Hachuel, Leticia
Boggio, Gabriela
Borra, Virginia
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

USO DE MODELOS LOGIT MIXTOS PARA EL ESTUDIO DEL BAJO PESO AL NACER EN ROSARIO*

1. INTRODUCCIÓN

La modelización estadística tiene como objetivo capturar los aspectos principales del proceso empírico bajo investigación. El primer paso consiste en elegir la variable respuesta y a partir de allí considerar el proceso que generó los datos y el posible conjunto de variables explicativas. El objetivo es explicar la variabilidad de la respuesta, a veces denominada heterogeneidad observada, en términos de dichas variables explicativas.

Resulta crucial, entonces, para formular un modelo estadístico apropiado considerar el proceso que pudo haber generado las respuestas recurriendo por ejemplo a un modelo lineal generalizado. Pero en la práctica no siempre se observan todas las covariables relevantes, lo que da lugar a una heterogeneidad no observada.

Frecuentemente esa heterogeneidad no observada deriva del hecho de que las respuestas provienen de individuos que comparten ciertas condiciones, lo cual provoca una propensión o predisposición a presentar el evento en estudio. No tener en cuenta este tipo de heterogeneidad, la cual no alcanza a ser explicada por las covariables consideradas, conduce a inferencias incorrectas.

Una forma de contemplar este tipo de heterogeneidad en el modelo de regresión es incluir un coeficiente aleatorio. En general si se puede considerar que en los datos existen niveles de agrupamiento o jerarquías, se incluyen efectos aleatorios para cada nivel de agrupamiento. Esta clase de modelos que incluyen efectos fijos y aleatorios, es decir modelos mixtos, se la conoce también bajo la denominación de modelos de niveles múltiples.

Precisamente, dentro de la línea de investigación que sustenta el proyecto "Modelos de niveles múltiples para respuesta categórica" (proyecto de la Secretaría de Ciencia y Tecnología de la Universidad Nacional de Rosario - ECO29) se intenta dar una respuesta metodológica al objetivo de analizar información en la que es posible reconoceruna cierta jerarquía. En particular se intenta responder cómo los contextos sociales afectan los resultados y riesgos de salud individuales mediante la modelización estadística de datos sobre bajo peso al nacer en la ciudad de Rosario.

A continuación se presenta brevemente el problema del bajo peso en el recién nacido. En la Sección 3 se describen los modelos lineales generalizados mixtos. En la Sec-

* Este trabajo forma parte de un proyecto de intercambio según Convenio entre la Escuela de Estadística de la Facultad de Ciencias Económicas y Estadística de la UNR y el Instituo de la Salud Juan Lazarte/Maestría en Salud Pública del Centro de Estudios Interdisciplinarios de la UNR.

ción 4 se analizan los resultados encontrados al aplicar la metodología multinivel y finalmente se discuten dichos resultados.

2. EL PROBLEMA DEL BAJO PESO AL NACER

El tema del bajo peso del recién nacido (BPN) es desde el punto de vista epidemiológico un problema trazador para la identificación de desigualdades en el proceso salud/enfermedad/atención ya que es sensible a diferentes condiciones de vida (Luppi et al, 2006). Al respecto hay estudios que afirman que cuanto más elevada es la proporción de BPN en una comunidad, mayor es la participación de los determinantes sociales y la subordinación de los factores que afectan la salud materna a los condicionantes sociales de la población donde ocurre el nacimiento (Leal, M. C. et al, 2006).

La perspectiva de análisis que se aborde debe por lo tanto tener en cuenta cierta estratificación de la sociedad, considerando en lo posible condiciones de vida de sus habitantes. La idea es determinar cómo los contextos sociales afectan los resultados y los riesgos de salud individuales (Diez-Roux, A. 2004).

En términos estadísticos esta estratificación concibe un agrupamiento de los individuos, hecho que podría originar esa heterogeneidad no observada antes descripta (Agresti 2002; Agresti et al, 2000).

Por tal razón para el estudio del BPN en una maternidad pública de Rosario durante el año 2005, se opta por un modelo lineal generalizado mixto, de modo de incorporar la dimensión poblacional o contextual teniendo en cuenta además factores de riesgo individuales.

Para el abordaje de este trabajo se realiza un estudio de corte transversal, de fuente secundaria, con información provista por la Secretaría de Salud Pública de la Municipalidad de Rosario, que lleva un registro de acuerdo al Sistema Informático Perinatal (SIP). El mismo consiste en una base de datos con los nacimientos ocurridos en las dos maternidades públicas de la ciudad, con mayoría de variables referidas a condiciones obstétricas y biológicas de la madre y del niño.

