

Modelos de Regresión Binaria Bayesiana Power y Reciprocal Power

Aplicación: La Calidad del Servicio de Salud Pública en la Ciudad de La Paz

Lic. Omar Chocotea Poca
 ✉ omarchp@outlook.com

Resumen. En este artículo, se presenta a dos nuevos modelos de regresión binaria generales con perspectiva bayesiana, al power y al reciprocal power, e incluyen al power probit y al power logit, y al reciprocal power probit y al reciprocal power logit, respectivamente —aún no disponibles en *software* comercial—, y tienen como casos especiales al probit y al logit. El resultado de la aplicación en la data de la Encuesta Calidad de Vida en la Ciudad de La Paz en su tópico salud indica que los cuatro modelos tienen un buen performance.

Palabras clave: Modelos de regresión binaria generales; perspectiva bayesiana.

1. Introducción

Los modelos de regresión binaria son utilizados para predecir la probabilidad de una respuesta binaria (0/1) en función de diversas variables explicativas. La aplicación se encuentra en un rango más amplio de escenarios de investigación que el análisis discriminante.

Bliss (1935) introduce el primer modelo de regresión binaria, el probit, aún es utilizado, sin olvidarnos de sus contendientes históricos, el logit (Berkson, 1944) y el cloglog (Gumbel, 1958), pero el investigador no debe ser estricto con su afinidad o recomendación, más bien debe ubicar alternativas para comparar y seleccionar al mejor (Chocotea, 2014).

2. Links simétricos y asimétricos

Definición 1. Sea $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ un vector de n variables aleatorias independientes binarias (0/1), $\mathbf{x}_i = (1, x_{i2}, \dots, x_{ik})^\top$ un vector de diseño y $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^\top$ un vector de coeficientes de regresión. El modelo de regresión binaria, está dado por

$$p_i = \Pr(y_i = 1) = F(\mathbf{x}_i^\top \boldsymbol{\beta}), \quad i = 1, 2, \dots, n, \tag{1}$$

donde $F(\cdot)$ denota a la fda (función de distribución acumulada), su inverso $F^{-1}(\cdot)$ de acuerdo a la teoría del modelo lineal general es llamado link.

Un link resulta ser simétrico cuando la fda procede de una fdp (función de densidad de probabilidad) simétrica. Por supuesto, un link resulta ser asimétrico cuando la fda procede de una fdp asimétrica. También, un link asimétrico puede reducirse a un link simétrico. Chen *et al.* (1999), Bazán *et al.* (2014) y Chocotea (2014) argumentan que, cuando la probabilidad de una respuesta binaria se aproxima a 0 en una tasa diferente que cuando se aproxima a 1, los modelos con link asimétrico son adecuados.

La siguiente definición es una generalización de los resultados presentados por Bolfarine y Bazán (2010), Pewsey *et al.* (2012), Martínez–Flórez *et al.* (2013, 2014), Bazán *et al.* (2014), y Chocotea (2014), y es central para nuestro desarrollo.

Definición 2. Denotando por $F(\cdot)$ a la fda de una fdp simétrica alrededor del origen con soporte en la recta real. Sean

$$\mathcal{F}_A = \{F_1(\eta; \lambda) = [F(\eta)]^\lambda, \lambda \in \mathbb{R}_+\} \quad \text{y} \quad \mathcal{F}_B = \{F_2(\eta; \lambda) = 1 - [F(-\eta)]^\lambda, \lambda \in \mathbb{R}_+\} \tag{2}$$

dos clases de fda's.

Los modelos definidos en (2) tienen propiedades atractivas convenientes: (a) cuando $\lambda = 0$ se reducen a un modelo con link simétrico; (b) la asimetría puede ser determinada por λ ; y (c) $F_1(\eta; \lambda) + F_2(-\eta; \lambda) = 1$. El modelo de regresión binaria power o recíprocal power, es caracterizado por

$$p_i = \Pr(y_i = 1) = F_\lambda(\mathbf{x}_i^\top \boldsymbol{\beta}), \tag{3}$$

donde $F_\lambda(\cdot)$ denota a la fda $F_1(\cdot; \lambda)$ o $F_2(\cdot; \lambda)$ de (2), respectivamente.

