en Pinnacle, Marco Blume, estuvo en el podcast *The Buisness of Betting* recientemente explicando el proceso con el que Pinnacle esatblece sus líneas y es un buen vistazo a cómo funciona la industria [23].

Blume: "La mayoría de nuestras líneas de apertura son francamente pobres. Intentamos hacer nuestra mejor estimación en las líneas de apertura, pero sabemos que son pobres. (...) Tenemos que ajustar exitosamente para poder ganar dinero. Intentamos tener la mejor línea disponible y utilizar toda la información que se nos presenta en forma de apuestas para mejorarla. Por eso, no discriminamos contra apostadores ganadores."

El mercado funciona así: las líneas de apertura son la mejor estimación que tiene la casa de apuesta sobre el partido. Por supuesto, es complicado que la casa tenga información perfecta cuando están publicando líneas para decenas de países y, en ocasiones, múltiples ligas por país. Por tanto, las líneas de apertura son las menos exactas. La casa abre estas líneas con límites de apuesta bajos y comienza a recibir apuestas. Si una línea es mala, los apostadores profesionales comenzarán a entrar al mercado y atacarla. Esto le da información a la casa de que su línea está fuera de rango y necesita ajustarla. Por ejemplo, un equipo podría abrir -105 y si muchas apuestas entran

hacia ese equipo, moverse rápidamente hacia -110 o -115.

La casa ajusta por dos razones: primero, para detener las apuestas unilaterales, haciendo el precio menos atractivo; segundo, como ese precio se ajusta hacia abajo, el resto de los precios sube un poco, por lo que el otro lado de la apuesta se vuelve más llamativo, intentando atraer apostadores a ese lado también. Este proceso continúa por algún tiempo, hasta que la línea se estabiliza en un precio. La casa tiene ahora mucha mejor información de qué cree el mercado sobre este partido y dónde debería estar la línea.

En ese momento, la casa ajusta para atraer a los apostadores grandes. Puede hacerlo de dos maneras: primero, puede reducir el *vig* que tenía la línea inicial, haciendo que todas las líneas se conviertan en un poco más interesantes. Segundo, abre un poco más los límites para poder recibir apuestas más grandes. Esto atrae a aquellos apostadores a los que no les convenía entrar temprano al mercado, puesto que los límites no eran llamativos para la escala de sus apuestas y no están dispuestos a decirle a la casa - en forma de apuestas - qué apuesta les interesa hasta no poder apostar mucho más que lo que permiten las líneas iniciales. Tras la apertura de límites, hay otro pequeño momento donde la línea se moverá rápido y la casa debe ajustar. Esto continúa hasta el inicio del partido: la casa recibe apuestas - que puede usar como información - y ajusta su línea.

Por todo lo anterior, la línea de cierre es lo más cercano a una probabilidad real que existe: es el reflejo de las opiniones de todo el mercado. En consecuencia, es importante atacar la línea temprano, cuando es más ineficiente y hay menos información disponible para la casa.

En cuanto a balancear la línea, es decir, intentar recibir la misma cantidad de dinero en ambos lados de un juego, Blume declaró: “Si pasa [un balance de apuestas] es bueno para nosotros; si las apuestas son de un sólo lado pero era algo que anticipábamos, es bueno para nosotros; si las apuestas son de un sólo lado pero no era algo que anticipábamos, es muy malo para nosotros. La apertura no es tan importante, es simplemente una conjetura. Sabemos que nuestros

clientes son mejores acertando que nosotros. Por esto, las líneas al principio se mueven mucho y cerca del inicio del partido casi no se mueven. No tenemos miedo en desbalancear una línea si creemos que tenemos una ventaja de información."

3.3.3. El mercado global de apuestas

El mercado global de apuestas es enorme, estimado en tres mil millones de dólares - la estimación es de gananacia, el total de dinero apostado es mucho mayor - de los cuales alrededor de 90 % son apostados ilegalmente. El debate sobre si las apuestas deportivas deben ser legales y su efecto en las ligas deportivas sigue abierto.

En Estados Unidos, es ilegal apostar en deportes, exceptuando Nevada, Delaware, Orgeon y Montana. Sin embargo, las opciones en los últimos tres estados son muy limitadas. Por ello, la inmensa mayoría de los apostadores profesionales vive en Nevada (específicamente, en Las Vegas).

La legislación sobre apostar en línea es un poco menos clara. Es totalmente ilegal recibir apuestas (tanto en línea, como en persona). Pero la ley no es muy clara sobre si realizar apuestas en línea es ilegal. Ningún estadounidense ha sido arrestado jamás por apostar en línea. Recientemente, New Jersey ganó en la Suprema corte de justicia en Estados Unidos el derecho de poder ofrecer apuestas legalmente en su estado y se cree que esta decisión es el inicio de la legalización de las apuestas por completo en Estados Unidos.

Como todo mercado ilegal, el hecho de que la mayoría de las apuestas sean ilegales genera algunos problemas de crimen. El más grande es que las personas que reciben apuestas, conocidas coloquialmente como *bookies*, reciben apuestas en crédito. Cuando las apuestas son contra una casa, la misma recibe el dinero de la apuesta en el momento en que ésta se hace. Sin embargo, como un *bookie* no puede tener una localización física, ni papeleo para garantizar las apuestas, éstas se hacen con crédito hacia el jugador y en un sistema de confianza. Si el jugador acumula una deuda grande con el *bookie* que no puede pagar, esto podría generar problemas de vio-

lencia. Desde hace algunos años, el comisionado de la NBA Adam Silver ha empujado por la legalización de las apuestas en EUA, lo cual ha dado un impulso a las pláticas y discusión al respecto. La NBA está negociando su postura respecto a cómo funcionaría un sistema legal de apuestas, donde la liga se llevaría un porcentaje de todas las apuestas realizadas. Esto podría ser un ingreso adicional sustancial para la liga.

En México, las apuestas deportivas son totalmente legales bajo los estatutos publicados en la reforma al Reglamento de la ley federal de juegos y sorteos, publicado en 2014. Bajo la ley mexicana el organizador está obligado a retener el impuesto sobre la renta y el impuesto sobre loterías rifas, sorteos o concursos, a quien tenga ganancias en apuestas. Por ello, cualquier apuesta realizada en una casa física (ya sea de deportes o un hipódrimo), de ganar, devuelve un valor menor al contemplado en el boleto de apuesta. El importe exacto depende de la legislación local, pero el impuesto federal es del 1 % sobre el total de la apuesta. Los detalles se encuentran en el Artículo 138 del Impuesto sobre la renta.

Capítulo 4

Estrategias óptimas para sistemas de apuestas y El Criterio de Kelly

Una vez que nuestro modelo ha identificado apuestas con valor esperado positivo, surge la pregunta natural: ¿cuánto apostar? ¿se debe apostar lo mismo en todas las apuestas? ¿se debe apostar más en algunas situaciones?

Hay muchos enfoques que devuelven distintas estrategias, dependiendo de lo que el jugador quiere obtener de las apuestas. Por ejemplo, se puede buscar minimizar la probabilidad de ruina o maximizar la probabilidad de alcanzar una ganancia arbitraria tras una cierta ventana de tiempo. Por otro lado, se puede definir una función de utlidad, que depende de alguna medida y refleja la actitud del jugador ante el riesgo.

En este capítulo, exploramos el Criterio de Kelly, una estrategia compatible con probabilidad de ruina cero que maximiza el crecimiento logarítmico esperado. Primero, discutimos una perspectiva histórica de las funciones de utilidad. Posteriormente, el origen y propiedades del Criterio de Kelly. Finalmente, su uso en el contexto de apuestas deportivas.

4.1. Estretegias de apuesta en el tiempo

La pregunta de cómo usar estrategias de apuesta en favor del apostador data al menos del siglo XVIII, cuando Nicolas Bernoulli planteara la que, 25 años después, sería conocida como la Paradoja de San Petersburgo, tras la solución ofecrida por su primo Daniel Bernoulli.

4.1.1. La Paradoja de San Petersburgo

La paradoja es la siguiente:

Un casino ofrece un juego donde, en cada turno, se tira una moneda. Inicialmente, el premio son 2 dólares. El premio se duplica cada vez que la moneda muestra cara. Cuando aparece cruz por primera vez, el juego termina y el jugador gana lo que sea que haya de premio en ese momento. Por ejemplo, si la secuencia de lanzamientos fuera cara, cara, cruz; el jugador ganaría 8 dólares, pues el premio se duplicó dos veces. Es sencillo ver que el jugador gana 2^K dólares, donde K es el lanzamiento donde apareció cruz por primera vez. ¿Cuál sería un precio justo para pagar al casino y entrar a este juego?

Para contestar, necesitamos saber cuál es la ganancia esperada para el jugador. El jugador gana 2 dólares con $\mathbb{P} = \frac{1}{2}$, gana 4 dólares con $\mathbb{P} = \frac{1}{4}$ y así sucesivamente. Por lo tanto, el valor esperado de G, la ganancia, es:

$$\mathbb{E}\left[\mathrm{G}\right] = \sum_{k=1}^{\infty} \mathbb{P}\left[\text{Juego acabe en K volados}\right] * 2^K = \sum_{k=1}^{\infty} \frac{1}{2^K} * 2^K = \sum_{k=1}^{\infty} 1 = \infty$$

O, visto de otra manera:

$$\mathbb{E}[\text{Ganancia jugador}] = \frac{1}{2}\ 2 + \frac{1}{4}\ 4 + \frac{1}{8}\ 8 + \cdots = 1 + 1 + 1 + \cdots = \infty$$

Esto, por supuesto, presupone que el casino tiene recursos ilimitados y que el juego puede continuar mientras siga cayendo cara. Considerando entonces que el valor esperado de la ganancia es infinito, el jugador debería estar dispuesto a pagar cualquier cantidad

para jugar este juego. Sin embargo, la mayoría de la gente no pagaría mucho para entrar al juego y ahí radica la paradoja.

4.2. Optimización con respecto a Funciones de utilidad

La solución propuesta por Daniel Bernoulli (1738, traducida al inglés por Sommer 1954) [24] es un precursor directo de muchos conceptos que ahora se utilizan en economía, como la Ley de la Utilidad marginal decreciente. Bernoulli: "La determinación del valor de un objeto no debe estar basada en el precio del mismo, sino en la utilidad que éste genera. (...) No hay duda alguna que una ganancia de mil ducats es más significativa para el pobre que para el rico, aunque ambos ganan la misma cantidad."

Bernoulli sugiere una función de utilidad logarítmica $U(w) = ln(w)$, donde la utilidad está dada por el logaritmo de la riqueza total del apostador w. Tener una función de utilidad cóncava, como $U(w) = ln(w)$, refleja una posición adversa ante el riesgo: cuando la cantidad crece, los cambios de misma magnitud devuelven una utilidad menor. Esto expresa, matemáticamente, la idea de Bernoulli: \$1000 para una persona con menor riqueza son mucho más útiles que para la misma persona si tuviera mayor riqueza. Por ejemplo, para alguien con \$1000 pesos de riqueza, otros \$1000 representan, bajo una función de utilidad logarítmica, $ln(2000) - ln(1000) \approx .693$ unidades de utilidad extra; mientras que alguien con \$500000 habrá obtenido solamente $ln(501000) - ln(500000) \approx .002$ unidades.

Aunque esta función resuelve la paradoja para la estructura de pagos antes propuesta, es fácil ver que modificando la función pagos en el ejemplo de los volados de 2^K a e^{2^K}, la paradoja vuelve a aparecer, puesto que la esperanza es nuevamente infinita.

Este problema es tan inherentemente interesante, que desató una rama matemática completamente nueva: la Teoría de la decisiones, que terminara por formalizarse en 1947, cuando John von Neumann y Oskar Morgenstern probaron el Teorema de la utilidad von

Neumann-Morgenstern. Éste muestra que bajo ciertos axiomas de comportamiento racional, un tomador de decisiones - en la versión original de Bernoulli, el apostador - intentará maximizar su utilidad esperada para una función definida, llamada Función de utilidad. Por supuesto, como en la economía tradicional, los axiomas propuestos por von Neumann y Morgenstern tienen muchas situaciones donde son muy controversiales. Esto llevería algunas décadas después a Amos Tversky y Daniel Kahneman a desarollar la Teoría de la Perspectiva, la cual le ganaría al segundo en 2002 el Premio en memoria de Alfred Nobel en Economía - Tvesrky murió de cancer en 1996 y el premio no se entrega póstumamente. La historia de la relación entre ambos está narrada magníficamente en Lewis (2017). [25]

4.2.1. Claude Shannon y la Teoría de la comunicación: un antecedente al Criterio de Kelly

La función de utilidad logarítmica propuesta para ofrecer una "solución" a la Paradoja por Bernoulli tiene, sin embargo, propiedades muy interesantes que fueron exploradas por Kelly (1956) [26] en una publicación extraordinaria. Éste logra aplicar los conceptos que propone Shannon (1948) [27], que hoy se convirtieron en el principio lo que llamamos la Teoría de la Comunicación y en los que están cimentados todas las telecomunicaciones modernas, en un contexto totalmente diferente. Bajo éste, Kelly encuentra una estrategia de apuesta óptima.

En la idea básica de la Teoría de la Comunicación, se quiere comunicar un mensaje entre un emisor y un receptor a través de un canal. La teoría de Shannon establece que dígitos binarios se pueden codificar y transmitir por dicho canal con una probabilidad de error arbitriamente pequeña. Para Kelly, sin embargo, queda pendiente "encontrarle significado a un sistema de comunicación con una tasa de error no despreciable."
Además, Kelly sugiere que es fácil asignar una función de costo, es decir, de utilidad, a pares de símbolos recibidos por el canal. Esto se convierte, bajo la escala de la función de utilidad, en un sistema aleatorio con resultados deseables e indeseables. De esta forma, el

receptor se puede beneficiar del resultado del mensaje y de una buena forma de predecir qué mensaje vendrá en el canal.

Para Kelly, este planteamiento general no tiene porque estar limitado a un canal de comunicación y podría ser extendido a cualquier situación donde un evento aleatorio (el mensaje emitido a través del canal, con errores) tiene resultados positivos y negativos (utilidad) para un receptor (podría ser un apostador), que puede utilizar información acerca del canal (la distribución de los errores) y las entradas (por ejemplo, parámetros). Ofrece un ejemplo para ilustrar sus ideas:

Ejemplo 8. Imagine un canal de comunicación sin ruido, utilizado para transmitir los resultados de una sucesión de partidos de béisbol, antes de que estos resultados se sepan públicamente, permitiendo que se pueda apostar en los momios originales del partido. Los partidos son entre dos equipos equilibrados, lo cual permite apostar por cualquiera de los dos con momio 2 a 1, aunque el apostador ya sabe el resultado del partido. ¿Cuánto debería apostarse en cada partido?

Como el apostador ya conoce el resultado, resulta lógico que apueste todo lo que tiene en cada partido. De esta manera, su capital crece por un factor 2^n tras n juegos apostados. Esta estrategia sólo tiene sentido pues no hay error alguno en la transmisión de resultados por el canal. Si éste tuviera cualquier ruido asociado y hubiera una probabilidad no cero de error en el mensaje, habría problemas.

La idea del Criterio de Kelly es hacer una apuesta en cada repetición para maximizar $\mathbb{E}[log(X)]$ donde X es la variable aleatoria que representa el capital. Sus propiedades se pueden ilustrar fantásticamente con un ejemplo, tomado de Thorp (2008) [28], dentro de MacLean, Thorp, Ziemba (2011): [29]

Ejemplo 9. Imagine que vamos a tirar volados independientes entre sí contra un oponente de capital ilimitado, donde ambos jugadores arriesgan la misma cantidad de dinero en cada volado. Para cada uno, la probabilidad de ganar es $p > \frac{1}{2}$, la probabilidad de perder es $q = 1 - p < \frac{1}{2}$.
Nuestro capital inicial es X_0. Sea X_n nuestro capital tras tirar n

volados. Si nuestra meta es maximizar el capital esperado tras n volados, es decir, $\mathbb{E}(X_n)$, ¿cuánto debería ser la apuesta en el k-ésimo volado, B_k?

Para cada k volado, sea $V_k = 1$ si hemos ganado el volado y $V_k = -1$ si lo hemos perdido.
Es claro ver que $X_k = X_{k-1} + V_k B_k$; es decir, nuestro capital tras el k-ésimo volado es simplemente la suma de nuestro capital un volado antes, más el resultado de la apuesta en el k-ésimo volado. Esta es una Cadena de Markov, pues sólo depende de la innovación y el estado de la cadena a un paso anterior. De esta forma:

$$X_n = X_0 + \sum_{k=1}^{n} V_k B_k$$

Por lo tanto:

$$\mathbb{E}[X_n] = \mathbb{E}[X_0 + \sum_{k=1}^{n} V_k B_k] = X_0 + \sum_{k=1}^{n} \mathbb{E}[V_k B_k] = X_0 + \sum_{k=1}^{n} (p-q) E[B_k] \tag{4.1}$$

Como $p > \frac{1}{2}$, entonces $p - q > 0$ y el juego tiene esperanza positiva.

Dada la forma de la ecuación (4.1), para poder maximizar $\mathbb{E}[X_n]$, necesitaríamos maximizar $\mathbb{E}[B_k]$ en cada volado. La ganancia se maximiza, entonces, si apostamos todos nuestros recursos en cada volado. Por lo tanto, $B_1 = X_0$. De ganarse esta apuesta, $B_2 = 2\,X_0$ y así sucesivamente. Sin embargo, bajo esta estrategia, el jugador se arruina si pierde un solo volado. Esta probabilidad está dada por $1 - p^n$ tras n volados - es, simplemente, el complemento de haber ganado esos n volados. Como $p < 1$ (ningún oponente accedería a un juego en donde pierde seguramente en cada realización):

$$\lim_{n \to \infty} 1 - p^n = 1$$

por lo que la ruina se da casi seguramente. Esta es, claramente, una estrategia indeseable. En el ejemplo anterior, si los resultados del partido de béisbol transmitidos por el canal tuvieran un error no

cero de transmisión, entonces el jugador se arruinaría con probabilidad uno al apostar todo su capital en cada partido.

Por otro lado, la única manera de garantizar que el jugador nunca se arruina (i.e. que no ocurra la barrera absorbente $X_k = 0$ en algún momento k) es apostando \$0 en cada volado, lo cual minimiza simultáneamente la ganancia esperada a \$0 y hace que el jugador se quede simplemente con X_0.

4.3. El Criterio de Kelly

Nos gustaría encontrar, entonces, una estrategia intermedia entre maximizar $\mathbb{E}[X_n]$, lo cual aseguraría la ruina, y minimizar la probabilidad de ruina, lo cual asegura una ganancia cero.

Puesto que las probabilidades y los pagos para cada lanzamiento de la moneda son los mismos - son realizaciones independientes e idénticamente distribuidas - es plausible pensar que una estrategia óptima apostaría siempre el mismo porcentaje del capital disponible.

