MODELOS MATEMÁTICOS PARA EL ESTUDIO DE PROBLEMAS MUSCULARES

UNIVERSIDAD
COMPLUTENSE
MADRID

TRABAJO DE FIN DE GRADO

GRADO EN INGENIERIA MATEMÁTICA

Curso 2019/2020

Víctor Navas Iglesias

Tutora: Ana Carpio Rodríguez

Departamento de Análisis Matemático y Matemática Aplicada (AMMA)

Facultad de Ciencias Matemáticas

Madrid, 29 de septiembre de 2020

RESUMEN

En este trabajo se muestran distintos modelos para el análisis de un estudio matemático sobre lesiones musculares debidas a esfuerzos físicos realizados en el deporte. En dicho trabajo, se intenta dar una perspectiva sobre las diversas técnicas matemáticas para todo aquel que intente realizar cualquier investigación, en particular lesiones deportivas, conozca qué técnica se adecua mejor para su investigación.

Primero, se presenta una introducción formal del trabajo y sus motivos de realización junto a un contexto de los objetivos y metas. Después, se pasa a explicar diversos factores que influyen en lesiones musculares dentro del mundo del deporte como, por ejemplo, la duración de los entrenamientos u horas de competición. A continuación, se desarrolla una teoría simple, pero concisa, de los distintos tipos de regresiones que se disponen para realizar cualquier experimento y cómo se deben elegir en función de los resultados que se quiera obtener. Para así, en último lugar, aplicar dichas técnicas a bases de datos con características de deportistas, utilizando el entorno de programación estadístico RStudio empleado también para generar todas las gráficas del trabajo y sacar conclusiones.

PALABRAS CLAVE

Base de datos. Especificidad. Lesiones deportivas. Matriz de confusión. Regresión lineal. Regresión logística. Sensibilidad. Variable categórica. Variable dicotómica. Variable dummy. Variable numérica. Variable predictora.

ABSTRACT

In this academic work we revise different models for the mathematical study of muscular injuries due to physical efforts in sports. We try to give an outlook about miscellaneous mathematical techniques in case somebody wants to make a research, in particular on sports injuries.

Firstly, we give a formal introduction of the work and motivation with the objective and context. After that we discuss several factors that influence muscular injuries in sports. Then, we present a simple and concise theory about the different regression models that can be used in an experiment and how they should be used depending the results you want to achieve. Finally, these models are applied in a database with sports features using the statistical software RStudio, also used to generate all graphs, in order to achieve several conclusions.

KEYWORDS

Database. Specificity. Sport injuries. Confusion matrix. Linear regression. Logistic regression. Sensitivity. Categorical variable. Dichotomous variable. Dummy variable. Numeric variable. Predictor variable.

Tabla de contenido

RESUMEN 2
PALABRAS CLAVE 2
ABSTRACT 3
KEYWORDS 3
1 INTRODUCCIÓN 5
2 FACTORES QUE INFLUYEN EN LESIONES DEPORTIVAS 6
3 MODELOS MATEMÁTICOS 7
3.1 Regresión lineal simple 12
3.2 Regresión multilineal 18
3.3 Regresión logística 21
4 ANÁLISIS DE LOS MODELOS 26
4.1 Hipótesis de Gauss – Márkov 27
4.2 Bondad de ajuste 28
4.3 Variables dummy 28
4.4 Matriz de confusión: sensibilidad y especificidad. 30
4.5 Aplicación de los modelos 32
4.5.1 ¿Se puede predecir si un paciente tendrá más de 10 días de lesión según su edad, su altura, su IMC, su peso y si ha tenido lesiones previas? 33
4.5.2 ¿Depende de las variables edad, peso, altura e IMC que una lesión sea muscular o no? 33
4.5.3 Sabiendo el tipo de lesión, mecanismo e IMC, ¿se puede saber si una lesión es grave entendiendo por grave una lesión de 2 o más semanas? 34
5 CONCLUSIONES 34
BIBLIOGRAFÍA 36
ANEXOS 37
Código 37
Gráficas 44

1 INTRODUCCIÓN

Una fuente de investigación en auge en nuestros tiempos son las lesiones físicas de deportistas, debido al interés creciente individual y colectivo acerca de la práctica o el seguimiento de actividades deportivas profesionales o no. Sería muy conveniente tanto para deportistas como para equipos o instituciones como colegios, clubs deportivos o gimnasios saber cómo prevenir la aparición de lesiones musculares, conocer qué ejercicios de recuperación son los más adecuados para el tratamiento de ellas, estimar la duración que va a tener la lesión o incluso predecir cuantas horas debe el atleta entrenar para evitarlas. Para ello, se han realizado numerosos estudios clínicos desde radiografías hasta simulaciones reales con el objetivo de entender su causa, pero una herramienta que ya se ha explorado y, además de ser objeto de este trabajo, puede ayudar a esclarecer con mayor eficiencia y precisión esta problemática son las técnicas de regresión lineal y logística. Actualmente, el procedimiento usado para realizar los estudios es medir las características físicas de los pacientes y, posteriormente, realizar un análisis de la información como se hace en los siguientes estudios [1] y [2].

Asimismo, se han realizado diversas prácticas matemáticas para este estudio como KNN (*K-Nearest Neighbour*), análisis discriminatorio, árboles de decisión, *Random Forest*, análisis multivariante, etc. En general, todas ellas relacionadas con *data mining* y *big data*, debido a que, en definitiva, se quiere buscar, a partir de una base de datos conocida, relaciones entre las características de los pacientes, que llamaremos durante todo el trabajo variables. Es decir, intentaremos predecir el valor de la variable que desconocemos y queremos saber del paciente, a partir de las variables que sí conocemos de él y que sabemos que se relacionan con la variable que desconocemos. De tal forma, si tenemos una base de datos de pacientes en las que se explican ciertas variables como, por ejemplo, la altura, el peso, el índice de masa corporal (IMC), las horas entrenadas y si ha sufrido lesión, y se encuentra que la altura, el peso y las horas entrenadas se relacionan de forma logística (en este caso logística múltiple) con la variable 'Si ha sufrido o no una lesión'. De manera que a cualquier paciente futuro que llegue se le podría predecir si va a tener o no una lesión si conocemos su altura, su peso y sus horas dedicadas al entrenamiento.

En este caso, hemos hablado de predecir una lesión, pero las técnicas de regresión se pueden utilizar para cualesquiera variables, siempre y cuando se tenga claro que nuestro

objetivo con ellas es encontrar una relación entre una o un conjunto de variables explicativas y la variable que queremos predecir. Por ello, en este trabajo se intenta enseñar, de forma básica, el concepto de regresión como técnica matemática para el análisis de datos y la distinción de los dos fundamentales tipos de regresión, además de realizar diversos estudios sobre una base de datos asociada al mundo del deporte aplicando dichas técnicas.

2 FACTORES QUE INFLUYEN EN LESIONES DEPORTIVAS

A la hora de prevenir una lesión se miden ciertas propiedades, que se sabe por estudios previos del mismo tema, que pueden influir en la lesión. Aquellos factores que se utilizan para predecir un suceso se llaman factores predictivos. En el deporte, estos elementos sirven para asegurar la prevención de una lesión física y el rendimiento de un entrenamiento específico, entre otros.

Los factores predictivos dependiendo de su naturaleza se dividen en dos clases:

- Factores extrínsecos: afectan a la predicción por causas externas. Normalmente, este tipo de factores se deben al entorno de entrenamiento o por llevar a cabo la actividad física de mala forma.
- Factores intrínsecos: afectan a la predicción debido a la genética del propio individuo. Por ejemplo, enfermedades crónicas, problemas musculares o de tendones de nacimiento debido al desgaste por practicar deporte, etc.

Un ejemplo de factor extrínseco se puede encontrar en el siguiente articulo [3] que explica la relación entre el hecho de estirar antes y después de un entrenamiento o de una prueba física con la aparición de lesiones.

Otro ejemplo, pero esta vez de factor intrínseco, se estudió en [4] que habla de la relación existente entre la morfología de la planta del pie y las lesiones en futbol.

Por tanto, el objetivo de cualquier trabajo de este estilo es encontrar factores predictivos dentro de una población del mundo deportivo para intentar sacar conclusiones útiles para evitar lesiones indeseadas futuras como se lleva a cabo en [5].

3 MODELOS MATEMÁTICOS

Cuando se busca información de una base de datos existen relaciones que no se ven a simple vista y que pueden resultar realmente decisivas en la investigación. Esta información oculta viene dada por relaciones entre las variables que componen la base de datos y que pueden aportar gran valor al estudio. Para encontrar las relaciones entre las variables que componen la base de datos se utilizan diversas técnicas matemáticas como, por ejemplo, análisis discriminatorio, *KNN*, *clustering*, *random forest*, etc. Una de las técnicas más elementales y sencillas de comprender son los modelos de regresión matemáticos, debido a su forma tan visual de presentar dichas relaciones [6].

Los modelos de regresión se utilizan tanto para evaluar cómo cambia una variable con respecto a otra variable, como también para predecir los valores futuros de una variable en función de otra. En el primer escenario, nos encontraríamos con un estudio con una finalidad explicativa, ya que lo único que se quiere poner en relieve es la relación existente entre dos o más variables, debido a la dificultad de cálculo o la necesidad de explicar la variable ya conocida. En el segundo escenario, se tendría un estudio con finalidad predictiva porque, a diferencia del anterior, se busca, a partir de la información que ya se tiene, predecir futuros valores de la variable, debido a la necesidad de conocer ahora valores futuros como, por ejemplo, enfermedades o lesiones físicas. Aunque sean dos enfoques distintos, en el uso de la regresión son compatibles el uno con el otro, de tal forma que se suele realizar primero un análisis explicativo y después un análisis predictivo para comprobar la fiabilidad del análisis previo realizado.

Por ejemplo, según [1], en su estudio utiliza la regresión logística en un conjunto de jugadores de voleibol con una finalidad claramente explicativa con respecto a la aparición de lesiones en el tendón rotuliano, dependiendo de las características físicas de los jugadores. A partir de esta investigación, se podría realizar un análisis predictivo en otros jugadores de voleibol (o incluso otros deportes) que tuvieran las variables explicativas usadas en el artículo, para así comprobar cómo de fiable es este estudio usando distintas muestras.

A pesar de la sencillez de los modelos de regresión, no solo existe un amplio abanico de ellos, sino que también las diferencias entre unos y otros son sutiles. Fundamentalmente,

esta desemejanza de modelos se basa en el tipo de variable dependiente y en el número de variables utilizadas para explicarla.

Las variables que forman una base de datos no son todas iguales entre si. Existen diferencias significativas entre ellas dependiendo del dato que expliquen:

- Variable cuantitativa o numérica: el tipo de dato que explican es un número cuantificable como su nombre indica. Así, la altura o el peso de una persona son dos buenos ejemplos de variables numéricas.
- Variable cualitativa o categórica: los datos explicados por ellas son cualidades, es decir, no son cuantificables y están agrupados de forma excluyente. Un ejemplo de este tipo de variable sería los estados de un semáforo (rojo, ámbar y verde), puesto que son cualidades y excluyentes las unas con las otras, porque al estar el semáforo en verde no puede estar en ninguno de los otros dos estados.

Dentro del primer conjunto, cabe distinguir entre discretas o continúas dependiendo de si los valores numéricos son números enteros o reales. Un ejemplo para entender esta subdivisión sería la variable "altura" y la variable "número de hijos". La primera toma cualquier valor real, pero la segunda solo toma valores enteros positivos que son cuantificables.

Por otro lado, en la segunda división existe un tipo de variable cualitativa muy importante llamada variable dicotómica. Este subtipo se distingue de cualquier otra variable por tomar solo dos valores posibles del estilo "Sí" o "No", "Hombre" o "Mujer", "Existe" o "No existe", etc. Por ejemplo, la variable "paciente con lesión lumbar" que indica si un atleta tiene o no tiene lesión lumbar sería una variable dicotómica.

Por último, aunque no influya en la regresión, dentro de las variables cualitativas que no son dicotómicas se hayan dos subgrupos:

- Nominales: no indican ningún orden. Por ejemplo, la variable "colores de ojos".
- Ordinales: indican un orden o jerarquía. Por ejemplo, la variable "notas académicas".

En la siguiente tabla se muestra, de forma resumida, todos los grupos explicados arriba:

Tipos de variables		Ejemplo
Cuantitativas o numéricas	Discretas	Número de hijos de una familia
	Continuas	Índice de masa corporal (IMC) de un atleta
Cualitativas o categóricas	Dicotómicas	Aparición de una lesión en un atleta
	Ordinales	Notas de un examen
	Nominales	Estados de un semáforo

Tabla 3.0.1: Tipos de variables

En el caso de que la variable dependiente sea numérica, ya sea discreta o continua, se utilizará un modelo de regresión lineal o no lineal. En cambio, si nuestra variable explicada es dicotómica, se usará un modelo de regresión logística, porque, como se verá más adelante para este modelo, se necesita explicarlo utilizando probabilidades.

Ahora que se ha explicado la finalidad de uso de esta técnica y la distinción principal entre sus dos variantes más características, se verá la estructura general de los modelos de regresión.

