XII Jornadas Argentinas de Estudios de Población. Asociación de Estudios de Población de la Argentina, Bahía Blanca, 2013.

Una aproximación cuantitativa al grado de exposición a riesgo a la mortalidad infantil.

Pablo Caviezel.

Cita:

Pablo Caviezel (2013). *Una aproximación cuantitativa al grado de exposición a riesgo a la mortalidad infantil. XII Jornadas Argentinas de Estudios de Población. Asociación de Estudios de Población de la Argentina, Bahía Blanca.*

Dirección estable: https://www.aacademica.org/xiijornadasaepa/69

ARK: https://n2t.net/ark:/13683/edrV/1QG

Esta obra está bajo una licencia de Creative Commons.
Para ver una copia de esta licencia, visite
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es.

Acta Académica es un proyecto académico sin fines de lucro enmarcado en la iniciativa de acceso abierto*. Acta Académica fue creado para facilitar a investigadores de todo el mundo el compartir su producción académica. Para crear un perfil gratuitamente o acceder a otros trabajos visite: https://www.aacademica.org*.

UNA APROXIMACIÓN CUANTITATIVA AL GRADO DE EXPOSICIÓN AL RIESGO A LA MORTALIDAD INFANTIL

Pablo Caviezel
Maestría en Demografía Social, Universidad Nacional de Luján
pablocaviezel@yahoo.es

INTRODUCCIÓN

La mortalidad infantil fue y es un tema que preocupa a la sociedad en tanto y en cuanto revela aspectos intrínsecos de su desarrollo y de la dirección de las políticas públicas de quienes la dirigen. En tal sentido, el objetivo del trabajo es presentar un modelo estadístico que permita determinar cuantitativamente la probabilidad de que un nacimiento –que presenta determinadas características iniciales– se encuentre expuesto a alto riesgo de fallecer antes de cumplir un año de edad. Las características o condiciones iniciales se presentan, para este modelo, como un vector de valores asociados a distintas variables que son tanto cuantitativas como categóricas. Asimismo, el trabajo pretende introducir al modelo un análisis de sensibilidad que permita evaluar cuantitativamente el riesgo extra al que está expuesto un nacimiento con características específicas diferentes respecto de aquél tomado como parámetro al establecer las condiciones iniciales y permitir conocer cómo impacta en la probabilidad de que un nacido vivo fallezca antes de cumplir un año distintas hipótesis acerca del comportamiento de distintas variables.

EL MODELO DE REGRESIÓN LOGÍSTICA: SUPUESTOS Y CONDICIONES DE APLICABILIDAD

Cuando estudiamos el modelo lineal clásico de regresión, intentamos encontrar el vector de estimadores que mejor represente un modelo con *k* variables explicativas de la forma:

(1) $$y=b_o + b_1x_1 + b_2x_2 +\ldots+ b_kx_k + \varepsilon$$

Si la variable dependiente es una proporción o una probabilidad y, por tanto, su recorrido el intervalo denso real [0; 1], por más robustas que resulten nuestras estimaciones, podríamos predecir un valor para la variable explicada que quede fuera del rango prestablecido. Nos interesa entonces predecir o estimar el valor probabilístico *p* que puede interpretarse como la probabilidad de éxito de que una variable aleatoria dicotómica presente un atributo o característica definida.

Dada una probabilidad p distinta de cero, cuyo valor se encuentra incluido en el intervalo real (0;1), se deduce que:

- $1 - p$ se encuentra en el intervalo real (0;1)
- $\frac{p}{1 - p}$ se encuentra en el intervalo real (0;+∞)
- $\ln\left[\frac{p}{1 - p}\right]$ se encuentra en el intervalo real (-∞;+∞)

Convenimos en llamar «odds absoluto» a la expresión $\theta \frac{p}{1 - p}$ y en llamar «logit» a la expresión $L = \ln\left[\frac{p}{1 - p}\right]$, entendiéndolas como construcciones matemáticas que adecuan el dominio de la variable dependiente al uso de la clásica regresión lineal.

Así, para un evento altamente probable (digamos, con probabilidad mayor a 0,50), resultara que su *odds absoluto* es superior a 1 y, por lo tanto, su *logit* positivo. De manera análoga, para un evento poco probable

(digamos, con probabilidad inferior a 0,50), resultará que su *odds absoluto* es un número real comprendido entre 0 y 1 y, por lo tanto, su *logit* negativo.

