Nure Inv. 21 (131)
https://doi.org/10.58722/nure.v21i131.2553
nure investigación

Reflexión Metodológica **Open Access**

Regresión logística binaria para clínicos poco amantes de las matemáticas

Binary logistical regression for clinicians with little loves of mathematics

Autores: *Javier Piury-Pinzón (1), Lucía Cayuela-Rodríguez (2), Aurelio Cayuela-Domínguez (3), Manuel Ortega-Calvo* (1).*

* ***Dirección de contacto:*** *106mayorque104@gmail.com*

Médico de Familia. Centro de Salud "Las Palmeritas". Distrito Sanitario de Atención Primaria (Sevilla, España).

Resumen

La conceptualización de los modelos matemáticos es una tarea difícil para los profesionales sanitarios que se dedican a tareas clínicas. Uno de los más utilizados es la regresión logística binaria (RLB). Nuestro objetivo en este artículo es tratar de acercar y de hacer comprender a los clínicos, la historia, el substrato matemático y la utilidad predictiva de la RLB. La función logística no es más que un modelo lineal, en el que a diferencia de la ecuación de la recta, la variable resultado es el logaritmo natural de una *odds ratio* observada, y las variables predictoras poseen unos coeficientes "β" que también son los logaritmos naturales de las *odds ratio* observadas en el trabajo científico que estemos llevando a cabo. La función logística se adapta muy bien a fenómenos clínico-biológicos, por ejemplo, la aparición de la epidemia de SIDA en los Estados Unidos de América o la curva de disociación de la hemoglobina. Cuando el ordenador nos ofrece los valores de los coeficientes para nuestras variables predictoras, nos está dando una función de tipo logístico que nos permite predecir si las observaciones futuras van a poder ser clasificadas en una u otra categoría de la variable resultado binaria. Cada modelo de RLB tiene su equivalente en un área bajo la Curva ROC determinada (Teoría de las Pruebas diagnósticas). Cuando generamos un diagnóstico médico o enfermero, realmente estamos aplicando un modelo de RLB. En Framingham se utilizaron por primera vez dentro de la investigación epidemiológica.

Palabras clave

Modelos lineales; Modelos Logísticos; Función de verosimilitud; Historia de la Medicina.

Abstract

The conceptualization of mathematical models is a hard duty for healthcare professionals engaged in clinical tasks. One of the most used is binary logistic regression (BLR) . Our objective in this paper is to try to bring closer and make clinicians understand the history, the mathematical groundwork and the predictive usefulness of BLR. The logistic function is nothing more than a linear model, in which, unlike the equation of the line, the result variable is the natural logarithm of an observed odds ratio, and the predictor variables have coefficients "β" that are also the natural logarithms of the odds ratios observed in the scientific work that we are carrying out. The logistic function adapts very well to clinical-biological phenomena, for example, the emergence of the AIDS epidemic in the United States of America or the hemoglobin dissociation curve. When the computer offers us the values of the coefficients for our predictors, it is giving us a logistic-type function that allows us to predict whether future observations will be able to be classified into one or another category of the binary outcome variable. Each BLR model has its equivalent in an area under the determined COR Curve (Diagnostic Tests Theory). When we generate a medical or a nursing diagnosis, we are really applying an BLR model. In Framingham they were used for the first time in epidemiological research.

Keywords

Linear Model; Logistic Models; Likelihood Function; History of Medicine.

INTRODUCCIÓN

La prestación de atención clínica implica una dedicación intensiva de tiempo y energía por parte del personal sanitario, tanto médicos como enfermeros, cuyas responsabilidades clínicas primarias a menudo ocupan la mayor parte de su jornada laboral. En este contexto, la comprensión y aplicación de métodos de análisis estadístico, especialmente aquellos utilizados en la investigación biosanitaria, puede resultar desafiante debido a las demandas de tiempo y a la complejidad inherente de estos métodos.

Entre los métodos estadísticos ampliamente empleados en la investigación sanitaria destaca la regresión logística binaria (RLB). Esta técnica se utiliza para modelar la relación entre una variable de resultado binaria y una serie de variables predictoras, tanto categóricas como continuas. A través de la aplicación de la RLB, los investigadores pueden analizar y predecir eventos binarios, como la presencia o ausencia de una enfermedad, en función de múltiples factores clínicos y biológicos.

