Universidad de Cantabria

Facultad de Ciencias

REGRESIÓN NO PARAMÉTRICA
Análisis de la evolución en el rendimiento de ajedrecistas a lo largo de su vida deportiva

Trabajo de Fin de Grado
para acceder al
GRADO EN MATEMÁTICAS

Autor: Javier Berrueta Irigoyen

Director: Juan Antonio Cuesta Albertos

Septiembre - 2024

Agradecimientos

No quiero comenzar este trabajo sin antes dar las gracias a todas las personas que han sido importantes en mi vida académica.

A todos los profesores y profesoras con quienes he coincidido, cada uno de los cuales me ha enseñado algo, también a nivel personal. Especialmente a Juan Cuesta, por el tiempo y esfuerzo que ha dedicado a este trabajo y lo que he aprendido en las reuniones que hemos mantenido. También quiero agradecer al profesor D. Santos Domínguez Menchero, de la Universidad de Oviedo, a quien no he tenido el placer de conocer pero que sugirió el análisis de los resultados separados por género.

No me puedo olvidar tampoco de mis amigos y amigas, tanto de los que he hecho a lo largo de la vida universitaria como de quienes me acompañan desde antes de esta etapa. Sin ellos, sería mucho más difícil superar las dificultades que aparecen, tanto en el aspecto académico como personal.

Por último, pero no menos importante, a mi familia, especialmente a mis padres y hermanos, que me han guiado y acompañado durante toda mi vida. Aunque durante la carrera haya vivido lejos de casa, siempre los he sentido cerca.

"De ninguna de mis victorias
he aprendido tanto como de
la mayoría de mis derrotas."
José Raúl Capablanca.

Resumen

El ajedrez es un deporte muy peculiar, por la importancia capital de las habilidades cognitivas en su desemepeño. En este trabajo, se pretende analizar la importancia relativa que tienen el aspecto físico y el aspecto mental en este deporte, así como describir el proceso de evolución de los jóvenes que acaban llegando a la élite del mismo. Además, se busca explicación a la brecha de nivel observada entre hombres y mujeres. Con este fin, se hace uso de técnicas de regresión no paramétrica por núcleos, así como cross validation y bootstrap para optimizar la regresión y estimar su error.

Palabras clave: ajedrez, rendimiento, deporte, género, regresión no paramétrica, núcleo, ventana, cross-validation, validación cruzada, bootstrap.

Abstract

Chess is a very distinctive sport, due to the important role the cognitive abilities play in it. In this work, the aim is to analyse the relative importance of the physical and mental aspects in this sport and to describe the developmental process of young players who eventually reach the elite level. Moreover, an explanation for the observed level gap between men and women is sought. To this goal, kernel-based non-parametric regression techniques are employed, along with cross-validation and bootstrapping to optimize the regression and estimate its error.

Key words: chess, performance, sport, gender, non-parametric regression, kernel, bandwidth, cross-validation, bootstrap.

Índice general

1 Introducción **1**
- 1.1 Descripción del trabajo . . . 1
- 1.2 Objetivos . . . 2
- 1.3 Organización de la memoria . . . 3

2 Marco teórico **5**
- 2.1 Regresión . . . 5
- 2.2 Regresión no paramétrica . . . 6
  - 2.2.1 Ventana de suavizado . . . 7
  - 2.2.2 Núcleo . . . 8
- 2.3 Cross Validation . . . 12
  - 2.3.1 k-folds CV . . . 13
- 2.4 Bootstrap . . . 14
  - 2.4.1 Bootstrap Naive . . . 16
  - 2.4.2 Wild Bootstrap . . . 16
  - 2.4.3 Región de confianza . . . 17

3 Aplicación práctica **19**
- 3.1 Elo en ajedrez . . . 19
  - 3.1.1 Cálculo del resultado esperado . . . 21
- 3.2 Rendimiento . . . 22
- 3.3 Muestra utilizada . . . 23

4 Resultados **27**

5 Discusión **31**
- 5.1 Evolución y pico de rendimiento . . . 31
- 5.2 Periodo de aprendizaje . . . 32

6 Conclusiones **35**

Capítulo 1

Introducción

1.1. Descripción del trabajo

Las matemáticas son una ciencia que desempeña un papel crucial en numerosos ámbitos, como la física, la biología, las ciencias sociales... Una de sus aplicaciones más destacadas es la modelización de la relación entre variables, proporcionando información que facilita la comprensión y el análisis de los datos. En este trabajo, se utilizan técnicas de regresión no paramétrica para modelizar la evolución temporal del rendimiento de ajedrecistas.

El rendimiento deportivo no es constante a lo largo de la vida de los deportistas, sea cual sea su disciplina, ya que el cuerpo humano experimenta transformaciones que afectan a la fuerza, resistencia o agilidad mental, entre otros aspectos. También la propia experiencia del deportista tiene un impacto en sus resultados que, a fin de cuentas, son la herramienta que se usa para medir el rendimiento. Aunque estos cambios afectan a todas las personas, en los estudios sobre el envejecimiento y el deporte se suelen utilizar los datos de profesionales [1], de los que se puede suponer que dan su máximo a lo largo de toda su carrera, por lo que la evolución de su nivel se debe netamente al envejecimiento.

Por el momento se ha hablado de deportes en general, pero, de igual manera que las distintas disciplinas exigen diferentes cualidades, tanto físicas como mentales, el envejecimiento les afecta de distinta manera. De hecho, en [2] se hace una categorización entre deportes, en función de si requieren más explosividad, resistencia... y se demuestra que el rendimiento óptimo se alcanza a distintas edades en función de ello. Por ejemplo, en deportes como el ciclismo el pico de rendimiento se encuentra en torno a los 40 años [3], mientras que en otros, como el baloncesto, se produce antes, entre los 25 y los 30 años [4] . Incluso se puede ir más allá y estudiar distintos aspectos de un deporte, como el fútbol, en el que el envejecimiento hace que disminuya la velocidad pero que aumente la precisión de los pases [5].

Asimismo, las diferencias fisiológicas entre hombres y mujeres hacen que sus cuerpos evolucionen de diferente manera [6], [7], por lo que es probable que también su rendimiento deportivo se vea afectado de forma distinta por la edad. Esto hace que sea de interés diferenciar por sexos el análisis, algo que no es muy frecuente en la literatura.

Como se ha visto, existen artículos científicos que analizan cómo el envejecimiento afecta al rendimiento en distintos deportes físicos, detallando incluso su influencia sobre parámetros concretos del juego. Sin embargo, apenas se encuentran estudios que aborden el impacto del envejecimiento en el ajedrez, un deporte bastante particular, en el que las capacidades cognitivas juegan un papel decisivo y las físicas parecen perder relevancia. Es por ello que surgen preguntas

como las siguientes:

- Sabiendo que la edad afecta negativamente a la resistencia física, ¿qué importancia tiene esta en ajedrez?
- ¿Cómo evolucionan las habilidades mentales de una persona a lo largo del tiempo?
- Con los años se gana experiencia, ¿cuál será su impacto en el rendimiento ajedrecístico?

Ya dijo Anatoly Karpov (ex-campeón del mundo de ajedrez) que "el ajedrez es deporte, ciencia y arte", y precisamente un análisis de estas preguntas puede ayudar a discernir hasta qué punto el ajedrez incorpora estos tres aspectos, y si alguno de ellos domina frente a los otros dos.

Hay otro fenómeno muy curioso en este deporte y es que, en el campo profesional, los hombres superan ampliamente a las mujeres, tanto históricamente como en la actualidad. Si bien esto es habitual en deportes en los que la fuerza o la velocidad son determinantes y sucede a la inversa cuando son aspectos como la flexibilidad los que se requieren, se podría pensar que el ajedrez sería un deporte bastante *igualitario*. Aunque explicar los motivos de esta circunstancia es una tarea que correspondería a las ciencias biológicas o sociales, en este trabajo se pretende, al menos, describir el fenómeno desde el punto de vista matemático.

En resumen, este trabajo consiste en el uso de técnicas de regresión no paramétrica para describir la evolución temporal del rendimiento ajedrecístico en profesionales, con la intención de obtener conclusiones relativas a los aspectos ya mencionados. Para ello, se utilizarán las puntuaciones obtenidas con el sistema Elo (puntos Elo o Elo en lo sucesivo) de 32 ajedrecistas y, para cada uno, se determinará su *rendimiento personal* en cada momento de su carrera, de forma que se puedan comparar jugadores de diferente *nivel* y hacer un promedio entre ellos. A lo largo del trabajo, se entenderá que *nivel* es una magnitud intrínseca al jugador, no afectada por el paso del tiempo, mientras que *rendimiento* se referirá a las oscilaciones temporales de sus puntos Elo.

1.2. Objetivos

Este trabajo tiene como objetivo principal el análisis del impacto del envejecimiento en el rendimiento de ajedrecistas de ambos sexos. Para ello, se han establecido varios objetivos específicos que permitirán abordar distintas facetas de esta cuestión.

En primer lugar, se ubicará en qué edad se encuentra el pico del rendimiento ajedrecístico. Esto podrá ser usado para categorizar el ajedrez dentro de las distintas modalidades de deportes descritas en [2].

Otro objetivo será estudiar la constancia del rendimiento a lo largo del tiempo, si se producen altibajos, mesetas, caídas pronunciadas, etcétera. Evaluar estas variaciones permitirá entender mejor cómo evoluciona la capacidad competitiva de los ajedrecistas y compararla con los patrones típicos de rendimiento deportivo y cognitivo.

Además, se analizarán las diferencias entre hombres y mujeres en su evolución como ajedrecistas, especialmente de jóvenes, buscando una explicación a la diferencia de nivel observada, aunque sea de manera preliminar y poco rigurosa. Este estudio pretende ofrecer una visión inicial que podría servir como base para investigaciones futuras más detalladas

y fundamentadas sobre las razones detrás de estas disparidades en el nivel ajedrecístico por género.

Por último, se determinará a partir de qué edad se define el nivel de los ajedrecistas, es decir, el momento en que el rendimiento de los jugadores se estabiliza en torno a su valor máximo, más allá de pequeños altibajos. Esta información es crucial para comprender el proceso de desarrollo y maduración de las habilidades en el ajedrez, y cómo factores como la experiencia y el aprendizaje continuo influyen en el rendimiento a largo plazo.

En resumen, los objetivos de este trabajo buscan aportar conocimiento sobre la evolución del rendimiento en el ajedrez, además de entender mejor las características intrínsecas de este deporte.

1.3. Organización de la memoria

Esta memoria contiene 6 capítulos, en cada uno de los cuales se aborda un aspecto del estudio realizado, y su estructura es la siguiente.

El Capítulo 2 presenta el marco teórico, donde se presentan las bases matemáticas que permiten el análisis. En esta sección se aborda la regresión, particularmente la local o no paramétrica, explicando detalladamente sus aspectos fundamentales, que será la herramienta usada para el ajuste de los datos a una curva. Además, se explica la técnica de *cross validation*, cuya función será optimizar el funcionamiento de la regresión, y el *bootstrap*, que permitirá dar una estimación del error en el ajuste.

El Capítulo 3 se centra en el problema particular al que se le aplican las técnicas matemáticas mencionadas, es decir, se detalla el origen de los datos y su significado. Concretamente, se hablará del sistema de puntuación Elo en ajedrez y el tratamiento por el cual se puede derivar el rendimiento de un jugador a partir de él. También se detallan los jugadores y jugadoras escogidos y el periodo de su carrera analizado.

En el Capítulo 4 se presentan los resultados del trabajo. Contiene las curvas de regresión obtenidas, incluyendo regiones de confianza, y un pequeño análisis de las mismas.

El Capítulo 5 está dedicado a la discusión, que estará dividida en distintos apartados, de forma que a partir de los resultados obtenidos se pueda reflexionar sobre aspectos como la forma de las curvas, relacionada con el tipo de deporte que es el ajedrez, el periodo de aprendizaje y la determinación del nivel personal, y la comparación entre hombres y mujeres.

Finalmente, el Capítulo 6 presenta las conclusiones principales extraídas del estudio, y su relación con los objetivos marcados. Además, se sugieren posibles líneas de investigación futuras.

Capítulo 2

Marco teórico

2.1. Regresión

La regresión es una técnica estadística utilizada para analizar y modelar la relación entre variables. En general, consiste en la predicción del valor esperado de una variable a la vista de los valores que toman otras. Para su uso, es necesario un conjunto de observaciones a partir del cual se pueda inducir la relación existente entre ellas. Por simplicidad, y porque en el problema práctico que vamos a analizar contamos con solo dos variables, en este capítulo nos restringiremos a ese caso, donde X e Y serán sendas variables aleatorias reales (v.a.'s). Así, el objetivo de las técnicas de regresión es encontrar una función $r : \mathbb{R} \longrightarrow \mathbb{R}$ tal que [8]

$$Y = r(X) + \epsilon. \tag{2.1}$$

Aquí, ϵ es otra v.a. denominada error, que incluye la variabilidad de Y debida a factores ajenos a X. Es muy habitual [9] suponer que ϵ es normal de media 0 e independiente de X, aunque el mínimo requerimiento es que $r(X)$ y ϵ sean incorreladas. Obsérvese que, si X y ϵ son v.a. independientes, entonces $r(x) = E[Y|X = x]$. Como ya se ha mencionado, hará falta un conjunto de observaciones, es decir, una muestra del par de variables aleatorias (X, Y) que aporte información sobre r. Por ello, supondremos que se dispone de una muestra de observaciones $\{(X_i, Y_i) : i = 1, 2, ...n\}$ i.i.d. y mantendremos esta asunción a lo largo de todo el capítulo.