Sean entonces S y F el número de éxitos y fracasos respectivamente, en n experimentos. $S + F = n$. Si apostáramos una fracción fija del capital en cada realización, es decir:

$$B_i = f\, X_{i-1} \text{ con } 0 \leq f \leq 1$$

el capital tras n lanzamientos está dado por:

$$X_n = X_0\,(1+f)^S(1-f)^F$$

De esta manera, bajo la restricción $0 < f < 1$, la probabilidad de ruina, $\mathbb{P}[X_n = 0]$ es cero, aunque el capital disponible para el apostador se puede hacer muy chico. Por ello, reinterpretamos la probabilidad de ruina como la probabilidad de que el capital se vuelva menor a un umbral predefinido ε. Es decir:

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq \varepsilon] = 1$$

Definimos además la tasa de crecimiento exponencial del capital como:

$$G = \lim_{N\to\infty} \frac{1}{N}\, log\left(\frac{X_N}{X_0}\right)$$

donde X_N es el capital del apostador tras N apuestas y X_0 es el capital inicial.

Para el ejemplo anterior, como $X_n = X_0\,(1+f)^S\,(1-f)^F$:

$$G = \lim_{N\to\infty}\left[\frac{S}{N}\log(1+f) + \frac{F}{N}\log(1-f)\right] = q\,\log(1+f)+p\,\log(1-f)$$

donde la última igualdad se da con probabilidad uno: a largo plazo, la probabilidad de éxito y fracaso se estabilizan a sus probabilidades reales.

Kelly decide maximizar el valor esperado del coeficiente de crecimiento - que definimos como $g(f)$ - puesto que:

$$g(f) = \mathbb{E}\left[\log\left(\frac{X_n}{X_0}\right)^{\frac{1}{n}}\right] = \mathbb{E}\left[\frac{S}{n}\log(1+f) + \frac{F}{n}\log(1-f)\right] = p\,\log(1+f)+q\,\log(1-f)$$

Por otro lado, notemos que:

$$g(f) = \mathbb{E}\left[\log\left(\frac{X_n}{X_0}\right)^{\frac{1}{n}}\right] = \frac{1}{n}\,\mathbb{E}\left[\log(X_n)\right] - \frac{1}{n}\log(X_0)$$

Así que, para n fijo, maximizar $g(f)$ es lo mismo que maximizar $\mathbb{E}[\log(X_n)]$; nuestra intención inicial. La ventaja de trabajar con $g(f)$ es que tiene una derivada sencilla de encontrar:

$$g'(f) = \frac{p}{1+f} - \frac{q}{1-f} = \frac{p-q-f}{(1+f)(1-f)} = 0 \iff f = f^* = p-q$$

Para mostrar que éste es en verdad el máximo, encontramos la segunda derivada:

$$g''(f) = -\frac{p}{(1+f)^2} - \frac{q}{(1-f)^2} < 0$$

Esto pues:

$$-\frac{p}{(1+f)^2} < 0 < \frac{q}{(1-f)^2}$$

Por otro lado:

$$g(0) = 0 \qquad \lim_{f \to q^-} = -\infty$$

Esto quiere decir que existe un único número f_c tal que:

$$f_c > 0 \qquad g(f_c) = 0 \qquad 0 < f^* < f_c < 1$$

La importancia de f_c será explicada más adelante.

Notemos además que $g(f^*) = p\ log(p) + q\ log(q) + log(2) > 0$. Es decir, el crecimiento exponencial en el punto óptimo es mayor a cero.

4.3.1. Propiedades del Criterio de Kelly

Las propiedades más importantes del criterio están listadas a continuación. Las demostraciones se pueden encontrar en Breiman (1961) [30] y Thorp (1969) [31]:

1. Si $g(f) > 0 \Longrightarrow \lim_{n\to\infty} X_n = \infty$ casi seguramente

Es decir, para cada M fija:

$$\mathbb{P}\left[\lim_{n\to\infty} X_n > M\right] = 1$$

2. Conversamente, si $g(f) < 0 \Longrightarrow \lim_{n\to\infty} X_n = 0$ casi seguramente

Es decir, para cada $\varepsilon > 0$ fija:

$$\mathbb{P}[\lim_{n\to\infty} X_n < \varepsilon] = 1$$

El punto 1. muestra que, salvo por un número finito de trayectorias - puesto que la convergencia se da casi seguramente - X_n, la fortuna del jugador, excederá cualquier umbral fijo M, cuando f se elige en el intervalo $(0, f_c)$. Pero 2. muestra que, si $f > f_c$, la ruina es entonces casi segura.

3. Si $g(f) = 0$, entonces:

$$\limsup_{n\to\infty} X_n = \infty \text{ casi seguramente} \qquad \liminf_{n\to\infty} X_n = 0 \text{ casi seguramente}$$

El punto 3. demuestra que, si $f = f_c$, X_n oscilará aleatoriamente - casi seguramente - entre ∞ y 0.

4. Dada una estrategia de apuesta Φ^* que maximiza $\mathbb{E}[\log(X_n)]$ y cualquier otra estrategia esencialmente diferente Φ; es decir, una estrategia que no necesariamente apuesta una fracción del capital disponible en cada evento, se cumple que:

$$\lim_{n\to\infty} \frac{X_n(\Phi^*)}{X_n(\Phi)} = \infty \text{ casi seguramente.}$$

5. El tiempo esperado para que el capital actual X_n llegue a cualquier nivel arbitrario C es, asintóticamente, el mínimo con una estrategia que maximiza $\mathbb{E}[\log(X_n)]$.

Los puntos 4. y 5. establecen que la estrategia propuesta por Kelly, maximizar $\mathbb{E}[\log(X_n)]$, es asintóticamente óptima bajo dos criterios importantes para cualquier estrategia de apuestas: en crecimiento bruto y en tiempo de crecimiento.

Para precisar, por 'estrategia escencialmente diferente', nos referimos a una tal que la diferencia $\mathbb{E}[\log(X_n^*)] - \mathbb{E}[\log(X_n)]$ entre X_n^*, el capital bajo la estrategia de Kelly y X_n, el capital bajo la otra estrategia esencialmente diferente, crece más rápido que la desvación estándar de $\log(X_n^*) - \log(X_n)$ (asegurando así que $\mathbb{P}[\log(X_n^*) - \log(X_n) > 0] \to 1$).

6. Suponga que el rendimiento al apostar una unidad en el i-ésimo evento es una variable aleatoria binomial U_i; más aún, suponga que la probabilidad de éxito es p_i con $\frac{1}{2} < p_i < 1$.

Entonces $\mathbb{E}[\log(X_n)]$ se maximiza al elegir, en cada evento, la fracción de apuesta $f_i^* = p_i - q_i$. Ésta maximiza $\mathbb{E}[\log(1 + f_i U_i)]$

Por último, 6. afirma la validez de utilizar el criterio de Kelly, eligiendo f_i^* en cada realización - aún si las probabilidades cambian entre una realización y otra - para poder maximizar $\mathbb{E}[\log(X_n)]$.

Ejemplo 10. Un jugador apuesta contra un oponente de riqueza infinita. El jugador gana volados con $p = .52$ y los pagos son dos a uno. El jugador tiene un capital inicial X_0. Entonces, usando 6:

$$f_i^* = f^* = .52 - .48 = .04$$

Es decir, el jugador debe apostar 4 % de su capital en el momento de cada volado para que X_n tenga el mayor crecimiento posible, compatible con probabilidad de ruina cero, en el sentido de que la riqueza se vuelva menor a ε. Si la apuesta fuera menor a 4 %, puesto que el jugador tiene una esperanza positiva en cada lanzamiento, X_n también crecería a infinito, pero con una taza menor; más lentamente. El coeficiente de crecimiento está dado por:

$$g(f^*) = g(.04) = .52 \ log(.52) + .48 \ log(.48) + log(2) \approx 0.0008$$

De tal manera que, después de n apuestas sucesivas, la esperanza del logaritmo de la riqueza tiende a $0.0008 * n$. Podemos calcular el momento esperado en que los fondos se duplican, haciendo que $0.000800213 \ n = log(2)$. Esto nos da un tiempo esperado de $n \approx 866$ volados.

Además, podemos usar la forma de $g(f) = .52 \log(1 + f) + .48 \log(1 - f)$ para encontrar los dos ceros de la función: 0 (que ya conocíamos) y $f_c = 0.0799147$. Es decir, si el jugador apuesta más de 8 % del capital en cada volado, aunque temporalmente la taza de crecimiento podría ser mayor, las fluctuaciones del proceso harán que eventualmente $X_n \to 0$

4.3.2. El Criterio de Kelly para pagos desiguales

El criterio de Kelly se puede extender fácilmente a situaciones con pagos que no son iguales para los dos jugadores; es decir, distintos de dos a uno. Suponga que el jugador A gana b unidades por cada apuesta unitaria. Más aún, suponga que la probabilidad de ganar es

$p > 0$ y que $p\,b - q > 0$ para que el juego presente una ventaja para el jugador A - en caso contrario, éste no apostaría.

Análogamente al desarrollo anterior, donde los pagos eran iguales para ambos jugadores, obtenemos la función de crecimiento para pagos desiguales, dada por:

$$g(f) = \mathbb{E}\left[\log\left(\frac{X_n}{X_0}\right)\right] = p\ log(1+bf) + q\ log(1-f)$$

Nuevamente, podemos derivar y encontrar el máximo $f^* = \frac{bp-q}{b}$. Esta es la fracción óptima del capital disponible que deberá ser apostada en cada iteración para maximizar el coeficiente de crecimiento $g(f)$.

La intuición coincide perfectamente con $\frac{bp-q}{b}$:
Con b la posible ganancia fija, si la probabilidad de ganar p crece, entonces q decrece y el tamaño de la apuesta crece. El converso se da también, si p decrece, q crece y el tamaño de la apuesta decrece. Por último, con p fija, si la ganancia posible b crece, el tamaño de la apuesta crece también.

Como el Criterio de Kelly maximiza la taza de crecimiento esperado asintóticamente, también es conocida como la Estrategia de crecimiento óptimo.

4.3.3. Consideraciones del Criterio de Kelly para apuestas deportivas

La vida real, por supuesto, es mucho más complicada que la teoría. Hasta ahora, implícitamente, hemos asumido que conocemos p_i, la probabilidad de éxito en cada una de las realizaciones del proceso. Bajo esta hipótesis, fuimos capaces de obtener f_i. Sin embargo, para todas las aplicaciones interesantes y, en particular, la que compete a esta tesis, p_i no es conocida; está siendo estimada por un modelo.

Sea $m_t = p_t - q_t$ donde p_t, q_t son las probabilidades reales y $m_e = p_e - q_e$ donde p_e, q_e son las probabilidades estimadas por el

modelo.
Como m_e es una estimación con incertidumbre de m_t, es bueno asumir que $m_t < m_e$; es decir, que la ventaja real que tiene el jugador es menor que la estimada por el modelo. Por ende, asumimos también que $f < f_e$. Esto, para prevenir $g \leq 0$ y por tanto, la ruina.

Aún si el modelo predictivo estuviera perfectamente especificado - que en realidad, nunca lo estará - el fenómeno modelado nunca es estático. En particular, el futbol cambia mucho: cada temporada los equipos venden y compran jugadores, algunos equipos ascienden/-descienden a la división superior/inferior, los estadios cambian, los balones no son los mismos, las estrategias y objetivos de los equipos evolucionan, etcétera.

MacLean, Thorp, Ziemba (2011) [29] muestran que, sujeto a la condición anterior, escojer f en el rango $\frac{1}{2}f_e^* \leq f < f_e^*$ ofrece una protección contra $g \leq 0$ que probablemente no reduzca el crecimiento por más de 25 %.

Al aplicar estos modelos a apuestas deportivas, los apostadores usualmente usan una fracción del Criterio de Kelly, dependiendo del tamaño muestral donde el modelo ha sido exitoso. Por ejemplo, pueden empezar usando una fracción tan baja como $\frac{1}{10}$ Kelly al comenzar a utilizar modelo, pero, tras establecer su éxito y valor predictivo, abrir sus apuestas hasta $\frac{1}{2}$ Kelly. Esto, por supuesto, depende enteramente de la aversión al riesgo del apostador y la confianza que le tenga al modelo.

Capítulo 5

Un análisis exploratorio de la naturaleza aleatoria de marcadores y el mercado de apuestas

En este capítulo, hacemos un análisis del mercado de apuestas y su relación con las probabilidades reales de ocurrencia. Posteriormente, analizamos evidencia empírica para explorar si la naturaleza de los marcadores en futbol soccer está bien representada por una distribución Poisson.

5.1. Probabilidades implícitas y eficiencia del mercado de apuestas para futbol soccer

Siguiendo la línea de investigación realizada por Lopez, Matthews, Baumer (2017) [22], discutida en el capítulo Apuestas deportivas, queremos analizar si las líneas de apuesta en el mercado son efectivamente predictivas de lo que pasa en el campo de juego de futbol soccer, como lo son para la NBA, NFL, NHL y MLB.

Para ello, hacemos un análisis del mercado de los momios. La casa de apuesta elegida para el análisis es Pinnacle, que es la más grande en el mercado de futbol, por lo que recibe la mayor cantidad de apuestas profesionales. Se usa la línea de cierre por razones que ya

han sido previamente explicadas en el capítulo Apuestas deportivas: ésta es la línea que ha sido ajustada por las opiniones del mercado, en forma de apuestas, hasta llegar a línea que teóricamente es lo más cercano a la probabilidad real de ocurrencias en el partido. Y eso es exactamente lo que queremos medir: cómo se comparan las probabilidades implícitas en los momios contra la ocurrencia real de eventos. Se usan las cinco temporadas terminadas más recientes - desde la 2012-2013 hasta la 2016-2017 - de las cuatro principales ligas europeas; es decir, la English Premier League inglesa, la Serie A italiana, la Bundesliga alemana y La Liga española.

Queremos obtener probabilidades implícitas en cada una de las líneas de apuesta, por lo que necesitamos remover la comisión de la casa. El proceso se ilustra nuevamente con un ejemplo:

El juego celebrado el 26/11/2016, en donde el Liverpool recibió al Sunderland, tuvo una línea de cierre con momios decimales 1.19/9/15 para victoria local, empate y victoria visitante respectivamente. Para los tres eventos, calculamos la probabilidad implícita por la línea simplemente sacando el inverso multiplicativo. Esto resulta en probabilidades de ocurrencia 84.03 %/11.11 %/6.66 %. Éstas suman 101.8 %, pues incluyen la comisión de la casa.

Para remover la comisión, simplemente normalizamos el vector para que ahora sume 1, dividiendo entre la suma total de las probabilidades. Así, las probabilidades implícitas por la línea son 82.54 %/10.91 %/6.55 % respectivamente.

Una vez obtenidas las probabilidades implícitas de cada partido, las sumamos. Posteriormente, contamos cuántas veces ocurrió en realidad una victoria local, un empate y una victoria visitante en los mismos datos y los comparamos con las probabilidades implícitas.

En la figura 5.1 se muestra un resumen de la información. Al agregar los 7230 juegos en los datos, las líneas de cierre tienen probabilidades implícitas promedio de $(.4560, .2478, .2962)$ y probabilidades empíricas $(.4617, .2437, .2946)$. Aunque las probabilidades son muy cercanas, hay una muy pequeña distorsión en el mercado: los momios para el empate son medio punto porcentual más atractivos respecto a la probabilidad de ocurrencia; por el contrario, los

momios para los equipos locales son menos atractivos respecto a la realidad.

Liga	Juegos Totales	Implícitos L%	Reales L%	Implícitos E%	Reales E%	Implícitos V%	Reales V%
EPL	1900	45.0493%	45.3158%	24.9674%	24.7368%	29.9833%	29.9474%
Serie A	1900	45.2324%	45.7368%	25.8360%	25.3158%	28.9315%	28.9474%
La Liga	1900	47.0803%	47.5263%	23.8994%	23.2632%	29.0203%	29.2105%
Bundesliga	1530	44.9385%	46.0784%	24.3355%	24.1176%	30.7260%	29.8039%
Totales	7230	45.6077%	46.1687%	24.7813%	24.3707%	29.6110%	29.4606%

Figura 5.1: Resumen de la probabilidad implícita en los momios de cierre y la probabilidad empírica de ocurrencia para la victoria local, empate y victoria visitante en Inglaterra, Italia, España y Alemania

Aunque la distorsión cercana a 0.5 % existe, no es suficiente para ser redituable, pues la comisión promedio en las líneas de los datos es, aproximadamente, 2.06 %. Esto es consistente con lo encontrado por Deschamps, Gergaud (2007)[16], los cuales concluyen que no hay una estrategia redituable solamente basada en los momios, pero que, si se apostara ciegamente alguno de los tres posibles resultados en cada partido, el empate devuelve el mejor rendimiento (aproximadamente $-7\,\%$).

En resumen, el mercado de apuestas de futbol, al cierre, refleja casi perfectamente la realidad de los marcadores observados. A continuación, hacemos un análisis similar sobre la distribución Poisson y cómo refleja ésta los marcadores.

5.2. Comparación empírica de los marcadores en futbol a la distribución Poisson

Como hemos visto anteriormente, la literatura previa sobre modelación de futbol ha discutido extensamente si la distribución Poisson es un reflejo adecuado de lo que ocurre en partidos de futbol, o si se debe utilizar alguna otra distribución, como la Binomial Negativa.

En la figura 5.2 se muestran los histogramas para los goles locales, visitantes y la suma de ambos marcadores en los partidos de las últimas tres temporadas completas (2014-2017) de las cuatro principales ligas europeas; es decir, la English Premier League inglesa,

la Serie A italiana, la Bundesliga alemana y La Liga española. Sobrepuesto al histograma, se encuentra una distribución Poisson cuyo parámetro es la media muestral. Estas medias son, respectivamente: 1.581, 1.190 y 2.771.

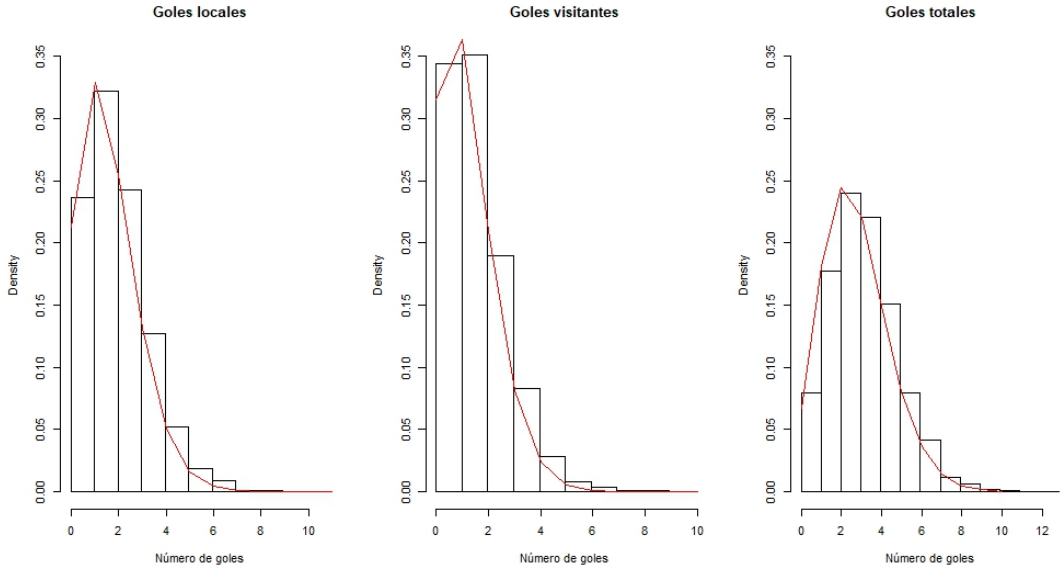

Figura 5.2: Histogramas para goles locales, visitantes y suma de marcadores; 2014-2017 en Inglaterra, España, Italia y Francia. Sobrepuesta una distribución Poisson cuyo parámetro es la media muestral (1.581, 1.190 y 2.771).

