Trabajo Fin de Máster

Modelos de regresión multivariante y su aplicación en datos médicos

Sara Rodríguez Pastoriza

Máster en Técnicas estadísticas

2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Modelos de regresión multivariante e a súa aplicación en datos médicos
Título en español: Modelos de regresión multivariante y su aplicación en datos médicos
English title: Multivariate regression models and their application in medical data
Modalidad: Modalidad B
Autora: Sara Rodríguez Pastoriza, Universidade de Santiago de Compostela
Directora: Ana Clavería Fontán, Instituto de Investigación sanitaria Galicia Sur.
Tutor: Javier Roca Pardiñas, Universidade de Vigo.
Breve resumen del trabajo: Se realizará un análisis de diferentes modelos de regresión multivariante y de su aplicación en casos médicos reales. En primer lugar, se profundizará en la metodología estadística de cada uno de los modelos considerados para después estudiar su aplicación y utilidad en casos reales de la medicina más actual.

Doña Ana Clavería Fontán, epidemióloga especialista en Atención Primaria y Comunitaria responsable del Instituto de Investigación Sanitaria Galicia Sury don Javier Roca Pardiñas, profesor y estadístico de la Universidade de Vigo, informan que el Trabajo Fin de Máster titulado

Modelos de regresión multivariante y su aplicación en datos médicos

fue realizado bajo su dirección por doña Sara Rodríguez Pastoriza para el Máster en Técnicas estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Vigo, a 27 de Enero de 2023.

La directora:
Doña Ana Clavería Fontán

El tutor:
Don Javier Roca Pardiñas

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Me gustaría dedicar este trabajo a mi abuelo, por sus ganas infinitas de que me formase y aprendiese todo lo posible, de que me convirtiese en la persona que realmente buscase ser y de que aprovechase todas las oportunidades habidas y por haber.
Agradecimiento absoluto a mis padres y a mi abuela, por todo el apoyo incondicional, por haberme acompañado en cada paso de esta etapa y de todas las anteriores, y por haber confiado en mi en tantos momentos en los que dudaba de mi misma. Esto es gracias a vosotros, ¡gracias!

Gracias a Ana Clavería Fontán, directora de este trabajo fin de máster, por iniciarme en el mundo de la investigación sanitaria con tanta paciencia, vocación y cariño, por descubrirme un mundo sinfín que desconocía y que realmente me apasiona y por transmitirme tantos conocimientos que sin duda serán de gran utilidad en mi futuro, gracias por esta gran oportunidad. También a Javier Roca Pardiñas, tutor académico de este trabajo, por orientarme en la planificación de esta memoria, por todas las recomendaciones y por transmitirme su amplio conocimiento estadístico.

Gracias también a Manuel Modroño Beires-Días, por todos sus consejos en el análisis estadístico y por el apoyo incondicional e ilimitado, tanto a nivel académico como personal, durante toda esta etapa, ¡gracias de corazón!

A todos mis compañeros de trabajo, por haberme hecho sentir como en casa desde el primer día. En especial, a Macarena Chacón Docampo y a Wilfredo Guanipa Sierra, por adoptar una actitud tan profesional, didáctica y cercana hacia mi y de la que tanto aprendo día a día, ¡gracias!

Gracias también a mis amigas Sonia Gil Coello, Beatriz María Pérez González, María Fuhong Toyos Pedrayes y Julia García Piñeiro, que se han mantenido ahí durante toda la vida y que se han convertido en una parte permanente e imprescindible para mi, ¡gracias por siempre! Finalmente, a los que me han acompañado desde mi primera entrada en la Facultad de Matemáticas de Santiago: Andrés Pérez Rodríguez, Iria Rodríguez Acevedo y Brais Ramos Pérez, con los que he compartido grandes momentos y los que, sin duda, llegarán muy alto.

Índice general

Resumen

Resumen en español

En este trabajo estudiaremos los principales modelos de regresión multivariante, en los cuales se considera una variable resultado y más de una variable predictora. Así, presentaremos en detalle los modelos lineales (LM), los modelos lineales generalizados (GLM), los modelos aditivos generalizados (GAM) y los modelos aditivos generalizados para posición, forma y escala (GAMLSS), estudiando sus principales características. Además, se estudiarán las métricas de evaluación más relevantes en la actualidad, tanto para evaluar la bondad de ajuste, entre las que destacan el criterio de información de Akaike o Bayesiano, como para la capacidad de clasificación, entre las que destaca el área bajo la curva característica de funcionamiento del receptor (AUC). Finalmente, presentaremos dos casos médicos, uno terapéutico y otro clínico. El primero de ellos es un estudio sobre la efectividad y seguridad de la vacuna antitetánica en pacientes tratados con anticoagulantes orales, y el segundo un caso clínico sobre el riesgo de sobrepeso u obesidad en niños y adolescentes a partir del porcentaje de masa grasa, considerando medidas antropométricas. En ambos, la metodología estadística aplicada fue imprescindible para llegar a conclusiones relevantes en el mundo de la investigación sanitaria.

English abstract

In this paper we will study the main multivariate regression models, in which one outcome variable and more than one predictor variable are considered. Thus, we will present in detail linear models (LM), generalized linear models (GLM), generalized additive models (GAM) and generalized additive models for position, shape, and scale (GAMLSS), studying their main characteristics. In addition, the most relevant evaluation metrics at present will be studied, both to evaluate the goodness of fit, among which the Akaike or Bayesian information criterion stands out, and for the classification capacity, among which the area under the evaluation stands out. receiver operating characteristic curve (AUC). Finally, we will present two medical cases, one therapeutic and the other clinical. The first of these is a study on the effectiveness and safety of the tetanus vaccine in patients treated with oral anticoagulants, and the second is a clinical case on the risk of being overweight or obese in children and adolescents based on the percentage of fat mass, considering measures anthropometric. In both, these statistical techniques were very useful for reaching relevant conclusions in the world of health research.

Introducción

La estadística es una ciencia matemática que tiene una gran utilidad en la mayoría de las ciencias humanas y naturales, dando solución a multitud de problemas de la vida real. En particular, las diferentes técnicas estadísticas que hoy en día se conocen juegan un papel fundamental en el campo de la medicina, siendo utilizadas, por ejemplo, para predecir el desarrollo de un tipo de cáncer, para evaluar nuevas estrategias de prevención y control de una determinada patología, orientar los procesos de toma de decisiones, y planificar y ayudar en la gestión y el cuidado de la salud de las personas. Este vínculo entre la estadística y la medicina constituye una verdadera necesidad social, que requiere de la búsqueda y estudio de nuevas herramientas encaminadas a prolongar la vida de las personas, mejorar su calidad y promover su adecuación.

El primer evento en el que se relacionó la medicina con la estadística tuvo lugar a mediados del siglo XVII de la mano del demógrafo John Graunt, considerado además el fundador de la bioestadística y el precursor de la epidemiología. En paralelo a un notable avance científico y tecnológico que se estaba desarrollando por aquel entonces, Graunt analizó los registros de nacimientos y muertes de la ciudad de Londres durante 50 años previos, observando un patrón constante en las causas de las muertes diferenciando entre zonas demográficas (rural y urbana). Así, estimó las causas de las muertes de la población y a qué edad o en qué estación del año ocurrían con mayor frecuencia. También observó que nacían más hombres que mujeres y que el 36 % de la población se moría antes de los 6 años. Con este trabajo nació la idea de lo que hoy conocemos como estudio científico.

En los siguientes años, se diseñaron las conocidas como "leyes de enfermedad", elaboradas por los científicos de la época y cuyo origen, curiosamente, surgió del análisis de la distribución de los nacimientos. Estas leyes informaban acerca de la probabilidad de enfermar a una determinada edad, la probabilidad de permanecer enfermo durante un número específico de días o la probabilidad de morir dependiendo de cada causa determinada.

Este ambiente de avance científico y elaboración de estudios de investigación continuó en el siglo XVIII. Así, fueron publicados multitud de trabajos que también usaban la enumeración estadística, como por ejemplo el de Daniel Bernouilli en 1760, que aseguraba que la vacunación protegía de la viruela confería inmunidad de por vida. Los esfuerzos de los matemáticos de este siglo, como Pierre Laplace o Adolphe Quetelet, dieron inicio a lo que hoy conocemos como la estadística moderna.

A principios del siglo XIX, los emblemáticos Alexander Louis, Laplace, Poisson, Quetelet, Galton y Pearson pensaron y diseñaron técnicas estadísticas muy novedosas, como el método de mínimos cuadrados para encontrar la función continua que mejor se ajuste a los datos, cuyo mérito se atribuyó a Gauss (1809). Además, continuaron las "leyes de la enfermedad", contribuyendo al desarrollo de la estadística moderna. Ya por aquel entonces, se empezó a emplear el término de "estadísticas vitales" para referirse al recuento y almacenamiento sistemático de las características más esenciales, como los nacimientos, las muertes, los matrimonios y los divorcios.

Fue a principios del sigo XX cuando se comenzaron a recoger datos más completos sobre los indivi-

duos, con el objetivo de que mejorase el servicio de salud que se ofrecía. Así, se recogía una información mucho más completa de los mismos, más allá de las características vitales. Estos datos demográficos, económicos y sociales pertenecientes a un tiempo específico y de un área demográfica determinada fueron llamados censos poblacionales. Con el tiempo, surgieron alternativas a los censos, principalmente, las encuestas a sólo una muestra de la población y la posterior generalización de los resultados obtenidos para ella.

Desde entonces, el desarrollo y aplicación de la estadística en el campo de la medicina ha ido en aumento, convirtiéndose hoy en día en algo indispensable para la salud y cuidado de la población.

Además de responder a preguntas de interés sobre los datos o demostrar determinadas hipótesis, algo que resulta de vital importancia en salud es modelar las relaciones entre determinadas patologías y los posibles factores de riesgo asociados a ellas. En este contexto, juegan un papel fundamental los modelos de regresión multivariante, los cuales tendrán mucho protagonismo a lo largo de esta memoria. Estos permiten explicar el comportamiento de una determinada variable dependiente o resultado, denotada por Y, utilizando la información que proporcione un determinado conjunto de variables independientes o predictoras, denotadas por $x_1, x_2, ..., x_p$ siendo p el número total.

Existen diferentes tipos de modelos de regresión multivariante, entre los que destacan en primer lugar los modelos lineales o "linear model" (LM). Estos modelos consideran una serie de hipótesis, las cuales pueden llegar a ser restrictivas en determinadas situaciones, dado principalmente a que la variable respuesta no siempre sigue una distribución normal. Como alternativa, se consideran los modelos generalizados, entre los que se encuentran los modelos lineales generalizados o "generalized linear model" (GLM) y los modelos aditivos generalizados o "generalized additive models" (GAM). Estos modelos no exigen que la variable respuesta siga una distribución normal, ampliando considerablemente las posibilidades a pesar de restringirse a distribuciones pertenecientes a la familia exponencial. Por último, los modelos aditivos generalizados para localización, escala y forma o "generalized additive lodels for location, scale and shape" (GAMLSS) permiten expresar cada uno de los parámetros de la distribución de la variable respuesta a partir de las variables predictoras, mientras que en los modelos anteriores se suponían constantes salvo la media.

Una vez seleccionado el modelo de regresión que consideremos más oportuno, es de vital importancia llevar a cabo una evaluación del mismo a partir de métricas de evaluación, que son muy específicas dependiendo de si la variable respuesta es de naturaleza cuantitativa o cualitativa. Estas técnicas han ganado mayor peso en los últimos años, sobre todo en el caso de modelos de clasificación, con la aparición de la inteligencia artificial.

Respecto a la organización de la memoria, esta se dividirá de la siguiente forma.

En el capítulo 1 estudiaremos los modelos multivariantes más relevantes, explicando con detalle la formulación, las hipótesis, las posibles distribuciones que debe seguir la variable respuesta, el ajuste por mínimos cuadrados y, en el caso de los GLM, los Odds Ratio (OR). Finalmente se explicarán las diferentes métricas de evaluación, las cuales serán diferentes dependiendo de la naturaleza de la variable respuesta. Así, las dividimos en el caso de estar evaluando la bondad del ajuste o la capacidad de clasificación. Para la bondad del ajuste, incluímos el criteiro de información akaike y bayesiano, el R^2 y R^2 ajustado y los gráficos worm (worm plot). Para la capacidad de clasificación, consideramos la sensibilidad, la especificidad, la precisión, el likelihood ratio positivo y negativo, las curvas características operativa del receptor (ROC) y el área bajo la curva asociada (AUC), la puntuación F_1 y los valores predictivos positivos y negativos.

En el capítulo 2, presentamos dos casos médicos reales en los que se aplicó la metodología estadística explicada en el capítulo 1. El primero de ellos, consiste en un ensayo aleatorizado y terapéutico en

el que se quiso evaluar la eficacia y seguridad de la vacuna antitetánica en pacientes tratados con anticoagulantes orales, comparando las vías intramuscular y subcutánea. Por otro lado, el segundo caso es de tipo clínico, un estudio observacional transversal sobre el riesgo de sobrepeso u obesidad a partir del porcentaje de masa grasa en niños y adolescentes, considerando medidas antropométricas de diferentes partes del cuerpo.

Se consultaron las referencias [1], [2], [3] y [4].

Capítulo 1

Metología estadística

En este primer capítulo estudiaremos los modelos de regresión multivariante más utilizados en el análisis estadístico, junto con las métricas de evaluación más relevantes. Veremos en detalle el modelo lineal (LM), el modelo lineal generalizado (GLM), el modelo aditivo generalizado (GAM) y el modelo aditivo generalizado para posición, forma y escala (GAMLSS). Después, analizaremos las métricas de evaluación más conocidas, diferenciando entre la evaluación según la bondad de ajuste y la capacidad de clasificación, este último caso solo si la naturaleza de la variable respuesta es cualitativa binaria.

A lo largo del capítulo haremos referencia a la bibliografía empleada en cada una de las secciones, pero se consultó para la estructura del mismo la referencia [5], el capítulo 6 de la referencia [6] y el capítulo 19 de la referencia [7].

1.1. Introducción a la regresión multivariante

La regresión múltivariante es aquella en la que existe más de una variable predictora. Así, los modelos de regresión multivariante, también conocidos como modelos de regresión múltiple, estudian la relación entre una variable de interés o variable dependiente (respuesta) y un conjunto de variables independientes o explicativas (predictoras), que denotaremos a lo largo de esta memoria por Y y $x_1, x_2, ..., x_p$ respectivamente, donde p será el número total de variables predictoras.

La diferencia principal con el caso más simple en el se estudia la relación entre tan solo dos variables, es decir, considerando una única variable explicativa y una única variable respuesta, es el proceso de búsqueda del modelo que implica un procedimmiento de selección de variables para determinar cuales son las que mejor describen la variable resultado.

A continuación estudiaremos en detalle los modelos más conocidos en el contexto de la regresión multivariante.

1.2. Modelo lineal múltiple (LM)

En este apartado hemos seguido la referencia [8].

El modelo de regresión lineal simple puede generalizarse permitiendo que la variable respuesta dependa de más de una variable explicativa, dando lugar a lo que se conoce como modelo lineal múltiple. Cabe destacar que las variables predictoras de este tipo de modelos pueden ser transformaciones simples de las variables predictoras originales.

1.2.1. Formulación del modelo

El modelo lineal múltiple tiene como fórmula general la siguiente expresión

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \tag{1.1}$$

donde μ denota la media de la variable respuesta $\mathbb{E}(Y)$, $x_i, i = 1, ..., p$, las variables explicativas, $\beta_0, ..., \beta_p$ los coeficientes asociados y ϵ los errores del modelo. Teniendo en cuenta esto, lo que supone realmente el modelo lineal es que la media de la variable respuesta Y se relaciona de forma lineal con las variables explicativas.
Como la información muestral se puede presentar como

$$\begin{matrix} y_1 & x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ y_2 & x_{21} & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & & & \cdot \\ & & & & & \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ y_n & x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{matrix},$$

siendo n el tamaño de la muestra, de la expresión matemática del modelo de regresión lineal múltiple se deduce

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, ..., n,$$

donde μ_i es cada uno de los elementos del vector de medias de la variable respuesta y estamos asumiendo que los errores $\epsilon_1, \epsilon_2, ..., \epsilon_n$ son independientes y siguen una distribución normal de media 0 y varianza σ^2. Además, las variables explicativas son independientes entre sí.

Podemos plantear la formulación del modelo en forma matricial desarrollando la ecuación 1.1 como sigue

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ 1 & x_{21} & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 1 & x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix},$$

que, en notación matricial, se expresa como

$$Y = X\beta + \epsilon.$$

1.2.2. Estimador de mínimos cuadrados

Una vez formulado el modelo, el objetivo será estimar, a partir de la muestra, los coeficientes del modelo, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$, y la varianza del error, $\hat{\sigma}^2$.
Para estimar el vector de parámetros β existen diferentes criterios de estimación y dependiendo del criterio escogido se obtendrán diferentes estimaciones, pero en nuestro caso consideraremos el método de mínimos cuadrados.

En primer lugar, los errores o residuos del modelo, que denotamos por ϵ, se calculan como la diferencia entre los valores observados de la variable respuesta y los valores estimados de dicha variable, es decir,

$$\hat{\epsilon} = Y - \hat{Y} = X\beta - X\hat{\beta}. \tag{1.2}$$

El estimador de mínimos cuadrados minimiza la suma de los cuadrados de los residuos, conocida como la suma residual, la cual, a partir de la ecuación 1.2, se puede reescribir como sigue

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = Y^t Y - 2\hat{\beta}^t X^t Y + \hat{\beta}^t X^t X \hat{\beta}.$$

Por tanto, para obtener los estimadores de los parámetros del modelo bastaría con resolver el siguiente problema de optimización

$$\min_{\hat{\beta}} \hat{\beta}^t \hat{\beta} = \min_{\hat{\beta}} \left(Y^t Y - 2\hat{\beta}^t X^t Y + \hat{\beta}^t X^t X \hat{\beta} \right).$$

Para eso, es suficiente con derivar respecto al vector de parámetros β e igualar a cero, obteniendo un sistema de ecuaciones. La solución a este sistema es lo que recibe el nombre de estimador de mínimos cuadrados ordinario, el cual viene dado en forma matricial como

$$\hat{\beta} = \left(X^t X \right)^{-1} X^t Y.$$

El estimador mínimos cuadrados ordinario escoge, de entre todas las posibles, la recta que minimiza la suma de los cuadrados de las distancias entre cada punto de la nube generada por las observaciones muestrales y el asignado por la recta.

Sin pérdida de generalidad, podemos restringirnos al caso particular en el que $p = 1$, tal y como se hace en la referencia [9], teniendo la siguiente expresión

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ . & . \\ . & . \\ . & . \\ 1 & x_{n1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}.$$

Los residuos asociados van a venir dados para cada observación por

$$\hat{u_i} = y_i - \left(\hat{\beta}_0 - \hat{\beta}_1 x_i \right), \quad i = 1, ..., n.$$

El objetivo sería minimizar la siguiente expresión

$$\sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 - \hat{\beta}_1 x_i 1\right)\right)^2, \quad i = 1, ..., n.$$

La solución a este problema de optimización en el caso particular $p = 1$ será $\hat{\beta}_0, \hat{\beta}_1$, el estimador de mínimos cuadrados ordinarios. Derivando la expresión anterior respecto a β_0 y β_1 e igualándolas a cero, obtenemos lo siguiente

$$-2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0,$$
$$-2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0,$$

siendo su matriz de segundas derivadas

$$\begin{pmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix},$$

y bastaría despejar los coeficientes para obtener los estimadores de mínimos cuadrados ordinarios. Cabe destacar que cuando exista un problema de multicolinealidad, el determinante de la matriz $X^t X$ es cero y por tanto no existe la inversa de esta matriz, por lo que en este caso el estimador $\hat{\beta}$ no podrá ser calculado.

Todos los modelos lineales se pueden expresar de esta forma, incluso los modelos en los que las observaciones están divididas en varios grupos, cuya media se supone distinta. Bastaría con reescribir la notación considerando variables indicadoras resultado de una recodificación.

1.2.3. Hipótesis del modelo

Para llevar a cabo una buena interpretación del modelo, nuestro estudio debe ir acompañado de una buena diagnosis y validación del mismo. Para esto, debemos estudiar si las observaciones cumplen los siguientes supuestos

- **Linealidad**. Dado que la función de regresión en este caso es una recta, se tendrá que cumplir la hipótesis de linealidad. Esta hipótesis significa que la media de la variable respuesta Y crecerá una cantidad fija β_1 cada vez que x incrementa en una unidad. Además, esta suposición hace que estemos ante un modelo paramétrico ya que supone que la función de regresión es una recta, dejando libertad al valor concreto de la pendiente β_1 y la ordenada en el origen β_0, cuyo valor tendremos que estimar a partir de la muestra. Esta suposición se podría relajar en caso de modelo polinómicos o incluso empleando modelos no paramétricos.
- **Homocedasticidad**. Otra de las hipótesis del modelo es que la varianza de los errores es constante, es decir, la varianza del error es la misma cualquiera que sea el valor de la variable explicativa, que en lenguaje matemático se escribiría como:

$$\text{Var}(\epsilon | X = x) = \sigma^2 \quad \text{para todo } x.$$

- **Normalidad**. El error debe tener distribución normal, es decir,

$$\epsilon \in \text{N}(0, \sigma^2).$$

- **Independencia**. Las variables aleatorias que representan los errores, $\epsilon_1, ..., \epsilon_n$, son mutuamente independientes, siendo n el número de observaciones de la muestra. Esta suposición dice que los n errores serían mutuamente independientes.

1.3. Modelo lineal generalizado (GLM)

En esta sección consultamos el capítulo 2 de la referencia [5].

El modelo de regresión lineal supone que la variable resultado sigue una distribución gaussiana. Este supuesto excluye muchos casos. Por ejemplo, la variable resultado podría ser el padecer una determinada enfermedad o no (variable dicotómica), el número de niños en una población (un recuento) o el tiempo que pasa hasta que ocurre un determinado suceso (tiempo que pasa hasta que dispositivo falla).

