COMPARACIÓN DE MODELOS DE REGRESIÓN ALEATORIA APLICADOS A DATOS LONGITUDINALES DE PRODUCCIÓN DE LECHE UTLIZANDO FACTORES DE BAYES

P. López-Romero, Rekaya, R[(1)], M.J. Carabaño
Departamento de Mejora Genética y Biotecnología. INIA
[(1)] Department of Animal Sciences, University of Wisconsin, Madison, WI - USA

Introducción

En estudios previos (López-Romero y Carabaño, 2000) se compararon modelos de regresión aleatoria (MRA) que mantenían constante el componente residual a lo largo de la lactación. Se utilizó REML para la estima de componentes de varianza y criterios clásicos para la comparación de modelos. En este trabajo, se plantea la comparación de MRA con varianza residual heterogénea dentro de un contexto bayesiano, utilizando Factores de Bayes (FB) para establecer los criterios de comparación.

Material y Métodos

Se han utilizado PDC procedentes de primeras lactaciones completas (11 controles) registradas en Navarra durante los años 1988 -1998. Los datos fueron seleccionados entre los 5 y 335 días de la lactación, exigiéndose una edad al parto entre los 18 y 40 meses, un intervalo entre el parto y el primer control entre 5 y 67 días y un intervalo entre controles sucesivos entre los 26 y 67 días. Tras este proceso se originó un archivo con 47.982 PDC pertenecientes a 4769 lactaciones, con una producción media de 25,73 ± 7,09 kg producidos en una duración media de 321,42 ± 8,25 días. El número de animales en el archivo de genealogía fue 10.068.

Se han *estimado parámetros genéticos para PDC* utilizando MRA con varianza residual (VR) heterogénea en función de 3 intervalos arbitrarios a lo largo de la lactación definidos entre los días: 5-75, 76-265 y 266-335. Basándonos en un estudio preliminar (López-Romero y Carabaño, 2000), sólo hemos analizado 6 MRA cuya ecuación general es:

$$y_{ijkl} = RFC_i + EE_j + \sum_{m=1}^{4} \beta_{jm} X_{klm} + u_k + \sum_{m=1}^{r} \alpha_{km} Z_{1,klm} + p_k + \sum_{m=1}^{s} \omega_{km} Z_{2,klm} + e_{ijkl} \qquad (1)$$

donde RFC_i es el grupo de comparación rebaño-fecha de control *i*, EE_j es el factor edad-época de parto *j*, β_{jm} es el coeficiente de regresión fijo *m* anidado a EE_j que se asocia a la covariable X_{klm} definida por la función lineal de Ali-Schaeffer (1987). u_k es el término independiente (TI) del efecto aditivo del animal *k*, p_k es el TI del efecto ambiental permanente de la vaca *k*, α_{km} es el coeficiente *m* de regresión aleatorio anidado a u_k, ω_{km} es el coeficiente de regresión aleatorio *m* anidado a p_k, y $Z_{1,klm}$ y $Z_{2,klm}$ son covariables dependientes del tiempo, definidas por funciones lineales. Los MRA utilizados difieren entre sí únicamente en las funciones que definen estas covariables. Se utilizaron las funciones lactacionales de Ali-Schaeffer (1987) para el modelo MRA-A y Wilmink (1987) para el modelo MRA-W, y Polinomios Ortogonales de Legendre (Kirkpatrick y Heckman, 1989). Se utilizaron los 4 modelos de Legendre siguientes: $L_a(3) + L_p(3) = A3$; $L_a(3) + L_p(5) = C3$; $L_a(5) + L_p(5) = C5$; $L_a(3) + L_p(6) = D$, donde en $L_a(n)$ y $L_p(m)$, *n* y *m* representan el orden de ajuste sobre la parte aditiva y permanente, respectivamente.

La *distribución condicional de los datos dados los parámetros* de los modelos especificados en la ecuación general (1) y las *distribuciones a priori* asumidas son las siguientes:

$$\mathbf{y/b,u,p,}\Sigma_u,\Sigma_p,\sigma^2_{e1},\sigma^2_{e2},\sigma^2_{e3} \sim \mathbf{NMV(Xb + Z_x u + W_x p\ ,\ R)} \text{ con } \mathbf{R} = \text{diag}\{\sigma^2_{ei\ (i=1,\ldots,3)}\}$$

$$\mathbf{b} \sim \mathbf{NMV(0,I}\sigma^2_b\mathbf{)} \qquad \text{con } \sigma^2_b = 10^6$$

$$\mathbf{u/}\Sigma_u, \mathbf{A} \sim \mathbf{NMV(0,G)} \qquad \text{con } \mathbf{G} = \Sigma_u \otimes \mathbf{A}$$

