***Revista de la Universidad Industrial de Santander. Salud* Vol.48 No.1 Enero - Marzo de 2016**

Regresiones aplicadas al estudio de eventos discretos en epidemiología

Regressions applied to the study of discrete events in epidemiology

Fredi Alexander Diaz-Quijano[1]

Forma de citar: Diaz-Quijano FA.Regresiones aplicadas al estudio de eventos discretos en epidemiología. Rev Univ Ind Santander Salud. 2016; 48(1): 9-15. **DOI:** http://dx.doi.org/10.18273/revsal.v48n1-2016001

RESUMEN

En este manuscrito se revisan algunos aspectos básicos de la utilización de regresiones en los estudios epidemiológicos, haciendo énfasis en aquellas aplicadas al estudio de eventos discretos. De esta manera se hace una introducción a los modelos lineales generalizados, cuya estructura es una extensión de una ecuación lineal para analizar desenlaces discretos. De este modo podemos estimar medidas de asociación como la razón de tasas usando la regresión de Poisson, o bien, el riesgo relativo (o la razón de prevalencias) usando la regresión log-binomial. En cada caso es esencial conocer la naturaleza de la variable dependiente, su distribución y reconocer las limitaciones de cada una de las herramientas de análisis.

Palabras clave: Modelos lineales generalizados, Regresión de Poisson; Regresión Binomial, Razón de tasas, Riesgo Relativo, Razón de Prevalencias.

ABSTRACT

Some basic aspects about using regressions in epidemiological studies are reviewed. Particularly, this manuscript focused on those applied to the study of discrete events. Generalized lineal models, such as Poisson and log-binomial, have a structure that is an extension of a lineal equation to analyze discrete outcomes. Thus, we can estimate association measures as the incidence rate ratio, using the Poisson regression, or the relative risk (or prevalence ratio), using log-binomial regression. In each case it is essential to know the nature of the dependent variable, as well as, its distribution and recognize the limitations of each analysis tool.

Keywords: Generalized Lineal Models; Poisson Regression; Binomial Regression; Incidence Rate Ratio; Relative Risk; Prevalence Ratio.

1. Departamento de Epidemiologia, Faculdade de Saúde Pública, Universidade de São Paulo. Brasil.
Correspondencia: Fredi Alexander Diaz Quijano. Dirección: Departamento de Epidemiologia, Faculdade de Saúde Pública da Universidade de São Paulo, Av. Dr. Arnaldo, 715, Cerqueira César, CEP 01246-904, São Paulo, SP, Brasil. Correo electrónico: frediazq@msn.com. Teléfono: +55 11 3061-7738

Recibido: Noviembre 4 de 2015 **Aprobado:** Diciembre 19 de 2015

INTRODUCCIÓN

Los estudios epidemiológicos se basan en la comparación de grupos que tienen diferentes distribuciones de uno o más factores de riesgo para una enfermedad o resultado de interés. Por lo tanto, para estimar correctamente el efecto de la exposición sobre la ocurrencia de un desenlace, se requieren métodos para controlar el efecto de otros factores de riesgo que actúan como variables de confusión para la asociación de interés[1,2]. Para controlar los factores de confusión podemos emplear estrategias como el diseño del estudio, la exclusión de grupos con categorías de riesgo diferenciales, el pareamiento (matching) y, en el caso de los estudios experimentales, la aleatorización. Sin embargo, en la mayoría de los estudios observacionales el control de la confusión requiere un abordaje durante la fase de análisis de los datos[2,3].

Entre las herramientas analíticas para controlar la confusión tenemos a la estandarización, la cual se usa con frecuencia para controlar el efecto de variables como la edad y el sexo en la comparación de tasas. Otra herramienta de análisis es la estratificación, con la cual se puede controlar el efecto de confusión de variables categóricas. Sin embargo, estas estrategias tienen limitaciones importantes como la dificultad para realizar ajustes por variables continuas y la pérdida de eficiencia para ajustar simultáneamente por un número elevado de variables[3]. El uso de las regresiones ha permitido lidiar con estas dificultades facilitando la estimación de medidas de asociación, ajustadas por múltiples variables de diversa naturaleza. El objetivo de este artículo es hacer una revisión de algunos aspectos básicos de la utilización de regresiones en los estudios epidemiológicos, haciendo énfasis en aquellas empleadas en el análisis de eventos discretos.