El impacto y la relevancia de la RLB en la investigación clínica se reflejan en la cantidad significativa de literatura científica dedicada a este método. Además, la RLB se ha integrado con éxito en el ámbito de la inteligencia artificial (IA), lo que ha amplificado considerablemente su utilidad e impacto en la investigación y la práctica clínica actual (1, 2).

En resumen, este artículo tiene como objetivo proporcionar una comprensión más profunda de la historia, los fundamentos matemáticos y la utilidad práctica de la RLB en el contexto sanitario. Esperamos que esta información sea de utilidad para los clínicos (3).

ECUACIÓN DE LA RECTA

Vamos a comenzar por lo básico y por algo conocido por todos, la ecuación de la recta.

$$Y = a + b\,X$$

En donde "Y" es la variable dependiente, "X" es la variable predictora y "a" es la ordenada en el origen esto es, el valor que toma la variable dependiente (Y) cuando la variable predictora (X) toma el valor 0. El parámetro "b" es la pendiente de la recta de regresión (Figura 1). Como pequeña aclaración, variable predictora y variable independiente son términos sinónimos. X e Y son las coordenadas de un punto en el plano Euclidiano y finalmente, variable resultado y variable dependiente también son términos sinónimos (Tabla 1).

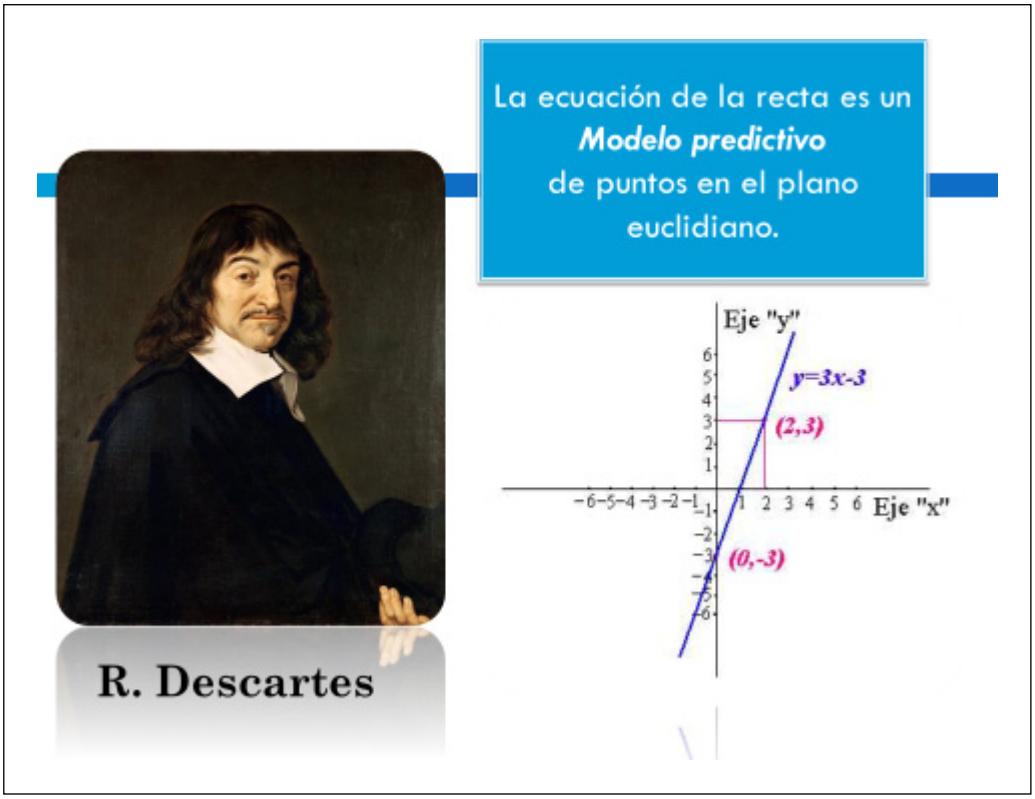

Figura 1. La ecuación de la recta como modelo matemático que predice coordenadas de puntos en el plano Euclidiano.

Hasta ahora todo es conocido incluso para sanitarios con una formación matemática elemental. Pero, reflexionemos un poco a partir de este concepto clave.

