Universidad Católica de Valencia San Vicente Mártir

TRABAJO FINAL DE GRADO PARA OPTAR AL TÍTULO DE

"GRADO DE MEDICINA"

TÍTULO:

UTILIDAD DE LA REGRESIÓN LINEAL MÚLTIPLE EN ESTUDIOS DE CIENCIAS DE LA SALUD

Presentado por:

CARMEN ZAPATA CARRATALÁ

Tutor:

FRANCISCO JAVIER ARTEAGA MORENO

AGRADECIMIENTOS

En primer lugar, mi más profundo agradecimiento al Dr. Francisco Javier Arteaga Moreno, por su guía, consejos y dedicación, sin los cuales este trabajo no habría sido posible.

A la Universidad Católica de Valencia, por darme la oportunidad de estudiar medicina y realizar las estancias que han permitido formarme no solo en lo académico, sino también en tantos otros aspectos.

A mis compañeros de carrera, porque su compañía durante estos años ha sido siempre un impulso de motivación.

Y a mi familia, por su apoyo incondicional en cada momento, sois la base sólida que me da seguridad para avanzar.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

I. RESUMEN

Introducción: A lo largo de la historia se ha ido desarrollando el conocimiento científico paralelamente a la evolución cultural de las sociedades. Esta información ha ido creciendo con el paso de los siglos de manera que, especialmente en la Era Digital, se ha hecho inabarcable. Para garantizar una correcta atención asistencial surge la medicina basada en la evidencia. Paralelamente, y con el estudio del genoma surge otra tendencia, la medicina personalizada. Para poder relacionar conceptos y sacar conclusiones nos ayudamos de los modelos de predicción clínica, en concreto nos centraremos en el Modelo de Regresión Lineal.

Material y métodos: La Regresión Lineal es una herramienta estadística que nos indica la relación entre variables. Para hacer una correcta aplicación del modelo deberos seguir un orden metódico y dejar constancia de ello.

Resultados: En las publicaciones científicas se suelen reportar los resultados del análisis estadístico en tablas de resultados, pero en raras ocasiones se define paso a paso la aplicación del modelo.

Conclusiones: El Modelo de Regresión Lineal es un modelo valioso para aplicar en el campo de las ciencias de la salud por su capacidad de predecir, interpretar y explicar los fenómenos que nos interesen estudiar de manera cuantitativa. Es importante que los métodos estadísticos sean adecuados a los objetivos de investigación que nos planteamos. La estadística permite estudiar relaciones complejas entre variables en un ambiente de incertidumbre.

Palabras clave: Regresión Lineal Múltiple, Análisis de Regresión, modelos estadísticos.

ABSTRACT

Introduction: Throughout history, scientific knowledge has been developed in parallel to the cultural development of societies. Over the centuries, this information has grown in an immesurable way, especially during the Digital Age. Evidence-based medicine emerges to guarantee that this information is dealt with care. Following from the advent of genomics, personalized medicine emerges as yet another trend. In order to relate concepts and draw conclusions, we use clinical prediction models, specifically we will focus on the Linear Regression Model.

Material and methods: Linear Regression is a statistical tool that indicates the relationship between variables. To apply the model correctly, we must follow a methodical order and record it.

Results: Scientific publications usually report the results of statistical analysis in tables of results, but seldom how the application of the model is defined step by step.

Conclusion: The Linear Regression Model is a valuable model to apply in the field of health sciences due to its capability to predict, to interpret and to explain the phenomena that we are interested in studying in a quantitative way. Statistics allow to study complex relationship between variables in an uncertainty atmosphere.

Key words: Multiple Linear Regression, Regression Analysis, statistical models.

II. INTRODUCCIÓN

1. CONTEXTO HISTÓRICO Y MOTIVACIÓN

El desarrollo de las matemáticas ha ido ligado a los aspectos culturales e intelectuales de la sociedad a lo largo de la historia, incluido el conocimiento médico. Podríamos decir que hace 44.000 años la especie humana ya utilizaba herramientas para predecir fenómenos biológicos, como se cree que fue el **Hueso de Lebombo**. Un peroné de babuino con 29 muescas que se piensa, entre otras teorías, que pudo haber sido utilizado como calendario para el seguimiento del ciclo menstrual (1).

Desde los tiempos antiguos se han hecho diversos censos y registros poblacionales, pero no fue hasta 1662 con las aportaciones de los métodos de **John Graunt** a la política que nace la demografía moderna, convirtiéndose así en el padre de la bioestadística (2).

También es destacada la figura de **William Farr** quien estudiaba relaciones entre enfermedades y ciertas características del entorno con el fin de hacer predicciones. Cabe resaltar su estudio junto a John Snow sobre el brote de cólera de 1849 en Londres, mediante el cual concluyeron que había mayor incidencia de cólera en las poblaciones asentadas más próximas a aguas contaminadas (3).

No solo por su calidad asistencial, sino también por su estudio sobre la mortalidad en la guerra de Crimea y la innovación que supuso utilizar diagramas para la explicación de datos estadísticos sanitarios, es reconocida la figura de **Florence Nightingale** (4).

Todos ellos junto a otros tantos han ido completando lo que hoy en día conocemos y aplicamos como Bioestadística. Actualmente la importancia que tiene esta rama del conocimiento nos ha guiado a la toma de decisiones clínicas basadas en la evidencia. La **medicina basada en la evidencia** consiste en el uso de la evidencia de la literatura médica para garantizar una atención sanitaria de calidad para el paciente. El objetivo de la unificación de los criterios asistenciales es la reducción de la variabilidad en la práctica clínica y asegurar

que los pacientes reciban la atención de manera igualitaria (5). Así, la medicina basada en la evidencia se utiliza para:

- Recopilar las últimas investigaciones sobre un área de la medicina.
- Seleccionar las investigaciones más relevantes y válidas cuyos resultados puedan suponer un cambio en la práctica clínica.
- Resumir las nuevas investigaciones sobre un tema del que ya se tiene conocimiento previo.
- Interpretar la evidencia obtenida en la práctica clínica diaria.

A pesar de todas estas aplicaciones, nos encontramos con un gran problema a la hora de abordar la búsqueda de conocimiento científico de calidad: mantenerse al día con la cantidad de información que se publica continuamente (6). En 2008 un estudio de la Universidad de Ottawa concluyó que "Medline añade alrededor de 1 millón de nuevas publicaciones cada año" y que "se estima que hasta el 7% de las conclusiones clínicas de las revisiones sistemática cambian cada año" (7). Otro estudio de 2010 afirmaba que al día se publican 75 ensayos y 11 revisiones sistemáticas, esto nos muestra la imposibilidad de mantenerse actualizado con toda la información que se publica. "No hay señales de que esto descienda, pero todavía hay 24 horas en un día" (8). Esta situación del creciente número de publicaciones ha seguido aumentando con los años y se ha acentuado notablemente con la pandemia de la COVID-19. Alrededor de 100.000 artículos sobre la pandemia del coronavirus se han publicado en 2020, un 4% de las publicaciones totales (9). Solamente en el mes de enero de 2021 ha habido 40.221 publicaciones al respecto (10).

Paralelamente, con el auge del estudio del genoma ha surgido la **medicina personalizada**, tendencia que defiende la necesidad de adaptar la atención sanitaria a las características genéticas, moleculares, psicológicas y del entorno, todas ellas únicas que cada paciente posee. Sin embargo, este modelo no rechaza al anterior y el profesional sanitario deberá encontrar un equilibrio entre la evidencia conocida y las condiciones personales del paciente para dar una atención de calidad (5).

Aquí es donde entra la importancia de los **modelos de predicción clínica**, que nos sirven para poder relacionar conceptos. Estos tienen en cuenta la población

de riesgo, el uso de datos relevantes, la delimitación temporal y los predictores, todo ello analizado por medio de un modelo matemático para llevarlo finalmente a la aplicación clínica. El modelo matemático más utilizado son las técnicas de regresión, como la regresión lineal múltiple o la regresión logística (11).

2. MODELOS DE PREDICCIÓN CLÍNICA

Los modelos de predicción clínica cada vez son más populares y se han extendido a áreas más allá de la medicina. Las principales ramas médicas en las que podemos aplicar estos modelos son: la salud pública, la práctica clínica y la investigación médica.

En la salud pública para predecir la futura aparición de enfermedades y así diseñar planes preventivos.

En la rama de la práctica clínica para:

- Decidir si son necesarias más pruebas predictivas de la enfermedad.
- Decidir si es necesario el inicio de un tratamiento, intensificar su uso, su coste-efectividad o si es más conveniente retrasar el inicio de este. Iniciaremos un tratamiento tras el diagnóstico confirmado si la probabilidad del diagnóstico es mayor que el umbral de tratamiento (la probabilidad en la que el beneficio esperado del tratamiento es igual a la de evitar el tratamiento).
- Decidir si una cirugía es necesaria, sopesando los riesgos a corto y largo plazo.

En la rama de la investigación clínica se puede utilizar para seleccionar a los participantes apropiados y el ajuste covariable en ensayos aleatorios controlados o para el ajuste de confusores y de mezcla de casos en estudios observacionales.

Para desarrollar modelos de predicción válidos se han de cumplir tres condiciones:

- Unas consideraciones generales para las que se deberá tener en cuenta la cuestión de investigación, la aplicación prevista, el resultado, los predictores, el diseño del estudio, el modelo estadístico y el tamaño muestral.
- Seguir 7 pasos para diseñar el modelo:
    1) Fase preliminar, en la que nos plantearemos el problema de predicción.

2) Codificación de predictores, considerando el tipo de variables de las que disponemos.
3) Especificación del modelo, analizando si cumple las condiciones del modelo planteado, que predictores incluiremos…
4) Estimar los parámetros del modelo.
5) Rendimiento del modelo, analizar si el modelo es útil para el problema planteado.
6) Validación del modelo, comprobar que el modelo es válido para nuestra muestra y extrapolable a la población.
7) Presentación del modelo, reportando los resultados relevantes para su interpretación.

Validación interna, estudiando si existe sobreajuste, y externa, estudiando si es viable la generalización del modelo (12).

Los **modelos estadísticos** son utilizados tanto para realizar la estimación de nuestra predicción como para probar la hipótesis del estudio. Estos sintetizan los patrones de los datos disponibles para el análisis. La elección del modelo se define principalmente a partir de la variable respuesta, además, los datos deberán cumplir ciertas condiciones para que este pueda ser aplicado. Los modelos de regresión son los más ampliamente utilizados en el área de las ciencias de la salud.

Cuando la variable respuesta sea continua, el modelo de referencia es la **regresión lineal**, este es el modelo en el que nos centraremos a lo largo de este trabajo. En el caso de que la variable respuesta continua sea sobre datos económicos, se recomienda utilizar la media como un buen descriptor de esta.

Una variación del modelo de regresión son los **modelos aditivos generalizados**, que es más flexible para variables explicativas continuas y efectos potencialmente no lineales estimados por medio de regresión polinómica.

En el caso de que la variable respuesta sea binaria, el modelo más ampliamente utilizado es la **regresión logística**. Sus variables explicativas pueden ser categóricas y continuas, además, se pueden realizar transformaciones no

lineales e interacciones entre sus elementos. La regresión logística es una importante herramienta para hacer análisis discriminantes.

Otros modelos para respuestas binarias son el **teorema de Naïve Bayes**, utilizado para predecir la probabilidad condicional de padecer una enfermedad, o los modelos de **regresión multivariantes aditivos en spline**, cuyo objetivo es encontrar una estructura aditiva de bajo orden e interacciones entre los factores de riesgo.

Para las variables respuesta categóricas se puede utilizar la **regresión logística multinomial** o aplicar consecutivamente modelos dicotómicos multivariantes de regresión logística.

En el caso de que la respuesta sea ordinal, el método más utilizado es el modelo logístico de **odds ratio**, que es una extensión del modelo de regresión logística.

Cuando el elemento de estudio es la supervivencia, la variable respuesta suele ser la muerte, el modelo más empleado es la **regresión de Cox**, también es una extensión de la regresión logística y nos indica el riesgo de que ocurra el evento que estudiamos, normalmente la muerte, durante el seguimiento realizado en el estudio. Otro modelo que también se utiliza en el estudio de la supervivencia es el modelo de **Kaplan-Meier** y el modelo de **Weibull** (13).

3. ESTRUCTURA DEL TRABAJO

En este trabajo nos centraremos en el Modelo de Regresión Lineal Múltiple. Explicaremos la utilidad que tiene en estudios de medicina, cumpliendo los objetivos que quedan redactados a continuación en la sección IV.

En el apartado V de material y métodos, comenzaremos explicando en qué consiste el modelo de Regresión Lineal Múltiple; definiendo los pasos a seguir a la hora de aplicarlo, las variables que se pueden incluir, sus usos, la recta de regresión, la obtención del modelo de regresión lineal múltiple por el método de mínimos cuadrados, también, aplicaremos la inferencia y expondremos los métodos para decidir qué variables añadir o no en el modelo definitivo.

A continuación, estudiaremos la validación del modelo por medio del análisis de las condiciones que debe cumplir, los problemas con la hipótesis del modelo, el efecto de cada variable por su significación e interpretación, y el reporte de los resultados.

Seguidamente, pondremos en práctica el modelo con un ejemplo en el que describiremos con detalle cada paso que se ha de seguir; desde la elección de la población y la muestra, pasando por el estudio de cada variable, hasta el estudio de los residuos. Veremos también las adaptaciones que puede sufrir el modelo cuando prescindimos de algunas medidas que contemplábamos inicialmente.

En los resultados expondremos el análisis de dos artículos que utilizan la regresión lineal para el análisis de sus resultados, veremos si han seguido todos los pasos y si reportan toda la información necesaria.

Finalmente, dispondremos de las conclusiones del trabajo, seguido de la bibliografía utilizada.

En los anexos se podrá consultar la base de datos utilizados en el diseño del modelo.

III. OBJETIVOS

Cabe subrayar que este trabajo no se hace con el objetivo de investigar, sino con el fin de ilustrar la aplicación del Modelo de Regresión Lineal Múltiple.

El principal objetivo de este trabajo es subrayar la importancia de la Regresión Lineal Múltiple en estudios de ciencias de la salud.

Esto lo haremos por medio de los siguientes objetivos específicos:

- Explicar el Modelo de Regresión Lineal y valorar su utilidad en ciencias de la salud.
- Aplicar el Modelo de Regresión en un ejemplo concreto.
- Analizar el uso de esta herramienta en dos ejemplos sacados de la literatura médica.

IV. MATERIAL Y MÉTODOS

1. EL MODELO DE REGRESIÓN LINEAL

El modelo de Regresión Lineal Simple es una herramienta estadística que sirve para estudiar la relación entre una variable respuesta y una variable explicativa. Pero suele ocurrir que más de una variable explicativa está relacionada con la variable respuesta, es entonces cuando aplicamos el modelo de Regresión Lineal Múltiple.

1.1. Pasos en un estudio de Regresión

Los pasos para seguir para un estudio de regresión están ilustrados en la figura 1 y son los siguientes:

1. **Definición del problema:** Lo primero que debemos hacer es identificar la finalidad y el contexto del estudio.
2. **Formulación del modelo:** Seleccionamos la variable respuesta y las variables explicativas. Escogemos la forma analítica que mejor se ajuste a nuestros objetivos, definimos cuáles van a ser los parámetros y nos planteamos la necesidad de agregar transformaciones o interacciones.
3. **Recogida de datos:** Deberemos dejar constancia de donde proceden los datos que vamos a utilizar. Pueden proceder de la recopilación directa de datos sacados de la historia clínica. Se puede realizar también un muestreo para obtener información sobre una población. Otra manera de obtener los datos sería de forma experimental por medio de un ensayo clínico.
4. **Estimación del modelo:** Para realizar la estimación del modelo deberemos utilizar un paquete estadístico, en este trabajo el programa que hemos utilizado para analizar los datos es *f*-Sats. En esta fase obtendremos los valores de los parámetros del modelo más verosímiles posibles a partir de los datos. Además, se evaluará la precisión de las estimaciones obtenidas.
5. **Validación del modelo:** Comprobaremos que el modelo cumple las condiciones para que sea válido. Aquí analizaremos si los datos siguen una distribución normal, la existencia de datos anómalos que distorsionen los resultados del modelo o si hubiera sido más adecuado aplicar otro

modelo. Se deberá comprobar si el modelo representa los aspectos relevantes de la realidad estudiada mediante la abstracción y simplificación.

6. **Explotación del modelo:** Nos permitirá aplicar el modelo para sacar conclusiones y tomar decisiones (14).

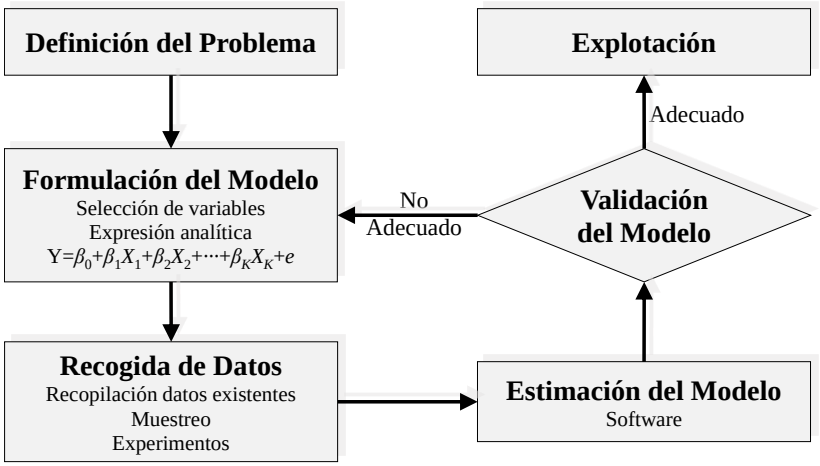


Figura 1. Pasos para seguir en el estudio de Regresión (14).