En esencia, las relaciones entre las variables independientes y la variable dependiente se explican mediante funciones matemáticas, es decir, dependiendo de la función que modelice la relación se tendrá un tipo de regresión. Por tanto, de forma genérica se puede definir un modelo de regresión de la siguiente forma:

$$Y = f(X) + \varepsilon \tag{3.0.1}$$

De la misma forma, si en el modelo existen varias variables independientes, se tendría la siguiente fórmula,

$$Y = f(X_1, \dots, X_l) + \varepsilon \tag{3.0.2}$$

donde Y representa la variable dependiente, es decir, la variable que se pretende explicar o predecir; X sería la variable independiente o $X_1, \dots, X_l$ si fueran varias variables independientes y l sería el número de variables independientes; f sería la función que

explica la relación y; por último, ε sería una variable aleatoria que representa el error cometido al usar el modelo con respecto a la realidad.

Así que esta relación se ha transformado en una función matemática que la modeliza.

Por ejemplo, si la función es lineal y solo utiliza una variable independiente, se trabajará con un modelo de regresión lineal cuya estructura será la siguiente:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.0.3}$$

Pero, por otro lado, si la función es lineal y se utilizan varias variables independientes, se usará el modelo de regresión multilineal, cuya estructura es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_l X_l + \varepsilon \tag{3.0.4}$$

Aunque los modelos (3.0.3) y (3.0.4) son los más usados, se debe tener en cuenta que la función es utilizada para explicar la relación, por lo cual, no tiene por qué ser siempre lineal. Esta función se define a partir de los datos con los que se trabaja, debido a que el objetivo de obtener la regresión es tener un modelo que se ajuste lo mejor posible a los datos y, por ello, la función puede ser lineal o no lineal.

Un ejemplo de regresión no lineal sería la regresión cuadrática, que tiene como modelo:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \varepsilon \tag{3.0.5}$$

Y se usaría para representar una relación cuadrática entre las dos variables. De todos modos, el uso de los distintos modelos lo determinan los datos y las variables con los que se esté trabajando.

Cuando los modelos están compuestos por dos variables se utiliza el diagrama de puntos que representa sus valores en el plano, para así determinar el modelo de regresión. Esto se debe a que se puede sospechar la relación entre las dos variables del modelo, de una forma muy visual.

En las siguientes gráficas se observan dos ejemplos de diagrama de puntos que permiten sospechar la relación entre las dos variables:

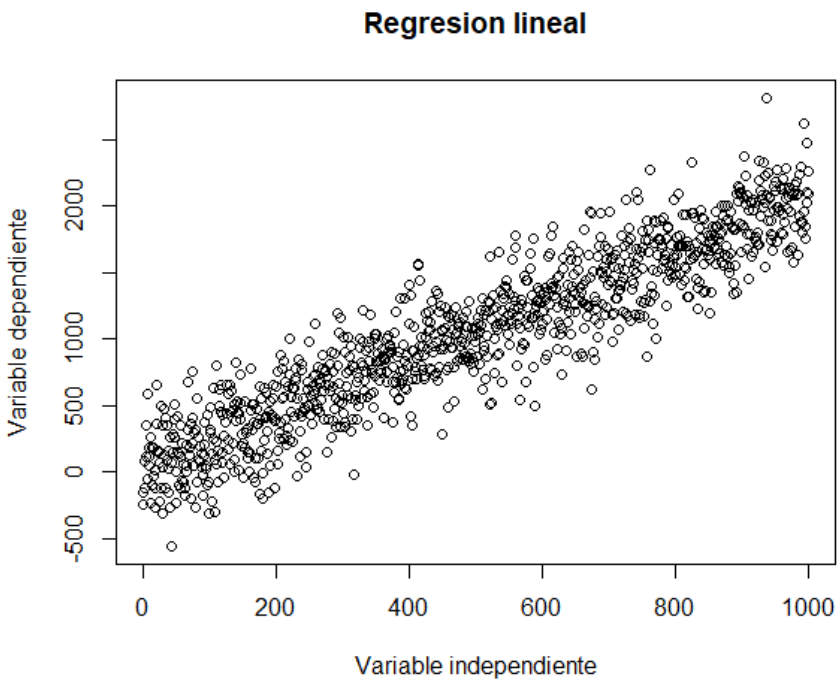


Figura 3.0.1

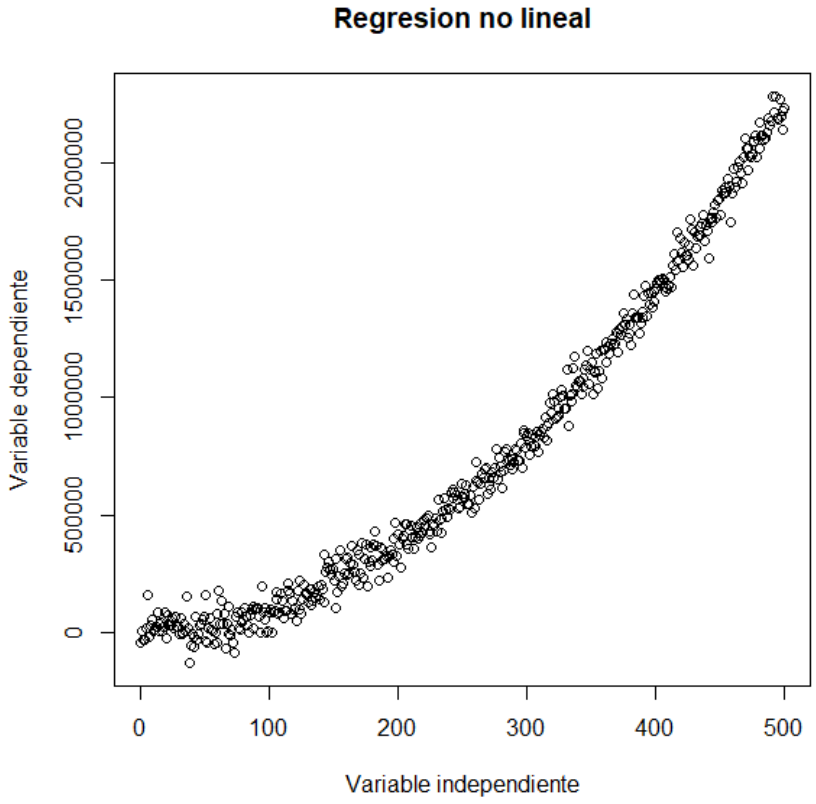


Figura 3.0.2

Como se observa en la Figura 3.0.1, cuando aumenta la variable X, la variable Y también aumenta de forma lineal. En cambio, en la Figura 3.0.2, a pesar de que sucede lo mismo que en la Figura 3.0.1, este aumento se lleva a cabo de una forma no lineal cuadrática.

Aunque es una forma muy clara de ver la relación entre dos variables, la gráfica se realiza como primer acercamiento al estudio del modelo más adecuado. En el apartado de análisis se mostrará qué herramienta se utiliza para elegir el modelo y qué factores externos se han de tener en cuenta a la hora de realizar la regresión y, en anexos un ejemplo de aplicación de regresión lineal.

3.1 Regresión lineal simple

Antes de tratar la relación lineal entre las dos variables, se usa frecuentemente la nube de puntos para visualizar de forma gráfica la posible dependencia entre las dos variables.

El diagrama de puntos o nube de puntos consiste en asignar a cada observación x_i e y_i como si fuera un punto en el plano $(x_i, y_i),\ \forall i \in \{1, \dots, n\}$ y representarlo gráficamente. Al hacerlo queda una gráfica con todas las observaciones representadas en el plano, lo que nos permite ver la existencia de una relación lineal entre ambas variables.

A continuación, se exponen dos ejemplos de diagramas de puntos en los que se ve la relación y la no relación lineal entre las variables:

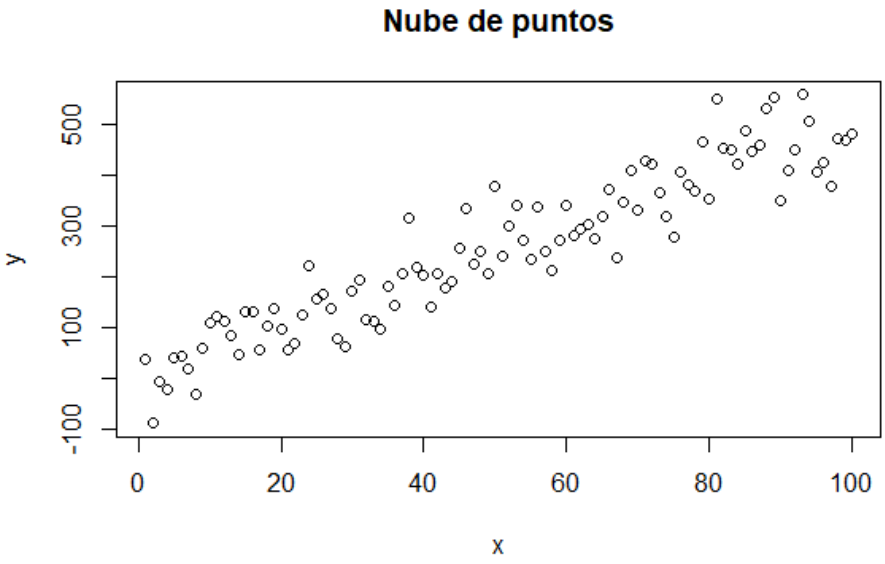


Figura 3.1.1

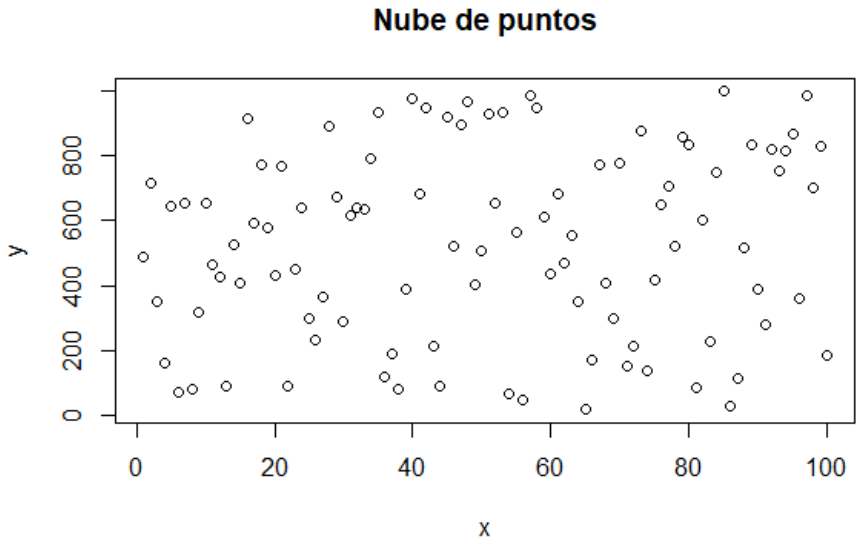


Figura 3.1.2

En la Figura 3.1.1 se puede intuir una dependencia lineal entre las variables X e Y, pero, por otro lado, en la Figura 3.1.2 se observa que no existe una relación aparente, aunque la nube de puntos no nos pueda asegurar qué modelo utilizar, nos sirve para poder desechar la existencia de dependencia entre dos variables.

Después de observar los puntos en el plano y sospechar sobre la existencia de una relación lineal entre ambas variables, se procede a utilizar dos herramientas matemáticas que ayudan, de manera más fiable y rigurosa, a esclarecer la dependencia lineal entre ambas variables. Estas herramientas son la covarianza y el coeficiente de correlación lineal.

La covarianza entre X e Y se define de este modo,

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3.1.1}$$

donde x_i e y_i son las observaciones de cada variable $\forall i \in \{1, \dots, n\}$ y $\bar{x}$ e $\bar{y}$ son las medias muestrales de las variables X e Y:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1.2}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.1.3}$$

Para ver por qué es tan útil la covarianza a la hora de estudiar la dependencia lineal entre X e Y, vamos a analizar el significado de la fórmula.

En la fórmula de la covarianza (3.1.1) al realizar $(x_i - \bar{x})$ e $(y_i - \bar{y})$, $\forall i \in \{1, \dots, n\}$ lo que se está haciendo es desplazar el origen de coordenadas $(0,0)$ del diagrama de puntos original al punto $(\bar{x}, \bar{y})$, el cual será el nuevo origen. Al desplazarlo, todos los puntos originales (x_i, y_i) pasan a ser $(x_i - \bar{x}, y_i - \bar{y})$.

