RECUPERACIÓN DEL SESGO DE SELECCIÓN EN LA ESTIMACIÓN DE COMPONENTES DE VARIANZA

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

Esta comunicación sugiere la aplicación de la técnica de variable aumentada para la estimación de componentes de varianza insesgados de una población en la que se ha realizado un proceso selectivo con pérdida de información. Básicamente, en el ámbito de la mejora genética animal este problema ha sido abordado bajo dos perspectivas. En primer lugar, Henderson propone una modificación de las ecuaciones del modelo mixto basada en resultados desarrollados por Pearson, mientras que Im et al.(1989), siguiendo las directrices de Little y Rubin (1987), incorporan la información del proceso selectivo en sus funciones de verosimilitud. En este trabajo, se concreta una resolución del segundo procedimiento basada en la aplicación de un algoritmo de Gibbs en el que se ha introducido una variable aumentada.

Introducción

La información disponible en mejora genética animal procede en la mayoría de las ocasiones de poblaciones sometidas a selección. En situaciones donde toda la información en la que se ha basado la selección es conocida, se han definido métodos para obtener estimadores insesgados de los parámetros genéticos de estas poblaciones. Sin embargo, cuando el proceso selectivo se ha realizado a partir de información no disponible, las estimas obtenidas son sesgadas si se ignora dicha información (Rubin, 1976).

Henderson (1990) destaca que las medias y varianzas de variables aleatorias sometidas a selección a partir de fenotipos no disponibles para el análisis son distintas de las obtenidas bajo las asunciones usuales del modelo lineal mixto. Por lo tanto, la resolución de las ecuaciones de dicho modelo puede generar estimadores y predictores sesgados. A pesar de ello, el mismo autor, fundamentándose en el análisis realizado por Pearson sobre variables condicionadas a funciones de selección lineales, señala que estas ecuaciones ofrecen resultados insesgados cuando se incorporan las variables utilizadas en la selección. En este sentido y a partir del análisis anterior, Schaeffer (1987) propone un algoritmo REML para la estimación de componentes de varianza.

Una alternativa al enfoque Pearsoniano consiste en realizar inferencias basadas en funciones de verosimilitud en las que se introduzca el proceso de pérdida de datos que ha tenido lugar en las poblaciones seleccionadas. Esta idea, desarrollada por Little y Rubin (1987), ha sido aplicada en el campo de la mejora genética animal por Im et al. (1989). Estos autores destacan la flexibilidad de esta opción frente a la anterior y definen las funciones de verosimilitud específicas de una serie de procesos selectivos en los que se produce pérdida de información, señalando las situaciones en las que es imprescindible considerar dicha pérdida. Sin

embargo, la complejidad de las mismas dificulta su uso en la resolución de problemas de estimación.

En esta línea, el objetivo de este trabajo es ofrecer un método que permita una fácil implementación de este tipo de funciones. Para ello introducimos en el modelo el vector de valores observados y perdidos como una variable aumentada, siendo este modelo posteriormente resuelto mediante muestreo de Gibbs. Este objetivo forma parte de un proyecto de mayor magnitud en el que se pretende analizar la recuperación del sesgo de selección en la estimación de componentes, mediante la información aprotada por genes marcadores, en situaciones informativas más drásticas que las aquí presentadas.

Modelo

Se asume un modelo animal donde y_{ij} es el fenotipo del individuo i en la generación j, μ es la media de la población, a_{ij} es el valor aditivo del correspondiente individuo y e_{ij} es el residuo.

$$y_{ij} = \mu + a_{ij} + e_{ij}$$

Se han simulado dos generaciones con 3000 individuos por generación. La generación parental se sometió a selección masal, siendo elegidos como reproductores el 20%.

El proceso de pérdida de datos afecta a los individuos de la generación filial. Los fenotipos de los individuos que no alcanzan el punto de truncamiento *t*, que se asume conocido, no son registrados. De la misma forma, el individuo no es seleccionado. Siguiendo el desarrollo de Gelfand et al. (1992), sea *z* la variable definida según:

$$z_{ij} = mis \;\; \forall \; y_{ij} \leq t$$
$$z_{ij} = y_{ij} \;\; \forall \; y_{ij} > t$$

donde y_{ij} es una variable latente (variable aumentada) incluida en el modelo. De tal forma que si y_{ij} es menor que el punto de truncamiento *t* el valor fenotípico es desconocido (z_{ij}=*mis*).