La ecuación de la recta no es más que un modelo matemático predictivo de las coordenadas de puntos en el plano Euclidiano (4). Nos sirve para predecir de forma exacta donde se ubican los infinitos puntos de una recta en el plano bidimensional (5). (Por ejemplo, en la recta de la Figura 1 sabemos perfectamente que habrá puntos en las coordenadas siguientes: 0 y -3 ; +1 y 0 ; +2 y +3 …).

La ecuación de la recta no solo es una herramienta fundamental para comprender y predecir relaciones lineales en el plano bidimensional, sino que también constituye una base importante para el análisis de datos y la toma de decisiones en diversas áreas del conocimiento. Sin embargo, es importante tener en cuenta que la ecuación de la recta es un modelo lineal puro, por lo que solo se ajusta a relaciones lineales entre las variables. En la mayoría de los casos, las relaciones biológicas y clínicas no son completamente lineales.

Imaginemos que deseamos predecir la probabilidad de que una persona desarrolle una enfermedad. La ecuación de la recta no sirve, ya que solo funciona para relaciones lineales estrictas. Pero la Regresión Logística Binaria (RLB) sí puede ayudarnos.

REGRESIÓN LOGÍSTICA BINARIA

En la parte superior de la Figura 2 está representada la expresión matemática de la denominada función logística que no es más que un tipo de modelo lineal (6), en el que a diferencia de la ecuación de la recta , la variable resultado es el logaritmo natural de una *odds ratio* (7) observada , y las variables predictoras poseen unos coeficientes " β " que también son logaritmos naturales de las *odds ratio* observadas en el ejercicio clínico científico que estemos desarrollando para nuestro trabajo fin de grado , fin de residencia o fin de Máster (Tabla 1).

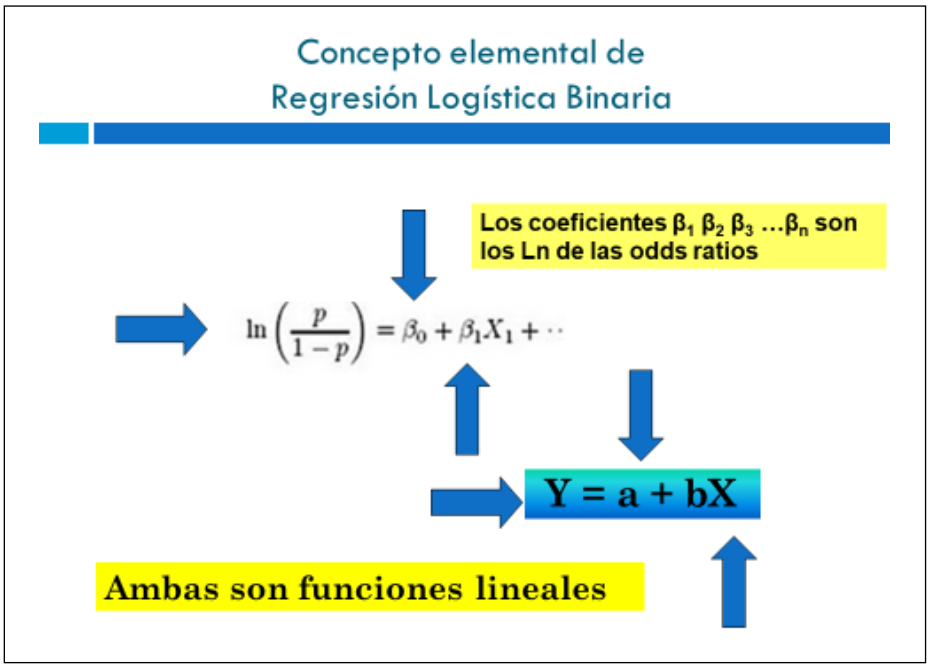

Figura 2. La función logística y la ecuación de la recta. Ln: Logaritmo natural.

Terminología básica en el modelo de RLB	
Variable Resultado o dependiente	Es la variable resultante del modelo. "Y " en la ecuación de la recta.
Variable independiente	Son las variables que predicen ("predictoras") a la variable resultado. "X" en la ecuación de la recta.
Odds Ratio	Logaritmo natural de los coeficientes numéricos de las variables predictoras.
Univariante	Modelo que contiene una sola predictora.

Tabla 1. Conceptos clave en el Modelo de Regresión Logística Binaria (RLB).

Se preguntará el lector porqué hemos elegido la función logística y no otra para complicarnos la vida. Pretendemos que la respuesta esté en las Figuras 3 y 4.