Para cada histograma tenemos 4116 observaciones. En las tres gráficas, el ajuste de la distribución es relativamente parecido al mostrado en los datos.

Maher (1982)[1] había notado entonces que utilizar dos distribuciones Poisson para modelar ambos marcadores independientemente, hacía que el modelo tuviera pequeños problemas de especificación. Maher observa que el modelo subestima el número de ocasiones en que se anotan 1 y 2 goles y sobrestima el número de veces en que se anotan 0 y 4 goles.

¡Sin embargo, en los histogramas podemos ver lo contrario!
Al modelar los goles locales y visitantes por separado, el modelo

está subestimando el número de ocasiones que se anotan 0 y sobrestimando el número de ocasiones en que se anotan 1 y 2 goles, sin sesgo alguno en 3 y 4.

Los motivos por los cuales esto sucede son inciertos. Podría ser que, tras 35 años, el futbol ha cambiado fundamentalmente. Como hemos discutido previamente, el sistema de puntuación cambió para recompensar al equipo ganador con 3 puntos en vez de 2, empezando por Inglaterra en 1981 (por lo que los datos usados por Maher son bajo el sistema de competencia previo) y adoptado por casi todas las ligas en 1995 [32].

En la figura 5.2, se puede observar que el sesgo sólo ocurre al modelar los goles locales / visitantes por separado: la distribución Poisson con media muestral como parámetro parece ajustar a la suma de los dos marcadores casi perfectamente.

Las figuras 5.3, 5.4 y 5.5 muestran los histogramas para goles locales, visitante y la suma de goles en las cuatro ligas por separado. Aunque los histogramas tienen cierta variabilidad natural dado que la muestra ha sido reducida, relativa a la de la figura 5.2 - por ejemplo, el modelo Poisson sobrestima la frecuencia con que el equipo visitante anota 1 gol en Inglaterra, mientras que la subestima un poco en España - no parece haber una diferencia sustancial entre los cuatro paises y, en los cuatro, el modelo Poisson hace un trabajo razonable modelando los marcadores. Además, aunque las medias muestrales son ligeramente distintas en cada país, la forma de la distribución es muy parecida, por lo que no tenemos ninguna razón para pensar que existen diferencias fundamentales de país a país.

En resumen, hemos visto que el mercado de apuestas en futbol refleja adecuadamente el resultado de los partidos en conjunto. Las líneas de cierre del mercado son lo más cercano que tenemos a probabilidades reales, por lo que son un buen termómetro para medir la eficiencia de predicción partido a partido de cualquier modelo. Además, hemos visto también que los marcadores parecen comportarse como una distribución Poisson, como había pensado Maher al investigar su comportamiento.

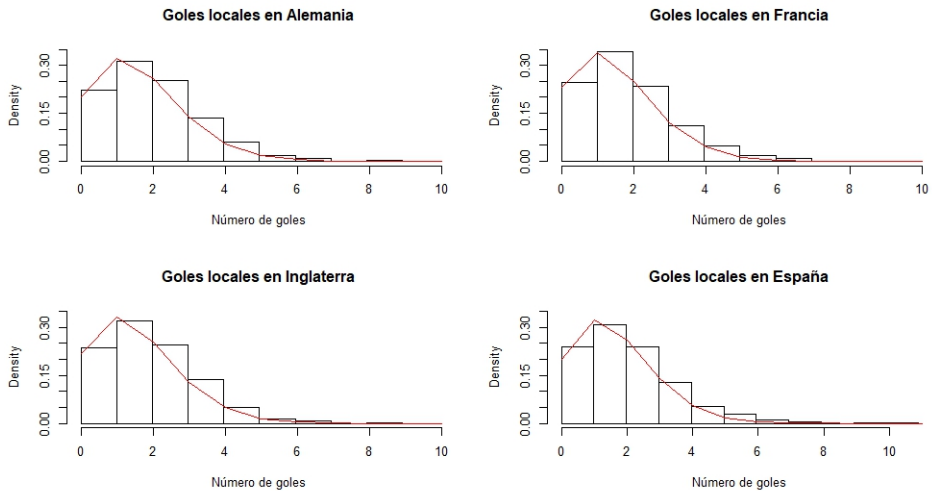

Figura 5.3: Histogramas para goles locales 2014-2017 en Francia, Alemania, Inglaterra y España. Sobrepuesta una distribución Poisson cuyo parámetro es la media muestral (1.603, 1.465, 1.521, 1.606)

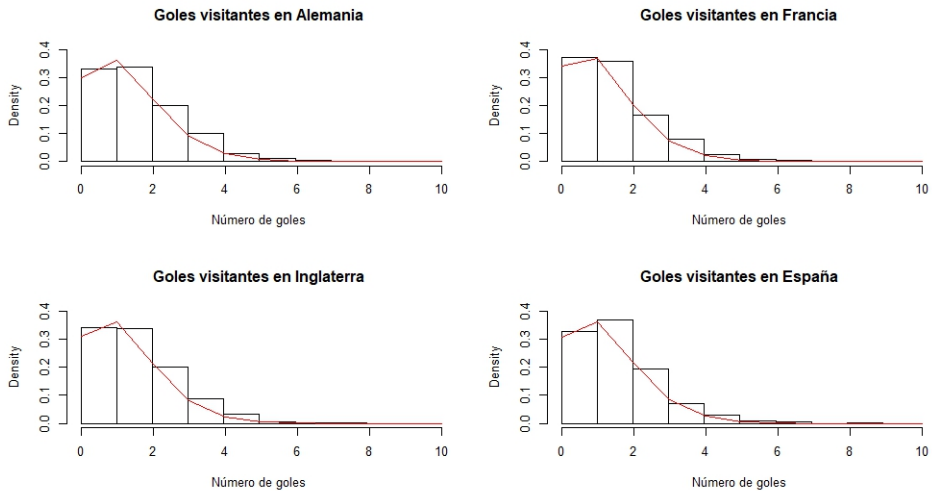

Figura 5.4: Histogramas para goles Visitantes 2014-2017 en Francia, Alemania, Inglaterra y España. Sobrepuesta una distribución Poisson cuyo parámetro es la media muestral (1.214, 1.080, 1.168, 1.175)

A continuación, mostramos cuáles son las debilidades del modelo de Maher y cómo es que los autores posteriores atacan el problema,

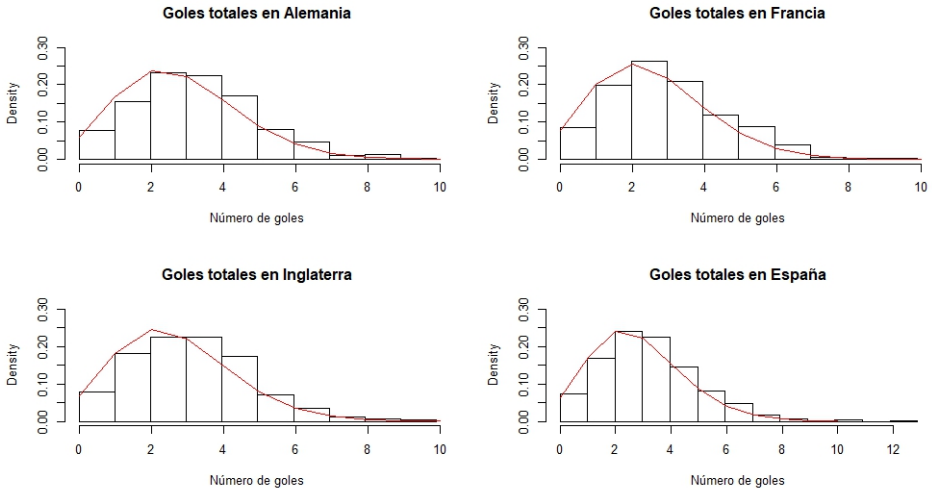

Figura 5.5: Histogramas para la suma de marcadores en un partido, 2014-2017, en Francia, Alemania, Inglaterra y España. Sobrepuesta una distribución Poisson cuyo parámetro es la media muestral (2.817, 2.545, 2.689, 2.781)

intentando mejorarlo y analizamos si logran su cometido satisfactoriamente.

Capítulo 6

Un análisis exploratorio de las alternativas al modelo Poisson bivariado en la literatura

Hasta ahora, hemos discutido algunas de las bondades del modelo original de Maher, que utiliza una distribución Poisson para cada marcador independiente del otro: es un modelo fácil de implementar, que sólo necesita los resultados de partidos previos para calibrar los parámetros máximo verosímiles. Además, los estimadores tienen una forma cerrada, sencilla y fácil de encontrar bajo un estimador inicial y un proceso iterativo.

Sin embargo, la propuesta original de Maher, que modela el marcador local y visitante en un partido como variables aleatorias independientes tiene dos problemas fundamentales:

1.- Maher mismo observa que, aunque la distribución Poisson hace un muy buen trabajo modelando los marcadores por separado, parece haber una pequeña subestimación en la probabilidad de empate en el partido, puesto que los marcadores tienen una covarianza positiva. Para ello, propone un modelo Poisson bivariado.

2.- El trabajo de Maher no resuelve el problema de obtener parámetros con información parcial de una temporada. Maher estaba intere-

sado en ver si la naturaleza del futbol era Poisson, y, por lo tanto, utiliza información de una temporada completa simplemente para ver si la estructura de los resultados asemeja a lo predicho por el modelo. Sin embargo, nunca intenta utilizar información de parte de la temporada para predecir partidos posteriores.

Estas limitaciones detonaron una larga línea de trabajos en modelación de partidos, intentando ajustar el modelo original, haciéndolo dinámico y ajustando las pequeñas diferencias que parecía haber entre el modelo y la realidad.

A continuación, exploramos cómo la literatura posterior a Maher (1982) intenta resolver estos dos problemas.

6.1. El modelo Poisson bivariado y la subestimación de empates

Después de ajustar el modelo Poisson doble, Maher se da cuenta que su modelo subestima la probabilidad de que los equipos anoten 1 o 2 goles, mientras que sobrestima la probabilidad de anotar 0 o 4+ goles, por lo que cree necesario usar una distribución distinta.

Para explorar la diferencia en los marcadores, define $Z_{i,j} = X_{i,j} - Y_{i,j}$, pues si $Z_{i,j} > 0$, ha gando el equipo local ($X_{i,j} > Y_{i,j}$); si $Z_{i,j} < 0$, ha gando el equipo visitante ($X_{i,j} < Y_{i,j}$). Maher nota que el modelo subestima sistemáticamente la probabilidad de empate ($Z_{i,j} = 0$).

Al explorar los datos, concluye que hay una correlación entre los dos marcadores $X_{i,j}$,$Y_{i,j}$ y sugiere como alternativa de modelar cada uno de los marcadores como Poisson independientes, un modelo Poisson bivariado.

Sean X_1, X_2, X_3 variables aleatorias Poisson con parámetro λ_1, λ_2, λ_3 respectivamente. Definimos $X = X_1 + X_3$, $Y = X_2 + X_3$. Entonces, decimos que el vector aleatorio (X, Y) tiene distribución Poisson Bivariada, BP$(\lambda_1, \lambda_2, \lambda_3)$, con función de densidad:

$$\mathbb{P}[X = x, Y = y] = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1}{k_1!} \frac{\lambda_2}{k_2!} \sum_{k=0}^{min(k_1,k_2)} \binom{k_1}{k} \binom{k_2}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k \tag{6.1}$$

Es decir, Maher decide modelar los goles locales ($X_{i,j}$) y visitantes ($Y_{i,j}$) para un partido donde se enfrentan el equipo local i y el equipo visitante j bajo un esquema Poisson bivariado, con:

$$X_{i,j} = X_1 + X_3 \qquad Y_{i,j} = X_2 + X_3 \tag{6.2}$$

$$X_1 \sim \text{Poisson}(\lambda_1 - \lambda_3) \quad X_2 \sim \text{Poisson}(\lambda_2 - \lambda_3) \quad X_3 \sim \text{Poisson}(\lambda_3) \tag{6.3}$$

$$\lambda_1 = \alpha_i \, \beta_j \quad \lambda_2 = \alpha_j \, \beta_i \, \hat{k}^2 \quad \lambda_3 = Cov(X_{i,j}, Y_{i,j}) = \rho \sqrt{\lambda_1 \, \lambda_2} \tag{6.4}$$

Maher prueba con distintos valores para la correlación ρ y encuentra que el más adecuado para resolver la subestimación de los empates es cercano a $\rho = 0.2$. Esto mejora el ajuste del modelo, bajo la estadística Z para bondad de ajuste de una χ^2 en los tres años de datos utilizados por el autor.

6.1.1. Modificaciones del modelo Poisson bivariado en la literatura

En esta sección, exploramos a detalle las modificaciones posteriores al modelo Poisson bivariado de Maher (1982) [1], que intentan mejorar el ajuste e introducir un enfoque dinámico en el tiempo.

▷ **Modelo de Dixon/Coles (1997)**

Tras la investigación de Maher (1982), Dixon/Coles (1997)[10] hacen un análisis granular de cómo se comporta el modelo de Maher en marcadores específicos. Su intención es validar o desacreditar la hipótesis del modelo original Poisson doble de independencia entre los marcadores local y visitante.

Hacen un análisis exploratorio de cómo se comparan los marcadores predichos por el modelo independiente con lo observado en la realidad en sus datos. Por ejemplo, para la victoria local 3-1, cuentan la frecuencia del par (3,1) en todos los marcadores. Después, en

todos los marcadores locales, cuentan la frecuencia de 3; en todos los marcadores visitantes cuentan la frecuencia de 1 y multiplican estas frecuencias. Si el supuesto de independencia fuera cierto, estas dos cantidades deberían ser muy parecidas.

Encuentran que, bajo la prueba del cociente y respecto a errores estándar Bootstrap, sólo hay cuatro marcadores con una diferencia significativa entre el producto de las distribuciones empíricas marginales y la distribución empírica conjunta: 0-0, 1-0, 1-1 y 0-1

Para ilustrar, en los datos utilizados por los autores, un equipo local anota 0 goles 22.1 % de los partidos y un equipo visitante anota 0 goles 33.4 %. Esto quiere decir que esperaríamos ver un empate 0-0 bajo total intependencia en 7.38 % de los partidos. Sin embargo, en los datos se presenta un empate 0-0 8.2 %, que dada la frecuencia de ocurrencia de un empate sin goles - los marcadores con pocos goles son los más comunes - es una diferencia significatvia en el cociente.

La figura 6.1 muestra un resumen de las diferencias encontradas por los autores en los cuatro marcadores. Parece entonces que la diferencia real de predicción en cuanto a marcadores, es que el empate está subestimado por aproximadamente 1.5 %.

Marcadores	Frecuencia esperada bajo independencia	Frecuencia Observada	Diferencia	Cociente (errores estándar bootstrap)
0 - 0	7.38%	8.20%	0.82%	1.115 (.00352)
1 - 0	11.02%	10.30%	-0.72%	.937 (.0243)
0 - 1	8.04%	7.40%	-0.64%	0.92 (.0287)
1 - 1	12.01%	12.70%	0.69%	1.057 (.02)

Figura 6.1: Diferencia en marcadores modelados y empíricos, encontrada por Dixon, Coles (1997)

La modificación de Dixon/Coles hace que los parámetros ya no tengan una forma cerrada. Los autores resuelven esto creando una

pseudo-máxima verosimilitud para su modificación del modelo (recordemos que modifican las probabilidades para los marcadores 0-0, 1-0, 0-1, 1-1; con una función discutida en el capítulo Apuestas deportivas) y tienen que utilizar una optimización numérica para encontrar todos los parámetros sobre ésta. Además del parámetro de la modificación, los autores tienen que estimar uno adicional para la función de pesos sobre los partidos, lo que complica aún más la ecuación y la obtención de los parámetros.

Es importante recordar ahora que Koopman, Lit (2015)[14] encuentran que su modelo de apuestas necesita un valor esperado mucho mayor a 0 para generar ganancia consistentemente y que 0 no esté en el intervalo de error. La figura 6.2 muestra lo encontrado por los autores, donde τ es la ventaja esperada en cada apuesta realizada por el modelo. Los autores encuentran que el modelo empieza a generar ganancias medias significativamente distintas a cero cuando tienen valores esperados por apuesta mayores a 12 %.

En una investigación posterior, Dixon, Pope(2004)[11] hacen un análisis del mercado de momios para los resultados local, empate y visitante. La figura 6.3 es el histograma de momios de una casa de apuesta y las probabilidades obtenidas por el modelo Dixon, Coles (1997) para esos partidos. Se puede observar que el rango de probabilidades implícitas para los momios de empate se concentra casi todo en $20\,\% - 40\,\%$.

La poca variabilidad de los momios de empate, hará muy difícil que el modelo encuentre una ventaja suficientemente grande - en el sentido de la τ de Koopman, Lit - para que el modelo indique apostar a un empate. Por esto, el posible error de subestimación 1.5 % sugerido por Dixon, Coles (1997) no hará mucha diferencia en términos reales de apuesta. Es decir, aunque al comparar las predicciones en conjunto, el modelo podría estar subestimando la probabilidad de empate, las probabilidades de ocucrrencia no generarán un cambio significativo al evaluar el modelo partido a partido.

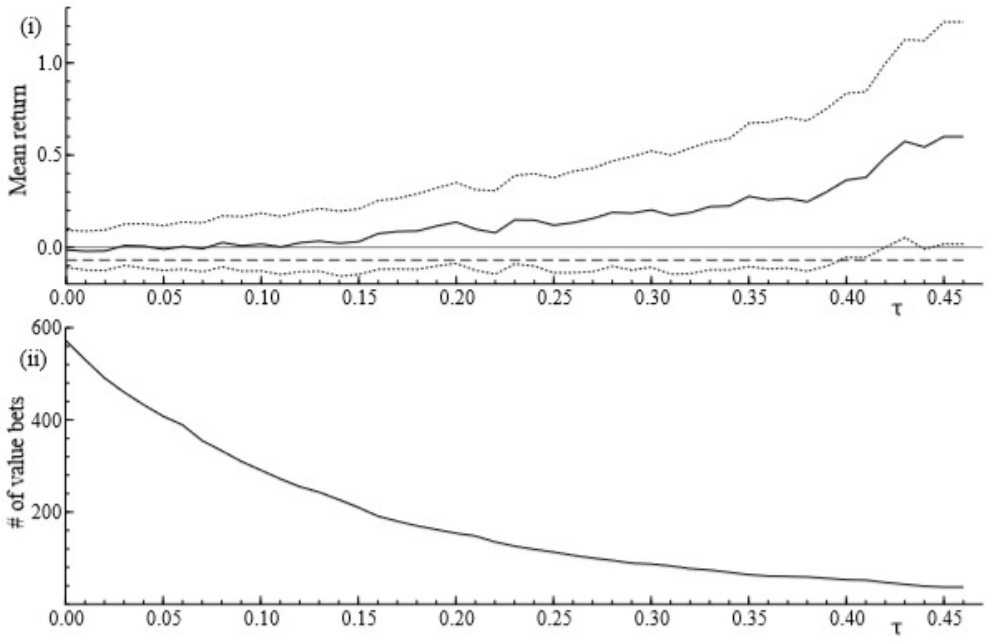

Figura 6.2: Rendimiento medio y número de apuestas bajo el modelo Koopman,Lit (2012) para distintos umbrales de ventaja τ.

