APLICACION DEL ALGORITMO EN EL ANALISIS BAYESIANO DEL DISEÑO DE WARNER

CARLOS ARIAS MARTIN
JESUS BASULTO SANTOS
JAVIER BUSTO GUERRERO

Universidad de Sevilla

1. INTRODUCCION

Cuando se realizan encuestas, es frecuente que se presenten problemas al hacer preguntas íntimas a los sujetos (por ejemplo, preguntas sobre abortos, fraude fiscal, etc.). Los sujetos, para ocultar la información que se les solicita, suelen no contestar la pregunta, o de forma intencionada, ofrecer una respuesta falsa. Así, es difícil hacer inferencias sobre preguntas sensibles cuando se realizan encuestas en las que las preguntas se hacen directamente.

Un intento de evitar las dificultades señaladas anteriormente ha sido abordado a partir del desarrollo de los diseños de respuesta aleatorizada, los cuales introducen un elemento probabilístico en la selección de las preguntas. El entrevistado selecciona una pregunta por medio de un instrumento aleatorio de un conjunto de preguntas ofrecidas. El resultado del instrumento aleatorio no es conocido por el entrevistador, y así, el entrevistado puede contestar de forma honesta sin tener que revelar información sobre su intimidad. Como las propiedades del instrumento aleatorio son conocidas por el experimentador, las respuestas dadas por los entrevistados aportan suficiente información sobre la pregunta sensible de forma que permitan hacer inferencias sobre parámetros globales de la población.

Se han desarrollado varios diseños de respuesta aleatorizada, el trabajo de Horvitz, Greenberg y Abernathy (1976) proporciona un resumen comprensible de dichos diseños. El primer diseño de respuesta aleatorizada fue propuesto por Warner (1965). Los estudios de inferencia basados en el diseño de Warner han utilizado fundamentalmente los métodos clásicos, siendo el trabajo de Winkler y Franklin (1979) el primero que utiliza los métodos bayesianos.

Los problemas de inferencia que plantean los diseños de respuesta aleatorizada han entrado en una fase de resolución gracias al reciente trabajo de Bourke y Morán (1988). Estos autores aplican el algoritmo EM, que puede verse en Dempster, Laird y Rubin (1977), al problema de hacer inferencias clásicas con los diseños de respuesta aleatorizada.

El objetivo del presente trabajo es aplicar el algoritmo EM al modelo bayesiano utilizado en el diseño de Warner, observando que la inferencia bayesiana aplicada al diseño de Warner se reduce, en cierta forma, a un problema de inferencia bayesiana usual.

Hemos estructurado el presente trabajo en un primer apartado en el que se expone el modelo de Warner, en el siguiente se resume la aproximación bayesiana, en el tercero se describe la aplicación del algoritmo EM a la aproximación bayesiana, y, por último, se ilustra la aplicación anterior por medio de ejemplos numéricos.

2. MODELO DE RESPUESTA ALEATORIZADA DE WARNER

El modelo de Warner supone una partición de la población en dos grupos no solapados y cuya unión es la población. El primer grupo lo forman los elementos de la población que poseen la característica A (grupo A), mientras que el segundo lo forman aquellos elementos que no poseen la característica A (grupo $\bar{A}$).

Así pues, se construyen dos declaraciones alternativas, como son:

(1) Pertenezco al grupo A.
(2) Pertenezco al grupo $\bar{A}$.

De esta forma, cada uno de los entrevistados selecciona aleatoriamente una de estas dos declaraciones, sin que el entrevistador conozca el resultado de la selección, y contesta en función de su situación real simplemente Si o No a la declaración que haya seleccionado.

Todo ello nos conduce a que si denominamos por p a la probabilidad de que el mecanismo de aleatorización selecciona la declaración "Pertenezco al grupo A", y 1−p "Pertenezco al grupo $\bar{A}$", esto es, construimos la variable aleatoria de tipo Bernouilli:

$$z_i \begin{cases} 1 \text{ si el i–ésimo elemento muestral selecciona (1).} \\ 0 \text{ si el i–ésimo elemento muestral selecciona (2).} \end{cases}$$

donde $P(z_i = 1) = p$
$p(z_i = 0) = 1 - p$
p es elegido por el experimentador, y sin pérdida de generalidad supondremos que $p > 0{,}5$.

el problema sea estimar la proporción de entrevistados que poseen la característica A en la población, que denotamos por π.