1.2. Variables

La **variable respuesta**, es la variable que queremos predecir. También es conocida como variable dependiente, *outcome*, y se representará en el eje de ordenadas (*y*). Para el modelo de regresión lineal deberá ser una variable continua.

Para estimarla nos basamos en el resto de variables recogidas, las **variables explicativas**, independientes o *inputs*. Estas variables podrán ser continuas, binarias o categóricas.

Sin embargo, para poder añadir una variable categórica con k niveles tenemos que codificarla, empleando $k-1$ variables *dummies*. Las categorías de las variables categóricas deberán ser excluyentes entre sí. Si, por ejemplo, queremos estudiar el continente de procedencia entre Europa, Asia y África de una serie de individuos, un individuo procedente de África, no podrá hacerlo de Europa ni de Asia al mismo tiempo. En este caso tenemos 3 niveles ("Europa", "África" y "Asia") por lo que necesitamos 2 variables *dummies*:

Tabla 1. Construcción de una variable dummy.

	Variables indicadoras		
	África	Europa	Procedencia
1	1	0	África
2	0	1	Europa
3	0	0	Asia

La categoría omitida al crear las *dummies*, "Asia" en este caso, actúa como base. De esta manera, estimaremos el efecto de cada una de las categorías con *dummy* asignada, en relación a la base elegida.

Esto también lo aplicaremos a las variables binarias, como el sexo, si un individuo es mujer se excluye de ser hombre. De este modo, en vez de tener dos variables sexo, usaremos, por ejemplo, la variable "hombre", así, a los individuos masculinos le asignaremos el valor 1 y a los femeninos 0 (15).

1.3. Usos

La RLM se utiliza para:

- **Predecir:** Para una combinación específica de las variables explicativas podemos obtener la estimación de la variable respuesta correspondiente. De esta manera respondemos a la pregunta: ¿qué valor esperamos para la variable respuesta para un individuo dado del que conocemos el valor de las variables explicativas?
- **Explicar:** El modelo explica parte de la variabilidad de la variable respuesta a partir de la variabilidad de las variables explicativas. Responde a la pregunta: ¿por qué la variable respuesta toma distintos valores para los diferentes individuos? Al responder a esta pregunta relacionamos dicha variabilidad con la de las variables explicativas.
- **Interpretar:** La ecuación del modelo describirá cuánto cambia, en promedio, la variable respuesta conforme cambian las variables explicativas. Responde a la pregunta: ¿qué efecto tiene sobre la variable respuesta un cambio en el valor de una de las variables explicativas, manteniendo el resto de las variables contantes? (16).

Para introducir algunos de los elementos básicos del Modelo de Regresión Lineal empleamos una versión simplificada del mismo: el Modelo de Regresión Lineal Simple, en el tenemos una única variable explicativa.

1.4. El Modelo de Regresión Lineal Simple: Recta de regresión

Para estudiar una posible relación entre dos variables, podemos representar lo datos en un diagrama de dispersión. Si dicha relación es lineal, es decir, si los puntos se distribuyen alrededor de una recta, podremos modelar la relación mediante la recta. Esta relación se denomina **recta de regresión** y, como modelo, nos indica, para cada valor de X el valor esperado de Y. De esta manera, el Modelo de Regresión Lineal Simple se puede expresar mediante la ecuación de una recta:

$$E[Y|X = x] = \alpha + \beta x$$

El parámetro β de la ecuación, la pendiente de la recta de regresión, representa el **coeficiente de regresión**, que indica lo que aumenta en promedio la variable respuesta por cada unidad que aumenta la variable explicativa. Si la pendiente fuese cero, el valor esperado de la variable respuesta no variaría al variar el de la variable explicativa, indicando la ausencia de relación lineal entre ambas variables.

El parámetro α del modelo, la **ordenada de origen**, coincide con el valor esperado de la variable respuesta cuando la variable explicativa vale cero, es decir: $\alpha = E[Y|X = 0]$. Debemos tener en cuenta que, en ocasiones, que el valor de la variable explicativa sea cero no tiene sentido físico, pero sí matemático, que será cuando el rango de los valores posibles para la variable explicativa no incluye al cero (17).

1.5. El Modelo de Regresión Lineal Múltiple

Expresar la variable respuesta como función de una única variable explicativa es una limitación importante del modelo de regresión lineal simple. Una extensión natural del modelo de regresión es emplear varias variables explicativas, dando lugar al **Modelo de Regresión Lineal Múltiple**, que se puede expresar de la siguiente manera:

$$E[Y|X_1 = x_1, X_2 = x_2, \dots, X_K = x_K] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

La anterior expresión nos proporciona el valor esperado para la variable respuesta, dada una combinación de valores para las variables explicativas. Para una observación real $(y, x_1, x_2, \dots, x_K)$ el valor de la variable respuesta no

coincidirá exactamente con el valor esperado proporcionado por el modelo. La diferencia entre el valor real y el estimado a partir del modelo se denomina **residuo** y se denota con la letra e: $e = y - \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. Los residuos así definidos pueden entenderse como errores de predicción.

El modelo de regresión lineal múltiple se completa con las siguientes hipótesis: el valor de las variables explicativas se conoce sin error, los residuos correspondientes a las diferentes observaciones son **independientes** entre sí e independientes de las variables explicativas, **normales** con media cero y **varianza constante**: $e \sim N(0, \sigma^2)$. Hay que notar que no se exige la normalidad de las variables explicativas ni de la variable respuesta.

Los parámetros del Modelo de Regresión Lineal Múltiple se estiman a partir del **método de los mínimos cuadrados**, que consiste en minimizar la suma de los cuadrados de los residuos del conjunto de datos empleado para estimar el modelo. Si tenemos *n* observaciones completas para estimar el Modelo de Regresión Lineal Múltiple, para una combinación de valores de los parámetros $(b_0, b_1, \ldots, b_K)$ y para una de las observaciones (la *i*-ésima), el residuo correspondiente será:

$$e_i = y_i - (b_0 + b_1\, x_{1i} + b_2\, x_{2i} + \cdots + b_k x_{ki})$$

La estimación de los parámetros, el vector $(\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_K)$, será aquella combinación de valores para la que se minimiza la siguiente expresión:

$$Min\left\{\sum_{i=1}^{n} e_i^2\right\}$$

Al valor estimado para la variable respuesta del individuo *i*-ésimo se le denota $\hat{y}_i$. Con lo anterior, podemos escribir: $e_i = y_i - \hat{y}_i$ (18).

El error expresa el efecto que tienen sobre la variable respuesta los factores que nos son explicados por medio de las variables explicativas del modelo(11).

Aparte de poder usar más de una variable explicativa, también podemos extender el modelo de regresión lineal considerando **efectos no lineales** de las variables explicativas, a menudo modelables mediante funciones polinómicas. Otra extensión del modelo es considerar **interacciones** entre dos o más variables explicativas, que se da cuando el efecto de una variable explicativa

sobre el valor medio de la variable respuesta cambia al hacerlo el de otra u otras variables explicativas. También podemos incluir variables explicativas cualitativas, modeladas mediante variables *dummy*, anteriormente explicado (14).

1.6. Inferencia

Un modelo bien diseñado nos permitirá sacar conclusiones de la población a partir de una muestra representativa con un margen de error conocido reducido, es decir, nos permitirá hacer **inferencia**. La inferencia es un proceso por el que, a partir de unas observaciones se hace una deducción de las consecuencias lógicas. Para asegurar que las conclusiones a las que hemos llegado son válidas, el proceso deberá tener un efecto significativo.

Será necesario entonces, conocer la significación global del ajuste de mínimos cuadrados. Para ello nos fijaremos en las siguientes medidas de variabilidad:

- Variabilidad total: Mide la dispersión de los valores observados de *y* en torno a su media. También es conocida como suma cuadrados total: SC_T.

$$SC_T = \sum (y_i - \bar{y})^2$$

- Variabilidad explicada: La variabilidad total de los valores estimados por la recta de regresión. Es llamada suma de los cuadrados del modelo o explicada:SC_M.

$$SC_M = \sum (\hat{y} - \bar{y})^2$$

- Variabilidad inexplicada: La dispersión de los valores observados de *y* en torno a la recta de regresión. Es la suma del cuadrado de los residuos: SC_R.

$$SC\ total = SC\ explicada + SC\ residual = \sum (y_i - \hat{y})^2$$

Es fácil comprobar la siguiente relación entre los tres tipos de variabilidad definidos:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y_i - \hat{y})^2$$

Es decir:

$$SC_T = SC_M + SC_R$$

El **coeficiente de determinación** es la medida de la eficacia del ajuste de la ecuación de regresión de la muestra a los valores observados de la variable respuesta. Se define como la proporción de la variabilidad de la variable respuesta explicada a partir del modelo. Lo podremos calcular de la siguiente manera:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SC_M}{SC_T} = 1 - \frac{SC_R}{SC_T}$$

Cuanto mayor sea R^2, más acertado será el ajuste. El valor máximo que puede tomar R^2 es 1, que significará que toda la variación de la variable respuesta es explicada por el modelo. Por el contrario, el valor mínimo que puede tomar es 0, lo que significaría que ninguna variación de la misma sería explicada por el modelo de regresión (15).

1.7. Elección de variables

En las ocasiones en las que tenemos varias variables explicativas deberemos hacer una selección para saber qué combinación de ellas es la óptima para calcular el modelo más ajustado. Para hacer esta selección utilizaremos el método de ***stepwise regression***. Este método evalúa si dejar o eliminar una variable explicativa y el efecto que tendrá sobre las demás variables al hacerlo. Se basará en criterios predefinidos como el R^2, $PRESS$, $p - value$, $R^2_{i \cdot Otras}$ y FIV_i; que serán explicados más adelante. La mayoría de paquetes estadísticos obtienen el modelo aplicando este método de manera automática.

Si queremos aplicar este método de manera manual podremos abordarlo mediante dos procesos:

- ***Forward* o *step-up*:** Se inicia diseñando un modelo con una variable explicativa, a continuación, iremos añadiendo otra variable explicativa, estudiando sus efectos paso a paso.
- ***Backward* o *step-down*:** Es el modelo contrario al anterior. Comenzamos con un modelo en el que se han incluido todas las variables, y vamos eliminando una a una las variables no significativas hasta obtener un modelo óptimo (16).

2. VALIDACIÓN DEL MODELO

2.1. Condiciones del modelo

Para asegurarnos de que aplicamos el Modelo de Regresión Lineal Múltiple correctamente, deberemos comprobar que se cumplen las siguientes condiciones:

- **Suposición de linealidad.** El modelo ha de ser lineal, es decir, debe existir una relación lineal entre cada variable independiente y la dependiente.

  Como hemos mencionado anteriormente, el modelo también puede aplicarse cuando la relación entre la variable explicativa y la variable respuesta no es lineal, si podemos encontrar una transformación de una de ellas o de ambas, de manera que la relación entre las variables transformadas sea lineal. Esto complica la explotación del modelo, pero le da generalidad. En este trabajo no consideraremos las transformaciones de las variables para linealizar la relación, por exceder el alcance previsto para el mismo.
- Las variables independientes no pueden estar correlacionadas entre sí. La violación de esta hipótesis de denomina **multicolinealidad.**
- Los residuos, la diferencia entre el valor observado y el estimado, deben ser **independientes** entre sí.
- Los residuos no deben estar relacionados con las variables independientes y su varianza ha de ser constante para diferentes valores de las variables independientes (hipótesis de **homocedasticidad**). Esto lo podremos deducir observando el diagrama de dispersión en el que los puntos estarán repartidos de una manera más o menos constante sobre la línea.
- Los residuos seguirán una **distribución normal** (19–21)**.**

2.2. Problemas con las hipótesis del modelo

El cumplimiento de los requisitos hipotetizados para la construcción del Modelo de Regresión Lineal Múltiple se exige con cierta flexibilidad. Por ejemplo, modelos de Regresión Lineal con una distribución claramente alejada de la normalidad, frecuentemente producen resultados válidos. Mientras que modelos

de Regresión Lineal con distribución muy próxima a la normalidad en sus residuos no aseguran la validez del modelo.

También es usual la flexibilidad en el requisito de la independencia entre las variables explicativas: si varias variables explicativas están relacionadas con la variable respuesta, hasta el punto de resultar útiles dentro del modelo, es muy frecuente que las primeras estén relacionadas entre sí. De esta manera, sólo se considera que hay un problema de **multicolinealidad** cuando la relación lineal entre ellas es muy acusada, en cuyo caso se produce un gran aumento en la variabilidad de los estimadores, llevándolos a no ser significativamente diferentes de cero. Para ello se suele recurrir a ver en qué medida cada variable explicativa está explicada por el resto, mediante el estadístico $R^2_{i \cdot Otras}$ o, equivalentemente, podemos recurrir a la medida en que se multiplica la varianza de la estimación del parámetro asociado a una variable explicativa, mediante el estadístico FIV_i (Factor de Inflación de la Varianza). Esto nos ayudará a decidir si debemos o no eliminar una variable explicativa del modelo (22).

Cuando no podemos aceptar que la varianza de los errores es constante tenemos un problema de **heterocedasticidad**.

Los modelos también pueden verse afectados por la presencia de ***outliers*** o valores extremos, que pueden tener un efecto desproporcionado sobre los parámetros estimados. En este trabajo emplearemos la distancia de Cook como una medida de disonancia de un individuo con el resto de la muestra. El criterio habitual es rechazar los individuos cuya distancia de Cook al modelo sea mayor que 1,, asegurándonos así de que ninguna distancia es considerablemente mayor al compararse con el resto de ellas (23).

Sobreajuste

El Modelo de Regresión, concebido como una herramienta inferencial, se estima a partir de una muestra, pero debe ser útil para su explotación sobre la población de origen. Si el número de parámetros del modelo es demasiado grande, puede suceder que el modelo pierda capacidad para predecir la respuesta de nuevas observaciones, pese a aumentar el valor del coeficiente de determinación, que no disminuirá al añadir nuevas variables. Una manera de corregir esto es ajustar el coeficiente de determinación, para corregirlo por los grados de libertad. Esto

se consigue dividiendo las sumas de cuadrados residual y total por sus respectivos grados de libertad (cambiamos las sumas de cuadrados por cuadrados medios) en la expresión: $R^2 = 1 - \frac{SC_R}{SC_T}$, de manera que obtenemos el **coeficiente de determinación ajustado**: $R^2_{adj} = 1 - \frac{SC_R/(n-1-K)}{SC_T/(n-1)} = 1 - \frac{CM_R}{CM_T}$. De esta manera, aunque siempre que añadimos una variable, la SC_R se reduce. Si la reducción es pequeña, el denominador $(n-1-K)$ se reduce en mayor medida y el CM_R aumenta, con lo que el R^2_{adj} sí se reduciría. Esto sirve para evidenciar que al añadir parámetros a estimar, supone el coste de perder grados de libertad residuales (16).

2.3. Significación del efecto de cada variable

Dado el modelo:

$$E[Y|X_1 = x_1, X_2 = x_2, \dots, X_K = x_K] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K$$

La variable X_k influye sobre el valor esperado de la variable respuesta cuando $\beta_k \neq 0$. Por lo tanto, para estudiar si la variable X_k tiene un efecto significativo tenemos que contrastar si el valor poblacional del parámetro asociado (β_k) es o no distinto de 0. El **contraste de hipótesis** asociado es el siguiente:

$$\left.\begin{matrix} H_0: \beta_k = 0 \\ H_1: \beta_k \neq 0 \end{matrix}\right\}$$

Sabemos que, si $\beta_k = 0$, entonces $\frac{b_k}{S_{b_k}} \sim t_{n-1-K}$, mientras que, si $\beta_k \neq 0$, el cociente anterior tiende a ser, en valor absoluto, mayor que una t de Student:

$$Si\ \beta_k \neq 0 \Longrightarrow \left|\frac{b_k}{S_{b_k}}\right| > t_{n-1-K;\alpha/2}$$

En las expresiones anteriores, S_{b_k} es la desviación típica del estadístico b_k (su **error estándar**).

Lo usual es retirar del modelo las variables que no tengan un efecto significativo sobre el valor esperado de la variable respuesta. Sin embargo, esto debe hacerse con cuidado, ya que una variable puede tener o no un efecto significativo según otras variables estén o no presentes en el modelo, por lo que retiraremos o añadiremos variables al modelo de una en una y estudiando cada resultado,

ya que cualquier cambio en una de las variables puede provocar cambios en la significatividad del efecto del resto de las variables.

Cabe destacar que, aunque en la Regresión Lineal Simple de la variable respuesta con cada variable explicativa nos dé un efecto significativo, este no será un criterio para incluir dicha variable en el Modelo de Regresión Lineal Múltiple. Ya que es posible que al combinarse con el resto de variables, esta pierda la significación de su efecto.

2.4. Interpretación del efecto de una variable

Cuando decimos que una variable explicativa tiene un **efecto significativo** sobre el valor esperado de la variable respuesta tenemos que ser capaces de explicar el significado de dicha expresión, es decir, tenemos que interpretar en qué consiste dicho efecto. Para ello mostraremos cómo se puede separar algebraicamente dicho efecto del del resto de las variables. Ilustrativamente, lo haremos para la variable X_1 en un modelo con sólo tres variables explicativas, ya que es directamente generalizable para cualquier variable y cualquier número de variables explicativas.

