Int. J. Morphol.,
43(5):1545-1552, 2025.

La Regresión Logística y su Aplicación en la Investigación Biomédica

Logistic Regression and its Application in Biomedical Research

Lydia Lera[1]; Bárbara Leyton[2] & Pablo A Lizana[3]

LERA, L.; LEYTON, B. & LIZANA, P. A. La regresión logística y su aplicación en la investigación biomédica. *Int. J. Morphol, 43(5)*:1545-1552, 2025.

RESUMEN: La regresión logística es una técnica estadística ampliamente empleada en la investigación biomédica para modelar la relación entre variables independientes y una variable dependiente categórica. A diferencia de la regresión lineal, esta metodología se basa en la función logística, lo que permite predecir probabilidades entre 0 y 1 sin requerir supuestos estrictos como normalidad, homocedasticidad o linealidad. Una de sus principales fortalezas es la interpretación de los coeficientes en términos de odds ratio (OR), lo que permite cuantificar la fuerza de asociación entre exposición y resultado. La regresión logística constituye una herramienta robusta y flexible en estudios biomédicos, siempre que su aplicación sea rigurosa y se acompañe de una adecuada interpretación contextual de los resultados. La correcta especificación del modelo y la experiencia del investigador son elementos clave para garantizar la validez de las inferencias. Este trabajo presenta los conceptos, aplicaciones y ejemplos clave de la regresión logística, una herramienta importante en lainvestigación biomédica. El objetivo es facilitar la comprensión y correcta aplicación de esta técnica en diversos estudios, contribuyendo así a generar evidencia científica sólida y relevante para la toma de decisiones en el área de la salud y ciencias biológicas.

PALABRAS CLAVE: Odds Ratio; Modelos logísticos; Métodos epidemiológicos; Intervalos de confianza.

INTRODUCCIÓN

La regresión logística es una técnica estadística ampliamente utilizada en las ciencias biomédicas para analizar la relación entre variables predictoras y una variable de resultado binaria, nominal u ordinal (Stoltzfus, 2011; Ranganathan *et al*., 2017; Schober & Vetter, 2021; Vega-Fernández *et al*., 2022; Castro & Ferreira, 2023; Lizana *et al*., 2024). El objetivo fundamental de esta técnica es modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores, así como estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico). En este estudio se profundiza en el caso en que la variable dependiente es binaria, siendo la regresión logística el modelo más utilizado para el análisis de este tipo de datos donde el resultado de la respuesta para cada sujeto es un "éxito" o un "fracaso" (Agresti, 2007). Esta técnica permite predecir la probabilidad que un individuo pertenezca a una de dos categorías mutuamente excluyentes, como, por ejemplo, la presencia o ausencia de una enfermedad (King, 2008; Schober & Vetter, 2021; Harris, 2021).

La regresión logística es una de las herramientas estadísticas con mejor capacidad y más utilizada para el análisis de datos en la investigación biomédica en particular en la investigación clínica y epidemiológica. En la década de 1970 la aplicación de la regresión logística a los estudios de caso y control fue un avance importante, así como el enfoque de aprendizaje automático condicional para el ajuste de modelos en dichos estudios (Agresti, 2007). Recientemente, muchos investigadores se han enfocado en ajustar modelos de regresión logística a respuestas correlacionadas para datos agrupados.

A diferencia de la regresión lineal, la regresión logística no asume una relación lineal entre las variables independientes y la variable dependiente, sino que, en su lugar modela la relación entre las variables a través de una función logística (DeMaris, 1995; Stoltzfus, 2011), función de la familia de las curvas S y que describe la forma matemática en la que se basa el modelo logístico. El objetivo de este estudio es facilitar la comprensión y correcta

[1] Latin Division, Keiser University, Fort Lauderdale, Florida, USA.
[2] Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile, Santiago, Chile.
[3] Laboratory of Epidemiology and Morphological Sciences, Instituto de Biología, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.
FUNDED. DI Asociativo 039.302/2024, DI Regular 039.704/2025, Vicerrectoría de Investigación, Creación e Innovación de la Pontificia Universidad Católica de Valparaíso, Chile.

Received: 2025-05-20 Accepted: 2025-07-11

aplicación de esta técnica en diversos estudios, contribuyendo así a generar evidencia científica sólida y relevante para la toma de decisiones en el área de la salud y ciencias biológicas, enfatizando el papel que juegan los investigadores al utilizar correctamente este método.

