EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Fundamentos de medicina basada en la evidencia

Regresión logística múltiple

Carlos Ochoa Sangrador[1], Manuel Molina Arias[2], Eduardo Ortega Páez[3]
[1]Servicio de Pediatría. Hospital Virgen de la Concha. Complejo Asistencial de Zamora. Zamora. España.
[2]Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.
[3]Pediatra. Unidad de Gestión Clínica Góngora. Distrito Granada-Metropolitano. Granada España.

Correspondencia: Manuel Molina Arias: mma1961@gmail.com

Palabras clave en español: análisis multivariante; estadística; inferencia estadística.
Palabras clave en inglés: multivariate analysis; statistics; statistical inference.

Fecha de recepción: 11 de septiembre de 2023 • **Fecha de aceptación:** 21 de septiembre de 2023
Fecha de publicación del artículo: 27 de septiembre de 2023

Evid Pediatr. 2023;19:34.

CÓMO CITAR ESTE ARTÍCULO

Ochoa Sangrador C, Molina Arias M, Ortega Páez E. Regresión logística múltiple. Evid Pediatr. 2023;19:34.

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2023;19:34.
 • ISSN: 1885-7388

Regresión logística múltiple

Carlos Ochoa Sangrador[1], Manuel Molina Arias[2], Eduardo Ortega Páez[3]
[1]*Servicio de Pediatría. Hospital Virgen de la Concha. Complejo Asistencial de Zamora. Zamora. España.*
[2]*Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.*
[3]*Pediatra. Unidad de Gestión Clínica Góngora. Distrito Granada-Metropolitano. Granada España.*

Correspondencia: Manuel Molina Arias: mma1961@gmail.com

INTRODUCCIÓN

En un documento previo sobre análisis multivariante abordamos los fundamentos del análisis multivariante, los tipos de análisis y el proceso de modelización multivariante. Asimismo, mostramos este proceso para la regresión lineal múltiple. En este documento repasaremos el proceso para la regresión logística múltiple.

REGRESIÓN LOGÍSTICA MÚLTIPLE

La regresión logística es una técnica de análisis que utilizamos cuando se desea conocer la relación entre una variable dependiente cualitativa dicotómica y una o más variables explicativas independientes, ya sean cualitativas o cuantitativas. Existen variantes de esta técnica para cuando la variable dependiente es nominal politómica, esto es, tiene más de dos categorías (regresión logística multinomial o politómica), y para cuando es ordinal (regresión logística ordinal).

El objetivo de la regresión logística binaria no es, como en la regresión lineal, predecir el valor de la variable dependiente (Y) a partir de una o varias variables independientes (X), sino predecir la probabilidad de que ocurra el evento que caracteriza la variable dependiente (éxito, enfermedad, etc.), conocidos los valores de las variables independientes.

La variable dependiente va a adoptar los valores 1 y 0 que representan, respectivamente, la presencia y ausencia del evento que clasifica dicha variable. Para explorar la contribución de las variables independientes a la estimación de la probabilidad de la variable dependiente, tendríamos que construir un modelo similar al empleado en regresión lineal:

$$a + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

Sin embargo, el resultado de esta ecuación no va a dar estimaciones de probabilidad (p) entre 0 y 1, porque para determinados valores de las variables independientes sus resultados podrán estar entre menos y más infinito. Para resolver este problema la regresión logística hace dos transformaciones. Primero, transforma cada probabilidad para cada combinación de valores de las variables independientes en sus *odds* (p/[1-p]), que adoptan valores entre cero e infinito. En segundo lugar, calculan los logaritmos neperianos de las *odds*, ln(p/[1-p]), que adoptan valores entre menos y más infinito y que además siguen una distribución lineal.

En la **Figura 1** se presentan las distribuciones de las probabilidades de enfermar en función de la edad, que siguen una distribución sigmoidea, y a su derecha las transformaciones en logaritmos naturales de las *odds* de dichas probabilidades, que siguen una distribución lineal.

Con este nuevo parámetro ya podemos emplear una ecuación lineal. La fórmula de la regresión logística quedaría:

$$Ln(p/[1-p]) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

De la que podemos calcular directamente las probabilidades con la fórmula:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n)}}$$

En la que P(Y) es la probabilidad de que Y adopte el valor 1 (éxito, enfermedad, etc.), b_o es la constante del modelo, b_i los coeficientes de cada variable independiente y X_i sus valores.