Dada la expresión del modelo:

$$E[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Es fácil ver que:

$$E[Y|\boldsymbol{X_1} = \boldsymbol{x_1} + \mathbf{1}, X_2 = x_2, X_3 = x_3] - E[Y|\boldsymbol{X_1} = \boldsymbol{x_1}, X_2 = x_2, X_3 = x_3] = \beta_1$$

Es decir, $\boldsymbol{\beta_1}$ es lo que aumenta en promedio el valor de la variable respuesta cuando X_1 aumenta una unidad, manteniéndose constante el valor de X_2 y X_3.

Aplicando esto para cualquier variable explicativa en un modelo con un número arbitrario de variables explicativas, podemos decir, para la variable i-ésima:

$\boldsymbol{\beta_i}$ es lo que aumenta en promedio el valor de la variable respuesta cuando X_i aumenta una unidad, manteniéndose constante el valor de las demás variables explicativas.

De esta manera, podemos estudiar el **efecto marginal** de cada variable explicativa sobre la variable respuesta, bloqueando el efecto de las demás variables explicativas. Tenemos que notar que, dado que las variables explicativas normalmente no son totalmente independientes, no siempre es

posible garantizar la posibilidad de incrementar una unidad una de las variables explicativas sin que se modifique el valor del resto.

En el caso de una variable *dummy*, un incremento unitario supone pasar de cero (la base) a uno (la categoría asociada a la *dummy*), es decir, podemos estudiar el efecto de cada categoría en relación con el de la base.

También hay que notar que cuando el modelo presenta interacciones u otros efectos no lineales (potencias, raíces, logaritmos, …) la interpretación del efecto de las variables individuales se complica notablemente.

2.5. Reporte de los resultados

Aunque hayamos tenido en cuenta diferentes medidas en el análisis realizado para el Modelo de Regresión, no se suelen reportar todas estas. Los principales datos que se recomienda que sean expuestos junto al estudio realizado son los siguientes:

- La descripción de todas las variables, subrayando cuál de ellas es la variable respuesta.
- El coeficiente de determinación lineal.
- Cada coeficiente estimado acompañado de su intervalo de confianza y/o $p - value$.
- Se recomienda también, aunque es menos frecuente que se haga, representar el diagrama de dispersión de la variable dependiente con cada variable independiente, añadiendo también su recta de regresión.
- Aclarar si se han verificado las condiciones del modelo: ¿tiene sentido que el modelo sea lineal? ¿Existe multicolinealidad? ¿Se ha detectado algún *outlier* o punto influyente? ¿Los errores siguen una distribución normal, son independiente entre sí y con las variables independientes? (16).

3. APLICACIÓN DEL MODELO

A continuación, ilustraremos el modelo de Regresión Lineal Múltiple con un ejemplo, explicando cada paso a desarrollar. Comenzaremos definiendo la población, la muestra y las variables que vamos a utilizar. Seguiremos con el estudio sobre cómo se distribuye cada variable por separado, para después aplicar la Regresión Lineal Simple a cada par de variables. Finalmente, calcularemos el modelo de Regresión Lineal Múltiple más adecuado para nuestros datos, analizando el resultado y viendo otras posibles aplicaciones.

Cabe remarcar que el objetivo por el que se han recopilado estos datos no es un estudio de investigación, si no ilustrar de una manera práctica el uso del modelo de Regresión Lineal Múltiple.

Sin embargo, en la medida en que consideremos que la muestra es representativa de la población hipotética, podríamos sacar conclusiones sobre la población a partir de la muestra con un margen de error conocido reducido, es decir, hacer inferencia.

Los cálculos y gráficos se han realizado por medio del programa *f*-Stats para Excel.

3.1. Población y muestra

En una **población** dada, de **deportistas de alto rendimiento**, observamos que los diferentes individuos pesan diferente, es decir, constatamos variabilidad en el peso de los individuos de la población. Ante este hecho nos preguntamos: ¿de qué depende que unos deportistas tengan un peso más elevado que otros? Para responder a la pregunta, hacemos un estudio que pretende estimar el peso a partir de ciertas medidas antropométricas en una población de jóvenes deportistas de ambos sexos. Este estudio permitirá determinar el efecto que una variación en cada una de las medidas tiene sobre el peso esperado de los deportistas, bloqueando el efecto de variaciones en el resto de las medidas, es decir, estimaremos el efecto marginal para cada medida.

La **muestra**: tomamos una muestra homogénea con 186 socios de una escuela de alto rendimiento deportivo; 100 mujeres y 86 hombres, con edades entre 20 y 25 años. Disponíamos de mayor volumen de datos, pero hemos seleccionado solamente a los individuos en la franja de edad reseñada, para homogeneizar la

muestra, ya que no estamos interesados en el efecto de la edad sobre el peso esperado.

Como hemos dicho, este es un ejemplo ilustrativo del Modelo de Regresión. La base de datos que hemos utilizado en el siguiente análisis procede de distintas series de medidas antropométricas de militares, publicadas por el ejército de Estados Unidos (24,25).

3.2. Variables

De cada uno de ellos hemos obtenido las siguientes medidas (figura 2):

- **Peso** en *kg*. Esta es el objeto de nuestro estudio, la **variable respuesta**. Queremos relacionar la variabilidad del peso a partir de la variabilidad del resto de variables.
- **Edad** en años.
- **Sexo**. Para hacer una regresión lineal todas las variables han de tener un valor numérico. Al ser el sexo una variable categórica (hombre o mujer), hemos asignado el valor 1 para hombre y el 0 para mujer, creando así la variable *dummy* hombre.
- **Altura** en *cm* medida por medio de una báscula a la vez que el peso con la persona descalza y en postura recta.
- **Diámetro biacromial**, para tomar esta medida primero debemos palpar con las manos cada acromion y seguidamente colocar el antropómetro.
- **Diámetro biilíaco**, debemos localizar la espina ilíaca anterior superior de cada lado y realizar la medida.
- **Diámetro bitrocantéreo**, localizaremos los trocánteres mayores de cada fémur y mediremos.
- **Diámetro pectoral**, se realizará al nivel de los pezones en la mitad de la espiración.
- **Profundidad pectoral**, esta será la distancia que existe entre el esternón y la columna vertebral al nivel de los pezones. También se medirá en la mitad de la espiración.
- **Diámetro del codo**, será la suma de la distancia entre los epicóndilos lateral y medial de cada húmero.

- **Diámetro de la muñeca**, se mide localizando la epífisis estiloides del cúbito y del radio y sumando la de cada brazo.
- **Diámetro de la rodilla**, se obtiene al sumar la distancia entre los epicóndilos lateral y medial de cada fémur.
- **Diámetro del tobillo**, es la suma de la distancia del maléolo medial de la tibia y el maléolo lateral del peroné de ambos tobillos.

Los diámetros han sido medidos en *cm* por medio de antropómetros de hoja ancha, así como la profundidad pectoral (26–28).

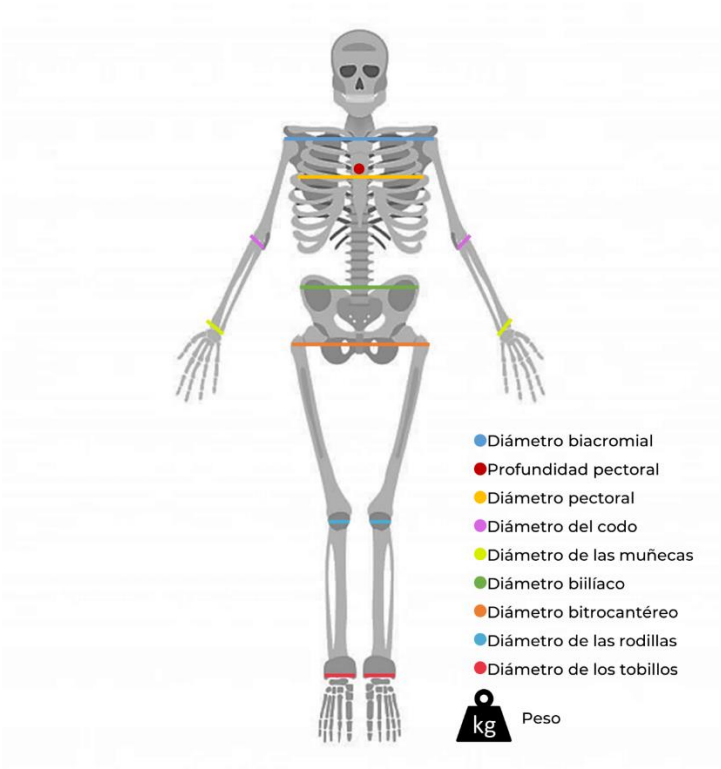


Figura 2. Variables del modelo.

3.3. Estudio univariante

Lo primero que debemos hacer para estudiar la influencia que pueden tener sobre el peso diferentes variables, es analizar cada una de ellas por separado.

Para ilustrar el estudio que hemos realizado con cada una, escogemos la variable altura, peso y diámetro bitrocantéreo.

Peso

Tenemos una muestra con 186 observaciones del peso de los individuos, con una media de 66,79 ($95\%IC = [65{,}01; 68{,}58]$); una mediana de 65,35 y una desviación típica es de 12,34 ($95\%IC = [11{,}2; 13{,}74]$). Al ser la media mayor que la mediana, la distribución de los datos sugieren una ligera asimetría positiva. Esto también se ve en el valor positivo del coeficiente de asimetría de Fisher.

No existen datos anómalos por defecto, pero sí hallamos dos datos anómalos por exceso: 101,4 y 108,6.

Representamos los datos con un histograma de 5 intervalos y un diagrama de caja en la figura 3.

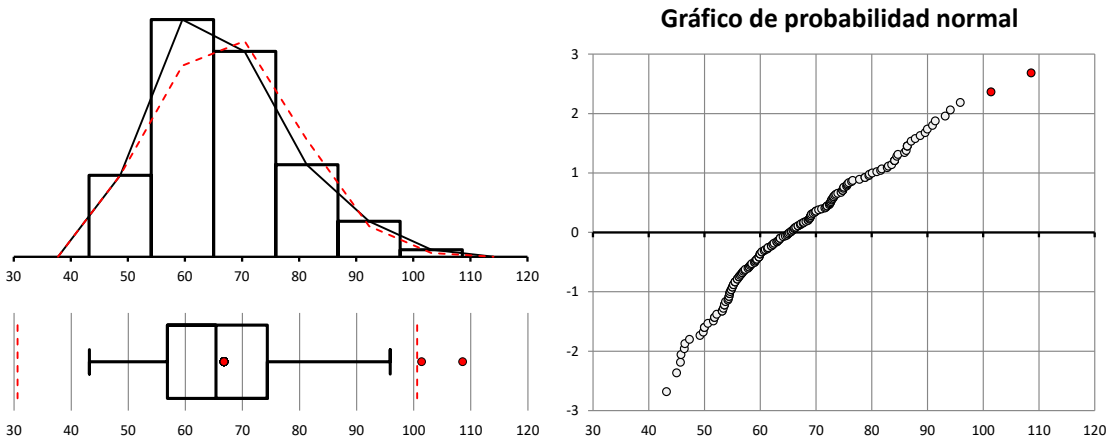


Figura 3. Distribución de los datos de la variable peso.

La línea roja del histograma representa la forma que daría el polígono de frecuencias en una distribución normal con la misma media y varianza que las obtenidas de la muestra. De manera visual, apreciamos una ligera asimetría positiva. Esta asimetría también se puede apreciar en el diagrama de caja al ver que el cuadrado derecho es algo más grande que el izquierdo. Sin embargo, podemos considerar que no hay una gran desviación de la normalidad.

A continuación, vemos el análisis de la variable peso para mujeres en rojo comparándose con sus valores para los hombres en azul (figura 4). Podemos comprobar que hay una diferencia significativa ($p - value < 0{,}001$) para la media de los hombres 75,70 ($95\%IC = [73{,}40; 78]$) y la de las mujeres 59,14 ($95\%IC = [57{,}63; 60{,}64]$).

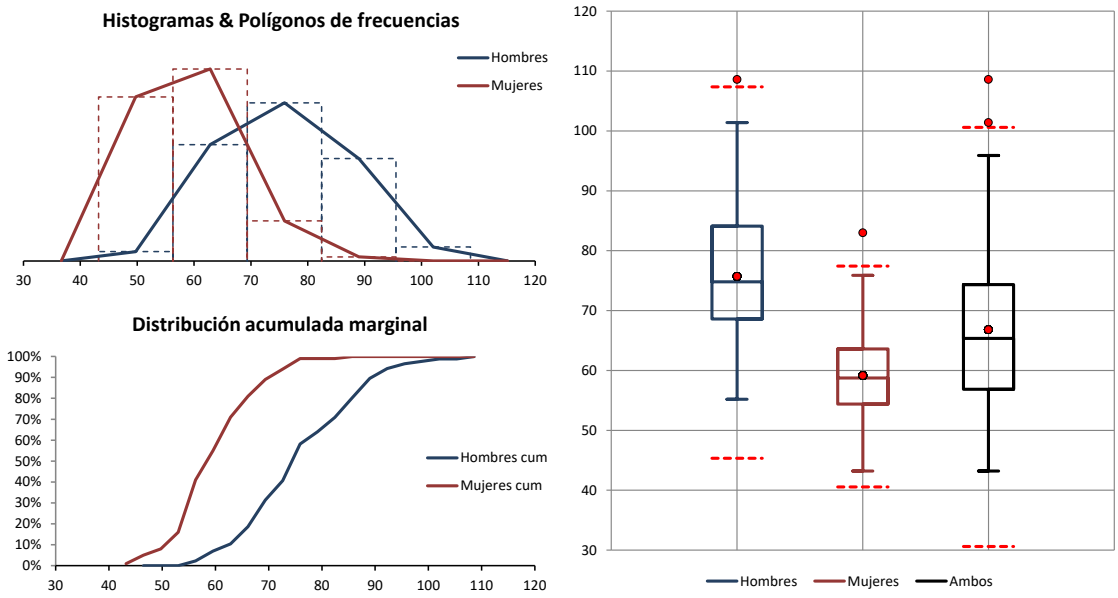


Figura 4. Distribución de los datos del peso para hombre y para mujeres.

Diámetro bitrocantéreo

Al analizar la variable del diámetro bitrocantéreo observamos que la mediana de 31,85 es ligeramente mayor que la media de 31,70 ($95\% IC = [31{,}43; 31{,}98]$); lo que es indicio de una ligera asimetría negativa en la distribución de los datos que también vemos reflejada en los gráficos de la figura 5.

Identificamos dos datos anómalos por defecto: 26,3 y 26,8.

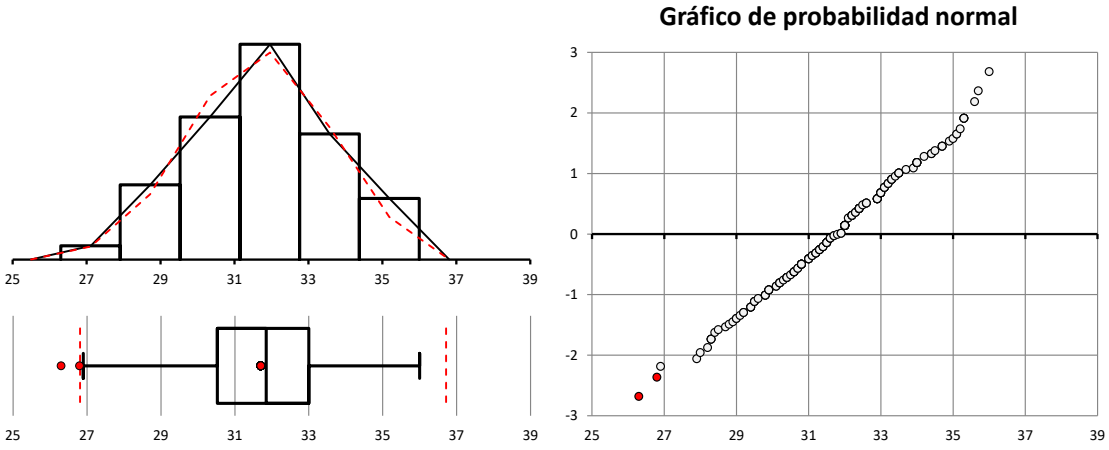


Figura 5. Distribución de los datos del diámetro bitrocantéreo.

A pesar de la ligera asimetría negativa que observamos, podemos decir que los datos se ajustan de manera adecuada a la normalidad.

Si separamos las medidas para hombres y para mujeres (figura 6) obtenemos que existe una diferencia significativa ($p - value\ 0{,}001$) para la media de los hombres 32,19 ($95\%IC[31{,}80; 32{,}59]$) y la de las mujeres 31,28 ($95\%IC[30{,}91; 31{,}65]$).