Queda claro por otra parte que, habiendo definido $L = \ln\left[\frac{p}{1-p}\right]$, a partir de la exponenciación directa resulta:

$$e^{L} = \frac{p}{1-p}$$

O su expresión equivalente:

(2) $$e^{L} = \theta$$

De esta manera, podemos reexpresar el modelo de regresión (1) como:

$$L = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

Y el modelo lineal final estimado:

(3) $$\hat{L} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \ldots + \hat{b}_kx_k$$

En el caso de que las variables independientes sean variables binarias, utilizaremos la letra z como notación para distinguirlas. Estas variables toman sólo los valores 0 y 1 para referirse a ausencia o presencia de determinado atributo y la fórmula (3) quedaría expresada de la siguiente manera:

$$\hat{L} = \hat{b}_0 + \hat{b}_1z_1 + \hat{b}_2z_2 + \ldots + \hat{b}_kz_k$$

Donde -a partir de la estimación del valor L- podemos, utilizando la fórmula $\hat{L} = ln \frac{\hat{p}}{1-\hat{p}}$ obtener, mediante sencillo manejo algebraico, la probabilidad estimada $\hat{p}$ de ocurrencia del evento, dada la muestra, como:

(1) $$\hat{p} = \frac{e^{\hat{f}}}{1 + e^{\hat{f}}}$$

La relación unívoca entre un valor *p* de probabilidad de éxito y su *odds absoluto* θ permite comparar los *odds* de dos eventos diferentes. En efecto, supongamos que un evento A_1 cuya probabilidad de éxito es p_1 resulta más probable que otro evento A_2 cuya probabilidad de éxito es p_2. Resulta claro que $p_1 > p_2$

Luego, puesto que $p_1 > p_2$ resulta multiplicando miembro a miembro por -1 que $-p_1 < -p_2$ y, sumando 1 miembro a miembro que $1 - p_1 < 1 - p_2$.

De esa manera, resulta que la relación de los *odds absolutos* es:

$$\frac{p_1}{1 - p_1} > \frac{p_2}{1 - p_2}$$

$$\frac{\frac{p_1}{1 - p_1}}{\frac{p_2}{1 - p_2}} > 1$$

La expresión de la izquierda, que es el cociente de los odds absolutos, la llamaremos *odds relativo* y la simbolizaremos $\theta_{1/2}$. En definitiva, si un evento A_1 es más probable que otro evento A_2 entonces el *odds relativo* $\theta_{1/2}$ resulta superior a la unidad. Más aún, si tomamos logaritmo natural a ambos miembros, verificamos entonces que el logaritmo natural de $\theta_{1/2}$ debe ser necesariamente mayor a cero.

En el caso particular donde todas nuestras variables explicativas sean variables aleatorias binarias (es decir, que toman 0 ante ausencia de algún atributo particular y 1 en presencia), la interpretación de los coeficientes de la regresión lineal resulta muy clara.

En efecto, volviendo a (3):

$$\hat{L} = \hat{b}_o + \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k$$

Si todas las variables explicativas binarias tomaran simultáneamente el valor 0 (ausencia de atributo), resulta:

(2) $$\hat{L}_0 = \hat{b}_0$$

Exponenciando a ambos miembros:

(3) $$e^{\hat{L}_0} = e^{\hat{b}_0}$$

Pero, como habíamos notado que el *logit* tomaba la forma:

$$L = \ln\left(\frac{p}{p\text{-}1}\right)$$

exponenciando sobre esta última fórmula obtenemos: $e^{\hat{L}_0} = \hat{\theta}_0$ y, puesto que la relación (5) establecía $\hat{L}_0 = \hat{b}_0$ surge inmediatamente la útil relación:

(4) $$e^{\hat{L}_0} = \hat{\theta}_0$$

Donde $\hat{\theta}_0$ hace referencia al *odds absoluto* trivial, es decir, correspondiente a aquella combinación que surge de la ausencia de atributo para todas las variables explicativas y que, en la literatura tradicional, se lo denomina «grupo de referencia». Algebraicamente, el grupo de referencia está caracterizado por la evaluación del vector nulo en la ecuación de la regresión.

De manera análoga podríamos razonar si todas las variables explicativas binarias tomaran simultáneamente el valor 0 (ausencia de atributo), excepto una de ellas que tomara el valor 1 (presencia de

atributo). En este caso, dicha variable independiente, asumamos Z_j, ocupará el j-ésimo puesto en la definición de variables independientes y el vector de valores asociado será un vector canónico con un 1 en la j-ésima posición. De forma tal que, la expresión (3):

$$\hat{L} = \hat{b}_0 + \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_j z_j + \ldots + \hat{b}_k z_k$$

Resulta, especializando en dicho vector canónico:

$$\hat{L}_j = \hat{b}_0 + \hat{b}_j$$

Exponenciando a ambos miembros:

(5) $$e^{\hat{L}_j} = e^{\hat{b}_0 + \hat{b}_j}$$

Si dividimos miembro a miembro la expresión (8) por la expresión (6) $e^{\hat{L}_0} = e^{\hat{b}_0}$ ya enunciada anteriormente, resulta:

$$\frac{e^{\hat{L}_j}}{e^{\hat{L}_0}} = \frac{e^{\hat{b}_0 + \hat{b}_j}}{e^{\hat{b}_0}}$$

Por analogía con la fómula (6) resulta:

$$\theta_{j/0} = e^{\hat{b}_j}$$

En este sentido, $e^{\hat{b}_j}$ nos provee de una estimación para el *odds relativo* del grupo j respecto de aquél tomado como referencia. Es importante mencionar que un b_j positivo se traduce en un valor de odds relative mayor a la unidad y, por lo tanto, en una mayor probabilidad de éxito del grupo en cuestión respecto del grupo de referencia. De la misma manera, un b_j

negativo implica que la probabilidad de éxito es menor para el grupo en cuestión que para el grupo de referencia.

