Modelos de regresión para variables expresadas como una proporción continua

Aarón Salinas-Rodríguez, Psic, M en C,[1] Ricardo Pérez-Núñez, MC, M en C,[2]
Leticia Ávila-Burgos, MC, M en C, Dra en C.[2]

**Salinas-Rodríguez A, Pérez-Núñez R, Avila-Burgos L.
Modelos de regresión para variables expresadas como una proporción continua.
Salud Publica Mex 2006;48:395-404.**

Resumen

Objetivo. Describir algunas de las alternativas estadísticas disponibles para el estudio de proporciones continuas y comparar los distintos modelos que existen para evidenciar sus ventajas y desventajas, mediante su aplicación a un ejemplo práctico del ámbito de la salud pública. **Material y métodos**. Con base en la Encuesta Nacional de Salud Reproductiva realizada en el año 2003, se modeló la proporción de cobertura individual en el programa de planificación familiar –propuesta en un estudio previo realizado en el Instituto Nacional de Salud Pública en Cuernavaca, Morelos, México (2005)– mediante el uso de los modelos de regresión normal, gama, beta y de quasi-verosimilitud. La variante del criterio de información de Akaike (AIC) que propusieron McQuarrie y Tsai se utilizó para definir el mejor modelo. A continuación, y mediante simulación (enfoque Monte Carlo/cadenas de Markov), se generó una variable con distribución beta para evaluar el comportamiento de los cuatro modelos al variar el tamaño de la muestra desde 100 hasta 18 000 observaciones. **Resultados**. Los resultados muestran que la mejor opción estadística para el análisis de proporciones continuas es el modelo de regresión beta, de acuerdo con sus supuestos y el valor de AIC. La simulación mostró que a medida que aumenta el tamaño de la muestra, el modelo gama y, en especial, el modelo de quasi-verosimilitud se aproximan en grado significativo al modelo beta. **Conclusiones**. Para la modelación de proporciones continuas se recomienda emplear el enfoque paramétrico de la regresión beta y evitar el uso del modelo normal. Si se tiene un tamaño de muestra grande, el uso del enfoque de quasi-verosimilitud representa una buena alternativa.

Palabras clave: proporciones continuas; modelos de regresión; estadística; México

**Salinas-Rodríguez A, Pérez-Núñez R, Avila-Burgos L.
Regression models for variables expressed as a continuous proportion.
Salud Publica Mex 2006;48:395-404.**

Abstract

Objective. To describe some of the statistical alternatives available for studying continuous proportions and to compare them in order to show their advantages and disadvantages by means of their application in a practical example of the Public Health field. **Materials and Methods.** From the National Reproductive Health Survey performed in 2003, the proportion of individual coverage in the family planning program –proposed in one study carried out in the National Institute of Public Health in Cuernavaca, Morelos, Mexico (2005)– was modeled using the Normal, Gamma, Beta and quasi-likelihood regression models. The Akaike Information Criterion (AIC) proposed by McQuarrie and Tsai was used to define the best model. Then, using a simulation (Monte Carlo/Markov Chains approach) a variable with a Beta distribution was generated to evaluate the behavior of the 4 models while varying the sample size from 100 to 18 000 observations. **Results.** Results showed that the best statistical option for the analysis of continuous proportions was the Beta regression model, since its assumptions are easily accomplished and because it had the lowest AIC value. Simulation evidenced that while the sample size increases the Gamma, and even more so the quasi-likelihood, models come significantly close to the Beta regression model. **Conclusions.** The use of parametric Beta regression is highly recommended to model continuous proportions and the normal model should be avoided. If the sample size is large enough, the use of quasi-likelihood model represents a good alternative.

Keywords: regression models; continuous proportions; statistics; Mexico

(1) Centro de Investigación en Salud Poblacional, Instituto Nacional de Salud Pública. Cuernavaca, Morelos, México
(2) Centro de Investigación en Sistemas de Salud, Instituto Nacional de Salud Pública. Cuernavaca, Morelos, México

Fecha de rebido: 5 de enero de 2006 • **Fecha de aprobado:** 7 de junio de 2006
Solicitud de sobretiros: Ricardo Pérez-Núñez. Centro de Investigación en Sistemas de Salud, INSP. Av. Universidad 655, Col. Sta. María Ahuacatitlán. 62508 Cuernavaca, Morelos, México.
Correo electrónico: rperez@correo.insp.mx o riquiperez@lycos.com

Estudios de las más diversas disciplinas se encuentran con cierta frecuencia ante la necesidad de explicar una variable expresada como una proporción, porcentaje, tasa o fracción en el rango continuo (0,1). En economía, por ejemplo, se han estudiado los factores que influyen en la proporción de hogares que se suscriben a la televisión por cable. De manera paralela, la proporción de impurezas en los compuestos químicos es de interés cotidiano para la física y la química. Mientras que en estudios sobre preferencias electorales se analizan las tasas de participación ciudadana y las variables que puedan explicarlas, en el ámbito educativo y de desempeño académico se intenta explicar la proporción de aciertos en pruebas o *tests* estandarizados. También el área de la salud pública se ha enfrentado a la necesidad de modelar la proporción de cobertura en programas de salud con el fin de identificar las características sociodemográficas y económicas relacionadas con el hecho de que una mujer esté cubierta.* Una descripción detallada de estos y otros usos para una variable expresada como proporción puede encontrarse en Johnson y colaboradores,[1] Hviid y Villadsen[2] y Bury.[3]

Johnson y colaboradores[1] expusieron las propiedades de la distribución de probabilidad de este tipo de variables; estos investigadores muestran que la distribución beta puede usarse para modelar proporciones, ya que su densidad puede tomar diferentes formas según sean los valores de los dos parámetros de forma que indexan a la distribución. Sin embargo, ni en este ni en otros textos de probabilidad se describen situaciones en las que es necesario imponer una estructura de regresión a este tipo de variables. Dada la complejidad que representa el análisis estadístico de estas mismas, los investigadores de la salud deben conocer las alternativas estadísticas disponibles para modelarlas, así como los supuestos bajo los cuales es válida la aplicación de estas alternativas. En este sentido, el objetivo de este trabajo es describir algunas de las alternativas estadísticas disponibles para el estudio de las proporciones continuas y comparar los distintos modelos que existen para evidenciar sus ventajas y desventajas, mediante su aplicación a un ejemplo práctico de salud pública relacionado con el análisis de la cobertura del programa de planificación familiar.