▷ **Karlis, Ntzoufras (2003, 2005)**

Karlis, Ntzoufras (2003 y 2005)[3, 12] toman un acercamiento muy distinto a Dixon, Coles y proponen inflar el modelo Poisson bivariado por la diagonal para resolver el problema de la pequeña subestimación del empate en sus datos. Por inflar la diagonal, los autores buscan modificar con otra distribución los eventos en la distribución conjunta que devuelven el mismo valor, es decir, $\mathbb{P}(X = k, Y = k)$. En el futbol soccer, dichos eventos son empates en el partido. Nuevamente, los detalles se encuentran en el capítulo Apuestas deportivas.

Crean un paquete estadístico para la especificación del modelo y lo utilizan con datos de la Serie A italiana 1991 -1992. Como la temporada se jugó bajo el sistema anterior de competencia, los datos tienen una covarianza alta $\lambda_3 = Cov(X, Y) = 0.21$.

Encuentran que la distribución mezcla para inflar la diagonal es

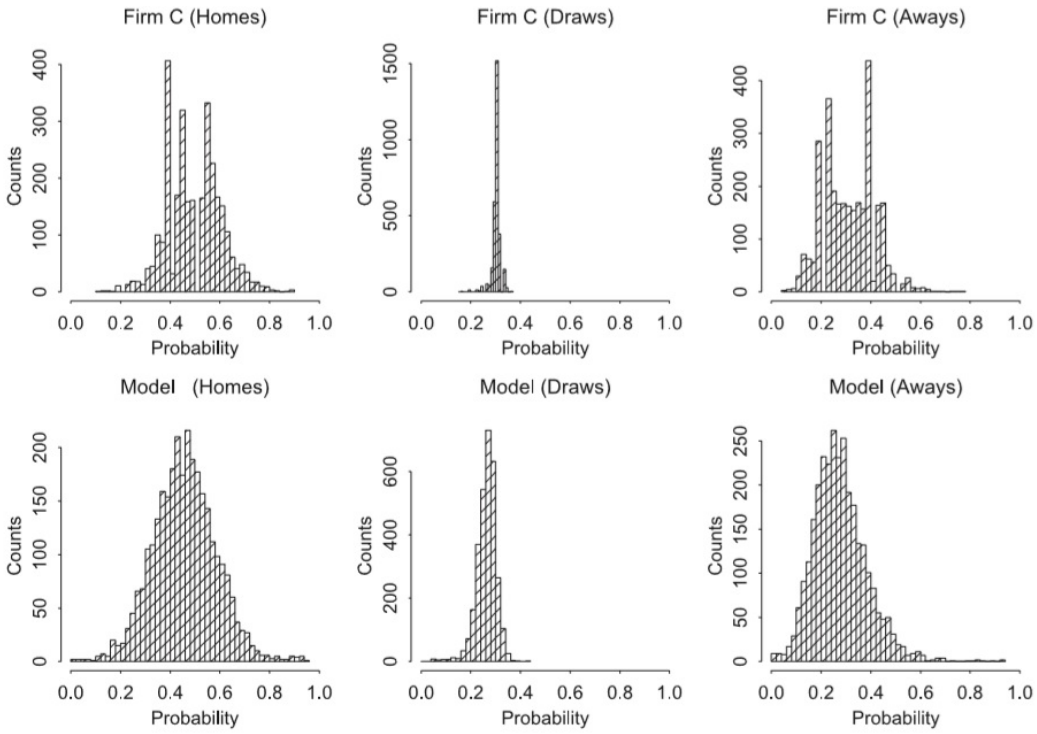

Figura 6.3: Histogramas para local, empate y visitante en Dixon, Pope (2004). Arriba, el histograma de los momios para una casa de apuesta. Abajo, el histograma de las probabilidades encontradas por el modelo de los autores.

una degenerada que sólo hace más probable que el equipo anote 1 gol, con un parámetro de mezcla pequeño (.09). Esto pues, bajo el sistema de competencia anterior, 1-1 fue un marcador muy frecuente en Italia ese año.

Como en Dixon, Coles; la modificación de Karlis, Ntzoufras tiene un costo sustancial al poder encontrar los parámetros, al grado que los autores tienen que diseñar un paquete en el lenguaje de programación R para poder optimizar bajo un Algoritmo esperanza-maximización y encontrar los parámetros necesarios.

Los problemas con las adaptaciones al modelo Poisson bivariado

Hemos mostrado que los problemas de estimación que observó Maher son opuestos en nuestros datos: el modelo sobrestima la ocurrencia de 0, mientras subestima la ocurrencia de 1 y 2 en ambos los marcadores local y visitante.

Por otro lado, en la figura 6.4 se muestran las covarianzas para las últimas tres temporadas de las cuatro ligas más importantes en el mundo. Las 12 temporadas tienen una covarianza negativa entre marcadores. Esto hace que el modelo Poisson bivariado se vuelva inutilizable, ya que no puede modelar covarianzas negativas: como $X_3 \sim$ Poisson (λ_3), el modelo está restringido a $\lambda_3 = Cov(X_{i,j}, Y_{i,j}) > 0$,

Temporada \ País	Alemania	España	Francia	Inglaterra
2014 - 2015	-0.01311475	-0.2167754	-0.01311475	-0.0411054
2015 - 2016	-0.1927676	-0.2894042	-0.1927676	-0.03133593
2016 - 2017	-0.0788171	-0.1168727	-0.0788171	-0.2110749

Figura 6.4: Covarianza para las primeras divisiones en las temporadas 2014-2017, ordenadas por país. Las doce covarianzas son negativas.

Esto podría coincidir nuevamente con el cambio del sistema de puntuación y los incentivos que tienen los equipos para ir por la victoria. Por ejemplo, en los 306 partidos de la temporada 1991-1992 en la Serie A italiana, 111 terminaron en empate o 36.27 %. En los 380 partidos de la temporada 2016-2017 en Italia, solo 80 terminaron en empate o 21.05 %. Por otro lado, el cambio pordría estar dado por la fuerza relativa de los equipos y la tendencia en el futbol moderno de tener sólo algunos equipos muy superiores al resto de la liga.

Cualquiera que sea la razón, la disminución en la ocurrencia de empates en el futbol genera una covarianza negativa en los datos y hace obsoleto al modelo Poisson bivariado.

Además, hemos visto que las modificaciones al modelo Poisson bivariado de Maher propuestas por Dixon, Coles y Karlis, Ntzoufras tienen costos significativos al encontrar los parámetros, pues éstos pierden su forma cerrada y simple. Más aún, ambas modificaciones parecen estar sobreajustados a casos particulares de los datos. Al intentar usar el algoritmo desarrollado por Karlis, Ntzoufras para ajustar los parámetros utilizando una de las temporadas recientes con covarianza negativa, por ejemplo, las iteraciones del algoritmo hacen que el parámetro $\lambda_3 = Cov(X, Y)$ se acerque cada vez más y más a cero, hasta llegar tan cerca del mismo como el cambio deseado especificado al algoritmo. Es decir, bajo covarianza negativa, el algoritmo mismo parece determinar que un modelo Poisson doble - que es un caso particular del modelo Poisson bivariado cuando $\lambda_3 = 0$ - es la mejor manera de atacar el problema.

En todo caso, las modificaciones no parecen aportar mucho en la modelación partido a partido y tienen un costo computacional significativo, lo cual nos da una pista que el modelo original de Maher, que modela cada uno de los marcadores como Poisson independientes, podría ser más adecuado que el Poisson bivariado.

En la siguiente sección, discutimos cómo distintos autores proponen modelos dinámicos en el tiempo, derivados del modelo Maher (1982), para que éste pueda funcionar durante toda la temporada.

6.2. Adaptaciones dinámicas al modelo Poisson bivariado en la literatura

Hasta ahora, hemos visto que el modelo de Maher parece tener un valor predectivo interesante en su forma original, pues los marcadores observados ajustan bien bajo una distribución Poisson. Además, hemos mostrado ya que el modelo Poisson bivariado propuesto en la literatura previa y sus modificaciones, no ofrecen una solución satisfactoria al problema. Primero, porque la modificaciones propuestas no tienen una mejora sustancial a la predictibilidad partido a partido, haciéndolas irrelevantes al medir el modelo contra momios de apuesta. Segundo, porque al modificar el modelo, la forma cerrada

y sencila de obtener de los parámetros se pierde. Y tercero y más importante, los marcadores de futbol en la actualidad presentan una covarianza negativa, misma que no puede ser modelada por un Poisson bivariado, pues éste está restringido por $\lambda_3 = Cov(X, Y) > 0$.

Hay un aspecto más en los trabajos derivados de Maher (1982) que no ha sido resuelto y es la dinámica del modelo. Maher obtiene parámetros utilizando la información de toda la temporada, pues su único objetivo era ver si la distribución que ajustaba mejor a los marcadores de futbol era la Poisson. En ningún momento Maher intenta utilizar la información disponible dentro de la temporada para predecir el resto, por lo que es un modelo estático.

A continuación, exploramos las adaptaciones dinámicas al modelo en la literatura:

Como hemos visto ya en la seccion 6.1.1 Dixon, Coles (1997)[10] y Karlis, Ntzoufras (2003, 2005)[3, 12] atacan el problema de la obtención de parámetros bajo sus modificaciones al modelo Poisson bivariado optimizando: Dixon, Coles usa una pseudo-verosimilitud, pues modifican la probabilidad de cuatro marcadores específicos y Karlis, Ntzoufras propone un modelo general lineal y un Algoritmo esperanza-maximización para encontrar los parámetros. El costo computacional de ambas soluciones es alto y parece estar totalmente especializados a casos particulares en los datos.

6.2.1. ▷ El modelo Poisson bajo parámetros autoregresivos: Koopman, Lit (2015)

El último estudio interesante que intenta hacer dinámico el modelo Poisson es Koopman, Lit (2015)[14], que intenta generar un modelo que utilice la Poisson bivariada y sea dinámico en el tiempo. Su solución es hacer que los parámetros varíen ligeramente en el tiempo, al hacerlos autoregresivos de primer orden con innovaciones normales. Los detalles del modelo están discutidos en el capítulo Modelación histórica de futbol soccer.

Para ayudar a la estimación, los autores suponen que los procesos son independientes y estacionarios, lo cual es difícil de justificar: los equipos están cambiando de jugadores constantemente, ya sea por

lesión o compra-venta, además de cambio de técnicos y formas de juego. Por ello, el proceso subyacente no es estático.
Además, introducir coeficientes autoregresivos e innovaciones al proceso, hace que los autores sólo puedan obtener sus parámetros a través de un proceso bayesiano y estimaciones Monte Carlo para la distribución posterior, lo que genera un costo computacional sustancial.

Es de resaltar que ambos coeficientes del proceso autoregresivo encontrados por los autores son casi uno ($\phi_\alpha = 0.9985, \phi_\beta = .9992$), lo cual sugiere una alta persistencia en las habilidades de los equipos a través del tiempo. Además, aunque el modelo propuesto por los autores tiene el menor valor en la función de pérdida para medir la eficiencia, el modelo con dos Poisson independientes devuelve resultados similares. De hecho, bajo la prueba estadística que realizan los autores para predecir fuera de la muestra, ¡no pueden rechazar la hipótesis nula de que el modelo Poisson doble sea tan efectivo como el Poisson bivariado!

Los mismos autores sugieren que el parámetro de dependencia en el modelo poisson bivariado $Cov(X, Y) = \lambda_3$ no tiene un impacto grande al predecir valores fuera de la muestra, a pesar de que tiene una fuerte significancia dentro de la muestra. Esto coincide con lo que hemos visto hasta ahora, donde la pequeña subestimación del modelo a los empates no parece ser de gran relevancia al predecir partidos individuales.

Todo lo anterior sugiere que una adaptación dinámica al modelo Poisson doble propuesto por Maher, que pudiera predecir con la información que tiene disponible, podría ser un modelo exitoso al explotar las líneas de apuesta. En el siguiente capítulo, exploramos esta modificación al modelo

Capítulo 7

El Modelo Poisson doble dinámico en el tiempo

La propuesta original de Maher (1982)[1] es modelar los goles anotados en un partido con dos distribuciones Poisson independientes para cada uno de los marcadores. A continuación, recordamos el modelo detalladamente.

En un partido donde el equipo i recibe al equipo j, los goles locales $X_{i,j}$ y los goles visitantes $Y_{i,j}$, que son independientes entre ellos, están dados por:

$$X_{i,j} \sim \text{Poisson}(\alpha_i \beta_j) \quad Y_{i,j} \sim \text{Poisson}(\alpha_j \beta_i k^2) \tag{7.1}$$

Los parámetros α_i, β_i están asociados al equipo local y modelan, respectivamente, su fuerza de ataque y defensa. De la misma manera, α_j, β_j modelan las habilidades de ataque y defensa del equipo visitante. De tal manera que, los goles anotados por el equipo local, dependen de qué tan bueno es su ataque (α_i) y qué tan buena es la defensa visitante (β_j). Análogamente, los goles visitantes dependen de qué tan buena es la defensa local (β_i) y de la habilidad del ataque visitante (α_j).

El parámetro restante, k^2, modela la ventaja que tiene un equipo al jugar en casa.

Bajo la condición de independencia entre marcadores, dado que la distribución subyacente de los mismos es Poisson, las formas máximo

verosímiles de los parámetros para cada equipo están dadas por:

$$\hat{\alpha}_i = \frac{\sum_{j \neq i}(x_{i,j} + y_{j,i})}{(1+\hat{k}^2)\sum_{j \neq i}\hat{\beta}_j} \qquad \hat{\beta}_j = \frac{\sum_{i \neq j}(x_{i,j} + y_{j,i})}{(1+\hat{k}^2)\sum_{i \neq j}\hat{\alpha}_i} \qquad \forall i, j \tag{7.2}$$

Los parámetros son sorprendentemente intuitivos. En el parámetro de ataque, $\sum_{j \neq i}(x_{i,j} + y_{j,i})$ son, simplemente, los goles anotados por el equipo en la temporada, tanto de local como de visitante. $\sum_{j \neq i}\hat{\beta}_j$ en el denominador compila la fuerza de las defensas de los otros equipos, ajustando la dificultad de los oponentes enfrentados para cada equipo. En el parámetro de defensa, $\sum_{i \neq j}(x_{i,j} + y_{j,i})$ son los goles permitidos por el equipo j en todos sus partidos, mientras que $\sum_{i \neq j}\hat{\alpha}_i$ ajusta por la fuerza de los ataques enfrentados.

El parámetro de localía está dado por:

$$\hat{k}^2 = \frac{\sum_i \sum_{j \neq i} y_{i,j}}{\sum_i \sum_{j \neq i} x_{i,j}} \tag{7.3}$$

$\hat{k}^2$ es un factor en el parámetro de la distribución para los goles visitantes, lo cual a simple vista parece raro. Sin embargo, su trabajo es reducir la intensidad con la que el equipo visitante anota goles. Intrínsecamente, no hay nada que restrinja a que $\hat{k}^2 < 1$. Sin embargo, el nominador en la expresión de $\hat{k}$, que es la suma de todos los goles anotados como visitante por todos los equipos en la liga es siempre menor que el denominador, que son todos los goles anotados como local por todos los equipos en la liga. Esta es, simplemente, una realidad del futbol: los equipos locales tienen una ventaja por jugar en casa que se refleja en el marcador.

De la forma de los estimadores, se sigue que:

$$\sum_i \sum_{j \neq i} \hat{\alpha}_i \hat{\beta}_j = \sum_i \sum_{j \neq i} x_{i,j} \qquad \sum_i \sum_{j \neq i} \hat{k}^2 \hat{\alpha}_j \hat{\beta}_i = \sum_i \sum_{j \neq i} y_{i,j} \tag{7.4}$$

Eso significa que la suma de las medias del modelo Poisson ajustado a los datos es igual al número de goles anotados observados totales.

Y es aquí donde radica el mayor problema del modelo de Maher. La magnitud de los parámetros depende enteramente de la cantidad de goles observados al momento de calibrarlos.

Aunque los parámetros reflejan la fuerza relativa de los equipos, no representan lo que pasa en un partido en cuanto a la cantidad de goles. La figura 7.1 muestra los parámetros que obtiene el modelo tras la primera mitad de la temporada de English Premier League 16-17. Es decir, al modelo le damos información solamente de las primeras 18 semanas. Es posible comprarar estos parámetros con los de la figura 7.2, que son los parámetros de la misma temporada, pero calibrados con todos los datos.

Al contrastar las figuras, se puede notar que ambos juegos de parámetros coinciden en muchas de las fuerzas relativas entre los equipos. Por supuesto, no en todo, pues media temporada de información es importante para los parámetros. Pero, por ejemplo, el modelo en ambos momentos coincide en cuáles son los mejores cinco ataques, aunque las magnitudes son totalmente desproporcionadas como causa del problema antes descrito. Al utilizar la información de toda la temporada, los mejores cinco ataques y sus respectivas α son: Tottenham (1.98), Chelsea (1.97), Manchester City (1.87), Liverpool (1.83) y Arsenal (1.81); con la información de media temporada, los mejores cinco ataques y sus respectivas α son: Liverpool (1.52), Chelsea (1.36), Arsenal (1.35), Man City (1.28), Tottenham (1.20).

Hay un par de cosas importantes escondidas aqui:

1.- Aunque el modelo no tenga la información de la temporada completa , está reflejando en buena medida las habilidades relativas de ataque y defensa de los equipos. Esto sugiere, que será útil en la predicción de resultados.