EL CONCEPTO DE NACIMIENTO EN RIESGO Y LAS HIPÓTESIS SUBYACENTES

¿Por qué ocurre una defunción infantil? Pregunta difícil de contestar si las hay, mismo aún por la dificultad de entender cuál es el alcance de la pregunta "por qué". Respuestas muy simples serían: "porque el niño nació con bajo peso" o "porque el niño nació prematuro" pero éstas, a su vez, exigen repreguntas acerca de por qué entonces se dieron estas situaciones. Las respuestas a estas preguntas, a su vez, se hallan dentro del campo de estudio más bien biológico y no tanto demográfico. Siguiendo a Aguirre (2009), los hijos de mujeres que se hallan en los extremos del período reproductivo, los de orden alto (del cuarto en adelante) y los que presentan bajo peso al nacer tienen mayor riesgo de morir antes de cumplir un año de edad. Marmot (2005), por su parte, agrega factores sociales al explicar que entre los países no sólo la mortalidad en la primera infancia es mayor entre los hogares más pobres, sino que además, a mayores niveles socioeconómicos, menores tasas de mortalidad infantil. En todos estos análisis se logra encontrar variables que se relacionan con la mortalidad infantil, sin cabalmente explicarla. Se trata entonces de establecer hipótesis de comportamiento y de relaciones de causalidad entre distintas variables.

Cuando se analizan estos factores que inciden en el nacimiento de un niño con alto riesgo de fallecer antes de cumplir un año de edad, se pueden clasificar los mismos en dos grandes grupos: a) aquellos que se conocen con anterioridad al nacimiento del niño, tales como la edad de la madre, el número de orden del nacimiento, el nivel socioeconómico del hogar, etc. y b) aquellos que se conocen sólo al momento de nacimiento del niño, tales como las semanas de gestación y el peso al nacer. En virtud de este agrupamiento, y –a los efectos de este trabajo- vamos a definir "nacido vivo con alto riesgo" de acuerdo a tres definiciones alternativas:

a) Tipo I: nacido vivo prematuro (menos de 37 semanas de gestación)

b) Tipo II: nacido con bajo peso al nacer (hasta 2.500 gramos)

Este trabajo intenta, como se detallará más adelante, la probabilidad de que un nacido vivo sea de alto riesgo, a partir de las variables relevadas en el Informe Estadístico de Nacido Vivo, de las Estadísticas vitales de la República Argentina.

FUENTE DE DATOS: SUS CARACTERÍSTICAS Y EVALUACIÓN

A partir del completamiento del Informe Estadístico de Nacido Vivo para los nacimientos registrados en la Ciudad de Buenos Aires durante el año 2012, se cuenta con una base de 81.527 nacidos vivos y sus características relevadas en el informe mencionado. Se trata de nacimientos que han ocurrido en la Ciudad de Buenos Aires y corresponden a madres que pueden o no residir en dicha ciudad. Las variables que se estudiarán del Informe se presentan en el siguiente cuadro:

Cuadro 1. Variables seleccionadas del Informe Estadístico de Nacido Vivo. Ciudad de Buenos Aires. Año 2012.

Variable	Unidad de medida	Tipo	Dominio
Tiempo de gestación	Semanas	Cuantitativa	Números naturales
Peso al nacer	Gramos	Cuantitativa	Números naturales
Sexo del nacido vivo	-	Nominal	Varón / Mujer
Edad de la madre	Años cumplidos	Cuantitativa	Números naturales
Tipo de parto	-	Nominal	Binaria: {0; 1}
Hijos nacidos vivos	Nacidos vivos	Cuantitativa	Números naturales
Nivel de instrucción	-	Ordinal	Nivel

Fuente: Elaboración propia sobre la base de Dirección General de Estadística y Censo (Ministerio de Hacienda GCBA). Estadísticas vitales.

Para evaluar la calidad de los datos, se hará hincapié en tres dimensiones:

- Oportunidad: contar con la mayor cantidad de nacimientos ocurridos en el año 2012 y que hayan sido registrados en tiempo.
- Cabalidad: comprobar que no haya respuestas ignoradas en las variables en estudio en la base de datos.
- Consistencia: verificar la coherencia de la información relevada para cada variable.