* Pérez-Núñez R, Salinas-Rodríguez A, Avila-Burgos L, Mojarro-Íñiguez MG, Medina-Solís CE, Schiavon R *et al*. Cobertura y financiamiento de la Planificación Familiar en México: hallazgos de la Encuesta Nacional de Salud Reproductiva 2003. Documento no publicado.

Varias propuestas metodológicas se han elaborado para analizar variables en el rango continuo (0,1). A continuación se exponen los fundamentos teóricos de los enfoques propuestos al respecto. En virtud de que este trabajo está destinado a investigadores de la salud, el nivel de complejidad y notación técnica se mantiene al mínimo para permitir al lector seguir la secuencia de la exposición; empero, donde sea necesario se hará uso de algunas expresiones o fórmulas, o ambas cosas.

Modelos de regresión alternativos

Distribución normal

Por mucho, la práctica más común para modelar proporciones continuas ha sido la aplicación del método de estimación de regresión por mínimos cuadrados ordinarios (OLS, por sus siglas en inglés). Sea o no que se utilice asumiendo los supuestos distribucionales, se aduce con regularidad un argumento asintótico para su aplicación, en el sentido de que tamaños de muestra *grandes* permiten generar cuantificaciones válidas y confiables. Sin embargo, como apunta Godfrey,[4] cuando se analizan la prueba *t* o la prueba *F* correspondientes se asume una distribución normal sin importar cuál sea el tamaño de muestra, lo mismo que al emplear algunas pruebas de heteroscedasticidad (como la de Breusch-Pagan o Cook-Weisberg). Además del supuesto distribucional, el modelo de regresión lineal requiere el supuesto de homoscedasticidad y linealidad. En términos conceptuales, este enfoque es erróneo por varios motivos. Primero, es obvio que una proporción no está definida sobre el dominio de los números reales, que es el dominio sobre el cual se define la distribución normal. Segundo, como se usa una variable acotada en el intervalo (0,1) la función de la esperanza condicional no es lineal; dicho en otras palabras, implica que no existe una relación lineal entre la media de la variable de interés y las variables predictoras.[5] Tercero, por la misma razón, la varianza es heteroscedástica ya que se acerca a cero cuando la media se aproxima a los límites de su dominio, es decir, la varianza depende de la media,[6] lo que subestima los estimadores puntuales de los coeficientes.[4] Cuarto, como apuntan Ferrari y Cribari-Nieto,[7] si la variable de respuesta está restringida al intervalo (0,1), el método de OLS podría generar valores ajustados que excedan las cotas inferior y superior, además de que la distribución de probabilidad de las proporciones es casi siempre asimétricas y, por lo tanto, las inferencias efectuadas sobre el supuesto de normalidad podrían ser erróneas. Como puede observarse, las condiciones bajo las cuales los resultados del modelo de regresión li-

neal son válidos no se aplican cuando se tiene una proporción continua.

Transformación de la variable de respuesta

En un capítulo de su libro, Atkinson[8] describe varias transformaciones para porcentajes y proporciones, y aplica sobre ellas el método de OLS. Atkinson recomienda dos transformaciones . La primera es la transformación *logit*, en la que:

$$\ln\left(\frac{y}{1-y}\right)=x'\beta+\varepsilon \qquad (1)$$

donde ln(y/(1-y)) representa la transformación logit de la variable dependiente.

La segunda es una transformación logarítmica para generar una variable que sólo tenga valores no negativos, es decir, que su dominio se halle en el intervalo (0,∞):

$$-\ln(y)= x'\beta+\varepsilon \qquad (2)$$

En ambos casos el análisis implica la utilización de la metodología OLS. Atkinson señala que estas transformaciones generalmente logran *linealizar* la relación entre la variable de respuesta y los predictores; sin embargo, y como lo resalta Aitchison,[9] cuando se usan con proporciones continuas casi nunca se consigue *estabilizar* la varianza y el efecto es el uso inapropiado de este enfoque. Además, existe el inconveniente de que los valores límite, cero y uno, deben modificarse por una pequeña constante para que no se generen valores perdidos o *missings*.[5] Con todo, se preserva el supuesto según el cual el error (ε) se distribuye normalmente y por ello todos los supuestos destacados en la sección anterior deben cumplirse para tener la seguridad de que las inferencias son válidas. Una alternativa para la segunda transformación expresada en (2) consiste en utilizar un modelo de regresión gama, tras asumir que la variable transformada, expresada en el rango (0,∞), sigue una distribución gama. Esta distribución es en particular útil para modelar variables que son estrictamente no negativas, ya que es muy flexible para modelar distintas *formas* de la variable de respuesta por los dos parámetros que indexan la distribución[10] y no impone el supuesto de homoscedasticidad, aunque sí el de un coeficiente de variación constante.[11] Cualquiera de los dos enfoques puede ser de utilidad si se satisfacen los supuestos que cada uno impone y el modelo gama debe preferirse si se tienen sólo valores no negativos y la distribución de la variable de respuesta no es simétrica. El principal inconveniente de ambos enfoques es que los parámetros del modelo no pueden interpretarse con facilidad en términos de la escala original de la variable; la segunda desventaja es que se requieren muestras *grandes* para lograr que la aproximación sea adecuada.

