UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2017, Vol. 10, No. 2, pp. 225–246

Elicitación de una distribución *a priori* para el modelo logístico[1]

Eliciting an *a priori* distribution for a logistic model

Jenny Andrea Tangarife Quintero[a]
jatanga@unal.edu.co

Juan Carlos Correa Morales[b]
jccorrea@unal.edu.co

Resumen

En muchas situaciones resulta útil cuantificar información subjetiva que uno o varios expertos conocen acerca de un tema de interés, por esto, una parte importante dentro de la estadística bayesiana es la construcción de métodos de elicitación para hallar distribuciones de probabilidad.

Con el fin de contribuir al desarrollo en este campo,se desarrolló una metodología para elicitar los parámetros de la regresión logística con una sola covariable. El método que se plantea requiere que se fijen unos niveles de la covarible y en estos se asume una distribución binomial; para cada nivel se elicita el parámetro de interés mediante la metodología indirecta de muestras hipotéticas.

***Palabras clave*:** distribución *a priori*, distribución beta, distribución binomial, estadísica bayesiana, metodología indirecta.

Abstract

In many situations it is useful to quantify subjective information that one or more experts know about a topic of interest, therefore, an important part of Bayesian statistics is building elicitation methods for finding probability distributions.

In order to contribute to the development in this field, a methodology was developed to elicit the parameters of the logistic regression with a single covariate. The proposed method requires to determite the levels of the covarible, at each level a binomial distribution is assumed and the parameter of interest is elicited by using the indirect method hypothetical samples.

***Keywords*:** *a priori* distribution, bayesian statistic, beta distribution, binomial distribution, indirect methodology.

[1]DOI: http://dx.doi.org/10.15332/s2027-3355.2017.0002.03
Tangarife, J., Correa, J. (2017) Elicitación de una distribución *a priori* para el modelo logístico. *Comunicaciones en Estadística*, **10**(2), 225-246.

[a]Estudiante maestría, Escuela de estadística, Universidad Nacional de Colombia, sede Medellín

[b]Profesor asociado, Escuela de estadística, Universidad Nacional de Colombia, sede Medellín

1. Introducción

Una parte importante dentro de la estadística bayesiana es la construcción de métodos de elicitación para hallar distribuciones de probabilidad, dichas probabilidades subjetivas son una cuantificación del conocimiento de un experto acerca de un tema de interés (De Finetti 1937). El proceso de expresar conocimiento en términos de probabilidades no es simple y ha demostrado estar sujeto a algunos tipos de errores repetibles (Hora 2007) y es por esto que un protocolo de elicitación estructurado correctamente puede mejorar sustancialmente la calidad de los juzgamientos (Shephard & Kirkwood 1994). Muchos trabajos en este campo se han encaminado hacia la construcción de métodos de elicitación para los modelos más populares (Hamada et al. 2001), y otros a la comparación de estos diferentes métodos (Umesh 1998).

En un proceso de elicitación, el primer cuestionamiento a abordar es qué significa tener una elicitación exitosa así, una elicitación exitosa es aquella que logra representar fielmente la opinión del experto, esto sin importar si el conocimiento de este es cierto o no, es importante diferenciar entre la calidad del conocimiento del experto y la exactitud con la que la distribución de probabilidad construida refleja el conocimiento elicitado. Si el experto es un estadístico o está muy familiarizado con los conceptos estadísticos, entonces no será de gran necesidad direccionar esfuerzos a la construcción de métodos de elicitación para que estos sean de fácil entendimiento para la persona elicitada, pero esto es poco frecuente en la práctica y hace que la obtención de probabilidades subjetivas sea un proceso complejo que requiere de una serie de habilidades (Garthwaite & O'Hagan 2005). El uso de un facilitador entrenado es otro punto importante a considerar, puesto que este puede ayudar a traducir en probabilidades el conocimiento elicitado, lo cual es finalmente el objetivo de la elicitación.

Un método de elicitación es el puente entre las evaluaciones de un experto y la expresión de estas evaluaciones en una forma estadísticamente útil (Garthwaite & O'Hagan 2005), y es por esto que se debe prestar especial atención no solo a las cantidades que se elicitan, sino tambien al cómo estas cantidades son elicitadas (Kynn 2008). Cuando se diseña un cuestionario de elicitación es importante tomar en cuenta las consideraciones desde el campo sicológico (estudios sobre las Huerísticas y sesgos); adicionalmente, la selección de una técnica pueden ser una decisión crucial, antecedentes de éxito de otros investigadores con diversas técnicas puede proporcionar una guía para la selección. El método de elicitación se debe seleccionar en función de su costo, el experto y la forma de su conocimiento en la materia, si un experto se siente familiarizado y cómodo con una técnica en particular seráa una buena razón para elegir dicha técnica (Chesley 1975). Un registro de la elicitación debería llevarse, idealmente para todas las preguntas que se formulen, junto con las respuestas de los expertos, así como el proceso por el cual una distribución de probabilidad se ajustó a esas respuestas.

2. Elicitación en la regresión logística

Para la construcción de un modelo de regresión en el paradigma bayesiano, el conocimiento experto es introducido especificando una distribución *a priori* para el vector de parámetros β; en el caso del modelo logístico en el cual la distribución condicional de la variable respuesta Y sigue una distribución binomial con probabilidad dada por la media condicional $\pi(x)$, la distribución beta es una opción natural debido a que es su familia conjugada.

2.1. Distribución beta

La distribución beta es posible para una variable aleatoria continua que toma valores en el intervalo [0,1], lo que la hace muy apropiada para modelar proporciones. Por esta razón, es una familia conjugada natural para la distribución binomial. Esta distribución tiene dos parámetros, α y β.