Se define como población objetivo la totalidad de los nacimientos ocurridos en una de dichas maternidades en el año 2005 y se selecciona un conjunto de variables que permitan valorar la relación entre el peso de los recién nacidos y atributos de nivel individual, de carácter social, obstétricos y de cuidado prenatal conjuntamente con el contexto de residencia de la madre.

Se define como variable respuesta dicotómica el peso al nacer: bajo - inferior a 2500 gramos- o no. En base a los resultados hallados en la caracterización socioepidemiológica de madres atendidas en maternidades municipales en el año 1999 (Informe estadístico de la gestión de Salud Pública en la ciudad de Rosario - Secretaría de Salud Pública, 2001) se seleccionaron como variables explicativas de nivel individual: edad, nivel educacional máximo alcanzado, situación de convivencia, cantidad de controles realizados durante el embarazo, condición de primípara y forma de terminación del parto.

Para la operacionalización del nivel contextual se recurre a la variable indicadora del área geográfica, el distrito municipal correspondiente al domicilio de la madre en función de la disponibilidad de información de fuente secundaria. Se elige este indicador proxy a pesar de reconocer su debilidad como variable de estratificación.

En el punto siguiente se detalla la metodología estadística elegida, enfocada para el caso de una respuesta binaria.

3. MODELOS LINEALES GENERALIZADOS MIXTOS

La premisa básica de los modelos lineales generalizados mixtos es que la correlación entre las unidades de un mismo grupo puede pensarse que surge por el hecho de compartir un conjunto de efectos aleatorios.

Estos modelos también se suelen denominar modelos a niveles múltiples. Desde esta perspectiva, si en la estructura de la información se pueden reconocer por ejemplo dos niveles, se supone que el agrupamiento de las unidades de nivel 1 (dentro de las unidades de nivel 2) puede tenerse en cuenta a través de la consideración de heterogeneidad a través de las unidades del nivel 2 en un subconjunto de coeficientes de regresión de un modelo lineal generalizado. Condicional sobre los efectos aleatorios, las observaciones del nivel 1 se suponen independientes y con una distribución perteneciente a la familia exponencial.

La formulación de los modelos a dos niveles siguen la especificación de los modelos lineales generalizados mixtos salvo con los subíndices asociados al nivel 1 y 2 manejados en forma inversa (Goldstein, 2003; Fitzmaurice et al, 2004; Hedeker & Gibbons, 2006).

Sea Y_{ij} la respuesta para la i-ésima unidad del nivel 1 en el j-ésimo grupo del nivel 2, pudiendo ser continua, binaria o de conteo. Asociado con cada Y_{ij} hay un vector (fila) X_{ij} de covariables de dimensión 1xp, las cuales pueden estar definidas en cada uno de los dos niveles. Es decir la respuesta se obtiene sobre unidades del nivel más bajo (nivel 1) pero la información sobre las covariables puede ser medida en cualquier nivel.

Para el caso particular en que Y_{ij} es una respuesta binaria, que toma los valores 0 y 1, en el i-ésimo individuo del j-ésimo grupo, un modelo logit para Y_{ij} a dos niveles con interceptos aleatorios se especifica:

1.- Condicional sobre un único efecto aleatorio, b_j, las Y_{ij} son independientes y tienen una distribución Bernoulli, con

$Var(Y_{ij}/b_j)= E(Y_{ij}/b_j)\ \{1- E(Y_{ij}/b_j)\}$ (ϕ=1).

2.- La media condicional de Y_{ij} depende de efectos fijos y aleatorios a través del siguiente predictor lineal:

$$\eta_{ij} = X_{ij}\beta + b_j$$

con

$$\ln\left\{\frac{Pr(Y_{ij}=1/b_j)}{1-Pr(Y_{ij}=1/b_j)}\right\} = \eta_{ij} = X_{ij}\beta + b_j .$$

Es decir la media condicional de Y_{ij} se relaciona con el predictor lineal a través del enlace logit.

3.- El único efecto aleatorio b_j se supone que tiene una distribución normal univariada con media cero y variancia σ_b^2.

La inferencia en este tipo de modelos presenta desafíos debido a la necesidad de aplicar métodos numéricos para la integración de la función de verosimilitud. En este sentido

están bien documentadas las dificultades en la estimación de parámetros de los modelos de regresión logística con efectos aleatorios (Bellamy et al, 2005).

