Métodos de comparación de modelos: un estudio de simulación.

Luis Varona.
Area de Producció Animal. Centre UdL-IRTA. 25198. Lleida.

Introducción

En mejora genética animal, la comparación de modelos es una herramienta clave tanto en los procedimientos de valoración de reproductores, como en las técnicas estadísticas ligadas a la genética molecular (detección de QTL, estudios de asociación o análisis de datos de expresión génica).

La aproximación clásica al contraste de hipótesis en modelos jerarquizados se resuelve mediante el contraste del cociente de verosimilitudes (LRT). Asintóticamente, y siempre que la hipótesis nula sea cierta, -2 ln LRT se distribuye como una Chi-cuadrado con tantos grados de libertad como diferencia exista entre el número de parámetros de ambos modelos.

Desde una perspectiva bayesiana, la comparación entre modelos se resuelve mediante el cálculo de la probabilidad posterior de cada modelo dados los datos -$p(M_i|y)$-. Este procedimiento no tiene ningún requisito de jerarquía entre los modelos candidatos, y se presenta como un criterio claro y mesurable.

En general:

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)}\frac{p(M_1)}{p(M_2)}$$

donde $\frac{p(M_1|y)}{p(M_2|y)}$ es el cociente entre las probabilidades posteriores de los modelos candidatos, $\frac{p(y|M_1)}{p(y|M_2)}$ es el factor de Bayes entre ambos modelos, y $\frac{p(M_1)}{p(M_2)}$ es el cociente entre las probabilidades a priori.

Por lo tanto, el Factor de Bayes (BF) puede considerarse el criterio de comparación entre modelos si se asume igual probabilidad a-priori para los dos modelos candidatos. Sin embargo, el cálculo de la probabilidad marginal de los datos $p(y|M_i)$, imprescindible para el Factor de Bayes, implica la integración a lo largo de todos los parámetros del modelo:

$$p(y|M_1) = \int p(y|M_1, q_1)p(q_1|M_1)dq_1$$

Esta integración no es sencilla en modelos complejos, y, además, es claramente dependiente de estas distribuciones a-priori (Sorensen y Gianola, 2002), que en muchas ocasiones están definida de manera vaga.

Como solución a este problema se han presentado múltiples alternativas, tanto analíticas como numéricas. En este trabajo se pretende comparar algunos de estos procedimientos en el marco de un modelo padre de componentes de varianza.

SIMULACIÓN

Se han simulado poblaciones consistentes en 10, 30 y 50 machos con 10 hijos con registros fenotípicos (100, 300 y 500 registros fenotípicos, respectivamente). Cada una de estas poblaciones se simuló con heredabilidades de 0.00, 0.10, 0.20, 0.30, 0.40 y 0.50. Para cada combinación de tamaño de población y heredabilidad se generaron 1000 réplicas. La media y la varianza poblacional se asumieron conocidas y fijadas en 100.

MÉTODOS

Los modelos a comparar fueron, en todos los casos de simulación, un modelo con $h^2=0$ – M_1- y un modelo con $h^2>0$ –M_2-. Se han utilizados los siguientes procedimientos:

1. Contraste de cociente de verosimilitudes –LRT - (Peña, 1995).

$$-2\ln\frac{L\left(\mathbf{y}\middle|\hat{h}^2\right)}{L\left(\mathbf{y}\middle|h^2=0\right)}\sim c_1^2$$

   Se han calculado el porcentaje de réplicas que superan los umbrales al 5 y al 1%.

2. Factor de Bayes (Aproximación para modelos jerárquicos) –HMBF- (Garcia-Cortes et al., 2001;Varona et al., 2001)

$$BF=\frac{p\left(h^2=0\middle|M_2\right)}{p\left(h^2=0\middle|y,M_2\right)}$$

3. Factor de Bayes (Aproximación la media armónica) – NRBF-, (Newton y Raftery, 1994)

$$NRBF=\frac{p\left(y\middle|M_1\right)}{p\left(y\middle|M_2\right)}$$

$$p(y|M_2)\cong\frac{n}{\sum_{i=1}^{n}\frac{1}{p(y|h_i^2,M_2)}}$$

   donde h^2_i son muestras obtenidas al azar de la distribución posterior $p(h^2|y,M_2)$

4. Factor de Bayes Posterior - PBF-(Aitkin, 1991)

$$PBF=\frac{p_1(y|M_1)}{\int p_2(y|h^2,M_2)\,p_2(h^2|y,M_2)dq}$$

5. Akaike Information Criterium –AIC-(Akaike, 1973). El modelo con un AIC inferior es seleccionado.

$$AIC_2 = -2\log\left\{p\left(y|\hat{h}^2\right)\right\}+2$$

6. Bayesian Information Criterium – BIC- (Swartz, 1978). Se selecciona el modelo con un BIC inferior.

$$BIC_2 = -2\log\left\{p\left(y|\hat{h}^2\right)\right\}+\log(n)$$

donde *n* es el numero de observaciones

7. Deviance Information Criterium –DIC- (Spiegelhalter et al., 2002)

$$DIC = \overline{D}\left(h_i^2\right)+p_d$$

donde $\overline{D}(M)$ es la devianza ($-2\log\left\{p\left(y|h_i^2\right)\right\}$) media de las observaciones de la distribución posterior (h^2_i) obtenida mediante MCMC, y

$$p_d = \overline{D}\left(h_i^2\right)-D(\hat{h}^2)$$

es la diferencia entre la devianza media de las observaciones de la distribución posterior y devianza observada en la media posterior de h^2. Por otra parte, p_d se puede entender como el número efectivo de parámetros del modelo.