Al evaluar la oportunidad, es menester considerar que se consideran nacimientos registrados en el año 2012. Como es de esperar, existen nacimientos que habrían ocurrido en 2011 (o incluso antes) y que son registrados en el año 2012. Se considera que estos casos se compensan con aquellos nacimientos que ocurren a finales de 2012 y serán (son) registrados en 2013. Tal como se especificó al inicio del presente acápite, se cuenta con 81.527 nacidos vivos y con sus características relevadas.

Cuadro 2. Cabalidad de las variables seleccionadas del Informe Estadístico de Nacido Vivo. Ciudad de Buenos Aires. Año 2012.

Variable	Casos			Porcentaje de casos válidos
	Total	Válidos	Respuesta ignorada	
Tiempo de gestación	81.527	79.859	1.668	98,0
Peso al nacer	81.527	80.800	727	99,1
Sexo del nacido vivo	81.527	81.527	-	100,0
Edad de la madre	81.527	81.259	268	99,7
Tipo de parto	81.527	81.527	-	100,0
Hijos nacidos vivos	81.527	80.311	1.216	98,5
Nivel de instrucción	81.527	80.140	1.387	98,3

Fuente: Elaboración propia sobre la base de Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA). Estadísticas vitales.

El cuadro 2 permite conocer la cabalidad de los datos, es decir, el porcentaje de respuestas no ignoradas para cada variable. Vemos que en todos los casos el porcentaje es mayor o igual a 98 % y, por ende, las respuestas ignoradas representan a lo sumo el 2 % en cada variable. Es menester, sin embargo, tener en cuenta que cuando se trabaje con varias variables en simultáneo se constatará que ninguna de ellas presente casos ignorados. Para cada conjunto de variables siempre se establecerá el total de casos válidos y el porcentaje que representa sobre los 81.527 casos de la base.

Respecto a la consistencia de los datos, la misma Dirección General de Estadística y Censos de la Ciudad realiza un seguimiento y análisis de la coherencia y validez de las respuestas de cada variable. Por tanto, para este trabajo, se considera que dicho análisis ya fue efectuado y sólo se han evaluado las frecuencias de cada variable para no encontrar casos con valores no posibles de variable.

APLICACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

El primer trabajo consiste en crear variables dicotómicas (o dummies); es decir, variables que toman valor 0 cuando el atributo no está presente y 1 cuando está presente. Para ser consistente y sencillos en la interpretación, se ha intentando mantener el 1 como indicador de un evento que favorece la presencia de riesgo para el recién nacido.

El cuadro 3 presenta una síntesis de las variables y categorías utilizadas en los distintos modelos de regresión que se plantearán:

Cuadro 3. Variables dicotómicas para el modelo de regresión logística. Ciudad de Buenos Aires. Año 2012

Variable	Símbolo	Valor	
		0	1
Tiempo de gestación	Y_1	Normal (37 o más semanas)	Bajo (Menos de 37)
Peso al nacer	Y_2	Normal (2500 g o más)	Bajo (Menos de 2500 g)
Sexo del nacido vivo	Z_1	Mujer	Varón
Madre joven	Z_2	No (20 años o más)	Sí (Hasta 19 años)
Madre mayor	Z_3	No (Hasta 39 años)	Sí (40 años o más)
Tipo de parto	Z_4	Simple (un nacido vivo)	Múltiple (dos o más nacidos vivos)
Cantidad de hermanos	Z_5	Sin hermanos	Con hermanos
Riesgo educativo	Z_6	Sin riesgo (primario completo y más)	Con riesgo (hasta primario incompleto)

Nota: Las variables Y se consideran dependientes y las variables Z se consideran independientes para el modelo.
Fuente: Elaboración propia sobre la base de Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA). Estadísticas vitales.

El vector transpuesto $Z^T = [Z_1\ Z_2\ Z_3\ Z_4\ Z_5\ Z_6]$ resulta en un vector que contiene sólo ceros y unos, de acuerdo al grupo especificado. Por ejemplo:

Z^T [0 0 1 0 0 1] refiere a un nacido vivo de sexo femenino, de una madre mayor de 40 años con riesgo educativo (sin educación o con primario incompleto a lo sumo), nacida de parto simple y primera hija.

Z^T = [1 1 0 1 1 0] refiere a un nacido vivo de sexo masculino, de una madre menor de 20 años sin riesgo educativo (primario completo o más), nacido de parto múltiple y con hermanos.

Z^T = [1 0 0 0 1 0] refiere a un nacido vivo de sexo masculino, de una madre entre 20 y 39 años cumplidos sin riesgo educativo (primario completo o más), nacido de parto simple y con hermanos.