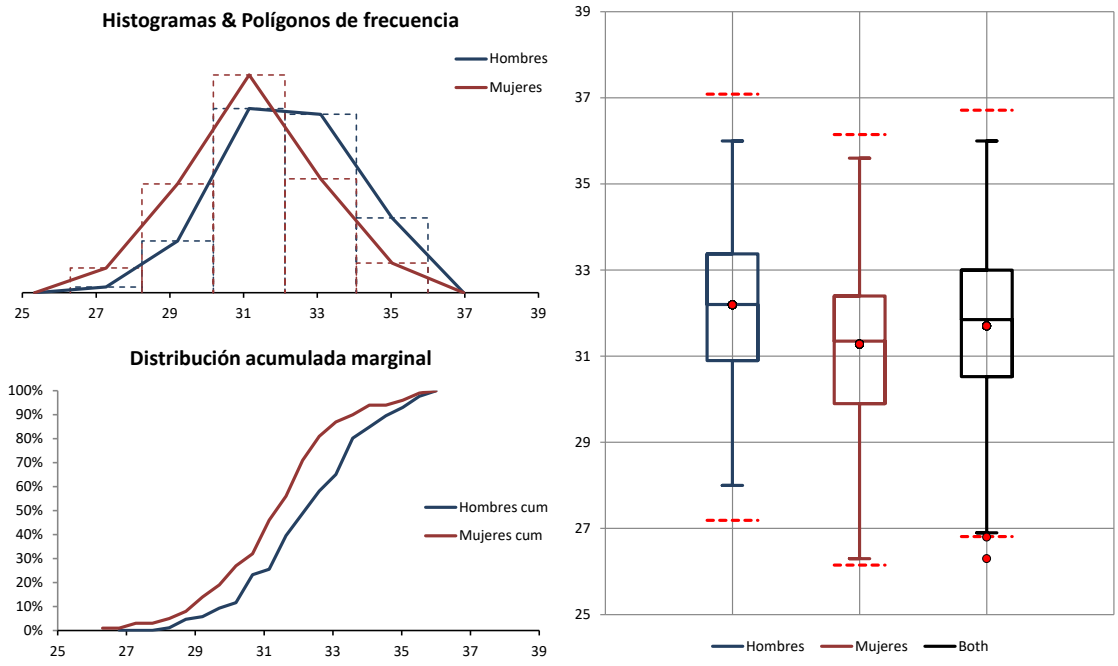


Figura 6. Distribución de los datos del diámetro bitrocantéreo para hombres y para mujeres.

Altura

En este caso la media de 171,02 ($95\%IC = [169{,}67; 172{,}37]$) es mayor a la mediana de 170,2. Por lo tanto sugiere una ligera asimetría positiva que también podemos observar en las imágenes de la figura 7, aunque no supone una desviación de la normalidad.

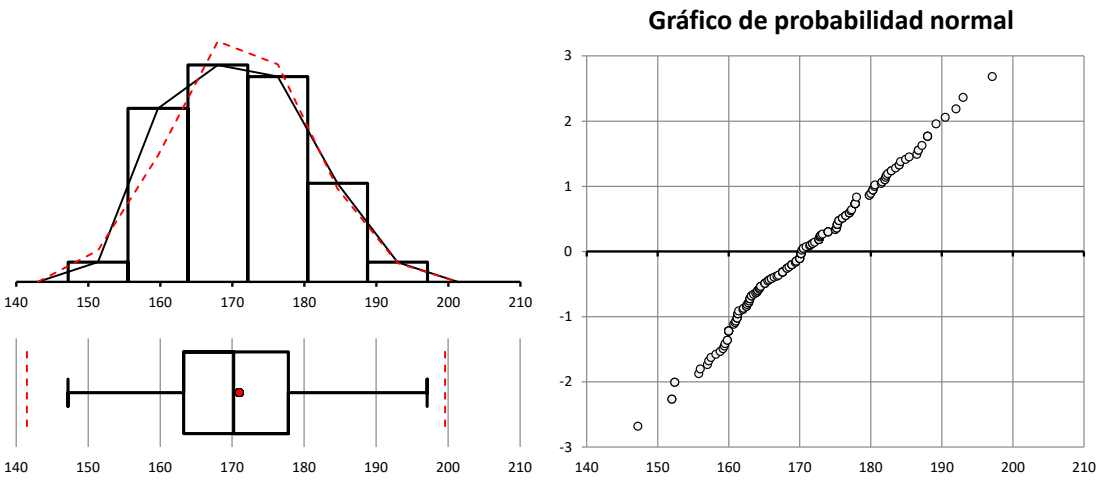


Figura 7. Distribución de los datos de la variable altura.

En promedio la altura de los hombres con media 177,74 ($95\% IC = [176{,}15; 179{,}34]$) es significativamente ($p - value > 0{,}001$) superior a la de las mujeres 165,24 ($95\% IC = [163{,}96; 166{,}56]$), como vemos representado en la figura 8.

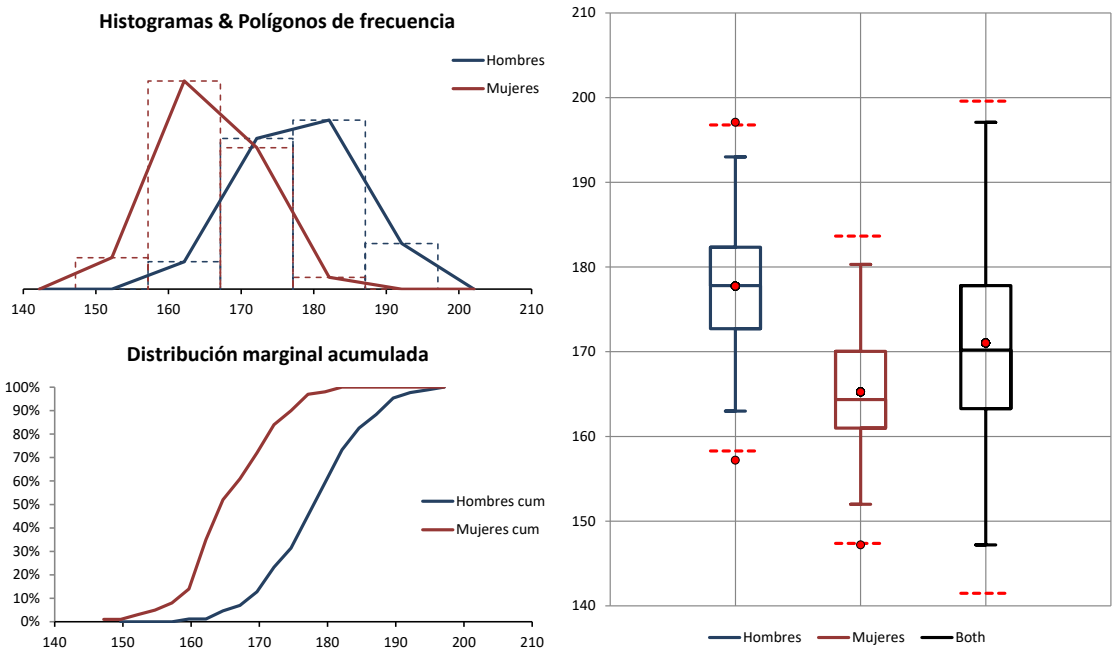


Figura 8. Distribución de los datos de la altura para hombre y para mujeres.

Como hemos visto que la media de los hombres es significativamente mayor a la de las mujeres en estas variables, queremos comprobar si pasa lo mismo con el resto. De manera que representamos todas las variables, separando a hombres de mujeres en la tabla 2.

Tabla 2. Análisis univariante para población de hombres y de mujeres.

	Hombres (86)		Mujeres (100)					
	μ_H	S_H^2	μ_M	S_M^2	$p-value$	$\mu_H-\mu_M$	$95\%CI(\mu_H-\mu_M)$	
Peso	75,7	10,74	59,14	7,58	0,000	16,56	13,90	19,22
Altura	177,74	7,43	165,24	6,48	0,000	12,50	10,49	14,51
D. biacromial	41,51	2,39	36,59	1,66	0,000	4,92	4,33	5,51
D. biilíaco	27,77	2,10	27,11	2,08	0,034	0,66	0,05	1,27
D. bitrocantéreo	32,19	1,84	31,28	1,85	0,001	0,91	0,37	1,45
Prof. pectoral	20,17	2,21	17,35	1,51	0,000	2,82	2,28	3,36
D. pectoral	29,57	2,22	25,89	1,63	0,000	3,68	3,12	4,24
D. codo	14,44	0,91	12,27	0,83	0,000	2,17	1,92	2,42
D. muñeca	11,13	0,65	9,78	0,61	0,000	1,35	1,17	1,53
D. rodilla	19,55	1,15	17,86	0,95	0,000	1,69	1,39	1,99
D. tobillo	14,53	0,91	12,83	0,87	0,000	1,70	1,44	1,96

Vemos que la media de los hombres es significativamente mayor que la de las mujeres en estas medidas, pues los $p-values$ son todos menores de 0,05 y los intervalos de confianza no contienen el 0. De modo que podríamos decir que tenemos dos poblaciones diferenciadas. Sin embargo, esto no alterará nuestro estudio de regresión lineal ya que, al ser el sexo una variable, el modelo dará resultados diferentes para hombres y para mujeres.

3.4. Regresión Lineal Simple: Estudio bivariante

Antes de calcular el modelo de regresión múltiple, deberemos estudiar de forma marginal cada variable explicativa con la variable respuesta y cada par de variables explicativas entre ellas. Como hemos hecho anteriormente, mostraremos la relación entre algunos pares de ellas para ilustrar el proceso. El objetivo de graficar estos datos es comprobar que existe una relación lineal entre las variables.

Relación del peso con cada variable explicativa:

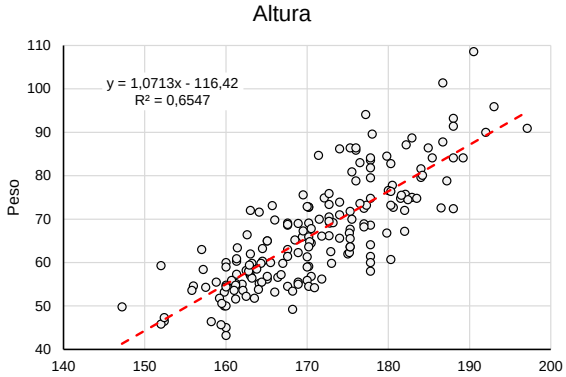


Figura 9. Relación de la altura con el peso.

Por cada *cm* que aumente la altura, el peso aumentará 0,7 *kg* en promedio.

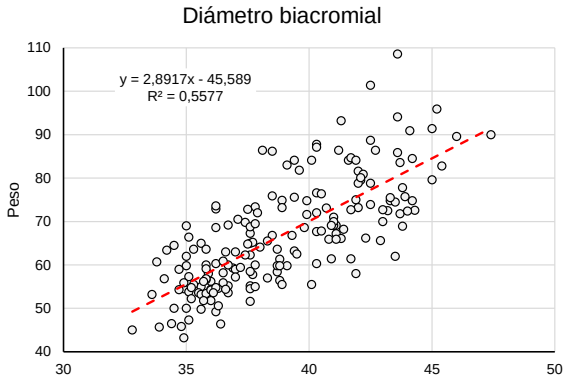


Figura 10. Relación del diámetro biacromial con el peso.

Por cada *cm* que aumente el diámetro biacromial, el peso aumentará 2,89 *kg* en promedio.

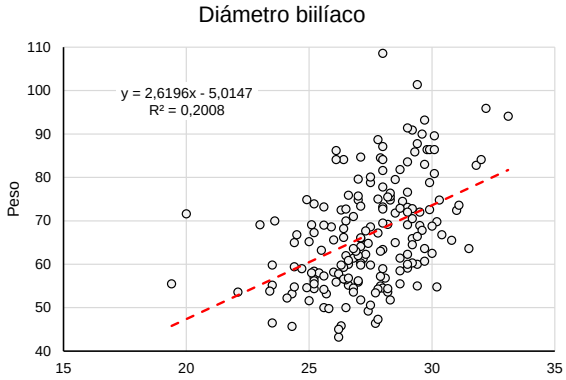


Figura 11. Relación del diámetro biilíaco con el peso.

Por cada *cm* que aumente el diámetro biilíaco, el peso aumentará 2,62 *kg* en promedio.

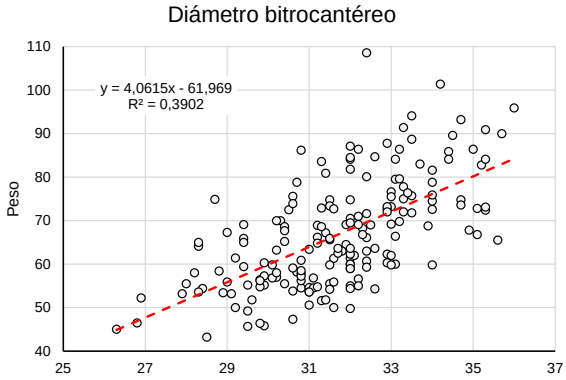


Figura 12. Relación del diámetro bitrocantéreo con el peso.

Por cada *cm* que aumente el diámetro bitrocantéreo, el peso aumentará 4,06 *kg* en promedio.

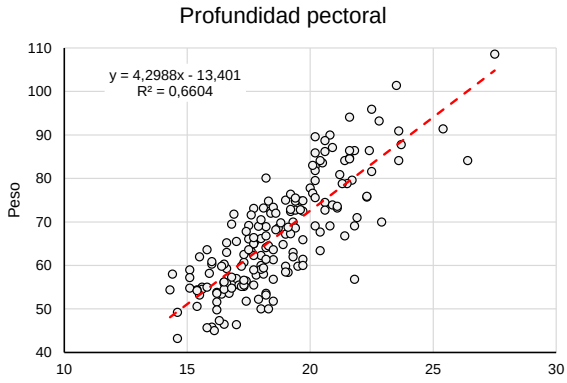


Figura 13. Relación de la profundidad pectoral con el peso.

Por cada *cm* que aumente la profundidad pectoral, el peso aumentará 4,30 *kg* en promedio.

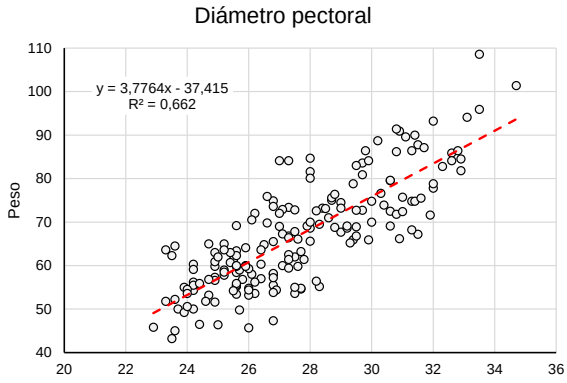


Figura 14. Relación del diámetro pectoral con el peso.

Por cada *cm* que aumente el diámetro pectoral, el peso aumentará 3,78 *kg* en promedio.

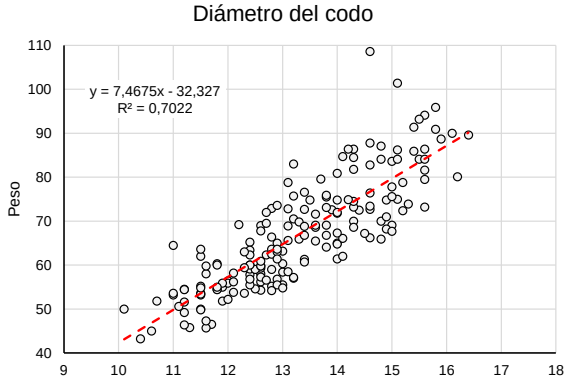


Figura 15. Relación del diámetro del codo con el peso.

Por cada *cm* que aumente el diámetro del codo, el peso aumentará 7,47 *kg* en promedio.

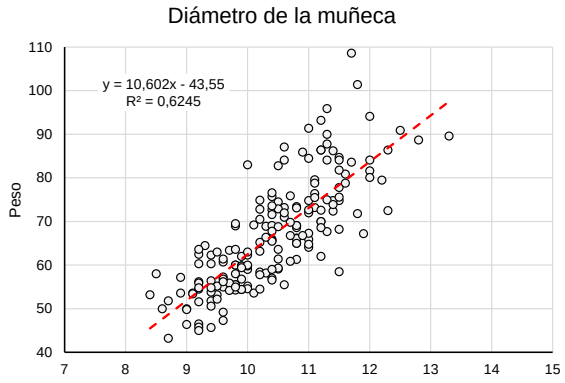


Figura 16. Relación del diámetro de la muñeca con el peso.

Por cada *cm* que aumente el diámetro de la muñeca, el peso aumentará 10,60 *kg* en promedio.

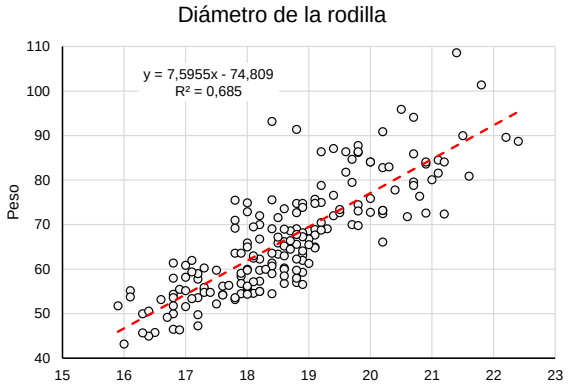


Figura 17. Relación del diámetro de la rodilla con el peso.

Por cada *cm* que aumente el diámetro de la rodilla, el peso aumentará 7,60 *kg* en promedio.

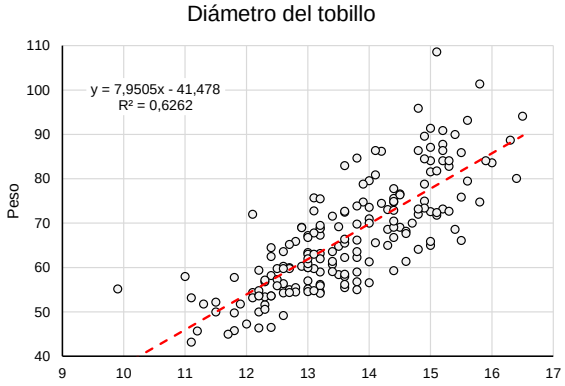


Figura 18. Relación del diámetro del tobillo con el peso.

Por cada *cm* que aumente el diámetro del tobillo, el peso aumentará 7,95 *kg* en promedio.

Todas estas regresiones tienen un $p - value$ menor de 0,001 por lo que cada variable respuesta tiene una relación lineal significativa con el peso. Esto no quiere decir que todas ellas vayan a ser significativas en el modelo de regresión múltiple.