Si utilizamos el ejemplo Figura 3.1.2, con el cuál sospechábamos que existía una relación lineal por cómo era el diagrama de dispersión, vemos qué ocurre al cambiar los puntos:

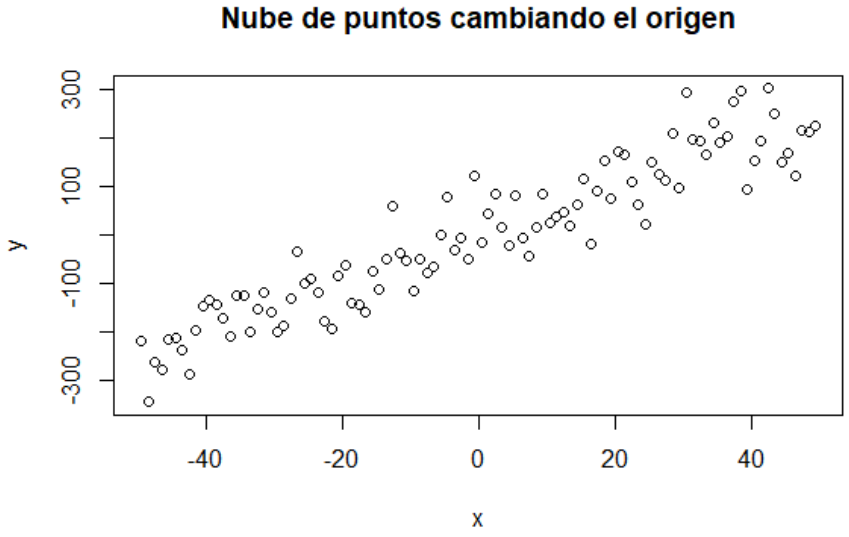


Figura 3.1.3

En esta imagen se puede observar cómo los puntos de la gráfica de la Figura 3.1.1 pasan a estar centrados en el origen en la Figura 3.1.3.

Por último, al realizar el producto se está midiendo qué puntos del diagrama de dispersión caen en cada cuadrante. Si el producto es un número positivo, significa que ha caído en el primer o tercer cuadrante. Sin embargo, si el producto da un número negativo significa que cayó en el segundo o en el cuarto cuadrante. Al realizar la suma de todos estos productos, se mide en qué pares de cuadrantes han caído el mayor número de puntos, por lo que, si la covarianza es positiva, entonces puede existir una relación lineal positiva y, por lo tanto, si la covarianza es negativa, puede existir una relación lineal negativa.

A pesar de ser correcto, presenta dos grandes inconvenientes:

1. Depende demasiado de las unidades de las variables.
2. Los valores que toma la covarianza no están acotados y no se puede cuantificar cómo de buena es la relación lineal.

Por tanto, el estudio de la covarianza entre X e Y no es la herramienta más fiable para estudiar la dependencia, debido a que lo único que nos explica con seguridad es si es positiva o negativa la relación lineal de las variables. El coeficiente de correlación lineal soluciona estos dos grandes problemas que presenta la covarianza.

Se define así el coeficiente de correlación lineal entre X e Y,

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.1.4}$$

donde σ_X y σ_Y son las desviaciones típicas de las variables:

$$\sigma_X = \sqrt{Var(X)} = \sqrt{Cov(X,X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.1.5}$$

$$\sigma_Y = \sqrt{Var(Y)} = \sqrt{Cov(Y,Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.1.6}$$

Este factor soluciona los dos problemas que tenía la covarianza:

1) No tiene unidades, es decir, es adimensional.
2) $\rho_{XY} \in [-1,1]$

Es muy sencillo ver que es un coeficiente adimensional, debido a que las unidades del numerador y las del denominador son las mismas y, por tanto, se anulan entre ellas.

Para demostrar 2) hay que ver la fórmula del coeficiente de correlación como el producto escalar de dos vectores. Sabiendo que todo producto escalar de 2 vectores cumple las siguientes fórmulas,

$$\langle \vec{u}, \vec{v} \rangle = \sum_{i=1}^{n} u_i v_i \tag{3.1.7}$$

$$\langle \vec{u}, \vec{v} \rangle = \|u\|\|v\| cos\varphi_{XY} \tag{3.1.8}$$

donde $\|u\|$ y $\|v\|$ son los módulos de los vectores n dimensionales u y v, y φ_{XY} es el ángulo que forman u y v.

Ahora vamos a observar que la fórmula del coeficiente de correlación en realidad es un producto escalar. Por un lado, se tiene que las fórmulas de la covarianza se pueden escribir,

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\langle \overrightarrow{d_x}, \overrightarrow{d_y} \rangle \tag{3.1.9}$$

$$Var(X) = \frac{1}{n}\langle \overrightarrow{d_x}, \overrightarrow{d_x} \rangle = \frac{1}{n}\|d_x\|^2 \tag{3.1.10}$$

$$\mathrm{Var}(Y) = \frac{1}{n}\langle \overrightarrow{d_Y}, \overrightarrow{d_Y} \rangle = \frac{1}{n}\|d_Y\|^2 \tag{3.1.11}$$

donde d_X y d_Y son dos vectores n dimensionales.

Juntando las fórmulas (3.1.4) y (3.1.9), (3.1.10) y (3.1.11) se tiene:

$$\rho_{XY} = \frac{\langle \overrightarrow{d_x}, \overrightarrow{d_y} \rangle}{\|d_x\|\|d_x\|} \tag{3.1.12}$$

Por tanto,

$$\rho_{XY} = cos\varphi_{d_X d_Y} \tag{3.1.13}$$

Y como la función coseno esta acotada entre [-1,1], entonces el coeficiente de correlación también lo está. Con esto quedaría demostrado 2).

Este coeficiente se utiliza para medir el comportamiento lineal entre ambas variables. Dependiendo de los valores que tome el coeficiente, nos dará la siguiente información:

- Si los valores del coeficiente de correlación son cercanos a 1, entonces existe una relación lineal positiva entre X e Y.
- Si los valores del coeficiente de correlación son cercanos a -1, entonces existe una relación lineal negativa entre X e Y.
- Si los valores del coeficiente de correlación son cercanos a 0, entonces no existe una relación lineal entre X e Y.

Cuando se sabe que existe dicho comportamiento lineal entre ambas variables, se procede a calcular los coeficientes de la regresión lineal mediante la técnica de los mínimos cuadrados.

Como se está tratando con una relación lineal, la nube de puntos ha de verificar la siguiente ecuación,

$$y = \beta_0 + \beta_1 x \tag{3.1.14}$$

donde $\beta_0, \beta_1 \in \mathbb{R}$.

Por tanto, cada punto de la nube tiene que cumplir:

$$y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \dots, n \tag{3.1.15}$$

(3.1.15) es un sistema de n ecuaciones y 22 incógnitas. Por el Teorema de Rouché-Frobenius, este sistema es incompatible, luego no existen valores β_0 y β_1 que verifiquen todas las ecuaciones, es decir, no existirán valores de los coeficientes que se ajusten de forma exacta a los valores reales de la nube de puntos. Luego, se buscará una estimación de los coeficientes que mejor se ajuste al diagrama de dispersión, lo que dará lugar a un error llamado residuo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, n \tag{3.1.16}$$

Ahora en (3.1.16) se consigue un sistema de n ecuaciones y $n + 2$ incógnitas. A continuación, por el Teorema de Rouché-Frobenius esta vez será un sistema compatible indeterminado con un número infinito de soluciones.

Para entender que son los residuos, se debe tener en mente la finalidad de la regresión: encontrar un modelo a partir de unas observaciones reales que logre predecir, lo mejor posible, cierto dato a partir de información conocida. Luego, por un lado, estarán las observaciones y, por otro, las predicciones usando el modelo que denotaremos:

$$\widehat{y_i} = \beta_0 + \beta_1 x_i, \qquad i = 1, \dots, n \tag{3.1.17}$$

Los residuos serán la diferencia entre el valor real de la variable y la predicción:

$$\varepsilon_i = y_i - \widehat{y_i} = y_i - (\beta_0 + \beta_1 x_i), i = 1, \dots, n \tag{3.1.18}$$

Luego, el objetivo es encontrar entre ese número infinito de soluciones la mejor solución posible, es decir, se busca la recta que se ajusta lo mejor posible a la nube de puntos. Lo que matemáticamente se quiere es que la suma de los residuos sea lo menor posible, pero sin que se compensen los unos con los otros, porque como se ve en (3.1.18) adquieren del mismo modo valores negativos.

Existen dos formas de buscar la suma mínima de residuos: usando el valor absoluto o mínimos cuadrados. Si se utiliza la primera:

$$\min \sum_{i=1}^{n} |\varepsilon_i| \tag{3.1.19}$$

No obstante, calcular los coeficientes minimizando esta expresión es difícil y, además, no penaliza el hecho de que sea grande el error. Por ello, se usa para minimizar:

$$\min \sum_{i=1}^{n} \varepsilon_i^2 \tag{3.1.20}$$

Esta técnica es llamada mínimos cuadrados. Si desarrollamos (3.1.20) se tendría,

$$\min \sum_{i=1}^{n} \varepsilon_i^2 \;=\; \min \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 = \min g(\beta_0, \beta_1) \tag{3.1.21}$$

donde se quieren encontrar los valores de β_0 y β_1 que minimizan la función g (3.1.21). Para ello, estudiaremos los valores que anulan la derivada y hallaremos qué es un mínimo global. Derivando y resolviendo el sistema de dos ecuaciones y dos incógnitas resultante, se obtiene que los coeficientes son:

$$\widehat{\beta_1} = \frac{n \sum x_i y_i \;-\; \sum x_i \sum y_i}{n \sum x_i^2 \;-\; \left(\sum x_i\right)^2} = \frac{Cov(X,Y)}{Var(X)} \tag{3.1.22}$$

$$\widehat{\beta_0} = \frac{\sum x_i^2 \sum y_i \;-\; \sum x_i \sum x_i y_i}{n \sum x_i^2 \;-\; \left(\sum x_i\right)^2} = \bar{y} \;-\; \beta_1 \bar{x} \tag{3.1.23}$$

Por último, faltaría estudiar si para los valores (3.1.22) y (3.1.23) se tiene un mínimo absoluto, lo cual es obvio ya que la función (3.1.21) es un paraboloide y, por tanto, solo posee un extremo relativo (mínimo) el cual será un mínimo absoluto.

3.2 Regresión multilineal

Como se ha visto en (3.0.4), este modelo es prácticamente análogo a la regresión lineal. En él no se busca explicar una relación directa entre dos variables, sino que se intenta hallar una influencia de varios factores sobre una variable de forma lineal. Normalmente, se desconoce que tan siquiera existe dicha relación, ya que al trabajar con varias variables explicativas no se distingue de forma clara cuáles afectan y cuáles no al modelo. Por ello, como se puede observar en [7], existen algoritmos y técnicas que nos ayudan a escoger

las variables cuya relación con la variable dependiente es mayor y desechar aquellas que no se relacionan.

El objetivo de los algoritmos nombrados previamente es seleccionar, de un conjunto de variables, un subconjunto de estas que se relacione con mayor fuerza con la variable predictora.

Estos algoritmos se llaman algoritmos de selección de variables y se emplean no solo en regresión multilineal, sino que también en cualquier modelo de regresión que posea un conjunto de dos o más variables explicativas. Y aunque en esencia y en finalidad todos sean lo mismo, existen diferencias en los algoritmos debidas a cómo llevan a cabo la tarea de selección:

- Backward: se basa en estudiar el modelo usando todas las variables independientes e ir eliminando una a una aquellas que tengan menor relación con la variable predictora, es decir, que se ajusten peor. Cada vez que se elimina una variable, se vuelve a estudiar el modelo y se busca la variable que peor se ajusta y se elimina. Esto se realiza de manera sucesiva hasta que no se puedan eliminar más variables.
- Forward: se basa en empezar con el modelo sin ninguna variable e ir añadiéndolas. Cuando se encuentra una variable que se relaciona bien a la variable dependiente, se añade al modelo. Después de añadirla, se vuelve a buscar otra variable que se ajuste bien, pero esta vez se tiene en cuenta la variable añadida previamente y así sucesivamente hasta que no se pueda añadir más variables.
- Stepwise: este algoritmo mezcla ambas técnicas anteriores. A cada paso que se toma se estudia si entra una variable y si sale otra variable del modelo.

De todas formas, estos algoritmos son orientativos, por lo que siempre prevalecerá por encima de cualquier técnica los criterios de selección que el investigador considere apropiado de estudiar.

Una vez se han seleccionado las variables que se relacionan de forma lineal, el modelo utilizado sería:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_l x_l \qquad (3.2.1)$$

Como se puede apreciar en (3.2.1) es análogo a (3.1.14), salvo que esta vez no existe una sola variable, sino que un número l de ellas. Este modelo, al igual que en regresión lineal, no se ajusta de forma exacta y, por ello, existe un residuo que se denota:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_l x_l + \varepsilon \tag{3.2.2}$$

Ahora si aplicamos a (3.2.2) las observaciones reales nos darían el siguiente sistema,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_l x_{il} + \varepsilon_i, \qquad i = 1, \dots, n \tag{3.2.3}$$

donde los subíndices de x_{ij} denotan la observación i de la variable j.

Este sistema, añadiendo los residuos, se convierte en un sistema de n ecuaciones y $n + l + 1$ incógnitas. De modo que por el teorema de Rouché-Frobenius, sería un sistema compatible indeterminado con infinitas soluciones. Para escribir dicho sistema y por comodidad, se utiliza la siguiente notación matricial:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1l} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nl} \end{pmatrix} \tag{3.2.4}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_l \end{pmatrix} \tag{3.2.5}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3.2.6}$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{3.2.7}$$

De tal forma que el sistema de (3.2.3) se puede reescribir de una forma más sencilla y compacta:

$$Y = X\beta + \varepsilon \tag{3.2.8}$$

Ahora como se hizo en la regresión lineal simple, hay que encontrar los coeficientes que minimizan la suma de los errores al cuadrado. Utilizando la notación matricial de (3.2.6) la fórmula (3.1.20) que queremos minimizar quedaría:

$$\min \varepsilon^t \varepsilon \tag{3.2.9}$$

Y hallando de forma análoga los coeficientes de la regresión, se anulan las derivadas con la salvedad de que son $l + 1$ coeficientes y el vector de coeficientes de la regresión escrito de forma matricial sería,

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{3.2.10}$$

donde X e Y son las formas matriciales de las observaciones (3.2.4) y (3.2.7).