Si se define una distribución *a priori* beta para el parámetro π de la distribución binomial, se tiene que $\pi|\alpha,\beta \sim beta(\alpha,\beta)$

$$P(\pi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{1}$$

Siempre que se define una familia conjugada como una distribución *a priori*, la distribución posterior pertenece a la misma familia de distribuciones, por lo tanto, la distribución posterior para el parámetro π es :

$$\pi|k,\alpha,\beta \sim Beta(\alpha+k,\beta+n-k) \tag{2}$$

Donde k corresponde al número de éxitos en n ensayos Bernoulli.

Finalmente, la estimación puntual de π corresponde a la media de una distribución beta con parámetros $\alpha = \alpha + k$ y $\beta = \beta + n - k$:

$$E(\pi) = \frac{\alpha+k}{\alpha+\beta+n} \tag{3}$$

2.2. Distribución normal truncada

La distribución beta es la conjugada natural de la distribución binomial y por esto frecuentemente es usada como su distribución *a priori*, en algunas situaciones cuando las probabilidades de éxito son muy bajas o muy altas producen parámetros elicitados de la distribución beta menores que 1 y en este caso esta distribución no es unimodal y con colas pesadas. Bajo esta situación, la distribución normal

truncada podría usarse y garantizaría la unimodalidad.

La distribución normal truncada es particularmente popular en casos donde se requiere describir patrones no negativos y un límite superior también es necesario (aunque la distribución beta es muy flexible).

Sea $X \sim N(\mu, \sigma^2)$ y su distribución condicional de $X \in [a, b] \subset \mathbb{R}$. La distribución condicional de X sobre el intervalo $[a, b]$ es la distribución normal truncada. La densidad condicional es:

$$f(x|x \in [a,b]) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{4}$$

Para $a \leq x \leq b$ donde ϕ y Φ representan la densidad y la CDF de una normal estándar respectivamente.

Los primeros trabajos en elicitación para modelos lineales generalizados (GLM) fueron propuestos por Bedrick et al. (1996)Bedrick et al. (1997), ellos propusieron un método de elicitación en el que la distribución predictiva es elicitada en diferentes puntos de diseño y luego combinada para formar una distribución *a priori*. Algunas formas especifícas de GLM, entre ellas la regresión logística han recibido especial atención e importantes desarrollos se han dado en el área de la ecología; ejemplo de esto, son los métodos presentados a continuación.

- Kynn (2005): herramienta gráfica interactiva de metodología indirecta llamada elicitor (complemento del *software* `WinBUGS`). Se pregunta al experto por dos puntos cualquiera y la mediana, luego se grafica la relación univariante entre la variable respuesta y una covarible (manteniendo todas las otras constantes). Inicialmente se usó para elicitar distribuciones normales *a priori* para un modelo de regresión logística, con el fin de estimar la presencia de especies en un ecosistema. Este método fue inspirado por Bedrick et al. (1996) y Garthwaite & Dickey (1988).
- Martin et al. (2005): método directo para elicitar opinión de expertos, usando cuestionarios, a partir de multiples expertos; y en dicha ocasión sólo se considerá una covariable para la regresión Poisson. O'Leary et al. (2008) adaptó este enfoque para uno o múltiples expertos y múltiples covariables en el contexto de la regresión logística; para cada covarible se preguntó a los expertos si el efecto sobre la variable respuesta incrementaba, disminuía o no existía. Este método no requiere conocimiento acerca de probabilidad o distribuciones.
- Garthwaite & Al-Awadhi (2006): desarrollaron un método en el área de la ecología que modela la distribución muestral mediante un modelo de regresión logística continuo lineal por partes, el cual es más flexible que el modelo

de regresión logística estándar, además se usaron gráficos interactivos para realimentar al experto.

- Denham & Mengersen (2007): método indirecto para el modelamiento ambiental, este procedimiento hace uso de la naturaleza geográfica de estos problemas e incorpora un sistema de información geográfico (SIG) para suministrar información acerca de la vegetación, tipos de rocas, precipitaciones, temperatura, etc. La elicitación de expertos en este caso se usó para relacionar todas estas variables con la probabilidad de presencia/ausencia de una especie en peligro de extinción. Durante este ejercicio, en lugar de especificar puntos de diseño como números, cada punto de diseño fue una ubicación real en Queensland.

- James et al. (2010): diseñaron el *software* elicitator e hicieron una aplicación a través de un estudio que tiene como objetivo desarrollar un modelo de regresión logística para predecir la distribución geográfica de una especie en un contexto ecológico; esta herramienta extiende el trabajo hecho por Denham & Mengersen (2007), puesto que soporta una variedad de aplicaciones y usos, e igualmente se usó un método indirecto. Se pide al experto para cada caso k con covariables $X_{1k}, X_{2k}, ..., X_{jk}$ conocidas, estimar la probabilidad de éxito Z_k, el rango de valores con probabilidad variable (percentiles) y su mejor estimación (moda), esta información se utiliza para estimar numéricamente μ_k y γ_k en $P(Z_k|x_k)$, posteriormente se proporciona realimentación al experto y se le da la oportunidad de modificar sus creencias. Esto se repite para $k = 1, 2, ..., K$. La información suministrada por el experto para todas las covaribles puede ser combinada para formar el modelo del experto y se utiliza una regresión beta para relacionar los datos del experto Z_k a las covariables (Choy et al. 2009).

$$Z_K \sim Beta(\mu_k, \gamma_k), \quad logit(\mu_k) = x_k\beta \tag{5}$$

  Por medio de la función "Link" el parámetro de forma a_k y el parámetro escala b_k para la probabilidad esperada de éxito es $\mu_k = a_k/\gamma_k$ y el tamaño de muestra efectivo del experto es $\gamma_k = a_k + b_k$.

