Economía Vol. XXXI, N° 62, semestre julio-diciembre 2008, pp. 17-32 / ISSN 0254-4415

Una clasificación de modelos de regresión binaria asimétrica: el uso del BAYES-PUCP en una aplicación sobre la decisión del cultivo ilícito de hoja de coca*

Jorge Luis Bazán**
Óscar Millones**
PUCP

RESUMEN

En modelos econométricos clásicos de regresión binaria tradicionalmente se emplea la regresión logística, que se basa en el enlace simétrico logito. El propósito de este trabajo es presentar modelos de regresión binaria que, más bien, tengan enlaces asimétricos —aún no disponibles en *software* comercial—, cuando esta asimetría es más conveniente al investigador. Además, haciendo uso de un enfoque bayesiano con el programa WinBUGS, se implementa el programa BAYES-PUCP, que facilitará la escritura de la sintaxis necesaria para implementar los modelos revisados. El BAYES-PUCP genera tanto las sintaxis de los modelos revisados así como de la estructura de los datos. El método es ilustrado con el caso de una muestra de agricultores que consideran la decisión de erradicar cultivos ilícitos de hoja de coca y, al mismo tiempo, se exploran factores asociados a esta decisión.

Palabras clave: enlaces asimétricos, regresión binaria, inferencia bayesiana, modelos econométricos de elección discreta, WinBUGS.

ABSTRACT

In classical econometric binary regression models, the logistic regression model has been used associated to the logit symmetric link. The purpose of this paper is to present binary regression models with rather asymmetric links —not yet available as commercial *software*—, when this asymmetry is more appropriate to the researcher. In addition, a utility program called BAYES-PUCP is implemented. BAYES-PUCP uses a bayesian approach with the program WinBUGS and will facilitate the writing of the syntax for the models reviewed. It also generates both the syntax and the structure of the data. The method is illustrated for a sample of farmers who consider the decision to eradicate illegal crops of coca leaf. Factors associated with this decision are also explored.

Keywords: asymmetric links, binary regression, bayesian inference, discrete choice econometric models, WinBUGS.

* Una versión preliminar de este trabajo fue presentada en la *12th Escola de Séries Temporais e Econometría*, Gramado, Porto Alegre, Brasil, entre el 31 de julio y el 3 de agosto de 2007. Este trabajo ha recibido apoyo financiero de la Dirección Académica de Investigación de la Pontificia Universidad Católica del Perú (PUCP), Proyecto DAI 3412. Asimismo ha recibido apoyo de los Departamentos de Ciencias y de Economía de la PUCP.

** jlbazan@pucp.edu.pe y omillones@pucp.edu.pe.

INTRODUCCIÓN

Los modelos de comportamiento de elección discreta aleatoria se emplean en la literatura econométrica (ver, por ejemplo, Greene 2008, Agresti 2002). La regresión binaria se aplica, por ejemplo, para evaluar riesgos de obtener o no una tarjeta de crédito, la posibilidad de otorgar un préstamo, o no hacerlo, comprar o no comprar un bien durable, o elegir o no un servicio de atención de salud formal. Este tipo de situación es común también en otras áreas de aplicación como la biológica, educativa y médica. Entre los modelos de elección discreta aleatoria, la regresión binaria usando regresión logística es el método usado por defecto.

Algunos enlaces asimétricos para modelos de regresión binaria han sido propuestos considerando situaciones donde este enlace es más apropiado que el simétrico, sin embargo, no existe un esquema general que caracterice las formas de construcción de enlaces. En este trabajo damos una caracterización general de los enlaces asimétricos y también presentamos un método general de inferencia bayesiana para la estimación de parámetros en la regresión binaria. Para la implementación de los modelos se desarrolla un programa que contiene los diversos modelos revisados, que pueden ser usados para el desarrollo de diversas aplicaciones econométricas, pero también por diferentes usuarios como médicos, psicólogos, educadores, entre otros. El método será ilustrado considerando información para una muestra de agricultores cocaleros en su decisión de producir cultivo ilícito. Se utilizará las sintaxis del BAYES-PUCP para su implementación usando el programa WinBUGS.

En la sección 1 se describe el caso investigado en este estudio. En la sección 2 se presenta la metodología de este trabajo que comprende: a) la justificación de enlaces asimétricos, b) la clasificación de los enlaces asimétricos, c) la formulación del modelo general y d) la inferencia bayesiana. En la sección 3 se presenta el programa desarrollado para la implementación de diversos modelos y se muestran los resultados obtenidos con los datos del estudio. Finalmente, en la sección 4 se presentan los comentarios finales del estudio.

1. EL CASO DE LA ERRADICACIÓN DEL CULTIVO ILÍCITO DE HOJA DE COCA

Después de una reducción del cultivo de hoja de coca en la década de 1990, la tendencia empezó a revertirse ligeramente en el año 2002 (ONU 2003). Por otro lado, la superficie total cultivada ha aumentado en 30% entre 1999 y 2004, la producción total se ha incrementado en 127% para el mismo periodo, llegando a 109,900 TM en 2004 (Glave y Rosemberg 2005). La erradicación de este cultivo es manual, desde que la erradicación por medios químicos está prohibida (DS 004-2000-AG, artículo 1). Por otro lado, se estima que el 80% del cultivo lo realizan pequeños productores de menos de una hectárea

de extensión (Bedoya 2003, ONU 2003), lo que sugiere una práctica de diseminación en zonas de producción. En los Andes peruanos la producción se concentra en el Alto Huallaga y en el Valle del Río Apurímac y Ene (VRAE). Recientemente, los gobiernos de Perú y Colombia han reaccionado ante crecientes brotes de este cultivo también en la zona limítrofe del valle del río Putumayo y Napo.