En particular, el grupo de referencia al que llamaremos Z^T_0 es:

Z^T_0= [0 0 0 0 0 0] refiere a un nacido vivo de sexo femenino, de una madre entre 20 y 39 años cumplidos sin riesgo educativo (primario completo o más), nacido de parto simple y sin hermanos.

Se supone, y ésta es una hipótesis que podrá ser refutada, que el grupo de referencia describe la situación en la que el nacido vivo presenta menor riesgo de fallecer antes de cumplir un año de edad. Finalmente, es importante destacar, antes de presentar los distintos modelos, que si bien la regresión logística no se basa en supuestos distribucionales para las variables aleatorias independientes intervinientes, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y, por lo tanto, a conclusiones erróneas. Se trabajará con el supuesto de independencia estocástica de las variables explicativas.

La primera regresión que se hará considera como variable dependiente el tiempo de gestación del nacido vivo e incorpora el análisis de todas las variables regresoras. Para este modelo $\widehat{L}_1 = ln\left(\frac{p_1}{1-p_1}\right)$

y p_1 representa la probabilidad de que un nacido vivo nazca prematuro y, por ende, tenga riesgo de fallecer antes de cumplir un año de edad. En otras palabras, p_1 representa la probabilidad de que el recién nacido tenga riesgo tipo I, de acuerdo con la definición de la página 7. En el análisis de las potenciales variables regresoras deben excluirse todos aquellos casos que contengan algún valor ignorado (respuesta ignorada) en alguna de las variables. Así, nos quedamos con una muestra cabal de 78.146 casos, que representa casi el 96 % del total de casos de la base.

Cuadro 4. Distribución porcentual de los nacimientos por tiempo de gestación según variables seleccionadas. Ciudad de Buenos Aires. Año 2012.

Características seleccionadas	Tiempo de gestación (en semanas)		
	Total	Hasta 36	37 y más
Total	100,0	8,6	91,4
Sexo			
Varón	100,0	8,6	91,4
Mujer	100,0	8,5	91,5
Edad de la madre			
Hasta 19	100,0	7,9	92,1
20 - 39	100,0	8,3	91,7
40 y más	100,0	14,6	85,4
Tipo de parto			
Simple	100,0	7,0	93,0
Múltiple	100,0	58,7	41,3
Número de hermanos			
0	100,0	7,2	92,8
1 o más	100,0	9,8	90,2
Nivel de instrucción			
Hasta primario incompleto	100,0	7,4	92,6
Secundario incompleto	100,0	7,6	92,4
Secundario completo y más	**100,0**	9,1	90,9

Fuente: Elaboración propia sobre la base de Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA). Estadísticas vitales.

¿Qué vemos en el cuadro 4? Vemos estructuras (distribuciones porcentuales) para las distintas categorías de cada una de las variables

explicativas propuestas. El objetivo de este cuadro es verificar si existen diferencias significativas entre las categorías de las variables al momento de intentar explicar las diferencias en los tiempos de gestación. Si bien este análisis puede realizarse con pruebas de hipótesis de independencia de atributos, el uso de distribuciones porcentuales permite efectuar comparaciones y dimensionar los fenómenos. En primer lugar, vemos el tiempo de gestación del niño recién nacido no es una variable que presente diferencias significativas por sexo y, en el caso de la edad de la madre, la influencia de las madres mayores de 40 años puede producir cambios en el tiempo gestacional. Por otra parte, y sin duda, el parto múltiple favorece la corta gestación y el nivel educativo no presenta grandes diferencias en la estructura de los tiempos de gestación. La variable nivel de instrucción puede presentar algún tipo de multicolinealidad con la edad de la madre y –como se explicó anteriormente- distorsionar algún resultado. No son terminantes los guarismos del citado cuadro para determinar si existe relación entre la cantidad de hermanos y la edad gestacional, por lo que se optará por incluir esta variable en el modelo. En síntesis, conviene quedarnos con las variables Z_3, Z_4 y Z_5. Al plantear el modelo es importante tener en cuenta que ahora nuestra base de datos va a contener todos aquellos casos que no presentan valores ignorados para este nuevo conjunto de variables seleccionados. Así, quedan seleccionados 78.660 casos, que representan un 96,5 % del total de casos de la base. La ecuación de regresión planteada será entonces del tipo:

$$\hat{L}_1 = \hat{b}_0 + \hat{b}_3 z_3 + \hat{b}_4 z_4 + \hat{b}_5 z_5$$

El resultado del modelo, de acuerdo con la estimación máximo-verosímil de los parámetros es:

$$L_1 = -0{,}8 + 0{,}504 z_3 + 2{,}930 z_4 - 0{,}055 z_5$$

Si se aplicara este modelo, para predecir cuántos nacimientos serían prematuros, en función de los valores que se observaron para una de las 78.660 ternas [Z_3 Z_4 Z_5], 72.309 nacimientos hubiesen nacido tal como anticipaba el modelo (92 %) mientras que los restantes casos corresponden a 5.392 nacimientos prematuros que el modelo habría pronosticado a término (7 %) y 959 nacidos en término que el modelo habría considerado prematuros (1 %).

