# Investigación clínica XX
# Del juicio clínico a la regresión logística múltiple

Ricardo Berea-Baltierra,[a] Rodolfo Rivas-Ruiz,[b] Marcela Pérez-Rodríguez,[b] Lino Palacios-Cruz,[c] Jorge Moreno,[d] Juan O. Talavera[b]

**Clinical research XIX. From clinical judgment to multiple logistic regression model**

The complexity of the causality phenomenon in clinical practice implies that the result of a maneuver is not solely caused by the maneuver, but by the interaction among the maneuver and other baseline factors or variables occurring during the maneuver. This requires methodological designs that allow the evaluation of these variables. When the outcome is a binary variable, we use the multiple logistic regression model (MLRM). This multivariate model is useful when we want to predict or explain, adjusting due to the effect of several risk factors, the effect of a maneuver or exposition over the outcome. In order to perform an MLRM, the outcome or dependent variable must be a binary variable and both categories must mutually exclude each other (i.e. live/death, healthy/ill); on the other hand, independent variables or risk factors may be either qualitative or quantitative. The effect measure obtained from this model is the odds ratio (OR) with 95 % confidence intervals (CI), from which we can estimate the proportion of the outcome's variability explained through the risk factors. For these reasons, the MLRM is used in clinical research, since one of the main objectives in clinical practice comprises the ability to predict or explain an event where different risk or prognostic factors are taken into account.

| Keywords | Palabras clave |
|---|---|
| Logistic models | Modelos logísticos |
| Causality | Causalidad |
| Biomedical research | Investigación biomédica |

**Recibido:** 05/02/2014 **Aceptado:** 10/02/2014

Desde que el ser humano tuvo conciencia de los peligros y oportunidades que a su alrededor podrían suceder, se originó la necesidad de determinar qué factores podrían estar relacionados y, finalmente, ser predictores o favorecedores; sobre todo, se interesó en aquellos que podrían ser controlados o modificados. Finalmente, en nuestra introspección y a través de nuestra capacidad de observación y de evaluación, los seres humanos nos dimos cuenta de que la mayor parte de los fenómenos que analizamos y muchas veces estudiamos están determinados por distintos factores distales y proximales, que en conjunto originan la presencia o ausencia de algún resultado.

Uno de los principales objetivos de la práctica clínica es predecir o explicar un evento en el que se tomen en cuenta diferentes factores de riesgo o pronósticos, de tal forma que nos permita estimar con un mayor grado de certeza un diagnóstico o un suceso a futuro, por ejemplo, al tratar de establecer la probabilidad de neumonía nosocomial en pacientes de reciente ingreso hospitalario a partir de múltiples variables, como antecedentes de diabetes mellitus, enfermedad pulmonar obstructiva crónica (EPOC), discapacidad funcional, edad del paciente, etcétera. Otro ejemplo de este carácter multifactorial en la predicción de un fenómeno consistiría en intentar predecir en un grupo de adolescentes con obesidad, a 10 años, cuáles son los factores más importantes (edad, actividad física, patrón de dieta, antecedentes hereditarios, etcétera) para el desarrollo de diabetes mellitus. En estos casos, si intentáramos aproximarnos matemáticamente, una alternativa adecuada podría ser aplicar el modelo de regresión logística múltiple.

En este modelo, la variable dependiente tiene dos valores posibles: la ausencia o la presencia de una característica (desarrollo o no de neumonía, diabetes presente o ausente). A su vez, las variables independientes o predictoras pueden ser continuas (edad), ordinales (estadio de Tanner, actividad física leve, moderada o intensa) o dicotómicas (discapacidad funcional, presente o ausente). Como ya lo mencionamos, en la vida diaria como en la práctica clínica, en pocas ocasiones existe una causa única para desarrollar una enfermedad o explicar un fenómeno; casi siempre coexisten múltiples factores de riesgo que hacen a un sujeto proclive a desarrollar o no una enfermedad.