Regresión Logística: Conceptos Básicos

La regresión logística es una técnica de modelado multivariable que se emplea cuando la variable dependiente es binaria (por ejemplo, presencia o ausencia de obesidad), nominal u ordinal. Es similar al modelo de regresión lineal con la diferencia del tipo de escala de la variable dependiente. Otra diferencia con la regresión lineal es que la regresión logística utiliza la función logística, también conocida como sigmoidea, para garantizar que las predicciones estén restringidas al rango de 0 a 1. Esta propiedad de la regresión logística es especialmente relevante en el campo de las ciencias biomédicas, donde con frecuencia se trabaja con características binarias, como la presencia o ausencia de una enfermedad (DeMaris, 1995; King, 2008; Harris, 2021). Además, la regresión logística no requiere que las variables predictoras se distribuyan normal, ni que haya homocedasticidad (varianza constante) o linealidad entre las variables predictoras y la variable de respuesta. Esto la convierte en una herramienta flexible y robusta para el análisis biomédico, donde las suposiciones de normalidad y linealidad pueden no cumplirse (King, 2008).

¿Qué variables incorporo al modelo de regresión logística?

La regresión logística, al igual que la regresión lineal puede tener múltiples variables explicativas las que pueden ser categóricas. Se ha descrito que el modelo de regresión logística es un modelo multivariante, por lo tanto, es tentador para el investigador incorporar todas las variables que tenga a disposición y visualizar cuales son significativas. Ese tipo de procedimientos poco rigurosos solo generará resultados espurios (Sperandei, 2014).

Antes de realizar un análisis de regresión logística se debe hacer un análisis exploratorio de las variables involucradas en el estudio. Se pueden calcular las distribuciones de frecuencias para la variable dependiente y las independientes, tablas de asociación entre la variable dependiente y cada una de las independientes (chi cuadrado, por ejemplo) y también se debe justificar la inclusión de las variables que se consideren de ajuste. Para la justificación podemos realizar un análisis estratificado de la asociación entre las variables dependiente e independientes con las variables de ajuste. Una vez realizado el análisis exploratorio se estima el modelo de regresión logística.

También se recomienda, al realizar un análisis de regresión logística multivariante, llevar a cabo un análisis de regresión logística univariante. El análisis univariante permite identificar qué variables independientes se asocian de manera significativa con la variable dependiente de interés. Este paso ayuda a seleccionar las variables que se incluirán en el modelo multivariante, reduciendo así el riesgo de sobreajuste y mejorando la parsimonia del modelo. Como regla general, si contamos con un tamaño de muestra grande, digamos que tenemos al menos diez individuos por variable, podemos intentar incluir todas las variables explicativas en el modelo completo y aplicar un método de selección de variables. Sin embargo, si el tamaño de muestra es limitado en comparación con el número de variables candidatas, se debe realizar una preselección. Una forma de hacerlo es probar todas las variables previamente, utilizando modelos con una sola variable explicativa a la vez, e incluir en el modelo multivariado todas aquellas que han mostrado un valor p relajado ($p \leq 0,25$). No es necesario preocuparse por un criterio de valor p riguroso en esta etapa, ya que se trata solo de una estrategia de preselección y no se derivará ninguna inferencia de este paso. Este criterio de valor p relajado permitirá reducir el número inicial de variables en el modelo, lo que disminuye el riesgo de omitir variables importantes que podrían generar confusión o distorsión de la variable dependiente (Sperandei, 2014). También en modelos de regresión logística con múltiples variables independientes, se puede establecer el modelo de regresión con las variables independientes incluidas y se pueden eliminar las variables no significativas ($p> 0,25$). Por otra parte, si la variable independiente obtiene un $p>0,05$ y $< p\ 0,25$ se debe observar el cambio que realiza en el OR, por lo general, cambios entre 10 /15 % la variable está generando una confusión, por lo tanto, debe permanecer en el modelo (Stoltzfus, 2011).

Detección de Multicolinealidad

El análisis univariante también permite identificar posibles problemas de multicolinealidad, situación en la cual dos o más variables independientes presentan alta correlación entre sí. Comprender estas relaciones es fundamental para tomar decisiones informadas sobre la inclusión o exclusión de variables en el modelo multivariante. Cabe destacar que muchas variables pueden estar altamente correlacionadas, lo que podría dificultar la interpretación precisa de las contribuciones individuales en las predicciones del modelo (Midi *et al*., 2010).

¿Qué tipo de variables independientes se pueden incluir en un modelo de regresión logística?

Lo interesante de un modelo de regresión logística es que puede incluir diversos tipos de variables independientes.