En muchos casos, en vista de la muestra o de información previa sobre el problema, se supone que r tiene una determinada forma, esto es, que sea una función lineal, polinómica de grado conocido, exponencial, etc. Esta técnica se conoce como regresión paramétrica, y se basa en lo siguiente: Dado un conjunto de parámetros $\Theta \subseteq \mathbb{R}^k$ para algún $k \geq 1$ y una familia de funciones $\{r_\theta : \theta \in \Theta\}$, donde se supone que existe $\theta_0 \in \Theta$ tal que r_{θ_0} satisface (2.1), el objetivo es estimar θ_0. Para ello, se puede elegir un $\hat{\theta} \in \Theta$ que optimice en algún sentido el ajuste

$$Y_i \approx r_{\hat{\theta}}(X_i). \tag{2.2}$$

Por ejemplo, en el caso de la regresión lineal, esto es, con $\Theta = \mathbb{R}^2$ y $\theta = (a, b)$, se supone que la relación entre variables es la siguiente [10]

$$Y = aX + b + \epsilon. \tag{2.3}$$

Una posibilidad bastante habitual es utilizar el método de mínimos cuadrados, que se usa para calcular los valores de los parámetros para los cuales el error cuadrático es mínimo.

$$\chi^2_\theta = \sum_{i=1}^{n} (Y_i - r_\theta(X_i))^2. \tag{2.4}$$

La regresión paramétrica es sin duda una técnica muy útil en los casos en que se dispone de información sobre el problema y la relación entre variables. De hecho, en estos casos es muy frecuente que los parámetros tengan algún significado concreto, especialmente en física [11], por lo que su estimación es muy importante. Otra de sus ventajas más destacadas es la posibilidad de extrapolar, es decir, de hacer predicciones más allá de los límites de la muestra de la que se dispone.

Sin embargo, hay muchos problemas en los que no se puede asumir la pertenencia de r a ninguna familia paramétrica concreta, ya sea por falta de información o porque influyen demasiados factores para poder tenerlos en cuenta. Esto es más habitual en ciencias sociales, donde el comportamiento de los individuos es dificilmente modelizable [12]. Existe otra técnica de ajuste, la regresión local o no paramétrica, que sirve para determinar la relación entre variables cuando no se dispone de un modelo. Es la que se utilizará en este trabajo y su base teórica se muestra a continuación.

2.2. Regresión no paramétrica

Estamos trabajando con el conjunto de observaciones $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. Denotaremos $X_{(1)}, X_{(2)}, ..., X_{(n)}$ a la muestra ordenada, es decir, donde imponemos $X_{(i)} \leq X_{(i+1)}$.

La regresión no paramétrica, también conocida como regresión local, es una técnica estadística para la estimación de la función de (2.1) en los casos en que r es continua. Como hemos indicado, en este trabajo nos limitaremos a dos variables reales. Su fundamento básico consiste en estimar el valor de la curva en cada punto de las abscisas utilizando las observaciones más cercanas a ese punto. Para ello, se supone que se verifica la relación (2.1), con la salvedad de que ahora r no pertenece a una familia de funciones paramétricas, sino que puede ser aproximada localmente por una función paramétrica [13].

La idea es que, para cada $x \in \mathbb{R}$, la estimación de $r(x)$ solo dependerá de las observaciones (X_i, Y_i) próximas a x. Desde el punto de vista teórico, y para mantener el formalismo matemático, consideramos que $r(u)$ puede ser aproximado por un polinomio mientras u esté lo suficientemente cerca de x.

$$r(u) = a_0 + a_1(u - x) + ... + a_n(u - x)^n + O(u - x)^{n+1}. \tag{2.5}$$

En la práctica, será necesario decidir hasta qué orden del polinomio tomamos, lo cual supone un compromiso entre precisión y coste computacional. Nadaraya [14] y Watson [15] asumen que la función es continua en todos los puntos y toman el polinomio de orden 0, es decir, la aproximación

$$r(u) \approx r(x) = a_0. \tag{2.6}$$

El estimador de Nadaraya-Watson en cada $x \in \mathbb{R}$ es un promedio de los Y_i con pesos que dependen de las distancias $|x - X_i|$.

$$\hat{r}_h(x) = \sum_{i=1}^{n} w_i(x) Y_i. \tag{2.7}$$

Los pesos w_i vienen dados por la elección de una *ventana* y un *núcleo*. En las siguientes subsecciones se explica detalladamente el papel que juegan y cómo se eligen. Por ahora, basta decir que el núcleo es una función K con ciertas características, por ejemplo, puede ser la

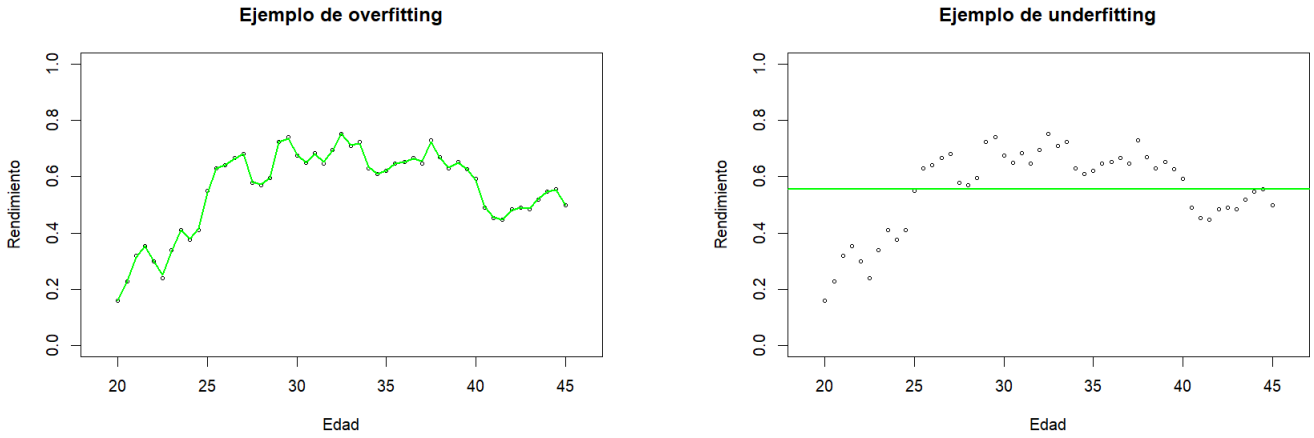

Figura 2.1: (a). Ventana es demasiado pequeña: solo los puntos más cercanos influyen en la curva de regresión, siendo prácticamente una interpolación lineal. (b). Ventana demasiado grande: todos los puntos influyen en la curva de regresión. La curva es constante igual a la media muestral.

función de densidad de una normal $\mathcal{N}(0, \sigma^2)$, y la ventana h un parámetro positivo que permite ajustar la importancia relativa que damos a las observaciones en función de su distancia al punto x.

$$w_i(x) \propto K\left(\frac{|x - X_i|}{h}\right).$$

2.2.1. Ventana de suavizado

La ventana de suavizado es el único parámetro que interviene en la regresión local. Controla el peso de las observaciones en la estimación del valor de la curva de regresión en cada punto. Hablando un poco imprecisamente, podemos decir que cuanto más pequeña sea, menos observaciones afectan, por lo que la curva seguirá más a las mediciones y, si es demasiado pequeña, se habla de *overfitting*. Ver Figura 2.1 a). Si la ventana es grande, se tendrán en cuenta observaciones muy lejanas y la curva será más suave. Cuando es excesivamente grande, se produce *underfitting*. Ver Figura 2.1 b).

Se comentan a continuación algunas de las técnicas más frecuentes [13] para la determinación de h.

La más sencilla es tomar $h(x) = h$ para todo $x \in \mathbb{R}$. Esta elección es lógica en los problemas donde las observaciones están equiespaciadas en el eje de abscisas, es decir, cuando tenemos un diseño fijo con la distancia entre $X_{(i)}$ y $X_{(i+1)}$ igual para cualquier i. En este caso, la cantidad de observaciones que afectan a la estimación en el punto x es la misma para cualquier x. El problema aparece cuando tenemos distribuciones no uniformes, ya que, fijado un h constante, los x de las regiones más densas se verán afectados por muchas observaciones, mientras que los que estén en la situación opuesta corren el riesgo de que ninguna observación esté dentro de su ventana. Incluso en el caso en que nos aseguremos de que la ventana nunca está vacía de observaciones, en estas regiones la estimación sería muy inestable por su dependencia de muy pocas observaciones. Afortunadamente, en este trabajo se trabajará con una distribución equiespaciada en X, por lo que se podrá utilizar la ventana constante sin sufrir los problemas descritos.

En general, para entender la influencia del valor de h, hay que tener claro lo siguiente:

Cuando h es grande, esto es, muchos puntos influyen en las estimaciones, disminuye la varianza en cada punto porque hay más puntos que intervienen en la estimación, pero aumenta el sesgo, ya que observaciones alejadas adquirirán importancia. Por el contrario, valores bajos de h nos llevan a una situación en que la varianza será alta pero el sesgo bajo. En consecuencia, el valor óptimo de h será diferente para cada conjunto de observaciones.

Con todo lo expuesto, se hace manifiesta la importancia capital de la elección del tamaño de la ventana en la regresión local. La forma de calcularla será la búsqueda del h que minimice el error en la estimación. Hay distintas maneras de determinar el error, una de ellas es el MISE (*Mean Integrated Square Error*), definida en (2.8), donde $\hat{r}_h$ hace referencia al estimador utilizando una ventana h y r es la verdadera relación subyacente entre las variables [16].

$$E\left[\int_{-\infty}^{+\infty} (r(x) - \hat{r}_h(x))^2 \, dx\right]. \tag{2.8}$$

Como r es desconocida, será necesario utilizar otra manera de calcular el error. En este proyecto se utilizará el MSE (*Mean Squared Error*), que elimina la integral anterior y además utiliza la variable Y, de la cual tenemos una muestra $\{Y_i\}$:

$$\text{MSE}(h) := E\left[(\hat{r}_h(X) - Y)^2\right]. \tag{2.9}$$

Por lo tanto la ventana óptima será

$$h_{\text{opt}} = \arg \min_{h \in \mathbb{R}^+} \text{MSE}(h). \tag{2.10}$$

Utilizaremos la técnica de *cross-validation* para aproximar h_{opt} (ver Sección 2.3) En [17] se demuestra la optimalidad asintótica de la solución obtenida por este procedimiento.

2.2.2. Núcleo

El concepto de núcleo (*kernel* en inglés) es fundamental cuando queremos aplicar la técnica de regresión local por núcleos [13], como es el caso de este proyecto. En esta subsección se explicará detalladamente el concepto y su relevancia en el cálculo de la curva de regresión.

Un núcleo es una función $K : \mathbb{R} \longrightarrow \mathbb{R}^+$ con las siguientes características:

- Es simétrica, esto es, $K(u) = K(-u)$ para todo $u \in \mathbb{R}$.
- Es no creciente en $(0, +\infty)$.
- Es integrable; de hecho, podemos suponer que está normalizada con $\int_{-\infty}^{+\infty} K(u)du = 1$.

Como observación, estas características hacen que la función alcance su máximo en 0 y además se verifique

$$\lim_{u \longrightarrow \pm\infty} K(u) = 0.$$

La función que cumple el núcleo en la regresión local es la de *pesar los datos*, es decir, determinar los parámetros $\{w_i(x)\}$ en (2.7). Supongamos que queremos calcular el valor de cada w_i para un punto $x \in [X_{(1)}, X_{(n)}]$. La idea es que, para cada i, cuanto menor sea $|x - X_i|$ mayor peso tendrá que tener Y_i en el estimador, es decir, mayor deberá ser w_i. Es más, para tener un estimador razonable, las observaciones más alejadas deberían tener un peso muy cercano a 0 o, incluso, 0. Aquí entran en juego las características del núcleo K, que tiene un

valor mayor cuanto más cercano a 0 es su argumento, manteniendo la simetría y tendiendo a 0 cuando este aumenta en valor absoluto.

Parece, pues, que el peso de cada observación puede ser la evaluación de K en las distancias entre x y la correspondiente observación, o algo parecido. Por tanto, una primera aproximación de la relación entre la función núcleo y los pesos es

$$w_i(x) = K\left(X_i - x\right).$$

Esta primera idea tiene varios problemas. Para empezar, no permite controlar la escala, es decir, si K es suficientemente decreciente, no obtendríamos el mismo resultado para una distribución uniforme de 10 puntos entre 1 y 10 que 10 puntos entre 0.1 y 1. Esto es absurdo desde el punto de vista matemático. Además, en la misma línea, no nos permite jugar con la cantidad de puntos que queremos que intervengan en cada estimación. Y aquí es donde entra en juego la ventana h, que soluciona ambos problemas ya que servirá de *lupa* para reescalar las distancias entre puntos. Como se verá a continuación, muchos núcleos tienen como soporte el intervalo $[-1, 1]$, por lo que solo tendrán peso no nulo aquellas observaciones para las cuales el argumento de la función K esté entre -1 y 1. Y como h era precisamente la distancia máxima hasta la cual las observaciones debían tener peso no nulo, parece buena idea utilizar la expresión (2.11) para poder modular, a través de h, qué región (o cuántas observaciones) nos interesan.

$$w_i(x) = K\left(\frac{X_i - x}{h}\right). \tag{2.11}$$

Y si bien esta expresión ya es suficientemente flexible y cumple los requisitos que buscábamos, falta un último detalle: la normalización de los pesos $\{w_i\}_{i=1,2,\ldots,n}$, de tal forma que su suma sea la unidad. La solución es sencilla y basta dividir cada uno por la suma de todos ellos. Esta es la definición definitiva de los pesos:

$$w_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}. \tag{2.12}$$

Con ello se consigue que la estimación de r en cada punto $x \in [X_{(1)}, X_{(n)}]$ sea un promedio de las observaciones. Además, si n es grande, usualmente, h será pequeño por lo que solo las observaciones próximas a x tendrán influencia en la estimación. De aquí el adjetivo de local.

Ejemplos de núcleos

Se muestran a continuación algunos núcleos frecuentes en la literatura [13]. Aquí, la función, $I_{[-1,1]}(u)$ es la función indicador de $[-1, 1]$ evaluada en u, que toma valor 1 si $u \in [-1, 1]$ y 0 en caso contrario. Hay que notar que tanto la continuidad como el orden de derivabilidad del estimador coinciden con los del núcleo, por lo que su elección tiene una importancia destacada.