## Modelo de regresión logística múltiple

En el modelo de regresión logística múltiple se busca explicar o predecir la probabilidad de que ocurra o no un evento, el cual se identifica como la variable dependiente o $Y$; se utiliza la ecuación de regresión, en la que se conoce a las variables $X$ ($X_1$, $X_2$,… $X_k$) como

**Resumen**

La complejidad del fenómeno de causalidad en la práctica clínica implica que el resultado de una maniobra no se deba únicamente a esta, sino a la interacción con otros factores del estado basal o variables que ocurran durante la maniobra. Esto requiere diseños metodológicos que permitan evaluar estas variables. Cuando el resultado es dicotómico, se usa la regresión logística múltiple (RLM). La RLM es un modelo multivariado útil cuando se requiere predecir o explicar, al ajustar por el efecto de distintos factores de riesgo, el efecto de una maniobra o exposición sobre el desenlace. Para realizar la RLM se requiere que el desenlace (o la variable dependiente) sea dicotómico y mutuamente excluyente (por ejemplo, vivo/muerto, enfermo/sano); las variables independientes o factores de riesgo pueden ser cuantitativas o cualitativas. La asociación que se obtiene es la razón de probabilidades, también llamada razón de momios (RM), con intervalos de confianza (IC) del 95 % y con estas medidas se estima el porcentaje de la variabilidad del desenlace que se explica a partir de los factores de riesgo. Por estas razones, este modelo es el más usado en la investigación clínica, ya que uno de los principales objetivos de la práctica clínica es poder predecir o explicar un evento en el que se tomen en cuenta diferentes factores de riesgo.

variables independientes y se corresponden con las variables predictivas.

Este es un modelo multivariado muy popular y atractivo para resolver preguntas de investigación clínica. Una de sus características más importantes es que puede incluir variables predictivas de todos los tipos, es decir, continuas, ordinales o categóricas (figura 1).

Ahora bien, si en el análisis de nuestras variables de estudio decidimos utilizar la regresión logística múltiple, debemos tomar en cuenta los siguientes supuestos:

Las $X_i$ son variables no aleatorias (fijas), es decir, son posibles predictoras de *Y*. La relación de las variables debe ser clara y permitir que se establezcan los factores de riesgo que puedan producir la enfermedad o el desenlace (*Y*). Es importante tener en cuenta que no deben utilizarse variables en los casos en los que no se conozca su relación con la enfermedad (*Y*).

La variable dependiente (*Y*) debe ser dicotómica y mutuamente excluyente, por ejemplo: neumonía (tiene o no tiene) o estado al final del tratamiento (vivo o muerto).

Los valores de *X* son independientes y no dependen de la presencia de *Y*; por ejemplo, la presencia de neumonía nosocomial (*Y*) no es un factor de riesgo para el desarrollo de diabetes mellitus o de EPOC (*X*).

Una vez que conocemos que nuestra variable dependiente es dicotómica, se le codifica como 0 o 1. La probabilidad (*odd* en inglés) de que el evento suceda entre la probabilidad de que no suceda se representa con la ecuación *Y / (1-Y)*. El modelo logístico es formado por el logaritmo natural (*ln*) de esta probabilidad, es decir, *ln [Y/ (1-Y)]*. Al igual que en el cálculo de los intervalos de confianza de 95 % de la razón de probabilidades o de momios (RM), utilizaremos el *ln* con el fin de normalizar esta variable dicotómica (el desenlace).

El formato básico de la ecuación de la regresión logística es similar al del modelo de regresión líneal:

$$ln\,[Y/(1-Y)] = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Donde *Y* es el logaritmo natural (*ln*) de las probabilidades (*odds*), *a* es la constante de regresión, e *Y* es