Los modelos lineales generalizados (GLM) son una generalización flexible de la regresión lineal ordinaria, permitiendo variables respuesta cuyos errores de sus modelos asociados siguen una distribución que pertenece a la familia exponencial, ampliando las posibilidades de la variable respuesta. Así, la media de la variable respuesta ya no está relacionada directamente con las variables explicativas, sino que se relaciona a partir de una función de enlace, conocida como "link function".

1.3.1. Formulación del modelo

Sea n el tamaño muestral, un modelo lineal generalizado tiene una estructura básica de la forma

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i = X_i\beta + \epsilon_i, \quad i = 1, ..., n,$$

donde $\mu_i = E(y_i)$, g es la función link, que será una función suave[1], X_i es la i-ésima fila de la matriz del modelo X, β es un vector de parámetros desconocidos y ϵ es el vector de residuos del modelo.
La formulación del modelo lineal generalizado es entonces la misma que para los modelos lineales, excepto por la presencia de la función link y por las distribuciones de la variable respuesta. Es por esto por lo que muchos de los conceptos y de las ideas generales de los modelos lineales continúan en este caso. Si nos restringimos al caso más simple en el que la función link fuese la identidad y la distribución la Normal, obtendríamos el modelo lineal ordinario.
Como observación, uno de los costes que conlleva esta generalización es que el ajuste del modelo se debe hacer de manera iterativa y los resultados son aproximados, siendo exactos en el caso lineal.

1.3.2. Hipótesis del modelo

Las hipótesis de los modelos GLM son entonces las siguientes:

- Las y_i son independientes para todo i.
- La relación entre la media de la variable respuesta debe ser lineal y constante.
- $y_i \sim$ alguna distribución de la familia exponencial.

La familia exponencial incluye muchas distribuciones que resultan de utilidad en multitud de ocasiones para el entrenamiento de modelos como, por ejemplo, la Poisson, la Binomial, la gamma y la distribución Normal, que explicaremos en detalle a continuación.

1.3.3. Distribuciones de la familia exponencial

Como comentamos anteriormente, en este tipo de modelos y a diferencia de los modelos lineales, ya no es necesario que la variable respuesta siga una distribución normal. Por definición, una distribución pertenece a la familia exponencial de distribuciones si su función de densidad se pueda escribir como

$$f_\theta(y) = \exp[(y\theta - b(\theta))/a(\phi) + c(y, \phi)], \tag{1.3}$$

[1] Una función suave o infinitamente diferenciable es la función que admite derivadas de cualquier orden y, por tanto, todas sus derivadas son continuas.

donde b, a y c representan funciones arbitrarias, ϕ es un parámetro de escala arbitrario, θ es un parámetro canónico de la distribución e y un valor concreto de la variable aleatoria de respuesta. Cabe destacar que θ dependerá completamente del vector de parámetros del modelo.
Por ejemplo, es fácil ver que la distribución normal pertenece a la familia exponencial. Sabemos que la distribución normal tiene como función de densidad

$$\begin{aligned} f_\mu &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - log\left(\sigma\sqrt{2\pi}\right)\right] \\ &= \exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - log\left(\sigma\sqrt{2\pi}\right)\right]. \end{aligned}$$

Bastaría tomar $\theta = \mu, b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}, a(\phi) = \phi = \sigma^2$ y $c\left(\phi, y\right) = \frac{-y^2}{2\phi} - log\left(\sqrt{\phi 2\pi}\right) = \frac{-y^2}{2\sigma^2} - log\left(\sigma\sqrt{2\pi}\right)$ para que quede demostrado lo propuesto.

Es posible obtener expresiones generales para la media y la varianza de distribuciones de la familia exponencial en términos de a, b y ϕ. Expresando la función de densidad en forma genérica (ecuación 1.3), el logaritmo de la función de verosimilitud de θ se obtiene sin más que aplicar el logaritmo a esta expresión, es decir,

$$l(\theta) = [y\theta - b(\theta)]/a(\phi) + c(y, \phi)$$

y, derivando respecto a θ,

$$\frac{\partial l}{\partial \theta} = [y - b'(\theta)]/a(\phi). \tag{1.4}$$

Si consideramos l como una variable aleatoria, sin más que reemplazar la observación particular y por la variable aleatoria Y, podemos expresar el valor esperado de $\frac{\partial l}{\partial \theta}$ de manera que

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = [\mathbb{E}(Y) - b'(\theta)]/a(\phi).$$

Considerando el resultado genérico de que $\mathbb{E}(\frac{\partial l}{\partial \theta}) = 0$, tenemos finalmente que

$$\mathbb{E}(Y) = b'(\theta), \tag{1.5}$$

es decir, la esperanza de cualquier variable aleatoria que siga una distribución de la familia exponencial viene dada por la primera derivada de b evaluada en θ, donde b dependerá de la distribución escogida. Esta ecuación es clave para enlazar los parámetros de un GLM, β, con los parámetros canónicos de la familia exponencial. En un GLM, los parámetros β determinan la media de la variable respuesta, además de determinar también los parámetros canónicos de cada observación en la respuesta.

Una vez formulada la expresión genérica para la media, podemos obtener también la expresión genérica para la varianza. Para eso lo que haremos será diferenciar de nuevo la primera derivada de la función l (ecuación 1.4) obteniendo

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi),$$

y, considerando el resultado general que dice que $\mathbb{E}\left(\partial^2 l/\partial\theta^2\right) = -\mathbb{E}[(\partial l/\partial\theta)^2]$, llegamos a que

$$b''(\theta)/a(\phi) = \mathbb{E}[(y - b'(\theta))^2]/a(\phi)^2,$$

que finalmente da lugar al segundo resultado

$$var(Y) = b''(\theta)a(\phi).$$

Cabe destacar que a podría ser en principio cualquier función de ϕ, y no habría dificultades para manejar cualquier forma de a siendo ϕ conocido. En caso de que dicho parámetro fuese desconocido, las cosas podrían complicarse considerablemente, a menos que podamos escribir $a(\phi)$ de la forma $\frac{\phi}{\omega}$, donde ω fuese una constante conocida. Esta forma cubre todos los casos prácticos de interés, pero en la mayoría de los casos ω es igual a la unidad. Por todo lo anterior, llegamos finalmente a la siguiente expresión para la varianza

$$var(Y) = b''(\theta)\phi/\omega. \tag{1.6}$$

1.3.4. Estimador de mínimos cuadrados ponderados

Recordemos que un modelo GLM ajusta un vector n-dimensional de observaciones independientes de la variable respuesta, Y, donde $\mu = \mathbb{E}(Y)$, tal que

$$g(\mu_i) = X_i\beta + \epsilon$$

e

$$Y \sim F_{\theta_i}(y_i),$$

con $F_{\theta_i}(y_i)$ la distribución de la familia exponencial de parámetros canónicos θ_i. Estos parámetros se determinan a partir de las μ_i y de β. Dado un vector respuesta y, es decir, una observación de la variable aleatoria Y, es posible obtener la estimación de máxima verosimilitud de los parámetros β. Como los Y_i son mutuamente independientes, la función de máxima verosimilitud de β viene dada por

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i).$$

Teniendo en cuenta la expresión para la función de densidad 1.3, el logaritmo de la función de máxima verosimilitud de β es de la forma

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} log[f_{\theta_i}(y_i)] \\ &= \sum_{i=1}^{n} \left[[y_i\theta_i - b_i(\theta_i)]/a_i(\phi) + c_i(\phi, y)\right], \end{aligned}$$

donde la dependencia del lado derecho sobre β es a partir de la dependencia de θ_i sobre β. Nótese que las funciones a, b y c varían en función del subíndice i.

Por comodidad, consideraremos sin pérdida de generalidad solo los casos en los que se puede escribir $a_i(\phi) = \frac{\phi}{\omega_i}$, donde ω_i es una constante conocida, en cuyo caso tendríamos

$$l(\beta) = \sum_{i=1}^{n} \omega_i \left[y_i\theta_i - b_i(\theta_i)\right]/\phi + c_i(\phi, y).$$

Con el fin de optimizar el procedimiento, se puede diferenciar l respecto a cada elemento del vector β, igualando después a cero las expresiones obtenidas y despejando β para obtener la estimación de máxima verosimilitud correspondiente. Así, tenemos que

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}\right),$$

y por la regla de la cadena

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j},$$

por lo que la primera derivada respecto a cada elemento del vector β queda de la forma

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\omega_i\left(y_i\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j} - b_i'(\theta_i)\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}\right). \tag{1.7}$$

Diferenciando respecto a cada elemento de θ (θ_i) la expresión genérica de la esperanza de una variable aleatoria que obtuvimos anteriormente 1.5, llegamos a que

$$\frac{\partial\mu_i}{\partial\theta_i} = b_i\text{“}(\theta_i) \Rightarrow \frac{\partial\theta_i}{\partial\mu_i} = \frac{1}{b_i\text{“}(\theta_i)},$$

y sustituyendo esta expresión en la ecuación 1.7,

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{[y_i - b_i'(\theta_i)]}{b_i\text{“}(\theta_i)/\omega_i}\frac{\partial\mu_i}{\partial\beta_j}.$$

Considerando ahora las expresiones para la media y varianza obtenidas con anterioridad 1.5, 1.6, sustituyéndolas en la expresión anterior e igualando a cero llegamos a la ecuación cuya solución es la estimación óptima de β

$$\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{var(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j} = 0 \quad \forall j. \tag{1.8}$$

Sin embargo, estas expresiones son exactamente las mismas que habrían resultado si hubiésemos aplicado el método de mínimos cuadrados ponderados no lineales para calcular la estimación de máxima versoimilitud de β. Si los pesos $var(\mu_i)$ fuesen conocidos e independientes de β, el objetivo de mínimos cuadrados sería:

$$S = \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{var(\mu_i)}, \tag{1.9}$$

donde las μ_i dependen de manera no lineal de β y los pesos $var(\mu_i)$ son tratados como valores fijos. Para encontrar las estimaciones por el método de mínimos cuadrados debe resolverse

$$\partial S/\partial\beta_j = 0 \quad \forall j,$$

un sistema de ecuaciones equivalente al obtenido anteriormente (ecuación 1.8), siempre que se consideren como valores fijos $var(\mu)$. Esto sugiere emplear un método alternativo para resolver el sistema de ecuaciones 1.8, el cual explicamos paso a paso a continuación.

Sea $\beta^{[k]}$ el vector de parámetros estimado en la k-ésima iteración y $\eta^{[k]}$ y $\mu^{[k]}$ los vectores cuyos elementos vienen dados por $\eta_i^{[k]} = X_i\beta^{[k]}$ y $\mu_i^{[k]} = g^{-1}(\eta_i)^{[k]}$, respectivamente, donde $g^{-1}(\cdot)$ es la función inversa de la función link. El proceso se inicia con una estimación aproximada de un parámetro, $\beta^{[0]}$, y se realizan los siguientes pasos iterativamente hasta que la secuencia de $\beta^{[k]}$ converge. Los pasos a seguir son los siguientes

- Paso 1. Calcular los términos $var\left(\mu_i^{[k]}\right)$ implícitos en la $\beta^{[k]}$ actual.
- Paso 2. Dadas estas estimaciones, se utiliza el método no lineal de mínimos cuadrados para minimizar 1.3.4 con respecto a β, obteniendo así $\beta^{[k+1]}$. Cabe destacar que los $var\left(\mu_i^{[k]}\right)$ son siempre tratados como fijos y no como funciones de β.
- Paso 3. Establecer k en $k+1$.

Este método es más lento de lo necesario. Esto es debido a que el segundo paso del procedimiento implica el número de iteraciones necesarias del método no lineal de mínimos cuadrados hasta que se alcance la convergencia, pero esto no tiene sentido hacerlo antes de que $var\left(\mu_i\right)$ haya convergido. Es por esto por lo que el paso 2 se suele sustituir por usar el valor del $\beta^{[k]}$ correspondiente como valor inicial y realizar solo una iteración para obtener $\beta^{[k+1]}$.

De la aplicación de esta alternativa resulta un esquema ordenado y compacto. Así, escribamos en primer lugar el problema de mínimos cuadrados no lineales en forma matricial.

Sea $V_{[k]}$ la matriz diagonal donde $V_{[k_{ii}]} = var\left(\mu_i^{[k]}\right)$, la ecuación 1.3.4 se convierte en

$$S = \left|\left|\sqrt{V_{[k]}^{-1}}\left[y - \mu(\beta)\right]\right|\right|.^2$$

y, siguiendo el método no lineal de mínimos cuadrados, el valor de $\mu(\beta)$ es reemplazado por su expansión de Taylor de primer orden alrededor $\beta^{[k]}$, de modo que, sin más que sustituir esta expansión en la expresión, obtenemos

$$S = \left|\left|\sqrt{V_{[k]}^{-1}}\left[y - \mu^{[k]} - J\left(\beta - \beta^{[k]}\right)\right]\right|\right|^2,$$

donde J es la matriz jacobiana cuyos elemtnos vienen dados por $J_{i,j} = \partial\mu_i/\partial\beta_j|_{\beta^{[k]}}$. Ahora, como el modelo es un GLM, se cumple que

$$g(\mu_i) = X_i\beta \Rightarrow g'(\mu_i)\frac{\partial\mu_i}{\partial\beta_j} = X_{ij}$$

y por lo tanto

$$J_{ij} = \frac{\partial\mu_i}{\partial\beta_j}|_{\beta^{[k]}} = \frac{X_{ij}}{g'\left(\mu_i^{[k]}\right)}.$$

Así, definiendo G como una matriz diagonal con elementos $G_{ii} = g'(\mu_i^{[k]})$ y $J = G^{-1}X$, llegamos a la siguiente expresión para S

$$\begin{aligned} S &= \left|\left|\sqrt{V_{[k]}^{-1}}G^{-1}\left[G\left(y - \mu^{[k]}\right) + \eta^{[k]} - X\beta\right]\right|\right|^2 \\ &= \left|\left|\sqrt{W^{[k]}}\left(z^{[k]} - X\beta\right)\right|\right|^2, \end{aligned}$$

siendo los elementos del vector $z^{[k]}$ de la forma

$$z_i^{[k]} = g'\left(\mu^{[k]}\right)\left(y_i - \mu_i^{[k]}\right) + \eta_i^{[k]},$$

y los elementos de la matriz diagonal de pesos $W^{[k]}$ de la forma

$$W_{ii}^{[k]} = \frac{1}{V\left(\mu_i^{[k]}\right)g'\left(\mu_i^{[k]}\right)^2}.$$

Por lo tanto, los siguientes pasos se iteran para lograr la convergencia

1. Calcular los $z^{[k]}$ y los pesos $W^{[k]}$ en cada iteración usando $\mu^{[k]}$ y $\eta^{[k]}$ de manera recurrente.
2. Minimizar la suma de cuadrados $\left|\left|\sqrt{W^{[k]}}\left(z^{[k]} - X\beta\right)\right|\right|^2$ con respecto al vector de parámetros β, con el objetivo de obtener $\beta^{[k+1]}$, y luego en consecuencia $\eta^{[k+1]} = X\beta^{[k+1]}$ y $\mu^{[k+1]}$.
3. Pasar de k a $k+1$.

La estimación de β que resulta convergente resuelve la ecuación 1.8, siendo por tanto la estimación de máxima verosimilitud de dicho parámetro. Nótese que este algoritmo converge en la mayoría de situaciones, aunque en la práctica hay excepciones, como por ejemplo modelos deficientes.

Cabe destacar que en este método solo se necesitan como iterantes iniciales los valores $\mu^{[0]}$ y $\eta^{[0]}$, pero no $\beta^{[0]}$. Por tanto, generalmente el proceso se inicia se inicia tomando $\mu_i^{[0]} = y_i$ y $\eta_i^{[0]} = g(\mu_i^{[0]})$, con cuidado en el parámetros $\mu_i^{[0]}$ para que no provoque un valor infinito de $\eta_i^{[0]}$.

El modelo lineal de trabajo a través del método de mínimos cuadrados reponderados iterativamente (IRLS) no es simplemente un medio para encontrar el estimaciones de máxima verosimilitud de los parámetros. En la convergencia, la constante

$$S = -\frac{1}{2\phi} \left|\left| \sqrt{W} \left(z - X\beta\right) \right|\right|^2$$

es también una aproximación cuadrática al logaritmo de la función de máxima verosimilitud del modelo en las inmediaciones de la estimación $\hat{\beta}$, tal y como demostraremos a continuación. Por un lado, las primeras derivadas con respecto a β_j del logaritmo de la función de máxima verosimilitud y de S coinciden y todas ellas son cero.
Respecto a la segunda derivada de S, viene dada por $-XWX/\phi$, que coincide con la segunda derivada esperada del logaritmo de la función de máxima verosimilitud y, por la ley de los grandes números, con la propia segunda derivada en el límite cuando n tiende a infinito.
Para probar esto, definimos u como el vector de derivadas del logaritmo de los parámetros del modelo, por lo que $u_i = \partial l/\partial\beta_i$ y, en consecuencia, se pueden reescribir las ecuaciones de 1.8 en forma vectorial como sigue

$$u = X^T G^{-1} V^{-1} \left(y - \mu\right)/\phi.$$

Luego, haciendo cuentas, tenemos que

$$\begin{aligned} \mathbb{E}(uu^T) &= X^T G^{-1} V^{-1} \mathbb{E}\left[(Y-\mu)(Y-\mu)^T\right] V^{-1} g^{-1} x/\phi^2 \\ &= X^T G^{-1} V^{-1} V V^{-1} G^{-1} X/\phi \\ &= X^T W X/\phi, \end{aligned}$$

ya que $\mathbb{E}\left[\left(Y-\mu\right)\left(Y-\mu\right)^T\right] = V\phi$, siendo $W = G^{-1}V^{-1}G^{-1}$. A continuación demostraremos un resultado general que será clave para demostrar que S es una aproximación cuadrática de la función logaritmo de la máxima verosimlitud. Sabemos que se cumple que

$$\mathbb{E}(H) = -\mathbb{E}\left(uu^T\right),$$

donde H denota la matriz hessiana.

Por un lado, siendo l el logaritmo de la función máxima verosimilitud, tenemos que

$$\mathbb{E}_0\left(\frac{\partial l}{\partial\theta}\,|_{\theta_0}\right) = 0,$$

es decir, donde haya suficiente regularidad para que el orden de diferenciación e integración se pueda cambiar tendremos que

$$\begin{aligned} \mathbb{E}_0\left(\frac{\partial l}{\partial\theta}\right) &= \mathbb{E}_0\left(\frac{\partial}{\partial\theta} log\left[f(Y,\theta)\right]\right) = \int \frac{1}{f(y,\theta_0)} \frac{\partial f}{\partial\theta} f(y,\theta_0) dy \\ &= \int \frac{\partial f}{\partial\theta} dy = \frac{\partial}{\partial\theta}\int f dy = \frac{\partial 1}{\partial\theta} = 0 \end{aligned}$$

El resultado anterior implica entonces que

$$\int \frac{\partial log(f)}{\partial \theta} f dy = 0.$$

Aplicando la regla del producto podemos derivar de nuevo la expresión anterior respecto a θ, llegando a lo siguiente

$$\int \frac{\partial^2 log(f)}{\partial \theta^2} f + \frac{\partial log(f)}{\partial \theta} \frac{\partial f}{\partial \theta} dy = 0,$$

pero

$$\frac{\partial log(f)}{\partial \theta} = \frac{1}{f} \frac{\partial f}{\partial \theta},$$

por lo que

$$\int \frac{\partial^2 log(f)}{\partial \theta^2} f dy = -\int \left[\frac{\partial log(f)}{\partial \theta} \right]^2 f dy,$$

que por definición es lo mismo que

$$\mathbb{E}_0 \left[\frac{\partial^2 l_i}{\partial \theta^2} |_{\theta_0} \right] = -\mathbb{E}_0 \left[\left(\frac{\partial l_i}{\partial \theta} |_{\theta_0} \right)^2 \right].$$

Esta relación con las segundas derivadas es suficiente para demostrar que S es una aproximación cuadrática del logaritmo de la función de máxima verosimilitud en un entorno de $\hat{\beta}$ y, por la consistencia de los MLEs, también en un entorno de los valores verdaderos del vector β.

1.3.5. Odds Ratio

En modelos con variable respuesta binaria y, en especial, en el campo de la medicina, es bastante usual el cálculo de los llamados Odds Ratio (OR). El Odds Ratio representa la probabilidad de que ocurra un resultado cuando se está expuesto a otra característica, en comparación con las probabilidades de que ocurra el resultado en ausencia de esta característica.
Cuando se construye un modelo con variable respuesta logística, los coeficientes asociados representan el aumento estimado en las probabilidades logarítmicas del resultado por unidad de aumento en el valor de la exposición, es decir, si aplicamos la exponencial a la estimación de los coeficientes, obtenemos la razón de probabilidad asociada con un aumento de una unidad en la exposición [10]. Así, el valor de los OR va a depender del punto de corte 1, de manera que:

- Si OR=1, entonces la exposición no va a influir en las probabilidades de resultado.
- Si OR<1, en exposición cuanto más bajas sean las probabilidades de resultado.
- Si OR >1, en exposición cuanto más altas sean las probabilidades de resultado.

Los intervalos de confianza al 95 % asociados a los OR son de vital importancia a la hora de determinar si la exposición correspondiente va a influir en la probabilidades de resultado, lo cual solo ocurrirá cuando no contenga el valor 1.

1.4. Modelos aditivos generalizados (GAM)

Seguiremos el capítulo 3 de la referencia [5].

Aunque los modelos lineales generalizados ofrecen diferentes posibilidades, siguen teniendo la limitación de que la relación entre la media de la variable respuesta y las variables explicativas debe ser

lineal y constante. Esto no siempre se adapta bien a todas las posibles situaciones, por lo que aparecieron los conocidos como modelos aditivos, los cuales son una extensión de los modelos lineales en la que la relación entre la media de la variable respuesta con las variables explicativas se hace a través de funciones $f_i(x_i)$. A medida que se fueron estudiando este tipo de modelos, apareció una generalización de los mismos, los modelos aditivos generalizados (GAM). Estos permitían a mayores incorporar relaciones no lineales entre la media de la variable respuesta y las variables explicativas. Así, la relación entre la media de la variable respuesta $g(\mu)$ con cada predictor se hacía a través de las funciones $f_i(x_i)$.

