Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud

Aarón Salinas-Rodríguez, M en C,[1] Betty Manrique-Espinoza, M en C,[1] Sandra G Sosa-Rubí, Dr en C.[1]

Salinas-Rodríguez A, Manrique-Espinoza B, Sosa-Rubí SG. Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. Salud Publica Mex 2009;51:397-406.

Resumen

Objetivo. Describir algunos de los modelos estadísticos para el estudio de variables expresadas como un conteo en el contexto del uso de los servicios de salud. **Material y métodos**. Con base en la Encuesta de Evaluación del Seguro Popular (2005-2006) se calculó el efecto del Seguro Popular sobre el número de consultas externas mediante el uso de los modelos de regresión Poisson, binomial negativo, binomial negativo cero-inflado y Hurdle binomial negativo. Se utilizó el criterio de información de Akaike (AIC) para definir el mejor modelo. **Resultados**. La mejor opción estadística para el análisis del uso de los servicios de salud resultó ser el modelo Hurdle, de acuerdo con sus presuposiciones y el valor del AIC. **Discusión**. La modelación de variables de conteo requiere el empleo de modelos que incluyan una medición de la dispersión. Ante la presencia de exceso de ceros, el modelo Hurdle es una opción apropiada.

Palabras clave: variables de conteo; uso de servicios de salud; modelos de regresión; estadística; México

Salinas-Rodríguez A, Manrique-Espinoza B, Sosa-Rubí SG. Statistical analysis for count data: Use of healthcare services applications. Salud Publica Mex 2009;51:397-406.

Abstract

Objective. To describe some of the statistical models for the study of count variables in the context of the use of health services. **Material and Methods.** We used the Seguro Popular Evaluation Survey to estimate the effect of Seguro Popular on the frequency of use of outpatient health services, using Poisson regression models and negative binomial, zero-inflated negative binomial and the hurdle negative binomial models. We used the Akaike Information Criterion (AIC) to define the best model. **Results.** Results show that the best statistical approach to model the use of health services is the hurdle model, taking into account both the main theoretical assumptions and the statistical results of the AIC. **Discussion.** The modelling of count data requires the application of statistical models to model data dispersion; in the presence of an excess of zeros, the hurdle model is an appropriate statistical option.

Key words: health services; statistical models; statistics; Mexico

(1) Centro de Investigación en Evaluación y Encuestas, Instituto Nacional de Salud Pública. Cuernavaca, Morelos, México.

Fecha de recibido: 16 de febrero de 2009 • **Fecha de aprobado:** 17 de junio de 2009
Solicitud de sobretiros: Betty Manrique-Espinoza. Centro de Investigación en Evaluación y Encuestas, INSP. Av. Universidad 655, col. Santa María Ahuacatitlán. 62100 Cuernavaca, Morelos, México.
Correo electrónico: bmanrique@insp.mx, bs_manrique@yahoo.com.mx

El uso y análisis de variables que se expresan en la forma de un conteo (variable con un valor entero no negativo, $y = 0,1, \ldots$) es frecuente en el ámbito de la salud pública. Esto es en particular cierto para el caso específico del uso de servicios de salud, en los cuales y para el contexto específico de los modelos de regresión, en muchos de los análisis se utiliza como variable de respuesta una variable de conteo; pueden mencionarse los siguientes: el número de visitas al médico (algunas veces desagregadas por especialidad del médico),[1,2] el número de días de estancia hospitalaria,[3] el número de fármacos prescritos,[4] o el número de visitas a los servicios de emergencia,[5] entre otros.

En general, un enfoque estándar para analizar variables de conteo es el modelo de regresión *Poisson*.[6,7] Sin embargo, son conocidas las limitaciones de este enfoque que se desprenden del supuesto restrictivo de que la media y la varianza para la distribución *Poisson* deben ser iguales.[6-8] Por esta razón se ha planteado en la literatura una serie de alternativas, incluidos el uso de la regresión binomial negativa,[9] los modelos de clase latente[5] o los modelos de ecuaciones estructurales.[10]

En este último sentido, el objetivo de este trabajo es describir algunas de las alternativas estadísticas disponibles para el análisis de las variables de conteo y comparar los distintos modelos expuestos para evidenciar sus ventajas y desventajas, todo ello en el contexto general del uso de los servicios de salud.

En la descripción que sigue se exponen los fundamentos de algunos de los modelos de regresión alternativos para variables de conteo y en seguida se ilustra su aplicación mediante el empleo de datos referentes a la utilización de servicios de salud en el contexto de la evaluación de impacto del Seguro Popular (SP). Es importante mencionar que, dado que el texto está dirigido a investigadores de la salud, el nivel de complejidad y notación técnica se mantendrá al mínimo para permitir al lector seguir la secuencia de la exposición; empero, donde sea necesario, se utilizan algunas expresiones o fórmulas y, de manera particular, notación matricial para denotar un vector de valores (para las variables independientes o los coeficientes de regresión) mediante el uso de símbolos con una raya inferior.

Modelos de regresión para datos de conteo

Regresión Poisson

Como se mencionó, el modelo de regresión común para datos de conteo es el de regresión Poisson. Este modelo se ha descrito ampliamente en la bibliografía estadística en general[6-8] y en textos específicos para datos de conteo,[9,11] por lo que aquí sólo se enuncian sus principales características, así como su especificación dentro de una estructura de regresión y en términos de su función de verosimilitud.

