Publicación Electrónica **ISSN 2718- 6636**

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales
"Investigaciones en la Facultad"
Ciencias Económicas y Estadística

Boggio, Gabriela
Harvey, Guillemina
Instituto de investigaciones Teóricas y aplicadas de la Escuela de Estadística

DIFERENTES ALTERNATIVAS PARA LA INTERPRETACIÓN DE EFECTOS EN MODELOS PARA RESPUESTA BINARIA[1]

Resumen

El modelo más utilizado para el caso de respuesta binaria es el modelo de regresión logística, sin embargo la interpretación de los efectos estimados puede resultar dificultosa para los usuarios no estadísticos. En el presente trabajo se evalúa el uso de enlaces alternativos propuestos por Agresti y Tarantola, los cuales permiten interpretar dichos efectos de forma más intuitiva. Los mismos se ponen a prueba en el estudio de la demanda de consultas pediátricas en centros de salud de la ciudad de Rosario, Argentina, durante 2019. La interpretación de los resultados hallados se ve favorecida con el uso de los enlaces identidad y logaritmo, permitiendo interpretar los efectos de las variables explicativas en términos de medidas como la diferencia y la razón de probabilidades.

Palabras clave: Modelos Lineales Generalizados, Enlace Identidad, Enlace Logaritmo, Efectos Marginales Promedio

Abstract

The logistic regression model is widely used with binary responses. However, the interpretation of the estimated effects can be difficult for nonstatisticians. In this work the use of alternative link functions proposed by Agresti and Tarantola is evaluated. These alternatives allow to interpret the effects in a more intuitive way. This is illustrated by the analysis of pediatric consultations demand in health centers in the city of Rosario, Argentina, during the year 2019. The identity and log link provided results that were simpler to interpret than logistic regression, through difference and ratio of probabilities.

Keywords: Generalized Linear Models; Identity Link; Log Link; Average Marginal Effect

Introducción

El modelo de regresión logística es el modelo lineal generalizado (MLG) más utilizado cuando la variable respuesta es binaria debido fundamentalmente a la posibilidad de interpretar los efectos de las variables explicativas en términos de razones de odds (RO). Sin embargo, para los usuarios no estadísticos su interpretación no es intuitiva. Agresti y Tarantola (2021) proponen el uso de los enlaces alternativos identidad y logaritmo, según el valor de la

[1] Trabajo elaborado en el marco del Proyecto ECO215, titulado: “Enfoques estadísticos alternativos para el estudio de la ocurrencia de eventos según tiempos de exposición”, dirigido por Gabriela S. Boggio.

proporción observada para el evento de interés. Estos enlaces facilitan la interpretación de los efectos de las variables explicativas en la escala de las probabilidades. Es por ello que resulta de interés poner a prueba estas diferentes opciones en el estudio de la demanda de consultas en la guardia pediátrica de centros de salud ubicados en diferentes distritos de Rosario.

Metodología

Dado un conjunto de variables explicativas $x_j\ j = 1, \dots, p$, el MLG para respuesta binaria, con componente aleatoria Binomial, se pueden expresar de la siguiente manera:

$$g(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

donde $\pi = P(Y = 1)$ y $g(.)$ es la denominada función de enlace (Agresti, 2015).

El modelo de regresión logística en el cual la función de enlace es la denominada *logit*: $g(\pi) = ln\left[{}^{\pi}/_{1-\pi}\right]$, permite interpretar el efecto de las variables explicativas sobre la respuesta en términos de RO ya que: $RO^{Y,x_j} = exp(\beta_j)$.

Agresti *et al* (2021) proponen el uso del MLG con enlace identidad: $g(\pi) = \pi$ cuando los valores de probabilidad se encuentran en el intervalo [0,20; 0,80] ya que a través de un estudio por simulación comprueban que los ajustes derivados del enlace identidad y *logit* son prácticamente idénticos.

Ello trae aparejada una importante ventaja respecto de la interpretación del efecto de las variables explicativas directamente en la escala de las probabilidades. Es decir, β_j proporciona el incremento en la $P(Y = 1)$ ante incrementos unitarios en la variable x_j manteniendo en valores constantes el resto de las variables explicativas.

Por su parte, Agresti y Tarantola (2021) mostraron también en base a estudios por simulación que el enlace logaritmo produce un ajuste muy similar al *logit* cuando $P(Y = 1)$ se encuentra por debajo de 0,25. Su interpretación también puede expresarse en escala de las probabilidades ya que $exp(\beta_j)$ es igual a la razón de las probabilidades del evento de interés ante incrementos unitarios en la variable x_j.

En la práctica, el analista habituado al uso del modelo de regresión logística, puede complementar la interpretación del modelo en base a RO, en términos de efectos marginales promedios obtenidos como *diferencias promedio de probabilidades estimadas* a través de los n individuos.