A continuación, se describen las variables con ejemplos típicos utilizados en ciencias biomédicas:

- **Variables continuas:** Estas variables pueden tomar cualquier valor dentro de un rango. Ejemplos: edad, presión arterial, porcentaje de grasa corporal, ingresos, temperatura (Stoltzfus, 2011). En la Figura 4 se ha incluido en el modelo la variable edad como variable continua.
- **Variables categóricas:** Representan grupos o categorías distintas. Se subdividen en:
    - **Binarias o dicotómicas:** Variables con solo dos categorías (ej. vacunado/no vacunado, fumador/no fumador, presencia/ausencia de una enfermedad). Suelen codificarse como 0 y 1.
    - **Nominales:** Variables con más de dos categorías sin un orden inherente (ej. color de ojos, grupo sanguíneo). Requieren variables dummy (también llamadas variables indicadoras) para su inclusión en el modelo, más adelante se describe con más detalle las variables dummy.
    - **Ordinales:** Variables con categorías que tienen un orden significativo (ej. nivel educativo, gravedad de una enfermedad, respuestas en una escala Likert).
- **Términos de Interacción:** Representan el efecto combinado de dos o más variables independientes sobre la variable dependiente. Por ejemplo, el efecto de la edad podría variar según el sexo biológico; un término de interacción captura esta diferencia.

Cuando se utiliza la regresión logística en un diseño de muestreo retrospectivo se tiene que la variable independiente es aleatoria, en lugar de la variable dependiente. El uso de estos diseños retrospectivos la mayoría de las veces se debe a la ocurrencia poco frecuente de una de las categorías de respuesta por lo que en un diseño prospectivo no podrían seleccionarse casos suficientes para tener una buena estimación, siendo común su uso en estudios de casos y controles biomédicos (Agresti, 2007).

Es fundamental seleccionar las variables independientes basándose en investigaciones previas y conocimiento clínico o sustantivo del tema (Stoltzfus, 2011). La inclusión de variables de confusión es esencial para evitar estimaciones sesgadas (Skelly *et al*., 2012).

Variables Dummy

En regresión logística, las variables *dummy* se utilizan para incorporar variables categóricas en el modelo. Una variable *dummy* es una variable binaria que representa las categorías de una variable categórica original. Usando el ejemplo de obesidad, se crearían dos variables *dummy* ya que tenemos tres categorías: normal, sobrepeso y obeso. Acá se pueden crear tantas variables dummy menos uno que el número de categorías.

Dummy_sobrepeso: 1 si la persona tiene sobrepeso, 0 si no.

Dummy_obeso: 1 si la persona es obesa, 0 si no.

La categoría omitida, en este caso "normal", se convierte en la categoría de referencia. El principio fundamental de las variables *dummy* es transformar una variable nominal u ordinal en variables numéricas que pueden ser incluidas en modelos de regresión. Además, es importante destacar que los resultados de los coeficientes de regresión asociados con las variables *dummy*, se interpretan en relación a la categoría de referencia [ejemplo: (Lizana & Lera, 2022) (Fig. 5)].

Odds Ratio: Significado e Interpretación

Un aspecto fundamental de la regresión logística es la relación que guarda con un parámetro de cuantificación de riesgo conocido como "*odds ratio*" (OR). Primero podemos definir ¿qué es un odds? es la división entre un evento de interés o característica y la no ocurrencia de este evento o característica (Gregoire, 2014; Sperandei, 2014; Sánchez-Villegas *et al*., 2014). Por ejemplo, tenemos 100 personas con hipertensión que recibieron un fármaco y 60 de ellos lograron ser normotenso, en este caso un odds sería la división entre los que se curaron (60) *versus* los que no (40), que sería 1,5. Lo que se interpreta que este fármaco tiene una ventaja terapéutica (razón de posibilidades) de éxito 1,5 veces mayor que el fracaso.

Sin embargo, en el contexto de la regresión logística, el OR es una medida clave que se utiliza para interpretar los coeficientes del modelo ($OR=\exp(\beta)$; β coeficiente del modelo). El odds ratio es la división entre dos odds. Siguiendo el ejemplo anterior, ahora tenemos 140 participantes con un tratamiento de ejercicio físico programado sobre la hipertensión, de ellos 100 lograron ser normopeso. Por lo tanto, el odds sería 120/20=6. Ahora, el OR sería dividir el odds de un tratamiento (fármaco) y el odds del otro tratamiento (ejercicio físico programado) que sería igual a 4 (6/1,5). En este caso, la interpretación sería, que el ejercicio físico programado ofrece una ventaja terapéutica 4 veces mayor que el fármaco. En la Figura 1, se presenta cómo calcular el OR en los ejemplos descritos.