Métodos de quasi-verosimilitud

Los dos enfoques anteriores han asumido, explícita o implícitamente, alguna familia específica de distribuciones para analizar la distribución condicional de la proporción bajo estudio. Cox,[6] Papke y Wooldridge[12] optan por una vía diferente y utilizan un enfoque de quasi-verosimilitud[11] que, de modo sinóptico, se explica a continuación.

En la mayoría de las investigaciones empíricas se sabe que los datos siguen alguna distribución de probabilidad. Si es continua y simétrica, por ejemplo, se puede asumir la distribución normal; si es discreta y dicotómica, la distribución binomial. No obstante, existen ocasiones en las que es menor la certidumbre respecto de la distribución exacta de una variable. Un ejemplo de esta situación es el análisis de los costos de hospitalización, que se sabe son positivos e invariablemente sesgados a la derecha. Quizás, y con un poco más de experiencia en el análisis de esta variable, se podría determinar que la varianza se incrementa con la media e incluso establecer la velocidad de este incremento. Sin embargo, podría resultar difícil, si no imposible, especificar la distribución exacta para esta variable sin hacer uso de algún argumento asintótico. El problema es que, de no hacerlo, no se podría construir la verosimilitud ni emplear técnicas como la máxima verosimilitud o la prueba del cociente de verosimilitud. McCullagh y Nelder[11] evidenciaron que era posible construir un método de cuantificación, la quasi-verosimilitud, sin la necesidad de especificar ningún supuesto distributivo, y que aun así fuese capaz de generar procedimientos de inferencia confiables. La quasi-verosimilitud permite hacer inferencias con cierto grado de *robustez* a partir de dos condiciones importantes.[13] Primero, no se necesita imponer un supuesto distribucional y, segundo, sólo es preciso especificar la relación entre la media y la varianza, que se determina mediante una constante de proporcionalidad que puede calcularse a partir de los datos observados.

Cox[7] ha desarrollado un modelo que analiza la especificación para los primeros dos momentos de la distribución condicional de variables observadas en el rango continuo (0,1). De forma específica, examina el uso de las funciones liga *logit* y *log-log complementaria* junto con dos especificaciones para la función va-

rianza a las que él llama *canónica* y *ortogonal*. Según este investigador, la liga logit y la función varianza ortogonal son las más adecuadas en términos asintóticos, es decir, Cox propone utilizar las siguientes relaciones para la media y la varianza de la distribución de la proporción continua:

$$\mu(y)=1/(1+e^{-\beta}) \quad (3)$$

$$v(\mu)=\mu^2(1-\mu^2) \quad (4)$$

Papke y Wooldridge[12] recurren a un enfoque similar, sólo que en una problemática distinta. Están interesados en la especificación de un modelo de regresión de quasi-verosimilitud para proporciones continuas y usan la siguiente especificación de la log-verosimilitud, la cual, apuntan, está bien definida para $0 <G(\bullet)<1$.

$$\ell_i=y_i \ln[G(x_i)]+(1-y_i)\ln[1-G(x_i)] \quad (5)$$

Aunque discuten diversas especificaciones para $G(\bullet)$, prefieren la función logística en su análisis (equivalente a la liga logit de Cox). De manera adicional, utilizan un enfoque más robusto para la estimación de los errores estándar de los coeficientes, a partir de la idea de que las proporciones estimadas pueden exhibir algún grado de correlación dentro de alguna especificación o combinación de las covariables observadas; por dicha razón este será el enfoque empleado en este artículo.

Distribución beta

La distribución beta puede usarse de manera efectiva en situaciones donde la variable está restringida al intervalo continuo (0,1). Ferrari y Cribari-Nieto[7] han propuesto un modelo de regresión para variables que tienen una distribución beta. Antes de describir su propuesta, se define la distribución beta y sus dos primeros momentos. La función de una variable beta se expresa como:

$$f(y)= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \quad (6)$$

donde p y q son parámetros de forma ($p, q >0$) y $\Gamma(\bullet)$ es la función gama. La media y la varianza de la distribución son:

$$E(y)= \frac{p}{(p+q)} \quad (7)$$

$$var(y)= \frac{pq}{(p+q)^2(p+q+1)} \quad (8)$$

Puesto que el propósito de Ferrari y Cribari-Nieto[7] es especificar un modelo de regresión, utilizan una diferente *parametrización* del modelo, lo cual les permite especificar la media así como un parámetro de dispersión para el modelo. Si se definen la media $\mu= p/(p+q)$ y el parámetro de dispersión $\phi= p+q$, es decir, $p= \mu \phi$ y $q= (1-\mu) \phi$, entonces la media y la varianza de la variable según (7) y (8) son:

$$E(y)= \mu \quad (9)$$

$$var(y)= \frac{V(\mu)}{1+\phi} \quad (10)$$

en donde $V(\mu)= \mu(1-\mu)$, de modo que μ es la media de la variable de respuesta y ϕ puede interpretarse como un parámetro de dispersión, en el sentido de que, para valores fijos de μ, mientras más grande sea el valor ϕ, menor es la varianza de *y*. A partir de esta *parametrización* se puede especificar la función de probabilidad de la variable de respuesta como sigue:

$$f(y,\mu,\phi)= \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad (11)$$

con $0 <y <1$, $0 <\mu <1$ y $\phi>0$. Entonces, el modelo especifica que la media de *y* puede escribirse como:

$$g(\mu)= \sum_{i=1}^{k} x_i\beta_i \quad (12)$$

donde β es un vector de parámetros de regresión desconocidos y las *x* son observaciones para las *k* covariables y se asumen fijas y conocidas. La correspondiente función de log-verosimilitud se expresa como sigue:

$$\ell(\beta,\phi)=\sum \ell_i(\mu_i,\phi) \quad (13)$$

donde:

$$\ell_i(\mu_i,\phi)=\ln\Gamma(\phi)-\ln\Gamma(\mu_i\phi)-\ln\Gamma((1-\mu_i)\phi) +(\mu_i\phi-1)\ln y_i+[((1-\mu_i)\phi-1)]\ln(1-y_i) \quad (14)$$

