FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

ANÁLISIS ECONOMÉTRICO DE LAS APUESTAS EN EL MUNDO DEL FÚTBOL

Curso 2021/2022

Autor: Alejandro Martínez Ibáñez

Doble Grado en Derecho y Administración y Dirección de Empresas

Econometría

Tutor: Ángel Sánchez Barbie

Elche

Junio 2022

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Las apuestas deportivas se sostienen sobre la base de la estadística y la probabilidad, algo que es desconocido por la gran mayoría de los apostadores. Su funcionamiento es más complejo del que aparenta, y su conocimiento puede ayudar a ser conscientes de la dificultad que estas conllevan.

En este sentido, a través del presente trabajo se explica cómo funcionan las apuestas, intentando ajustar un modelo que permita predecir las probabilidades de que ocurra un suceso en un evento deportivo a través de una serie de variables que consideramos relevantes. Este modelo es analizado gráfica y analíticamente, se comprueba su aceptación a través del análisis de los residuos y se comparan los diferentes modelos obtenidos. Los resultados obtenidos difieren de los esperados, y nos permiten demostrar la dificultad de crear un modelo de estas características.

Palabras clave: apuestas, regresión lineal múltiple, modelo, análisis econométrico.

ABSTRACT

Sports betting is based on statistics and probability, something that is unknown to the vast majority of bettors. Its operation is more complex than it appears, and its knowledge can help to be aware of the difficulty that these entail.

In this sense, through the present work it is explained how betting works, trying to adjust a model that allows predicting the probabilities of an event occurring in a sporting event through a series of variables that we consider relevant. This model is analyzed graphically and analytically, its acceptance is verified through the analysis of the residuals and the different models obtained are compared. The results obtained differ from those expected, and allow us to demonstrate the difficulty of creating a model with these characteristics.

Keywords: gambling, multiple linear regression, model, econometric analysis

CAPÍTULO 1: INTRODUCCIÓN Y OBJETIVOS

1.1 INTRODUCCIÓN

Las apuestas deportivas han crecido de forma muy acentuada estos últimos años, especialmente con la llegada del apuestas por internet, que han experimentado un crecimiento de 15% en comparación con el año anterior[1]. Estos son unos datos alarmantes para la sociedad principalmente por dos motivos.

En primer lugar, porque este crecimiento de las apuestas lleva aparejado un aumento aún mayor de los problemas de ludopatía y dependencia del juego que de la misma se derivan, creciendo de forma drástica[2].

En segundo lugar, porque cada vez hay más gente que apuesta, lo que es especialmente preocupante cuando se trata de gente joven, o incluso, menores de edad que no según la legislación española no pueden apostar legalmente. En este sentido, es identificativo el dato de que el 28% de los jugadores comenzaron a apostar cuando aún eran menores de edad[3].

A nuestro entender, una de las principales razones por las que esto sucede es porque las personas no conocen realmente qué son las apuestas deportivas. Ven como fácil y simple algo que realmente es muy complicado, y que detrás lleva mucho estudio y análisis.

De alguna forma, se infravalora el poder de las casas de apuestas, y mucha gente se cree capaz de ganarles. Sin embargo, la realidad es diferente: de los 1,36 millones de españoles que apostaron online en el año 2019, el 80% perdió dinero, y únicamente el 0,81% obtuvo más de 3.000€ de ganancias netas[4].

[1]Consejo Empresarial del Juego (2021) Anuario del Juego en España.
[2] El Confidencial (2020). La ludopatía vuelve a crecer durante el periodo de pandemia del coronavirus. https://www.elconfidencialdigital.com/articulo/tendencias/ludopatia-vuelve-crecer-periodo-pandemia-coronavirus/20201001132515168320.html
[3] Dirección General de Ordenación del Juego (2015). Estudio sobre prevalencia, comportamiento y características de los usuarios de juegos de azar en España 2015.
[4] Dirección General de Ordenación del Juego (2019). Memoria anual.

Y es que, cada año, las casas de apuestas ganan cantidades ingentes de dinero a costa de esta gente, y con ella, perfeccionan sus métodos de estimación de probabilidades y recogida de información, vitales para poder seguir ganando.

La gran mayoría de gente no sabe, que, desde el momento que empiezas a jugar, -y tal como explicaremos y comprobaremos en el presente trabajo-, la casa ya parte con ventaja, y su probabilidad de ganar en el largo plazo es mucho mayor que la que pueda tener cualquier apostante.

Por ello, podemos decir las apuestas no son un juego. Es algo muy serio que puede llegar a tener muchas implicaciones tanto sociales, como económicas o incluso psicológicas. Por tanto, entender y comprender la dificultad que entrañan las apuestas deportivas puede llegar a ser esencial para que no se traten las apuestas como un simple juego, y provoque que la participación en las mismas sea lleva a cabo de forma consciente e informada. Lo cual intentaremos demostrar con el presente trabajo.

2.1 OBJETIVOS

Son objetivos del presente trabajo:

1. Entender en profundidad cómo funcionan las apuestas deportivas.
2. Identificar variables que puedan afectar a la determinación de la probabilidad de que suceda un determinado resultado en un evento deportivo.
3. Crear un modelo econométrico que permita predecir la probabilidad que ocurra un determinado resultado en un evento deportivo, a partir de unas variables y parámetros.

CAPÍTULO 2: APROXIMACIÓN AL FUNCIONAMIENTO DE LAS APUESTAS DEPORTIVAS

El objetivo de esta primera parte del trabajo es entender de una manera profunda el mundo de las apuestas deportivas, de forma que nos permita conocer su funcionamiento y los factores que dependen de ella, para posteriormente poder ser aplicados al análisis econométrico. Como veremos, no es un funcionamiento sencillo, lo que dificulta su conocimiento por parte de la población.

2.1 EL JUEGO

El juego[5], en la forma que a nosotros nos interesa y estudiaremos a lo largo de este trabajo, es una actividad recreativa donde intervienen uno o más participantes. Su principal función es proporcionar diversión y entretenimiento a los jugadores, y pueden cumplir diversos roles: educativo, ayudar al estímulo mental y físico, y contribuir al desarrollo de las habilidades prácticas y psicológicas, e incluso, obtener unas ganancias.

En el libro "*Anatomía del Juego: Un análisis comparativo de las posibilidades de ganar en los diferentes juegos de azar" (*Miguel Córdoba Bueno, 2013), su autor divide los juegos en 4 grandes categorías:

1. Los juegos pasivos dependientes exclusivamente del azar: las loterías. Son pasivos en tanto en cuanto el jugador no participa físicamente en los mismos, puesto que son juegos de observación; y dependen exclusivamente del azar puesto que no depende de la habilidad o destreza de los jugadores, sino únicamente de la suerte.

2. Los juegos pasivos dependientes del resultado de eventos deportivos: las apuestas deportivas. No dependen de la suerte, sino de las habilidades y destrezas de otros actores distintos al jugador, que son los que participan en el evento. Y es pasivo puesto que el jugador no participa físicamente, sino que basa su juego en la observación de dicho evento.

[5] https://definicion.de/juegos-recreativos/

3. Los juegos iterativos dependientes exclusivamente del azar: los juegos del casino, como la ruleta, los dados o las máquinas tragaperras. En este caso, a pesar de que el resultado depende exclusivamente de la suerte, el jugador puede participar de manera física cuantas veces desee, puesto que el juego se repite constantemente.

4. Los juegos competitivos: como podría ser el tres en raya o "el mentiroso". El resultado depende exclusivamente de las destrezas y habilidades del jugador y las de su/s adversario/s.

2.2 LAS APUESTAS

Según la RAE, una de las acepciones de la palabra "apostar" es:

> *"Arriesgar cierta cantidad de dinero en la creencia de que algo, como un juego, una contienda deportiva, etc., tendrá tal o cual resultado; cantidad que en caso de acierto se recupera aumentada a expensas de las que han perdido quienes no acertaron".*

Se trata de una definición que nos permite comprender el concepto de manera clara a través de diferentes elementos clave que incorpora:

- *Se arriesga una cierta cantidad de dinero*: para ganar dinero, tienes que arriesgar también dinero, puesto que en el lado contrario, también hay alguien (normalmente en el modo en que lo abordaremos en este trabajo, las casas de apuestas) que arriesga dinero para ganar en caso de que el apostador pierda, y el cual pierden para el caso de que el jugador gane.

- *Creencia de un resultado*: a pesar de que este no depende del completamente del azar, como hemos visto anteriormente, tampoco depende del apostador. Por lo que el apostador únicamente puede creer que ocurrirá una cosa u otra, sin poder llevar a cabo nada para favorecer ese resultado creído. Evidentemente, cuanto más datos tenga y más exhaustivo

sea su análisis, mayor va a ser la creencia de ese resultado, y más se ajustará a la realidad de que ocurra.

- En caso de acertar, *la cantidad de dinero arriesgada se recupera de forma aumentada a expensas de lo que han perdido quienes no acertaron:* es decir, el beneficio de los que aciertan viene determinado por el riesgo que asumen los que fallan, como explicaremos y trataremos en los siguientes apartados.

2.2.1 Tipos de apuestas

En nuestro caso, la modalidad que nos interesa y que analizaremos son los apuestas deportivas, especialmente aquellas que se pueden realizar en partidos de fútbol. Estas pueden clasificarse en diversos tipos a tenor de distintos criterios.

2.2.1.1 Según el momento: pre-partido o en directo.

- Apuestas pre-partido: son apuestas realizadas antes del inicio del evento. Vendrán determinadas por la probabilidad que las casas de apuestas asignen a cada suceso, y como analizaremos posteriormente, por el comportamiento de los apostadores con respecto a ese suceso.

- Apuestas en directo (apuestas *live*): son apuestas que se realizan durante la celebración del evento: desde que empieza hasta que finaliza. En este caso, las cuotas varían con mucha más frecuencia, tanto en cantidad como en número de veces, condicionadas por los distintos sucesos que vayan ocurriendo en el evento.

Por ejemplo, en el caso del fútbol, dependerá de la cantidad de disparos que durante el partido vayan haciendo los equipos, los disparos a puerta, los córners, las tarjetas rojas, el resultado en un determinado momento, etc...

A lo largo del trabajo, trabajaremos y analizaremos apuestas pre-partido.

2.2.1.2 Según la casa de apuestas: de intercambio u ordinarias.

- Apuestas de intercambio: son las que se llevan a cabo en las casas de apuestas de intercambio, que se explicará en el apartado *"1.4.1 Tipos de casas de apuestas: de intercambio u ordinarias"*.

- Apuestas ordinarias: son las apuestas que se llevan a cabo en las casas de apuestas ordinarias, que se explicará en el apartado "*1.4.1 Tipos de casas de apuestas: de intercambio u ordinarias".*

A lo largo del trabajo, trabajaremos con las apuestas ordinarias.

2.2.1.3 Según las opciones: de 2 opciones, 3 opciones o abiertas.

- Apuestas de 2 opciones: el resultado sólo puede ser positivo o negativo, es decir, sólo hay dos posibilidades, las cuales son excluyentes y no pueden darse a la vez.

Por ejemplo, en un partido de tenis: o gana un tenista, o gana el otro. No existe la posibilidad de empate. En el fútbol también existen este tipo de apuestas, por ejemplo, las que consistente en que en un partido haya más o menos de 2,5 goles.

	Resultado: 3-0	**Resultado: 0-0**
Apuesta: + 2,5 goles	**Apuesta ganada**	*Apuesta perdida*
Apuesta: - 2,5 goles	*Apuesta perdida*	**Apuesta ganada**

Tabla 1. Tipos de apuestas con 2 opciones. Elaboración propia.

- Apuestas de 3 opciones: el resultado puede tener 3 opciones distintas, de forma que habrá una opción ganadora, y 2 opciones perdedoras.

Es el caso, por ejemplo, de apostar al resultado de un partido de fútbol: podrá ganar el local, empatar o ganar el visitante.

	Resultado: 1-0	Resultado: 1-1	Resultado: 0-1
Apuesta: gana local	**Apuesta ganada**	*Apuesta perdida*	*Apuesta perdida*
Apuesta: empate	*Apuesta perdida*	**Apuesta ganada**	*Apuesta perdida*
Apuesta: gana visitante	*Apuesta perdida*	*Apuesta perdida*	**Apuesta ganada**

Tabla 2. Apuestas de 3 opciones. Elaboración propia

- Apuestas abiertas: son apuestas en los que únicamente un resultado será el ganador, entre muchas opciones.

Es el caso, por ejemplo, de apostar a un resultado exacto en un partido de fútbol. Si el resultado final es de 1-1, únicamente ganará el que haya apostado a ese resultado, y perderán todos los que hayan apostado a un resultado diferente (1-0, 2-0, 2-1, 0-0, 2-2, 0-1, 0-2, 1-2, etc...).

A lo largo del trabajo, nos centraremos en las apuestas de 2 y 3 opciones.

2.2.2.4 Según la cantidad de apuestas: individuales o múltiples.

- Apuestas individuales: la apuesta consiste en una única predicción. Por ejemplo: en el partido Betis vs Levante, se apuesta a que el resultado final será de empate.

- Apuestas múltiples: la apuesta está formada por distintas predicciones, que puede ir de 2 a 14 (normalmente es el máximo permitido por las casas de apuestas).

Se trata de sucesos independientes en los que, para ser ganadora la apuesta múltiple, todas las apuestas deben resultar ganadoras por separado. De esta forma, la probabilidad de que ocurra disminuye cuanto más apuestas seleccionemos, a la par que aumenta el beneficio potencial en caso de resultar ganadora. Cabe mencionar que, puesto que se trata de sucesos independientes, cada una de las apuestas que se ponen en conjunto deben pertenecer a partidos diferentes.

Al tratarse de sucesos independientes:

$$P\ (A \cap B) = P\ (A) * P(B)$$

Como veremos a continuación, la probabilidad de un suceso en las apuestas deportivas viene representada por la cuota (riesgo que asume la casa de apuestas en caso de que el apostador gane la apuesta).

Escogemos dos partidos al azar, en el que las apuestas y las cuotas con las siguientes:

Apuesta	Cuota	Probabilidad
Gana Real Madrid	1,78	0,5618

Apuesta	Cuota	Probabilidad
Gana Celta de Vigo	2,1	0,4761

Una vez tenemos la probabilidad de cada suceso, calculamos la probabilidad de que sucedan ambos sucesos:

$$P\ (Gane\ R.Madrid \cap Gane\ Celta\ ) = P\ (Gane\ R.Madrid) * P(Gane\ Celta)$$
$$P\ (Gane\ R.Madrid \cap Gane\ Celta\ ) = 0{,}5618 * 0{,}4761 = 0{,}2675$$

	Cuotas	Probabilidad
Gana Real Madrid	1,78	0,5618
Gana Celta de Vigo	2,1	0,4761
TOTAL	3,738	0,2675

Tabla 3. Relación cuotas y probabilidad. Elaboración propia

A lo largo del trabajo, trabajaremos sobre las apuestas individuales.