Cuadro 1 Características de los nuevos modelos

Clase	$F(\cdot)$	$\lambda > 0$	Modelo	Link
\mathcal{F}_A	$\Phi(\cdot)$	$\lambda \neq 1$	power probit	asimétrico
		$\lambda = 1$	probit	simétrico
	$\Psi(\cdot)$	$\lambda \neq 1$	power logit	asimétrico
		$\lambda = 1$	logit	simétrico
\mathcal{F}_B	$\Phi(\cdot)$	$\lambda \neq 1$	recíprocal power probit	asimétrico
		$\lambda = 1$	probit	simétrico
	$\Psi(\cdot)$	$\lambda \neq 1$	recíprocal power logit	asimétrico
		$\lambda = 1$	logit	simétrico

$\Phi(\cdot)$ denota a la fda procedente de la fdp normal estándar

$\Psi(\cdot)$ denota a la fda procedente de la fdp logística estándar

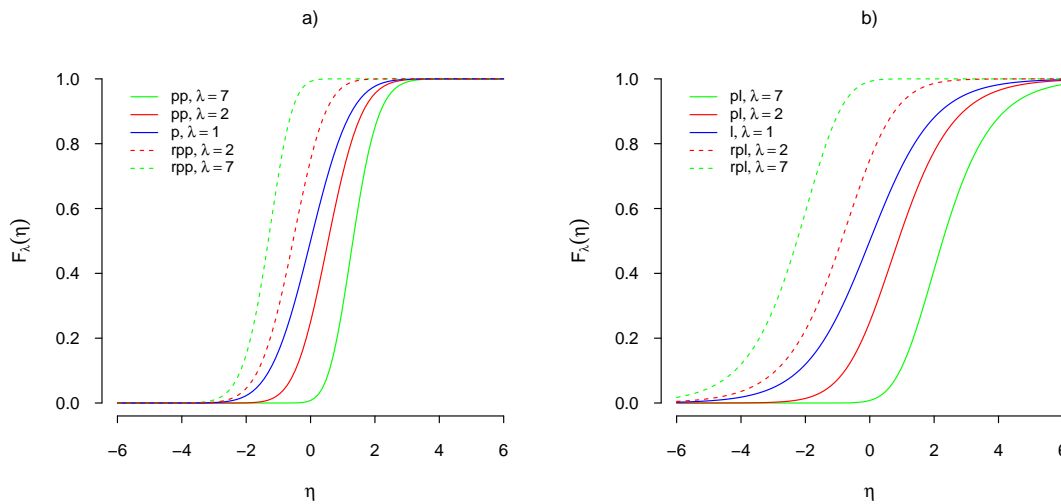


Figura 1 Curvas de probabilidad de los modelos a) power probit (pp), probit (p), recíprocal power probit (rpp), b) power logit (pl), logit (l) y recíprocal power logit (rpl).

Para la estimación e inferencia, se puede utilizar la data aumentada (Bolfarine y Bazán 2010; Chocotea, 2014; Kroese y Chan, 2014). Detalles importantes de modelos de variables latentes en el análisis de datos categóricos aparecen en Agresti y Kateri (2014). Las variables latentes evitan trabajar con la verosimilitud tipo Bernoulli (Bolfarine y Bazán, 2010; Chocotea, 2014).

3. Análisis bayesiano

Asumiendo independencia entre las previas (Bazán *et al.*, 2014), la estructura jerárquica está dada por

$$y_i | \mathbf{x}_i, \boldsymbol{\beta}, \lambda \sim \text{Ber}(F_\lambda(\mathbf{x}_i^\top \boldsymbol{\beta})) \quad \boldsymbol{\beta} \sim \pi_1(\boldsymbol{\beta}) \quad \lambda \sim \pi_2(\lambda). \tag{4}$$

Cuando se tiene varios modelos alternativos, surge la pregunta inmediata ¿cómo los comparo y selecciono al mejor? Lo primero que no se debe olvidar es que, no siempre existe un mejor modelo, y se puede tener un conjunto de modelos que ostenten un desempeño similar. El criterio principal de bondad de ajuste para comparar modelos Bayesianos es, el Criterio de Información de la Devianza (*Deviance Information Criterion*: DIC). Simplificando, un modelo es mejor que otro cuando su DIC es menor.

