III Congreso Internacional de Estadística

MINICURSO

MODELOS DE REGRESIÓN PARA RESPUESTA DICOTÓMICA

(versión preliminar)

Dr. Jorge Luis Bazán

jlbazan@icmc.usp.br

Trujillo-Perú, Octubre del 2015

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo
São Carlos/São Paulo/Brasil

Conteúdo

Lista de Figuras

Lista de Tabelas

1. Introducción

Una variable aleatoria es considerada de respuesta binária o de respuesta dicotómica cuando puede tomar dos posibles valores o categorías. Por ejemplo suceso o falla, resultado positivo o resultado negativo, correcto o incorrecto. Ese tipo de datos son comunes en muchas áreas entre las cuales destacamos las ciencias sociales, médicas, agricultura, genética, educación y psicología. En este caso, se puede asumir que esta respuesta sigue una distribución Bernoulli o Binomial. Cuando este tipo de variable quiere ser explicada considerando otras variables, llamadas variables explicativas (las cuales no son aleatorias y son fijas o se asumen conocidas), nos encontramos en el caso de la regresión binaria. Así, los modelos de regresión binaria son modelos estadísticos usados para predecir la probabilidad de una respuesta binaria en función de diversas variables explicativas o predictores. Como indicado en Bazán y Bayes (2010), conjuntos de datos que requieren este tipo de análisis se encuentran en áreas tan diversas como ingeniería, ciencias naturales, educación, etc. Por ejemplo, el hecho de que un paciente sobreviva o no a una enfermedad puede ser explicado por variables como el tratamiento aplicado, edad, etc; o el resultado de un examen de admisión (aprueba o no aprueba) puede ser influenciado por el nivel de conocimiento del alumno en matemáticas, lenguaje, etc.

En este tipo de modelo se considera una función de enlace entre los predictores y la probabilidad se suceso que representa el error aleatório subyacente asociado con la respuesta 1 o la respuesta 0. Los modelos más conocidos de regresión binaria son la regresión logística, cuando el enlace es el logit (el error latente tiene distribución de probabilidad logística estándar), y la regresión probit, cuando dicho enlace es el probit (el error latente tiene distribución de probabilidad normal estándar). En ambos casos, la función de enlace (el error latente) es simétrica y determina que la curva correspondiente sea también simétrica, es decir, ideal para datos balanceados de unos o ceros. En este trabajo presentaremos estos casos con mayor detalle.

Sin embargo este tipo de suposiciones son restrictivas y no aplicables cuando se tiene una mayor frecuencia de una de las respuestas binarias, es decir, datos desbalanceados. Varios autores, entre ellos Collet (2003), Czado y Santner (1992), Chen, Dey y Shao (1999), Bazán y Millones (2008) mostraron que enlaces asimétricos pueden ser más apropiados que enlaces simétricos en situaciones específicas. Este ocurre, como dicho por Bazán y Bayes (2010), por ejemplo cuando se quiere modelar o explicar el resultado de la probabilidad de tener un paro cardíaco en una población considerando una serie de características individuales como si fuma o no, nivel de colesterol, etc, donde en general la proporción de personas con este mal es muy baja. Pero también por ejemplo para explicar la probabilidad de default (no pago), la probabilidad de uso de un determinado seguro, credit escoring, entre otros ejemplos en el ámbito financiero. Por ello diversos enlaces fueron propuestos en la literatura en los últimos 30 años.

Como ha sido presentado en Bazán y Millones (2008), Bazán y Bayes (2010), el estudio de la regresión binaria es una área importante para el modelamiento estadístico y por tanto es conveniente tener modelos alternativos, por ejemplo diferentes usando enlaces simétricos, que puedan ser estudiados e implementados fácilmente, pero también enlaces asiméticos que puedan ser analizados como alternativos a los enlaces tradicionales.

En este documento preparado para un minicurso del III Congreso Internacional de Estadística 2015 desarrollado en la UNT presentamos principalmente los modelos más conocidos de la regresión binaria tanto considerando estimación clásica como estimación bayesiana. Mate-

rial complementario se encuentra disponible en `http://www.icmc.usp.br/pessoas/jlbazan/presentations.html`.

Si bien mostraremos algunos nuevos enlaces recomendamos fuertemente el libro de Bazán y Bayes (2010) para quienes tienen interés en estudiar e implementar diversos modelos de regresión binaria asimétricos propuestos en la literatura reciente.

El resto del documento está organizado de la siguiente manera. En el capítulo 1 presentaremos la regresión binaria. En el capítulo 2 presentamos la estimación clásica en regresión binaria, considerando el método de estimación de máxima verosimilitud, y temas relacionados como prueba de hipótesis para la significación de los efectos de las variables explicativas, análisis de desvio y selección de modelos, significación de los efectos de las variables explicativas, diagnóstico del modelo ajustado, análise de residuos y métodos auxiliares para el ajuste del modelo como el poder predictivo del modelo ajustado incluyendo las llamadas curvas ROC. (Receiver Operating Characteristic en inglés, o Característica Operativa del Receptor en español). En el capítulo 3 se presenta la Inferencia Bayesiana en Regresión Binaria incluyendo otros tópicos de Inferencia Bayesian como las condiciones para la existencia de la distribución posterior de los parámetros de la regresión y se presentan diversas alternativas para la comparación de modelos desde la perspectiva bayesiana. En el capítulo 4 se muestran dos aplicaciones principales, una en la que se emplea datos del transbordador espacial Challenger considerando tanto el enfoque clásico cuanto el enfoque bayesiano. La segunda aplicación corresponde al uso de diferentes modelos de regresión binaria para predecir la probabildad de erradicar el cultivo de coca entre agricultores peruanos. Finalmente en Anexo presentamos los programas empleados incluyendo códigos en los programas de uso libre R y WinBUGS.

2. Modelo de Regresión Binaria

Considere: $y = (y_1, y_2, \ldots, y_n)^T$ observaciones de una v.a. dependiente Y_i, $x_i = (x_{i1}, \ldots, x_{ik})'$ variables explicativas, $i = 1, 2, \ldots, n$. En el modelo de regresión binaria se asume que:

$$\begin{aligned} Y_i &\sim Bernoulli(p_i) \\ p_i &= F(\eta_i) = F\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) \\ i &= 1, 2, \ldots, n \end{aligned}$$

donde

- Y_i una variable binaria tal que $Y_i = 1$ ocurre con probabilidad p_i.
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})^T$ un vector con los valores de k variables explicativas.
- $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)^T$ un vector de k coeficientes de regresión.
- $F(.)$ denota una función de distribución acumulada (fda) definida para valores en la recta. La función inversa $F^{-1}(.)$ es comummente denominada función de enlace.
- $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_1 + \beta_2 x_{2i}, ..., \beta_k x_{ki}$ es el i-ésimo predictor linear con $x_{1i} = 1$.

Cuando F es una fda de una distribución simétrica la función de enlace resultante es simétrica y tiene una forma simétrica alrededor de $p_i = 0{,}5$. En el caso que F sea la fda de una normal estándar tenemos el enlace probit, $F(t) = \Phi(t)$ y en el caso de que F sea la fda de una distribución logística obtenemos el enlace logit, $F(t) = \frac{e^t}{1+e^t}$.

Específicamente en la regresión Logística tenemos que

$$F\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) = \frac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

y en la regresión probito usando la fda de la distribución Normal estándar, tenemos

$$F\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) = \Phi\Big(x_i' \beta\Big)$$

Cuando $F(t) = 1 - exp(-exp(t))$ corresponde a la fda da Gumbel, se tiene la regresión binaria de de valor extremo. En este caso $F^{-1}(.)$ es llamada de función de enlace loglog complementar. Cuando consideramos $F(t) = exp(-exp(-t))$ cque corresponde a la fda de Gumbel reversa, se tiene una regresión binaria de valor extremo reverso y este caso $F^{-1}(.)$ es llamada función de enlace loglog. En la figura 1 presentamos las correspondientes curvas para los diferentes enlaces presentados.

Un resumen de enlaces tradicionales para datos con respuesta binaria son mostrados en la siguiente tabla

donde $\phi(.)$ denota la función de la distribución Normal(0,1), $arctg$ = arco tangente y $F(.)$ denota la función de la distribución Cauchy(0,1) o t de student com 1 grado de libertad $\sim t_{(1\ g.l.)}(0, 1)$

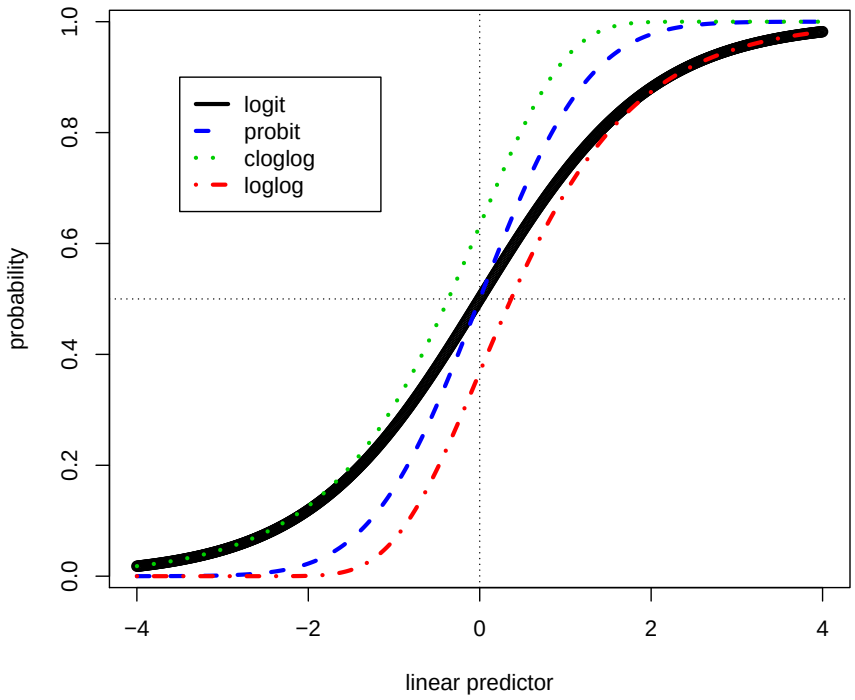

Figura 1 *Funciones de enlace en Regresión Binária. Logito y Probito son simétricos al rededor de p=0.5 y el preditor linear = 0, que es diferente para Cloglog e Loglog.*

$p_i = \theta(x) = F\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$	enlaces parametricos alternativos
$\frac{\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}$	logito$\Rightarrow \ln\left(\frac{\theta(x)}{1-\theta(x)}\right)$
$\Phi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$	probito$\Rightarrow \Phi^{-1}\left(\theta(x)\right)$
1-exp $\left\{-\exp\left\{\boldsymbol{x}_i^T\boldsymbol{\beta}x\right\}\right\}$	cloglog$\Rightarrow \ln\left(-\ln\left(1-\theta(x)\right)\right)$
$\exp\left\{\exp\left\{\boldsymbol{x}_i^T\boldsymbol{\beta}\right\}\right\}$	loglog$\Rightarrow \ln\left(\ln\left(\theta(x)\right)\right)$
$\frac{1}{2}+\frac{\arctan\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{\pi}$	cauchy$\Rightarrow F^{-1}\left(\theta(x)\right)$

Cuadro 1 *Enlaces tradicionales en regresión binaria*

En este caso tenemos que los enlaces simétricos son logito, probito y cauchy, y como enlace asimetrico: complemento log-log y loglog.

2.1. Enlaces Asimétricos en Regresión Binaria

Chen et al (1999) 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. Un ejemplo muy popular es el enlace log-log complementario o cloglog mostrado arriba, 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. Otra forma sin embargo es considera la fda de distribuciones asimétricas más generales para construir enlaces asimétricos.