1.4.1. Formulación del modelo

La formulación de este tipo de modelos viene dada por tanto de la siguiente forma

$$g(\mu) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_k(x_ip),$$

con

$$\mu = \mathbb{E}(Y) \quad \text{e} \quad Y \sim \text{alguna distribución de la familia exponencial,}$$

donde Y representa la variable respuesta, x_ip son los p valores de la fila i-ésima de la matriz del modelo para cualquier componente del modelo estrictamente paramétrica y las f_j son cualquier función, tanto lineal como no lineal, que toma como argumentos los valores de las variables explicativas x_k. En la práctica, las funciones más empleadas son las conocidas como suaves.

1.4.2. Hipótesis del modelo

En cuanto a las hipótesis de este tipo de modelos, al igual que en los modelos generalizados lineales, la variable respuesta tiene que seguir una distribución perteneciente a la familia exponencial. La condición que requieren a mayores es que se debe definir la relación flexible (lineal o no) que se establece entre la variable respuesta y las explicativas en términos de funciones suaves en lugar de en función de relaciones paramétricas detalladas, evitando así la construcción de modelos complejos y difíciles de interpretar.
Esta flexibilidad y facilidad tiene el coste de dos nuevos problemas teóricos. Por un lado, es necesario representar las funciones suaves que vayamos a aplicar a cada variable explicativa. Por otro, se tiene que decidir cuánto de suaves serán estas funciones.

En la siguiente sección, veremos cómo se pueden representar los modelos GAM utilizando los conocidos como splines de regresión penalizados, cuya estimación es resultado de aplicar métodos de regresión penalizados. Veremos también cómo se puede estimar el grado óptimo de suavidad para las f_j a partir de los datos y haciendo uso del método de validación cruzada. El modelo sobre el que llevaremos a cabo las explicaciones será un GAM simple con dos funciones univariadas, con el fin de no perder la simplicidad del enfoque con detalles técnicos. Además, los métodos presentados no son aquellos que son más convenientes para el uso práctico en general, sino más bien los que permitan explicar de forma simple el marco básico de este tipo de modelos.

1.4.3. Funciones suaves univariadas

En primer lugar introduciremos la representación de funciones suaves, considerando el caso particular de un modelo que contiene tan solo una función suave de una variable explicativa, el cual se representa de la siguiente forma

$$y_i = f(x_i) + \epsilon_i, \tag{1.10}$$

donde y_i es la variable respuesta, x_i es la covariable, f es una función suave y los ϵ_i son variables aleatorias independientes e identicamente distribuidas que siguen una distribución $N(0, \sigma^2)$. Supongamos por comodidad que $x_i \in [0, 1]$ sin pérdida de generalidad.

Representación de una función suave: splines

Para estimar f es necesario que esta sea representada de tal manera que la ecuación 1.10 se convierta en un modelo lineal. Esto se puede hacer definiendo un espacio de funciones, las cuales serán llamadas bases, del cual la función f, o una aproximación de esta, sea un elemento de este espacio. El proceso consiste entonces en elegir una base de funciones, que serán tratadas como conocidas por completo. Sea $b_i(x)$ es i-ésima función de la base, la función f puede expresarse de la siguiente forma

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i,$$

para cualquier valor de los parámetros desconocidos β_i. Sustituyendo la anterior expresión en la ecuación 1.10 obtenemos claramente la expresión de un modelo lineal.

Grado de suavizado con splines penalizados

Existen varias alternativas para elegir el grado de suavizado de las funciones. Una de ellas consiste en mantener la dimensión de la base fija, en un tamaño un poco mayor de lo que se considere que podría ser necesario, para después agregar una penalización por ondulaciones al objetivo del ajuste a través del método de mínimos cuadrados y controlar así la suavidad del modelo. Es decir, en lugar de ajustar el modelo minimizando

$$||y - X\beta||^2,$$

podría ser ajustado minimizando

$$||y - X\beta||^2 + \lambda \int_0^1 [f''(x)]^2 \, dx,$$

donde el cuadrado integrado de la segunda derivada penaliza los modelos que tienen demasiadas "ondulaciones". Vemos que a este término lo acompaña el parámetro λ, que controla el equilibrio entre el ajuste del modelo y la suavidad del mismo. Nótese que si $\lambda \longrightarrow \infty$, la estimación para f sería una línea recta ya que la penalización sería infinita, mientras que, en caso contrario, si $\lambda = 0$ la estimación sería un spline no penalizado (el término de penalización se haría 0).

Como la función a estimar f es lineal en los parámetros β_i, la penalización siempre se puede escribir como forma cuadrática en β de la siguiente forma

$$\int_0^1 [f''(x)]^2 \, dx = \beta^T S \beta,$$

donde S es una matriz de coeficientes conocidos.
Por lo tanto, el problema del ajuste del spline de regresión penalizado se reduce a minimizar la siguiente expresión

$$||y - X\beta||^2 + \lambda \beta^T S \beta.$$

El problema de conocer la estimación del grado de suavidad de la función f consistirá entonces en estimar el parámetro de penalización λ.

Antes de abordar la aproximación del valor de λ, consideremos conocido este parámetro y calculemos la estimación del vector de parámetros β.
Derivando la expresión respecto a β e igualando a 0 tenemos que el estimador de mínimos cuadrados penalizado de este parámetro viene dado por

$$\hat{\beta} = \left(X^T X + \lambda S\right)^{-1} X^T y.$$

Por tanto, teniendo en cuenta la definición de matriz hat en este caso viene dada por

$$A = X\left(X^T X + \lambda S\right)^{-1} X^T.$$

Cabe destacar que estas expresiones no son cómodas a la hora de hacer cálculos, por lo que se suelen emplear los conocidos como métodos ortogonales, que ofrecen una mayor estabilidad numérica. Por tanto, para el cálculo tendremos en cuenta que

$$\left|\left|\begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda}B \end{bmatrix} \beta \right|\right|^2 = ||y - X\beta||^2 + \lambda\beta^T S\beta,$$

donde B es cualquier raíz cuadrada de la matriz S, es decir, se cumple que $B^T B = S$. Cabe destacar que la suma de cuadrados en el lado derecho es solo un objetivo de mínimos cuadrados para un modelo en el que la matriz del modelo ha sido aumentada por una raíz cuadrada de la matriz de penalización.

Grado de suavizado por validación cruzada

Sabemos que si el parámetro de suavizado λ es demasiado alto, los datos se suavizarán en exceso, y si es demasiado bajo, los datos se suavizarán por debajo de lo usual. En ambos casos lo que ocurrirá es que la estimación spline de la función f, $\hat{f}$, no va a estar cerca de dicha función. Lo ideal sería encontrar el valor de λ tal que $\hat{f}$ sea lo más cercano posible a f. Uno de los cirterios más utilizados sería escoger λ tal que se minimice la siguiente expresión

$$M = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i - f_i\right)^2,$$

donde se denota por $\hat{f}_i$ a la estimación de f aplicada sobre la covariable correspondiente, es decir, $\hat{f}_i \equiv \hat{f}(x_i)$ y f_i significa la función f sobre la covariable x_i correspondiente, es decir, $f_i \equiv f(x_i)$.

Nótese que el término M 1.4.3 no se puede usar directamente por ser la función f desconocida. A pesar de esto sí es posible obtener una estimación del error cuadrático esperado al predecir una nueva variable, el cual viene dado por $E(M) + \sigma^2$.
Si denotamos por $\hat{f}^{[-i]}$ al modelo ajustado a todos los datos excepto al dato i-ésimo y_i, la conocida por puntuación ordinaria de validación cruzada dejando uno fuera vendrá dada por la siguiente expresión

$$V_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}^{[-i]} - y_i\right)^2. \tag{1.11}$$

Este nuevo concepto $\hat{f}^{[-i]}$ es resultado de omitir cada uno de los datos uno a uno por turnos, de manera que el modelo se ajusta a los datos restantes y se calcula, con cada ajuste, la diferencia al cuadrado entre el dato faltante y su valor predicho. Después se calcula el promedio del cuadrado de estas diferencias sobre todos los datos.

Si sustituímos $y_i = f_i + \epsilon_i$ en la expresión 1.11 llegamos a la siguiente expresión

$$\begin{aligned} V_0 &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i - \epsilon_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i\right)^2 - 2\left(\hat{f}_i^{[-i]} - f_i\right)\epsilon_i + \epsilon_i^2. \end{aligned}$$

Como por un lado $(\epsilon_i) = 0$ y por otro ϵ_i y $\hat{f}_i^{[-i]}$ son independientes para todo valor de i, al calcular la media de la expresión anterior tenemos que el segundo término del sumatorio del lado derecho es 0, es decir, tendríamos que

$$\mathbb{E}(V_0) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i\right)^2\right) + \mathbb{E}(\epsilon_i^2).$$

Ahora bien, por definición de la varianza tenemos que $\sigma^2 = \mathbb{E}(\epsilon_i^2) - (\mathbb{E}(\epsilon_i))^2$, y como $E(\epsilon_i) = 0$, llegamos a que

$$\mathbb{E}(V_0) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i\right)^2\right) + \sigma^2.$$

Como además es claro que se cumple que $\hat{f}^{[-i]} \approx \hat{f}$, con igualdad cuando el tamaño de la muestra tiende a infinito ($n \to \infty$), llegamos finalmente a que $E(V_0) \approx E(M) + \sigma^2$, de nuevo con igualdad cuando n tiende a infinito. Esto demuestra que la elección del parámetro λ con el objetivo de minimizar V_0 es un enfoque adecuado si lo que se busca es minimizar M.

Nótese que la técnica de validación cruzada ordinada consiste precisamente en escoger λ para minimizar V_0. Esta técnica ordinaria tiene un enfoque razonable en la gran mayoría de los casos, incluso sin justificación del error cuadrático medio. Esto es debido a que la elección del modelo se lleva a cabo para maximizar la capacidad de predecir datos a los que no fue ajustado el modelo y no solo para maximizar la capacidad que tengan para ajustarse a los datos a partir de los cuales se estimaron, en cuyo caso los modelos más complejos siempre serán seleccionados sobre los más sencillos.

A pesar de todo lo anterior, el procedimiento de calcular V_0 omitiendo un dato de cada vez y ajustando el modelo para cada uno de los n conjuntos de observaciones restantes no se considera eficiente, pero se puede demostrar que

$$V_0 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_i\right)^2 / \left(1 - A_{ii}\right)^2,$$

donde $\hat{f}_i$ es la estimación del ajuste a todos los datos y A es la matriz hat correspondiente. El término $1 - A_{ii}$ es reemplazado en la práctica por su media, es decir, por la expresión $tr(I - A)/n$ con el fin de llegar a la puntuación de validación cruzada generalizada, que viene dada por

$$V_g = \frac{n\sum_{i=1}^{n}\left(y_i - \hat{f}_i\right)^2}{[tr(I - A)^2]}.$$

La validación cruzada generalizada (GCV) tiene ventajas computacionales sobre la validación cruzada ordinaria (OCV) aunque se puede demostrar que también minimiza $E(M)$ cuando el tamaño muestral n tiende a infinito.

Cabe destacar que hemos hecho uso de la técnica de validación cruzada dejando uno fuera, pero que existen más versiones de esta metodología, como la validación cruzada de K iteraciones que utilizaremos en el capítulo 2.

1.4.4. Varias funciones suaves univariadas

Supongamos ahora que tenemos un modelo aditivo simple con dos variables explicativas, x y z, y una variable respuesta y. Su formulación será entonces la siguiente

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i, \tag{1.12}$$

donde f_1 y f_2 son funciones suaves y ϵ_i son variables aleatorias independientes e identicamente distribuídas que siguen una distribución $N(0, \sigma^2)$. Por comodidad y sin pérdida de generalidad supondremos

que z_i y x_i se encuentran en el intervalo $[0, 1]$.

Sobre este modelo formulado en la ecuación 1.4.4 podemos destacar las siguientes observaciones:

- En primer lugar, la suposición de efecto aditivo es considerable, por ser $f_1(x) + f_2(z)$ un caso bastante restrictivo de la función suave general con dos variables $f(x, z)$.
- En segundo lugar, como ahora el modelo que estamos considerando contiene más de una función, aparece un problema de identificabilidad: f_1 y f_2 son estimables cada uno dentro de una constante aditiva. En estos casos hay que tener en cuenta que cualquier constante podría sumarse a f_1 y restarse de f_2 simultáneamente sin cambiar las predicciones del modelo. Por tanto, es de gran importancia indicar las restricciones de identificabilidad al modelo antes de ejecutarlo.

En los casos en los que se presente el problema de la identificabilidad, el modelo aditivo se puede representar a partir de splines de regresión penalizados, cuyas estimaciones se llevan a cabo a partir del criterio de mínimos cuadrados penalizados y cuyo grado de suavizado estimado se calcula haciendo uso de la técnica de validación cruzada, siguiendo el procedimiento análogo al modelo univariado simple.

Representación spline de regresión penalizada de un modelo aditivo

Cada función suave considerada en un modelo GAM se puede representar a partir de una base de splines de regresión penalizados. Así, restringiéndonos de nuevo al caso en el que se consideren dos funciones suaves univariadas, estas se pueden expresar de la siguiente forma:

$$f_1(x) = \delta_1 + x\delta_2 + \sum_{j=1}^{q_1-2} R(x, x_j^*)\delta_{j+2},$$

$$f_2(z) = \gamma_1 + z\gamma_2 + \sum_{j=1}^{q_2-2} R(z, z_j^*)\gamma_{j+2},$$

donde δ_j y γ_j son parámetros desconocidos para f_1 y f_2 respectivamente, q_1 y q_2 son el número de parámetros desconocidos para f_1 y f_2 respectivamente y x_j^* y z_j^* son las ubicaciones de los nodos para estas dos funciones.

En caso de existir un problema de identificabilidad en un modelo GAM, lo que va a ocurrir es que los parámetros δ_1 y γ_1 se confunden. La solución que se suele emplear es fijar uno de ellos a cero, por ejemplo γ_1. Bajo esta suposición, es fácil ver que el modelo aditivo se puede expresar con la misma forma que un modelo lineal, es decir, $y = X\beta + \epsilon$, donde la i-ésima fila de la matriz del modelo vendrá dada por

$$X_i = \left[1, x_i, R\left(x_i, x_1^*\right), R\left(x_i, x_2^*\right), ..., R\left(x_i, x_{q_1-2}, z_i\right), ..., R\left(z_i, z_{q_2-1}^*\right)\right]$$

y el vector de parámetros desconocidos por

$$\beta = \left[\delta_1, \delta_2, ..., \delta_{q_1}, \gamma_2, \gamma_3, ...\gamma_{q_2}\right]^T.$$

La suavidad de las funciones también se puede medir a partir de las siguientes integrales

$$\int_0^1 f\text{“}_1(x)^2 dx = \beta^T S_1 \beta,$$

$$\int_0^1 f_2\text{“}(x)^2 dx = \beta^T S_2 \beta,$$

teniendo en cuenta que S_1 y S_2 son cero salvo los términos $S_{1i+2,j+2} = R\left(x_i^*, x_j^*\right)$ para $i, j = 1, ..., q_1 - 2$ y $S_{2i+q_1+1,j+q_1+1} = R\left(z_i^*, z_j^*\right)$ para $i, j = 1, ..., q_2 - 2$.

Por supuesto, se pueden utilizar cualesquiera de un gran número de bases alternativas en lugar de la base spline de regresión que acabamos de usar en esta memoria dado que la idea general será la misma en todos los casos.

1.4.5. Estimador de mínimos cuadrados penalizados

Siguiendo el procedimiento que llevamos a cabo en el caso en el que existía una única función suave univariada, el ajuste del modelo aditivo (ecuación 1.4.4) por mínimos cuadrados penalizados consiste en minimizar la siguiente expresión

$$||y - X\beta||^2 + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta,$$

donde los parámetros de suavizado λ_1 y λ_2 controlan el peso que se le a cada una de las funciones f_1 y f_2, en relación con el objetivo de ajustar la variable respuesta. Supongamos en un primer momento que dichos parámetros son conocidos. Si denotamos por S a la suma $\lambda_1 S_1 + \lambda_2 S_2$, el objetivo se puede reescribir como

$$||y - X\beta||^2 + \beta^T S \beta = \left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \beta \right|\right|^2,$$

donde B es cualquier raíz cuadrada de la matriz S, por lo que se cumple entonces que $B^T B = S$. Si nos fijamos en el lado derecho de la igualdad anterior, vemos que es exactamente el objetivo de mínimos cuadrados no penalizado para una versión aumentada del modelo y los datos de respuesta correspondientes, es decir, el modelo se puede ajustar mediante regresión lineal estándar.

1.4.6. Modelos aditivos generalizados

Cabe esperar que el modelo aditivo generalizado es en general más natural que el modelo aditivo. La principal diferencia entre ambos es que, mientras el modelo aditivo fue ajustado por mínimos cuadrados penalizados, el modelo aditivo generalizado puede ajustarse mediante la maximización de la probabilidad penalizada, esto es, la técnica de mínimos cuadrados iterativos penalizados. Esta técnica de ajuste consiste en llevar a cabo el conocido como Penalized Iteratively Reweighted Least squares (P-IRLS), un procedimiento en donde cada iteración se estructura de la siguiente forma:

1. Dadas las estimaciones actuales de los parámetros $\beta^{[k]}$ y vector correspondiente a la media estimada de la variable respuesta $\mu^{[k]}$, calcular los términos:

$$w_i = \frac{1}{V\left(\mu_i^{[k]}\right) g'\left(\mu_i^{[k]}\right)} \tag{1.13}$$

y

$$z_i = g\left(\mu_i^{[k]}\right)\left(y_i - \mu_i^{[k]}\right) + X_i \beta^{[k]} \tag{1.14}$$

donde $var\left(Y_i\right) = V\left(\mu^{[k]}\right)\phi$ por la sección 2.1.2 y X_i es la i-ésima fila de la matriz X.

2. A continuación, minimizar

$$||\sqrt{W}(z - X\beta)||^2 + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta,$$

donde W es una matriz diagonal tal cuyos elementos son los términos w_i calculados en el paso anterior.

Cabe destacar que este paso es equivalente a minimizar la siguiente expresión

$$\left\| \begin{bmatrix} \sqrt{W} & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} z \\ o \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \beta \right) \right\|^2$$

donde B denota la matriz raíz cuadrada tal que se cumple $B^T B = \lambda_1 S_1 + \lambda_2 S_2$. En el caso particular por ejemplo del modelo log-link de errores gamma, la función enlace que denotamos por g, es la función logaritmo, por lo que $g'(\mu_i) = \mu_i^{-1}$, mientras que por seguir los errores una distribución gamma, $V(\mu_i) = \mu_i^2$. Por tanto, si sustituímos estos valores en las ecuaciones 1.13 y 1.14, obtenemos lo siguiente

$$w_i = 1$$

y

$$z_i = \frac{\left(y_i - \mu_i^{[k]}\right)}{\mu_i^{[k]}} + X_i \beta^{[k]}.$$

1.5. Modelos aditivos generalizados para localización, escala y forma (GAMLSS)

Seguiremos como referencia en este apartado [11].

Los modelos lineales generalizados (GLM) y los modelos generalizados aditivos (GAM) tienen como hipótesis que la distribución que sigue la variable respuesta Y tiene que pertenecer a la familia exponencial, cuya media se expresa en función de las variables explicativas. Nótese que en este tipo de modelos los demás parámetros, como son la varianza, la asimetría y la kurtosis, se suponen constantes. Existen casos en los que esta suposición puede no ser adecuada, por lo que modelar los parámetros restantes en función de los predictores puede ser muy oportuno. De esto se encargan los modelos aditivos generalizados para localización, escala y forma (GAMLSS), que permiten establecer la relación entre los distintos parámetros que determinan la distribución de la variable respuesta y los predictores ofreciendo resultados considerablemente buenos en la mayoría de los casos. Además, recordemos incorporan a mayores nuevas y posibles distribuciones para la variable respuesta.

1.5.1. Formulación del modelo

La formulación general para este tipo de modelos es la siguiente:

$$\begin{aligned} Y &\sim D(\mu, \sigma, \gamma, \tau) \\ g_1(\mu) &= X^T\beta + f_1(x_1) + f_2(x_2) + ... + f_p(x_p), \\ g_2(\sigma) &= X^T\beta + f_1(x_1) + f_2(x_2) + ... + f_p(x_p), \\ g_3(\gamma) &= X^T\beta + f_1(x_1) + f_2(x_2) + ... + f_p(x_p), \\ g_4(\tau) &= X^T\beta + f_1(x_1) + f_2(x_2) + ... + f_p(x_p), \end{aligned}$$

donde $D(\mu, \sigma, \gamma, \tau)$ denota la distribución que sigue la variable respuesta Y, cuyos parámetros denotamos por μ, σ, γ y τ, X contiene los términos lineales del modelo, β es el vector de parámetros lineales, los términos $f_i(x_i)$ son las funciones de suavizado no lineales asociadas a cada predictor y x_i denota cada una de las filas de la matriz asociada al modelo.

Como observación, la implementación actual en R de este tipo de modelos permite especificar alrededor de 100 distribuciones discretas, continuas y mixtas, que se determinan con la opción "GAMLSS.family". Además, cualquier distribución que se considere en esta opción se puede mezclar para crear una nueva mezcla finita de distribuciones, además de que se pueden crear también distribuciones continuas discretizadas para modelar variables respuesta discretas. También se pueden llevar a cabo tranformaciones sobre una distribución definida en $(0, \infty)$ o $(0, 1)$ usando los argumentos tipo con opciones log o logit respectivamente.

1.5.2. Hipótesis del modelo

Los modelos aditivos generalizados para localización, escala y forma ya no requieren que la variable respuesta tenga que seguir una distribución perteneciente a la familia exponencial, así como lo hacían los GLM y los GAM, aumentando considerablemente el rango de posibilidades para la distribución de la variable respuesta.
Dado que los GAMLSS permiten modelar los diferentes parámetros de la distribución de la variable respuesta a partir de las variables explicativas, son muchas sus ventajas, debido a ser incluso más flexibles que los modelos GAM. Sin embargo, existen ciertas incertidumbres asociadas. Una de ellas por ejemplo que no pueden ser calculados los intervalos de confianza asociados a las estimaciones de los coeficientes y a las predicciones que se deseen calcular.