Ferrari y Cribari-Nieto[7] proponen expresiones para maximizar esta función y encontrar los estimadores de máxima verosimilitud, así como para construir procedimientos de inferencia. Como ellos indican, este enfoque tiene algunas características deseables. Por un lado, no se necesita transformar la variable de

respuesta, lo que supone que la interpretación de los coeficientes de regresión sea directa sobre el valor esperado de la variable, esto es, sobre la media de la proporción. Por otro lado, la varianza de la variable de respuesta es una función de la media, por lo cual no se impone el supuesto de homoscedasticidad; y, por último, los parámetros de la distribución beta permiten modelar diversas formas de la distribución, sobre todo aquellas que muestran asimetría o relaciones no lineales.

Además de estos enfoques se han utilizado algunos otros procedimientos como el modelo *tobit* para muestras censuradas,[14] que no se discute en este estudio en virtud de que en una proporción continua no existe censura: los valores fuera del intervalo (0,1) no están definidos y por consiguiente su uso sería inapropiado. Otro enfoque empleado, aunque desde una perspectiva distinta, es la aplicación de la estadística bayesiana,[15] en la que se hace uso de la distribución beta de manera regular; empero, tampoco se considera debido a la poca difusión que este enfoque ha recibido dentro del área de la salud pública y a que su introducción requiere un mayor espacio para presentar sus fundamentos.

Material y métodos

Con el objetivo de comparar los resultados de las diferentes aproximaciones, se tomó como referencia información sobre la cobertura en el programa de planificación familiar obtenida a partir de la Encuesta Nacional de Salud Reproductiva (ENSAR). Esta encuesta se realizó a principios de 2003 bajo la coordinación del Centro Nacional de Equidad de Género y Salud Reproductiva. El diseño metodológico, trabajo de campo, captura y procesamiento de la información estuvieron a cargo del Centro Regional de Investigaciones Multidisciplinarias de la Universidad Nacional Autónoma de México, con la supervisión y el control de la calidad del levantamiento de la información, captura y procesamiento de los datos por parte del Instituto Nacional de Salud Pública.[16] Los comités de ética de dichas instancias aprobaron su ejecución.

Para el levantamiento de la información se utilizaron tres tipos de cuestionarios: uno de hogar, otro para mujeres de 15 a 49 años y uno más de localidad (aplicado a localidades menores de 2 500 habitantes). Para la selección de la muestra se empleó un muestreo probabilístico, polietápico y estratificado. Se trata de una encuesta con representatividad nacional, para ámbitos rurales y urbanos y para ocho estados del país, que cuenta con un total de 19 498 cuestionarios individuales completos (tasa de no respuesta a nivel individual de 6.6%). El cuestionario individual se aplicó sólo a mujeres en edad fecunda y recaba información detallada sobre ocho áreas de interés: características sociodemográficas, fecundidad y antecedentes gineco-obstétricos, anticoncepción, atención materno-infantil, exposición al riesgo de concebir, infecundidad y menopausia, sexualidad y violencia doméstica, así como infecciones de transmisión sexual.[16]

A partir de la ENSAR, Pérez-Núñez y colaboradores* intentaron identificar las variables sociodemográficas y económicas vinculadas con la cobertura de la planificación familiar. Para ello se construyó un indicador compuesto de cobertura en el nivel individual a partir de 13 indicadores de naturaleza dicotómica (0= no cubierto, 1= cubierto).‡ Sin embargo, y en virtud de que no todos los indicadores se aplican a todas las mujeres, se construyó un indicador que resumiera el nivel de cobertura individual de cada mujer, que debía considerar sólo los indicadores en los que ésta formaba parte del denominador. La expresión siguiente resume la construcción del indicador compuesto:

$$CI = \frac{\sum_{i=1}^{k} x_i}{\sum_{j=1}^{r} x_j} \qquad (15)$$

donde:

CI= Cobertura individual
i= Intervenciones que recibió o en las que está cubierta
j= Intervenciones que necesita o en las que debió estar cubierta

Como se puede advertir, el indicador compuesto está acotado al intervalo continuo (0,1) y entra al modelo como la variable de respuesta. Las variables empleadas para modelar la proporción de cobertura en planificación familiar son edad, religión, índice de riqueza, escolaridad, estatus indígena, estado civil, aseguramiento médico, lugar de residencia antes de los 12 años, disponibilidad de la cartilla de salud de la mujer, edad de inicio de vida sexual activa, edad de

* Pérez-Núñez R, Salinas-Rodríguez A, Avila-Burgos L, Mojarro-Íñiguez MG, Medina-Solís CE, Schiavon R *et al*. Cobertura y financiamiento de la planificación familiar en México: hallazgos de la Encuesta Nacional de Salud Reproductiva 2003. Documento no publicado.

‡ Para mayor detalle sobre los indicadores utilizados para evaluar la cobertura en el nivel individual se sugiere consultar: Pérez-Núñez R, Salinas-Rodríguez A, Avila-Burgos L *et al*. Cobertura y financiamiento de la planificación familiar en México: hallazgos de la Encuesta Nacional de Salud Reproductiva 2003. Artículo enviado a la revista de la Organización Panamericana de la Salud en octubre de 2005.

inicio laboral, número de hijos, paridad satisfecha, riesgo reproductivo, lugar de residencia actual (rural o urbano), edad en el primer embarazo y ocupación. En dicho estudio se utilizó a la totalidad de las mujeres con entrevista completa para integrar un modelo de regresión gama. Sin embargo, en el análisis que se presenta en este artículo se generó una muestra aleatoria simple de 2 000 observaciones para la construcción y comparación de los cuatro modelos ya descritos.