Este vector de coeficientes, al igual que en la regresión lineal, es donde se alcanza el mínimo absoluto de la función (3.2.9).

De todos modos, aunque los coeficientes de los modelos no se suelen calcular debido a la dificultad de cálculos tanto para regresiones lineales o no lineales como logística, hay que saber de dónde vienen dichos coeficientes para entender los modelos de regresión de forma adecuada. En la práctica, se utilizan programas estadísticos que los calculan de manera automática como, por ejemplo, RStudio, Python o SAS. De manera que se deja al investigador solo la parte de análisis e interpretación de los resultados.

3.3 Regresión logística

Como se introdujo previamente al inicio de este apartado, este modelo es usado para predecir o explicar el comportamiento que sufre una variable que solamente toma dos valores a partir de una o varias variables independientes.

Este modelo es muy útil para responder preguntas del estilo “Sí” o “No”, “Ocurre” o “No ocurre”, etc. Fundamentalmente, es el tipo de regresión más usada debido a su versatilidad en cualquier ámbito, aunque como aplicación más desarrollada actualmente son los estudios clínicos y farmacológicos, de los cuales se trasladó al ámbito deportivo.

Aunque parezcan similares tanto la regresión lineal como logística, hay que entender por qué son distintas y por qué no es aplicable un modelo lineal cuando se trabaja con una variable dicotómica. Para ver esto, se utilizará una variable dicotómica de la base de datos que se usará en el apartado de análisis. Su diagrama de puntos es el siguiente:

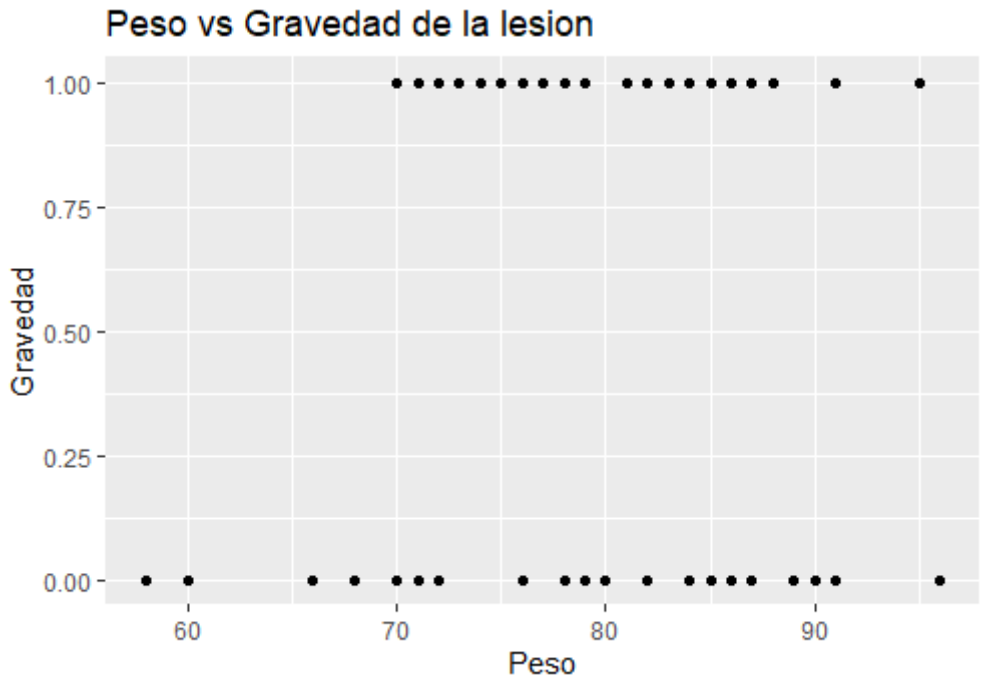


Figura 3.3.1

Visualmente se observa que un modelo lineal no se ajustaría a Figura 3.3.1, ya que los puntos no están dispuestos según una distribución lineal. Por ello, si se hiciera dicha regresión lineal quedaría de la siguiente manera:

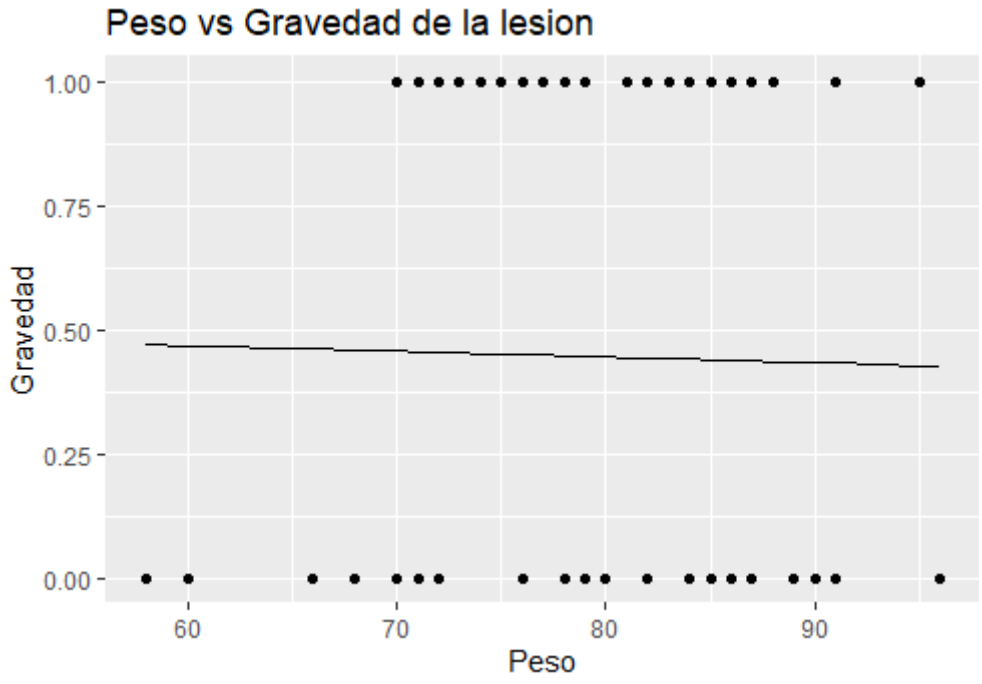


Figura 3.3.2

Para valores de la recta en Figura 3.3.2 positivos grandes no tendría sentido los valores que predijera la recta, puesto que no se ajusta correctamente. De forma análoga sucedería en el caso de varias variables independientes.

Por tanto, de este ejemplo se deduce que matemáticamente en un modelo logístico cuya variable dependiente es dicotómica es posible aplicar un modelo lineal, aunque si se aplicara no tendría sentido lógico ni funcional para su objetivo, el cual es explicar o predecir el comportamiento de la variable respuesta.

Luego, la dificultad de la regresión logística reside en la sutileza conceptual y no en los cálculos matemáticos que se realicen, que como se verá más adelante son prácticamente análogos a la regresión lineal.

Antes de seguir con el modelo, es necesario detenerse para realizar dos pasos previos. El primero de ellos es transformar los valores dicotómicos de la variable en dos valores numéricos 0 y 1, de tal forma que los valores "Sí" se transformen en 1 y los valores "No", en 0. El segundo paso que hay que tener en cuenta es otra transformación, pero esta vez no es estética, sino que es un cambio en la variable respuesta. En vez de trabajar con dos valores, se tratará con la probabilidad de que un suceso ocurra. Al estar dichos valores comprendidos entre 0 y 1 por ser probabilidades, se podrá fijar una probabilidad (normalmente p = 0.5) a partir de la cual, si la probabilidad dada por la regresión es mayor que la probabilidad fijada, se toma como resultado 1, es decir, "Sí", y en caso contrario, se toma como "No".

Para esta transformación se utiliza la razón de probabilidades o más comúnmente conocida como "razón de momios" o "odds ratio", que aparece de forma natural en la propia ecuación del modelo cuya fórmula es:

$$Odds\ Ratio = \frac{p}{1-p} \tag{3.3.1}$$

La fórmula (3.3.1) que se puede leer en el artículo [8] expresa la proporción de casos favorables que existen con respecto a aquellos desfavorables y su funcionamiento en cualquier estudio. Por ejemplo, si la probabilidad de tener lesión realizando una sesión intensiva de entrenamiento es p = 0.75, entonces la probabilidad de no tenerla será 1 – p = 0.25 y la razón de probabilidad será el cociente cuyo resultado es 3, lo cual quiere decir que se esperan 3 lesionados por cada caso no lesionado.

El problema fundamental que tiene la regresión logística es crear un modelo que se ajuste de forma correcta y eficaz a las futuras predicciones, debido a la propia naturaleza de la variable respuesta. Al igual que en los modelos lineales, se utiliza una ecuación que modelice dicho modelo logístico de la forma más apropiada. La función más acorde a lo buscado es la función sigmoidal:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.3.2}$$

La función de la ecuación (3.3.2) se ajusta de forma apropiada a la gráfica de una nube de puntos de una variable dicotómica, debido al comportamiento de su límite cuando se acerca a ∞ y a $-\infty$:

$$\lim_{x \to \infty} f(x) = \lim_{x \to \infty} \frac{1}{1 + e^{-x}} = 1 \tag{3.3.3}$$

$$\lim_{x \to -\infty} f(x) = \lim_{x \to \infty} \frac{1}{1 + e^{-x}} = 0 \tag{3.3.4}$$

De tal forma que como se puede ver en (3.3.3), cuando los números de la variable independiente sean positivos grandes se aproximará a 1 y, de forma análoga, en (3.3.4) para valores negativos grandes acercándose a 0. Esta función soluciona los problemas que tenía la regresión lineal para ajustarse a la nube de puntos, ya que esta vez no existirá la oportunidad de que los valores que prediga sean valores sin sentido.

Haciendo los cambios oportunos y sustituyendo por el concepto de regresión en la ecuación (3.3.2) se obtiene,

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{3.3.5}$$

de donde despejando de la ecuación (3.3.5) da lugar a una expresión muy parecida a la obtenida en (3.1.14),

$$\log \frac{y}{1 - y} = \beta_0 + \beta_1 x \tag{3.3.6}$$

donde en (3.3.6) el cociente dentro del logaritmo sería la razón de probabilidades (3.3.1).

Luego, en vez de trabajar con la variable dicotómica, se tratará con una variable que marque la probabilidad de que el suceso ocurra. Se marcará un umbral a partir del cual, si se excede dicha probabilidad previamente seleccionada, se tomará como afirmativo y en caso contrario, negativo.

Normalmente, se denota esta nueva variable de probabilidad como p, y se calcula despejando de la ecuación (3.3.6), dando lugar a:

$$p = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3.3.7}$$

En el caso en el que existieran varias variables independientes que intervengan en la variable respuesta, la fórmula, al igual que en la regresión multilineal con respecto a la lineal simple, sería una ampliación trivial de coeficientes, dando lugar a la siguiente expresión:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots\beta_l x_l}}{1 + e^{\beta_0+\beta_1 x_1+\cdots\beta_l x_l}} \tag{3.3.8}$$

Para terminar esta sección, habría que estudiar cuáles y cómo se calcularían los coeficientes de la regresión logística. A diferencia de la regresión lineal, no se minimiza la suma de los residuos al cuadrado (3.1.20), ya que en este caso no tendría sentido tan siquiera hablar de residuos por ser un modelo logístico. Por ello, el estimador usado para calcular los coeficientes es el de máxima verosimilitud, el cual se calcula maximizando la función de verosimilitud. Estos coeficientes no se pueden calcular de forma analítica, debido a dificultad de cálculo, pero se puede ver de dónde salen dichos coeficientes. A continuación, se verá cómo llegar a la expresión de los coeficientes en el caso general con varias variables independientes.

Partimos de un modelo Bernoulli que solo puede tomar dos valores, ya que estamos en regresión logística y nuestra variable es dicotómica:

$$P(Y = 1 \mid X) = p \tag{3.3.9}$$

$$P(Y = 0 \mid X) = 1 - p \tag{3.3.10}$$

Suponiendo la independencia de los sucesos (hipótesis trivial), se consigue la expresión de la función de máxima verosimilitud de un modelo Bernoulli:

$$P(Y) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{3.3.11}$$

Tomando logaritmos de la expresión (3.3.11):

$$\log P(Y) = \sum_{i=1}^{n} y_i \log(\frac{p_i}{1 - p_i}) + \sum_{i=1}^{n} \log(1 - p_i) \tag{3.3.12}$$

Ahora expresamos las probabilidades de (3.3.12) como función de los valores independientes de las variables X:

$$p = P(Y = 1\,|X) = G(\beta_0 + \beta_1 x_1 + \cdots \beta_l x_l) \tag{3.3.13}$$

En el caso que estamos viendo, la función G de (3.3.13) sería:

$$G(\beta_0 + \beta_1 x_1 + \cdots \beta_l x_l) = \frac{e^{\beta_0+\beta_1 x_1+\cdots\beta_l x_l}}{1 + e^{\beta_0+\beta_1 x_1+\cdots\beta_l x_l}} \tag{3.3.14}$$

Por tanto, utilizando la función de (3.3.14) en la ecuación (3.3.12) daría lugar a la siguiente expresión:

$$\begin{aligned}\log P(Y) = &\sum_{i=1}^{n} y_i \log(\beta_0 + \beta_1 x_1 + \cdots \beta_l x_l) \\ &- \sum_{i=1}^{n} \log(1 + e^{\beta_0+\beta_1 x_1+\cdots\beta_l x_l})\end{aligned} \tag{3.3.15}$$

La ecuación (3.3.15) habría que derivarla con respecto a cada coeficiente β_i y anularla, porque estamos buscando un máximo de dicha ecuación. Una vez calculadas todas las derivadas, solo quedaría resolver el sistema de ecuaciones resultante, el cual es no lineal, por lo que habría que utilizar técnicas de métodos numéricos para resolverlo.