Como cualquier modelo de regresión, el de regresión *Poisson* requiere una correcta especificación de la media condicional, es decir, que la distribución condicional para la variable de respuesta sea correctamente especificada así como el parámetro relacionado con su valor esperado. Para la regresión *Poisson* se asume que la distribución condicional de y_i dado $\underline{x}_i$ se distribuye como una variable aleatoria *Poisson* con función de densidad:

$$f(y_i \mid \underline{X}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \qquad (1)$$

y el parámetro para la media condicional:

$$E[y_i \mid \underline{X}_i] = \mu_i = \exp(\underline{x}'_i \beta) \qquad (2)$$

Si la especificación para la distribución condicional de la variable de respuesta, así como la de la media condicional, es correcta, y bajo el supuesto de que se tienen observaciones independientes, entonces se puede utilizar la siguiente función de log-verosimilitud para obtener estimadores consistentes de β:

$$L(\beta) = \sum_{i=1}^{n} \{y_i \underline{x}'_i \underline{\beta} - \exp(\underline{x}'_i \underline{\beta}) - \log(\underline{y_i}!)\} \qquad (3)$$

donde $L(\cdot)$ representa la función de log-verosimilitud.

Por lo regular, los paquetes estadísticos ofrecen estimadores de $\underline{\beta}$ que maximizan esta función. Sin embargo, y para obtener inferencias válidas respecto de $\underline{\beta}$, todavía es necesario verificar el supuesto de que la media y varianza condicionales para este modelo son iguales. Se ha demostrado que, aunque este supuesto no se cumpla (lo cual sucede la mayor parte de las veces), el estimador puntual de $\underline{\beta}$ es aún válido, pero no así el estimador de su error estándar, y por tanto las inferencias respecto de $\underline{\beta}$.[9,11]

Por ello se han propuesto algunas alternativas que conservan el supuesto de una distribución condicional *Poisson*, pero que suavizan el supuesto de que la media y la varianza son iguales, esencialmente al ajustar los errores estándar ante la presencia de sobredispersión (que la varianza sea más grande que la media) o subdispersión (que la varianza sea más pequeña que la media). En particular se ha propuesto el uso de errores

estándar robustos (algo también conocido como método de estimación de *pseudomáxima verosimilitud*),[6] el empleo de un enfoque de quasiverosimilitud[7] o la utilización de errores estándar *bootstrap*.[12]

Aun así, se han detallado bien las limitaciones del modelo de regresión *Poisson*, en particular para modelar datos relativos al uso de servicios de salud. Cameron y Trivedi[11] han mostrado que el modelo no es adecuado por lo siguiente: a) no se cumple el supuesto de equidispersión, debido sobre todo a la presencia de heterogeneidad no observada; b) existe un número excesivo de ceros, esto es, una frecuencia observada de ceros que no es consistente con el modelo *Poisson*; y c) hay multimodalidad, ya que si las observaciones se toman de distintas poblaciones, la distribución observada puede ser multimodal. Sin embargo, esto puede corregirse si el efecto de las covariables es el mismo para las distintas poblaciones.

Regresión binomial negativa

Una de las principales razones por las que el modelo *Poisson* falla es la heterogeneidad no observada. Esto significa que hay factores no observados, en especial características de los individuos, que ejercen alguna influencia sobre la variabilidad relacionada con la variable de respuesta.

El problema es que la heterogeneidad no observada puede tener algunas consecuencias para los procesos de inferencia estadística.[9,13] En primer término, puede introducir sobredispersión y, en segundo, un número excesivo de ceros. Esta heterogeneidad, ignorada por el modelo *Poisson,* puede modelarse de manera explícita mediante el uso de la regresión binomial negativa.[9,11]

Existen al menos dos formas en las que la distribución binomial negativa puede derivarse;[7,9] la más común consiste en asumir que se está ante la presencia de una mezcla de distribuciones, en la cual los datos observados se distribuyen como una *Poisson*, pero se presupone un elemento de heterogeneidad individual no observado (que sigue una distribución gamma en su formulación clásica) que refleja el hecho de que la verdadera media no se ha medido perfectamente.[9] La segunda asume que existe una forma particular de dependencia entre eventos, en el sentido de que la ocurrencia de un evento incrementa la probabilidad de ocurrencia de otros posteriores,[11] aunque esto último sólo puede dilucidarse en estudios longitudinales. Para fines de este artículo se utiliza la primera derivación.

La función de densidad para la distribución binomial negativa se determina por la siguiente expresión:

$$f(y_i \mid \underline{x}_i)= \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \qquad (4)$$

donde α es un parámetro de dispersión ($\alpha \geq 0$) y $\Gamma(\cdot)$ es la función gamma. El parámetro de dispersión α es el que ayuda a definir la relación entre la media y la varianza condicionales, conocida en términos estadísticos como *función varianza*. Si, por ejemplo, $\alpha = 0$, entonces la media y la varianza son iguales y se tiene el modelo *Poisson*. Por otro lado, las funciones más comunes para la relación media-varianza de la distribución binomial negativa son la lineal y la cuadrática.[7]

La función de log-verosimilitud para el modelo de regresión binomial negativa, con función de varianza cuadrática, después de alguna manipulación algebraica de la función de densidad, es la siguiente:

$$L(\underline{\beta},\alpha)= \sum_{i=1}^{n}\left\{\sum_{j=0}^{yi-1}\log(j+\alpha^{-1}) - \log(y_i!)-(y_i+\alpha^{-1})\log (1+\alpha^{*}\exp(\underline{x}'_i\underline{\beta})+y_i\log(\alpha+y_i\underline{x}'_i\underline{\beta})\right\} \qquad (5)$$

que es la utilizada por la paquetería estadística para encontrar los estimadores de β.