Para evaluar modelos estadísticos, el criterio usado y aceptado de modo más amplio para la selección de modelos es el Criterio de Información de Akaike (AIC).[5,17] No obstante, existen diversas variaciones de este criterio que dependen del modelo de regresión evaluado y el tamaño de muestra. En este trabajo se empleó la variante de McQuarrie y Tsai[18] que toma en cuenta la función de verosimilitud y el tamaño de muestra, así como el número de parámetros en el modelo; esta formulación, según señalan estos especialistas, es útil para incorporar modelos de regresión de quasi-verosimilitud o no normales. El criterio se define de la forma siguiente:

$$AIC=\ln(\hat{\sigma})+\frac{n+k}{n-k-2} \qquad (16)$$

donde $\ln(\hat{\sigma})$ es el logaritmo natural del error cuadrático medio del modelo de regresión, n es el número de observaciones y k el número de parámetros. Al igual que otras formulaciones de esta expresión, el criterio de selección es del estilo "pequeño, mejor", ya que para un menor valor de AIC, un mejor ajuste del modelo, esto es, ante un número fijo de observaciones y parámetros, se elige el modelo con el menor error cuadrático medio.

Por último, es necesario aclarar que se utilizaron como variables independientes o de control, o ambas, aquellas que se habían incorporado anteriormente en el modelo original propuesto (Pérez-Núñez), aunque en este caso varias de ellas no tienen un valor p significativo, con el objetivo de efectuar un análisis comparativo con los resultados que se obtuvieron en dicho trabajo. Además, para otorgarle sentido a la comparación entre los estimadores de los coeficientes para cada uno de los cuatro modelos se decidió reportar el valor de la derivada parcial de la función de la esperanza condicional, en relación con cada una de las covariables, evaluado en su media muestral (Kieschnick y McCullough[5]).

En una segunda etapa de análisis se realizó un ejercicio de simulación para modelar una variable de respuesta con distribución beta a partir de una variable independiente que sigue una distribución normal, con un tamaño de muestra variable entre 100 y 18 000 observaciones. Para simular los datos se utilizó un enfoque de Monte Carlo/Cadenas de Markov,[19] tal y como opera en WINBUGS. Se generó, como variable dependiente, una variable con distribución beta (parámetros: p= 0.5, q= 0.5), y como variable independiente una variable con distribución normal (parámetros: μ= 2, σ^2= 6). Los análisis estadísticos y de simulación se llevaron a cabo con la ayuda de los programas R,* SAS‡ y WINBUGS.§

Resultados

Comparación de modelos

Los resultados para cada uno de los cuatro enfoques descritos con anterioridad se muestran en el cuadro I. Como puede observarse, con excepción de la variable *riesgo reproductivo*, en el cual ninguno de los modelos (gama, quasi-verosimilitud, normal) coincide con el modelo beta, cuyo valor p reportado no encuentra una diferencia significativa, para todas las demás variables al menos coinciden dos de los modelos con los resultados de la regresión beta en cuanto a los valores p reportados. Sin embargo, si se realiza un análisis de los estimadores puntuales, se encuentran marcadas diferencias entre los cuatro modelos. Al tomar como referencia el modelo beta, el modelo normal subestima los valores de los coeficientes para todas las variables, el modelo de quasi-verosimilitud los sobrestima, mientras que el modelo gama tiende a subestimar algunos y sobrestimar a otros.

Si se examina el supuesto de normalidad para los residuos, el modelo normal, único de los modelos que lo impone, tampoco lo cumple (prueba de Anderson-Darling, p <0.005). Asimismo, el cuadro I muestra que los valores predichos por el modelo normal están fuera del rango observado para la variable de respuesta (0,1), que es otro inconveniente de utilizar este modelo para ajustar proporciones, como ya se había adelantado, ya sea que éstas sean continuas o que se generen porque la variable de respuesta es dicotómica.

La figura 1 muestra los gráficos de normalidad para los residuos generados por los cuatro modelos. Como se puede observar, los modelos beta, gama y de quasi-verosimilitud muestran un buen grado de ajuste y, tal y como se mencionó, tienden a ser asimétricos. En cambio, a partir del modelo normal, aunque es muy pareci-

* R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: 2005. Disponible en: http://www.R-project.org.

‡ SAS. STAT Software, Version 9.1.3. Cary, NC: SAS Institute Inc, 2004.

§ Spiegelhalter D, Thomas A, Best N, *et al*. WinBUGS User Manual Version 1.4. Cambridge, UK: MRC Biostatistics Unit, 2003.

Cuadro I
COMPARACIÓN DE MODELOS[≠,‡]