1.5.3. Distribuciones para la variable respuesta

Como comentamos anteriormente, los modelos GAMLSS permiten que la variable respuesta siga una distribución que no necesariamente tiene que pertenecer a la familia exponencial. Aunque ya son muchas las opciones que permiten para la variable respuesta este tipo de modelos, pueden ser extendidas considerando alguna de las siguientes ideas que explicamos brevemente a continuación.

Aplicar transformaciones

Pueden considerarse diferentes transformaciones a aplicar sobre las distribuciones continuas de la familia gamlss, las cuales están definidas en el intervalo $(-\infty, \infty)$. Las más empleadas son las transformaciones logarítimica y esponencial. Así, si por ejemplo Z fuese una variable aleatoria que sigue una distribución continua perteneciente a la familia gamlss y está definida en el intervalo $(-\infty, \infty)$, podría ser transformada de manera que $Y = exp(Z)$. Así, dicha variable pasaría a estar definida en un rango positivo.

Distribuciones truncadas

También podemos cambiar los límites de una distribución, de manera que dicho truncamiento se puede establecer a la izquierda, a la derecha o en ambas colas de la variables respuesta.

Distribuciones censuradas

A pesar de su ligera similitud, no debemos confundir entre una distribución truncada y una distribución censurada. En la primera, los valores por encima o por debajo de un determinado valor no existen, mientras que en la segunda sí existen pero no pueden ser observados. Existen paquetes para el diseño de esta situación, tanto si la variable respuesta está censurada por la izquierda como si está censurada por la derecha, como es el paquete gamlss.cens [12].

Mixturas finitas de las distribuciones de la familia gamlss

Por último otra de las alternativas que podemos contemplar es el ajustar mezclas finitas de distribuciones, que en R se traduce a emplear el paquete gamlss.mx [13]. Una mezcla finita de distribuciones que pertenezcan a al familia gamlss tendrá la siguiente forma:

$$f_Y(y|\psi) = \sum_{k=1}^{K} \pi_k f_k\left(y|\theta_k\right),$$

donde $f_k(y|\theta_k)$ denota la función de densidad de la variable respuesta y para componente k y π_k, cuyos valores se encuentran en el intervalo $[0, 1]$, es la probabilidad previa (o de mezcla) de la k-ésima componente, para $k = 1, 2, ..., K$. Además, se cumple que $\sum_{k=1}^{K} \pi_k = 1$ y $\psi = (\theta, \pi)$ donde $\theta = (\theta_1, \theta_2, ..., \theta_k)$ y $\pi = (\pi_1, \pi_2, ..., \pi_K)$. En esta opción pueden ser consideradas cualquier tipo de distribución, tan continuas como discretas, de la familia gamlss.

A continuación, explicamos más detalladamente la última de las opciones, dado que en nuestro caso en particular fue imprescindible.

1.5.4. Distribuciones de mixturas finitas

En esta sección estudiaremos el caso particular de que la variable respuesta se distribuya de acuerdo a una mezcla de distribuciones finitas. Existen dos casos posibles: el caso en el que exista algún parámetro en común entre ambas distribuciones y el caso en el que todos sean comunes. Cabe destacar que esta consideración es muy importante a la hora de ajustar distribuciones multimodales a los datos, tal y como veremos en el capítulo 2.

La función gamlss.mx

En la práctica, una vez detectada la presencia de más de una moda en el histograma de los datos, debe llevarse a cabo la implementación del modelo en R. Para eso, el paquete gamlss.mx[13] cuenta con la función *gamlss.mx*, que permite ajustar mezclas finitas sin parámetros en común. Explicaremos brevemente cuales son sus argumentos y la aplicaremos en el capítulo 2 sobre un caso médico real.

Los argumentos de dicha función son los siguientes:

- **formula:** Este argumento debe ser una sola fórmula o una lista de fórmulas de longitud K, donde K es el número de distribuciones consideradas en la mezcla. Si se usa una sola fórmula, los componentes de la mezcla K tienen el mismo predictor para μ, pero diferentes parámetros en sus predictores ya que no hay parámetros en común para dos o más de los componentes K.
- **pi.formula:** Con esta opción podemos modelar el predictor de probabilidades previas en función de las variables explicativas en el modelo multinomial. La opción predeterminada considera que las probabilidades previas son constantes.
- **family:** Con esta opción le especificamos la distribución de la variable respuesta, la cual debe pertenecer al conjunto de distribuciones de la familia de los modelos GAMLSS. También se puede especificar una lista de K distribuciones. Para facilitar la interpretación del modelo es recomendable, pero no necesario, que los parámetros sean comparables.
- **weights:** Esta opción permite determinar manualmente los pesos previos.
- **K**: Podemos especificar el número de componentes en la mezcla finita manualmente. Por defecto considera $K = 2$.
- **prob**: También podemos establecer valores iniciales para las probabilidades previas.

- **data**: Debemos especificar el conjunto de datos que contiene las variables que determinan el modelo.

1.6. Métricas de evaluación

Una vez tenemos construido el modelo que hayamos considerado adecuado, es de vital importancia realizar una evaluación rigurosa del mismo haciendo uso de diferentes métricas y teniendo en cuenta qué evalúa cada una de ellas. Teniendo en cuenta esto, las métricas de evaluación más conocidas se pueden clasificar en dos grandes grupos según si evalúan la bondad del ajuste o la capacidad de clasificación del modelo. Es claro que las métricas de evaluación de este segundo grupo solo se aplicarán a modelos cuya variable sea cualitativa binaria, puesto que no tendría ningún sentido aplicarlas en el resto de los casos.

1.6.1. Bondad de ajuste

La bondad de ajuste de un modelo indica cuánto de bien se ajusta el modelo a los datos observados. Estas medidas muestran el grado de discrepancia existente entre los valores predichos por el modelo y los valores observados. A continuación estudiaremos en detalles algunos de los más destacados.

Criterio del AIC y del BIC

Una de las herramientas más utilizadas en estadística para evaluar la bondad de ajuste de un modelo son los criterios de información, entre los que destacan el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC). Todos ellos surgen de la necesidad de escoger el modelo óptimo entre los que hayamos considerado para explicar nuestros datos. De forma general, sabemos que, cuantas más variables explicativas considere el modelo y cuanto más complejo sea, mejor describirá el proceso que estamos analizando, pero a la vez perderá generalidad, es decir, será muy preciso a la hora de predecir sobre nuestro conjunto de datos pero perderá precisión cuando los datos sean desconocidos. Por tanto, el objetivo siempre va a ser encontrar un modelo lo suficientemente complejo como para que describa bien nuestros datos pero no tanto como para que pierda validez general y no funcione adecuadamente sobre unos datos desconocidos.

Los indicadores de información miden el balance entre el ajuste y la complejidad del modelo, de manera que todos ellos vienen dado por la diferencia entre ajuste y complejidad. En el caso del AIC y del BIC, los cuales estudiaremos a continuación en detalle, utilizan la máxima verosimilitud como criterio de bondad de ajuste y el número de parámetros del modelo como medida de complejidad.

- **Criterio de Información de Akaike.** La fórmula del AIC viene dada por

$$AIC = 2 \cdot k - 2 \cdot ln(l),$$

  donde de nuevo k es el número de parámetros y l es la función de máxima verosimilitud. En este caso la medida de bondad de ajuste es $2ln(l)$. Si nos fijamos en la función l, esta va a venir dada por el producto de las probabilidades de cada dato condicionado al modelo, por lo que se obtiene multiplicando n valores entre 0 y 1. Como la función logaritmo es creciente, cuando l tiende a infinito, $-2ln(l)$ tiende a un valor muy pequeño y, en consecuencia, el AIC decrece a medida que aumenta la bondad de ajuste. En comparación con el BIC, que veremos a continuación, se quedará con un modelo más complejo y menos abstracto que hace predicciones con mayor detalle.

- **Criterio de Información Bayesiano.** Por su parte, la fórmula del BIC viene dada por

$$BIC = k \cdot ln(n) - 2 \cdot ln(l),$$

siendo k es el número número de parámetros, l es la función de máxima verosimilitud y n es el número de datos. En comparación con el AIC, que veremos acabe destacar que incorporamos el término $ln(n)$, por lo que al aumentar la complejidad los efectos son mayores que en el AIC, es decir, el BIC penaliza más la complejidad que AIC, buscando el modelo más sencillo con predicciones en un contexto más amplio.

Una vez explicados ambos criterios de información, podríamos preguntarnos por qué existen otros indicadores. El problema principal viene de que ambos consideran como única medida de complejidad el número de parámetros k, pues su efecto suele pasar desapercibido comparado con el valor que toma $-2ln(l)$ en el caso del AIC por ejemplo. Esto quiere decir que se perdona que el modelo sea muy complejo siempre y cuando tenga muchos datos para avalar cada parámetro, y se debe a que el AIC tiene como objetivo seleccionar el modelo que mejores predicciones hace dentro de un conjunto de datos, lo cual es muy conservador, pero puede que no sea lo más adecuado en todo momento. Esto implica que ambos criterios puedan quedarse cortos en algunas situaciones puntuales, por ejemplo en el caso en el que existan iteraciones entre parámetros, si existe colinealidad o si existen estructuras jerárquicas de factores aleatorios, aunque ambos funcionan considerablemente bien en la gran mayoría de situaciones y son muy utilizados en la práctica.

R cuadrado (R^2) y R cuadrado ajustado (R^2 ajusted)

Para este apartado se consultó la referencia [9].

El conocido por coeficiente de determinación y denotado por R^2, es un indicador de bondad de ajuste del modelo. El R^2 es un indicador sin unidades que toma valores entre 0 y 1, de manera que será más próximo a 1 a medida que se incorporen más variables, aunque las variables no sean significativas. Este es un problema que el R^2 ajustado o coeficiente de determinación ajustado soluciona, tal y como veremos a continuación.

Podemos expresar en primer lugar la diferencia entre la observación i-ésima de la variable respuesta y el valor medio de la misma como sigue:

$$y_i - \bar{y} = (\hat{y_i} - \bar{y_i}) + (y_i - \hat{y_i}) = (\hat{y_i} - \bar{y_i}) + \hat{u_i},$$

de manera que llegamos a que esta diferencia puede expresarse a partir de la distancia entre el valor ajustado $\hat{y_i}$ y su media más el residuo $\hat{u_i}$ correspondiente. Como además esta distancia puede ser mayor o menor que y_i, el residuo puede ser mayor o menor que cero. Esta igualdad demuestra que "la desviación total respecto a la media puede escribirse como la suma de la desviación explicada y el residuo".
Elevando al cuadrado la igualdad anterior y sumando todas las observaciones llegamos a la siguiente expresión

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 + \sum_{i=1}^{n}\hat{u_i}^2 + 2\sum_{i=1}^{n}(\hat{y_i} - \bar{y})\hat{u_i}.$$

Se puede demostrar que el término $\sum_{i=1}^{n}(\hat{y_i} - \bar{y})\hat{u_i} = 0$, por lo que teniendo esto en cuenta llegamos finalmente a que

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 + \sum_{i=1}^{n}\hat{u_i}^2,$$

o lo que es lo mismo, denotando $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$,

$$nS_y^2 = nS_{\hat{y}^2} + nS_{\hat{u}^2}.$$

Esto indica que n veces la varianza de la variable Y se puede expresar o descomponer como la suma explicada por la regresión estimada, que se corresponde con el término $nS_{\hat{y}^2}$, más la suma no explicada,

que se corresponde con la suma de los residuos al cuadrado, es decir, $nS_{\hat{u}^2}$.
A partir de estos términos podemos obtener el coeficiente de determinación como sigue:

$$R^2 = 1 - \frac{\text{Variación no explicada en Y}}{\text{Variación total en Y}} = \frac{\text{Variación explicada en Y}}{\text{Variación total en Y}}.$$

A la vista de la fórmula, se puede ver que, en efecto, el coeficiente de determinación nunca disminuye al añadir variables explicativas al modelo. Esto implica que no tiene sentido comparar modelos a partir del valor de este coeficiente, debido a que siempre tendrá mayor valor el modelo con mayor número de variables explicativas. Con el fin de solventar este problema, apareció el conocido como coeficiente de determinación ajustado o R^2 ajustado, cuya expresión es la siguiente:

$$R^2\text{ajustado} = (1 - R^2)\frac{n-1}{n-k},$$

siendo n el número de observaciones y k el número de variables expplicativas.
Así, al añadir variables explicativas al modelo, el término $(1 - R^2)$ disminuirá y el término $\frac{n-1}{n-k}$ aumentará.

Worm plot

Una forma de evaluar de manera visual la precisión de un modelo GAMLSS es a partir de los llamados gráficos worm (worm plot), que representan los residuos del modelo correspondiente. Esta representación es similar a la de un gráfico Q-Q (Q-Q plot) pero incluyen a mayores la línea teórica que representa la situación ideal en la cual los residuos deberían ser nulos. También se muestra dos curvas discontinuas en color negro que indican el intervalo de confianza al 95 %. El modelo es aceptable si solo entorno al 5 % de las observaciones salen fuera del intervalo de confianza. Esta representación se implementa en R con la función *wp()*, que muestra por defecto un ajuste cúbico (curva continua roja) y ayuda a identificar así la tendencia de los residuos.

1.6.2. Capacidad de clasificación

En medicina, una prueba diagnóstica se define como cualquier procedimiento quiera determinar la presencia de una determinada condición en un paciente, como por ejemplo, la presencia de una determinada patología. Esta es uno de los casos más comunes en este ámbito, considerando como respuesta dicotómica enfermo o sano (sí/no), que se conocen como pruebas diagnósticas. En este ámbito, es de gran importancia evaluar este tipo de pruebas, para lo que existen criterios que estudiaremos en profundidad a continuación.
Así, con el fin de seleccionar un modelo que clasifique de manera óptima según una variable dicotómica, veremos en detalle las métricas para evaluar la capacidad de clasificación de un modelo. Todas ellas se calcularán a partir de la conocida como matriz de confusión, de la cual hablaremos en primer lugar para dar paso a continuación a las métricas calculadas a partir de ella.

Matriz de confusión

Se consultó para esta sección la referencia [14].
En primer lugar, definiremos la llamada matriz de confusión, una matriz o tabla que representa la distribución de los valores predichos de la variable respuesta frente a los valores observados de la misma. Para poder calcular esta matriz se deben dividir los registros de datos originales en dos subconjuntos: el subconjunto de entrenamiento y el subconjunto de test, que normalmente representan el 80 % y 20 % respectivamente. Así, la matriz de confusión se define como la representación matricial de los resultados de las predicciones que se utiliza a menudo para describir el rendimiento del modelo de clasificación sobre un conjunto de datos de test cuyos valores reales son conocidos. Su estructura se muestra a continuación (Tabla 1.1).

	Clase real positiva	**Clase real negativa**
Clase predicha positiva	Verdaderos positivos (TP)	Falsos positivos (FP)
Clase predicha negativa	Falsos negativos (FN)	Verdaderos negativos (TN)

Tabla 1.1: Matriz de confusión para clasificación binaria.

Así, cada una de las predicciones del modelo se clasifica en uno de estos cuatro grupos dependiendo de si coinciden o no con el valor real correspondiente:

- Verdadero Positivo (TP): verdadero en la realidad y predicho por el modelo como verdadero.
- Verdadero Negativo (TN): falso en la realidad y predicho por el modelo como falso.
- Falso Positivo (FP): falso en la realidad y predicho por el modelo como verdadero.
- Falso Negativo (FN): verdadero en la realidad y predicho por el modelo como falso.

A partir de estos cuatro parámetros de la matriz de confusión podemos obtener las medidas de rendimiento para conseguir información sobre el modelo que se va a escoger. Las métricas que estudiaremos en nuestro caso serán la especificidad, la sensibilidad, la precisión (o accuracy), el likelihood positivo y negativo y el F1 score.

Sensibilidad

La sensibilidad es la tasa de lo que ha sido identificado correctamente como positivo entre todo lo que realmente es positivo. Se calcula a partir de la siguiente fórmula:

$$\text{Sensibilidad} = \frac{TP}{FN + TP},$$

y sus valores se encuentran entre 0 y 1.

Especificidad

La especificidad es la tasa de lo que ha sido identificado correctamente como negativo entre todo lo que realmente es negativo. Se calcula a partir de la siguiente fórmula:

$$\text{Especificidad} = \frac{TN}{FP + TN},$$

y sus valores se encuentran entre 0 y 1.

Precisión (accuracy)

La precisión o accuracy mide el porcentaje de casos que el modelo ha predicho bien,es decir, es el cociente entre el número de casos clasificados correctamente entre el total. Esta métrica nos informa de si el modelo de predicción presenta muchos falsos positivos. El cálculo de este parámetro se lleva a cabo a partir de la siguiente expresión:

$$\text{Precisión} = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}} = \frac{(TP + TN)}{(TP + FP + TN + FN)}.$$

Likelihood ratio positivo y negativo

En medicina son muy utilizadas como métrica de evaluación para modelos de clasificación las razones de probabilidades o likelihod ratio (LR) [15]. Estos se calculan a partir de la matriz de confusión y existe el LR positivo y el LR negativo, los cuales detallamos a continuación:

- **Likelihood ratio positivo**: Este se calcula como el cociente entre los enfermos que el modelo predijo como enfermos y los no enfermos que se clasificaron como enfermos,es decir, considerando las etiquetas de la matriz de confusión:

$$LR+ = \frac{TP}{FP}.$$

- **Likelihood ratio negativo**: Este es el caso contrario al anterior. Se calcula como el cociente entre los que sí estaban enfermos pero el modelo los clasificó como no enfermos y los que no estaban enfermos y el modelolos clasificó como no enfermos, es decir,

$$LR- = \frac{FN}{TN}.$$

Cabe destacar que el modelo será mejor cuanto mayor sea el LR positivo y cuanto menor sea el LR negativo.

Curvas ROC y AUC.

Se siguieron las referencias [16], [17] y la sección 15.2.2 de la referencia [18].

Las curvas Características Operativas del Receptor (ROC) son el conjunto de posibles valores de falsos positivos (1-especificidad) y verdaderos positivos (sensibilidad) que se obtienen al determinar un punto de corte c a partir de un modelo de clasificación. Es una técnica muy utilizada para cuantificar el valor diagnóstico de una prueba en toda su GAMa de posibles puntos de corte consecutivos para la probabilidad predicha de una variable resultado binaria. En el caso particular de que se considere un punto de corte $c = 0$, implica que todos los individuos se clasificarían como positivos, de manera que la sensibilidad sería igual a 1 (el 100 %) y la especificidad igual a 0 (el 0 %) y viceversa en caso de que se tomase el punto de corte máximo.

Para la representación de las curvas ROC se considera como eje x el valor de 1-especificidad y como eje y la sensibilidad. Así, la prueba será más exacta cuánto más cercana esté la curva ROC de la esquina superior izquierda. Sin embargo, si la curva ROC se encuentra sobre la diagonal la capacidad de discriminación de la prueba será nula. En la Figura 1.6.2 podemos ver un ejemplo de la representación de una curva ROC, asociada en este caso a un grupo de pacientes en estudio que se realizaron un test diagnóstico correspondiente a imágenes de tomografía computarizada[2] [16].

[2]Una tomografía computarizada (TC) es un procedimiento médico on imágenes que usa un equipo especial de rayos X para crear imágenes detalladas o exploraciones de regiones internas del cuerpo.

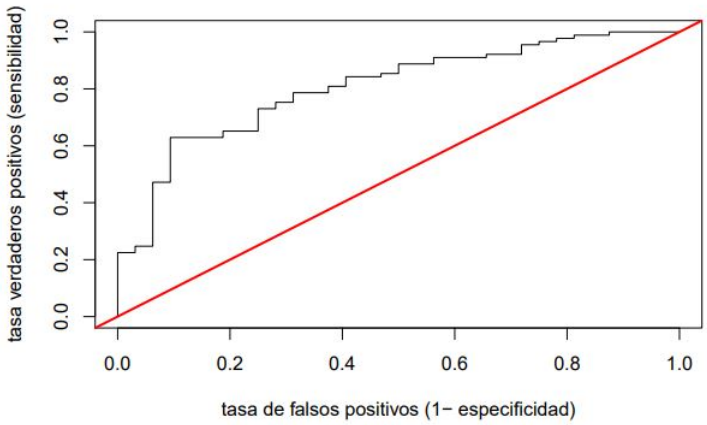

Figura 1.1: Representación de la curva ROC asociada a un caso médico de pacientes sometidos a un test diagnóstico correspondiente a imágenes de tomografía computarizada.

Siguiendo con este ejemplo en el que se quiere clasificar como enfermos y sanos, el área bajo la curva ROC (AUC) se considera una medida efectiva de la validez de una prueba diagnóstica, que variará entre 0 y 1. Este área se puede expresar entonces como sigue:

$$AUC = \int_0^1 ROC(t)dt,$$

siento ROC la función que denota a la curva ROC correspondiente. Así, la interpretación asociada a cada AUC será la siguiente:

- Si el AUC toma valor 1, esto se interpretaría como que la prueba diagnóstica que estamos analizando es perfecta en cuanto a diferenciar la enfermedad entre individuos enfermos y sanos. Este caso no ocurre en la práctica.
- Si el AUC toma valor 0, significa que el modelo no ha acertado con las predicciones en ninguno de los casos.
- Si el AUC toma valores entre 0 y 1, se interpretará como que, cuanto más cerca esté de 1 el valor del AUC, mejor clasificará la prueba diagnóstica.

Si nos fijamos en la representación de la curva ROC asociada al ejemplo (Figura 1.6.2), la diagonal de izquierda a derecha que divide al cuadrado de lado 1 en dos partes iguales tiene área 0,5 en cada una de sus partes. En caso de que la curva ROC fuese esta recta, habría el 50 % de probabilidades de que la prueba discrimine correctamente los sujetos enfermos y no enfermos. El valor 0,5 es, por tanto, el valor mínimo de AUC, puesto que un AUC nulo significaría que clasificó a todos los pacientes con enfermedad como negativos (no enfermos) y a todos los pacientes sin enfermedad como positivos (enfermos).