Conceptos preliminares

Con el término *evento* nos referiremos a aquellos desenlaces o resultados de interés que pueden ser identificados en participantes en los estudios epidemiológicos. Estos pueden corresponder a enfermedades o episodios que pueden ser identificados como casos nuevos, durante el seguimiento de una cohorte; o bien, como casos prevalentes, en los estudios transversales. En el análisis estadístico, los eventos o resultados corresponden a las variables dependientes. Con el término *evento discreto* se hará referencia a una variable dependiente que corresponde a una dicotómica (por ejemplo, enfermos vs no enfermos) o, si es cuantitativa, correspondería a una variable que sólo puede adoptar valores enteros (por ejemplo, número de hospitalizaciones). Las variables discretas se diferencian de las continuas, pues en estas últimas siempre se puede hallar un valor intermedio entre dos valores posibles (por ejemplo, peso, presión arterial, glucosa en sangre, etcétera).

REGRESIONES

La regresión es una herramienta que permite describir una relación entre variables de tal manera que, si hay una asociación entre ellas, se pueda predecir (con algún margen de error) el valor de la variable dependiente dado que se conoce el valor de al menos una variable independiente[4-6]. Una regresión lineal simple es una técnica estadística que evalúa si la relación entre dos variables es lineal[5,6]. En el caso de regresión lineal al menos la variable dependiente es cuantitativa. Por lo tanto, la relación lineal entre la variable dependiente "Y" y una independiente "X_1" podría resumirse con la fórmula:

$$Y = \beta_0 + \beta_1 X_1 + e$$

donde β_0 sería la intercepción o el valor esperado de la variable cuando la variable dependiente es igual a cero. Por otra parte, β_1 sería el coeficiente de la regresión para la variable independiente X_1 y se interpreta como el aumento esperado de la variable dependiente "Y", con cada aumento de la variable independiente "X_1". El término "e" de la fórmula se refiere al error en la predicción, el cual se asume como aleatorio. Sin este término de error no podríamos definir el valor exacto de Y en la fórmula y sólo podríamos hablar del valor esperado o predicho. Este último podría ser representado por $\hat{Y}$ llevando implícito la aceptación de la existencia del error. Entonces tendríamos la fórmula:

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

La Figura 1, presenta un ejemplo de distribución de individuos en relación a dos variables: "X" y "Y". En este caso, una regresión puede resumir la relación de estas variables con la ecuación:

$$\hat{Y} = -1{,}09 + 0{,}53X$$

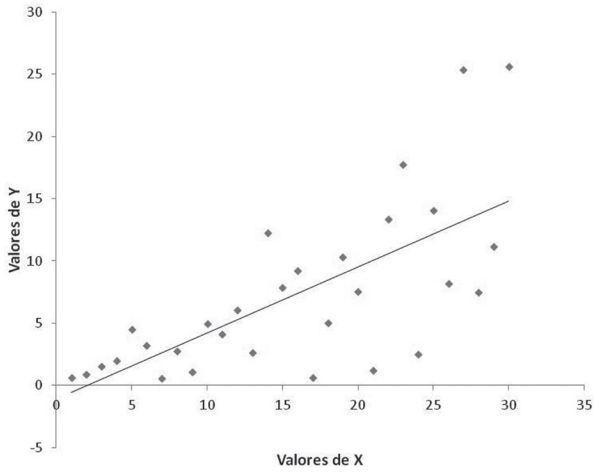

Figura 1. Regresión lineal de Y en X.