La función logística, mejor dicho, su expresión gráfica con una sola predictora ("*monovariante*") (Tabla 1), se adapta muy bien a fenómenos biológicos, por ejemplo, la aparición de la epidemia de SIDA en los Estados Unidos de América o la curva de disociación de la hemoglobina (Figura 4).

Podemos aceptar por lo tanto la hipótesis de que la RLB es un "buen" modelo matemático de tipo lineal para predecir fenómenos clínicos y biológicos. Se puede aplicar también en Economía y en Psicología.

Por la naturaleza matemática de este modelo distinto al de la ecuación de la recta, la variable dependiente o resultado es el logaritmo natural de una *odds ratio* medida en una variable categórica binaria (razón de ventaja de sufrir una neoplasia de colon o no sufrirla, de ser obeso o no serlo …). Es lo que se denomina análisis categórico en bioestadística y epidemiologia.

En resumen, la RLB ha tenido una evolución histórica notable, desde sus antecedentes en la teoría de la probabilidad hasta su aplicación en diversas áreas del conocimiento. Su desarrollo ha sido impulsado por la necesidad de analizar datos binarios y estimar la probabilidad de un evento.

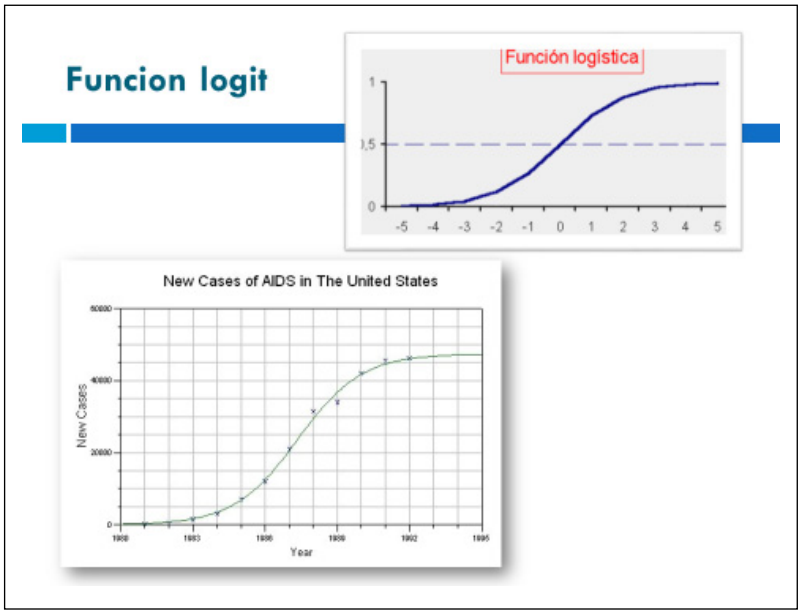

Figura 3. Semejanza entre la función logística monovariante y la gráfica de aparición de la epidemia de SIDA en los EEUU.

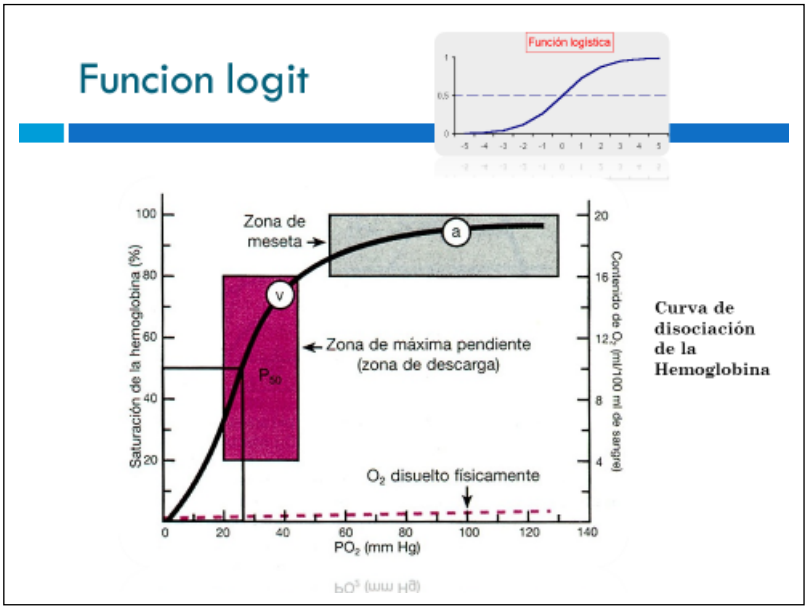

Figura 4. Semejanza entre la función logística monovariante y la gráfica de disociación de la hemoglobina.