2.3 LAS CUOTAS

Es un concepto con gran relevancia en las apuestas, que podemos definir desde 2 puntos de vista diferentes:

1) Desde el punto de vista de los apostadores: la cuota es la ganancia que tendremos en caso de acertar nuestra apuesta, y que variará dependiendo de la cantidad de dinero que nosotros hemos apostado.

- En caso de fallar (no acertar) la apuesta, la cuota es irrelevante para el apostador, ya que tan sólo importa la cantidad apostada, que es lo que se perderá.

- En caso de acertar la apuesta, para el apostador sí son relevantes tanto la cuota como la cantidad apostada, ya que ambas tienen relación directamente proporcional con las ganancias, y de este modo, con los beneficios, que vienen determinados de la siguiente forma:

$$Beneficio = (Cantidad\ apostada * Cuota) - Cantidad\ apostada$$

Cuota	Cantidad Apostada	Ganancia	Beneficio
1,3	10€	13€	3€
2,3	10€	23€	13€
3,3	10€	33€	23€

Tabla 4. Cuotas y cantidad apostada (1)

Por tanto, comprobamos para una misma cuota: a mayor cantidad apostada, mayor ganancia y mayor beneficio.

Cuota	Cantidad Apostada	Ganancia	Beneficio
2,3	10€	23€	13€
2,3	20€	46€	26€
2,3	30€	69€	39€

Tabla 5. Cuotas y cantidad apostada (2)

Por tanto, comprobamos que para una misma cantidad apostada: a mayor cuota, mayor ganancia y mayor beneficio.

2) Desde el punto de visto de las casa de apuestas es el caso completamente opuesto al anterior: la cuota es el riesgo que asume la propia casa si el apostador gana su apuesta.

- En caso de que el apostador falle (no acierte) la apuesta, la cuota es de nuevo irrelevante, importando únicamente la cantidad apostada, que es lo que ganará la casa de apuestas.

- En caso de que el apostador acierte la apuesta, sí son relevantes tanto la cuota como la cantidad apostada, teniendo en este caso una relación directamente proporcional con las pérdidas.

De esta forma, podemos decir que:

$$Ganancia\ de\ la\ casa\ de\ apuestas = Riesgo\ del\ apostador$$
$$Ganancia\ del\ apostador = Riesgo\ de\ la\ casa\ de\ apuestas$$

De manera gráfica, esto puede ser representado de la siguiente forma:

	Cuota	**Cantidad Apostada**	**Beneficio Casa de Apuestas**	**Beneficio del Apostador**
Apostador acierta	2,3	10€	-23€	23€
Apostador falla	2,3	10€	10€	-10€

Tabla 6. Beneficios según acierto o fallo

Así, en este ejemplo, podemos decir que:

· Desde el punto de vista de la casa de apuestas, arriesga 23€ (cantidad apostada x cuota) para ganar 10€ (cantidad apostada por el apostador).

· Desde el punto de vista del apostador, arriesga 10€ (cantidad apostada) para ganar 23€ (cantidad apostada x cuota).

2.3.1 Probabilidad

Objetivamente, la cuota que la casa de apuestas ofrece es una medida inversa de la probabilidad [6] de un resultado.

$$Probabilidad\ (x) = \frac{1}{Cuota\ (x)}$$

Por tanto,

$$Cuota\ (x) = \frac{1}{Probabilidad\ (x)}$$

Lo normal es que la suma de las probabilidades de que ocurran todos los sucesos contrarios sea igual a 1. Sin embargo, esto no ocurre en las casas de apuestas.

Por ejemplo, cogemos las cuotas de un partido al azar en una casa de apuestas:

	Cuota	**Probabilidad**	**Probabilidad acumulada**
Gana Local	1,4	0,7142	0,7142
Empate	4,33	0,2309	0,9451
Gana Visitante	9	0,1111	1,0562

Tabla 7. Probabilidad y probabilidad acumulada

Así pues, comprobamos que la probabilidad teórica que la casa de apuestas le asigna al suceso es mayor que 1, concretamente 1,0562.

¿Cómo es esto posible? Y sobre todo, ¿qué significa que la probabilidad teórica que la casa de apuestas le asigna a un suceso sea mayor que 1?

[6] *A priori real, pero como veremos más adelante, no es así.*

Este hecho viene explicado por la rentabilidad de las casas de apuestas, y su esperanza matemática, y tiene una consecuencia muy importante: la casa de apuestas juega con una ventaja probabilística con respecto a los apostadores. Toman así especial importancia, además de los anteriores conceptos, otros como la probabilidad corregida y las cuotas corregidas, que pasaremos a explicar a continuación.

2.3.2 Probabilidad real o corregida

Como hemos comprobado, las cuotas, a pesar de que reflejan la probabilidad de que ocurra un determinado suceso, no reflejan su probabilidad real, debido al sobreprecio que las casas de apuestas aplican para aumentar su rentabilidad y mantenerla estable en el largo plazo.

Es importante, por ello, determinar la probabilidad corregida del suceso, para de esta forma conocer su probabilidad real y por tanto, la cuota real que tendría que tener en base a esta.

$$Probabilidad\ corregida\ (x) = \frac{\frac{1}{Cuota\ (x)}}{\sum_{i=1}^{n} \frac{1}{Cuota\ (i)}}, para\ n = 1, X, 2$$

Siguiendo el mismo ejemplo de antes (TABLA X), para la victoria del equipo local se le asigna una cuota de 1,40, que llevaba implícita una probabilidad del 0,7142. Sin embargo, la cuota corregida se determinaría de la siguiente forma:

$$PC\ (vict.local) = \frac{\frac{1}{Cuota\ (vict.local)}}{\frac{1}{Cuota\ (vict.local)} + \frac{1}{Cuota\ (empate)} + \frac{1}{Cuota\ (vict.vis)}}$$

$$PC\ (vict.local) = \frac{\frac{1}{1{,}40}}{\frac{1}{1{,}40} + \frac{1}{4{,}33} + \frac{1}{9}}$$

$$PC\ (vict.local) = \frac{\frac{1}{1{,}40}}{\frac{1}{1{,}40} + \frac{1}{4{,}33} + \frac{1}{9}}$$

$$PC\ (vict.local) = \frac{0{,}7142}{0{,}7142 + 0{,}2309 + 0{,}1111}$$

$$PC\ (vict.local) = \frac{0{,}7142}{1{,}0562} = 0{,}6762$$

De igual forma se determinarán la probabilidad corregida de los otros dos sucesos (empate y victoria visitante), de forma que la tabla con las probabilidades corregidas quedaría de la siguiente forma:

	Cuota	**Probabilidad**	**Probabilidad corregida**
Gana Local	1,4	0,7142	0,6762
Empate	4,33	0,2309	0,2187
Gana Visitante	9	0,1111	0,1051
TOTAL		1,0562	1,0000

Tabla 8. Probabilidad y probabilidad corregida

Una vez determinada esta probabilidad corregida, podemos determinar las cuotas corregidas.

2.3.3 Cuotas corregidas o reales

Son las cuotas asociadas a la probabilidad corregida, es decir, a la probabilidad real sobre 1 (y no sobre la probabilidad asignada por las casas de apuestas, la cual recordamos que es en su conjunto >1). Así, la determinación de las cuotas corregidas viene determinada por la siguiente fórmula:

$$Probabilidad\ corregida\ (x) = \frac{1}{Cuota\ corregida\ (x)}$$

Por tanto,

$$Cuota\ corregida\ (x) = \frac{1}{Probabilidad\ corregida(x)}$$

De esta forma, siguiendo el ejemplo anterior:

	Cuota	Probabilidad	Probabilidad corregida	Cuota corregida
Gana Local	1,4	0,7142	0,6762	1,48
Empate	4,33	0,2309	0,2187	4,57
Gana Visitante	9	0,1111	0,1051	9,51
TOTAL		1,0562	1,0000	

Tabla 9. Cuotas corregidas

Por tanto, si hemos establecido que las cuotas tiene 2 posibles visiones, tanto desde el punto de vista del apostador como de las casas de apuestas, podemos decir que el hecho de que la cuota corregida sea siempre mayor que la cuota real supone que:

- Para la casa de apuestas, al representar la cuota la cantidad que perdería, supone un hecho positivo, puesto que perderá menos de lo que realmente debería.

- Para el apostador, al representar la cuota la cantidad que ganaría, supone un hecho negativo, puesto que ganará menos de lo que realmente debería.

2.3.4 Rentabilidad de las casas de apuestas

A priori, las apuestas serían lo que se conoce como "juego de suma cero". Es decir, si una persona gana, lo hace en detrimento de otra u otras personas que, por sí solas o en su conjunto, pierden una cantidad equivalente a la obtenida por el jugador ganador.

Sin embargo, esto no es así, puesto que como hemos comprobado en los apartados anteriores, las casas de apuestas juegan con la probabilidad en su favor: en primer lugar, porque asignan una probabilidad conjunta al hecho mayor que 1 (es decir, mayor que la

probabilidad real de que ocurra el suceso); y en segundo lugar, porque tienen el poder de variar el mercado de cuotas para que esta rentabilidad sea constante y no varíe, de forma independiente al resultado.

Es por ello que la rentabilidad de la casa de apuestas en el largo plazo es siempre positiva (en detrimento de la rentabilidad de los apostadores), puesto que su esperanza matemática es también positiva (en detrimento también de la de los apostadores, que es negativa).

2.3.5 Esperanza matemática

La esperanza matemática [7] (EM), también llamada valor esperado (VE) o Expected Value (EV, en inglés) es la medida de cuánto esperas ganar o perder en el largo plazo, y viene definida por la siguiente función:

$$VE = (Beneficio\ \ neto\ por\ apuesta * Probabilidad\ real\ ganar) - (Pérdida\ neta\ por\ apuesta * Probabilidad\ real\ de\ perder)$$

Para que el juego sea justo y equitativo, la esperanza matemática tiene que ser igual a 0. En caso de ser inferior, la esperanza matemática será negativa, por lo que en el largo plazo se tiende a la pérdida; mientras que si es superior, la esperanza matemática será positiva, y por tanto, ganadora en el largo plazo.

Para la explicación de este concepto suele utilizarse el ejemplo del lanzamiento de una moneda. Si yo lanzo una moneda al aire, pueden salir 2 resultados: cara o cruz, por lo que la probabilidad de que ocurra cada uno de los sucesos es del 50% (0,50, es decir, 1/2). Para que el valor esperado fuera 0, la cuota debería ser 2. De esta forma, si yo apuesto 10€, mis beneficios serían 10€ (20€ ganancia – 10€ apostados), la esperanza matemática:

$$EM\ (VE) = (10 * 0{,}5) - (10 * 0{,}5) = 5 - 5 = 0$$

[7] https://apuestasev.com/que-es-y-como-calcular-la-ev-o-esperanza-matematica/

Sin embargo, como venimos explicando, esto no ocurre en las apuestas, puesto que las casas de apuestas cuentan con una ventaja probabilística en el largo plazo, al ser su esperanza matemática <0; mientras que los apostadores cuentan con desventaja, al ser su esperanza matemática >0. De forma gráfica, podemos verlo en el siguiente ejemplo.

Supongamos, siguiendo el ejemplo que venimos utilizando, que un apostador desea apostar 10€ a la victoria del equipo local, siendo sus cuotas y probabilidades reales (en base a las cuotas corregidas) las siguientes:

	Cuota	**Probabilidad real gana local**	**Probabilidad real NO gana local**
Gana Local	1,4	0,6762	0,3238

Tabla 10. Probabilidad de suceso y probabilidad de no suceso

A. Para la casa de apuestas: si no gana el equipo local, obtendría 10€ netos (cantidad apostada); mientras que si gana, perdería 4€ netos. De esta forma:

$$VE\ (casa\ apuestas) = (10 * 0{,}3238) - (4 * 0{,}6762) = 3{,}238\ - 2{,}7048 = 0{,}5332$$

Así, para la casa de apuestas: VE = 0,5332 > 0, y por tanto, rentable en el largo plazo.

B. Para el apostador: si gana el equipo local, obtendría 4€ netos (14€ beneficio – 10€ apostados); mientras que si no gana, perdería 10€ netos (cantidad apostada). De esta forma:

$$VE\ (apostador) = (4 * 0{,}6762) - (10 * 0{,}3238) = 2{,}7048 - 3{,}238 = -\ 0{,}5332$$

Así, para el apostador: VE = -0,5332 < 0, y por tanto, no rentable en el largo plazo.

De esta forma, podemos concluir que las casas de apuestas van a ganar en el largo plazo, en detrimento de los apostadores. De hecho:

$$VE\ (casa\ apuestas) + VE\ (apostador) = 0$$

2.3.6 Tipos de cuotas

2.3.6.1 Según su presentación: decimales, fraccionales o americanas.

- **Decimales** (también llamada cuota europea): viene representada por un número entero o con decimales. Es la cuota que utilizaremos a lo largo de todo el trabajo, puesto que es la más frecuente en nuestro país y la más intuitiva: con cada euro apostado ganarás la cantidad que establece la cuota multiplicado por el dinero apostado.

Por ejemplo, si apuestas 1€ a la victoria del equipo local que se paga a 1,90: ganarás 1,90€ (la cuota), pero tu beneficio será 0,90€ (la ganancia menos el dinero apostado).

- **Fraccionales**: la cuota viene determinada por una fracción. En el numerador encontramos la cantidad que ganaríamos para el caso de que apostamos lo establecido en el denominador. Es menos intuitiva y suele utilizarse en los países anglosajones.

Por ejemplo, en un partido tenemos las siguientes cuotas decimales:

	Cuota fraccional
Gana Local	20/21
Empate	11/4
Gana Visitante	13/5

Tabla 11. Cuota fraccional

De esta forma, si queremos ganar 20€ en caso de la victoria del equipo local, deberemos apostar 21€.

Se puede establecer una relación entre las cuotas decimales y las fraccionales, que viene determinada de la siguiente forma:

$$Cuota\ decimal = Cuota\ fraccional + 1$$

Para el caso, por ejemplo, del empate:

$$Cuota\ decimal = \left(\frac{11}{4}\right) + 1 = 2{,}75 + 1 = 3{,}75$$

De esta forma, las cuotas decimales que corresponderían a las cuotas fraccionales anteriores son las siguientes:

	Cuota fraccional	**Cuota decimal**
Gana Local	20/21	1,95
Empate	11/4	3,75
Gana Visitante	13/5	3,60

- **Americana**: es la menos utilizada en nuestro país y es propia de los países americanos, como puede desprenderse de su propio nombre. Se expresa con un nombre positivo o negativo, y se toma como referencia la base de 100% dólares.

De esta forma[8], cuando la cuota aparece con un signo positivo "+" indica el beneficio neto que se obtiene apostando 100 dólares. Por ejemplo, una cuota +200 significa que ganaremos 200 dólares por cada 100 que se apuesten. Cuando la cuota aparece con signo negativo "-", se refiere a la cantidad que es necesaria apostar para obtener 100 dólares. Una cuota -200 significa que es necesario apostar 200 dólares para ganar 100.