Propuestas de otros enlaces simétricos en la literatura pueden ser revisados en Prentice (1976), Aranda-Ordaz (1981), Guerrero y Johnson (1982), Stukel (1988), Czado e Santner (1992a,b), Nagler (1994), Chen et al., (1999), Basu y Mukhopadhyay (2000), Haro-López, et al. (2000). Sin embargo, enlaces simétricos podem ser inadequadas e mal especificadas como puede ser apreciado en la Figura 2 y 3. Ver también Collet (2003). Algunos enlaces probito asimétricos fueron propuestos por Czado (1994), Chen et al. (1999) y Bazán et al. (2005). Mayores detalles en Bazán, Bolfarine y Branco (2010).

Entre los enlaces asimétricos que nos gustaría destacas están aquellos basados en las siguientes fdas

$$F(t) = 1 - (1 + e^{t})^{-\lambda} \text{ y } F(t) = (1 + e^{-t})^{-\lambda} \quad \lambda > 0$$

estos enlaces son logit asimetrizados y son conocidos como scobit y power logit, respectivamente, e incluyen al enlace logit como caso especial cuando el parámetro $\lambda = 0$. Para una revisión de estos enlaces ver Prentice (1976) y Nagler (1994).

En la Figura 2, mostramos la curva, histograma, densidad y boxplot de datos simulados de un modelo de regresión binaria usando el enlace power logístico com $lambda = 4$, um modelo logístico (power logístico com $lambda = 0$) y un modelo power logístico com $\lambda = 0{,}25$. Observamos que únicamente tenemos datos balanceados solamente en el caso logístico.

También en la Figura 3, mostramos el sesgo que se produce cuando se usa el modelo de regresión binaria con enlace logístico (modelo mal especificado) cuando los datos siguen realmente un modelo de regresión binaria con otros enlaces. Nosotros analizamos el caso de datos usandi los enlaces logit, cauchy, scobit, power probit, cloglog y loglog. Nosotros encontramos que las estimativas de los coeficientes de regresión (intercepto y pendiente) son estimados sesgadamente si usamos un modelo incorrecto o mal especificado, especialmente si la muestra es pequeña.

otro modelo

Otro grupo de enlaces asimétricos son los siguientes tres que se basan en el uso de la fda de una distribución normal asimétrica Bazán, Bolfarine y Branco (2006 y 2010), la cual puede ser representada de manera general la siguiente manera:

$$F(t; \boldsymbol{\theta}) = 2\Phi_2(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Omega})$$

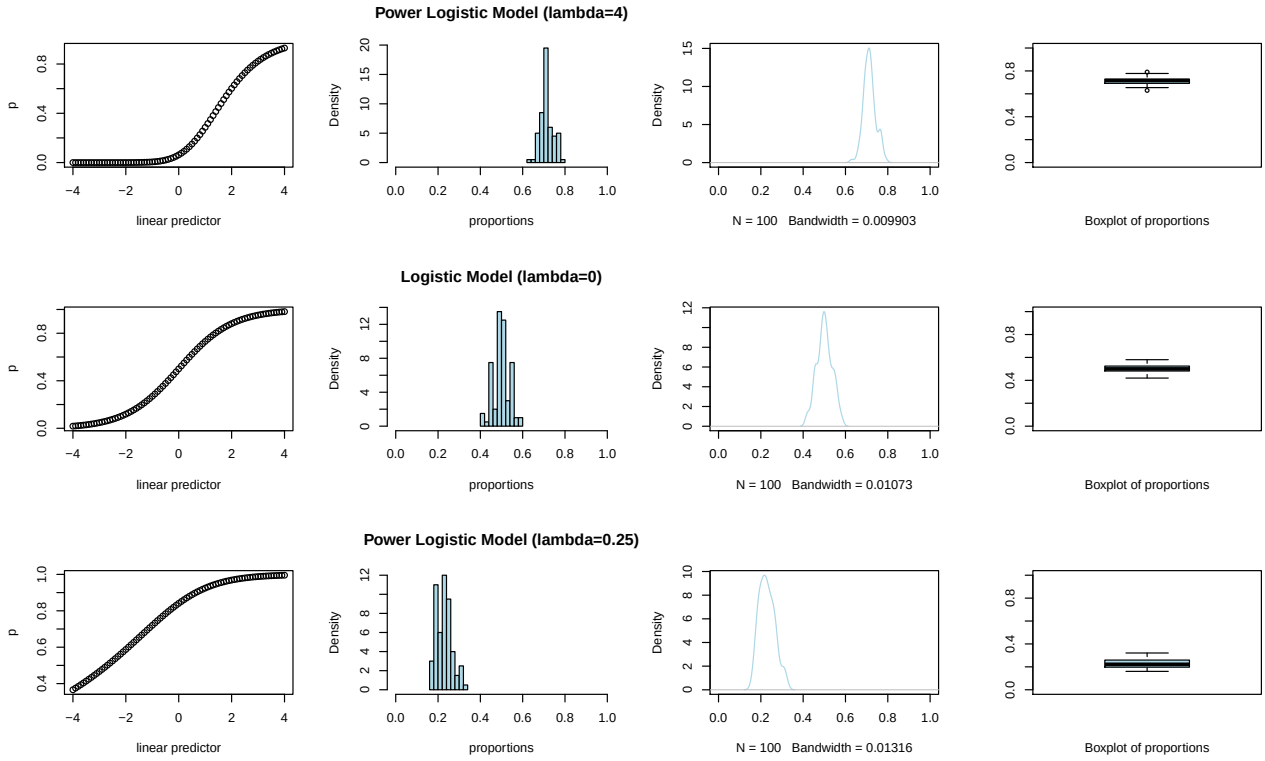

Figura 2 *Distribución de 1′s e 0′s (n=81) en 100 muestras simuladas considerando el enlace Power logit (PL) considerando un predictor fijo. Note que PL es más adecuado para altos e bajos valores de proporción observada e similar que el modelo logit para proporciones entorno de 0.5*

donde $\boldsymbol{x} = (t, 0)^{'}$; $\boldsymbol{\theta} = (\mu, \sigma^2, \lambda)^{'}$; $\Phi_2(.)$ representa la distribución acumulada de una distribución normal bivariada con parámetros $\boldsymbol{\mu} = (\mu, 0)^{'}$ y $\boldsymbol{\Omega} = \begin{bmatrix} \sigma^2 & -\delta \\ -\delta & 1 \end{bmatrix}$; y $\delta = \dfrac{\lambda}{\sqrt{1+\lambda^2}}$,

Considerando los siguientes casos tenemos tres grupos de diferentes enlaces

- Si $\boldsymbol{\theta} = (0, 1+\lambda^2, -\lambda)$, se obtiene el enlace probit asimetrizado propuesto en Chen et al (1999) denominado CDS skew probit.

- Si $\boldsymbol{\theta} = (0, 1, \lambda)$, se obtiene el enlace propuesto por Bazán, Branco y Bolfarine (2006) denominado BBB skew probit.

- Si $\boldsymbol{\theta} = \left(-\dfrac{\sqrt{2}\delta}{\sqrt{\pi - 2\delta^2}}, \dfrac{\pi}{\pi - 2\delta^2}, \lambda\right)$, obtenemos el enlace denominado estándar probit asimetrizado (Bazán, Bolfarine y Branco 2006 y 2010), denominado Standard skew probit.

En estos tres enlaces, λ es el parámetro que controla la asimetría, así tenemos que para valores negativos (positivos) de λ tenemos asimetría negativa (positiva).

Esta clase de modelos puede verse también como perteneciente a la clase de mezclas de dis-

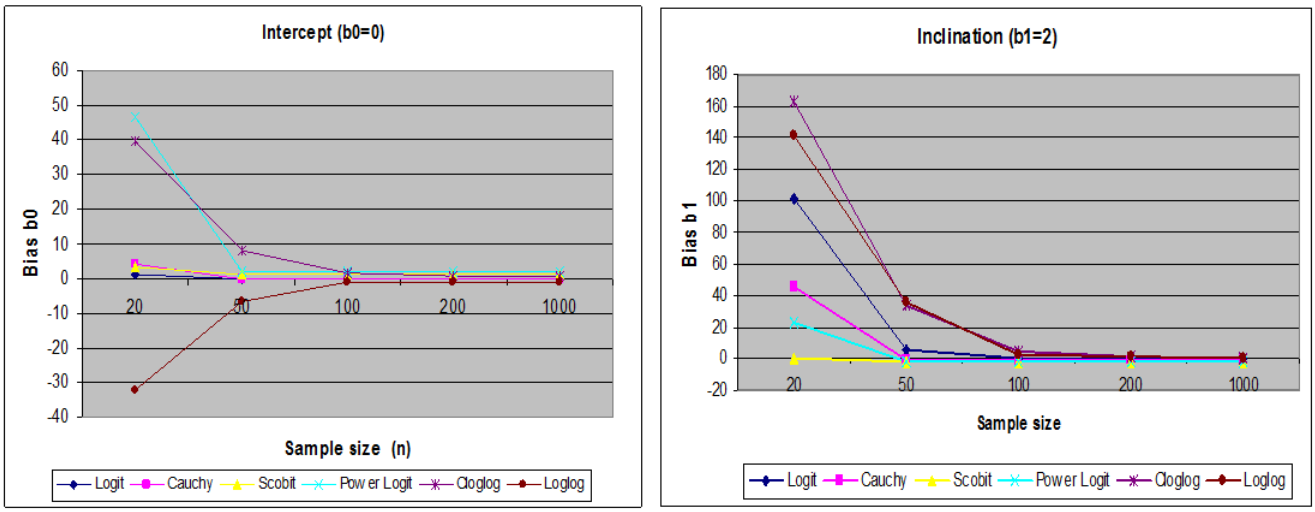

Figura 3 *Sesgo en la estimación de los coeficientes de regresión binaria simple con diferentes enlaces cuando el enlace logito es usado como estándar (modelo mal especificado). Note que cuando o tamaño de la muestra se incrementa el sesgo disminuye, sin embargo áun es importante*

tribuciones elípticas propuestas por Basu y Mukhopadhyay (2000) dada por:

$$F = \left\{ F(.) = \int_{[0.\infty]} H(. \mid v) dG(v) \right\}$$

donde G es la función de distribución acumulada $[0, \infty >$ y H es una distribución elíptica. En este caso el CDS skew probit considera una clase de mezclas de normales donde la medida de mezcla es la distribución normal positiva con función de densidad dada por $g(x) = 2\phi(x)$, $x > 0$, con $\phi(.)$ siendo la función de densidad de la normal estándar. Otro caso interesante cuando se mezcla la normal positiva con H la función de distribución acumulada de la distribución logística es conocida como skew logistic o skew logit (ver Chen, Dey y Shao, 2001).

3. Inferencia Clássica en Regresión binaria

En este capítulo abordamos la inferencia clássica de regresión binaria considerando el método de máxima verosimilitud. También estudiamos otros temas como la prueba de hipótesis para la significación de los efectos de las variables explicativas, análisis de desvio y selección de modlos, significación de los efectos de las variables explicativas, diagnóstico del modelo ajustado, análise de residuos y métodos auxiliares para el ajuste del modelo como el poder predictivo del modelo ajustado incluyendo las llamadas curvas ROC.

