INFERENCIA EN POBLACIONES DONDE SE DESCONOCE LA CANTIDAD DE INFORMACIÓN PERDIDA

G. Yagüe-Utrilla, C. Moreno, L.A. García-Cortés, J. Altarriba.
Unidad de Genética Cuantitativa y Mejora Animal
Facultad de Veterinaria, Universidad de Zaragoza
C/ Miguel Servet, 177 Zaragoza 50013

Resumen: Se estudia la realización de inferencias en poblaciones con pérdida de información, dependiente de un proceso selectivo, en los casos en que se desconoce la magnitud de esa pérdida. Se han aplicado dos procedimientos: por un lado se modelizó directamente este proceso a través de la utilización del modelo truncado Tobit. Por otro, se analizaron los efectos de la inclusión del número de datos faltantes como una variable del modelo.

Introducción

Con el objeto de realizar inferencias correctas en una población sometida a un proceso selectivo que provoca una pérdida de información, suele requerirse la inclusión de dicho proceso en el modelo de estimación. Según la teoría de datos perdidos, "missing data theory" (Rubin, 1976) una forma flexible de abordar este problema consiste en la introducción del proceso de pérdida de datos en la función de verosimilitud, idea que ya ha sido aplicada en el campo de la mejora genética animal por Im et al., (1989).

Bajo la teoría bayesiana, y en el supuesto de que la cantidad de información faltante sea conocida, es posible conseguir estimadores adecuados en estas poblaciones mediante el muestreo de Gibbs (Geman y Geman, 1984). Además, la inclusión en el modelo del vector de datos observados y perdidos como una variable aumentada (Tanner, 1993) facilita la implementación del muestreo de Gibbs. Esto se debe a que las condicionales que se obtienen con este procedimiento tienen una forma distribucional conocida (Gelfand, 1992; Chib, 1992; Sorensen et al., 1998).

Sin embargo, la resolución de situaciones en las que sólo se dispone de los datos observados no es inmediata. Nuestro objetivo es realizar inferencias correctas bajo este supuesto, presentando a continuación las dos estrategias con las que hemos abordado este problema.

Descripción del modelo.

Se utilizó un modelo animal: $y_i^* = \mu + a_i + e_i$, donde y_i^* son los datos en el caso de información completa, lo que implícitamente incluye la información genealógica. Consecuentemente, μ es la media general; a_i y e_i son, respectivamente, el valor genético y el residuo asignados al individuo i. Por último, las distribuciones de las dos variables aleatorias son $\mathbf{a} \sim N(0, \mathbf{A}\sigma_a^2)$ y $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. A efectos de simplificar el análisis consideramos únicamente dos generaciones: generación parental (G_0) y generación filial (G_1). Sobre dicha población actúa un proceso de pérdida de información que queda reflejado en la variable $\mathbf{z}$, de tal forma que:

$$z_i = \begin{cases} y_i^* \;\; si \;\; y_i^* \in G_0 \\ y_i^* \;\; si \;\; y_i^* \in G_1 \;\; , y_i^* \geq t \end{cases}$$

$$z_j = miss \;\; si \;\; y_i^* \in G_1, \;\; y_i^* < t$$

Analizamos la realización de inferencias en esta población bajo los siguientes procedimientos:

- El primer enfoque supone la modelización directa de la pérdida de información. El patrón de pérdida con el que tratamos es un truncamiento de los datos de la generación filial a partir de un umbral conocido t, por lo que esta modelización pasa por la aplicación del modelo Tobit (Amemiya, 1985).
- La alternativa a esta estrategia surge como consecuencia de la introducción del número de datos faltantes (nf) como una variable del modelo, lo que conduce a una distribución posterior de tipo mezcla (Robert, 1996).

Aplicación del modelo Tobit

Los datos observados junto con el patrón de pérdida de información componen nuestro conocimiento acerca de la población. Como ya se ha dicho, pensamos que sería posible modelizar directamente esta situación dado que conocemos el proceso de pérdida de información y que ésta se debe al truncamiento por la izquierda de los fenotipos de la generación filial a partir de un umbral conocido t.

En un supuesto de una distribución normal truncada por la izquierda, con una media μ y una varianza σ^2, la verosimilitud del modelo Tobit truncado (Amemiya, 1985) es:

$$L = \prod_{i=1,n} \sigma^{-1} \frac{\exp\left[\frac{(y_i - \mu)^2}{\sigma^2}\right]}{1 - \Phi\left(\frac{t-\mu}{\sigma}\right)}.$$

Aplicando este modelo a nuestro patrón de pérdida de información la distribución posterior de las variables será:

$$p\left(\mu, \mathbf{a}, \sigma_a^2, \sigma_e^2 \middle| \mathbf{z}_{obs}\right) \propto p\left(\mathbf{z}_{obs} \middle| \mu, \mathbf{a}, \sigma_a^2, \sigma_e^2\right) p\left(\mathbf{a} \middle| \mathbf{A}\sigma_a^2\right).$$