2.- Mientras que las comparaciones relativas están bien reflejadas

Equipo	Arsenal	Bournemouth	Burnley	Chelsea	Crystal Palace
Alpha Media Temporada	1.351894782	0.876053335	0.704130965	1.367209683	0.98443481
Beta Media Temporada	0.645959658	1.024081508	0.948165954	0.441639649	1.162143323
Equipo	Everton	Hull City	Leicester	Liverpool	Manchester City
Alpha Media Temporada	0.828515687	0.548451715	0.807629666	1.521148379	1.289067622
Beta Media Temporada	0.756707696	1.327401635	1.018820467	0.720159001	0.710470224
Equipo	Manchester United	Middlesbrough	Southampton	Stoke City	Sunderland
Alpha Media Temporada	0.95419305	0.562134404	0.628452035	0.742208986	0.576664366
Beta Media Temporada	0.630412102	0.713726962	0.716559967	1.049290445	1.136926693
Equipo	Swansea	Tottenham	Watford	West Brom	West Ham
Alpha Media Temporada	0.725766542	1.207621395	0.779749695	0.82983212	0.778339991
Beta Media Temporada	1.441733986	0.471912197	1.117784401	0.758393949	1.084905685

Figura 7.1: Parámetros tras 18 jornadas (media temporada) 2016 - 2017 de English Premier League bajo el modelo Poisson doble de Maher (1982)

Equipo	Arsenal	Bournemouth	Burnley	Chelsea	Crystal Palace
Alpha	1.813465028	1.323296597	0.925903978	1.977954919	1.196803204
Beta	1.074653254	1.601933543	1.293119768	0.810652768	1.496929841
Equipo	Everton	Hull City	Leicester	Liverpool	Manchester City
Alpha	1.457082765	0.900006447	1.148903809	1.831179678	1.872883824
Beta	1.056752163	1.87733662	1.493861422	1.025274419	0.954071175
Equipo	Manchester United	Middlesbrough	Southampton	Stoke City	Sunderland
Alpha	1.250818963	0.639552958	0.967273865	0.975083499	0.698231824
Beta	0.691160593	1.231853428	1.130934008	1.320057984	1.608255824
Equipo	Swansea	Tottenham	Watford	West Brom	West Ham
Alpha	1.08595183	1.989801159	0.963207079	1.018408783	1.127890304
Beta	1.658022399	0.639963668	1.602869084	1.204973301	1.519036716

Figura 7.2: Parámetros finales para la temporada 2016 - 2017 de English Premier League bajo el modelo Poisson doble de Maher (1982)

en los parámetros, la realidad de éstos al interactuar en un partido no. Los vectores de parámetros $\vec{\alpha}, \vec{\beta}$ calculados con la información de media temporada tienen media 0.9032 y 0.8939 respectivamente. Esto quiere decir que un partido modelado bajo estos parámetros, tendría, en promedio 1.41 goles. A comparación, los vectores de parámetros $\vec{\alpha}, \vec{\beta}$ calculados con la información de la temporada completa, tienen media 1.258 y 1.265 respectivamente. Bajo estos parámetros, un partido modelado tendría en promedio 2.78 goles. La temporada 2016 - 2017 tuvo en promedio 2.8 goles por partido. Es decir, modelar partidos con los parámetros de media temporada, devolvería, en promedio, partidos con la mitad de goles. Esta es una buena pista de cómo podría arreglarse el modelo.

En la siguiente sección, exploramos cómo se comporta el modelo de Maher con la información de toda la temporada para intentar probar que la estructura de los marcadores es, en efecto, Poisson.

7.1. El modelo de Maher bajo información perfecta

Tras el estudio de la sección anterior, surge naturalmente la pregunta: ¿Qué tan bueno es el modelo de Maher para predecir resultados en una temporada? ¿es la estructura de los marcadores Poisson o algo más?

Una manera de probar si la estructura subyacente de los marcadores es realmente Poisson, es dándole al modelo original de Maher (1982), que modela los marcadores como dos distribuciones Poisson independientes, una por marcador, información perfecta y comparándolo contra el mercado; que, ya sabemos, predice muy bien los marcadores conjuntamente. Es decir, al modelo Poisson doble le damos información de toda la temporada, misma que usamos para calibrar los parámetros. Posteriormente, usamos éstos para intentar predecir los marcadores de la misma temporada.

Si la estructura fuera realmente Poisson, al modelo debería irle bien prediciendo la temporada con la que fue calibrada. Si la estructura no se pareciera a la propuesta por el modelo, entonces, incluso utilizando toda la temporada como información, el modelo debería tener problemas generando una ganancia.

Hacemos la prueba con la temporada 2016 - 2017 de la English Premier League. La liga tiene 20 equipos y cada uno se enfrenta a todos los demás, una vez como local y una vez como visitante. Esto quiere decir que, en total, hay 380 juegos. Bajo el modelo de Maher, los estimadores máximos verosímiles toman la forma:

$$\hat{\alpha_i} = \frac{\sum_{j \neq i}(x_{i,j} + y_{j,i})}{(1+\hat{k}^2)\sum_{j \neq i}\hat{\beta}_j} \qquad \hat{\beta}_j = \frac{\sum_{i \neq j}(x_{i,j} + y_{j,i})}{(1+\hat{k}^2)\sum_{i \neq j}\hat{\alpha}_i} \qquad \forall i, j \tag{7.5}$$

El parámetro de localía está dado por:

$$\hat{k}^2 = \frac{\sum_i \sum_{j \neq i} y_{i,j}}{\sum_i \sum_{j \neq i} x_{i,j}} \tag{7.6}$$

Maher sugiere que se pueden usar estimadores iniciales de la forma:

$$\hat{\alpha_i} = \frac{\sum_{j \neq i} x_{i,j}}{\sqrt{S_x}} \qquad \hat{\beta}_j = \frac{\sum_{i \neq j} x_{i,j}}{\sqrt{S_x}} \qquad S_x = \sum_i \sum_{j \neq i} x_{i,j} \tag{7.7}$$

y calcular los estimadores máximo verosímiles iteradamente a partir de éstos, lo cual hacemos.

Los parámetros obtenidos utilizando la información de la temporada completa se observan en la figura 7.2.

Tras calibrar los parámetros de final de temporada, podemos obtener las probabilidades implícitas para cualquier partido. Esto, pues el modelo permite la especificación de la probabilidad para todos los posibles marcadores. Ilustramos con un ejemplo:

El 27/08/2016, Watford recibía a Arsenal, en un partido correspondiente a la jornada 3 de la temporada.
Como se puede ver en la figura 7.2 los parámetros alpha/beta de Arsenal son 1.813 y 1.074. Para Watford, los parámetros alpha/beta son 0.963 y 1.602. De esta manera, bajo el modelo de Maher, sabemos que, en un partido donde el equipo i recibe al equipo j, los goles locales $X_{i,j}$ y los goles visitantes $Y_{i,j}$ están dados por:

$$X_{i,j} \sim \text{Poisson}(\alpha_i \beta_j) \quad Y_{i,j} \sim \text{Poisson}(\alpha_j \beta_i k^2) \tag{7.8}$$

Como alfabéticamente Arsenal es el 1er equipo y Watford el 18vo:

$$X_{18,1} \sim \text{Poisson}(0.963 * 1.074) \quad Y_{18,1} \sim \text{Poisson}(1.813 * 1.602 * 0.752) \tag{7.9}$$

Por lo tanto:

$$X_{18,1} \sim \text{Poisson}(1.035) \quad Y_{18,1} \sim \text{Poisson}(2.188) \tag{7.10}$$

Ahora, como la probabilidad del marcador (x, y) está simplemente dada por el producto de $\mathbb{P}[X = x] * \mathbb{P}[Y = y]$, cuyas distribuciones ya conocemos, podemos calcular la probabilidad implícita por el modelo para cualquier marcador. En la figura 7.3 se muestra una matriz teórica para este partido. Los marcadores más comunes son 1-2, 0-2 y 1-1, en ese orden. Aunque en este trabajo sólo evaluamos dos posibles apuestas - el resultado del partido y si el partido tiene más o menos de 2.5 goles - la matriz de marcadores permite, sumando las probabilidades relevantes, obtener estimaciones para otras apuestas ofrecidas en el mercado. Por ejemplo, podríamos apostar que un equipo gana y el otro no anota, o apostar si hay al menos 5 goles en el partido. En toda la literatura previa, además, no se había evaluado el modelo contra apuestas de total de goles. Esta es otra gran ventaja del modelo Poisson doble.

Local \ Visitante	0	1	2	3	4	5	6	7	8	9	10
0	3.981%	8.713%	9.534%	6.955%	3.805%	1.665%	0.607%	0.190%	0.052%	0.013%	0.003%
1	4.121%	9.019%	9.869%	7.199%	3.939%	1.724%	0.629%	0.197%	0.054%	0.013%	0.003%
2	2.133%	4.668%	5.108%	3.726%	2.038%	0.892%	0.325%	0.102%	0.028%	0.007%	0.001%
3	0.736%	1.611%	1.762%	1.286%	0.703%	0.308%	0.112%	0.035%	0.010%	0.002%	0.001%
4	0.190%	0.417%	0.456%	0.333%	0.182%	0.080%	0.029%	0.009%	0.002%	0.001%	0.000%
5	0.039%	0.086%	0.094%	0.069%	0.038%	0.016%	0.006%	0.002%	0.001%	0.000%	0.000%
6	0.007%	0.015%	0.016%	0.012%	0.007%	0.003%	0.001%	0.000%	0.000%	0.000%	0.000%
7	0.001%	0.002%	0.002%	0.002%	0.001%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
8	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
9	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
10	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%

Figura 7.3: Matriz teórica de los marcadores modelados para el partido Watford - Arsenal del 27/08/2016. Los renglones corresponden al marcador local, las columnas al visitante.

Una vez obtenidas todas las probabilidades, sólo tenemos que sumar las probabilidades pertinentes para cada evento que nos interesa. Por ejemplo, el juego termina en victoria local si el equipo local anota más goles quel visitante. La probabilidad de dicho even-

to estará dada por la suma de todos los marcadores debajo de la diagonal en la matriz teórica.

Como podemos especificar las probabilidades para todos los marcadores, usamos el modelo para, además del resultado, modelar el total de goles en el partido. Como discutimos en el capítulo Apuestas deportivas, ésta es la segunda apuesta más popular en futbol soccer. Involucra predecir la suma de los marcadores. La línea más común y la que está disponible en nuestra base de datos es 2.5. De esta manera, si se apuestan las altas (*over 2.5*) y el juego tiene tres o más goles, la apuesta gana; si se apuestan las bajas (*under 2.5*) y el juego tiene dos o menos goles, la apuesta gana. Nuevamente, lo único que tenemos que hacer es sumar las probabilidades para todos los marcadores pertinentes.

El modelo devuelve entonces un vector de probabilidades estimadas por el modelo; las tres asociadas al resultado: victoria local, empate, victoria visitante; y las dos asociadas al total de goles: over 2.5, under 2.5. Las probabilidades calculadas para este partido son:

$$\mathbb{P}_{\text{estimadas}} = (\hat{\mathbb{P}_H}, \hat{\mathbb{P}_D}, \hat{\mathbb{P}_A}, \hat{\mathbb{P}_{O2.5}}, \hat{\mathbb{P}_{U2.5}}) = (0.1682, 0.1959, 0.63582; 0.6249, 0.3750)$$

Es importante notar que, teóricamente, la probabilidad de cualquier marcador positivo siempre es distinta de cero, por el dominio de la distribución Poisson. Sin embargo, aunque hemos limitado la matriz teórica de marcadores a un máximo de 10 por equipo, ambos vectores de probabilidad están a una distancia menor a 10^{-5} de 1. En efecto, los marcadores con muchos goles son muy raros. Como hemos podido ver en los histogramas de las figuras **??** y **??**, en las principales ligas europeas no ha habido un sólo equipo que haya anotado 11 goles en las últimas tres temporadas. Más aún, sólo un equipo ha anotado 10 goles en un partido: El 20/12/2015, el Real Madrid venció al Rayo Vallecano 10-2.

Tras calcular las probabilidades dadas por el modelo, comparamos éstas con los momios de apuesta en busca de una apuesta favorable. Todas las bases de datos utilizadas provienen de Football-Data.co.uk, que compila probabilidades de las casas de apuesta en archivos descargables.

Para comparar, utilizamos las líneas de apertura de Pinnacle para el resultado, que, como hemos discutido previamente, son las más ineficientes y vulnerables.

Para el total, utilizamos los momios promedio compilados por BetBrain, que están en la base de datos de FootballData. Esto pues FB no compila las líneas de Pinnacle para totales, sólo tienen disponibles las líneas promedio de BetBrain - que se calcula como la línea promedio entre todas las casas de las que compilan datos - o la línea máxima dentro de estas mismas casas.

Con el método descrito en el capítulo Apuestas deportivas, calculamos la probabilidad implícita en las líneas para el partido, la cual está dada por:

$$\mathbb{P}_{\text{implícitas}} = (\mathbb{P}_H, \mathbb{P}_D, \mathbb{P}_A, \mathbb{P}_{O2.5}, \mathbb{P}_{U2.5}) = (0.1828, 0.2557, 0.58139, 0.5235, 0.5347)$$

Recordamos aquí que los vectores de probabilidad de la casa siempre suman más de uno, pues incluyen la comisión que cobra la casa; en este caso, el vector de resultados suma 1.0199 y el vector del total de goles en el partido suma 1.0583. Es importante notar aquí que las líneas de Pinnacle suelen ser mucho más competitivas que la línea promedio del mercado. Por ello, la comisión para la línea del total es casi 6 %, mientras que la comisión para el resultado es menor a 2 %. En consecuencia, al correr el modelo más adelante a tiempo real, usaremos las líneas de BetBrain máximas y no las promedio para el total de goles.

Ahora podemos comparar los dos vectores para determinar dónde tenemos una ventaja:

$$\Delta = \mathbb{P}_{\text{estimadas}} - \mathbb{P}_{\text{implícitas}} = (-0.01458, -0.05982, 0.05443; 0.10141, -0.1597)$$

Por último, determinamos cuáles de estas apuestas queremos hacer. La regla más sencilla es que el modelo apueste si la ventaja que tiene es mayor a una constante τ. Recordemos que Koopman, Lit (2015) [14] encontraron que su modelo comenzaba a tener ganancias

significativamente mayores a 0 cuando las apuestas realizadas tenían un valor esperado mayor a 12 %. Más adelante hacemos un estudio exhaustivo en búsqueda de la τ que maximiza la ganancia.

Por ahora, como estamos usando información perfecta, podemos bajar un poco el umbral y decidimos apostar a cualquier apuesta tal que la ventaja $\Delta > 5\,\%$. Es decir, $\tau = 0.05$. Para este partido hay dos de estas apuestas: el modelo tiene una ventaja estimada de 5.43 % apostando al visitante Arsenal y una ventaja de 10.41 % apostando que hay 3 o más goles en el partido.

Una vez decididas las apuestas que el modelo hará, tenemos que determinar cuánto apostar. Para ello, utilizamos el Criterio de Kelly. Recordemos que el tamaño de apuesta bajo el Criterio de Kelly en un juego con pagos desiguales está dada por:

$$\frac{bp - q}{b}$$

donde b es la posible ganancia, p es la probabilidad de ganar; en este caso, estimada por el modelo y $q = 1 - p$ es la probabilidad de perder.

En las apuestas deportivas, b está dada por la línea de apuesta o momio, de la cual ya sacamos la probabilidad de quiebre o probabilidad implícita para comparar contra la probabilidad estimada para el modelo.

Los momios para las dos apuestas en que el modelo está interesado son 1.72 para la victoria visitante (Arsenal) y 1.91 para las altas. Recordemos que, a lo que la apuesta paga bajo momios decimales, hay que quitarle lo apostado originalmente para calcular la ganancia. De esta manera, las posibles ganancias buscadas son $b_{\text{away}} = .72$ y $b_{\text{over 2.5}} = .91$. Así, utilizando las probabilidades calculadas por el modelo $\hat{p}$, determinamos el tamaño de la apuesta para el modelo bajo el Criterio de Kelly:

$$\text{Apuesta}_{\text{ Away}} = \frac{(.72)(0.63582) - (1 - 0.63582)}{.72} = 0.1300$$

$$\text{Apuesta}_{\text{O2.5}} = \frac{(.91)(0.5235) - (1 - 0.5235)}{.91} = 0.2127$$

Esto quiere decir que, usando el Criterio de Kelly, deberíamos apostar 13 % y 21.27 % del capital disponible respectivamente. Como hemos discutido previamente, es conveniente suponer que la ventaja del jugador es menor a la estimada por el modelo. Aunque esto reduce parte de la ganancia potencial, logra combatir la incertidumbre asociada a especificaciones erróneas en el modelo y suaviza el proceso de ganancia para el modelo.

Por esta razón, en vez de apostar lo que sugiere el Criterio, apostamos una fracción. Como el umbral asociado a la selección de apuestas del modelo es muy bajo, el modelo tendrá muchas apuestas. Así, elegimos usar una fracción pequeña de Kelly, $\frac{1}{8}$. Las apuestas ajustadas se convierten en:

$$\text{Apuesta}_{\text{Away}} = \frac{1}{8}0.13 = 0.01625 \qquad \text{Apuesta}_{\text{O2.5}} \approx \frac{1}{8}0.2127 \approx .02660$$

Tras determinar el tamaño de las apuestas, el modelo revisa qué ha sucedido en el partido. En el partido del ejemplo, Arsenal dominó el juego desde temprano. Con un codazo en el área de Nordin Amrabat sobre Alexis Sánchez, Arsenal tomaría una ventaja 0-1 de penal antes de los 10 minutos. Con velocidad en las bandas, Arsenal logró generar las mejores oportunidades del partido, anotando dos veces más en los últimos 5 minutos del primer tiempo, en un juego que ganarían cómodamente con marcador final de 1-3, por lo que ambas apuestas han ganado.

$$\text{Rendimiento}_{\text{A}} = (0.01625)(1.72) = 0.02795 \quad \text{Gananacia}_{\text{A}} = 0.02795 - 0.01625 = 0.0117$$

$$\text{Rendimiento}_{\text{O2.5}} = (.02660)(1.91) = 0.0508 \quad \text{Ganancia}_{\text{O2.5}} = 0.0508 - .02660 = 0.0242$$

De esta manera, el modelo ha ganado 3.59 % del capital disponible en las apuestas.

Todo esto, por supuesto, lo hace un sistema automatizado creado por el autor, de tal forma que las operaciones completas toman fracciones de segundo. Esto permite correr el modelo en temporadas completas sin ninguna complicación.

A continuación, exploramos las implicaciones de utilizar el Criterio de Kelly para encontrar el valor de las apuestas.

Las bondades de Kelly bajo un modelo con ventaja

Las propiedades del Criterio de Kelly son fascinantes. Por un lado, apostar una fracción del capital disponible hace que, en sentido estricto, el jugador nunca se arruine, pues nunca lo estamos apostando todo y, por ello, la fortuna jamás podrá llegar a cero. Incluso bajo la definición alternativa de ruina - que el capital termine debajo de un umbral predeterminado - el Criterio de Kelly sigue siendo la mejor opción para el crecimiento del capital.

Pero, por otro lado, apostar una parte del capital disponible al momento de la apuesta hace que los efectos del modelo se amplifiquen enormemente. Si el modelo está perdiendo, las apuestas se reducirán de tamaño considerablemente para responder a los resultados. Pero si el modelo está ganando, apuesta por apuesta, el tamaño de las mismas sigue creciendo. Esto hace que haya un un efecto muy parecido al interés compuesto.

En las figura 7.4 se puede ver perfectamente la diferencia de este efecto. En ambas trayectorias, se grafica la riqueza del jugador con las apuestas sugeridas por el modelo para la English Premier League 2016-2017 con información perfecta. Las apuestas que se realizan son exactamente las mismas en ambas situaciones: el modelo sugiere realizar 366 apuestas en la temporada, a un promedio de casi una apuesta por partido, pues hay 380 partidos en la temporada. Recordemos aquí que en cada partido, el modelo tiene cinco posibles apuestas: local, empate, visitante, altas y bajas.