Regresión Poisson cero-inflado

Si bien es cierto que el modelo binomial negativo se ha desarrollado para modelar de modo explícito la heterogeneidad no observada, también es cierto que esa misma heterogeneidad es originada por un número *excesivo* de ceros. Es decir, observar más ceros que los que son consistentes con el modelo *Poisson*, cuestión que llega a ser común cuando se analizan datos relativos al uso de los servicios de salud, en especial para la utilización de los servicios hospitalarios.

De manera particular, es posible que el mecanismo aleatorio que dio origen a los datos de conteo muestre una mayor concentración para algún valor específico, que puede ser el cero (como ocurre con el uso de los servicios de salud) o cualquier otro valor positivo. Esto implica que dicho valor tiene una mayor probabilidad de ocurrencia que la especificada por la distribución *Poisson* o cualquier otra distribución.

Para el caso específico de los ceros –y en el contexto del uso de los servicios de salud– es posible que los ceros tengan un doble origen. Por ejemplo, si se pregunta ¿cuántas veces en los últimos 12 meses acudió a solicitar servicios de salud como paciente externo? Los ceros

observados pueden ser originados porque la persona *no utiliza los servicios de salud nunca* o porque *en esos 12 meses* no utilizó los servicios de salud. Esto significa que se tiene una mezcla de distribuciones, por lo que no sería adecuado asumir en esta instancia que los ceros y no ceros se han generado por un mismo proceso.

En todo caso, si se tiene información respecto del origen de los ceros, se pueden estimar los parámetros relacionados con una distribución con valores concentrados en cero. De manera particular, los modelos para datos de conteo cero-inflado le confieren mayor peso a la probabilidad de que la variable de conteo sea igual a cero, al incorporar un mecanismo que divide a los sujetos con valor en cero, y cuya probabilidad es $p(\underline{x}_{1i}\underline{\beta}_1)$, e individuos con valor positivo y probabilidad $1 - p(\underline{x}_{1i}\underline{\beta}_1)$. En consecuencia, la función de probabilidad para un modelo de regresión *Poisson* cero-inflado es una mezcla de un modelo *Poisson* estándar y una distribución con función de masa concentrada en cero.

De manera más formal, y de acuerdo con Lambert,[14] el modelo de regresión *Poisson* cero-inflado puede especificarse como sigue. Primero se definen las siguientes cantidades $\phi_i = \Pr(y_i = 0)$ y $\mu_i = \mu(x_i,\beta)$; la primer a cantidad es la probabilidad de que la variable de conteo sea igual a cero y, la segunda, el valor esperado cuando la variable de conteo asume un valor positivo.

En primera instancia, debe hallarse una manera estadística para expresar ϕ_i, de tal modo que se obtengan sólo valores no negativos. Para esa razón, Lambert propuso una *parametrización* para ϕ_i con base en la función logística y la ubicó en una estructura de regresión, de tal forma que un vector de covariables $\underline{z}_i$ podía utilizarse para modelar ϕ_i, es decir:

$$\begin{aligned} &y_i=0 && \text{con probabilidad } \varphi_i \\ &y_i \sim \Pr(\mu_i) && \text{con probabilidad } (1-\varphi_i) \\ &\varphi_i=\frac{\exp(\underline{z}_i\underline{\gamma})}{1+\exp(\underline{z}_i\underline{\gamma})} \end{aligned} \qquad (6)$$

En tal caso, y en términos de un modelo de regresión, el interés se centra en calcular $(\underline{\gamma},\underline{\beta})$. Si se define de manera adicional una variable indicadora para denotar que y toma el valor de 1 si $y_i = 0$, y cero en cualquier otro caso, entonces la función de log-verosimilitud conjunta, después de omitir las constantes, es:

$$\begin{aligned} L(\underline{\beta},\underline{\gamma})=&\sum_{i=1}^{n}1(y_i=0)\log(\exp(\underline{z}'_i\underline{\gamma})+\exp(-\exp(\underline{x}'_i\underline{\beta}) \\ &+\sum_{i=1}^{n}(1-1(y_i=0))(y_i\underline{x}'_i\underline{\beta}-\exp(\underline{x}'_ib) \\ &-\sum_{i=1}^{n}\log(1+\exp(\underline{z}_i\underline{\gamma}) \end{aligned} \qquad (7)$$

Esta función puede utilizarse para encontrar los estimadores de $\underline{\beta}$ (de interés primordial), así como $\underline{\gamma}$. Vale la pena señalar que aunque aquí se ha hecho la derivación de la función de log-verosimilitud en términos del modelo de regresión *Poisson* cero-inflado, es también posible hacerlo para el modelo de regresión binomial negativa cero-inflado.

Modelos de regresión Hurdle

Al igual que el modelo de regresión para datos de conteo cero-inflado, el modelo *Hurdle* asume la presencia de una mezcla de distribuciones, sólo que dicho modelo tiene una interpretación en dos partes.* La primera se refiere a un modelo con variable de respuesta binaria y la segunda a un modelo de datos de conteo truncado-en-cero. En consecuencia, esta configuración en dos partes permite la interpretación de que los valores positivos se generan toda vez que el umbral (*Hurdle*) en cero se ha cruzado. Por consiguiente, la primera parte modela la probabilidad de que el umbral sea cruzado, mientras que la segunda modela el valor esperado de los valores positivos.