APLICABILIDAD

Cuando el ordenador nos ofrece los valores de los coeficientes para nuestras variables predictoras, nos está dando una función lineal de tipo logístico que nos permita predecir si las observaciones futuras van a poder ser clasificadas en una u otra categoría de la variable resultado binaria.

VEROSIMILITUD Y DIAGNÓSTICO CLÍNICO

La probabilidad o verosimilitud *("likelihood")* (8) de que la predicción sea acertada dependerá de circunstancias diversas –tamaño de la muestra analizada (9, 10), errores alfa y beta utilizados, grado de colinealidad de las predictoras (11, 12)…–.

Una de las ventajas visuales de la RLB es que se le puede emplear en la metodología de las curvas ROC (13). Cada modelo de RLB tiene su equivalente en un área bajo la Curva determinada (Teoría de las Pruebas diagnósticas). Hace algunos años, algunos de nosotros, pudimos desarrollar un modelo predictivo de neoplasia maligna de colon con RLB. El área bajo la curva ROC está en la Figura 5 (14).

El Profesor Silva (15) ha comentado alguna vez que cuando generamos un diagnóstico médico o enfermero, realmente estamos aplicando un modelo de RLB (Figura 6) (16).

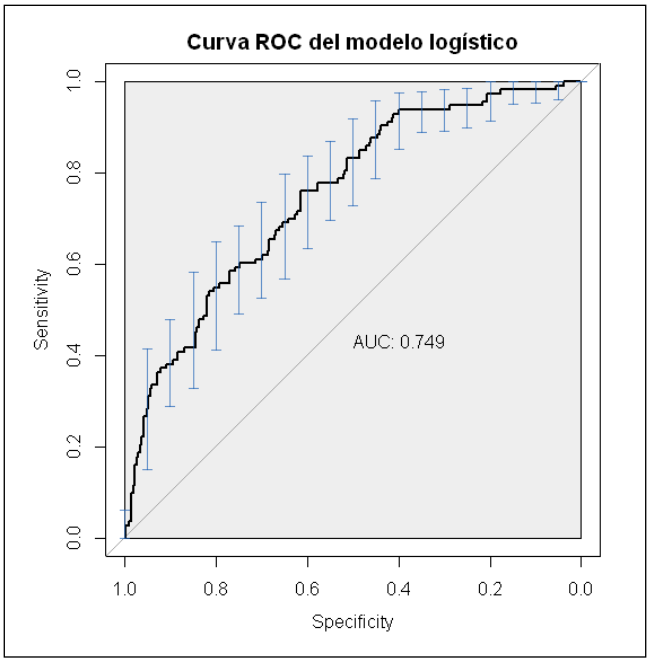

Figura 5. Área bajo la Curva ROC de un modelo de RLB (14).

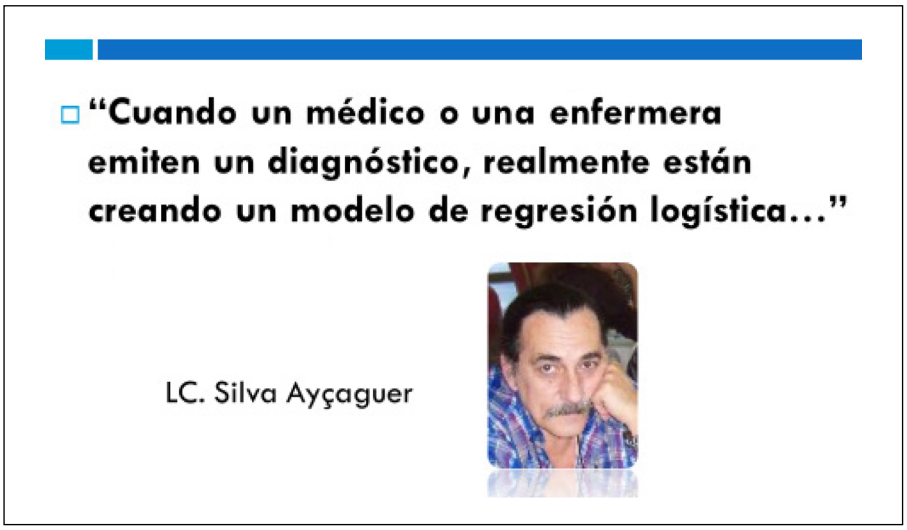

Figura 6. Aforismo de metodología investigadora.