Así por ejemplo si se desea interpretar el efecto de una variable explicativa binaria z sobre la respuesta se puede calcular la siguiente medida:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{P}(Y_i = 1/z_i = 1, \boldsymbol{x}_i) - \hat{P}(Y_i = 1/z_i = 0, \boldsymbol{x}_i)\right],$$

la cual es provista por el paquete mfx (Fernihough, 2014) de R (R Core Team, 2021).

Es de esperar que esta medida reproduzca muy cercanamente el valor del coeficiente de regresión asociado a la variable explicativa z en el correspondiente modelo con enlace identidad.

En el caso de variables cuantitativas se puede reportar la diferencia promedio de probabilidades estimadas entre dos valores extremos.

Otra medida, útil cuando las probabilidades del evento bajo estudio son muy pequeñas, es el

logaritmo de las razones de probabilidades promedio, ya que su exponenciación se puede interpretar como un riesgo relativo. Esta medida toma la siguiente forma para el caso de una variable explicativa binaria:

$$\frac{1}{n}\sum_{i=1}^{n} log\left[\frac{\hat{P}(Y_i = 1/z_i = 1, \boldsymbol{x}_i)}{\hat{P}(Y_i = 1/z_i = 0, \boldsymbol{x}_i)}\right]$$

Se espera que su valor resulte cercano al coeficiente de regresión del modelo con enlace logaritmo.

Aplicación

Se analizan las consultas médicas de los niños inscriptos en 5 centros de salud de la ciudad de Rosario, Argentina, durante el año 2019.

El objetivo, en particular, es modelar la demanda de consultas en la guardia pediátrica de centros de salud ubicados en diferentes distritos de Rosario. Como variables explicativas se considera el sexo del paciente y la cantidad de consultas realizadas en los consultorios externos durante el mencionado año, teniendo en cuenta posibles diferencias entre centros de salud.

Para ello, se indaga si los pacientes requieren al menos una consulta en la guardia y, con el propósito de estudiar la necesidad reiterada de recurrir al servicio de guardia ante algún problema de salud, se decide analizar también si los pacientes requirieron de tres o más consultas a la guardia durante el año.

Ello da lugar a la consideración de dos variables respuesta:

. *Realización de al menos una consulta en el servicio de guardia del centro de salud de pertenencia* (1: si, 0: no) – Respuesta 1.

. *Realización de tres o más consultas a la guardia* (1: si, 0: no) – Respuesta 2.

En relación a la Respuesta 1, dado que la proporción observada (0,283) se encuentra en el rango 0,20-0,80, se procede al ajuste de los modelos con enlace *logit* e identidad a los efectos de su comparación.

Tabla 1: Estimaciones de los modelos ajustados para la Respuesta 1

	Modelo con enlace *logit*		Modelo con enlace identidad	
	Coeficiente estimado	Prob. asociada	Coeficiente estimado	Prob. asociada
Sexo masculino	-0,025	0,7809	-0,006	0,3656
Cantidad de consultas en consultorio	0,167	<0,0001	0,017	<0,0001
Centro C. Namuncurá	-2,034	<0,0001	-0,331	<0,0001
Centro Dr. Mazza	-3,710	<0,0001	-0,384	<0,0001
Centro Itatí	0,578	<0,0001	0,127	<0,0001
Centro Luis Pasteur	0,043	0,6951	-0,006	0,8289

Categoría de referencia para Centro: Mauricio Casals.

Se puede apreciar en la Tabla 1 que, tal como era de esperar, el efecto del sexo no resulta significativo sobre la probabilidad de realizar al menos una consulta en el servicio de guardia bajo ninguno de los enlaces, pero sí está asociada esta probabilidad con la cantidad de consultas realizadas en consultorio externo y se puede apreciar un efecto diferencial según sea el centro de salud considerado.

Dada la relación entre los coeficientes de regresión y la medida de asociación RO, se puede decir que para niños de un mismo centro de salud, la chance de requerir consulta en el servicio de guardia aumenta un 18% (exp(0,167)) a medida que los niños realizan una consulta más en consultorio externo.

Una forma de interpretación más intuitiva sería ver en cuánto aumenta la probabilidad de consultar en el servicio de guardia al aumentar la cantidad de consultas en consultorio externo. Es decir, dar una medida en la escala de las probabilidades. Ello surge directamente de la estimación del coeficiente de regresión correspondiente en el modelo con enlace identidad. Se puede decir, entonces, que la probabilidad de realizar consultas en la guardia aumenta aproximadamente en 0,02 al aumentar en una unidad la cantidad de consultas anuales en consultorio externo.

Una medida en esta escala se podría obtener también bajo el modelo con enlace *logit*, la cual no surge directamente a partir de los coeficientes del modelo, sino que se obtiene promediando las probabilidades estimadas para cada individuo bajo el supuesto de que todos ellos realizaron "x" consultas en consultorio externo y bajo el supuesto de que todos ellos realizaron "x+1" consultas. Ello conduce a una diferencia promedio en las probabilidades estimadas igual a 0,026, valor comparable con el obtenido fácilmente a partir del modelo con enlace identidad.