- **Rectangular** (2.13). El más básico. Da el mismo peso a todos los datos dentro de la ventana. La curva de regresión se convierte en una función escalonada que en cada punto es la media de las observaciones que están dentro de la ventana.

$$K_{rect}(u) = \frac{1}{2} I_{[-1,1]}(u). \tag{2.13}$$

- **Triangular** (2.14). En este caso, el peso de las observaciones decae linealmente desde el 0 hasta los puntos ± 1, donde vale 0. Al ser una función continua, la curva de regresión no tendrá saltos.

$$K_{tria}(u) = (1 - |u|) I_{[-1,1]}(u). \tag{2.14}$$

- **Cuadrático** (2.15). Se puede interpretar como una variación del núcleo triangular, donde el decaimiento del peso es cuadrático, en vez de lineal, lo que provoca que la curva final sea más suave. Sin embargo, la función no es derivable en el extremo, por lo que la mejoría no es tan destacada.
$$K_{cuad}(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u). \tag{2.15}$$

- **Tricúbico** (2.16). Esta función es un poco más compleja que las anteriores y, además de cumplir la condición de continuidad, tiene la ventaja de ser derivable en todo su dominio, lo cual es una importante diferencia.
$$K_{tcub}(u) = \frac{70}{81}(1-|u|^3)^3 I_{[-1,1]}(u). \tag{2.16}$$

- **Gaussiano** (2.17). Este caso es un poco diferente, ya que el soporte de la función es $\mathbb{R}$. Por lo tanto, la ventana pierde en cierta medida su significado de que *solo las observaciones a distancia menor o igual a h afectan a la estimación*. Sin embargo, un sencillo cálculo nos indica que el peso de una observación en el límite de la ventana ($u = 1$) es menor que el 5 % del peso de una observación en el centro ($u = 0$). Se puede considerar que, aunque ahora todas las observaciones entran en juego, solo las que se encuentran a distancia menor que h lo hacen de forma significativa. De hecho, más del 99 % del peso total está dentro de la ventana.
$$K_{gauss}(u) = e^{-(2.5u)^2/2}. \tag{2.17}$$

- **Exponencial** (2.18). De forma similar al núcleo gaussiano, esta función también está definida y es no nula en toda la recta real. En este caso, una observación en el límite de la ventana pesa en torno a un 7.5 % del peso en el centro de la ventana, lo cual es algo más que en el caso gaussiano pero en cualquier caso no es demasiado. Además, el 95 % del peso total se encuentra dentro de la ventana, así que de nuevo podemos interpretar que las observaciones que se encuentran a distancia mayor o igual que h tienen una influencia residual en el estimador para cada x.
$$K_{exp}(u) = \frac{3}{2}e^{-3|u|}. \tag{2.18}$$

Por otra parte, los coeficientes que aparecen en casi todos los núcleos son totalmente innecesarios en el cálculo, ya que los pesos son normalizados según (2.12). Si aparecen es solo para que la integral del núcleo sea la unidad, lo cual ayuda a comparar los núcleos entre ellos. Por ejemplo, si se desea ver qué núcleo da más imporatancia relativa y cuál menos a una observación en el centro de la ventana basta tomar $u = 0$ en todos ellos y comparar los valores obtenidos.

En la Figura 2.2 se muestran gráficamente los núcleos expuestos. Aquí cobra importancia la normalización de los mismos, de forma que el área que se extiende debajo de todos ellos es la unidad. Esto nos permite comprobar, por ejemplo, que aunque el núcleo gaussiano y el exponencial se comportan de manera similar en los extremos de la ventana, el exponencial crece de forma más acentuada en el centro, dando mucha más importancia que el gaussiano a las observaciones más cercanas. Esto hará necesariamente que la curva de regresión sea más inestable cuando haya pocas observaciones, para un tamaño fijo de ventana.

Otra observación que puede hacerse al respecto de estas funciones es su continuidad y derivabilidad en los extremos de la ventana. El núcleo rectangular, como ya se ha mencionado, no es continuo y, por tanto, la estimación obtenida con él tampoco lo será. Los núcleos triangular

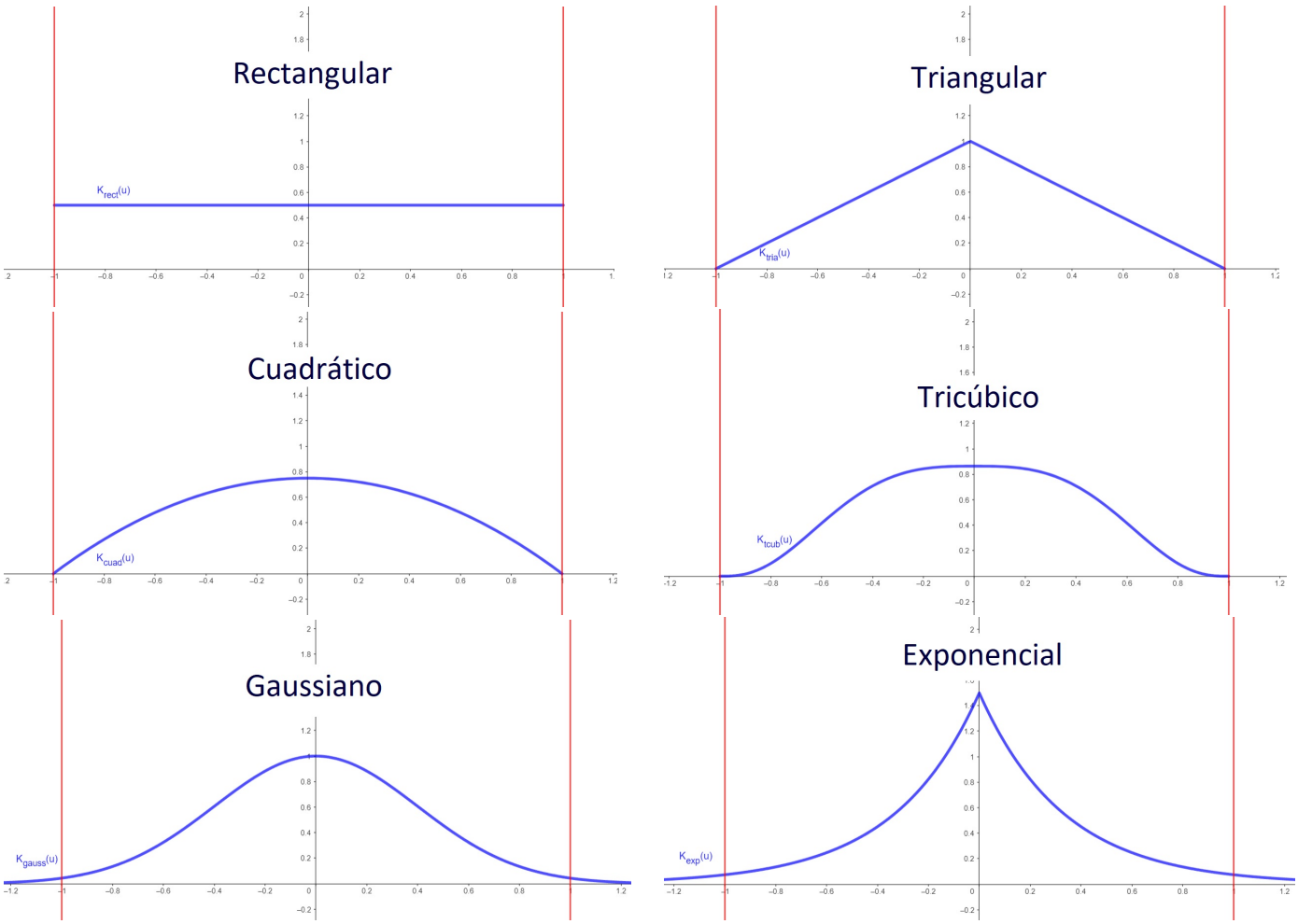

Figura 2.2: Representación gráfica de los 6 núcleos mencionados. Las líneas rojas representan los límites de la ventana. Los cuatro primeros toman valor nulo fuera de la ventana.

y cuadrático, sin embargo, sí que son continuos en los extremos, lo cual permite obtener una curva de estimación continua, pero es evidente que no son derivables en $u = \pm 1$, ya que en la zona interior de la ventana tienen pendiente no nula. Esto no supone un gran problema para la regresión, pero desde luego no es lo ideal, ya que la *entrada de las observaciones en la ventana* no será demasiado suave. El núcleo exponencial, por su parte, no es derivable en el centro $u = 0$, lo cual, de nuevo, es una pequeña incomodidad desde el punto de vista matemático. Por tanto, los únicos núcleos continuos y derivables en todo su dominio son el tricúbico y el gaussiano, este último por ser una función continua y derivable definida en todo $\mathbb{R}$ y el tricúbico porque en los puntos de empalme $u = \pm 1$ cumple que sus derivadas laterales son 0.

En este trabajo se utilizará el núcleo tricúbico (2.16), ya que, además de ser continuo y derivable en $\mathbb{R}$, tiene la pequeña ventaja de que su soporte sea exactamente la ventana de suavizado, a diferencia del núcleo gaussiano. Aunque esta es una 'ventaja' puramente anecdótica, las pruebas con ambos núcleos no muestran diferencias apreciables en la curva de regresión generada, por lo que la elección de uno u otro no es muy relevante.

2.3. Cross Validation

La *cross-validation* o validación cruzada es una técnica de selección de parámetros en problemas estadísticos. En nuestro contexto, se utiliza para elegir la ventana más apropiada.[18]

Como se ha explicado en la sección 2.2.1, la elección de la ventana es clave y el objetivo será minimizar el error cuadrático medio (MSE), definido en (2.9). El problema con esta medida del error es que se trata de la esperanza de una variable aleatoria con distribución desconocida, ya que solo disponemos de una muestra de ella. La técnica de cross-validation hace una estimación de esa esperanza a partir de la muestra disponible.

Una primera idea sobre cómo hacer esta estimación sería seleccionar una rejilla de valores de h que cubra el rango de todos los posibles candidatos y, a continuación, para cada uno de estos valores, calcular el error cuadrático total con la sencilla expresión

$$\mathrm{EC}(h) = \sum_{i=1}^{n} (Y_i - \hat{r}_h(X_i))^2, \tag{2.19}$$

y elegir la ventana que la minimice. Esto sería una mala idea, ya que se estarían utilizando para probar la precisión del ajuste los mismos puntos que intervienen en su cálculo, lo cual haría que el h óptimo fuese el más bajo posible dando lugar a un *overfitting* con $\mathrm{EC} = 0$. Utilizando, por ejemplo, el núcleo rectangular, el resultado sería un h de valor menor que la distancia entre puntos, de tal forma que para cada X_i solo influiría el valor de Y_i.

La conclusión que se obtiene es evidente: Para estimar correctamente el MSE hay que tener en cuenta también el error cometido en las regiones en las que no se dispone de observaciones. Sería ideal poder calcular la regresión para distintos h y, después, tomar nuevas observaciones y ver el error cometido en ellos. Esto es imposible, ya que solo contamos con la muestra original, por lo que hay que pensar alguna forma de adaptar esta idea a nuestras posibilidades. La solución no es compleja, y consiste en trabajar *usando una muestra reducida* y, una vez hechas las regresiones, utilizar las observaciones restantes para estimar el error en ellas. Más precisamente, el procedimiento de cross validation es el siguiente:

1. Elegir una rejilla con los valores posibles para h, $\{h_j\}_{j=1,2,\ldots,m}$.
2. Para cada h_j, utilizar algunas observaciones para construir la curva de regresión (*puntos de entrenamiento*).
3. Calcular la suma de los errores cuadráticos cometidos en el resto de observaciones (*puntos de test*).
4. Repetir los pasos 2[1] y 3 cambiando los conjuntos de test y entrenamiento hasta que todos las observaciones han sido *puntos de test* una única vez.
5. El h óptimo será aquel h_j cuyo error sea mínimo, según se indica en (2.10).

Hay distintas variantes de la cross-validation, entre las que destacan la *Leave-One-Out cross validation* (LOO) y la *k-folds cross validation* [13]. Habitualmente, cuando se habla de cross-validation, la variante referida es la LOO. La diferencia entre ambas radica en la cantidad de puntos incluidos en los conjuntos de entrenamiento y de test, siendo realmente la LOO un

[1]No es necesario construir toda la curva de regresión, basta calcular el estimador (2.7) en los puntos X_i que hagan de test.

caso particular de *k-folds*. Por ello, en la subsección 2.3.1 se explica en detalle el método de *k-folds* y la influencia del valor de k en el mismo. Asimismo, se decidirá razonadamente el valor más apropiado para nuestro problema.

2.3.1. k-folds CV

La base de esta técnica, como su propio nombre indica, es la partición en k partes (*folds*) del conjunto de datos. Desarrollemos este concepto: Una familia de folds, denotada $\{\mathcal{P}_l\}_{l=1,2,\ldots,k}$ es una partición **al azar** del conjunto de observaciones, donde cada (X_i, Y_i) pertenece a un único $\mathcal{P}_l$ y los folds son del mismo tamaño (o tan parecido como sea posible). Ahora, para cada h_j haremos k iteraciones, en cada una de las cuales un fold será el conjunto de test y el resto de folds conformarán el conjunto de entrenamiento. De esta forma, cada observación formará parte del conjunto de test una única vez para cada h_j. Para mayor claridad, en Algoritmo 2.3.1 se muestra el algoritmo de *k-folds cross validation* en pseudocódigo. Aquí, la notación $r_{h_j}^{-l}(X_i)$ hace referencia al estimador de Nadaraya-Watson (2.7) en el punto X_i utilizando las observaciones que **no** están en $\mathcal{P}_l$ para un tamaño de ventana h_j. Notar que la observación (X_i, Y_i) siempre está en $\mathcal{P}_l$, por lo que nunca afecta a $r_{h_j}^{-l}(X_i)$. Por su parte, el peso de la observación t en el punto de abscisas X_i, esto es, $w_t(X_i)$, se calcula según lo indicado en (2.12).