Los primeros investigadores de Framingham han influido decisivamente no solo en la epidemiología cardiovascular sino también en el desarrollo del pensamiento clínico entre los siglos XX y XXI (17). Paul Dudley White (Figura 7), Jerome Cornfield y William B. Kannel (Figura 8) fueron capitales en ese sentido (18). Como es bien sabido, la muerte prematura del presidente Franklin D. Roosevelt por enfermedad cardiovascular impulsó el desarrollo y la implementación del estudio Framingham.

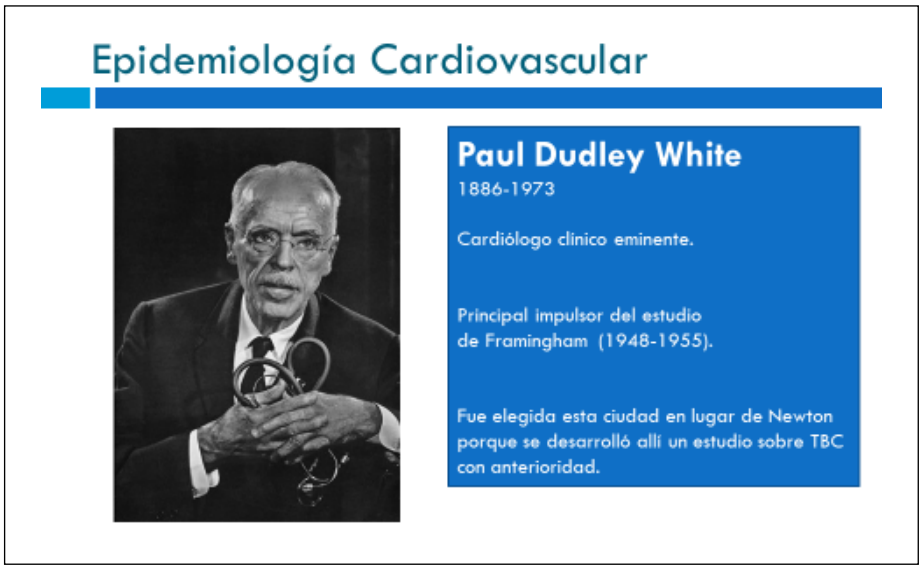

Figura 7. Paul Dudley White impulsor del estudio de Framingham.

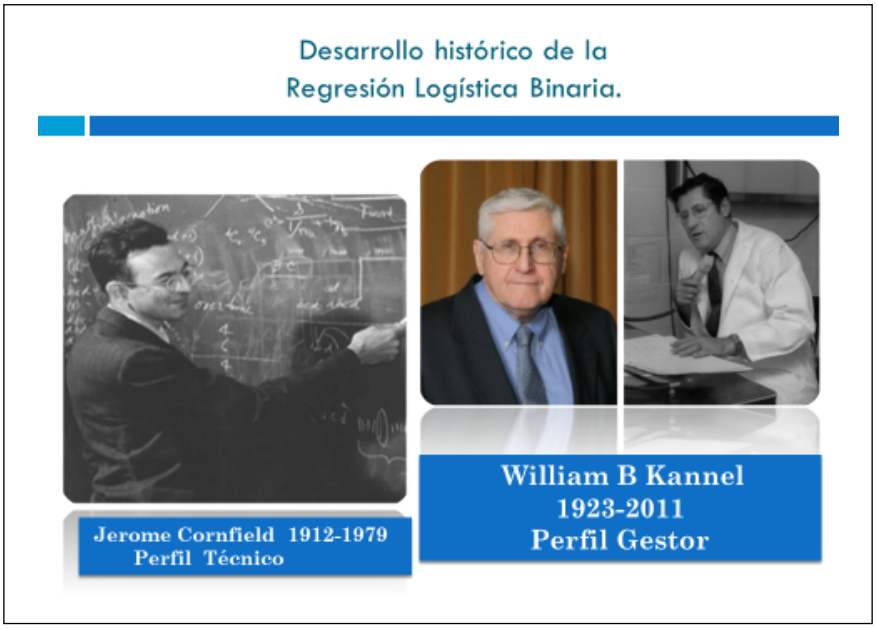

Figura 8. Jerome Cornfield y William B Kannel.

