Rev Argent Cir. 2025;117(2):e-cymcaa- http://dx.doi.org/10.25132/raac.v117.n1.cymcaa

Regresión Logística: lo que autores y revisores no pueden ignorar

Logistic regression analysis: what authors and reviewers should not ignore

*Cristian Agustín Angeramo**

"Los hechos no dejan de existir porque se los ignore."
Aldous Huxley, Un mundo feliz.

Introducción

Cuando se analizan datos, es habitual querer determinar si existe una relación entre dos variables. Para ello se utilizan herramientas estadísticas como la correlación y la regresión, que pueden parecer similares, pero responden a objetivos distintos.

La correlación cuantifica la fuerza y la dirección de la relación entre dos variables. Se trata de una medida simétrica, lo que implica que el orden de las variables no altera el resultado: la correlación entre A y B es idéntica a la de B con A. No obstante, la correlación no implica causalidad ni permite realizar predicciones[1]. Por otro lado, la regresión también analiza la relación entre variables, pero no de forma simétrica. Aquí, se distingue entre una variable dependiente, que se desea predecir o explicar, y una o más variables independientes, que actúan como predictores. Esta distinción permite modelar cómo las variables independientes influyen sobre la dependiente, posibilitando predicciones y la evaluación del peso relativo de cada predictor. Sin embargo, al igual que la correlación, la regresión tampoco implica causalidad[2].

¿Cuándo se usa la regresión logística?

La regresión logística se utiliza muy frecuentemente en investigación clínica cuando la variable dependiente (o variable de respuesta) es dicotómica, es decir, presenta solo dos posibles categorías (por ejemplo, "presencia" o "ausencia" de una enfermedad)[3]. Esa es la distribución logística binaria o binomial, pero hay otras que escapan al objetivo de este artículo (multinomial, ordinal)[4].

El objetivo de la regresión logística es analizar el efecto que tienen una o más variables predictoras (independientes) sobre la variable dependiente[3]. Esto permite: 1) identificar factores de riesgo o de protección: por ejemplo, determinar si una característica clínica (como la edad o el tabaquismo) aumenta o disminuye la posibilidad o chance de presentar una enfermedad, 2) estimar posibilidades: calcular la posibilidad o chance de que ocurra un evento. Por ejemplo, cuántas veces más es la chance que un paciente tiene de desarrollar una enfermedad en función de sus comorbilidades. Estas posibilidades pueden representarse visualmente en nomogramas o implementarse en calculadoras en línea para comunicar de manera clara y accesible el riesgo de un evento[5], 3) hacer predicciones: predecir la categoría más probable de la variable dependiente para un paciente determinado (por ejemplo, clasificar a un paciente como de alto o bajo riesgo de desarrollar una enfermedad)[3].

Se dice posibilidad o chance porque la regresión logística brinda Odds ratio y no Riesgos Relativos, que sí son cálculos de probabilidades.

Selección de variables independientes: cómo evitar el sobreajuste y el subajuste

Un modelo de regresión logística múltiple permite evaluar el efecto de dos o más variables independientes sobre un evento (variable dependiente). A diferencia del modelo con un solo predictor, este ajusta por factores de confusión y estima asociaciones más precisas. Uno de los aspectos más importantes al construirlo es la adecuada selección de variables independientes, tanto en número como en relevancia. Incluir demasiadas puede generar sobreajuste ("overfitting"), donde el modelo pierde capacidad para generalizar a nuevos datos. Por el contrario, excluir variables relevantes puede producir subajuste ("underfitting"), lo que deriva en una falta de precisión. Por ejemplo, en un estudio para predecir el riesgo de desarrollar diabetes, incluir variables irrelevantes (como el color de ojos) podría causar sobreajuste, mientras que omitir variables importantes (obesidad o historia familiar) podría llevar a un subajuste[3].