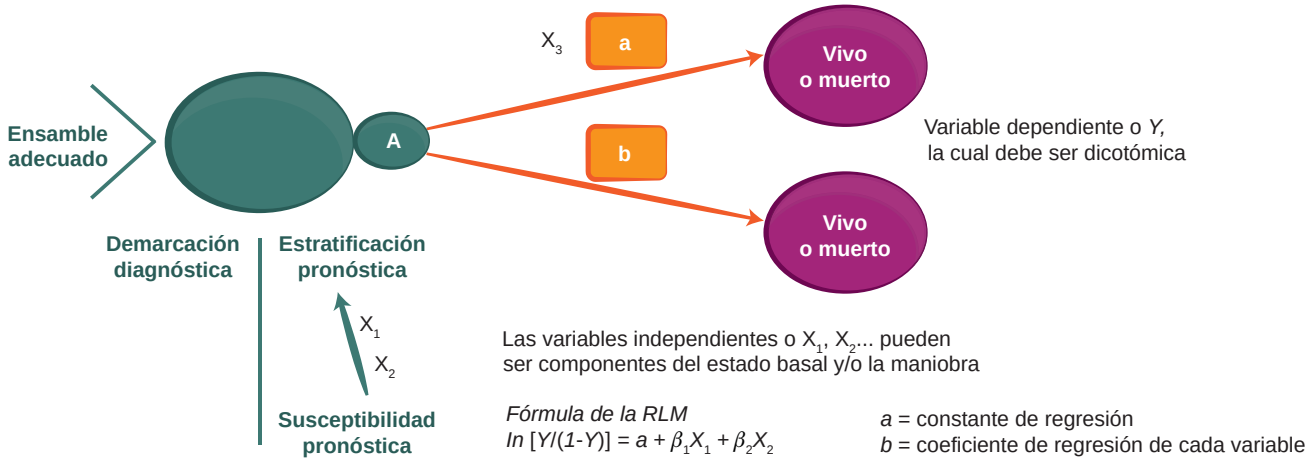


**Figura 1** Modelo de regresión logística múltiple dentro del modelo arquitectónico de la investigación

el valor de *Y* cuando *X* es igual a cero, lo que significa la probabilidad de la enfermedad cuando no se considera ningún factor de riesgo (y se corresponde con la prevalencia o incidencia de la enfermedad), por ejemplo, la prevalencia de la diabetes mellitus en la población general (16 %). Asimismo, $\beta$ es el coeficiente de regresión de cada variable y se refiere a la magnitud de cambio en *Y* por unidad de cambio en *X*, sin que esto signifique causalidad.

Retomando el ejemplo inicial, supongamos que queremos predecir la posibilidad de que se presente neumonía nosocomial en pacientes de una clínica geriátrica y los catalogamos como sin neumonía (codificados como 0) o con neumonía nosocomial (codificados como 1). Idealmente, en la práctica clínica ubicamos las características basales que al ingreso del paciente pudieran predecir el riesgo de presentar dicha complicación, a fin de tomar medidas preventivas adecuadas. Para esto, se interroga su edad, si tiene otras enfermedades (EPOC, DM, etcétera) o el estado funcional medido con la escala de Karnofsky.

Lo mismo sucede en el análisis de regresión logística múltiple, en el que se recomienda utilizar las variables que en análisis bivariados previos se hubieran mostrado con al menos cierto grado de relación como factores de riesgo ($p < 0.25$ o menor). Debe quedar claro que en el modelo final se pueden incluir todas las variables que han demostrado relación con el desenlace, incluso variables que en el análisis bivariado de nuestros datos no hayan resultado significativas.

También es deseable que cada variable en el modelo cuente con un mínimo de 10 a 20 casos/individuos por cada evento en el menor de los grupos (se consideran dos grupos, el grupo de los sujetos que desarrollan el desenlace y el grupo de los sujetos que no lo desarrollan).

Tomando nuestro ejemplo, si la variable dependiente es neumonía y tuvimos 47 eventos en 121 pacientes, no deberíamos incluir en nuestro modelo de regresión más de cuatro variables (47/10 = 4.7). Esta estrategia se utiliza para estabilizar los datos y es llamada comúnmente como *eventos por variable* (*event per variable* o *EPE value*). En caso de incluir más variables, los datos pueden volverse inestables y podemos obtener resultados estadísticamente significativos, sin que en realidad lo sean.

Las preguntas en nuestro ejemplo serían: ¿qué características de los pacientes predicen la aparición de neumonía?, ¿cuál es la probabilidad de que un individuo que ingresa a la clínica geriátrica presente neumonía nosocomial, dada la presencia combinada de dichos factores?