2.3.6.2 Según el momento: apertura vs cierre

- Cuotas de apertura: son las primeras cuotas que el mercado ofrece para un resultado concreto en un evento concreto. Son determinadas directamente por las casas de apuestas, mediante algoritmos que tienen en cuenta los resultados pasados para calcular la probabilidad de los resultados.

[8] https://www.casasdeapuestas.com/glosario/cuotas/

- Cuotas de cierre: son las cuotas inmediatamente anteriores al inicio del evento. Estas cuotas no solo vienen determinadas por la probabilidad calculada por las casas de apuestas, sino que también están condicionados por el propio mercado, su comportamiento respecto a la cantidad de dinero apostada en favor de cada suceso, y como las casas desplazan las cuotas.

2.3.7 Desplazamiento de las cuotas

Implica que la probabilidad asignada por la casa de apuestas a un resultado, y por tanto, su cuota, difiere de la probabilidad que el mercado de apostadores piensa que tiene realmente ese suceso.

Por ejemplo: la casa de apuestas ofrece una cuota 1,80 a la Victoria del Villarreal en un partido de Liga. De esta forma, estima que la probabilidad de que suceda ese resultado es del 0,5555.

Sin embargo, el mercado de apostadores considera que la probabilidad de que el Real Madrid venza ese partido es del 0,6666, y por tanto, la cuota que debería corresponder a esta probabilidad sería 1,50.

Al ofrecerse una cuota mayor de la que corresponde según la probabilidad, los jugadores apostarán mucho más dinero a que ese suceso ocurrirá, de lo que lo harían si la cuota reflejara fielmente la probabilidad. De esta forma, la casa de apuestas bajará la cuota a favor del Real Madrid para desincentivar su victoria (es decir, ofreciendo menos dinero para el supuesto de que se acierte), al mismo tiempo que incentivará los otros dos resultados, el empate y la victoria del equipo contrario de forma que se incentive el empate y la victoria del otro equipo, subiendo ambas cuotas (es decir, ofreciendo más dinero en cas de que se acierte).

Esta bajada de cuotas supone el ajuste a la probabilidad estimada por el mercado, y la situación de equilibrio que debe tener la casa de apuestas, para conseguir que, independientemente del resultado, consiga mantener la rentabilidad (que recordamos es la cantidad superior a 1 de la suma de la probabilidad implícita de las cuotas ofrecidas).

2.3.8 Características de las cuotas

En base a todo lo anterior, podemos establecer unas características básicas de las cuotas:

- La cuota siempre será > 1. Si fuera cuota 1, no tendría sentido apostar nada, puesto que la ganancia sería igual a la cantidad apostada, y en ese caso, el beneficio = 0. Y si fuera menor, menos sentido tendría, puesto que perderías siempre, aún en caso de acertar el resultado apostado

- La cuota es directamente proporcional a los beneficios (si aumenta uno, crecen los otros; y viceversa) e inversamente proporcional a la probabilidad (a mayor probabilidad, menor cuota; y viceversa).

- Son variables hasta la conclusión del evento. Es decir, desde que se publican las cuotas, estas puede variar dependiendo de multitud de factores que influyen en la probabilidad de un resultado, y por tanto, en las cuotas.

- Son invariables una vez realizada la apuesta. A pesar de que las cuotas varían, una vez realices la apuesta a una determinada cuota, esa cuota ya es invariable, fija, inamovible. Es decir, tu apuestas 10€ a que en un partido de futbol el resultado es empate, y la cuota para dicho resultado es de 3.20. Una vez realizada, esa será la cuota que en todo momento regirá en tu apuesta, manteniéndose de forma independiente a que en el mercado la misma pueda subir o bajar:

  · Por un lado, si apuestas 10€ a una cuota 3.2, y finalmente la cuota antes del inicio del evento es de 3.1, significa que a priori (sin conocer el resultado del evento), tu rentabilidad será mayor, puesto que en caso de ganar recibirás más dinero (22€ netos), que alguien que haya apostado la misma cantidad a la cuota de cierre de mercado (que recibirá 21€).

  · Por el contrario, si apuestas 10€ a una cuota 3.2, y finalmente la cuota antes del inicio del evento es de 3.3, significa que a priori (sin conocer el resultado del evento), tu rentabilidad será menor, puesto que en caso de ganar recibirás menos

dinero (22€ netos), que alguien que haya apostado la misma cantidad a la cuota de cierre de mercado (que recibirá 23€).

2.4 LAS CASAS DE APUESTAS

Teniendo claros todos los conceptos anteriores, podemos definir casa de apuestas como la compañía que, en base a un probabilidad que calcula y a un margen que establece, propone unas cuotas para determinados eventos, acepta apuestas para los mismos y asume un riesgo, con el fin de conseguir unas ganancias, obtenidas de las pérdidas de los apostadores para el caso de que no se produzca dicho resultado.

2.4.1 Tipos de casas de apuestas: de intercambio u ordinarias.

- De intercambio: por un lado, existen casas de apuestas las cuáles actúan simplemente como una mera intermediaria entre los jugadores que desean apostar a favor de que ocurrirá un suceso en un determinado evento, y los jugadores que por el contrario, desean apostar en contra de que ese determinado suceso no ocurrirá. La más conocida es Betfair Exchange, pero existen otras como Smarkets, Betdaq o Matchbook.

Por tanto, se necesitará que alguien te "iguale" la apuesta, es decir, que asuma el riesgo que tu has asumido a favor, en contra (o viceversa). Si nadie "iguala" la apuesta, esa apuesta quedará sin efecto. Por ello, en caso de haber apostado a favor, lo que el apostador deberá hacer es bajar la cuota, de forma que quien quiera apostar en contra gane lo mismo con menos riesgo; y al contrario sucede con las apuestas en contra. Es decir, es el propio mercado el que se regula (sin necesidad de la participación de la casa de apuestas moviendo las cuotas). Por ejemplo, partimos del siguiente caso extraído de la página de Betfair Exchange:

3 Selecciones		Favor		Contra		
Oporto		1.87 €275	1.88 €150	1.95 €113	1.96 €239	2.02 €85
Lazio		4.4 €6	4.5 €116	4.9 €235	5.1 €5	
Empate		3.6 €7	3.65 €150	3.85 €111	3.9 €110	3.95 €511

Figura 1. *Cuotas Betfair Exchange*

Como vemos, podemos apostar tanto a favor como en contra:

· A favor: el funcionamiento es exactamente igual que las apuestas ordinarias. Apuestas a que ocurrirá un suceso: si aciertas, ganarás la apuesta y tendrás beneficios, que vendrán determinados por la cuota, como hemos explicado en los apartados anteriores.

· En contra: en este caso, el apostador es el que de alguna forma "actúa como las casas de apuestas". Apuesta en contra de que sucederá un evento: si el evento ocurre, perderá; pero si el evento no ocurre, ganará la apuesta.
En este caso, la cuota es el riesgo que asume el apostador en caso de que ocurra el suceso por el que ha apostado en contra (al igual que lo que suponen para las casas de apuestas).

Por ejemplo, apuesta en contra del empate, y la cuota es de 3.85. En este caso, el apostador deberá introducir la cantidad de dinero que desea ganar, 10€. Así:

- En caso de victoria del equipo local o victoria visitante, el apostador ganará, puesto que no ha ocurrido el suceso por el que ha apostado en contra (es decir, apuesta a favor de los otros dos sucesos diferentes), y su beneficio será 10€.

- En caso de empate, el apostador perderá la cantidad que deseaba ganar multiplicada por el riesgo que había corrido, menos la cantidad apostada: (10€ x 3,85) – 10 = 28,5€ perderá.

Insistimos en que, para que todo esto ocurra, debe haber gente que apuesta a favor de un suceso a una determinada cuota, y gente que este dispuesta a apostar en contra y arriesgue, para conseguir 1€, la cantidad establecida por esa misma cuota (uno gana lo que el otro pierde, y viceversa).

En este tipo, la casa de apuestas únicamente proporciona la plataforma que permite a los apostadores llevar a cabo sus apuestas, sin asumir ningún riesgo. Por ello, la casa de apuestas cobra un porcentaje sobre las ganancias netas obtenidas (por ejemplo, Betfair Exchange aplica un porcentaje del 2% sobre los beneficios).

- Ordinarias: por otro lado, existen casas de apuestas, donde podemos ubicar la inmensa ellas, en las cuales el apostador compite directamente contra la casa de apuestas, pudiendo únicamente apostar a un suceso a favor.

En este caso, la casa de apuestas automáticamente igualará la apuesta que has hecho a favor, en contra. A diferencia de las anteriores, la casa de apuestas aquí sí que asume riesgo, por lo que, como es lógico, exigirá una mayor rentabilidad. Y esto las casas de apuestas lo consiguen jugando con la probabilidad de los sucesos, su rentabilidad y la esperanza matemática, tal y como hemos explicado anteriormente.
Entre las más conocidas, encontramos: Bet365, Bwim, Betfair Sportsbook, WilliamHill, Kirolbet, 888sport, Winamax, etc...

2.4.2 El mercado de las apuestas deportivas

El mercado viene definido por el deporte al cual se quiere apostar. Encontramos diversos mercados: fútbol, tenis, baloncesto, balonmano, etc....

Dentro de estos, encontramos otros mercados, en base a las distintas ligas y divisiones de cada uno de los deportes. Así, en el caso concreto del fútbol podemos encontrar diversos mercado: Liga Santander (1ª división española); Liga Smartbank (2ª divisón española); Ligue 1 (1ª división francesa); Bundesliga (1ª división alemana); Serie A (1ª división italiana), Premier League (1ª divisón inglesa), etc...

Y dentro de los mismos, encontramos otros submercados, que se refieren a los mercados disponibles para cada partido de las ligas, es decir, los diferentes tipos de apuestas que puedes realizar en cada partido: resultado final, número de goles, número córners, resultado exacto, etc...

2.4.3 La liquidez en el mercado de las apuestas deportivas

Una vez entendido el concepto de mercado, es importante explicar el concepto de liquidez de los mercados.

En general, la liquidez hace referencia a la capacidad de convertir algo en dinero. A mayor liquidez, con mayor facilidad podrá convertirse ese algo en dinero, y viceversa.

En el caso de las apuestas deportivas, y siguiendo en esta línea, la liquidez[9] vendrá marcada por el dinero disponible en un determinado mercado. De esta forma, cuantos más apostadores jueguen en un mercado, y por tanto, más dinero haya en circulación, más líquido será un mercado.
Como es lógico, serán más líquidos los mercados más conocidos, puesto que son los aquellos de los que los apostadores disponen más información para tomar decisiones y se sienten "más seguros" jugando ahí su dinero. De esta forma:

- Serán más líquidos los mercados de las primeras ligas de países punteros, que aquellas ligas inferiores de estos países o ligas de países menos seguidos. Así, tendrán mayor liquidez los mercados de la Premier League que los mercados de la 7ª división inglesa o de la 1ª división de Albania.

- Dentro de cada partido, serán más líquidos los mercados más conocidos por los apostadores. Así, tendrá mayor liquidez los mercados en los que se apueste al resultado de un partido o al número de goles que se marcarán, que en los que se apueste por qué jugador marcará o el resultado exacto de un partido.

La mayor liquidez de un mercado tiene dos implicaciones fundamentales para los apostadores:

- Menos variaciones de cuotas se producirán: cuanto más dinero haya en circulación, la distribución de las apuestas se acercará más a la probabilidad de las mismas, de forma que estas sufrirán menos variaciones.

- Menos márgenes impondrán las casas de apuestas y mayor será la esperanza matemática de los apostadores: como hemos analizado, las casas buscan mantener su rentabilidad a través del mantenimiento estable de sus márgenes. De esta forma,

[9] https://apuestas.marathonbet.es/glosario-apuestas/liquidez-del-mercado-apuestas/

en un mercado cuya liquidez total ascienda a 1.000€, el margen de la casa deberá ser mayor para mantener una rentabilidad constante que en un mercado con una liquidez total de 100.000€.

Por tanto, la liquidez de los mercados se relaciona inversamente tanto con las variaciones de cuotas como con los márgenes de ganancia de las casas de apuestas. Lógicamente, al apostador le convendrá llevar a cabo apuestas en mercados líquidos, en lugar de hacerlo en mercados no líquidos.

CAPÍTULO 3: DELIMITACIÓN DEL MODELO

3.1 MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Para proceder a este estudio, deberemos elaborar un modelo que sea adecuado para estudiar correctamente las variables que determinan la probabilidad de victoria de un determinado equipo en un partido de fútbol. Para ello, llevaremos a cabo la construcción de un modelo basado en la regresión lineal múltiple.

El modelo de regresión múltiple estudia la relación entre una variable de interés *Y* (variable respuesta o dependiente) y un conjunto de variables x1, x2, x3,....xn (variables explicativas o regresoras).

El modelo de regresión lineal múltiple se representa de la siguiente forma:

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_k \cdot X_k + \varepsilon$$

donde:

· X_1, X_2...X_n: son las variables independientes.
· a: es la ordenada del punto de intersección con el eje Y.
· b_1, b_2 ...b_n: es el coeficiente de regresión, es decir, el cambio neto de Y por cada cambio unitario de una X_n, manteniendo las demás X_n constantes.

3.2 LAS VARIABLES DEL MODELO

V	NAME	EXPLICACIÓN
Y	CCLOCI	Cuotas iniciales corregidas para la victoria del equipo local
x1	U3CL	Puntos del equipo local en los últimos 3 partidos como local (sobre 9)
x2	U3L	Puntos del equipo local en los últimos 3 partidos (sobre 9)
x3	U1CL	Puntos del equipo local en el último partido como local (sobre 3)

x4	U1L	Puntos del equipo local en el último partido (sobre 3)
x5	TCL	Partido del equipo local jugado tras un partido de otra competición (1=no ha jugado; 2=jugado y perdido; 3=jugado y empatado; 4=jugado y ganado)
x6	%PTOSL	% puntos totales del equipo local hasta la jornada actual
x7	DGL	Diferencia de goles del equipo local
x8	U1V	Puntos del equipo visitante en el último partido (sobre 3)
x9	U1CV	Puntos del equipo visitante en el último partido como visitante (sobre 3)
x10	U3V	Puntos del equipo visitante en los últimos 3 partidos (sobre 9)
x11	U3CV	Puntos del equipo visitante en los últimos 3 partidos como visitante (sobre 9)
x12	TCV	Partido del equipo visitante jugado tras un partido de otra competición (1=no ha jugado; 2=jugado y perdido; 3=jugado y empatado; 4=jugado y ganado)
x13	%PTOSV	% puntos totales del equipo visitante hasta la jornada actual
x14	DGV	Diferencia de goles del equipo visitante

Tabla 12. Variables del modelo

3.3 DATOS Y DESCRIPCIÓN DE LA MUESTRA

Para llevar a cabo el análisis, se han extraído las cuotas de los partidos a analizar (temporadas y jornadas) a través de la página web: www.oddsportal.com.