De manera un poco más formal, el modelo *Hurdle* puede especificarse como sigue. Puesto que se asume una mezcla de dos distribuciones, los momentos de dichas distribuciones difieren de una distribución *Poisson* y pueden especificarse como sigue:

$$\begin{aligned} &P(y=0)=f_1(0) \\ &P(y=j)=\frac{1-f_1(0)}{1-f_2(0)}\ f_2(y), \qquad j>0 \end{aligned} \qquad (8)$$

el cual se colapsa al modelo *Poisson* sólo si $f_1(\cdot) = f_2(\cdot)$. En otras palabras, no se asume que los procesos que generaron los ceros y los valores positivos sean iguales.

En términos más simples, el modelo *Hurdle* es una mezcla finita generada mediante la combinación de una función de densidad que origina los ceros, y otra función de densidad que produce los valores positivos. De allí que los momentos del modelo *Hurdle* están determinados por la probabilidad de cruzar el umbral y por los momentos de la función de densidad truncada-en-cero, es decir:

$$E(y|\underline{x})=P[y>0|\underline{x}]E_{y>0}[y|y>0,\underline{x}] \qquad (9)$$

Por lo tanto, el modelo global, con ceros y valores positivos, se utiliza para estimar los parámetros vinculados con ambas densidades.

* Por esa razón esta clase de modelos se conoce comúnmente como *two-part hurdle models*.

Para estimar los parámetros del modelo *Hurdle* se puede utilizar alguna distribución de probabilidad para variables de conteo (*Poisson* o binomial negativa), sólo que sin perder de vista que se tiene una variable truncada-en-cero.

Si se utiliza, por ejemplo, la distribución binomial negativa con función varianza cuadrática y se especifican las siguientes cantidades:

$\mu_{1i} = exp(x_i, \beta_1)$ media condicional para el caso de que el conteo sea igual a cero
$\mu_{2i} = exp(x_i, \beta_2)$ media condicional para el caso de que el conteo sea positivo, j = {1,2, …}
$1[y_i \in J] = 1$ Si y_i es positiva
$1[y_i \in J] = 0$ Si $y_i = 0$

se pueden calcular entonces las siguientes probabilidades:

$$P(y_i=0|\underline{x}_i)=(1+\alpha_1\mu_{1i})^{-1/\alpha 1} \quad (10)$$

$$1-P(y_i=0|\underline{x}_i)=\sum_{yi\in J} h(y_i|\underline{x}_i)=1-(1+\alpha_1\mu_{1i})^{-1/\alpha 1} \quad (11)$$

$$P(y_i|\underline{x}_i,y_i>0)=\frac{\Gamma(y_i+\alpha_2^{-1})}{\Gamma(y_i+1)\Gamma(\alpha_2^{-1})}\left(\frac{1}{(1+\alpha_2\mu_{2i})^{1/\alpha 2}-1}\right)^{\underline{x}}\left(\frac{\mu_{2i}}{\mu_{2i}\alpha_2}\right)^{yi} \quad (12)$$

En (10) se define la probabilidad de que la variable de conteo sea igual a cero, mientras que en (11) se define la probabilidad de que el umbral sea cruzado. Por último, (12) es la distribución binomial negativa, con función de varianza cuadrática, truncada-en-cero.

Si se emplean estas probabilidades se puede especificar la función de log-verosimilitud en dos partes, a saber:

$$L_1(\underline{\beta}_1,\underline{\alpha}_1)=\sum_{i=1}^{n}[1-1(y_i\in J)]\log[P(y_i=0|\underline{x}_i)] + \sum_{i=1}^{n}1(y_i\in J)\log[1-P(y_i=0|\underline{x}_i)] \quad (13)$$

y

$$L_2(\underline{\beta}_2,\underline{\alpha}_2)=\sum_{i=1}^{n}1[y_i\in J]\log[P(y_i|\underline{x}_i y_i>0] \quad (14)$$

así que

$$L(\underline{\beta}_1, \underline{\beta}_2,\underline{\alpha}_1,\underline{\alpha}_2)= L_1(\underline{\beta}_1,\underline{\alpha}_1)+ L_2(\underline{\beta}_2,\underline{\alpha}_2) \quad (15)$$

Aquí, $L(\underline{\beta}_1,\underline{\alpha}_1)$ es la log-verosimilitud para el proceso binario que divide a las observaciones en ceros y valores positivos; y $L(\underline{\beta}_2,\underline{\alpha}_2)$ es la log-verosimilitud relacionada con la parte truncada-en-cero con una distribución binomial negativa para los valores positivos. Dado que se asume que estos dos procesos son independientes,[11] la función global puede maximizarse al trabajar por separado con las dos respectivas funciones de log-verosimilitud.

Mullahy[15] fue quien primero desarrolló esta clase de modelos *Hurdle* en el contexto del consumo diario de bebidas, aunque lo hizo mediante el uso de las distribuciones *Poisson* y geométrica, mientras que los resultados presentados aquí se han elaborado a partir de la distribución binomial negativa según Cameron y Trivedi.[11]

Material y métodos

Con el objetivo de ilustrar la aplicación de los modelos descritos hasta aquí, se utilizará información concerniente al uso de servicios de salud dentro de la Encuesta de Evaluación del Seguro Popular (2005-2006). King y colaboradores han descrito los detalles sobre el diseño y características de la encuesta.[16] El estudio recibió la aprobación de las comisiones de Ética y Bioseguridad del Instituto Nacional de Salud Pública.

Entre otros motivos, la Encuesta de Evaluación del Seguro Popular se diseñó para determinar si la afiliación al Seguro Popular tendría algún efecto sobre el uso de los servicios de salud, en el entendido de que la oferta de servicios del SP de manera gratuita* podría permitir que las familias afiliadas obtuvieran un mayor acceso a los servicios de salud.