La realización de un modelo de regresión logística múltiple es un proceso complejo, motivo por el que es necesario apoyarnos en programas estadísticos como SPSS, SAS o Stata, entre otros. En este ejemplo, nos apoyaremos en el programa estadístico SPSS para realizar el análisis.

El primer paso es hacer el análisis bivariado de los datos con el que contrastamos cada una de las posibles variables involucradas contra la variable de desenlace (aparición de neumonía). Las variables dicotómicas (EPOC y diabetes mellitus) se contrastan con $\chi^2$ (capítulo XVII de esta serie), se busca su medida de asociación con el cálculo de la razón de momios (o exponente de β: Exp[β]) y el intervalo de confianza (IC 95 %) (capítulo VI de esta serie), mientras que la edad se contrasta con *t* de Student para grupos independientes (capítulo XV de esta serie).

Una vez que identificamos las variables que se asocian, procederemos a realizar el modelo multivariado (preseleccionar las variables que se van a incluir a través de un análisis bivariado es solo una estrategia, pero bien pueden incluirse todas las variables) e intentar definir cuál de estas variables se relaciona de modo independiente y cuál no.

Como primer paso para realizar la regresión logística múltiple en el programa SPSS, seleccionaremos en la barra de herramientas la opción *Analizar*, posteriormente la pestaña de *Regresión* y finalmente la opción *Logística binaria* (figura 2).

En el cuadro *Dependientes* colocaremos nuestra variable dependiente, en este caso la presencia o no de neumonía. En la sección *Covariables* agregaremos las variables independientes que queremos utilizar para nuestro modelo de predicción. En *Método* asignamos cómo queremos que se lleve a cabo el análisis: *Introducir* (analiza todas las variables incluidas y las deja en el modelo final aun cuando no sean estadísticamente significativas), *Hacia adelante* (automáticamente se irán agregando al modelo una a una las variables significativas estadísticamente) o *Hacia atrás* (del total de variables se irán eliminando automáticamente las que no contribuyan al modelo por falta de significación estadística). En este ejemplo utilizaremos el método *Hacia atrás*.

Posteriormente, tenemos que especificarle al programa las variables que son categóricas, especificar el tipo de contraste (en este caso, simple, en el que cada categoría —ordinal o dicotómica— se contrasta contra la categoría de referencia) y poner de acuerdo con lo observado si la categoría de referencia es la primera o la última (el orden afectará la manera como se representa la RM en la tabla de resultados, ya sea como factor de riesgo o de protección, pero no modifica el resultado del modelo).

Al realizar el análisis, como primer dato se presenta una tabla de clasificación aún sin valores pronosticados y que establece el porcentaje de pacientes que no presentaron neumonía (61.2 %).