Algoritmo 2.3.1 Pseudocódigo de *k-folds cross validation* para elegir la ventana óptima.

```
Crear una rejilla de m ventanas desde h_1 hasta h_m
Crear una lista de k folds desde P_1 hasta P_k
for cada h_j desde h_1 hasta h_m do
    MSE(h_j) ⟵ 0
    for cada P_l desde P_1 hasta P_k do
        for cada i ∈ P_l do
            Calcular r_{h_j}^{-l}(X_i) = Σ_{t∉P_l} w_t(X_i) Y_t
            MSE(h_j) ⟵ MSE(h_j) + (Y_i − r_{h_j}^{-l}(X_i))^2
        end for
    end for
end for
h_opt = arg mín_{h_j=h_1,h_2,...,h_m} MSE(h_j)
```

En la Tabla 2.1 se muestra esquemáticamente el proceso de cálculo de errores y su suma para entender mejor el procedimiento. El h_j óptimo será el correspondiente al mínimo valor de la última fila. En caso de empate, se tomará el mayor valor de h, ya que la curva será más simple.

En este punto, solo queda preguntarse cuál debe ser el valor de k para el correcto funcionamiento de este método. Estudiemos los casos límite para hacernos una idea de la influencia que tiene parámetro k, teniendo en cuenta que $k = 1$ no tiene sentido, ya que hace falta al menos un conjunto de entrenamiento y otro de test.

- $k = 2$. En este caso, las observaciones están dividas en solo dos conjuntos, por lo que, para cada h_j, se utilizará la mitad de los puntos para construir el estimador en la otra mitad y se medirá el error cuadrático en estos. A continuación, se invertirán los papeles de entrenamiento y test y se sumarán los errores cuadráticos. El problema con tan pocos folds es que pueden existir regiones con todos sus puntos en un fold y ninguno en el otro. Así, cuando ese fold sea el conjunto de test, el estimador en esos puntos será bastante

	h_1	...	h_m
$\mathcal{P}_1$	$\sum_{t\in\mathcal{P}_1}(Y_t - r_{h_1}^{-1}(x_t))^2$	...	$\sum_{t\in\mathcal{P}_1}(Y_t - r_{h_m}^{-1}(x_t))^2$
$\mathcal{P}_2$	$\sum_{t\in\mathcal{P}_2}(Y_t - r_{h_1}^{-2}(x_t))^2$	...	$\sum_{t\in\mathcal{P}_2}(Y_t - r_{h_m}^{-2}(x_t))^2$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathcal{P}_k$	$\sum_{t\in\mathcal{P}_k}(Y_t - r_{h_1}^{-k}(x_t))^2$	...	$\sum_{t\in\mathcal{P}_k}(Y_t - r_{h_m}^{-k}(x_t))^2$
	$\sum_{l=1}^{k}\left(\sum_{t\in\mathcal{P}_l}(Y_t - r_{h_1}^{-l}(x_t))^2\right)$	...	$\sum_{l=1}^{k}\left(\sum_{t\in\mathcal{P}_l}(Y_t - r_{h_n}^{-l}(x_t))^2\right)$

Tabla 2.1: Esquema del proceso de cross validation *k-folds*, con m tamaños de ventana y k folds.

arbitrario. La ventaja de tener pocos folds es el coste computacional, ya que, aunque en cualquier caso hay que construir el estimador en el total de n puntos, este cómputo es más barato cuantos menos puntos de entrenamiento tenemos, y este número aumenta con k según $n - n/k$, ya que n/k es, redondeando, el número de puntos de test en cada iteración.

- $k = n$. Este es el método *Leave-One-Out* (dejar uno fuera), y es sencillo entender el motivo del nombre. Al tener n folds disjuntos y no vacíos, cada fold contiene una única observación, por lo que el conjunto de test es de tamaño uno en todo momento, es decir, que en cada cálculo del estimador se utilizan todos los puntos menos uno (se deja uno fuera). Es un método muy utilizado, ya que permite utilizar la mayor cantidad posible de puntos $(n - 1)$ en el estimador. Sin embargo, tiene el problema de ser el más costoso computacionalmente.

También hay que tener en cuenta, a la hora de seleccionar k, el tamaño n de la muestra, ya que, por ejemplo, el coste computacional se deberá tener más en cuenta cuando n sea grande. En este trabajo se dispone de muestras de en torno a 40 o 50 observaciones, lo cual no es demasiado, por lo que no será necesario preocuparse en exceso por el coste. En [19] se analiza detalladamente el rendimiento de distintos k para distintos tamaños de muestras, y la conclusión que se obtiene es que $k = 5$ y $k = 10$ son los valores óptimos, particularmente para muestras de pequeño tamaño como $n = 50$. Por ello, y porque en la literatura es muy frecuente su uso [20], [21], [22], en este trabajo se utilizará el método de *k-folds cross validation* con $k = 10$. Sin embargo, se hicieron pruebas utilizando LOO y los resultados eran muy parecidos, es decir, el tamaño óptimo de la ventana no cambiaba sustancialmente con respecto al obtenido usando *k-folds* con $k = 10$.

2.4. Bootstrap

En estadística, y en ciencia en general, la estimación del error en los resultados es una parte fundamental del trabajo. Cuando los resultados son mediciones directas de una variable, es posible repetir la medición tantas veces como se desee y utilizar la media como resultado y la desviación típica (o algún otro parámetro de dispersión) como estimación del error. Hay otros muchos casos, como el que nos ocupa, en los que no es posible efectuar el cálculo repetidamente, ya que solo disponemos de un conjunto de datos y el resultado, en este caso, la

curva de regresión, está directamente determinado por ellos.

Para solucionar este problema, en estadística se utilizan métodos de remuestreo (*resampling*). Estos utilizan la muestra aleatoria de la que se dispone para, mediante diferentes técnicas, obtener muestras *ficticias* que se originan con una distribución muy similar a la original, que es desconocida. De este modo, es posible replicar el cálculo del resultado para estas nuevas muestras y, a través de ellas, aproximar la distribución del estadístico que se está utilizando, en nuestro caso, la curva de estimación.

Uno de los métodos de remuestreo más utilizados es el *bootstrap*, propuesto por Efron en [23]. La idea general en este caso es utilizar los residuos de la muestra[2], esto es, la diferencia en cada punto entre la observación y la estimación, junto con la curva obtenida, para generar nuevas muestras bootstrap. A continuación, se usan esas nuevas muestras para crear otras curvas de regresión estimadas. Repitiendo el proceso suficientes veces se genera suficiente base estadística para poder establecer una región de confianza para la curva de regresión. Hay múltiples estudios que sugieren que 1000 repeticiones proporcionan una buena precisión en la estimación de intervalos de confianza [24] [25], además de que computacionalmente es factible hacerlo. De hecho, es habitual en la literatura utilizar este valor.

Se detalla a continuación el procedimiento, concretándolo para el caso que nos ocupa de una muestra bidimensional $\{(X_i, Y_i)\}_{i=1,2,...,n}$. Una vez elegidos el núcleo y la ventana, usando los procedimientos explicados en las secciones 2.2 y 2.3, y $B \in \mathbb{N}$ se procede del siguiente modo:

1. Se calcula la curva de regresión, de forma que para cada $x \in (X_{(1)}, x_{(n)})$ el estimador es $\hat{r}_h(x)$ según lo calculado en (2.7).

2. Los residuos $\{\hat{\epsilon}_i\}_{i=1,2,...,n}$ son las diferencias entre las observaciones Y_i y las estimaciones.

$$\hat{\epsilon}_i = Y_i - \hat{r}_h(X_i), \qquad i = 1, 2, ..., n. \tag{2.20}$$

3. Aquí se produce el tratamiento de residuos, que será distinto en función de la variante del bootstrap usada. En las subsecciones 2.4.1. y 2.4.2. se explicará en profundidad este punto para las variantes *naive* y *wild* del bootstrap, muy utilizadas. Por el momento basta decir que se obtienen B vectores de *residuos bootstrap*, de tamaño n. Se les denotará como $\{\epsilon_i^{*,b}\}_{i=1,2,...,n}^{b=1,2,...,B}$.

4. Utilizando los B vectores de residuos bootstrap, se generan B *muestras bootstrap* bidimensionales, que se denotarán $\{(X_i, Y_i^{*,b})\}_{i=1,2,...,n}^{b=1,2,...,B}$.

$$Y_i^{*,b} := \hat{r}_h(X_i) + \epsilon_i^{*,b}, \qquad i = 1, 2, ..., n \text{ y } b = 1, 2, ..., B. \tag{2.21}$$

5. Para cada $b \in \{1, 2, ..., B\}$, se calcula la curva de regresión $\hat{r}_h^b$ utilizando la muestra $\{(X_i, Y_i^{*,b})\}_{i=1,2,...,n}^{b=1,2,...,B}$ según (2.7), con lo que se obtienen B curvas de regresión.

6. Dado un $\alpha \in (0, 1)$, es posible obtener una región de confianza al $100 \cdot (1 - \alpha)\,\%$ para la curva de regresión. Esto se explicará con más detalle en la subsección 2.4.3.

En las dos siguientes subsecciones se analizan las variantes *naive* y *wild* del bootstrap, incluyendo el motivo de la elección de una u otra, sus limitaciones y su algoritmo. La validez de esta técnica de remuestreo para la regresión no paramétrica está demostrada en [26].

[2]La forma de *utilizarlos* es clave y se explicará más adelante.

2.4.1. Bootstrap Naive

Esta es la versión original del bootstrap [23]. Para su correcto funcionamiento se necesita que los errores estén igualmente distribuidos a lo largo de la muestra. En particular, la muestra ha de ser homoscedástica. La manera de generar los B conjuntos de residuos bootstrap a partir de los residuos originales es la siguiente:

1. Del conjunto $\{\hat{\epsilon}_i\}_{i=1,2,...,n}$ se seleccionan aleatoriamente n elementos con reemplazamiento, generando así el conjunto $\{\epsilon_i^*\}_{i=1,2,...,n}$, donde cada $\hat{\epsilon}_i^*$ es igual a un ϵ_j con i y j no necesariamente iguales. Notar que al tomar los residuos con reemplazamiento normalmente algunos ϵ_i's no serán elegidos y otros aparecerán varias veces.

2. Repetir el paso anterior B veces, con lo que tenemos B vectores de residuos bootstrap $\{\epsilon_i^{*,b}\}_{i=1,2,...,n}^{b=1,2,...,B}$.

Esta es una técnica bastante sencilla, con muy poco coste computacional, y que funciona realmente bien en muchos ámbitos [26]. Sin embargo, es muy importante comprobar que los errores están igualmente distribuidos, y el motivo es muy claro: Si en una región la varianza de las observaciones es muy pequeña y en otra muy grande, la región de confianza deberá ser mayor en esta segunda zona. Sin embargo, esta técnica distribuye aleatoriamente los residuos, por lo que producirá regiones de confianza de anchura más o menos constante.

2.4.2. Wild Bootstrap

Esta variante fue propuesta por Wu en [27], con la idea de poder utilizar la técnica del bootstrap cuando los errores no son i.i.d. La novedad frente a la anterior es que cada nuevo error generado ϵ_i^* depende únicamente del original $\hat{\epsilon}_i$. Se explica a continuación cómo se obtienen los B conjuntos de residuos bootstrap:

- Generar una muestra de tamaño n, $\{z_i\}_{i=1,2,...,n}$ de la siguiente distribución, propuesta por Mammen en [28] y que tiene media 0 y varianza 1.

$$z = \begin{cases} -\frac{\sqrt{5}-1}{2} & \text{con probabilidad } \frac{\sqrt{5}+1}{2\sqrt{5}}. \\ +\frac{\sqrt{5}+1}{2} & \text{con probabilidad } \frac{\sqrt{5}-1}{2\sqrt{5}}. \end{cases} \tag{2.22}$$

- Para cada $i \in \{1, 2, ..., n\}$, el error bootstrap será calculado según

$$\epsilon_i^* = z_i \epsilon_i. \tag{2.23}$$

- Repetir los dos pasos anteriores B veces, de forma que se obtienen los B conjuntos de errores bootstrap.

Como se puede observar, este método mantiene la distribución de la varianza de los datos, por lo que regiones con errores más pequeños tendrán regiones de confianza menores. Por otra parte, como la distribución de (2.22) tiene media 0, la región de confianza no estará sesgada. Además, el hecho de que su varianza sea 1 hace que los nuevos residuos no sean ni mayores ni menores que los originales, en promedio. En conclusión, esta variante es una buena alternativa al bootstrap naive cuando las condiciones no son favorables para este último.

2.4.3. Región de confianza

La obtención de una región de confianza es el motivo por el que se realiza todo el proceso bootstrap. En esta subsección se detalla el procedimiento a seguir una vez ya se tienen las B curvas bootstrap calculadas.

1. Para cada $i = 1, 2, ..., n$ y para cada $b = 1, 2, ..., B$ se calcula el estimador $\hat{r}_h^b(X_i)$.

2. Para cada $i = 1, 2, ..., n$, se aplica el método del Elipsoide de Mínimo Volumen (MVE) al conjunto $\{\hat{r}_h^b(X_i)\}_{b=1,2,...,B}$. Este método encuentra el intervalo de menor longitud que contiene el $100 \cdot (1 - \alpha)\,\%$ de las observaciones. De este modo, se pueden determinar los límites inferior y superior del intervalo de confianza para cada X_i, denotados como $\hat{r}_h^{b_{min}}(X_i)$ y $\hat{r}_h^{b_{max}}(X_i)$, respectivamente.

3. El intervalo de confianza al $100 \cdot (1 - \alpha)\,\%$ para $r(X_i)$ será $[\hat{r}_h^{b_{min}}(X_i), \hat{r}_h^{b_{max}}(X_i)]$.

4. Ahora ya tenemos el intervalo de confianza para cada $r(X_i)$, con $i = 1, 2, ..., n$. Interpolando linealmente los límites obtenidos entre X_i y X_{i+1}, con $i = 1, 2, ..., n-1$, se obtienen los intervalos de confianza para cada $r(x)$.