		Modelos			
		BETA	*Gama*	*QV*[#]	*Normal*
Edad (20-35 años[§])	Menores de 20 años	-0.4237	-0.0812	-0.5457	-0.1183
		(0.0000)	(0.0964)	(0.0000)	(0.0000)
	Mayores de 35 años	0.0349	0.0677	0.0449	0.0095
		(0.6427)	(0.2996)	(0.5710)	(0.6190)
Escolaridad (sin instrucción[§])	Primaria incompleta	0.4107	0.1912	0.4865	0.1140
		(0.0012)	(0.0012)	(0.0010)	(0.0000)
	Primaria completa	0.6112	0.3080	0.7152	0.1672
		(0.0000)	(0.0000)	(0.0000)	(0.0000)
	Secundaria o más	0.8330	0.4289	1.0403	0.2352
		(0.0000)	(0.0000)	(0.0000)	(0.0000)
Religión (católica[§])	Sin religión	-0.1006	-0.1074	-0.1408	-0.0299
		(0.3990)	(0.2199)	(0.2880)	(0.3250)
	Otras religiones	-0.1358	-0.0884	-0.1442	-0.0327
		(0.1023)	(0.2311)	(0.1430)	(0.1220)
Edad de inicio laboral (antes de 18 años[§])	No ha trabajado	-0.1599	-0.1239	-0.1948	-0.0455
		(0.0153)	(0.0059)	(0.0110)	(0.0070)
	18 años o más	0.1051	0.1704	0.1377	0.0296
		(0.0949)	(0.0095)	(0.0560)	(0.0640)
Indigenismo	Sí	-0.2032	-0.1169	-0.2592	-0.0582
		(0.0140)	(0.0190)	(0.0080)	(0.0060)
Aseguramiento médico	Sí	-0.1788	-0.1519	-0.2060	-0.0467
		(0.0020)	(0.0100)	(0.0020)	(0.0010)
Índice de riqueza	Índice	0.0803	0.0468	0.1055	0.0233
		(0.0000)	(0.0051)	(0.0000)	(0.0000)
Cartilla de salud de la mujer	Sí	0.2300	0.1619	0.3081	0.0661
		(0.0000)	(0.0012)	(0.0000)	(0.0000)
Edad de inicio de vida sexual activa (no ha iniciado[§])	Antes de 20 años	0.4609	0.2320	0.5703	0.1272
		(0.0003)	(0.0366)	(0.0000)	(0.0000)
	20 años o más	0.4228	0.3263	0.5021	0.1116
		(0.0018)	(0.0161)	(0.0010)	(0.0010)
Número de hijos (sin hijos[§])	Menos de 4	0.2121	0.2795	0.2384	0.0544
		(0.4873)	(0.2646)	(0.4210)	(0.4840)
	4 o más	0.1246	0.0522	0.1360	0.0308
		(0.6939)	(0.8439)	(0.6620)	(0.7020)
Tiene pareja	Sí	0.0933	0.1453	0.1156	0.0260
		(0.3237)	(0.1506)	(0.2670)	(0.2800)
Paridad satisfecha	Sí	0.1285	0.1446	0.1792	0.0401
		(0.0381)	(0.0111)	(0.0080)	(0.0110)
Riesgo reproductivo	Sí	0.1229	0.2087	0.2060	0.0432
		(0.1215)	(0.0256)	(0.0110)	(0.0320)
Residencia antes de los 12 años (en la ciudad[§])	En un rancho	-0.0565	-0.0631	-0.0697	-0.0146
		(0.4511)	(0.3361)	(0.4140)	(0.4440)
	En un pueblo	-0.0478	-0.0234	-0.0520	-0.0112
		(0.4487)	(0.6913)	(0.4830)	(0.4870)
Religión * índice de riqueza (católica[§])	Sin religión	0.0246	0.0335	0.0311	0.0057
		(0.6934)	(0.3914)	(0.6450)	(0.7200)
	Otras religiones	-0.0152	-0.0123	-0.0303	-0.0056
		(0.7152)	(0.7351)	(0.5310)	(0.5970)
Valores ajustados	AIC[&]	-1.5385	-0.4247	0.1482	0.1369
	mínimo	0.1322265	0.000000042	0.0837034	-0.0212431
	máximo	0.8366405	0.657702610	0.8953044	0.9693555

≠ Los coeficientes expresan el efecto marginal; entre paréntesis se registra el valor *p*
‡ Los coeficientes se han ajustado por lugar de residencia actual, edad al primer embarazo y ocupación
§ Grupo de referencia
# Quasi-verosimilitud
& Criterio de información de Akaike