La gran mayoría de los métodos mencionados en la recopilación anterior, coinciden en el uso de una metodología indirecta de elicitación, a excepción del método presentado por Martin et al. (2005). En el caso de un modelo de regresión, el uso de un método directo requeriría que el experto cuantificara el impacto de un cambio en el valor de la covariable sobre la varible respuesta, siendo aún más complicado el caso de la regresión logística puesto que esta relación no es lineal. En la práctica es poco probable que el experto sea capaz de hacer una estimación directa sobre los parámetros del modelo , incluso si están bien informados de la relación que se está modelando (Huson & Kinnersley 2008). Desde el punto de vista del modelista estadístico, un enfoque directo puede resultar más fácil, pero este podría producir resultados menos precisos en comparación con un enfoque indirecto, especialmente para expertos un poco ajenos a los conceptos de probabilidad. Un

enfoque indirecto resulta más fácil para el experto; estos se sienten más cómodos estimando cantidades observables, que en un modelo de regresión equivale a una estimación de la variable respuesta, para diferentes valores de las covariables (Choy et al. 2009). Pero a menudo requiere más esfuerzo del modelista en el diseño del método de elicitación y la codificación para transformar respuestas de los expertos en la forma requerida (James et al. 2010). Kadane & Wolfson (1998) recalcan que el objetivo de la elicitación es que sea lo más fácil posible para los expertos en la materia, en términos probabilísticos, al tiempo que se reduce la necesidad de un conocimiento acerca de la teoría de probabilidad.

La propuesta de elicitación para determinar la *a priori* conjunta para el modelo de regresión logística se basa en el uso de la metodología indirecta de muestras hipotéticas. El modelo a elicitar es:

$$logit(\pi) = \beta_0 + \beta_1 x \tag{6}$$

Se pretende elicitar la distribución conjunta para β_0 y β_1, que finalmente será una distribución normal bivarible.

3. Algoritmo propuesto

1. Se fijan los niveles de la covariable adecuados $x_1, x_2, ..., x_k$, estos puntos se deben elegir en consenso con el experto. Chaloner & Larntz (1989) concluyen que para un modelo de regresión logística el número mínimo de puntos de diseño es igual número de parámetros a estimar.

   Recomendación: los puntos de diseño deben ser tomados siguiendo las recomendaciones dadas a continuación, para evitar tomar puntos donde la probabilidad de éxito es muy cercana a cero o muy cercana a uno.

   - El primer punto debe ser tomado de tal forma que corresponda a una probabilidad de éxito aproximadamente de 0.5, se debe seleccionar de tal manera que sea igualmente probable que el verdadero valor sea mayor o menor que este punto.
   - El segundo punto debe ser tomado de tal forma que la probabilidad de éxito corresponda aproximadamente al 0.25, se debe seleccionar de tal manera que si el valor verdadero está por debajo de la mediana, sea igualmente probable que sea por encima o por debajo de este valor.
   - El tercer punto debe ser tomado de tal forma que la probabilidad de éxito corresponga aproximadamente al 0.75, se debe seleccionar de tal manera que, si el valor verdadero está por encima de la mediana, es igualmente probable que sea por encima o por debajo de este valor.

2. Para cada nivel se procede así:

- Se fija un n y se pide al experto dar el número de éxitos que él esperaría se den en una muestra hipotética de tamaño n, dígase X_0, calcule $E(\pi) = X_0/n$.
- Para el mismo n se pide al experto dar el número máximo de éxitos que él esperaría aceptable, X_L, calcule $\pi_L = X_L/n$.
- Para el mismo n se pide al experto dar el número mínimo aceptable, X_I, calcule $\pi_I = X_I/n$.

  Se puede repetir este paso las veces que se consideren necesarias, esto sirve para evaluar la consistencia del experto con las diferentes muestras hipotéticas.

3. A los valores elicitados en el punto dos. se ajusta una distribución beta para estimar los parámetros α y β

   Sean:

   $E(\pi) = x_0/n$
   $P(\pi \geq X_L/n) = 0.05$
   $P(\pi \leq X_I/n) = 0.05$

   Los valores α y β se obtienen de minimizar la siguiente función:

   $$f(\alpha, \beta) = (\pi_I - qbeta(0.05, \alpha, \beta))^2 + (\pi_L - qbeta(0.95, \alpha, \beta))^2 + (\pi + \alpha/(\alpha+\beta))^2$$

4. Calcule el N equivalente, esto permite cuantificar el conocimiento del experto en términos de tamaño muestral, este tamaño representa realmente el nivel de conocimiento sobre el parámetro que el experto tiene; donde un tamaño muestral pequeño indica un menor conocimiento y un tamaño muestral grande indica un mayor conocimiento (Sedlmeier 1999). El N equivalente se halla usando la ecuación de un intervalo de confianza para la proporción basado en el teorema central del límite:

   $$\left(\hat{\pi} - Z_{(\alpha/2)} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}, \hat{\pi} + Z_{(\alpha/2)} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \right) \tag{7}$$

   Donde $\hat{\pi}$ es dado por el experto como el número de éxitos más probable.

   Sean a y b el límite inferior y superior del intervalo de confianza respectivamente:

   $$a = \hat{\pi} - Z_{(\alpha/2)} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{8}$$

$$b = \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{9}$$

Luego tomando la diferencia entre (4-3) y (4-4),

$$b - a = 2Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

$$N = \frac{4Z^2_{(\alpha/2)}\hat{\pi}(1-\hat{\pi})}{(b-a)^2} \tag{10}$$

Dado que se calcula un N equivalente por nivel de la covariable, el N equivalente del experto será el N del punto diseño que tenga asociado el menor valor.

5. Para cada nivel repita los siguientes pasos m veces:

   - Genere un valor de la beta con α_i y β_i hallados en el punto tres.

   $$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{bmatrix}$$

   En este paso se obtiene un vector de tamaño k.
   - Genere una muestra de valores y de la distribución binomial.

   $$\begin{bmatrix} y_1^{(1)}, & x_1 \\ y_2^{(1)}, & x_1 \\ \vdots & \\ y_{n_{eq}}^{(1)}, & x_1 \\ y_1^{(2)}, & x_2 \\ y_2^{(2)}, & x_2 \\ \vdots & \\ y_{n_{eq}}^{(2)}, & x_2 \\ \vdots & \\ y_1^{(k)}, & x_k \\ y_2^{(k)}, & x_k \\ \vdots & \\ y_{n_{eq}}^{(k)}, & x_k \end{bmatrix}$$

   En este paso se obtiene una matriz de tamaño $(n_{eq} * k)$x2, donde n_{eq} es el N equivalente hallado en el punto cuatro y $x_1, x_2, ...x_k$ son los niveles de la covarible.