Es importante destacar que, en este enfoque, no obtenemos una región de confianza continua sobre todo el dominio de X, sino intervalos de confianza para cada X_i individual, y su interpolación permite aproximar el intervalo de confianza en cada x.

Una verdadera región de confianza requeriría aplicar el MVE a las curvas bootstrap en su conjunto, en lugar de aplicarlo para los n puntos en los que existen observaciones. En este caso, el método MVE consideraría cada curva como un punto en un espacio multidimensional, y calcularía el elipsoide de menor volumen que contiene el $100 \cdot (1-\alpha)\,\%$ de los puntos, esto es, de las curvas. A continuación, borrando las curvas que quedan fuera del elipsoide, obtendríamos la región de confianza. El problema es que la dimensión es demasiado alta (coincide con n, el número de puntos de evaluación), por lo que el método necesita demasiadas curvas para funcionar correctamente, y su obtención es muy costosa. Por lo tanto, se optó por aplicar el MVE en cada punto X_i de manera individual, lo que permite trabajar en dimensión 1, convirtiendo así el elipsoide en un intervalo y permitiendo su cálculo con relativamente pocas iteraciones bootstrap.

Capítulo 3

Aplicación práctica

3.1. Elo en ajedrez

La Federación Internacional de Ajedrez (FIDE) es el organismo más importante que regula el ajedrez en el mundo, y clasifica a todos los jugadores federados en función de su rendimiento. Esta clasificación es el Rating Elo, en el cuál los jugadores ganan o pierden puntos en función de los resultados que obtienen en las partidas oficiales. El sistema Elo, basado en técnicas estadísticas, se instauró en los años 70 del siglo XX, y fue desarrollado por Arpad Elo (de quien tomó su nombre) en [29]. En esta sección se pretende explicar brevemente los fundamentos de este sistema, aunque para el trabajo actual basta saber que los puntos de cada jugador en cada momento son una medida de su rendimiento ajedrecístico.

Las puntuaciones Elo de los jugadores son actualizadas por la FIDE periódicamente. Actualmente la periodicidad es mensual, pero originalmente se actualizaba solo dos veces al año, y durante los primeros años de este siglo cada tres meses. Esto es importante, ya que en cada actualización la nueva puntuación depende de la anterior, de los resultados durante el periodo medido y de la puntuación de los rivales enfrentados. Así, los cambios en el Elo tienden a ser más suaves ahora que las actualizaciones son más frecuentes.

La idea básica detrás de este sistema es la siguiente: La variación del Elo de cada jugador tras cada partida depende de *la diferencia entre el resultado obtenido y el resultado esperado* (3.1). Desarrollemos esta idea. En ajedrez, el ganador de una partida obtiene 1 punto, el perdedor 0 puntos y en caso de tablas los jugadores se reparten 1/2 punto cada uno, es decir, el resultado obtenido W será 1, 1/2 o 0. Es en el resultado esperado W_e donde se complican las cosas, ya que esto es más difícil de determinar y depende de los Elos de los dos jugadores. El cálculo del resultado esperado se explicará con más detalle en la Subsección 3.1.1 ya que, aunque no tiene influencia directa en este trabajo, es una técnica estadística muy interesante desde el punto de vista matemático. Por el momento, basta decir que W_e tomará un valor en el intervalo $[0, 1]$ y que la suma de los resultados esperados de los dos jugadores que participan en una partida será 1. El nuevo Elo se calcula con la fórmula

$$R = R_0 + K \cdot (W - W_e), \tag{3.1}$$

donde R es el nuevo rating Elo del jugador en cuestión, R_0 es el rating inicial y $K \in \mathbb{R}$ es un factor que permite ajustar la importancia que se da a la puntuación original y la que se da a la última partida.

En realidad, (3.1) es una simplificación del proceso real, para el caso de una sola partida. En la realidad, como se ha mencionado antes, las actualizaciones son mensuales, por lo que

la variación del Elo depende, en general, de más de una partida. La generalización sería la mostrada en (3.2), suponiendo que el jugador en cuestión ha jugado n partidas en el último mes, cuyos resultados han sido $\{W_i\}_{i=1,2,\dots,n}$ y los esperados eran $\{W_{e,i}\}_{i=1,2,\dots,n}$.

$$R = R_0 + K \cdot \sum_{i=1}^{n} (W_i - W_{e,i}). \tag{3.2}$$

En este caso, R_0 se refiere al Elo con el que empezó el mes el jugador analizado y R es el que tendrá en el nuevo mes. Hay que destacar que los resultados esperados se calculan utilizando el Elo con que cada jugador comienza el mes, sin actualizarse tras cada partida.

Por último, hay que dar un valor al factor K que, a fin de cuentas, determinará la estabilidad del Elo de cada jugador. Esto es muy sencillo de entender, ya que un K bajo implica que en cada mes influye mucho el Elo inicial y poco las partidas jugadas, haciendo que cada jugador tenga un rating muy estable y un K alto otorga mucha importancia a los últimos resultados, lo que haría que las evoluciones tengan muchos altibajos. La FIDE establece diferentes valores de K en función del nivel y del tramo de la carrera de cada jugador [30], los cuales se muestran a continuación:

- $K = 40$ para jugadores que han jugado menos de 30 partidas.
- $K = 40$ para jugadores menores de 18 años que tengan menos de 2300 puntos Elo.
- $K = 20$ para jugadores que no forman parte de ningún caso anterior y nunca han alcanzado 2400 puntos Elo.
- $K = 10$ en el resto de casos.

Para dar contexto a quien no esté familiarizado con el Elo en ajedrez, hoy en día un jugador profesional suele está por encima de los 2400 puntos, y se considera élite a los jugadores que superan los 2700. Menos de 20 personas han superado los 2800 puntos y la máxima puntuación Elo de un jugador en la historia son 2882 puntos (Magnus Carlsen, abril 2014). Una discusión frecuente en ajedrez es el tema de la *inflación del Elo*, que dificulta la comparación de nivel entre jugadores de distintas épocas. La explicación es sencilla: Según (3.1), y dado que en cada partida en que juegan los jugadores A y B se cumple

$$W_e(A) + W_e(B) = W(A) + W(B) = 1.$$

Entonces, si suponemos que K es igual para los dos jugadores (lo cual es cierto para jugadores profesionales), el Elo total de los dos se conserva:

$$\begin{aligned} R(A) + R(B) &= R_0(A) + K \cdot (W(A) - W_e(A)) + R_0(B) + K \cdot (W(B) - W_e(B)) \\ &= R_0(A)+R_0(B)+K\cdot(W(A)+W(B)-W_e(A)-W_e(B)) = R_0(A)+R_0(B). \end{aligned}$$

Por lo tanto, si se crease un grupo de jugadores con el mismo K, por grande que sea, tal que solo jugasen entre ellos, el Elo se intercambiaría entre sí pero no se crearía ni se destruiría. De esta forma, si uno de estos jugadores alcanza en un determinado momento la puntuación máxima histórica de ese grupo se podría decir que ha llegado a ser el mejor del grupo.

La situación real del ajedrez, sin embargo, es diferente, ya que no es posible definir un grupo de jugadores, por enorme que sea, que solo jueguen entre sí, ya que cualquier persona puede apuntarse a torneos y recibe automáticamente una puntuación Elo distinta de 0, con lo cual se

introduce Elo en el sistema. Si cuando este nuevo jugador deja de jugar torneos de ajedrez su puntuación es mayor que la que recibió en un inicio, se habrá destruido Elo, ya que este jugador se habrá llevado puntos del resto de jugadores. Es lo que sucede, por ejemplo, cuando se retiran jugadores profesionales, que destruyen cientos de puntos Elo en su retirada. Si, por el contrario, el jugador se retira con menor puntuación que la que recibió al inicio, se habrá creado Elo en el sistema. Esto es lo que sucede, a pequeña escala, cuando alguien participa puntualmente en un torneo en el que pierde puntos y no vuelve a apuntarse a torneos. Los puntos introducidos en este caso no son muchos, pero es algo muy habitual y al final esa cantidad de Elo introducido por miles de personas acaba *subiendo* en la pirámide de nivel hasta los jugadores de élite.

No es fácil determinar si es superior la creación de puntos o su destrucción, pero se observa que los grandes jugadores de la actualidad tienen más puntos que los de hace unas décadas, lo cual parece indicar que cada vez hay más Elo en la población ajedrecística. Los detractores de esta teoría de la *inflación del Elo* afirman que el motivo por el que la élite tiene cada vez más Elo es que el nivel ajedrecístico sube generación tras generación, y que realmente el Elo es un fiel reflejo del mismo. Es decir, que los puntos Elo *valen lo mismo* ahora que cuando Arpad Elo creó el sistema.

En cualquier caso, en este trabajo se ha decidido utilizar una manera diferente de medir el rendimiento de los jugadores, tomando como referencia los niveles extremos de cada jugador (ver sección 3.2) para estudiar su evolución.

3.1.1. Cálculo del resultado esperado

Para entender cómo se determina el resultado esperado de una partida, en primer lugar hay que comprender que el nivel que un jugador mostrará en una partida, que llamaremos *actuación*, es una variable aleatoria que se ve afectada por factores ajenos a su nivel real, que suponemos medido por su puntuación Elo. El efecto de estos factores se modela por una normal de media 0 y varianza σ^2, por lo que la actuación esperada de un jugador con R puntos Elo será un valor obtenido de la distribución $\mathcal{N}(R, \sigma^2)$. Al desarrollar el sistema en [29], Arpad Elo determinó $\sigma = 200$ mediante el estudio de jugadores y partidas anteriores. Desde entonces se mantiene ese criterio.

Ahora bien, cuando dos jugadores A y B, de Elos R_A y R_B, se enfrentan, su resultado dependerá de sus respectivas actuaciones. Más concretamente, de la diferencia de estas, ya que quien tenga una mejor actuación será el ganador. Como suponemos que las actuaciones tomarán valores aleatorios siguiendo sendas normales $\mathcal{N}(R_A, \sigma^2)$ y $\mathcal{N}(R_B, \sigma^2)$ (ver Figura 3.1), la diferencia de ellas será la variable aleatoria Δ, que seguirá la distribución $\mathcal{N}(D, \sigma'^2)$ (ver Figura 3.2), donde $D = R_A - R_B$ y $\sigma'^2 = 2 \cdot \sigma^2$. Su función de densidad será, por tanto,

$$f_\Delta(x) = \frac{1}{\sigma'\sqrt{2\pi}} e^{\frac{-(x-D)^2}{2\sigma'^2}}. \tag{3.3}$$

Ahora, si Δ toma un valor mayor que 0, entonces el jugador A habrá sido mejor que B y habrá ganado, mientras que si Δ toma un valor menor que 0, el ganador habrá sido B. Por tanto, este modelo adjudica con valor de 0 a la probabilidad de empate y, consecuentemente,

$$W_e(A) = P[\Delta > 0] \qquad \text{y} \qquad W_e(B) = P[\Delta < 0]. \tag{3.4}$$

A modo de orientación, una diferencia de hasta 3 puntos se traduce en un resultado esperado de 0.5 para cada jugador. Con 100 puntos Elo de diferencia, el resultado esperado es

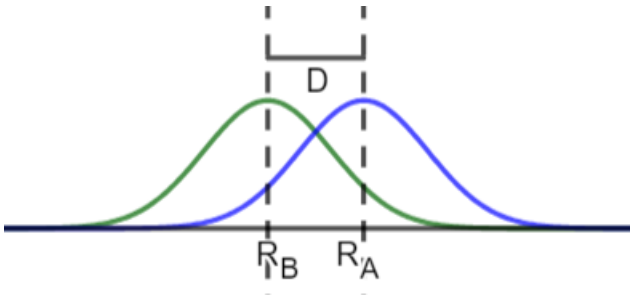

Figura 3.1: Ejemplo de las funciones de densidad de las actuaciones de dos jugadores de Elos R_A y R_B.

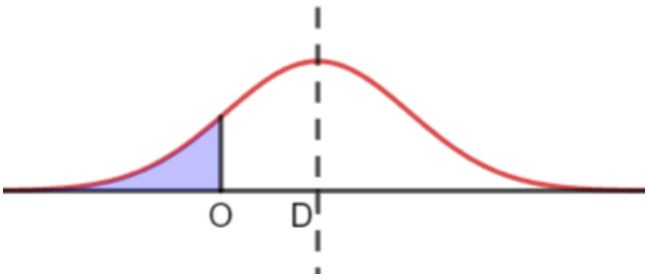

Figura 3.2: Función de densidad de Δ cuando la diferencia de Elos es $D > 0$. La región sombreada representa la probabilidad de victoria del jugador de menor Elo.

0.64 para el jugador con mayor puntuación, y con $D = 400$ se tiene $W_e(A) = 0.92$.

3.2. Rendimiento

Una vez entendido cómo funciona el sistema Elo de ajedrez, en esta sección se explica cómo se van a utilizar las puntuaciones Elo de distintos jugadores para estudiar los aspectos explicados en el Capítulo 1. La clave aquí es que se medirá el rendimiento de cada jugador en distintos momentos de su carrera en relación a sus rendimientos máximo y mínimo. Algunas de las ventajas de hacer esto en lugar de utilizar directamente la evolución de su puntuación Elo son las siguientes:

- Permite comparar jugadores de diferentes épocas eliminando el problema de la inflación.
- Elimina la influencia de la habilidad intrínseca de cada jugador.
- Facilita el estudio del proceso de aprendizaje de los jugadores jóvenes ya que todos llegan al mismo punto, su nivel máximo.

La metodología utilizada para calcular la evolución del rendimiento relativo de un jugador es la siguiente:

1. Tomamos sus puntuaciones Elo en n momentos de su carrera.
2. Si R_i es su Elo en el momento i, definimos $R_{min} = \text{mín}\{R_i : i = 1, 2, ..., n\}$ y $R_{max} = \text{máx}\{R_i : i = 1, 2, ..., n\}$.
3. Ahora, para cada $i = 1, 2, ..., n$, su rendimiento individual será

$$T_i = \frac{R_i - R_{min}}{R_{max} - R_{min}}.$$

Es fácil comprobar que el rendimiento del jugador tomará valores entre 0 y 1, siendo 0 en su peor momento y 1 en el mejor.