4. Aplicación

¿Me siento satisfecho con la calidad del servicio de salud pública en general?, es una de las preguntas que se hacen miles de paceños y paceñas, pero ¿cómo podemos analizar los motivos de satisfacción o insatisfacción respecto a un amplio espectro de situaciones que condicionan su calidad de vida y su felicidad?.

El Observatorio La Paz Cómo Vamos el 2012 realizó la segunda Encuesta de Percepción Ciudadana sobre la Calidad de Vida en la ciudad de La Paz. Ilustremos el análisis a partir del tópico salud, las variables son: *tiempo*, e indica la insatisfacción (1) o satisfacción (0) con el tiempo de demora para conseguir una consulta médica en el servicio público; *trato*, e indica la insatisfacción (1) o satisfacción (0) con la calidad del trato en los centros de salud y hospitales públicos; *infraequi*, e indica la insatisfacción (1) o satisfacción (0) con la calidad de la infraestructura y equipamiento de los centros de salud y hospitales públicos; *esfurespo*, e indica la insatisfacción (1) o satisfacción (0) con el esfuerzo y responsabilidad de los paceños para cuidar la salud de la familia; y *servicio*, e indica la insatisfacción (1) o satisfacción (0) con la calidad del servicio de salud pública en general. Las variables explicativas serán *tiempo*, *trato*, *infraequi* y *esfurespo*, y la variable dependiente será *servicio*. Debido a la no respuesta en algunas preguntas, la data se reduce a 955.

Cuadro 2 ➔ Sumario descriptivo de variables

Variable	Insatisfecho	Satisfecho	Media	DE
tiempo	783 (82.0 %)	172 (18.0 %)	0.820	0.384
trato	781 (81.8 %)	174 (18.2 %)	0.818	0.386
infraequi	782 (81.9 %)	173 (18.1 %)	0.819	0.385
esfurespo	730 (76.4 %)	225 (23.6 %)	0.764	0.424
servicio	776 (81.3 %)	179 (18.7 %)	0.813	0.390

DE: Desviación Estándar

Comparemos a siete modelos de regresión binaria, al probit (\mathcal{M}_1), al logit (\mathcal{M}_2), al cloglog (\mathcal{M}_3), al power probit (\mathcal{M}_4), al reciprocal power probit (\mathcal{M}_5), al power logit (\mathcal{M}_6) y al reciprocal power logit (\mathcal{M}_7). Las fdp's a posteriori son difíciles de obtener, por lo tanto las buenas aproximaciones son dadas bajo los métodos de Monte Carlo vía Cadenas de Markov (*Markov Chain Monte Carlo*: MCMC). Sus sintaxis pueden ser implementadas en los softwares RStudio, R, OpenBUGS, WinBUGS, o SAS, o también online, la nueva interfaz se llama WebBUGS y solo es necesario registrarse, la dirección de la web es <http://WebBUGS.psychstat.org>.

Cuadro 3 ➔ Sumario de la inferencia para los parámetros de regresión y valores del DIC

Parámetro	\mathcal{M}_1	\mathcal{M}_2	\mathcal{M}_3	\mathcal{M}_4	\mathcal{M}_5	\mathcal{M}_6	\mathcal{M}_7
β_1	-1.705	-3.076	-2.280	-6.560	-1.008	-8.698	-3.502
β_2	1.046	1.895	1.028	2.088	1.813	3.167	1.490
β_3	0.512	0.906	0.585	1.148	0.898	1.841	0.784
β_4	1.003	1.763	1.054	2.091	1.736	3.304	1.420
β_5	1.223	2.275	1.081	2.364	2.212	3.369	1.732
λ				0.129	0.264	0.311	2.474
DIC	416.5	421.1	417.4	411.5	398.2	417.0	403.6

El valor del DIC determina que, el modelo con mejor desempeño o el más apropiado es el recíproco power probit (\mathcal{M}_5).

