MODELO DE REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIAS APLICADOS A JUGADORES DE PÁDEL PROFESIONALES

Grado en Ciencias de la Actividad Física y el Deporte

Manuel Jesús Lobo Mateos

MODELO DE REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIAS APLICADOS A JUGADORES DE PÁDEL

Manuel Jesús Lobo Mateos

Estudio presentado como parte de los requisitos para la obtención de título de Grado en Ciencias de la Actividad Física y Deportes por la Universidad de Sevilla

Tutorizada por

Prof. José María Fernández Ponce

.

RESUMEN

Este estudio tiene dos objetivos principales. En primer lugar, crear un marco teórico donde desarrollar un modelo estadístico para pronosticar el comportamiento de una variable en función de otra. Estos modelos son por un lado la regresión lineal simple, con el coeficiente de correlación de Pearson y sus test de hipótesis respectivos, y por otro lado tablas de contingencias con el test de Chi-cuadrado de Pearson y el test exacto de Fisher. Y en segundo lugar, llevar a la práctica dichos modelos para jugadores de pádel profesionales. Utilizando el modelo de regresión lineal para estudiar la relación entre la altura y el golpeo de remate, dando como resultado un modelo cuyo *p-value* es inferior a 0.05 para la relación entre la altura y los remates ganados y continuados para el grupo de hombres. Para el grupo de las mujeres todos los modelos obtienen un *p-value* superior al 0.05. En el caso de las variables categóricas, se estudia la relación entre las variables sexo, lateralidad y posición en la pista, obteniendo un *p-value* superior a 0.05, lo cual indica independencia entre ellas.

ABSTRACT

This study has got two main goals. Firstly, creating a theoretical frame where we can develop an statistical model to predict the behaviour of one variable in accordance to another one. These models are on the one hand, a simple lineal regression, with Pearson product-moment correlation coefficient and its hypothesis tests, and on the other hand; contingency tables with Pearson's chi-square test and Fisher's exact test. Secondly, implementing those models on professional paddle tennis players. Using the simple lineal regression in order to study the relationship between height and smash, the result for men is a p-value under 0.05, in the relationship between height and smashes which score and those which continue the game. The result for women is a p-value over 0.05 in all the smash types. In the case of categoric variables, the relationship among the variables of sex, laterality and position on court are studied and a p-value over 0.05 is obtained what means that these variables are independent among them.

ÍNDICE

RESUMEN 5

ABSTRACT 5

ÍNDICE 7

CAPÍTULO 1. 10

INTRODUCCION 10

SALUD 10

ENTRENAMIENTO Y RENDIMIENTO 10

CAPÍTULO 2. 13

MODELO DE REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIAS 13

2.1. REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIA 13

2.1.1. REGRESIÓN LINEAL SIMPLE Y COEFICIENTE DE CORRELACIÓN 13

2.1.1.1. COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON 15

2.1.1.2. REGRESIÓN LINEAL 18

CONCEPTOS BÁSICOS EN REGRESIÓN LINEAL SIMPLE 19

IDENTIFICACIÓN DEL MODELO 20

VALORACIÓN DEL MODELO 24

2.1.2. TABLAS DE CONTINGENCIA 26

2.2. CONTRASTES DE HIPÓTESIS 28

2.2.1. TEST DE HIPÓTESIS EN LA REGRESIÓN LINEAL 28

CÁLCULO DEL ESTADÍSTICO *t* Y DEL *p-value* 29

2.2.2. TEST DE HIPÓTESIS PARA LA CORRELACIÓN 30

2.2.3. TEST EXACTO DE FISHER 2X2 31

2.2.4. TEST CHI-CUADRADO X^2 DE INDEPENDENCIA 32

CAPÍTULO 3. 34

MODELO DE REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIAS PARA JUGADORES DE PÁDEL 34

3.1. METODOLOGÍA 34
3.2. ANÁLISIS ESTADÍSTICO 36
3.2.2. ANÁLISIS DEL REMATE EN FUNCION DE LA ALTURA 36
3.2.3. ANÁLISIS DE LAS VARIABLES SEXO, LATERALIDAD Y POSICIÓN .. 37
3.3. RESULTADOS 38
3.3.1. MODELO LINEAL DE REMATES EN FUNCION DE ALTURA 38
3.3.2. RELACIÓN ENTRE SEXO, LATERALIDAD Y POSICIÓN 47
3.4. DISCUSIÓN 48
3.5. CONCLUSIONES Y LÍNEAS FUTURAS 49
Bibliografía 50

CAPÍTULO 1.

INTRODUCCION

Se ha realizado una revisión bibliográfica de los estudios existentes relativos al deporte del pádel. Tras esta revisión se han catalogado los distintos estudios en dos categorías, salud y entrenamiento y rendimiento.

SALUD

En esta categoría se enmarcan varios artículos, el primero es un estudio descriptivo donde se valora los hábitos saludables en jugadores no profesionales de pádel, utilizando para ello el software SPSS y llegando a la conclusión de los malos hábitos de vida saludables en estos jugadores y observando un mayor porcentaje de sobrepeso en hombres que en mujeres. (Parrón Sevilla, Nestares Pleguezuelo, & De Teresa Galván, 2015).

Otros de los artículos encuadrados en esta categoría, analizan las diferentes lesiones sufridas en los jugadores de pádel y se concluye que las más frecuentes se sufren en los miembros inferiores y superiores seguidas de la espalda (Castillo-Lozano & Alvero-Cruz, 2016) y se realiza una descripción y clasificación de las lesiones en el miembro inferior y su recuperación. (García Navarro, López Martínez, De Prado Campos, & Sánchez Alcaraz Martínez, 2016).

Por último se encuentra un estudio donde se realiza una revisión bibliográfica de las lesiones sufridas en los deportes de raqueta y más concretamente en pádel. Estudian los procesos de readaptación de las diferentes lesiones y hacen valer el dato de que muchos procesos de readaptación llevan ligados tratamientos de fisioterapia y de actividad física. (Martínez Maqueda & Sarabia Cachadiña, 2018).

ENTRENAMIENTO Y RENDIMIENTO

En este marco se encontraron numerosos estudios, se han destacado algunos de ellos. El primero es un estudio sobre el tipo de esfuerzo que provoca el pádel en función de las modificaciones bioquímicas. Se realizó un análisis de sangre para hallar las modificaciones y se utilizó el software SPSS para el análisis de los datos recogidos. En él se concluye que el pádel provoca una situación catabólica, principalmente a nivel muscular y agudo. Su principal ruta metabólica es aeróbica, ver (Pradas, y otros, 2015).

Posteriormente en 2016 se publicó otro estudio donde se analizó las capacidades de la condición física que predominan en el pádel, para ello se utilizó una batería de test para capacidades físicas y el software SPSS para el análisis de los datos. Este artículo concluye que el pádel es un deporte complejo que conlleva muchos cambios de ritmo, lo que implica a las diferentes rutas metabólicas. No destaca una capacidad física sobre otra, más bien es la coordinación inter e intramuscular lo que permite generar los necesarios picos de fuerza y velocidad, y la obtención de un buen rango de movimiento facilita la realización del gesto técnico, ver (Pradas, Castellar, Quintas, & Arracó, 2016).

Un año más tarde en 2017 se publica un artículo que valora y compara la condición física en jugadores de pádel principiantes y avanzados. Se valora entre otras cosas las capacidades relacionadas con la condición física y la frecuencia con que se practica el pádel. Los resultados muestran que a mayor nivel de juego los jugadores presentan una menor fuerza en el tren inferior, igualmente cuanto mayor es la frecuencia de práctica, independientemente del nivel, se observa una menor capacidad cardiorrespiratoria. Esto es debido a la falta de entrenamiento específico y la necesidad de ello, ya que se observa que la práctica únicamente de pádel no es suficiente para la mejora de las capacidades físicas, ver (Herrera & Courdel.Ibáez, 2017).

Por último, el siguiente artículo que se destaca fue publicado en 2019, en este artículo se realiza un análisis técnico-táctico del pádel donde se estudian los diferentes golpeos que se realiza, la posición desde donde se golpea y la efectividad de los mismos. En él se menciona que los golpes más realizados son derecha y revés desde el fondo de la pista y que los golpes que mayores errores provocan al rival son los golpeados desde la red, como son voleas o remates. Estos últimos son los elegidos para analizar en este artículo, ver (Melledo-Arbelo, Baige Vidal, & Vivés Usón, 2019)

El último apartado de este capítulo trata de explicar conceptos desarrollados en los capítulos posteriores.

En el capítulo 2 se desarrolla los conceptos teóricos básicos de los modelos estadísticos empleados para el análisis de los datos. Estos modelos son, la regresión lineal simple, el coeficiente de correlación de Pearson y los distintos test de hipótesis utilizados para ello. También se desarrolla la teoría de las tablas de contingencias y los test del coeficiente de Chi-cuadrado de Pearson y el test exacto de Fisher.

En el capítulo 3 se lleva a la práctica los conceptos desarrollados en el capítulo 2. Previamente se recopilan datos sobres los jugadores profesionales de pádel y se crean las

siguientes variables: altura, remates ganados por partido, remates fallados por partidos, remates continuados por partido, sexo, lateralidad, posición donde juega, las cuales están explicadas en el capítulo 3. Por último se añade un capítulo de conclusiones finales y líneas futuras

CAPÍTULO 2.

MODELO DE REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIAS

A continuación se crea un marco teórico donde se desarrollan los modelos estadísticos utilizados para el análisis de los datos recopilados. En este marco se explica el modelo de regresión lineal simple y las tablas de contingencias.

2.1.REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIA

A menudo los entrenadores quieren conocer cómo será la progresión de un jugador en según qué posición, o bien llega un jugador nuevo al equipo y quieren saber cuál sería su posición ideal. Pongamos un ejemplo, a un equipo de vóley llegan 2 jugadores nuevos y miden 1,86 metros y 1,78 metros respectivamente. A priori cabe pensar que el jugador con mayor altura ocupará la posición de rematador y el de menos altura puede ocupar la posición de líbero, pero cuánto de cierto hay en esta afirmación o dicho de otro modo ¿cuál es la relación entre la altura de un jugador de vóley y su posición en el campo?

Desde un punto de vista matemático se puede demostrar la existencia o ausencia de relación entre dos o más variables. Estas relaciones pueden ser de muchos tipos (lineales, cuadráticas, cúbicas, logarítmicas, etc.). Nosotros nos centraremos en las relaciones de tipo lineal, ya que son estas las que con más frecuencia aparecen dentro del campo de las ciencias sociales donde podemos enmarcar las ciencias de la actividad física y el deporte.

Este capítulo trata de la regresión lineal donde se define el coeficiente de correlación, y sobre las tablas de contingencia y el test de Chi-Cuadrado de independencia.

2.1.1. REGRESIÓN LINEAL SIMPLE Y COEFICIENTE DE CORRELACIÓN

Como se menciona anteriormente a los entrenadores les interesa saber cuál es la posición que mejor se adapta a su jugador, pero ¿cómo podríamos saberlo? Para ello se debe comprobar si existe o no relación entre la variables físicas del jugador y las variables que definen la posición que puede ocupar. Por ejemplo se podría comparar la envergadura de un jugador de balonmano y el número de defensas que realiza.

Se dice que la relación entre dos variables, *y* (criterio) y *x* (predictora) es lineal si adoptan la siguiente forma:

$$y = a + bx$$

Donde *a* y *b* son dos constantes. Esta ecuación dibuja una recta en el eje de coordenadas, por esto a las variables que adoptan esta forma se les llaman lineales. La constante *a* es el origen de la recta, es decir es el valor de *y* cuando *x* vale 0. Por lo tanto es el punto donde la recta corta el eje de ordenadas. La constante *b* es la pendiente de la recta. La pendiente es el grado de inclinación de la recta con respecto al eje de abscisas. Si $b>0$ la función es creciente, es decir para valores más altos de *x* obtendremos valores más altos de *y*. Sin embargo si $b<0$ la función es decreciente para valores altos de *x* obtendremos valores bajos en *y*.

Todas las rectas que es posible representar en unos ejes de coordenadas pueden ser expresadas como una función lineal. Lo que identifica a cada una de ellas es el origen y la pendiente, es decir las constantes *a* y *b*.

Sin embargo, en ciencias sociales no disponemos de una función lineal que nos relaciones dos variables, sino que tenemos las puntuaciones de una muestra de elementos en dos variables que cuando la representamos en unos ejes de coordenadas nos dan una nube de puntos (diagrama de dispersión). Esta representación nos puede ayudar a evaluar qué tipo de función matemática describe la relación entre nuestras variables, **ver figura1**.