De acuerdo con la interpretación de los coeficientes o, mejor dicho, de la exponenciación de sus coeficientes, podemos afirmar:

$\theta_{3/0}$ = 1,655 es el *odds relativo* entre las madres de 40 años o más y las menores de esa edad.

$\theta_{4/0}$ = 18,728 es el *odds relativo* entre los nacimientos de parto múltiple y los nacimientos de parto simple.

$\theta_{5/0}$ = 0,946 es el *odds relativo* entre los nacidos con hermanos y los primeros hijos.

Del estudio de los *odds relativos* aprendemos que el hecho de que el parto sea múltiple afecta sobremanera el tiempo de gestación del niño (mucho más que la edad de la madre) y que partos de madres primerizas tienen más riesgo de que un niño nazca prematuro que aquellas que no lo son. La evidencia se puede volver a apreciar en el cuadro 4, aunque con nuestro modelo de regresión tenemos dimensionado el riesgo relativo de la presencia de cada variable frente al grupo de referencia.

Al realizar las pruebas de hipótesis de significatividad individual sobre los coeficientes de la regresión logística, no puede rechazarse la hipótesis nula de un coeficiente nulo para la variable Z_5, es decir, para la variable que determina si existieron o no hermanos.

Si quitamos esta variable, nos quedamos con 79.623 casos ya que no nos interesan aquellos casos con número de hermanos ignorados. Para esta variante el resultado del modelo, de acuerdo con la estimación máximo-verosímil de los parámetros es: L_1 = -0,788 + 0,4884z_3 + 2,900z_4 y con todos sus coeficientes significativamente estadísticos de acuerdo con las

pruebas de hipótesis de significatividad individual. Si se aplicara este modelo, para predecir cuántos nacimientos serían prematuros, en función de los valores que se observaron para una de los 79.623 pares ordenados [z_3 z_4], 73.173 nacimientos hubiesen nacido tal como anticipaba el modelo (92 %) mientras que los restantes casos corresponden a 5.488 nacimientos prematuros que el modelo habría pronosticado a término (7 %) y 962 nacidos en término que el modelo habría considerado prematuros (1 %).

De acuerdo con la interpretación de los coeficientes o, mejor dicho, de la exponenciación de sus coeficientes, podemos afirmar:

$\theta_{3/0}$ = 1,629 es el *odds relativo* entre las madres de 40 años o más y las menores de esa edad.

$\theta_{4/0}$ = 18,174 es el *odds relativo* entre los nacimientos de parto múltiple y los nacimientos de parto simple.

Vemos que los resultados no difieren respecto del modelo presentado anteriormente.

La otra regresión que se hará considera como variable dependiente el peso al nacer del nacido vivo e incorpora el análisis de todas las variables regresoras.

Para este modelo $L_1 = ln\left(\frac{p1}{1-p_1}\right)$, y p_1 representa la probabilidad de que un nacido vivo nazca con menos de 2500 gramos y, por ende, tenga riesgo de fallecer antes de cumplir un año de edad. En otras palabras, p_1 representa la probabilidad de que el recién nacido tenga riesgo tipo II, de acuerdo con la definición de la página 7. En el análisis de las potenciales variables regresoras deben excluirse todos aquellos casos que contengan algún valor ignorado (respuesta ignorada) en alguna de las variables. Así, nos quedamos con una muestra cabal de 79.047 casos, que representa casi el 97 % del total de casos de la base.

Cuadro 5. Distribución porcentual de los nacimientos por peso al nacer según variables seleccionadas. Ciudad de Buenos Aires. Año 2012.

Características seleccionadas	Peso al nacer (en gramos)		
	Total	Hasta 2500	2500 y más
Total	100,0	7,5	92,5
Sexo			
Varón	100,0	6,9	93,1
Mujer	100,0	8,2	91,8
Edad de la madre			
Hasta 19	100,0	8,0	92,0
20 - 39	100,0	7,2	92,8
40 y más	100,0	12,0	88,0
Tipo de parto			
Simple	100,0	6,0	94,0
Múltiple	100,0	59,1	40,9
Número de hermanos			
0	100,0	6,8	93,2
1 o más	100,0	8,2	91,8
Nivel de instrucción			
Hasta primario incompleto	100,0	7,2	92,8
Secundario incompleto	100,0	7,0	93,0
Secundario completo y más	100,0	7,8	92,2

Fuente: Elaboración propia sobre la base de Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA). Estadísticas vitales.