Regresión lineal múltiple

El modelo de regresión múltiple es una extensión para diversas variables del modelo de regresión simple. Esto se aplica cuando existe más de una variable independiente[3,7]. En este caso, podemos construir la ecuación:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$$

Para cada variable independiente X_i el modelo considera un coeficiente de regresión β_i. Este coeficiente es interpretado como el cambio esperado en la variable dependiente, por un cambio de una unidad en la variable independiente correspondiente (X_i : X_1, X_2, X_3 o X_k), siempre y cuando se mantengan constantes las demás variables independientes. En síntesis, la regresión lineal analiza una variable dependiente de naturaleza continua y ésta es modelada directamente como el resultado de una fórmula lineal. En este caso, la medida de asociación obtenida y de interés en epidemiología es el cambio (o delta) en el valor predicho de dicha variable dependiente.

Modelos lineales generalizados

Los modelos lineales generalizados (MLG, o GLM por sus siglas en inglés) son una extensión de la estructura de la regresión lineal ordinaria con la que se pretende analizar variables dependientes cuya distribución del error es diferente de la normal. Los MLG fueron formulados por John Nelder y Wedderburn Robert como una forma de unificar varias técnicas estadísticas incluyendo las regresiones logística y de Poisson[8]. Un MLG permite relacionar un conjunto de variables independientes a través de una ecuación similar al modelo lineal pero utilizando una función de enlace (*link function*) para predecir la variable dependiente. Como consecuencia, la magnitud de la varianza estimada es una función de su valor predicho[9,10]. Los MLG se pueden expresar como una ecuación similar a la de una regresión lineal pero el resultado es una versión transformada de la variable dependiente, usualmente el logaritmo natural (Ln) de la misma. En este caso, la medida estadística de la asociación es el cambio (delta) predicho en el Ln de la variable dependiente (Tabla 1).

Tabla 1. Características de modelos de regresión frecuentemente usados en epidemiología.

Regresión	Tipo de Variable dependiente	Función de enlace (Link)	Resultado estadístico	Medida Epidemiológica
Linear	Continua	Identidad	Delta promedio	Delta promedio
Poisson	Conteo	Ln (Taxa [T] o conteo)	Delta Ln(T)	Razón de Tasas
Binomial	Dicotómica	Ln (Proporción[P])	Delta Ln(P)	Riesgo Relativo o Razón de Prevalencias
Logística	Dicotómica	Ln (Odds)*	Delta Ln(Odds)	Razón de Chances (Odds Ratio)

* La función de enlace de la regresión logística es conocida como *logit* y corresponde al Ln del Odds = $Ln\left(\frac{p}{1-p}\right)$

Sin embargo, para su utilización en epidemiología, este resultado suele convertirse obteniendo el anti-logaritmo correspondiente. Para obtener los estimados estos modelos suelen utilizar el método de máxima verosimilitud, el cual es el procedimiento estándar en la mayoría de los programas estadísticos[9-11]. De forma general, podemos decir que este método busca, mediante iteraciones sucesivas, los coeficientes que construyen el modelo cuyos valores predichos sean lo más cercano posible a los valores observados en la muestra analizada. Entre los MLG, la regresión logística ha tenido un destaque especial y ha sido ampliamente descrita en textos didácticos[12,13]. Por lo anterior, en lo sucesivo de este manuscrito haremos énfasis en la regresión de Poisson y la binomial (o log-binomial).

DOI: http://dx.doi.org/10.18273/revsal.v48n1-2016001

REGRESIÓN DE POISSON

La distribución de Poisson describe la frecuencia esperada de un conjunto de probabilidades para una variable discreta[13]. En cada punto de esta distribución, se representa la probabilidad de que un determinado número de eventos ocurra durante un periodo de tiempo en un espacio o población especificada (Figura 2). Esta distribución se aplica principalmente en el estudio de eventos con probabilidades muy pequeñas o "eventos raros".