En ese sentido, la encuesta cuenta con información relativa al uso de los servicios de salud tanto para la consulta externa como para los datos de hospitalización, además de contar con dos mediciones, la basal y la de seguimiento, llevada a cabo 11 meses después. En particular, y para los análisis aquí presentados, se pretende evaluar el efecto del SP sobre la probabilidad de uso de los servicios de salud (consulta externa) y sobre la intensidad de uso de dichos servicios (número de consultas), mediante la aplicación del estimador de diferencias-en-diferencias.[17] Para ello se tiene una muestra de 21 537 individuos que cuentan con información sobre estas variables.

Para los modelos de regresión utilizados se han tomado en cuenta, además del estatus de afiliación al SP, aquellas covariables consideradas en las publicaciones como relevantes para explicar el uso de los servicios de

* Es importante mencionar que el costo del SP es nulo para las familias que se ubiquen en los dos primeros deciles de ingreso, mientras que para las demás familias se realiza una aportación proporcional al ingreso.

salud, tales como a) nivel individual: sexo, edad, nivel educativo, condición indígena, estado de unión, condición laboral, condición de aseguramiento, enfermedad crónica, percepción del estado de salud; b) nivel hogar: tamaño del hogar, nivel socioeconómico, afiliación a *Oportunidades*; y c) nivel localidad: condición de ruralidad.

Por último, para comparar los modelos analizados se utiliza el criterio de información de Akaike (AIC).[6,7] Y, con la finalidad de homogeneizar la comparación entre modelos en relación con el estimador vinculado con la afiliación al SP, los resultados se presentan en términos de efectos marginales. Todos los análisis incorporan el hecho de que se tiene un diseño pareado a nivel localidad mediante la inclusión de un efecto fijo (variable indicadora) a nivel de localidad, y se realizaron en el paquete estadístico R.[18]

Resultados

La figura 1 muestra la distribución para el número de consultas externas. Es posible observar una elevada concentración de valores en cero, un total de 10 552 observaciones (38.89%), lo cual puede introducir mayor variabilidad de la que nominalmente asume el modelo *Poisson*. De manera adicional, los datos contienen más ceros de los que serían predichos por un modelo *Poisson* con media de 1.58. Es decir, bajo la presuposición de una distribución *Poisson* se esperaría observar un total de $27\,135 * e^{1.58} = 5\,589$ ceros, 47% menos que la frecuencia observada.

En el cuadro I se presentan los resultados para los modelos *Poisson*, binomial negativo (función de varianza lineal, NB1) y binomial negativo (función de varianza cuadrática, NB2). Para los tres modelos, el efecto del SP sobre el número de consultas externas es altamente significativo. Por otro lado, los resultados muestran, según el valor de AIC, que los modelos de regresión binomial negativa son una mejor opción que el modelo *Poisson*, lo cual se corrobora también al llevar a cabo la prueba del cociente de verosimilitudes, LR_{NB1}=12,145, LR_{NB2}=12,115, ambas significativas (p<0.001). Además, los estimadores del parámetro de dispersión (α) para NB1 y NB2 son estadísticamente distintos de cero (p<0.001), lo que significa que existe evidencia de sobredispersión. Finalmente, al comparar estos dos modelos, el valor de AIC muestra que es preferible el modelo NB1.

Los resultados de los modelos de regresión binomial negativa cero-inflado y *Hurdle*, con función de varianza lineal, se muestran en el cuadro II. Como se mencionó en la introducción, estos modelos tienen una especificación en términos de mezcla de distribuciones, razón por la cual los resultados de ambos modelos se encuentran divididos en dos secciones.

Para el modelo cero-inflado, la primera sección incluye los resultados al modelar la probabilidad de ocurrencia de los ceros y se han incluido aquellas variables que pueden explicar dicha ocurrencia.* Es posible

* Aunque se ha incluido todo el vector de covariables utilizadas para modelar el valor esperado del conteo E(Y), no existe restricción alguna para incluir un subconjunto del total de covariables, aunque es importante notar que la elección de estas covariables debe realizarse, como siempre sucede, con un marco teórico que justifique su elección.

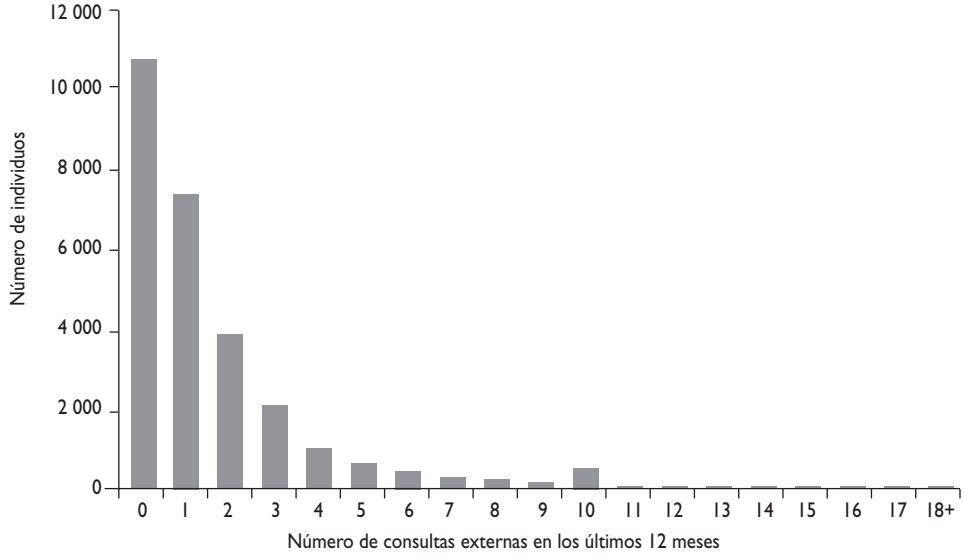

Figura 1. Distribución para el número de consultas externas. Encuesta de Evaluación de Impacto del Seguro Popular. México 2005-2006