La lectura del cuadro 5 es análoga a la del cuadro 4. En este caso vemos que el tipo de parto es la más relevante o influyente en el peso al nacer, así

como lo era para el tiempo de gestación. La edad de la madre parece influir más cuando ésta tiene 40 años o más y el creciente número de hermanos favorece el nacimiento con bajo peso al nacer. Por otra parte, el sexo del niño no presenta grandes diferencias aunque entre las mujeres es más frecuente el bajo peso al nacer y el nivel de instrucción no resulta relevante a la luz del estudio del peso al nacer. Del cuadro 5 se optará por quedarnos con las variables Z_1, Z_3, Z_4 y Z_5. Al plantear el modelo es importante tener en cuenta que ahora nuestra base de datos va a contener todos aquellos casos que no presentan valores ignorados para este nuevo conjunto de variables seleccionados. Así, quedan seleccionados 79.579 casos, que representan un 97,6 % del total de casos de la base. La ecuación de regresión planteada será entonces del tipo:

$$\hat{L}_1 = \hat{b}_0 + \hat{b}_1 z_1 + \hat{b}_3 z_3 + \hat{b}_4 z_4 + \hat{b}_5 z_5$$

El resultado del modelo, de acuerdo con la estimación máximo-verosímil de los parámetros es:

$$\hat{L}_1 = -0{,}632 - 0{,}194 z_1 + 0{,}413 z_3 + 3{,}223 z_4 - 0{,}269 z_5$$

Las pruebas de hipótesis de significatividad de los coeficientes han dado todas exitosas, rechazando la hipótesis nula de coeficientes nulos. Si se aplicara este modelo, para predecir cuántos nacimientos presentarían bajo peso, en función de los valores que se observaron para una de los 79.579 cuaternas [z_1 z_3 z_4 z_5], 73.972 nacimientos hubiesen nacido tal como anticipaba el modelo (93 %) mientras que los restantes casos corresponden a 4.654 nacimientos de bajo peso que el modelo habría pronosticado con peso normal (6 %) y 953 nacidos con peso normal que el modelo habría considerado de bajo peso (1 %).

De acuerdo con la interpretación de los coeficientes o, mejor dicho, de la exponenciación de sus coeficientes, podemos afirmar:

$\theta_{1/0} = 0{,}8236$ es el *odds relativo* entre los varones respecto a las mujeres.

$\theta_{3/0}$ = 1,5113 es el *odds relativo* entre las madres de 40 años o más y las menores de esa edad.

$\theta_{4/0}$ = 25,103 es el *odds relativo* entre los nacimientos de parto múltiple y los nacimientos de parto simple.

$\theta_{5/0}$ = 0,764 es el *odds relativo* entre los nacidos con hermanos y los primeros hijos.

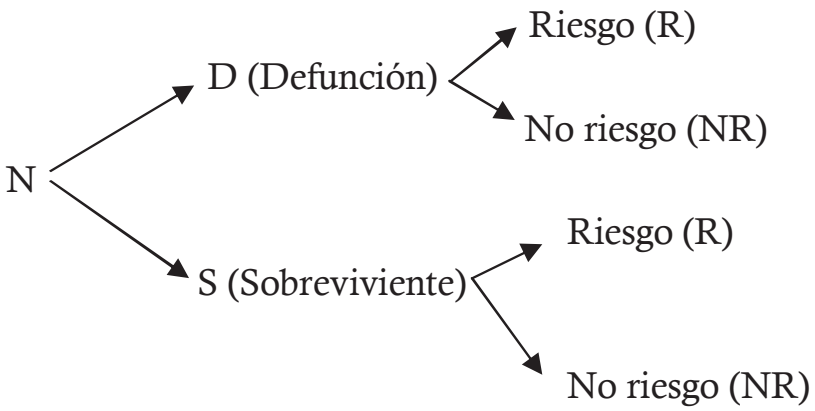
Del estudio de los *odds relativos* aprendemos que el hecho de que el parto sea múltiple afecta sobremanera al peso del niño (mucho más que la edad de la madre) y aún mucho más que al tiempo de gestación y que partos de madres primerizas tienen menor riesgo de que un niño nazca con bajo peso que aquellas que no lo son. Por otra parte, ser varón disminuye el riesgo de nacer con bajo peso. La evidencia se puede volver a apreciar en el cuadro 5, aunque con nuestro modelo de regresión tenemos dimensionado el riesgo relativo de la presencia de cada variable frente al grupo de referencia y además podemos comparar las incidencias con la variable peso al nacer.