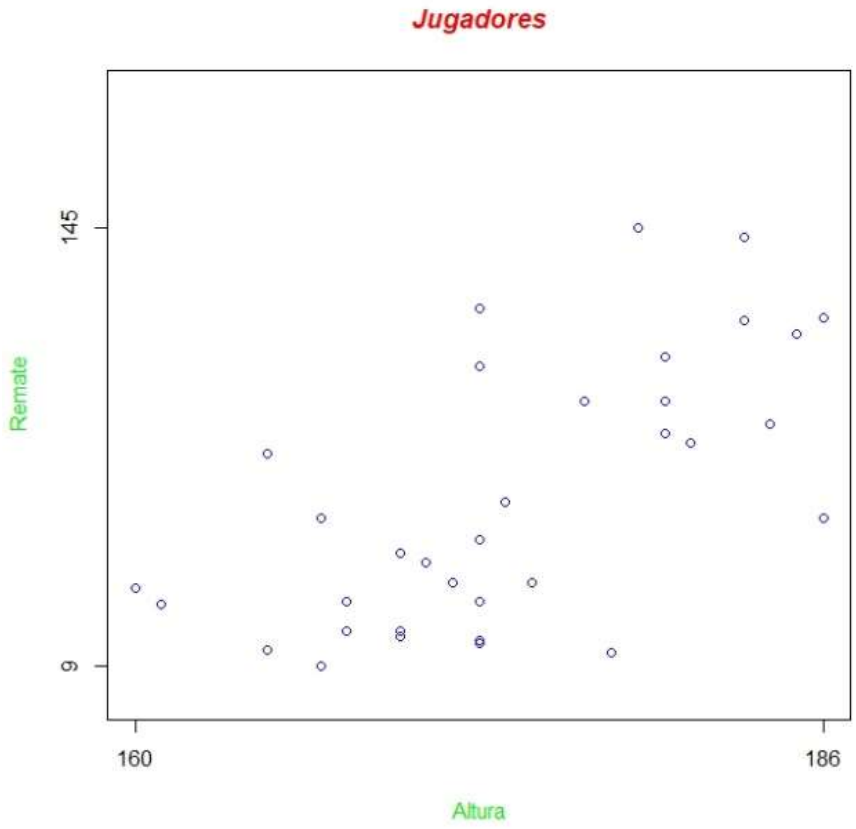

Figura 1: Diagrama de dispersión.

La regresión lineal simple trata de encontrar la función lineal que mejor se ajuste a esa nube de puntos. Es decir busca la función lineal que mejor describa la relación entre nuestras puntuaciones. Pero antes se debe determinar si nuestras variables se ajustan a un modelo lineal y de qué tipo o si, por el contrario, se podría decir que las variables son linealmente independientes. El estadístico que se utiliza para cuantificar el grado de relación lineal entre dos variables cuantitativas se llama coeficiente de correlación de Pearson.

2.1.1.1. COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

A continuación se explica que tipos de relaciones lineales existen y cómo calcular el coeficiente de correlación de Pearson y cuál es su significado.

Si para los valores bajos un variables corresponde los valores más bajos de la segunda variables, para los intermedios corresponden los valores intermedios de la segunda y para los valores más altos corresponden los valores más altos de la otra variable, diremos que mantienen una relación lineal directa o una relación lineal positiva, **ver figura 1**.

Si por el contrario los valores bajos de la primera variable corresponden con los valores altos de la segunda variables, los intermedios coinciden con los valores intermedios de la otra variables y los valores más altos de la primera variables corresponden a los valores más bajos de la segunda variables, diremos que nuestras variables mantienen una relación lineal inversa o una relación lineal negativa.

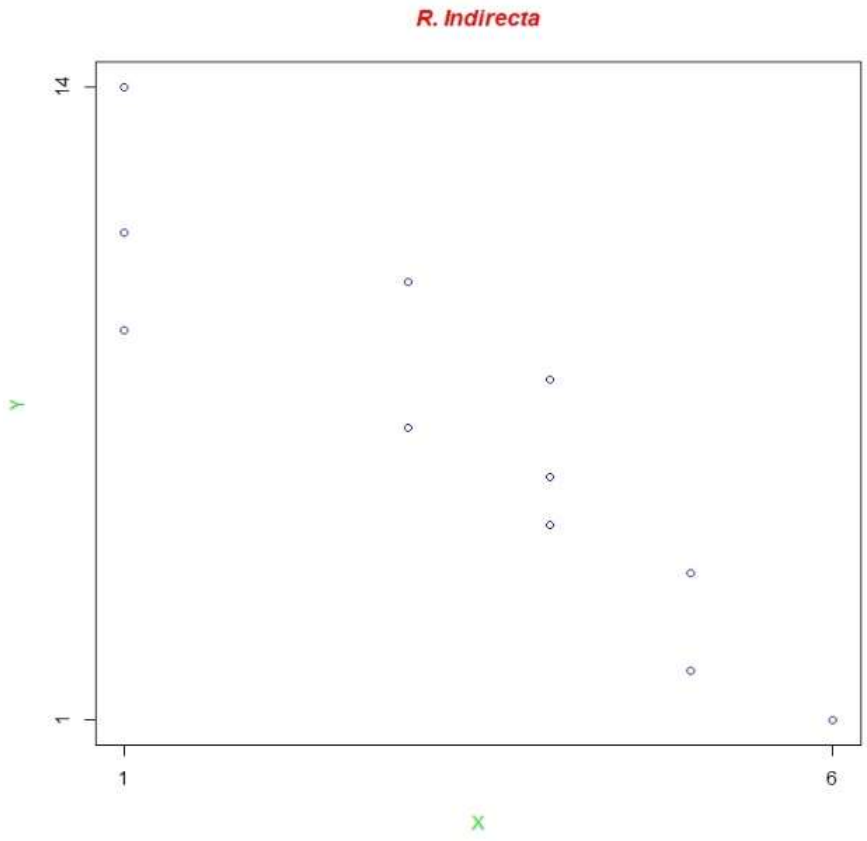

Figura 2: Diagrama de dispersión.
Variables con relación lineal inversa

Por último, si los valores altos, medios y bajos de una variable están emparejados por igual con valores altos, medios y bajos de otra variable decimos que hay ausencia de relación lineal o que las variables son linealmente independientes.

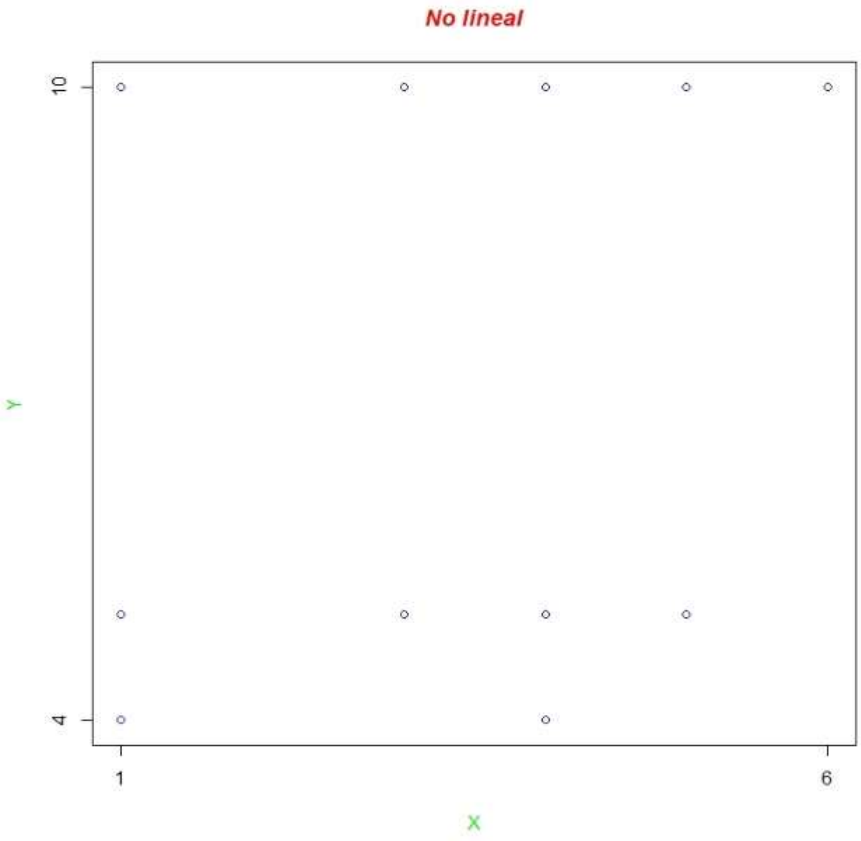

Figura 3: Diagrama de dispersión.
Variables con ausencia de relación lineal

El cálculo del coeficiente de correlación de Pearson se realiza con la siguiente expresión:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Donde y_i y x_i representan los valores obtenidos en las variables y n es el número total de observaciones.

Veamos un ejemplo de cómo calcular este coeficiente:

	X_i	Y_i	$X_i \cdot Y_i$	X_i^2	Y_i^2
	8	7	56	64	49
	9	8	72	81	64
	7	4	28	49	16
	6	5	30	36	25
	4	2	8	16	4
TOTAL	34	26	194	246	158

Tabla 1: Tabla de frecuencias de dos variables ejemplo x e y.

$$r_{xy} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}} =$$

$$= \frac{5 \cdot 194 - 34 \cdot 26}{\sqrt{5 \cdot 246 - 34^2}\sqrt{5 \cdot 158 - 26^2}} = \frac{86}{\sqrt{74}\sqrt{114}} = 0{,}936332.$$

Pero ¿qué significado tiene este coeficiente? pues bien, una de las propiedades de r_{xy} es que es un coeficiente adimensional, es decir, que es independiente de las unidades en que están expresadas las variables. Su valor está comprendido entre [-1 y 1]. El signo de este coeficiente marca el modelo lineal que siguen nuestras variables, si el signo es positivo las variables siguen un modelo lineal directo y si es negativo el modelo es inverso. Y el valor indica el grado de relación que mantienen las variables, es decir cuanto más se acerque a -1 o 1, mayor es el grado de relación entre las variables. Decimos que dos variables tienen una relación lineal positiva perfecta si el valor de r_{xy} es igual a 1, de igual modo mantienen una relación lineal inversa perfecta si el valor es -1, y por el contrario si el valor es igual a 0 podemos afirmar la ausencia de relación lineal entre las variables. Esta ausencia de relación lineal no significa que no exista otro tipo de relación entre las variables, tan solo que no existe una función lineal que pueda expresar una variable en función de la otra. Otra propiedad importante de este coeficiente es que la correlación entre *x* e *y* es igual que entre *y* y *x*.

2.1.1.2. REGRESIÓN LINEAL

Una vez descubierto que tipo de modelo de relación mantienen nuestras variables, en nuestro caso un modelo lineal, toca hallar la función que las describen. Para hallar dicha función utilizaremos la regresión lineal.

La regresión lineal es una técnica estadística que trata de determinar relaciones de dependencia de tipo lineal entre una variable criterio o endógena, respecto de una o varias variables predictoras o exógenas. Esto quiere decir que a partir de las puntuaciones en una o varias variables vamos a predecir, pronosticar, las puntuaciones en otra variable a partir de una ecuación lineal.

Pongamos un ejemplo que nos ayude a entender este concepto. Supongamos que en un estudio previo hemos hallado que existe un coeficiente de correlación de -0,93 entre el dominio del bote en baloncesto y la altura de los jugadores. Esto quiere decir que los jugadores con menor altura tienen un mejor dominio del bote que los jugadores más altos. Ahora supongamos que entrenamos por primera vez a un equipo de baloncesto del cual tenemos la información

antropométrica de los jugadores pero desconocemos su habilidad con el balón. Para la posición del base necesitamos al jugador que mejor maneje el bote, pero ¿cómo podemos saber quién es ese jugador? pues con la altura de los jugadores podríamos pronosticar cual será el dominio del bote de nuestros jugadores. Esto es lo que haremos con la regresión lineal simple: a partir de las puntuaciones de los sujetos en una variable vamos a predecir, pronosticar, las puntuaciones que obtendrían en otra variable a partir de una ecuación lineal.

Evidentemente del rendimiento del jugador que ocupa la posición de base no sólo se define por su dominio del bote, sino que intervienen otras muchas variables. En este caso lo que se hace es predecir la variable a partir de un conjunto de variables con las que está relacionada. A esto se le llama regresión lineal múltiple.

Nosotros nos centraremos en la explicación y cálculo de la regresión lineal simple.

CONCEPTOS BÁSICOS EN REGRESIÓN LINEAL SIMPLE

Como ya sabemos la relación entre dos variables es lineal si es de la forma:

$$y = a + bx$$

En regresión se le llama variables predictora a la variable utilizada para hacer los pronósticos (x) y variable criterio a la variable que se pretende hallar (y). Se denomina recta de regresión de y sobre x a la recta que nos permite predecir y a partir de los valores de x, es decir:

$$y' = a + bx$$

Con esta recta de regresión no se hallan los valores de y sino los valores que se pronostican en la variable y. A estos valores se les denominan pronósticos y'.

Siguiendo con nuestro ejemplo de los jugadores de baloncesto, el dominio del bote sería nuestra variable criterio y la variable predictora sería la estatura.