F1 score

La métrica F_1 se calcula a partir de la sensibilidad y especificidad. En particular, se calcula como la media armónica entre estas dos métricas mencionadas tal y como podemos ver a continuación

$$F_1 = 2 \cdot \frac{\text{precisión} \cdot \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}}.$$

La fórmula anterior solo se puede usar si la precisión y la sensibilidad tienen el mismo peso, es decir, ambas métricas tienen para nosotros igual importancia. La expresión en este caso sería de la forma

$$F_\beta = (1+\beta^2)\frac{\text{precision} \cdot \textit{sensibilidad}}{\beta^2\text{precision} + \text{sensibilidad}}$$

En medicina, cabe destacar que según estemos en un hospital o en un centro de salud, los pesos para la sensibilidad y especificidad serán muy diferentes.

Valores predictivos positivos y negativos

Otra de las propiedades que se pueden evaluar de un modelo de clasificación son los valores predictivos positivos y negativos, que nos informan de la probabilidad de que un positivo esté enfermo (valor predictivo positivo) o de que un negativo no lo esté (valor predictivo negativo). Cabe destacar que suelen emplearse cuando desconocemos si padece o no la enfermedad de estudio.

Capítulo 2

Aplicación en datos médicos

Como bien afirman diferentes estudios, la aplicación de la estadística a la investigación en salud es una necesidad social que requiere de la búsqueda de alternativas encaminada a mejorar su aplicación en la investigación, aportando soluciones a multitud de problemas de salud [19].

En este segundo capítulo presentamos dos problemas médicos reales en los que fue necesaria la aplicación de diferentes técnicas estadísticas, con el objetivo de obtener resultados y llegar a conclusiones relevantes en el mundo de la investigación sanitaria. El primero de ellos es un caso terapéutico relacionado con la vacuna antitetánica en pacientes tratados con anticoagulantes orales y el segundo es un caso clínico relacionado con el diagnóstico de sobrepeso y obesidad a partir del porcentaje de masa grasa en niños y adolescentes.

2.1. Caso terapéutico de la vacuna antitetánica

Presentaremos en primer lugar un estudio para evaluar la efectividad y seguridad de la vacuna antitetánica en pacientes tratados con anticoagulantes orales, que trató de determinar si existían diferencias significativas entre la vía intramuscular y subcutánea. Este trabajo fue publicado en la revista Frontiers y se puede consultar [20].

2.1.1. Introducción al problema médico

A pesar de que el tétanos es una enfermedad de poca incidencia en España, es un importante problema de salud pública debido a la alta tasa de mortalidad asociada. Los informes anuales de casos registrados muestran una caída gradual, con una media de 10 casos por año entre 2009 y 2015, resultando letales el 18,4 %.

El tétanos es una enfermedad que puede ser controlada por completo y se puede prevenir por vacunación. A pesar de esto, no es erradicable, principalmente debido a que la presencia del microorganismo que provoca esta enfermedad, que recibe el nombre de Clostridium Tetani, está muy extendido en el medio ambiente [21]. La eficacia de la vacuna es considerablemente alta, brindando protección a largo plazo y es recomendada para la población en general, aunque se debe administrar una dosis de refuerzo después de completar la vacunación primeria para mantener dicha inmunidad. Cabe destacar que la mayoría de los casos de tétanos aparecen en adultos que no han sido vacunados previamente, en particular en los mayores de 64 años. De los casos registrados en España entre el 2009 y el 2015, el 73,5 % no habían recibido ninguna dosis de vacuna y el 25,3 % sólo había recibido una.

El estudio de seroprevalencia publicado entre los años $2017 - 2018$ en España demostró que la inmunidad contra el tétanos supera el 90 % en edades de entre 6 y 49 años. En los pacientes de 50 años

en adelante, el porcentaje de población susceptible aumenta paulatinamente, llegando a un 40 % en el grupo de edad de entre 70 y 80 años [22]. Actualmente, la mayoría de los adultos mayores de 50 años tienen la vacunación incompleta o directamente no están vacunados, lo cual refleja el hecho de que el calendario de vacunación con 5 dosis antitetánicas no se introdujera hasta principios de la década de 1970 en España.

La edad media de la mayoría de los pacientes anticoagulados en Atención Primaria y que reciben tratamiento por fibrilación auricular es de más de 74 años, lo que significa que su cobertura de vacunas es probablemente baja. En pacientes anticoagulados no suele ser aconsejable el uso de la vía intramuscular (IM), por el hipotético riesgo de sangrado tras la punción, por lo que se recomienda la vía subcutánea (SC), incluso para vacunas administradas habitualmente por vía intramuscular, como la vacuna contra el tétanos.

Existen varios estudios que analizan la eficacia de la vacuna, pero solo considerando la vía intramuscular, y otros que comparan la efectividad de las dos vías, aunque no hay uniformidad de resultados. Además, las reacciones locales en la mayoría de los casos fueron más frecuentes con la vía subcutánea que con la intramuscular [23], aunque, cabe destacar que además de la vía en sí, el tamaño de la aguja también puede tener influencia [24].

Si revisamos literatura de estos aspectos en relación a otras enfermedades con evolución similar, en las vacunas contra la hepatitis B y la influenza, se ha demostrado la seguridad de la vía intramuscular en pacientes con alteraciones de la coagulación y, por este motivo, en el año 2006 la guía de los Centros para el Control y la Prevención de Enfermedades (CDC) recomienda la vía intramuscular para la vacuna antitetánica, que previene el tétanos y la difteria, sujeta en cada caso a la opinión y juicio del médico. Respecto a la vacuna contra la influenza, existe una revisión bibliográfica que compara la eficacia de las vacunas vía intramuscular y vía subcutánea, pero no existe ningún estudio en la literatura que evalúe la seguridad y eficacia de los tipos de vía intramuscular y subcutánea para la vacuna antitetánica en pacientes que reciben terapia anticoagulante oral (OAT).

2.1.2. Material y métodos

Previo al estudio

Se realizó un taller de capacitación para todos los investigadores participantes en el ensayo clínico, cubriendo técnicas, recolección de datos y medición de variables de estudio.

Diseño y período

Consistió en un ensayo clínico aleatorizado (ECA) independiente, doble ciego, en el que se compararon dos grupos de pacientes tratados todos ellos con anticoagulantes orales y el período de estudio comenzó en enero de 2009, con una previsión inicial de participación de 24 meses.

Objetivo

Fueron dos los principales objetivos del estudio: comparar la eficacia y seguridad de las vías de administración subcutánea e intramuscular de la vacuna DTP (mezcla de tres vacunas que inmuniza contra tres enfermedades causadas por bacterias: la difteria, bordetella pertussis (la tos ferina/tos convulsa) y el tétanos) en adultos anticoagulados.

Población

La población estaba formada por todos aquellos pacientes de 15 centros de salud del Área de Atención Primaria de Vigo en tratamiento monitorizado con anticoagulantes orales. Cada grupo recibió

dosis de vacuna antitetánica por diferente vía, intramuscular o subcutánea. La aleatorización se realizó dentro de una estratificación de 3 niveles, en función del número de dosis de vacuna requeridas para una inmunización exitosa. Dentro de cada nivel, se realizó una aleatorización simple utilizando una hoja de cálculo. La toma de muestras fue realizada por la Fundación Galicia Sur (EOXI Vigo), a la que se dio acceso telefónico a todos los investigadores.
El médico desconocía la vía de administración original en la visita de control para detectar efectos secundarios.

Criterios de inclusión y exclusión

Los criterios de inclusión y exclusión fueron los siguientes:

- Criterios de inclusión: Pacientes indicados para recibir al menos una dosis de vacuna antitetánica y tratados con anticoagulantes orales. Este criterio se aplicó a aquellos cuyo registro de vacunación era desconocido o incierto, o no habían sido vacunados. Personas que dieron su consentimiento informado por escrito para recibir la vacuna y participar en el estudio.

- Criterios de exclusión: Reacción local grave a dosis anteriores, afectando toda la zona de la extremidad donde se había inyectado la vacuna. Trastornos neurológicos periféricos causados por dosis previas. Reacción anafiláctica severa por dosis previas o por alguno de sus componentes. Mal control hematológico en los dos meses anteriores. Personas con estados de enfermedad terminal, enfermedad grave, afectados negativamente por patología crónica, inmovilizados o en estado inmunosupresor. Mujeres embarazadas o lactantes.

Los pacientes que cumplían los criterios de inclusión fueron reclutados por sus médicos de familia (MF) en los centros de salud de atención primaria correspondientes. En la primera visita, se registraba el estado de vacunación de cada paciente, consultando la historia clínica del mismo o preguntando en los casos en los que no estaba registrada esta información. Posteriormente, el médico evaluaba si el paciente había sido adecuadamente vacunado, en cuyo caso se le excluía del estudio, o si el paciente necesitaba completar o iniciar el proceso de vacunación. Las guías utilizadas fueron las emitidas por el Ministerio de Sanidad español en el año 2008 [25].

Variables

Las variables explicativas se dividían en sociodemográficas y terapéuticas, como, por ejemplo, la aparición de síntomas generales (fiebre, malestar general, cefalea, debilidad, artralgias) y la ocurrencia de cualquier efecto adverso grave, fatal o potencialmente mortal para el paciente, que resulte en incapacidad o requiera hospitalización. De entre todas las recogidas se consideraron para el estudio las siguientes:

- Sexo (hombre/mujer): toma valor 0 si el paciente es hombre y valor 1 si es mujer.

- Edad: edad de cada paciente en años.

- Tipo de vía (intramuscular/subcutánea): toma valor 0 si la vacuna fue administrada por vía intramuscular y valor 1 si fue administrada por vía subcutánea.

- Serología inicial: nivel de anticuerpos en sangre de cada paciente en el momento previo a recibir las dosis correspondientes.

- Serología final: nivel de anticuerpos en sangre de cada paciente en el momento final del procedimiento, posterior a recibir las vacunas correspondientes.

- Dosis requeridas: número de dosis de vacuna antitetánica que el paciente requería. Toma valores entre cero y tres dosis.

- Dosis aplicadas: número de dosis de vacuna antitetánica que el paciente finalmente recibió. Toma valores entre 0 y 3 dosis.

Como variables resultado para evaluar la seguridad de las dosis, se consideraron las variables dicotómicas:

- Dolor (sí/no): toma valor 0 si el paciente no sintió dolor y 1 en caso contrario.
- Eritema (sí/no): toma valor 0 si el paciente no sufrió eritemas y 1 en caso contrario.
- Tumefacción (sí/no): toma valor 0 si el paciente no sufrió tumefacciones y 1 en caso contrario.
- Hematoma (sí/no): toma valor 0 si el paciente no sufrió hematomas y 1 en caso contrario.
- Granuloma (sí/no): toma valor 0 si el paciente no sufrió granulomas y 1 en caso contrario.

Para evaluar la eficacia de las vacunas se consideró como variable resultado el aumento de anticuerpos, calculada como la diferencia entre los anticuerpos contra el tétanos al inicio y final del proceso de vacunación.

Tamaño muestral

Respecto al tamaño de la muestra, suponiendo que el porcentaje de efectos secundarios locales para la vía intramuscular fue de 30 %, y que el aumento esperado en los efectos secundarios locales para la vía subcutánea utilizando un enfoque bilateral con un intervalo de confianza de 95 % fue de 18 % y un riesgo beta de 0,20, calculamos que serían necesarios 115 pacientes para cada grupo. En vista de las posibles pérdidas de datos de 15 %, el tamaño final de la muestra se fijó en 135 pacientes para cada grupo. Basándonos en este tamaño de muestra, estimamos que se podría detectar una diferencia media de 3 UI/ml en los niveles de anticuerpos.

Análisis estadístico

En el análisis estadístico, tras la lectura de los datos y siguiendo el protocolo usual, se realizó un análisis exploratorio inicial.

Se estudió la normalidad de las variables cuantitativas con el test de Shapiro Wilk considerando un nivel de significación $\alpha = 0{,}05$. Se empleó para eso la función *shapiro.test()* del paquete stats [26].

En el caso de existir valores perdidos, se imputaron las observaciones correspondientes por el valor de la mediana en aquellas variables cuantitativas que no fuesen gaussianas y por el valor de la media en las variables gaussianas.

Tras el análisis exploratorio inicial, la imputación de los datos y el estudio de la normalidad, llevamos a cabo el análisis descriptivo considerando la media aritmética y la desviación estándar para variables cuantitativas normales, la mediana, el percentil 25 y 75 para variables cuantitativas que no cumpliesen el test de normalidad y frecuencia y porcentaje para variables cualitativas.

Se llevaron a cabo gráficos oportunos, como diagramas de dispersión, que se construyeron a partir de la función *xyplot* del paquete lattice [27], y representaciones de funciones de densidad e histogramas, que se representaron con la función *ggplot()* del paquete ggplot2 [28].

Se llevó a cabo el análisis de la correlación entre las variables a partir de la matriz de correlaciones, con el fin de no incluir información redundante en el análisis multivariante. Para la representación gráfica de esta matriz hicimos uso del comando *heatmap* la librería seaborn del entorno Python.

Se llevó a cabo un análisis bivariante según la variable tipo de vía, por un lado de las variables explicativas, tanto sociodemográficas como terapéuticas y, por otro, de las variables resultado. Se aplicó la la prueba ji cuadrado para las variables cualitativas y la prueba U de Mann-Whitney para las variables cuantitativas.

En el análisis multivariante, se construyeron modelos lineales generalizados (GLM) a partir de la función *GLM()* del paquete stats [26] cn las variables resultado consideradas para evaluar la seguridad y modelos GAMLSS para mixtura de normales con la función *gamlss.mx()* del paquete gamlss.mx [13] para evaluar la efectividad, comparando ambas posibilidades para las variables explicativas con la función *GAIC()* del paquete gamlss [29].

Empleamos la función *texreg()* del paquete texreg [30] para visualizar las estimaciones de los coeficientes de cada uno de los modelos, tanto para los de seguridad como para los de efectividad, y los Odds Ratio asociados a las variables explicativas cualitativas de los modelos de seguridad.

Para el estudio de la efectividad, se seleccionó como variable resultado el aumento de anticuerpos tras recibir las dosis de vacuna antitetánica, la cual se construyó como diferencia entre la serología final la serología inicial.

Resultados

Como resultados, se detectaron valores perdidos en algunas de las variables del dataset.

Los correspondientes a las variables serología final y serología inicial fueron imputados por la mediana correspondiente. En las variables cualitativas de interés con valores perdidos, como son las dosis aplicadas, los valores perdidos eran en proporción insignificantes, por lo que decidimos prescindir de ellos.

Para los niveles de significación usuales y, en particular, para $\alpha = 0{,}05$, existen evidencias significativas para rechazar la hipótesis nula de normalidad en todos los casos de variables cuantitativas. Por tanto, en el análisis descriptivo se consideraron mediana y percentiles 25 y 75 (Tabla 2.1).

	Total[#]	**NA**
N	234	
Serología inicial	531,0 (40,0-1205,8)	
Serología final	4214 (1962-5001)	
Edad	73,0 (66,5-79,0)	
Tipo de vía (subcutánea)	117 (50,0)	
Sexo (hombre)	132 (56,4)	
Dolor (sí)	77 (34,4)	10
Eritema (sí)	27 (12,1)	10
Tumefacción (sí)	31 (13,8)	10
Hematoma (sí)	6 (2,7)	10
Granuloma (sí)	5 (2,2)	10

[#]Los datos se expresan como mediana (P_{25}-P_{75}) para variables cuantitativas y en frecuencia (%) para variables cualitativas.

Tabla 2.1: Análisis descriptivo de las variables de interés.

Podemos ver que existe una diferencia notable entre la mediana de la serología inicial y final, y este aumento tiene sentido debido a la aplicación de las dosis de vacuna antitetánica, a pesar de que existen personas que no se aplicaron ninguna dosis. Se consideró de interés visualizar el diagrama de dispersión de la variable serología inicial frente a la variable serología final (Figura 2.1.2).

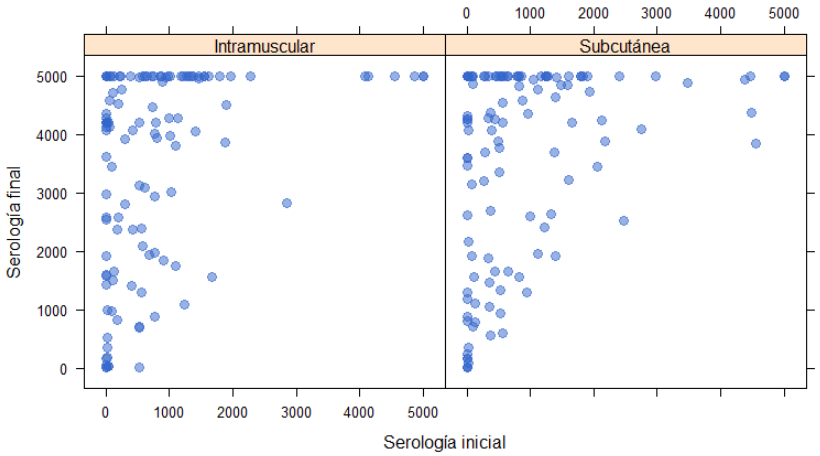

Figura 2.1: Diagrama de dispersión de las variables serología inicial y serología final según el del tipo de vía.

Observamos que la gran mayoría de individuos aumentaron su nivel de anticuerpos, a pesar de que sabemos que existen algunos con número de dosis aplicadas igual a cero. El historial de estos pacientes fue revisado y se comprobó que los registros eran correctos. Clínicamente no se contempla que pueda ser posible que un individuo aumente su nivel de anticuerpos de manera natural, sin haber recibido dosis de la vacuna correspondiente. La explicación a la que llegaron los investigadores principales del estudio fue que estos individuos pudieron vacunarse por otra vía ajena al estudio y no haber quedado registrado en el historial clínico. Por tanto se tomaron como valores perdidos aquellas dosis aplicadas igual a cero.

La representación de las correlaciones entre variables explicativas se puede ver en la Figura 2.2.

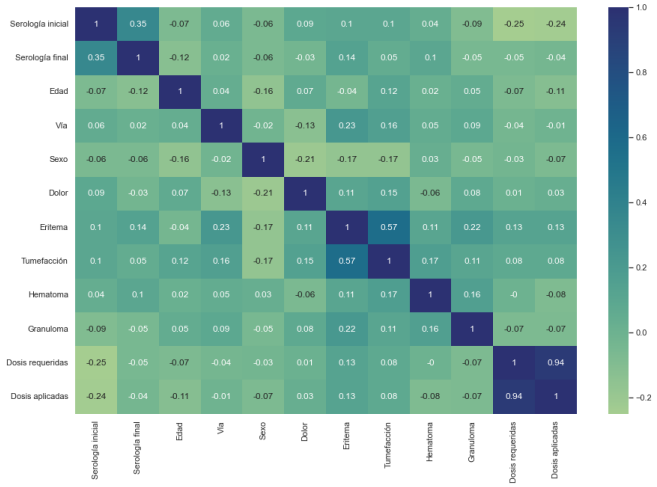

Figura 2.2: Representación de las correlaciones entre las variables consideradas.

Se puede ver que, tomando como referencia el valor 0,7, las únicas variables correladas fueron las dosis aplicadas y las dosis requeridas, por lo que seleccionaremos tan solo una de ellas en el análisis.

En las Tablas 2.2 y 2.3 podemos ver los resultados obtenidos del análisis bivariante según el tipo de vía, de las variables sociodemográficas y terapéuticas por un lado y de las variables resultado por otro.

	Tipo de vía[#]				
	Intramuscular	NA	**Subcutánea**	NA	p valor
Edad	73,00 (67-79)		73,62 (66-80)		0,930
Sexo					0,895
Mujer	50 (42,74)		52 (44,44)		
Hombre	67 (57,26)		65 (55,66)		
Serología inicial	531,00 (30-1025)	4	531,00 (71-1276)	1	0,363
Dosis requeridas					0,388
1	88 (75,21)		94 (80,34)		
2	10 (8,55)		5 (4,27)		
3	19 (16,24)		18 (15,38)		

[#] Los datos se expresan como mediana ($P_{25} - P_{75}$) o frecuencia (%).

$^{***}p$ <0,001; $^{**}p$ <0,01; $^{*}p$ <0,05.

Tabla 2.2: Análisis bivariante según el tipo de vía.

Se puede ver que ninguna de las variables sociodemográficas resultó significativa respecto al tipo de vía para los niveles de significación usuales y, en particular, para $\alpha = 0{,}05$. Veamos ahora qué ocurre con las variables resultado.

	Tipo de vía[#]		
	Intramuscular	**Subcutánea**	p valor
Dolor (sí)	45 (58,44)	32 (41,56)	0,074
Eritema (sí)	5 (18,52)	22 (81,48)	0,001**
Tumefacción (sí)	9 (29,03)	22 (70.97)	0,023*
Hematoma (sí)	2 (33,33)	4 (66,67)	0,695
Granuloma (sí)	1 (20,00)	4 (80,00)	0,377
Serología final	4214,00 (1978/5001)	4260,00 (1956/5001)	0,803

[#] Los datos se expresan como mediana ($P_{25} - P_{75}$) o frecuencia (%).

$^{***}p$ <0.001; $^{**}p$ <0.01; $^{*}p$ <0.05.

Tabla 2.3: Análisis bivariante según tipo de vía para las variables resultado.

Para los niveles de significación usuales y, en particular, para $\alpha = 0{,}05$, existen evidencias significativas para rechazar las hipótesis de independencia entre la variable eritema (sí/no) y tipo de vía y tumefacción (sí/no) y tipo de vía

En el análisis multivariante, se consideraron diferentes variables resultado según se analizase la seguridad o la efectividad.

- **Seguridad**

  Se obtuvieron las estimaciones de los coeficientes asociados a cada uno de los modelos, los cuales se pueden ver en la Tabla 2.4.

