REFLEXIÓN ACERCA DE LA REGRESIÓN LOGÍSTICA Y LAS DECISIONES CLÍNICAS

Víctor Patricio Díaz-Narváez*[1,2]; Aracelis Calzadilla Núñez[3]; Alejandro Reyes Reyes[4]

1 Facultad de Salud. Universidad Bernardo O'Higgins. Santiago. Chile.
2 Investigador Adjunto. Universidad Autónoma de Chile. Santiago. Chile.
3 Departamento de Investigaciones. Universidad Gabriela Mistral. Santiago. Chile.
4 Carrera de Psicología. Universidad Santo Tomás. Concepción. Chile.
Profesor correspondencia: Dr. Víctor Patricio Díaz Narváez (PhD). Facultad de Salud. Universidad Bernardo O'Higgins. Av. Viel 1497. Santiago. Chile.
Conflicto de Intereses: No existen conflictos de intereses.
Ética de la Investigación. Este trabajo no emplea seres humanos o animales de experimentación.

Resumen

Introducción. La aplicación del estudio de riesgos mediante la regresión logística múltiple implica necesariamente realizar tales estimaciones con la necesaria información al clínico acerca de todas las limitaciones que dicho trabajo tiene. **Problema.** Es posible que los trabajos que aplican la regresión logística múltiple no incluyan el uso de las pruebas de Bondad de Ajuste o del Coeficiente de Determinación o ambas. **Desarrollo**: pueden existir altos valores de riesgo relativo u odds ratio junto a modelos no ajustados o bajos valores del coeficiente de determinación o ambos al mismo tiempo. Como consecuencia, si el profesional clínico no posee esta información, en relación a una entidad dada, no puede saber si las estimaciones de altos valores de riesgo tienen o no importancia clínica y, por tanto, la ausencia de reportes de estos estimadores podría implicar una intervención terapéutica cuyos resultados podrían no ser los esperados. **Conclusión:** Los investigadores que aplican el modelo de la regresión logística deben obligatoriamente informar de los resultados de los estimadores que le dan consistencia a dicho modelo.
Palabras Claves: Regresión Logística, Estimadores del Modelo, Decisiones Clínicas.

REFLECTION ON LOGISTIC REGRESSION AND CLINICAL DECISIONS

Summary

Introduction. *The application of risk study through multiple logistic regression necessarily involves making such estimates with the necessary information to the clinician about all the limitations that such work has.* ***Problem.*** *It is possible that works that apply multiple logistic regression do not include the use of the tests of Goodness of Adjustment or the Coefficient of Determination or both.* ***Development:*** *there may be high values of relative risk or odds ratio together with unadjusted models or low values of the coefficient of determination or both at the same time. Therefore, if the clinician does not have this information, in relation to a given entity, it cannot know whether the estimates of elevated risk values are of clinical importance or not and, therefore, the absence of reports of these estimators could imply a therapeutic intervention whose results may not be as expected.* ***Conclusion:*** *Researchers applying the logistic regression model must necessarily report the results of the estimators that give consistency to said model.*
Key Words: *Logistic Regression, Model Estimators, Clinical Decisions.*

* vpdiaz@tie.cl; vicpadina@gmail.com

Introducción

Se conoce que los modelos de regresión en general tratan acerca del conocimiento de la influencia que ciertas variables independientes cuyo producto es la modificación de las variables dependientes [1,2]. La regresión lineal simple es un caso particular del modelo de regresión multivariada (lineal o no lineal). Esta última requiere de ciertos supuestos como son los de normalidad multivariada y homocedasticidad [1,3] (entre otros), todo lo cual imprime dificultades en el análisis cuándo los datos no cumplen las condiciones nombradas. La regresión logística es una herramienta estadística de gran utilidad para la investigación clínica y epidemiológica debido a su capacidad de modelar la probabilidad de un suceso (con datos no distribuidos normalmente y no homocidásticos) con el objeto de evaluar asociaciones entre variables y, por tanto, en la observación de ciertos factores considerados como riesgos. [3,4] Estos riesgos tienen el atributo de que pueden ser cuantificados [4,5]. La cuantificación de los riesgos (con significación estadística, mediante intervalos de confianza: IC) permiten estimar dos probabilidades (Riesgo Relativo: RR) y Odds Radio ("Razón de productos cruzados": OR), entre otros estadígrafos de este tipo. Sin embargo, a pesar de que la regresión logística no requiere los supuestos antes nombrados para la regresión lineal simple o multivariada, es necesario cumplir con los supuestos y exigencias que son propios de este tipo de prueba y, sobre todo, informar el grado de intensidad asociativa entre las variables.