En la regresión lineal simple se deben seguir los siguientes pasos. La identificación del modelo, que supone encontrar la recta de regresión que mejor represente la relación entre las variables bajo estudio, es decir, la recta que mejor nos permita pronosticar y a partir de x. Valorar el modelo, es decir, evaluar como de bueno es el modelo para predecir y a partir de x. Aplicación del modelo para predecir y a partir de x. Para explicar estos pasos tomaremos el siguiente ejemplo, donde a 5 sujetos se les ha medido el porcentaje de grasa corporal con pesaje

hidrostático (%GC), los pliegues cutáneos (Pli), índice cintura y cadera (Ci/Ca), y el índice de estatura al cuadrado (IMC). Expresados en la siguiente tabla:

Sujeto	*%GC*	*Pli*	*Ci/Ca*	*IMC*
1	11,5	39,78	23,3	0,83
2	10,7	28,74	21,4	0,82
3	14	52,86	24,7	0,87
4	11,2	35,41	22,1	0,81
5	12,7	45,21	24,1	0,85

Tabla 2: Medición antropométrica

IDENTIFICACIÓN DEL MODELO

Empezaremos dibujando una matriz de diagramas de dispersión para nuestras variables. En ella podemos identificar qué modelo de relación mantienen entre si las diferentes mediciones.

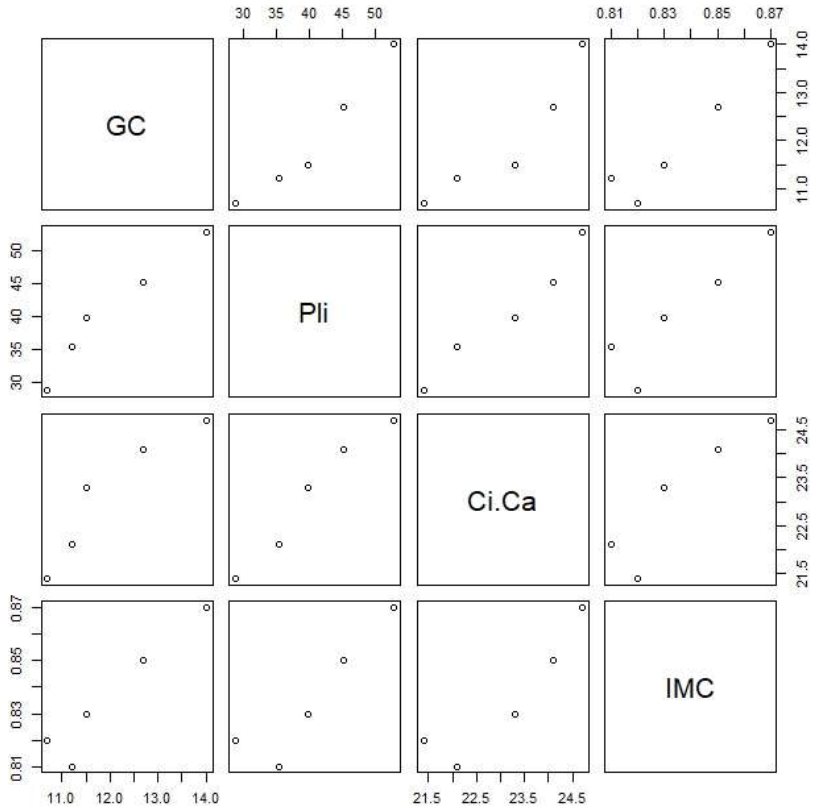

Figura 4: Matriz de dispersión.
Medidas antropométricas.

Como podemos observar existe una relación lineal entre el porcentaje de grasa y la medición de los pliegues cutáneos. Pues la identificación del modelo consiste en encontrar la recta que

mejor describe esta relación. ¿Cómo identificamos dicha recta? Para ello utilizamos el criterio de mínimos cuadrados. Con este criterio tratamos de hacer mínimos los errores, es decir, tratamos de minimizar las distancias entre las puntuaciones obtenidas en *y* y sus pronósticos *y'*.

Lo que tratamos de buscar es la recta con la que cometamos el menor error posible de forma global para todos los sujetos. Si hacemos la media de los errores, los errores positivos se compensarían con los negativos, para evitar este problema y hacernos una idea del error global cometido, calcularemos la media de los errores al cuadrado, de esta forma se anularían los signos negativos de los errores. Expresado formalmente:

$$S_{y\cdot x}^2 = \frac{\sum(y_i - y_i')^2}{n}.$$

Esta expresión se denomina error cuadrático medio. La recta que seleccionaremos es aquella que haga mínimo el error cuadrático medio, a esto se le denomina criterio de mínimo cuadrados.

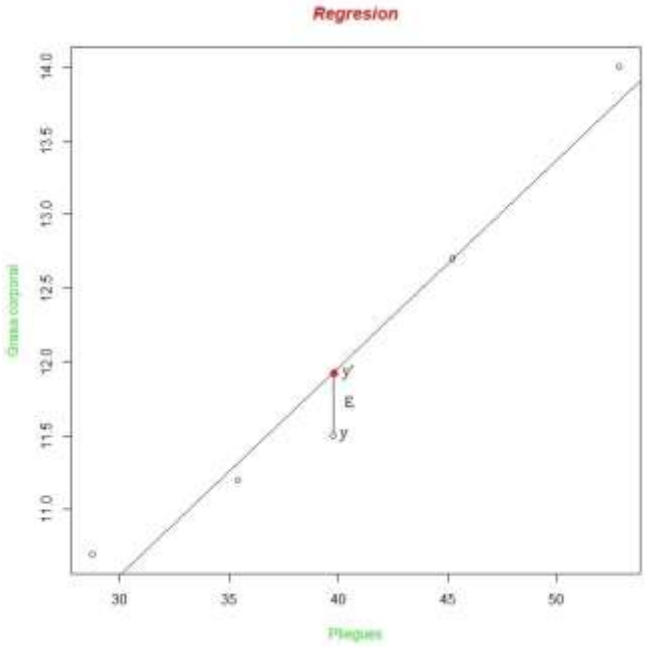

Figura 5: Representación gráfica del error cometido del pronóstico y' sobre y.

$$min \sum E_i^2 = \min \sum (y_i - y'_i)^2$$

La recta que hace mínimo el error cuadrático medio tiene como pendiente:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad \text{o} \qquad B = r_{xy}\frac{S_y}{S_x}$$

Y como origen:

$$a = \bar{y} - b \cdot \bar{x},$$

donde $\bar{y}$ y $\bar{x}$ representan la media de *y* y *x*, respectivamente.

Calculemos la recta que representa la relación entre el porcentaje de grasa y la medición de pliegues, y su error cuadrático medio.

Sujeto	*%GC(y_i)*	*Pli(x_i)*	$x_i y_i$	x_i^2
1	11,5	39,78	457,47	1582,4484
2	10,7	28,74	307,518	825,9876
3	14	52,86	740,04	2794,1796
4	11,2	35,41	396,592	1253,8681
5	12,7	45,21	574,167	2043,9441
∑	60,1	202	2475,787	8500,4278

Tabla 3: Porcentaje de grasa corporal y medición de pliegues cutáneos. Cálculos para la función lineal.

Calculemos la pendiente y el punto de corte con el eje *y* de nuestra recta:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{5 \cdot 2475{,}787 - 202 \cdot 60{,}1}{5 \cdot 8500{,}4278 - 202^2} = 0{,}1406$$

$$a = \bar{y} - b\bar{x} = \frac{60{,}1}{5} - 0{,}1406\frac{202}{5} = 6{,}3398$$

La ecuación de regresión que nos permite pronosticar la grasa corporal a partir de la medición de los pliegues cutáneos es:

$$y' = 6{,}3398 + 0{,}1406x$$

A continuación dibujaremos nuestra recta:

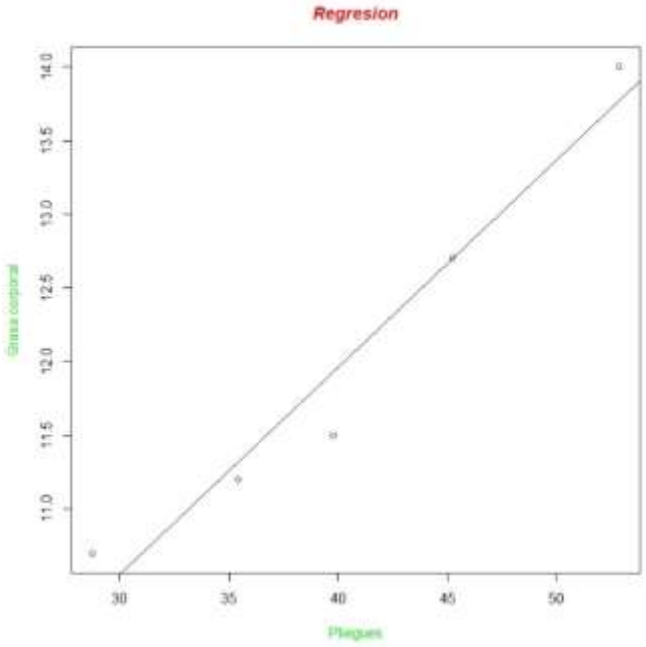

Figura 6: Representación de la función lineal del porcentaje de grasa corporal sobre la medición de los pliegues cutáneos.

Como podemos observar no todos los puntos recaen sobre nuestra recta. Calculemos cual es el error cuadrático medio:

Sujeto	*%GC(Yi)*	*Pli(Xi)*	*Y'*	*Y-Y'*	*$(Y-Y')^2$*
1	11,5	39,78	11,9329	-0,4329	0,1874
2	10,7	28,74	10,3806	0,3194	0,1020
3	14	52,86	13,7719	0,2281	0,0520
4	11,2	35,41	11,3184	-0,1184	0,0140
5	12,7	45,21	12,6963	0,0037	1,34983E-05
∑	60,1	202			0,3554

Tabla 4: Porcentaje de grasa corporal y medición de pliegues cutáneos. Cálculos para el error cuadrático medio.

$$S_{y\cdot x}^2 = \frac{\sum(y_i - y_i')^2}{n} = \frac{0{,}3554}{5} = 0{,}0711$$

Una vez obtenida la recta y calculado el error cuadrático medio debemos valorar nuestro modelo.

VALORACIÓN DEL MODELO

Como hemos comentado anteriormente, nuestro objetivo es encontrar la función lineal que nos permita pronosticar una variable a partir de los resultado obtenidos en otra, pero cometiendo el menor error posible. Pero, ¿cuándo la suma de los errores al cuadrado es pequeña y cuándo es grande? Esta valoración se realiza a partir del cuadrado de la correlación de Pearson, que en el contexto de regresión recibe el nombre de coeficiente de determinación.

Para poder realizar la valoración del modelo debemos explicar que es la descomposición de la varianza criterio. La descomposición de la varianza criterio expresa que la varianza de la variable criterio es igual a la suma de la varianza de los pronósticos realizados a partir de la recta de regresión más el error cuadrático medio o varianza de los errores. Expresado formalmente:

$$S_y^2 = S_{y'}^2 + S_{y \cdot x}^2$$

En esta ecuación aparece dos criterios para ver cómo se asemeja nuestro modelo a la variable criterio. Uno es la varianza de los pronósticos y el otro la varianza de los errores. Nuestro modelo será mejor cuanto menor sea la varianza de los errores, es decir que la varianza de la variable criterio sea igual a la varianza de los pronósticos.

Si dividimos ambos miembros por la varianza de *y*:

$$\frac{S_y^2}{S_y^2} = \frac{S_{y'}^2}{S_y^2} + \frac{S_{y \cdot x}^2}{S_y^2} = 1$$

De esta ecuación obtenemos la proporción de varianza de la variable criterio explicada por el modelo de regresión y la proporción de error que cometeríamos al utilizar la recta de regresión. Además se puede demostrar que:

$$\frac{S_{y'}^2}{S_y^2} = r_{xy}^2$$

Podemos deducir que nuestro modelo no cometerá errores cuando el coeficiente entre la varianza de los pronósticos y la de la variable criterio, es decir, la proporción de varianza explicad por *x*, sea 1. Este caso se dará cuando la correlación sea 1 o -1. Esto quiere decir que el ajuste de la recta es perfecto.

Por lo contario cometeremos el máximo error cuando la varianza de los errores sea igual a la de la variable criterio. Esto ocurrirá cuando la correlación sea 0 e indica la ausencia de relación lineal entre las variables.

La correlación al cuadrado, denominada coeficiente de determinación nos indica la proporción de varianza de la variable criterio que queda explicada por el modelo lineal. Se suele multiplicar por 100 para expresarla en términos de porcentaje.