3.1. Estimación de máxima verosimilitud

Similar al caso de la regresión linear, en la regresión binaria también el interés es modelar $E(Y \mid \mathbf{x})$. Note que en este caso $E(Y_i \mid \mathbf{x}_i) = P(Y_i = 1\mathbf{x}_i) = p_i \in [0,1]$. Note también, como vimos en el capítulo anterior la relación entre x_i y $E(Y_i \mid \mathbf{x}_i)$ es en forma de S que como vimos puede ser muy bien representada usando una distribución acumulada de alguna distribución continua con soporte en los números reales. El caso más popular es la regresión logística donde tenemos que

$$p_i = \frac{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}, \quad 1-p_i = \frac{1}{1+exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}$$

Adicionalmente la transformación logito es

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

y la razón de chances

$$odds = \frac{p_i}{1-p_i} = exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$$

Con el propósito de estimar un modelo de regresión binaria, consideramos la función de verosimilitud definida como

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{x}_i) = \prod_{i=1}^{n} F\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^{y_i}\left(1-F\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right)^{1-y_i}$$

y considerando el logaritmo de la verosimilitud tenemos que

$$l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} y_i log(F\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)) + (1-y_i)log(1-F\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right))$$

Esta función puede ser maximizada usando algún procedimiento computacional para los valores de $\boldsymbol{\beta}$. En este caso los valores que maximizam $L(\boldsymbol{\beta})$ son denotadas por $\hat{\boldsymbol{\beta}}$. Usando resultados de teoría asintótica es posible obtener que $\hat{\boldsymbol{\beta}}$ sigue una distribución aproximadamente normal, es decir $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sum(\boldsymbol{\beta}))$. En este caso $\sum(\boldsymbol{\beta}) = [I(\boldsymbol{\beta})]^{-1}$ es la matriz de variancias-covariancias y $I(\boldsymbol{\beta})$ es la matriz conteniendo el negativo de las derivadas parciales de segundo orden de $l(\boldsymbol{\beta})$. Estimadores se obtiene para evaluar $\sum(\boldsymbol{\beta})$ en $\hat{\boldsymbol{\beta}}$

3.2. Significación de los efectos de las variables explicativas

Una prueba de hipótesis relativa a la significación conjunto de todos los parámetros β_j, $= 1, 2, \ldots, k$ es la Prueba de Razón de Verosimilitudes (TRV) que es definida como

$$TRV = -2log\left(\frac{L_S}{L_C}\right) = 2log(L_C) - 2log(L_S) \sim \chi^2(q)$$

donde L_S es la función de verosimilitud asociada al modelo sin incluir las variables que están siendo investigadas. Esto es llamado también modelo de refrencia, L_C es la función de verosimilitud asociada al modelo incluyendo las variables que están siendo investigadas. Esto es llamado también modelo sobre investigacipon y q es la diferencia de parámetros entre ambos modelos. En este caso $2log(L_C)$ y $2log(L_S)$ son llamadas funciones de desvió (deviance en inglés) y la estadística TRV puede ser vista como diferencias de desvíos entre el modelo sobre investigación (con las variables a ser analizadas) y el modelo de referencia (sin incluir estas variables). De esta manera, este tipo de análisis es llamado también análisis de desvío y puede ser usado para seleccionar modelos

3.3. Análisis de desvio y selección de modelos

Considere que está siendo analizado la influencia de dos covariables categóricas x_1 y x_2 con dos categorías cada una y de sus interacciones $x_1 \times x_2$ en una respuesta binaria. Los resultados pueden ser presentados en el siguiente cuadro

Modelos	gl	Desvios	TRV	q
Nulo	n	D_n		
x_1	$n-1$	D_1	$D_n - D_1$	1
x_2	$n-2$	D_2	$D_1 - D_2$	1
$x_1 \times x_2$	$n-3$	D_3	$D_2 - D_3$	1

Cuadro 2 *Desvios em modelos anidados o secuenciales*

En presencia de datos faltantes, el tamanho de la muestra en los modelos secuenciales dependerá de las variables X_k que lo componen y entonces la estadística TRV presentará problemas.

3.4. Significación de los efectos de las variables explicativas

En este caso se puede usar el Test de Wald es

a) Para probar la hipótesis relativa a un parámetro

$$H_0 : \beta_j = 0, j = 1, \ldots, k$$

$$W = \frac{\hat{\beta_j}^2}{var(\hat{\beta})} \sim \chi_1^2$$

b) Para probar hipótesis relativas a $k \geq 2$ parámetros

$$H_0 : \beta^* = \mathbf{0},$$

β^* es un vector $k \geq 1$

$$W = (\hat{\beta}^*) \left(\hat{\sum} (\hat{\beta}^*) \right)^{-1} (\hat{\beta}^*) \sim \chi^2_1 \sim \chi^2_q$$

3.5. Diagnóstico del modelo ajustado

Para evaluar el ajuste del modelo pueden ser consideradas dos medidas de bondad de ajuste: el test chi-cuadrado de pearson Q_P y la estadística de razón de verosimilitud chi-cuadrado Q_L.

Sobre la hipótesis H_0: El modelo es satisfactorio. Se puede usar las siguientes estadísticas que resumen la concordancia entre los valores observados y los valores predichos para el modelo.

$$Q_P = \sum_{i=1}^{s} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_m$$

$$Q_L = 2 \sum_{i=1}^{s} \sum_{j=1}^{2} n_{ij} log(\frac{n_{ij}}{e_{ij}}) \sim \chi^2_m$$

donde las frecuencias esperadas o conteos sobre el modelo ajustado que son definidas como $e_{ij} = n_{i+}\hat{\theta}(\mathbf{x_i})$, $j = 1$ y $e_{ij} = n_{i+}1 \hat{-} \theta(\mathbf{x_i})\, j = 2$

- n_{i+} son sujetos de la i-ésima subpoblación formada de la tabla de datos $s \times 2$ y j son los grupos para una de las respuestas binarias.
- $\hat{\theta}(\mathbf{x_i})$ es la probabilidad $p(Y = 1 \mid \mathbf{x_i})$ predicha por el modelo ajustado considerando los valores estimados $\hat{\beta}^*$.
- m es el número de subpoblaciones menos el número de parámetros del modelo ajustado.

Es recomendable que para cada uno de los grupos $n_{i+} > 10$, también que al menos el 80 % de los valores esperados satisfagan $e_{ij} > 5$ y el resto debe ser al menos de 2. Cuando esto no ocurre las estadísticas anteriores son limitadas. También en presencia de variables continuas tenemos una frecuencia muy pequena para la gran mayoria de las s sub poblaciones, esto hace menos válido el uso de Q_L y Q_P. Por esta razón, Hosmer y Lemeshow (1989) propusieron una estadística alternativa Q_{HL} la cual es obtenida calculandose la estadística Chi cuadrado de Pearson a partir de una tabla $g \times 2$ de frecuencias observadas y predichas descrita a continuación.

- Inicialmente, las n observaciones son ordenadas en orden creciente de probabilidades $\theta\,(x)$ predecidas por el modelo.
- Tales observaciones son, divididas en g grupos ($g = 10$ por ejemplo). En el primer grupo quedan las n_1 observaciones con probabilidades estimadas $< 0,1$ y en el ultimo, las n_g observaciones con probabilidades $\geq 0,9$

$$Q_{HL} = \sum_{i=1}^{g} \frac{\left(o_i - n_i \bar{\theta}\,(x_i)\right)^2}{n_i \bar{\theta}\,(x_i) \left(1 - \bar{\theta}\,(x_i)\right)} \sim \chi^2_{(g-2)}$$

n_i = frecuencia de observaciones en el grupo i.
o_i = frecuencia de respuesta $Y = 1$ en el grupo i.
$\bar{\theta}(x_i)$ = probabilidad media estimada de respuesta $Y = 1$ en el grupo i.

3.6. Análisis de residuos

Como limitaciones de las estadisticas Q_P y $Q_L \Rightarrow$ podemos indicar que un unico valor es utilizado para resumir una cantidad considerable de información. Por esta razón, Pregibon(1981) extendio los métodos de diagnóstico de regresión lineal para la regresión logistica, haciendo uso de los componentes individuales de las estadisticas Q_P y Q_L. En este caso propone usar

$$c_i = \frac{n_{i1} - (n_{i+})\,\hat{\theta}(x_i)}{\sqrt{(n_{i+})\,\hat{\theta}(x_i)\left(1 - \hat{\theta}(x_i)\right)}},\ i = 1, \ldots, s$$

donde los componentes c_i son denominados residuos de Pearson, pues la suma de ellos al cuadrado resulta en Q_P, esto es:

$$Q_P = \sum_{i=1}^{s} (c_i)^2$$

Analogamente, los componentes d_i son denominados residuos deviance o de desvios, pues la suma de ellos al cuadrado resulta en Q_L, esto es:

$$Q_L = \sum_{i=1}^{s} (d_i)^2$$

donde

$$d_i = \pm \left[2n_{i1} \ln\left(\frac{n_{i1}}{e_{i1}}\right) + 2\,(n_{i+} - n_{i1}) \ln\left(\frac{n_{i+} - n_{i1}}{n_{i+} - e_{i1}}\right) \right]^{1/2}$$

com $e_{i1} = (n_{i+})\,\hat{\theta}(x_i)$ para $i = 1, \ldots, s$. y signo de $d_i \Rightarrow$ es definido a partir de las diferencias $(n_{i1} - e_{i1})$.

En este caso la distribución aproximada de los residuos c_i y $d_i es \sim N(0, 1)$ y entonceslos residuos excediendo $\pm 2,5$ pueden indicar:

- posible falta de ajuste del modelo
- presencia de outliers
- patrones sistematicos de variación

Asumiendo que los residuos d_i siguen distribución aproximadamente Normal $\Rightarrow$, es posible construir un gráfico normal Q-Q plot con el envelope simulado (Davison e Gigli, 1989) y entonces si los residuos estuviesen dentro del envelope simulado $\Rightarrow$ entonces tendremos evidencias favorables al modelo ajustado.

Ben y Yohai (2004) argumentan que para algunos modelos, la distribución de los residuos puede estar lejana de la normalidad. Por esta razón propusieron una estimativa de la distribución de los residuos d_i, de modo que en el Q-Q plot tales residuos son graficados versus los cuantiles de la distribución estimada.

3.7. Métodos auxiliares para ajuste del modelo: Poder predictivo

Para evaluar el poder predictivo del modelo es necesario establecer un punto de corte ($0 < pc < 1$), tal que:

a) Probabilidades predecidas por el modelo $\geq pc \Rightarrow Y = 1$

b) Probabilidades predecidas por el modelo $< pc \Rightarrow Y = 0$

Respuesta Observada	Respuesta predecida por el modelo		
	Y=1(+)	Y=0(-)	Totales
Y=1(+)	a	b	(a+b)
Y=0(-)	c	d	(c+d)
Totales	(a+c)	(b+d)	n

Cuadro 3 *Observeado vs Predicho*

Usando el cuadro 2, podemos encontrar las siguientes medidas

- Sensibilidad = $\frac{a}{a+b}$ = tasa de verdaderos +
- Especificidad = $\frac{d}{c+d}$ = tasa de verdaderos −
- Valor predecido = $\frac{a+d}{n}$ = proporcion general de aciertos

Para diversos puntos de corte ⇒ podemos construir la llamada Curva ROC considerando

- Pares (x,y) = (1 - especificidad,sensitividad).
- Modelo con discriminacion perfecta ⇒ (x,y)=(0,1).
- Puntos de corte proximos al canto superior izquierdo, produciran los mayores porcentajes de aciertos (V+ y V-).
- Cuando mas proxima de 1 sea el area bajo la curva, mejor el poder de prediccion del modelo.

4. Inferencia bayesiana en Regresión binaria

En este capítulo abordamos la inferencia bayesiana de regresión binaria así como presentamos diferentes modelos basados en diferentes enlaces simétricos y asimétricos en código BUGS. También estudiamos las condiciones de existencia de la posteriori para los parámetros de regresión β cuando se consideran diferentes prioris impropias. Adicionalmente se presentan diferentes indicadores para la comparación de modelos en la perspectiva bayesiana.

4.1. Inferencia Bayesiana

Considerando la distribución Bernoulli para la variable respuesta, podemos escribir una forma general de la función de verosimilitud del model de regresión binaria dada por

$$L(\beta, \theta_1, \gamma) \prod_{i=1}^{n} \left[F_\theta(m(x_i^T\beta, \gamma))\right]^{y_i} \left[1 - F_\theta(m(x_i^T\beta, \gamma))\right]^{1-y_i}$$

donde $F_\theta(m(x_i^T\beta, \gamma))$ es la función de distribución acumulada de una distribución que puede ser indexada por el parámetro θ que no necesariamente es unidimensional y $m(.)$ es una función continua del predictor lineal $x_i^T\beta$ que también incluye la función identidad con γ como un parámetro de forma. Los enlaces logit, probit, cloglog, scobit y power logit comentados en el capítulo 2 consideran esta función de verosimilitud. Para otros enlaces como el skew probit y el skew logit también presentados en el capítulo 2, se debe considerar otras versiones de la función de verosimilitud llamadas de versiones aumentadas que son discutidas en las referencias especificas de estos modelos.