Ahora bien, si consideramos que extraemos de la población una muestra aleatoria simple con reemplazamiento (o sin reemplazamiento, siempre que la población sea suficientemente grande para considerarla infinita) de tamaño n, se tiene que cada observación muestral genera una variable aleatoria de tipo Bernouilli, es decir:

$$y_i \begin{cases} 1 \text{ si el i–ésimo elemento muestral responde Sí.} \\ 0 \text{ si el i–ésimo elemento muestral responde No.} \end{cases}$$

De esta forma, la probabilidad de que el i–ésimo elemento muestral responda "Sí", es:

$$P(y_i = 1) = P(y_i = 1/z_i = 1) \cdot P(z_i = 1) + P(y_i = 1/z_i = 0) \cdot P(z_i = 0) = \pi \cdot p + (1 - \pi) \cdot (1 - p) = \lambda$$

Bajo el supuesto que el número de respuestas afirmativas en la muestra es r, esto es:

$$r = \sum_{i=1}^{n} y_i$$

se tiene que la función de verosimilitud de la muestra, L, adopta la siguiente expresión:

$$L = \lambda^r \cdot (1 - \lambda)^{n-r} \text{ para } 1 - p \leq \lambda \leq p.$$

que en términos de π, es:

$$(3)\quad L = [(2 \cdot p - 1) \cdot \pi + (1 - p)]^r \cdot [p - (2 \cdot p - 1) \cdot \pi]^{n-r} \text{ para } 0 \leq \pi \leq 1.$$

El estimador maximoverosímil de λ:

$\hat{\lambda} = p$ para $r/n > p$

$= r/n$ para $1-p \leq r/n \leq p$

$= 1-p$ para $r/n < 1-p$

siendo entonces el estimador maximoverosímil de π:

$$\hat{\pi} = [\hat{\lambda} - (1 - p)] / (2p - 1)$$

La restricción de que λ pertenezca al intervalo [1−p , p] se impone para evitar la posibilidad de que la estimación de π sea negativa o mayor que 1.

3. APROXIMACION BAYESIANA AL MODELO DE WARNER

En el trabajo de Winkler y Franklin (1979), se propone como distribución inicial para el parámetro π, la distribución beta, cuya función de densidad viene dada por:

$f(\pi/\alpha,\beta) = [B(\alpha,\beta)]^{-1} \cdot \pi^{\alpha-1} \cdot (1-\pi)^{\beta-1}$

donde $B(\alpha,\beta) = \Gamma(\alpha) \cdot \Gamma(\beta) / \Gamma(\alpha+\beta)$

para $\alpha > 0$, y $\beta > 0$.

A partir de esta distribución inicial y la función de verosimilitud definida en (3), se obtiene, como consecuencia del teorema de Bayes, la siguiente distribución final para el parámetro π:

$f(\pi/r,n) \propto \pi^{\alpha-1} \cdot (1-\pi)^{\beta-1} \cdot [(2 \cdot p-1) \cdot \pi + (1-p)]^{r} \cdot [p-(2 \cdot p-1) \cdot \pi]^{n-r}$

para $\pi \in [0 , 1]$.

Ahora bien, aplicando el teorema de la probabilidad total, tenemos que:

$$(4)\quad f(\pi/r,n) = \sum_{t=o}^{n} f(t/r,n) \cdot f(\pi/t,n)$$

donde:
t, cantidad no observada, es el número de entrevistados que pertenecen al grupo A. Es decir, si construimos la variable aleatoria de tipo Bernouilli:

$$x_i \begin{cases} 1 \text{ si el i−ésimo elemento muestral pertenece al grupo A.} \\ 0 \text{ si el i−ésimo elemento muestral pertenece al grupo } \bar{A}. \end{cases}$$

$$\text{se tiene que: } t = \sum_{i=1}^{n} x_i$$

$f(\pi/t,n)$ es una función de densidad beta con parámetros $\alpha+t$ y $\beta+n-t$.
$f(t/r,n) \propto f(t/n,\alpha,\beta) \cdot f(r/t,n)$
donde:
$f(t/n,\alpha,\beta)$ es una distribución betabinomial cuya función de cuantía es:

$$\begin{bmatrix} n \\ t \end{bmatrix} \cdot B(\alpha+t,\beta+n-t)/B(\alpha,\beta)$$

para t = 0, 1, 2, ..., n.

$$f(r/t,n) = \sum_{j=\alpha}^{b} \begin{bmatrix} t \\ j \end{bmatrix} \cdot \begin{bmatrix} n-t \\ r-j \end{bmatrix} \cdot p^{n-t-r+2j} \cdot (1-p)^{t+r-2j}$$

donde a es el máximo de (0 , t−n+r) y b es el mínimo de (r , t).
para r = 0, 1, 2, ..., n.