Donde $\mathbf{z}_{obs}$ se refiere a la información registrada. Esta expresión se divide en:

$$p\left(\mathbf{z}_{obs} \middle| \mu, \mathbf{a}, \sigma_a^2, \sigma_e^2\right) \propto \prod_{i \in G0} \left\{ \sigma_e^{-1/2} \exp\left[\frac{(z_i - \mu - a_i)^2}{-2\sigma_e^2}\right] \right\} \prod_{i \in G1} \left\{ \frac{\sigma_e^{-1/2} \exp\left[\frac{(z_i - \mu - a_i)^2}{-2\sigma_e^2}\right]}{1 - \Phi\left(\frac{t - \mu - a_i}{\sigma_e}\right)} \right\},$$

y $\quad p\left(\mathbf{a} \middle| \sigma_a^2, \mathbf{A}\right) \propto \sigma_a^{-n/2} \exp\left(\frac{\mathbf{a}'\mathbf{A}\mathbf{a}}{-2\sigma_a^2}\right).$

Se decide llevar a cabo un muestreo de Gibbs para realizar inferencias sobre dicha población. Sin embargo, las condicionales obtenidas no se corresponden con distribuciones estándar, y por esta razón se ha utilizado un Metrópolis (Tanner, 1993) dentro de Gibbs para realizar el muestreo de valores.

A continuación se muestran los resultados de la aplicación de este modelo con distintos grados de pérdida de información. Se han considerado tres situaciones: pérdidas del 20%, 50%, 80%. Los parámetros verdaderos fueron en todas las ocasiones de: $m=100$; $s_a^2=10$; $s_e^2=10$.

Tabla 1: Resultados de las estimaciones de m, s_a^2 *y* s_e^2 *en tres casos de pérdida de información. Se muestra la media de las esperanzas obtenidas en las 15 repeticiones realizadas. Entre paréntesis se presentan la desviación típica de dichas esperanzas.*

	Valores verdaderos	pérdida=20%	pérdida=50%	pérdida=80%
m	100	100.122 (0.242)	100.121 (0.181)	100.049 (0.140)
s_a^2	10	10.630 (1.345)	6.339 (2.445)	4.154 (1.808)
s_e^2	10	10.481 (1.400)	15.211 (2.163)	16.528 (1.885)

Se destaca que con pérdidas leves de información este modelo ofrece resultados acordes con los valores simulados, sin embargo, no es capaz de ello en situaciones más restrictivas. Es necesario señalar que la condicional para los valores genéticos de los individuos de la generación parental no se ve modificada con respecto a la obtenida en el caso en el que no se tenga en cuenta la información perdida. Esto provoca que los hijos conocidos sean los únicos que aparezcan en dicha condicional, lo que parece indicar que de alguna manera no se tiene en cuenta toda la información perdida.

Utilización de una verosimilitud mezcla .

Con el propósito de poner de manifiesto las dificultades que entraña este procedimiento, compararemos el caso de número de datos faltantes (nf) conocido (caso I), donde la dimensión del vector $\mathbf{z}$ es n, con el caso de nf desconocido (caso II) en el que no sabemos ni siquiera la dimensión de este vector, sino sólo la parte correspondiente a los datos observados.

a) Caso I:

En este caso tratamos con un modelo censurado en vez de truncado (Gauderman, 1994). Con el fin de simplificar el análisis aplicamos la técnica de la variable aumentada (Tanner, 1987), mediante la introducción de $\mathbf{y}^*$ en el modelo. Con ello se consiguen unas distribuciones condicionales de las que es sencillo muestrear (Sorensen, 1998). La distribución posterior de los parámetros en el caso I es entonces:

$$p(\mu,\mathbf{a},\sigma_a^2,\sigma_e^2,\mathbf{y}^*|\mathbf{z},nf) \propto p(\mathbf{z}|\mathbf{y}^*)p(\mathbf{y}^*|\mu,\mathbf{a},\sigma_a^2,\sigma_e^2)\sum_{k=1}^{m}\left[w_k\, p_k(\mathbf{a}\,|\,\mathbf{A}_k\sigma_a^2)\right].$$

Cabe destacar que debido a la pérdida producida debemos tener en cuenta todas las genealogías posibles, lo que da como resultado una distribución tipo mezcla "***mixture***" expresada en el sumatorio de dichas genealogías (Robert, 1996). Siendo m el número total de genalogías distintas posibles, de tal forma que $w_k = 1/m$, dado que todas ellas son igualmente probables ***a priori***.