Para evitar el sobreajuste se recomienda un mínimo de 10 a 20 eventos por variable (EPV). Esto significa que –si se tienen 100 eventos en un estudio (por ejemplo, 100 pacientes con una enfermedad)– el modelo no debería incluir más de 10 variables independientes (usando el criterio de 10 EPV) o 5 variables (usando el criterio de 20 EPV)[6]. Por otro lado, para evitar el subajuste, se debe realizar una correcta selección de variables independientes. Se sugiere: 1) basarse

*Departamento de Investigación, Hospital Alemán de Buenos Aires. ORCID ID: 0000-0001-7833-9416. E-mail: cangeramo@hospitalaleman.com

en la literatura: variables que, según la evidencia científica y la experiencia clínica, tengan una posible relación con el evento estudiado, 2) evitar variables altamente correlacionadas entre sí (multicolinealidad): la inclusión de variables independientes con alta correlación puede distorsionar los resultados. Por ejemplo, durante la pandemia de COVID-19, un modelo mostró erróneamente al tabaquismo como factor protector debido a su colinealidad con la enfermedad pulmonar crónica, que absorbía su efecto real negativo[7], 3) selección guiada por el análisis simple (un predictor y la variable dependiente): se sugiere incluir, en el modelo múltiple, las variables que en el análisis univariado presenten un p-valor ≤ 0,20-0,25, ya que un umbral más estricto (p < 0,05) podría descartar factores potencialmente relevantes. Sin embargo, la inclusión final debe basarse en la relevancia clínica, más allá de la significación estadística[3,8].

Ejemplo práctico en R[9]

Supongamos que estamos estudiando los factores que influyen en la chance de que un paciente desarrolle diabetes. Utilizamos un modelo de regresión logística en R, donde la variable dependiente es "Diabetes" y las variables independientes son "Edad", "IMC"(Índice de masa corporal), "Historial familiar", "Actividad física" y "Consumo de alcohol". En aras de la brevedad, no se presentan aquí los datos originales del análisis, pero a continuación se resumen los principales resultados obtenidos.

Call:*
glm(fórmula = Diabetes ~ Edad + IMC + Historial_familiar + Actividad_fisica + Consumo_alcohol,
family = binomial, data = datos)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.12345	0.67890	-6.073	1.26e-09 ***
Edad	0.04567	0.01234	3.701	0.000215 ***
IMC	0.12345	0.03456	3.572	0.000354 ***
Historial_Familiar	1.23456	0.23456	5.263	1.42e-07 ***
Actividad_Física	-0.56789	0.12345	-4.601	4.23e-06 ***
Consumo_Alcohol	0.01234	0.09876	0.125	0.900

**Formato de salida (output) literal del software.*

Estimate (Coeficiente β): indica el efecto de cada variable independiente sobre la chance del evento (tener diabetes). Valores positivos implican mayor chance, y negativos, menor chance. Valor z y Pr (>|z|) (valor p): evalúan si el coeficiente es significativamente distinto de cero. Si p < 0,05, se considera que la variable tiene significación estadística. En nuestro ejemplo, edad, IMC e Historia familiar se asociaron significativamente con mayor riesgo de diabetes (coeficientes positivos y p < 0,05). La actividad física resultó ser un factor protector (coeficiente negativo y p < 0,05). El consumo de alcohol no mostró asociación significativa (p = 0,90).

En una regresión logística, el odds ratio (OR) se obtiene exponenciando el coeficiente β del modelo: OR = exp(β). Esta medida indica cómo cambia el odds (razón entre la posibilidad de que ocurra un evento y la de que no ocurra) por cada unidad de cambio en la variable explicativa[10]. Un OR = 1 implica ausencia de asociación entre la variable y el evento. Un OR > 1 indica un aumento en el odds del evento por cada unidad de cambio en la variable, mientras que un OR < 1 indica una disminución. La interpretación específica depende del tipo de evento evaluado: si el evento es perjudicial (por ejemplo, complicaciones o muerte), un OR > 1 sugiere un mayor riesgo y un OR < 1 sugiere un efecto protector. En cambio, si el evento es beneficioso (por ejemplo, curación o respuesta al tratamiento), un OR > 1 refleja un efecto favorable, y un OR < 1 indica una menor chance de lograr ese resultado deseado. Por lo tanto, el contexto clínico es clave para una correcta interpretación del OR[11]. Cuando el evento es poco frecuente (<10%), el riesgo relativo (RR) y el OR suelen ser similares. Sin embargo, a medida que aumenta la frecuencia del evento, un OR >1 tiende a sobreestimar y un OR <1 a subestimar el riesgo en comparación con el RR. Por esta razón, en estudios prospectivos, es recomendable ajustarlo a RR cuando la frecuencia del evento de interés supera el 10%[12].