$$\mathbf{p/}\Sigma_p, \sim \mathbf{NMV(0,P)} \qquad \text{con } \mathbf{P} = \Sigma_p \otimes \mathbf{I}$$

$$\sigma^2_{ei}/\upsilon_{ei}, s^2_{ei} \sim \chi^{-2}\ (\upsilon_{ei},\ \upsilon_{ei}s^2_{ei}) \text{ con } i = 1,\ldots,3$$

$$\Sigma_u\ /\upsilon_u, \mathbf{S^2_u} \sim \mathbf{W^{-1}}\ (\upsilon_u,\ \upsilon_u \mathbf{S^2_u})$$

$$\Sigma_p\ /\upsilon_p, \mathbf{S^2_p} \sim \mathbf{W^{-1}}\ (\upsilon_p,\ \upsilon_p \mathbf{S^2_p})$$

donde υ_x y $\mathbf{S^2_x}$ son, respectivamente, los grados de libertad y el factor de escala de la distribución correspondiente. Se ha asignado un valor de 4 para υ_{ei} en todos los modelos, y un valor igual al número de orden de ajuste en el modelo para los parámetros υ_u y υ_p, reflejando un bajo conocimiento de la información a priori. Los valores para los escalares s^2_{ei} (i= 1,...,3), y para las matrices $\mathbf{S^2_u}$ y $\mathbf{S^2_p}$ fueron obtenidos de estimas REML previas (López-Romero, Carabaño, 2000), excepto para el MRA-A que se obtuvieron de Rekaya (1999). A partir de la *distribución posterior conjunta* se derivaron las

distribuciones condicionales completas necesarias para aplicar el algoritmo del muestreo de Gibbs siguiendo a Sorensen (1996). Las distribuciones condicionales obtenidas son:

$\mathbf{b/y,u,p,}\Sigma_u,\Sigma_p,\sigma^2_{el\ (l=1,...,3)} \sim \mathbf{NMV(}\hat{b}\mathbf{\ , (X'R^{-1}X + I}\sigma^{-2}_b\mathbf{\ )^{-1})}$ con $\hat{b} = \mathbf{(X'R^{-1}X + I}\sigma^{-2}_b\mathbf{)^{-1}X'R^{-1}(y - Z_xu - W_xp)}$

$\mathbf{u/y,b,p,}\Sigma_u,\Sigma_p,\sigma^2_{el\ (l=1,...,3)} \sim \mathbf{NMV(}\hat{u}\mathbf{\ ,(Z'R^{-1}Z + G^{-1})^{-1})}$, con $\hat{u} = \mathbf{(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y-Xb - W_xp)}$

$\mathbf{p/\ y,b,u,}\Sigma_u,\Sigma_p,\sigma^2_{el\ (l=1,...,3)} \sim \mathbf{NMV(}\hat{p}\mathbf{\ , (W'R^{-1}W + P^{-1})^{-1})}$, con $\hat{p} = \mathbf{(W'R^{-1}W + P^{-1})^{-1}\ W'R^{-1}(y- Xb - Z_x\ u)}$

$\sigma^2_{ei}\mathbf{/y,b,u,p,}\ \Sigma_u,\Sigma_p,\sigma^2_{ej(j\neq i)} \sim \chi^{-2}\ (N_i + \upsilon_{ei},\ \upsilon_{ei}S^2_{ei} + \mathbf{e_l'e_l})$ para i=1,...,3, siendo N_i el número de PDC contabilizados en el tramo de lactación asignado a la varianza σ^2_{ei}, y $\mathbf{e_l'e_l}$ la suma de residuos al cuadrado correspondiente al mismo tramo de lactación.

Σ_u **/ y,b,u,p,**$\Sigma_p,\sigma^2_{e1},\sigma^2_{e2},\sigma^2_{e3} \sim \mathbf{W^{-1}\ (}q + \upsilon_u,\ \mathbf{U'A^{-1}U} + \upsilon_u\mathbf{S^2_u)}$, donde q es el número de animales en genealogía, y $\mathbf{U}_{(q\ x\ r+1)}$ es la matriz que contiene los $r+1$ vectores columna $\mathbf{u}_i$, con las soluciones del coeficiente de regresión aleatorio i para los q animales: $\mathbf{U = (u_1, ..., u_i, ..., u_{r+1})}$

Σ_p **/ y,b,u,p,**$\Sigma_u,\sigma^2_{e1},\sigma^2_{e2},\sigma^2_{e3} \sim \mathbf{W^{-1}\ (}c + \upsilon_p,\ \mathbf{P'P} + \upsilon_p\mathbf{S^2_p)}$. donde c es el número de animales con dato y $\mathbf{P}$ es la matriz que contiene las soluciones para los $s+1$ coeficientes de regresión aleatorios del permanente.