Cuadro I

EFECTO DEL SP SOBRE EL NÚMERO DE CONSULTAS EXTERNAS: MODELOS *POISSON*, BINOMIAL NEGATIVO (FUNCIÓN DE VARIANZA LINEAL) NB1, BINOMIAL NEGATIVO (FUNCIÓN DE VARIANZA CUADRÁTICA) NB2.[§] ENCUESTA DE EVALUACIÓN DE IMPACTO DEL SEGURO POPULAR. MÉXICO, 2005-2006

	Poisson	*NB1*	*NB2*
Seguro Popular	0.19203***	0.20530***	0.19213***
	(0.02799)[&]	(0.02807)	(0.03007)
Nivel individual			
Sexo (Mujer)	0.40837***	0.40848***	0.45240***
	(0.03245)	(0.02722)	(0.02980)
Edad	-0.00351	-0.00006	-0.00586*
	(0.00295)	(0.00265)	(0.00294)
Edad 2	0.00008**	0.00005	0.00011***
	(0.00003)	(0.00003)	(0.00003)
Condición indígena	-0.21898***	-0.20787***	-0.21851***
	(0.04775)	(0.04662)	(0.04807)
Estado de unión	0.12985***	0.10770***	0.14160***
(unido o no unido)	(0.02149)	(0.01787)	(0.02169)
Condición laboral	-0.06461**	-0.05075**	-0.06354**
	(0.02333)	(0.01917)	(0.02206)
Condición de aseguramiento	0.17224***	0.15657***	0.17524***
	(0.02632)	(0.02322)	(0.02515)
No. de enfermedades crónicas	0.19345***	0.16871***	0.21500***
	(0.01057)	(0.00993)	(0.01210)
Percepción de estado de salud	0.12601***	0.12556***	0.13004***
(regular/malo/muy malo o bueno/muy bueno)	(0.01769)	(0.01575)	(0.01745)
Nivel hogar			
No. de integrantes en el hogar	-0.03344***	-0.02853***	-0.03264***
	(0.00460)	(0.00393)	(0.00467)
Índice de bienes	0.05444***	0.05388***	0.06317***
	(0.01349)	(0.01170)	(0.01300)
Afiliación a *Oportunidades*	0.10030***	0.13161***	0.10749***
	(0.02275)	(0.02122)	(0.02249)
Nivel localidad			
Condición de ruralidad	0.37697*	0.49604**	0.39114**
	(0.15148)	(0.18499)	(0.14079)
Intercepto	-0.55608***	-0.73394***	-0.58360***
	(0.16204)	(0.19505)	(0.15008)
α (parámetro de dispersión)		1.39145***	0.82674***
		(0.06662)	(0.04070)
log-verosimilitud	-50698.77	-44625.98	-44641.11
AIC	101535.5	89391.97	89422.22

[§] Modelos ajustados por número basal de consultas externas (2005), nivel educativo, condición basal de exposición (control o tratamiento) y pareamiento.
[&] El error estándar aparece entre paréntesis.
* $p<0.05$; ** $p<0.01$; *** $p<0.001$.

advertir que el hecho de ser mujer implica una menor probabilidad de *no utilizar* los servicios de salud, al igual que tener una mala/muy mala percepción del estado de salud, entre otros datos. En cuanto a los valores positivos, que en este contexto tienen una interpretación de intensidad de uso, los datos muestran que el efecto del SP es altamente significativo para incrementar el uso de servicios de consulta externa. Por su valor de AIC, este modelo debe preferirse sobre los modelos *Poisson*, NB1 y NB2. En cuanto al parámetro de dispersión, su valor es significativamente distinto de cero, lo que evidencia una vez más la presencia de sobredispersión.

Por su parte, los resultados del modelo *Hurdle* están presentados en dos partes; en la primera se modela la probabilidad de uso, mientras que en la segunda se modela la intensidad de uso. Aquí, el efecto del SP es significativo tanto para la probabilidad de uso de los servicios de consulta externa como para incrementar el número promedio de consultas externas. También es significativo (distinto de cero) el valor del parámetro de dispersión, y el valor de AIC de este modelo es el más bajo de los cinco modelos comparados, lo cual sugiere que el modelo *Hurdle* es el que mejor se ajusta a los datos observados.

Por último, en el cuadro III figuran los resultados de la comparación entre modelos en relación con el efecto del SP en términos de efectos marginales. Los modelos *Poisson*, NB1, NB2 y NB cero-inflado muestran estimadores similares, en los cuales se puede observar en general que el número promedio de visitas de consulta externa para familias beneficiarias del SP es 0.3 veces más grande respecto de las familias no beneficiarias, en tanto que para el modelo *Hurdle* esta cantidad es de 0.15.

Discusión

Las variables expresadas como un conteo pueden hallarse prácticamente en cualquier ámbito de aplicación de la salud pública. En el contexto particular del uso de servicios de salud, su aplicación se ha extendido, ya que numerosos estudios utilizan como variable de respuesta una variable de conteo. A pesar de que el modelo *Poisson* es el más empleado cuando se tiene una variable de conteo, no existe una práctica generalizada en cuanto al tipo del modelo de regresión que debe utilizarse, aun cuando existen una gran variedad de modelos que pueden utilizarse.[9,11] En este artículo se han revisado algunas de las medidas más empleadas para construir un modelo de regresión cuando la variable de interés es un conteo.