Un método simple de estimación es el enfoque de cuasiverosimilitud penalizada, utilizado en el procedimiento GLIMMIX de SAS. Sin embargo, varios autores han notado que los efectos de las covariables estimados por este método pueden presentar un sesgo importante. Un procedimiento alternativo que brinda SAS es el NLMIXED, que obtiene estimadores máximo verosímiles de los parámetros del modelo aproximando numéricamente la logverosimilitud.

Para la interpretación de los coeficientes de las covariables es necesario tener en cuenta que el modelo es condicional sobre el efecto aleatorio para grupo, y dentro de ese grupo, el coeficiente de una covariable representa la magnitud del cambio en el logaritmo del odds de respuesta positiva que uno debería esperar ante un valor particular de la covariable versus otro valor de la misma. Debido a que el modelo especifica que ese coeficiente es el mismo para todos los grupos, se estima combinando la información de diferentes grupos, es decir promediando sobre todos los grupos de acuerdo a la distribución de ese efecto aleatorio para grupo.

4. RESULTADOS

El objetivo principal del ajuste de modelos logísticos mixtos es realizar inferencias acerca de los efectos fijos. La inclusión de efectos aleatorios en el modelo es un mecanismo para caracterizar cómo la correlación positiva se presenta entre las observaciones dentro de un grupo. De esta manera el error estándar del intercepto aleatorio de un modelo constituye un resumen útil del grado de heterogeneidad de la población en estudio (Agresti, 2000).

Para intentar explicar el bajo peso del recién nacido en términos de condiciones maternas individuales se considera el ajuste de modelos de regresión logística mixto de acuerdo a la siguiente definición de las variables.

Variable Respuesta: Bajo peso al nacer: sí (entre 500 y 2500 g.) , no (>2500 g.).

Se excluyen abortos (peso menor que 500 g. o edad gestacional menor que 20 semanas), muertes intrauterinas, malformaciones y embarazos múltiples.

Variables explicativas del nivel individual:

- Convivencia: con pareja (casada o unión consensual), sin pareja (soltera, separada o viuda).
- Nivel educacional máximo alcanzado: ninguno, primaria, secundaria, universitario
- Controles durante el embarazo: ≤ 4 controles, entre 5 y 7 controles, ≥ 8 controles.
- Edad: ≤ 19 años, entre 20 y 39 años, ≥ 40 años.
- Condición de primípara: no, si.
- Forma de Terminación del Parto: espontáneo, no espontáneo (fórceps, cesárea, otro)

Variable del nivel contextual:

- Distrito municipal correspondiente al domicilio de la madre: centro, norte, noroeste, oeste, sudoeste, sur.

Se busca de esta forma recuperar en el análisis estadístico la diferenciación de niveles, individual –de la madre– y contextual –grupal según ámbito socioespacial de pertenencia–.

Se evaluaron diferentes modelos considerando efectos fijos para las variables definidas e interacciones dobles entre ellas. El modelo más simple hallado es el que incluye efectos principales significativos para edad, condición de primípara, controles durante el embarazo y forma de terminación del parto. Los resultados se presentan en la siguiente tabla.

Tabla 1: Estimaciones de los parámetros del modelo logit con efecto aleatorio para cada distrito

Parámetro	**Estimación**	**Error Estándar**	**P-asoc.**
Intercepto	-2.566	0.193	**<0.0001**
Edad:			
entre 20 y 39 años	0.176	0.170	0.302
≥ 40 años	1.175	0.420	**0.005**
Primípara	0.455	0.161	**0.005**
Controles prenatales:	-0.3067	0.0318	**<0.0001**
entre 5 y 7 controles	-0.751	0.152	**<0.0001**
≥ 8 controles	-1.719	0.208	**<0.0001**
Parto no espontáneo	0.932	0.142	**<0.0001**
Variancia del efecto aleatorio	0.020	0.262	0.4698

El valor estimado de la variancia del efecto distrito, asumido aleatorio en el modelo, resulta igual a 0.020 con alta probabilidad asociada (p=0.4698), por lo cual el grado de heterogeneidad no es importante ni resulta significativo. Sin embargo bajo este modelo no tiene significación estadística la interacción entre primípara y forma de terminación, la cual hubiese sido incluida en el modelo si los datos se hubieran ajustado por un modelo de regresión logística convencional.

Estos resultados concuerdan con la posible reducción en la significación estadística de los efectos cuando se tiene en cuenta el agrupamiento de las respuestas según distrito, aún cuando la heterogeneidad entre ellos sea leve. La inclusión de efectos aleatorios asociados a agrupamientos según el lugar de residencia condujo a resultados más confiables que los obtenidos con el enfoque convencional.