En la Inferencia bayesiana, a diferencia de la inferencia clásica, los parámetros de interés β, θ, γ se asumen como variables aleatorias y así se establecen diferentes distribuciones de probabilidad a priori que reflejan nuestro conocimiento previo de su conducta. Combinando la función de verosimilitud y las distribuciones a priori podemos obtener la distribución posterior de los parámetros de interés.

Estos parámetros tienen significados diferentes. Los parámetros θ y γ están asociados con el enlace, y el parámetro β corresponde a los datos observados y no depende del modelo escogido.

En nuestro trabajo, consideramos prioris vagas (prioris propias con distribuciones conocidas con varianza grande)

Asumimos independencia entre las prioris, esto es:

$$f(\beta, \theta, \gamma) = f(\beta)f(\theta)f(\gamma)$$

Usamos prioris para β comunes en la literatura incluyendo prioris normales. Especificaciones para $f(\theta)$ y $f(\gamma)$ dependen de la elección particular del modelo considerando un intervalo de variación. En muchas situaciones esos intervalos son determinados de acuerdo a la literatura.

Una vez especificada la distribución a priori tenemos que la posteriori viene dada por:

$$f(\beta, \theta, \gamma \mid \boldsymbol{y}) = L(\beta)f(\beta)f(\theta)f(\gamma).$$

La inferencia (Bayesiana) para los modelos de regresión binaria, especialmente para los modelos citados antes, puede ser facilitada por la simulacion MCMC implementada en el programa WinBUGS. Usando una programación mínima es posible implementar todos los modelos pre-

sentados. Para una revisión de los métodos MCMC revisar Chen, Shao y Ibrahim, J. G (2000) y Gamerman y Lopes (2006). Los códigos para estos modelos pueden ser generados usando el programa BRMUW como es descrito en Bazán y Bayes (2010).

En la siguiente sección será presentada la estimación bayesiana del modelo de regresión binaria considerando como función de enlace la función logit y daremos la sintaxis generada en BRMUW.

4.2. Inferencia Bayesiana en el Modelo de Regresión Logística

Considere un modelo de regresión binaria

$$Y_i \sim Bernoulli(p_i)$$
$$p_i = F\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$$

donde

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

entonces tenemos que la función de verosimilitud será dada por

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{y_i} \left(\frac{1}{1+e^{x_i^T\beta}}\right)^{1-y_i} \tag{0.1}$$

y si consideramos que la priori β sigue una distribución normal multivariada con vector de medias $\mathbf{0}$ y matriz de covarianza Σ,

$$f(\beta) = \left(\frac{1}{2\pi}\right)^{n/2} e^{\beta\Sigma^{-1}\beta}$$

Luego la distribución a posteriori viene dada por

$$f(\beta \mid \boldsymbol{y}) \propto \frac{e^{\sum_{i=1}^n y_i x_i^T\beta}}{\prod_{i=1}^n 1+e^{x_i^T\beta}} e^{\beta\Sigma^{-1}\beta}$$

Como podemos observar en esta caso la distribución a posteriori no presenta una forma conocida por lo que se hace necesario utilizar métodos MCMC como los implementados en los programas WinBUGS y OpenBUGS por lo que usando una programación mínima es posible implementar estos métodos. En este caso, la programación en WinBUGS u OPENBUGS para $N = 200$ sujetos y dos variables predictoras ($k = 3$ coeficientes de regresión) queda dado por:

```
model
  {
   for(i in 1:N) {
         y[i] ~ dbern(p[i])
         logit(p[i]) <- m[i]
                m[i] <- beta[1]+beta[2]*x1[i]+beta[3]*x2[i]
                 }
  for (j  in 1:k) {beta[j] ~ dnorm(0.0,1.0E-3)
                   }
```

```
  }

Inits
list(beta=c(0,0,0,0,0))

Data
list(N=200,k=3)
```

Para una revisión de las sintaxis de otros modelos de regresión binaria tradicionales como el probit y cloglog, sugerimos el libro de Congdon (2005). Las diferentes sintaxis de los modelos de regresión binaria implementados en BRMUW pueden ser revisadas en el propio programa. En algunos casos se requiere una revisión adicional para la especificación de prioris en los modelos asimétricos que incluyen un parámetro diferente del coeficiente de regresión. En este caso, para cada modelo implementado la sintaxis del modelo indica las referencias pertinentes en la especificación de prioris del modelo.

4.3. Condiciones de existencia de la posteriori para los parámetros de regresión

En el caso que consideremos una priori no informativa para β del tipo uniforme esto es

$$f(\beta) \propto 1$$

debemos comprobar la existencia de la distribución a posteriori. Estos aspectos han sido abordados por Chen y Shao (2000) y son reseñados a continuación.
Consideraremos el caso en que la función de enlace $F(.)$ no está indexada por otros parámetros, en este caso la función de verosimilitud solo es una función de β y es dada por

$$L(\beta) = \prod_{i=1}^{n} \left[F(x_i^T\beta)\right]^{y_i} \left[1 - F(x_i^T\beta)\right]^{1-y_i}$$

Asi tenemos que la distribución a posteriori de β es dada por

$$f(\beta \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} \left[F(x_i^T\beta)\right]^{y_i} \left[1 - F(x_i^T\beta)\right]^{1-y_i}$$

Desde que una distribución a posteriori impropia hace que la inferencia bayesiana sea imposible, es importante que estudiemos si la distribución a posteriori $f(\beta \mid \boldsymbol{y})$ es propia. Debe quedar claro que esta distribución será propia solamente si

$$\int_{\Re^k} \prod_{i=1}^{n} \left[F(x_i^T\beta)\right]^{y_i} \left[1 - F(x_i^T\beta)\right]^{1-y_i} < \infty$$

Para obtener condiciones necesarias y suficientes para que la posteriori de β sea propia, sea $z_i = -1$ si $y_i = 0$ y $z_i = 1$ si $y_i = 1$, $\boldsymbol{X}$ es una matriz de diseño $n \times k$ con filas x_i^T y definamos $\boldsymbol{X}^*$ una matriz con filas $z_i x_i^T$.

Teorema 4.1. *Asumiendo que las siguiente condiciones son verdaderas:*

(C1) La matriz de diseño $\boldsymbol{X}$ es de rango completo

(C2) Existe un vestor positivo $a = (a_1, ..., a_n)^T \in \Re^n$*, esto es, que cada componente* $a_i > 0$*, tal que*

$$\boldsymbol{X}^{*T} a = 0$$

(C3) $\int_{-\infty}^{\infty} |u|^k dF(u) > \infty$

entonces

$$\int_{\Re^k} L(\beta) d\beta < \infty$$

El teorema 1 garantiza la existencia de la distribución a posteriori desde que las condiciones (C1)-(C3) sean satisfechas. El siguiente teorema extiende el teorema 2.1 para la existencia de momentos a posteriori.

Teorema 4.2. *Asumiendo que las condiciones (C1) y (C2) son verdaderas. Si* $\int_{-\infty}^{\infty} |u|^{k+p} dF(u) > \infty$ *para algun* $p \geq 0$*, entonces*

$$\int_{\Re^k} \|\beta\|^p L(\beta) d\beta < \infty$$

donde $\|.\|$ *denota la norma euclideana*

Los modelos logístico, probit y cloglog cumplen la condición (C3) del teorema 1 entonces solo se deben verificar las condiciones (C1) y (C2) referente a la matriz de diseño para saber si la posteriori β será propia.

Este tipo de estudio también ha sido realizado para distribuciones que tengan otros parámetros además de β, por ejemplo en Chen y Shao (1999) y recientemente para modelos que consideran enlaces asimétricos basados en la distribución normal asimétrica Bazán, Branco y Bolfarine (2006).

4.4. Comparación de Modelos

Como hemos visto podemos considerar diferentes modelos para un conjunto de datos binarios con solo cambiar la función de enlace, en esta sección revisaremos diferentes critérios para la comparación de modelos que nos ayudaran a decidir que modelo es más apropiado.

Existen una seria de metodologías para comparar modelos alternativos, entre los principales criterios para comparación de modelos en la inferencia bayesiana tenemos: *(*deviance information criterion) (DIC) propuesto por Spiegelhalter et al. (2002), el esperado del criterio de información de Akaike (EAIC) and y el esperado del criterio de información de Schwarz o Bayesiano (EBIC) estos dos últimos propuestos en Carlin and Louis (2000) y Brooks (2002). Estos criterios son basados en media a posteriori del *desvío* $E\Big[D(\mathbf{a}, \mathbf{b}, \lambda, \theta)\Big]$, donde

$$D(\beta, \lambda, \theta) = -2ln(p(\mathbf{y}|\beta, \lambda, \theta) = -2\sum_{i=1}^{n} lnP(Y_{ij} = y_{ij}|\beta, \lambda, \theta),$$

que es una medida de ajuste que puede ser aproximada utilizando la salida de la simulación MCMC de la distribución a posteriori, esta aproximación es dada por

$$Dbar = \frac{1}{G}\sum_{i=1}^{G} D(\beta^g, \lambda^g, \theta^g),$$

donde el índice g indica el $g-$ésimo valor simulado de un total de G simulaciones.
El EAIC, EBIC y DIC pueden ser estimados de la siguiente manera

$$\widehat{EAIC} = Dbar + 2p,$$

$$\widehat{EBIC} = Dbar + plogN,$$

y

$$\widehat{DIC} = Dbar + \widehat{\rho_D} = 2Dbar - Dhat,$$

respectivamente donde p es el número de parámetros en el modelo, N es el total de observaciones y ρ_D es el *número efectivo de parámetros* y es definido como

$$\rho_D = E\Big[D(\beta, \lambda, \theta)\Big] - D\Big[E(\beta, E(\lambda), E(\theta)\Big],$$

donde $D\Big[E(\beta), E(\lambda), E(\theta))\Big]$ es el *desvio de la media a posteriori* obtenido cuando evaluamos la función desvio en la média a posteriori de los parámetros, el cual es estimado por

$$Dhat = D\Big(\frac{1}{G}\sum_{i=1}^{G}\beta^g, \frac{1}{G}\sum_{i=1}^{G}\lambda^g, \frac{1}{G}\sum_{i=1}^{G}\theta^g\Big)$$

Para comparar dos o más modelos alternativos, el modelo que presente mejor ajuste al conjunto de datos será el modelo que presente el menor valor de DIC, EAIC y EBIC. En el EAIC y EBIC $2p$ y $plogN$ son valores fijos que penalizan a la media a posteriori del *desvío*. Desde que, no existe consenso en la literatura de que criterio sea el mejor, el uso de más de un criterio parece ser apropiado para realizar comparación de modelos.

Dbar y DIC son reportados en WinBUGS directamente cuando se requiere durante el proceso de simulación. $EAIC$ y $EBIC$ pueder ser derivados a partir del valor obtenido de *Dbar* considerando las expresiones presentadas.

Información de como implentar la estimación bayesiana de la regresión binaria usando los enlaces cloglog, probit y logit en WinBUGS u OpenBUGS puede ser vista en el Ejemplo Beetles:logistic,probit and extreme value models del Manual. Sin embargo la regresión binaria bayesiana considerando otros enlaces como los discutidos en Bazán, Bolfarine y Branco (2006 y 2010) actualmente no se encuentran disponibles comercialmente pero estos enlaces pueden ser ejecutados usando inferencia bayesiana. Especificamente usando el paquete WinBUGS y generadores de códigos como BRMUW (Bazán y Bayes, 2010).

En resumen, los modelos de regresión binaria implementados en BRMUW clasificados según sus enlaces son

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

5. Aplicaciones

5.1. Aplicación 1 Inferencia Classica para datos del Challenger

El 28 de enero de 1986 el transbordador *Challenger*, tuvo una falla catastrófica debido la falla de un de las juntas especiales selladas con anillos de goma (O-rings), diseñadas para prevenir el escape de combustible muy caliente producido durante la ignición. En total hay seis de tales anillos en cada nave (tres en cada uno de los dos cohetes).