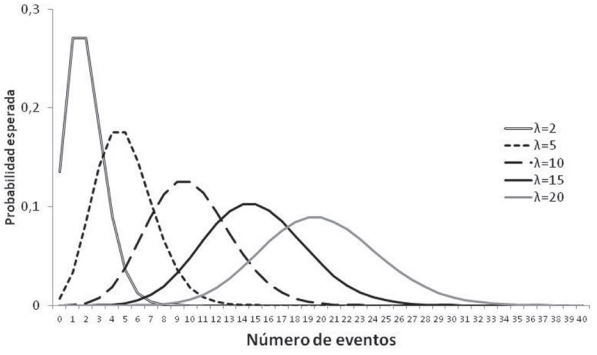

Figura 2. Distribución del número de eventos esperados según la media (λ) de una variable tipo Poisson

De esta forma, una variable tipo Poisson consiste en el número de veces que un acontecimiento ocurre en un tiempo, espacio y población definida. En consecuencia, estas variables se presentan como tasas de un evento de interés. Por ejemplo, el número de muertes por cáncer gástrico en personas entre 50 a 70 años en una población específica, podrá expresarse como una tasa por millón de habitantes en ese grupo etario, durante el periodo de observación. En este caso, la unidad de observación estaría delimitada en términos de tiempo, grupo poblacional y lugar. Como características importantes de este tipo de variables, se destacan las siguientes:

a) No adopta valores negativos, pues el conteo de eventos siempre tendrá resultados superiores o iguales a cero. De lo anterior se deriva que la suma de resultados nunca disminuye.
b) Las tasas suelen ser bajas y se asumen como constantes dentro de cada unidad de observación.
c) Independencia de eventos, es decir, que la adición de un evento en una unidad de observación no depende del número pasado o presente de eventos.
d) En una observación dada, el promedio de la variable Poisson es igual a su varianza.

La figura 2 presenta una simulación de la distribución de las probabilidades del número de eventos de una variable tipo Poisson, según cinco posibles valores del promedio (media) de la misma. En ésta se aprecian características de este tipo de variables, tales como la ausencia de valores negativos y el incremento de la variabilidad conforme aumenta el parámetro de la media (λ) del número de eventos. Debido a las características de este tipo de variables, es inadecuado analizar una variable de conteo de eventos utilizando una regresión lineal. Entre las razones está el hecho de que la regresión lineal podría predecir valores negativos, los cuales carecerían de sentido para una variable basada en un conteo de eventos. Además, la regresión lineal asume que la varianza es constante para todos los valores de la variable dependiente, esto traería pérdida de precisión y de eficiencia en las estimaciones. Por lo anterior, contamos con herramientas estadísticas para este tipo de variables. Específicamente, la regresión de Poisson está diseñada para definir una ecuación lineal cuyo resultado directo es el Ln del conteo de eventos. En consecuencia, un modelo resultante puede definirse por la siguiente fórmula:

$$Ln(T) = \beta_0 + \beta_1 X_1$$

donde $Ln(T)$ es el logaritmo natural de la tasa o del número de eventos por unidad de observación y los otros términos son análogos a los del modelo lineal. Considerando variables independientes adicionales (X_1, X_2, X_3, ..., X_k), tendríamos la siguiente fórmula para representar un modelo de Poisson múltiple:

$$Ln(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$$

La regresión de Poisson asume que los eventos son independientes en el sentido de que la ocurrencia de uno de ellos no hace más o menos probable la ocurrencia de otro[14,15]. Sin embargo, la probabilidad por unidad de observación puede estar relacionada con las características de la misma, es decir con las variables independientes que la definen.

Aplicaciones de la regresión de Poisson

Con un modelo de Poisson, podemos predecir una tasa a partir de la presencia o ausencia de exposiciones. En la forma más simple, consideremos un modelo de Poisson con una variable independiente (X_1) con la siguiente ecuación:

$$Ln\,(T) = \beta_0 + \beta_1 X_1$$

Donde X_1 adopta el valor de cero "0" en el grupo no expuesto y el valor de 1 en el grupo expuesto. Con esta fórmula podemos estimar la tasa predicha para cada uno de los grupos de exposición:

- Tasa en no expuestos a X_1
  $(T_0) \rightarrow Ln\ (T_0) = \beta_0 + \beta_1 * 0 = \beta_0 \rightarrow T_0 = \boldsymbol{e^{\beta_0}}$
- Tasa en expuestos a X_1
  $(T_1) \rightarrow Ln\ (T_1) = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1 \rightarrow T_1 = \boldsymbol{e^{(\beta_0+\beta_1)}}$

La predicción de tasas se extiende para los modelos múltiples, en los que se puede calcular la tasa esperada según diferentes niveles de exposición para múltiples variables independientes.