Los resultados muestran que los estimadores puntuales de los cinco modelos son similares, aunque difieren en el valor de los estimadores de sus errores estándar, y por tanto en los niveles de significancia, sobre todo para el modelo *Poisson*. Este fenómeno se reconoce en la bibliografía[7,9,11] y es atribuible a la presencia de sobredispersión; en consecuencia, los resultados concuerdan con lo esperado desde el punto de vista teórico.

En este sentido, todos los modelos que han incorporado un parámetro adicional para modelar la dispersión tienen un valor de AIC menor al del *Poisson*, lo que confirma la importancia de tomar en cuenta el hecho de que la media y varianza para una variable de conteo no son casi nunca iguales.[9] Se debe resaltar que en este artículo se ha determinado la significancia del valor del parámetro de dispersión α a partir de la estimación hecha por los modelos mismos de regresión binomial negativa, aunque en las publicaciones estadísticas existen diversas pruebas de hipótesis relacionadas con este estimador.[19-21]

Los resultados muestran que el modelo *Hurdle* es el que exhibe un mejor ajuste para los datos de uso de servicios de consulta externa, de acuerdo con sus supuestos y el valor de AIC, y que de los otros cuatro modelos, el binomial negativo cero-inflado podría ser una opción viable, sin perder de vista que puede sobreestimar el efecto de las covariables de interés. Es decir, y de acuerdo con los resultados en relación con los valores de AIC, es posible pensar que el efecto del SP lo sobre-estiman los modelos *Poisson*, NB1, NB2, y NB cero-inflado, pero aun así el efecto es significativo bajo el modelo *Hurdle*; esto indica que el SP incrementa el número de consultas externas, además de la probabilidad de uso de estos servicios de consulta externa. Lo anterior es consistente con los hallazgos notificados en la bibliografía, en la cual se analiza si el modelo *Hurdle* es un mejor punto de partida que los modelos de regresión binomial negativa.[1,3,4]

Por último, esta presentación se ha restringido a cinco modelos de regresión específicos para datos de conteo y han quedado fuera otras propuestas, muchas de ellas detalladas en las revisiones monográficas de Cameron-Trivedi[11] y Hilbe.[9] No obstante, hay tres que en particular vale la pena mencionar. Primero, y sin dejar de reconocer la complejidad vinculada con el uso de servicios de salud, se ha hecho uso de los modelos de ecuaciones estructurales,[10,22,23] los cuales han mostrado cierta flexibilidad para modelar las relaciones de causalidad entre las variables relacionadas con el uso de servicios de salud. En segundo lugar aparecen los modelos de clase latente o mezclas finitas, los cuales se han aplicado también en el estudio del uso de servicios de salud.[24,25] De manera específica, estos modelos son útiles porque permiten modelar de forma directa la heterogeneidad no observada; de ahí la presencia de sobredispersión o subdispersión. Por último, en fecha

Cuadro II

EFECTO DEL SP SOBRE EL NÚMERO DE CONSULTAS EXTERNAS: MODELOS, BINOMIAL NEGATIVO (FUNCIÓN DE VARIANZA LINEAL) CERO-INFLADO, HURDLE BINOMIAL NEGATIVO (FUNCIÓN DE VARIANZA LINEAL) NB HURDLE.[§] ENCUESTA DE EVALUACIÓN DE IMPACTO DEL SEGURO POPULAR. MÉXICO, 2005-2006.

	NB cero-inflado		NB *Hurdle*	
	Prob(Y=0)	E(Y)	Prob(Y>0)	E(Y)
Seguro Popular	-0.39591	0.15653 ***	0.37934***	0.11174**
	(0. 24916)	(0.03061)	(0.05680)	(0.03556)
Nivel individual				
Sexo (Mujer)	-1.23354***	0.33612***	0.69861***	0.32819***
	(0.18702)[&]	(0.03288)	(0.04107)	(0.04037)
Edad	-0.01540	-0.00624	0.00328	-0.01275**
	(0.02980)	(0.00331)	(0.00490)	(0.00422)
Edad 2	-0.00001	0.00009 **	0.00005	0.00016***
	(0.00029)	(0.00003)	(0.00005)	(0.00004)
Condición indígena	0.48029	-0.18384***	-0.32643***	-0.19187***
	(0.30899)	(0.04151)	(0.08095)	(0.05220)
Estado de unión	-0.12280	0.12689 ***	0.14218***	0.15310***
(unido o no unido)	(0.15961)	(0.02366)	(0.03072)	(0.03081)
Condición laboral	0.28543	-0.04454	-0.06850*	-0.05864
	(0.17538)	(0.02399)	(0.03457)	(0.03177)
Condición de aseguramiento	-0. 81767**	0.12917***	0.21176***	0.15949***
	(0. 27704)	(0.02850)	(0.04450)	(0.03594)
No. de enfermedades crónicas	-0.28999	0.20269***	0.26391***	0.23727***
	(0.15055)	(0.01252)	(0.02695)	(0.01591)
Percepción del estado de salud	-0.17841*	0.10710***	0.20559***	0.09969***
(regular/malo/muy malo o bueno/muy bueno)	(0.21888)	(0.01883)	(0.03260)	(0.02259)
Nivel hogar				
No. de integrantes en el hogar	0.03308	-0.03079***	-0.03988***	-0.03454***
	(0.03337)	(0.00463)	(0.00724)	(0.00693)
Índice de bienes	-0. 27540*	0.03984*	0.11248***	0.03809*
	(0. 12465)	(0.01582)	(0.02196)	(0.01916)
Afiliación a *Oportunidades*	-0.59550**	0.06755**	0.31506***	0.00128
	(0. 22541)	(0.02267)	(0.04408)	(0.02840)
Nivel localidad				
Condición de ruralidad	-0.63582	0.33649*	0.96210**	-0.01126
	(0.56287)	(0.14320)	(0.31624)	(0.03913)
Intercepto	0.34177	-0.29146	-1.61700***	-0.12166
	(0.79014)	(0.15831)	(0.32567)	(0.12897)
α (parámetro de dispersión)		0.71278 ***		1.60801***
		(0.05196)		(1.13062)
log-verosimilitud		-44387.59		-44108.72
AIC		88969.19		88495.44