Cuadro 4  Sumario de la inferencia para los parámetros de regresión de \mathcal{M}_5

Parámetro	A posteriori			Intervalo HPD 95 %	
	Media	Mediana	DE	Inferior	Superior
β_1	-1.008	-1.008	0.381	-1.720	-0.259
β_2	1.813	1.777	0.496	0.968	2.832
β_3	0.898	0.869	0.376	0.214	1.718
β_4	1.736	1.719	0.455	0.940	2.621
β_5	2.212	2.204	0.534	1.261	3.200
λ	0.264	0.197	0.187	0.100	0.847

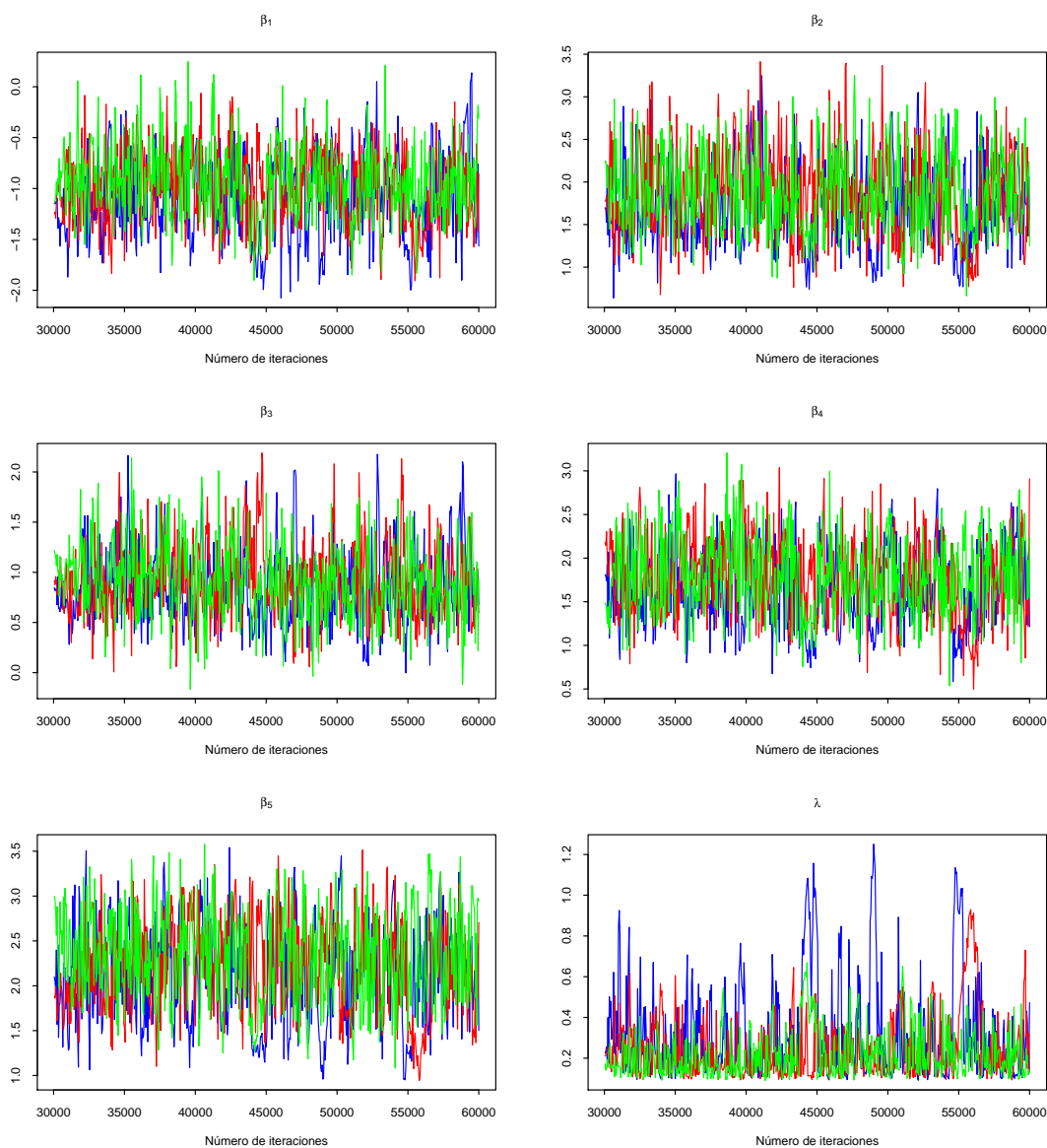



Figura 2  Historial de las cadenas de Markov de las fdp's a posteriori de \mathcal{M}_5 (1.^a línea azul, 2.^a línea roja y 3.^a línea verde).