Coincidimos con Calderón y de los Godos [4] con el hecho que la aplicación masiva de esta técnica, junto al uso de programas estadísticos, puede conducir a un proceso de empleo inadecuado y mecánico de la misma, que podría evitarse con el conocimiento de las ventajas y desventajas que posee y de la correcta interpretación de los estadígrafos asociados a esta técnica. De hecho, podrían influir elementos subjetivos, tanto en la elección de los modelos, de las propias variables que se incluyan en el estudio, así como de las inferencias que se realizan cuando, precisamente, no se entienden las implicancias de los resultados de las técnicas estadísticas o por la omisión de la estimación de algunos de los parámetros de estas mismas pruebas que les darían consistencia a los resultados. No es garantía de objetividad el uso de la matemática por sí misma para asegurar la obtención de hechos científicos, [6] en general, y especialmente aquellos de importancia en la práctica clínica. El principal problema, desde el punto de vista médico, es que se pueden asumir ciertas inferencias estadísticas como hechos científicos y sean trasladados a la práctica clínica cuando tienen poca o ninguna incidencia en las patologías (crónicas o no crónicas). Por tales razones, el propósito de este artículo es describir teóricamente las condiciones esenciales que son necesarias e imprescindibles para proporcionar una información eficaz a la práctica clínica y contribuir a una correcta a interpretación clínica de la regresión logística, de forma asequible para el profesional médico y sin complicaciones matemáticas innecesarias.

Desarrollo

Modelo general e Interpretación de los coeficientes.

La mayor de todas las ventajas de la regresión logística es el hecho de estimar si cierto acontecimiento pudo o no ocurrir. Esta situación da origen a un valor binomial (o multinomial, según el caso) que constituye la variable dependiente y que la regresión logística tiende a predecir el comportamiento de esta variable, sobre la base de la probabilidad de que un suceso se produzca o no a partir de la evaluación de las variables independientes. Estas últimas, pueden tener distribución binomial o cuantitativa discreta o continua. Si la predicción resulta mayor a el valor de probabilidad de 0,50 implica que la variable independiente es capaz de predecir de forma positiva la ocurrencia de cierto fenómeno y viceversa en la misma medida que dicha predicción se acerca al valor de 1. Tal procedimiento nos conduce a estimar el coeficiente logístico, el cual nos permite calcular la razón de dos probabilidades: la ocurrencia y la no ocurrencia de un fenómeno. Esta razón de probabilidades (odds ratio) se expresa:

$$\text{Prob}_{(\text{suceso})} / \text{Prob}_{(\text{no suceso})} = e^{(\text{Exp})\,\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}$$

Donde X_1, X_2,…..,X_n son las variables independientes y "e" es la base del logaritmo neperiano.

Los coeficientes estimados (β_0, β_1, β_2,….,β_n) constituyen medidas de los cambios que se producen en la razón de probabilidades en la variable dependiente. Desde el punto de vista práctico y concreto, estos coeficientes están precedidos por signos: un coeficiente positivo indica un aumento de la probabilidad y un negativo disminuye la probabilidad de que ocurra un fenómeno. Como consecuencia, el signo del coeficiente es una forma de conceptualizar el tipo de asociación que existe entre la variable independiente con respecto a la variable dependiente.

Tabla I. Resultados ficticios de la evaluación de tres variables independientes sobre una variable dependiente.

						Intervalo de Confianza (95%)		
Variable	**β**	**DS**	**Wald**	**g.l.**	**Sig**	**Exp (β)**	**LI**	**LS**
Constante	1,80	0,28	34,65	1	0,0001			
V_1	-0,89	0,12	45,33	1	0,0001	0,389	0,311	0,522
V_2	-0,13	0,008	0,084	1	0,812	0,987	0,980	1,065
V_3	-0,23	0,029	53,71	1	0,0001	0,801	0,759	0,839

DS: Desviación Estándar del coeficiente de regresión correspondiente. **g.l.**: Grados de libertad; **Exp (β):** Razón de probabilidades o riesgo. **LI**: Límite Inferior; **LS**: Límite Superior.