El ajuste del modelo se realizó utilizando el procedimiento GLIMMIX de SAS. El procedimiento de estimación que emplea suele producir estimaciones sesgadas de los parámetros de regresión sobre todo en el caso de variable respuesta binaria y cuando el número de elementos por grupo es pequeño. Si bien la cantidad de mujeres por grupo –distrito- es importante, se reprocesaron los datos con el procedimiento alternativo disponible en SAS, NLMIXED, a fin de corroborar los resultados. La desventaja de este procedimiento es el elevado tiempo de procesamiento y la necesidad de asignar buenos valores iniciales a los

parámetros a estimar. Los resultados hallados con ambos procedimientos fueron muy similares, despejando de esta manera las dudas que podrían surgir como consecuencia del método de estimación elegido.

Los coeficientes de las covariables en el modelo logit presentado se interpretan en términos de razones de odds condicionales. Éstas constituyen una medida aproximada del riesgo de que las mujeres residentes en un determinado distrito tengan un niño de bajo peso al nacer según sea el valor asumido por cada covariable en particular manteniendo constante el valor de las restantes.

Por lo tanto, para un mismo distrito la chance estimada de bajo peso en el recién nacido:

- es el 58% (exp(0.455)=1.58) mayor para las madres primíparas que para las que no lo son;
- es más del doble (exp(0.932)=2.54) cuando la terminación del parto no fue espontánea, en el sentido que requirió de cesárea, fórceps o alguna otra intervención;
- disminuye a la mitad (exp(-0.751)=0.47) para las madres que realizan entre 5 y 7 controles en comparación con las que realizan un número insuficiente de controles (4 ó menos). Esta disminución es de más de un 80% (exp(-1.719)=0.18) cuando se compara el grupo que realiza 8 o más controles con el grupo de mayor riesgo.
- Aumenta al triple (exp(1.175)=3.24) cuando la madre tiene 40 o más años de edad que si tiene menos de 20. Sin embargo tienen igual chance de bajo peso los recién nacidos de madres muy jóvenes (menos de 20 años) que las de edad intermedia (entre 20 y 39 años).

Si se recurre a la construcción de intervalos de confianza para la interpretación de las razones de odds se observa que el intervalo para la condición de primípara ($IC_{95\%}$: 1.15; 2.16) muestra que la chance de bajo peso puede llegar a ser sólo un 15% mayor para las mujeres primíparas que para las que no lo son. En cambio con respecto a la forma de terminación del parto, el intervalo de confianza ($IC_{95\%}$: 1.92; 3.35) indica que el aumento en la chance de bajo peso al nacer es, como mínimo, casi del doble cuando el parto no termina en forma espontánea y puede llegar a ser algo más que el triple.

En relación al número de consultas, el intervalo de confianza para un número de consultas entre 5 y 7 en comparación con un bajo número de consultas, ($IC_{95\%}$: 0.35; 0.64), muestra que la disminución en la chance de bajo peso es como mínimo de un 36%. En cambio esta disminución es como mínimo de un 73% ($IC_{95\%}$: 0.12; 0.27) cuando el número de consultas supera las 7.

Por último con respecto a la edad las chances de bajo peso de las madres de edad intermedia versus las muy jóvenes resultan similares ($IC_{95\%}$: 0.85; 1.66). A diferencia de ello, esta chance puede llegar a ser 7 veces mayor cuando la edad de la madre supera los 40 años que cuando es muy joven ($IC_{95\%}$: 1.42; 7.38).

Resulta importante señalar, y los resultados así lo corroboran, que la utilización del distrito municipal como aproximación para representar grupos geográficos homogéneos en cuanto a la dinámica social, dista mucho de ser satisfactoria. En este sentido, se está trabajando en la elaboración de indicadores más finos de la micro-área de residencia de la mujer (Luppi et al, 2006).

Merece un comentario particular la comparación de estos resultados con los hallados en un análisis similar realizado en base a los partos acontecidos en la misma maternidad en el año 2003 (Hachuel et al, 2004). En ambos cortes transversales -2003 y 2005 – se verifica la

significación de la condición de primípara y de la forma de terminación del parto sobre el BPN. También resulta significativa en ambos estudios la cantidad de controles prenatales, aunque se piensa que su consideración en categorías utilizada en el 2005 es más apropiada. Por último en el año 2005 aparece como efecto significativo sobre el BPN la edad de la madre, que en esta oportunidad se categoriza según grupos etáreos definidos para diferenciar entre embarazadas adolescentes, añosas y de edades intermedias, reconocido este último grupo como el de menor riesgo.