Para aumentar la base estadística, se han analizado diversos jugadores y jugadoras, en distintas etapas de su carrera, de forma que con lo que se ha trabajado en la regresión ha sido con los promedios de sus rendimientos individuales. El procedimiento ha sido el siguiente:

1. Elegimos J jugadores y n edades y tomamos las puntuaciones Elo $R_{i,j}$ de cada jugador $j = 1, 2, ..., J$ en cada momento X_i, con $i = 1, 2, ..., n$.

2. Para cada jugador j, utilizamos el método anterior para determinar su rendimiento $T_{i,j}$ en cada i.

3. Para cada $i = 1, 2, ..., n$, calculamos el promedio Y_i del rendimiento individual a esa edad exacta según
$$Y_i = \frac{1}{J} \sum_{j=1}^{J} T_{i,j}.$$

Dado que todos los jugadores analizados tienen mediciones en todos los instantes, es equivalente analizar los n puntos $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ que los $n \times J$ puntos $\{(X_1, T_{11}), ..., (X_1, T_{1J}), (X_2, T_{21}), ..., X_n, T_{nJ})\}$.

De esta forma se consiguen los pares (X_i, Y_i) de los que se ha hablado en el capítulo 2, siendo X_i la edad de los jugadores. En función de lo que se quiera estudiar, se tomarán unos jugadores u otros. En la siguiente sección se expone qué jugadores concretos se han analizado en cada caso.

3.3. Muestra utilizada

Con el fin de lograr los objetivos fijados en el Capítulo 1, se ha decidido tomar medidas de la puntuación Elo de jugadores de ambos sexos a lo largo de toda su carrera deportiva. No es sencillo determinar qué intervalo de edades comprende la carrera deportiva de un ajedrecista, especialmente, cuándo se inicia, ya que cuando alcanzan el nivel de profesionales ya llevan años dedicándose al ajedrez. De hecho, como uno de los objetivos es estudiar el proceso de aprendizaje, lo óptimo sería poder analizar el nivel de los jugadores desde el inicio, es decir, desde que aprenden a jugar. Como es evidente, esto no es posible, ya que la FIDE no mide el Elo de todos los niños que juegan al ajedrez. Por lo tanto, hay que buscar un criterio que determine cuándo se inicia una carrera ajedrecística.

Una opción sería tomar el momento en el que el jugador alcanza un cierto nivel de Elo como el inicio de su carrera. Esta idea, sin embargo, tiene varios problemas. El primero de ellos es que cada jugador comenzaría su carrera a una edad diferente, lo cual nos generaría un dilema a la hora de determinar el promedio de rendimientos: ¿Comparamos los datos en función de la edad o en función del tiempo pasado desde el inicio de la carrera? Además, aunque pudiéramos responder razonadamente a esa pregunta, este criterio tiene un problema, y es que, como cada jugador alcanza un nivel máximo diferente, el *Elo de inicio* que fijamos arbitrariamente tiene un significado distinto para cada jugador. Es decir, quizás un jugador muy talentoso alcance este nivel sin haber comenzado a tomarse en serio el ajedrez, mientras que otro jugador con menos capacidad intrínseca ya lleva años evolucionando para alcanzarlo, y estamos perdiendo información sobre su progreso.

Por ello, se ha decidido elegir otro criterio, también algo arbitrario, pero que parece más útil para los objetivos que se persiguen. Consiste en fijar una edad, $x_1 = 11$ años, como inicio de la carrera ajedrecística. El motivo de escoger esta edad es que con 12 años ya hay jugadores que alcanzan el título de Gran Maestro, el máximo que se puede tener en ajedrez, por lo que, su carrera debe comenzar, al menos, en ese momento. Sin embargo, hasta los 10 años apenas hay datos oficiales de los jugadores, por lo que sería muy díficil conseguir información fiable hasta esa edad.

La edad de fin de la carrera deportiva tampoco es sencilla de determinar, ya que la retirada de los jugadores no es tan clara como en otros deportes, especialmente los colectivos, en los que cuando un jugador deja de pertenecer a un equipo se le puede considerar retirado. En nuestro caso, hay muchos jugadores que a partir de cierta edad dejan de participar en torneos de *élite*, que requieren mucha preparación, pero sí participan en otros de nivel intermedio, o exhibiciones. También hay jugadores veteranos que participan en muy pocos torneos al año, de forma que su Elo se mantiene prácticamente constante por no jugar partidas, y otros que incluso pasan unos años sin jugar una sola partida oficial y vuelven a jugar siendo muy mayores. En este trabajo, se ha fijado los 45 años como el límite, ya que a partir de esa edad el rendimiento parecía estabilizarse en la mayoría de los casos, lo que parece indicar, al menos, una reducción de las partidas oficiales disputadas.

Hemos llegado a la conclusión de que nos gustaría tener datos de jugadores con edades entre los 11 y los 45 años. Aquí aparece un problema, ya que para conseguir esto necesitamos, evidentemente, jugadores de más de 45 años de edad. Y necesitamos también que existan datos de cuando estos jugadores tenían 11 años. Sin embargo, hace 30 años era más difícil que en la actualidad conseguir Elo oficial, ya que la FIDE no tenía en cuenta los torneos menores, por lo que prácticamente nadie lo tenía hasta los 18 años, al menos. La solución adoptada ha sido la siguiente: Dividimos la carrera ajedrecística (y el estudio realizado) en dos partes, una de ellas referida al proceso de aprendizaje, que abarca entre los 11 y los 30 años, y otro referida a la evolución del rendimiento en el más alto nivel, que abarca entre los 20 y los 45 años. De este modo, utilizamos a los jugadores ya retirados para estudiar su carrera completa en la élite, y a jugadores de entre 30 y 40 años para estudiar su proceso de aprendizaje, ya que para ellos sí que existen datos desde los 11 años.

Se han elegido 8 hombres y 8 mujeres para cada rango de edades, es decir, un total de 16 jugadores y 16 jugadoras, que se muestran en la siguiente tabla.

Tramo de edad	Sexo	Nombre	Año de nacimiento	Elo máximo
Joven	Hombre	Magnus Carlsen	1990	2882
Joven	Hombre	Fabiano Caruana	1992	2828
Joven	Hombre	Hikaru Nakamura	1987	2814
Joven	Hombre	Ian Nepomniachtchi	1990	2789
Joven	Hombre	Ding Liren	1992	2811
Joven	Hombre	Wesley So	1993	2810
Joven	Hombre	Sergey Karjakin	1990	2786
Joven	Hombre	Anish Giri	1994	2798
Joven	Mujer	Hou Yifan	1994	2680
Joven	Mujer	Humpy Koneru	1987	2623
Joven	Mujer	Kateryna Lagno	1989	2560
Joven	Mujer	Mariya Muzychuk	1992	2552
Joven	Mujer	Anna Muzychuk	1990	2594
Joven	Mujer	Dronavalli Harika	1991	2539
Joven	Mujer	Elisabeth Paehtz	1985	2493
Joven	Mujer	Alina Kashlinskaya	1993	2505
Veterano	Hombre	Garry Kasparov	1963	2851
Veterano	Hombre	Viswanathan Anand	1969	2817
Veterano	Hombre	Vladimir Kramnik	1975	2809
Veterano	Hombre	Veselin Topalov	1975	2813
Veterano	Hombre	Peter Svidler	1976	2768
Veterano	Hombre	Peter Haine Nielsen	1973	2681
Veterano	Hombre	Alexei Shirov	1972	2751
Veterano	Hombre	Evfeniy Najer	1977	2706
Veterana	Mujer	Maia Chiburdanidze	1961	2551
Veterana	Mujer	Antonaeta Stefanova	1979	2560
Veterana	Mujer	Pia Cramling	1963	2544
Veterana	Mujer	Monika Socko	1978	2499
Veterana	Mujer	Alisa Galiamova	1972	2560
Veterana	Mujer	Elina Danielian	1978	2521
Veterana	Mujer	Almira Skripchenko	1976	2501
Veterana	Mujer	Tatiana Melamed	1974	2386

Capítulo 4

Resultados

Como se ha indicado en el capítulo anterior, hay 4 grupos de jugadores divididos por el tramo de su carrera estudiado y por su sexo. Dependiendo del aspecto que se pretende analizar, se utiliza la curva de regresión de uno u otro grupo. Para hacer más sencilla la interpretación, las curvas referidas a hombres están en rojo, las referidas a las mujeres en verde y las que incluyen datos tanto de hombres como de mujeres en azul.

En este capítulo se muestran las curvas de regresión realizadas, su motivación y se describen brevemente, dejando para el Capítulo 5 un análisis y una comparación más exhaustiva.

Dos de los objetivos definidos en el Capítulo 1 eran determinar la ubicación del punto de máximo rendimiento de los jugadores, es decir, a qué edad se es mejor jugador de ajedrez, y estudiar las crestas y valles en general. Para ello, se han utilizado los datos de todos los jugadores (hombres y mujeres) veteranos, cuya carrera en la élite ya ha terminado. Los resultados aparecen en la Figura 4.1. Como se observa en la gráfica, hay una fuerte subida entre los 20 y los 26 años, aunque escalonada, edad con la que se alcanza lo que podríamos considerar la meseta de rendimiento, que se mantiene estable hasta algo antes de los 40 años. Aunque el punto más alto se alcanza con 37 años, el rendimiento es tan constante a lo largo de la meseta que no se puede asegurar que esa sea la edad óptima para la práctica del ajedrez.

En la Figura 4.2 se muestra la misma información que en la anterior pero separada por sexos, de forma que se perciben las diferencias en la evolución ajedrecística de hombres y mujeres. Y el resultado muestra que, aunque en la Figura 4.1 parecía que los jugadores mantenían un rendimiento constante, esto era un efecto del promedio realizado. En la Figura 4.2 se comprueba que, aunque los jugadores de ambos sexos frenan su crecimiento con 26 años, más o menos, en torno a la edad de 33 los hombres son capaces de subir un poco más su rendimiento, mientras que las mujeres sufren una bajada en el mismo. Es la combinación de la subida y la bajada lo que provocaba la ilusión del rendimiento constante en la figura anterior.

Las curvas anteriores servirán para analizar en profundidad, y diferenciando por sexos, la evolución del rendimiento ajedrecístico en la madurez de los jugadores. Sin embargo, el proceso de aprendizaje comienza mucho antes de los 20 años, por lo que resulta conveniente hacer también las curvas de regresión de los jugadores más jóvenes, para estudiar las diferencias en su crecimiento ajedrecístico desde la infancia.

La Figura 4.3 muestra la evolución de hombres y mujeres, por separado, entre los 11 y los 30 años. En ells se advierte que hacia los 22-23 años el nivel queda prácticamente definido en ambos sexos, ya que desde esa edad hasta los 30 apenas hay variaciones significativas. Comparando

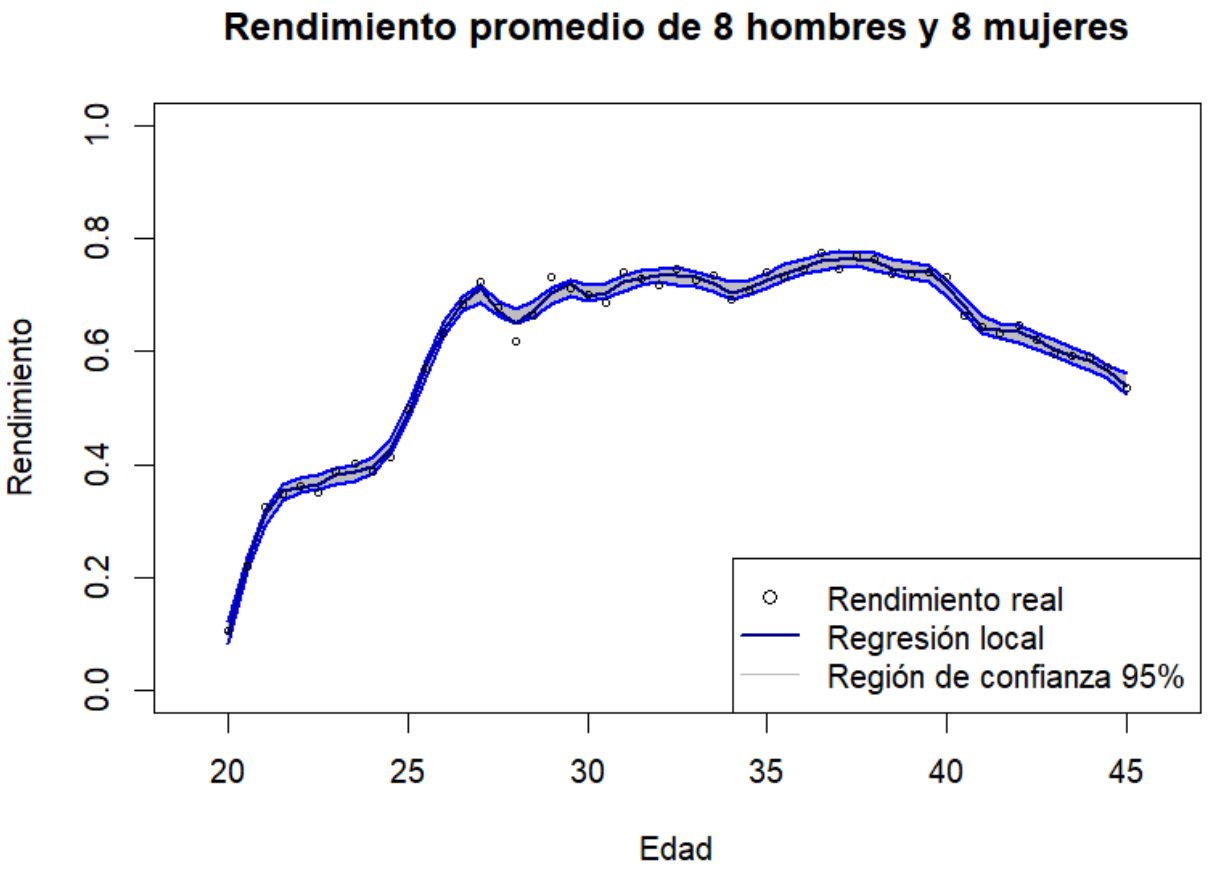

Figura 4.1: Rendimiento promedio de 16 jugadores entre 20 y 45 años.

la curva verde de las mujeres con la roja de los hombres, se puede percibir que la curva verde tiene un crecimiento inicial más pronunciado, y sufre una ralentización en el mismo entre los 15 y los 18 años, para continuar a partir de aquí de modo similar a la roja, mientras que en la roja no se observa ese efecto. De esto se hablará más en profundidad en el siguiente capítulo, pero podría ser un motivo que explique la diferencia de nivel que se observa entre hombres y mujeres en la élite.