Por último, cabría señalar, aunque se haya hecho mención durante el capítulo, que dicho modelo de regresión logística tiene su forma ampliada en el modelo logístico múltiple. Este modelo funcionaría exactamente igual que el modelo simple, con la única salvedad que, al igual que en el modelo multilineal, habría una fase previa de selección de variables.

4 ANÁLISIS DE LOS MODELOS

En este apartado se enseñará inicialmente unos cuantos conceptos de análisis muy importantes previos a la regresión, como la comprobación de las hipótesis de Gauss-Márkov, el estudio de la bondad de ajuste del modelo, el tratamiento de variables cualitativas dentro del conjunto de variables independientes e incluso el concepto de matriz de confusión.

Por último, se estudiará un ejemplo de aplicación de la regresión logística múltiple en una base de datos de índole deportiva con el fin de responder ciertas preguntas planteadas más adelante.

4.1 Hipótesis de Gauss – Márkov

Durante todo el capítulo anterior se ha supuesto la validez de los modelos sobre las observaciones, pero dichas observaciones deben cumplir ciertos requisitos para que los modelos vistos funcionen de forma eficaz.

Los requisitos previos necesarios para llevar a cabo la regresión lineal simple y multilineal se llaman hipótesis de Gauss-Márkov y están basados en las propiedades que han de cumplir los residuos, ya que a la hora de estimar los coeficientes de la regresión en el caso lineal se centra en el método de los mínimos cuadrados sobre ellos.

El teorema de Gauss-Márkov establece que, si se verifican las hipótesis, entonces el estimador calculado por mínimos cuadrados es el mejor estimador lineal insesgado o también llamado en inglés BLUE (*Best Linear Unbiased Estimator*), es decir, los modelos de regresión lineal son correctos.

Las hipótesis de Gauss-Márkov son las siguientes:

1) Linealidad: los valores de la variable dependiente están relacionados de forma lineal con las variables explicativas:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_l X_l \tag{4.1.1}$$

2) Medida nula: la esperanza matemática de los residuos es nula:

$$E[\varepsilon_i] = 0, \qquad \forall i \tag{4.1.2}$$

3) Homocedasticidad: todos los residuos poseen la misma varianza:

$$Var[\varepsilon_i] = \sigma^2, \qquad \forall i \tag{4.1.3}$$

4) No correlación entre los residuos: los residuos dos a dos son incorrelados:

$$Cov[\varepsilon_i, \varepsilon_j] = 0, \qquad \forall i \neq j \tag{4.1.4}$$

5) Normalidad: la variable aleatoria que modeliza los residuos se distribuye con una distribución normal:

$$\varepsilon \sim N(0, \sigma^2) \tag{4.1.5}$$

4.2 Bondad de ajuste

Una vez diseñado el modelo, se necesita una forma para comprobar cómo se ajusta el modelo a las observaciones. Para llevar a cabo esto se utiliza el p-valor.

Antes de hablar de él, hay que entender qué buscamos. El objetivo es tomar una decisión acerca de nuestro modelo, es decir, decidir si es bueno o malo. Para ello, se realiza un contraste de hipótesis de la siguiente manera:

$$\begin{aligned} H_0 &\equiv \beta_1 = \cdots = \beta_l = 0 \\ H_1 &\equiv \beta_i \neq 0, \qquad i = 1, \dots l \end{aligned} \tag{4.2.1}$$

Si se acepta la hipótesis nula, entonces significa que el modelo es mejor tomando nulos los coeficientes de las variables y, por tanto, la regresión quedaría como una recta horizontal constante en el valor β_0. Si no se acepta, se reconoce que los coeficientes calculados se ajustan bien a la muestra.

Este contraste tiene de forma intuitiva mucho sentido, ya que la base de la regresión son los coeficientes e indicar si es mejor o no tenerlos, explica claramente que sea bueno o malo el modelo calculado a partir de las observaciones.

Una vez entendido el contraste, quedaría pasar a estudiar el concepto de p-valor. El p-valor es un valor usado en los contrastes de hipótesis para determinar la probabilidad de que ocurra la hipótesis nula. En el caso de la regresión, indicaría la probabilidad de que los coeficientes no independientes sean nulos, entonces cuanto más pequeño sea el p-valor, mejor será el modelo, puesto que se ajustará mejor. Normalmente, el modelo de regresión calculado se acepta cuando el p-valor < 0.05 y se rechaza en caso contrario.

4.3 Variables dummy

En cualquier modelo de regresión, se utilizan como variables explicativas no solo aquellas que toman valores numéricos, sino que también variables categóricas que pueden ser dicotómicas o no, es decir, tener dos o más estados cualitativos.

El tratamiento de estas variables en los modelos es distinto al de otras variables y dependerá del número de casos que tome. En el caso de que la variable cualitativa sea dicotómica, el único tratamiento necesario será transformar sus valores cualitativos en 0 y en 1. A esta variable se le puede considerar en sí misma variable dummy. Por otro lado,

al tratar con una variable que toma t estados con $t > 2$, se necesita transformar la variable en $t - 1$ variables auxiliares dicotómicas llamadas dummy. De forma precisa, cada variable dummy representa si se cumple uno de los estados de la variable original, es decir, la variable dummy que representa dicho estado toma el valor 1 y el resto toman el valor 0. Al ser $t - 1$ variables representan $t - 1$ estados de la variable original, por lo que restaría uno, el cual viene representado en el coeficiente independiente de la regresión al tomar todas las variables dummy el valor nulo.

Por tanto, los coeficientes de la regresión que acompañen a cada variable dummy generada indicarán cuánta relación tendrá la variable dependiente cuando el estado de la variable original que representa ocurre.

Para ilustrar el concepto de variable dummy se va a estudiar un ejemplo. Supongamos que ya sabemos que existe una relación lineal entre el número de días lesionado y el número de horas de entrenamiento y el tipo de lesión.

$$Y \equiv \text{Número de días lesionado} \tag{4.3.1}$$

$$X_1 \equiv \text{Horas de entrenamiento} \tag{4.3.2}$$

$$X_2 \equiv \text{Tipo de lesión} \tag{4.3.3}$$

La variable X_2 es cualitativa y puede tomar los siguientes estados: ósea, muscular y ligamentosa. Las variables X_1 e Y son variables numéricas.

Por saber que existe relación lineal, todas estas variables tienen que cumplir el modelo multilineal (3.2.2):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{4.3.4}$$

Al ser X_2 categórica, se tiene que transformar en $t - 1$ variables dummy, es decir, en 2 variables dicotómicas (porque tiene 3 estados), de tal forma que el modelo final resultante quedaría,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon \tag{4.3.5}$$

donde la variable dummy D_1 representa si ocurre una lesión de tipo óseo o no y D_2, una lesión muscular o no. Si ambas toman el valor nulo, entonces nos encontraríamos con una lesión ligamentosa y vendría representado por β_0.

4.4 Matriz de confusión: sensibilidad y especificidad.

Al realizar una regresión logística con el objetivo de predecir los valores de la variable respuesta, se suele llevar a cabo una partición de los datos. Una parte de las observaciones se utilizan para crear el modelo logístico y la otra parte, para comprobar si el modelo calculado predice correctamente. Normalmente en esta partición se utiliza el 80% de las observaciones como entrenamiento para el modelo y el 20% restante como prueba.

La matriz de confusión es una tabla que representa y enfrenta aquellos datos de prueba, de los cuales, por ser datos de la base, sabemos frente a los datos interpretados y calculados por el modelo. De tal forma que se puede interpretar cómo de bueno es el modelo logístico, al igual que hicimos con el modelo lineal.

Para entender el funcionamiento de una matriz de confusión, vamos a ilustrar un ejemplo en la siguiente tabla:

		Predicciones		
Observaciones		No lesionados	Lesionados	Total
	No lesionados	12	7	19
	Lesionados	3	50	53
	Total	15	57	72

Tabla 4.4.1

En esta tabla se representan los valores reales de la base de datos en las filas y los valores predichos por el modelo en las columnas. Por tanto, nuestra base de datos ficticia de 72 pacientes de Tabla 4.1.1 posee 53 sujetos lesionados y 19 no lesionados, mientras que, por otro lado, el modelo predijo 57 sujetos lesionados y 15 no lesionados. Otra forma de analizar el modelo sería estudiando cada celda de la matriz, de donde se puede observar que nuestro modelo de 53 observaciones de lesionados ha predicho 50 lesionados y 3 no lesionados. De forma análoga, se puede ver que de 19 no lesionados el modelo predijo 7 pacientes como lesionados y 12 como no lesionados.

De forma más clara para entender la tabla, se hace la siguiente traducción a la información que aporta:

	Negativos	Positivos	Total
Negativos	Verdadero negativo (VN)	Falso positivo (FP)	Total negativos
Positivos	Falso negativo (FN)	Verdadero positivo (VP)	Total positivos
Total	Negativos predichos	Positivos predichos	Total observaciones

Tabla 4.4.2

Por consiguiente, a raíz de los datos proporcionados por la matriz de confusión se puede calcular la precisión de acierto del modelo:

$$Exactitud = \frac{VP + VN}{Total} \tag{4.4.1}$$

El ejemplo de la Tabla 4.1.1 tendría una capacidad de acierto de $^{62}/_{72} = 0.8611$, es decir, de un 86.11%.

Por otro lado, dos valores relevantes a la hora de decidir la viabilidad del modelo logístico son la sensibilidad y especificidad. A pesar de la importancia de conocer la exactitud del modelo, recae sobre él un gran inconveniente: la incapacidad de distinguir si ambos VP o VN se están prediciendo correctamente o solo uno de ellos dos, ya que en un modelo cuya mayoría de datos son positivos y la restante minoría son negativos, predecir bien los positivos daría lugar a un porcentaje alto de exactitud sin tener por qué predecir correctamente los negativos.

Como se puede observar en el artículo [9] ambas especificidad y sensibilidad son fundamentales para las pruebas diagnósticas en medicina y, en general, para cualquier estudio de regresión logística, ya que explican:

- Sensibilidad: es la probabilidad de positivos entre todos los sujetos que deberían predecir como positivos, es decir, de entre los verdaderos positivos y los falsos positivos.

$$Sensibilidad = \frac{VP}{VP + FN} \tag{4.4.2}$$

- Especificidad: es la probabilidad de negativos de entre todos los sujetos que deberían predecirse como negativos, es decir, de entre los verdaderos negativos y los falsos positivos.

$$Especificidad = \frac{VN}{VN + FP} \tag{4.4.3}$$

Por tanto, ambas explican cada uno de los aspectos para validar el modelo logístico, es decir, la sensibilidad indica la proporción de verdaderos positivos de entre todos los positivos de las observaciones, lo cual indica cómo de bien se están prediciendo los positivos y, de forma análoga, la especificidad con los negativos.

4.5 Aplicación de los modelos

A continuación, se aplicarán los conceptos explicados a lo largo de todo el trabajo en una base de datos orientada al mundo del deporte. Dicha base de datos está compuesta por 95 observaciones, de las cuales habría que rechazar 14 por ser repetidas y 3 por ser previamente excluidas. Por tanto, la base quedaría con un total de 78 observaciones, las cuales representan a pacientes deportistas.

Sobre cada observación se han medido ciertos datos que representan las variables de la tabla, teniendo un total de 16 variables el sistema. De entre ellas, se descartaron varias variables debido a dos motivos claros: no aportaban información a pesar de que fueran variables importantes debido a su nula variabilidad (Sexo, Dominancia, etc.) o no tenían interpretabilidad clara (Unidad, Especificidad, Empleo, etc.). Una vez realizados todos los descartes por la falta de información, quedarían un total de 8 variables, siendo una de ellas adaptada a una variable dicotómica:

- Edad: variable que indica la edad del paciente.
- Peso: variable que indica el peso del paciente.
- Altura: variable que indica la altura del paciente.
- IMC: variable que indica el índice de masa corporal del paciente.
- Antecedentes: variable transformada en dicotómica que indica si un paciente ha padecido una lesión previa.
- Tipo: variable categórica que indica el tipo de lesión que ha sufrido el paciente.
- Mecanismo: variable categórica que indica el movimiento que imposibilita la lesión del paciente.
- Días recuperación: variable que indica el número de días que el paciente va ha estado lesionado.