La *comparación de modelos* se ha basado en la utilización de *Factores de Bayes*. Además, se estudió la forma en la que los diferentes modelos predecían las varianzas aditiva y permanente diaria y las correlaciones genéticas a lo largo de la lactación. *El Factor de Bayes (FB)* de un modelo M_1 frente a otro M_2, dados los datos $\mathbf{y}$, se define formalmente como la razón de los *"odds"* a posteriori entre los *"odds"* a priori.(Wasserman, 1997, Kass y Raftery, 1995, Raftery, 1998,Gelfand, 1998). (*Odd es la razón de probabilidad de un suceso frente a su complementario)*

$$FB = \frac{\pi(M_1 / y)/\pi(M_2 / y\ )}{\pi(M_1)/\pi(M_2)}$$,donde $\pi(M_i / y)$ y $\pi(M_i)$ son, respectivamente, las distribuciones posterior y a priori del modelo M_i.

El FB se puede interpretar como el peso relativo de la evidencia de un modelo frente al otro suministrada por los datos (Gelfand, 1998) y se calcula como el cociente entre las densidades marginales de los datos (o verosimilitudes marginales) para cada modelo (Gamerman, 1997, Kass y Raftery, 1995, Gelfand, 1998). Un aspecto importante en la comparación de modelos es la evaluación de estas densidades marginales, densidades que se calculan integrando sobre todo el espacio paramétrico (Kass y Raftery, 1995), y que no son más que las constantes de integración de las distribuciones posteriores (Gamerman, 1995) de los parámetros Φ dados los datos en cada uno de los modelos. La densidad marginal de los datos bajo el modelo M_i, es:

$$f(y / M_i) = \int_{\Phi} f(y / \Phi_i, M_i)\pi(\Phi_i / M_i)d\Phi_i$$

Esta expresión no se puede evaluar analíticamente debiéndose utilizar métodos de aproximación. El estimador que hemos utilizado, se basa en el cálculo de la media harmónica de los valores de la verosimilitud en cada iteración.

$$\hat{f}(y / M_i) = \left[\frac{1}{H}\sum_{i=1}^{H}\frac{1}{f(y/\Phi_i)}\right]^{-1}$$, donde H es el número de iteraciones una vez superado el calentamiento, y $f(y/\Phi_i)$ es la densidad de los datos dados los parámetros (o verosimilitud) evaluada en la iteración i como: $f(y/\Phi_i) = \prod_{j=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2_{ei}}}\exp(-\frac{1}{2\sigma^2_{ei}}\sum_{j=1}^{N}e_i^2)$, siendo N el número total de datos, y e_i^2 el cuadrado del residuo Este estimador converge al correcto valor de $f(y/M_i)$ *si* H tiende a infinito. (Newton y Raftery, 1994).

Resultados y Discusión

Las inferencias se han realizado a partir de las *distribuciones marginales posteriores* de cada uno de los parámetros de interés, a partir del algoritmo de muestreo de Gibbs. El periodo de calentamiento que se estimo como correcto para considerar que las cadenas habían convergido a la distribución estacionaria, se obtuvo a partir de la visualización de sus trazas. El número de iteraciones totales y salto entre muestras para la obtención de las distribuciones marginales, se basó en el criterio de Raftery y Lewis (R&L) que se incluye en el programa CODA (1995). No obstante, decidimos ser conservadores en cuanto al numero total de iteraciones, y aumentamos hasta el especificado en la

Tabla 1, de forma que se asegurase que se visitan todas las regiones del espacio paramétrico en el proceso de muestreo.

Tabla 1: Descripción de los parámetros descriptivos de las cadenas de Markov, Densidad Marginal de los datos y VR para los distintos modelos

Modelo M_k	Salto	Calentamiento		Iteraciones criterio Raftery &Lewis	Iter. totales efectuadas	Ln[f(y/M_k)]	varianza residual		
							Tramo en días		
		R&L	traza				5-75	76-265	266-335
A3	11	682	10.000	83.655	140.000	-4.906.278.388	15,5363	10,0677	9,2462
C3	12	564	18.000	69.600	100.000	-4.798.162.364	13,0874	9,3079	8,3045
C5	12	132	10.000	16.320	100.000	-4.796.109.392	13.1235	9,4166	8.3856
D3	12	468	10.000	57.552	140.000	-4.781.651.620	12,6760	9,3079	8,2951
MRA-W	19	532	10.000	64.144	100.000	-5.663.074.763	21,9477	17,1151	28,0783
A3	11	431	10.000	56.541	100.000	-4.912.407.781	10,9055		
C3	10	321	10.000	50.211	100.000	-4.781.258.021	9,7812		
D3	10	380	10.000	46.960	100.000	-4.770.190.572	9,5803		