Para los 23 vuelos espaciales antes del desastre de la mision del Challenger en 1986, la tabla 4.10 (Agresti, 2007), muestra la temperatura (F) en el momento del vuelo y si alguna pieza sufrio desastre térmico o no.

Table 4.10. Data for Problem 4.5 on Space Shuttle

Ft	Temperature	TD	Ft	Temperature	TD
1	66	0	13	67	0
2	70	1	14	53	1
3	69	0	15	67	0
4	68	0	16	75	0
5	67	0	17	70	0
6	72	0	18	81	0
7	73	0	19	76	0
8	70	0	20	79	0
9	57	1	21	75	1
10	63	1	22	76	0
11	70	1	23	58	1
12	78	0			

Note: Ft = flight no., TD = thermal distress (1 = yes, 0 = no).

Source: Data based on Table 1 in S. R. Dalal, E. B. Fowlkes and B. Hoadley, *J. Am. Statist. Assoc.*, **84**: 945–957, 1989. Reprinted with the permission of the American Statistical Association.

Figura 4 *Datos del Challenger (Ejercicio 4.5 en Agresti, 2007)*

Entonces tenemos una clásica respuesta binaria, hubo o no hubo daño en los anillos, que en este caso puede ser explicada por la temperatura ambiente en el momento del lanzamiento. A continuación veremos como utilizar un modelo de regresión binaria para calcular la probabilidad de falla del transbordador.

Consideremos el siguiente modelo

$$Falla_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F(\eta_i)$$

$$\eta_i = \beta_0 + \beta_1 Temperatura_i, \quad i = 1, 2, ..., 23$$

donde $F(.)$ es la acumulada de una distribución con soporte real y F^{-1} es la función de enlace.

a) Para el problema anterior proponga cuatro modelos usando los enlaces logito, probito, cloglog y cauchit para modelar el efecto de la temperatura en la probabilidad de desastre térmico, usando el enfoque clásico. Presente los modelos ajustados e interprete los coeficientes de regresión estimados.
Primero, se realizara un analisis descriptivo con los datos del Challenger.

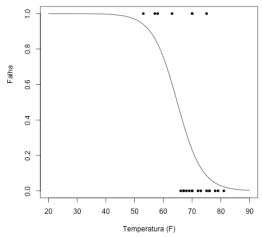

Figura 5 *Falla vs Temperatura. Datos del Challenger*

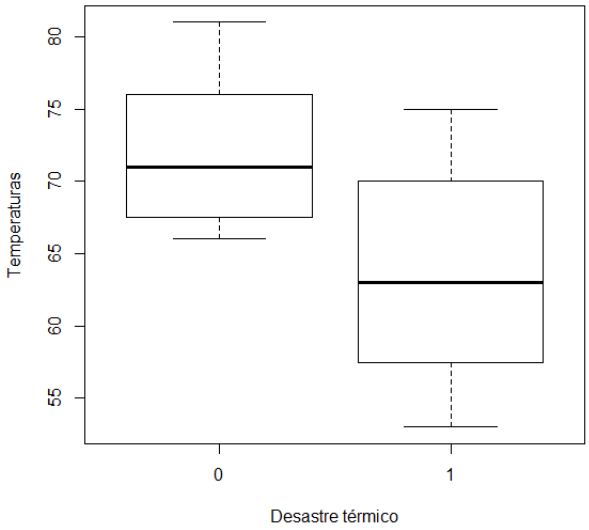

Figura 6 *Box plot de Temperatura por Falla. Datos del Challenger*

Por los gráficos se puede observar que las temperaturas de los lanzamientos anteriores a la explosion estaban entre 65^oF e 80^oF. Se puede apreciar que cuanto mas frio, mayores las chances de que exista algun problema.

Para una evaluacion mas concreta, se modelara el efecto de la temperatura en la probabilidad del desastre termico a partir de cuatro modelos: logito, probito, cloglog e cauchit que son dados, respectivamente, por:

- $\theta(x) = \frac{exp(\eta)}{1+exp(\eta)}$, com $\eta = \beta_0 + x\beta_1$
- $\theta(x) = \Phi(\eta)$
- $\theta(x) = 1 - exp(-exp(\eta))$
- $\theta(x) = \frac{1}{2} + \frac{arctg(\eta)}{\pi}$

Ajustando cada uno de ellos con el comando glm de R, obtenemos:

- Logito

```
> summary(ajuste1)

Call:
glm(formula = y ~ temp, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
temp         -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

Figura 7 *Modelo Logito. Datos del Challenger*

Se observa que a través del p-valor los coeficientes son significativos a un nivel $\alpha = 0,05$ de significancia. De esta forma, el modelo ajustado por el enlace logito es dado por:

$$\theta(Temperatura) = \frac{exp(15,0429 - 0,2322Temperatura)}{1 + exp(15,0429 - 0,2322Temperatura)}$$

Se percibe que cuanto mas alta la temperatura, menor la probabilidad de desastre térmico, debido al valor del coeficiente $\beta_1 = -0,2322$.

- Probito

Nuevamente los coeficientes son significativos. El modelo ajustado es:

$$\theta(Temperatura) = \Phi(8,7749 - 0,1351Temperatura)$$

```
Call:
glm(formula = y ~ temp, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0831  -0.7930  -0.3747   0.4413   2.2081

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.77490    3.87231   2.266   0.0234 *
temp        -0.13510    0.05646  -2.393   0.0167 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.378  on 21  degrees of freedom
AIC: 24.378

Number of Fisher Scoring iterations: 6
```

Figura 8 *Modelo Probito. Datos del Challenger*

```
Call:
glm(formula = y ~ temp, family = binomial(link = "cloglog"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0358  -0.7361  -0.3891   0.1729   2.2050

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 12.30215    5.19483   2.368   0.0179 *
temp        -0.19583    0.07809  -2.508   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 19.531  on 21  degrees of freedom
AIC: 23.531

Number of Fisher Scoring iterations: 8
```

Figura 9 *Modelo Cloglog. Datos del Challenger*

- Cloglog

  El modelo ajustado para el enlace cloglog es dado por:

$$\theta(Temperatura) = 1 - exp(-exp(12,3021 - 0,1958Temperatura))$$

- Cauchit

  El modelo ajustado para el enlace Cauchit:

```
Call:
glm(formula = y ~ temp, family = binomial(link = "cauchit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9020  -0.5899  -0.3944   0.4450   2.2386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  23.1933    18.4607   1.256    0.209
temp         -0.3601     0.2779  -1.296    0.195

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 19.637  on 21  degrees of freedom
AIC: 23.637