- Con la tabla de datos construida en el punto anterior estime los parámetros de la regresión logística. Guarde los resultados.
  $\texttt{betas}^{(1)} < -glm(Y \sim X, family = "binomial")\$coef$

$$\begin{bmatrix} \beta_0^{(1)}, & \beta_1^{(1)} \\ \beta_0^{(2)}, & \beta_1^{(2)} \\ \vdots & \\ \beta_0^{(m)}, & \beta_1^{(m)} \end{bmatrix}$$

  En este paso se obtiene una matriz de tamaño mx2.

6. Ajuste la normal bivariable a los betas hallados en el paso anterior.

4. Uso de la distribución normal trucada

En muchos casos donde la probabilidad de éxito es muy baja o muy alta, los parámetros de la distribución beta están por debajo de uno, en esta situación dicha distribución no es unimodal y tiene colas pesadas. En este caso se recomienda usar la distribución normal truncada entre [0,1] y aqué se garantizaría la unimodalidad. El algoritmo presentado a continuación es una modificación al presentado anteriormente y permitiría ajustar como distribución *a priori* para cada punto de diseño la distribución normal truncada.

1. Se fijan los niveles de la covariable adecuados $x_1, x_2, ..., x_k$. Chaloner & Larntz (1989) concluyen que para un modelo de regresión logística el número mínimo de puntos de diseño es igual número de parámetros a estimar.

   Recomendación: los puntos de diseño deben ser tomados siguiendo las recomendaciones dadas a continuación, para evitar tomar puntos donde la probabilidad de éxito es muy cercana ac cero o muy cercana a uno.

   - El primer punto debe ser tomado de tal forma que corresponda a una probabilidad de éxito aproximadamente de 0.5.
   - El segundo punto debe ser tomado de tal forma que la probabilidad de éxito corresponda aproximadamente al 0.25
   - El tercer punto debe ser tomado de tal forma que la probabilidad de éxito corresponga aproximadamente al 0.75

2. Para cada nivel se procede así:

   - Se fija un n y se pide al experto dar el número de éxitos que el esperaría se den en una muestra hipotética de tamaño n, dígase X_0, calcule $E(\pi) = X_0/n$.

- Para el mismo n se pide al experto dar el número máximo de éxitos que él esperaría aceptable, X_L, calcule $\pi_L = X_L/n$.
- Para el mismo n se pide al experto dar el número mínimo aceptable, X_I, calcule $\pi_I = X_I/n$.

  Se puede repetir este paso las veces que se consideren necesarias, esto sirve para evaluar la consistencia del experto con las diferentes muestras hipotéticas.

3. Los valores elicitados en el punto dos. permiten estimar los parámetros la media y la desviación típica de una distribución normal.

   Sean:

   $E(\pi) = X_0/n$
   $P(\pi \leq X_L/n) = 0.95$
   $P(\pi \leq X_I/n) = 0.05$

   El valor de σ se obtienen así:

   $dt1 = (\pi_I - \pi)/Q_{(0.05)}$
   $dt2 = (\pi_L - \pi)/Q_{(0.95)}$
   $\sigma = (dt1 + dt2)/2$

   Donde $Q_{(0.05)}$ y $Q_{(0.95)}$ son cuantiles teóricos de la distribución normal.

4. Calcule el N equivalente, esto permite cuantificar el conocimiento del experto en términos de tamaño muestral, este tamaño representa realmente el nivel de conocimiento sobre el parámetro que el experto tiene; donde un tamaño muestral pequeño indica un menor conocimiento y un tamaño muestral grande indica un mayor conocimiento (Sedlmeier 1999). El N equivalente se halla usando la ecuación de un intervalo de confianza para la proporción basado en el teorema central del límite:

$$\left(\hat{\pi} - Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}, \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}\right) \tag{11}$$

   Donde $\hat{\pi}$ es dado por el experto como el número de éxitos más probable.

   Sean a y b el límite inferior y superior del intervalo de confianza respectivamente:

$$a = \hat{\pi} - Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{12}$$

$$b = \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{13}$$

Luego tomando la diferencia entre (4-3) y (4-4),

$$b - a = 2Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

$$N = \frac{4Z^2_{(\alpha/2)}\hat{\pi}(1-\hat{\pi})}{(b-a)^2} \tag{14}$$

Dado que se calcula un N equivalente por nivel de la covariable, el N equivalente del experto será el N del punto diseño que tenga asociado el menor valor.

5. Para cada nivel repita los siguientes pasos m veces:

- Genere un valor de la distribución normal truncada en el intervalo $[0, 1]$ con μ_i y σ_i hallados en el punto tres.

$$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{bmatrix}$$

En este paso se obtiene un vector de tamaño k.

- Genere una muestra de valores y de la distribución binomial.

$$\begin{bmatrix} y_1^{(1)}, & x_1 \\ y_2^{(1)}, & x_1 \\ \vdots & \\ y_{n_{eq}}^{(1)}, & x_1 \\ y_1^{(2)}, & x_2 \\ y_2^{(2)}, & x_2 \\ \vdots & \\ y_{n_{eq}}^{(2)}, & x_2 \\ \vdots & \\ y_1^{(k)}, & x_k \\ y_2^{(k)}, & x_k \\ \vdots & \\ y_{n_{eq}}^{(k)}, & x_k \end{bmatrix}$$

En este paso se obtiene una matriz de tamaño $(n_{eq} * k)$x2, donde n_{eq} es el N equivalente hallado en el punto cuatro y $x_1, x_2, ...x_k$ son los niveles de la covarible.