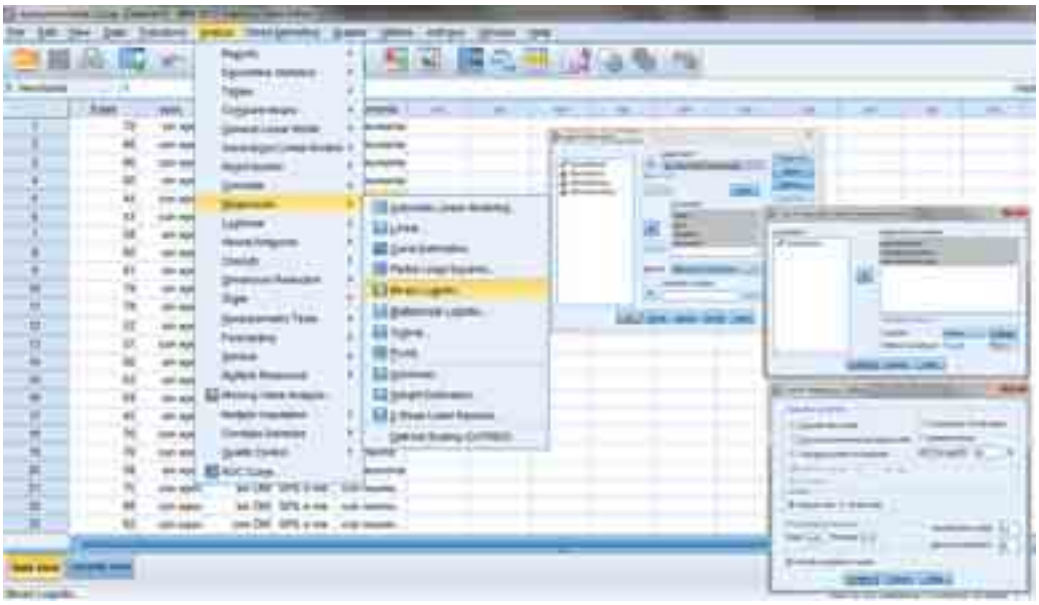

**Figura 2** Interfaz de inicio para regresión logística en el programa estadístico SPSS

En la figura 3 se muestran las variables de la ecuación. Tenemos las columnas del coeficiente de regresión o *B* (indica el incremento en la variable dependiente con cada unidad de la variable independiente), de la significación estadística y de la RM o Exp(β) (indica la probabilidad de desarrollar el evento), con su respectivo intervalo de confianza del 95 %. En este ejemplo tenemos que el logaritmo de posibilidades disminuiría –0.022 con cada año de edad (es decir, a menor edad, mayor probabilidad de desarrollar neumonía durante la hospitalización), aunque al ver que el valor de *p* es > 0.05 y que el intervalo de confianza del Exp(β) o la RM atraviesa la unidad, notamos que carece de significación estadística. Por otro lado, tenemos el caso de la EPOC: en los pacientes que tienen esta patología, sus probabilidades aumentan 2.203 (Exp[B]), lo que en la clínica se traduce como RM = 9.05 (IC 95 % = 3.44-23.76). Por lo tanto, se puede interpretar esta RM como que los que tienen EPOC tienen 8.05 veces más riesgo de tener neumonía en comparación con los que no la padecen (capítulo VI de esta serie) y que esta asociación es estadísticamente significativa, y se relaciona independientemente de la edad, de la presencia de diabetes mellitus y del Karnofsky.

En este primer paso encontramos que ni la edad, ni el antecedente de tener DM se relacionan con el desarrollo de la neumonía nosocomial y que las variables que se relacionan son tener EPOC y el estado funcional medido con el Karnofsky (figura 3).

En un segundo paso de la regresión logística se pueden eliminar variables con las que no se explique la variable dependiente. Esto se hace al eliminar la variable con mayor valor de *p*, en este caso la presencia de diabetes ($p = 0.611$), por lo que los resultados se pueden ver en la tabla de clasificación y los resultados de variables de la figura 4. Con esto se mejora el porcentaje global de efectividad a 70.2 %. Nosotros recomendamos no eliminar las variables clínicas que tengan lógica biológica (en este ejemplo, la edad), con el fin de ajustar el modelo.

Variables de la ecuación

| | | B | E.T. | Wald | Sig. | Exp(β) | IC 95% para Exp(β) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Inferior | Superior |
| Paso 1 | Edad | -0.022 | 0.019 | 1.418 | 0.234 | 0.978 | 0.943 | 1.014 |
| | EPOC(1) | 2.203 | 0.492 | 20.027 | 0.000 | 9.053 | 3.449 | 23.761 |
| | Diabetes(1) | .301 | 0.593 | 0.258 | 0.611 | 1.351 | 0.423 | 4.318 |
| | Karnofsky(1) | 1.933 | 0.677 | 8.157 | 0.004 | 6.912 | 0.834 | 26.049 |
| | Constante | 1.429 | 1.182 | 1.461 | 0.227 | 4.175 | | |

**Figura 3** Resultados del modelo de regresión logística múltiple, en el que la variable dependiente es el desarrollo de neumonía nosocomial

Por ejemplo, si realizamos un tercer paso y se procede a eliminar la variable con *p* mayor, en este caso la variable edad ($p = 0.195$), se logra un porcentaje global de explicación de 71.1 % (figura 5). Sin embargo, este no es clínicamente mayor al modelo anterior que predecía un 70 %, pero incluía una variable que clínicamente resultaba muy significativa como la edad, y que en estudios previos había demostrado impactar en el desenlace. Si en este estudio no resultó significativa, habrá que ver si tenemos un grupo con una edad muy compacta o si no incluimos grupos de edades similares a los estudios en los que sí resultó significativa.