Respecto a los datos referentes a las variables, han sido extraídos de diferentes páginas webs: www.laliga.com, www.flashscore.es, www.bdfutbol.com/es, www.oddspedia.com, www.understat.com/league/La_liga

Tras haber sido depurada eliminando aquellos partidos de los que no teníamos datos completos (10 partidos más las primeras 4 jornadas) la muestra está formada por 337 partidos, cada uno de los cuales tiene asociadas sus 14 variables.

3.4 R STUDIO

El programa utilizado para el análisis econométrico de los datos es R Studio[10].

[10] www.rstudio.com/products/rstudio/download/#download.

CAPÍTULO 4: APLICACIÓN DEL MODELO

4.1 OBTENCIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

En primer lugar, obtenemos el modelo de regresión que expresa la probabilidad de victoria local (y) en función de las diferentes variables con las que trabajamos (x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14).

Para ello, introducimos los datos a través de la hoja Excel que los contiene:

```
install.packages("openxlsx")
library(openxlsx)
Modelo <- read.xlsx("VARIABLESTFG.xlsx", sheet = "VARIABLES")
summary(Modelo)
attach(Modelo)
Modelo<-data.frame(Y, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14)
head(Modelo)
regresionmultiple <- lm(Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14)
```

Y ajustamos el modelo de regresión lineal múltiple:

```
regresionmultiple <- lm(Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14)
summary(regresionmultiple)
```

Obtenemos los siguientes resultados:

```
Call:
lm(formula = Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14)

Residuals:
     Min       1Q   Median       3Q      Max
-0.225582 -0.049535 -0.006548  0.047617  0.275819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4194274  0.0297398  14.103  < 2e-16 ***
x1          -0.0004329  0.0027211  -0.159    0.874
x2           0.0015181  0.0036133   0.420    0.675
x3          -0.0015988  0.0052544  -0.304    0.761
x4          -0.0014150  0.0050406  -0.281    0.779
x5           0.0007120  0.0046551   0.153    0.879
x6           0.3666068  0.0589183   6.222 1.52e-09 ***
x7           0.0037450  0.0005652   6.625 1.45e-10 ***
x8           0.0054661  0.0046949   1.164    0.245
x9           0.0019846  0.0052144   0.381    0.704
x10         -0.0030147  0.0033952  -0.888    0.375
x11          0.0018902  0.0029009   0.652    0.515
x12         -0.0031121  0.0052251  -0.596    0.552
x13         -0.3623641  0.0576287  -6.288 1.05e-09 ***
x14         -0.0034044  0.0005702  -5.971 6.24e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08424 on 322 degrees of freedom
Multiple R-squared:  0.7595, Adjusted R-squared:  0.7491
F-statistic: 72.65 on 14 and 322 DF,  p-value: < 2.2e-16
```

En consecuencia, el modelo ajustado es:

$$Y = 0.4194274 - 0.0004329 \cdot x1 + 0.0015181 \cdot x2 - 0.0015988 \cdot x3 - 0.0014150 \cdot x4 + 0.0007120 \cdot x5 + 0.3666068 \cdot x6 + 0.0037450 \cdot x7 + 0.0054661 \cdot x8 + 0.0019846 \cdot x9 - 0.0030147 \cdot x10 + 0.0018902 \cdot x11 - 0.0031121 \cdot x12 - 0.3623641 \cdot x13 - 0.0034044 \cdot x14$$

4.2 BONDAD DEL AJUSTE

Por bondad de ajuste entendemos el grado de acoplamiento que existe entre los datos originales y los valores teóricos que se obtienen de la regresión[11].

De esta forma, procedemos ahora a verificar si este modelo proporciona un buen ajuste a la hora de explicar la variable Y (variable respuesta o dependiente), que en nuestro caso es la cuota inicial corregida para la victoria del equipo local (es decir, la probabilidad de que el equipo local gane el partido).

Para ello, analizaremos las tres medidas que nos permiten cuantificar la bondad del ajuste:

1. El coeficiente de determinación (R^2). Nos indica qué proporción de la varianza es explicada por la recta de regresión. En nuestro modelo $R^2 = 0.7595$, debemos interpretarlo de la siguiente manera: el 75,95% de la variabilidad de la variables *Y* es explicada por el modelo de regresión (en funciones de las variables). Así, podemos decir que el modelo es bastante bueno. Incluso si comparamos el R^2 original con el R^2 ajustado, vemos que los resultados son muy parecidos (0.7595 y 0.7491, respectivamente), de forma que hecho a tener una gran cantidad de variables no distorsiona el R^2 original.

De hecho, si hacemos raíz R^2 obtenemos el coeficiente de correlación múltiple (R), que mide la fuerza de la asociación entre la variable dependiente y dos o más variables independientes. De esta forma, R = 0,8714, por lo que, al encontrarse próximo a 1, podemos decir que existe una correlación relativamente grande.

2. Error estándar residual. Se prefieren modelos que tengan un menos error estándar residual. En nuestro caso, en modelo parece bueno en base a este criterio, puesto que obtenemos un resultado muy bajo de este parámetro = 0.08424.

[11] Universidad de Valencia. Bondad del ajuste. Consultado en: https://www.uv.es/ceaces/base/regresion/bondad.htm

3. Tabla ANOVA. ¿Hay relación lineal significativa entre la variable *Y*, y las variables explicativas del modelo (x1, x2 ... x14), con un nivel de confianza del 95%? Para ello, resolvemos el siguiente sistema:

H0 :β1 =β2 =β3 = β4 =β5 =β6 = β7 =β8 =β9= β10 =β11 =β12= β13 =β14 = 0
H1 :βi≠0, para algún i=1,2,3,4,5,6,7,8,9,10,11,12,13,14

P-valor = $2{,}2 \cdot 10^{-16}$. Como es menor 0,05, se rechaza H0. En consecuencia, concluimos que existe una relación lineal estadísticamente significativa a un nivel de significación del 5% entre la variable Y, y alguna(s) de las variables explicativas del modelo. Es decir, desde un punto de vista práctico, las variables independientes (x1, x2...x14) tienen la capacidad de explicar la variación en la variable dependiente (probabilidad inicial corregida de victoria local).

4.3 DIAGNOSIS DEL MODELO

Un residuo es la diferencia entre el valor real de Y, y su valor pronosticado. La diagnosis del modelo se llevará a cabo a través del análisis de los residuos, que deberá cumplir las siguientes condiciones[12]:

- Linealidad. Los valores de la variable dependiente están generados por el siguiente modelo lineal:

$$Y = X * B + U$$

- Homocedasticidad. Todas las perturbaciones tienen las misma varianza:

$$V(ui) = \sigma^2$$

- Independencia: las perturbaciones aleatorias son independientes entre sí:

$$E(u_i \cdot u_j) = 0, \forall i \neq j$$

- Normalidad: la distribución de la perturbación aleatoria tiene distribución normal:

$$U \approx N(0, \sigma^2)$$

- Colinealidad.

[12] Abuín, J. R. (2007). Regresión lineal múltiple. IdEyGdM-Ld Estadística, Editor, 32.

4.3.1 Linealidad

Para estudiar gráficamente la linealidad, debemos:

- En primer lugar, realizar el gráfico de cada variable explicativa frente a la variable dependiente.

```
plot(datos, main="Diagrama de dispersión")
```

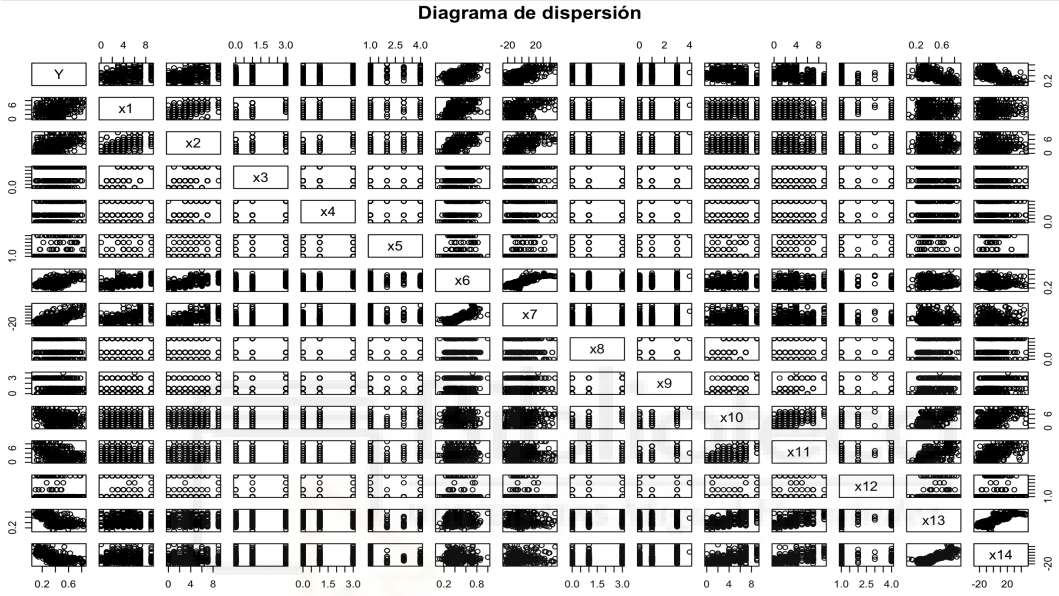

Figura 2. Gráfico de cada variable explicativa frente a la variable dependiente del modelo 1

- En segundo lugar, el gráfico de residuos frente a valores predichos. Para ello, debemos calcular previamente estos valores valores predichos y los residuos.

```
valores_predichos2 <- regresionmultiple$fitted.values
residuos2 <- regresionmultiple$residuals
plot(residuos2 ~ valores_predichos2, main="Diagrama de dispersión")
abline(glm(residuos ~ valores_predichos), col= "darkgreen")
```

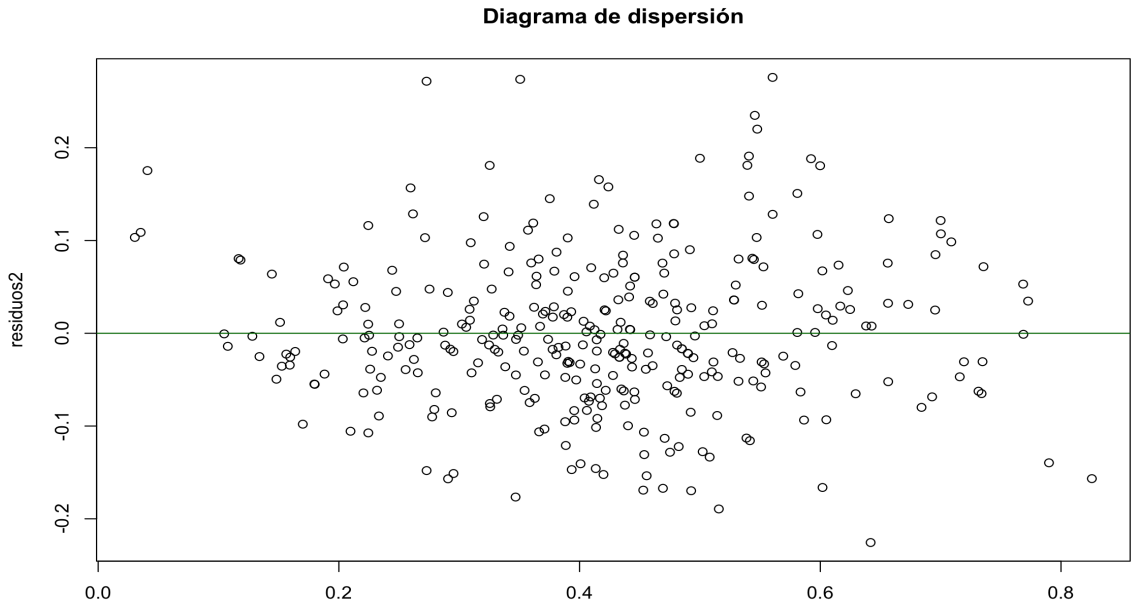

Figura 3. Diagrama de dispersión del modelo 1.

A simple vista, no se aprecia ni una tendencia no lineal clara. Pero esta debe ser comprobada analíticamente, usando el test Rest de Ramsey.

```
resettest(regresionmultiple)
```

```
RESET test

data:  regresionmultiple
RESET = 6.1883, df1 = 2, df2 = 320, p-value = 0.002307
```

Se obtiene un p-valor = 0,002307. Como p-valor < 0,05 concluimos que no hay linealidad con un nivel de significación del 5%.

4.3.2 Homocedasticidad

Gráficamente, atendiendo al gráfico de la *Figura 3*, podemos decir que, a simple vista, se aprecia una cierta heterocedasticidad, ya que observa como la varianza de los errores aleatorios tiene una cierta tendencia, y por tanto, no es constante. Para comprobarlo, debemos estudiar la homocedasticidad analíticamente, utilizando para ello el test de Breusch-Pagan.

```
library(zoo)
library(lmtest)
bptest(Y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14,
studentize=TRUE, data=Modelo)
```

```
studentized Breusch-Pagan test

data:  Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14
BP = 35.854, df = 14, p-value = 0.001097
```

Se obtiene un p-valor = 0,001097. Como p-valor < 0,05, concluimos que no hay homocedasticidad con un nivel de significación del 5%. Por tanto, comprobamos que hay heterocestadicidad, tal y como habíamos intuido de forma gráfica.

4.3.3 Normalidad

Para un estudio gráfico de la normalidad, debemos obtener:
- En primer lugar, el histograma

```
hist(residuos2, freq=TRUE, col="lightyellow")
```

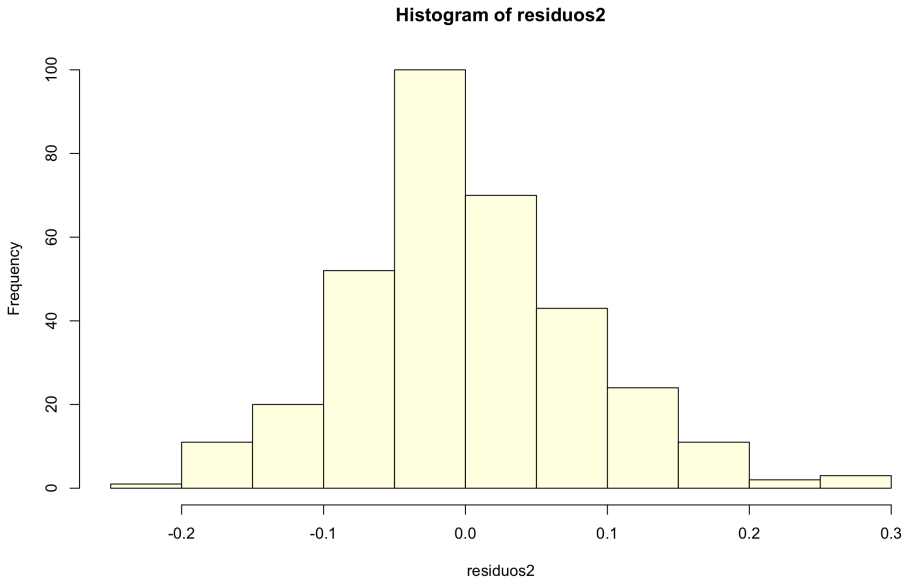

Figura 4. Histograma de residuos del modelo 1. Elaboración propia.