Ocurre que al juntar muchas variables en un mismo modelo estas interactúan entre sí, ya que hacen la misma función y hace que sean menos significativas. Esto es la **multicolinealidad**. Esta se evitaría si las variables fueran independientes. Aunque es recomendable que las variables explicativas sean independientes en la Regresión Lineal Múltiple, es algo complicado de conseguir ya que, al estar todas las variables explicativas relacionadas con la variable respuesta, fácilmente también estarán relacionadas entre sí. En nuestro caso, al ser todas las variables medidas antropométricas, es lógico pensar que al aumentar el diámetro del pecho también lo hará la profundidad de este, al igual que la altura.

Relación entre pares de variables explicativas:

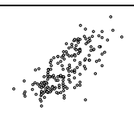
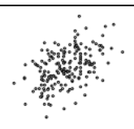
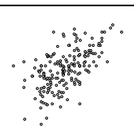
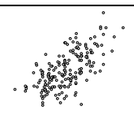
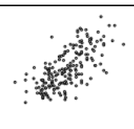
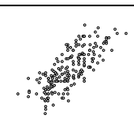
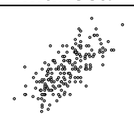
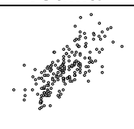
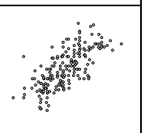
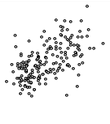
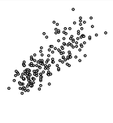
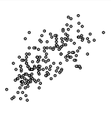
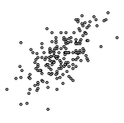
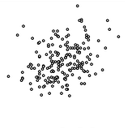
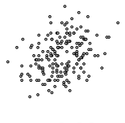
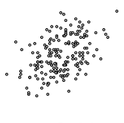
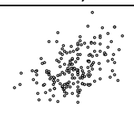
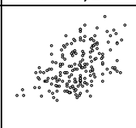
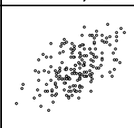
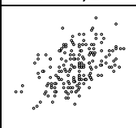
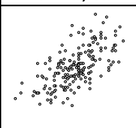
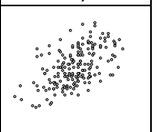
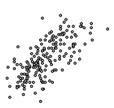
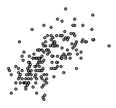
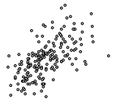
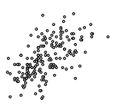
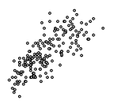
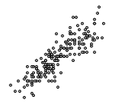
	Diámetro biacromial	Diámetro biilíaco	Diámetro bitrocantéreo	Profundidad pectoral	Diámetro pectoral	Diámetro del codo	Diámetro de la muñeca	Diámetro de la rodilla	Diámetro del tobillo
Altura	 *r* = 0,74	 *r* = 0,45	 *r* = 0,58	 *r* = 0,66	 *r* = 0,68	 *r* = 0,76	 *r* = 0,70	 *r* = 0,72	 *r* = 0,72
Diámetro biacromial		 *r* = 0,30	 *r* = 0,52	 *r*= 0,58	 *r* = 0,80	 *r* = 0,78	 *r* = 0,74	 *r* = 0,72	 *r* = 0,67
Diámetro biilíaco			 *r* = 0,63	 *r* = 0,30	 *r* = 0,26	 *r* = 0,31	 *r* = 0,29	 *r* = 0,46	 *r* = 0,44
Diámetro bitrocantéreo				 *r* = 0,42	 *r* = 0,46	 *r* = 0,53	 *r* = 0,47	 *r* = 0,63	 *r* = 0,56
Profundidad pectoral					 *r* = 0,64	 *r* = 0,69	 *r* = 0,63	 *r* = 0,63	 *r* = 0,60
Diámetro pectoral						 *r* = 0,76	 *r* = 0,73	 *r* = 0,66	 *r* = 0,63
Diámetro del codo							 *r* = 0,86	*r* = 0,80	*r* = 0,85
Diámetro de la muñeca								*r* = 0,78	*r* = 0,79
Diámetro de la rodilla									*r* = 0,81

Figura 19. Relación de cada variable explicativa con el resto de variables explicativas.

3.5. Regresión Lineal Múltiple: *Stepwise backward*

Finalmente queremos calcular el modelo de regresión óptimo para estudiar los datos recopilados con el peso como variable respuesta.

Para ello utilizaremos el método de *stepwise backward* que consiste en comenzar calculando un modelo con todas las variables explicativas de las que disponemos, siendo este el resultado:

Obtenemos un R^2 de 90,69%, lo que nos haría pensar que tenemos un modelo muy potente. El $PRESS$ (3045,55 en este caso) es la suma de los cuadrados de los errores de predicción. Por sí mismo, el $PRESS$ no tiene una interpretación directa, pero, al ser una suma de cuadrados de errores de predicción, el modelo es mejor cuanto menor sea su $PRESS$ asociado. Por este motivo, es una medida útil para comparar diferentes modelos.

A continuación, obtenemos una tabla con la que podemos analizar cada variable por separado:

Tabla 3. Resultados de la estimación del Modelo de Regresión Lineal Múltiple con todas las variables.

	Coef	StdErr	t-stat	sign	IC 95%		FIV_i	$R^2_{i\cdot Otras}$
Cte	-106,09	7,42	14,30	0,000	-120,74	-91,45		
Altura	0,20	0,07	3,59	0,000	0,09	0,31	3,39	70,51%
Hombre	-2,02	1,21	1,67	0,096	-4,40	0,36	4,45	77,53%
D. biacromial	-0,19	0,20	0,97	0,331	-0,57	0,20	4,92	79,65%
D. biilíaco	0,20	0,19	1,08	0,281	-0,16	0,57	1,87	46,59%
D. bitrocant.	0,38	0,26	1,47	0,143	-0,13	0,88	2,91	65,67%
Prof. pectoral	1,63	0,18	8,85	0,000	1,27	1,99	2,27	55,99%
D. pectoral	1,40	0,20	7,05	0,000	1,01	1,79	3,42	70,80%
D. codo	0,56	0,57	0,97	0,332	-0,57	1,69	7,70	87,02%
D. muñeca	0,58	0,67	0,87	0,386	-0,74	1,91	4,68	78,65%
D. rodilla	1,67	0,44	3,77	0,000	0,79	2,54	4,33	76,92%
D. tobillo	1,15	0,51	2,23	0,027	0,13	2,16	4,88	79,52%

La tabla nos muestra la información con la que decidiremos dejar o eliminar las variables en el modelo, sin embargo, no es usual ponerla completa a la hora de publicar un estudio que haya sido analizado por medio de la regresión lineal múltiple.

En la primera columna vemos la estimación de los parámetros del modelo (coeficientes estimados), cuando el valor del coeficiente es cercano a 0, querrá decir que no influye mucho sobre el peso.

Aquí vemos que, para varias variables, su contraste de significatividad del efecto tiene un $p - value$ mayor de 0,05, es decir, su efecto sobre el valor esperado del peso no llega a ser significativo. Las variables cuyo efecto ha resultado significativo son: sexo, diámetro biacromial, diámetro biilíaco, diámetro bitrocantéreo, diámetro del codo, diámetro de la muñeca, diámetro del tobillo.

En la segunda columna de la Tabla 3 se muestra la desviación típica de la estimación de cada parámetro, es decir, su error estándar.

Antes de eliminar todas estas variables nos fijaremos qué porcentaje de cada variable es explicado por el resto de variables explicativas, para ello utilizamos el valor de $R^2_{i\cdot Otras}$. Por ejemplo, el 87,02% de la variabilidad del diámetro del codo está explicado por el resto de las variables explicativas, por lo que el resto de las variables pueden reconstruir gran parte de la información que perdemos si la eliminamos. Además, tiene un $p - value$ de 0,332 lo que nos indica que no tiene un efecto significativo sobre el peso; de modo que elegimos eliminarla del modelo.

Esta información nos la da de una manera similar el FIV_i que, cuanto más elevado sea, querrá decir que una mayor parte de esa variable es explicada por las otras. Así vemos que para el diámetro del codo tiene un valor de 7,70; el más elevado de todos. El FIV_i y el R^2_i nos muestra que existe multicolinealidad, siendo que gran parte de la variable diámetro del codo es explicada por las demás. Esto nos decantará por eliminar esta variable del modelo.

Además, quitamos las variables diámetro biacromial, diámetro de la muñeca y diámetro biilíaco. En el modelo obtenido, el $p - value$ del diámetro de la rodilla será de 0,053; al ser no significativa nos inclinamos por eliminar también esta variable. Pero podría pasar que para otro estudio esta variable fuera esencial, así que estará en manos del investigador decidir si conviene dejarla para tener en cuenta más variables o eliminarla y obtener un modelo más parsimonioso. Nosotros decidimos eliminarla y así obtenemos:

> El coeficiente de determinación a 90,27%. Por el contrario, vemos una mejora del modelo al reducir el $PRESS$ a 2984,03. Los $R^2_{i\cdot Otras}$ han disminuido, ya que al eliminar variables cada una de las que dejamos

aportan más información sobre el estudio, como se puede ver reflejado en la tabla 4.

Tabla 4. Resultados del Modelo de Regresión Lineal por el método stepwise backward.

	Coef	StdErr	t-stat	sign	IC 95%		FIV_i	$R^2_{i\cdot Otras}$
Cte	-106,69	7,17	14,87	0,000	-120,84	-92,53		
Altura	0,2	0,05	4,58	0,000	0,14	0,34	2,90	65,51%
Hombre	-3,02	0,92	3,29	0,001	-4,84	-1,21	2,56	60,88%
Prof. pectoral	1,70	0,18	9,44	0,000	1,35	2,06	2,14	53,22%
D. pectoral	1,44	0,17	8,33	0,000	1,10	1,78	2,55	60,77%
D. rodilla	2,04	0,40	5,09	0,000	1,25	2,83	3,50	71,46%
D. tobillo	1,78	0,44	4,03	0,000	0,91	2,65	3,55	71,82%

En esta ocasión podemos considerar que todas las variables de nuestro modelo son significativas. La tabla 4 nos da los resultados que calcula el modelo, sin embargo, se suelen reportar solamente los valores de la tabla 5.

Tabla 5. Resultados reportados del Modelo de Regresión Lineal Múltiple.

	Coeficientes	StdErr	p-value
Cte	-106,69	7,17	0,000
Altura	0,24	0,05	0,000
Hombre	-3,02	0,92	0,001
Prof pectoral	1,70	0,18	0,000
D. pectoral	1,44	0,17	0,000
D. rodilla	2,04	0,40	0,000
D. tobillo	1,78	0,44	0,000

La recta de regresión que nos da el modelo es la siguiente:

$$peso = -106{,}687 + 0{,}240 \times altura - 3{,}024 \times hombre + 1{,}702 \times prof.pectoral + 1{,}440 \times d.pectoral + 2{,}039 \times d.rodilla + 1{,}776 \times d.tobillo$$

Con este modelo podemos **explicar** el 90,27% de la variabilidad del peso a partir de la variabilidad de las variables explicativas. Esto responde a nuestra pregunta inicial: ¿por qué unos deportistas pesan más que otros? En un 90,27% esta variación es explicada por la altura, el sexo, la profundidad pectoral, el diámetro pectoral, los diámetros de las rodillas y los diámetros de los tobillos de cada deportista de nuestra muestra.

Se podría **predecir** el peso en promedio de 62,22 *kg*, para una persona con una altura de 164 *cm*, mujer, una profundidad pectoral de 17 *cm*, un diámetro pectoral de 31 *cm*, una rodilla con un diámetro de 17 *cm* y un diámetro del tobillo de 12 *cm*.

También podemos **interpretar** que por cada *cm* que aumenta la altura, siendo el resto de las medidas antropométricas constantes, el peso aumentará 20,24 *kg* en promedio.

Por cada *cm* que aumenta la profundidad pectoral, siendo el resto de las medidas constantes, el peso aumentará 1,70 *kg* en promedio.

Por cada *cm* que aumenta el diámetro pectoral, siendo el resto de las medidas constantes, el peso aumentará 1,44 *kg* en promedio.

Por cada *cm* que aumenta el diámetro de las rodillas, siendo el resto de las medidas constantes, el peso aumentará 2,04 *kg* en promedio.

Por cada *cm* que aumenta el diámetro de los tobillos, siendo el resto de las medidas constantes, el peso aumentará 1,78 *kg* en promedio.

Para un mismo valor de todas las medidas, si el deportista es hombre su peso será 3,02 *kg* inferior que si fuera mujer. Este es el efecto del sexo bloqueando el efecto del resto de variables.

Este dato nos llama la atención, ya que anteriormente hemos visto que la media de los hombres para todas las variables es significativamente superior a la de las mujeres. Sin embargo, ante un hombre y una mujer con las mismas características en el resto de las medidas, el peso de la mujer será 3 *kg* superior al del hombre. Esto implicaría que la densidad media de las mujeres es superior a la de los hombres, es decir, podemos concluir que la composición del cuerpo de los hombres es diferente a la de las mujeres.

Este hallazgo es de una gran relevancia, ya que demuestra que, sin haber medido la densidad, esta es superior en mujeres que en hombres. La estadística nos ha llevado a una conclusión que el diseño del estudio no contemplaba en un principio. Esto tiene sentido, ya que las mujeres tienen un porcentaje de grasa corporal mayor que la de los hombres. La grasa es más densa que el músculo y otros tejidos, por tanto aumentará el peso en menos volumen (29).

De esta manera ilustramos como el modelo cumple las 3 funciones principales de la regresión lineal múltiple.

3.6. Distancia de Cook

Otro dato que debemos tener en cuenta, aunque no se suele reportar, es la distancia de Cook. Esta mide la influencia que tiene un solo punto sobre la ecuación de mínimos cuadrados. Nos ayuda a detectar posibles datos atípicos que nos harán valorar si es conveniente eliminar ese dato de la muestra. Para nuestra muestra sería el ejemplo de una persona con una pierna muy ancha y una estatura muy baja.

Sin embargo, en nuestra estimación vemos que todos sus valores son menores de 1, que es el valor con el que consideramos que el dato es anómalo; por lo que consideramos que ningún dato influyente.

Si realizamos la representación gráfica son fáciles de identificar, como podemos observar en la figura 20.

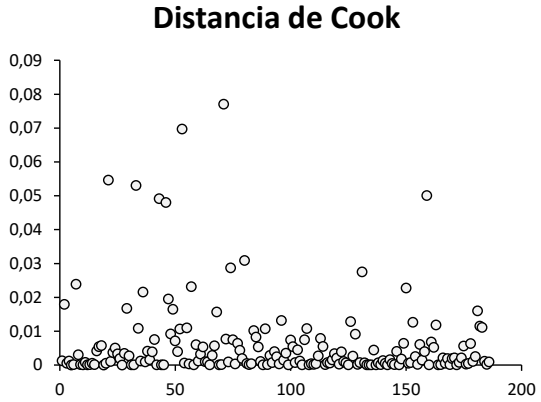


Figura 20. Distancia de Cook.

3.7. Estudio de los residuos

Como hemos comentado anteriormente, para que el modelo sea válido los errores tienen que cumplir ciertas características:

- Los errores son independientes de las variables explicativas. Así, en la figura 21, representamos el diagrama de puntos de los errores con las variables explicativas. Vemos que en la nube de puntos no se forma ninguna forma, por lo que no hay relación, no observamos heterocedasticidad.

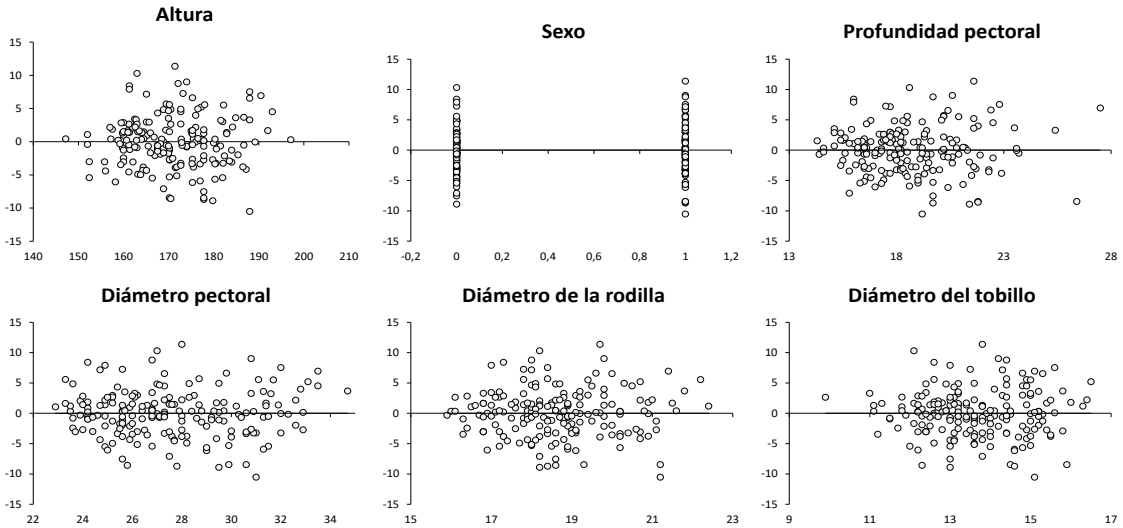


Figura 21. Relación entre los errores y las variables explicativas.

También observamos la relación entre los residuos y el peso estimado (figura 22), no observamos ninguna relación entre ellos.