La proporción de error sería igual a 1 menos el coeficiente de determinación. Este nos indica la proporción de la varianza de la variable criterio que no queda explicada por el modelo lineal. Se expresa así:

$$\frac{S^2_{y\cdot x}}{S^2_y} = 1 - r^2_{xy}$$

Volvamos a nuestro ejemplo y veamos qué porcentaje de la varianza de nuestra variable queda explicada mediante nuestro modelo de regresión.

Ya habíamos calculado nuestra ecuación y el error cuadrático medio:

$$y' = 6{,}3398 + 0{,}1406x \qquad S^2_{y\cdot x} = \frac{\sum(y_i - y'_i)^2}{n} = \frac{0{,}3554}{5} = 0{,}0711$$

Ahora calculemos la varianza del criterio (%GC) y de los pronósticos (*y'*):

Sujeto	*%GC(Yi)*	*Pli(Xi)*	*Y'*	*(Y-Y')²*	*Y²*	*Y'²*	*X²*
1	11,5	39,78	11,9329	0,1874	132,25	142,3933	1582,4484
2	10,7	28,74	10,3806	0,1020	114,49	107,7578	825,9876
3	14	52,86	13,7719	0,0520	196	189,6657	2794,1796
4	11,2	35,41	11,3184	0,0140	125,44	128,1072	1253,8681
5	12,7	45,21	12,6963	1,34983E-05	161,29	161,1967	2043,9441
∑	60,1	202	60,1002	0,3554	729,4700	729,1207	8500,4278

Tabla 5: Porcentaje de grasa corporal y medición de pliegues. Cálculo del coeficiente de determinación.

$$S^2_y = \frac{\sum y_i^2}{n} - \bar{y}^2 = \frac{729{,}47}{5} - \left(\frac{60{,}10}{5}\right)^2 = 133{,}874$$

$$S^2_{y'} = \frac{\sum y_i'^2}{n} - \bar{y}'^2 = \frac{729{,}1207}{5} - \left(\frac{60{,}1002}{5}\right)^2 = 133{,}8029$$

Con la descomposición de la varianza:

$$S^2_y = S^2_{y'} + S^2_{y\cdot x}$$

En nuestro ejemplo:

$$133{,}874 = 133{,}8029 + 0{,}0711$$

El coeficiente de Pearson:

$$r_{xy} = \frac{5 \cdot 2475{,}7870 - 202 \cdot 60{,}1}{\sqrt{5 \cdot 8500{,}4278 - (202)^2} \cdot \sqrt{5 \cdot 729{,}47 - (60{,}1)^2}} = 0{,}9745$$

El coeficiente de determinación:

$$r_{xy}^2 = 0{,}9497$$

Con este modelo se explica el 94,97 por 100 de la varianza criterio.

Una vez valorado nuestro modelo sólo nos queda aplicarlo

2.1.2. TABLAS DE CONTINGENCIA

Hasta ahora se ha explicado cómo se relacionan dos variables cuantitativas, a continuación se estudiará la independencia entre dos variables cualitativas. Para ello utilizaremos las tablas de contingencias.

"Sean *X* e *Y* dos variables categóricas de respuesta, *X* con *I* categorías e *Y* con *J* categorías. Un sujeto puede venir clasificado en una de las *I*×*J* categorías, que es el número posible de categorías que existe. Las respuestas (*X*.*Y*) de un sujeto elegido aleatoriamente de alguna población tiene una distribución de probabilidad. Una tabla rectangular que tiene *I* filas para las categorías de *X* y *J* columnas para las categorías de *Y* muestra esta distribución". (Millán Díaz, 2017).

Dos variables son independientes cuando la distribución de una variable es similar sea cual sea el nivel que examinemos de la otra. Esto se traduce en una tabla de contingencias en la que las frecuencias de las filas y columnas son aproximadamente proporcionales. Para reconocerlo se pueden observar en la tabla de contingencia si los porcentajes por filas o columnas son similares.

Pongamos un ejemplo, tenemos dos variables nominales como son la comunidad autónoma donde el jugador o jugadora está federado y su lateralidad (si es diestro o zurdo), y queremos saber si existe independencia entre ellas, ya no podemos calcular un coeficiente de correlación que indique si a mayor comunidad autónoma, mayor o menor lateralidad. Que dos variables nominales sean independientes, significa que las proporciones de *x* serán iguales en cada

categoría de *y*. Si *x* e *y* no son independientes, entonces las proporciones de *x* serán diferentes en las distintas categorías de *y*.

Esto significa que si las variables de nuestro ejemplo son independientes, ser diestro o zurdo no depende de la comunidad donde están federados los jugadores o jugadoras, y que la proporción de diestros y zurdos será similar en cada comunidad autónoma. Pongamos un ejemplo:

	ANDALUCIA	
Diestro	750	0,75
Zurdo	250	0,25
TOTAL	1000	

	EXTREMADURA	
Diestro		
Zurdo		
TOTAL	600	

Tabla 6: Tabla de frecuencias de dos variables cualitativas. Comunidad autónoma y lateralidad del jugador/a. Comunidad de Extremadura sin completar. (Variables independientes)

Supongamos que nuestras variables, comunidad autónoma y lateralidad, son independientes, ¿qué frecuencia de diestros y zurdos tendría que haber en Extremadura? Calculémoslo:

	ANDALUCIA	
Diestro	750	0,75
Zurdo	250	0,25
TOTAL	1000	

	EXTREMADURA	
Diestro	450	0,75
Zurdo	150	0,25
TOTAL	600	

Tabla 7: Tabla de frecuencias de dos variables cualitativas. Comunidad autónoma y lateralidad el jugador/a. (Variables independientes)

Como hemos supuesto que son independientes, sabemos que la proporción debe ser similar y podemos calcular el número de diestros y zurdos federados en Extremadura. En este caso 450 diestros y 150 zurdos. Esto se expresa en una tabla con *i* filas para cada categoría de *x* y *j* columnas para cada categoría de *y*. En nuestro ejemplo:

	ANDALUCIA	*EXTREMADURA*	
Diestro	750	450	1200
Zurdo	250	150	400
	1000	600	1600

Tabla 8: Tabla de contingencias. Comunidad autónoma y lateralidad del jugador/a. (Variables independientes)

La independencia entre esas dos variables implica que cada frecuencia absoluta conjunta es igual al producto de sus frecuencias marginales dividido entre la frecuencia total:

Andaluces diestros=1200·1000/1600=750

Andaluces zurdos=400·100/1600=250

Extremeños diestros=1200·600/1600=450

Extremeños zurdos=400·600/1600=150

Estas frecuencias son las que esperaríamos obtener en caso de que *x* e *y* fuesen independientes y las llamamos frecuencias esperadas (e_{ij}, es la frecuencia esperada en la fila *i* y columna *j*).

Si existe dependencia entre las variables, en nuestro ejemplo comunidad autónoma y lateralidad, las proporciones de *x* no serían iguales en las distintas categorías de *y*. En este caso las frecuencias esperadas (suponiendo independencia) no coincidirían con las frecuencias observadas (o_{ij}) en la tabla. Pongamos un ejemplo:

	ANDALUCIA	*MURCIA*	
Diestro	750	220	970
Zurdo	250	410	660
TOTAL	1000	630	1630

Tabla 9: Tabla de contingencias. Comunidad autónoma y lateralidad del jugador/a. (Variables dependientes)

Calculemos la frecuencia esperada para diestros en Andalucía:

Diestros en Andalucía=1000·970/1630=596,09 $\neq$ 750

Vemos como la frecuencia esperada es distinta a la frecuencia observada. Para valorar la independencia de las variables nominales utilizamos el estadístico coeficiente chi-cuadrado de Pearson (χ^2).

2.2. CONTRASTES DE HIPÓTESIS

2.2.1. TEST DE HIPÓTESIS EN LA REGRESIÓN LINEAL

El estudio de regresión se aplica a una muestra, pero su objetivo es obtener un modelo lineal que explique la relación entre las dos variables en toda la población. Esto significa que el

modelo generado es una estimación de la relación poblacional a partir de la relación que se observa en la muestra y, por tanto, está sujeta a variaciones. Para cada uno de los parámetros de la ecuación de regresión lineal simple (*a* y *b*) se puede calcular su significancia (*p-value*) y su intervalo de confianza. El test estadístico más empleado es el *t-test*.

El test de significancia para la pendiente *b* del modelo lineal considera como hipótesis:

- H_0: No hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es cero, $b=0$.
- H_α: Sí hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es distinta de cero, $b\neq0$.

De esta misma forma también se aplica a *a*.

CÁLCULO DEL ESTADÍSTICO *t* Y DEL *p-value*

$$t = \frac{\hat{b}-0}{SE(\hat{b})}; t = \frac{\hat{a}-0}{SE\ (\hat{a})}$$

El error estándar de *a* y *b* se calcula con las siguientes ecuaciones:

$$SE(\hat{a})^2 = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{\bar{x}})^2})$$

$$SE(\hat{b})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

La varianza del error σ^2 se estima a partir del *Residual Estándar Error* (*RSE*), que puede entenderse como la diferencia promedio que se desvía la variable respuesta de la verdadera línea de regresión. En el caso de regresión lineal simple, *RSE* equivale a:

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)}$$

Los grados de libertad es igual al número de observaciones menos el número de predicciones menos 1 y el *p-value* es igual a la probabilidad de que el valor absoluto de t sea mayor al valor calculado de t. Expresado formalmente:

- Grados de libertad (*df*) = número observaciones – número predictores – 1
- *p-value* = P(|t| > valor calculado de t)

Una vez hallado el valor *t*, este se compara con su valor crítico. El valor crítico de *t* se halla con la tabla de distribución *t*, en función de los grados de libertad (*n* - 2), el valor de α del intervalo de confianza y en este caso es una prueba de dos colas.

- Si |*t*| > El valor critico de *t*, se rechaza H_0 y por lo tanto existe relación lineal entre ambas variables.
- Si |*t*| < El valor critico de *t*, no se rechaza H_0 y por lo tanto no existe evidencias estadísticas para afirmar que exista una relación lineal entre ambas variables.

2.2.2. TEST DE HIPÓTESIS PARA LA CORRELACIÓN

Existen dos métodos para las pruebas de hipótesis para la correlación. Las hipótesis son:

- H_0: $p = 0$ No hay relación lineal entre ambas variables
- H_α: $p \neq 0$ Sí hay relación lineal entre ambas variables

Uno de los estadísticos usado para estas pruebas es el *t-test.* El procedimiento es similar al anterior. Para hallar *t* se utiliza esta fórmula:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Una vez hallado su valor, se vuelve a comparar con el valor crítico de *t*.

- Si |*t*| > El valor critico de *t*, se rechaza H_0 y por lo tanto existe relación lineal entre ambas variables.
- Si |*t*| < El valor critico de *t*, no se rechaza H_0 y por lo tanto no existe evidencias estadísticas para afirmar que exista una relación lineal entre ambas variables.

Otro estadístico utilizado es *r*, el coeficiente de correlación de Pearson. Una vez hallado este coeficiente, debemos compararlo con su valor crítico. El valor crítico de *r* se halla con la tabla de valor crítico del coeficiente *r* de correlación de Pearson. Para ello hace falta saber el tamaño de la muestra *n* y el valor de *α* del intervalo de confianza.

- Si |*r*| > El valor critico de *r*, se rechaza H_0 y por lo tanto existe relación lineal entre ambas variables.
- Si |*r*| < El valor critico de *r*, no se rechaza H_0 y por lo tanto no existe evidencias estadísticas para afirmar que exista una relación lineal entre ambas variables.

2.2.3. TEST EXACTO DE FISHER 2X2

El test exacto de Fisher permite analizar si dos variables categóricas (dicotómicas, binarias) están asociadas cuando la muestra a estudiar es demasiado pequeña y no se cumplen las condiciones necesarias para la aplicación del test del coeficiente Chi-Cuadrado de Pearson.

El test exacto de Fisher consiste en calcular la probabilidad asociada a cada una de las tablas 2 x 2 que se pueden formar manteniendo los mismo totales de filas y columnas que tiene la tabla observada, a esto se le denomina prueba de permutaciones. Todas estas posibles tablas y sus probabilidades se obtiene bajo la hipótesis nula de independencia de las dos variables que se están considerando.

- H_0: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.
- H_α: Las variables son dependientes, una variables varía entre los distintos niveles de la otra variable.

Es decir se calcula la distribución de todas las tablas posibles bajo independencia y se contrasta con la observada.

VARIABLE y	*VARIABLE x*		
	x_1	x_2	TOTAL
y_1	a	b	a + b
y_2	c	d	c + d
TOTAL	a + c	b + d	n

Tabla 10: Tabla genérica de dos variables categóricas binarias.

Es igual a:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+b)!\,(a+c)(b+d)}{a!\,b!\,c!\,d!\,n!}$$

En esta fórmula se calculan todas las posibles formas en las que podemos disponer *n* sujetos en una tabla 2 x 2 de modo que los totales de filas y columnas estén fijos, (a + b), (c + d), (a + c) y (b + d).