Instituto Nacional de Salud Pública, Cuernavaca, Morelos, México, 2005

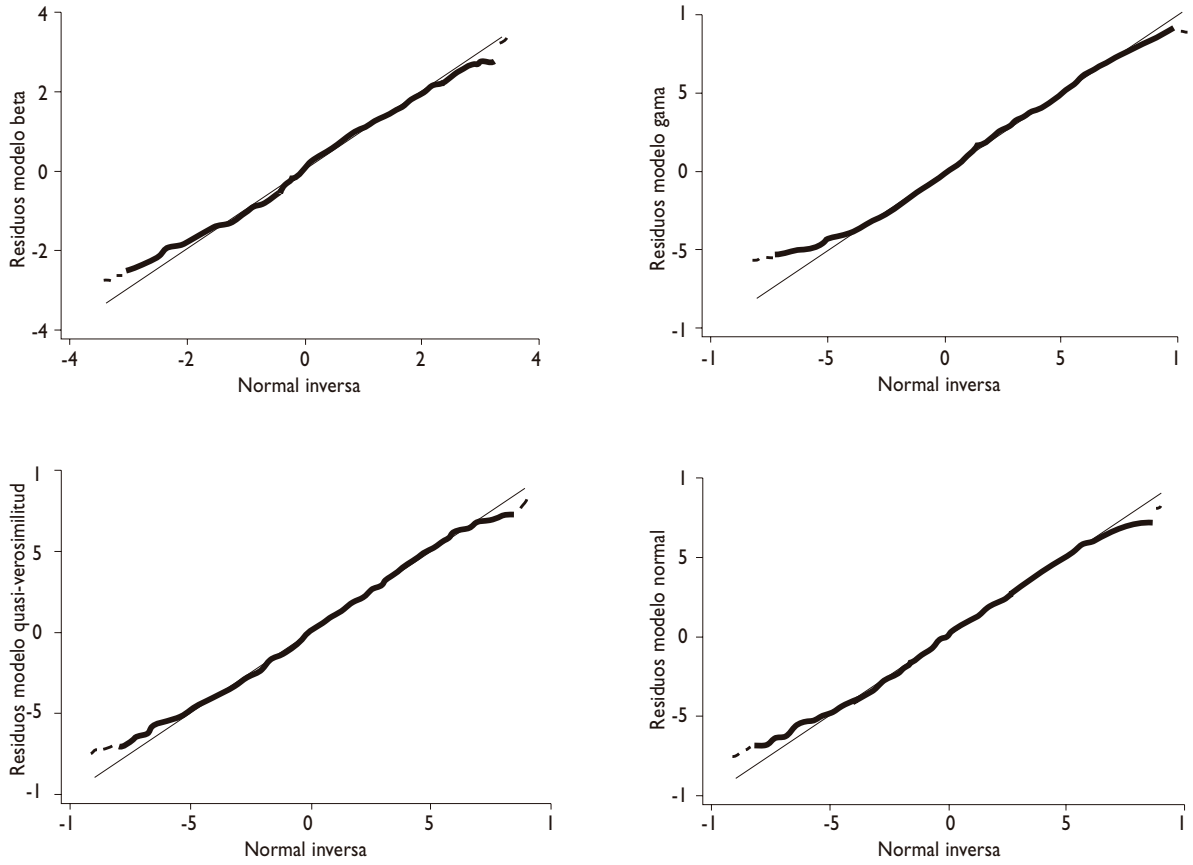

Figura 1. Comparación de modelos: residuos de devianza

do a los otros modelos, no puede concluirse que sus residuos sigan una distribución normal, lo cual confirma lo que había detectado la prueba de Anderson-Darling. También se puede reconocer en el cuadro I que los valores predichos por estos tres modelos (beta, gama y de quasi-verosimilitud) se encuentran dentro del rango observado para la variable de respuesta (0,1).

Por último, a partir del valor del AIC se puede identificar que el modelo beta es el que mejor se ajusta a los datos, seguido por el modelo gama (cuadro I). El modelo normal y el de quasi-verosimilitud muestran el valor de AIC más grande y, en consecuencia, son los que *peor* ajuste poseen; sin embargo, y como se verá a continuación, en cuanto al enfoque de quasi-verosimilitud, cuando se incrementa el tamaño de muestra es el que más se aproxima al modelo beta.

Comparación de modelos según el efecto del tamaño de la muestra

En la figura 2 se muestra el comparativo de los cuatro modelos descritos y su comportamiento al incremen-

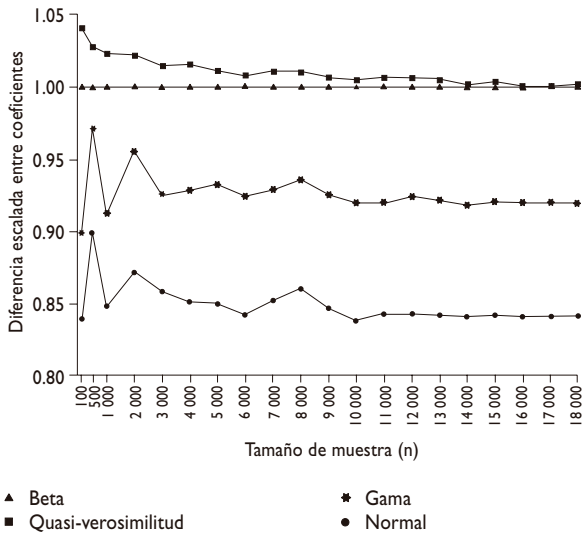

Figura 2. Comparación de modelos de acuerdo con el tamaño de la muestra

tar el tamaño de muestra. Al igual que en la sección anterior, los coeficientes están expresados como efectos marginales. Una vez más, si se toma el modelo beta como referencia (la línea que permanece fija en 1), se puede advertir que los modelos gama y normal subestiman de manera consistente a los coeficientes sin importar si el tamaño de la muestra crece; en cambio, el modelo de quasi-verosimilitud sobrestima a los coeficientes, pero cuando el tamaño de muestra se incrementa ($n \rightarrow \infty$) los coeficientes de este modelo se aproximan de manera notoria a los del modelo beta.

Discusión

La presencia de una variable dependiente que está expresada como una proporción en el intervalo continuo (0,1) es cada vez más común en estudios y aplicaciones de distintas disciplinas, tal y como lo han documentado Johnson y colaboradores.[1] Varios autores han destacado también la necesidad de generar modelos con una estructura de regresión que ayuden a explicar la *variabilidad* de este tipo de variables.[6,7,12] Pese a ello, no existe una práctica aceptada y difundida para esta clase de modelos. Kieschnick y McCullough[5] efectuaron una revisión, no exhaustiva, de las distintas propuestas que se han generado para modelar una proporción continua.

En este artículo se han comparado las cuatro formas más difundidas para construir un modelo de regresión cuando la variable de interés es una proporción continua. Se han tomado en cuenta los siguientes aspectos. Primero, que la variable de interés esté acotada en el intervalo continuo (0,1); segundo, que la función de su esperanza condicional sea no lineal; tercero, que sea de naturaleza heteroscedástica; y cuarto, que la aplicación de cualquier modelo estadístico suponga la existencia de ciertas condiciones (*supuestos*) bajo las cuales los resultados son válidos, entre ellos el supuesto distribucional.

La semejanza entre los resultados obtenidos a partir de los diferentes modelos de regresión bajo estudio, reflejan, como lo ha notado Cox,[6] que los estimadores de la varianza, vinculados con los coeficientes de un modelo de regresión cuando la variable de respuesta es una proporción continua, son *consistentes*, sobre todo si se utiliza el método de estimación de máxima verosimilitud. Con todo, no debe dejar de destacarse que las conclusiones e inferencias dependen del modelo que se elija para el análisis.