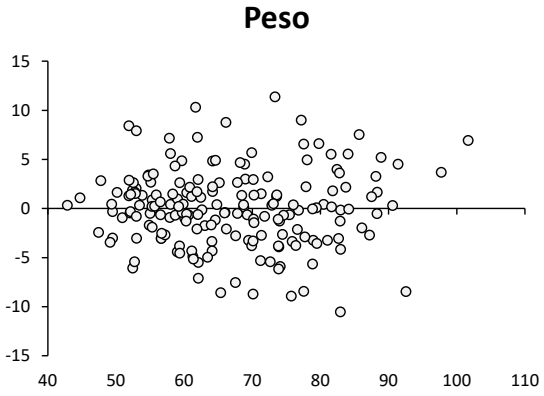


Figura 22. Relación entre los residuos y el peso estimado

- El error sigue una distribución normal. Representamos los errores en la figura 23, como hemos hecho anteriormente con las variables para comprobarlo y vemos que tiene la media en 0.

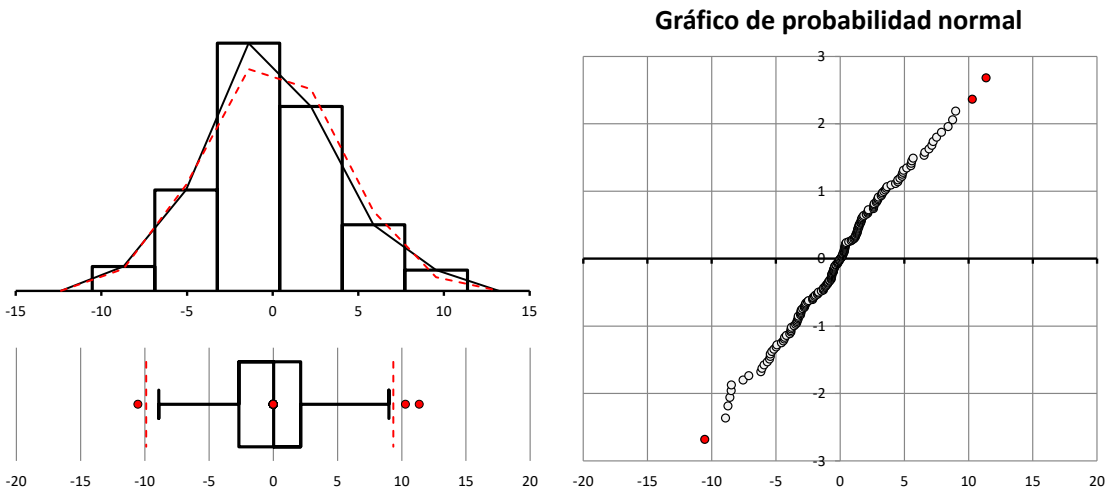


Figura 23. Distribución de los datos de los errores.

- Los errores son independientes entre sí.

  Esta es una característica que no tiene sentido aplicarla a nuestro modelo, ya que los individuos no siguen ningún orden en particular ni siguen una serie temporal.

3.8. Adaptaciones del modelo

3.8.1. Aplicación en medicina forense

Aunque este es el modelo óptimo para los datos que tenemos, debemos tener en cuenta que, en ocasiones, no podremos disponer de todos ellos para hacer una predicción con el modelo.

Este sería el caso de una investigación forense, en la que se investiga el encuentro de un cuerpo desmembrado. De manera que queremos estimar el peso de la persona, pero no disponemos de todas las medidas del modelo que hemos obtenido, para poder hacer esto, deberemos calcular un nuevo modelo.

Partimos de las medidas: 45,2 diámetro biacromial; 28 diámetro biilíaco; 22 profundidad pectoral; 30,6 diámetro pectoral, 14,5 diámetro del codo y hombre. Por medio de *stepwise backward* llegamos a la siguiente conclusión:

PRESS: 3858,3455
R2: 87,19%

Tabla 6. Estimación del Modelo de Regresión Lineal Múltiple.

	Coeficientes	StdErr	p-value
Cte	-71,0099	4,9057	0,000
D. biilíaco	0,9736	0,1650	0,000
Prof. pectoral	1,8506	0,2022	0,000
D. pectoral	1,4647	0,1948	0,000
D. codo	2,7253	0,4002	0,000

Estimamos un modelo en el que las variables que tienen un efecto significativo sobre el peso son el diámetro biilíaco, la profundidad pectoral, el diámetro pectoral y el diámetro del codo. Con un coeficiente de determinación lineal de 87,19%. La recta de regresión es la siguiente:

$$Peso = -71{,}01 + 0{,}97 \times d.biilíaco + 1{,}85 \times prof.pect + 1{,}46 \times diam.pect + 2{,}73 \times d.codo$$

A partir de los datos obtenidos del cadáver podemos predecir que el peso de la víctima será en promedio de 81,11 *kg* para una variabilidad del 87,19%.

3.8.2. Aplicación en antropología

También podemos adaptar el modelo cambiando la variable respuesta.

En las excavaciones arqueológica en la que encontramos un conjunto de huesos, podríamos estar interesados en predecir el sexo de esta muestra. Aún con el esqueleto completo solo dispondríamos de las siguientes medidas: 23,2 diámetro biilíaco; 18,6 de profundidad pectoral; 24,2 de diámetro pectoral; 11 el diámetro del codo y 17,9 el diámetro de la rodilla.

Cabe destacar que el modelo más adecuado para predecir una variable respuesta binaria, como es el sexo, sería la regresión lineal logística. Pero el objetivo de este trabajo es ver todas las utilidades que podemos sacar de la regresión lineal múltiple, por lo que haremos este cálculo con el modelo estudiado.

$$\text{PRESS: } 17{,}275578$$
$$\text{R2: } 64{,}42\%$$

Tabla 7. Estimación del Modelo de Regresión Lineal Múltiple.

	Coeficientes	StdErr	p-value
Cte	-3,1357	0,3302	0,0000
D. biilíaco	-0,0248	0,0110	0,0256
D. pectoral	0,0453	0,0127	0,0005
D. codo	0,2283	0,0248	0,0000

$$sexo = -3{,}135743 - 0{,}024844 \times d.biiliaco + 0{,}045253 \times d.pectoral + 0{,}228301 \times d.codo$$

Con la recta de regresión estimamos una aproximación al sexo de -0,11; lo cual significa que es más probable que se trate de una mujer que de un hombre con una variabilidad del 64,42%.

Al deducir que es una mujer estaríamos haciendo un **análisis discriminante**, usando la regresión lineal para separar en dos categorías.

V. RESULTADOS

A continuación, analizaremos el uso del Modelo de Regresión Lineal en dos artículos de publicaciones científicas.

En este apartado aparecen las tablas del artículo correspondiente, citadas con el título de publicación al inicio de cada sección.

1. ARTÍCULO 1

"Linking childhood emotional abuse and depressive symptoms: The role of emotion dysregulation and interpersonal problems" (30)**.**

Este artículo estudia la relación entre el abuso emocional infantil y los síntomas depresivos.

Comienza describiendo la muestra del estudio, las variables que han utilizado y cómo han sido medidas estas variables.

Las variables explicativas son variables continuas que, por medio de un escore, miden el abuso emocional infantil (CEA), el abuso físico infantil (CPA) y el abuso sexual infantil (CSA). Menciona que hay 3 variables respuesta que son escores que miden síntomas depresivos, desregulación emocional y problemas interpersonales. Se remarca que, al haber 3 variables dependientes diferentes, se realizarán 3 modelos de regresión lineal múltiple distintos, uno por cada una de ellas.

Queda reflejado que se han tenido en cuenta las condiciones del modelo de regresión, pero no se facilita ninguna medida o representación que lo demuestre.

La tabla 3 expone un análisis descriptivo en el que nos facilita la media, el rango y la desviación típica de cada medida, pero no especifica si existe asimetría, la distribución, ni la existencia de datos anómalos. También se indica la correlación entre variables por medio del coeficiente de Pearson.

A continuación, realiza un análisis de regresión lineal entre las variables explicativas y los tres tipos de abuso infantil. Además de reflejar los resultados en la tabla 4 del artículo, se facilita el coeficiente de regresión, el coeficiente de determinación lineal y nos indica si los parámetros son significativos por medio del estadístico t y el $p - value$.

Table 4. Results of univariate linear regression analyses for predicting depressive symptoms, emotion dysregulation, and interpersonal problems in female college students ($N = 276$).

	QIDS-SR		DERS		IIP-32	
	B (SE)	*p*	*B (SE)*	*p*	*B (SE)*	*p*
CEA	0.38 (0.07)	**< .001**	1.14 (0.34)	**.001**	1.13 (0.24)	**< .001**
CPA	0.33 (0.13)	**.015**	0.80 (0.67)	.235	1.23 (0.46)	**.008**
CSA	0.09 (0.09)	.338	0.21 (0.46)	.640	0.49 (0.32)	.128

Note. CEA = childhood emotional abuse; CPA = childhood physical abuse; CSA = childhood sexual abuse; QIDS-SR = Quick Inventory of Depressive Symptomatology; DERS = Difficulties in Emotion Regulation Scale; IIP-32 = Inventory of Interpersonal problems.
P-values indicating significance at the .05 level are shown in bold.

La regresión lineal simple se ha realizado entre las variables explicativas con sus variables respuesta, sin embargo, no se deja constancia de haberlo hecho entre las variables explicativas para estudiar si están relacionadas entre ellas.

Después, aplica la regresión lineal múltiple para estudiar el efecto de varias variables al mismo tiempo. En este caso nos facilita la información del $ratio\ F$, el $p - value$ y el coeficiente de determinación lineal de cada modelo estimado. Además, nos indican que todos los valores de FIV_i son menores de 1,5, indicando que no existe multicolinealidad. Los resultado se exponen en la tabla 5 del artículo.

Table 5. Results of multiple linear regression analyses for predicting depressive symptoms, emotion dysregulation, and interpersonal problems in female college students ($N = 276$).

	QIDS-SR		DERS		IIP-32	
	B [99% CI]	*p*	*B* [99% CI]	*p*	*B* [99% CI]	*p*
CEA	0.42 [0.26, 0.57]	**< .001**	1.33 [0.52–2.14]	**.001**	1.09 [0.54–1.65]	**< .001**
CPA	-0.03 [-0.32, 0.27]	.866	-0.32 [-1.82–1.19]	.681	0.23 [-0.80–1.26]	.662
CSA	-0.11 [-0.29, 0.08]	.254	-0.37 [1.32–0.59]	.452	-0.08 [-0.74–0.58]	.809

Note. CEA = childhood emotional abuse; CPA = childhood physical abuse; CSA = childhood sexual abuse; QIDS-SR = Quick Inventory of Depressive Symptomatology; DERS = Difficulties in Emotion Regulation Scale; IIP-32 = Inventory of Interpersonal problems.
P-values indicating significance at the .05 level are shown in bold.

El abuso emocional infantil (CEA) es el único que sale significativamente asociado a los síntomas depresivos, la desregulación emocional y a los problemas interpersonales.

Aunque no refleja específicamente el estudio de los errores, asumimos que ha sido realizado ya que al inicio el artículo ha mencionado que cumplía con las condiciones del modelo de regresión. Ningún diagrama de dispersión ha sido representado en este artículo.

Finalmente, el estudio expone sus limitaciones. Destaca que los resultados deben ser analizados con precaución, ya que se muestra que existe relación, pero esto no significa que haya causalidad. Además, relata los problemas para realizar la inferencia poblacional de estos resultados.

2. ARTÍCULO 2

"A Short Screener Is Valid for Assessing Mediterranean Diet Adherence among Older Spanish Men and Women" (31)**.**

Este artículo trata de demostrar la validez del score MEDAS como instrumento de medida para el estudio PREDIMED sobre la adherencia a la dieta mediterránea.

El estudio comienza explicando sus objetivos y contexto. Describe las variables que va a analizar y cómo han sido medidas. Se realiza un análisis de las variables que va a utilizar que queda representado en la tabla 1 de la publicación, en la que se muestra el valor de cada quintil seguido de su desviación estándar o intervalo de confianza.

TABLE 1 Characteristics of participants according to quintile distribution of the PREDIMED score derived by the 14-point MEDAS[1]

	Quintile distribution of the PREDIMED SCORE derived from the MEDAS					
	1st	2nd	3rd	4th	5th	*P*-linear trend[2]
n	1875	1383	1380	1315	1193	
Men, %	40.4 (38.2, 42.7)	42.7 (40.1, 45.3)	42.7 (40.1, 45.3)	41.9 (38.2, 43.6)	48.4 (43.6, 51.2)	<0.001
Age, *y*	66.8 ± 6.2	67.1 ± 6.2	67.3 ± 6.1	66.9 ± 6.2	66.7 ± 6.1	0.331
BMI, kg/m^2	30.3 ± 3.9	30.1 ± 3.7	29.9 ± 3.9	29.7 ± 3.8	29.5 ± 3.7	<0.001
Current smokers,[3] %	21.4 (19.4, 23.4)	21.1 (18.8, 23.5)	18.1 (15.7, 20.5)	19.7 (17.3, 22.2)	16.9 (14.3, 19.5)	0.039
Educational level,[4] %	21.6 (19.7, 23.5)	22.1 (19.9, 24.3)	21.2 (19.0, 23.4)	24.8 (22.5, 27.1)	26.3 (23.9, 28.7)	0.005
LTPA,[5] METs · min/d	210 ± 229	209 ± 211	232 ± 236	252 ± 261	264 ± 265	<0.001
Alcohol consumption, *g/d*	6.6 ± 13.1	8.0 ± 14.9	8.7 ± 14.7	9.3 ± 14.5	11.0 ± 14.3	<0.001

[1] Values are mean ± SD or proportion (95% CI).
[2] The polynomial contrast and chi square test were used to determine *P*-linear trend for continuous and categorical variables, respectively.
[3] Current or ex-smokers (up to 1 y).
[4] More than primary school.
[5] Leisure-time physical activity.

Nos indica que, mediante el Modelo de Regresión Múltiple, se quiere estudiar la relación de ciertas variables con el escore MEDAS PREDIMED, por lo que el resultado de este será la variable respuesta. Así, las variables explicativas serán: el índice de masa corporal, el perímetro de la cintura, el nivel de triglicéridos (TG), la glucosa en ayunas, el colesterol total, el HDL-C, la relación TG:HDL-c y la relación colesterol:HDL-c, además del riesgo de enfermedad coronaria a los 10 años.

No se menciona el estudio de multicolinealidad entre las variables respuestas por medio de la regresión lineal simple entre pares de variables, ni se facilitan los valores de FIV_i y $R^2_{i\cdot Otras}$ que indicaría la existencia de multicolinealidad.

Tampoco se reporta ningún diagrama de dispersión ni otro tipo de representación gráfica sobre estos análisis.

Se menciona que, al realizar la regresión lineal múltiple, existe una asociación inversa entre el escore MEDAS y el resto de las variables, excepto con el HDL-C que tiene una relación directa, lo que también podemos observar con el signo de los coeficientes. El resto de los resultados del Modelo de Regresión Múltiple son expresados en la tabla 4 de este artículo.

TABLE 4 Multiple adjusted regression coefficients and 95% CI of the association between the Mediterranean diet score derived by the 14-point MEDAS with cardiovascular risk variables and 10-y coronary risk (n = 4675)[1]

	Mediterranean diet score (1 unit)		
Dependent variable	Regression coefficient	95% CI	*P*
BMI, *kg/m*^{2}	−0.146	−0.191, −0.100	<0.001
WC, *cm*	−0.562	−0.689, −0.435	<0.001
HDL-C,[2] *mmol/L*	0.010	0.005, 0.014	<0.001
TG,[2] *mmol/L*	−0.005	−0.009, −0.002	<0.001
Fasting blood glucose,[2] *mmol/L*	−0.003	−0.005, −0.001	0.004
Total cholesterol:HDL-C[2]	−0.016	−0.031, −0.001	0.038
TG:HDL-C[2]	−0.009	−0.012, −0.005	<0.001
10-y coronary risk[3]	−0.001	−0.002, 0.001	<0.001

[1] Adjusted for sex, age, smoking status, leisure time physical activity, marital status, and educational level.
[2] Log transformed.
[3] Adjusted for sex, age, leisure time physical activity, marital status, and educational level.

En la tabla se reportan los coeficientes de regresión acompañados de su intervalo de confianza y su $p - value$. En el pie de la tabla se indica que para el HDL-C, los TG, la glucosa en ayunas, la relación colesterol total:HDL-C y TG:HDL-C se ha realizado una transformación logarítmica para ajustar el modelo.

Se menciona que el coeficiente de Pearson muestra una correlación moderada de 0,52 junto a su $p - value$ <0,001.

El artículo no menciona el cumplimiento de las condiciones del modelo, tampoco hace referencia al análisis de los errores y no facilita ningún diagrama de dispersión, así como tampoco la ecuación de regresión.

VI. CONCLUSIONES

La importancia de la Regresión Lineal Múltiple en el campo de las ciencias de la salud reside en su capacidad de interpretar, predecir y explicar fenómenos expresables de manera cuantitativa, relacionándolos con una o más variables explicativas.

Es importante notar que la estadística no es simplemente un conjunto de herramientas y técnicas para analizar datos y obtener conclusiones. La estadística permite estudiar relaciones complejas entre variables, en un ambiente de incertidumbre. Permite, en esencia, separar la información del ruido. La utilidad de la estadística depende, en último extremo, de lo adecuado de los modelos que se emplean al enfrentarse a cada problema. Es importante que los métodos estadísticos sean adecuados a los objetivos de investigación que nos planteamos. El trabajo estadístico no empieza después de disponer de los datos. La estadística debe ayudarnos a generar la muestra, ya que los objetivos de nuestro estudio implican unas variables que deben medirse y las herramientas a aplicar dependen de la naturaleza de dichas variables. Además, la variabilidad de las variables implicadas también condiciona el tamaño muestral, ya que este debe ser suficiente para contrarrestarla.