La literatura sobre la problemática de la hoja de coca ha girado en torno a estudios sobre producción (oferta), la distribución (comercialización) y el consumo (demanda);[1] por ejemplo, ver Durand (2004), Onudd-Devida (2006) y León y Castro de la Mata (1989). Los estudios detallados desde el lado de la producción han sido relativamente limitados en información, basados en muestras o estudios de casos, desde que el levantamiento de la información del proceso productivo mismo es de extrema dificultad, debido mayormente a factores de seguridad.[2]

Por otro lado, estudios más cualitativos a nivel de microdatos pueden enfocar la racionalidad y conducta del productor. Al respecto se argumenta que entre la multiplicidad de factores demográficos, económicos y ecológicos, su producción es el resultado del aseguramiento del nivel de subsistencia dentro de la parcela familiar (Aramburú 1989). Se destaca también la conducta de diversificación del riesgo, en un esquema de transacción intertemporal —«hipótesis de caja chica»— (Bedoya 2003) y otros temas relacionados al uso tradicional, consumo y de hábitos de salud (Rospigliosi 2005, Durand 2004).

Por el lado de las actitudes y percepciones del productor, se han adaptado recientemente modelos para la evaluación de cambio de comportamiento de diversos actores que incluyen agricultores productores, líderes de opinión y generadores de opinión pública, en el marco de programas de desarrollo alternativo (ver Bardales 2004). En este esquema de comportamiento, por el lado del productor-agricultor, se ha sugerido un proceso en cadena que va desde la percepción de los beneficios de erradicar (y riesgos de no hacerlo) hasta la decisión final de mantener libres sus cultivos ilícitos de hoja. Este estudio incluyó la elaboración de una metodología para estimar un conjunto de indicadores de actitudes y percepciones de los agricultores y familiares para ser aplicados a una muestra para el año 2004. Algunos de los índices propuestos por Bardales (2004) son empleados en este trabajo (ver más adelante). La muestra comprende la aplicación de una encuesta estructurada a 1,947 agricultores productores de hoja de coca en áreas cocaleras donde parcialmente se han implementado programas de desarrollo alternativo.

---

[1] El trabajo de Córdova (2007) es más integral que esta caracterización. En este trabajo se vinculan las fuerzas del mercado de coca, en sus diversas fases desde el productor primario hasta el mercado de productos derivados o finales, donde, por el lado de la oferta se modela el rol del monopolio estatal para controlar la oferta de coca, al mismo tiempo de la acción represiva de erradicación de los cultivos de coca.

[2] Estadísticas parciales recientes incluyen Fonafe (2005), Onudd-Devida (2006), Dirección Nacional Antidrogas - Dirandro (2007), recopilación de fuentes y encuestas esporádicas en zonas específicas por Cedro, Cepes (2004a, 2004b), PDA (2004), entre otros.

En los esquemas de comportamiento propuestos en Bardales (2004), es central la estimación de proporciones a favor o en contra en percepciones y decisiones en cada etapa del proceso potencial de erradicación, así como la exploración de las condiciones que determinan estas proporciones. Siguiendo esta línea, se realizó un estudio exploratorio sobre los factores asociados al comportamiento del productor en su decisión de erradicar; así Millones (2005) estima factores significativos asociados con ese tipo de decisión haciendo uso del modelo logístico. Las regresiones indican que la sensibilidad de este indicador de erradicación se asocia positivamente con participación comunal, confianza en el futuro, percepción de daño a la naturaleza y pobreza, por otro lado, se asocia negativamente con consumo personal de coca, mala percepción de servicios públicos y nivel de educación formal.

En el presente trabajo, en un primer paso, nos restringimos a estudiar si los campesinos desean erradicar o continuar la siembra de hoja de coca (variable *sierr*), dado un conjunto de factores asociados que incluyen sus características demográficas y aspectos de los programas sociales que han sido implementados en recientes años. La proporción observada que desea erradicar el cultivo de hoja de coca es 0.58%, es decir, los campesinos son favorables a erradicar el cultivo de hoja de coca.