- Con la tabla de datos construida en el punto anterior estime los parámetros de la regresión logística. Guarde los resultados.
  $\texttt{betas}^{(1)} < -glm(Y \sim X, family = binomial)\$coef$

$$\begin{bmatrix} \beta_0^{(1)}, & \beta_1^{(1)} \\ \beta_0^{(2)}, & \beta_1^{(2)} \\ \vdots & \\ \beta_0^{(m)}, & \beta_1^{(m)} \end{bmatrix}$$

En este paso se obtiene una matriz de tamaño mx2.

6. Ajuste la normal bivariable a los betas hallados en el paso anterior.

5. Aplicación: modelo logístico para el cáncer de próstata en Colombia

5.1. Introducción

Desde 1990 el cáncer de próstata en Colombia ha venido en aumento, entre 1990 y 2013 el número de nuevos casos de tumores malignos de próstata al año pasí de 3.200 a 13.200, así lo reveló el estudio "La carga mundial del cáncer 2013" elaborado por el consorcio internacional de investigadores del instituto para medición y evalucación de la salud.

Es importante entender que este tipo de tumor está relacionado con el envejecimiento, es decir, a mayor edad, mayor probabilidad de desarrollarlo; por otra parte, el cáncer de próstata es más frecuente en los hombres afroamericanos que en los blancos, los hombres de raza negra presentan un mayor riesgo de padecer este tipo de cáncer que los de raza blanca. también tienen más probabilidades de desarrollar cáncer de próstata a una edad más temprana y de tener tumores agresivos, de crecimiento rápido. Se desconocen los motivos exactos de estas diferencias, los cuales pueden estar vinculados con factores socioeconómicos y de otros tipos. Los hombres hispanos tienen un menor riesgo de desarrollar cáncer de próstata y de morir por la enfermedad que los hombres de raza blanca. El cáncer de próstata se produce con más frecuencia en América del Norte y el norte de Europa. también parece que el cáncer de próstata está aumentando entre los asiáticos que viven en áreas urbanizadas, como Hong Kong, Singapur, y ciudades de América del Norte y de Europa, particularmente, entre aquellos que llevan un estilo de vida más occidental.

En este artículo se aplica el método de elicitacón propuesto a la relación que existe entre el cáncer de próstata y la edad en hombres de raza negra, además se pretende determinar la prevalencia de este tipo de cáncer por edad. Conocer la prevalencia del cáncer de próstata por grupos de edad es importante, debido a que en personas de menor edad el diagnóstico puede ser mas tardío, puesto que se tiene la concepción de que este tipo de cáncer se presenta con mayor frecuencia sólo en hombres de edad avanzada, como se comentó anteriormente, este supuesto puede no cumplirse para hombres de raza negra; además, en algunos casos este tipo de cáncer no presenta síntomas muy evidentes y un hombre por debajo de los 50 años probablemente no se haga chequeos rutinarios.

5.2. Metodología

En este proceso de elicitación se siguieron las etapas recopiladas por Jenkinson (2005) para lograr un proceso de elicitación éxitoso:

- Inicialmente se día al experto una contextualización en el tema, donde se presentó el objetivo de la investigación, se le explicó que todas las preguntas estarían basadas en muestras hipotéticas de hombres de raza negra de diferentes edades, además se validó que estas fueran entendidas, que estuviera en capacidad y se sintiera cómodo dando este tipo de información.

- Como experto se tiene al doctor Manuel García, profesor de la Universidad Sur Colombiana, que cuenta con investigaciones en biología prostática y biología de la reproducción, además, un posdoctorado de la universidad de Sao Paulo, el cual titula ¿Analisis de microRNA que regula el Receptor de Androgeno en el cáncer de próstata?.

- En la etapa de estructuración descomposición y entrenamiento en probabilidad, se le da al experto una introducción sobre la distribución binomial y su relación con el modelo logístico, información acerca de las heurísticas y sesgos a los cuales se tendría que enfrentar. Adicionalmente, se le especifica al experto que la cantidad incierta que se quiere elicitar es el número de hombres de raza negra con cáncer de próstata en diferentes niveles de edad.

- Aplicación del método,

    1. Se pide al experto dar un intervalo de la edad en el cual sea de interés conocer la prevalencia del cáncer de próstata, además, ubicarse en puntos cercanos al cuantil 25, 50 y 75. Luego de llegar a un consenso con el experto, se eligieron los niveles de edad: 50, 60, 65, 70. El experto dio cuatro puntos, aunque para estimar los parámetros de una regresión logística con una sola covarible son suficientes dos puntos, se decidío elicitar en estos cuatros puntos dado que el proceso de elicitación es de fácil aplicación.

2. Para cada uno de los niveles de edad hallados en el punto anterior, se da al experto diferentes muestras hipotéticas de hombres de raza negra y se le pide dar el número de casos de cáncer de prostata que el espera encontrar y el número mínimo y máximo de casos que el considera aceptable. Se repite este proceso tres veces con tres diferentes muestras hipotéticas, se calculan las proporciones en cada nivel de la edad y se verifica que el experto haya sido consistente con sus respuestas.
3. Se calcula el N equivalente del experto, reemplazando (4-5) los valores elicitados en el punto anterior. Se obtuvo un N equivalente de 203.
4. En esta etapa se realiza la simulación y se estiman los parámetros β_0, β_1 del modelo logístico para la prevalencia de cáncer de próstata en hombres de raza negra en relación con la edad.

- Una vez se estimaron los coeficientes del modelo, se mostraron al experto y se le explicó su significado e implicaciones, con el fin de validar si dichos resultados reflejan sus creencias. En caso de que el experto estuviera de acuerdo con el modelo se daba por finalizado el proceso, en caso contrario, se ajustarían los casos de cáncer de próstata para las diferentes muestras hipotéticas.