Pero la aplicación más común de los modelos de regresión de Poisson es la estimación de tasas relativas o razones de tasas (RT), bien sea crudas o ajustadas por diversas variables independientes. En este caso, la RT puede estimarse como una razón de funciones teniendo en el numerador la ecuación para la categoría considerada de exposición y en el denominador la ecuación correspondiente al grupo de referencia (o no expuesto). Así por ejemplo, la RT para una variable $X_1 (RT_{x_1})$ en un modelo múltiple, se estimaría de la siguiente forma:

$$RTx_1 = \frac{T_{expuestos}(X_1=1)}{T_{no\ expuestos}(X_1=0)} = \frac{e^{\beta_0} e^{\beta_1(X_1=1)}\, e^{\beta_2 X_2}\, e^{\beta_3 X_3}\, e^{\beta_k X_k}}{e^{\beta_0} e^{\beta_1(X_1=0)}\, e^{\beta_2 X_2}\, e^{\beta_3 X_3}\, e^{\beta_k X_k}}$$

Simplificando la anterior expresión tenemos que: $RT_{x_1} = e^{\beta_1}$

De forma genérica, podemos decir que la RT para una variable independiente X_i es el antilogaritmo de su coeficiente correspondiente β_i. De esta forma, se calcularía la medida de asociación que sería el factor por el cual se multiplicaría la tasa al cambiar de categoría, desde un estado de no expuesto a uno de expuesto. En el caso de las variables independientes cuantitativas, la RT indicaría el factor por el que se multiplica la tasa con cada aumento en una unidad en la escala de exposición.

Magnitud de la población expuesta

La regresión de Poisson permite modelar el conteo de eventos en relación a una determinada unidad de observación, es decir, permite analizar dicho conteo dividido por alguna medida de la unidad de exposición (exposure)[10,14,15]. Por ejemplo, los biólogos podrían estar interesados en analizar el número de especies de árboles por unidad área de bosque, en este caso la unidad de observación podría ser espacial (kilómetro cuadrado). En epidemiología, con frecuencia utilizamos como denominador a la población expuesta durante un periodo de tiempo. De esta manera, podemos calcular las tasas de un evento como (muertes por cáncer u hospitalizaciones por dengue) utilizando el denominador poblacional de *personas-año*.

Con frecuencia este denominador puede variar entre las unidades de observación, bien sea por diferencias en el número de personas expuestas o en el tiempo de observación. En la regresión de Poisson, estas diferencias en el denominador de las tasas pueden modelarse pasando la medida de exposición al lado derecho de la ecuación. De esta manera, si consideramos que la tasa (T) es una razón entre el conteo de eventos (n) y la magnitud de la exposición (Exp), entonces tenemos que:

$$Ln\ (T) = Ln\left(\frac{n}{Exp}\right) = Ln(n) - \boldsymbol{Ln(Exp)}$$

Por lo anterior, la regresión de Poisson, puede modelar el logaritmo del número de eventos por unidad de exposición definiendo la siguiente ecuación:

$$Ln(n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \boldsymbol{Ln(Exp)}$$

Sobredispersión y cero-inflación

Una característica de la distribución de Poisson es que la tasa promedio y su varianza son iguales. Sin embargo, en ciertas circunstancias la varianza es mayor, fenómeno conocido como sobredispersión (overdispersion)[10,14,15]. Lo anterior sugiere que el modelo no es apropiado y suele llevar a que la regresión de Poisson subestime el error estándar, dando lugar a valores de p sesgados (demasiado pequeños) e intervalos de confianza muy estrechos.