Modelo de la regresión logística.

Supongamos un ejemplo ficticio donde observamos en la Tabla I los resultados de una regresión logística múltiple:

Esta ecuación tiene varias columnas y nos referiremos a las relacionadas estrictamente con el objetivo del presente trabajo. En la primera columna aparecen los coeficientes con sus respectivos signos. Si la variable independiente está constituida por la "presencia" de alguna característica (por ejemplo: con tratamiento) le damos el valor 1 y a la "ausencia" de la misma (por ejemplo: sin tratamiento) le damos el valor 0, entonces el signo negativo en la variable V_1 nos indicará que la presencia de esa característica disminuye la probabilidad de que se presente la característica estudiada en la variable dependiente (por ejemplo: condición de enfermo con valor 0 y sano con valor 1). En el ejemplo, todos los coeficientes tienen signos negativos y, por tanto, la ausencia de la característica en las variables independientes nos indica que disminuye la probabilidad de la presencia de la condición de sano en la variable dependiente.

La tercera columna contiene los resultados del estadígrafo de Wald que somete a prueba si los valores de β son iguales a 0 (hipótesis estadística nula) o diferentes de este valor (hipótesis estadística alternativa). Este estadígrafo sigue una distribución χ^2 y podemos encontrar su significación situada en la columna 5. Luego (para un $\alpha \leq 0,05$) es posible afirmar que los coeficientes de las variables V_1 y V_3 tienen coeficientes diferentes de 0. Como consecuencia, estas dos variables están asociadas de alguna manera a la variable dependiente (específicamente, surgen como variables protectoras). Que estos coeficientes sean diferentes de 0 no significa necesariamente una alta asociación entre la variable independiente y dependiente.

En la columna 6 se sitúan los valores de probabilidad de un evento y en las últimas columnas el intervalo de confianza. Si el valor de 1 se encuentra dentro del intervalo, se dice que la razón de probabilidades o riesgo no es significativo y si está fuera del intervalo se dice que es significativo. La interpretación de este riesgo está debidamente explicada.[5]

Lo antes expresado es lo que usualmente se emplea para la confección de artículos científicos que son enviados a publicar y, normalmente, si cumplen ciertos estándares, estos son aceptados . Sin embargo, la puede ocurrir que existan artículos que no estiman (o no reportan) algunos estadígrafos que son cruciales para determinar si el riesgo calculado (relativo, odds ratio u otro) tiene alguna significación clínica o es capaz de entregar la información necesaria para que el propio clínico pueda evaluar su real importancia desde las leyes teóricas y empíricas de su disciplina. Para superar esta posible limitación es necesario que los clínicos pongan su atención en los siguientes principales estadígrafos y en su correspondiente interpretación cuando se encuentren frente a trabajos que aplican la regresión logística múltiple:

1. Ajustes del modelo y Pruebas de Bondad de Ajuste.

Estas son pruebas que permiten saber si los datos que tratamos están bien representados por el modelo de regresión. Un buen ajuste indicará finalmente que los

riegos calculados constituyen una buena estimación de su valor estadístico.

a) Prueba de Ómnibus. Garantiza que el modelo tenga capacidad descriptiva. Si la prueba de χ^2, asociada a la de Ómnibus, es significativa quiere decir que las variables consideradas en el modelo son capaces de predecir el comportamiento probabilístico de la variable dependiente.

Un ejemplo de la importancia de la prueba de Ómnibus puede ser el siguiente (Tabla II):

Tabla II. Resultados ficticios de una Prueba de Ómnibus.