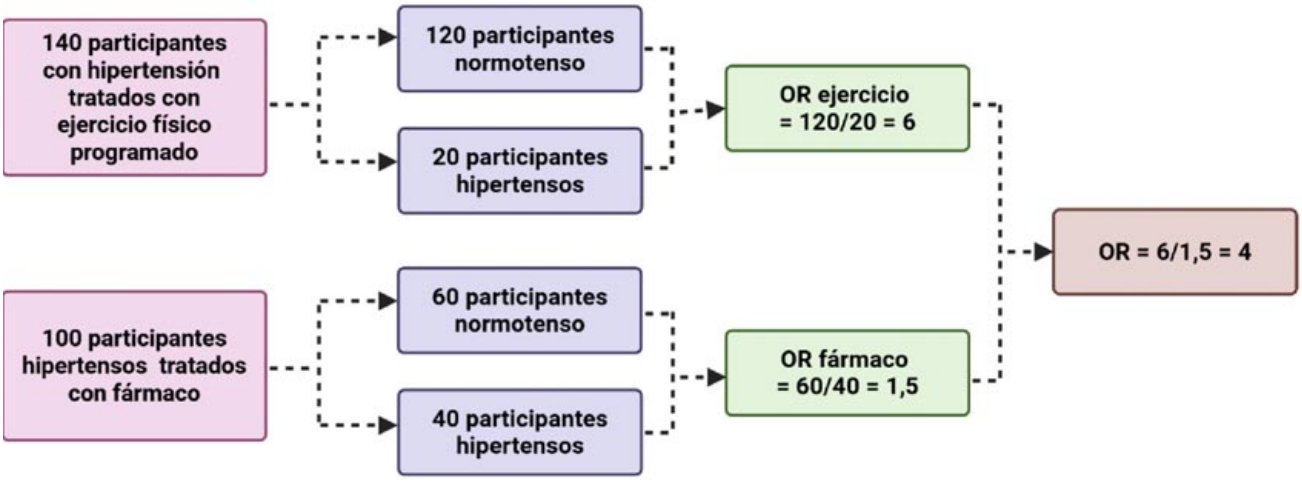

Fig. 1. Representación de la obtención de los odds ratio, que se obtienen de la división de un odds y otro odds. En los cuadros rosados se describen una serie de pacientes hipertensos, tratados con ejercicio físico y otros con un fármaco. En los cuadros morados se obtienen los resultados de los tratamientos en frecuencias, luego en los cuadros verdes se obtienen los odds se cada tratamiento y finalmente en el cuadro naranjo se obtiene el odds ratio (división entre dos odds), donde el ejercicio físico ofrece un tratamiento 4 veces superior al fármaco frente a la hipertensión. Figura creada en BioRender.

Interpretación del Odds Ratio

Luego de definir un OR, es importante interpretarlo. En este contexto, se presentan 3 escenarios que se pueden presentar al obtener la estimación de la regresión logística en OR.

OR > 1: Indica que a medida que la variable independiente aumenta, las posibilidades (razón de posibilidades) del evento de interés también aumentan. Por ejemplo, un OR de 2 sugiere que el evento es dos veces más probable de ocurrir con cada unidad de incremento en la variable independiente. Si observamos la Figura 2 donde se analiza el SIMCE de ciencias naturales de segundo medio se puede observar que los estudiantes que hacen más de 4 horas a la semana de actividad física tienen un significativo mejor rendimiento el SIMCE de ciencias naturales en comparación a los estudiantes que realizan menos de 4 horas de actividad física a la semana ($p<0,002$). Por lo tanto, el OR de 4,8 se interpreta de la siguiente manera, un estudiante que realiza más de 4 horas de actividad física a la semana tiene 4,8 veces más posibilidades (razón de posibilidades) de tener un buen rendimiento en el SIMCE de ciencias naturales comparado con los que realizan menos de 4 horas de actividad física a la semana.

		SIMCE (>p50)			
		OR	[Intervalo de confianza 95%]		p valor
Actividad física semanal	2-4 horas de actividad física semanal	Referencia	-	-	-
	>4 horas de actividad física semanal	4,827	1,821	12,791	0,002

Fig. 2. Análisis obtenido de una regresión logística donde la variable dependiente es puntaje obtenido en una prueba estandarizada (SIMCE, variable dicotómica: >p50 mejores puntajes / ≤ p50 peores puntajes) y la actividad física semanal de escolares (variable dicotómica: > a 4 horas de actividad física semanal / escolares que realizan actividad física semanal entre 2 y 4 horas semanales). La interpretación del OR de este resultado indica que los estudiantes que realizar más de 4 horas de actividad física semanal aumenta 4,8 la posibilidades (razón de posibilidades) de tener un puntaje SIMCE sobre el p50, comparado con los estudiantes que realizan actividad física entre 2 a 4 horas a la semana. Abreviaturas: OR, odds ratio, p percentil. Figura creada en BioRender.