Tal como señalan los autores en su trabajo, si t fuese conocido, la distribución final (4), se reduciría a una distribución beta con parámetros α+t y ß+n−t. Precisamente, porque t es desconocido, la distribución final (4) resulta ser una combinación convexa de distribuciones betas con parámetros α+t y ß+n−t para t = 0, 1, 2, ..., n.

4. APLICACION DEL ALGORITMO EM AL MODELO DE WARNER

En el trabajo de Bourke y Morán (1988), los autores introducen el algoritmo EM en los diseños de respuesta aleatorizada, para proporcionar un cálculo simple de una buena aproximación al estimador de máxima verosimilitud.

En el caso de un modelo de Bernouilli, los pasos para aplicar este algoritmo son:

I) Tomar un valor de π_1.

II) Imputar el valor desconocido t por $\hat{t}$, donde:

$$\hat{t} = E\,[t/r,\pi_1]$$

es decir, la esperanza de t condicionado a r y π_1.

III) Calcular la función de verosimilitud del parámetro π, dado el "dato" $\hat{t}$, es decir:

$$\hat{L} \propto \pi^{\hat{t}} \cdot (1-\pi)^{n-\hat{t}}$$

IV) Calcular el máximo de $\hat{L}$, que resulta ser:

$$\hat{\pi} = \hat{t}/n$$

V) Tomar $\pi_1 = \hat{\pi}$ y repetir todos los pasos anteriores.

El algoritmo EM nos asegura que la sucesión de los valores (π_1) converge al estimador de máxima verosimilitud de la función verosímil (3).

En el presente trabajo, el algoritmo EM nos permite obtener la moda de la distribución final (4). Los pasos son los siguientes:

A) Realizar los pasos I y II anteriores.

B) Calcular la distribución final bajo el supuesto de que el dato es $\hat{t}$. En nuestro caso se trata de una distribución beta de parámetros α+$\hat{t}$ y ß+n−$\hat{t}$.

C) Calcular la moda $\hat{\pi}_M$ de dicha distribución beta, es decir:

$$\hat{\pi}_M = (\alpha+\hat{t}-1)/(\alpha+ß+n-2)$$

D) Tomar $\pi_1 = \hat{\pi}_M$ y repetir los pasos anteriores.

La aplicación del algoritmo EM nos asegura la convergencia de la sucesión (π_1) a la moda de la distribución final (4).

Para calcular el valor de $\hat{t}$ basta obtener, en primer lugar, para un individuo que ha contestado "sí", es decir, y = 1, la probabilidad de que pertenezca al grupo A; y en segundo lugar, para un individuo que ha contestado "no", es decir, y = 0, la probabilidad de que pertenezca al grupo A. La primera de estas probabilidades, aplicando el teorema de Bayes, es:

$$P(x=1/y=1) = P(x=1) \cdot P(y=1/x=1)/P(y=1) = \pi_1 \cdot p/[\pi_1 \cdot p + (1-\pi_1)\cdot(1-p)] = p_1$$

y, análogamente, la segunda adopta la siguiente expresión:

$$P(x=1/y=0) = \pi_1 \cdot (1-p)/(1-[\pi_1 \cdot p + (1-\pi_1)\cdot(1-p)]) = p_2$$

A partir de estas probabilidades, se tiene que:

$$\hat{t} = E\,[t/r,\pi_1] = \sum_{i=1}^{n} E\,[x_i / y_i,\pi_1] = \sum_{i=1}^{n} P(x_i=1/y_i,\pi_1) = r\cdot P(x=1/y=1,\pi_1) + (n-r)\cdot P(x=1/y=0,\pi_1) = r\cdot p_1 + (n-r)\cdot p_2$$

El conocimiento de la moda de la distribución final (4) no es suficiente para tener una idea sobre la variabilidad de dicha distribución. En el presente trabajo vamos a dar un valor aproximado de la varianza, σ^2, de la distribución final (4), a partir de la siguiente fórmula (Lindley, 1965):

$$(5)\ \hat{\sigma}^2 = -\left[\frac{\delta}{\delta\,\pi^2}\ln\left[\hat{\lambda}^r\cdot(1-\hat{\lambda})^{n-r}\right]\right]^{-1} = \left[(2\cdot p-1)^2\cdot[(r/\hat{\lambda}^2) + (n-r)/(1-\hat{\lambda})^2]\right]^{-1}$$

siendo $\hat{\lambda} = \hat{\pi}\cdot p + (1-\hat{\pi})\cdot(1-p)$, y donde $\hat{\pi}$ es el valor obtenido por el algoritmo EM.