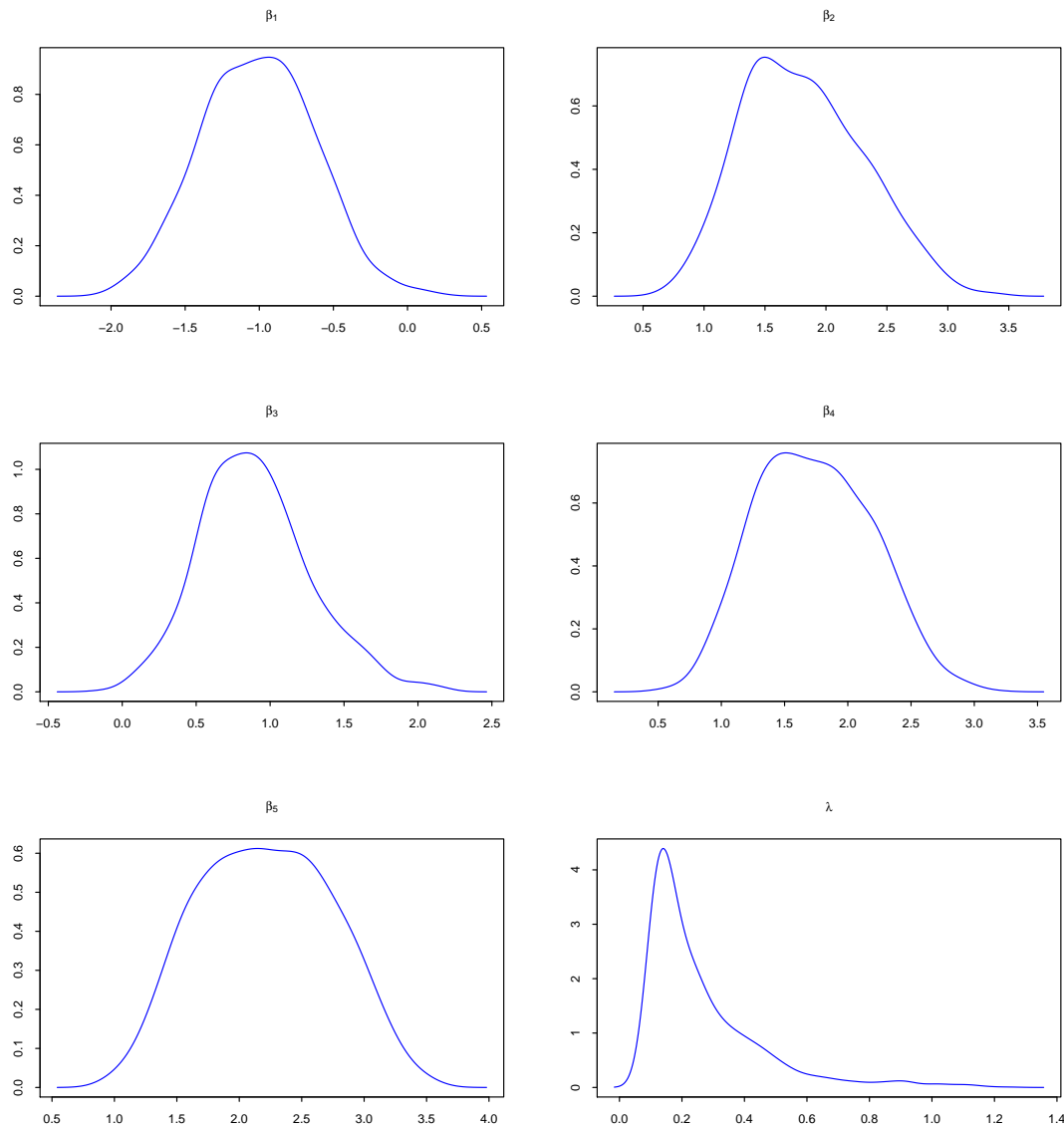


Figura 3 ▶ Plots de las fdp's a posteriori de \mathcal{M}_5 (1.^a cadena).

El modelo seleccionado para el análisis, está dado por

$$\Pr(\text{servicio}_i = 1) = 1 - [\Phi(1.008 - 1.813\text{tiempo}_i - 0.898\text{trato}_i - 1.736\text{infraequ}_i - 2.212\text{esfurespo}_i)]^{0.264},$$

donde $\Phi(\cdot)$ denota a la fda de la $\mathcal{N}(0,1)$.

La insatisfacción con la calidad del servicio de salud pública en general se debe en primer lugar a la insatisfacción con el esfuerzo y responsabilidad de los pacaños para cuidar la salud de la familia y en segundo lugar a la insatisfacción con el tiempo de demora para conseguir una consulta médica en el servicio público.

5. Conclusiones

Está claro que, la actualización del investigador va coherente con los nuevos aplicativos estadísticos para tomar mejores decisiones con los mejores modelos. Por ahora los cuatro modelos resaltaron por tener mejor desempeño.

Un estudio fortificante para una buena teoría, es el de las fdp's de las distribuciones power normal y power logística. Las extensiones van hacia los modelos de teoría de la respuesta al ítem (TRI).

Referencias

- [1] Agresti, A., y Kateri, M. (2014) Some Remarks on Latent Variable Models in Categorical Data Analysis. *Communications in Statistics – Theory and Methods* **43**, 801–814.
- [2] Bazán, J. L., Romeo, J. S., y Rodrigues, J. (2014) Bayesian Skew–Probit Regression for Binary Response Data. *Brazilian Journal of Probability and Statistics* **28**(4), 467–482.