Además, en ciencias biomédicas un OR mayor a 1 se describe como un factor de riesgo entre la exposición y el resultado. En el ejemplo de la Figura 3, se observa que la presencia de trastornos musculoesqueléticos es un factor de riesgo en presentar una peor calidad de vida en el componente mental (p<0,001). Por lo tanto, se puede interpretar como: una persona que tiene trastornos musculoesqueléticos presenta 4,5 veces más posibilidades de tener una peor calidad de vida en el componente mental, comparado con los que no presentan trastornos musculoesqueléticos.

		Calidad de Vida - Componente mental (<p50)			
		OR	[Intervalo de confianza 95%]		p valor
Trastornos Musculo-esqueléticos	Sin trastornos musculo-esqueléticos	Referencia	-	-	-
	Con trastornos musculo-esqueléticos	4,500	2,231	9,081	<0,001

Fig. 3. Análisis obtenido de una regresión logística donde la variable dependiente es calidad de vida en su componente mental (variable dicotómica: <p50 peor calidad de vida / ≥p50 mejor calidad de vida) y la presencia o ausencia de trastornos musculoesqueléticos. La interpretación del OR de este resultado indica que la presencia de trastornos musculoesqueléticos aumenta 4,5 veces el riesgo de tener una peor calidad de vida en el componente mental, comparado con los que no presentan trastornos musculoesqueléticos. Abreviaturas: OR, odds ratio, p percentil. Figura creada en BioRender.

OR = 1: Indica que la variable independiente no tiene efecto sobre las posibilidades del evento de interés. En la práctica es inusual obtener un OR exactamente igual a 1 por lo que no se profundizará en este caso a través de ejemplos.

OR < 1: Indica que a medida que la variable independiente aumenta, las posibilidades del evento de interés disminuyen. Por ejemplo, un OR de 0,5 sugiere que el evento es la mitad de probable de ocurrir con cada unidad de incremento en la variable independiente. En la Figura 4 se presenta un ejemplo en personas que cumplen las recomendaciones de actividad física de la organización mundial de la salud (OMS) (Szumilas, 2010) versus las que no lo cumplen y su asociación con presentar hipertensión arterial, se puede observar que las personas que cumplen las recomendaciones de la OMS tienen un OR=0,13, lo cual indica que las personas que realizan actividad física tienen un 87 % menor posibilidad (1 - 0,13 = 0,87) de presentar hipertensión en comparación con quienes no realizan actividad física. Esta interpretación puede también expresarse en términos de razones de odds:

La razón de odds de presentar hipertensión es 0,18 veces (es decir, significativamente menor) en quienes realizan actividad física, comparado con quienes no la realizan.

		Hipertensión arterial			
		OR	[Intervalo de confianza 95%]		p valor
Actividad física según la Organización Mundial de la Salud	No cumple las recomendaciones	Referencia	-	-	-
	Cumple las recomendaciones	0,126	0,062	0,253	0,001

Fig. 4. Análisis obtenido de una regresión logística donde la variable dependiente es hipertensión arterial (variable dicotómica: presencia de hipertensión / ausencia de hipertensión) y el cumplimiento de la actividad física según la organización mundial de la salud para adultos. La interpretación del OR de este resultado indica que el cumplimiento de las recomendaciones de actividad física son un factor protector de presentar hipertensión arterial, comparado con los que no cumplen las recomendaciones. Abreviatura: OR, odds ratio. Figura creada en BioRender.

Es crucial tener en cuenta que la interpretación del OR debe realizarse en el contexto específico del estudio y considerando las demás variables incluidas en el modelo. También, es importante incluir los intervalos de confianza del 95 % para el OR (Szumilas, 2010).