5. CONSIDERACIONES FINALES

Un concepto clave en epidemiología es que todos los determinantes y condicionantes de las enfermedades no pueden ser conceptualizados sólo como atributos a nivel individual, por lo que resulta necesario, cuando se quieren estudiar problemas de salud-enfermedad, considerar aspectos de los grupos a los cuales pertenecen los individuos.

El reconocimiento de esta diferenciación en niveles jerárquicos de los condicionantes del fenómeno en estudio provoca desafíos metodológicos. Los modelos de niveles múltiples constituyen un enfoque apropiado para el análisis de este tipo de datos.

La puesta a prueba de esta clase de modelos en datos del campo de la salud ilustra cómo concebir las condiciones sociales y medio ambientales como una dimensión explicativa adicional en el estudio de un problema de salud a nivel individual. En particular para el estudio del bajo peso al nacer, la aplicación de los modelos de niveles múltiples, más concretamente un modelo logístico-normal, proviene de pensar al lugar de residencia de la embarazada como una posible fuente adicional de variabilidad. La interpretación de los coeficientes de efectos fijos se puede realizar, como en toda regresión logística, en términos de razones de odds pero en estos modelos están referidas a cada grupo –distrito- en particular.

En relación a los efectos aleatorios, si bien no son directamente observables se los piensa como que representan propensiones producto de las condiciones socio-ambientales compartidas. Si se tiene en cuenta que en la concepción de una ciudad en distritos subyace la idea de subdividir la misma en mini-comunidades que reproduzcan en cierta forma las mismas situaciones y relaciones que el municipio en su conjunto parece no objetable el resultado encontrado sobre la no significación estadística del efecto aleatorio distrito. Sin embargo, es sabido que existen diferencias a nivel territorial las cuales podrían detectarse a partir de indicadores más finos que la “proxy” distrito utilizada en este trabajo. Precisamente, una línea de investigación que se está siguiendo es la búsqueda de indicadores más eficientes de la micro-área de residencia de la madre.

6. REFERENCIAS

AGRESTI, A. 2002. *Categorical Data Analysis*, 2nd ed. John Wiley & Sons.

AGRESTI, A.; BOOTH, J.; HOBERT, J.; CAFFO, B. 2000. Random-effects modeling of categorical response data. *Sociological Methodology*, 30: 27-81.

BELLAMY, L.; LI, Y.; LIN, X.; RYAN, L. 2005. “Quantifying PQL Bias Estimating Cluster-level Covariate Effects in Generalized Linear Mixed Modelsfor Group-randomized Trials”. Statistica Sinica, 15: 1015 -1032.

DIEZ-ROUX, A. 2004. The study of group-level factors in epidemiology: rethinking variables, study designs and analytical approaches. *Epidemiology reviews*, 26:104-111.

FITZMAURICE, G.; LAIRD, N.; WARE, J. 2004. *Applied longitudinal analysis* . John Wiley & Sons.

GOLDSTEIN, H. 2003. *Multilevel Statistical Models.* 3rd edition. Kendall's Library of Statistics. London.

HACHUEL, L.; BOGGIO, G.; WOJDYLA, D. 2004. Modelos logit mixtos: una aplicación en el área de salud. Novenas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística. www.fcecon.unr.edu.ar.

HEDEKER, D.; GIBBONS, R. D. 2006. *Longitudinal data analysis.* John Wiley & Sons.

LEAL, M. C.; NOGUEIRA DA GAMA, S. G.; BRAGA DA CUNHA, C. *Rev Saúde Pública* V 40 nº 3.

LUPPI, I.; HACHUEL, L. BOGGIO, G, BORRA, V. 2006. Desigualdades en salud: estudio del bajo peso del recien nacido en Rosario. Décimo Congreso de la salud en el municipio de Rosario. 3° Jornadas Nacionales de Epidemiología y 4° Jornadas de Economía y Gestión de Salud. Rosario.

SAS Institute, Inc. 2004. SAS/STAT user's guide, version 9.1.3. Cary, NC, USA.

SECRETARÍA DE SALUD PÚBLICA. Municipalidad de Rosario, 2001. Informe estadístico de la gestión de Salud Pública en la ciudad de Rosario. *Boletín de Bioestadística*. Año 4 N° 1.