	Dolor	**Eritema**	**Tumefacción**	**Hematoma**	**Granuloma**
Intercepto	0,15	-2,56***	-1,96***	-3,59***	-3,87***
	(0,25)	(0,48)	(0,38)	(0,41)	(1,05)
Tipo de vía (subcutánea)	-0,60*	1,65**	1,01*		1,46
	(0,29)	(0,52)	(0,43)		(1,13)
Sexo (mujer)	-0,95**	-1,09*	-0,99*		
	(0,29)	(0,44)	(0,41)		
Serología inicial					-0,00
					(0,00)
AIC	279,71	151,79	173,78	57,28	47,35
BIC	289,95	162,02	184,01	60,69	57,58
Log likelihood	-136,86	-72,89	-83,89	-27,64	-20,67
Deviance	273,71	145,79	167,78	55,28	41,35
Num. obs.	224	224	224	224	224

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

AIC: Criterio de información Akaike

BIC: Criterio de información Bayesiano

Tabla 2.4: Regresión multivariante logística con dolor, eritema, tumefacción, hematoma, y granuloma consideradas como respuesta.

En esta tabla resumen que nos devuelve la función *screenreg()*, vemos las estimaciones de los coeficientes originales, es decir, el logaritmo de la conocida como odd del modelo, pero la interpretación de los mismos debería hacerse una vez aplicada la exponencial.

En la Tabla 2.5, se muestran los odds ratio asociados a los modelos anteriores.

	OR (IC 95 %)			
	Dolor	**Eritema**	**Tumefacción**	**Granuloma**
Tipo de vía (subcut.)	0,55 (0,31; 0,97) *	5,21 (2,01; 16,14) **	2,75 (1,22; 6,62) *	4,31 (0,62; 85,78)
Sexo (mujer)	0,39 (0,22; 0,68) **	0,34 (0,13; 0,79) *	0,37 (0,16; 0,82) *	

$^{***}p$ <0.001 ; $^{**}p$ <0.01 ; $^{*}p$ <0.05

Tabla 2.5: Coeficientes OR e intervalos de confianza para las variables predictoras cualitativas.

A la vista de los odds ratio, comparándolos con el valor 1, concluímos lo siguiente:

- El paciente tiene más probabilidades de sufrir dolor por vía intramuscular que por vía subcutánea y siendo mujer (valor 0) que siendo hombre (valor 1).
- El paciente tiene más probabilidades de sufrir eritema si la vía es subcutánea y si el paciente es mujer.
- El paciente tiene más probabilidades de sufrir tumefacciones si la vía es subcutánea y si el paciente es mujer.
- El paciente tiene más probabilidades de sufrir granuloma si la vía es subcutánea.

- **Efectividad**

Para evaluar la variable resultado asociado a la efectividad de las vacunas, se representó el histograma del aumento de anticuerpos junto con las funciones de densidad de la misma en función del tipo de vía (Figura 2.3).

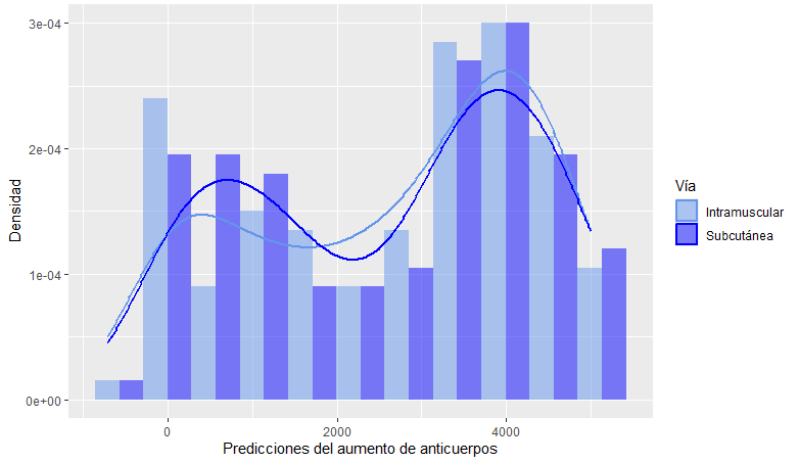

Figura 2.3: Histograma y funciones de densidad de la variable aumento de anticuerpos según el tipo de vía.

Observamos valores negativos para la variable aumento de anticuerpos. Los 8 casos en los que ha disminuido la cantidad de anticuerpos final comparado con la cantidad inicial, no presentan ninguna característica común. Posiblemente se justifique por la variabilidad personal en el sistema inmunitario a lo largo del tiempo o bien por incidencias en la gestión de muestras y/o determinación analítica. Además, hemos comprobado que en promedio la diminución de anticuerpos en

estos pacientes es de -199, clínicamente no significativo.

Respecto a la distribución que sigue la variable resultado de interés, vemos dos campanas bien diferenciadas. Esto sugiere la aplicación de un modelo GAMLSS para mixturas normales explicadas, explicado en la sección 1.5.4. Cabe destacar que en nuestro caso $K = 2$ por visualizar dos modas en el histograma (Figura 2.3).

Una de las preguntas que podría surgirnos a la vista del gráfico anterior es si estas dos campanas estarán separadas por alguna razón o motivo. Pudiese ser el caso, por ejemplo, de que se separasen según el sexo (hombres y mujeres), o según el dolor (sí no). Para estudiar estas hipotéticas situaciones, dividimos la muestra en dos grupos según el aumento de anticuerpos, considerando como punto de corte $c = 2500$. Así, construimos una nueva variable con la etiqueta 1 para aquellos pacientes con aumento de anticuerpos menor o igual a 2500 (primera campana) y con etiqueta igual a 2 para el caso contrario (segunda campana). Con el fin de conocer cual era la variable que mejor explicaba estos dos grupos, construimos un modelo GLM considerando como variable respuesta la que indicaba el número de etiqueta y como variables predictoras todas las demás. Se excluyó como variable explicativa el número de dosis aplicadas por estar correlada con el número de dosis requeridas.
En la Figura 2.4 podemos ver la importancia de las variables del modelo construido.

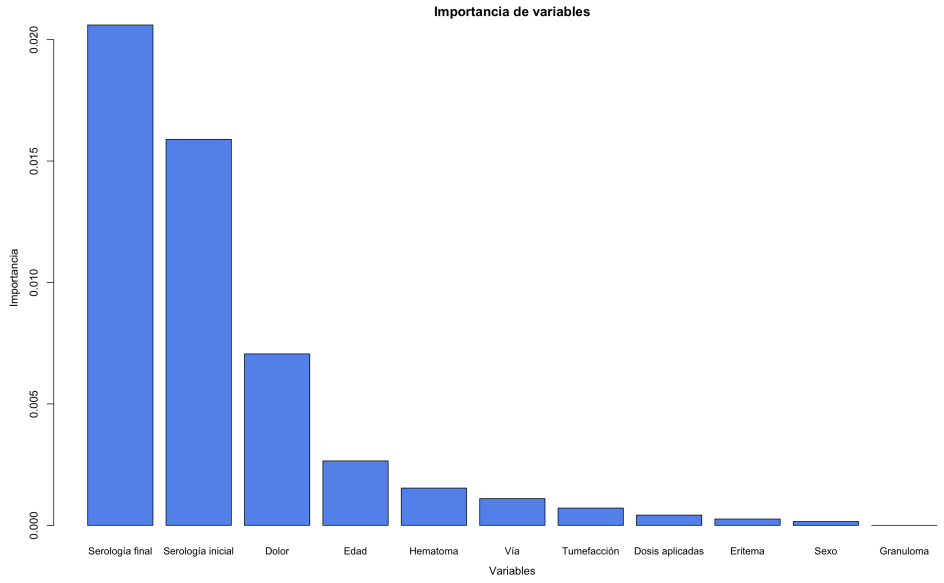

Figura 2.4: Diagrama de barras que mide la importancia de variables en el modelo con variable respuesta el número de etiqueta asignado.

Como era de esperar por la construcción de la variable respuesta, las variables que más influyen en la variable resultado fueron la serología final y la serología inicial, en orden descendente. La siguiente variable que mejor explica las dos campanas es el dolor, pero si representamos el aumento de anticuerpos según hayan sufrido dolor o no, no se diferencian estas dos modas que vimos anteriormente. Lo mismo ocurre con las demás variables explicativas.
A la vista de los resultados obtenidos, podemos concluir que la variable que determina la presencia de estas dos modas no es ninguna de las consideradas en el estudio, por lo que sería de mucho interés el detectar cual podría ser esta característica en futuras investigaciones.

Como la distribución es bimodal, todo indica que es resultado de combinar dos distribuciones

distintas. Así, ajustamos un modelo mixto con dos componentes (K=2) en el que cada componente sigue una distribución normal (campanas simétricas).

Como variables predictoras consideraremos, después de hacer pequeños análisis previos de las mismas, la serología inicial, el tipo de vía, la edad y el sexo. Cabe destacar que a la variable edad le aplicamos el cuadrado basándonos en que esta transformación suele mejorar el comportamiento del modelo y, siguiendo el mismo razonamiento, consideraremos la raíz cuadrada de la serología inicial.

Nuestro objetivo será, por tanto, construír modelos adecuados para explicar nuestra variable respuesta diferencia y cual es el modelo más adecuado de entre los candidatosque se presentan a continuación:

- Modelo 1: como variables explicativas serología inicial, sexo, tipo de vía y edad.
- Modelo 2: como variables explicativas serología inicial, sexo y tipo de vía.
- Modelo 3: como variables explicativas serología inicial y tipo de vía.
- Modelo 4: como variables explicativas serología inicial.

El modelo con menor AIC resultó ser el cuarto, cuya variable explicativa fue la serología inicial. Una vez seleccionado, llevamos a cabo la representación de los residuos con los gráficos worm plot, a partir de la función *wp()* del paquete gamlss [29]. En la Figura 2.5 se representa el gráfico worm asociado al modelo seleccionado, en el cual podemos ver que las observaciones que se encuentran fuera de las bandas de confianza son menos del 5 %.

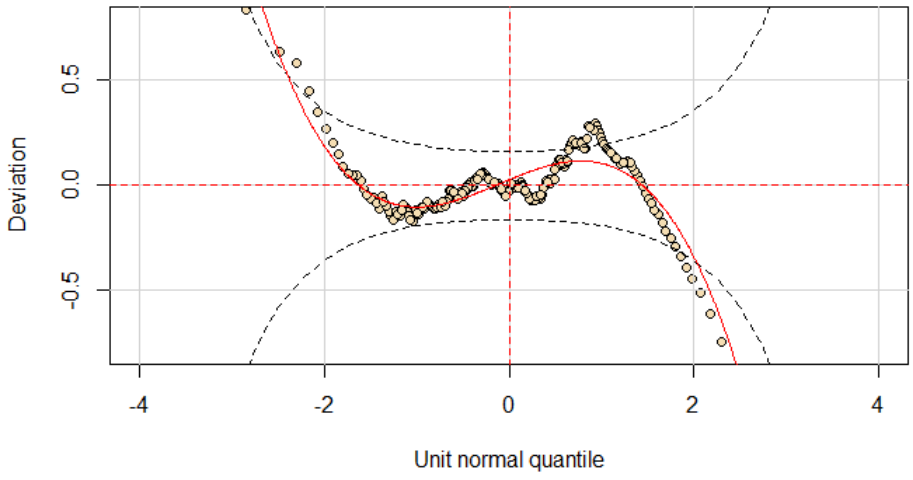

Figura 2.5: Gráfico worm asociado al modelo con variable respuesta aumento de anticuerpos y variable explicativa serología inicial.

En la Tabla 2.6 podemos ver las estimaciones para los coeficientes y demás resultados del modelo escogido según el criterio del AIC.

Mixing Family: c("NO". "NO")
Fitting method: EM algorithm
Call: GAMLSSMX(formula = dif ~sqrt(inisero). family = NO.
K = 2. data = datasetaux1. control = MX.control(plot = FALSE))
Mu Coefficients for model: 1
(Intercept) sqrt(inisero)
1289.208 -8.693
Sigma Coefficients for model: 1
(Intercept)
6.862
Mu Coefficients for model: 2
(Intercept) sqrt(inisero)
4716.84 -33.13
Sigma Coefficients for model: 2
(Intercept)
6.115
Probabilidades estimadas: 0.4577287 0.5422713
Degrees of Freedom for the fit: 7 Residual Deg. of Freedom 227
AIC: 3988.63
BIC: 4012.82
AIC: Criterio de Información Akaike
BIC: Criterio de Información Bayesiano

Tabla 2.6: Modelo aditivo generalizado de mixtura de normales con variable dependiente el aumento de anticuerpos.

Todos los análisis se realizaron utilizando el paquete de software estadístico R Studio versión 4.1.3.[31].

2.1.3. Discusión y conclusiones

Este estudio encontró un aumento similar en las tasas de anticuerpos para ambos tipos de vía, intramuscular y subcutánea, sin diferencias en términos de efectividad. Este hallazgo está en línea con otros estudios clínicos publicados hasta la fecha, que observan respuestas inmunes similares inducidas por inmunizaciones por vía subcutánea e intramuscular, en varios tipos de vacunas, que incluyen entre otras la hepatitis A; la hepatitis B; la influenza; el VIH o la meningocócica. Los ensayos clínicos en la población pediátrica del Reino Unido, Suecia y EE. UU. observaron también respuestas inmunogénicas muy similares para ambas vías de administración, en las vacunas antitetánicas. Sin embargo, la inoculación[1] por vía subcutánea de las vacunas contra la hepatitis A, hepatitis B, rabia, influenza y virus

[1]Una inoculación es la introducción voluntaria o accidental del virus o el principio material de una enfermedad.

del papiloma humano, genera una mayor respuesta inmune que la vía intramuscular con dosis equivalentes, que oscilan, según estudios, entre el 10 % y el 60 % de la utilizada por la vía intramuscular [32].

Además de la vía de administración, se describen otros factores externos que potencialmente modificarían la inmunogenicidad, como el sitio anatómico de inoculación, técnica y tamaño de la aguja, intervalo entre inmunizaciones, estrategia de vacunación e ingesta concomitante de fármacos. A mayores, se ha descrito la influencia de factores intrínsecos en la respuesta inmune a la vacunación, en el sentido de que los ancianos tendrían una menor respuesta y una mayor tasa de eliminación de anticuerpos [33]. Los hombres obtendrían un mayor grado de seroprotección con la vacuna contra el tétanos que las mujeres [33]. Existe una clara influencia genética en las respuestas a las vacunas, con un grado estimado de heredabilidad para el tétanos del 44 % [34]. Otro factor es el nivel preexistente de inmunidad: las personas que tienen títulos de anticuerpos contra el tétanos más altos antes de la vacunación, tienen tasas de seroprotección más altas después de la vacunación de refuerzo. Entre los adultos mayores, un estado mental positivo en el momento de la inoculación induciría mejores respuestas, mientras que el estrés crónico produciría respuestas de anticuerpos más bajas [35].

Respecto a la seguridad, los efectos locales, como eritema e hinchazón, fueron significativamente más pronunciados con la administración subcutánea. Otros estudios en población infantil serían consistentes con estos hallazgos, y describen mayor enrojecimiento e inflamación con la administración de la vacuna antitetánica por vía subcutánea que por vía intramuscular [23]. También resultó en nuestro estudio que la vía intramuscular produce más dolor que la vía subcutánea, algo plausible teniendo en cuenta que el número de terminaciones nerviosas nociceptivas es mayor en el músculo que en el espacio subcutáneo [36]. Aun así, otros estudios afirman lo contrario: más dolor con la vía subcutánea [23]. En general, la vía subcutánea puede estar asociada con irritación local, hinchazón y endurecimiento, decoloración de la piel, inflamación y formación de granulomas. La administración intramuscular se recomienda especialmente para las vacunas inactivadas con adyuvante, como la vacuna contra el tétanos [32]. El tamaño de la aguja también influiría en la aparición de efectos secundarios para las inoculaciones por vía intramuscular. Además, en este estudio se encontró que las mujeres presentan más efectos secundarios que los hombres, y no es posible corroborarlo con literatura alguna.

En los ensayos clínicos con la vacuna antigripal en población anticoagulada, en los cuales se compararon las dos vías de administración al igual que en el nuestro, se muestran resultados similares de manera que no se encuentran diferencias en cuanto a efectividad entre ambas vías [37].

Los obstáculos que se presentaron a lo largo del estudio estuvieron relacionados con la limitada financiación, la compleja coordinación debido principalmente a la participación de un gran número de investigadores y los problemas de adherencia experimentados por los pacientes participantes, en su gran mayoría ancianos dependientes en una variedad de de circunstancias como el transporte o la situación familiar. Además, en el momento en el que se inició el estudio eran pocos todavía los pacientes anticoagulados en la población ya que se estaba implementando este tipo de medicación en atención primaria. A pesar de esto, muchos de ellos no cumplían con los criterios de vacunación, otros no tenían interés en participar y algunos dependían de familiares, vecinos o amigos para asistir a las visitas médicas, algo que hizo más laborioso el proceso de selección y retrasó la consecución del tamaño muestral propuesto y, en consecuencia, la finalización del estudio.

Como limitación principal, la medida de la eficacia se verificase mediante el aumento del nivel de anticuerpos, dado que pequeñas variaciones en las concentraciones de anticuerpos entre grupos de personas pueden no ser clínicamente relevantes en lo que respecta a la protección que aporta la vacuna. Lo importante es la calidad de respuesta de los anticuerpos, pues solo un subgrupo de todos ellos puede neutralizar a los patógenos, además de que la respuesta innata, celular y de citocinas también influye en la eficacia de las vacunas. Es claro que estos complejos mecanismos e interacciones del sistema inmunitario a la respuesta de las vacunas no están todavía bien determinados y no existen marcadores

conocidos para monitorear esto. Lo que hicimos fue evaluar el aumento de anticuerpos, detectando valores negativos para esta variables: ocho casos con una cantidad de anticuerpos disminuida con respecto a la basal, sin características comunes en las variables analizadas. En promedio, se encontró que la disminución de anticuerpos en estos pacientes fue de 199. Esto posiblemente se explica por la variabilidad personal en el sistema inmunológico a lo largo del tiempo o, alternativamente, por incidentes en el manejo de muestras y/o determinación analítica.

Como fortalezas, se realizó en atención primaria, un ámbito donde los ECA adquieren mayor relevancia. Esto es debido a que es el entorno donde se aplican la mayoría de los tratamientos y donde se puede demostrar más la efectividad de las terapias, realizar una evaluación de forma realista. Muchos autores mantiene que la mayoría de las vacunas deben administrarse por vía intramuscular basándose en que existe evidencia de un alto grado de reactogenicidad reducida, y esto optimiza la inmunogenicidad [38]. A pesar de esto, se mantiene la pauta de vacunación subcutánea para grupos especiales con alto riesgo de hemorragia, por el peligro de hematomas secundarios tras la inyección. La contribución de este estudio es un punto fuerte, a pesar de que los resultados obtenidos no pueden extrapolarse a pacientes con trastornos hemorrágicos congénitos o a aquellos tratados con anticoagulantes orales de acción directa (ACOD).

A la vista de los resultados obtenidos, no existen diferencias en la inmunogenicidad de la vacuna antitetánica entre el tipo de vía de administración, por lo que la vía intramuscular puede estar preferible, dado que el riesgo de hematoma es mínimo en pacientes tratados con anticoagulados. Mientras que la reactogenicidad local de la vía subcutánea es mayor, la inyección intramuscular produce más dolor, con variaciones según el sexo, por lo que es recomendable incorporar a los pacientes en la toma de decisiones.

2.2. Caso clínico de medidas antropométricas

A continuación, presentamos un segundo caso médico, en particular, un caso clínico para predecir el riesgo de sobrepeso u obesidad a partir del porcentaje de masa grasa con medidas antropométricas. Este estudio todavía no fue publicado dado que se encuentra en período de revisión.

2.2.1. Introducción al problema médico

Tras la relajación de las medidas sanitarias adoptadas durante la pandemia de COVID-19, trastornos como la obesidad infantil han resurgido vigorosamente, facilitado, en gran medida medida, por la exposición forzada de los niños a las pantallas, ya sea para la educación o el ocio, y la reducción de la actividad física diaria como consecuencia del confinamiento. Además, los cambios en la dieta debido a problemas financieros han cambiado los hábitos de compra y consumo hacia alimentos ultraprocesados, que son más baratos, menos perecederos y tardan menos en preparar. De hecho, este desarrollo ha causado tal preocupación que el término se ha determinado como “covibesidad” para identificar este cambio [39].

La prevalencia de la obesidad ha aumentado en los últimos 25 a 30 años a tal punto que la Organización Mundial de la Salud (OMS) la ha definido como la pandemia del siglo XXI. En la actualidad, más de mil millones de personas en todo el mundo son obesas: de este número, 340 millones son adolescentes y 39 millones son niños. Este aumento hace no solo afecta a los países de ingresos altos, sino que también está presente en los países de ingresos bajos y medios. países. Si las tendencias actuales continúan, el número de lactantes y niños pequeños con sobrepeso Se estima que los niños aumentarán a 70 millones para 2025 (OMS) [40].

En clínica, el Índice de Masa Corporal (IMC) es el índice utilizado para evaluar el sobrepeso y la

obesidad en la infancia y adolescencia. dado a su fácil aplicación, su bajo coste y por no ser invasivo. Sin embargo, como el IMC no mide directamente la grasa corporal, puede sobrestimar adiposidad en niños con masa muscular aumentada y, por el contrario, subestimarla en niños con masa muscular reducida. Además, su la relación con la masa grasa varía con el sexo y la edad. Mientras que el IMC aumenta linealmente con la edad, la masa grasa se estabiliza o incluso disminuye entre los 8 y los 12 años de edad.

Un metanálisis de estudios realizados en niños y los adolescentes revelaron una baja sensibilidad para detectar la obesidad cuando solo se utilizó el valor de corte del IMC, y la evaluación de la adiposidad es importante para determinar en qué medida la obesidad puede afectar la fisiopatología, la respuesta al tratamiento y/o la salud resulta en una variedad de condiciones.