- En segundo lugar, el gráfico QQ:

```
library(car)
qqPlot(residuos2, xlab="Cuantiles teoricos", ylab="Cuantiles empiricos")
```

```
[1] 89 28
```

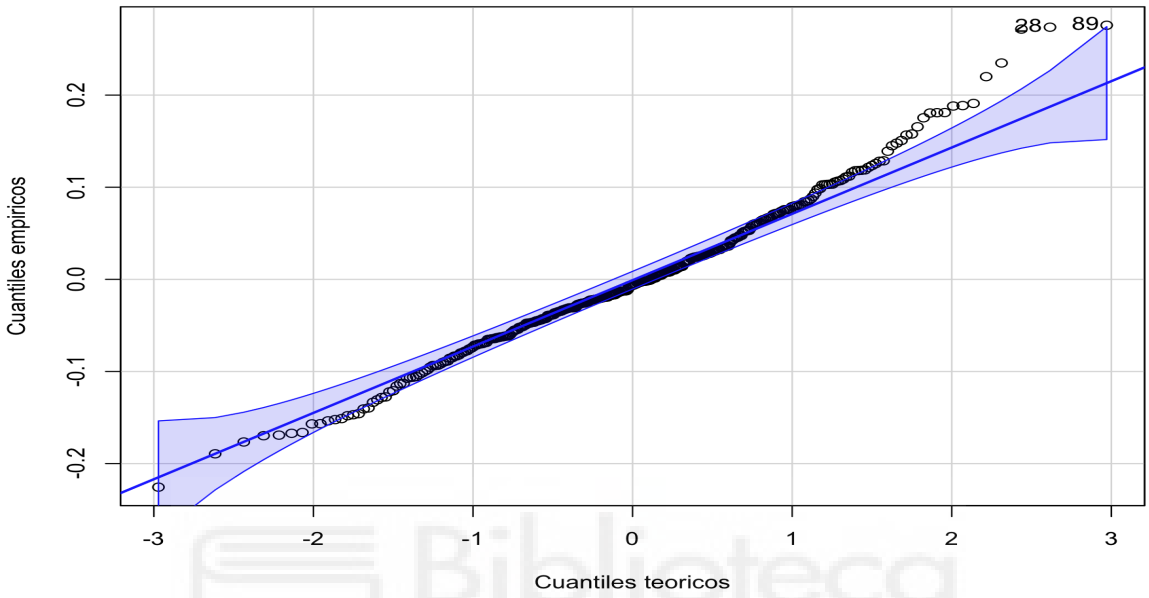

Figura 5. Gráfico QQ del modelo 1

Gráficamente, podemos apreciar la existencia de puntos que se encuentran relativamente alejados de la diagonal, especialmente en la cola derecha. Esto nos permite, de forma gráfica, intuir la no normalidad de las datos, de forma que deberemos comprobar analíticamente si se encuentran dentro de los límites tolerables.

Analíticamente, el estudio de la normalidad se lleva a cabo a través del contraste de Shapiro-Wilk:

```
shapiro.test(residuos)
```

```
Shapiro-Wilk normality test

data:  residuos
W = 0.98534, p-value = 0.001687
```

P-valor = 0,001687. Como p-valor < 0,05, se rechaza la hipótesis de normalidad para un nivel de significación del 5%.

4.3.4 Independencia

Para estudiar analíticamente la independencia de los residuos, utilizaremos el contraste de autocorrelación de Durbin-Watson:

```
library(zoo)
library(lmtest)
dwtest(Y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14,
alternative="two.sided", data=Modelo)
```

```
Durbin-Watson test

data:  Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +     x12 + x13 + x14
DW = 1.7342, p-value = 0.01424
alternative hypothesis: true autocorrelation is not 0
```

P-valor = 0.01424. Como p-valor < 0.05, rechazamos la hipótesis de independencia de los residuos para un nivel de significación del 5%.

4.3.5 Colinealidad

Para el estudio de la colinealidad, se utilizará:

- La matriz de correlaciones

```
datos <- data.frame(Y,x1, x2, x3, x4, x5,x6, x7, x8, x9, x10,x11, x12, x13, x14)
cor(datos)
```

De los datos obtenidos por la matriz, observamos que la diagonal principal es 1, mientras que el resto de valores sí que encontramos valores muy cercanos a 0 (por ejemplo, 0.009904150) y a 1 (por ejemplo, 0.815592586). Por tanto, en base a la matriz de correlaciones, podemos decir que tendemos a la existencia de multicolinealidad.

- El Factor de Inflación de la Varianza (VIF):

```
library(car)
vif(regresionmultiple)
```

```
      x1       x2       x3       x4       x5       x6       x7
2.120705 3.607248 2.171202 1.966014 1.160745 4.312387 3.116047

      x8       x9      x10      x11      x12      x13      x14
1.735620 2.046003 3.128622 2.257618 1.209278 4.198296 3.271244
```

El valor de los VIF para todas las variables x1 hasta x14 están muy por debajo de 10, por lo que podemos decir que no existen indicios de multicolinealidad.

De esta forma, las hipótesis planteadas para este primer modelo se pueden resumir de la siguiente forma:

	P-valor	**Conclusión**
Linealidad	0,002307 < 0.05	No hay linealidad
Homocedasticidad	0,001097 < 0.05	Hay heterocedasticidad
Normalidad	0,001687. < 0.05	No hay normalidad
Independencia	0.01424 < 0.05	No hay independencia
Colinealidad	Todo xn < 10	No hay multicolinealidad

Tabla 13. Resumen hipótesis modelo 1

En conclusión, a excepción de la linealidad, de la cual podemos decir que no existen indicios de multicolinealidad, no se cumple ninguna de las hipótesis planteadas para el modelo: linealidad, homocedasticidad, normalidad, independencia e independencia lineal. Por ello, no aceptamos el modelo.

Sin embargo, es preciso matizar en este punto que, a pesar de que no aceptamos el modelo puesto que los residuos no cumplen las condiciones, no significa que el modelo no sea bueno (de hecho, su coeficiente de determinación R2 es 0.7595, por lo que es grande), sino que es menos efectivo que si las cumpliera.

4.4 SELECCIÓN DE UN NUEVO MODELO

Puesto que el modelo anterior no cumplía las condiciones de los residuos, intentaremos ajustar este modelo, quedándonos únicamente con aquellas variables influyentes. Para ello, utilizaremos el proceso secuencial hacia adelanta (*forward*) y el proceso secuencial hacia atrás (*backward*).

4.4.1 Proceso secuencial paso a paso hacia adelante

Puesto que se trata del proceso secuencial hacia adelante, debemos partir del modelo más simple y ver si se van incluyendo variables. Es decir, la primera variable que se introduce es la de mayor correlación con la variable dependiente que cumpla el criterio de entrada. De esta forma, cuando ya no queden variables que puedan cumplir este criterio, el procedimiento termina. Para ello, se utilizará el criterio BIC.

```
nulo <- lm(Y ~ 1, data=Modelo)
step(nulo, scope=Y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14,
direction="forward")
```

```
Step:  AIC=-1670.14
Y ~ x7 + x14 + x6 + x13

      Df Sum of Sq    RSS     AIC
<none>              2.3037 -1670.1
+ x8    1 0.0050571 2.2986 -1668.9
+ x11   1 0.0040869 2.2996 -1668.7
+ x12   1 0.0025630 2.3011 -1668.5
+ x9    1 0.0017629 2.3019 -1668.4
+ x1    1 0.0010093 2.3027 -1668.3
+ x3    1 0.0002801 2.3034 -1668.2
+ x4    1 0.0001864 2.3035 -1668.2
+ x5    1 0.0001804 2.3035 -1668.2
+ x2    1 0.0000633 2.3036 -1668.1
+ x10   1 0.0000166 2.3037 -1668.1

Call:
lm(formula = Y ~ x7 + x14 + x6 + x13, data = Modelo)

Coefficients:
(Intercept)          x7         x14          x6         x13
   0.418892    0.003718   -0.003264    0.373718   -0.366710
```

El modelo resultante es el siguiente:

$$y = 0.418892 + 0.373718 \cdot x6 + 0.418892 \cdot x7 - 0.366710 \cdot x13 - 0.003264 \cdot x14$$

De esta forma, el nuevo modelo consta únicamente de 4 variables: x6, x7, x13 y x14.

4.4.2 Proceso secuencial paso a paso hacia atrás

Puesto que se trata de un proceso secuencial hacia atrás, debemos partir del modelo completo. Es decir, se incluyen todas las variables y se va analizando si resulta interesante excluir alguna de las variables. Para ello, se utilizará el criterio BIC.

```
numero_observaciones <- nrow(Modelo)
step(regresionmultiple, direction="backward", k=log(numero_observaciones))
```

```
Step:  AIC=-1670.14
Y ~ x7 + x14 + x6 + x13

       Df Sum of Sq    RSS     AIC
<none>              2.3037 -1670.1
+ x8    1 0.0050571 2.2986 -1668.9
+ x11   1 0.0040869 2.2996 -1668.7
+ x12   1 0.0025630 2.3011 -1668.5
+ x9    1 0.0017629 2.3019 -1668.4
+ x1    1 0.0010093 2.3027 -1668.3
+ x3    1 0.0002801 2.3034 -1668.2
+ x4    1 0.0001864 2.3035 -1668.2
+ x5    1 0.0001804 2.3035 -1668.2
+ x2    1 0.0000633 2.3036 -1668.1
+ x10   1 0.0000166 2.3037 -1668.1

Call:
lm(formula = Y ~ x7 + x14 + x6 + x13, data = Modelo)

Coefficients:
(Intercept)         x7        x14         x6        x13
   0.418892    0.003718   -0.003264    0.373718   -0.366710
```

El nuevo modelo resultante es el siguiente:

$$y = 0.418892 + 0.373718 \cdot x6 + 0.418892 \cdot x7 - 0.366710 \cdot x13 - 0.003264 \cdot x14$$

Podemos comprobar, por tanto, que con ambos métodos llegamos a la misma conclusión, al mismo modelo, compuesto por cuatro variables: x6, x7, x13, x14.

4.5 EL NUEVO MODELO

```
Call:
lm(formula = Y ~ x6 + x7 + x13 + x14, data = Modelo)

Residuals:
     Min       1Q   Median       3Q      Max
-0.223841 -0.053619 -0.007042  0.049239  0.277044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4188917  0.0287271  14.582  < 2e-16 ***
x6           0.3737181  0.0493343   7.575 3.60e-13 ***
x7           0.0037182  0.0005533   6.720 7.94e-11 ***
x13         -0.3667104  0.0488778  -7.503 5.78e-13 ***
x14         -0.0032644  0.0005442  -5.999 5.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0833 on 332 degrees of freedom
Multiple R-squared:  0.7576, Adjusted R-squared:  0.7546
F-statistic: 259.4 on 4 and 332 DF,  p-value: < 2.2e-16
```

De esta forma, el nuevo modelo vendrá definido de la siguiente forma:

$$y = 0.418892 + 0.373718 \cdot x6 + 0.418892 \cdot x7 - 0.366710 \cdot x13 - 0.003264 \cdot x14$$

4.6 EXPLICACIÓN DE LAS VARIABLES DEL NUEVO MODELO

4.6.1 Porcentaje de puntos totales del equipo local hasta la jornada actual (x6)

La variable x6 da valores correspondientes al porcentaje de puntos totales del equipo local hasta la jornada en la que se disputa el partido, con valores que pueden ir entre 0 (no ha conseguido ningún punto) o 1 (ha conseguido todos los puntos).

Analíticamente, presenta los siguientes parámetros:

Media	0.44666
Mediana	0.40351
1er cuartil	0.3333
3er cuartil	0.55072
Mínimo	0.08333
Máximo	1
Varianza	0.0262352
Desviación típica	0.00068828

Tabla 14. Análisis variable x6

Gráficamente, puede ser representada mediante un gráfico de caja (Box Plot), que resalta aspectos de la distribución de las observaciones. Este gráfico comienza en el primer cuartil (25%) y termina en el tercero (75%), por lo que representa el 50% de los datos centrales, con una línea dentro que representa la mediana. A cada lado de la caja se dibuja un segmento con los datos más lejanos sin contar los valores atípicos (outliers) del Box Plot, que en caso de existir, se representan con círculos.

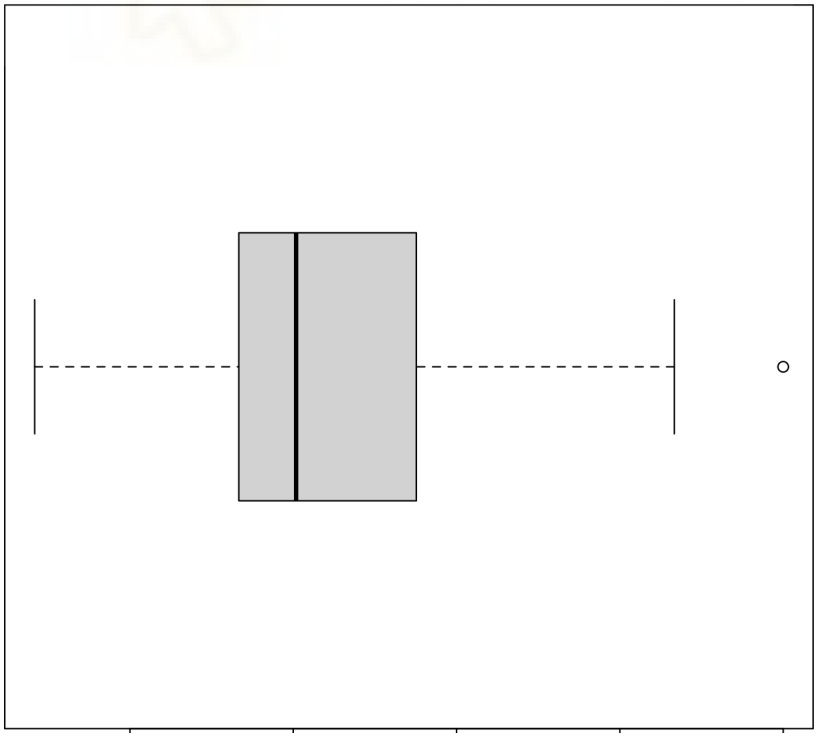

Figura 6. Gráfico de cajas de la variable x6

4.6.2 Diferencia de goles del equipo local (x7)

La variable x7 da valores correspondientes a la diferencia de goles del equipo local hasta la jornada en la que se disputa el partido, con valores que pueden ir entre - ∞ hasta + ∞, con saltos de 1 en 1.