El cálculo del logaritmo neperiano de la densidad marginal de los datos, Ln[*f(y/M$_i$)*], se efectuó a partir de la iteración 40.000 para todos los modelos que convergieron a la distribución de equilibrio. Para el MRA-A se lanzaron dos cadenas desde puntos distintos, no alcanzando ninguna de las dos cadenas la convergencia tras 200.000 y 480.000 iteraciones. También se detectaron problemas de convergencia en el MRA-A en análisis previos con métodos de máxima verosimilitud. El modelo C5, de igual número de parámetros que el MRA-A, convergió rápidamente por tratarse de polinomios ortogonales, que hacen que se disminuya la correlación entre muestras consecutivas acelerando la convergencia de la cadena de Markov. Se realizó un análisis de sensibilidad sobre los modelos A3 y C5, utilizándose distintas matrices en el factor de escala de la distribución a priori Wishart invertida para Σ_u y Σ_p. En ambos casos se obtuvieron medias posteriores similares, siendo las varianzas aditivas y permanentes diarias que se obtienen con ellas, prácticamente coincidentes excepto en los extremos. Una cosa que podría explicar esto es un problema de *"ill-conditioning"* sobre el problema de calcular funciones de covarianza, de tal forma que pequeñas diferencias en los valores de entrada de Σ_u y Σ_p, originan grandes diferencias en las varianzas diarias. Las diferencias observadas en los extremos también se podrían explicar por una falta de información en esas zonas. Estas diferencias fueron más manifiestas en el modelo de mayor número de parámetros, C5. No obstante, se puede concluir que la información a priori utilizada no condiciona el resultado final obtenido, recayendo todo el peso sobre la información que aportan los datos.

En la Tabla 1, se presentan los logaritmos neperianos de las densidades marginales de los datos para los distintos modelos, Ln[f(y/M_k)], siendo el Ln(BF)= Ln[f(y/M_1)]- Ln[f(y/M_2)]. En todos los casos el Ln(BF) es mayor de 150, significando una *evidencia muy fuerte* de M_1 sobre M_2 en la escala de Jeffreys (Kass y Raftery, 1995). El Ln[f(y/M_k) también puede ser utilizado para establecer un ranking de modelos, siendo el peor de los analizados el MRA-w y el mejor el D3. Estos resultados coinciden con los obtenidos por REML (López-Romero y Carabaño, 2000). Este criterio también se ve apoyado por el hecho de verse reducida la VR según se asciende en el ranking de modelos (Tabla 1). También se observó una fuerte deficiencia en las predicciones de las correlaciones genéticas en los modelos C5, y MRA-w, al igual que ocurrió en análisis previos (López-Romero y Carabaño, 2000). Se hizo un análisis adicional para los modelos A3, C3 y D3 manteniendo la VR constante, observándose un mejor comportamiento para los modelos de VR heterogénea sólo en el caso del A3. Se necesitan mas estudios al respecto de la modelización de la VR.

Referencias

Ali, T.T. , Schaeffer, L.R. , 1987. J. Anim. Sci. 67:637

Best, N. G. ,Cowles, M. K., Vines, S. K., 1995. CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampler Output: Version 3.0. Technical report, Biostatistics Unit- MRC, Cambridge, UK.

Gamerman, D. 1997. Markov Chain Monte Carlo .Chapman y Hall

Gelfand, A.E., 1996. Gilks, W.R. et al (Eds.),Markov Chain Monte Carlo in Practice. Champan y Hall -145-162

Kass, R. E., Raftery, A. E. 1995.. Journal of the American Statistical Association. Vol 90, No 30, 773.

Kirkpatrick, M., Heckman,N.,1989.. J. Math. Biol. 27:429.

Newton, M. A., Raftery, A. E., 1994. J. R. Statis. Soc. B 56, No 1:3

López-Romero, P., Carabaño, M.J., 2001- ITEA, Vol. 96A nº 3.(Enviado)

Raftery, A.E., 1996 Gilks, W.R. et al (Eds.),Markov Chain Monte Carlo in Practice. Champan y Hall-163,187

Rekaya, R., M.J. Carabaño, M.A. Toro. 1999. Livest. Prod. Sci. 57:203

Sorensen, D.1996. Internal report no. 82 from the Danish Institute of Animal Science.

Wilmink J.B.M., 1987.. Livest. Prod. Sci. 16:335

Wasserman, L., 1997. Bayesian Model Selection and Model Averaging. Presented at the Mathematical Psycholgy Sympsosium on "Methods for Model Selection" held on August 3-4 1997 in Bloommington, Indiana.