Existen diferentes alternativas que intentan solventar este problema, como por ejemplo diferentes modelos para determinar la composición corporal [41].
Dentro de los modelos bicompartimentales, existen métodos clásicos que hacen referencia a distintas ecuaciones, cuyo cálculo se basa en medidas antropométricas tales como pliegues cutáneos. Las medidas de perímetros y pliegues cutáneos han sido ampliamente utilizadas en la evaluación del estado nutricional. En estos modelos, se detectó una falta de precisión en la detección de alteraciones que ocurren en períodos cortos de tiempo, además de una menor precisión en la obesidad [42].
Respecto a los modelos tricompartimentales, estos dividen la composición corporal en masa grasa, masa magra y hueso. El método utilizado para medir los tres compartimentos es la radiografía de energía dual. absorciometría (DXA). DXA se puede utilizar para medir la masa grasa corporal, la masa magra, contenido mineral óseo, así como el porcentaje de grasa corporal de todo el cuerpo y subregiones, como los brazos y los piernas. A diferencia de los modelos multicompartimentales, DXA es relativamente preciso y, en comparación con los métodos antropométricos (modelos bicompartimentales), tiene la ventaja adicional de proporcionar medidas de composición corporal total y regional. Dicho esto, sin embargo, DXA no es portátil, lleva mucho tiempo y conlleva la exposición a radiación, limitando así su aplicación en la práctica clínica.
Los modelos multicompartimentales describen la composición corporal con mayor precisión, pero se limitan al entorno de la investigación, siendo inadecuados para el seguimiento o cribado de grandes grupos de pacientes. Sumado a esto, su uso es limitado debido a su costo y exposición a la radiación [44].

Por otro lado, la resonancia magnética cuantitativa (QMR) ha demostrado ser también una técnica a la altura de las anteriores, precisa, rápida y válida, con la ventaja adicional de mostrar también la masa muscular [45]. Esta no utiliza radiaciones ionizantes y no dependen de la hidratación de la masa magra. Parecería correlacionarse bien con DXA, particularmente para determinar el tejido graso y magro, pero se reserva para fines de investigación, debido a su alto costo.

Finalmente, otra alternativa es el análisis de impedancia bioeléctrica, un método más nuevo basado en el principio físico de impedancia. Proporciona estimaciones de agua corporal total determinada por impedancia, a partir de la cual se elaboran modelos predictivos luego se usa para estimar la masa magra [43].

Algunos autores [54] han sostenido que existe una fuerte correlación entre las mediciones individuales del grosor de los pliegues cutáneos y la masa grasa medida por DXA en población infantil, y que las ecuaciones antropométricas más precisas son aparentemente aquellas que incluyen los pliegues cutáneos tricipital y subescapular [47]. En base a esto, hemos querido llevar a cabo un estudio descriptivo observacional en búsqueda de una prueba diagnóstica para evaluar el sobrepeso y obesidad en niños y adolescentes.

2.2.2. Material y métodos

Previo al estudio

Previo al inicio del trabajo de campo, los investigadores asistieron a un taller de capacitación para adquirir las habilidades necesarias y estandarizar el procedimiento para la obtención de medidas antropométricas. Una vez seleccionadas las escuelas, se contactó a los responsables para informarles sobre el estudio. Si aceptaban participar, se planificaba la estrategia y se delineaba el cronograma, los procedimientos y las acciones. En esta reunión se entregaron al colegio los folletos informativos y consentimiento informado para ser entregados a los alumnos, quienes debían hacerlos diligenciar y firmar antes de poder proceder a la medición de sus datos antropométricos.

Las explicaciones en relación con el estudio fueron dadas a las personas responsables y/o a cualquier docente que las solicitara. Los números de teléfono de los investigadores también se proporcionaron en los folletos informativos, para que los padres/tutores pudieran resolver cualquier consulta o duda que pudieran tener en relación con el estudio.

El procedimiento se acordó con cada escuela, teniendo en cuenta las modalidades más adecuadas sugeridas por las personas a cargo. En las fechas acordadas, tres miembros del equipo de investigación visitaron la escuela: dos tomaron y archivaron las medidas, mientras que el tercer miembro de apoyo ajustó el ritmo de reclutamiento y acompañó a los alumnos desde sus aulas hasta la sala de medición.

Diseño y período

El estudio consistió en un estudio observacional transversal para evaluar las medidas antropométricas de la población escolar de Vigo, iniciado en Mayo-Junio de 2009.

Objetivo

Dado que las medidas antropométricas son totalmente accesibles en la atención primaria ambiental, fáciles de implementar, seguras y bien tolerado por los niños [48], este estudio buscó generar una ecuación predictiva para evaluar el porcentaje de masa grasa en una población infantil a partir de valores antropométricos, con el fin de clasificar el riesgo de sobrepeso u obesidad en niños y adolescentes de entre 11 y 17 años.

Población

La población de estudio estuvo formada por escolares de ambos sexos de 11 a 17 años del área metropolitana de Vigo que asistiesen a los colegios de Vigo, tanto públicos como concertados, un total de 10.747. El desglose arroja un 60 % en colegios concertados y un 40 % en colegios públicos, con 2.741 en 1º de Educación Secundaria Obligatoria, 2.789 en el segundo año, 2.735 en el tercer año y 2.482 en el cuarto año. Finalmente, el tamaño de las muestra fue de 577 escolarizados.

La aleatorización por conglomerados se realizó con las escuelas como unidades de muestreo. Un total de 343 alumnos procedían de colegios concertados y 228 de colegios públicos. Los alumnos de cada uno de los cuatro cursos anteriores representaron el 25 % de las muestras respectivas.

Criterios de inclusión y exclusión

Se excluyeron los alumnos sin consentimiento informado firmado por los padres/tutores, así como aquellos que padecieran enfermedades o trastornos que afectaran los valores antropométricos. La infor-

mación escrita sobre las medidas tomadas se proporcionó a las familias de los alumnos que lo solicitaron.

Variables

Las variables de estudio fueron tanto sociodemográficas como clínicas, estas últimas relacionadas con el sobrepeso y obesidad. Las explicaremos en detalle a continuación.
Como variables sociodemográficas se consideraron:

- Sexo: toma valor 0 si el paciente es hombre y valor 1 si es paciente es mujer.
- Edad: edad de cada paciente en años.

Como variables clínicas relacionadas con el sobrepeso y obesidad fueron de interés:

- Peso: peso de cada paciente en kilogramos.
- Altura: altura en centímetros de cada paciente.
- Masa grasa: porcentaje de masa magra de cada paciente. Se definen como sujetos obesos aquellos con porcentajes superiores al 25 % en hombres y al 33 % en mujeres. Se consideran límites valores entre el 21 y el 25 % en hombres y entre el 31 y el 33 % en mujeres. Los valores normales son del orden de 12 a 20 % en hombres y de 20 a 30 % en mujeres [49], [50].
- Riesgo de sobrepreso u obesidad a partir del porcentaje de masa grasa: toma valor 1 si el paciente está en riesgo de sobrepeso u obesidad según el porcentaje de masa grasa y valor 0 en caso contrario. Se consideraron en riesgo a los hombres con porcentaje de masa grasa mayor al 25 % y a las mujeres con porcentaje de masa grasa superior al 31 %.
- Masa magra: porcentaje de masa magra de cada paciente.
- Agua: porcentaje de agua de cada paciente.
- Índice de Masa Corporal (IMC): Índice de Masa Corporal obtenido aplicando la fórmula

$$\text{IMC} = \frac{\text{Peso (kg)}}{\text{Altura}^2(m^2)}.$$

- Clasificación según el IMC: se aplicaron los criterios de la Organización Mundial de la Salud (OMS): peso normal, IMC 18,5-24,9 kg/m^2; sobrepeso, IMC 25-29,9 kg/m^2; obesidad grado I, IMC 30-34,9 kg/m^2; obesidad grado II, IMC 35-39,9 kg/m^2 y obesidad grado III, IMC $\geq$ 40 kg/m^2) [40], [49].
- Índice Cintura Cadera (ICC): Índice Cintura Cadera obtenido aplicando la fórmula:

$$\text{ICC} = \frac{\text{Perímetro cintura (cm)}}{\text{Perímetro cadera (cm)}}.$$

- Índice Cintura Altura (ICA): Índice Cintura-Altura obtenido aplicando la fórmula:

$$\text{ICA} = \frac{\text{Perímetro cintura (cm)}}{\text{Altura (cm)}}.$$

Como medidas antropométricas se analizaron 3 diámetros, 7 perímetros y 8 pliegues:

- Diámetros: diámetro biepicondíleo del húmero, diámetro biestiloideo del radio y diámetro biepicondíleo del fémur.

- Perímetros: perímetro del brazo contraído, del brazo relajado, de la cadera, de la cintura, cefálico, de la muñeca y de la pierna.

- Pliegues: pliegue tricipital, bicipital, suprailíaco anterior, subescapular, abdominal, pectoral, del muslo anterior y del medial de la pierna.

Cabe destacar que para los pliegues se tomaron tres medidas, una detrás de otra, y se calculó el valor medio correspondiente.

Tamaño muestral

Calculamos el tamaño de la muestra sobre la base de una prevalencia de sobrepeso estimada del 17 %, un intervalo de confianza del 95 % de diferencia del 3 % y una población escolar total de 10.747 estudiantes, de los cuales 571 debían ser reclutados ("Calculadora de tamaño de muestra QuestionPro", disponible en: https://www.questionpro.com/es/calculadora-de-muestra.html)

Análisis estadístico

Se llevó a cabo un análisis exploratorio inicial de los datos. Se representaron diagramas de cajas con la función *boxplot()* del paquete graphics [31] para detectar valores atípicos y se comprobó cuales de ellos debían ser excluídos del estudio por encontrarse fuera del rango posible de valores.
Por otro lado, se construyeron modelos lineales generalizados con la función *glm()* del paquete stats [26] para analizar la distancia de cook y estudiar en qué observaciones esta distancia era mayor a 0,5, considerando como variable resultado la masa grasa y como independientes cada una de las demás ajustadas con edad y sexo.

Se estudió la normalidad de las variables cuantitativas con el test de Shapiro Wilk considerando un nivel de significación $\alpha = 0{,}05$ a partir de la función *shapiro.test()* del paquete stats [26].

En el análisis bivariante se consideró la variable sexo para analizar si existían diferencias relevantes según el sexo de las variables sociodemográficas y de las medidas antropométricas. Para eso, se llevaron a cabo gráficos bivariantes con la función *ggplot()* del paquete ggplot2 [28] según la edad y el sexo (Anexo 1).

Para el análisis multivariante se consideró como variable principal el riesgo de sobrepeso u obesidad a partir del porcentaje de masa grasa.

Dada la naturaleza de esta variable, se construyeron modelos GAM con la función *gam()* del paquete mgcv de respuesta binaria, con el fin encontrar una fórmula de índices antropométricos que explicase adecuadamente el riesgo por sobrepreso u obesidad según el porcentaje de masa grasa. Para la selección de variables en Modelos Aditivos Generalizados, propusimos un algoritmo basado en la técnica de validación cruzada de K iteraciones, cuyo procedimiento se basa en el siguiente esquema:

1. Consideramos cada una de las posibles combinaciones de p variables explicativas por separado, pudiendo especificar un número máximo de variables.

2. Para cada subconjunto de variables explicativas, se construye un modelo GAM por el método de validación cruzada de K iteraciones, considerando el 70 % de la muestra inicial como muestra de entrenamiento y el 30 % como muestra de test.

3. Calculamos la media de las "area under the curve (AUC)" obtenidas en las K iteraciones del método de validación cruzada.

4. Se comparan las medias de los AUC obtenidas para cada combinación de variables explicativas y se selecciona la mayor.

Una vez seleccionado el modelo con AUC óptimo, se construyeron otros tres modelos GAM con otros tres cuyas variables son el Índice de Masa Corporal (SEIMC), el Índice Cintura Cadera (SEICC) y el Índice Cintura Altura (SEICA), respectivamente, junto con la edad y el sexo.

Para su comparación, se calcularon y representaron las curvas de características de operación del receptor (ROC) para estimar la probabilidad de falsos positivos y falsos negativos en la predicción de cada uno de los cuatro modelos a partir de la función *roc* de la librería pROC, además de calcular el AUC correspondiente a cada uno con el objetivo de comparar lo bien que clasificaban el riesgo de sobrepeso u obesidad (sí/no) según el porcentaje de masa grasa. Se representaron las curvas ROC asociadas a cada modelo con la función *plot.roc()* del paquete pROC [51] y, una vez comparadas gráficamente, se llevó a cabo una comparación analítica con la función *roc.test* del paquete pROC [51], junto con los intervalos de confianza calculados con la función *ci.auc* de esta misma librería.

También calculamos los siguientes valores para cada modelo: la sensibilidad; la especificidad; la precisión; los valores predictivos positivos (VPP); los valores predictivos negativos (VPN); la tasa de verdaderos positivos (TPR), la cual viene dada por verdaderos positivos entre positivos; la tasa de verdaderos negativos, la cual viene dada por verdaderos negativos entre negativos; la tasa de falsos positivos, la cual viene dada por falsos positivos entre negativos; la tasa de falsos negativos, la cual viene dada por falsos negativos entre positivos. Para determinar estos valores se hizo uso de la función *ci.coords* de la librería pROC.
También se calcularon los likelihood negativos (NLR) y los likelihood positivos (PLR).

Finalmente, se diseñó una calculadora online con el paquete shiny [52] de R para su posible uso en clínica.

Resultados

En el análisis exploratorio inicial de los datos se detectaron signos de puntuación inusuales, por lo que fueron sustituídos por puntos (indicador decimal).

En el análisis de valores atípicos e influyentes, de setectaron 16 casos de datos influyentes, 19 registros con valores fuera de los rangos habituales , y 52 con valores perdidos. Todos ellos fueron excluídos del estudio, representando un 15,08 % del total.

A mayores, los investigadores consideraron motivo de exclusión no tener todas las variables completas, por lo que también fueron eliminados del estudio.

Tras llevar a cabo la depuración de los datos, se llevó a cabo el análisis descriptivo de los mismos considerando las variables que fuesen de interés, el cual podemos ver en la Tabla 2.7.

	Total*	NA		Total*	NA
N	490		Perímetros (cm)		
Sexo (mujer)	281 (48,70)		Cefálico	55,77 (4,81)	2
Edad	13,64 (1,41)	9	Brazo contraído	26,99 (8,93)	2
Peso (kg)	57,07 (12,23)	2	Brazo	25,36 (12,79)	2
Altura (cm)	165,90 (60,53)	2	Muñeca	16,26 (16,81)	3
Masa grasa (%)	20,71 (8,38)	7	Cintura	69,20 (9,64)	2
Riesgo (sí)	80 (16,3)	7	Cadera	88,46 (9,57)	2
Masa magra (%)	45,14 (14,59)	8	Pierna	39,29 (108,39)	14
Agua (%)	33,26 (8,21)	9	Pliegues (mm)		
Índice de masa corporal (kg/m^2)	21,31 (4,87)	2	Pectoral	8,41 (4,35)	2
Clasificación (%)		2	Bicipital	9,51 (5,35)	2
Bajo-peso	92 (18,8)		Abdominal	17,53 (8,62)	1
Normo-peso	334 (68,2)		Suprailíaco	12,32 (7,15)	1
Sobrepeso	53 (10,8)		Muslo	22,03 (10,58)	1
Obesidad grado I	10 (2,0)		Pierna	16,06 (7,62)	1
Obesidad grado II	1 (0,2)		Subscapular	11,49 (5,81)	1
Obesidad grado III	0 (0,0)		Tricipital	15,78 (6,73)	1
Diámetros (cm)		3			
Humero	6,30 (3,18)				
Radio	4,97 (0,50)				
Fémur	8,80 (0,93)				

*Los datos se expresan en media (DE) o número (%).

Tabla 2.7: Análisis descriptivo tras la depuración de los datos.

Para el nivel de significación establecido, existieron evidencias significativas para rechazar la normalidad en todas las variables cuantitativas consideradas en el estudio, exceptuando el diámetro del fémur y el perímetro de la cadera para los cuales se aceptó la hipótesis nula de normalidad.

En la Tabla 2.2.2 podemos ver las estimaciones de los coeficientes asociadas a cada modelo construido en el análisis multivariante, junto con su desviación típica y su significación.

	SEPA2 + 1	**SEIMC**	**SEICC**	**SEICA**
Intercepto	51,41***	-8,86***	-30,56***	-4,22
	(11,49)	(1,91)	(3,79)	(6,12)
Sexo	-1,40***	-0,66	0,47	0,17
	(0,40)	(0,34)	(0,35)	(0,34)
Edad		-0,58***	-0,53***	-0,23
		(0,13)	(0,12)	(0,13)
Peso	3,22***			
	(0,56)			
Altura	-5,80***			-2,62***
	(1,01)			(0,56)
Perímetro de cintura			3,81***	4,58***
			(0,57)	(0,48)
Perímetro de cadera			0,37	
			(0,49)	
Perímetro de pierna	-1,39			
	(0,81)			
Pliegue pectoral	0,45			
	(0,34)			
Pliegue abdominal	0,79**			
	(0,30)			
IMC		0,68***		
		(0,07)		
Num, obs,	490	490	490	490
Nagelkerke R^2 ajust.	0.50	0.46	0.34	0.42
Generalized AIC	234.99	250.11	304.51	280.08

$^{***}p < 0{,}001$; $^{**}p < 0{,}01$; $^{*}p < 0{,}05$.

Tabla 2.8: Estimaciones para los coeficientes con sus respectivas desviaciones típicas para cada uno de los predictores.

En la Tabla 2.9 se pueden ver los Odds ratio asociados a las variables explicativas cualitativas de los modelos anteriores.

	SEPA2 + 1$^{\#}$	**SEIMC**$^{\#}$	**SEICC**$^{\#}$	**SEICA**$^{\#}$
Sexo (hombre)	0,25 (0,40)***	0,52 (0,34)	1,60 (0,35)	1,19 (0,34)

$^{***}p < 0{,}001$; $^{**}p < 0{,}01$; $^{*}p < 0{,}05$.

$^{\#}$Los datos se expresan como OR(DE).

Tabla 2.9: Odds ratio (desviación típica) asociados a las estimaciones de los coeficientes de las variables predictoras cualitativas.
Vemos que en los dos primeros modelos ser hombre es protector respecto al riesgo de sobrepeso u obesidad, mientras que en los dos modelos restantes sucede lo contrario.

En la Tabla 2.10 podemos ver la comparación de los modelos en función de las métricas de evaluación calculados y los intervalos de confianza correspondientes.

	SEPA2 + 1$^{\#}$	**SEIMC**$^{\#}$	**SEICC**$^{\#}$	**SEICA**$^{\#}$
Sensibilidad	0,91 (0,83; 0,98)	0,90 (0,80; 0,96)	0,79 (0,68; 0,90)	0,84 (0,66; 0,95)
Especificidad	0,87 (0,79; 0,93)	0,87 (0,80; 0,95)	0,84 (0,72; 0,92)	0,81 (0,70; 0,96)
Valores predictivos positivos	0,58 (0,47; 0,72)	0,57 (0,47; 0,76)	0,49 (0,37; 0,63)	0,47 (0,37; 0,79)
Valores predictivos negativos	0,98 (0,96; 0,99)	0,98 (0,96; 0,99)	0,95 (0,93; 0,97)	0,96 (0,94; 0,99)
Tasa de verdaderos positivos*	91,25 (84,78; 97,06)	90,00 (82,80; 95,52)	78,75 (73,47; 87,10)	83,75 (73,79; 87,10)
Tasa de verdaderos negativos*	86,83 (81,62; 92,39)	87,07 (82,59; 93,98)	83,90 (76,33; 89,10)	80,98 (76,40; 95,05)
Tasa de falsos positivos*	13,17 (7,61; 18,38)	12,93 (0,06; 17,41)	16,10 (10,91; 23,67)	19,02 (4,95; 23,60)
Tasa de falsos negativos*	8,75 (2,94; 15,22)	10,00 (4,48; 17,20)	21,25 (12,90; 26,53)	16,25 (7,02; 26,21)
Precisión	0,88 (0,82; 0,92)	0,82 (0,77; 0,91)	0,83 (0,74; 0,88)	0,87 (0,82; 0,93)
Likelihood ratio positivo	7,19 (3,98; 14,81)	6,96 (4,00; 18,81)	4,97 (2,41; 10,85)	4,35 (2,24; 25,97)
Likelihood ratio negativo	0,10 (0,03; 0,22)	0,11 (0,04; 0,25)	0,25 (0,11; 0,45)	0,20 (0,05; 0,48)
AUC	0,937 (0,908; 0,966)	0,928 (0,896; 0,960)	0,861 (0,813; 0,910)	0,890 (0,847; 0,934)

$^{\#}$ Se expresan como estimaciones de los coeficientes (intervalo de confianza al 95 %).
* Estos datos se expresan como porcentajes.

Tabla 2.10: Sensibilidad, especificidad, precisiónm verdaderos negativos y verdaderos positivos, falsos negativos y falsos positivos, valores predictivos negativos y valores predictivos positivos, likelihood positivos y negativos y AUC para cada modelo.

En la Figura 2.6 podemos ver la representación de las curvas ROC asociadas a cada modelo junto con sus intervalos de confianza.

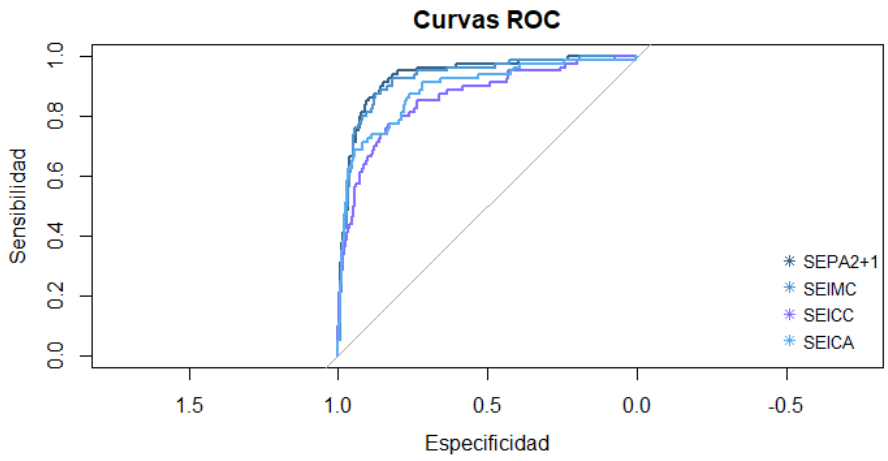

Figura 2.6: Representación de las curvas ROC asociadas a cada uno de los cuatro modelos.