Analíticamente, presenta los siguientes parámetros:

Media	-0.2819
Mediana	-3
1er cuartil	-9
3er cuartil	7
Mínimo	-24
Máximo	47
Varianza	205.9649
Desviación típica	42421.56

Tabla 15. Análisis variable x7

Gráficamente, puede ser representada mediante el siguiente gráfico de caja (Box Plot):

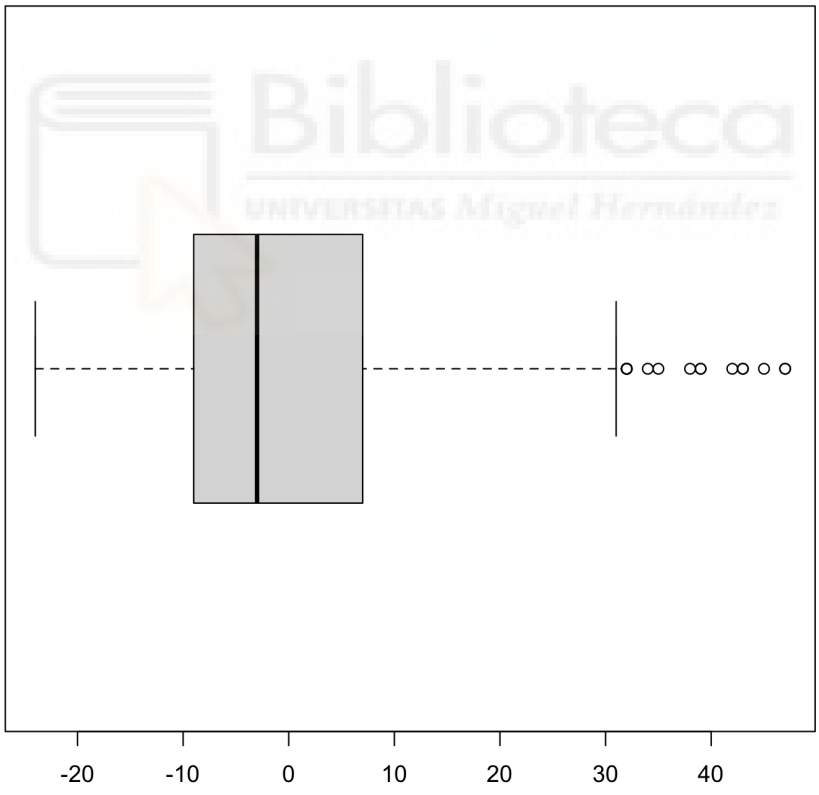

Figura 7. Gráfico de cajas de la variable x7

4.6.3 Porcentaje de puntos totales del equipo visitante hasta la jornada actual (x13)

La variable x13 da valores correspondientes al porcentaje de puntos totales del equipo visitante hasta la jornada en la que se disputa el partido, y puede ser cualquier valor entre 0 (no ha conseguido ningún punto) o 1 (ha conseguido todos los puntos).

Analíticamente, presenta los siguientes parámetros:

Media	0.45404
Mediana	0.40741
1er cuartil	0.3333
3er cuartil	0.56863
Mínimo	0.08333
Máximo	0.86667
Varianza	0.026697
Desviación típica	0.0007127

Tabla 16. Análisis variable x13

Gráficamente, puede ser representada mediante el siguiente gráfico de caja:

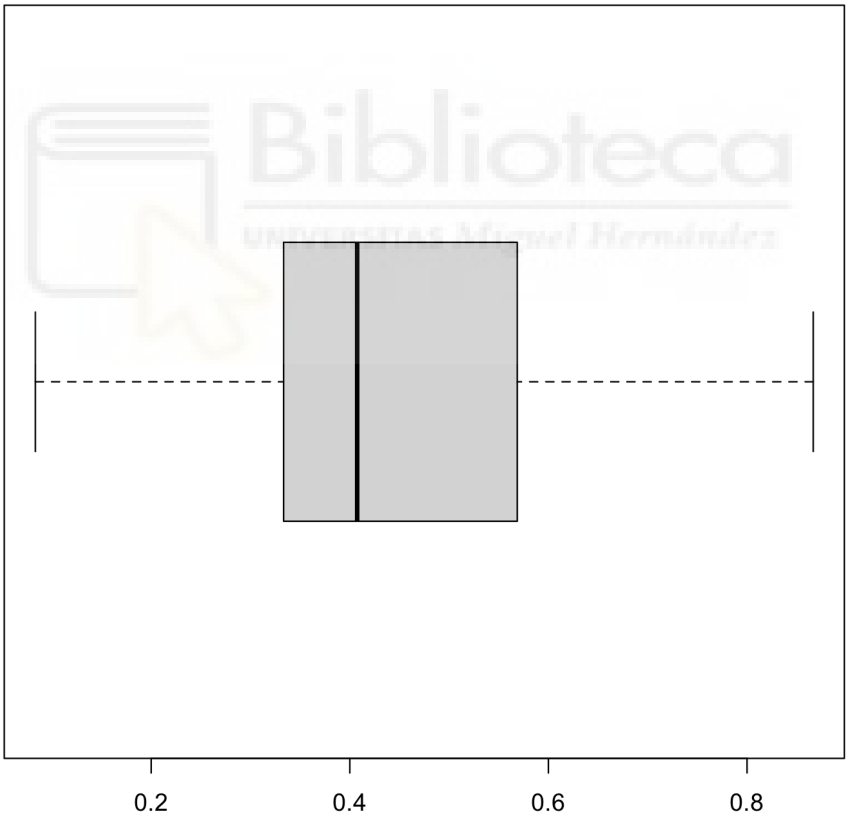

Figura 8. Gráfico de cajas de la variable x13

4.6.4 Diferencia de goles del equipo visitante (x14)

La variable x14 da valores correspondientes a la diferencia de goles del equipo visitante hasta la jornada en la que se disputa el partido, con valores que pueden ir entre - ∞ hasta + ∞, con saltos de 1 en 1.

Analíticamente, presenta los siguientes parámetros:

Media	0.04748
Mediana	-3
1er cuartil	-10
3er cuartil	6
Mínimo	-25
Máximo	47
Varianza	212.4977
Desviación típica	45155.29

Tabla 17. Análisis variable x14

Gráficamente, puede ser representada mediante el siguiente gráfico de caja:

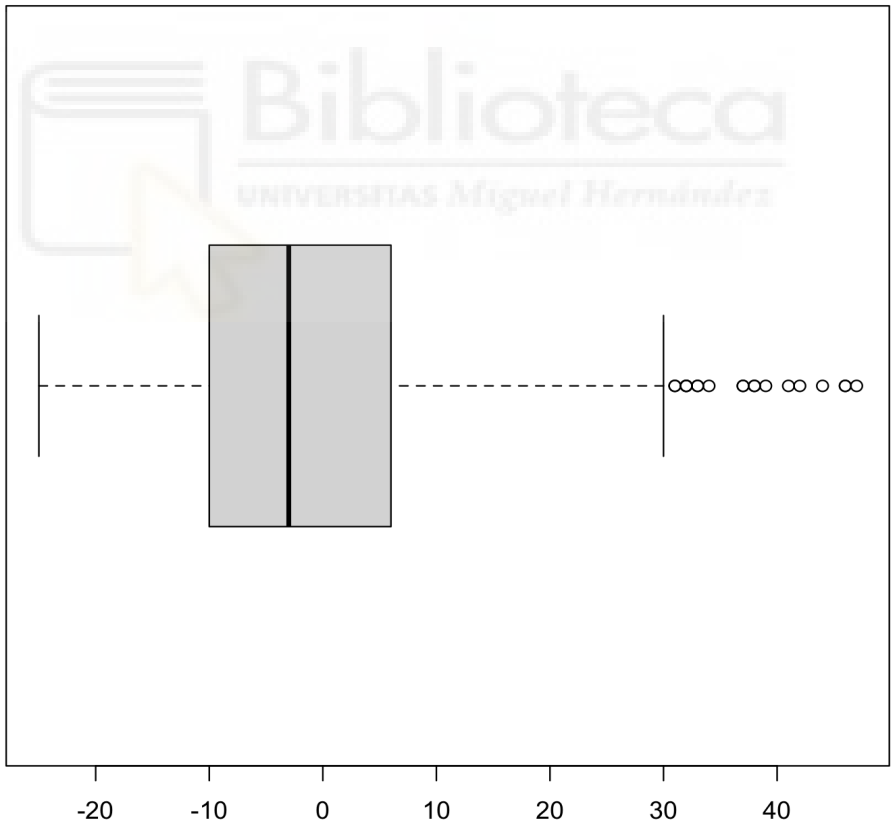

Figura 9. Gráfico de cajas de la variable x14

4.7 BONDAD DEL AJUSTE DEL NUEVO MODELO

Al igual que en el anterior modelo, la bondad del ajuste se llevará a cabo a través del análisis de tres medidas:

1. El coeficiente de determinación (R2). En este nuevo modelo, R2 = 0.7576, lo que interpretamos como que el 75,76% de la variabilidad de la variable *Y* es explicada por el modelo de regresión (en funciones de las 4 variables). Es muy parecido al coeficiente de determinación del modelo anterior.

2. Error estándar residual. Es 0.0833, por lo que podemos decir que el modelo parece buen en base a este criterio, debido a que su resultado es muy bajo. Aunque es parecido al error estándar residual del anterior modelo, el de este nuevo es más bajo, por lo que parece indicar que este modelo es mejor.

3. Tabla ANOVA. ¿Hay relación lineal significativa entre la variable *Y*, y las variables explicativas del modelo (x6, x7, x13 Y x14), con un nivel de confianza del 95%? Para ello, resolvemos el siguiente sistema:

H0 :$\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = \beta7 = \beta8 = \beta9 = \beta10 = \beta11 = \beta12 = \beta13 = \beta14 = 0$
H1 :$\beta i \neq 0$, para algún i=1,2,3,4,5,6,7,8,9,10,11,12,13,14

P-valor = $2{,}2 \cdot 10^{\wedge -16}$. Como es menor 0,05, se rechaza H0. En consecuencia, concluimos que existe una relación lineal estadísticamente significativa a un nivel de significación del 5% entre la variable Y, y alguna(s) de las variables explicativas del modelo. Es decir, desde un punto de vista práctico, las variables independientes (x6, x7, x13 Y x14) tienen la capacidad de explicar la variación en la variable dependiente (probabilidad inicial corregida de victoria local).

4.8 DIAGNOSIS DEL NUEVO MODELO

Para realizar la diagnosis del nuevo modelo, se procederá, al igual que en el primer modelo, al estudio de la linealidad, homocedasticidad, normalidad, independencia e independencia lineal.

El lenguaje de R Studio que utilizaremos será el mismo que utilizamos en el primer modelo, pero ajustado a este.

4.8.1 Linealidad

RESET test

data: Y_671314
RESET = 6.0211, df1 = 2, df2 = 330, p-value = 0.002702

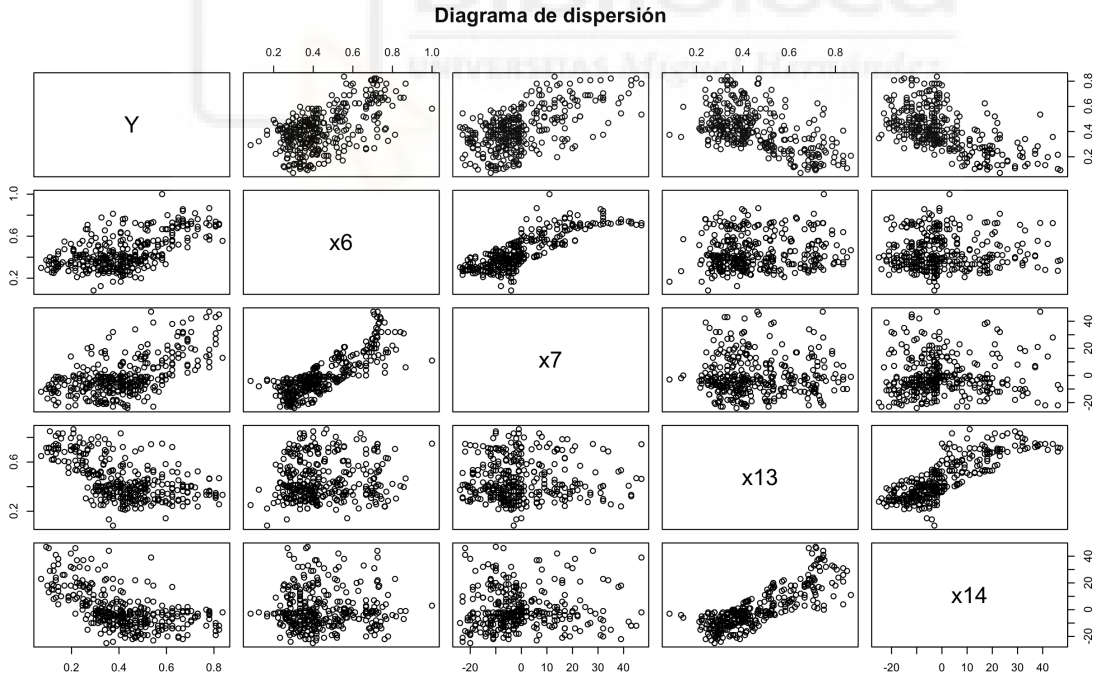

Figura 10. Gráfico de cada variable explicativa frente a la variable dependiente del modelo 1

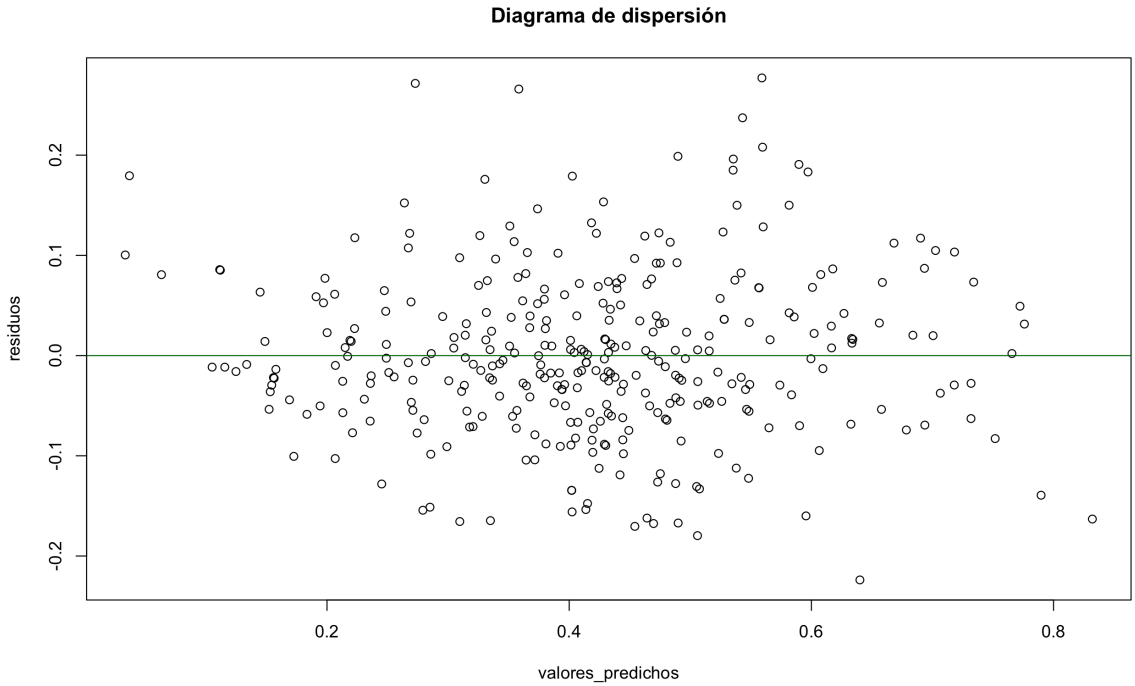

Figura 11. Diagrama de dispersión del modelo 2

P-valor = 0,002702. Como p-valor < 0,05, podemos decir que no hay linealidad con un nivel de significación del 5%.