	Chi cuadrado	gl	Significación
Paso	8,669	3	0,034
Bloque	8,669	3	0,034
Modelo	8,669	3	0,034

Estos resultados significan que fue suficiente analizar todas las variables en un solo bloque y que no es necesario hacerlo paso a paso. Si así no fuera, estaríamos obligados a evaluar qué tipo de entrada es la mejor de todas y, por tanto, emplear métodos diferentes de introducción de las variables en el modelo hasta encontrar el que mejor se ajuste.

b) Valor de la verosimilitud (-2LL ó –2 veces el logaritmo del valor de la verosimilitud). Los valores que tengan un mayor ajuste deberán tener necesariamente un menor valor de -2LL; de tal forma que si -2LL = 0 (verosimilitud =1), el ajuste es perfecto.[7] Por ejemplo, valores que son inferiores a 100, en este estadígrafo, pueden considerarse satisfactorios para decidir que el modelo tiene buen ajuste. Mientras más se acerque a 0 será mejor el ajuste.

c) Prueba de Hosmer y Lebeshow (H-L). Permite determinar si el modelo propuesto puede explicar los datos del fenómeno que se estudia. Ejemplo: Si χ^2 asociada a esta prueba tuviese un valor de 11,13 (con 7 grados de libertad), entonces p=0,133. Esto quiere decir que no existen diferencias entre valores observados y esperados y, por tanto, la prueba nos orienta de la existencia de un buen ajuste entre el modelo y los datos.[8]

d) Tabla de clasificación: La Tabla III (resultados ficticios) permite también interpretar el ajuste del modelo a los datos sobre la base del valor de corte de 0,50 en un análisis de regresión logística realizado en un solo paso. Constituye una descripción práctica muy útil para evaluar el ajuste. Valores altos de falsos positivos o falsos negativos constituyen un dato empírico irrefutable de la calidad del ajuste. Por ejemplo, si nos encontramos con los siguientes resultados ficticios:

Tabla III. Resultados ficticios de una posible clasificación de los sujetos estudiados en función de su condición observada con el pronóstico de dicha condición realizada por el modelo.

			Pronosticado		
			Variable dependiente		Porcentaje correcto
			Enfermo	Sano	
Observado	**Variable dependiente**	**Enfermo**	30	18	62,5
		Sano	19	29	60,4
	Porcentaje global				61,5

Como se puede apreciar, el valor observado de enfermos fue de 30 y de sano 29 y entrega entonces un porcentaje de clasificación de 62,5% y un 60,4 % de correcta clasificación. Un resumen de toda esta información es: **Sensibilidad**: % de sanos que son clasificados por el modelo como sanos: (29/48) =60,4%. **Especificidad**: % de enfermos que son clasificados por el modelo como enfermos: (30/48) =62,5%. **Tasa de falsos positivos**: (18/48) =37,5%. **Tasa de falsos negativos**: (19/48) = 39,5%. Por tanto, existe un 60,4% de posibilidades de ser sano a partir del comportamiento de las variables independientes estudiadas, valor que no es muy alto. Por otra parte, los valores de falsos positivos y falsos negativos son demasiado altos para aceptarlos sin reticencia clínica. Con estos resultados, tal vez, es posible dudar de que los valores de riesgo que podrían haberse calculado en la aplicación de la regresión logística realmente denoten el verdadero valor estadístico de este riesgo en la población. Luego, sin esta información no se puede conocer si los valores de riesgo estimados sean importantes para la clínica.

2. Coeficiente de determinación (R^2). Por mientras el estadígrafo de Wald muestra la significación de la variable, es decir, indica la existencia de asociación entre la variable dependiente y la independiente, el coeficiente de correlación **mide el grado de asociación** entre dos o más variables. Luego, el estadígrado de Wald podría ser altamente significativo (p=0,001), pero al mismo tiempo puede ocurrir que el grado de asociación sea muy bajo. Esto tiene una importancia crucial para el clínico, pues a partir de esta información tomará la decisión de introducir o no a la práctica clínica de forma operativa la información observada.