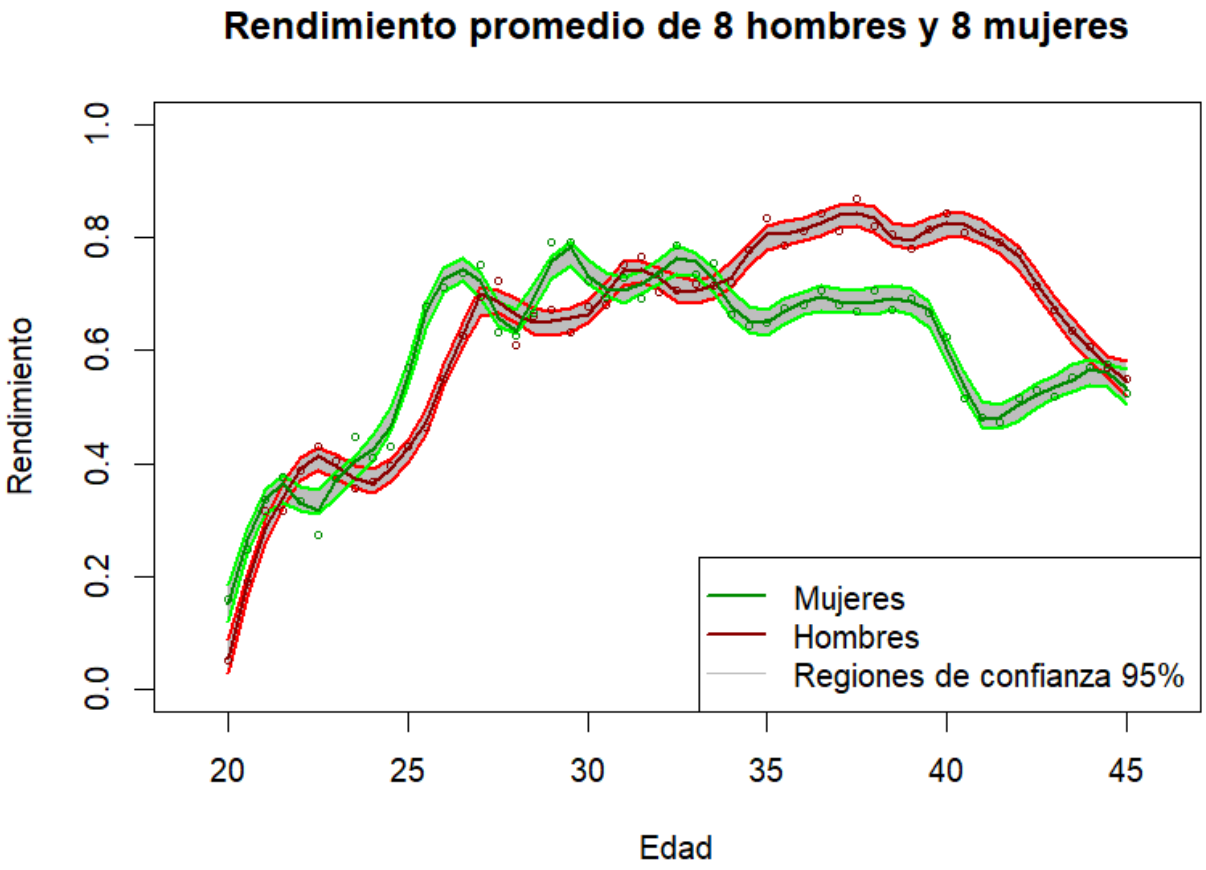

Figura 4.2: Rendimiento promedio de 8 hombres y 8 mujeres entre 20 y 45 años.

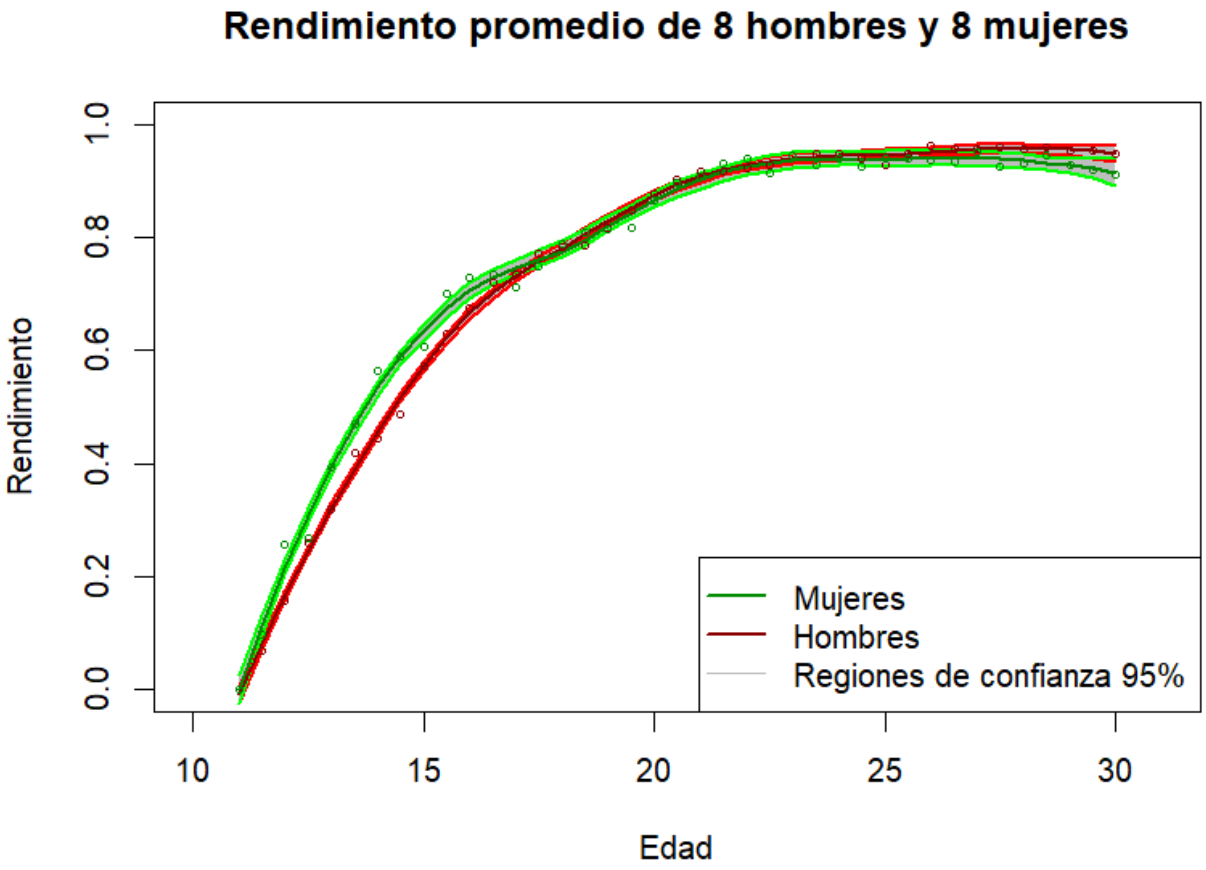

Figura 4.3: Rendimiento promedio de 8 hombres y 8 mujeres entre 11 y 30 años.

Capítulo 5

Discusión

5.1. Evolución y pico de rendimiento

La Figura 4.2 revela patrones distintos en la evolución del rendimiento de los ajedrecistas de ambos sexos, pero también es posible encontrar similitudes entre ellos. La trayectoria tanto de hombres como de mujeres se caracteriza por un importante frenazo en su crecimiento entre los 22 y 25 años. Una posible explicación de esto es que, hasta esas edades, los jugadores tienen la posibilidad de jugar torneos para jóvenes, en los cuales no se enfrentan a jugadores de élite y ganan la mayor parte de sus partidas. Esto podría provocar que su Elo en esos momentos esté ligeramente 'inflado', por lo que cuando salen del circuito juvenil, aunque sigan mejorando, les cuesta más subir su puntuación, ya que el sistema Elo funciona de tal forma que, cuanto mayor puntuación tiene un jugador mejores resultados debe obtener para seguir subiendo.

Hasta los 33 años, más o menos, tanto hombres como mujeres sostienen un rendimiento en torno al 70 % de su máximo, con crestas y valles, pero relativamente constante. Sin embargo, a esa edad sucede algo muy curioso, ya que los hombres logran mejorar ligeramente, hasta superar el 80 % de su rendimiento máximo y lo sostienen durante cerca de 5 años, mientras que las mujeres, por el contrario, sufren una ligera bajada, que se acentúa a partir de los 40. Por este motivo, analizaremos por separado la ubicación del pico de rendimiento en cada sexo.

En el caso de los hombres, está claro que su mejor momento a nivel ajedrecítico se ubica entre los 35 y los 40 años, ya que a partir de entonces la caída es pronunciada. No se puede determinar la edad exacta, ya que aunque el máximo de la curva de regresión se produce en 37, la región de confianza es demasiado ancha para que se pueda determinar con tanta precisión la ubicación del pico. Buscar dónde está el pico, más allá de ser curioso, nos ayuda a caracterizar el ajedrez, comparando este resultado con los que se obtienen en [2]. En este trabajo, los autores obtuvieron una expresión para la edad en que se alcanza el pico de rendimiento en función del tipo de deporte, el sexo de los atletas y la duración de los eventos, en este caso, las partidas. El ajedrez necesariamente debe ser considerado un deporte de resistencia y no de explosividad, que son las categorías descritas en [2], por lo que utilizaremos la ecuación (5.1), expuesta en el citado artículo, que tiene un error de ± 1.8 años.

$$\text{Edad (años)} = 28.7 + 6.0 \cdot \log_{10}[\text{duración (horas)}]. \tag{5.1}$$

Considerando que una partida de ajedrez clásico (la modalidad analizada en este trabajo) dura en torno a las 5 horas, si el ajedrez fuese puramente un deporte de resistencia se esperaría que el pico fuese alcanzado con 32.9 ± 1.8 años. El hecho de que esto se retrase en la realidad debe estar relacionado con la influencia que las capacidades cognitivas tienen en el ajedrez. Según [31], las habilidades mentales no sufren retroceso alguno a lo largo de la vida profesional

de las personas, incluso pasados los 50 años. Parece razonable pensar que, en una disciplina en la que son importantes tanto la resistencia física como la habilidad cognitiva, existirá un descenso del nivel, debido al físico, pero no será tan temprano, como efectivamente se observa en el ajedrez masculino.

En el caso de las mujeres, como se ha mencionado antes, su mejor momento se produce antes, entre los 29 y los 33 años. De nuevo, podemos acudir a [2] para ver qué relación hay en el caso de las mujeres, y nos encontramos con la siguiente ecuación:

$$\text{Edad (años)} = 28.4 + 6.4 \cdot \log_{10}[\text{duración (horas)}], \tag{5.2}$$

cuya incertidumbre es de ± 2.3 años. Suponemos de nuevo que las partidas duran unas 5 horas, y el resultado es que el pico se espera en 32.9 ± 2.3 años. En este caso, como vemos, el resultado esperado coincide con lo obtenido a través del análisis de datos del ajedrez. Esto parece indicar que para las mujeres la resistencia física es más importante que para los hombres, que se benefician más de la experiencia adquirida. Por supuesto, este análisis está sujeto a cuestiones ajenas al propio deporte, como podrían ser circunstancias socio-culturales que hagan que las mujeres, especialmente pasada la treintena, dediquen menos tiempo al ajedrez.

Otro aspecto a comentar, en el análisis del rendimiento femenino, es el repunte que se observa en torno a los 41 años, después del fuerte descenso de nivel que sufren entre los 38 y los 40. Una posible explicación para este fenómeno sería que, debido a su bajada de rendimiento, comparado con el que tenían 10 años atrás, y a que a partir de ciertas edades al cuerpo le cuesta más recuperarse tras los esfuerzos, muchas jugadoras podrían haber empezado a participar en menos torneos, o a torneos de menor nivel, alejándose así de la élite. Esto les permitiría rendir mejor en cada partida, frente a lo que les habría sucedido en los años anteriores. El hecho de observarlo solo en mujeres podría deberse a que la bajada de nivel de los hombres es más suave y tardía, por lo que con 45 años todavía no estarían en ese punto de "retirarse" paulatinamente del ajedrez.

5.2. Periodo de aprendizaje

Es fácil ver que la evolución de los jóvenes es muy parecida para ambos sexos, ya que ambas curvas de la Figura 4.3 muestran una gran subida de nivel entre los 11 y los 22 años, momento en el que tanto hombres como mujeres estabilizan su rendimiento. Esto parece ser contradictorio con lo visto en las gráficas anteriores, que mostraban la evolución del rendimiento entre 20 y 45. Sin embargo, hay que tener en cuenta que el 0 del rendimiento es muy diferente en cada caso, ya que al representar el rendimiento de los veteranos, el mínimo rendimiento lo tienen con 20 años, mientras que en el caso de los jóvenes el mínimo se produce con 11 años, y con 20 ya están cerca del 90 % del máximo. Por lo tanto, está claro que las figuras 4.1 y 4.2 muestran evoluciones de rendimiento para jugadores ya asentados en la élite, mientras que la figura 4.3 representa el aumento de nivel desde la infancia hasta llegar a la élite. De hecho, el último tramo de estas curvas, cuando ya están estabilizadas, no aporta información más allá de la comprobación de que los jugadores ya han llegado al nivel máximo a partir de los 22 años.