La probabilidad anterior deberá calcularse para todas las tablas de contingencia que puedan formarse con los mismos totales marginales que la tabla observada. Posteriormente, estas probabilidades se usan para calcular el *p-value* asociado al test exacto de Fisher. Este *p-value*

indicará la probabilidad de obtener una diferencia entre grupos mayor o igual a la observada, bajo la hipótesis nula de independencia. Si esta probabilidad es pequeña (*p-value* < 0.05) se deberá rechazar la hipótesis de partida y deberemos asumir que las dos variables no son independientes, sino que están asociadas. En caso contrario, se dirá que no existe evidencia estadística de asociación entre ambas variables.

2.2.4. TEST CHI-CUADRADO X^2 DE INDEPENDENCIA

Esta medida se basa en calcular la diferencia entre las frecuencias observadas en nuestra tabla y las frecuencias esperadas si se diese la condición de independencia de las variables. Expresado formalmente:

$$X^2 = \frac{\sum_i (o_{ij} - e_{ij})^2}{e_{ij}}$$

Donde o_{ij} representa la frecuencia conjunta observada en la fila *i* y columna *j* de nuestra tabla y e_{ij} representa la frecuencia conjunta esperada para ese dato suponiendo que exista independencia entre las variables. La frecuencia conjunta esperada se expresa:

$$e_{ij} = \frac{(frecuencia\ marginal\ de\ la\ fila\ i) \cdot (frecuencia\ marginal\ de\ la\ columna\ j)}{n}$$

Las hipótesis de partida son:

- H_0: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.
- H_α: Las variables son dependientes, una variables varía entre los distintos niveles de la otra variable.

A continuación hallaremos el coeficiente Chi-Cuadrado para los dos ejemplos anteriores, para explicar su interpretación.

	ANDALUCIA	*EXTREMADURA*	
Diestro	750	450	1200
Zurdo	250	150	400
	1000	600	1600

Tabla 11: Tabla de contingencias. Comunidad autónoma y lateralidad del jugador/a. (Variables independientes)

$$X^2 = \frac{(750-750)^2}{750} + \frac{(450-450)^2}{450} + \frac{(250-250)^2}{250} + \frac{(150-150)^2}{150} = 0$$

	ANDALUCIA	*MURCIA*	
Diestro	750	220	970
Zurdo	250	410	660
TOTAL	1000	630	1630

Tabla 12: Tabla de contingencias. Comunidad autónoma y lateralidad del jugador/a. (Variables dependientes).

F. ESPERADAS	*ANDALUCIA*	*MURCIA*
Diestro	595,09	374,91
Zurdo	404,91	255,09

Tabla 13: Tabla de contingencias con las frecuencias esperadas. Comunidad autónoma y lateralidad del jugador/a. (Variables dependientes)

$$X^2 = \frac{(750 - 595{,}09)^2}{595{,}09} + \frac{(250 - 404{,}91)^2}{404{,}91} + \frac{(220 - 374{,}91)^2}{374{,}91} + \frac{(410 - 255{,}09)^2}{255{,}09} = 257{,}66$$

Como observamos en los dos casos nos ha dado valores diferentes del coeficiente. En el primer caso el valor de χ^2=0, esto implica la existencia de independencia entre las variables. Y en el segundo caso el valor de χ^2>0 lo que implica que las variables no son independientes.

CAPÍTULO 3.

MODELO DE REGRESIÓN LINEAL Y TABLAS DE CONTINGENCIAS PARA JUGADORES DE PÁDEL

En este capítulo se lleva a la práctica todo lo desarrollado en el capítulo anterior. Para ello se ha extraído los datos de jugadores de pádel profesional, que compiten en el torneo más importante de este deporte el WPT (World Padel Tour).

3.1. METODOLOGÍA

Para realizar este proyecto recogimos los datos mediante dos procesos. Para los datos de los jugadores/as se recurrió a sus fichas personales de la web oficial de WPT https://www.worldpadeltour.com/ y para el resto de variables se recogieron mediante el visionado de videos. Estos videos están colgados en la plataforma YouTube en el canal oficial de WPT, https://www.youtube.com/user/WorldPadelTourAJPP. Se eligió tres de los diecinueve torneos disputados en la temporada 2019 (debido a la Pandemia producida por el Covid-19 la temporada 2020 se vio suspendida temporalmente) que fueron Open de Menorca, Open de Córdoba y Master-Final que se disputó en Barcelona. Las fases de competición de las que se recogieron datos fueron a partir de cuartos de final en los tres torneos, visionando así un total de 38 partidos.

Los videos fueron visionados durante los meses de marzo y abril de 2020 para el posterior vuelco de datos a un fichero Excel. El fichero Excel fue importado al software R para el análisis de datos.

Las variables recogidas son: la altura de los jugadores/as, sexo o género, remates totales que realizan, remates ganados, remates fallados, remates continuados, partidos jugados, mano dominante y posición en pista.

A continuación se define las variables que pueden generar más duda como son los diferentes remates y la posición en pista.

Remate según la RAE significa: “en el fútbol y otros deportes, acción y efecto de rematar”, rematar: “en el fútbol y otros deportes, dar término a una serie de jugadas lanzando el balón hacia la meta contraria”. De la definición de la RAE se concluye que el remate es un golpeo de ataque, que se utiliza para finalizar la jugada y con la intención de conseguir la meta.

Para definir el remate en pádel se acude a un manual específico de este deporte, “El remate se considera como el golpe de definición por antonomasia, el cual ejecutaremos para definir un punto, ya sea sacando la bola de la pista o golpeando para traernos la bola a nuestro campo sin que los contrarios puedan devolverla.” “…impacto en el punto más alto posible, lo que hará que el golpe adquiera un ángulo lo suficientemente amplio como para que, después de golpear en la pared de fondo, suba hacia arriba y evitemos la devolución del contrario.” (Moyano Vázquez, 2016).

Uniendo ambas definiciones, el remate es un golpe de ataque que se realiza por encima de la cabeza, con una trayectoria descendente y con la intención de finalizar el punto, ya sea trayendo la bola hacia nuestro campo o sacándola de la pista por encima de las paredes.

Una vez definido el remate, se explica las diferentes categorías que se le ha dado para la creación de las variables.

- Remate ganado: es aquel que cumple su objetivo sin la necesidad de realizar más golpeos.
- Remate fallado: es aquel que tras realizar la ejecución provoca la pérdida del punto sin que el rival tenga que intervenir.
- Remate continuado: es aquel que aunque la ejecución es correcta, el rival consigue mantener la bola en juego.

El pádel es un deporte de pareja por lo tanto distinguimos dos posiciones. Estas las vamos a definir en función de la mano dominante del jugador, si es diestro o zurdo, y el lado desde donde resta. Las dos posiciones son: Drive, son los jugadores que restan desde el mismo lado de su mano dominante, es decir un diestro que resta desde el lado derecho de su pista o un zurdo que resta desde el lado izquierdo, y Revés, son los jugadores que restan desde el lado contrario a su mano dominante, es decir un zurdo que resta desde el lado derecho y de la pista y viceversa.

Tras la recogida de datos se realizó tres análisis de variables cuantitativas, buscando el modelo lineal que las relacionasen. Siendo la variable predictora la altura y las variables criterio los diferentes tipos de remates. Este análisis se realizó con el conjunto global de los datos y luego se realizó un segundo análisis distinguiendo por la variable sexo. También se realizaron tres análisis de tres variables categóricas buscando la existencia de independencia entre ellas. Las variables son sexo, lateralidad y posición en la pista.

Para la manipulación y cálculos de los resultados se ha utilizado el software libre R. A continuación se detallan los comandos utilizados para la creación y los gráficos y los cálculos de los modelos.

- pairs(), se utiliza para crear la matriz de correlación, ver figura 7.
- plot(), se utiliza para crear diagramas de dispersión, ver figura 8.
    - xlim, ylim, para establecer la escala de los ejes.
    - xlab, ylab y main, para crear la leyenda de los ejes y el título del gráfico.
        - .col para establecer el color de las letras.
        - .font para elegir el tipo de fuente.
    - abline(lm()), para crear la recta de regresión en el gráfico, ver figura 11.
- summary(lm()), para ver todos los datos de la regresión.
- fisher.test(), para aplicar el test exacto de Fisher.

3.2. ANÁLISIS ESTADÍSTICO

Para realizar este estudio se han analizado los datos de un total de 34 jugadores, de los cuales 18 eran hombres y 16 mujeres. En el grupo había 30 jugadores diestros, 16 hombres y 14 mujeres, y 4 jugadores zurdos, 2 en cada grupo. En la siguiente tabla se muestra un estudio descriptivo de las variables que se van a analizar posteriormente:

GRUPO	*Altura*	*Remates ganados por partido*	*Remates fallados por partido*	*Remates continuados por partido*
GLOBAL	Media: 175,2 ±7 cm. Q_2: 170 cm.	Media: 7,74 ±4,86 Q_2: 6,75	Media: 0,67 ±0,65 Q_2: 0,54	Media: 5,52 ±3,42 Q_2: 5,07
HOMBRES	Media: 178,56 ±5,99 cm. Q_2: 180 cm.	Media: 10,02 ±4,94 Q_2: 8,95	Media: 0,69 ±0,83 Q_2: 0,5	Media: 6,24 ±4,12 Q_2: 5,58
MUJERES	Media: 169,31 ±453 cm. Q_2: 170,5 cm.	Media: 5,17 ±3,31 Q_2: 4,17	Media: 0,65 ±0,41 Q_2: 0,67	Media: 4,71 ±2,7 Q_2: 4

Tabla 14: Análisis descriptivo del conjunto de datos para las variables: altura y remates ganados, fallados y continuados por partido.

3.2.2. ANÁLISIS DEL REMATE EN FUNCION DE LA ALTURA

En primer lugar se hallará el coeficiente de correlación de Pearson entre la altura y los tres tipos de remates, ganados, fallados y continuados. Para posteriormente realizar el *t-test* para el test de hipótesis para la correlación. Las hipótesis son:

- H_0: $P = 0$ No hay relación lineal entre la altura y el remate (ganado, fallado y continuado)
- H_α: $P \neq 0$ Sí hay relación lineal entre la altura y el remate (ganado, fallado y continuado)

Y en segundo lugar se calcula el modelo lineal que mejor se ajuste a los datos utilizando el criterio de mínimos cuadrados. Luego se le aplica el *t-test* para testear la hipótesis para el punto en el origen y la pendiente de la regresión lineal. Las hipótesis para los tres casos son:

- H_0: No hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es cero, b=0.
- H_α: Sí hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es distinta de cero, $b \neq 0$.
- H_0: No hay relación lineal entre ambas variables por lo que el punto en el origen del modelo lineal es cero, a=0.
- H_α: Sí hay relación lineal entre ambas variables por lo que el punto en el origen del modelo lineal es distinta de cero, $a \neq 0$.

3.2.3. ANÁLISIS DE LAS VARIABLES SEXO, LATERALIDAD Y POSICIÓN

Para el análisis de estas variables se realizó el test exacto de Fisher 2x2 para cada par de variables, ya que cada una de las variables solo cuenta con 2 categorías. Las hipótesis son:

- H_0: Las variables sexo y lateralidad son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.
- H_α: Las variables sexo y lateralidad son dependientes, una variables varía entre los distintos niveles de la otra variable.
- H_0: Las variables sexo y posición son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.
- H_α: Las variables sexo y posición son dependientes, una variables varía entre los distintos niveles de la otra variable.
- H_0: Las variables posición y lateralidad son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.
- H_α: Las variables posición y lateralidad son dependientes, una variables varía entre los distintos niveles de la otra variable.

3.3. RESULTADOS

3.3.1. MODELO LINEAL DE REMATES EN FUNCION DE ALTURA

En la figura 6 podemos observar la matriz de dispersión de las variables altura, remates ganados, remates fallados y remates continuados del conjunto total de los datos.