Existen dos estimaciones que son las más frecuentes en la regresión logística: R^2 de Cox y Snell y R^2 de Nagelkerke. Cualquier coeficiente de determinación pretende estimar en qué grado o magnitud una variable dependiente o un conjunto de ellas puede explicar la varianza de la variable dependiente, así: a) El R^2 de Cox y Snell es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes), se basa en la comparación del log de la verosimilitud (LL) para el modelo respecto al log de la verosimilitud (LL) y sus valores teóricamente fluctúan entre 0 y 1, pero en la práctica no llega a 1, lo cual implica una seria limitación metodológica y b) El R^2 de Nagelkerke es una transformación del R^2 de Cox y Snell. Este estadígrafo corrige la escala del estadístico para cubrir el rango completo de 0 a 1. Si en un ejemplo ficticio el R^2 de Cox y Snell estimado fuera de 0,021 implicaría que las variables independientes empleadas en el modelo de regresión solamente explican el 2,1 % de la varianza de la variable dependiente. Dicho de otra forma, solo explica el 2,1 % de la condición o evento que se estudia (entidad) y si el R^2 de Nagelkerke fuese estimado en un 0,028 tendría la misma interpretación que la anteriormente descrita, pero en esta oportunidad explicaría sólo el 2,8% de la variación de la variable dependiente. El clínico y solo él puede decidir si esta información puede serle útil o no. Como consecuencia, el investigador tiene la obligación de informar el coeficiente de determinación para que el clínico, a su vez, pueda tomar decisiones pertinentes. Cuando no se reportan estos indicadores en los resultados de la aplicación de la técnica de regresión logística, los resultados de la investigación están incompletos y las conclusiones presentadas en el artículo pueden estar sesgadas, casi siempre, con la tendencia a hiperbolizar la importancia de los valores de riesgo encontrados.

Conclusión

Un análisis de regresión logística puede arrojar el siguiente escenario: a) pueden existir altos valores de riesgo relativo u odds ratio junto a modelos no ajustados o bajos valores del coeficiente de determinación o ambos al mismo tiempo, los cuales (en el contexto analizado) son desconocidos para el clínico y b) valores altos de riesgo estimados en condiciones de un modelo ajustado y, además, con altos valores del coeficiente de determinación los cuales podrían constituir una información del mayor interés para el clínico (pero que, en el mismo contexto, son desconocidos para él). Como consecuencia, si el clínico no posee esta información, tampoco puede saber (sólo con las estimaciones de altos valores de riesgo y su correspondiente significación) si tal situación necesariamente tiene o no importancia clínica y, por tanto, la ausencia de reportes de estos estimadores constituye un serio error metodológico en el empleo de la técnica de regresión logística. Los editores de las revistas y los árbitros deben estar atentos a la situación descrita y exigir por norma la presentación de los indicadores antes señalados y los valores asociados a su estimación.

Bibliografía

1. Díaz-Narváez VP. *Metodología de la Investigación Científica y Bioestadística para Profesionales y Estudiantes de Ciencias de la Salud. 2da. Ed*. RiL Editores, Santiago . 67-79, 2009.
2. Reding A, Zamora M, López JC. *¿Cómo y cuándo realizar un análisis de regresión lineal simple? Aplicación e interpretación.* Dermatol Rev Mex 55: 395-402, 2011.
3. Navarro E, Verbel A, Robles D, Hurtado K R. *Regresión Logística Ordinal Aplicada a la Identificación de Factores de Riesgo para Cáncer de Cuello Uterino*. Ingeniare 9:87-105, 2014.
4. Calderón JP, de los Godos *LA. Regresión Logística Aplicada a la Epidemiología*. Rev Salud, Sexualidad y Sociedad. 1: 78-84, 2009.
5. Cerda J, Vera C, Rada G. *Odds ratio: aspectos teóricos y prácticos*. Rev Med Chile 141:1329-1335, 2013.
6. Díaz-Narváez VP, Calzadilla-Núñez A, López Salinas H. *Una Aproximación al Concepto de Hecho Científico. Rev Austral* de *Ciencias Sociales 8*:3-16, 2004.
7. Hair JF, Anderson RE, Tatham RL, Black WC. *Análisis multivariante. 5ª ed*. Prentice Hall. Madrid: 283, 2001
8. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Wiley Ed. New York, 1989.

Es mal indicio que a un ictérico se le endurezca el hígado. Las mujeres no padecen gota hasta llegar a la menopausia.

Si los pechos de una embarazada disminuyen súbitamente, es señal de aborto.

Los que se enferman de tétanos mueren al cuarto día. Si logran pasar el cuarto día, sanan.

En todas las estaciones del año aparecen enfermedades de todo tipo, pero algunas son más frecuentes y graves en unas estaciones que en otras.

Los obesos tienen más posibilidad de sufrir muerte súbita que los delgados.

Se debe preferir un alimento y una bebida agradable aunque menos sanos, a un alimento y una bebida más sanos pero menos agradables.

La aplopejía ataca entre los cuarenta y sesenta años de edad.

Hipócrates (Circa 460-395 A de C)