Si lo comparamos con el p-valor del modelo anterior (0,002307), podemos observar como el p-valor del nuevo modelo es mayor. Por tanto, a pesar de que en ambos casos concluimos que se rechaza la linealidad, el hecho de que el p-valor del nuevo modelo sea mayor, nos indica que este nuevo modelo mejora un poco el anterior, aunque la mejora no sea significativa.

4.8.2 Homocedasticidad

A pesar de que gráficamente se observa una tendencia, al igual que en modelo anterior, esta parece menos clara, por lo que debemos estudiar la homocedasticidad de forma analítico a través del test Breusch-Pagan.

```
studentized Breusch-Pagan test

data:  Y ~ x6 + x7 + x13 + x14
BP = 22.432, df = 4, p-value = 0.0001644
```

Se obtiene un p-valor = 0,0001644. Como p-valor < 0,05, concluimos que no hay homocedasticidad con un nivel de significación del 5%. Por tanto, comprobamos que hay heterocestadicidad, tal y como habíamos intuido de forma gráfica.

A diferencia de lo que ocurría con la linealidad, el p-valor del nuevo modelo es menos que el p-valor del otro modelo, por lo que, a pesar de que en ambos rechacemos la homocedasticidad, este modelo empeora el anterior.

4.8.3 Normalidad

Para un estudio gráfico de la normalidad, debemos obtener:

- En primer lugar, el histograma:

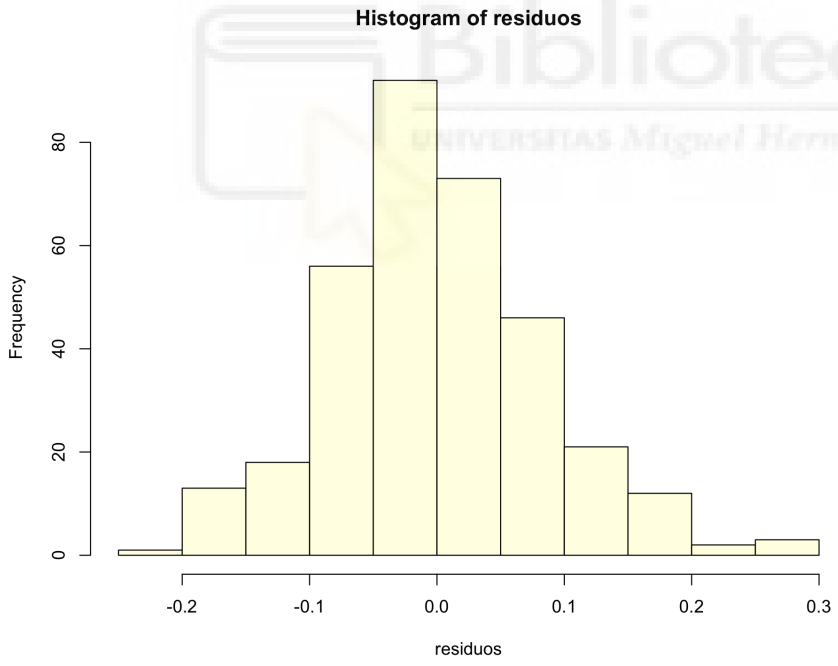

Figura 12. Histograma de residuos del modelo 2

- En segundo lugar, el gráfico QQ:

```
[1] 89 331
```

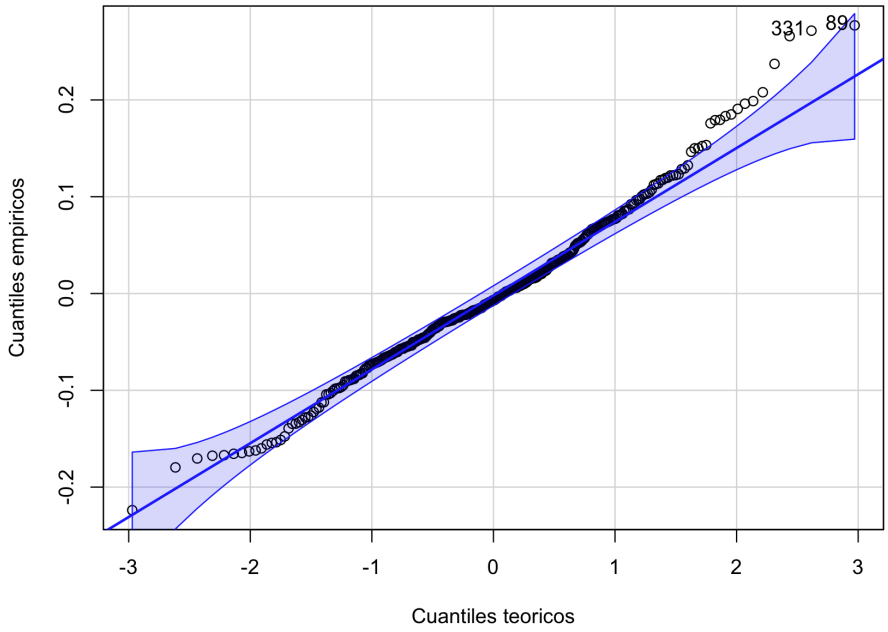

Figura 13. Gráfico QQ del modelo 2

Al igual que ocurría en el modelo anterior, podemos intuir la no normalidad de los datos, debido a la existencia de puntos en las colas, especialmente la derecha, que se encuentran alejados de la diagonal. Para ello, deberemos comprobar analíticamente esta condición a través del test Shapiro-Wilk:

```
Shapiro-Wilk normality test

data:  residuos
W = 0.98534, p-value = 0.00168
```

P-valor = 0,00168. Como p-valor < 0,05, se rechaza la hipótesis de normalidad para un nivel de significación del 5%.

El p-valor en el test Shapiro-Wilk de ambos modelos es igual, por lo que ni mejora ni empeora con respecto al otro.

4.8.4 Independencia

Durbin-Watson test

data: Y ~ x6 + x7 + x13 + x14
DW = 1.7428, p-value = 0.01839
alternative hypothesis: true autocorrelation is not 0

P-valor = 0.01424. Como p-valor < 0.05, rechazamos la hipótesis de independencia de los residuos para un nivel de significación del 5%.

Puesto que el p-valor de este nuevo modelo es superior, podemos decir que este modelo es un poco mejor, a pesar de que en ninguno de los dos aceptemos la hipótesis de la independencia de los residuos.

4.8.5 Colinealidad

Para llevar a cabo el estudio de la colinealidad, utilizaremos:

- La matriz de correlaciones:

	Y	x6	x7	x13	x14
Y	1.0000000	0.58476597	0.60149447	-0.55337668	-0.56208878
x6	0.5847660	1.00000000	0.81579636	0.07999068	0.01958666
x7	0.6014945	0.81579636	1.00000000	0.01548338	0.01393465
x13	-0.5533767	0.07999068	0.01548338	1.00000000	0.81559259
x14	-0.5620888	0.01958666	0.01393465	0.81559259	1.00000000

Así, podemos observar como, mientras la diagonal principal adopta el valor 1, en el resto de valores encontramos tanto valores muy cercanos a 0 (como 0.07999068 o 0.01393465), así como valores muy cercanos a 1 (0.81559259 o 0.81579636). De esta forma, podemos intuir con esta matriz de correlaciones, que hay una tendencia a la existencia de multicolinealidad.

- El Factor de Inflación de la Varianza (VIF):

```
      x6       x7       x13      x14
3.091963 3.053679 3.088426 3.046997
```

Sin embargo, a pesar de que con la matriz de correlación podamos intuir una multicolinealidad, podemos concluir que la misma no existe, en tanto que el valor de los VIF para las variables de este nuevo modelo (x6, x7, x13, x14) están muy por debajo de 10. De esta forma, a través del Factor de Inflación de la Varianza, podemos concluir la no existencia de multicolinealidad.

De este modo, podemos resumir las hipótesis pleanteadas y analizadas para este segundo modelo, en la siguiente tabla:

	P-valor	**Conclusión**
Linealidad	0,002702< 0.05	No hay linealidad
Homocedasticidad	0,0001644< 0.05	Hay heterocedasticidad
Normalidad	0,00168. < 0.05	No hay normalidad
Independencia	0.01424< 0.05	No hay independencia
Colinealidad	Todo xn < 10	No hay multicolinealidad

Tabla 18. Tabla resumen del análisis del modelo 2

En conclusión, y al igual que sucedía con el modelo anterior, a excepción de la linealidad (la independencia lineal, de la cual podemos decir que no existen indicios de multicolinealidad), no se cumple ninguna de las hipótesis planteadas para el modelo: linealidad, homocedasticidad, normalidad e independencia.

Por ello, y a pesar de que como hemos comprobado que algunas hipótesis arrojan valores que evidencian una mejora de este nuevo modelo respecto al anterior, las mismas no son suficientes para aceptar el nuevo modelo.

De igual forma, también debemos tener en cuenta que a pesar de que no aceptamos el modelo, e igual que sucedía con el modelo anterior, no significa que es modelo sea malo

de por sí, puesto que si nos fijamos en su coeficiente de determinación, R2 = 0.7576, lo que interpretamos como que el 75,76% de la variabilidad de la variable Y es explicada por el modelo de regresión.

Por tanto, y al igual que concluíamos con el anterior modelo (cuyo R2 = 0,7595, prácticamente idéntico al del nuevo modelo), que no cumpla las hipótesis analizadas no significa que el modelo deba de ser malo, sino que, simplemente, es peor que si cumpliera estas condiciones.

4.9 PUNTOS PALANCA, PUNTOS INFLUYENTE Y VALORES ATÍPICOS

4.9.1 Puntos palanca

Serán puntos palanca aquellas observaciones para las cuales:

$$hii > \frac{3 * 5}{337} = 0{,}4451038576$$

Para obtener los puntos palanca, introduciremos el siguiente código:

```
hatvalues <- hatvalues(Y_671314)
pmasuno <- length(Y_671314$coefficients)
n<-length(hatvalues)
hatvalues[which(hatvalues>3*pmasuno/n)]
```

Obtenemos los siguientes puntos palanca:

```
        1         7         8        10        13
0.05539654 0.04459199 0.05240039 0.08531262 0.04875446
       38       252       287       300       310
0.05011474 0.04550517 0.05074725 0.06498282 0.04541564
      321       330
0.04728669 0.05257067
```

Gráficamente, podemos ubicar los puntos palanca de la siguiente forma:

```
plot(index(x6), hatvalues, col = "blue", xlab = "Observaciones", ylab =
"Hatvalues")
abline(a = 3 * pmasuno/n, b = 0, col = "red")
```

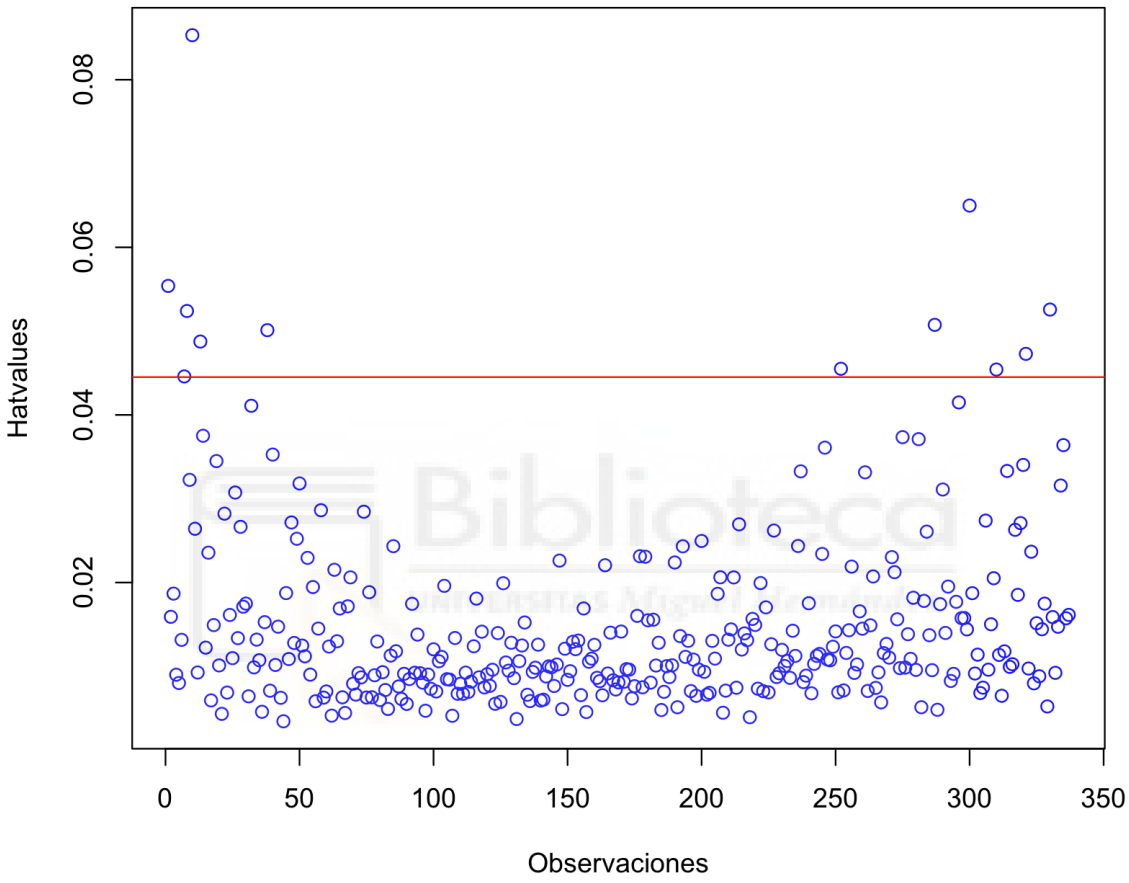

Figura 14. Gráfico de puntos palanca del modelo 2

Podemos observar gráficamente los 12 puntos palanca que hemos obtenido analíticamente.

Estos 12 puntos palanca son las observaciones: 1, 7, 8, 10, 13, 38, 252, 287, 300, 310, 321, 330.