La diferencia es la siguiente: en la trayectoria de color azul de la figura 7.4, apostamos bajo el mismo Criterio que antes, usando $\frac{1}{8}$ del Criterio de Kelly para determinar las apuestas. Pero en vez

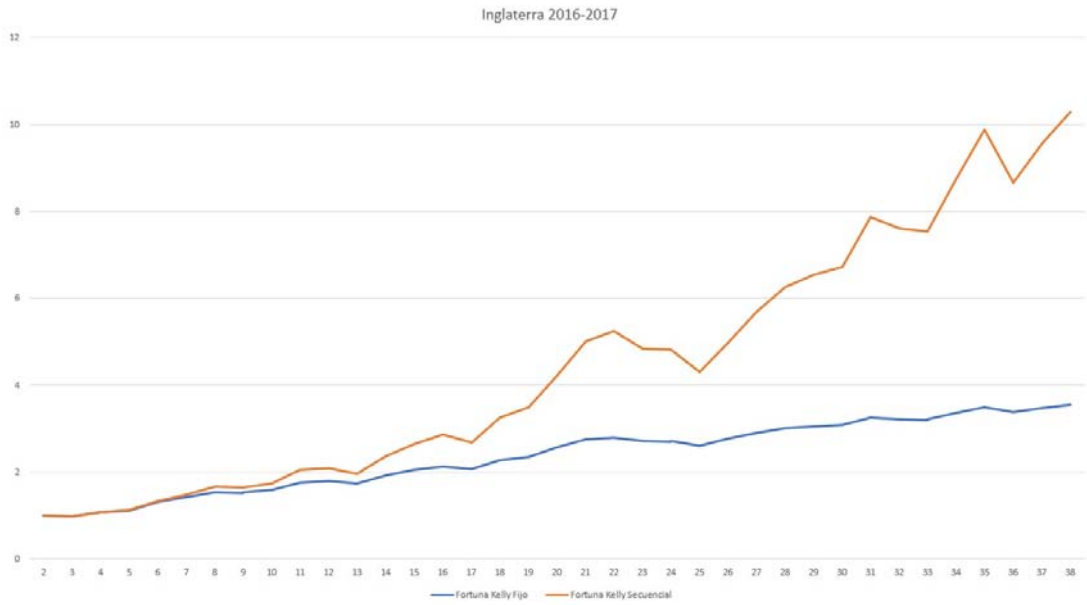

Figura 7.4: Ganancias del Modelo Poisson doble bajo información perfecta, umbral de apuesta $\tau = .05$, apostando bajo $\frac{1}{8}$ del Criterio de Kelly. En azul, utilizando el Criterio fijo a la fortuna inicial. En naranja, utilizando el Criterio sobre el capital disponible al apostar.

de apostar el porcentaje sugerido por el Criterio a partir del capital disponible al momento de la apuesta, apostamos el porcentaje sugerido de la riqueza original. En la trayectoria de color naranja, apostamos bajo el mismo $\frac{1}{8}$ del Criterio de Kelly, pero apostamos el porcentaje sugerido por el Criterio sobre el capital disponible al inicio de la jornada. De las 366 apuestas, el modelo gana 205 y pierde 161.

Un ejemplo podría ayudar a aclarar las diferencias. Supongamos que el jugador empieza con $\$100,000$. En la semana 12, el modelo sugiere apostar que habrá 3 o más goles en el partido en el que Tottenham recibe a West Ham. Tras hacer todos los cálculos, la fracción del Criterio de Kelly utilizado sugiere apostar 3.19 %. Bajo el primer régimen, apostaríamos 3.19 % del capital inicial, o $3190. Bajo el segundo régimen, apostaríamos 3.19 % del capital disponible al inicio de la jornada 12, es decir, el acumulado hasta el final de la jornada 11. Como el modelo había comenzado bien la temporada, el capital

disponible al final de la jornada 11 era 2.04061 veces el original, o $\$204,061$. Esto quiere decir que la apuesta bajo el segundo régimen de apuestas sería $6509.

Podemos ver que aunque las trayectorias en las figura 7.4 son parecidas, la escala de los efectos no. Bajo el régimen que apuesta sobre el capital inicial, el jugador termina con 3.55385 veces su capital inicial al final de la temporada, es decir, con $\$355,385$ para un crecimiento de $+255\,\%$. Bajo el régimen que apuesta sobre el capital disponible, el jugador termina con 10.28243 veces el capital inicial al final de la temporada, es decir, con $\$1,028,243$ para un crecimiento de $+928\,\%$.

Es importante recordar aquí que estamos bajo información perfecta, es decir, utilizando los resultados de los partidos que el modelo está intentando predecir para calibrar los parámetros. En una situación real, no conocemos dichos marcadores para calibrar los parámetros. Por ello, los resultados del modelo son irreales y estarán muy inflados. Sin embargo, el hecho de que el modelo esté teniendo tales resultados deja entrever que la estructura Poisson del modelo podría ser la correcta: si intentáramos predecir con una distribución que no se asemejara en absoluto a la realidad, aún con información perfecta, el modelo debería tener problemas prediciendo.

Además, la magnitud de los efectos amplificados del Modelo de Kelly secuencial - donde el dinero está creciendo casi diez veces utilizando $\frac{1}{8}$ del Criterio - sólo son posibles bajo un modelo con información perfecta; en una única temporada usual, será complicado que el modelo duplique sus fondos. Sin embargo, la simplicidad del modelo nos permite adaptarlo a muchas situaciones y ligas distintas, por lo que, aunque el crecimiento de una temporada podría no ser muy grande, el crecimiento conjunto del modelo será suficientemente bueno para que valga el tiempo invertido.

Estrictamente hablando, bajo el Criterio de Kelly, hay una modificación adicional que tendría que hacerse al tamaño de las apuestas cuando estas son simultáneas. Los detalles se pueden ver en MacLean, Thorp, Ziemba (2011) [29]. La idea detrás es que, al realizar apuestas simultáneas - como podrían ser partidos de futbol de una

misma jornada - el tamaño de la apuesta debe ajustarse para reducir la probabilidad de ruina. No obstante, bajo ambas reglas de apuesta, ya hemos reducido la apuesta original del criterio por un factor de $\frac{1}{8}$ para mitigar buena parte de la varianza asociada al proceso y, por lo tanto, a la probabilidad de ruina. Es por ello que no consideramos esta reducción adicional.

Por último, es importante notar que establecer un régimen de apuesta secuencial como el de la figura 7.4 no es tan simple, por diversas consideraciones. Primero, para los apostadores profesionales, es difícil subir el tamaño de sus apuestas, puesto que las casas no están dispuestas a recibir la cantidad de dinero que a éstos les gustaría apostar. Además, no hay una manera estándar de lidiar con ligas distintas, lo cual podría causar un problema. Por ejemplo, se podría usar una sola fortuna para todas las ligas en la que se esté apostando o partir la fortuna total en diferentes bancos, uno por liga. En el caso fragmentado, el crecimiento exponencial se verá limitado, pero la probabilidad de una racha desfavorable de consideración es todavía menor, puesto que ahora tendría que haber una mala racha en cada uno de los bancos. En el caso de un único banco, el crecimiento exponencial tiene rienda suelta pero, al mismo tiempo, pequeñas rachas malas en diferentes ligas se pueden ver fuertemente amplificadas. Discutimos a detalle un poco más adelante.

A continuación, exploramos qué pasaría si utilizáramos información de temporadas pasadas para predecir la nueva temporada.

7.2. El modelo Poisson doble bajo una ventana fija de información

Como hemos visto, aunque el modelo de Maher hace un buen trabajo en reflejar las fuerzas relativas entre equipos, necesitamos encontrar una solución a la magnitud de los parámetros.

Al analizar el comportamiento de los parámetros en la sección 6.2, hemos visto que, los parámetros de final de temporada coinciden con el promedio de goles por partido, como deberían. Una solución

natural sería siempre usar una temporada completa de datos entonces. Esto presenta un problema grande: para utlizar una temporada completa de datos, necesitaríamos involucrar datos del año anterior para predecir la nueva temporada. Es decir, una ventana de datos de tamaño finjo donde, al agregar la semana más reciente, el modelo deje de utilizar la semana más antigua, pero siempre utilizando una temporada completa de datos.

Desafortunadamente, la inmensa mayoría de las ligas en el mundo presenta cambios de equipos de temporada a temporada bajo distintos sistemas de ascenso y descenso. Por ejemplo, la English Premier League cambia tres equipos por temporada. La predictibilidad para los equipos cambiados es muy complicada bajo el modelo, pues se necesitaría usar datos del equipo pasado como datos del equipo nuevo para no alterar los parámetros de los otros equipos.

No parece totalmente insensato pensar que los equipos descendiendo y ascendiendo tienen características similares y podrían usarse los datos de los equipos como intercambiables. No obstante, aún al intentar emparejar los equipos ascendiendo y descendiendo de distintas maneras (por ejemplo, hemos intentado emparejar los equipos por posición de ascenso/descenso, por goles anotados y por goles permitidos), el valor predictivo del modelo es bajísimo. Esto, porque las habilidades de los equipos cambiando de división son muy variables: muchos de ellos reciben una inversión de capital fuerte y renuevan a buena parte de su plantilla.
La figura 7.5 usa diez años de equipos ascendidos a la primer división de Inglaterra. Grafica los goles a favor/contra anotados en la división inferior contra los goles a favor/contra anotados en la primera división y una posible relación lineal entre los mismos, con y sin intercepto. Aunque éstos parecen estar débilmente relacionados, la variabilidad es tal que, para predecir a nivel partido por partido dentro del modelo Poisson, la relación no sirve en absoluto.

Por lo anterior, nos vemos obligados a utilizar solamente datos de la presente temporada para predecir partidos futuros. Con ellos queremos lograr encontrar parámetros cuya magnitud refleje correctamente la realidad de goles anotados partido a partido. A continuación, presentamos nuestra propuesta para hacer dinámico el modelo de Maher.

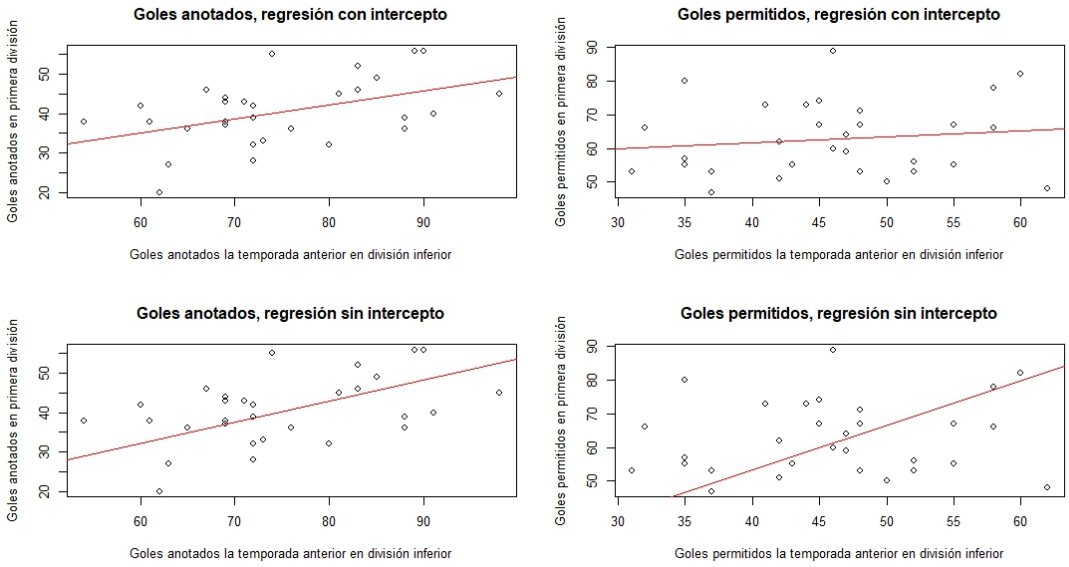

Figura 7.5: Dispersión de goles anotados y recibidos para equipos ascendiendo a English Premier League 2007 - 2017.

7.3. El modelo ∇-Poisson doble: una modificación dinámica

Como hemos visto en la introducción del capítulo 7, la forma de los parámetros máximo verosímiles hace que estos estén restringidos por las ecuaciones:

$$\sum_i \sum_{j \neq i} \hat{\alpha}_i \hat{\beta}_j = \sum_i \sum_{j \neq i} x_{i,j} \qquad \sum_i \sum_{j \neq i} \hat{k}^2 \hat{\alpha}_j \hat{\beta}_i = \sum_i \sum_{j \neq i} y_{i,j} \quad (7.11)$$

Eso significa que la suma de las medias del modelo Poisson ajustado a los datos es igual al número de goles anotados observados totales.

Nuevamente, la magnitud de los parámetros depende enteramente de la cantidad de goles observados al momento de calibrarlos. En

ambas igualdades, al obtener parámetros antes de que la temporada termine, muchas de las observaciones $x_{i,j}, y_{i,j}$ - los marcadores local y visitante, cuando se enfrentan el equipo local i al equipo visitante j - no existen, simplemente porque el partido i, j no se ha jugado. Para la suma, estos marcadores son tratados como ceros. Es decir, el modelo está calibrando los parámetros con información parcial, como si todos los marcadores para los partidos que aún no se juegan fueran $0 - 0$. Esto genera que, aunque los parámetros reflejan la fuerza relativa de los equipos, no representan lo que pasa en un partido en cuanto a la cantidad de goles.

Buscamos entonces una manera de transformar los vectores de parámetros máximos verosímiles obtenidos bajo la estructura del modelo de Maher, $\vec{\alpha}$ y $\vec{\beta}$, para que éstos puedan modelar no sólo la fuerza relativa entre los equipos, sino el resultado del partido y número de goles correctamente.

Utilizando las ecuaciones (7.11), se puede ver que:

$$\frac{\sum_i \sum_{j \neq i} x_{i,j}}{N(N-1)} = \frac{\sum_i \alpha_i}{N} \frac{\sum_{j \neq i} \beta_j}{N-1} \qquad \frac{\sum_j \sum_{i \neq j} y_{j,i}}{N(N-1)} = \frac{\sum_j \beta_j}{N} \frac{\sum_{i \neq j} \alpha_i}{N-1} k^2 \tag{7.12}$$

donde N es el número de equipos en la liga. $N(N-1)$ será entonces el número total de partidos en una temporada para dicha liga, pues un equipo no se puede enfrentar a si mismo. Por ejemplo, la English Premier League tiene 20 equipos, por lo que hay 380 partidos en una temporada. Por lo tanto, cada equipo juega $2(N-1)$ juegos, que es el número de jornadas en la liga. La English Premier League tiene $2\,(20-1) = 38$ jornadas.

De esta manera:

$$\frac{S_{xy}}{N(N-1)} = \frac{\sum_i \alpha_i \sum_{j \neq i} \beta_j (1 + k^2)}{N(N-1)} \qquad S_{xy} = \sum_i \sum_{j \neq i} x_{i,j} + y_{i,j} \tag{7.13}$$

donde S_{xy} representa el total de goles que se anota en la temporada. Es decir, $\frac{S_{xy}}{N(N-1)}$ representa el número de goles promedio

anotados por ambos equipos en un partido.

De tal manera que necesitamos encontrar una forma de estimar $\frac{S_{xy}}{N(N-1)}$ cuando no tenemos el total de goles de la temporada. Sea S^f_{xy} el número de goles anotado cuando se han jugado f jornadas. Notemos entonces que:

$$\frac{S_{xy}}{N(N-1)} \approx \frac{S^f_{xy}}{\frac{N}{2}f} \tag{7.14}$$

Primero, $S^f_{xy} \to S_{xy}$ cuando $f \to 2(N-1)$. Es decir, el número parcial de goles se convierte en el número total de goles cuando se han jugado todas las jornadas.

Segundo, $\frac{N}{2}f$ es el número de partidos se han jugado hasta la jornada f, pues cada equipo juega en cada jornada contra otro de la liga. Por ello, $\frac{N}{2}$: de otra manera, estaríamos contando dos veces los partidos. Así, $\frac{N}{2}f \to N(N-1)$ cuando $f \to 2(N-1)$, que es la última jornada.

Entonces, el promedio de goles anotados hasta la jornada f es un estimador consistente para el promedio de goles anotado durante toda la temporada, pues los dos coinciden cuando hay información perfecta.

Por ello, sucede que:

$$\frac{2S^f_{xy}}{Nf} \approx \frac{\sum_i \alpha_i \sum_{j\neq i} \beta_j(1+k^2)}{N(N-1)} \tag{7.15}$$

Y, por lo tanto:

$$\frac{2S^f_{xy}(N-1)}{f} \approx \sum_i \sum_{j\neq i} \alpha_i\beta_j(1+k^2) \tag{7.16}$$

Recordemos que, sin importar el momento en que se calculan los estimadores, éstos estan sujetos a las restricciones:

$$\sum_i \sum_{j\neq i} \hat{\alpha}_i\hat{\beta}_j = \sum_i \sum_{j\neq i} x_{i,j} \qquad \sum_i \sum_{j\neq i} \hat{k}^2\hat{\alpha}_j\hat{\beta}_i = \sum_i \sum_{j\neq i} y_{i,j}$$

De esta manera, definimos al ponderador dinámico ν como:

$$\nu = \frac{2(N-1)}{f} \tag{7.17}$$

En la ecuación 7.16, ajustar el total de goles parciales S_{xy}^{f} por ν hace que éstos se parezcan a los estimadores máximos verosímiles del final de la temporada. Por supuesto, no serán iguales: los partidos que aún no se han jugado contienen información de la habilidad de los equipos. Pero ya hemos visto que los parámetros calculados con información parcial de la temporada reflejan alguna parte de las fuerzas relativas entre equipos y que, entonces, podemos usar ésta para modelar el futuro. Así, utilizamos ν para poder modelar correctamente los partidos con información parcial.

Intuitivamente, ν tiene todo el sentido del mundo: $2(N-1)$ es el número de jornadas que tiene una temporada y f es el número de jornadas que ya ha transcurrido. De tal manera que, $\frac{1}{\nu} = \frac{f}{2(N-1)}$ representa la proporción de la temporada que ya ha transcurrido. Por lo tanto, al multiplicar por $\nu = \frac{2(N-1)}{f}$ estamos transformando los vectores a la magnitud correcta, como si reflejaran la información de una temporada entera, bajo las restricciones de la ecuación 7.3. Lo hacemos de la siguiente manera:

Sea $X_{i,j}, Y_{i,j}$ los marcadores local y visitante respectivamente cuando se enfrentan el equipo local i y el equipo local j. Sean $\alpha_i, \alpha_j \in \vec{\alpha}$ las fuerzas de ataque del equipo local y visitante; $\beta_i, \beta_j \in \vec{\beta}$ las fuerzas de la defensa y k^2 el parámetro de localía, obtenidos bajo las ecuaciones máximo verosímiles del modelo Poisson doble. Entonces sabemos que:

$$X_{i,j} \sim \text{Poisson}(\gamma_{H;i,j} = \alpha_i \beta_j) \qquad Y_{i,j} \sim \text{Poisson}(\gamma_{A;i,j} = \alpha_j \beta_i k^2) \tag{7.18}$$

Pero, como hemos visto antes, éstos parámetros no reflejarán la magnitud de goles anotados por partido correctamente. Por ello, los modificamos con ν para hacer dinámico al modelo:

$$X_{i,j} \sim \text{Poisson}(\gamma_{\nabla H;i,j} = \nu\alpha_i \beta_j) \quad Y_{i,j} \sim \text{Poisson}(\gamma_{\nabla A;i,j} = \nu\alpha_j \beta_i k^2) \tag{7.19}$$

La idea detrás de ν y los parámetros $\gamma_{\nabla H;i,j}, \gamma_{\nabla A;i,j}$ es que, sin importar la magnitud de los vectores $\vec{\alpha}$ y $\vec{\beta}$, los marcadores modelados con $\gamma_{\nabla H;i,j}, \gamma_{\nabla A;i,j}$ tendrán la magnitud correcta, correspondiente a un juego de la temporada en curso. La bondad adicional de ν es que no importa cuánta información tenga el modelo al momento de la calibración de los parámetros, el ajuste siempre devuelve partidos cuyas anotaciones estarán cercanas en conjunto al promedio de goles de la liga, pues la magnitud de los vectores $\gamma_{\nabla H;i,j}, \gamma_{\nabla A;i,j}$ ahora refleja el número de goles promedio por partido en la liga, mismo que se estabiliza rápidamente: el número de goles anotado por jornada no cambia drásticamente a través de una temporada.