En la **Figura 2** se muestran funciones logísticas con sus correspondientes estimaciones de coeficientes. El punto de inflexión de la función, que corresponde habitualmente al valor de la variable con una probabilidad del 50% (0,50), determina la constante (a) y la pendiente de la tangente en ese punto el coeficiente de la variable (b). En la parte inferior de la figura se presentan una serie de funciones con los coeficientes estimados.

Las covariables cualitativas deben ser dicotómicas, tomando valor 0 para su ausencia y 1 para su presencia. Si la covariable tuviera más de dos categorías debemos realizar una transformación de la misma en varias covariables cualitativas dicotómicas ficticias (variables indicadoras o *dummies*). Luego veremos un ejemplo. Al hacer esta transformación, cada categoría de la variable entraría en el modelo de forma individual, aunque su modelización y análisis debe ser conjunto.

Figura 1. Distribución de las probabilidades de enfermar en función de la edad y los logaritmos naturales de las *odds* de dichas probabilidades

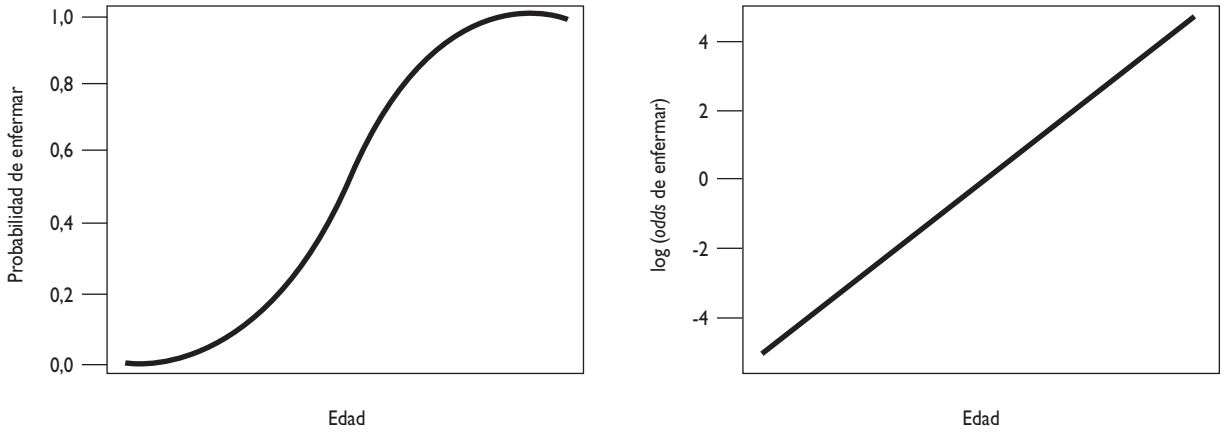
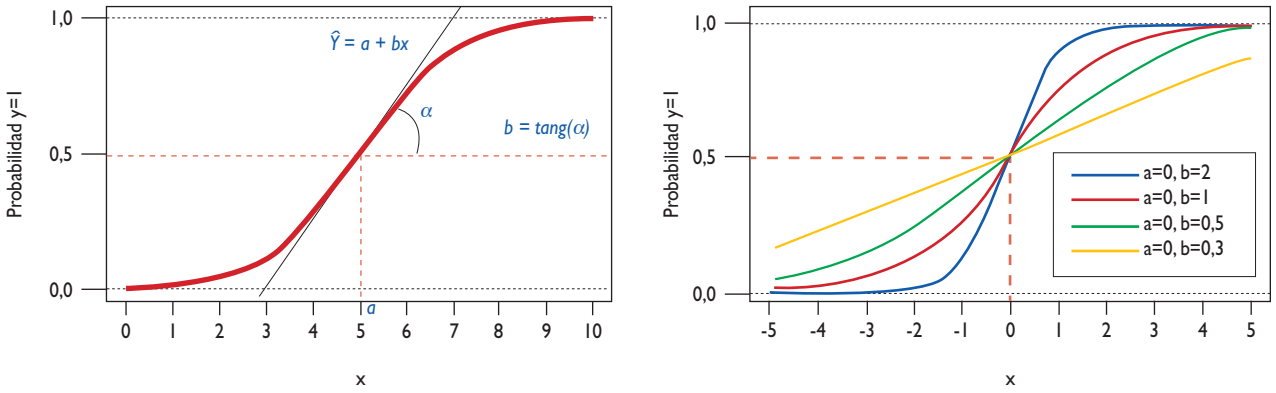


Figura 2. Funciones logísticas



Aunque las covariables cuantitativas pueden ser analizadas en su escala natural, no siempre el riesgo asociado a cada unidad de la variable alcanza la magnitud suficiente para mostrar significación clínica o estadística. Por ello, no es infrecuente tener que recodificar estas variables para que el modelo estime el riesgo asociado a determinados rangos o magnitudes de la variable.