5. EJEMPLOS NUMERICOS

Para ilustrar los procedimientos desarrollados en el cuarto apartado, vamos a considerar, siguiendo el trabajo de Winkler y Franklin (1979), cinco distribuciones iniciales de tipo beta para π: α=1, ß=1; α=2, ß=4; α=2, ß=8; α=10, ß=20 y α=10, ß=40. Además, vamos a tomar cinco tamaños muestrales (n=15, 75, 150, 300 y 450) y siendo el cociente n/r constante igual a 0,4, y p=0,7.

TABLA

α,ß		r=6 n=15	r=30 n=75	r=60 n=150	r=120 n=300	r=180 n=450
1, 1	π_M	0,250	0,250	0,250	0,250	0,250
	$\sigma(\hat{\sigma})$	0,226(0,316)	0,130(0,141)	0,098(0,100)	0,070(0,071)	0,058(0,058)
2, 4	π_M	0,250	0,250	0,250	0,250	0,250
	$\sigma(\hat{\sigma})$	0,155(0,316)	0,110(0,141)	0,088(0,100)	0,066(0,071)	0,055(0,058)
2, 8	π_M	0,142	0,187	0,210	0,227	0,234
	$\sigma(\hat{\sigma})$	0,114(0,301)	0,096(0,138)	0,082(0,098)	0,064(0,070)	0,054(0,057)
10,20	π_M	0,316	0,302	0,291	0,279	0,273
	$\sigma(\hat{\sigma})$	0,081(0,322)	0,071(0,143)	0,063(0,101)	0,053(0,071)	0,047(0,058)
10,40	π_M	0,189	0,196	0,203	0,213	0,220
	$\sigma(\hat{\sigma})$	0,055(0,309)	0,053(0,138)	0,050(0,098)	0,046(0,070)	0,042(0,057)

En la tabla recogemos la moda de la distribución final del parámetro π, obtenida por el algoritmo EM. Hemos comprobado que en todos los ejemplos, el número de iteraciones del algoritmo es pequeño, logrando rápidamente una aproximación exacta, con cuatro decimales, al verdadero valor de la moda de la distribución final. Debido a esto, no se recoge el verdadero valor, para cada ejemplo, de la moda de la distribución final.

En esta tabla, también se recogen la desviación estándar de la distribución final, así como su aproximación (5), que en la tabla aparece entre paréntesis. Se observa que esta aproximación es excelente cuando el tamaño muestral es grande y la distribución inicial no domina a la función de verosimilitud. Comparando nuestra tabla con la que ofrecen Winkler y Franklin (1979), se observa que mientras estos autores logran una buena aproximación de la desviación estándar, y una no muy buena aproximación de la media de la distribución final, nosotros logramos precisamente lo contrario, puesto que damos una buena aproximación de la moda de la distribución final, y no tan buena de la desviación estándar.

BIBLIOGRAFIA

BOURKE, P. D. and MORAN, M. A. (1988). "Estimating Proportions From Randomized Response Data Using the EM Algorithm". *Journal of the American Statistical Association,* 83, 964-968.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). "Maximun Likelihood Estimation From Incomplete Data Via the EM Algorithm". *Journal of Royal Statistical Society, Ser B,* 39, 1-38.

HORVITZ, D. G., GREENBERG, B. G. and ABERNATHY, J. R. (1976). "Randomized Response: A Data-Gathering Device for Sensitive Questions". *International Statistical Review,* 44, 181-196.

LINDLEY, D. V. (1965). *"Introduction to probability and Statistics from a Bayesian Viewpoint". Parte 2. Inference.* Cambridge University Press.

WARNER, S. L. (1965). "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". *Journal of the American Statistical Association,* 60, 63-69.

WINKLER, R. L. and FRANKLIN, L. A. (1979). "Warner's Randomized Response Model: A Bayesian Approach". *Journal of the American Statistical Association,* 74, 207-214.