RELACIONES CON EL MODELO BAYESIANO

El análisis bayesiano presenta la posibilidad de calcular las probabilidades de dos fenómenos estocásticamente dependientes invirtiendo la causalidad. Para el análisis que nos interesa, se puede, siguiendo una cohorte de N nacimientos en una misma jurisdicción, observarlos durante un año y analizar si ocurre el evento fallecimiento. A su vez, dadas las métricas usuales para los recién nacidos, se puede saber cuánto pesó cada bebé al nacer y con cuántas semanas nació. De esta manera, se puede calcular la probabilidad condicional de que un bebé que se sabe falleció antes de cumplir un año de edad haya nacido con riesgo: P(R/D). Si bien el Informe Estadístico de Defunción indaga en el peso al nacer del niño, resulta alta la cantidad de respuestas ignoradas y, por otra parte, nunca sabríamos el peso al nacer de quienes sobrevivieron al año de vida. Por otra parte, dicho informe nada nos permite conocer acerca de las semanas

de gestación. A partir del análisis bayesiano de las causas, es posible algebraicamente calcular la probabilidad que surge de invertir la causalidad: P(D/R) es decir, la probabilidad de que un nacido con riesgo fallezca antes de cumplir un año de edad. Es decir, sumaríamos al modelo anterior la probabilidad de que el niño nacido en riesgo efectivamente se convierta en una defunción infantil.

Así, nuestro estudio por cohorte presentaría el siguiente esquema:



Donde la probabilidad de que un niño que se sabe nació con riesgo fallezca estará dada por:

$$P(D \mid R) = \frac{P(D) \,.\, P(R \mid D)}{P(D) \,.\, P(R \mid D) + P(S) \,.\, P(R \mid S)}$$

Todos los cálculos intermedios para la expresión se obtendrían a partir del diagrama de árbol anterior y con los datos obtenidos en el estudio de la cohorte. Este resultado permitiría aumentar la prevención y establecer controles adecuados en una jurisdicción donde esta probabilidad resulte especialmente, puesto que las defunciones infantiles –y específicamente las neonatales, que son consecuencia mayoritariamente de las pocas semanas de gestación y/o el bajo peso al nacer–, resultan, en muchos casos, evitables.

COMENTARIOSFINALES

Las distintas alternativas de elección de variables y tipos de modelos son múltiples en un análisis de regresión. Si bien este trabajo no pretende cubrirlas exhaustivamente, el objetivo es presentar la herramienta del análisis de regresión logística como una alternativa a las tablas de contingencia, con una finalidad no sólo descriptiva sino también predictiva. La extensión de este trabajo no lo permitió pero sería interesante plantear modelos de regresión lineal para estimar, por un lado, el tiempo de gestación (en semanas) y, por otro lado, el peso al nacer (en gramos) y luego estudiar la proporción de nacidos prematuros y de nacidos con bajo peso obtenidas bajo el modelo para luego compararlas contra las proporciones observadas y poder utilizar el modelo como predictivo.

Por supuesto que todo modelo excluirá siempre variables que son explicativas y, por otra parte, al no considerar el término de perturbación estocástica, todo intento de crear un modelo determinístico a partir de un fenómeno aleatorio resultará en un error de predicción. Estas advertencias deben estar siempre en la cabeza del investigador pero no por ello debe el mismo dejar de explorar y recorrer el interesantísimo mundo de los modelos estadísticos sofisticados que siempre responde nuestras cuestiones y nos enseña a repensar otras.

BIBLIOGRAFÍA

Aguirre, Alejandro (2009), "La mortalidad infantil y la mortalidad maternal en el siglo XXI" en Papeles de Población, volumen 15, Núm.61, julio-septiembre. Universidad Autónoma de México, pp. 75-99.

Escudero, José C. y Massa, Cristina (2006), "Cifras del Retroceso: El Deterioro Relativo de la Tasa de Mortalidad Infantil de Argentina en la Segunda Mitad del Siglo XX", en *Salud Colectiva*, 2 (3), septiembre-diciembre, pp. 195-223.

Govea Basch, Julián (2010), "Lo que todavía debemos mejorar en el registro de las estadísticas vitales", en *Población de Buenos Aires*, año 7, número 11, abril, pp. 63-72.

Marmot, Michael (2005), "Social determinants of health inequalities" The Lancet 365:1099-1104, disponible en www.thelancet.com.

Mazzeo, V. (2004), *El registro de los hechos vitales de la Ciudad de Buenos Aires*, en Revista Población de Buenos Aires. Año 1, Nº 0. Buenos Aires, Dirección General de Estadística y Censos. Gobierno de la Ciudad de Buenos Aires.

Mazzeo, Victoria (2006), *La inequidad en la salud-enfermedad de la primera infancia. Las políticas de salud y capacidad resolutiva de los servicios en la Ciudad de Buenos Aires*. Tesis doctoral inédita. Costa Rica. Facultad Latinoamericana de Ciencias Sociales. Sede Académica Argentina.

Pagano, Marcello y Gauvreau, Kimberlee (2001), *Fundamentos de Bioestadística*. 2da edición. México, Ed. Thompson Learning.

Pizarro, J. (Editor y compilador) (2011), *Colección de ensayos sobre población y derechos humanos en América Latina*. Serie investigaciones 10. Asociación Latinoamericana de Población.