Number of Fisher Scoring iterations: 12
```

Figura 10 *Modelo cauchit. Datos del Challenger*

$$\theta(Temperatura) = \frac{1}{2} + \frac{arctg(23,1933 - 0,3601 Temperatura)}{\pi}$$

Se observa que en todos los modelos ajustados los coeficientes fueron significativos para $\alpha = 0,05$, y en todos los casos el coeficiente β_1 fue negativo, dando evidencia que cuanto mayor es la temperatura, menor es el riesgo de accidente.

b) **Interprete el efecto de la temperatura en las chances de desastre térmico. Pista: Derive la funcion de probabilidad de cada modelo em funcion de la temperatura.**

Podemos ver el efecto de la temperatura a traves de los gráficos de pontos y box-plot presentados anteriormente, en los que claramente se observa una influencia de la temperatura en la chance de desastre.

Otra manera, la sugerida, es derivar los modelos respecto a la variable temperatura.

- Derivada del modelo logito respecto a la variable temperatura.

$$\frac{d\theta(Temperatura)}{dTemperatura} = \frac{(exp(\eta)\beta_1)(1 + exp(\eta)) - exp(\eta)(1 + exp(\eta)\beta_1)}{(1 + exp(\eta))^2}$$

$$= \beta_1 \frac{exp(\eta)}{(1 + exp(\eta))^2}$$

- Derivada del modelo probito respecto a la variable temperatura.

$$\frac{d\theta(Temperatura)}{dTemperatura} = \phi(\eta)\beta_1$$

en que ϕ es la densidad de la normal.

- Derivada del modelo cloglog respecto a la variable temperatura.

$$\frac{d\theta(Temperatura)}{dTemperatura} = exp(-exp(\eta))(-exp(\eta))\beta_1$$

$$= \beta_1(-exp(-exp(\eta)\eta))$$

- Derivada del modelo Cauchit respecto a la variable temperatura.

$$\frac{d\theta(Temperatura)}{dTemperatura} = \frac{\beta_1}{\pi(\eta)^2 + 1}$$

(obs: derivadas obtenidas con Wolfram-alpha)

Podemos observar a traves de las derivadas que en todos los casos el coeficiente β_1, que es el termino que acompaña a la variable temperatura, está influenciando el valor de la derivada, esto complementa la sospecha de la gran influencia causada por la temperatura que se tuvo con los analisis gráficos.

c) Pruebe la hipótesis del efecto de la temperatura, usando (i) Test de Wald, (ii) test de razon de verosimilitud, en cada modelo.
Para probar la hipótesis del efecto de la temperatura vamos a testear su coeficiente, esto es:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 = \beta_1 \neq 0$$

Un primer test puede ser realizado usando el test de Wald, que tiene la estadística dada por:

$$W = \frac{\hat{\beta}_1}{\hat{Var}(\beta_1)} \sim N(0, 1)$$

Para el modelo logito $W = \frac{-0{,}2322}{0{,}1082} = -2{,}146$ y con valor-p = 0,0159

Otra alternativa es usar el test de razon de verosimilitud (TRV), con estatística:

$$TRV = 2(l(\theta_c) - l(\theta_s)) \sim \chi^2_q$$

en que $l(\theta_c)$ es la logverosimilitud con la covariable y $l(\theta_s)$ sin la covariable. Nuevamente para el modelo logito $TRV = 15{,}904$ con valor-p $= 6{,}66 \times 10^(-5)$.

Siendo asi, en ambos casos, tomando um nivel de significancia $\alpha = 0,05$, rechazamos la hipótesis nula, esto es, existen evidencias que el efecto de la temperatura es significativo.

De manera analoga a lo realizado para el modelo logito, calculamos esas estadísticas para los otros modelos, obteniendo el cuadro siguiente:

En ellas podemos observar que solamente no rechazariamos la hipótesis nula para el modelo Cauchit usando el test de Wald, en los restantes tenemos evidencias para rechazar H_0, es decir, admitir que el efecto de la temperatura es significativo a nivel $\alpha = 0,05$.

Modelo	Wald (valor-p)	TRV (valor-p)
Probito	2.3933 (0.008)	15.778 (7.12exp(-0.5))
cloglog	-2.5074 (0.006)	17.472 (2.92exp(-0.5))
Cauchit	-1.2958 (0.097)	17.26 (3.26exp(-0.5))

Cuadro 4 *Comparación de modelos. Datos del Challenger*

d) Compare las diferentes propuestas de enlace usando por ejemplo AIC, curva ROC y curvas previstas. Cual es el mejor modelo propuesto para los datos?

Con la finalidad de comparar las diferentes propuestas de enlace observemos primero los criterios AIC para los modelos ajustados.

Modelo	AIC
Logito	24.315
Probito	24.378
Cloglog	23.531
Cauchit	23.637

Cuadro 5 *Comparación de modelos. Datos del Challenger*

Con base en la tabla anterior, el mejor modelo es aquel que tiene menor AIC, en este caso el Cloglog. Pero los valores no se diferencian mucho, siendo asi podemos complementar la eleccion del mejor modelo visualizando el gráficos de bandas.

(obs: Para efectuar esos gráficos utilizaremos los comandos disponibles en Paula (2010))

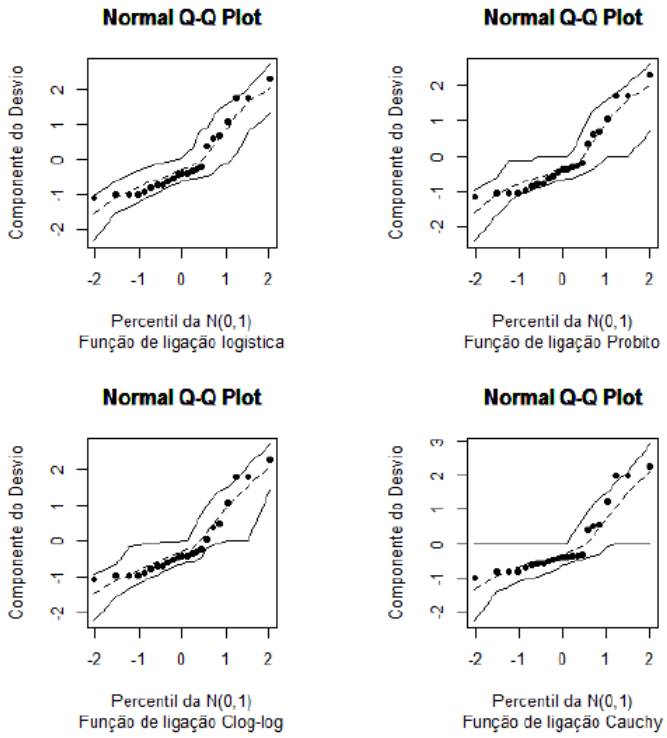

Figura 11 *Gráfico de Bandas para diferentes modelos. Datos del Challenger*

Los resultados muestran que todos los enlaces presentan ajuste adecuado, pero por el criterio AIC escogeríamos como mejor modelo el Cloglog. Como tarea dejamos calcular las correspondientes medidas de valor predictivo y curvas ROC.

e) **Usando el modelo escogido, estime la probabilidad de desastre térmico para 31 F, la temperatura en el momento del vuelo del Challenger. Comente sus resultados.**

Queremos estimar la probabilidad de desastre para cuando la temperatura estuviese en 31^oF, la temperatura en el momento del accidente.

Realizaremos esto con el modelo escogido, el cloglog.

Vimos que el modelo ajustado para este caso es dado por

$$\theta(Temperatura) = 1 - exp(-exp(12,3021 - 0,1958Temperatura))$$

Para $Temperatura = 31$ tenemos:

$$\theta(31) = 1 - exp(-exp(12,3021 - 0,1958(31)) = 1$$

Aqui vemos que el accidente era una certeza, pues la probabilidad de ocurrir un desastre era 1.

Como quedamos en duda del modelo logito, vamos a estimar la probabilidad para este modelo tambien.

$$\theta(31) = \frac{exp(15,0429 - 0,2322(31))}{1 + exp(15,0429 - 0,2322(31))} = 0{,}999$$

Confirmando el resultado que el accidente era cierto y que el lanzamiento no deberia haber ocurrido.

f) Usando el modelo escogido, en que temperatura la probabilidad estimada es igual 0,50? en aquella temperatura, de una aproximacion lineal para la alteracion en la estimacion de la probabilidad por aumento de un grado en la temperatura.

Queremos descubrir a cual temperatura la probabilidad estimada seria de $0,50$.

Para eso tenemos que despejar la variable $Temperatura$ en el modelo cloglog, obteniendo:

$$T = \frac{1}{\hat{\beta}_1}\Big(log(-log(1 - \theta(Temperatura))) - \hat{\beta}_0\Big)$$

Con los valores estimados, tenemos:

$$T = \frac{1}{-0{,}1958}\Big(log(-log(1 - 0{,}5))) - 12{,}3021\Big) = 64{,}69$$

Asi, tendriamos 0,50 de probabilidad de ocurrir accidente cuando la temperatura estuviese em $64{,}69^{o}F$, para el modelo cloglog.

Para la aproximacion lineal para la estimacion de la probabilidad por un aumento de un grado en la temperatura es dada por:

$$h(x) = \beta(\theta(x))(1 - \theta(x))$$

O sea,

$$h(Temperatura) = \hat{\beta}_1(1-exp(-exp(\beta_0+\beta_1(Temperatura)))(exp(-exp(\beta_0+\beta_1(Temperatura))))$$

Substituyendo por los valores encontrados anteriormente,

$$h(64{,}69) = -0{,}195\,(1 - \exp\{-\exp\{12{,}30 - 0{,}195(64{,}69)\}\})\,(\exp\{-\exp\{12{,}30 - 0{,}195(64{,}69)\}\})$$

$$h(64{,}69) = -0{,}049,$$

Por tanto, para el incremento de un grado en la temperatura, tenemos um decrecimiento de 0.049 en la probabilidad de una pieza sufrir un desaste térmico.

5.2. Aplicación 2 Inferencia Bayesiana para datos del Challenger

a) Para los datos del problema anterior. Ajuste modelos logístico, probito y cloglog asi como el modelo Power logístico y Power Probito bajo el enfoque bayesiano usando MCMC en WinBUGS. Considere 14000 iterações, Burnin=4000, y thin=10. Presente los modelos ajustados e interprete los coeficientes de regresión estimados.

Se hicieron los ajustes de los modelos Logístico, Probito y Complemento log log, asi como los ajustes de los modelos Power logístico y Power Probito bajoel enfoque bayesiano usando MCMC utilizando OpenBUGS, conforme sigue en el siguiente cuadro:

Cuadro 6 *Ajustes de los Modelos Presentados*

Modelo Probito								
	mean	**sd**	**MC error**	**val2.5pc**	**median**	**val97.5pc**	**start**	**sample**
beta[1]	7.206	2.321	0.405	1.618	7.666	10.31	4000	1000
beta[2]	-0.1125	0.03334	0.005815	-0.1559	-0.1197	-0.03387	4000	1000
DIC								
			Dbar	**Dhat**	**DIC**	**pD**		
		TD	20.8	20.53	21.07	0.2714		
		total	20.8	20.53	21.07	0.2714		

Modelo Logístico								
	mean	**sd**	**MC error**	**val2.5pc**	**median**	**val97.5pc**	**start**	**sample**
beta[1]	18.02	8.41	0.528	4.618	16.9	36.86	4000	1000
beta[2]	-0.2765	0.1237	0.007788	-0.5537	-0.2609	-0.07563	4000	1000
DIC								
			Dbar	**Dhat**	**DIC**	**pD**		
		TD	21.89	20.33	23.44	1.555		
		total	21.89	20.33	23.44	1.555		

Modelo Complemento Log Log								
	mean	**sd**	**MC error**	**val2.5pc**	**median**	**val97.5pc**	**start**	**sample**
beta[1]	11.89	2.968	0.5139	7.047	11.48	17.82	4000	1010
beta[2]	-0.1906	0.04414	0.007637	-0.2781	-0.1849	-0.1227	4000	1010
DIC								
			Dbar	**Dhat**	**DIC**	**pD**		
		TD	20.33	19.71	20.95	0.6199		
		total	20.33	19.71	20.95	0.6199		

Modelo Power Logístico								
	mean	**sd**	**MC error**	**val2.5pc**	**median**	**val97.5pc**	**start**	**sample**
beta[1]	32.04	17.5	1.252	5.132	29.65	71.95	4000	1000
beta[2]	-0.554	0.2944	0.02225	-1.212	-0.5221	-0.1037	4000	1000
lambda	0.449	0.4979	0.06632	0.06543	0.2703	2.147	4000	1000
DIC								
			Dbar	**Dhat**	**DIC**	**pD**		
		TD	20.85	20.29	21.41	0.5593		
		total	20.85	20.29	21.41	0.5593		

Modelo Power Probito								
	mean	**sd**	**MC error**	**val2.5pc**	**median**	**val97.5pc**	**start**	**sample**
beta[1]	13.7	8.042	0.5722	0.3734	12.99	32.19	4000	1000
beta[2]	-0.2366	0.1255	0.008637	-0.5269	-0.2224	-0.03989	4000	1000