Así, en cualquier punto, usamos toda la información de la misma temporada hasta ese punto para poder calibrar el modelo. Por ejemplo, si quisiéramos predecir la jornada 19, utilizamos los resultados de las primeras 18 jornadas para obtener los parámetros máximo verosímiles. El modelo modificado mantiene el proceso de obtención de parámetros del modelo original de Maher, que, ya vimos, hacen buen trabajo capturando las fuerzas relativas de los equipos. Tras obtener éstos, determinamos los parámetros para modelar los goles de cada partido $\gamma_{H;i,j} = \alpha_i \beta_j$ y $\gamma_{A;i,j} = \alpha_j \beta_i k^2$.

Multiplicamos éstos por ν para obtener $\gamma_{\nabla H;i,j} = \nu\gamma_{H;i,j}$, $\gamma_{\nabla A;i,j} = \nu\gamma_{A;i,j}$. Con estos últimos, obtenemos las probabilidades buscadas y determinamos las apuestas sobre todos los partidos de la jornada. El proceso se explica a detalle en el siguiente capítulo.

Capítulo 8

Una aplicación del modelo ∇-Poisson doble con información imperfecta al mercado de apuestas

A continuación, examinamos el funcionamiento del modelo ∇-Poisson doble a detalle: cómo obtiene las apuestas que realizará y el tamaño de la apuesta. Después, lo aplicamos al mercado de apuesta para ver si puede generar una ganancia consistentemente.

8.1. Un resumen del funcionamiento ∇-Poisson doble

Paso a paso, el proceso del modelo para una jornada cualquiera sigue el siguiente prodecimiento:

1.- El modelo obtiene los parámetros máximo verosímiles bajo la estructura Poisson doble de Maher (1982)[1], que modela ambos marcadores bajo distribuciones Poisson independientes. Utilizamos los estimadores iniciales y el proceso iterativo propuesto por Maher (1982), discutidos en el capítulo Modelación histórica de futbol soccer para encontrarlos.

2.- Tras encontrar los parámetros, hacemos el ajuste discutido

a detalle en la sección 7.3, para que los parámetros reflejen más adecuadamente la realidad de un partido.

3.- Una vez que el modelo tiene los parámetros ajustados, obtiene un vector de probabilidades $(\hat{\mathbb{P}_H}, \hat{\mathbb{P}_D}, \hat{\mathbb{P}_A}, \hat{\mathbb{P}_{O2.5}}, \hat{\mathbb{P}_{U2.5}})$ bajo la estructura Poisson doble - que modela independientemente ambos marcadores - para todos los partidos de la jornada.

4.- Utilizando las líneas de apuesta disponibles para la jornada, el modelo extrae un vector de probabilidades implícitas en los momios para las cinco apuestas posibles en cada partido $(\mathbb{P}_H, \mathbb{P}_D, \mathbb{P}_A, \mathbb{P}_{O2.5}, \mathbb{P}_{U2.5})$

5.- El modelo resta el vector de probabilidades calculado y el vector implícito en los momios para obtener Δ. Posteriormente, si una posible apuesta Δ_i tiene un valor esperado mayor a una ventaja predeterminada τ, se apuesta en dicho evento. De lo contratrio, el modelo no apuesta. Es decir, si $\hat{P}_i - P_i > \tau$, el modelo realiza esa apuesta.

6.- Tras determinar en qué apostará, el modelo obtiene la posible ganancia b de los momios y la probabilidad de ganar estimada $\hat{p}$, para determinar el tamaño de la apuesta bajo el Criterio de Kelly.

7.- El modelo ajusta la apuesta a una fracción del Criterio de Kelly predeterminada. En todos los casos de esta tesis, hemos usado $\frac{1}{8}$ de Kelly.

8.- El modelo consulta los resultados de los partidos en los que ha apostado, para determinar si las apuestas han ganado o perdido. Si han ganado, multiplica lo apostado por el momio para calcular el rendimiento de la apuesta y agregarlo a la cuenta. Como estamos haciendo backtesting, el modelo puede consultar los resultados del partido inmediatamente después de realizar las apuestas. Sin embargo, no conoce éstos al momento de calibrar los parámetros.

9.- El modelo calcula el récord de ganados y perdidos, además de la ganancia de la semana, restando al rendimiento lo originalmente apostado y lo añade a una base de datos con todas las apuestas realizadas y el total de los resultados jornada a jornada.

8.1.1. Consideraciones adicionales del modelo

Una vez que tenemos la estructura del modelo perfectamente determinada, sólo nos falta determinar dos cosas importantes:

Primero, necesitamos encontrar el valor de τ, el umbral de ventaja sobre el cual el modelo decide apostar. Queremos encontrar τ que maximice la ganancia del modelo. Cuando corrimos el modelo con información pefecta, utilizamos un umbral pequeño haciendo que $\tau = 0.05$. Sin embargo, sabemos de otros modelos similares que, al usar información incompleta, el umbral bajo el cual el modelo genera ganancias consistentes es cercano a $\tau = 0.12$.

Segundo, tenemos que determinar cuánta información necesita el modelo para empezar a predecir con precisión. Es decir, ¿cuántas semanas debemos utilizar al menos para calibrar el modelo y poder apostar? Una vez determinada la cantidad de información mínima, el modelo puede apostar en todas las jornadas subsecuentes, ganando una jornada de información en cada ronda de apuestas.

Por último, la simplicidad del modelo a utilizar nos da una ventaja adicional: como lo único que necesitamos de información para obtener parámetros es marcadores, el modelo se puede adaptar a cualquier liga en el mundo. Por lo mismo, evitamos los tres mercados grandes de apuesta: las primeras divisiones de Inglaterra, España e Italia. En estas ligas, los límites de apuesta son muy grandes, por lo que atraen a los principales apostadores profesionales. Ésto a su vez hace que las líneas sean muy eficientes; incluso las líneas de apertura, pues al poder apostar grandes cantidades - del orden de miles de dólares por apuesta - cualquier pequeña ventaja es significativa. Dada la adaptabilidad del modelo y la capacidad económica de un estudiante terminando su licenciatura, no hay motivo alguno por el cual competir contra las líneas más eficientes del mercado.

En vez, utilizamos estas cinco ligas para probar el modelo: la Bundesliga alemana, la Ligue 1 francesa y las segundas divisiones de los tres países antes mencionados; La Segunda en España, la Serie B en Italia y la English Championship en Inglaterra.

Esto nos permite mostrar un aspecto adicional de la adaptabi-

lidad del modelo: cada liga tiene distinto números de equipos: La Bundesliga tiene 18 equipos (306 partidos por temporada), la Ligue 1 tiene 20 equipos (380 equipos por temporada), La Segunda y Serie B tienen 22 equipos (462 partidos por temporada) y la English Championship tiene 24 equipos (526 partidos por temporada). Correremos el modelo en tres temporadas por país: 2014 - 2015, 2015 - 2016 y 2016 - 2017. Las bases de datos utilizadas - que incluyen líneas de apuesta y los resultados de los partidos - fueron obtenidas de Football-Data.co.uk; para los resultados utilizamos las líneas de apuesta de apertura de Pinnacle, para los totales los máximos de BetBrain. En los pocos casos donde no hay una línea de apertura en los datos, usamos la línea de cierre. En los todavía menos casos donde no hay línea de Pinnacle disponible, utilizamos la línea de 5Dimes - que ofrece líneas muy parecidas a Pinnacle - tomada del sitio OddsPortal.com para dichos partidos.

8.2. El umbral de apuestas τ para el modelo ∇-Poisson doble

A pesar de que tenemos una buena idea de dónde se encuentra el umbral ideal de apuestas por la exploración de Koopman,Lit ($\tau = .12$), queremos buscar el umbral que maximice la ganancia bajo la estructura de nuestro modelo específico. Para ello, corremos el modelo las tres temporadas de datos para las cinco ligas utilizadas. Como aún no sabemos cuál es la semana ideal para empezar a utilizar el modelo, tenemos que elegir una arbitrariamente. Sabemos que el modelo necesita alguna información para calibrar el modelo, pues las habilidades de cada equipo no se reflejan inmediatamente. En el análisis de la sección 6.2, vimos que los parámetros de media temporada en la English Premier League (es decir, tras 18 jornadas) reflejan relativamente bien la información al final de la temporada. Así, decidimos empezar a correr el modelo un poco antes, tras 15 jornadas de información. Es decir, haremos que el modelo apueste de la jornada 16 en adelante para las tres temporadas de datos en las cinco ligas elegidas. Para encontrar el umbral τ ideal, haremos que el modelo apueste a todas las apuestas donde tenga una ventaja mayor a 0.05 y luego filtraremos sobre éstas para encontrar el

umbral τ ideal.

Un resumen de la información encontrada tras correr el modelo se puede ver en las figuras 8.1, 8.2 y 8.3 que muestran el número de apuestas, la ganancia por apuesta y la ganancia total bajo distinos valores de τ. Al aumentar la ventaja necesaria para apostar, lógicamente, el número de apuestas que el modelo realiza disminuye, como se puede ver claramente en la figura 8.1.

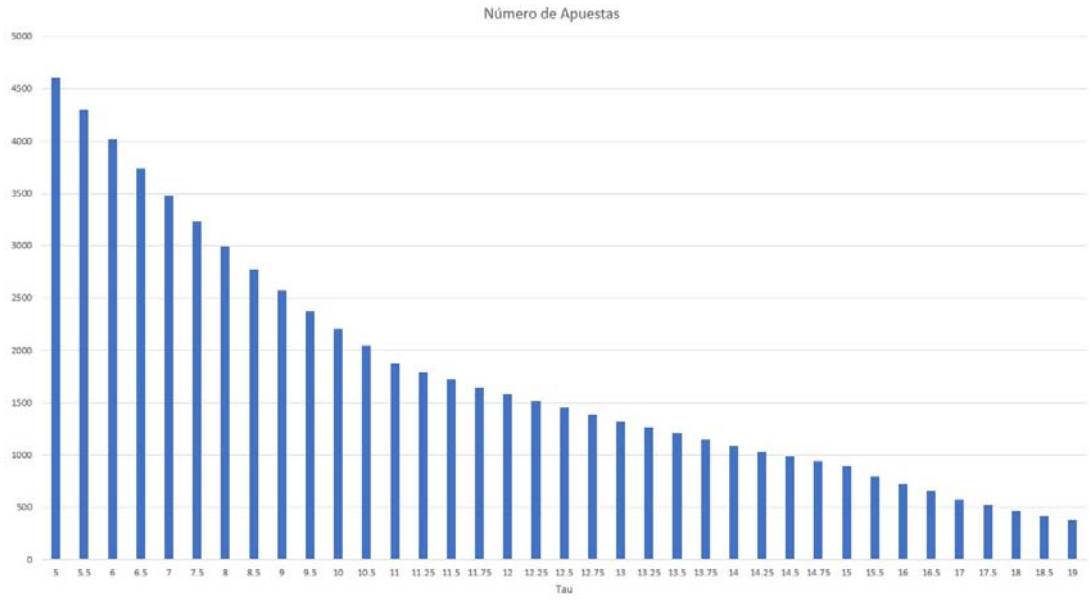

Figura 8.1: Número de apuestas para el modelo propuesto, 2014-2017, bajo distintos umbrales de apuesta τ

De la misma forma, cuando la ventaja del modelo crece, esperaríamos que la ganancia por apuesta aumenta también. Aunque la relación tiene un poco de ruido - a diferencia de la figura, 8.1 donde las apuestas bajo un umbral están anidadas en un umbral más pequeño - en la figura 8.2 podemos observar que efectivamente la ganancia por apuesta crece conforme la ventaja del modelo crece, alcanzando su máximo en $\tau = .175$, después de el cual la ganancia por apuesta disminuye un poco. Las razones por las que esto ocurre no están totalmente claras. Quizá las ventajas más grandes del modelo tengan una razón subyacente ajena al modelo. Por ejemplo, un equipo que ha perdido a su mejor jugador para el siguiente partido

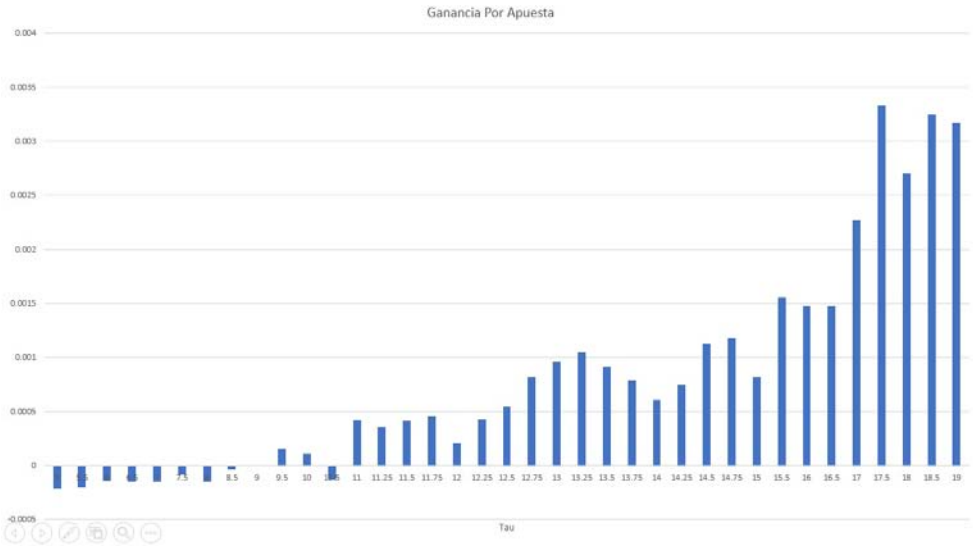

Figura 8.2: Ganancia promedio por apuesta para el modelo propuesto, 2014-2017, bajo distintos umbrales de apuesta τ

tendrá, en consecuencia, un momio mayor. Esto podría inflar artifi-

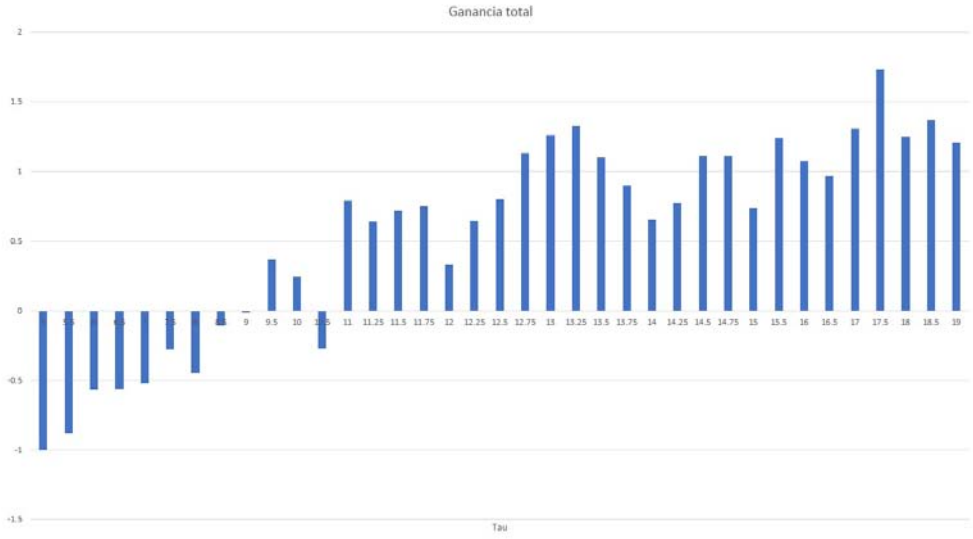

Figura 8.3: Ganancia total para el modelo propuesto, 2014-2017, bajo distintos umbrales de apuesta τ

cialmente la ventaja que el modelo cree tener en el partido.

Por último, en la figura 8.3 podemos observar que hay dos máximos con ganancias totales que sobresalen: en τ 13.25 % y 17.5 %; el segundo coincide con el máximo de ganancia por apuesta. Aunque el punto máximo de la ganancia se alcanza en $\tau = .175$, utilizar un umbral de ventaja tan grande reduce mucho las apuestas: en la figura 8.1 podemos observar que ir del umbral de apuestas $\tau = .1325$ al umbral de apuestas $\tau = .175$ reduce las mismas de 1263 a 574, es decir, en casi la mitad. Por ello, la varianza y suceptibilidad a malas rachas es mucho mayor bajo el segundo que bajo el primero. Así, decidimos dejar un poco de la posible ganancia en la mesa y utilizar el umbral de apuesta $\tau = .1325$.

8.3. Determinación de la semana de inicio ideal para apostar

Después de determinar la ventaja ideal para el modelo, queremos usar ésta para determinar la semana de inicio en la que el modelo tiene ya suficiente información para poder apostar y ganar dinero. El método utilizado es muy parecido al de la sección anterior: corremos el modelo con el umbral encontrado $\tau = .1325$ para las tres temporadas en las cinco ligas elegidas. Sabemos que el modelo tiene que tener algo de información, por lo que corremos el modelo empezando con la información de 6 semanas; es decir, dejamos que apueste de la semana 7 en adelante y filtramos a partir de ésta para buscar la semana de inicio ideal.

Hay una consideración adicional: en muchas de las ligas, los equipos ya no están dispuntando nada al final de la temporada. Más aún, hay muchas situaciones donde ambos equipos se pueden ver beneficiados por un empate, por lo que, sin explícitamente arreglar el partido, hay un acuerdo tácito entre ambos para no atacarse. Un ejemplo de lo discutido se puede ver en el partido Torino - Genoa de la Seria A italiana 2012 - 2013, celebrado el 08/05/2013. Los equipos llegaban al encuentro como 16vo y 17vo, apenas arriba de las posiciones de descenso (18-20) con una pequeña ventaja sobre

los equipos por debajo. Convenientemente para ambos, el juego terminó en un empate 0-0 en el que hubo solamente tres tiros a gol, todos sin oportunidades reales para anotar. Esta es una de las razones principales por la que a la base de datos le faltan algunas líneas de apertura: la casa puede decidir no abrir la línea a tiempo o no recibir apuestas para ese partido en absoluto.

Para evitar situaciones donde los intereses del equipo no se alínean con dar su mejor esfuerzo, o los equipos deciden usar suplentes puesto que ya no tienen nada por qué competir, el modelo no apostará en la semana final para ninguna de las ligas.