Intervalos de confianza en la regresión logística

En las Figuras 2-4, también se han presentado los intervalos de confianza ¿qué importancia pueden tener en la regresión logística? Partamos por conocer qué son: los intervalos de confianza proporcionan un rango de valores dentro del cual es probable que se encuentre el verdadero valor del parámetro poblacional con un cierto nivel de confianza (típicamente 95 %). Si observamos los resultados de la Figura 3 en una muestra de 100 participantes se obtiene un OR de 4,5 en la asociación entre trastornos musculoesqueléticos y calidad de vida-componente mental con un intervalo de confianza del 95 % entre 2,2 y 9,1 ¿qué interpretación le podemos otorgar a estos valores del intervalo de confianza? El intervalo de confianza al 95% calculado en el ejemplo refleja que, si se repitiera el estudio

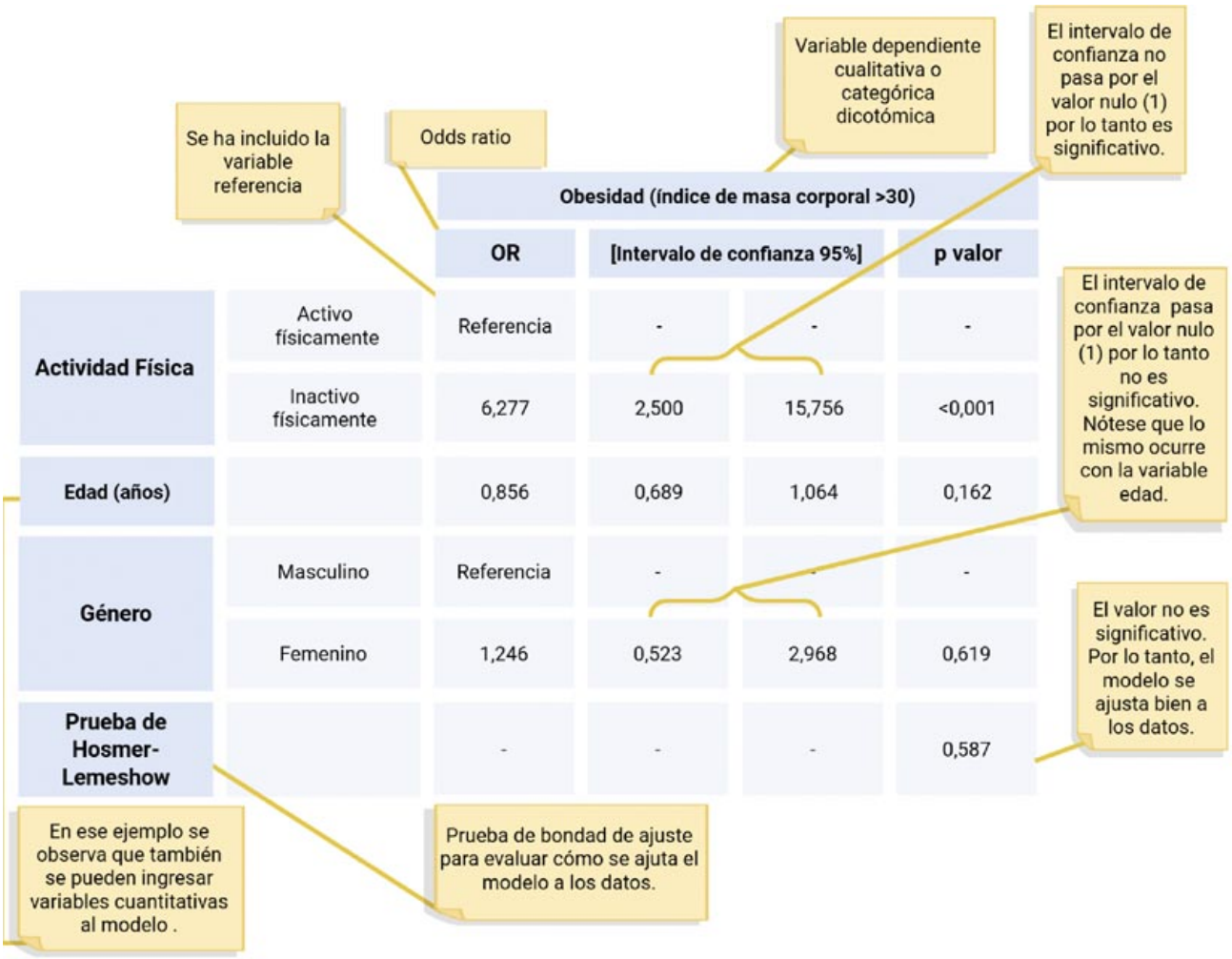

Fig. 5. Se representa los resultados de un modelo de regresión logística binaria, donde la variable dependiente es Obesidad (índice de masa corporal >30), que es dicotómica (presencia o ausencia). Se describen los Odds Ratio (OR), intervalos de confianza al 95 % (IC 95 %) para el OR y el p-valor para evaluar significancia estadística. Interpretación del análisis de regresión logística binaria presentado en la figura: La variable actividad física tiene por referencia a los activos físicamente, comparado con ellos los inactivos físicamente tienen un OR de 6,28. Es decir, los inactivos físicamente tienen 6,28 veces más riesgo de ser obesos que los activos físicamente. El IC 95 % de la variable actividad física es (2,500; 15,756), no incluye el 1, por lo que es estadísticamente significativo ($p < 0,001$). El modelo tiene la variable edad, que es variable continua, que no es estadísticamente significativa ($p = 0,162$). También se presenta la variable categórica género, donde la categoría masculina se presenta como referencia. Al igual que el caso de la variable edad, género tampoco es significativa ($p = 0,619$). Finalmente se presenta la prueba de Hosmer-Lemeshow con un $p = 0,587$, dado que el valor $p > 0,05$, no se rechaza la hipótesis nula de buen ajuste, lo que indica que el modelo se ajusta adecuadamente a los datos observados. Figura creada en BioRender.

en múltiples muestras de 100 participantes, el 95 % de estas muestras presentaría un valor de odds ratio entre 2,2 y 9,1. Esto sugiere que el verdadero valor del OR se encuentra probablemente dentro de este rango, con un nivel de confianza del 95 %. Los límites de este intervalo de confianza proporcionan una estimación del tamaño de efecto mínimo y máximo asociado a la exposición o intervención de interés, lo cual tiene una interpretación práctica y relevante.