Para efectos del presente estudio, consideramos entre las variables explicativas un índice de percepción acerca de que el cultivo de hoja coca produce daño al medio ambiente (variable *permedyc*), un índice de participación comunal (variable *partco*), un índice acerca de su consumo personal de hoja de coca (variable *concoca*) y, finalmente, los niveles de pobreza (variable *pobrez*). El modelo considerado tradicional es una regresión binaria logística dada por:

$$p_i = P(sierr = 1) = F[\eta_i], \tag{1}$$

$$\eta_i = \beta_1 + \beta_2 permedyc_i + \beta_3 partco_i + \beta_4 concoca_i + \beta_5 pobrez_i, i = 1,...,1947, \tag{2}$$

donde es la función de distribución acumulada de la distribución logística dada por:

$$F(t) = \frac{\exp(t)}{1+\exp(t)}$$

2. METODOLOGÍA

2.1. Justificación en enlaces asimétricos en la regresión binaria

Con la regresión binaria, se modela las probabilidades de una variable de respuesta que toma dos valores en función de otras variables explicativas considerando una función de enlace o *link*. Este enlace trata los datos dicotómicos como una variable de respuesta y explora su relación con otras variables explicativas, que son combinadas como un

predictor lineal. Es decir, los modelos de regresión binaria estiman la probabilidad de éxito de uno de los valores de la variable respuesta como función de un conjunto de predictores o regresores considerando un enlace entre estas variables. En la regresión binaria, los enlaces usados comúnmente son los enlaces probit y logit, que originan la regresión probit y la regresión logística, respectivamente. En ambos modelos (probit y logit), esta probabilidad tiene una forma simétrica de alrededor de 0.5.

Sin embargo, cuando hay probabilidades extremas, es decir, cuando hay presencia predominante de uno de los valores de la variable respuesta, los enlaces simétricos son inadecuados. Esto ocurre, por ejemplo, cuando se quiere establecer un modelo de atención de salud en una población considerando una serie de características individuales como edad, ingresos, niveles de pobreza, etcétera; donde, en general, la proporción de personas con problemas de salud es muy baja. Por ello, diversos enlaces asimétricos han sido propuestos en la literatura en los últimos treinta años.

Entre los argumentos a favor de que los enlaces logit y probit no son adecuados en situaciones donde hay probabilidades extremas están los de Nagler (1994) y Chen *et al.* (1999). Nagler (1994) indica que cuando se usa logit y probit se está asumiendo que un individuo con una probabilidad del 0.5 de éxito es más sensible a cambios en las variables independientes, por ejemplo, cambios en una unidad en el regresor, que alguien con un 0.3 o 0.7 de probabilidad de éxito. Sin embargo, Nagler sostiene que este no es necesariamente el caso en otras situaciones, donde un individuo con una probabilidad del 0.4 de éxito puede ser más sensible ante un cambio de una unidad en un regresor que otro que tiene una probabilidad del 0.5 de éxito. Si es así, la distribución es sesgada. Las respuestas no son simétricas alrededor de 0.5, por esta razón el empleo de los enlaces asimétricos es justificado en estos casos.

Chen *et al.* (1999) también sostienen que cuando la probabilidad de una respuesta binaria se aproxima a 0 en una tasa diferente que cuando se aproxima a 1, los enlaces simétricos para el ajuste de datos pueden ser inadecuados. En este caso, hay que considerar enlaces asimétricos.

Desde los trabajos de Prentice (1976) varios enlaces asimétricos han sido considerados en la literatura. Al mismo tiempo, varios textos reportan situaciones en donde un enlace asimétrico puede ser más apropiado que el simétrico. Sin embargo, estos enlaces asimétricos han estado construidos de diversos modos, pero sin un tratamiento unificado que pueda caracterizarlos en esta diversidad de construcciones.

Aun cuando la regresión binaria se ha discutido en la literatura en los últimos cincuenta años bajo una visión frecuentista clásica, el enfoque bayesiano ha sido tratado solo recientemente.[3]

Se ha demostrado que la metodología bayesiana es útil, especialmente bajo los métodos *Markov Chain Monte Carlo Methods* (MCMC). La aproximación bayesiana nos sirve

[3] Para el enfoque de econometría bayesiana, ver Lee (2007), Lancaster (2004) y Koop *et al.* (2007).

para conocer la distribución a posteriori de los parámetros de los modelos y también para observar medidas alternativas de bondad de ajuste. Esto es útil para comparaciones entre modelos alternativos.

2.2. Clasificación de los enlaces asimétricos para regresión binaria

Consideremos un vector $y=(y_1,y_2,...,y_n)'$ $n\times1$ de n variables aleatorias independientes binarias correspondientes a n observaciones de la variable respuesta, $x_i=(x_{i1},...,x_{ik})'$ un vector de dimensión $k\times1$ correspondiente a k variables explicativas para la observación i. Sea X la matriz diseño $n\times k$ con filas $x'_i, \beta=(\beta_1,...,\beta_k)'$ un vector $k\times1$ de coeficientes de regresión asociados a las variables explicativas. Considere que $y_i=1$ con probabilidad p_i y $y_i=0$ con probabilidad de $(1-p_i)$.

En los modelos de datos binarios es común asumir que:

$$p_i=F(\eta_i)=F(x_i'\beta), i=1,...,n, \tag{3}$$

en el cual $F(.)$ denota la función de distribución acumulada (FDA). La función inversa F^{-1} se denomina comúnmente función de enlace y $\eta_i=x_i'\beta$, es el i-*ésimo* predictor lineal. En el caso que F sea una FDA de una distribución simétrica, entonces el enlace resultante es simétrico y p_i tiene una forma simétrica alrededor de $p_i=0.5$. En el caso que $F(.)$ sea una FDA de una distribución normal estándar, nosotros tenemos el enlace probit, y en el caso que $F(.)$ sea la FDA de una distribución logística nosotros tenemos el enlace logit.