Tabla de clasificación[a]

| Observado | Pronosticado | | |
|---|---|---|---|
| | Neumonía | | |
| | sin neumonía | con neumonía | % correcto |
| Paso 2 | 51 | 23 | 68.9 |
| | 13 | 34 | 72.3 |
| | | | 70.2 |

[a] El valor de corte es 0.500

Variables en la ecuación

| | | B | E.T. | Wald | Sig. | Exp(β) | IC 95% para Exp(β) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Inferior | Superior |
| Paso 2 | Edad | -0.024 | 0.018 | 1.681 | 0.195 | 0.976 | 0.942 | 1.012 |
| | EPOC(1) | 2.203 | 0.492 | 20.053 | 0.000 | 9.055 | 3.452 | 23.753 |
| | Karnofsky(1) | 1.925 | 0.674 | 8.148 | 0.004 | 6.852 | 1.828 | 25.686 |
| | Constante | 1.624 | 1.115 | 2.121 | 0.145 | 5.075 | | |

**Figura 4** Segundo paso de la regresión logística

Con este modelo podríamos decir que los pacientes con EPOC y Karnofsky de 70 % tienen mayor riesgo de presentar neumonía durante una hospitalización, lo cual puede tener implicaciones en el pronóstico (o en el manejo) al ser considerados como pacientes con mayor riesgo que los que no tienen estas características clínicas.

## Resumen del modelo

En los resultados de la prueba de regresión logística se nos presentará el resumen del modelo (figura 6).

El valor de menos dos veces el logaritmo (–2 log) de la verosimilitud, o también llamado desviación, indica hasta qué punto el modelo se ajusta bien a los datos (en él se considera que cuanto más pequeño es el valor, mejor es el ajuste). 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 independientes (ver capítulo VI de esta serie). En este caso, el valor es de 0.213, lo cual significa que el 21 % de la variación de la variable dependiente (neumonía) se explica con las variables independientes incluidas en el modelo (edad, capacidad funcional y EPOC en este ejemplo). El $R^2$ de Nagelkerke es una versión corregida del $R^2$ de Cox y Snell.

Tabla de clasificación[a]

| Observado | Pronosticado | | |
|---|---|---|---|
| | Neumonía | | |
| | sin neumonía | con neumonía | Porcentaje correcto |
| Paso 3 | 45 | 29 | 60.8 |
| | 6 | 41 | 87.2 |
| | | | 71.1 |

[a] El valor de corte es 0.500

Variables en la ecuación

| | | B | E.T. | Wald | Sig. | Exp(β) | IC 95% para Exp(β) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Inferior | Superior |
| Paso 3 | EPOC(1) | 2.136 | 0.484 | 19.462 | 0.000 | 8.469 | 3.278 | 21.881 |
| | Karnofsky(1) | 2.076 | 0.668 | 9.655 | 0.002 | 7.972 | 2.152 | 29.532 |
| | Constante | 239 | 0.301 | 0.627 | 0/428 | 1.270 | | |

**Figura 5** Tercer paso de la regresión logística

## Prueba de Hosmer-Lemeshow

En la presentación final de los datos de regresión logística es deseable que figure algún tipo de bondad de ajuste, como la de Hosmer-Lemeshow. Esta prueba analiza, de acuerdo con deciles de riesgo, la presencia del evento contra la frecuencia esperada (probabilidad de neumonía < 10 %, < 20 % y hasta 100 %). Ambas distribuciones, la esperada y la observada, se contrastan con una prueba de $\chi^2$ (figura 7).