El objetivo de este apartado es responder a ciertas preguntas utilizando modelos de regresión, los cuales sustraen información de la base de datos para responderlas. Los códigos se encuentran en el apartado de Anexos.

4.5.1 ¿Se puede predecir si un paciente tendrá más de 10 días de lesión según su edad, su altura, su IMC, su peso y si ha tenido lesiones previas?

Para responder a esta pregunta inicialmente se ha transformado la variable 'Días lesión' en una variable dicotómica que representa si ha tenido más de 10 días de lesión o no. Una vez transformada, se ha planteado un modelo de regresión logística múltiple (debido a que existen más de una variable independiente) y, posteriormente, se ha descartado la variable 'Edad', ya que no era significativa. Por último, se ha calculado la matriz de confusión, en la cual se obtiene una precisión del modelo del 67,94%, una sensibilidad del 83,33%, es decir, explica de una forma muy precisa los pacientes que tendrán más de 10 días de lesión y una especificidad del 43,33%, o sea, no predice de forma muy acertada los pacientes que tendrán menos de 10 días de lesión.

El modelo que indicaría la probabilidad de que el paciente padezca más de 10 días de lesión sería:

$$P = \frac{e^{0.2648-12.8260*Peso+9.4568*Altura+9.4025*IMC+1.5723*Antecedentes}}{1+e^{0.2648-12.8260*Peso+9.4568*Altura+9.4025*IMC+1.5723*Antecedentes}} \quad (4.5.1)$$

4.5.2 ¿Depende de las variables edad, peso, altura e IMC que una lesión sea muscular o no?

Al igual que en la pregunta anterior, se ha transformado la variable 'Tipo' en una variable dicotómica que indicaría si se ha sufrido lesión muscular o no. Al tratarse de una variable que toma dos valores y tener varias variables independientes, se ha optado por utilizar la regresión logística múltiple. Calculada la matriz de confusión, se obtiene que el modelo tiene una precisión del 84,61%, una sensibilidad del 16,67% y una especificidad del 100%. Esta vez se tiene que ninguna de las variables del enunciado explica si se tendrá lesión muscular, pero sí de forma precisa si no se va a tener.

Se ve un gran contraste en la forma de predecir los valores correctos con respecto de los falsos, el cual se debe a que los coeficientes de las variables no son significativos, pero si lo es el valor del coeficiente independiente β_0.

Aunque sea un modelo que por norma general se rechazaría, su modelo resultante quedaría:

$$P = \frac{e^{-1.7790+0.6537*Edad+6.2665*Peso-4.4635*Altura-4.4487*IMC}}{1+e^{-1.7790+0.6537*Edad+6.2665*Peso-4.4635*Altura-4.4487*IMC}} \qquad (4.5.2)$$

4.5.3 Sabiendo el tipo de lesión, mecanismo e IMC, ¿se puede saber si una lesión es grave entendiendo por grave una lesión de 2 o más semanas?

Para plantear el modelo, se ha vuelto a transformar la variable 'Días lesión' en una variable dicotómica, pero con otro criterio, diferenciando los pacientes con más de 14 días de lesión o no. Al utilizar variables categóricas, el modelo logístico múltiple ha tenido que crear una serie de variables dummy para poder explicar el modelo. A diferencia de los dos casos anteriores, este modelo, aunque parezca que predice bien los valores, ninguno de sus coeficientes es significativo y, por tanto, no son válidos ni tan siquiera en el estudio.

5 CONCLUSIONES

Las técnicas de regresión, tal y como se ha visto a lo largo del trabajo, son muy útiles a la hora de sustraer información de una base de datos. En este caso se ha empleado regresión logística en una serie de datos relacionados con el deporte. Después de realizar un estudio sobre tres preguntas que se han planteado sobre la base, se han extraído las siguientes conclusiones para cada una de ellas:

1. La altura, el IMC y haber tenido lesiones previas afectan de forma directa positiva, mientras que el peso del atleta de forma inversa en el número de días de lesión del paciente, o sea, a mayor altura, IMC, al haber tenido lesiones previas y pesar menos se es más propenso a tener una lesión física de más de 10 días de recuperación.

2. La edad y el peso afectan de forma directa positiva y el IMC y la altura, de forma inversa en el tipo de lesión del paciente (muscular o no muscular), aunque en este apartado cabe señalar que ninguno de los coeficientes de las variables es significativo y, por tanto, a pesar de que la capacidad del modelo de predecir los pacientes sin lesión muscular es buena, no predice de forma adecuada cuando tienen lesión muscular. Por tanto, este modelo no sería apropiado.
3. El modelo no se ajusta correctamente y, por ello, se concluye que las variables 'Tipo lesión', 'Mecanismo' e 'IMC' no afectan en la gravedad de la lesión aparentemente.

BIBLIOGRAFÍA

[1] P. Malliaras, J. L. Cook, and P. M. Kent, “Anthropometric risk factors for patellar tendon injury among volleyball players,” *Br. J. Sports Med.*, vol. 41, no. 4, pp. 259–263, 2007, doi: 10.1136/bjsm.2006.030049.

[2] J. E. Gaida, J. L. Cook, S. L. Bass, S. Austen, and Z. S. Kiss, “Are unilateral and bilateral patellar tendinopathy distinguished by differences in anthropometry, body composition, or muscle strength in elite female basketball players?,” *Br. J. Sports Med.*, vol. 38, no. 5, pp. 581–585, 2004, doi: 10.1136/bjsm.2003.006015.

[3] R. D. Herbert and M. Gabriel, “Soreness and Risk of Injury : Systematic Review,” *Br. Med. J.*, vol. 325, no. August, pp. 1–5, 2002, doi: 10.1136/bmj.325.7362.468.

[4] L. E. Cain, L. L. Nicholson, R. D. Adams, and J. Burns, “Foot morphology and foot/ankle injury in indoor football,” *J. Sci. Med. Sport*, vol. 10, no. 5, pp. 311–319, 2007, doi: 10.1016/j.jsams.2006.07.012.

[5] B. Cueto-Martín, J. C. de la Cruz-Márquez, A. de la Cruz-Campos, and J. C. de la Cruz-Campos, “Prediction of Sports Injuries by Mathematical Models,” *Int. J. Sport Soc. Annu. Rev.*, vol. 3, no. 4, pp. 167–183, 2014, doi: 10.18848/2152-7857/cgp/v03i04/53955.

[6] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, vol. 821. John Wiley & Sons, 2012.

[7] A. Gonzalez, “Selección de variables: Una revisión de métodos existentes,” *Univ. da Coruña*, 2015.

[8] A. Schiaffino, M. Rodríguez, M. I. Pasarín, E. Regidor, C. Borrell, and E. Fernández, “¿Odds ratio o razón de proporciones? Su utilización en estudios transversales,” *Gac. Sanit.*, vol. 17, no. 1, pp. 70–74, 2003, doi: 10.1016/s0213-9111(03)71694-x.

[9] A. G. Lalkhen and A. McCluskey, “Clinical tests: Sensitivity and specificity,” *Contin. Educ. Anaesthesia, Crit. Care Pain*, vol. 8, no. 6, pp. 221–223, 2008, doi: 10.1093/bjaceaccp/mkn041.

ANEXOS

Código

Ejemplo aplicación regresión lineal simple

```
#Ahora vamos a aplicar utilizando las herramientas de RStudio la
#regresión lineal entre las variables que sospechamos por el
#diagrama de dispersión que puede existir una relación lineal entre
#ambas.

#Primero vamos a dividir nuestros datos en dos datasets: Training
#set y Test set
#install.packages('caTools')
library(caTools)
split <- sample.split(base_de_datos_sin_tratar$Altura,
                      SplitRatio = 0.8)

training_set <- data.frame(subset(base_de_datos_sin_tratar,
                                  split == TRUE))
test_set <- data.frame(subset(base_de_datos_sin_tratar,
                              split == FALSE))

############################################
#Escalo todas las variables (no haría falta ya que lm escala
#automaticamente)
#training_set <- scale(training_set)
#test_set <- scale(test_set)
#training_set <- data.frame(training_set)
#test_set <- data.frame(test_set)
############################################

regresion_lineal <- lm(formula = Altura ~ Peso,
                       data = training_set)
summary(regresion_lineal)
#Se ve que el p_valor<0.05. Por tanto, aceptamos la regresión.

#Una vez compruebo que la regresión tiene un p_valor < 0.05, es
#decir, se ajusta al modelo correctamente: voy a predecir usando la
#regresión lineal calculada arriba los datos dentro del conjunto
#test.

altura_pred <- predict(regresion_lineal, newdata = test_set)
View(altura_pred)
realidad_vs_prediccion <- data.frame(cbind(test_set$Altura,
                                           altura_pred))
View(realidad_vs_prediccion)

#Ahora voy a visualizar el diagrama de dispersión junto a la recta
#de regresion lineal calculada para mostrar cómo se ajusta al
#sistema:

#install.packages('ggplot2')
library(ggplot2)
```

```
#Grafica de la regresión lineal aplicada al conjunto de
#entrenamiento (Training set)
ggplot() +
  geom_point(aes(x = training_set$Peso,
                 y = training_set$Altura),
             colour = 'red') +
  geom_line(aes(x = training_set$Peso,
                y = predict(regresion_lineal, newdata =
                                                 training_set)),
            colour = 'blue') +
  ggtitle('Altura vs Peso (Training set)') +
  xlab('Peso') +
  ylab('Altura')

#Grafica de la regresión lineal aplicada al conjunto test (Test
#set)
ggplot() +
  geom_point(aes(x = test_set$Peso,
                 y = test_set$Altura),
             colour = 'red') +
  geom_line(aes(x = test_set$Peso,
                y = predict(regresion_lineal, newdata = test_set)),
            colour = 'blue') +
  ggtitle('Altura vs Peso (Test set)') +
  xlab('Peso') +
  ylab('Altura')
```

Tratamiento de la base de datos antes de realizar las preguntas

```
#Instalo la librería que voy a utilizar para leer el Excel con la
#base de datos.
install.packages('readxl')

#Abro la libreria para poder utilizarla
library(readxl)

#Cargo la base de datos donde se encuentre en el excel en el equipo
#ir a import dataset a la izquierda en enviroment:
base_de_datos_sin_tratar  <- read_excel("base de datos sin
tratar.xlsx")
#base_de_datos_sin_tratar <-
read_excel("C:/Users/Victor/Desktop/base de datos sin tratar.xlsx")
View(base_de_datos_sin_tratar)

#Ahora voy a cambiar los nombres de algunas variables para que sea
#más sencillo tratar con ellas.
names(base_de_datos_sin_tratar)

names(base_de_datos_sin_tratar)[4] <-  'Peso'
names(base_de_datos_sin_tratar)[5] <-  'Altura'
names(base_de_datos_sin_tratar)[12] <- 'Localizacion'
names(base_de_datos_sin_tratar)[16] <- 'DiasRecuperacion'

#Voy a eliminar algunas variables: no aportan información, la
#información que aportan es innecesaria o son difíciles de
#interpretar
base_de_datos_sin_tratar$Paciente <- NULL
```

```
base_de_datos_sin_tratar$Sexo <- NULL
base_de_datos_sin_tratar$Dominancia <- NULL
base_de_datos_sin_tratar$Empleo <- NULL
base_de_datos_sin_tratar$Especialidad <- NULL
base_de_datos_sin_tratar$Unidad <- NULL
base_de_datos_sin_tratar$Localizacion <- NULL
base_de_datos_sin_tratar$Lugar <- NULL


#Voy a eliminar observaciones repetidas y observaciones excluidas
#por la propia base de datos

repetidos <- c(79:93)
base_de_datos_sin_tratar <- base_de_datos_sin_tratar[-repetidos,]

excluidos <- c(79:82)
base_de_datos_sin_tratar <- base_de_datos_sin_tratar[-excluidos,]

#Voy a transformar las variables cualitativas para que sean
#más fáciles de interpretar.

#La variable Antecedentes la transformare en una variable
#dicotómica:

for (i in 1:length(base_de_datos_sin_tratar$Antecedentes)){
  if (is.na(base_de_datos_sin_tratar$Antecedentes[i]) == T){
    base_de_datos_sin_tratar$Antecedentes[i] <- '0'
  }
  else{
    base_de_datos_sin_tratar$Antecedentes[i] <- '1'
  }
}

base_de_datos_sin_tratar$Antecedentes <-
factor(base_de_datos_sin_tratar$Antecedentes, levels = c('0','1'),
                                              labels = c('0','1'))

base_de_datos_sin_tratar$Tipo <-
factor(base_de_datos_sin_tratar$Tipo, levels = c('Cartilaginosa',
                                                 'Ligamentosa',
                                                 'Múltiple',
                                                 'Muscular',
                                                 'Ósea',
                                                 'Tendinosa'),
                                      labels = c(1,2,3,4,5,6))

base_de_datos_sin_tratar$Mecanismo <-
factor(base_de_datos_sin_tratar$Mecanismo, levels=c('Combinación',
                                                    'Compresión',
                                                    'Flexión',
                                                    'Torsión'),
                                           labels = c(1,2,3,4))
summary(base_de_datos_sin_tratar)

#Ahora vamos a estudiar los diagramas de dispersión por pares
#de cada una de las variables para estudiar de forma grafica
#posibles relaciones entre variables:
pairs(base_de_datos_sin_tratar)
```