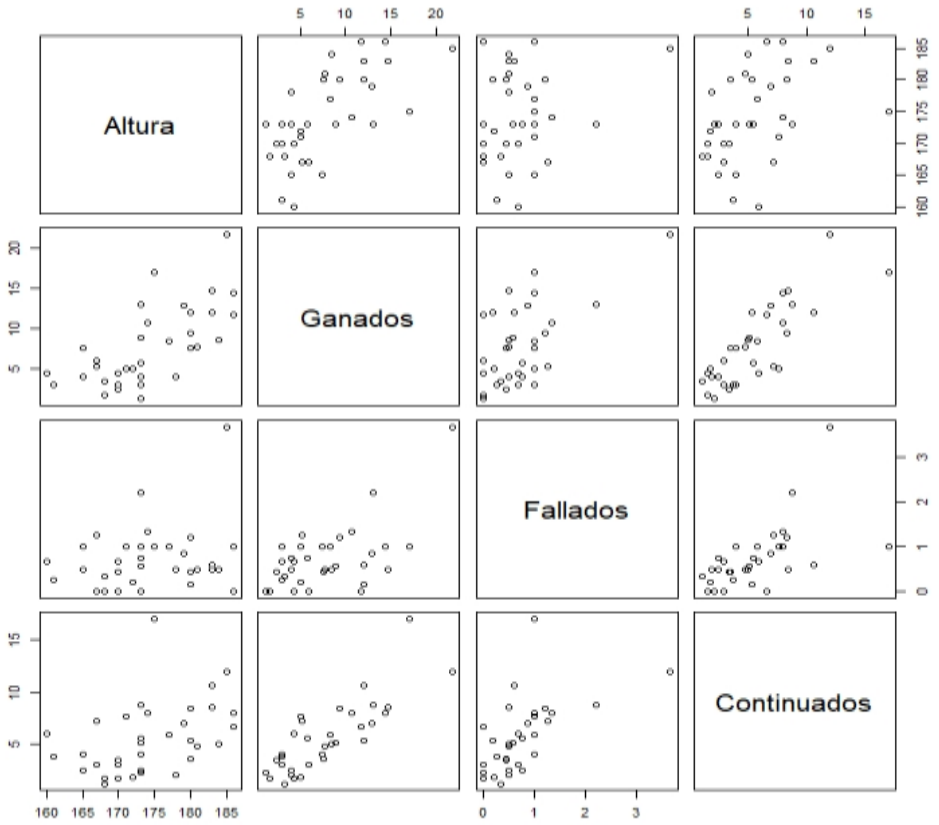

Figura 7: Matriz de dispersión. Relación entre la altura de los jugadores/as de pádel y el promedio de remates por partido.

A continuación se muestra los diferentes diagramas de dispersión que relaciona la altura de los jugadores/as de pádel con los remates ganados, tanto para el conjunto global de los datos, como el grupo dividido por el sexo.

Como se puede observar para la altura y los remates ganados mantienen una relación lineal directa, tanto para el conjunto de global, ver figura 8, como para mujeres, ver figura 9, y hombres, ver figura 10. Los coeficientes de correlación son, 0.672 para el global de datos, 0.625 para los hombres y 0.289 para las mujeres, ver tabla 14.

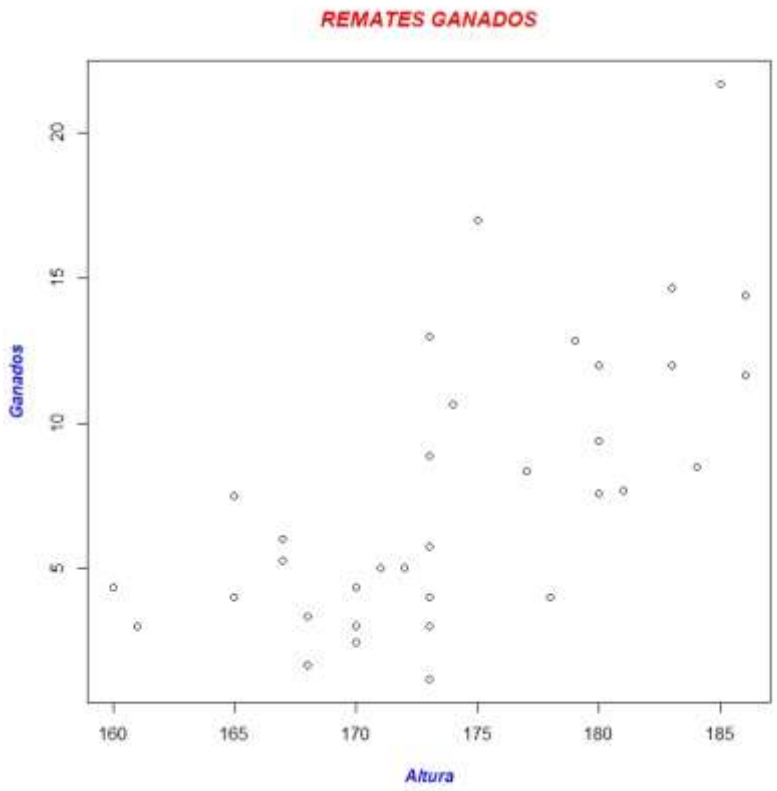

Figura 8: Diagrama de dispersión entre la altura y los remates ganados del conjunto global.

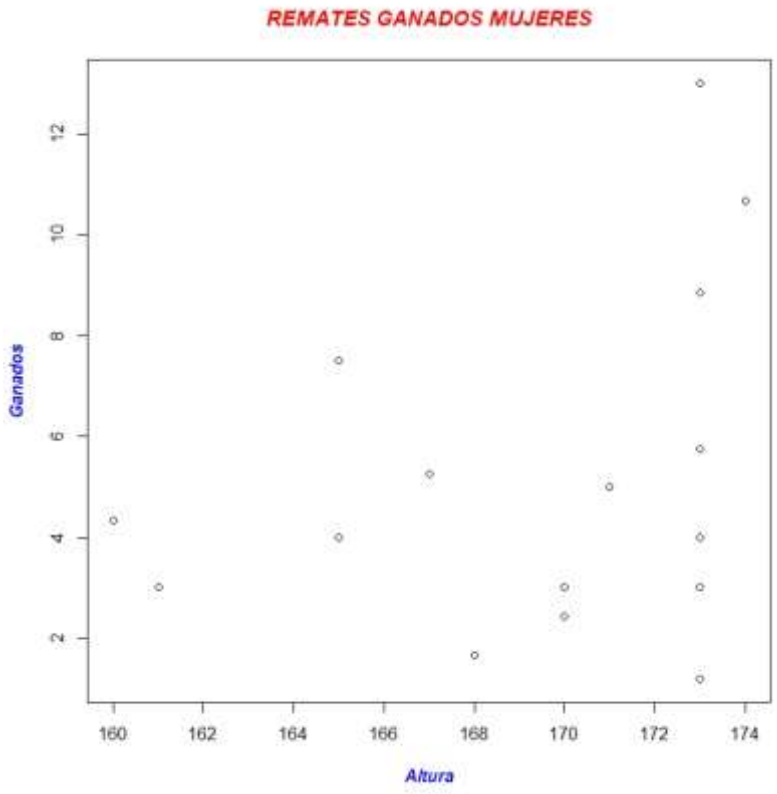

Figura 9: Diagrama de dispersión entre la altura y los remates ganados del grupo de hombres.

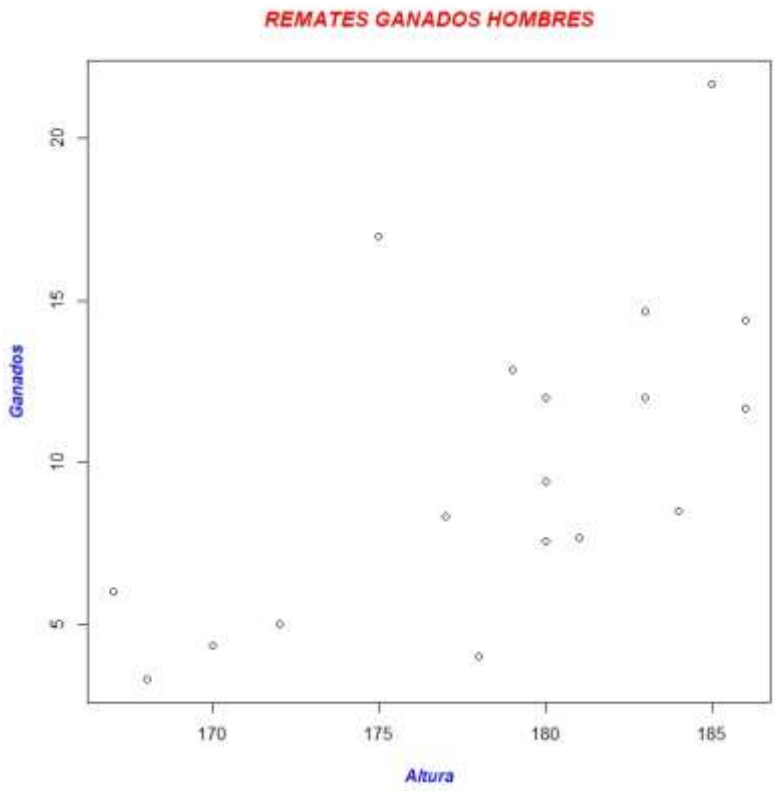

Figura 10: Diagrama de dispersión entre la altura y los remates ganados del grupo de mujeres.

A continuación se va a representar que modelo de regresión lineal es el que mejor representa la relación entre los tres casos y cuál es su coeficiente de determinación.

El modelo que representa la realción entre la altura de los jugadores y los remates ganados del global del grupo viene expresada por la siguiente ecuación, ver figura 11:

$$y' = -72{,}9147 + 0{,}463x$$

Siendo *y'* los remates ganados por partido que se estima a un jugador/a de *x* altura en cm. Seguidamente se calcula el mismo modelo pero para los grupos de hombres y mujeres por separado, ver figuras 12 y 13 respectivamente. Quedaría representado:

$$y'_{Hombres} = -82{,}0516 + 0.5157x$$

$$y'_{Mujeres} = -30{,}6283 + 0.2114x$$

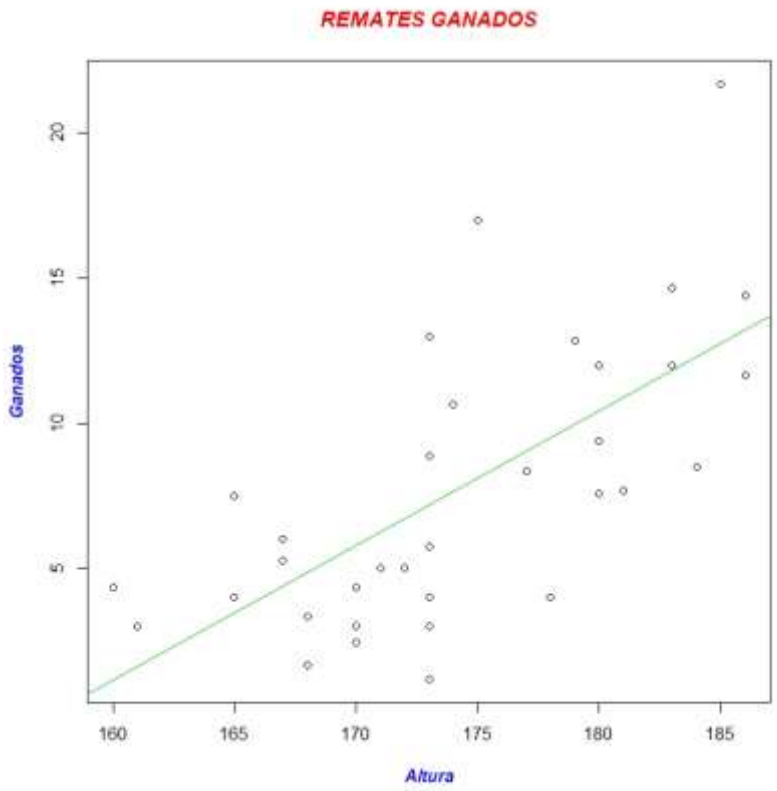

Figura 11: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates ganados del conjunto global.

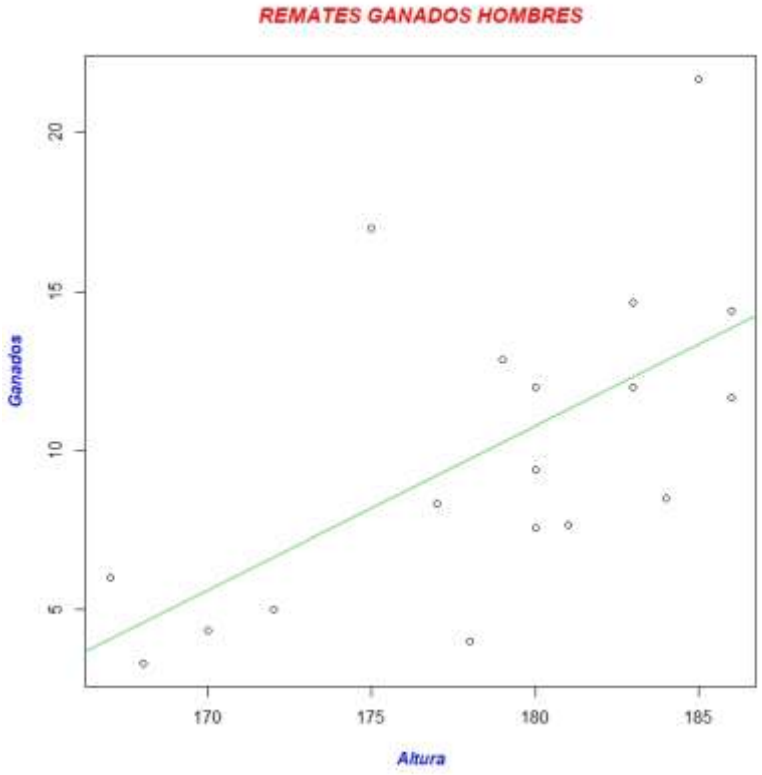

Figura 12: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates ganados del grupo de los hombres.