Bajo esta premisa de que a los 22 años ya está definido el nivel que un jugador mostrará en su vida, parece razonable estudiar el período anterior a esta edad en hombres y mujeres,

ya que podría explicar el motivo por el que la élite masculina tiene mucho más nivel que la femenina. La Figura 4.3 resulta de gran ayuda en esto, ya que enfrenta ambas curvas de crecimiento y permite compararlas con sencillez. En ella, podemos observar que la curva roja (referida a los hombres) crece de forma suave, acercándose asintóticamente al 1, que representa el rendimiento máximo. Lógicamente, cuanto mayor es el jugador y más puntos tiene, más lento es el crecimiento.

En el caso de la curva verde, sin embargo, sucede un fenómeno muy curioso. El rendimiento sube velozmente hasta los 15 años, más o menos, pero en el periodo entre los 16 y los 18, este crecimiento sufre una importante ralentización y, lo que es más sorprendente, luego se vuelve a acelerar hasta alcanzar el nivel máximo. Es decir, el frenazo en el crecimiento a los 16 no es por estar cerca del nivel máximo alcanzable, porque en ese caso el crecimiento no debería aumentar en los años posteriores. En términos matemáticos, la derivada de la curva roja es siempre positiva pero decreciente, mientras que la derivada de la curva verde, aunque también es siempre positiva, crece en torno a los 18 años. Esto indica que algún factor que solo afecta a las mujeres dificulta su proceso de aprendizaje en la adolescencia. Si damos por bueno que a partir de los 22 años el nivel está definido, esos pocos años que las mujeres "desaprovechan" para mejorar podrían acabar traduciéndose en un nivel inferior a largo plazo.

Determinar los motivos de este estancamiento, desde luego, se escapa de los objetivos y las posibilidades de este trabajo, pero podrían deberse a multitud de factores, incluyendo aspectos psicológicos y socioculturales. Es bien sabido que, especialmente en países en vías de desarrollo, no existe igualdad de oportunidades entre sexos, por lo que mujeres con un gran potencial podrían no disponer de entrenamiento ajedrecístico en esos años, lo que lastraría su desarrollo. Este es solo un ejemplo de las diferentes circunstancias que viven mujeres y hombres en la adolescencia, y que podrían dar lugar a esa ralentización del crecimiento ajedrecístico femenino en una etapa que, según se observa en este trabajo, es clave en la definición del nivel final de los jugadores.

En cualquier caso, este estancaminto podría no ser la única causa de la diferencia de nivel. Podría haber factores biológicos en los que los hombres tengan ventaja y ampliarían esta brecha, además de los factores socioculturales, que siguen teniendo impacto más allá de la adolescencia y también podrían dificultar a las mujeres dar su mejor nivel.

Capítulo 6

Conclusiones

Antes de exponer las conclusiones obtenidas, hay que destacar que el reducido tamaño de la muestra (no era fácil obtener datos fiables de muchos jugadores) hace que las conclusiones que siguen deban tomarse con suma cautela.

La utilidad de las matemáticas va mucho más allá de su estudio teórico y tienen multitud de aplicaciones. En este trabajo, se ha hecho uso de técnicas estadísticas para extraer información sobre un deporte como el ajedrez, así como sobre el desarrollo de la habilidad humana para practicarlo.

Para ello, se ha trabajado con la regresión no paramétrica como técnica de análisis de datos. Asimismo, se ha utilizado la *k-folds cross-validation*, una técnica muy empleada en la estadística moderna, para seleccionar el tamaño de ventana óptimo en cada regresión. Además, se ha construido una región de confianza al 95 % con la ayuda del *bootstrap*.

La utilidad de estas herramientas matemáticas se pone de manifiesto en vista de la información obtenida sobre el ajedrez en este trabajo, que se refleja en las siguientes conclusiones.

El estudio de la evolución ajedrecística de jugadores en la élite ha demostrado que no es un deporte convencional, en el sentido de que la condición física es fundamental, ya que se han producido mesetas en las que el rendimiento es máximo. Esto encaja con la cita de Anatoly Karpov a la que se hacía referencia en la introducción: "El ajedrez es deporte, ciencia y arte". Aunque es más difícil medir la cantidad de arte que encontramos en las partidas, se ha visto que es una disciplina que contiene ingredientes de deporte y de ciencia.

Curiosamente, en el caso de los hombres parece que la habilidad cognitiva, relacionada con el aspecto de *ciencia* cobra más relevancia frente al deporte, ya que a edades tan tardías como los 40 años aún rinden al máximo nivel. En el caso de las mujeres, sin embargo, este estudio sugiere que la resistencia física, es decir, la componente deportiva, parece ser más importante en términos relativos, ya que a partir de los 33 años su rendimiento disminuye sensiblemente. Por supuesto, esto es solo una hipótesis, ya que esta caída podría tener que ver con circunstancias de otra índole.

Por otro lado, este trabajo describe un posible motivo motivo (o, al menos, uno de ellos) de la diferencia entre el nivel de hombres y mujeres en el ajedrez de élite. La ralentización en el proceso de aprendizaje de las mujeres entre los 16 y 18 años ha demostrado ser clave, debido a que a partir de los 22 años el nivel queda estabilizado, por lo que aprovechar los años anteriores puede ser crucial para obtener el máximo nivel posible, como parecen hacer los

hombres. El motivo de esta ralentización, como se ha mencionado previamente, escapa de los objetivos de este trabajo y debería ser abordado desde la biología, la psicología o la sociología, pero sería muy interesante hacerlo en tanto en cuanto podría contribuir a reducir esta brecha.

Más allá de la comparación entre géneros, una de las conclusiones extraidas del trabajo es que tanto los hombres como las mujeres que acaban llegando a la élite evolucionan de forma muy rápida entre la infancia y la adolescencia, ya que a los 22 años han alcanzado su nivel definitivo. Esto es algo que se puede percibir hoy en día, con estudios matemáticos como el que se ha realizado en este trabajo y con la ayuda de la puntuación Elo, que permite ver el nivel de los jugadores. Sin embargo, es muy curioso que José Raúl Capablanca (1888 - 1942), uno de los mejores jugadores de ajedrez de la historia, ya había advertido esto y lo dejó plasmado en la siguiente sentencia: "Creo que a los 30 años el ajedrecista llega a su culminación. Si es suficientemente bueno para ser uno de los mejores del mundo, deberá mostrar sus aptitudes a los 14 años". Parece que con los 30 años se quedó algo corto (sobre todo teniendo en cuenta que en esa época el ajedrez era exclusivamente masculino), pero efectivamente, hoy en día podemos comprobar que, si a los 14 años un joven no ha "mostrado aptitudes", será muy difícil que llegue a ser uno de los mejores del mundo, ya que los futuros miembros de la élite ya han alcanzado el 40 % de su potencial para esa edad.

Por concluir, en este estudio se ha trabajado con el ajedrez clásico, y se han determinado sus características intrínsecas y cómo los jugadores rinden mejor en unas etapas o en otras. Sin embargo, sería muy interesante realizar un estudio similar con otros ritmos de juego del ajedrez, como son el *rapid* (partidas de en torno a 1 hora) y el *blitz* (partidas de unos 10 minutos o menos). En estos casos, especialmente en el *blitz*, se espera que la rapidez de reflejos sea crucial, por lo que los jugadores jóvenes tendrían ventaja frente a los veteranos y el pico de rendimiento debería ubicarse en años anteriores. Sin embargo, factores como la resistencia física juegan un papel menos importante, por lo que jugadores más veteranos podrían aguantar un torneo de ritmo *blitz* mejor que uno clásico. Sin embargo, este ritmo de juego requiere más agilidad mental que los otros, un aspecto en el que también influye el envejecimiento. En suma, no es fácil augurar que conclusiones se obtendrían, por lo que sería un estudio interesante.

Bibliografía

[1] N. J. Ronkainen, T. V. Ryba, and M. S. Nesti. The engine just started coughing! limits of physical performance, aging and career continuity in elite endurance sports. *Journal of Aging Studies*, 27(4):387–397, (2013). ISSN 0890-4065. doi: 10.1016/j.jaging.2013.09.001.

[2] S. Allen and W. Hopkins. Age of peak competitive performance of elite athletes: A systematic review. *Sports medicine (Auckland, N.Z.)*, 44, (2014). doi: 10.1007/s40279-014-0239-x.

[3] C. Capelli, J. Rittveger, P. Bruseghini, E. Calabria, and E. Tam. Maximal aerobic power and anaerobic capacity in cycling across the age spectrum in male master athletes. *European Journal of Applied Physiology*, 116(7):1395 – 1410, (2016). ISSN 1439-6327. doi: 10.1007/s00421-016-3396-9.

[4] N. Vaci, D. Cocić, B. Gula, and M. Bilalić. Large data and bayesian modeling—aging curves of nba players. *Behavior Research Methods*, 51(4):1544–1564, (2019). ISSN 1554-3528. doi: 10.3758/s13428-018-1183-8.

[5] E. Rey, M. Lorenzo-Martínez, R. López-Del Campo, R. Resta, and C. Lago-Peñas. No sport for old players. a longitudinal study of aging effects on match performance in elite soccer. *Journal of Science and Medicine in Sport*, 25(6):535–539, (2022). ISSN 1440-2440. doi: 10.1016/j.jsams.2022.03.004.

[6] E. Dworatzek, I. Baczkó, and G. Kararigas. Effects of aging on cardiac extracellular matrix in men and women. *Proteomics. Clinical applications*, 10, (2015). doi: 10.1002/prca.201500031.

[7] N. Fujii, G. W. McGarr, T. Amano, R. J. Sigal, P. Boulay, T. Nishiyasu, and G. P. Kenny. Ageing augments β-adrenergic cutaneous vasodilatation differently in men and women, with no effect on β-adrenergic sweating. *Experimental Physiology*, 105(10):1720–1729, (2020). doi: 10.1113/EP088583.

[8] Peck E.A. y Vining G.G. Montgomery, D.C. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 111 River Street, Hoboken, Nueva Jersey, (2013). ISBN 9781118627365. URL `https://books.google.es/books?id=lSyiRZh09oEC`.

[9] M. Lunt. Introduction to statistical modelling: linear regression. *Rheumatology*, 54(7):1137–1140, 04 2013. ISSN 1462-0324. doi: 10.1093/rheumatology/ket146.

[10] N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 111 River Street, Hoboken, Nueva Jersey, (1998). ISBN 9780471170822. URL `https://books.google.es/books?id=d6NsDwAAQBAJ`.

[11] A. de Backer, C.S. Becquart, P. Olsson, and C. Domain. Modelling the primary damage in fe and w: influence of the short-range interactions on the cascade properties: Part 2

– multivariate multiple linear regression analysis of displacement cascades. *Journal of Nuclear Materials*, 549:152887, (2021). ISSN 0022-3115. doi: 10.1016/j.jnucmat.2021.152887.

[12] C. Rask, A. Munkholm, L. Clemmensen, M. Rimvall, E. Ørnbøl, P. Jeppesen, and A. Skovgaard. Health anxiety in preadolescence - associated health problems, healthcare expenditure, and continuity in childhood. *Journal of abnormal child psychology*, 44, (2015). doi: 10.1007/s10802-015-0071-2.

[13] C. Loader. *Local Regression and Likelihood*. Springer, Nueva York, (1999).

[14] E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9 (1):141–142, 1964. doi: 10.1137/1109020.

[15] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, (1964). ISSN 0581572X. URL `http://www.jstor.org/stable/25049340`.

[16] W. Härdle. *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press, Cambridge, (1990).

[17] Philippe Vieu. Nonparametric regression: Optimal local bandwidth choice. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):453–464, (1991). ISSN 00359246. URL `http://www.jstor.org/stable/2345755`.

[18] D. Ouyang, D. Li, and Q. Li. Cross-validation and non-parametric k nearest-neighbour estimation. *The Econometrics Journal*, 9(3):448–471, (2006). doi: 10.1111/j.1368-423X.2006.00193.x.

[19] B. G. Marcot and A. M. Hanea. What is an optimal value of k in k-fold cross-validation in discrete bayesian network analysis? *Computational Statistics*, 36(3):2009–2031, (2020). ISSN 1613-9658. doi: 10.1007/s00180-020-00999-9.

[20] Y. Zhao and Y. Hasan. Machine learning algorithms for predicting roadside fine particulate matter concentration level in hong kong central. *Computational Ecology and Software*, 3: 61–73, (2013).

[21] P. Aguilera, A. Fernández, F. Reche-Lorite, and R Rumí. Hybrid bayesian network classifiers: Application to species distribution models. *Environmental Modelling and Software*, 25:1630–1639, (2010). doi: 10.1016/j.envsoft.2010.04.016.

[22] T. Booms, F. Huettmann, and P. Schempf. Gyrfalcon nest distribution in alaska based on a predictive gis model. *Polar Biology*, 33:347–358, (2009). doi: 10.1007/s00300-009-0711-5.

[23] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1 – 26, (1979). doi: 10.1214/aos/1176344552.

[24] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, Florida, USA, (1993).

[25] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, (1997).

[26] F. Ferraty, I. Van Keilegom, and P. Vieu. On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics*, 37(2):286–306, (2010). ISSN 03036898, 14679469. URL http://www.jstor.org/stable/41000880.

[27] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, (1986). ISSN 00905364. URL http://www.jstor.org/stable/2241454.

[28] E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):255–285, (1993). ISSN 00905364. URL http://www.jstor.org/stable/3035590.

[29] A. E. Elo. *The Rating of Chess players, Past and Present*. Arco Pub., New York, (1978). ISBN 0668047216 9780668047210. URL http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216.

[30] FIDE. Fide rating calculators. https://ratings.fide.com/calc.phtml?page=change. Último acceso: 04/06/2024.

[31] J. C. van Ours. Will you still need me when i'm 64? IZA Discussion Papers 4264, Institute of Labor Economics (IZA), (2009). URL https://ideas.repec.org/p/iza/izadps/dp4264.html.