5.3. Resultados

Las proporciones halladas para cada nivel de la edad son, presentadas en la tabla 1

Tabla 1: *Proporciones elicitadas. Fuente: elaboración propia.*

Edad	x_0	x	x_L
50	0.02	0.05	0.08
60	0.03	0.08	0.1
65	0.13	0.17	0.22
70	0.34	0.38	0.42

La estimación de los parámetros para el modelo logístico junto con su intervalo de probabilidad son, presentadas en la tabla 2

Tabla 2: *parámetros estimados. Fuente: elaboración propia.*

parámetro	estimación	LI	LS
β_0	-11.104	-12.449	7.430
β_1	0.149	0.126	0.220

así el modelo estimados es:

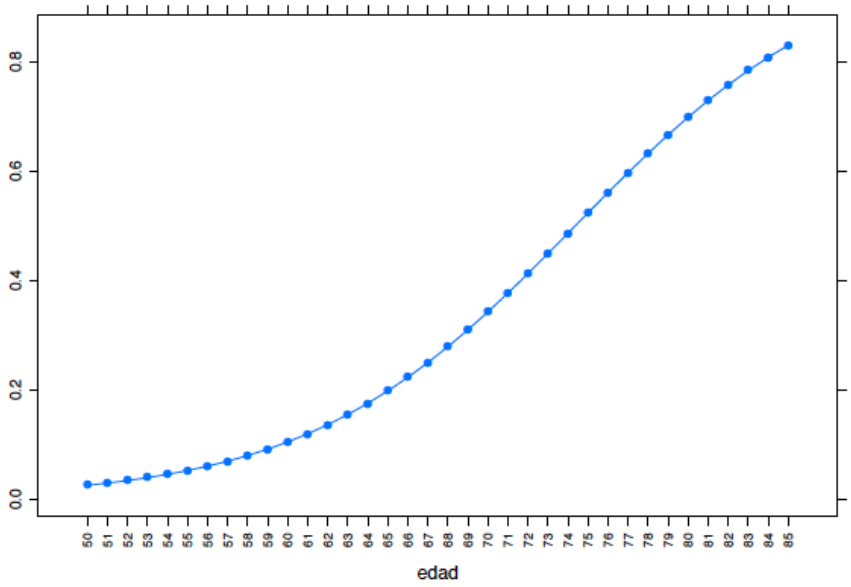

Figura 1: *Distribución logística para el cáncer de próstata en hombres de raza negra. Fuente: elaboración propia.*

$$ln(\pi/1-\pi) = -11.104 + 0.149 * Edad$$

Del signo de β_1 se puede concluir que la edad es un factor de riesgo para el cáncer de próstata, al aumentar la edad aumenta la probabilidad de sufrir la enfermedad. Conocer cómo es el comportamiento de la prevalencia de la enfermedad por edad es importante a la hora de diseñar políticas de prevención, como se ve puede ver en el gráfico anterior a la edad de 60 años la prevalencia de cáncer de próstata empieza a incrementar más rápidamente, es por esto que a esta edad se deben incrementar los chequeos rutinarios para una detección temprana de la enfermedad. Existen diferentes estudios en los que se ha mostrado la relación de esta enfermedad con la edad pero para poblaciones en general, si bien se conoce que es más probable en hombres de raza negra no hay muchos datos que indiquen a qué edad se deben incrementar los chequeos para una detección temprana.

6. Conclusiones

En este trabajo se estudiaron los diferentes puntos a considerar en el momento de afrontar un proceso de elicitación, con el fin de desarrollar un método que permita estimar los parámetros β_0 y β_1 de la regresión logística con una sola cova-

riable. El uso de un método indirecto es una alternativa amigable para el experto y aunque este tipo de metodología puede resultar un poco más complicada para el analista estadístico, el uso de muestras hipotéticas en particular disminuye este inconveniente, aunque el principal objetivo sea la fácilidad para el experto; estos dos puntos se traducen en un método de elicitación de fácil aplicación y de mayor facilidad para generar resultados. Desde este punto de vista, es un aporte importante al área debido a que se hace uso de una distribución *a priori* informativa y este método de elicitación no involucra apuestas.

Recibido: 18 de Julio de 2016
Aceptado: 3 de Octubre de 2017

Referencias

Barrera, C. (2015), 'Analysis of the elicited prior distributions using tools of functional', *Tesis Doctoral, Universidad Nacional de Colombia* .

Bedrick, E., Christensen, R. & Johnson, W. (1996), 'A new perspective on priors for generalized linear models', *The American Statistical Association* **91**, 1450–1460.

Bedrick, E., Christensen, R. & Johnson, W. (1997), 'Bayesian binomial regression: Predicting survival at a trauma center', *the American Statistical Association* **51**, 211–218.

Burgman, M., Fidler, F., McBride, M., Walshe, T. & Wintle, B. (2007), 'Eliciting expert judgments: Literature review', *University of Melbourne* .

Cannell, C. (1977), 'A summary of studies of interviewingmethodology', *Vital and Health Statistics* **2**, 69–72.

Chaloner, K. & Duncan, T. (1983), 'Assessment of a beta prior distribution: Pm elicitation', *The Statistician* pp. 174–180.

Chaloner, K. & Larntz, K. (1989), 'Optimal bayesian design applied to logistic regression experiments', *Statistical Planning and Inference* **21**, 191–208.

Chesley, G. (1975), 'Elicitation of sub active probabilities: A review', *The Accounting Review* **50**, 325–337.

Choy, L., James, A. & Mengersen, K. (2009), 'Expert elicitation and its interface with technology: a review with a view to designing elicitator', *The Accounting Review* **18**, 13–17.

De Finetti, B. (1937), 'La prevision: ses lois logigues, ses sources subjectives', *Annal es de l'Institut Henri Poincard* **7**, 1–68.