En nuestro análisis, el IMC mostró un OR de 1,13, lo que significa que, por cada unidad adicional de IMC, el odds de desarrollar diabetes fue 1,13 veces mayor. La actividad física presentó un OR de 0,56, lo que significa que el odds de desarrollar diabetes es 0,56 veces menor en quienes realizan actividad física comparado con quienes no la realizan. Cabe destacar que el valor del OR debe interpretarse junto con su intervalo de

■ TABLA 1

Informe de los OR, IC 95% y valor p de las variables independientes incluidas en el modelo de regresión logística múltiple

Variables	OR	IC 95%	p
Edad, años	1,04	1,02-1,07	<0,001
IMC, kg/m²	1,13	1,05-1,21	<0,001
Historia familiar	3,43	2,16-5,45	<0,0001
Actividad física	0,56	0,44-0,72	<0,0001
Consumo de alcohol	1,01	0,83-1,23	0,90

confianza (IC), generalmente del 95%; si este intervalo incluye el valor 1, no se considera que la asociación represente un riesgo significativo[13]. En nuestro ejemplo, el consumo de alcohol no se asoció significativamente con el evento estudiado (OR = 1,01; IC 95%: 0,83.1,23; p = 0,90), dado que el intervalo de confianza incluye el valor nulo (OR = 1). En la tabla 1 se informan los OR, IC 95% y valor p de nuestro ejemplo.

Los supuestos (assumptions) y la evaluación del modelo de regresión logística también exceden el límite de este artículo.

Conclusión

La regresión logística es una herramienta esencial en la investigación médica, que permite no solo identificar factores de riesgo o de protección, sino también estimar posibilidades o chances: es hacer predicciones útiles para la práctica clínica. Sin embargo, su correcta aplicación requiere una cuidadosa selección de variables y una interpretación crítica de los resultados, siempre considerando que, aunque útil para identificar asociaciones, no implica causalidad por sí misma.

■ ENGLISH VERSION

"Facts do not cease to exist because they are ignored".
Aldous Huxley, Brave New World.

Introduction

When analyzing data, it is common practice to determine whether there is a relationship between two variables. The utilization of statistical tools such as correlation and regression is optimal for achieving this objective. Despite their similarities, these tools serve distinct purposes.

Correlation measures the strength and direction of the relationship between two variables. It is a symmetrical measure, which means the order of the variables does not affect the result. The correlation between A and B is the same as the correlation between B and A. However, correlation does not imply causation, nor does it allow for the making of predictions[1]. On the other hand, regression also analyzes the relationship between variables, but not symmetrically. In regression analysis, it is essential to distinguish between a dependent variable, which is to be predicted or explained, and one or more independent variables, which act as predictor variables. This distinction enables the modeling of how the independent variables influence the dependent variable. This allows us to make predictions and evaluate the relative importance of each predictor variable. However, regression, like correlation, does not imply causation[2].

When is logistic regression analysis used?

Logistic regression is commonly used in clinical research when the dependent variable is dichotomous, meaning it has only two possible categories (e.g., "presence" or "absence" of a disease)[3]. This is the binomial or binary logistic distribution. There are other distributions beyond the scope of this article, such as the multinomial and ordinal distributions[4].

The objective of logistic regression is to analyze the effect of one or more predictor (independent) variables on the dependent variable[3], allowing for:

1) Identification of risk factors or protective factors. For instance, to determine whether a clinical characteristic (e.g., age or smoking habits) increases or decreases the likelihood or probability of presenting a disease.
2) Estimating the likelihood or probability of an event occurring. such as the likelihood of developing a disease based on comorbidities. These possibilities can be represented visually in nomograms or implemented in on-line calculators to clearly and accessibly communicate the risk of an event[5].
3) Making predictions: the primary objective is to predict the most likely category of the dependent variable for a given patient (e.g., classifying a patient as high or low risk of developing a disease)[3].