En nuestro caso, hemos detallado en qué casos es adecuado emplear el modelo de regresión lineal múltiple: cuando queremos explicar la variabilidad de una variable respuesta cuantitativa a partir de un conjunto de variables explicativas, siendo las observaciones independientes entre sí. También es útil si lo que queremos es predecir valores de dicha variable respuesta o cuando queremos interpretar el efecto de una o más variables explicativas sobre la variabilidad de la citada variable respuesta.

En este trabajo hemos explicado la utilidad de la regresión lineal múltiple describiendo sus elementos y las condiciones en que puede aplicarse. Para ilustrar la utilidad hemos recurrido a un ejemplo, detallando cada uno de los pasos que previamente habíamos explicado en los apartados anteriores, especialmente la verificación de las condiciones teóricas en que se basa el modelo, ya que, de no verificarse dichas condiciones, podríamos obtener conclusiones erróneas del análisis posterior.

Terminamos el trabajo mostrando el uso que se hace en la literatura médica del modelo de regresión lineal múltiple, recurriendo a dos artículos recientes. En ambos casos constatamos serios defectos en el reporte de los análisis y las conclusiones.

VII. ANEXO

1. Base de datos

	Peso	Edad	Altura	Hombre	D. biacromial	D. biiliaco	D. bitrocantéreo	Prof. Pecotral	D. pectoral	D. codo	D. muñeca	D. rodilla	D. tobillo
1	90	20	192	1	47,4	29,6	35,7	20,8	31,4	16,1	11,3	21,5	15,4
2	94,1	20	177	1	43,6	33,1	33,5	21,6	33,1	15,6	12	20,7	16,5
3	57	20	163	1	38,3	25,2	30,2	17	26,4	13,2	10,4	18,8	13
4	65,9	20	169	1	40,8	27,1	29,4	17,8	29,4	13,3	10,4	18,5	12,8
5	55,5	20	169	1	40,1	19,4	28	17,1	26,8	13	10,6	16,9	12,6
6	58,4	20	157	1	38,7	25,2	28,8	19,1	25,6	13	10,2	17,9	13,5
7	69,1	20	170	1	41,1	23	29,4	21,8	30,6	15	10,8	19,3	14,5
8	77,8	20	181	1	43,8	28	33,3	20	32	15	11,5	20,4	14,4
9	74,8	21	182	1	43,3	27	31,5	19,6	31,3	14	11,5	18,8	13,9
10	65,6	21	174	1	42,9	26	31,5	17,7	28	13,1	10,4	18,8	14,1
11	81,6	21	184	1	42	28	34	22,5	28	15,6	12	21,1	15
12	82,8	21	180	1	45,4	31,8	35,2	20,2	32,3	14,6	10,5	20,2	15,3
13	87,8	21	187	1	40,3	29,4	32,9	23,7	31,5	14,6	11,3	19,8	15,2
14	74,8	21	178	1	39,9	28,3	32	18,3	31,4	13,5	11,4	18,9	14,4
15	80,9	21	176	1	42,2	30,1	31,4	21,2	29,7	14	11,6	21,6	14,1
16	72,5	21	177	1	43,2	26,3	30,5	19,7	30,6	14,4	12,3	20,2	13,6
17	75,7	21	182	1	43,9	27	33,5	22,3	31	13,2	10,4	19,1	13,1
18	65	21	165	1	35,6	28,5	29,4	17,7	25,2	14	10,8	19,1	15
19	79,5	21	178	1	42,1	28,5	33,1	20,2	30,6	15,6	12,2	19,7	15,6
20	75	21	183	1	41,9	27,8	33,3	19	28,7	15,1	11,3	19,2	14,9
21	93,2	21	188	1	41,3	29,7	34,7	22,8	32	15,5	11,2	18,4	15,6
22	83,6	22	178	1	43,7	29	31,3	20,5	29,7	15	11,7	20,9	16
23	79,6	22	184	1	45	27	33,2	21,7	30,6	13,7	11,1	20,7	14
24	78,8	22	187	1	42,5	29,9	34	21,5	29,4	15,2	11,6	20,7	14,9
25	70	22	172	1	43	26,5	30,3	19,3	30	14,8	11,2	19,7	14,7
26	85,9	22	176	1	43,6	29,3	34,4	20,2	32,6	15,4	10,9	20,7	15,5
27	78,8	22	176	1	42	27,5	30,7	21,3	32	13,1	11,1	19,2	13,9
28	66,2	22	173	1	42,3	26,4	31,2	18	30,9	14,6	10,8	18,6	13,8
29	67,2	22	182	1	41,1	27,8	31,4	19	31,5	14,5	11,9	18,5	13
30	68,6	22	178	1	36,2	27,5	30,4	18,7	28	13,6	10,8	19	15,4
31	68,9	22	177	1	43,8	29,5	31,2	18,2	29,5	13,1	10,3	19,1	13,2
32	73,2	22	177	1	42	28	33	18,1	28,4	14,3	11,1	20,2	15,2
33	108,6	22	191	1	43,6	28	32,4	27,5	33,5	14,6	11,7	21,4	15,1
34	95,9	22	193	1	45,2	32,2	36	22,5	33,5	15,8	11,3	20,5	14,8
35	84,1	22	185	1	40,1	26,4	32	21,4	32,6	14,8	12	20	15,2
36	61,4	22	178	1	41,7	26,8	32	19,7	27,8	14	10,5	18,4	14,6
37	63,2	23	165	1	39,4	25,5	30,2	17,6	27,7	13	10,2	18,9	13,2

38	74,8	23	184	1	44,2	30,3	34,7	19,4	30	14,9	11	19,1	15,8
39	62	23	175	1	43,5	26,5	32,1	15,5	27,5	14,1	11,2	18,9	13,2
40	71,8	23	175	1	43,7	28,5	33,5	16,9	30,8	14	11,8	20,6	15,1
41	72,6	23	187	1	44,3	29,9	34	18,4	28,2	13,9	11,2	20,9	15
42	76,6	23	180	1	40,3	29	33	20,1	30,3	13,4	10,4	19,4	14,5
43	72,4	23	188	1	44	31	35,3	19,2	31	15,2	11,4	21,2	15,1
44	84,1	23	189	1	41,6	32	35,3	23,6	27	15,5	11,3	20,9	15
45	69,1	23	173	1	41	25,1	31,9	20,8	27,9	13,6	10,8	18,8	12,9
46	84,7	23	171	1	41,7	27,1	32,6	21,6	28	14,1	11,5	19,7	13,8
47	84,1	23	188	1	39,4	26,1	34,4	20,4	27,3	15,1	10,6	20	15,3
48	75,6	23	170	1	39,4	28,3	30,6	20,2	28,7	15	11,5	18,4	14,4
49	86,2	23	174	1	38,5	26,1	30,8	20,6	30,8	15,1	11,4	19,8	14,2
50	66,1	23	172	1	41,3	27,1	32,4	17,5	27,6	14,1	10,8	20,2	15,5
51	72	23	182	1	40,3	29,5	33,3	18,4	26,2	14	11	19,4	14,8
52	68,2	23	177	1	41,4	26,4	32,3	18,6	31,3	14,9	11,5	18,9	14,6
53	56,8	23	171	1	34,1	28,1	30,1	21,8	25,8	12,9	9,9	18,6	12,3
54	72,7	23	170	1	41,7	28	32,9	19,4	29,7	14,6	11	19,5	15,3
55	86,4	23	175	1	41,2	29,8	32,2	22,4	29,8	15,6	11,2	19,6	14,8
56	71	23	174	1	41	26,8	32,2	21,9	28,6	14,9	10,6	17,8	14
57	89,6	24	178	1	46	30,1	34,5	20,2	31,1	16,4	13,3	22,2	14,9
58	76,4	24	180	1	40,5	28,3	33,4	19,2	28,8	14,6	11,1	20,8	14,5
59	86,4	24	176	1	42,7	29,9	35	21,8	32,8	14,3	11,2	19,8	14,1
60	61,3	24	170	1	38,8	27,2	31,6	18,5	25,5	13,4	10,8	19	14
61	80,1	24	184	1	42,1	27,5	32,4	18,2	28	16,2	12	21	16,4
62	73,4	24	173	1	37,8	27,1	31,5	18,5	27,3	14,6	10,8	19,5	14,9
63	55,2	24	164	1	37,6	26,6	29,9	17,3	25,6	12,8	10	17	13
64	74,5	24	182	1	43,5	28,8	34	20,6	29	14,3	10,5	19,8	14,2
65	90,9	24	197	1	44,1	29,2	35,3	23,6	30,9	15,8	12,5	20,2	15,2
66	84,5	24	180	1	44,2	27,9	32	21,6	32,9	14,3	11	21,1	14,9
67	67,7	24	175	1	40,3	27,3	30,4	20,4	29	15	11,3	19,1	14,6
68	65,9	24	170	1	41,1	29,2	31,5	19,7	29,9	14,8	11	18	15
69	81,8	24	178	1	39,6	28,7	32	20,2	32,9	14,3	11,5	19,6	15,1
70	73,9	24	174	1	42,5	25,2	30,6	20,9	30,4	15,3	11,4	18,9	13,8
71	84,1	24	178	1	41,9	28	33,1	26,4	29,9	15,6	11,5	21,2	15,9
72	70	25	176	1	41	23,6	30,2	22,9	28	14,3	11,2	18,2	14
73	88,7	25	183	1	42,5	27,8	33,5	20,6	30,2	15,9	12,8	22,4	16,3
74	74,9	25	172	1	38,9	24,9	28,7	19,7	26,8	14,2	10,2	18	14,4
75	87,1	25	182	1	40,3	28	32	20,9	31,7	14,8	10,6	19,4	15
76	72,7	25	181	1	43	26,5	31,6	20,6	29,5	13,4	10,4	18,8	13,6
77	75,5	25	182	1	43,3	28,2	33	19,4	31,6	13,8	11,1	17,8	13,2
78	86,4	25	185	1	38,1	30,1	33,2	21,6	31,3	14,2	12,3	19,2	15,2
79	64,1	25	178	1	38	27,1	28,3	18,2	25,9	13,8	11	18,9	14,8
80	65	25	165	1	37,6	24,4	28,3	17,7	24,7	12,9	10,8	18	14,3
81	58	25	178	1	41,9	25,4	30,2	14,4	26,8	12,6	9,8	18,8	13,6
82	68,6	25	168	1	39,8	25,9	31,3	19,4	29,2	14,3	11,2	18,7	14,3
83	69,1	25	168	1	40,9	29	32,2	20,2	29,2	13,8	10,4	18,4	14,4

84	91,4	25	188	1	45	29	33,3	25,4	30,8	15,4	11	18,8	15
85	101,4	25	187	1	42,5	29,4	34,2	23,5	34,7	15,1	11,8	21,8	15,8
86	73,2	25	180	1	38,9	25,6	32,9	21,1	29	15,6	10,6	20,2	14,8
87	55	20	162	0	35,6	27,9	32	15,6	23,9	11,9	9,6	18,2	12,7
88	54,5	20	168	0	34,9	28	30,8	15,6	24,2	11,2	9,2	17,9	13,2
89	60,3	20	161	0	36,2	27,1	29,9	16	24,2	11,8	9,2	17,3	12,6
90	56,2	20	165	0	36	25,2	30,8	16,5	24,2	12,6	9,2	17,6	13,2
91	46,5	20	152	0	34,4	23,5	26,8	16,5	24,4	11,7	9,2	16,8	12,4
92	54,3	20	158	0	34,7	27,8	32,6	16,5	24,2	12,6	9,8	17,6	12,6
93	73,1	20	166	0	40,7	29	35,3	17,7	28,5	13,8	10,8	19,8	14,3
94	58,2	20	163	0	36,5	26	30,7	15,9	26,8	12,1	10,3	17	12,4
95	50	20	160	0	34,5	25,6	29,2	18	24,2	10,1	9	16,3	11,5
96	69,2	20	173	0	36,7	28,2	33	17,5	25,6	12,2	10,1	17,8	13,5
97	57,8	20	163	0	37,7	26,2	30,8	17,8	25,2	12,4	10,2	17,2	11,8
98	69	20	169	0	35	25,6	32,5	17,9	27	12,6	9,8	18,6	12,9
99	65,2	20	169	0	37,7	25	30,4	16,6	29,3	12,4	10,2	18,6	12,7
100	55,2	20	161	0	36,7	23,5	29,5	17,2	28,3	11,5	9,5	16,1	9,9
101	56,4	20	163	0	38,8	26,5	29,8	17,4	28,2	12	9,4	17,7	12,6
102	54,4	20	161	0	36	25,2	28,4	15,4	26,9	11,2	9,9	16,8	12,1
103	58	20	163	0	35,9	25,1	28,2	18,1	26,1	11,6	8,5	16,8	11
104	59,8	20	167	0	35	27,1	30,1	19,5	24,9	12,4	9,9	18,2	12,5
105	43,2	20	160	0	34,9	26,2	28,5	14,6	23,5	10,4	8,7	16	11,1
106	54,6	20	156	0	36,3	24,9	31	16,6	27,7	12,5	10	18,1	13
107	64,5	20	171	0	34,5	29,2	31,9	18,3	23,6	11	9,3	18,7	12,4
108	55,5	20	159	0	38,9	25,2	31,5	17,3	24,2	12,9	9,8	18	12,8
109	57,3	20	161	0	35,1	25,6	29,9	16,8	24,9	12,6	9,6	18,2	12,5
110	60	20	160	0	35,8	29,4	32	18	25,9	11,8	9,8	18	12,7
111	45,8	21	152	0	34,8	26,3	29,9	16	22,9	11,3	9,2	16,5	11,8
112	53,6	21	156	0	35,5	28,2	31	18,2	26,2	11,5	9,1	17,2	12,4
113	56,6	21	166	0	35,8	26,4	32,2	17,3	24,9	12,7	10,4	18,9	14
114	60,7	21	180	0	33,8	29,7	32,4	17,3	25,4	13,4	9,6	18,4	13,4
115	60	21	166	0	37,8	26,6	33,1	16,5	27,1	12,4	9,8	18,3	13
116	53,2	21	166	0	35,6	25,7	29,1	15,5	26	11,5	9,5	17,8	12,1
117	63,4	21	161	0	34,2	28	31	20,4	25,6	12,4	9,7	18,5	13
118	45	21	160	0	32,8	26,2	26,3	16,1	23,6	10,6	9,2	16,4	11,7
119	66,8	21	175	0	38,5	30,4	35,1	18,2	27,3	13,8	10,7	19	14,4
120	70,5	21	173	0	37,1	29,2	32	18	26,1	13,2	10,2	19,2	14,4
121	59,1	21	170	0	35,8	29	30,6	18	24,2	12,4	10,5	17,9	13,4
122	55,5	21	164	0	35,3	28,7	30,4	17,7	25,6	12,4	9,8	17,3	13,6
123	72,8	22	170	0	37,5	29,2	35,1	19,6	27,5	13,1	10,2	20	13,1
124	69,8	22	166	0	37,4	30,2	33,2	18,8	26,6	13,3	10,7	19,8	13,8
125	51,6	22	161	0	37,6	25	31,3	16,2	24,9	11,2	9,2	17	12,3
126	83	22	177	0	39,1	29,7	33,7	20,1	29,5	13,2	10	20,3	13,6
127	50	22	160	0	35	26,5	31,6	18,3	23,7	11,5	8,6	16,8	12,2
128	72,9	22	170	0	36,2	28,7	31,3	19,2	27,1	12,8	10,5	18	14,4
129	59,3	22	152	0	36,9	26,4	32,4	16,6	26	12,6	10,5	18,9	14,4