Ahora bien, a través de la observación del intervalo de confianza ¿se puede inferir que existe significancia estadística? La respuesta es sí, en el ejemplo de la Figura 3 el intervalo de confianza para la asociación entre trastornos musculoesqueléticos y calidad de vida-componente mental excluye el valor nulo (OR = 1; 2,2 y 9,1), lo que sugiere que existe una asociación estadísticamente significativa entre ambas variables en la población estudiada. Para visualizar el efecto contrario, se puede observar la misma Figura 5 en la asociación entre edad y obesidad, dónde el valor del intervalo de confianza incluye el valor nulo (es decir, el 1; 0,69-1,06), lo que indica que no se rechaza que el valor poblacional podría incluir el 1, lo que llevaría a concluir que no existe asociación significativa, en este caso entre edad y obesidad (Cerda *et al*., 2013)

¿Cómo se evalúa si el modelo se ajusta adecuadamente a los datos?

Para evaluar si el modelo se ajusta adecuadamente a los datos, existen varias pruebas y métricas que se pueden utilizar. Sin embargo, una prueba ampliamente utilizada y aceptada es la prueba de Hosmer-Lemeshow. La prueba de Hosmer-Lemeshow es una prueba estadística que evalúa si las probabilidades predichas por el modelo de regresión logística son similares a las probabilidades observadas en los datos. Esta prueba divide a los sujetos en grupos basados en sus probabilidades predichas y luego compara las probabilidades observadas en cada grupo con las probabilidades predichas. Por lo tanto, con una prueba de bondad de ajuste queremos predecir si la calibración del modelo no estará lejos de la realidad. Si el modelo se ajusta bien las diferencias entre lo que predice y lo que realmente sucede deben ser no significativas. Por el contrario, si el resultado es significativo podríamos decir que el modelo no se ajusta bien a los datos, lo que indica que puede haber problemas con la especificación/calibración del modelo.

Es crucial evaluar la adecuación del modelo para evitar inferencias erróneas (Hosmer *et al*., 1991). Para ver un ejemplo de la utilización de la prueba de Hosmer-Lemeshow, se muestra la Figura 5 donde se ha incluido el valor de la prueba de bondad de ajuste de Hosmer-Lemeshow.

CONCLUSIONES

La regresión logística binaria se establece como un método estadístico fundamental para modelar y analizar variables dependientes dicotómicas en el campo biomédico, proporcionando una poderosa herramienta para discernir asociaciones en datos categóricos. Además, la regresión logística permite analizar múltiples variables explicativas simultáneamente, reduciendo el efecto de los factores de confusión. Sin embargo, enfatizamos que los investigadores deben prestar atención a la construcción del modelo, evitando limitarse a ingresar datos brutos en el software y pasar directamente a los resultados. Numerosas decisiones sobre la construcción de modelos dependerán enteramente de la experiencia del investigador, la mejor elección de un modelo requiere un ejercicio de juicio, no sólo de cálculo.

AGRADECIMIENTOS. DI Asociativo 039.302/2024, DI Regular 039.704/2025, Vicerrectoría de Investigación, Creación e Innovación de la Pontificia Universidad Católica de Valparaíso, Chile.