La interpretación de los coeficientes en la regresión logística es distinta que la de los coeficientes de la regresión lineal; aquí se utilizan para calcular las *odds ratio* (OR) ajustadas de cada variable, que se obtienen exponenciando el coeficiente (e^b); el cálculo es proporcionado habitualmente por los paquetes estadísticos. Esta OR equivale a la OR combinada que obtendríamos en un análisis estratificado.

Para ilustrar el procedimiento de la regresión logística vamos a realizar un análisis con los datos de una serie de casos de gastroenteritis aguda. Queremos construir un modelo para predecir la gastroenteritis bacteriana. En la **Figura 3** se presentan las frecuencias de gastroenteritis bacteriana en función de algunas variables seleccionadas: presencia de sangre en heces, edad menor de 2 años y estación del año.

Podemos observar que en presencia de sangre en heces es mucho más frecuente la gastroenteritis bacteriana, que aparentemente no hay diferencias por edades, aunque el riesgo asociado a sangre en heces parece tener lugar principalmente en los menores de 2 años. Asimismo, observamos que en verano es mucho más frecuente la presencia de gastroenteritis bacteriana. Como vemos, hay varias variables implicadas, por lo que parece necesario recurrir a análisis multivariante.

Los pasos de la modelización son:

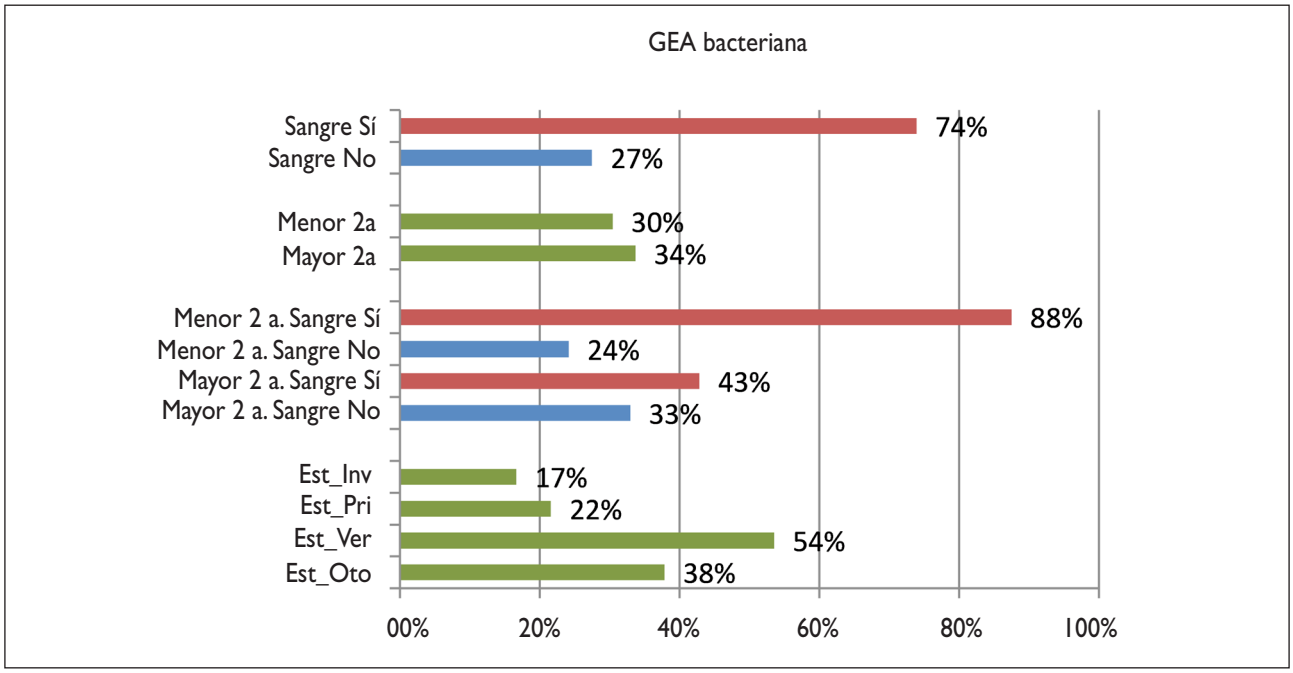
1. **Modelo máximo**

   Incluimos las variables descritas en la **Figura 3**, además de un término de interacción entre edad menor de 2 años y sangre en heces.