130	53,4	22	168	0	35,8	27,7	28,9	16,4	25,6	11,5	9,1	17,1	12,3
131	72	22	163	0	37,9	29	32,9	18,6	27	12,7	10,6	18,2	12,1
132	53,2	22	160	0	33,6	24,3	27,9	18,2	24,7	11	8,4	16,6	11,1
133	51,8	22	159	0	36	27,1	31,4	18,5	24,6	11,9	9,4	15,9	11,9
134	55,9	22	161	0	34,9	26,1	29	16,5	25,6	12	9,2	17,3	13,2
135	67,3	22	170	0	37,6	25,2	29	19,2	27,1	14	11	18,9	13,2
136	59	23	170	0	37	28	32	15,1	25,7	12,5	10	17,2	13,2
137	49,2	23	168	0	36,2	27,4	29,5	14,6	23,9	11,2	9,6	16,7	12,6
138	68,8	23	168	0	37,6	30	33,9	19,1	28,8	13,4	10,5	19,2	13,2
139	59,8	23	173	0	39,1	27,5	34	17,2	25,6	12,6	10	18	12,7
140	67,8	23	171	0	40,5	29,6	34,9	17,4	27,5	12,6	10,4	18,8	13,1
141	62,3	23	169	0	37,6	27,3	32	18	25,6	12,4	9,4	18,8	13,6
142	60,3	23	165	0	40,3	29,2	32,9	16,5	26,4	13	9,4	18,6	12,9
143	71,6	23	164	0	39,9	20	32,4	17,6	31,9	13,6	10,4	18,5	13,4
144	59,8	23	163	0	38,7	23,5	33	16,4	27,6	12,6	9,9	17,5	12,6
145	55,9	23	170	0	36,5	27	31,6	16,6	24,4	11,9	9,7	18	12,5
146	59,8	23	164	0	38,8	26,3	31,5	19	25,9	11,6	9,9	18,8	13,1
147	53,8	23	164	0	35,1	23,4	30,6	16,2	26,8	12,1	9,4	16,1	12,2
148	54,8	23	160	0	35,2	24,4	29,8	15,1	26	13	9,2	17,3	12,2
149	46,4	23	158	0	36,4	27,7	29,8	17	25	11,2	9	16,9	12,2
150	60	23	178	0	36,7	29	32	19,7	25,6	12,4	10	18,6	13
151	59	23	160	0	34,7	24,7	32	17,7	25,2	12,8	10,4	18,4	13,8
152	62,3	23	175	0	37,4	29	32,9	17,7	23,5	12,7	10	18,2	13,8
153	60,9	23	161	0	36,5	26,6	30,8	16	24,9	12,6	10,7	17	13
154	53,6	23	162	0	36,7	22,1	28,3	16,7	27,5	12,3	10,1	16,8	12,4
155	57,2	24	167	0	37	27,9	30,8	15,1	26,8	13,2	8,9	18,1	12,3
156	58,5	24	164	0	37,6	28,7	30,8	19	25,2	13,1	11,5	18,6	13,6
157	64,8	24	170	0	37,5	27,4	31,2	18,9	26,5	14	11	19,1	13,5
158	45,7	24	159	0	33,9	24,3	29,5	15,8	26	11,6	9,4	16,3	11,2
159	66,8	24	180	0	39,3	24,5	32,3	21,4	29,5	13,4	10,9	18,2	13
160	59,4	24	163	0	37,2	24,4	29,4	18,1	27,3	12,3	9,9	17,1	12,2
161	69,5	24	173	0	37,8	28	32	16,8	28,3	12,7	9,8	18,1	13,2
162	56,8	24	165	0	35,8	26,6	31,1	18,5	24,7	12,4	9,6	17,9	13,8
163	75,9	24	173	0	38,5	26,6	34	22,3	26,6	13,8	10,7	20	15,5
164	65,5	24	175	0	38,3	30,8	35,6	17	26,8	13,6	10,4	19	13,6
165	73,6	24	177	0	36,2	31,1	34,7	21,1	26,8	12,9	10,4	18,6	14
166	54,5	24	170	0	37,6	26,8	31,1	16,5	24	11,8	10,2	18,4	12,8
167	63,6	24	175	0	38,7	29,7	32	15,8	26,4	12,8	10,5	18,4	13,8
168	54,2	25	171	0	36	25,6	31,5	15,4	25,5	12,8	9,7	17,6	13,2
169	49,8	25	147	0	35,6	25,8	32	16,2	25,7	11,5	9	17,2	11,8
170	50,6	25	160	0	36,3	27,5	31	15,4	24	11,1	9,4	16,4	12,3
171	63	25	157	0	36,6	27,9	31,8	19,3	24,9	12,3	9,5	18,6	13
172	62,5	25	173	0	39,5	30	31,7	17,3	27,3	12,8	9,2	18,1	12,4
173	54,4	25	160	0	36,6	28,2	32	14,3	26	11,8	9,9	18	12,7
174	62	25	163	0	35	27	33	19,3	25	11,5	9,9	17,1	13
175	54,8	25	161	0	36,2	30,2	31,2	16,8	27,7	11,5	9,4	17,4	13,1

176	52,2	25	163	0	35,2	24,1	26,9	17,9	23,6	12	9,5	17,5	11,5
177	53,6	25	161	0	36,1	26,8	31	16,2	24	11	8,9	17,8	12,2
178	56,2	25	172	0	35,7	27	29,8	16,5	26,2	12,1	9,6	18	13,6
179	51,8	25	164	0	35,7	28,3	29,6	17,4	23,3	10,7	8,7	16,8	11,3
180	63	25	168	0	37	26,9	32,4	16,6	25,4	12,9	10	18,1	13,1
181	55	25	169	0	37,8	29,4	32,2	15,8	27,5	11,8	9,8	18,2	13,8
182	47,3	25	152	0	35,1	27,8	30,6	16,3	26,8	11,6	9,6	17,2	12
183	63,6	25	170	0	35,3	26,8	31,7	18,5	23,3	11,5	9,2	17,8	12,6
184	66,4	25	163	0	35,1	29,4	33,1	17,7	27,3	12,8	10,3	18,7	13,6
185	61,4	25	168	0	40,9	28,7	29,2	18,2	27,3	12,9	9,6	16,8	13,4
186	63,6	25	175	0	35,8	31,5	32,6	17,5	25,2	12,9	9,8	17,9	13,4

VIII. BIBLIOGRAFÍA

1. Pejlare J, Bråting K. Writing the History of Mathematics: Interpretations of the Mathematics of the Past and Its Relation to the Mathematics of Today. In: Sriraman B, editor. Handbook of the Mathematics of the Arts and Sciences [Internet]. Cham: Springer International Publishing; 2019 [cited 2021 Feb 17]. p. 1–26. Available from: http://link.springer.com/10.1007/978-3-319-70658-0_63-1

2. Mayer D. Essential evidence-based medicine. New York: Cambridge University Press; 2004. 381 p.

3. Bingham P, Verlander NQ, Cheal MJ. John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. Public Health. 2004 Sep;118(6):387–94.

4. Gill CJ, Gill GC. Nightingale in Scutari: Her Legacy Reexamined. Clin Infect Dis. 2005 Jun 15;40(12):1799–805.

5. Horwitz RI, Charlson ME, Singer BH. Medicine based evidence and personalized care of patients. Eur J Clin Invest. 2018 Jul;48(7):e12945.

6. Glasziou P, Aronson J. A brief history of clinical evidence updates and bibliographic databases. J R Soc Med. 2018 Aug;111(8):292–301.

7. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007 Aug 21;147(4):224–33.

8. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med. 2010 Sep 21;7(9):e1000326.

9. Else H. How a torrent of COVID science changed research publishing — in seven charts. Nature. 2020 Dec 24;588(7839):553–553.

10. COVID-19 Report: Publications, Clinical Trials, Funding [Internet]. Dimensions. 2021. Available from: https://reports.dimensions.ai/covid-19/

11. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction Models — Development, Evaluation, and Clinical Application. N Engl J Med. 2020 Apr 23;382(17):1583–6.

12. Chen L. Overview of clinical prediction models. Ann Transl Med. 2020 Feb;8(4):71–71.

13. Steyerberg EW. Clinical Prediction Models [Internet]. New York, NY: Springer New York; 2009 [cited 2021 May 7]. (Statistics for Biology and Health). Available from: http://link.springer.com/10.1007/978-0-387-77244-8

14. Rafael Romero Villafranca, Luisa Zúnica Ramajo. Estadística Diseño de experimentos Modelos de regresión.

15. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. Eleventh edition. Hoboken, NJ: Wiley; 2019. 1 p.

16. Campbell MJ, editor. Statistics at Square Two [Internet]. Oxford, UK: Blackwell Publishing Ltd; 2006 [cited 2021 Apr 14]. Available from: http://doi.wiley.com/10.1002/9780470755839

17. Edwards R. Statistics at Square One. J Epidemiol Community Health. 1997 Feb 1;51(1):104–104.

18. Bangdiwala SI. Regression: multiple linear. Int J Inj Contr Saf Promot. 2018 Apr 3;25(2):232–6.

19. Hess AS, Hess JR. Linear regression and correlation. Transfusion (Paris). 2017 Jan;57(1):9–11.

20. Shi R, Conrad SA. Correlation and regression analysis. Ann Allergy Asthma Immunol. 2009 Oct;103(4):S35–41.

21. Sedgwick P. Multiple regression. BMJ. 2013 Jul 5;347(jul05 2):f4373–f4373.

22. Schmidt AF, Finan C. Linear regression and the normality assumption. J Clin Epidemiol. 2018 Jun;98:146–51.

23. Marill KA. *Advanced Statistics:* Linear Regression, Part II: Multiple Linear Regression. Acad Emerg Med. 2004 Jan;11(1):94–102.

24. Clauser, C., Tucker, P., McConville, J., Churchill, E., Laubach, L., Reardon, J. Anthropometry of Air Force Women. Aeroesp Med Res Lab Wright-Patterson Air Force Base OH. 1972;AMRL-TR-70-5.

25. White, R. M., Churchill, E. The Body Size of Soldiers: U.S. Army Anthropometry - 1966. US Army Natick Lab Natick MA. 1971;75-51-CE (CPLSEL-94).

26. Heinz G, Peterson LJ, Johnson RW, Kerk CJ. Exploring Relationships in Body Dimensions. J Stat Educ. 2003 Jan;11(2):7.

27. Behnke AR, Wilmore JH. Evaluation and regulation of body build and composition. Englewood Cliffs, N.J: Prentice-Hall; 1974. 236 p. (International research monograph series in physical education).

28. Lohman TG, Roche AF, Martorell R, editors. Anthropometric standardization reference manual. Champaign, IL: Human Kinetics Books; 1988. 177 p.

29. Blaak E. Gender differences in fat metabolism: Curr Opin Clin Nutr Metab Care. 2001 Nov;4(6):499–502.

30. Christ C, de Waal MM, Dekker JJM, van Kuijk I, van Schaik DJF, Kikkert MJ, et al. Linking childhood emotional abuse and depressive symptoms: The role of emotion dysregulation and interpersonal problems. PloS One. 2019;14(2):e0211882.

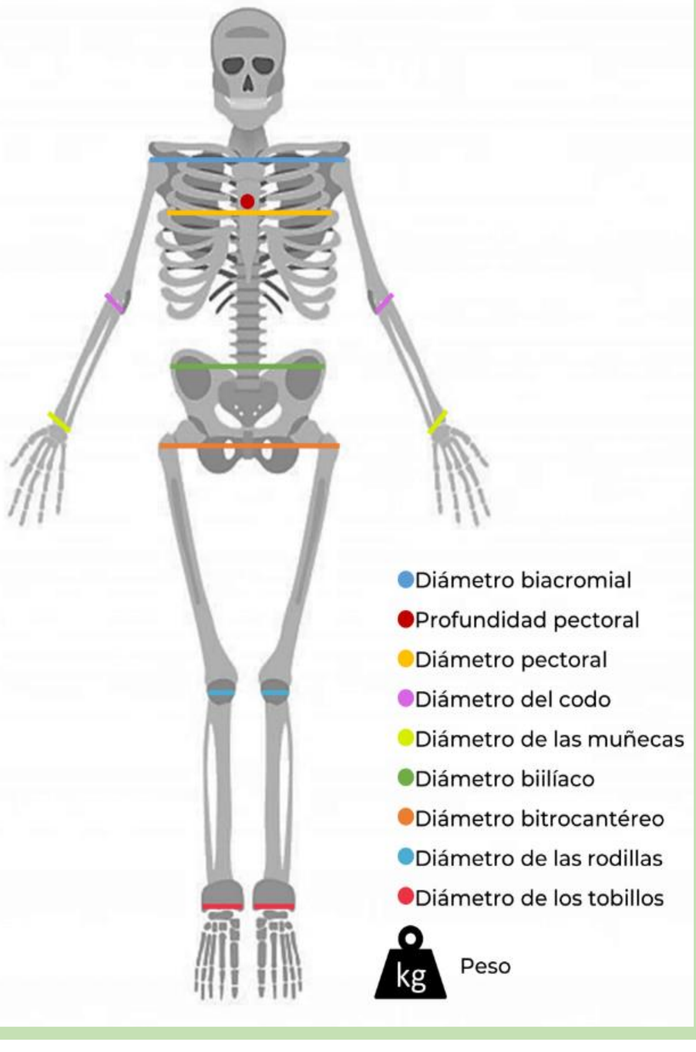
31. Schröder H, Fitó M, Estruch R, Martínez-González MA, Corella D, Salas-Salvadó J, et al. A Short Screener Is Valid for Assessing Mediterranean Diet Adherence among Older Spanish Men and Women. J Nutr. 2011 Jun 1;141(6):1140–5.

UTILIDAD DE LA REGRESIÓN LINEAL MÚLTIPLE EN ESTUDIOS DE CIENCIAS DE LA SALUD

¿Por qué unos deportistas pesan más que otros?

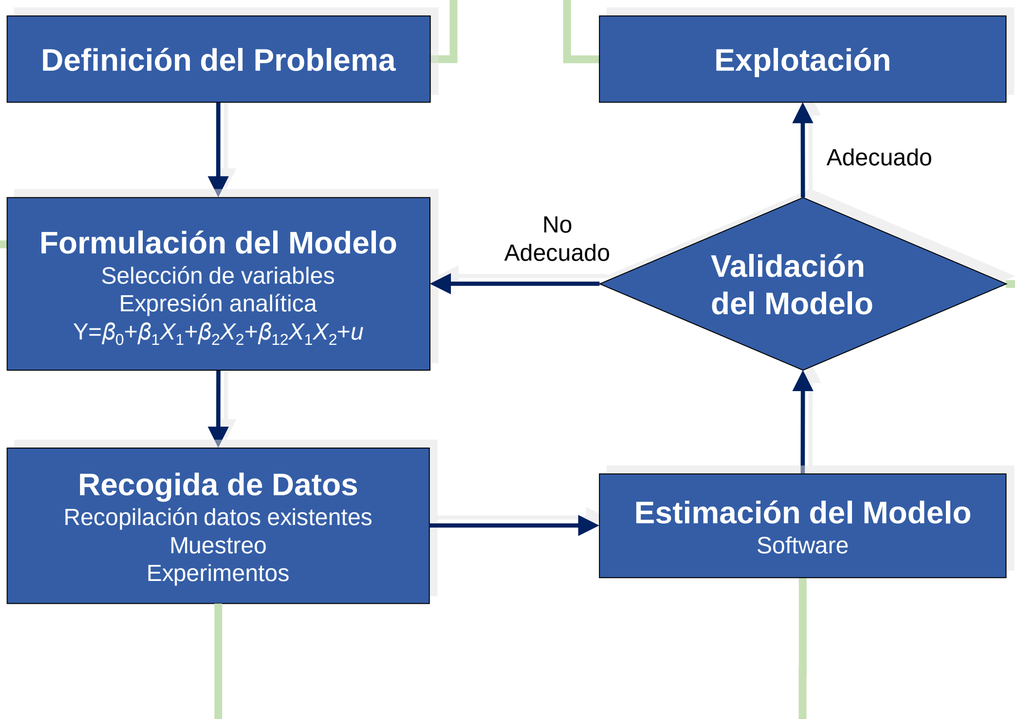
Población: Deportistas de alto rendimiento con edad entre 20 y 25 años.

Muestra: 186 socios de un centro de alto rendimiento (86 hombres y 100 mujeres)



Recogida de datos

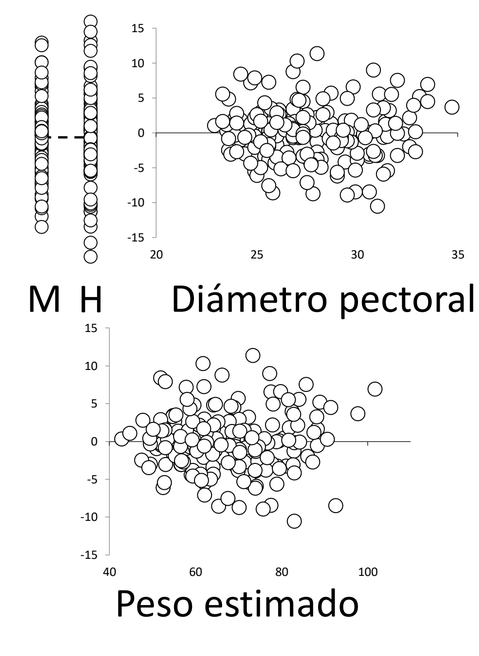
Hombre	Peso	Altura	D. biacromial	D. biilíaco	Prof. Pectoral	…
1	90,0	192,0	47,4	29,6	20,8	…
1	94,1	177,2	43,6	33,1	21,6	…
1	57,0	163,0	38,3	25,2	17,0	…
…	…	…	…	…	…	…



Explotación: Interpretar, Explicar, Predecir

Interpretar: para unas mismas medidas antropométricas, las mujeres pesan, en promedio, 3,0 Kg más que los hombres.
Explicar: el modelo explica el 90,3% de la variabilidad en el peso.
Predecir: el peso medio de las mujeres de 164 cm de altura, con una profundidad pectoral de 17 cm, un diámetro pectoral de 31cm, un diámetro de las rodillas de 17 cm y un diámetro de los tobillos de 12 cm; será de 62,2 kg.

Residuos



Modelo final estimado

f-Stats

	Coef	StdErr	t-stat	sign	IC 95%		FIV_i	$R^2_{i\cdot Otras}$
Cte	-106,69	7,17	14,87	0,000	-120,84	-92,53		
Altura	0,2	0,05	4,58	0,000	0,14	0,34	2,90	65,51%
Hombre	-3,02	0,92	3,29	0,001	-4,84	-1,21	2,56	60,88%
Prof. pectoral	1,70	0,18	9,44	0,000	1,35	2,06	2,14	53,22%
D. pectoral	1,44	0,17	8,33	0,000	1,10	1,78	2,55	60,77%
D. rodilla	2,04	0,40	5,09	0,000	1,25	2,83	3,50	71,46%
D. tobillo	1,78	0,44	4,03	0,000	0,91	2,65	3,55	71,82%

OBJETIVOS

Subrayar la importancia de la Regresión Lineal Múltiple en estudios de ciencias de la salud.

RESULTADOS

Se muestra el uso que se hace en la literatura médica del modelo de regresión lineal múltiple, recurriendo a dos artículos recientes. En ambos casos constatamos serios defectos en el reporte de los análisis y las conclusiones.

CONCLUSIONES

La importancia de la Regresión Lineal Múltiple en el campo de las ciencias de la salud reside en su capacidad de interpretar, predecir y explicar fenómenos expresables de manera cuantitativa, relacionándolos con una o más variables explicativas.

Presentado por: Carmen Zapata Carratalá **Tutor**: Francisco Javier Arteaga Moreno

Universidad Católica de Valencia San Vicente Mártir

Facultad de Medicina y Odontología