Por otra parte, $p(y_i^*|\mu,a_i,\sigma_a^2,\sigma_e^2,z_i = miss) \propto \Phi\left(\frac{t-\mu-a_i}{\sigma_e}\right)$, es decir, una normal truncada, de la que es sencillo muestrear. El resto de las condicionales tiene formas distribucionales conocidas, y por ello, en esta situación, la inferencia se desarrolla satisfactoriamente (Sorensen, 1998).

b) Caso II:

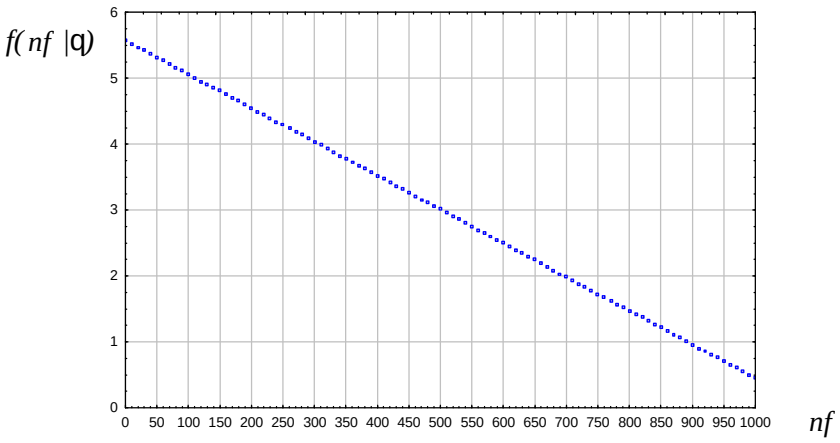
Sin embargo la distribución posterior se modifica cuando incluimos la cantidad de datos perdidos en el modelo en las ocasiones en que sólo se dispone de los datos registrados, dando lugar a:

$$p(\mu,\mathbf{a},\sigma_a^2,\sigma_e^2,\mathbf{y}^*,nf|\mathbf{z}) \propto \sum_{j=0}^{\infty}\left\{p(nf_j)p(\mathbf{z}|\mathbf{y}^*,nf_j)p(\mathbf{y}^*|\mu,\mathbf{a},\sigma_a^2,\sigma_e^2,nf_j)\sum_{k=1}^{m}\left[w_k\, p_k(\mathbf{a}\,|\,\mathbf{A}_k\sigma_a^2)\right]\right\}.$$

Se aprecia que nos encontramos con una distribución tipo mezcla que afecta incluso a la verosimilitud de los datos, ya que la dimensión de $\mathbf{z}$ depende de nf. El análisis de este modelo debería seguir básicamente los mismos pasos que el anterior, sin embargo, la inclusión de nf en el modelo presenta problemas.

Una de las principales dificultades con las que nos hemos encontrado es la asignación de una distribución ***a priori*** para el número de datos faltantes. Si consideramos una ***a priori*** no informativa, la condicional resultante de nf no tiene máximo (Gráfica 1), por lo que la cantidad perdida estimada siempre es igual a 0. Esto conduce a ignorar la pérdida de información, y como consecuencia, realizar inferencias incorrectas.

Gráfica 1: Forma distribucional de la condicional de datos faltantes (con la a priori no informativa).



En casos sencillos se ha comprobado que la utilización de una *a priori* para nf a partir de la cuál se obtenga una distribución condicional binomial permite la obtención de resultados aceptables.

Por otro lado, el continuo cambio de dimensión del modelo que produce la modificación de nf induce a contemplar a las estrategias de elección de modelos como otra alternativa para abordar este tipo de problemas.

Referencias

Amemiya T., Tobit models: a survey, Journal of Econometrics 24 (1984) 3,61.

Chib S., Bayes inference in the Tobit censored regression model, Journal of Econometrics 51 (1992) 79-99.

Gauderman W.J., Thomas D.C., Censored survival models for genetic epidemiology: a Gibbs sampling approach, Genet. Epidemiol. 11 (1994) 171-188.

Gelfand A.E., Smith F.M., Lee T., Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling, J. Am. Stat. Assoc. 87 (1992) 523-532.

Geman S., Geman D., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Analysis and Machine Intelligence 6 (1984) 721-741.

Im S., Fernando R.L., Gianola D., Likelihood inferences in animal breeding under selection: a missing-data theory view point, Genet. Sel. Evol. 21 (1989) 399-414.

Robert C. P., Mixtures of distributions: inference and estimation, in Gilks W.R., Richardson S., Spiegelhalter D.J. (Eds.), Markov Chain Monte Carlo in practice, Chapman and Hall. London, 1996, pp. 419-440.

Rubin D. B., Inference and missing data, Biometrika 63 (1976) 581-592.

Sorensen D.A., Gianola D., Korsgaard I., Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications, Acta Agric. Scand., Sect. A, Animal Sci. 48 (1998) 222-229.

Tanner M.A., Tools for statistical inference, Springer-Verlag. Berlin, 1993.

Tanner M.A., Wong W.H., The calculation of posterior distributions by data augmentation, J. Am. Stat. Assoc. 82 (1987) 528-540.