Pregunta 1: ¿Se puede predecir si un paciente tendrá más de 10 días de lesión sabiendo su edad, su peso, su altura, su IMC y si ha tenido antecedentes previos?

```
#Voy a copiar la base de datos en otra para responder la pregunta
#con las variables que se involucran en el enunciado.
datos <- data.frame(base_de_datos_sin_tratar$Edad,
                    base_de_datos_sin_tratar$Peso,
                    base_de_datos_sin_tratar$Altura,
                    base_de_datos_sin_tratar$IMC,
                    base_de_datos_sin_tratar$Antecedentes,
                    base_de_datos_sin_tratar$DiasRecuperacion)
names(datos)[1] <- 'Edad'
names(datos)[2] <- 'Peso'
names(datos)[3] <- 'Altura'
names(datos)[4] <- 'IMC'
names(datos)[5] <- 'Antecedentes'
names(datos)[6] <- 'DiasRecuperacion'

#Creo una nueva variable dicotómica 'DiasPredichos' utilizando la
#variable DiasRecuperacion de tal forma que valdrá 1 si supera los
#10 días de recuperación y 0 en caso contrario.
DiasPredichos <- c()
for (i in 1:length(datos$DiasRecuperacion)){
  if (datos$DiasRecuperacion[i] < 10){
    DiasPredichos[i] <- 0
  }
  else{
    DiasPredichos[i] <- 1
  }
}
datos <- data.frame(datos, DiasPredichos)

datos$DiasPredichos <- as.factor(datos$DiasPredichos)
summary(datos)

#Elimino la variable DiasRecuperacion ya que no se va a utilizar
#para la regresion logistica.
datos$DiasRecuperacion <- NULL

#Escalo los datos para que no haya problemas de ajuste
datos[,1:4] <- scale(datos[,1:4])

#Creo el modelo utilizando todas las variables
modelo_glm <- glm(formula = DiasPredichos ~ ., family = binomial,
                  data = datos)
summary(modelo_glm)

#Se ve claramente que el coeficiente de la Edad no es significativo
modelo_glm <- glm(formula = DiasPredichos ~ Peso + Altura + IMC +
Antecedentes,
                  family = binomial,
                  data = datos)
summary(modelo_glm)

library(MASS)
```

```
#Intervalo de confianza
confint(modelo_glm)

#Intervalo de confianza basado en el error estándar
confint.default(modelo_glm)

#Diferencia de los residuos
dif_residuos <- modelo_glm$null.deviance - modelo_glm$deviance
dif_residuos

#Grados de diferencia
df <- modelo_glm$df.null - modelo_glm$df.residual
df

#P-Valor
p_value <- pchisq(q = dif_residuos, df = df, lower.tail = F)
p_value


#Matriz de confusión
predicciones <- ifelse(test = modelo_glm$fitted.values > 0.5,
                       yes = 1,
                       no = 0)
matriz_confusion <- table(modelo_glm$model$DiasPredichos,
                          predicciones,
                          dnn = c('Obsevaciones', 'Predicciones'))
matriz_confusion
```

Pregunta 2: ¿Depende que la lesión sea muscular o no de la edad, IMC, peso y antecedentes previos?

```
datos2 <- data.frame(base_de_datos_sin_tratar$Edad,
                     base_de_datos_sin_tratar$Peso,
                     base_de_datos_sin_tratar$Altura,
                     base_de_datos_sin_tratar$Tipo,
                     base_de_datos_sin_tratar$IMC)
names(datos2)[1] <- 'Edad'
names(datos2)[2] <- 'Peso'
names(datos2)[3] <- 'Altura'
names(datos2)[4] <- 'TipoLesion'
names(datos2)[5] <- 'IMC'

#Ahora voy a transformar la variable TipoLesion en una variable
#dicotómica
#que indicara si tiene o no tiene lesión muscular un paciente, es
#decir,
#valdra 1 si el paciente tiene Lesión muscular o 0 si el paciente
#no la tiene.
```

```
LesionMuscular <- c()
for (i in 1:length(datos2$TipoLesion)){
  if (datos2$TipoLesion[i] == 4){
    LesionMuscular[i] <- 1
  }
  else{
    LesionMuscular[i] <- 0
  }
}
datos2 <- data.frame(datos2, LesionMuscular)

datos2$TipoLesion <- as.factor(datos$TipoLesion)

#Elimino la variable TipoLesion ya que no aporta nada más.
datos2$TipoLesion <- NULL

#Escalamos las variables
datos2[,1:4] <- scale(datos2[,1:4])

#Creamos el modelo de regresion logistica
modelo_glm2 <- glm(formula = LesionMuscular ~.,
                   family = binomial,
                   data = datos2)
summary(modelo_glm2)

#Intervalo de confianza
confint(modelo_glm2)

#Intervalo de confianza basado en el error estándar
confint.default(modelo_glm2)

#Diferencia de los residuos
dif_residuos2 <- modelo_glm2$null.deviance - modelo_glm2$deviance
dif_residuos2

#Grados de diferencia
df2 <- modelo_glm2$df.null - modelo_glm2$df.residual
df2

#P-Valor
p_value2 <- pchisq(q = dif_residuos2, df = df2, lower.tail = F)
p_value2

#Matriz de confusión
predicciones2 <- ifelse(test = modelo_glm2$fitted.values > 0.5,
                        yes = 1,
                        no = 0)
matriz_confusion <- table(modelo_glm2$model$LesionMuscular,
                          predicciones2,
                          dnn = c('Obsevaciones', 'Predicciones'))
matriz_confusion
```

Pregunta 3: Sabiendo el tipo de lesión, mecanismo e IMC, ¿se puede saber si una lesión es grave, entendiendo por grave una lesión que necesite 2 o más semanas de recuperación?

```
datos3 <- data.frame(base_de_datos_sin_tratar$IMC,
                     base_de_datos_sin_tratar$Tipo,
                     base_de_datos_sin_tratar$Mecanismo,
                     base_de_datos_sin_tratar$DiasRecuperacion)
names(datos3)[1] <- 'IMC'
names(datos3)[2] <- 'TipoLesion'
names(datos3)[3] <- 'Mecanismo'
names(datos3)[4] <- 'DiasRecuperacion'

#Vuelvo a crear una nueva variable dicotómica 'LesionGrave' a
#partir de la variable DiasRecuperacion de tal forma que valdrá 1
#si la lesión necesita más de 2 semanas de recuperación y 0 si no
#es el caso.
LesionGrave <- c()
for(i in 1:length(datos3$DiasRecuperacion)){
  if (datos3$DiasRecuperacion[i] > 14){
    LesionGrave[i] <- 1
  }
  else{
    LesionGrave[i] <- 0
  }
}
datos3 <- data.frame(datos3, LesionGrave)

datos3$LesionGrave <- as.factor(datos3$LesionGrave)


#Elimino la variable DiasRecuperacion ya que no aporta nada
datos3$DiasRecuperacion <- NULL
summary(datos3)

#Creo el modelo de regresión logística
modelo_glm3 <- glm(formula = LesionGrave ~.,
                   family = binomial,
                   data = datos3)
summary(modelo_glm3)

#Matriz de confusión
predicciones3 <- ifelse(test = modelo_glm3$fitted.values > 0.5,
                        yes = 1,
                        no = 0)
matriz_confusion <- table(modelo_glm3$model$LesionGrave,
                          predicciones3,
                          dnn = c('Obsevaciones', 'Predicciones'))
matriz_confusion
```

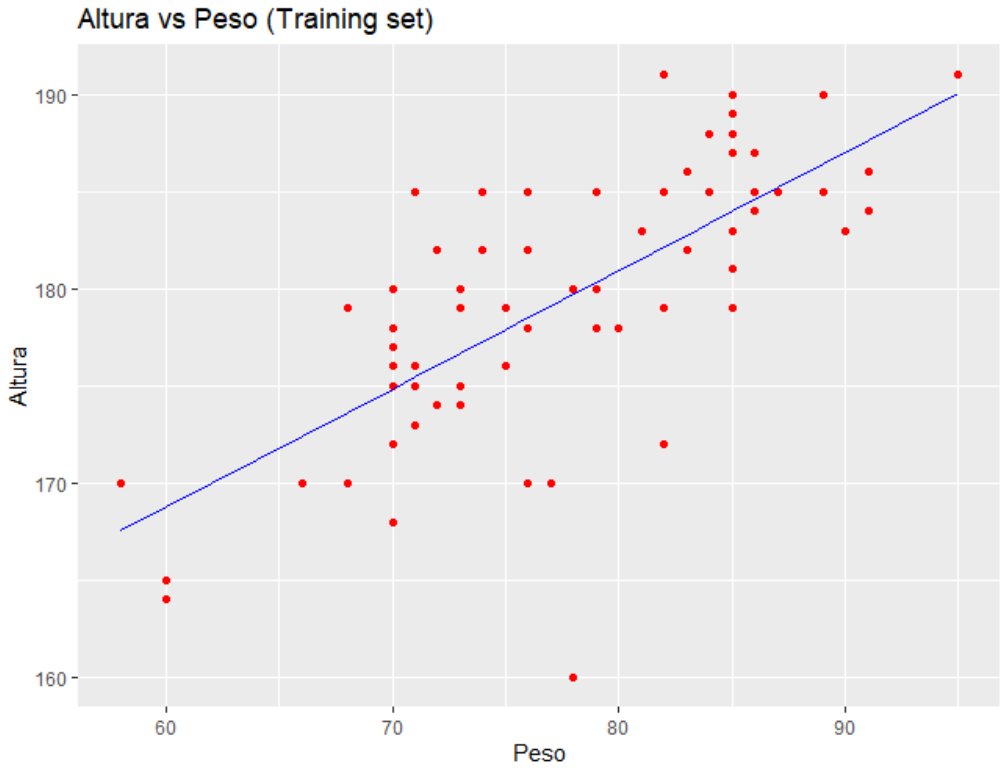
Gráficas

Ejemplo aplicación regresión lineal simple

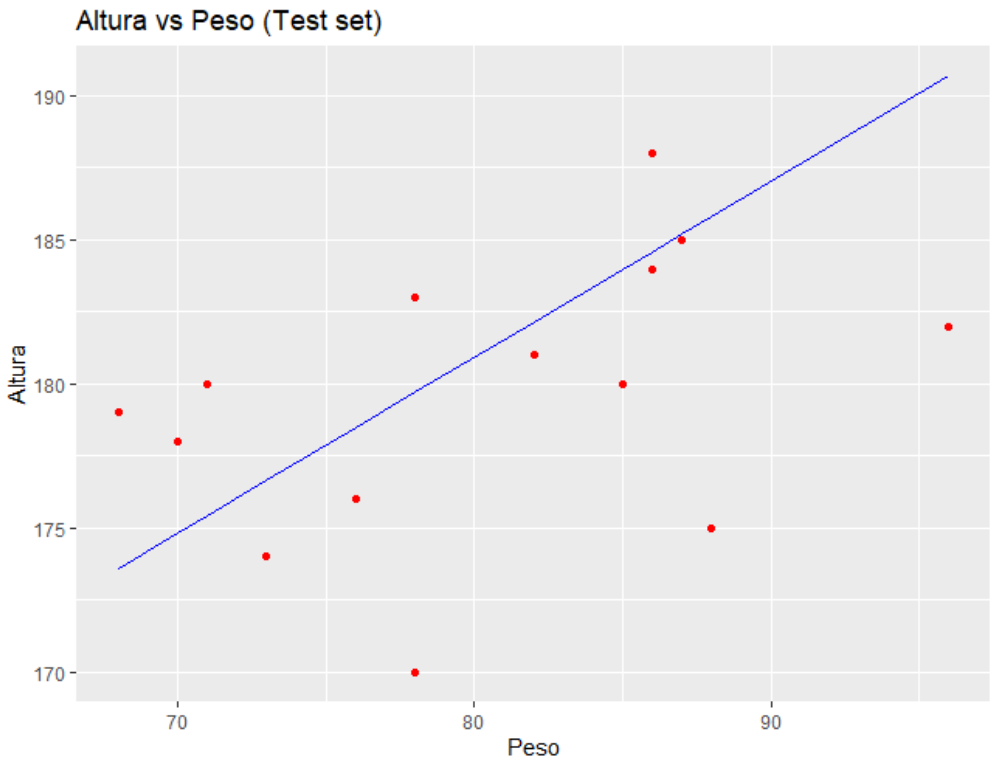
Gráfica A: Resumen información de la regresión lineal

```
Call:
lm(formula = Altura ~ Peso, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7108  -2.6638   0.2892   3.1645   9.5551

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 132.17656    5.70866  23.154  < 2e-16 ***
Peso          0.60941    0.07289   8.361 5.38e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.737 on 67 degrees of freedom
Multiple R-squared:  0.5106,    Adjusted R-squared:  0.5033
F-statistic:  69.9 on 1 and 67 DF,  p-value: 5.384e-12
```