RESULTADOS Y DISCUSIÓN

En las Tabla 1, 2 y 3 se presentan los resultados de la simulación. En los casos donde de simuló una heredabilidad nula, el LRT produjo una cantidad de falsos positivos similar al error de tipo I asumido, y, como era esperable, independiente del tamaño de muestra.

Por el contrario, en los procedimientos asociados con el Factor de Bayes (HMBF y NRBF), el porcentaje de falsos positivos se redujo a medida que el tamaño de población aumento. Los otros procedimientos aproximados (AIC, BIC, DIC y PBF) mantuvieron un porcentaje similar de falsos positivos, aún cuando se incremento este tamaño. El BIC se mostró como el procedimiento más conservador y el PBF como el más arriesgado a la hora de elegir el modelo complejo.

En cuanto a la potencia empírica de los contrastes, el PBF fue el procedimiento que selecciono en mayor número de ocasiones el modelo complejo, mientras que el BIC fue, de nuevo, el más conservador. El DIC y el AIC produjeron resultados simulares a los estimadores del factor de Bayes (HMBF y NRBF). Todos los procedimientos con excepción del BIC obtuvieron resultados superiores a los procedimientos basados directamente en el cociente de verosimilitudes (LRT).

Tabla 1. Porcentaje de réplicas que seleccionan el modelo complejo (h2>0) en el primer caso de simulación (n=100)

h^2	LRT 5%	LRT 1%	HMBF	NRBF	AIC	BIC	DIC	PBF
0.00	4.5	0.9	14.3	14.3	7.8	1.2	13.0	20.3
0.10	7.1	2.3	27.9	27.8	17.2	4.9	26.7	35.8
0.20	15.4	5.7	45.8	45.8	32.5	11.3	43.1	53.8
0.30	28.4	14.6	61.4	61.4	45.7	23.1	59.4	69.4
0.40	38.1	21.9	69.6	69.4	55.4	32.5	67.3	76.4
0.50	48.5	29.6	79.2	79.3	68.0	41.3	77.2	83.3

Tabla 2. Porcentaje de réplicas que seleccionan el modelo complejo (h2>0) en el segundo caso de simulación (n=300)

h^2	LRT 5%	LRT 1%	HMBF	NRBF	AIC	BIC	DIC	PBF
0.00	5.4	1.2	6.9	6.9	6.4	0.8	8.6	19.0
0.10	13.8	3.7	25.9	25.8	26.8	5.7	34.1	48.1
0.20	38.2	20.7	56.4	56.4	56.3	25.8	61.2	75.0
0.30	59.4	38.9	76.8	77.5	77.7	44.5	83.5	92.2
0.40	80.1	62.9	89.6	90.1	89.6	68.5	92.2	96.4
0.50	91.8	81.0	96.6	96.5	96.5	84.9	97.4	99.4

Tabla 3. Porcentaje de réplicas que seleccionan el modelo complejo (h2>0) en el tercer caso de simulación (n=500).

h^2	LRT 5%	LRT 1%	HMBF	NRBF	AIC	BIC	DIC	PBF
0.00	4.8	1.1	4.5	6.1	7.1	0.4	9.0	19.2
0.10	19.6	8.3	28.1	30.8	34.4	9.4	40.3	60.4
0.20	54.4	33.3	65.9	67.3	71.7	35.7	76.3	87.4
0.30	79.7	61.4	87.0	88.3	89.8	63.9	92.3	96.7
0.40	94.8	86.3	96.6	96.9	97.5	88.2	98.4	99.2
0.50	98.3	95.3	99.2	99.2	99.3	96.1	99.5	100.

Referencias

Aitkin, M. 1991. Posterior Bayes Factor (with discussion). J. R. Statist. Soc. B. 53:111-142.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle, in Proceedings 2nd international symposium information theory, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267-281.

García-Cortés, L. A., Cabrillo, C., Moreno, C., Varona, L. 2001. Hipothesis testing for the genetic background on quantitative traits. Genet. Sel. Evol. 33:3-16.

Kass, R. E., Raftery A. E. Bayes Factors. J. A. S. A. 90:773-795.

Newton, M. A., Raftery, A. E. 1994. Approximate Bayesian inference with the weighted likelihood boostrap. J. Royal Statistical Society B 56:3-48.

Schwartz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461-464.

Sorensen, D., Gianola, D. 2002. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. Springer-Verlag.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van der Linde, A. 2002. Bayesian measures of model complexity and fit. J. Royal Statistical Society B 64:583-639.

Varona, L. García-Cortés, L. A., Pérez-Enciso, M. 2001. Bayes Factor for detection of Quantitative Trait Loci. Genet. Sel. Evol. 33:133-152.