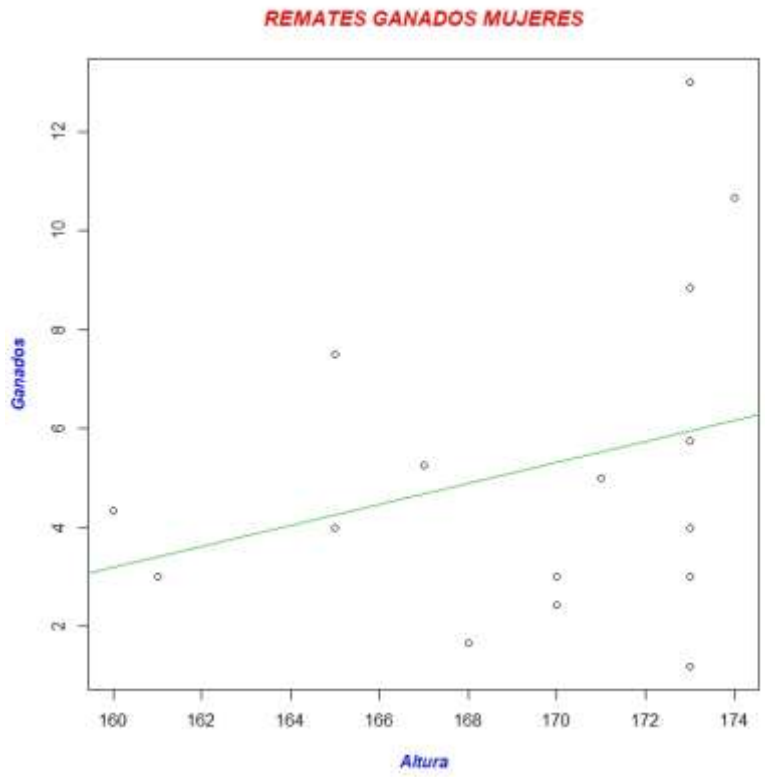

Figura 13: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates ganados del grupo de las mujeres.

Los coeficientes de determinación para estos tres modelos son, 0,4515 para el conjunto de datos completo, lo que indica que el modelo explica el 45,15% de *y* en función de *x*, 0,3907 para el grupo de hombres, quedaría así explicada el 39,07% de *y* en función de *x*, y 0,08377 para el grupo de las mujeres, explicando el 8,37% de *y* en función de *x*, ver tabla 14. Los *p-value* para estos tres modelos son:

GRUPO	***REMATES GANADOS***				
	CORRE-LACION	*p-value*	MODELO	*p-value*	COEFICIENTE DETERMINACION
GLOBAL	0,672	P<0.05	$y' = -72{,}9147 + 0{,}463x$	p<0.05	0,4515 (45,15%)
HOMBRES	0,625	P<0.05	$y' = -82{,}0516 + 0.5157x$	P<0.05	0,3907 (39,07%)
MUJERES	0,289	p>0.05	$y' = -30{,}6283 + 0.2114x$	p>0.05	0,0837 (8,37%)

Tabla 15: Tabla de coeficientes de correlación de Pearson, los modelos de regresión lineales y coeficientes de determinación que explican los remates ganados por partido en función de la altura. Para el grupo global, los hombres y las mujeres.

La siguiente tabla muestra un resumen con los datos para las siguientes variables, que son los remates fallados y continuados por partido de los jugadores y jugadoras de pádel.

GRUPO	*REMATES FALLADOS*				
	CORRE-LACION	*p-value*	MODELO	*p-value*	COEFICIENTE DETERMINACION
GLOBAL	0,2253	p>0.05	$y' = -3{,}175 + 0.02241x$	p>0.05	0,0508 (5,08%)
HOMBRES	0,386	p>0.05	$y' = -8{,}8622 + 0.0535x$	p>0.05	0,149 (14,9%)
MUJERES	0,2831	p>0.05	$y' = -5 + 0.0341x$	p>0.05	0,0802 (8,02%)

Tabla 16: Tabla de coeficientes de correlación de Pearson, los modelos de regresión lineales y coeficientes de determinación que explican los remates fallados por partido en función de la altura. Para el grupo global, los hombres y las mujeres.

GRUPO	*REMATES CONTINUADOS*				
	CORRE-LACION	*p-value*	MODELO	*p-value*	COEFICIENTE DETERMINACION
GLOBAL	0,4268	p<0.05	$y' = -30{,}48 + 0.207x$	p<0.05	0,1822(18,22%)
HOMBRES	0,4647	p<0.05	$y' = -50{,}76 + 0.319x$	p<0.05	0,216 (21,6%)
MUJERES	0,1599	p>0.05	$y' = -8{,}836 + 0.08x$	p>0.05	0,0256 (2,56%)

Tabla 17: Tabla de coeficientes de correlación de Pearson, los modelos de regresión lineales y coeficientes de determinación que explican los remates continuados por partido en función de la altura. Para el grupo global, los hombres y las mujeres.

Los siguientes diagramas representan los modelos anteriormente expuestos.

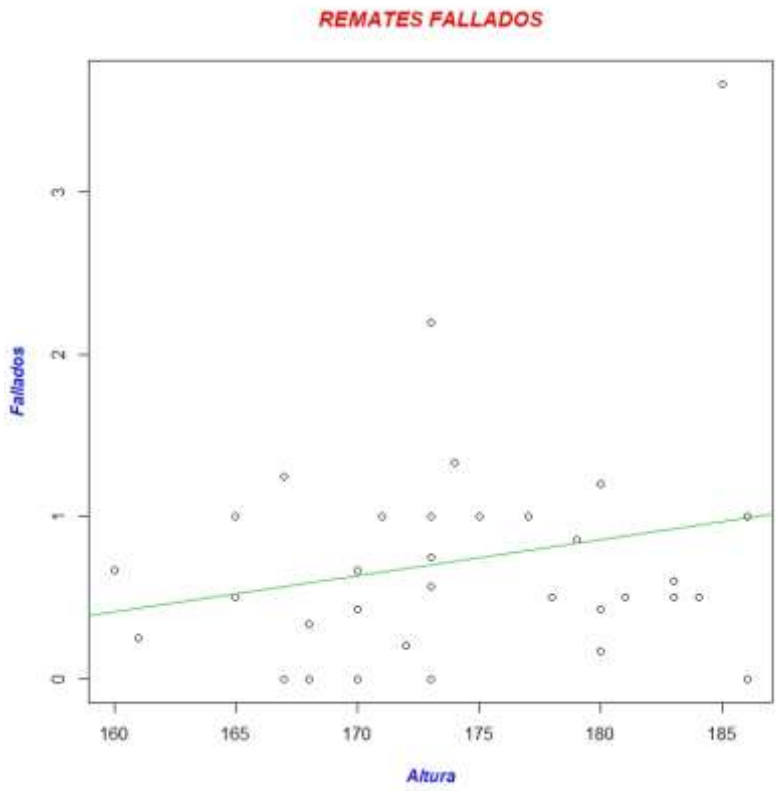

Figura 14: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates fallados del conjunto global.

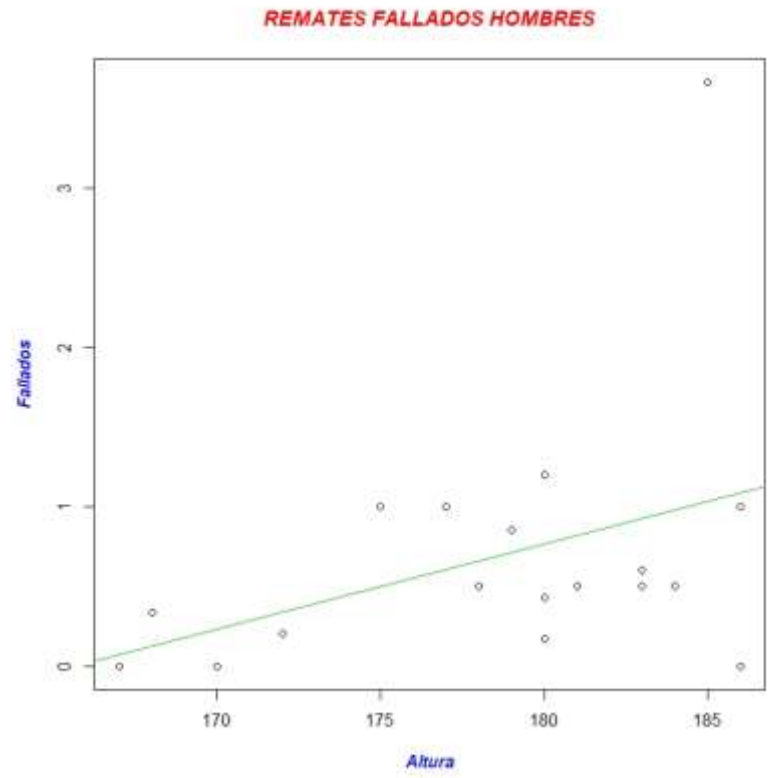

Figura 15: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates fallados del grupo de los hombres.

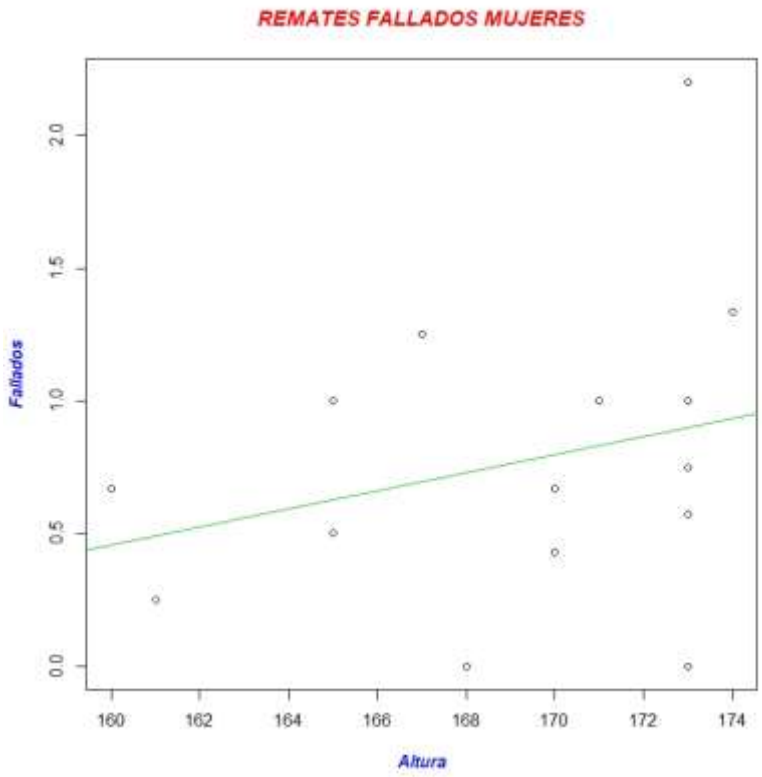

Figura 16: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates fallados del grupo de las mujeres.

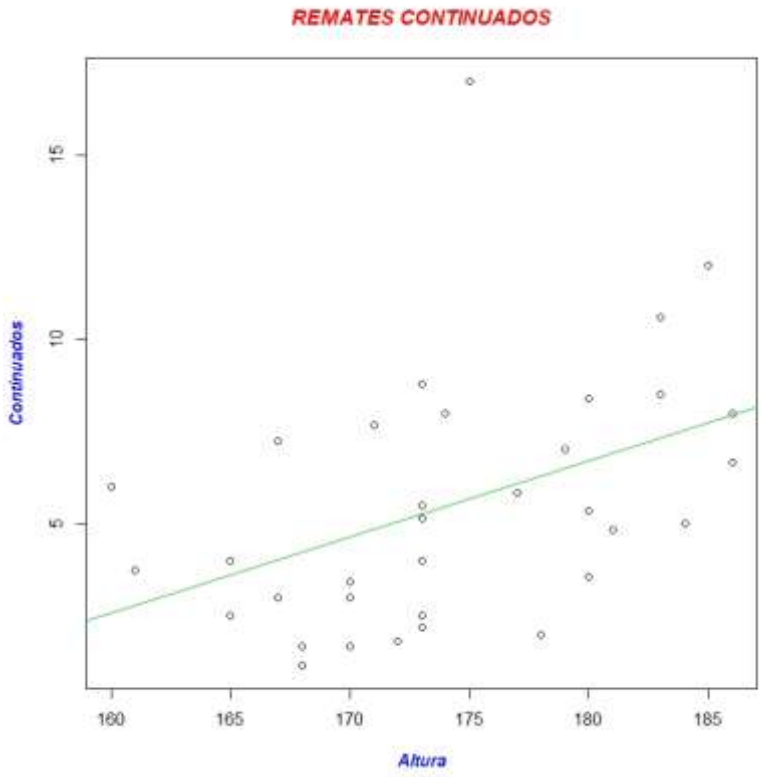

Figura 17: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates continuados del conjunto global.