En este sentido, las diferencias encontradas en los estimadores puntuales pueden tener su origen en dos causas principales. Primera, el tamaño de la muestra, ya que si se incrementa el número de observaciones los estimadores tienden a ser *consistentes*. Segunda, los supuestos que cada modelo impone. Con referencia al modelo normal, si bien es cierto que los valores *p* entre los modelos beta y normal son congruentes, en general, respecto de un nivel de significancia nominal, también lo es que el modelo normal tiende a mostrar valores más pequeños, es decir, *más significativos*, ya que sus errores estándar son también más pequeños. Esto concuerda con la crítica en cuanto al supuesto de homoscedasticidad del modelo normal, ya que al no cumplir con tal supuesto (prueba de Cook-Weisberg para heteroscedasticidad, $p < 0.001$), el modelo tiende a subestimar los errores estándar.

Además, la falta de cumplimiento del supuesto distribucional (normalidad de los residuos) lleva casi siempre a una subvaloración de los estimadores puntuales.[4] Como ya se comentó, de los cuatro modelos expuestos, el único que impone el supuesto de homoscedasticidad es el modelo normal, en tanto que el supuesto de normalidad para los residuos puede extenderse a los cuatro modelos, si es que se trata de los residuos de devianza, pero sin perder nunca de vista que para modelos que no asumen una distribución normal esto no es más que una aproximación, ya que las más de las veces muestran cierto grado de asimetría.[17]

Los resultados indican que el modelo beta es el que posee un mejor ajuste, de acuerdo con sus supuestos y el valor de AIC, y que de los otros tres modelos, el gama y el de quasi-verosimilitud podrían ser una opción viable, siempre y cuando el tamaño de muestra sea suficientemente *grande*. Del modelo normal los datos muestran que debe evitarse su aplicación y que, en el peor de los escenarios, debe utilizarse sobre una variable transformada, sea logit o logarítmica. Al final, se recomienda utilizar el enfoque paramétrico de la regresión beta de Ferrari y Cribari-Nieto[7] y, si se tiene un tamaño de muestra grande, el de quasi-verosimilitud de Papke y Wooldridge.[12] Los hallazgos encontrados en el enfoque de quasi-verosimilitud concuerdan con los datos de Kieschnick y McCullough[6] y confirman su naturaleza asintótica.

Dos factores deben incorporarse a las conclusiones, ya que no se han descrito en los párrafos anteriores. En primer lugar, es necesario llevar a cabo una comparación más amplia con modelos más *complejos,* por ejemplo el que desarrolló Jorgensen[20] para la distribución *simplex* o el modelo de McDonald y Xu[21] sobre la distribución beta generalizada. Más aún, es preciso comparar estos resultados con aquellos que se generen a partir de un enfoque bayesiano. En segundo lugar, también podría ser útil contrastar los resultados de un modelo de regresión beta para variables expresadas como una tasa y analizadas mediante un enfoque

de regresión *poisson* o un enfoque de regresión binomial (que incluya un término *offset*). Es concebible esperar resultados similares, siempre y cuando se cumplan los supuestos de los modelos *poisson* o binomial.

Por último, debe notarse que estos modelos han tenido poca difusión y aplicación en el ámbito de la salud pública, debido a la falta de una cultura estadística en el área que centra casi toda su atención en los modelos de regresión lineal y al uso limitado del *software* estadístico disponible para ajustar esta clase de modelos. En este sentido, todos los modelos presentados aquí pueden procesarse en cualquier *software* que cuente con un módulo para los modelos lineales generalizados, con excepción del modelo de regresión beta que requiere el uso de los programas R o SAS.

Referencias

1. Johnson NL, Kotz S, Balakrishnan N. Continuous univariate distributions, vol. 2. New York: Wiley, 1995.
2. Hviid M, Villadsen B. Beta distributed market shares in a spatial model with an application to the market for audit services. Review of Industrial Organization 1995;10:737-747.
3. Bury K. Statistical distributions in engineering. New York: Cambridge University Press, 1999.
4. Godfrey L. Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches. New York: Cambridge University Press, 1988.
5. Kieschnick R, McCullough BD. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Stat Model 2003;3(3):193-213.
6. Cox C. Nonlinear quasi-likelihood models: applications to continuous proportions. Comput StatDat Anal 1996;21:449-461.
7. Ferrari S, Cribari-Nieto F. Beta regression for modelling rates and proportions. J Appl Stat 2004;31(7):799-815.
8. Atkinson A. Plots, transformations and regression: an introduction to graphical methods of diagnostic regression analysis. New York: Oxford University Press, 1985.
9. Aitchison J. The statistical analysis of compositional data. New York: Chapman & Hall, 1986.
10. Hardin J, Hilbe J. Generalized linear models and extensions. Texas: Stata Press, 2001
11. McCullagh P, Nelder JA. Generalized linear models. New York: Chapman & Hall, 1989.
12. Papke LE, Wooldridge JM. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. J Appl Econom 1996;11(6):619-632.
13. McCulloch CH, Searle SR. Generalized, linear, and mixed models. New York: Wiley, 2001.
14. Barclay MJ, Smith CW. The determinants of corporate leverage and dividend policies. Journal of Applied Corporate Finance 1995;7:4-19.
15. Congdon P. Bayesian statistical modelling. Chichester, UK: John Wiley & Sons, 2001.
16. Programa Nacional de Población 2001-2006. Informe de ejecución 2003-2004 del Programa Nacional de Población 2001-2006:321-365.
17. Lindsey JK. Applying generalized linear models. New York: Springer-Verlag, 1997.
18. McQuarrie A, Tsai C. Regression and time series model selection. New Jersey: World Scientific Publishing Company, 1998.
19. Gilks WR, Richardson S, Spiegelhalter DJ eds. Markov chain Monte Carlo in practice. London, UK: Chapman and Hall, 1996.
20. Jorgensen B. The theory of dispersion models. New York: Chapman & Hall, 1997.
21. McDonald JB, Xu YJ. A generalization of the beta distribution with applications. J Econom 1995;66:133-52.