Gráfica B: Regresión lineal sobre el training set



Gráfica C: Regresión lineal sobre el test set

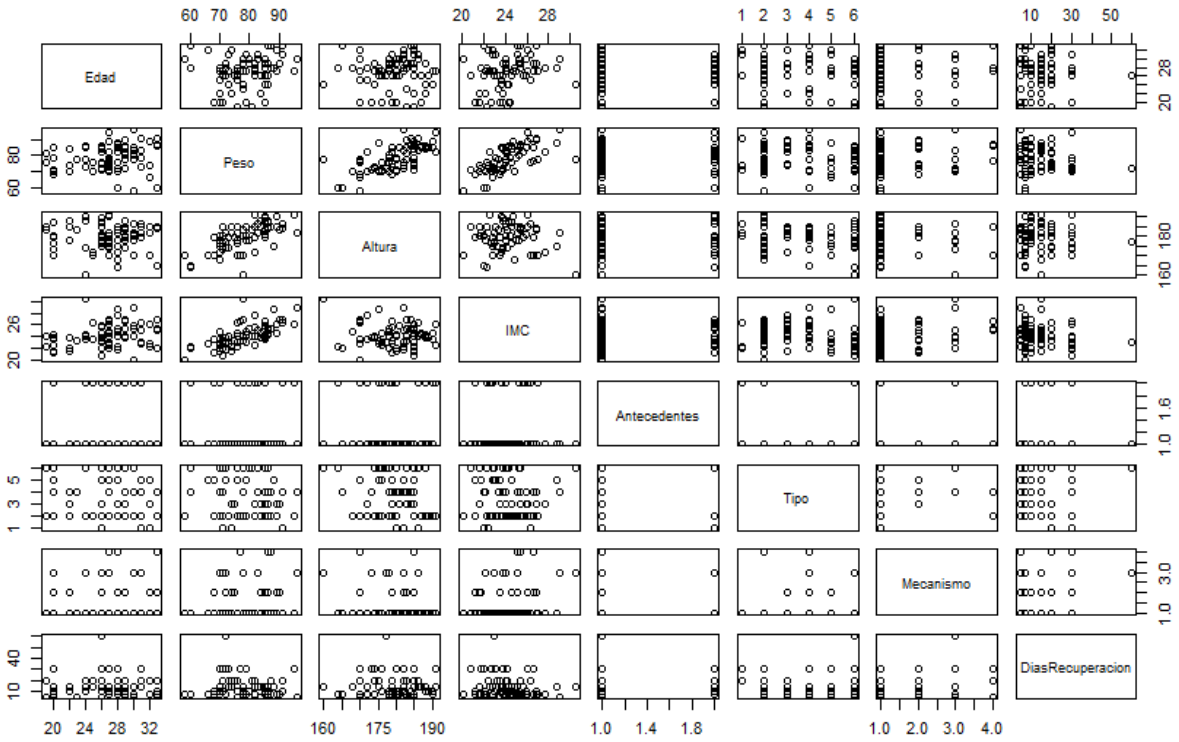


Tratamiento de la base de datos antes de realizar las preguntas

Gráfica D: Resumen de las variables involucradas en el estudio

Edad	Peso	Altura	IMC	Antecedentes	Tipo	Mecanismo	DiasRecuperacion
Min. :19.00	Min. :58.00	Min. :160.0	Min. :20.07	0:62	1: 3	1:57	Min. : 5.00
1st Qu.:25.25	1st Qu.:72.00	1st Qu.:176.0	1st Qu.:22.93	1:16	2:28	2:10	1st Qu.: 7.00
Median :27.00	Median :78.00	Median :180.0	Median :24.07		3: 9	3: 7	Median :10.00
Mean :26.88	Mean :78.41	Mean :179.8	Mean :24.24		4:14	4: 4	Mean :13.83
3rd Qu.:29.00	3rd Qu.:85.00	3rd Qu.:185.0	3rd Qu.:25.40		5: 8		3rd Qu.:18.75
Max. :33.00	Max. :96.00	Max. :191.0	Max. :30.47		6:16		Max. :60.00

Gráfica E: Diagrama de dispersión de todas las variables enfrentadas por pares



Pregunta 1

Gráfica F: Resumen variables

Edad	Peso	Altura	IMC	Antecedentes	DiasRecuperacion	DiasPredichos
Min. :19.00	Min. :58.00	Min. :160.0	Min. :20.07	0:62	Min. : 5.00	0:30
1st Qu.:25.25	1st Qu.:72.00	1st Qu.:176.0	1st Qu.:22.93	1:16	1st Qu.: 7.00	1:48
Median :27.00	Median :78.00	Median :180.0	Median :24.07		Median :10.00	
Mean :26.88	Mean :78.41	Mean :179.8	Mean :24.24		Mean :13.83	
3rd Qu.:29.00	3rd Qu.:85.00	3rd Qu.:185.0	3rd Qu.:25.40		3rd Qu.:18.75	
Max. :33.00	Max. :96.00	Max. :191.0	Max. :30.47		Max. :60.00	

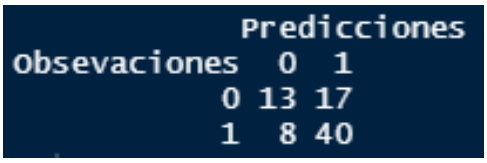
Gráfica G: Resumen del modelo logístico

```
Call:
glm(formula = DiasPredichos ~ ., family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3009  -1.1297   0.5394   0.9623   1.5254

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.26560    0.27348   0.971   0.3314
Edad          -0.09158    0.28960  -0.316   0.7518
Peso         -12.29304    5.47753  -2.244   0.0248 *
Altura         9.08284    3.92971   2.311   0.0208 *
IMC            9.03548    4.09725   2.205   0.0274 *
Antecedentes1  1.55160    0.78202   1.984   0.0472 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 103.939  on 77  degrees of freedom
Residual deviance:  90.231  on 72  degrees of freedom
AIC: 102.23

Number of Fisher Scoring iterations: 4
```

Gráfica H:Resumen del modelo logístico sin la variable edad

```
Call:
glm(formula = DiasPredichos ~ Peso + Altura + IMC + Antecedentes,
    family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2688  -1.1453   0.5474   0.9709   1.4982

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.2648     0.2731   0.970   0.3323
Peso         -12.8260     5.2162  -2.459   0.0139 *
Altura         9.4568     3.7498   2.522   0.0117 *
IMC            9.4025     3.9296   2.393   0.0167 *
Antecedentes1  1.5723     0.7813   2.012   0.0442 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 103.939  on 77  degrees of freedom
Residual deviance:  90.331  on 73  degrees of freedom
AIC: 100.33

Number of Fisher Scoring iterations: 4
```

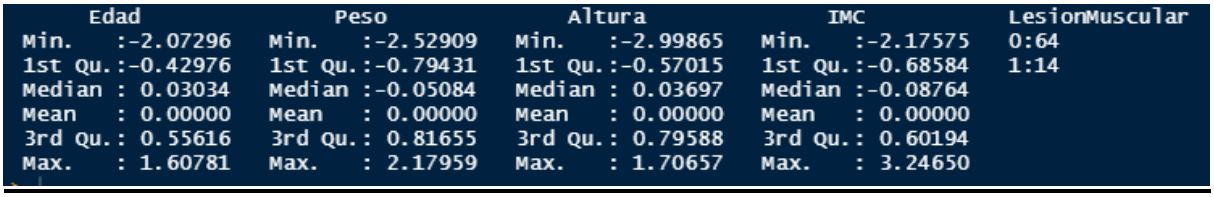
Gráfica I: Matriz de confusión del modelo logístico



```
                Predicciones
Obsevaciones  0  1
           0 13 17
           1  8 40
```

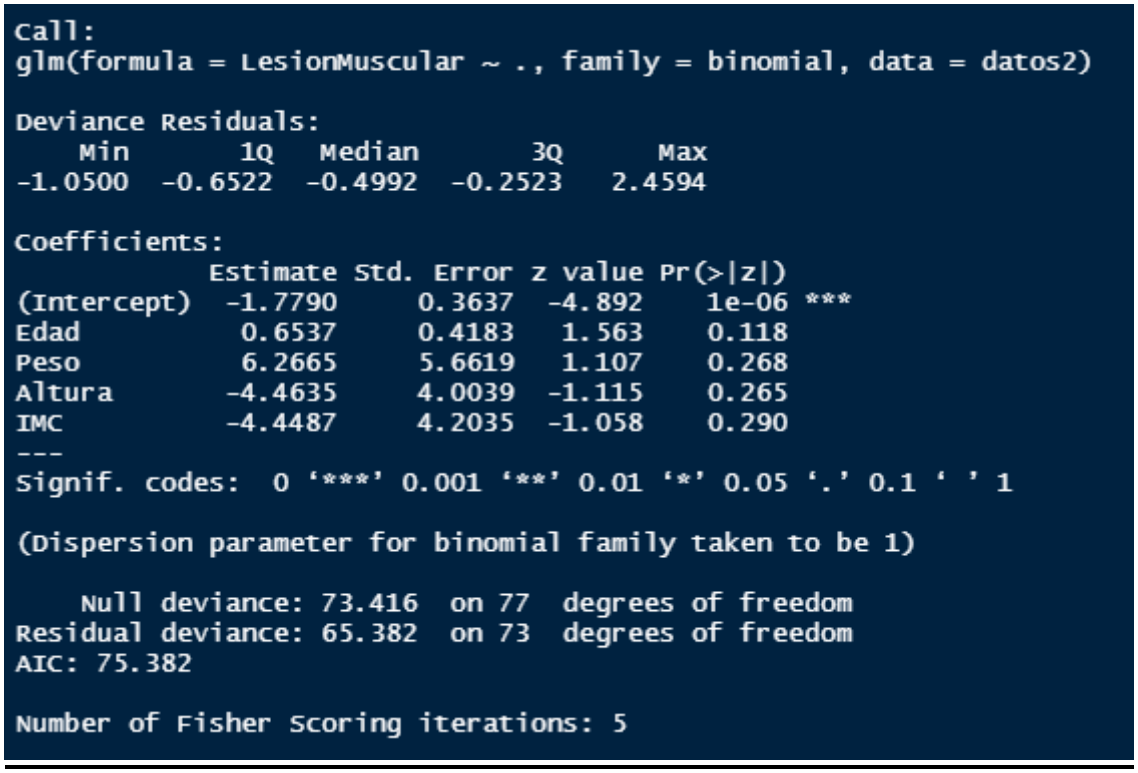
Pregunta 2

Gráfica J: Resumen de las variables



```
      Edad               Peso               Altura              IMC              LesionMuscular
 Min.   :-2.07296   Min.   :-2.52909   Min.   :-2.99865   Min.   :-2.17575   0:64
 1st Qu.:-0.42976   1st Qu.:-0.79431   1st Qu.:-0.57015   1st Qu.:-0.68584   1:14
 Median : 0.03034   Median :-0.05084   Median : 0.03697   Median :-0.08764
 Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.55616   3rd Qu.: 0.81655   3rd Qu.: 0.79588   3rd Qu.: 0.60194
 Max.   : 1.60781   Max.   : 2.17959   Max.   : 1.70657   Max.   : 3.24650
```

Gráfica K: Resumen del modelo logístico



```
Call:
glm(formula = LesionMuscular ~ ., family = binomial, data = datos2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0500  -0.6522  -0.4992  -0.2523   2.4594

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7790     0.3637  -4.892    1e-06 ***
Edad          0.6537     0.4183   1.563    0.118
Peso          6.2665     5.6619   1.107    0.268
Altura       -4.4635     4.0039  -1.115    0.265
IMC          -4.4487     4.2035  -1.058    0.290
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.416  on 77  degrees of freedom
Residual deviance: 65.382  on 73  degrees of freedom
AIC: 75.382

Number of Fisher Scoring iterations: 5
```

Gráfica L: Matriz de confusión del modelo logístico

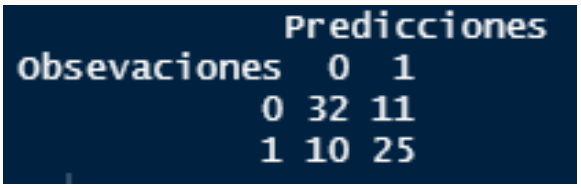
Obsevaciones	Predicciones 0	1
0	64	0
1	12	2

Pregunta 3

Gráfica M: Resumen de las variables

IMC	TipoLesion	Mecanismo	LesionGrave
Min. :20.07	1: 3	1:57	0:43
1st Qu.:22.93	2:28	2:10	1:35
Median :24.07	3: 9	3: 7	
Mean :24.24	4:14	4: 4	
3rd Qu.:25.40	5: 8		
Max. :30.47	6:16		

Gráfica N: Resumen del modelo logístico

```
Call:
glm(formula = LesionGrave ~ ., family = binomial, data = datos3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5315  -0.8668   0.0000   1.0118   1.7422

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  21.98393 6188.62727   0.004    0.997
IMC          -0.10218    0.16821  -0.607    0.544
TipoLesion2 -19.13088 6188.62597  -0.003    0.998
TipoLesion3 -20.14983 6188.62604  -0.003    0.997
TipoLesion4 -53.39227 7041.17092  -0.008    0.994
TipoLesion5 -20.11149 6188.62602  -0.003    0.997
TipoLesion6 -20.78275 6188.62600  -0.003    0.997
Mecanismo2    0.04818    1.07592   0.045    0.964
Mecanismo3   33.95825 3358.42190   0.010    0.992
Mecanismo4   17.39758 2334.59368   0.007    0.994

(Dispersion parameter for binomial family taken to be 1)
```

Gráfica O: Matriz de confusión del modelo logístico