USOS, UTILIDADES E HISTORIA

Por ejemplo, la RLB se ha utilizado con éxito para desarrollar herramientas de diagnóstico y pronóstico personalizadas, identificar biomarcadores de enfermedades y desarrollar sistemas de IA para la toma de decisiones clínicas. Los modelos de regresión logística también se manejan para calcular puntuaciones de propensión que luego pueden utilizarse para equilibrar los grupos en los estudios observacionales cuando el objetivo es comparar un resultado en dos cohortes (19, 20).

Los hitos en el desarrollo de la regresión logística binaria engloban varios acontecimientos relevantes. Joseph Berkson y David Cox introdujeron la teoría de la regresión logística en la década de 1950, como una extensión de la regresión lineal, con el propósito de modelar variables categóricas (21). Durante la década de 1970, se produjo un avance significativo en los métodos de estimación de parámetros para la regresión logística binaria, incluyendo el método de máxima verosimilitud. Además, el Estudio de Framingham representó un hito crucial en el desarrollo y aplicación de la RLB. Este estudio, reconocido por su contribución esencial a la comprensión de los factores de riesgo, no solo suministró datos valiosos para la elaboración de tablas de riesgo cardiovascular, sino que también marcó el comienzo del empleo generalizado de métodos de RLB en la investigación clínica (22, 23).

CONCLUSIÓN

En este artículo introductorio, hemos obviado las condiciones previas para aplicar la RLB, así como su capacidad para triangular análisis de fenómenos observados en la práctica sanitaria (24) o su aplicación en métodos de decisión más complejos (25). Nuestro objetivo se centró en que el lector comprendiera las bases matemáticas del método (26) y su evolución histórica. La RLB no es más que un modelo matemático de tipo lineal que sirve fundamentalmente para predecir hechos binarios.

DATOS AUTORES

(1) Médico de Familia. Centro de Salud "Las Palmeritas". Distrito Sanitario de Atención Primaria (Sevilla, España); (2) Médico Internista. Servicio de Medicina Interna. Hospital Universitario Severo Ochoa (Madrid, España); (3) Epidemiólogo. Área de Gestión Sanitaria Sur de Sevilla. Hospital Universitario Virgen de Valme (Sevilla, España).

BIBLIOGRAFÍA

1. Grigore M, Popovici RM, Gafitanu D, Himiniuc L, Murarasu M, Micu R. Logistic models and artificial intelligence in the sonographic assessment of adnexal masses - a systematic review of the literature. Med Ultrason. 2020 ;22:469-475. *https://medultrason.ro/medultrason/index.php/medultrason/article/view/2538/1680*
2. Issaiy M, Zarei D, Saghazadeh A. Artificial Intelligence and Acute Appendicitis: A Systematic Review of Diagnostic and Prognostic Models. World J Emerg Surg. 2023; 18(1):59. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10729387/*
3. Ortega Calvo M. La pedagogía clínica como una nueva rama del humanismo médico: aspectos de la no directividad Med Clin (Barc). 1993;100:107-9. *https://pubmed.ncbi.nlm.nih.gov/8426491/*
4. Jones ML. Descartes's Geometry as Spiritual Exercise. Critical Inquiry. 2001;28:40-71. Descartes's Geometry as Spiritual Exercise on JSTOR
5. Hernández VM. La geometría analítica de Descartes y Fermat: ¿Y Apolonio? *Apuntes de historia de las matemáticas, 2002*;1:32-45. *http://euler.mat.uson.mx/depto/publicaciones/apuntes/pdf/1-1-4-analitica.pdf*
6. Sánchez-Cantalejo Ramírez E. Regresión Logística en Salud Pública. Escuela Andaluza de Salud Pública. Granada. 2000. p.1- 173.
7. Martínez-González MA, de Irala-Estevez J, Guillén-Grima F. ¿Qué es una odds ratio? Med Clin (Barc). 1999;112:416-22.
8. Zurakowski D, Johnson VM, Lee EY : Biostatistics in clinical decision making for cardiothoracic radiologists. J Thorac Imaging. 2013;28:368-75. *https://journals.lww.com/thoracicimaging/abstract/2013/11000/biostatistics_in_clinical_decision_making_for.6.aspx*
9. Ortega Calvo M, Cayuela Domínguez A. Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. Rev Esp Salud Publica. 2002;76:85-93. *https://www.sanidad.gob.es/biblioPublic/publicaciones/recursos_propios/resp/revista_cdrom/vol76/vol76_2/RS762C_85.pdf*
10. González CG, Peña Rodríguez A, Salas Díaz I et al. Una concepción topológica del "bootstrap" permite la demostración del sesgo de Berkson en epidemiología nutricional. Nutrición Clínica: Dietética Hospitalaria. 2016; 36:134–142. *https://www.revistanutricion.org/articles/a-topological-conception-of-bootstrap-proofs-berkson-bias-in-nutritional-epidemiology.pdf*
11. Tu YK, Clerehugh V, Gilthorpe MS. Collinearity in linear regression is a serious problem in oral health research. Eur J Oral Sci. 2004;112:389-97. *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0722.2004.00160.x?sid=nlm%3Apubmed*
12. Nathanson BH, Higgins TL. An introduction to statistical methods used in binary outcome modeling. Semin Cardiothorac Vasc Anesth. 2008;12:153-66. *https://journals.sagepub.com/doi/10.1177/1089253208323415?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed*
13. Martínez Pérez JA, Pérez Martin PS. La curva ROC. Semergen. 2023;49:101821. *https://www.elsevier.es/es-revista-medicina-familia-semergen-40-articulo-la-curva-roc-S1138359322001952*