We speak of likelihood or odds because logistic regression provides odds ratios and not relative risks which are estimates of probabilities.

Selection of independent variables: how to avoid overfitting and underfitting.

A multiple logistic regression model allows for the evaluation of the effect of two or more independent variables on an event (dependent variable). Unlike the model with a single predictor variable, this model adjusts for confounding factors and estimates more precise associations. One of the most important aspects of constructing the model is to appropriately select the independent variables in terms of both number and relevance. Including too many variables can lead to overfitting, which causes the model to lose its ability to generalize to new data. Conversely, excluding relevant variables can lead to underfitting, resulting in a lack of precision. For example, including irrelevant variables (such as eye color) in a study to predict the risk of

developing diabetes could cause overfitting, while omitting important variables (such as obesity or family history) could lead to underfitting[3].

A minimum of 10 to 20 events per variable (EPV) is recommended to avoid overfitting. This means that if a study has 100 events (e.g., 100 patients with a disease), the model should include no more than 10 independent variables (using the 10 EPV criterion) or 5 variables (using the 20 EPV criterion)[6]. A proper selection of independent variables is necessary to avoid underfitting. We suggest the following: 1) Use variables that, according to scientific evidence and clinical experience, may have a possible relationship with the event under study, based on the literature. 2) Avoid using variables highly intercorrelated with each other (multicollinearity) which may distort the results. For example, during the COVID-19 pandemic, one model incorrectly showed smoking as a protective factor due to its collinearity with chronic obstructive pulmonary disease, which masked its actual negative effect[7]. 3) Select variables guided by simple analysis (one predictor variable and the dependent variable): it is recommended to include in the multivariate model those variables that in the univariate analysis present a p-value ≤ 0.20-0.25, since a stricter threshold ($p < 0.05$) could exclude potentially relevant factors. However, final inclusion should be based on clinical relevance, beyond statistical significance[3,8].

A Practical example in R9

Suppose we are analyzing the factors that influence a patient's likelihood of developing diabetes. We use a logistic regression model in R, where the dependent variable is "Diabetes" and the independent variables are "Age", "BMI" (Body Mass Index), "Family History", "Physical Activity" and "Alcohol Intake". For the sake of brevity, the original data of the analysis are not presented here, but the main results obtained are summarized below.

Call:*
glm(formula = Diabetes ~ Age + BMI + Family_History+ Physical_Activity + Alcohol_Intake,
family = binomial, data = data)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.12345	0.67890	-6.073	1.26e-09 ***
Age	0.04567	0.01234	3.701	0.000215 ***
BMI	0.12345	0.03456	3.572	0.000354 ***
Family_History	1.23456	0.23456	5.263	1.42e-07 ***
Physical_Activity	-0.56789	0.12345	-4.601	4.23e-06 ***
Alcohol_Intake	0.01234	0.09876	0.125	0.900

** This is the literal software output*

The β coefficient estimates the impact of each independent variable on the likelihood of the event (defined here as having diabetes) occurring. Positive values make the outcome event more likely, while negative values make it less likely. z-value and Pr(>|z|) (p-value): they evaluate if the coefficient is significantly different from zero. A p-value < 0.05 indicates that the variable has statistical significance. In our example, Age, BMI and Family History were significantly associated with increased risk of diabetes (positive coefficients and $p < 0.05$). Physical activity was a protective factor (negative coefficient and $p < 0.05$). Alcohol intake did not show significant association ($p = 0.90$).

In logistic regression, the odds ratio (OR) is obtained by exponentiating the β coefficient of the model: $OR = \exp(\beta)$. This measurement indicates how the odds (ratio of the probability of an event happening to the probability that event not happening) change for each unit change in the explanatory variable[10]. An OR = 1 implies absence of association between the variable and the event. An OR > 1 indicates an increase in the odds of the event occurring with each unit change in the variable. An OR < 1 indicates a decrease in the odds of the event occurring. The specific interpretation depends on the type of event being assessed. If the event is harmful, such as complications or death, an OR > 1 suggests an increased risk, while an OR < 1 suggests a protective effect. By contrast, if the event is favorable (e.g., a cure or a positive response to treatment), an OR > 1 reflects a favorable effect and an OR < 1 indicates a lower likelihood of achieving the desired outcome. Therefore, understanding the clinical context is essential for correctly interpreting the OR[11]. When the event is rare (<10%), the relative risk (RR) and the OR are usually similar. Yet, the more frequent the event becomes, the more the OR will overestimate the RR when it is > 1 or underestimate the RR when it is less than 1. For this reason, in prospective studies, it is convenient to adjust them to RR when the risk of the outcome of interest is greater than 10%[12].