Nosotros proponemos que los enlaces asimétricos pueden obtenerse considerando las siguientes modificaciones del modelo definido en (3):

(a) Tomando F como la FDA de una distribución asimétrica
(b) Considerando una modificación del predictor lineal
(c) Considerando F en una clase general de distribuciones de probabilidad, por ejemplo, en una clase de mezclas de distribuciones simétricas con asimétricas

Un ejemplo muy popular del caso (a) es el enlace log-log complementario o cloglog, donde la FDA usada en el enlace corresponde a la distribución de Gumbel. En este caso, la FDA está completamente especificada, no depende de ningún parámetro adicional desconocido y no presenta, como caso particular, un enlace simétrico. Otros ejemplos del caso (a) se obtienen, por ejemplo considerando las siguientes FDA:

$$F(\eta_i)=1-(1+e^{\eta_{ii}})^{-\lambda} \text{ y } F(\eta_i)=(1+e^{-\eta_{ii}})^{-\lambda}, \lambda>0.$$

Los enlaces que se obtienen de estas FDA fueron propuestos por Prentice (1976) y fueron popularizados en la literatura económica por Nagler (1994). Esos enlaces son logit asimetrizados y son llamados aquí como *scobit* y *power logit*, respectivamente, e incluyen al enlace logit como caso especial cuando el parámetro incluido es $\lambda=1$.

También cae en este caso el enlace probito asimetrizado (*skew probit*) propuesto por Bazán, Branco y Bolfarine (2006), denominado aquí BBB probito asimetrizado y otro denominado estándar probito asimetrizado (Bazán, Bolfarine y Branco 2006).

El caso (b) se obtiene cuando se mantiene a F como una distribución simétrica pero se considera una modificación del predictor lineal η_i por $m(\eta_i,l)$, donde $m(.)$ es una función no lineal y continua, y λ es el parámetro que controla la asimetría. Por ejemplo, en el modelo de Stukel (Stukel 1988), F corresponde a la distribución logística y la especificación de $m(\eta_i,l)$ es diferente para $\eta_i \geq 0$ y para $\eta_i < 0$. Es decir, $m(\eta_i,l)$ es una función partida para diferentes rangos de η_i y λ.

Otro caso más simple en el caso (b) se obtiene cuando el predictor lineal se reemplaza por una expresión polinomial, generalmente cuadrática o cúbica. Un caso conocido, denominado logit cuadrático, se obtiene cuando F corresponde a la distribución logística y el predictor lineal simple $\beta_1 + \beta_2 x$ es reemplazado por $\beta_1 + \beta_2 x + \beta_3 x^2$.

Un ejemplo del caso (c) ocurre cuando F está en la clase de mezcla de escala de distribuciones elípticas dada por:

$$F = \{F(.) = \int_{[0,\infty]} H(.\,|\,v) dG(v)\},$$

donde G es la FDA en $[0,\infty>$ y H es una distribución elíptica propuesta por Basu y Mukhopadhyay (2000). También cae en este caso el enlace probit asimetrizado propuesto en Chen *et al.* (1999), denominado aquí CDS probito asimetrizado, el cual considera una clase de mezcla de distribuciones normales, donde la medida de mezcla es la distribución normal positiva que tiene función de densidad de probabilidad (FDP) dada por $g(x) = 2\phi(x)$, $x > 0$, donde $\phi(.)$ es la FDP de la distribución normal estándar. Otro caso similar ocurre cuando se mezcla la normal positiva con la distribución logística que se conoce como logística asimetrizada o *skew logit*.

2.3. Formulación del modelo general de regresión binaria

Una manera de escribir de manera general el modelo de regresión binaria con enlaces asimétricos se obtiene considerando:

$$p_i = F_\theta[m(x_i\,'\beta, \lambda)], i = 1, ..., n, \tag{4}$$

donde $F_\theta(.)$ es la FDP de una distribución asimétrica indexada o una distribución simétrica asimetrizada por el parámetro θ, la cual no necesariamente es unidimensional y $m(.)$ es una función continua del predictor lineal que incluye también la función de identidad con λ como un parámetro de forma o asimetría que, tampoco, es necesariamente unidimensional.

Hay que notar que como casos particulares de esta formulación general se obtienen los enlaces revisados en la sección 2.2.

El conjunto de parámetros en esta formulación es $\beta, \theta, \lambda,$ donde β representa a los coeficientes de regresión asociados a las variables explicativas, θ es un parámetro de forma de la distribución de probabilidad asociada al enlace y λ es un parámetro de modificación no lineal del predictor. De acuerdo a esta formulación, el modelo puede estar mal especificado si los parámetros no pueden ser estimados a partir de los datos o si estos no son establecidos previamente.

2.4. Inferencia bayesiana en regresión binaria

Considerando una distribución Bernoulli para la variable respuesta y_i, la función de verosimilitud para el modelo formulado es dada por:

$$L\left(\beta,\theta,\lambda \middle| y, X\right) = \prod_{i=1}^{n}\left[F_\theta\left[m(x_i'\beta,\lambda\right]\right]^{y_i}\left[1-F_\theta\left[m(x_i'\beta,\lambda\right]\right]^{1-y_i} \tag{5}$$

En la inferencia bayesiana, a diferencia de la inferencia clásica, los parámetros de interés, β, θ y λ, son asumidos como variables aleatorias y para ello son consideradas distribuciones de probabilidad a priori, que reflejan nuestro conocimiento previo acerca de su comportamiento aleatorio, esto es, $\pi(\beta, \theta, \lambda)$.