2. **Codificación de variables y elección de los valores de referencia**

   Todas las variables se han recodificado a valores 0-1 y hemos creado tres variables indicadoras (*dummies*) para analizar la estación del año, siguiendo el esquema de la **Tabla 1**.

Figura 3. Frecuencia de gastroenteritis bacteriana en función de la presencia de sangre en heces, edad menor de 2 años y estación del año



Con este esquema cada una de las variables indicadoras, de valor 0 o 1, expresarán la diferencia de riesgo entre cada estación y la estación de referencia, en este caso invierno. En función del programa estadístico que empleemos, las variables indicadoras las tendremos que crear nosotros o las creará automáticamente el programa, pero siempre debemos controlar este proceso y conocer el esquema de recodificación, del que existen diversas alternativas, para poder interpretar los resultados. Asimismo, hemos creado un término de interacción para modelizar la diferente capacidad predictiva de la sangre en heces en función de la edad menor/mayor de 2 años (Edad × Sangre), multiplicando los valores de ambas variables.

3. Estrategia de modelización

Elegimos una estrategia "hacia atrás", considerando el valor Z asociado a cada variable (contraste de Wald), estimado a partir del cociente de cada coeficiente y su error estándar. Hay que recordar que las variables indicadoras entran y salen del modelo en bloque, aunque individualmente algunas de ellas no resulten significativas. Asimismo, el término de interacción entra con las variables relacionadas, pero puede eliminarse, dejando las relacionadas, si no resulta estadísticamente significativo y su eliminación no empeora el modelo en conjunto.

En la **Figura 4** se presenta el análisis de regresión logística múltiple realizado con RCommander. En la salida de resultados se presentan los coeficientes de cada variable (*Estimate*), sus errores estándar (*Std. Error*), el contraste individual (*Z value*) y su significación (Pr(>|z|)), más abajo están los coeficientes exponenciados (*Exponentiated coefficients* "*odds ratios*") y sus intervalos de confianza. Los coeficientes de las variables indicadoras (Primavera, Verano y Otoño) y sus correspondientes OR informan de la probabilidad de gastroenteritis bacteriana en comparación con el riesgo de la estación de referencia implícita (Invierno). Podemos ver que ingresar en verano se asocia a un riesgo 5,21 veces mayor que en invierno (categoría de referencia), con una confianza del 95% de que aumenta entre 2,25 y 12,97 veces. Aunque la presencia de sangre en heces está muy asociada a la etiología bacteriana, en este análisis se ve que solo es así en los

Tabla 1. Recodificación de variables indicadoras (valores nuevos)

	Valores originales de la variable estación del año			
Variables indicadoras	1 (invierno)	2 (primavera)	3 (verano)	4 (otoño)
E_Primavera	0	1	0	0
E_Verano	0	0	1	0
E_Otoño	0	0	0	1

Figura 4. Regresión logística para predicción de gastroenteritis bacteriana

```
> GeaBact <- glm(gea_bac1_0 ~ hec_sang1_0 + Edad.2a + Edad2San + (Primav + Verano
+ Otoño),
+ family=binomial(logit), data=GeaPed2)
> summary(GeaBact)

Call:
glm(formula = gea_bac1_0 ~ hec_sang1_0 + Edad.2a + Edad2San +
 (Primav + Verano + Otoño), family = binomial(logit), data = GeaPed2)

Deviance Residuals:
 Min  1Q Median  3Q  Max
-2.0440 -0.7050 -0.6228 1.0836 1.9865

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.6683  0.4192 -3.980 0.00019
SangreHec   0.3249  0.8531 0.381 0.70333
Edad<2a    -0.1552  0.3195 -0.486 0.62725
EdadxSangre  2.4617  1.1716 2.101 0.03563
Primav     0.4030  0.4497 0.896 0.37019
Verano    1.6511  0.4432 3.726 0.00019
Otoño     0.9937  0.5139 1.934 0.05314.
---
Null deviance: 319.55 on 255 degrees of freedom
Residual deviance: 274.60 on 249 degrees of freedom
AIC: 288.6

> Confint(GeaBact, level=0.95, type="Wald", exponentiate=TRUE)

Exponentiated Coefficients and Confidence Bounds
    Estimate  2.5 %   97.5 %
(Intercept) 0.1885715 0.07839043 0.4106513
SangreHec 1.3838716 0.23648993 7.4770289
Edad<2a  0.8562749 0.45955103 1.6144831
Edad2San 11.7250573 1.28167439 142.0545209
Primav   1.4962438 0.63228828 3.7459178
Verano  5.2126068 2.25530941 12.9799053
Otoño   2.7013233 0.99141590 7.5647231
```

menores de dos años. Como se comentó en el documento previo de análisis estratificado, los análisis deben darse por separado. En menores de dos años, el riesgo de tener bacterias en heces de los que tienen sangre en heces es 16,11 veces mayor que el de los que no la tienen; esta OR se calcula exponenciando la suma de los coeficientes de las variables Sangre en heces y la interacción (Exp[0,32 + 2,46]).