Denham, R. & Mengersen, K. (2007), 'Geographically assisted elicitation of expert opinion for regression models', *Bayesian Analysis* **2**, 99–136.

Garthwaite, P. & Al-Awadhi, S. (2006), 'Quantifying opinion about a logistic regression using interactive graphics', *Statistics Group* **6**.

Garthwaite, P. & Dickey, J. (1988), 'Quantifying expert opinion in linear regression problems', *The Royal Statistical Society* **29**, 462–474.

Garthwaite, P. & O'Hagan, A. (2005), 'Statistical methods for eliciting probability distributions', *The American Statistical Association* **100**, 680–700.

Hamada, M., Martz, H. F., Reese, C. S. & Wilson, A. G. (2001), 'Finding near-optimal bayesian experimental designs via genetic algorithms', *The American Statistician* **55**, 175–181.

Hora, S. (2007), 'Advances in desicion analysis: From foundations to applications', *Cambridge University Press* pp. 129–153.

Huson, L. & Kinnersley, N. (2008), 'ayesian fitting of a logistic dose-response curve with numerically derived priors', *John Wiley and Sons Inc* .

James, A., Low Choy, S. & Mengersen, K. (2010), 'Elicitator: an expert elicitation tool for regression in ecology', *Environmental Modelling and Software* **25**, 129–145.

Kadane, J. & Wolfson, L. (1998), 'Experiencies in elicitation', *The Statistician* **47**, 3–19.

Kynn, M. (2005), 'Designing elicitor: Software to graphically elicit expert priors for logistic regression models in ecology', *Department of Mathematics and Statistics, Fylde College, Lancaster University* .

Kynn, M. (2008), 'The heuristics and biases bias in expert elicitation', *The Royal Statistical Society* **171**, 239–264.

Martin, T., Kuhnert, P., Mengersen, K. & Possingham, H. (2005), 'he power of expert opinion in ecological models: a bayesian approach examining the impact of livestock grazing on birds', *Ecological Applications* **15**, 266–280.

O'Leary, R., Low Choy, S., Mengersen, K., Kynn, M., Kuhnert, P., Denham, R., Martin, T. & Murray, J. (2008), 'Comparison of expert elicitation methods for logistic regression for presence of endangered brush-tailed rock-wallaby petrolage penicillata', *Environmetrics* .

Sedlmeier, P. (1999), 'Improving statistical reasoning: theoretical models and practical', *implications, Lawrence Erlbaum, Mahwah, NJ* .

Shephard, G. & Kirkwood, C. (1994), 'Managing the judgmental probability elicitation process: A casestudy of analyst/manager interaction', *IEEE Transactions on Engineering Management* **41**, 414–425.

Umesh, G. (1998), 'Comparison of two elicitation methods for a prior for a binomial parameter', *Management Science* **34**, 784–790.

A. Códigos

A.1. Código: uso de la distribución beta como distribución *a priori*

```
# metodologia propuesta
library(XLConnect)
library(RODBC)
#
temp <- odbcConnectExcel2007("Datoselicitación.xlsx")
# Se leen los puntos de diseño dados por el experto
puntos <- sqlFetch(temp, "Puntos")
odbcCloseAll()
# Calculo n equivalente
n1 <- NULL
n_equ <- function(x){
for(i in 1:nrow(x)){
n <- round((4*1.96^2*x[i,3]*(1 - x[i,3]))/(x[i,4]-x[i,2])^2)
n1 <- rbind(n1,n)}
return(n1)
}
(n_exp <- min(n_equ(puntos)))
#
#
# ajuste de la distribución beta
ajuste.beta <- function(teta,valores= valores){
alfa<-teta[1]
beta<-teta[2]
cuantil0.05<-valores[1]
cuantil0.95<-valores[3]
media<-valores[2]
cuant1.teo<-qbeta(0.05,alfa,beta)
cuant2.teo<-qbeta(0.95,alfa,beta)
media.teo<-alfa/(alfa+beta)
res<-(cuantil0.05-cuant1.teo)^2
+(cuantil0.95-cuant2.teo)^2
+(media-media.teo)^2
return(res)
}
#
```

```
alfas1 <- NULL
par_betas <- function(x){
for(i in 1:nrow(x)){
valores <- x[i,2:4]
alfas <- optim(c(1,1),ajuste.beta,method="L-BFGS-B",
lower=c(1,1)/1000000,upper=c(10,10),
valores = valores)[[1]]
alfas1 <- rbind(alfas1,alfas)}
return(alfas1)
}
parametros_beta <- matrix(unlist(par_betas(x = puntos)), ncol = 2, byrow = F)
colnames(parametros_beta) <- c("alfa","beta")
#
puntos1 <- cbind(puntos,parametros_beta)
#
#
beta1 <- NULL
const1 <- NULL
for(i in 1:1000){
pi1 <- rbeta(n = 1, puntos1[1,5], puntos1[1,6])
pi2 <- rbeta(n = 1, puntos1[2,5], puntos1[2,6])
pi3 <- rbeta(n = 1, puntos1[3,5], puntos1[3,6])
pi4 <- rbeta(n = 1, puntos1[4,5], puntos1[4,6])

#
y1 <- sample(c(0,1), prob = c(pi1,1-pi1), n_exp, replace = T)
y2 <- sample(c(0,1), prob = c(pi2,1-pi2), n_exp, replace = T)
y3 <- sample(c(0,1), prob = c(pi3,1-pi3), n_exp, replace = T)
y4 <- sample(c(0,1), prob = c(pi4,1-pi4), n_exp, replace = T)

#
Y <- c(y1,y2,y3,y4) # se adicionan tanto y's como puntos de diseño
X <- c(rep(puntos1[1,1],n_exp),rep(puntos1[2,1],n_exp),
       rep(puntos1[3,1],n_exp),rep(puntos1[4,1],n_exp))
const <- glm(Y~X, family = "binomial")$coef[1]
beta <- glm(Y~X, family = "binomial")$coef[2]
#
#
const1 <- c(const1,const)
beta1 <- c(beta1,beta)
}
#
# Ajuste de la distribución normal multivariada
 a los parámetros Alfa y Beta hallados
library(MASS)
fitdistr(beta1,"normal")
```

```
fitdistr(const1,"normal")
#
histogram(beta1,xlab = "", ylab = "",
main = expression(paste('distribución del parámetro',sep = " ",beta,'1')))
#
histogram(const1, xlab = "", ylab = "" ,
main = expression(paste('distribución del parámetro',sep = " ",beta,'0')))
#
```