Dado lo anterior, podemos inferir que no existe diferencia estadística entre la distribución esperada y la que predice nuestro modelo, por lo que se puede considerar como adecuado para establecer el riesgo de ocurrencia del evento de interés.

Resumen del modelo

| Paso | -2 log de la verosimilitud | R cuadrado de Cox y Snell | R cuadrado de Nagelkerke |
|---|---|---|---|
| 1 | 132.492 | 0.214 | 0.291 |
| 2 | 132.754 | 0.213 | 0.288 |
| 3 | 134.476 | 0.201 | 0.273 |

**Figura 6** Resumen del modelo de regresión logística

Prueba de Hosmer-Lemeshow

| Paso | Chi cuadrado | gl | Sig. |
|---|---|---|---|
| 1 | 3.209 | 8 | 0.921 |
| 2 | 3.724 | 8 | 0.881 |
| 3 | 0.327 | 1 | 0.568 |

**Figura 7** Prueba de Hosmer-Lemeshow

## Comentarios

El modelo de regresión logística múltiple es una poderosa herramienta para el análisis multivariado y para ponderar una variable dependiente frente a otras con las cuales pudiera interactuar. Es uno de los modelos que más se parece al pensamiento clínico, en el que se conoce que no existe el modelo unicausal, sino que siempre hay múltiples causas y estas tienen distinto peso.

Cuando interpretamos o usamos la regresión logística múltiple debemos tomar en cuenta los siguientes puntos: *a*) la variable dependiente (*Y*) debe ser dicotómica y mutuamente excluyente; *b*) las variables independientes pueden ser de cualquier tipo, ya sean cualitativas o cuantitativas; *c*) los modelos finales deben diseñarse tanto a partir del análisis bivariado como por lógica biológica.

**Declaración de conflicto de interés:** los autores han completado y enviado la forma traducida al español de la declaración de conflictos potenciales de interés del Comité Internacional de Editores de Revistas Médicas, y no ha sido reportado alguno que esté relacionado con este artículo.

[a]Departamento de Medicina Interna, Hospital de Oncología
[b]Centro de Adiestramiento en Investigación Clínica (CAIC), Coordinación de Investigación en Salud
[c]Subdirección de Investigaciones Clínicas, Instituto Nacional de Psiquiatría "Dr. Ramón de la Fuente Muñiz", Secretaría de Salud
[d]Departamento de Urología, Hospital de Especialidades
[a,b,d]Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social

Distrito Federal, México

Comunicación con: Ricardo Berea-Baltierra
Correo electrónico: ricberbal@hotmail.com

### Referencias

1. Portney LG, Watkins MP. Logistic regression. En: Foundations of clinical research applications to practice. Third edition. New Jersey, USA: Pearson & Prentice Hall; 2009. p. 696-700.
2. Palacios-Cruz L, Pérez M, Rivas-Ruiz R, Talavera JO. Investigación clínica XVIII. Del juicio clínico al modelo de regresión lineal. Rev Med Inst Mex Seguro Soc. 2013;51(6):656-61.
3. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.
4. Dawson B, Trapp RG. Bioestadística médica. México: Manual Moderno; 2005. p. 239-41.
5. Talavera JO, Rivas-Ruiz R, Pérez-Rodríguez M. Investigación clínica VI. Relevancia clínica. Rev Med Inst Mex Seguro Soc. 2011;49(6):631-5.
6. Feinstein AR. Multivariable analysis: An introduction. New Haven: Yale University Press; 1996.
7. Talavera JO. Investigación clínica I. Diseños de investigación. Rev Med Inst Mex Seguro Soc. 2011; 49(1):53-8.
8. Rivas-Ruiz R, Castelán-Martínez OD, Pérez M, Talavera JO. Investigación clínica XVII. Prueba chi cuadrada, de lo esperado a lo observado. Rev Med Inst Mex Seguro Soc. 2013;51(5):552-7.
9. Rivas-Ruiz R, Pérez-Rodríguez M, Talavera JO. Investigación clínica XV. Del juicio clínico al modelo estadístico. Diferencia de medias. Prueba t de Student. Rev Med Inst Mex Seguro Soc. 2013;51(3): 300-3.