lambda	0.4664	0.5043	0.04326	0.04557	0.2737	1.889	4000	1000
DIC								
			Dbar	**Dhat**	**DIC**	**pD**		
		TD	23.24	21.81	24.66	1.424		
		total	23.24	21.81	24.66	1.424		

Considerando estos ajustes, observamos que todas las estimaciones de los modelos son significativas, pues el valor cero no pertenece al intervalo de credibilidad. Ademas, las estimaciones de β_2 son todas negativas, eso implica que con un aumento de temperatura hay una diminucion de la probabilidad de una pieza sufrir el desastre térmico.

b) Compare las diferentes propuestas de enlace usando por ejemplo el DIC y las curvas previstas. Cual es el mejor modelo propuesto para los datos?

Para el análisis del mejor modelo vamos a verificar los valores de DIC (Deviance Information Criterion), el cual segun Spiegelhalter et al (2002), utiliza el criterio de bondad de ajuste y nivel de complejidad de este para seleccionar el mejor modelo entre los analizados. Asi, segun este criterio de comparacion de modelos bayesianos, el mejor modelo de los analizados es el Complemento Log Log, pues tiene el menor DIC. Sin embargo, como hemos considerado una muestra MCMC pequeña para la simulación podemos decir que la convergencia necesita ser mejorada. pues observando las figuras abajo podemo observar que la distribución no es unimodal aún ni simétrica pero observando los quantiles de interacción observamos que las estimaciones caen dentro de las bandas de ajuste, aunque puede recomendarse hacer un burnin (descartar un número mayor que iteraciones iniciales).

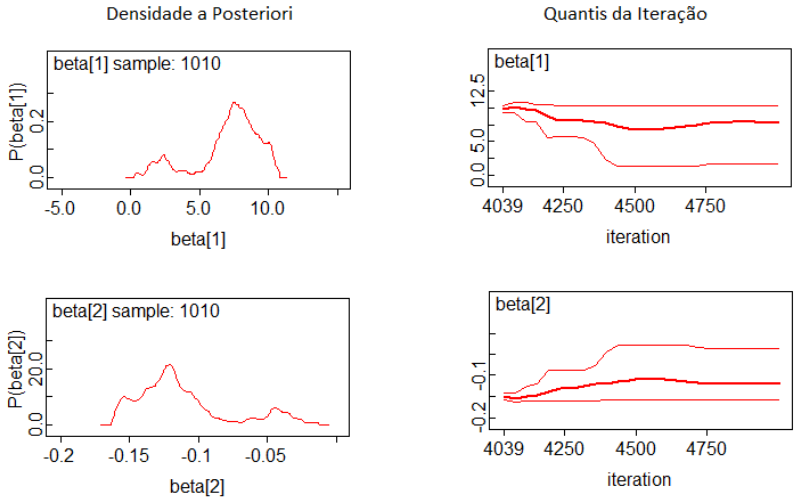

Figura 12 *Evaluación de convergencia. Modelo probito bayesiano*

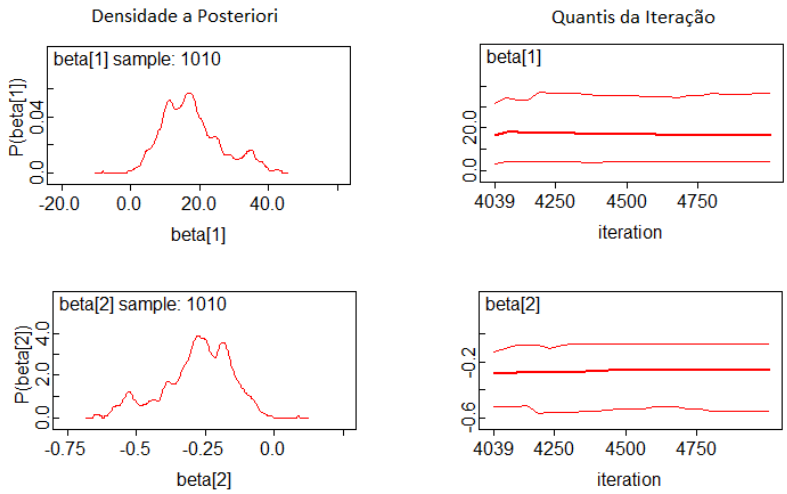

Figura 13 *Evaluación de convergencia. Modelo logístico bayesiano*

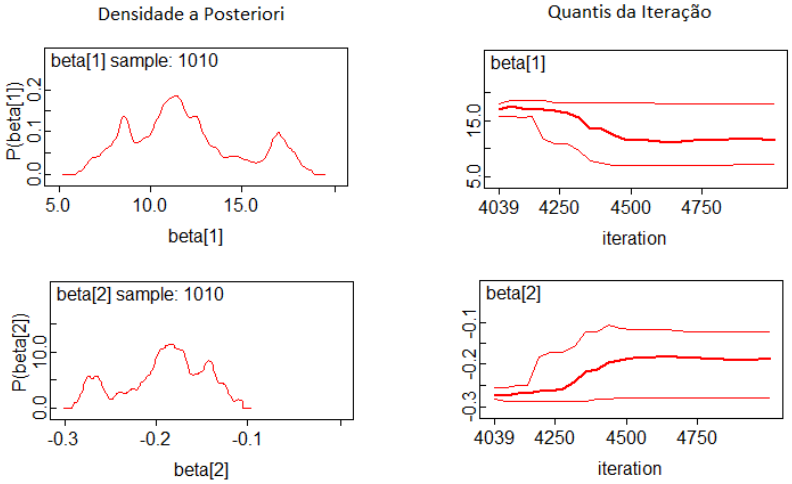

Figura 14 *Evaluación de convergencia. Modelo complemento loglog bayesiano*

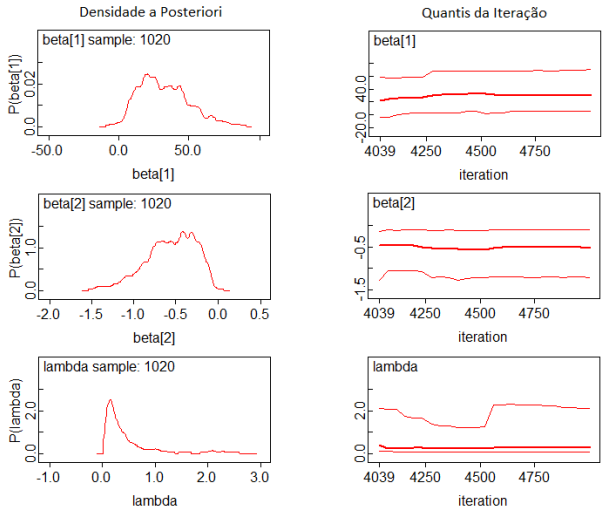

Figura 15 *Evaluación de convergencia. Modelo power logístico bayesiano*

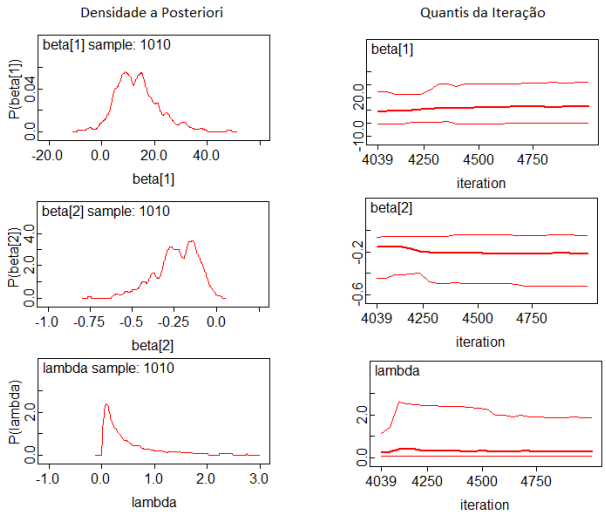

Figura 16 *Evaluación de convergencia. Modelo Power Probito Bayesiano*

c) Usando el modelo escogido, estime la probabilidad del desastre térmico para 31 F, la temperatura en el momento del vuelo del Challenger. Comente sus resultados.

Usando este modelo, estimaremos la probabilidad del desastre térmico para 31 F, la temperatura en el momento del vuelo del Challenger por:

$$\theta(31) = 1 - exp(-exp(12,3021 - 0,1958(31)) = 1$$

Aqui vemos, de acuerdo al ejercicio anterior, que la probabilidad de una pieza sufrer el desastre térmico era una certeza, pues tenia probabilidad 1 de ocurrir un desastre.

d) Usando el modelo escogido, a que temperatura, la probabilidad estimada es igual a 0,50? en aquella temperatura, brinde una aproximacion lineal para la alteracion en la estimacion de la probabilidad por aumento de un grado en la temperatura.

La temperatura para la cual la probabilidad estimada es igual 0,50 es:

$$\begin{aligned} 1 - exp(-exp(12,3021 - 0,1958t) &= 0,5 \\ exp(-exp(11,89 - 0,1906 * t)) &= 0,5 \\ exp(11,89 - 0,1906 * t) &= -log(0,5) \\ t = (11,89 - log(-log(0,5)))/0,1906 & \\ t = 64,30489 & \end{aligned}$$

Es decir, la temperatura para la cual la probabilidad de una pieza sufrer el desastre térmico es de $0,5$, fue de $64,30^oF$.

e) Presente una tabla comparativa de las estimaciones de los coeficientes de regresión bajo el enfoque bayesiano y clásico e incluya los criterios de comparacion de modelos AIC y DIC. Las estimaciones coinciden?, los criterios de comparacion de modelos coinciden al escoger el mejor modelo?. Comente sus resultados.

Presentamos por fin, una tabla comparativa de las estimaciones de los coeficientes de regresión bajo el enfoque bayesiano y clásico incluyendo los criterios de comparacion de modelos AIC y DIC.

Cuadro 7 *Comparacion de los Modelos Clásicos y Bayesianos*

	Coeficientes	**estimaciones**		**Critérios de Comparação**	
		Clássico	**Bayesiano**	**AIC**	**DIC**
ModeloProbito	Intercepto	8,7749	7.206	24,3780	21.07
	Temperatura	-0,1351	-0,1125		
Modelo Logistico	Intercepto	15,0429	18,02	24,3150	23.44
	Temperatura	-0,2322	-0,2765		
Modelo Complemento Log Log	Intercepto	12,3022	11,89	23,5310	20.95
	Temperatura	-0,1958	-0,1906		

En la Tabla 5.2, podemos observar que tanto por el criterio AIC como por el DIC, el mismo modelo Complemento Log Log fue escolgido, mientras que para la comparacion entre los otros modelos los criterios divergian. Las estimaciones pueden ser consideradas próximas para los modelos Probito y Complemento Log Log.

5.3. Aplicación 3: Datos de erradicación de cultivos de coca

En Bazán y Bayes (2010) se presenta una aplicación que ilustra las ventajas de la regresión binaria usando inferencia bayesiana. Como ejemplo consideramos una data que contiene algunas variables de un estudio con agricultores beneficiarios de un programa favorable a la erradicación de cultivos de coca (Bazán y Millones, 2008). La data se denomina `concoca.txt`. Esta data se encuentra disponible en el programa BRMUW dentro de la descarga del programa. Las variables en `concoca.txt` son

sierr	si se muestra favorable a erradicar el cultivo de coca
permedyc	índice de percepción de que el cultivo de coca produce daño al medio ambiente
partco	índice de participación comunal
concoca	índice acerca de si consume coca
pobrez	niveles de pobreza

El archivo de datos tiene la siguiente estructura

sierr	permedyc	partco	concoca	pobrez
1	2	2	1	2
0	0	6	1	2
⋮	⋮	⋮	⋮	⋮
1	2	9	0	3

Como un ejemplo de aplicación consideremos el siguiente modelo

$$sierr_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F(\eta_i)$$

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

donde para F consideraremos diferentes enlaces.

El programa BRMUW genera la sintaxis o códigos necesarios para la estimación bayesiana de varios modelos de regresión binaria, posteriormente estos códigos se usan en el programa WinBUGS (ver Spiegelhalter et al 1996) o OpenBUGS (Spiegelhalter et al 2007), usando diversos métodos MCMC. Para ello solo es necesario contar con un archivo de texto con los datos, generado de cualquier programa estadístico o de Excel. En las columnas usualmente aparece los nombres de las variables en la primera línea y la primera columna deberá contener la variable respuesta.

Utilizando el BRMUW para generar la sintaxis y el WinBUGS para simular de la distribución a posteriori obtuvimos los resultados presentados en las cuadros siguientes.

En este caso, considerando el cuadro 5, observamos que de acuerdo al DIC, todos los enlaces asimétricos con excepción del cloglog presentan mejor ajuste que los enlaces simétricos. El modelo

Cuadro 8 *Comparación de los modelos de regresión binaria para los datos de erradicación de cultivos de coca*

Enlaces	Modelos	Bur-in	Thin	Dbar	DIC
Simétricos	Probito	4000	5	2451.5	2456.8
	Logístico	4000	5	2450.9	2455.8
Asimétricos	Cloglog	4000	5	2451.6	2457
	Scobit	4000	25	2462.1	2441.2
	Power Logit	54000	100	2458.5	1794.1
	Skew logit	4000	25	2458.1	1708.4
	BBB sp	4000	35	2345.2	2252.5
	Standard sp	4000	15	1538.1	1751.7

que mejor se ajusta a los datos es aquel que considera el enlace *skew logit*. Mayores detalles sobre esta aplicación en Bazán y Millones (2008).

En los cuadros 9 y 10 presentamos los estimadores de los parámetros de los modelos logístico y *skew logit*. Aunque los resultados de los coeficentes de regresión son similarea en ambol modelos, la presencia del parámetro de forma $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ que toma valores en el intervalo $[0, 1]$, indica que una daecuada interpretación corresponde al modelo *skew logit*.

De acuerdo a los resultados del modelo *skew logit*, podemos comentar como indicado en Bazán y Bayes (2010), que la variable *percepción de que la producción ilegal de hoja de coca daña al medio ambiente* (*permedyc*) impacta en forma positiva al incrementar la posibilidad de la erradicación de este cultivo. De modo similar, también la variable *participación de los agricultores en actividades de la comunidad* (*partco*) (incluye por ejemplo, la ejecución de tareas en la comunidad, reuniones de la comunidad, participación en proyectos con grupos de trabajo de la comunidad), influyen favorablemente en la erradicación del cultivo de coca. Dos resultados interesantes que hemos encontrado en este análisis es que la *pobreza* (*pobrez*) incrementa la probabilidad de erradicación de los cultivo ilícitos, pero cuanto mayor es el *consumo personal de hojas de coca* (*concoca*) menor es la probabilidad de esta erradicación.

Cuadro 9 *Estimativas del los parámetros del modelo skew logístico para los datos de erradicación de cultivos de coca*

Parámetro	media	Desv. Est.	2.5 %	mediana	97.5 %
β_1	-2.84	1.08	-5.10	-2.70	-0.98
β_2	0.65	0.09	0.52	0.64	0.86
β_3	0.08	0.02	0.05	0.08	0.11
β_4	-0.21	0.05	-0.32	-0.21	-0.11
β_5	0.73	0.20	0.33	0.73	1.15
δ	0.14	0.59	-0.90	0.21	0.91

Cuadro 10 *Estimativas del los parámetros del modelo logístico para los datos de erradicación de cultivos de coca*

Parámetro	media	Desv. Est.	2.5 %	mediana	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

Adicionalmente, llevamos a cabo un análisis de la capacidad predictiva de ambos modelos siendo que el modelo logístico presenta un 64 % de buena clasificación, en contraste con el 95 % que se obtiene cuando se utiliza el modelo *skew logit*.

Capítulo 1

Anexo del Capítulo 4: Aplicaciones

En este anexo presentamos los códigos para las aplicaciones del do capítulo 4 usando principalmente o programa R.

Anexo Aplicación 1