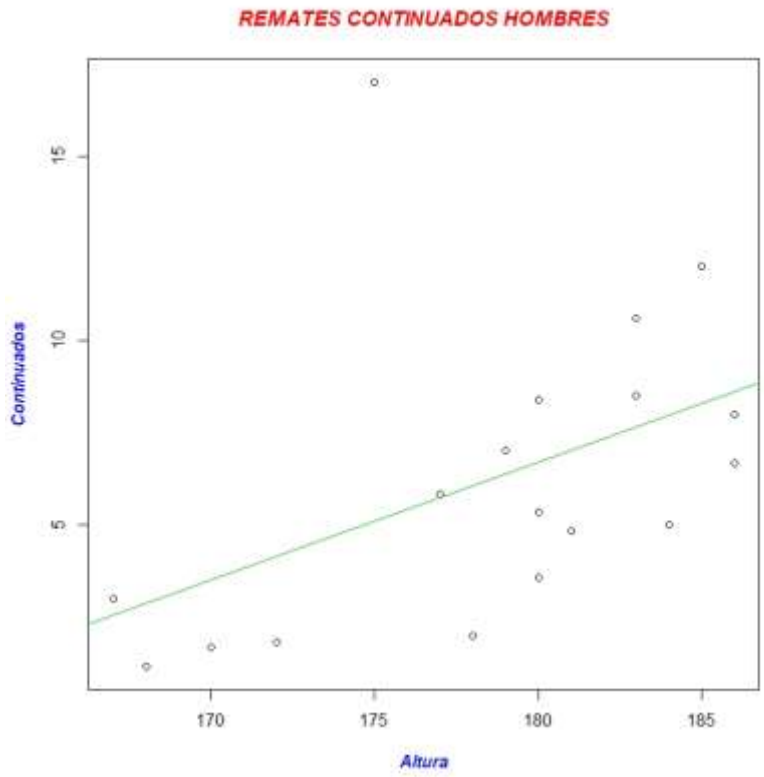

Figura 18: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates continuados del grupo de los hombres.

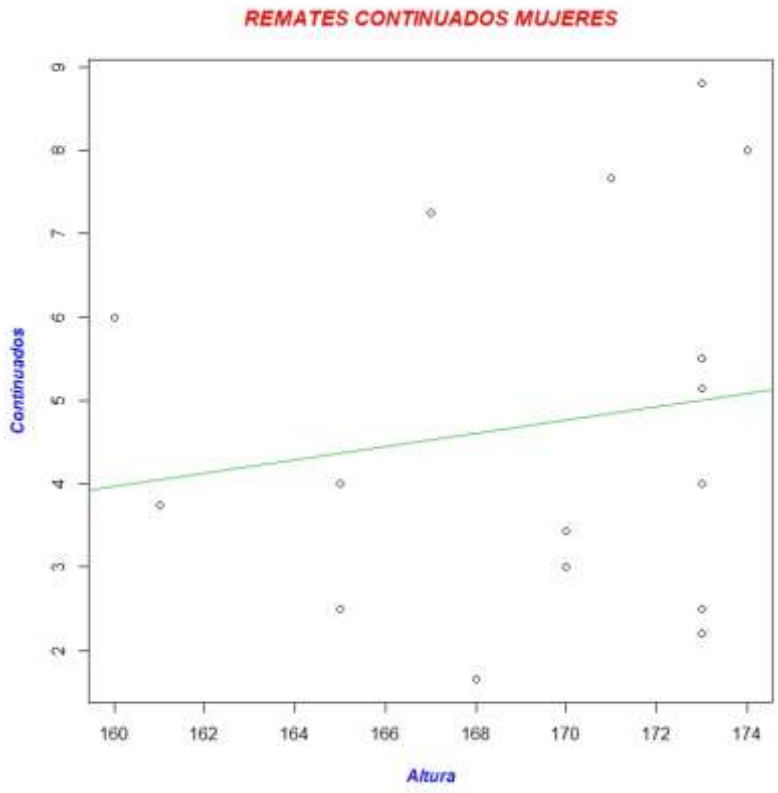

Figura 19: Diagrama de dispersión y representación del modelo lineal entre la altura y los remates continuados del grupo de las mujeres.

Como se observan en las figuras 11, 14 y 17, existen varios outliers que pueden afectar a los valores de los distintos modelos. Se ha realizado un segundo análisis omitiendo dichos valores y se han comparado los modelos. No se han hallado modificaciones en los modelos que alteren los *p-value* de cada uno de los estudios.

3.3.2. RELACIÓN ENTRE SEXO, LATERALIDAD Y POSICIÓN

A continuación se muestra una tabla de contingencia entre la variables sexo y lateralidad, cuyo *p-value* = 1

	DIESTRO/A	*ZURDO/A*	
HOMBRE	16	2	18
MUJER	14	2	16
	30	4	34

Tabla 18: Tabla de contingencia para las variables sexo y mano dominante.

La siguiente tabla de contingencia pertenece a las variables sexo y posición donde juega, cuyo *p-value* = 1

	DRIVE	*REVÉS*	
HOMBRE	7	11	18
MUJER	6	10	16
	13	21	34

Tabla 19: Tabla de contingencia para las variables sexo y posición que ocupa en la pista.

La siguiente tabla de contingencia pertenece a las variables lateralidad y posición donde juega, cuyo *p-value* = 0.1445

	DRIVE	*REVÉS*	
DIESTRO/A	13	17	30
ZURDO/A	0	4	4
	13	21	34

Tabla 20: Tabla de contingencia para las variables mano dominante y posición que ocupa en la pista.

3.4. DISCUSIÓN

Como se observa en los resultados anteriores los modelos de regresión lineal tienen un mejor ajuste para el grupo de hombres que para el grupo de mujeres. Se puede comprobar que tanto para el grupo global como para el grupo de hombres los *p-value* de la correlación y de los modelos lineales de las variables altura y remates ganados, ver tabla 15, son inferiores a 0.05 lo que significa que tienen una dependencia significativa y podemos decir que para el grupo de hombres el promedio de remates ganados por partido aumenta 0.02241 por cada centímetro del jugador. Sin embargo, para el grupo de mujeres estos *p-value* tienen un valor superior al 0.05, por lo cual no tienen una relación lineal significativa.

En el caso de los remates fallados y la altura, ver tabla 16, los valores de *p-value* en los tres grupos son superiores a 0.05. Por lo cual, se puede decir que la altura y los remates fallados no mantienen una relación lineal significativa.

Y por último, en el caso de los remates continuados y la altura, ver tabla 17, es similar a lo que ocurre con los remates ganados. Pero los valores de *p-value* de la correlación y del modelo lineal para los hombres son de 0.052 para ambas, es superior de 0.05 pero no se aleja mucho por lo que podemos rechazar nuestra hipótesis nula. Entonces, se observa que hay una relación lineal significativa entre los remates continuados y la altura para el global del grupo y para el grupo de hombres, pero no ocurre lo mismo para el grupo de mujeres.

En el caso de las variables categóricas, se observa que los valores de *p-value* son superiores a 0.05 en todos los casos, lo que indica que las variables son independientes entre sí. Llama la atención los valores entre el sexo y la lateralidad, ver tabla 18, y entre el sexo y la posición donde juega, ver tabla 19, ya que el valor es 1. Esto indica que existe una independencia absoluta entre estas variables.

CAPÍTULO 4

CONCLUSIONES Y LÍNEAS FUTURAS

Para concluir, este estudio puede servir de guía para estimar el rendimiento de nuestro jugador, en caso de que sea hombre, como se puede ver en los resultados en el caso de las mujeres el *p-value* del modelo lineal no da confianza para rechazar la hipótesis nula. En caso de que sea hombre se puede analizar los partidos del jugador y estudiar si su promedio de remates está o no al nivel de los mejores jugadores.

También se puede leer que las correlaciones obtenidas en las variables cuantitativas a pesar de ser positivas son bajas, se debe de realizar un diagnosis del estudio, pero esto queda fuera del objetivo del trabajo.

Por otro lado, se puede realizar un estudio similar pero analizando distintos golpeos de pádel como puede ser la bajada de pared, la bandeja, los golpeos defensivos. Para conocer el comportamiento de los mismo en función de las características de los jugadores y así poder trazar una estrategia de juego en función de nuestros jugadores y los rivales.

Y por último, en el caso de las variables categóricas, más concretamente en el par de lateralidad y posición donde juega, se observa como el p-value desciende. En esta muestra hay pocos casos de jugadores/as zurdos/as y una posible mejora del estudio puede ser aumentar la muestra recogiendo datos de la temporada completa o incluso de varias temporadas para estudiar el compartimiento entre estas dos variables.

Bibliografía

Amat Rodrigo, J. (Junio de 2016). *Correlación Lineal y Regreseión Lineal Simple*. Recuperado el 20 de Marzo de 2020, de RPubs: https://rpubs.com/Joaquin_AR/223351

Amat Rodrigo, J. (Enero de 2016). *Test estadísticos para variables cualitativas. Test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-cochran*. Recuperado el 15 de Abril de 2020, de Rpubs: https://rpubs.com/Joaquin_AR/220579

Barriopedro, M. I. (2012). *Análisis de datos en las ciencias de la actividad física y del deporte.* Madrid: Pirámide.

Castillo-Lozano, R., & Alvero-Cruz, J. R. (2016). Estudio epidemiológico de las principales lesiones músculo-esqueléticas en jugadores de pádel. (Wanceulen, Ed.) *Innovación e incestigación en pádel*, 21-38.

Fernández del Viso, D. S. (Noviembre de 2018). *Correlación*. Recuperado el 8 de Mayo de 2020, de RPubs: https://rpubs.com/dsfernandez/442629

Francisco, P. (2019). *Estadística y Machine Learning con R.* Recuperado el 22 de Febrero de 2020, de https://bookdown.org/content/2274/portada.html

Garavito, D. (21 de Octubre de 2018). *Introducción análisis datos Categóricos, bondad de ajuste y pruebas de independencia para dos variables categóricas.* Recuperado el 8 de Mayo de 2020, de RPubs: https://rpubs.com/bogotan/CategoDepend

García Navarro, J., López Martínez, J. J., De Prado Campos, F., & Sánchez Alcaraz Martínez, B. J. (2016). Lesiones músculo-tendinosas más frecuentes de miembro inferior en el pádel. *Innovación e incestigación en pádel*, págs. 63-75.

Gil Martínez, C. (Mayo de 2018). *Regresión lineal simple*. Recuperado el 12 de Maayo de 2020, de RPubs: https://rpubs.com/Cristina_Gil/Regresion_Lineal_Simple

Herrera, J., & Courdel.Ibáez, J. (2017). Valoración del estado de condición física en jugadores de pádel. Propuesta de batería de test basados en las demandas del deporte. *Revista andaluza de medicina del deporte, 10*(3), págs. 166-1667.

Martínez Maqueda, J. M., & Sarabia Cachadiña, E. (2018). *Lesiones más frecuentes en pádel y su readaptación.* Fundación Universitaria San Pablo CEU.

Melledo-Arbelo, O., Baige Vidal, E., & Vivés Usón, M. (2019). Análisis de las acciones de juego en pádel masculino profesional. *revista de ciencias de la actividad física y del deporte de la Universidad Católica de San Antonio,, 14*(42), págs. 191-201.

Millán Díaz, I. (2017). *Tablas de contingencias.* Recuperado el 20 de Abril de 2020, de https://idus.us.es/handle/11441/66971

Moyano Vázquez, J. (2016). *Pádel: sus golpes, entrenamiento y mas.* Sevilla: Wanceulen.

Parrón Sevilla, E., Nestares Pleguezuelo, T., & De Teresa Galván, C. (Diciembre de 2015). Valoración de los hábitos de vida saludables en jugadores de pádel. (Wanceulen, Ed.) *Innovación e investigación en pádel*, 13-20. Recuperado el 30 de Mayo de 2020

Pradas, F., Castellar, C., Blas, J., Garcias-Castañón, S., Otín, D., Llimiñana, C., & Puzo, J. (2015). Variaciones séricas de magnitudes bioquímicas en jugadores de pádel de alto nivel. *Innovación e investigación en pádel*, págs. 97-110. Recuperado el 30 de Mayo de 2020

Pradas, F., Castellar, C., Quintas, A., & Arracó, S. (2016). Análisis de la condición física de jugadores de pádel de elite. *Innovación e investigación en pádel*, págs. 79-96. Recuperado el 30 de Mayo de 2020

Sarrión Gavilán, M. D., Benítez Márquez, M. D., Iranzo Acosta, J. L., & Isla Castillo, F. (2012). *Estadísitca descriptiva.* Madrid: McGraw-Hill.