LERA, L.; LEYTON, B. & LIZANA, P. A. Logistic regression and its application in biomedical research. *Int. J. Morphol, 43(5)*:1545-1552, 2025.

SUMMARY: Logistic regression is a statistical technique widely used in biomedical research to model the relationship between independent variables and a categorical dependent variable. Unlike linear regression, this methodology is based on the logistic function, which allows predicting probabilities between 0 and 1 without requiring strict assumptions such as normality, homoscedasticity, or linearity. One of its main strengths is the interpretation of the coefficients in terms of odds ratio (OR), which allows quantification of the strength of association between exposure and outcome. Logistic regression is a robust and flexible tool in biomedical studies, provided its application is rigorous and accompanied by an adequate contextual interpretation of the results. The correct model specification and the researcher's experience are key elements to guarantee the validity of the inferences. This paper presents the concepts, applications, and key logistic regression examples, an important tool in biomedical research. The objective is to facilitate the understanding and correct application of this technique in various studies, thus generating solid and relevant scientific evidence for decision-making in the health and biological sciences.

KEY WORDS: Odds Ratio; Logistic models; Epidemiological methods; Confidence intervals.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. *An Introduction to Categorical Data Analysis*. 2nd ed. Hoboken, John Wiley & Sons, 2007.

Castro, H. M. & Ferreira, J. C. Linear and logistic regression models: when to use and how to interpret them? *J. Bras. Pneumol., 48(6)*:e20220439, 2023.

Cerda, J.; Vera, C. & Rada, G. Odds ratio: aspectos teóricos y prácticos. *Rev. Med. Chil., 141(10)*:1329-35, 2013.

DeMaris, A. A tutorial in logistic regression. *J. Marriage Fam., 57(4)*:956, 1995.

Gregoire, G. Logistic regression. *EAS Publ. Ser., 66*:89, 2014.

Harris, J. K. Primer on binary logistic regression. *Fam. Med. Community Health, 9(Suppl. 1)*:e001290, 2021.

Hosmer, D. W.; Taber, S. & Lemeshow, S. The importance of assessing the fit of logistic regression models: a case study. *Am. J. Public Health, 81(12)*:1630-5, 1991.

King, J. *Binary Logistic Regression*. In: Osborne, J. (Ed.). Binary Logistic Regression. New York, SAGE Publications, 2008. pp.358-84.

Lizana, P. A. & Lera, L. Depression, anxiety, and stress among teachers during the second COVID-19 wave. *Int. J. Environ. Res. Public Health, 19(10)*:5968, 2022.

Lizana, P. A.; Vilches-Gómez, V.; Barra, L. & Lera, L. Tobacco consumption and quality of life among teachers: a bidirectional problem. *Front. Public Health, 12*:1369208, 2024.

Midi, H.; Sarkar, S. K. & Rana, S. Collinearity diagnostics of binary logistic regression model. *J. Interdiscip. Math., 13(3)*:253-67, 2010.

Ranganathan, P.; Pramesh, C. S. & Aggarwal, R. Common pitfalls in statistical analysis: logistic regression. *Perspect. Clin. Res., 8(3)*:148-51, 2017.

Sánchez-Villegas, A.; Bes-Rastrollo, M. & Martínez-González, Á. *Regresión Logística*. En: Martínez-González, Á.; Sánchez-Villegas, A.; Toledo-Atucha, E.; Faulin-Fajardo, J. (Eds.). Bioestadística Amigable. Amsterdam, Elsevier, 2014.

Schober, P. & Vetter, T. R. Logistic regression in medical research. *Anesth. Analg., 132(2)*:365-6, 2021.

Skelly, A. C.; Dettori, J. R. & Brodt, E. D. Assessing bias: the importance of considering confounding. *Evid. Based Spine Care J., 3(1)*:9-12, 2012.

Sperandei, S. Understanding logistic regression analysis. *Biochem. Med. (Zagreb), 24(1)*:12-8, 2014.

Stoltzfus, J. Logistic regression: a brief primer. *Acad. Emerg. Med., 18(10)*:1099-104, 2011.

Szumilas, M. Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry, 19(3)*:227-9, 2010.

Vega-Fernández, G.; Olave, E. & Lizana, P. A. Musculoskeletal disorders and quality of life in Chilean teachers: a cross-sectional study. *Front. Public Health,* 10:810036, 2022.

Dirección para correspondencia:
Dr. Pablo Lizana Arce
Instituto de Biología
Laboratorio de Epidemiología y Ciencias Morfológicas
Pontificia Universidad Católica de Valparaíso
Valparaíso
CHILE

E-mail: pablo.lizana@pucv.cl