El propósito de la inferencia bayesiana es entonces, considerando la verosimilitud y la distribución a priori, obtener la distribución a posteriori de lo parámetros de interés usando el teorema de Bayes; esto es, se desea obtener:

$$p(\beta,\theta,\lambda \middle| y, X) = \frac{L\left(\beta,\theta,\lambda \middle| y, X\right)\pi(\beta,\theta,\lambda)}{p(y)},$$

donde $p(y)$ es la distribución marginal, no condicional, de la variable respuesta que no depende de los parámetros de interés.

Como ha sido expresado, los parámetros β, θ y λ tienen diferentes significados. Mientras θ y λ son parámetros estructurales asociados con el enlace, el parámetro β es un vector de parámetros inherente a los datos observados y no dependiente de la elección del modelo. Por tal razón, se puede considerar dos escenarios para la inferencia. Un primer escenario ocurre cuando todos los parámetros son estimados simultáneamente; en el segundo escenario, únicamente β varía y θ y λ son fijos o conocidos. Como indican Taylor *et al.* (1996), podemos referirnos a esos dos escenarios como no condicional y condicional, respectivamente.

La inferencia en el escenario condicional es fácil de implementar tanto considerando máxima verosimilitud como aproximación bayesiana, pues se está en el caso conocido de una regresión binaria común. En la aproximación no condicional, calcular estimadores de máxima verosimilitud usando la verosimilitud dada y distribuciones posteriores de β, θ y λ cuando se toman prioris uniformes impropias, no es simple y algunas condiciones

de existencia de estos estimadores deben considerarse. A los interesados en estas condiciones recomendamos la lectura del trabajo de Chen *et al.* (2000).

En este trabajo, nosotros consideramos prioris denominadas «vagas», prioris propias con distribución conocida pero con alta varianza. También asumimos independencia entre prioris, es decir:

$$\pi(\beta,\theta,\lambda)=\pi(\beta)\pi(\theta)\pi(\lambda) \tag{6}$$

Nosotros podemos usar para β, las prioris comúnmente consideradas en la literatura (ver, por ejemplo, Zellner y Rossi 1984), incluyendo prioris normales del tipo ($\beta_j \sim N(\mu_{\beta_j}, s^2_{\beta_j})$) o la priori uniforme ($\pi(\beta)$=1). Prioris de Jeffreys, como las que proponen Ibrahim y Laud (1991), también pueden ser consideradas.

Especificaciones para $\pi(\theta)$ y $\pi(\lambda)$ dependen del modelo particular elegido considerando su rango de variación. En muchas situaciones prioris para θ y λ son obtenidas de la literatura.

Siguiendo a Albert y Chib (1993), es posible obtener versiones de verosimilitud aumentada introduciendo variables auxiliares como las que ya han sido propuestas en la literatura para muchos de los modelos de regresión binaria con enlaces asimétricos citados en este trabajo. Nosotros recomendamos la lectura de los artículos donde se presentan estas formulaciones, especialmente Chen *et al.* (1999) y Bazán, Bolfarine y Branco (2006) para el caso de los modelos probito asimetrizados.

La inferencia bayesiana para modelos de regresión binaria y especialmente para los modelos mencionados se facilita al usar la simulación MCMC, implementada en el programa WinBUGS (ver Spiegelhalter *et al.* 1996). Este programa permite, usando una programación mínima, la implementación de diversos modelos estadísticos. Para una revisión se recomienda el libro de Congdon (2005).

3. APLICACIÓN

Para implementar la metodología propuesta en este estudio para los datos descritos en la primera sección proponemos el modelo de regresión binaria general dado por:

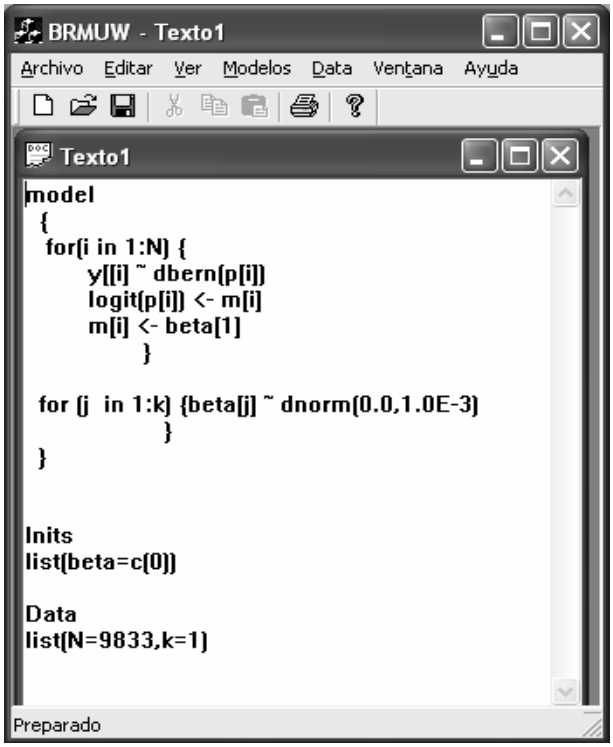
$$p_i = P(sierr=1) = F_\theta[m(\eta_i,\lambda)], \tag{7}$$

$$\eta_i = \beta_1 + \beta_2 permedyc_i + \beta_3 partco_i + \beta_4 concoca_i + \beta_5 pobrez_i, i = 1,...,1947, \tag{8}$$