4.9.2 Puntos influyentes

Serán puntos influyentes aquellas observaciones cuya Distancia de Cook sea > 1. Para ello, calculamos esta distancia de Cook para cada una de las observaciones del modelo:

```
cookdistance <- cooks.distance(Y_671314)
cookdistance[which(cookdistance>1)]
```

Obtenemos el siguiente resultado:

```
named numeric(0)
```

En consecuencia, podemos decir que no hay observaciones influyentes. Podemos comprobar esta conclusión gráficamente:

```
plot(index(x6), cookdistance, col = "blue", xlab = "Observaciones", ylab =
"Distancia de Cook")
abline(a = 1, b = 0, col = "red")
```

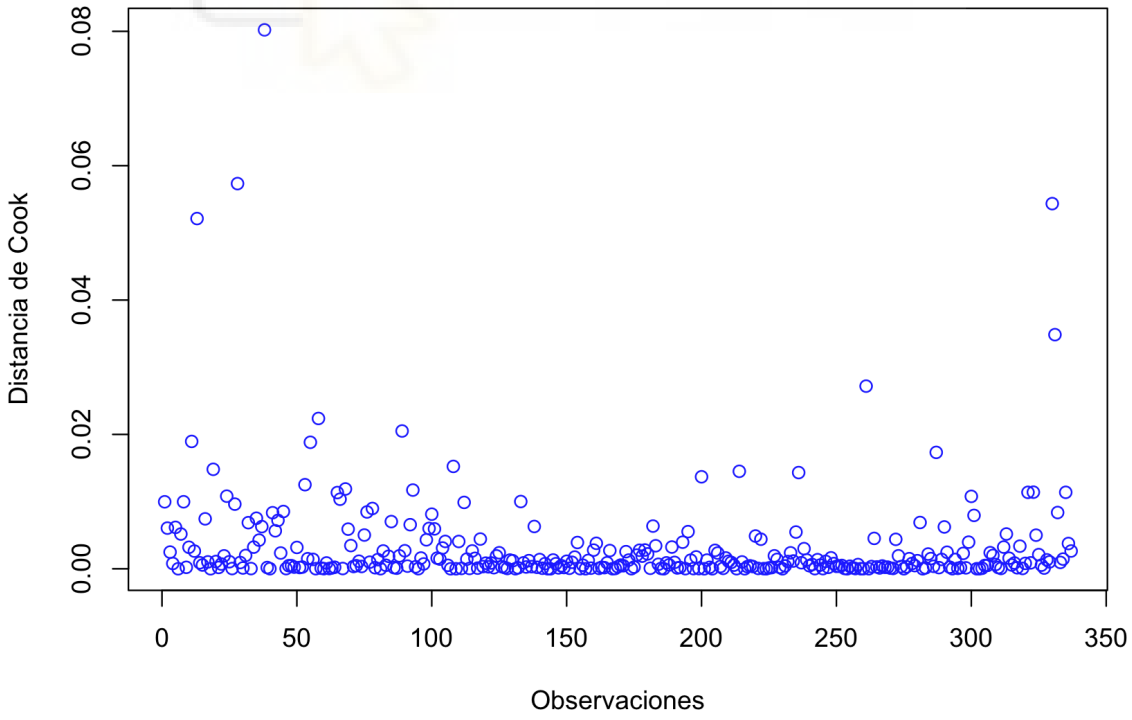

Figura 15. Puntos influyentes del modelo 2

4.9.3 Valores atípicos

Para estudiar la existencia de valores atípicos, utilizaremos el constraste de Bonferroni:

```
outlierTest(Y_671314)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferroni p
89  3.39361       0.00077351     0.26067
```

4.10 TRANSFORMACIÓN DE LAS VARIABLES

Las transformaciones son una de las posibles soluciones para mejorar el análisis estadístico, cuando los datos no presentan las características que se desean para la situación en la que se está trabajando (por ejemplo, que los datos no sean simétricos, no estén centrados en el origen o no sigan una distribución normal)[13].

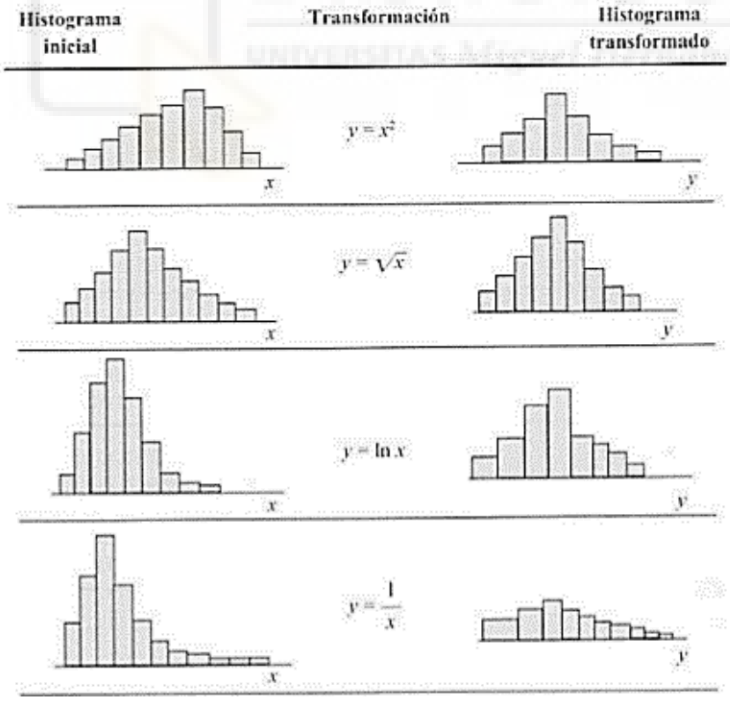

Figura 16. Efectos de las transformaciones (Peña, 1997)

[13] Sakia, R. M. (1992) The Box-Cox transformation technique: a review. The Statistician.

Las transformaciones recomendadas[14] con mayor frecuencia son: la raíz cuadrada, el logaritmo o la inversa.

Sin embargo, en el presente caso dichas transformaciones no pueden llevarse a cabo, debido a la existencia de valores negativos. De esta forma, no puede aplicarse ni la transformación logarítmica ni la raíz cuadrada, por lo que no se puede hacer una mejora de los modelos utilizando esta herramienta.

4.11 CAMBIO DE LA VARIABLE DEPENDIENTE

Puesto que la probabilidad corregida de que ocurra un suceso está formada por:

$$p\ (victoria\ local) + p\ (empate) + p(victoria\ visitante) = 1,$$

y teniendo en cuenta que los datos utilizados para calcular ambas probabilidades (los datos referentes a los partidos anteriores) son los mismos, comprobaremos si modificando la variable dependiente Y por las otras dos probabilidades (empate o victoria visitante) arrojan los mismos datos.

Llevando a cabo el mismo análisis que hemos aplicado anteriormente en el trabajo, obtenemos los siguientes datos:

4.11.1 Variable dependiente: cuotas iniciales corregidas para el empate

$$Y = 0.2985 - 0.0003749 \cdot x1 + 0.0004852 \cdot x2 - 0.0015988 \cdot x3 - 0.0002454 \cdot x4 + 0.00003443 \cdot x5 - 0.08067 \cdot x6 - 0.0007849 \cdot x7 + 0.001627 \cdot x8 - 0.0007852 \cdot x9 - 0.00199 \cdot x10 - 0.0001231 \cdot x11 - 0.001555 \cdot x12 + 0.03201 \cdot x13 - 0.0004437 \cdot x14$$

Los datos que extraemos del modelo cambiando la variable dependiente se pueden resumir en la siguiente tabla:

[14] Osborne, J. (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation.

	P-valor	Conclusión
Linealidad	0.08197 > 0.05	Hay linealidad
Homocedasticidad	4.622e-06 < 0.05	Hay heterocedasticidad
Normalidad	6.205e-12. < 0.05	No hay normalidad
Independencia	0.4468 > 0.05	Hay independencia
Colinealidad	Todo xn < 10	No hay multicolinealidad
R2	0.2896	
Error estándar residual	0.03968	
ANOVA	$2{,}2 \cdot 10^{\wedge -16} < 0.05$	Relación lineal

Tabla 19. Resumen modelo 1 con variable Y=empate

Obtenemos un nuevo modelo mediante el proceso secuencial:

$$y = 0.3026149 - 0.0826347 \cdot x6 - 0.0008175 \cdot x7$$

Las conclusiones de este nuevo modelo para Y= cuotas iniciales corregidas para el empate se pueden resumir de la siguiente forma:

	P-valor	**Conclusión**
Linealidad	0.0002645 < 0.05	No hay linealidad
Homocedasticidad	0.002543 < 0.05	Hay heterocedasticidad
Normalidad	4.508e-13 < 0.05	No hay normalidad
Independencia	0.281 > 0.05	Hay independencia
Colinealidad	Todo xn < 10	No hay multicolinealidad
R2	0.2697	
Error estándar residual	0.0395	
ANOVA	$2{,}2 \cdot 10^{\wedge -16} < 0.05$	Relación lineal

Tabla 20. Resumen modelo 2 con variable Y=empate

En conclusión, podemos decir que en el primer modelo (con todas las variables), a pesar de que no hay ni normalidad ni homocedasticidad, sí que hay linealidad e independencia de los residuos, por lo que mejora los anteriores modelos (pero seguimos sin aceptar todas las hipótesis). Sin embargo, el R2 es mucho menor (0,2896), por lo que únicamente el 28,96% de la variabilidad de Y viene explicada por el modelo de regresión.

Respecto al segundo modelo, únicamente con las variables x6 y x7, es peor que el anterior, puesto que además de no cumplirse ninguna de las condiciones de los residuos, su coeficiente de determinación es muy bajo, ya que es 0,2697 (muy cercano al anterior, pero incluso más bajo). Por ello, tampoco aceptamos este modelo.

4.11.2 Variable dependiente: cuotas iniciales corregidas para el empate

$$Y = 0.2749744 + 0.0002075 \cdot x1 - 0.0019657 \cdot x2 + 0.0038509 \cdot x3 + 0.0012061 \cdot x4 - 0.0003281 \cdot x5 - 0.2804572 \cdot x6 - 0.0028895 \cdot x7 - 0.0073295 \cdot x8 - 0.0009993 \cdot x9 + 0.0049779 \cdot x10 - 0.0017912 \cdot x11 + 0.0059143 \cdot x12 + 0.3336763 \cdot x13 + 0.0037772 \cdot x14$$

Los datos que extraemos del modelo cambiando la variable dependiente se resumen en la siguiente tabla:

	P-valor	Conclusión
Linealidad	7.225e-06 < 0.05	No hay linealidad
Homocedasticidad	0.002838 < 0.05	Hay heterocedasticidad
Normalidad	0.0386 < 0.05	No hay normalidad
Independencia	0.04174 < 0.05	No hay independencia
Colinealidad	Todo xn < 10	No hay multicolinealidad
R2	0.7536	
Error estándar residual	0.07756	
ANOVA	$2{,}2 \cdot 10^{\wedge -16} < 0.05$	Relación lineal

Tabla 21. Resumen modelo 1 con variable Y=victoria visitante

Obtenemos un nuevo modelo mediante el proceso secuencial:

$$y = 0.276681 – 0.2862557 \cdot x6 – 0.002849 \cdot x7 + 0.360355 \cdot x13 + 0003606 \cdot x14$$

Las conclusiones de este nuevo modelo para Y= cuotas iniciales corregidas para la victoria del equipo visitante se pueden resumir de la siguiente forma:

	P-valor	Conclusión
Linealidad	5.737e-06 < 0.05	No hay linealidad
Homocedasticidad	0.0003279 < 0.05	Hay heterocedasticidad
Normalidad	0.02913 < 0.05	No hay normalidad
Independencia	0.07076 > 0.05	Hay independencia
Colinealidad	Todo xn < 10	No hay multicolinealidad
R2	0.7491	
Error estándar residual	0.07709	
ANOVA	$2{,}2 \cdot 10^{\wedge -16} < 0.05$	Relación lineal

Tabla 22. Resumen modelo 2 con variable Y=victoria visitante

En conclusión, respecto al primer modelo (con todas las variables), podemos decir que tampoco se cumple ninguna de las hipótesis de los residuos. Por ello, rechazamos el modelo.

Respecto al segundo modelo (con las variables x6, x7, x13, x14), a pesar de que sí hay independencia, se rechazan las otras condiciones, por lo que, aunque mejora un poco los anteriores modelos, también debemos rechazarlo.

CAPÍTULO 5: CONCLUSIONES

A la luz de los análisis y las operaciones llevadas a cabo y explicadas a lo largo del presente trabajo, podemos concluir que no se ha podido encontrar un modelo explicativo que cumpla con las condiciones de los residuos, y que explique la probabilidad de que suceda un determinado suceso deportivo (victoria local, empate o visitante), en función de las variables seleccionadas. Si bien optimizando algún modelo obtenemos datos que lo mejoran, no podemos concluir que ninguno de ellos sea lo suficientemente bueno como para aceptarlo.

Esto no significa que de por sí el modelo sea malo, puesto que como hemos comprobado el coeficiente de determinación es elevado en la mayoría de ellos, sino que el modelo no es tan bueno como si cumpliera dichas condiciones.

Los resultados que hemos obtenido han podido ser causa del enfoque utilizado respecto a las variables, que creemos que no ha sido el adecuado. Quizás, con otro enfoque de las mismas, o eliminando e introduciendo otras variables distintas, el modelo podría haber dado buenos resultados. Por ello, sería conveniente un cambio de enfoque de las variables.

Por último, el trabajo nos ha permitido conocer el funcionamiento de las apuestas, de las cuotas y su íntima conexión -no evidente a simple vista- con la probabilidad. De este modo, nos ha ayudado a entender la dificultad de la determinación de la probabilidad, y comprender el laborioso y costoso trabajo que supone para las casas de apuestas este proceso, esencial para su obtención de rendimientos.

Así, sería conveniente que los apostadores fueran conscientes de las herramientas con las que cuentan las casas de apuestas, que les otorgan una ventaja sobre ellos en el largo plazo. Y esto podría ayudar a los apostadores a jugar de forma consciente e informada, y ver las apuestas como algo mucho más complejo de que lo parece en realidad, que les permita detener su conducta antes de que la misma pueda llevar a problemas.

BIBLIOGRAFÍA

[1] Abuín, J. R. (2007). Regresión lineal múltiple. IdEyGdM-Ld Estadística, Editor, 32.

[2] Consejo Empresarial del Juego (2021) Anuario del Juego en España.

[3] Dirección General de Ordenación del Juego (2015). Estudio sobre prevalencia, comportamiento y características de los usuarios de juegos de azar en España 2015.

[4] Dirección General de Ordenación del Juego (2019). Memoria anual.

[5] El Confidencial (2020). La ludopatía vuelve a crecer durante el periodo de pandemia del coronavirus. Consultado en: https://www.elconfidencialdigital.com/articulo/tendencias/ludopatia-vuelve-crecer-periodo-pandemia-coronavirus/20201001132515168320.html

[6] Osborne, J. (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation.

[7] Sakia, R. M. (1992) The Box-Cox transformation technique: a review. The Statistician.

[8] Universidad de Valencia. Bondad del ajuste. Consultado en: https://www.uv.es/ceaces/base/regresion/bondad.htm