Una razón común es la omisión de variables explicativas o de observaciones relevantes. Por lo anterior, lo primero que debe hacerse es revisar los datos y las variables consideradas en el análisis. Sin embargo, el problema de sobredispersión puede persistir y, en algunas circunstancias, puede resolverse usando otros modelos como el regresión binomial negativa[10,15]. Esta última regresión introduce un término (alfa: α) para modelar la sobredispersión, con lo que se podría alcanzar un modelo más ajustado a los valores observados.

Otro problema común con la regresión de Poisson es el exceso de ceros. Este problema se vuelve relevante cuando el exceso de ceros es causado porque en realidad hay dos procesos subyacentes: 1) un fenómeno que determina la presencia de un cero frente a un valor positivo; y 2) una vez que se alcanza un valor positivo, hay otro fenómeno que determina el conteo de eventos que se producen. Un ejemplo clásico sería la distribución de cigarrillos fumados en una hora por miembros de un grupo en el que algunos individuos no son fumadores. En estos casos, la regresión binomial negativa también podría corregir este problema, aunque también existen alternativas como la regresión cero-inflada, que permite modelar los fenómenos mencionados[15,16].

REGRESIÓN BINOMIAL

La distribución binomial se observa en las variables dicotómicas, donde el resultado sólo puede adoptar uno de dos valores posibles (por ejemplo, enfermar vs permanecer sano, o bien, morir vs sobrevivir). En estos casos, la regresión binomial (o log-binomial) modela el logaritmo de una proporción (p), que bien podría ser la incidencia o la prevalencia de una enfermedad[17-19]. El modelo tendría una estructura análoga a la de otros MLG:

$$Ln(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$$

A partir de este modelo, podemos obtener las medidas de asociación entre cada variable independiente y el desenlace de interés de la misma forma como obtiene con el modelo de Poisson, es decir, como el antilogaritmo del coeficiente correspondiente. De esta manera, para la variable independiente X_1, la medida de asociación sería: e^{β_1}. En este caso la medida de asociación sería el riesgo relativo (RR), si los datos fueron recolectados en estudios de cohorte, es decir si la proporción es una incidencia. Por otra parte, si el estudio corresponde a un corte transversal y la proporción es una prevalencia, la medida de asociación sería una razón de prevalencia (RP).

Log-binomial vs logística

Tanto la regresión log-binomial como la logística modelan un desenlace dicotómico utilizando una transformación de la variable dependiente basada en logaritmos. Sin embargo, la regresión logística consiste en la definición de un modelo que prediga el logaritmo del odds de la variable dependiente, enlace (link) también conocido como *logit*. De esta forma, un modelo de regresión logística podría representarse con la siguiente fórmula:

$$Ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$$

Si bien ambas regresiones son consideradas en la familia de la distribución binomial, la regresión logística modela odds y, por tanto, la medida de asociación es el Odds Ratio (o razón de chances) y no el RR o el RP. Este lleva a que existan diferencias importantes en la dimensión del efecto y en la selección de variables a ser incluidas en el modelo múltiple[17,18].

Otra diferenciación que es muy importante es que al contrario del logit, que puede adoptar cualquier valor de los números reales, el espectro de transformación del logaritmo de una proporción (usada en la log-binomial) no puede adoptar valores positivos pues esto implicaría predecir probabilidades superiores a uno (1), resultado que carecería de lógica.

En algunas ocasiones, la regresión log-binomial no consigue un modelo en el que todas las probabilidades predichas se mantengan dentro del intervalo de cero a uno (0 a 1)[19-21]. Esto puede ocurrir porque el modelo es inapropiado o por variación aleatoria en grupos con probabilidades cercanas a la unidad. Cuando esto ocurre, los programas acaban realizando iteraciones sucesivas de forma indefinida sin alcanzar una convergencia en un modelo razonable.