A partir de un enfoque bayesiano, se asumen distribuciones a priori no informativas para las variables μ, σ_a^2 y σ_e^2. Igualmente se asume que los valores aditivos se distribuyen según una $N(0,A\sigma_a^2)$. Por lo tanto, la distribución posterior conjunta de los parámetros es:

$$p\left\langle \mu, y, a, \sigma_a^2, \sigma_e^2 \middle| z, t \right\rangle \propto p\left\langle z \middle| \mu, y, a, \sigma_a^2, \sigma_e^2, t \right\rangle \, p\left\langle y \middle| \mu, a, \sigma_a^2 \sigma_e^2, z, t \right\rangle \, p\left\langle a \middle| \sigma_a^2 \right\rangle$$

donde $p(z|\mu,y,a,\sigma_a^2,\sigma_e^2,t)$ es una distribución degenerada, ya que el conocimiento de *y* implica el conocimiento exacto de *z*.

Este modelo se resuelve mediante muestreo de Gibbs, siendo las distribuciones condicionales de todas sus variables, excepto la correspondiente a la variable aumentada, idénticas a las obtenidas en situaciones sin pérdida de información. Por otra parte, la condicional de la variable latente es una normal truncada restringida al intervalo $y_{ij}<t$.

Resultados y discusión

En la tabla que se adjunta se muestran los resultados de la estimación de los componentes de varianza a partir del modelo anteriormente descrito. Estos resultados han sido obtenidos a partir de datos generados mediante simulación. Se han estudiado dos supuestos: en el primero no se considera la pérdida de información (S1), mientras que en el segundo se ha introducido este proceso de pérdida (S2). Estos dos supuestos han sido estudiados en cuatro situaciones donde el número de datos faltantes es creciente. Cada caso ha sido resuelto mediante un algoritmo de muestreo de Gibbs donde el periodo de quemado ha sido de 200 iteraciones y el número de puntos escogido de 2800, repitiéndose cada caso en 50 ocasiones. Así mismo, el valor simulado de σ_a^2 y σ_e^2 fue en todos los casos de 20 y 80, respectivamente.

% inf. perdida	S1		S2	
	σ_a^2	σ_e^2	σ_a^2	σ_e^2
0%	20,161 ± 1,753	79,873 ± 2,157	20,161 ± 1,753	79,873 ± 2,157
10%	29,872 ± 2,040	59,237 ± 2,220	20,205 ± 2,137	80,110 ± 2,407
30%	48,294 ± 2,051	36,612 ± 2,011	19,871 ± 1,911	80,282 ± 2,597
50%	67,426 ± 2,244	19,358 ± 1,976	20,332 ± 2,039	80,168 ± 3,006

Tal como muestra la tabla, en el primer supuesto el sesgo aumenta a medida que se incrementa el porcentaje de información perdida. De tal forma que se obtiene una sobreestimación de la varianza genética aditiva y una infraestimación de la varianza residual. Sin embargo, cuando se tiene en cuenta el efecto del proceso selectivo (segundo supuesto), introduciéndolo en la función de verosimilitud, el sesgo se corrige. Este resultado es totalmente acorde con las conclusiones de Im et al (1989).

Referencias bibliográficas

GELFAND, A.E., SMITH, F.M., LEE, T. 1992. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J.Am. Stat. Assoc. **87**:523-532

HENDERSON (1990) Accounting for selection and mating biases in genetic evaluations. En: GIANOLA, D.,HAMOND, K. (eds.) Advances in statitical methods for genetic improvement of livestock. Springer-Verlag.

IM, S.,FERNANDO R. L., GIANOLA, D. 1989. Likelihood inferences in animal breeding under selection: a missing data theory view point. Genet .Sel. Evol:**21**:399-414.

LITLLE R.J.A ,RUBIN D.B. 1987, Statitical analysis with missing data. Wiley, New York.

SCHAEFFER, L.R. 1987. Estimation of variance components under a selection model. J. Dayry. Sci. **70**: 661-671.