Gráficamente se puede ver que el mayor AUC es el asociado al primero de los modelos.

En la comparación analítica de las curvas ROC se llegó a que, para un nivel de significación $\alpha = 0{,}05$, no existían diferencias significativas entre las curvas ROC asociadas al modelo SEPA2+1 y al del IMC.

Finalmente, en la Figura 2.7, podemos ver la calculadora online para predecir el riesgo de sobrepeso u obesidad a partir de los dos mejores modelos (SEPA2+1 e SEIMC), a la cual se puede acceder en la siguiente dirección web: https://isaudegal.es/shiny/overweight/

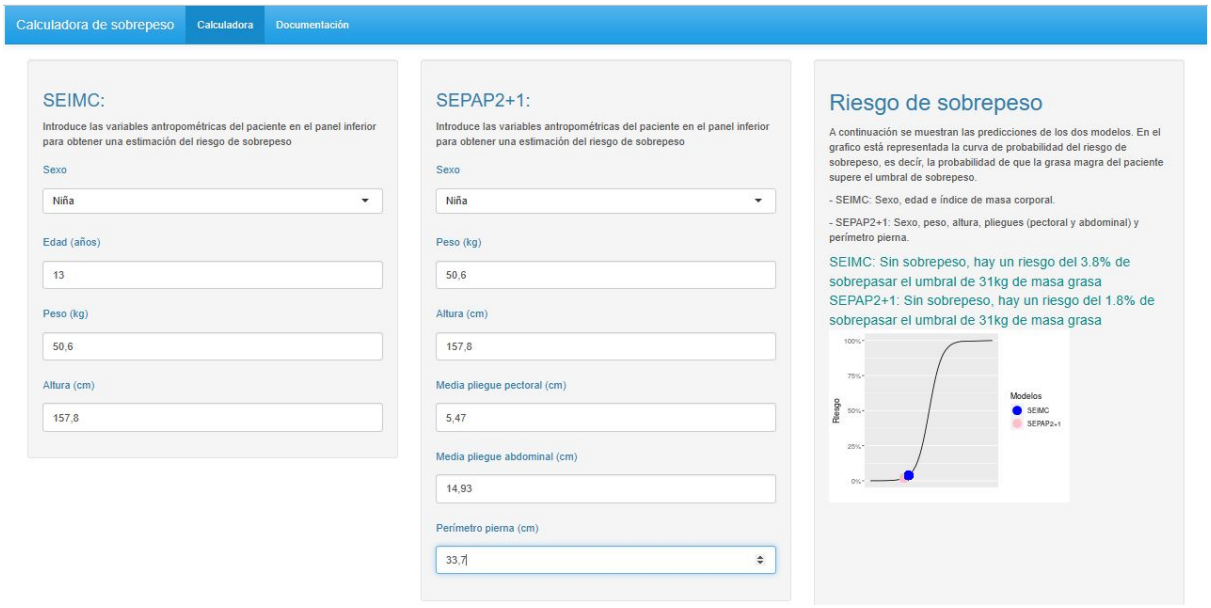

Figura 2.7: Calculadora online para predecir el riesgo de sobrepeso u obesidad a partir del porcentaje de masa grasa con modelos GAM.

2.2.3. Discusión y conclusiones

En la literatura existen una serie de ecuaciones para calcular la densidad corporal total [53]. Se basan en una agregación del grosor de los pliegues cutáneos, y con este fin, todos utilizan cuatro pliegues cutáneos (bicipital, tricipital, subescapular y suprailíaco), a diferencia de los modelos propuestos en este estudio. El modelo propuesto por Mi Goran en 1996 para estimar la masa grasa medida por DXA, que incluye entre las medidas el espesor de dos pliegues cutáneos (subescapular y tricipital) el peso, el sexo y la altura al cuadrado, y resultó tener un R^2 de 0,91 y una desviación estándar de 0,94 kilogramos de masa grasa en el análisis multivariante [54]. Ninguno de los modelos propuestos en los artículos científicos como alternativas al IMC utiliza medidas de los miembros inferiores, a pesar de que estas regiones son importantes en términos de funcionalidad en edades pediátricas. También hay que señalar que sólo algunos se basaron en muestras de tamaño reducido con individuos en edad pediátrica.

Wong et al [55] determinan la concordancia entre ocho pliegues cutáneos, utilizando un modelo multicompartimental para predecir la porcentaje de grasa corporal en 72 niñas blancas y 40 afroamericanas de 11 a 15 años años. En el análisis de Bland-Altman, las ecuaciones cuadráticas coincidieron más estrechamente con la ecuación de Slaughter, siendo este el que mejor estimó la masa grasa, y, a diferencia de los demás, consideró el pliegue de la pierna [53]. También se obtuvieron límites de concordancia al medir el pliegue tricipital y el pliegue de la pantorrilla en lugar de los pliegues subescapular y bicipital (ecuación [55] de Slaughter et al.).

Existen escasos datos de sensibilidad y especificidad de las diferentes ecuaciones publicadas para medidas antropométricas. En 2014, Wohlfahrt-Veje et al [56] analizó la concordancia entre una serie de medidas antropométricas y valores corporales de masa grasa estimados por DXA[2] en 2647 niños daneses.

Nuestros modelos, en comparación con los de Wohlfahrt-Veje, muestran generalmente una sensibilidad más alta y una especificidad ligeramente más baja. El modelo SEIMC incluye un menor número de variables explicativas, que, además, son más fáciles de obtener, no requieren entrenamiento especial, son menos susceptibles a las variaciones interobservador, y formar parte de la práctica rutinaria y sistemática en los exámenes clínicos en pediatría y en atención primaria. Por lo tanto, si nos centramos en la práctica clínica, parece más adecuado recomendar la ecuación del segundo modelo (SEIMC), dado que tiene un AUC cercano a 1, excelente sensibilidad y especificidad, y se obtiene mediante mediciones sencillas, haciéndolo accesible, fácil de usar y altamente eficiente. En casos específicos o en pacientes con riesgo de obesidad por factores genéticos (pacientes sindrómicos) o ambientales (como la obesidad de los padres), posiblemente sería recomendable evaluar el uso de la ecuación del modelo (SEPA2+1), que proporciona una mayor precisión.

Como limitaciones del estudio, se llevó a cabo en un entorno comunitario, con evaluación de escolares en sus escuelas, algo que limitaba en gran medida la elección del criterio de medición de la composición corporal. El método requerido tenía que ser portátil, fácil de usar, no invasivo, confiable, libre de radiación y asequible. Normalmente, el estándar de oro para la comparación de medidas corporales se basa en modelos multicompartimentales que incluyen estimaciones de peso, volumen corporal, densidad corporal, contenido mineral óseo y agua corporal total, algo que no se puede hacer de manera factible en entornos clínicos o comunitarios. Un método alternativo, que adquiere un valor especial a la hora de estimar la grasa corporal, tanto en individuos aislados como en grupos epidemiológicos, es el analisis de impedancia bioeléctrica o bioimpedancia. Es el método más utilizado, gracias a que su coste puede ser más o mens accesible, es fácil de usar y no es invasivo para determinar la masa grasa o magra. Además, se considera una herramienta válida para el seguimiento longitudinal.

[2]El DXA es una prueba de rayos X, no invasiva, que permite obtener una imagen para medir la densidad mineral ósea y valorar la salud de los huesos; así como medir la masa grasa y masa libre de grasa informando de la composición corporal.

Algunos autores encontraron excelentes grados de concordancia entre las bioimpedanciometría y DXA en la población pediátrica pero, validando sus hallazgos, utilizaron bioimpedanciometría eléctrica para validar datos antropométricos [57].

A la luz de esto, la bioimpedancia fue elegido como método de comparación.

Como fortalezas, se debe enfatizar que este estudio utilizó una muestra representativa de la población infantil con edades comprendidas entre los 11 y los 17 años, que nos ha permitido extrapolar los resultados a la población de origen y a otras poblaciones con características similares. Se eligió el rango de edad estudiado, teniendo en cuenta que los cambios en la composición corporal durante la adolescencia temprana (12-14 años) son más sensibles y significativas, e influyen directamente en la adolescencia tardía y etapas adultas [58].

Otra fortaleza que consideramos es que, usando un algoritmo interno que evaluó y comparó todas las combinaciones posibles de todas las variables de interés para lograr el modelo o ecuación con el mayor AUC posible, se elaboró un modelo optimizado para la estimación del riesgo de sobrepeso y obesidad obtenido. Esto se consideró así ya que asegura la mayor utilidad de diagnóstico posible. Además, el análisis se realizó sin imputación, ya que el valores perdidos fueron insignificantes, representaron menos del 5 % de los 577 participantes en la muestra, no se relacionaron con la variable resultado, y, además, los modelos GAM funcionan bien a pesar de la presencia de valores perdidos.

La calculadora online implementada que automatiza los cálculos y clasificación de niños y adolescentes en términos de sobrepeso y obesidad, optimiza el tiempo y mejorar la fiabilidad en la práctica clínica. La construcción y presentación de esta calculadora es un valor agregado de este estudio, ya que potencialmente puede tener un alto impacto en la eficiencia de las clínicas pediátricas en atención primaria y en otros contextos.

Bibliografía

[1] Pérez Hidalgo, M.D., Basulto Santos, J., Camúñez Ruiz, J.A. (2019). *Un antecedente histórico de regresión lineal la estimación mediana propuesta por Boscovich.* Gaceta de la Real Sociedad Matematica Española, ISSN 1138-8927, Vol. 22, Nº 2.

[2] Alexopoulos, E.C. (2010). *Introduction to multivariate regression analysis.*.PMID: 21487487; PMCID: PMC3049417.

[3] Lahera Rol, A., Pérez Olivarez, I., Hunte Roberts, V., & Ruiz Batista, E. (2018). *La estadística como necesidad en la investigación en salud.* Revista Información Científica, 97(4), 891-901. Disponible en: https://revinfcientifica.sld.cu/index.php/ric/article/view/1851/3836

[4] Matranga, D., Bono, F., Maniscalco, L. (2021). *Statistical Advances in Epidemiology and Public Health. Int. J. Environ. Res.* Public Health, 18, 3549. https://doi.org/ 10.3390/ijerph18073549

[5] Wood S.N. (2017). *Generalized additive models: an introduction with R.* Second Edition. Boco Raton: CRC Press.

[6] Izenman, A. J. (2013). *Multivariate regression. Modern multivariate statistical techniques* (pp. 159-194). Springer.

[7] Van Oijen, M. (2020). *Bayesian compendium, Cham: Springer Nature Switzerland AG.* The International Biometric Society. vol. 78(2), pages 813-815, June.

[8] Crujeiras Casas, R.M., Conde Amboage, M. (2019). *El modelo de regresión lineal simple.* Inferencia estadística-Grado en Matemáticas.

[9] Novales, A. (2010). *Análisis de regresión.* Departamento de Economía Cuantitativa de la Universidad Complutense de Madrid.

[10] Szumïlas, M. (2015). *Explaining odds ratios.* J Can Acad Child Adolesc Psychiatry.Erratum in: J Can Acad Child Adolesc Psychiatry. PMID: 20842279; PMCID: PMC2938757.

[11] Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R.* https://doi.org/10.1201/b21973

[12] Stasinopoulos, M., Rigby, B., Mortan, N. (2018). *Fitting an Interval Response Variable Using 'gamlss.family' Distributions.* R package version 5.0-1, ¡https://CRAN.R-project.org/package=gamlss.cens¿.

[13] Stasinopoulos M, Rigby B (2020). *Package gamlss.mx: Fitting Mixture Distributions with GAMLSS.*Disponible en https://CRAN.R-project.org/package=gamlss.mx

[14] Hossin, M. & S. M.N. (2015). *A Review on Evaluation Metrics for Data Classification Evaluations.* doi: 10.5121/ijdkp.2015.5201.

[15] Grimes, D. A., Schulz, K.F. (2005). *Refining clinical diagnosis with likelihood ratios.* The Lancet, Volume 365, pages 1500-1505, ISSN 0140-6736. Disponible en: https://www.sciencedirect.com/science/article/pii/S0140673605664227

[16] Roca Pardiñas, J. (2017). *Evaluación de pruebas diagnósticas y curvas ROC.*

[17] Kumar, R., Indrayan, A. (2011). *Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr.* doi: 10.1007/s13312-011-0055-4. PMID: 21532099.

[18] Steyerberg, E. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* 10.1007/978-0-387-77244-8.

[19] Lahera, Rol, A., Pérez Olivares, I., Hunte Roberts, V.G., Ruiz Batista E. (2018). *La estadística como necesidad en la investigación en Salud.* ISSN 1028-9933.

[20] Lago, F.I., Valladares Cabaleiro, M., Fernández Domínguez, M.J., Fernández Fernández, I., Clavería, A. Rodríguez Pastoriza, S., Roca Pardiñas, J. y Martín Miguel, M. V. (2022). *Effectiveness and safety of tetanus vaccine administration by intramuscular vs. subcutaneous route in anticoagulated patients: Randomized clinical trial in primary care.* Frontiers in Medicine. doi: 10.3389/fmed.2022.105498

[21] Yen, M.L. (2019). *Thwaites CL. Tetanus.* Lancet. 393:165768. doi: 10.1016/S0140-6736(18)33131-3

[22] Ministerio de Sanidad. (2020). *2º Estudio de Seroprevalencia en España.* Disponible en: *https://www.sanidad.gob.es/profesionales/saludPublica/prevPromocion/vacunaciones/comoTrabajamos/docs/EstudioSeroprevalencia_EnfermedadesInmunoprevenibles.pdf*

[23] Mark, A., Carlsson, R.M., Granström, M. (1999). *Subcutaneous versus intramuscular injection for booster DT vaccination of adolescents.* doi: 10.1016/s0264-410x(98)00410-1. PMID: 10217608.

[24] Jackson, L.A., Starkovich, P., Dunstan, M., Yu, O., Nelson, J., Dunn, J., Rees T., Zavitkovsky, A., Maus, D., Froeschle J.E. (2008). *Prospective assessment of the effect of needle length and injection site on the risk of local reaction to the fifth diphtheria-tetanus-acellular pertussis vaccination.* doi: 10.1542/peds.2007-1653

[25] *National Health System - Ministerio de Sanidad y Consumo.* (2008) [Internet] Madrid. Disponible en: http://www.mscbs.gob.es/en/organizacion/sns/docs/Spanish_National_Health_System.pdf

[26] The R Stats Package - R." https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

[27] Sarkar, D. (2008) *Lattice: Multivariate Data Visualization with R. Springer, New York.* ISBN 978-0-387-75968-5

[28] Wickham, H. (2016).*Elegant Graphics for Data.* Analysis. Springer-Verlag New York.

[29] Rigby, R.A., Stasinopoulos, D.M.. (2005). *Generalized additive models for location, scale and shape,(with discussion).* Appl. Statist., 54, part 3, pp 507-554.

[30] Leifeld, P. (2013). *Package texreg: Conversion of Statistical Model Output in R to LaTeX and HTML Tables.*

[31] R Core Team. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria. Disponible en: http://www.R-project.org/.

[32] Zhang, L., Wang, W., Wang, S. (2015). *Effect of vaccine administration modality on immunogenicity and efficacy.* Expert Rev Vaccines. doi: 10.1586/14760584.2015.1081067

[33] Hainz, U., Jenewein, B., Asch, E-, Pfeiffer, KP., Berger, P., Grubeck-Loebenstein B. (2005). *Insufficient protection for healthy elderly adults by tetanus and TBE vaccines. Vaccine.* 23(25):32325. doi:10.1016/j.vaccine.2005.01.085

[34] Newport, M.J., Goetghebuer, T., Weiss. H.A., Whittle H., Siegrist, C.A., Marchant A., et al. (2004). *Genetic regulation of immune responses to vaccines in early life. Genes Immun.* 5(2):1229. doi: 10.1038/sj.gene.6364051

[35] Zimmermann, P., Curtis, N. *Factors That Influence the Immune Response to Vaccination.* Rev. 32(2): e0008418. doi: 10.1128/CMR.00084-18

[36] Cook, I.F., (2020). Subcutaneous vaccine administration - an outmoded practice. Hum Vaccin Immunother. doi: 10.1080/21645515.2020.1814094.

[37] Delafuente, J.C., Davis, J.A., Meuleman, J.R., Jones. R.A. (1998). *Influenza vaccination and warfarin anticoagulation: a comparison of subcutaneous and intramuscular routes of administration in elderly men.* Pharmacotherapy. 18(3):6316

[38] Zuckerman, J. (2000). *The importance of injecting vaccines into muscle.* BMJ 321:12367. doi: 10.1136/bmj.321.7271.1237

[39] Khan M.A., Moverley Smith J.E. (2020). Çovibesity,.[a] new pandemic. Obes Med. doi: 10.1016/j.obmed.2020.100282. Epub 2020 Jul 21. PMID: 32835125; PMCID: PMC7371584.

[40] World Health Organization (WHO). (2021) *Obesity and overweight.* Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

[41] Phillips, S.M., Shulman, R.J. (2022). *Measurement of body composition in children. In: UpToDate, Hoppin AD (Ed), UpToDate, Waltham, MA.* Disponible en: http://www.uptodate.com/

[42] Casanova Román, M., Rodríguez Ruiz, I., Rico de Cos, S., Casanova Bellido, M. (2004). *Análisis de la composición corporal por parámetros antropométricos y bioeléctricos. Anales de Pediatría.* 1;61(1):2331. doi: 10.1016/S1695-4033(04)78349-6.

[43] Wan, C.S., Ward, L.C., Halim, J., Gow, M.L., Ho, M., Briody, J.N., Leung, K., Cowell, C.T. (2014). *Garnett SP. Bioelectrical impedance analysis to estimate body composition, and change in adiposity, in overweight and obese adolescents: comparison with dual-energy x-ray absorptiometry.* BMC Pediatr. 14:249. doi: 10.1186/1471-2431-14-249.

[44] Heymsfield, S.B., Wang ZM. (1993). *Measurement of total-body fat by underwater weighing: new insights and uses for old method. Nutrition.* ;9(5):4723.

[45] Borga, M., West, J., Bell, J.D., Harvey N.C., Romu, T., Heymsfield, S.B., et al. (2018). *Advanced body composition assessment: from body mass index to body composition profiling. J Investig Med.* 66(5):19.doi: 10.1136/jim-2018-000722.

[46] Goran, M.I., Driscoll, P., Johnson, R., Nagy, T.R., Hunter, G. (1996). *Cross-calibration of body-composition techniques against dual-energy X-ray absorptiometry in young children. Am J Clin Nutr.* 63(3):299305. doi: 10.1093/ajcn/63.3.299.

[47] Silva, D.R.P., Ribeiro, A.S., Pavão, F.H., Ronque, E.R.V., Avelar, A., Silva, A.M., et al. (2013). *Validity of the methods to assess body fat in children and adolescents using multicompartment models as the reference method: a systematic review.* Rev Assoc Med Bras. doi: 10.1016/j.ramb.2013.03.006.

[48] Jensen, N.S.O, Camargo, T.F.B. (2016). *BerGAMaschi DP. Comparison of methods to measure body fat in 7-to-10-year-old children: a systematic review.* Public Health. 133:313. doi: 10.1016/j.puhe.2015.11.025.

[49] Sociedad Española para el Estudio de la Obesidad (SEEDO). (2000).*Consenso SEEDO'2000 para la evaluación del sobrepeso y la obesidad y el establecimiento de criterios de intervención terapéutica.* medicina y Clínica.

[50] Bray, G., Bouchard, C., James, W.P.T. (1998) *Definitions and proposed current classifications of obesity.* En: Bray, G., Bouchard, C., James, W.P.T., editores. Handbook of obesity. Nueva York: Marcek Dekker

[51] Robin, A., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, JC. and Müller, M. (2011). *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 12, p. 77. DOI: doi: 10.1186/147121051277.

[52] Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B. (2022). *Application Framework for R_. R package version 1.7.3.*Disponible en https://CRAN.R-project.org/package=shiny

[53] Curilem Gatica, C., Almagià Flores, A., Rodríguez Rodríguez, F., Yuing Farias, T., Berral de la Rosa, F., Martínez Salazar, C. et al. (2016). *Evaluación de la composición corporal en niños y adolescentes: directrices y recomendaciones.* Nutr Hosp. 33:734-738. DOI: http://dx.doi.org/10.20960/nh.285.

[54] Goran, M.I., Driscoll, P., Johnson, R., Nagy, T.R., Hunter, G. (1996). *Cross-calibration of body-composition techniques against dual-energy X-ray absorptiometry in young children.* Am J Clin Nutr;63(3):299305. doi: 10.1093/ajcn/63.3.299.

[55] Wong, W.W., Stuff, J.E., Butte, N.F., Smith, E.O., Ellis, K.J. (2000). *Estimating body fat in African American and white adolescent girls: a comparison of skinfold-thickness equations with a 4-compartment criterion model.* Am J Clin Nutr. 72(2):348-54. doi: 10.1093/ajcn/72.2.348.

[56] Wohlfahrt Veje, C., Tinggaard, J., Winther, K., Mouritsen, A., Hagen, C.P., Mieritz, M.G. et al. (2014). *Body fat throughout childhood in 2647 healthy Danish children: agreement of BMI, waist perimeter, skinfolds with dual X-ray absorptiometry. Eur J Clin Nutr.* 68(6):664-70. doi: 10.1038/ejcn.2013.282.

[57] Sánchez Jaeger, A., Barón María, A. (2009). *Uso de la bioimpedancia eléctrica para la estimación de la composición corporal en niños y adolescentes.* 22(2): 105-110.

[58] Pombo Arias, M. et al. (1997). *Tratado de endocrinología pediátrica, 2ª edición.* Ed. Diaz de Santos.