En relación a la Respuesta 2, como la proporción observada es baja (igual a 0,099), se opta por considerar el modelo con enlace logaritmo además del clásico modelo de regresión logística (modelo con enlace *logit*) (Tabla 2).

Tabla 2: Estimaciones de los modelos ajustados para la Respuesta 2

	Modelo con enlace *logit*		Modelo con enlace log	
	Coeficiente estimado	Prob. asociada	Coeficiente estimado	Prob. asociada
Sexo masculino	0,083	0,5059	0,107	0,2898
Cantidad de consultas en consultorio	0,169	<0,0001	0,114	<0,0001
Centro C. Namuncurá	-1,245	<0,0001	-1,156	0,0002
Centro Dr. Mazza	-4,326	<0,0001	-4,110	<0,0001
Centro Itatí	0,658	<0,0001	0,557	<0,0001
Centro Luis Pasteur	0,242	0,1194	0,208	0,1147

Categoría de referencia para Centro: Mauricio Casals.

Al igual que en el caso de la Respuesta 1, el efecto del sexo no resulta significativo y la cantidad de consultas realizadas en consultorio sí influye sobre la probabilidad de realizar reiteradas consultas en el servicio de guardia, apreciándose también un efecto diferencial

según el centro de salud considerado tanto al considerar el enlace canónico como el enlace logaritmo.

De acuerdo a lo explicitado previamente, en lugar de recurrir a la estimación de RO, se puede aprovechar la directa interpretación que provee el modelo con enlace logaritmo. Así, la probabilidad estimada de realizar reiteradas consultas en la guardia aumenta un 12% (exp(0,114)) a medida que los niños realizan una consulta más en consultorio externo.

Es de notar que en los modelos ajustados para ambas variables respuesta, los efectos de los centros de salud se incluyeron como fijos. Sin embargo, se debe tener en cuenta que los centros considerados en este estudio constituyen una muestra de los centros del municipio según distrito y sería procedente incluirlos como efectos aleatorios en los modelos evaluados. En este sentido, Agresti *et al.* (2021) señalan que el uso de los enlaces no canónicos en este tipo de modelos puede constituir un desafío. A los efectos de una primera indagación se realiza una prueba del ajuste de un modelo con intercepto aleatorio asociado al centro de salud para la Respuesta 1 con enlace *logit* y con enlace identidad. El ajuste del modelo mixto con enlace *logit* provee resultados muy similares a los hallados en el respectivo modelo con efectos fijos. Sin embargo, al estimar el modelo mixto con enlace identidad se presentan problemas de convergencia. Vale mencionar que este inconveniente se presenta también en algunos modelos a efectos fijos con predictores diferentes a los mostrados en este trabajo. En este sentido resta avanzar en el estudio de las causas de dichos problemas de estimación.

Consideraciones finales

Dado que la interpretación de los efectos de MLG con enlaces no lineales puede ser difícil de comprender para los no estadísticos, se considera apropiado recurrir a enlaces alternativos según el rango en que se encuentre la probabilidad del evento de interés y, de este modo, obtener interpretaciones en la escala de probabilidad.

Vale destacar que el analista habituado al uso del modelo de regresión logística, puede complementar la interpretación del modelo en términos de efectos marginales promedios obtenidos como *diferencias o razones de probabilidades estimadas promedio*, según el rango en el que se encuentre la probabilidad del evento bajo estudio.

En la aplicación realizada se ha podido apreciar que el uso de los enlaces identidad y logaritmo proveen interpretaciones más intuitivas que las RO provistas por el modelo con enlace *logit*. Sin embargo, se reconocieron situaciones de falta de convergencia de algunos modelos bajo estos enlaces no canónicos. Es por ello que se pretende continuar indagando mediante estudios por simulación para analizar especialmente el comportamiento de modelos con efectos aleatorios.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* John Wiley & Sons, Inc.

Agresti, A.; Tarantola, C. (2021). Simple Ways to Interpret Effects in Modeling Binary and Ordinal Data. Conferencia en Facultad Ciencias Económicas y Estadística. Rosario, Argentina.

Agresti, A.; Tarantola, C.; Varriale, R. (2021). Interpreting effects in generalized linear modeling. Pages 1-8 in Statistical Learning and Modeling in Data Analysis, edited by S. Balzano, G. Porzio, R. Salvatore, D. Vistocco, and M. Vichi, Springer.

Fernihough, A. (2014). mfx: Marginal Effects, Odds Ratios and Incidence Rate Ratios for GLMs. R package version 1.1, URL http://cran.r-project.org/web/packages/mfx.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

FUENTES

Los datos utilizados fueron provistos por la Secretaría de Salud Pública de la Municipalidad de Rosario.