14. Villadiego-Sánchez JM, Ortega-Calvo M, Pino-Mejías R, Cayuela A, Iglesias-Bonilla P, García-de la Corte F, et al. Multivariate explanatory model for sporadic carcinoma of the colon in Dukes' stages I and IIa. Int J Med Sci. 2009;6:43-50. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2640476/*
15. Silva LC, Barroso IM. Regresión logística. Ed. La Muralla / Hespérides. Madrid. 2004.
16. González-García L, Chemello C, García-Sánchez F, Serpa-Anaya DC, Gómez-González C, et al. Aphorisms and short phrases as pieces of knowledge in the pedagogical framework of the andalusian school of public health. Int J Prev Med. 2012 ;3:197-210. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3309634/*
17. White PD. The tardy growth of preventive cardiology. Am J Cardiol. 1972 ; 29:886-8. *https://www.ajconline.org/article/0002-9149(72)90513-9/abstract*
18. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet. 2014;383(9921):999-1008. *https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(13)61752-3/abstract*
19. Kim DH, Pieper CF, Ahmed A, Colón-Emeric CS. Use and Interpretation of Propensity Scores in Aging Research: A Guide for Clinical Researchers. J Am Geriatr Soc. 2016;64 :2065-2073. *https://agsjournals.onlinelibrary.wiley.com/doi/10.1111/jgs.14253*
20. Raghunathan K, Layton JB, Ohnuma T, Shaw AD. Observational Research Using Propensity Scores. Adv Chronic Kidney Dis. 2016; 23:367-372. *https://www.akdh.org/article/S1548-5595(16)30138-0/fulltext*
21. Sagaró del Campo NM, Zamora Matamoros L. Evolución histórica de las técnicas estadísticas y las metodologías para el estudio de la causalidad en ciencias médicas. MEDISAN, 2019;23:534-556. *http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1029-30192019000300534&lng=es&tlng=es.*
22. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis. 1967;20:511-24. *https://www.sciencedirect.com/science/article/abs/pii/0021968167900823?via%3Dihub*
23. Walker SH, Duncan DB: Estimation of the probability of an event as a function of several independent variables. Biometrika. 1967;54:167-79.
24. González-García L, Gómez-González C, Chemello C, Cubiles-De La Vega M, Santos-Lozano J, Ortega-Calvo M. Triangulación de un estudio cualitativo mediante regresión logística. Index de Enfermería 2014;23(1-2):80-4. *https://scielo.isciii.es/scielo.php?pid=S1132-12962014000100017&script=sci_arttext&tlng=pt*
25. Trujillano J, Sarria-Santamera A, Esquerda A, Badia M, Palma M, March J: Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. Gac Sanit. 2008 Jan-Feb;22(1):65-72. *https://www.gacetasanitaria.org/es-aproximacion-metodologia-basada-arboles-decision-articulo-S0213911108712044?ref=busqueda&ant=S021391110876076X*
26. Krauss A. Homo methodologicus and the origin of science and civilisation. Heliyon. 2023;9(10):e20237. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10570580/*