Un resumen gráfico de lo encontrado se puede ver en las figuras 8.4 y 8.5. La figura 8.4 muestra la ganancias del modelo en todos los partidos apostados por jornada. La figura 8.5 muestra las ganancias totales del modelo si se empezara a apostar en esa semana. Por ser la semana de inicio donde se maximizan las ganancias totales bajo el umbral $\tau = .1325$, elegimos comenzar a apostar en la semana 10.

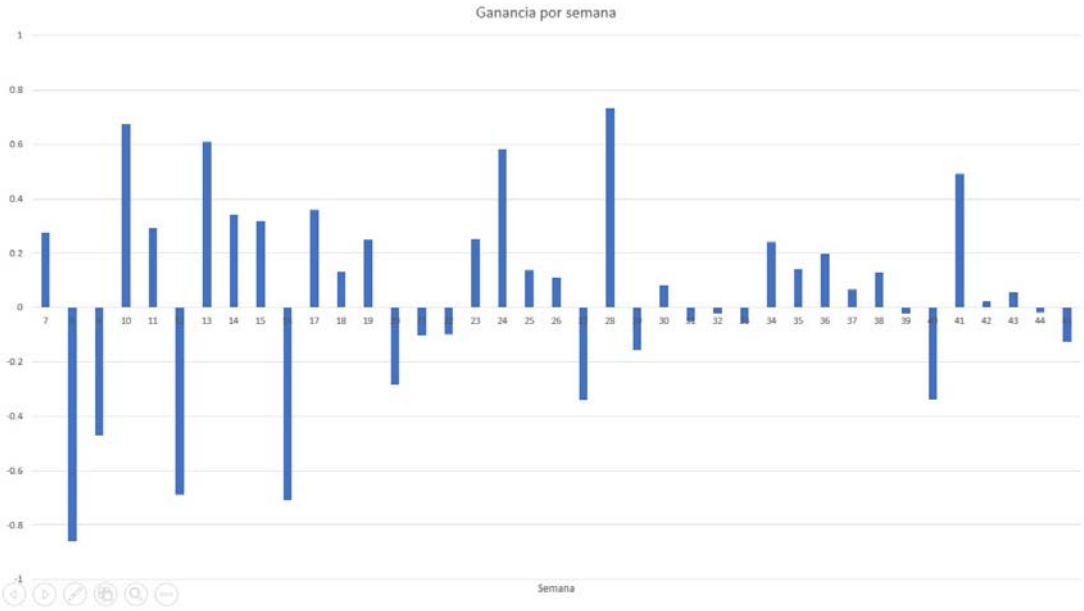

Figura 8.4: Ganancia en cada jornada a través para el modelo propuesto en el periodo 2014-2017, bajo un umbral de apuesta $\tau = .1325$

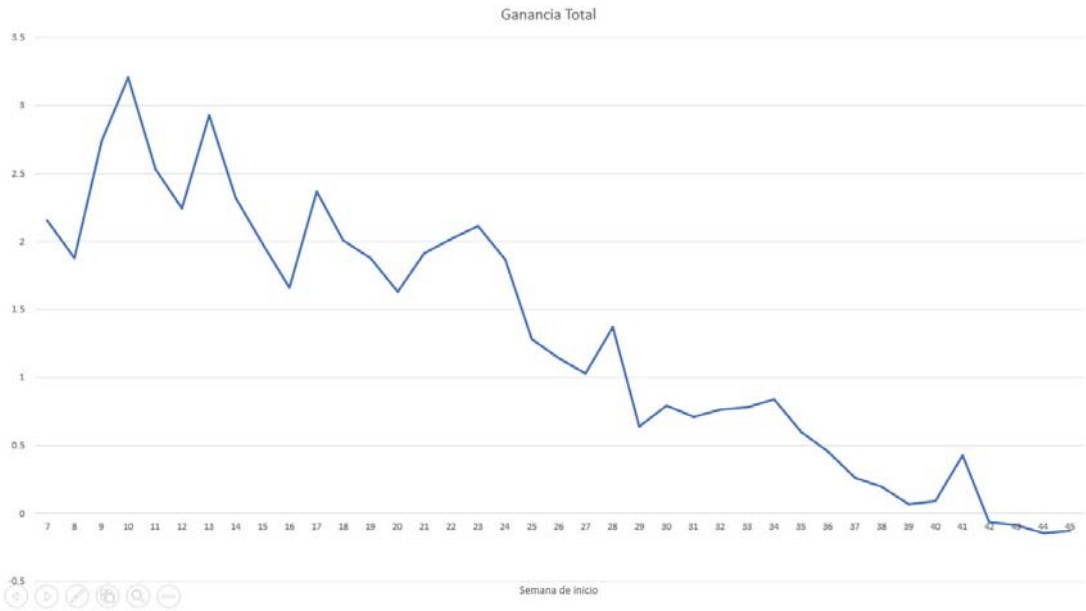

Figura 8.5: Ganancia total para el modelo propuesto en el periodo 2014-2017, bajo un umbral de apuesta $\tau = .1325$, variando la jornada de inicio.

8.4. Una aplicación del modelo ∇-Poisson doble al mercado y sus resultados.

Una vez que hemos elegido el umbral de apuestas $\tau = .1325$ y la jornada de inicio de apuesta (jornada 10) podemos finalmente determinar como le ha ido al modelo. Como recordatorio, el modelo está apostando en las temporadas del periodo 2014 - 2017 para la Bundesliga, Ligue 1, La Segunda, Serie B y English Championship.

En total, el número de juegos en nuestros datos son 6408. Dadas las restricciones al modelo - apuestas a partir de la jornada 10 y ninguna apuesta en la jornada final - el modelo tiene 4800 posibles juegos para apostar. Por juego, el modelo tiene cinco posibles apuestas: Local, Empate, Visitante, Over 2.5 y Under 2.5. Aunque bajo las restricciones del modelo es imposible que éste apueste en ambas altas y bajas, habrá casos en que el modelo apueste en dos resultados si uno de ellos está muy mal calibrado en las líneas. Por ejemplo, si el modelo cree que la probabilidad de victoria local es

mucho menor que en los momios, podría decidir apostar por ambos el empate y la victoria visitante. Éstos casos, sin embargo, son raros. Por otro lado, habrá casos en que el modelo apueste en el total de goles y el resultado de un mismo juego, como en el ejemplo de la sección 7.1. Por tanto, en los 4800 posibles juegos de apuesta, el modelo tiene 24000 apuestas posibles en el periodo especificado.

El modelo decide realizar 1809 apuestas, es decir, en promedio, un poco más de $\frac{1}{3}$ de apuesta por partido. De estas, el modelo gana 895 y pierde 913, con una ganancia total de +3.209 fortunas o 320 % de crecimiento. Es decir, si el apostador hubiera empezado en 2014 con $\$100,000$, habría terminado en 2017 con $\$320,903$.

Liga	Con Kelly Secuencial	Sin Kelly Secuencial
Segunda 1415	-0.235165	-0.163155
Segunda 1516	-0.359252	-0.356021
Segunda 1617	0.040160	0.148076
Francia 1415	-0.229526	-0.167112
Francia 1516	-0.252636	-0.094852
Francia 1617	0.811240	0.789383
EnglishC 1415	-0.395337	-0.286046
EnglishC 1516	1.045621	0.826183
EnclishC 1617	0.078354	0.184944
Serie B 1415	0.116770	0.295938
Serie B 1516	-0.030008	0.176876
Serie B 1617	0.561566	0.610027
Alemania 1415	0.774397	0.760483
Alemania 1516	0.228694	0.278321
Alemania 1617	0.026874	0.205987
Total	2.181752	3.209033

Figura 8.6: Ganancia para cada una de las temporadas apostadas, periodo 2014-2017, bajo un umbral de apuesta $\tau = .135$, empezando en la jornada 10, utilizando $\frac{1}{8}$ Criterio de Kelly. En la columna derecha, fijo a la fortuna inicial; en la columna izquierda apostando sobre el capital al inicio de la jornada.

Recordemos que las apuestas del modelo se hacen bajo $\frac{1}{8}$ de lo sugerido por el Criterio de Kelly. Cada apuesta se hace apostando un porcentaje del capital inicial, lo cual no es estrictamente lo sugerido por el Criterio de Kelly, bajo el cual se apuesta un porcentaje del capital al momento de hacer la apuesta. La distinción es igual a la de la sección 7.1. Esto presenta algunas consideraciones interesantes que ponderar:

1.- En todas las temporadas, estamos considerando que se empieza con exactamente el mismo capital inicial, para que el crecimiento porcentual a través de temporadas se pueda comparar correctamente. Esto, por supuesto, podría no ser así. Por ejemplo, a la conclusión de la temporada 2014 - 2015, podríamos decidir tomar todo el capital disponible como el nuevo capital inicial para la temporada 2015 - 2016, creando un efecto de crecimiento exponencial, por lo que los resultados podrían estar subestimados.

2.- Por otro lado, es importante recordar que para las apuestas sobre el total de goles, estamos usando la línea más alta disponible dentro de las casas compiladas por el sitio BetBrain. Esto podría hacer que el resultado esté un poco sobrestimado, ya que dicha línea de apuesta podría no estar disponible para nosotros en todos los partidos. Sin embargo, las casas con líneas atractivas para apostadores profesionales son pocas y suelen asemejar fuertemente a la mejor línea disponible, por lo que incluso teniendo cuentas en solamente 2 o 3 casas de apuesta, se puede conseguir la mejor línea o una muy cerca de ésta.

3.- Como hemos discutido en la sección 7.1, la implementación de un modelo bajo un Criterio de Kelly estricto - es decir, apostando el porcentaje sugerido por el Criterio bajo el capital disponible al momento de la apuesta - es complicado. Para el caso de este modelo, hay una consideración adicional: muchas de estas apuestas se estarán decidiendo simultáneamente a través de las distintas ligas, puesto que la inmensa mayoría de los partidos se juegan en fin de semana. Es por ello que es complicado determinar cuál es el capital disponible al momento de apostar. En la figura 8.6 se muestra una tabla con los resultados de todas las temporadas individuales en los datos, bajo un régimen Kelly fijo (que apuesta un porcentaje

del capital inicial) y un régimen Kelly secuencial (que apuesta un porcentaje del capital disponible al momento de apostar). La escala es relativa al capital inicial, es decir, fotunas ganadas o perdidas en la temporada. De interés, el resultado final parece ser mejor usando el sistema Kelly fijo que el sistema Kelly secuencial. Sin embargo, esto podría ser simplemente por la forma en que se compilan los resultados temporada a temporada. Es decir, tras una temporada cualquiera, el crecimiento se reinicia, por lo que los verdaderos efectos exponenciales bajo el sistema de Kelly secuencial se pierden a largo plazo. Sería interesante correr el sistema bajo un sistema de Kelly real, pero esto genera problemas computacionales, dado que las jornadas en distintas ligas se juegan en fechas distintas, por lo que determinar el capital disponible al momento de apostar es muy complicado.

4.- Este es un modelo que, probablemente, se beneficiaría de una persona ayudándolo a tomar decisiones semana a semana, por lo que los resultados podrían estar ligeramente subestimados. La causa es que, como hemos discutido ya, la gran fuerza del modelo es que no necesita mayor información que los marcadores hasta el momento que se quiere apostar para ser especificado. No toma en cuenta ningún factor externo. Y algunos de ellos serán significativos en la predicción, principalmente, la ausencia / presencia de algunos jugadores. Por ejemplo, si un equipo tiene a su mejor jugador suspendido o lesionado, o un conjunto de ausencias clave, el momio de apuiesta para ese equipo será, naturalmente, mayor en el mercado, relativo a su habilidad real, que es la que intenta representar el modelo, que no sabe de las ausencias y, por ello, encuentra una ventaja significativa para apostar, sin saber que su ventaja fue inflada artificialmente.

5.- Por otro lado, como la matriz de marcadores permite especificar todos los posibles resultados del partido, podría haber apuestas con mayor ventaja que las usadas por el modelo en este trabajo. Por ejemplo, si el modelo determinara que tiene ventaja en el empate y el equipo visitante, el modelo podría decidir, en vez, apostar al *Asian Handicap* +.5, que gana si ocurre cualquiera de los dos; o, en vez de apostar que hay tres o más goles en el mercado, apostar que hay cuatro o más goles en el partido. Es decir, como el modelo puede obtener probabilidades para un abanico muy amplio de po-

sibilidades, es probable que alguna de las apuestas no consideradas en este trabajo tenga una ventaja mayor.

Conclusiones

La intención de esta tesis era desarrollar un modelo de predicción para partidos de futbol, capaz de generar una ganancia en el mercado de apuestas.

Tras una larga exploración de los trabajos previos similares en la literatura, hemos concluido que los modelos recientes han sobrecomplicado la estructura subyacente de los marcadores, sin añadir mucha predictibilidad o haciendo que el modelo solamente sea útil en casos particulares.

Después de un análisis empírico de los datos, concluímos que la distribución Poisson hace un trabajo espléndido modelando tanto la suma de los marcadores en un partido, como los marcadores por sí mismos.

Por ello, hemos vuelto al principio y utilizado el primer modelo en la literatura por Maher (1982) [1], que asigna parámetros de ataque y defensa a cada equipo para poder modelar los goles anotados en un partido con distribuciones Poisson independientes, una por marcador.

Tras ver que, 35 años después, un modelo tan simple y hermoso como el de Maher sigue vigente, hemos resuelto su mayor problema: que no es dinámico en el tiempo. Esto nos ha llevado al desarrollo del modelo ∇-Poisson doble: una extensión del modelo original de Maher que ahora es dinámico en el tiempo y nos permite hacer predicciones dentro de una misma temporada.

El modelo ha sido probado en 15 temporadas de datos, apostando tanto en el resultado del partido, como en el total de goles del mismo, en el conocimiento del autor, la primera vez que esto se hace en la literatura. Además, por primera vez también, hemos integrado el Criterio de Kelly, usado en otros contextos, a las apuestas deportivas.

Los resultados del modelo son favorables en la muestra y parecen mostrar que el modelo puede ser adaptado y utilizado en más ligas en el mundo con éxito, pues el modelo ∇-Poisson doble, como el Poisson doble propuesto por Maher, no tiene ninguna condición que lo limite a una estructura de liga en particular.

Ha sido un placer trabajar en ésta tesis durante el último año y espero que usted haya disfrutado leyéndola tanto como yo escribiéndola.

Bibliografía

[1] M. J. Maher, "Modelling association football scores," *Statistica Neerlandica*, 1982.

[2] S. Kocherlakota and K. Kocherlakota, "Bivariate Discrete Distributions," in *Encyclopedia of Statistical Sciences*, 2006.

[3] D. Karlis and L. Ntzoufras, "Analysis of sports data by using bivariate Poisson models," *Journal of the Royal Statistical Society Series D: The Statistician*, 2003.

[4] M. Moroney, "Facts from figures," p. 472, 1953.

[5] M. D. Ugarte, A. F. Militino, and A. T. C. N. Q. U. . Arnholt, *Probability and statistics with R.* 2016.

[6] C. Reep, R. Pollard, and B. Benjamin, "Skill and Chance in Ball Games," *Journal of the Royal Statistical Society. Series A (General)*, 1971.

[7] C. Reep and B. Benjamin, "Skill and Chance in Association Football," *Journal of the Royal Statistical Society. Series A (General)*, vol. 131, no. 4, pp. 581–585, 1968.

[8] I. D. Hill, "Association Football and Statistical Inference," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1974.

[9] M. Thompson, "On Any Given Sunday: Fair Competitor Orderings with Maximum Likelihood Methods," *Journal of the American Statistical Association*, vol. 70, no. 351, pp. 536–541, 1975.

[10] M. J. Dixon and S. G. Coles, "Modelling association football scores and inefficiencies in the football betting market," *Journal*

of the Royal Statistical Society. Series C: Applied Statistics, 1997.

[11] M. J. Dixon and P. F. Pope, "The value of statistical forecasts in the UK association football betting market," *International Journal of Forecasting*, vol. 20, no. 4, pp. 697–711, 2004.

[12] D. Karlis and I. Ntzoufras, "Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R / SPLUS Models for Bivariate Poisson Data Bivariate Poisson Regression models," *Journal of Statistical Software to appear*, 2005.

[13] D. Sumpter, *Soccermatics: Mathematical Adventures in the Beautiful Game*. Bloomsbury Publishing, 2016.

[14] S. J. Koopman and R. Lit, "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, vol. 178, no. 1, pp. 167–186, 2015.

[15] A. J. Lee, "Modeling Scores in the Premier League: Is Manchester United Really the Best?," *CHANCE*, vol. 10, pp. 15–19, jan 1997.

[16] B. Deschamps and O. Gergaud, "Efficiency in Betting Markets: Evidence from English Football," *Journal of Prediction Markets*, vol. 1, no. 1, pp. 61–73, 2007.

[17] M. Cain, L. David, and D. Peel, "The favourite-longshot bias and market efficiency in UK football betting," *Scottish Journal of Political Economy*, 2000.

[18] "ESPN: nevada sportsbooks took in record bets and winnings." `http://www.espn.com/chalk/story/_/id/22273982/record-amounts-money-bet-lost-nevada-2017`. Accessed: 2018-04-12.

[19] "Twitter: nevada casinos' net win in 2017." `https://twitter.com/DavidPurdum/status/958729369178001408`. Accessed: 2018-04-12.

[20] "ESPN: nevada books eke out slim win off record super bowl betting." `http://www.espn.com/chalk/story/_/id/22337636/`

nevada-sportsbooks-report-record-numbers-bets-super-bowl-lii. Accessed: 2018-04-12.

[21] "The Guardian:the 5,000-1 payouts on leicester only tell part of premier league betting story." https://www.theguardian.com/football/2016/may/03/5000-1-outsider-leicester-city-bookmakers. Accessed: 2018-02-23.

[22] M. J. Lopez, G. J. Matthews, and B. S. Baumer, "How often does the best team win? A unified approach to understanding randomness in North American sport," 2017.

[23] "The Business of Betting Podcast. episode: 26 - pinnacle's director of trading." https://soundcloud.com/businessofbetting/ep-26-pinnacles-director-of. Accessed: 2018-01-29.

[24] D. Bernoulli, "Exposition of a New Theory on the Measurement of Risk," *Econometrica*, vol. 22, no. 1, pp. 23–36, 1954.

[25] M. Lewis, *The Undoing Project: A Friendship that Changed the World*. Penguin Books, Limited, 2017.

[26] J. L. Kelly, "A New Interpretation of Information Rate," *Bell System Technical Journal*, 1956.

[27] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, 1948.

[28] E. O. Thorp, "The Kelly Criterion in Blackjack Sports Betting, and the Stock Market," in *Handbook of Asset and Liability Management - Set*, 2008.

[29] W. T. Z. Leonard C. MacLean, Edward O. Thorp, *The Kelly Capital Growth Investment Criterion: Theory and Practice*. 2011.

[30] L. Breiman, "Optimal Gambling Systems for Favorable Games," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Fourth Berkeley Symposium on Mathematical Statistics and Probability, (Berkeley, Calif.), pp. 65–78, University of California Press, 1961.

[31] E. O. Thorp, "Optimal Gambling Systems for Favorable Games," *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, vol. 37, no. 3, pp. 273–293, 1969.

[32] "Wikipedia: year of adoption of 3-points-for-a-win." `https://en.wikipedia.org/wiki/Three_points_for_a_win#Year_of_adoption_of_3-points-for-a-win`. Accessed: 2018-03-05.