Como se ha mencionado, el modelo comúnmente usado para este tipo de estudio es el logístico. La sintaxis de este modelo bajo el enfoque bayesiano puede ser obtenida en BAYES-PUCP e implementada en el programa WinBUGS.[4]

[4] Véase <http://videos.pucp.edu.pe/videos/ver/db8373ad4703990c51fd196ef2500c9f>.

Gráfico 1
Sintaxis para el modelo de regresión binaria usando enlace logito obtenido en BAYES-PUCP



Fuente: BAYES-PUCP.
Elaboración: propia.

BAYES-PUCP (ver Spiegelhalter *et al.* 1996) es un programa desarrollado por nosotros, que contiene las diversas sintaxis en código del programa WinBUGS de los modelos de regresión binaria simétricos y asimétricos citados en este trabajo. El programa puede ser obtenido enviando un correo electrónico al primer autor. Es de uso libre desde que se cite la fuente utilizada. La versión preliminar del programa lleva el nombre de BRMUW (Bayesian Regression Model using WinBUGS).

Los modelos implementados en BAYES-PUCP son:

- Simétricos: probit, logit.
- Asimétricos: cloglog, scobit, power logit, skew logit, skew probit (CDS, BBB y standard).

Este programa implementa modelos de regresión binaria que no se encuentran en otros programas comerciales.

Una ilustración de cómo BAYES-PUCP puede ser utilizado para generar sintaxis de los modelos indicados, así como generar la sintaxis para la lectura de los datos, puede ser vista en <http://videos.pucp.edu.pe/videos/ver/b55ab3b7633c6dab0cad8eec47066e40>.

Para poder comparar diferentes modelos, en la inferencia bayesiana se emplea el criterio Deviance Infomation Criteria (DIC) y la media de los desvíos a posteriori (Dbar) propuestos por Spiegelhalter *et al.* (2002), los cuales indican que el mejor modelo es aquel que presenta el menor DIC o Dbar.

Cuadro 1
Comparación de modelos para el modelo explicativo de los campesinos que son favorables a erradicar el cultivo de hoja de coca

Enlaces	Modelos	Bur in	Thin	Dbar	DIC
Simétricos	Probito	4000	5	2451.5	2456.8
	Logito	4000	5	2450.9	2455.8
Asimétricos	Cloglog	4000	5	2451.6	2457.0
	Scobit	4000	25	2462.1	2441.2
	Power Logit	54000	100	2458.5	1794.1
	Logito asimetrizado	4000	25	2458.1	1708.4
	BBB probito asimetrizado	4000	35	2345.2	2252.5
	Estándar probito asimetrizado	4000	15	1538.1	1751.7

Fuente: cálculos con datos de PDA (2004) citados por Bardales (2004).
Elaboración: propia.
Nota: Basada en una cadena de tamaño 2000, obtenida después de retirar valores iniciales generados (b*ur in*) y luego de emplear un muestreo sistemático con saltos (*thin*). El modelo CDS probito asimetrizado presentó problemas de convergencia, por lo que no fue considerado.

Los resultados en el cuadro 1 indican que todos los enlaces asimétricos implementados, con excepción del cloglog, presentan mejor desempeño en los criterios de comparación de modelos a diferencia de los modelos simétricos tradicionales. El modelo con mejor desempeño es el logito asimetrizado.

En los cuadros 2 y 3 mostramos las estimaciones obtenidas para los parámetros del modelo logístico y logístico asimetrizado. Aunque los resultados de los coeficientes de regresión son similares en ambos modelos, la presencia del parámetro de forma $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, que cae en el intervalo *[-1,1],* indica que una adecuada interpretación corresponde al modelo logístico asimetrizado.

Según los resultados del cuadro 2, podemos comentar que la percepción de que el cultivo ilícito de hoja de coca causa daños al medio ambiente (*permedyc*) incide positivamente en aumentar la posibilidad de erradicar dicho cultivo. De igual forma, la participación que tienen los agricultores en las actividades de la comunidad (*partco*), como la ejecución de obras de la comunidad, participación en reuniones comunales

y otros cargos en grupos comunitarios, influyen favorablemente en la erradicación. Dos resultados interesantes que se han replicado en este análisis son que la mayor pobreza (*pobrez*) baja la probabilidad de erradicar el cultivo ilícito, así como el mayor consumo de hoja de coca (*concoca*) se asocia con reducir la posibilidad de tal erradicación.