Para lidiar con estos problemas de convergencia, se han planteado diversas alternativas para la estimación de RR o RP en modelos múltiples[19-22]. Entre ellas, una de las más populares es utilizar la misma regresión de Poisson incluyendo la variable dependiente dicotómica como si fuera un conteo (aunque sólo adopte dos valores) sin especificar una medida de la magnitud de la exposición. Cuando se realiza este truco, se recomienda utilizar una estimación robusta de la varianza para obtener intervalos de confianza similares a los que se obtendrían con la regresión log-binomial[19,22]. Alternativas como esta última arrojan medidas de asociación matemáticamente equivalentes a los RR obtenidos con la regresión log-binomial. Sin embargo, el investigador debe estar consciente de que está utilizando una herramienta creada para un fin diferente y debe interpretar con precaución los estadísticos post-estimación (por ejemplo, la bondad de ajuste y el pseudo-R cuadrado) para evitar extrapolaciones que entren en conflicto con la naturaleza de las variables.

CONCLUSIONES

A través de MLG utilizamos una estructura análoga a la regresión lineal para analizar desenlaces discretos, tales como resultados dicotómicos o conteos de eventos. De esta manera, podemos estimar medidas de asociación como la RT usando la regresión de Poisson, o bien, el RR o la RP usando la regresión log-binomial. En cada caso es esencial conocer la naturaleza de la variable dependiente, su distribución y reconocer las limitaciones de cada una de las herramientas de análisis.

REFERENCIAS

1. Gordis L. Epidemiology. 5th edition. Philadelphia, Elsevier Saunders, 2014: p. 177-303.
2. Rothman KJ, Greemland S, Lash TL. Modern Epidemiology. 3th edition. Philadelphia: Lippincott Williams & Wilkins 2008: p. 253-534.

3. Woodward M. Epidemiology. Study Design and Data Analysis. 2nd edition. New York: Chapman & Hall/CRC, 2005: p. 163-671.
4. Godfrey K. Simple lineal regression in medical research. N Engl J Med. 1985; 313(26): 1629-1636.
5. Zou KH, Tuncali K, Silverman SG. Correlation and simple lineal regression. Radiology. 2003; 227(3): 617-622.
6. Eberly LE. Correlation and simple lineal regression. Methods Mol Biol. 2007; 404: 143-164. DOI: 10.1007/978-1-59745-530-5_8.
7. Hamilton LC. Regression with graphics. Belmont, CA: Wadsworth, 1992.
8. Nelder J, Wedderburn R. Generalized Lineal Models. J Roy Stat Society. Series A (General) (Blackwell Publishing) 1972; 135(3): 370-384.
9. Kleinbaum DG, Kupper LL, Nizam, Muller KE. Applied regression analysis and other multivariate models. 4th edition. Belmont: Thompson Brooks/Cole, 2008.
10. Stata Corp. STATA Base Reference Manual, Volume 1, A–H, Release 11. Texas: Stata Press, 2009.
11. Miranda A, Rabe-Hesketh S. Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. Stata 2006; 6(3): 285–308.
12. Kleinbaum DG, Klein M. Logistic regression: A self-tearning text, Third Edition. New York: Springer, 2010.
13. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. New Jersey: John Wiley & Sons, 2013.
14. Hutchinson MK, Holtman MC. Analysis of count data using poisson regression. Res Nurs Health. 2005; 28(5): 408-418.
15. Coxe S, West SG, Aiken LS. The analysis of count data: a gentle introduction to poisson regression and its alternatives. J Pers Assess. 2009; 91(2): 121-136. DOI: 10.1080/00223890802634175.
16. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34(1): 1-14.
17. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol. 2003; 157(10): 940-943.
18. Pearce N. Effect measure in prevalence studies. Environ Health Perspect. 2004; 112(10): 1047-1050. DOI: 10.1289/ehp.6927.
19. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003: 3-21.
20. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. J Clin Epidemiol. 2007; 60(9): 874-882.
21. Diaz-Quijano FA. A simple method for estimating relative risk using logistic regression. BMC Med Res Methodol. 2012; 12:14. DOI: 10.1186/1471-2288-12-14.
22. Dwivedia AK, Mallawaarachchi I, Lee S, Tarwater P. Methods for estimating relative risk in studies of common binary outcomes. J Appl Statistics. 2014; 41(3): 484-500. DOI: 10.1080/02664763.2013.840772.