4. Evaluación de la fiabilidad del modelo

Una vez desarrollado el modelo, el siguiente paso es evaluar su fiabilidad, estimando las medidas de bondad de ajuste, medidas que informan del ajuste de los datos predichos por el modelo a los datos observados reales. En la **Tabla 2** se presenta la salida correspondiente al modelo de la **Figura 4**; incluye la tabla de clasificación de valores predichos y observados, que muestra una concordancia

Tabla 2. Medidas de ajuste del modelo de regresión logística

GEA bacteriana	Valor predicho		
Valor observado	**No**	**Sí**	**Porcentaje correcto**
No	172	3	98,29
Sí	64	17	20,99
Acierto global			73,83
-2 log-verosimilitud 274,60 R cuadrada de Cox & Snell 0,16 R cuadrada de Nagelkerke 0,23			

del 73,83%, y las estimaciones de la desviación (-2 logaritmo de la razón de verosimilitudes, en inglés *deviance*), y de los coeficientes de determinación R^2 (porcentaje de varianza explicada por el modelo). Ya comentamos para la regresión lineal múltiple la interpretación de R^2, que corresponde al porcentaje de varianza explicada por el modelo, que interesa sea lo mayor posible. Sin embargo, para la desviación nos interesa el modelo que la tenga menor, porque será el que mejor ajusta a los datos. Otro estimador proporcionado en el análisis es el criterio de información de Akaike (AIC), que ajusta la log-verosimilitud por el número de variables; resulta muy útil cuando hay muchas variables independientes; el mejor modelo será el que tenga el AIC más bajo.

Aunque alguna de las variables presentan contrastes no significativos, no podemos prescindir de ellas, porque están vinculadas a otras variables indicadoras (primavera y otoño) o por sustentar el término de interacción (sangre en heces y edad menor de 2 años). Por ello, el modelo máximo elegido al inicio va a ser el modelo final.

Los modelos de regresión logística no tienen los requerimientos de los de regresión lineal. Así, no requieren que exista una relación lineal entre las variables dependientes y las independientes, ni tampoco necesitan que haya normalidad ni homocedasticidad de los residuos. Sin embargo, sí requieren que la variable dependiente sea binaria (nominal dicotómica), que las observaciones sean independientes entre sí (no exista apareamiento o medidas repetidas) y que no exista colinealidad entre las variables independientes. Otro requisito es que los valores de las variables independientes tengan una relación lineal con los logaritmos de sus *odds* (probabilidad dividida por su complementario). También se requiere suficiente número de observaciones, al menos 10 eventos del resultado menos frecuente de la variable dependiente, por cada variable introducida en el modelo.

BIBLIOGRAFÍA

1. Abraira V, Pérez de Vargas Luque A. Métodos multivariantes en bioestadística. Ed. Centro de Estudios Ramón Areces. Madrid 1996.
2. Arezina R, Duolao W. Source and control of bias. En: Duolao W, Bakhai A (eds.). Clinical trials. A practical guide to design, analysis and reporting. Londres: Remédica; 2006. p. 55-64.
3. Argimón Pallás JM, Jiménez Villa J. Métodos de investigación clínica y epidemiológica. Barcelona: Elsevier, 2006.
4. Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Correlación. Modelos de regresión. Evid Pediatr. 2021;17:25.
5. Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Regresión lineal simple. Evid Pediatr. 2021;17:46.
6. Ochoa Sangrador C, Molina Arias M, Ortega Páez E. Métodos de ajuste de sesgos. Análisis estratificado. Evid Pediatr. 2022;18:31.
7. Ortega Páez E, Ochoa Sangrador C, Molina Arias M. Regresión logística binaria simple. Evid Pediatr. 2022;18:11.
8. Rosner B. Fundamentals of Biostatistics, 7th Edition. Boston: Brooks/Cole, Cengage Learning 2011.
9. Rothman K. J. Epidemiología Moderna. Madrid: Díaz de Santos, 1987.
10. Woodwark M. Epidemiology study design and data analysis. London; Chapman & Hall/CRC, 1999.