- [3] Berkson, J. (1944) Application of the Logistic Function to Bio–Assay. *Journal of the American Statistical Association* **39**, 357–365.
- [4] Bliss, C. I. (1935) The Calculation of the Dosage–Mortality Curve. *Annals of Applied Biology* **22**, 134–167.
- [5] Bolfarine, H., y Bazán, J. L. (2010) Bayesian Estimation of the Logistic Positive Exponent IRT Model. *Journal of Educational and Behavioral Statistics* **35**, 693–713.
- [6] Chen, M.–H., Dey, D. K., y Shao, Q.–M. (1999) A New Skewed Link Model for Dichotomous Quantal Response Data. *Journal of the American Statistical Association* **94**, 1172–1186.
- [7] Chocotea, O. (2014) *Modelos de Regresión Binaria Bayesiana Skew Probit*. Tesis de Licenciatura, Universidad Mayor de San Andrés, La Paz.
- [8] Gumbel, E. J. (1958) *Statistics of Extremes*. Columbia University Press, New York.
- [9] Kroese, D. P., y Chan, J. C. C. (2014) *Statistical Modeling and Computation*. Springer, New York.
- [10] Martínez–Flórez, G., Bolfarine, H., y Gómez, H. W. (2013) The Alpha–Power Tobit Model. *Communications in Statistics – Theory and Methods* **42**(4), 633–643.
- [11] ——— (2014) An Alpha–Power Extension for the Birnbaum–Saunders Distribution. *Statistics* **48**(4), 896–912.
- [12] Pewsey, A., Gómez, H. W., y Bolfarine, H. (2012) Likelihood–Based Inference for Power Distributions. *Test* **21**, 775–789.