```
###################
rm(list=ls())
y<-c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,
0,0,0,0,1,0,1)
temp<-c(66,70,69,68,67,72,73,70,57,63,70,
78,67,53,67,75,70,81,76,79,75,76,58)

falha = y
temperatura = temp

plot(temperatura,falha)
plot(temp, temp)
curve(temperatura,falha)
boxplot(y,temp)

#Logito
       ajuste1<-glm(y~temp,family=binomial(link="logit"))
    #   ajuste1
       anova(ajuste1,test="Chisq")
       summary(ajuste1)

#probabilidade predita
     plogis(15.0429-0.2322*31)


#Test Wald

w = -0.2322/0.1082
pnorm(w)

#TRV
```

```
trv = 2*(28.267 - 20.315)
valorp = 1 - pchisq(trv, 1)

#Probito
       ajuste2<-glm(y~temp,family=binomial(link="probit"))
      # ajuste2
       anova(ajuste2,test="Chisq")
       summary(ajuste2)

#Wald
w = -0.1351/0.05645
w
pnorm(w)

#TRV
trv = 2*(28.267-20.378)
trv
valorp = 1-pchisq(trv,1)

#Cloglog
       ajuste3<-glm(y~temp,family=binomial(link="cloglog"))
       #ajuste3
       anova(ajuste3,test="Chisq")
       summary(ajuste3)

#wald

w= -0.1958/0.07809
w
pnorm(w)

#TRV
trv = 2*(28.267-19.531)
valorp = 1-pchisq(trv,1)

##Estimativa para 31F

t = 31
x = 1 - exp(-exp(12.3021 - 0.1958*t))

#Cauchit
       ajuste4<-glm(y~temp,family=binomial(link="cauchit"))
       #ajuste4
       anova(ajuste4,test="Chisq")
       summary(ajuste4)
```

```
#Wald
w = -0.3601/0.2779
pnorm(w)
trv = 2*(28.267-19.637)
trv
valorp = 1-pchisq(trv,1)
valorp


#gráfico de curvas previstas de los modelos propuestos

       x<-seq(30,100,0.1)
       m1<-exp(15.0429-0.2322*x)/(1+exp(15.0429-0.2322*x))
       m2<-pnorm(8.77490 -0.13510*x)
       m3<-1-exp(-exp(12.30215 -0.19583 *x))
       m4<-pcauchy( 23.1933 -0.3601*x)


       par(mfrow=c(1,1))
       plot(temp,y,pch=16,ylab="falhas",xlab="Temperature"
,xlim=c(30,100),ylim=c(0,1.05))
       lines(x,m1, lty=2,lwd=2, col=2)
       lines(x,m2, lty=3,lwd=2, col=4)
       lines(x,m3, lty=6,lwd=2, col=3)
       lines(x,m4, lty=1,lwd=2, col=1)
       legend(32,0.8,lty=c(2,3,6,1),col=c(2,4,3,1),lwd=2,
c("logístico","probito","clog-log","cauchy"),bty="n")

       lines(c(64.78424,64.78424),c(0,0.50),lty=3)
       lines(c(0,64.78424),c(0.50,0.50),lty=3)
       legend(65,0.55,c("(64.78424, 0.5)"),bty="n",cex=0.8)



# QQ plot para los modelos
par(mfrow=c(2,2))
ntot<-c(rep(c(1),23))
dados <- data.frame(y,temperatura)
attach(dados)
fit.model <- m1
source("http://www.ime.usp.br/~giapaula/envelr_bino")
title(sub="Funcao de ligacao logistica")
fit.model <- m2
source("http://www.ime.usp.br/~giapaula/envelr_bino")
title(sub="Funcao de ligacao Probito")
```

```
fit.model <- m3
source("http://www.ime.usp.br/~giapaula/envelr_bino")
title(sub="Funcao de ligacao Clog-log")
fit.model <- m4
source("http://www.ime.usp.br/~giapaula/envelr_bino")
title(sub="Funcao de ligacao Cauchy")
```

0.4. Apêndice 2

Códigos utilizados en el Ejercicio 2:
Aproximacion Clásica en el R:

```
# Datos#
#######################################################################
rm(list=ls())
td<-c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,
            0,0,0,0,1,0,1)
temperatura<-c(66,70,69,68,67,72,73,70,57,63,70,
               78,67,53,67,75,70,81,76,79,75,76,58)
#######################################################################
# Ajuste de los Modelos
#Modelo Probito
  ajuste1 <- glm(td~temperatura, family=binomial(link="probit"))
  summary(ajuste1)

# Modelo Logístico
  ajuste2 <- glm(td ~ temperatura, family = binomial(link="logit"))
  summary(ajuste2)

#Modelo Complemento loglog
  ajuste3 <- glm(td~temperatura, family=binomial(link="cloglog"))
  summary(ajuste3)

#######################################################################

# Calculo de probabilidad del desastre térmico para 31°
t = 31
x = 1 - exp(-exp(11.89 - 0.1906*t))

# Calculo da temperatura dada a probabilidade de 0,5
exp(-exp(11.89 - 0.1906*t1)) = 0.5
t1 = (11.89  - log(-log(0.5))) / 0.1906
t1
```

Aproximacion Bayesiana con WingBUGS:

```
 #Logit Model
model {
for(i in 1:n) {
TD[i] ~ dbern(p[i])
logit(p[i]) <- m[i]
m[i] <- beta[1] + beta[2]*Temperature[i]
}

for (j  in 1:k) {
beta[j] ~ dnorm(0.0,1.0E-3)
                }


}

Inits
list(beta=c(0.0,0.0))


Data
list(n=23,k=2)


TD[] Temperature[]
0 66
1 70
0 69
0 68
0 67
0 72
0 73
0 70
1 57
1 63
1 70
0 78
0 67
1 53
0 67
0 75
0 70
0 81
0 76
0 79
```

```
1 75
0 76
1 58
END

#Probit Model

model {
for(i in 1:n) {
TD[i] ~ dbern(p[i])
p[i] <- phi(m[i])
m[i] <- beta[1] + beta[2]*Temperature[i]
}

for (j  in 1:k) {
beta[j] ~ dnorm(0.0,1.0E-3)
}

}

#Cloglog Model

model {
for(i in 1:n) {
TD[i] ~ dbern(p[i])
cloglog(p[i]) <- m[i]
m[i] <- beta[1] + beta[2]*Temperature[i]
}

for (j  in 1:k) {
beta[j] ~ dnorm(0.0,1.0E-3)
                }


}


#Power Logit Model

model {
for(i in 1:n) {
TD[i] ~ dbern(p[i])
logit(pl[i] ) <-m[i]
p[i]<-pow(pl[i],lambda)
m[i] <- beta[1] + beta[2]*Temperature[i]
}
```

```
for (j  in 1:k) {
beta[j] ~ dnorm(0.0,1.0E-3)
}
lambda ~dgamma(1,1)
}

Inits
list(beta=c(0.0,0.0),lambda=0.5)


##Power Probito

model{
for(i in 1:n){
TD[i] ~ dbern(p[i])
pl[i] <- phi(m[i])
m[i] <- beta[1] + beta[2]*Temperature[i]
p[i] <- pow(pl[i],lambda)
}

for (j in 1:k){
beta[j] ~ dnorm(0.0,1.0E-3)
  }
         lambda ~dgamma(1,1)

}
Inits
list(beta=c(0.0,0.0),lambda=0.5)
```

Bibliografía

[1] Agresti, A. (2007). An Introduction to Categorical Data Analysis. Second Edition.

[2] Albert, J. H. (2009). *Bayesian Computation with R.* Springer Verlag

[3] Basu, S. and Mukhopadhyay, S. (2000). Binary response regression with normal scale mixtures links, in *Generalized Linear Models: A Bayesian Perspective*,eds. D.K. Dey, S.K. Ghosh, and B.K: Mallick, New York: Marcel Dekker.

[4] Bazán, J. L., Bolfarine, H. y Branco, M. D. (2006) A generalized skew probit class link for binary regression. *Technical report* (RT-MAE-2006-05). Department of Statistics. University of São Paulo.

[5] Bazán, J. L. , Bolfarine, H. y Branco, D. M. (2010) A framework for skew-probit links in Binary regression (aceptado para publicacion *Communications in Statistics - Theory and Methods*)

[6] Bazán, J. L., Millones, O. (2008). A classification of binary asymmetric regression models: The use of BRMUW in an application to the decision to erradicate illegal crops of coca leaf. *Simposio Nacional de Probabilidade e Estatistica. SINAPE*, São Pedro, Julio 2008.

[7] Bazán, J. L., Millones, O (2008). 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. *Economía 29(62)*, 17-32. PUCP.

[8] Bazán, J., Bayes, C. (2010). Inferencia Bayesiana en Modelos de Regresión Binaria usando BRMUW. Reporte de Investigación. Serie B. Nro 25. Departamento de Ciencias. PUCP. `http://argos.pucp.edu.pe/~jlbazan/download/Reporte-25.pdf`

[9] Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002). *Journal of the Royal Statistical Society Series B*, *64*, 3,616-618.

[10] Ben MG, Yohai VJ (2004). Quantile-quantile plot for deviance residuals in the generalized linear model. *J. of Comput. and Graphical Statistics*, 13(1): 36-47

[11] Carlin, B.P. y Louis, T.A. (2001).*Bayes and Empirical Bayes Methods for Data Analysis Essays on Item Response Theory.* Second edition. New York: Chapman & Hall.

[12] Casella, G. y Berger, R. L (2002). *Statistical Inference*,Duxbury: Pacific Grove, CA.

[13] 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.

[14] Chen, M. H., Dey, D., and Shao, Q-M. (2001). Bayesian analysis of binary data using skewed logit models. *Calcutta Statistical Association Bulletin*, *51*, 201-202.

[15] Chen, M-H, Shao, Q. M, & Ibrahim, J. G (2000). Propriety of Posterior Distribution for Dichotomous Quantal Response Models. *Proceedings of the American Mathematical Society*, *129*, 283-302.

[16] Chen, M-H, Shao, Q. M, & Ibrahim, J. G (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer Verlag.

[17] Collet, D. (2003). *Modelling binary data*. Chapman & Hall/CRC, Second Edition, Boca Raton, USA.

[18] Congdon, P. (2005). *Bayesian Models for Categorical Data*, Wiley.

[19] Czado, C., and Santner, T. J. (1992). The effect of link misspecifiaction on binary regression inference. *Journal of Statistical Planning and Inference*, *33*, 213-231

[20] Davison, A.C. and Gigli, A. (1989). Deviance residuals and n ormal score plots. *Biometrika*, *76*, 211-221

[21] Dey, D. K., Ghosh, S. K., Mallick, B. K. (eds) (2000). *Generalized Linear Models: a Bayesian Perspective*, New York, Dekker.

[22] Dalal, S.R., Fowlkes, E.B., and Hoadley, B. (1989). Risk analysis of space shuttle : Pre-Challenger Prediction of Failure, *Journal of the American Statistical Association*, *84*, 945-957.

[23] Gamerman, D. y Lopes, H. F (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman and Hall/CRC".

[24] Gelfand, A.E. y Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc. 85*, 398-409.

[25] Geman, S. y Geman, D. (1984). Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images *IEEE Transactions of Pattern Analysis an Machine Intelligence*, *6*, 721-741.

[26] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains. *Biometrika 57*, 97-109.

[27] Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle o-ring data, *Journal of the American Statistical Association*, *86*, 919-922.

[28] Hosmer, D. W. y Lemeshow, S. (1989). Applied logistic regression. Wiley, New York.

[29] Nagler J. (1994) Scobit: an alternative estimator to logit and probit. *American Journal Political Science*, *38*, 230-255.

[30] Paula, G. A. (2014). Modelos de Regressão com Apoio Computacional. Instituto de Matemática e Estatística. Disponible en `http://www.ime.usp.br/~giapaula/livro.pdf`

[31] Prentice, R. L. (1976) A Generalization of the probit and logit methods for dose-response curves. *Biometrika*, *32*, 761-768.

[32] Roberts, C., P.(2002) *The Bayesian Choice: from decision-theoretic foundations to computational implementation*. 2nd ed. New york: Springer-Verlag.

[33] Ross, S. (1995). *Stochastic Processes*, Wiley: New York, NY.

[34] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. y van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*, 583-639.

[35] Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W.R.(1996). *BUGS 0.5 examples* (Vol. 1 Version i). Cambrigde, UK: University of Cambride.

[36] Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D (2007) *OpenBUGS User Manual version 3.0.2*. MRC Biostatistics Unit, Cambridge.

[37] Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Stat.* 22, 1701-1762.

Jorge Luis Bazán
Departamento de Matemática Aplicada e Estatística,
Universidade de São Paulo
e-mail: jlbazan@icmc.usp.br