Cuadro 2
Inferencia para los parámetros para el modelo explicativo de los campesinos que son favorables a erradicar el cultivo de la hoja de coca usando el enlace logito asimetrizado

	Media	Desviación estándar	Percentil 2.5	Mediana	Percentil 97.5
$\beta 1$	-2.84	1.08	-5.10	-2.70	-0.98
$\beta 2$	0.65	0.09	0.52	0.64	0.86
$\beta 3$	0.08	0.02	0.05	0.08	0.11
$\beta 4$	-0.21	0.05	-0.32	-0.21	-0.11
$\beta 5$	0.73	0.20	0.33	0.73	1.15
δ	0.14	0.59	-0.90	0.21	0.91

Fuente: cálculos con datos de PDA (2004) citados por Bardales (2004).
Elaboración: propia.

Cuadro 3
Inferencia para los parámetros del modelo explicativo de los campesinos que son favorables a erradicar el cultivo de la hoja de coca usando el enlace logito

	Media	Desviación estándar	Percentil 2.5	Mediana	Percentil 97.5
β_1	-2.80	0.55	-3.77	-2.80	-1.70
β_2	0.61	0.06	0.50	0.61	0.72
β_3	0.07	0.01	0.04	0.07	0.10
β_4	-0.20	0.05	-0.29	-0.20	-0.11
β_5	0.79	0.18	0.43	0.79	1.12

Fuente: cálculos con datos de PDA (2004) citados por Bardales (2004).
Elaboración: propia.

Adicionalmente, se realizó un análisis de la capacidad predictiva de ambos modelos encontrándose que el modelo logístico (usando el enlace logito) presenta un 64% de buena clasificación (no casos son pronosticados por el modelo como no casos y casos son pronosticados por el modelo como casos) frente a un 95% de buena clasificación usando el modelo skew-logístico (usando el enlace logito asimetrizado).

4. COMENTARIOS FINALES

El propósito de esta investigación fue presentar una caracterización general de los diversos enlaces asimétricos para la regresión binaria que se encuentran en la literatura, mas no así en los programas comerciales. Para ello se usó el enfoque bayesiano y especialmente el programa WinBUGS, los cuales vienen despertando interés en los trabajos de las publicaciones econométricas recientes, dado que es relativamente cómodo desarrollar nuevos modelos basados en sintaxis y usar la simulación MCMC. Con el propósito de facilitar la escritura de la sintaxis que se necesita para implementar los modelos revisados, se desarrolló un programa BAYES-PUCP, que genera tanto las sintaxis de los modelos revisados así como de la estructura de los datos.

Los valores de los parámetros del cuadro 2 son similares a los reportados en Millones (2005) en cuanto a su signo y significancia. Sin embargo, las diferencias metodológicas refieren que en dicho trabajo se usó solo el modelo logit y que se incluyeron más variables independientes asociadas al conocimiento y administración del Programa de Desarrollo Alternativo (PDA). En este trabajo, por el tamaño, se dispone de un modelo que ajusta convenientemente los datos, que además presenta mejor ajuste que los modelos simétricos tradicionales y tiene un óptimo desempeño en la predicción de los casos y no casos de erradicación.

REFERENCIAS

Agresti, Alan
2002 *Categorical Data Analysis*. New York: Wiley-Interscience.

Albert, James y Siddartha Chib
1993 «Bayesian Analysis of Binary and Polytomous Response Data». *Journal of the American Statistical Association*, Vol. 88, N° 2, pp. 669-679, New York.

Aramburú, Carlos
1989 «La economía parcelaria y el cultivo de la coca: el caso del Alto Huallaga». En Federico León y Ramiro Castro (editores). *Pasta básica de cocaína: un estudio multidisciplinario*. Lima: CEDRO, pp. 231-259.

Bardales, Alejandro
2004 *Esquemas de comportamiento para el análisis de las percepciones sobre el PDA y el cultivo ilícito de coca*. Informe presentado al Programa de Desarrollo Alternativo (PDA), Gerencia de Comunicaciones, Chemonics International INC.

Basu, Sanjib y Saurabh Mukhopadhyay
2000 «Binary Response Regression with Normal Scale Mixtures Links». En Dipak Dey, Sujit Ghosh y Bani Mallick (editores). *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker, pp. 231-242.

Bazán, Jorge, Marcia Branco y Heleno Bolfarine
2006 «A Skew Item Response Model». *Bayesian Analysis*, Vol. 1, N° 4, pp. 861-892.

Bazán, Jorge, Heleno Bolfarine y Marcia Branco
2006 *A Generalized Skew Probit Class Link for Binary Regression*. Technical report (RT-MAE-2006-05). São Paulo: Departamento de Estadística de la Universidad de São Paulo.

Bedoya, Eduardo
2003 «Las estrategias productivas y el riesgo entre los cocaleros del valle de los ríos Apurímac y Ene». En Carlos Aramburú y Eduardo Bedoya (editores). *Amazonía: procesos demográficos y ambientales*. Lima: Consorcio de Investigación Económica y Social, pp. 119-154.

Cepes – Centro Peruano de Estudios Sociales
2004a *Agrodata: rendimiento de hoja de coca seca al sol por cuenca y trimestre (2004 kg/ha).* Disponible en <http://www.cepes.org.pe>
2004b *Agrodata: precio mensual de la hoja de coca y la pasta básica en el Perú (1999-2004).* Disponible en <http://www.cepes.org.pe>

Córdova, Boris
2007 «El mercado ilegal de coca en el Perú». Tesis para optar el título de Licenciado en Economía. Lima: Pontificia Universidad Católica del Perú.