A.2. Código: uso de la distribución normal truncada como distribución *a priori*

```
# metodologia propuesta
#
library(XLConnect)
library(RODBC)
# se leen los puntos de diseño dados por el experto
temp <- odbcConnectExcel2007("Datoselicitación.xlsx")
puntos <- sqlFetch(temp, "Puntos")
odbcCloseAll()
#
# Calculo n equivalente
n1 <- NULL
n_equ <- function(x){
for(i in 1:nrow(x)){
n <- round((4*1.96^2*x[i,3]*(1 - x[i,3]))/(x[i,4]-x[i,2])^2)
n1 <- rbind(n1,n)}
return(n1)
}
(n_exp <- min(n_equ(puntos)))
#
#
ajuste.normal <- function(valores){
media<-valores[2]
cuantil0.05<-valores[1]
cuantil0.95<-valores[3]
#
# necesitamos calcular la desviación típica
dt1<-(cuantil0.05-media)/qnorm(0.05)
dt2<-(cuantil0.95-media)/qnorm(0.95)

desvi.tip<-(dt1+dt2)/2

return(c(media,desvi.tip))
}
```

```
param1 <- NULL
for(i in 1:nrow(puntos)){
valores <- puntos[i,2:4]
param <- ajuste.normal(valores)
param1 <- c(param1,param)
}
param1 <- matrix(unlist(param1), ncol = 2, byrow = T)
colnames(param1) <- c("media","desv")
puntos1 <- cbind(puntos,param1)
#
beta1 <- NULL
const1 <- NULL
for(i in 1:10000){
pi1 <- qnorm(runif(1,pnorm(0,mean = puntos1[1,5],sd = puntos1[1,6]),
pnorm(1, mean =puntos1[1,5] ,sd = puntos1[1,6])),
mean = puntos1[1,5],sd = puntos1[1,6])
#
pi2 <- qnorm(runif(1,pnorm(0,mean = puntos1[2,5],sd = puntos1[2,6]),
pnorm(1, mean =puntos1[2,5] ,sd = puntos1[2,6])),
mean = puntos1[2,5],sd = puntos1[2,6])
#
pi3 <- qnorm(runif(1,pnorm(0,mean = puntos1[3,5],sd = puntos1[3,6]),
pnorm(1, mean =puntos1[3,5] ,sd = puntos1[3,6])),
mean = puntos1[3,5],sd = puntos1[3,6])
#
pi4 <- qnorm(runif(1,pnorm(0,mean = puntos1[4,5],sd = puntos1[4,6]),
pnorm(1, mean =puntos1[4,5] ,sd = puntos1[4,6])),
mean = puntos1[4,5],sd = puntos1[4,6])
#
y1 <- sample(c(0,1), prob = c(1-pi1,pi1), n_exp, replace = T)
y2 <- sample(c(0,1), prob = c(1-pi2,pi2), n_exp, replace = T)
y3 <- sample(c(0,1), prob = c(1-pi3,pi3), n_exp, replace = T)
y4 <- sample(c(0,1), prob = c(1-pi4,pi4), n_exp, replace = T)
#
Y <- c(y1,y2,y3,y4) # se adicionan tanto y's como puntos de diseño
X <- c(rep(puntos1[1,1],n_exp),rep(puntos1[2,1],n_exp),
      rep(puntos1[3,1],n_exp),rep(puntos1[4,1],n_exp))
#
const <- glm(Y~X, family = "binomial")$coef[1]
beta <- glm(Y~X, family = "binomial")$coef[2]
#
const1 <- c(const1,const)
beta1 <- c(beta1,beta)
}
#
quantile(beta1, pr = c(0.25,0.975))
```

```
quantile(const1, pr = c(0.25,0.975))
#
library(MASS)
fitdistr(beta1,"normal")
fitdistr(const1,"normal")
#
histogram(beta1,  xlab = "", ylab = "",
main = expression(paste('distribución del parámetro',sep = " ",beta,'1')))
#
histogram(const1, xlab = "", ylab = "" ,
main = expression(paste('distribución del parámetro',sep = " ",beta,'0')))
#
0.1171922 + 2*0.01337853
# intervalo de confianza para b1
0.1493276076 + 1.96 *(0.0328542429/sqrt(n_exp))
0.1493276076 - 1.96 *(0.0328542429/sqrt(n_exp))
#
# para b0
pii <- NULL
edad <- 50:85
for(i in edad) {
pi <- exp(-11.10430093 + 0.1493276076*i)/(1+exp(-11.10430093 + 0.1493276076*i))
pii <- c(pii,pi)
}
pronosticos <-  unlist(pii)
pronosticos <- cbind(edad, pi = pronosticos)
plot(pronosticos[,2], ylab = "",xlab = "Edad",type = "l",
main = "Probabilidad para el cáncer de próstata \n en hombres de raza negra",
axes = "F")
axis(1, 1:36, 50:85, cex =0.6 )
axis(2)
box()
xyplot(pronosticos[,2] ~ edad, type = c("b"),
main = "distribución logística para el cáncer de próstata
        en hombres de raza negra",
ylab = "" ,pch = 19,
scale = list(rot = 90,x = list(at = 50:85, labels = 50:85,cex = .7)) )
```