[§] Modelos ajustados por número basal de consultas externas (2005), nivel educativo, condición basal de exposición (control o tratamiento) y pareamiento.
[&] El error estándar aparece entre paréntesis.
* $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Cuadro III

Comparación de modelos: efectos marginales del SP sobre el número de consultas externas.§ Encuesta de Evaluación de Impacto del Seguro Popular. México, 2005-2006.

	$d\lambda_i/dx$&	Error estándar	z	p	*IC*95%
Poisson	0.31	0.06	4.97	0.001	0.19-0.44
NB1	0.31	0.07	4.79	0.001	0.18-0.44
NB2	0.30	0.06	4.88	0.001	0.18-0.42
NB cero-inflado	0.31	0.06	5.04	0.001	0.19-0.43
NB *Hurdle*	0.15	0.05	3.22	0.001	0.06-0.25

§ Modelos ajustados por las covariables mostradas en el cuadro I.
& Efectos marginales calculados en el valor de la media para el vector de las covariables incluidas en los modelos de regresión.

reciente se ha propuesto el uso combinado de los modelos de clase latente con el modelo *Hurdle*.[26] En esta clase de modelos se pretende modelar la heterogeneidad no observada mediante la introducción de clases latentes, pero se conserva el hecho de que el uso de servicios de salud puede ser el resultado de un proceso en dos partes: la decisión de utilizar y la frecuencia del uso.

Agradecimientos

Los autores agradecen a la Secretaría de Salud, en particular a la Dirección General de Evaluación del Desempeño, por la autorización para el uso de la base de datos de la Encuesta de Evaluación del Seguro Popular, 2005-2006. Los resultados presentados constituyen un análisis y edición adicional de los autores. Los puntos de vista expresados no representan una posición oficial de la Secretaría de Salud, y son sólo responsabilidad de los autores.

Referencias

1. Pohlmeier W, Ulrich V. An econometric model of the two-part decision process in the demand for health. J Hum Res 1995;30(2):339-361.
2. Santos-Silva JMC, Windmeijer F. Two-part multiple spell models for health care demand. J Econo 2001;104(1):67-89.
3. Gerdtham UG. Equity in health care utilization: further tests based on Hurdle models and swedish micro data. Health Econo 1997;6:303-319.
4. Grootendorst PV. A comparison of alternative models of prescription drug utilization. Health Econo 1995;4:183-198.
5. Deb P, Trivedi PK. Demand for medical care by the elderly: a finite mixture approach. J Appl Econo 1997;12:313-336.
6. Hardin J, Hilbe J. Generalized linear models and extensions. 2nd ed. Texas: Stata Press, 2007.
7. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. New York: Chapman & Hall, 1989.
8. Dobson A, Barnett A. An introduction to generalized linear models. 3rd ed. New York: Chapman & Hall/CRC, 2008.
9. Hilbe J. Negative binomial regression. Cambridge: Cambridge University Press, 2007.
10. Congdon P, Almog M, Curtis S, Ellerman R. A spatial structural equation modelling framework for health count responses. Stat Med 2007;26:5267-5284.
11. Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge: Cambridge University Press, 1998.
12. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall/CRC, 1993.
13. Jones AM. Applied health economics. Oxford: Routledge, 2007.
14. Lambert D. Zero-inflated poisson regression with an application to defects in manufacturing. Technometrics 1992;34:1-14.
15. Mullahy J. Specification and testing of some modified count data models. J Econo 1986;33:341-365.
16. King G. A "politically robust" experimental design for public policy evaluation, with application to the Mexican Universal Health Insurance Program. J Pol Anal Manag 2007;26(3):479-506.
17. Ashenfelter O, Card D. Using the longitudinal structure of earnings to estimate the effect of training programs. Rev Econo Stat 1985;67(4):648-660.
18. R Development Core Team. [sitio de internet]. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2008. ISBN 3-900051-07-0. [Consultado 2009 feb 10]. Disponible en http://www.R-project.org.
19. Cameron AC, Trivedi PK. Econometric models based on count data: comparisons and applications of some estimators and tests. J Appl Econo 1986;(1):29-53.
20. Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the poisson model. J Econo 1990;46:347-364.
21. Dean C, Lawless JF. Tests for detecting overdispersion in poisson regression models. J Am Stat Ass 1989;84:467-472.
22. Ellencweig AY, Pagliccia N. Utilization patterns of cohorts of elderly clients: a structural equation model. Health Serv Res 1994;29:225-245.
23. Andersen AS, Laake P. A causal model for physician utilization: analysis of Norwegian data. Med Care 1983;21:266-278.
24. Deb P, Trivedi PK. The structure of demand for health care: latent class versus two-part models. J Health Econo 2002;21:601-625.
25. Jimenez-Martin S. Latent class versus two-part models in the demand for physician services across the European Union. Health Econo 2002;11:301-321.
26. Bago d'Uva T. Latent class models for health care utilization. Health Econo 2006;15:329-343.