Congdon, Peter
2005 *Bayesian Models for Categorical Data.* New York: Wiley.

Chen, Ming Hui, Dipak Dey y Qi-Man Shao
1999 «A New Skewed Link Model for Dichotomous Quantal Response Data». *Journal of the American Statistical Association*, Vol. 94, N° 448, pp. 1172-1186, New York.

Chen, Ming Hui y Qi-Man SHAO
2000 «Property of Posterior Distribution for Dichotomous Quantal Response Model». *Proceedings of the American Mathematical Society*, Vol. 129, N° 1, pp. 293-302.

Dirandro – Dirección Nacional Antidrogas
2007 Disponible en <http://www.drogasglobal.org.pe/estadisticas.php>

Durand, Francisco
2004 «El problema cocalero y el comercio informal para usos tradicional». *Debate Agrario*, N° 39, pp. 109-125, Lima.

Fonafe – Fondo Nacional de Financiamiento de la Actividad Empresarial del Estado
2005 *Perú: Oferta de hoja de coca. Estadística básica (2001-2004)*. Lima: FONAFE.

Glave, Manuel y Cristina Rosemberg
2005 *La comercialización de hoja de coca en el Perú: análisis del comercio formal.* Informe final de consultoría. Lima: GRADE.

Guerrero, Víctor y Richard Johnson
1982 «Use of the Box-Cox Transformation with Binary Response Models». *Biometrika*, Vol. 69, N° 2, pp. 309-314, London.

Greene, William
2008 *Econometric Analysis.* 6th edition. Upper Saddle River, NJ: Pearson

Ibrahim, Joseph y Purushottam Laud
1991 «On Bayesian Analysis of Generalized Linear Models Using Jeffrey's Prior». *Journal of the American Statistical Association*,Vol. 86, N° 4, pp. 981-986, New York.

INEI – Instituto Nacional de Estadística e Informática
2004 *Demanda anual de hoja de coca*. Disponible en <http://www.inei.gob.pe>

Koop Gary, Dale Poirier y Justin Tobias
2007 *Bayesian Econometric Methods (Econometric Exercises)*. Cambridge: Cambridge University Press.

Lancaster, Tony
2004 *An Introduction to Modern Bayesian Econometrics*. Malden, MA: Blackwell Publishing.

Lee, Sik-Yum
2007 *Structural Equation Modelling: A Bayesian Approach*. New York: Wiley.

León, Federico y Ramiro Castro de la Mata (editores)
1989 *Pasta básica de cocaína: Un estudio multidisciplinario*. Lima: Cedro.

Nagler, Jonathan
1994 «Scobit: An Alternative Estimator to Logit and Probit». *American Journal of Political Science*, Vol. 38, N° 1, pp. 230-255, Austin.

Millones, Óscar
2005 *La decisión de erradicar el cultivo ilegal de hoja de coca: explorando asociaciones con el modelo logístico*. Informe presentado al Programa de Desarrollo Alternativo (PDA), Gerencia de Comunicaciones, Chemonics International INC.

ONU – Naciones Unidas
2003 *Peru Coca Survey for 2002. ONU Office of Drugs and Crime*. Disponible en <http://www.unodc.org/pdf/publications/peru_coca-survey_2002.pdf>.

Onudd (Oficina contra la Droga y el Delito de las Naciones Unidas) y Devida (Comisión Nacional para el Desarrollo y Vida sin Drogas)
2006 *Monitoreo de cultivos de coca en el Perú 2005*. Lima: Onudd – Devida.

PDA – Programa de Desarrollo Alternativo
2004 *Encuesta a productores beneficiarios del PDA*.

Prentice, Ross
1976 «A Generalization of the Probit and Logit Methods for Dose-Response Curves». *Biometrics*, Vol. 32, N° 4, pp. 761-768, Washington.

Rospigliosi, Fernando
2005 «Coca legal e ilegal en el Perú». *Debate Agrario*, N° 39, pp. 81-107, Lima.

Spiegelhalter, David, Andrew Thomas, Nicola Best y Wally Gilks
1996 *BUGS 0.5 Examples.Vol. 1, version i.* Cambridge, UK: University of Cambridge.

Spiegelhalter, David, Nicola Best, Bradley Carlin y Angelika van der Linde
2002 «Bayesian Measures of Model Complexity and Fit». *Journal of the Royal Statistical Society. Series B (Statistical methodology)*, Vol. 64, N° 4, pp. 583-639, London.

Stukel, Therese
1988 «Generalized Logistic Models». *Journal of the American Statistical Association*, Vol. 83, N° 402, pp. 426-431, New York.

Taylor, Jeremy, Arminda Siqueira y Robert Weiss
1996 «The Cost of Adding Parameters to a Model». *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, N° 3, pp. 593-607, London.

Zellner Arnold y Peter Rossi
1984 «Bayesian Analysis of Dichotomous Quantal Response Models». *Journal of Econometrics*, Vol. 25, N° 3, pp. 365-393, Amsterdam.