In our analysis, the OR for BMI was 1.13. This means that for each additional unit of BMI, the odds of developing diabetes were 1.13 times higher. Physical activity presented an OR of 0.56, which means that the odds of developing diabetes are 0.56 times lower in those who engage in physical activity compared to those who do not. Note that the OR value should be interpreted with its 95% confidence interval. If this interval includes the value 1, the association is not considered significant. In our example, alcohol intake was not significantly associated with the event studied (OR = 1.01; 95% CI: 0.83-1.23; $p = 0.90$), since the confidence interval includes the null value (OR = 1). The OR, 95% CI and p-value for our example are detailed in Table 1.

■ TABLE 1

Odds ratio, 95% CI and p-value of the independent variables included in the multivariate logistic regression model

Variables	OR	95% CI	p
Age, years	1.04	1.02-1.07	<0.001
BMI, kg/m^2	1.13	1.05-1.21	<0.001
Family history	3.43	2.16-5.45	<0.0001
Physical activity	0.56	0.44-0.72	<0.0001
Alcohol intake	1.01	0.83-1.23	0.90

The assumptions and evaluation of the logistic regression model are also beyond the scope of this article.

Conclusion

Logistic regression is an essential tool in medical research. It allows researchers to identify risk or protective factors, estimate probabilities, and make useful predictions for clinical practice. However, its proper implementation requires a careful selection of variables and a critical interpretation of the results, always considering that, although this method is useful for identifying associations, it does not imply causation.

■ ENGLISH VERSION

Referencias bibliográficas /References

1. Iovaldi ML. Correlación: otra medida del tamaño del efecto. Rev Argent Cirug. 2025;117:1-2.
2. Tanni SE, Patino CM, Ferreira JC. Correlation vs. regression in association studies [Correlação vs. regressão em estudos de associação]. J Bras Pneumol. 2020;46(1):e20200030. doi:10.1590/1806-3713/e20200030.
3. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic Regression in Clinical Studies. Int J Radiat Oncol Biol Phys. 2022;112(2):271-7. doi:10.1016/j.ijrobp.2021.08.007.
4. Liang J, Bi G, Zhan C. Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. Ann Transl Med. 2020;8(16):982. doi:10.21037/atm-2020-57.
5. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008;26(8):1364-70. doi:10.1200/JCO.2007.12.9791.
6. van Domburg R, Hoeks S, Kardys I, et al. Tools and techniques-statistics: how many variables are allowed in the logistic and Cox regression models? EuroIntervention. 2014;9(12):1472-3. doi:10.4244/EIJV9I12A245.
7. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584(7821):430. https://doi.org/10.1038/s41586-020-2521-4.
8. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. Hoboken, NJ: Wiley; 2000.
9. R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org
10. Borracci R, Arribalzaga E. La medición de la magnitud del efecto al comparar tratamientos quirúrgicos. Rev Argent Cirug. 2004;87(3-4):123-9. https://revista.aac.org.ar/index.php/RevArgentCirug/2004.
11. Iovaldi ML. Riesgo relativo y odds ratio (razón de posibilidades): Conceptos básicos. Rev Argent Cirurg. 2023;115(4):310-5. https://revista.aac.org.ar/index.php/RevArgentCirug/article/view/626.
12. Ranganathan P, Aggarwal R, Pramesh CS. Common pitfalls in statistical analysis: Odds versus risk. Perspect Clin Res. 2015;6(4):222-4. doi:10.4103/2229-3485.167092.
13. Tan SH, Tan SB. The Correct Interpretation of Confidence Intervals. Proceedings of Singapore Healthcare. 2010;19(3):276-8. doi:10.1177/201010581001900316