Note

Threshold models with heterogeneous residual variance due to missing information

I. MISZTAL *,**, D. GIANOLA * and Ina HOESCHELE ***

* *University of Illinois, Department of Animal Sciences, Urbana, IL 61801, USA*

** *Warsaw Agricultural University, Department of Animal Sciences (SGGW-AR), Przejazd 4, 05-840 Brwinow, Poland*

*** *Virginia Polytechnic Institute and State University, Department of Dairy Science, Blacksburg, VA 24061, USA*

Summary

Threshold model equations are modified to account for unequal variances of residual effects in the underlying scale. Modifications are simple and can be easily incorporated in programs that conduct a threshold model analysis under the usual assumption of homoscedasticity.

Key words : threshold model, sire evaluation, heterogeneous variance.

Résumé

Les modèles à seuils à variance résiduelle hétérogène du fait d'une information incomplète

Les équations relatives au modèle à seuils peuvent être modifiées afin de prendre en compte des variances résiduelles inégales des effets mesurés sur l'échelle sous-jacente. Les modifications à apporter sont simples et peuvent être aisément incorporées dans les programmes effectuant une analyse par modèle à seuil sous l'hypothèse habituelle d'homoscédasticité.

Mots clés : modèle à seuil, évaluation des pères, variance hétérogène.

I. Introduction

Threshold model equations (GIANOLA & FOULLEY, 1983 ; HARVILLE & MEE, 1984) were originally derived assuming that the residuals of the model for the underlying normal variable have constant variance. This may not be true in general. Also, even if the assumption holds, there are certain genetic evaluation models where lack of some information leads to heterogeneity of residual variance. For example, consider a sire - maternal grandsire model (EVERETT *et al.*, 1979 ; QUAAS *et al.*, 1979). Here, the residual variance depends on whether or not the sire or maternal grandsire is identified. If any of these ancestors is not identified, its effect is not included in the model, but its variance is added to that of the residual effect. A similar problem arises in « reduced » animal models (QUAAS & POLLAK, 1980), when the dam is not identified.

The objective of this note is to present modifications of the threshold model equations needed to account for varying, but known, residual variance.

II. Methods

Consider, for example, a sire-maternal grandsire model. This can be written as :

$$y_{ijk} = \mathbf{x}_{ijk}'\boldsymbol{\beta} + s_i + \frac{1}{2} s_j + e_{ijk} \qquad [1]$$

where y_{ijk} is an observation on individual k, with sire i and maternal grandsire j. The scalars s_i and $\frac{1}{2} s_j$ are the random effects of sires and maternal grandsires, respectively, and $\boldsymbol{\beta}$ is a vector of fixed effects, which relate to y_{ijk} *via* the incidence vector $\mathbf{x}_{ijk}$. In practical applications, the pedigree may be incomplete so the identification of the sire or of the maternal grandsire may be missing. In these cases, one can define a « generalized » residual, ϵ_{ijk}, which can take the values :

$$\epsilon_{ijk} = \begin{cases} e_{ijk} \sim \mathrm{N}(0,\sigma_e^2), \\ s_i + e_{ijk} \sim \mathrm{N}(0,\sigma_s^2 + \sigma_e^2) \text{ if the sire is missing,} \\ \frac{1}{2} s_j + e_{ijk} \sim \mathrm{N}(0,\frac{1}{4}\sigma_s^2 + \sigma_e^2) \text{ if the maternal grandsire is missing.} \end{cases}$$

In the threshold model, due to non-observability of y_{ijk}, it is assumed that $\sigma_e^2 = 1$, so all parameters and random variables are expressed in units of residual standard deviation. Thus, depending on the situation :

$$\sigma_\epsilon^2 = \begin{cases} 1, \\ 1 + \sigma_s^2, \\ 1 + \frac{1}{4}\sigma_s^2. \end{cases}$$

With this in mind, the underlying variable in the threshold model can be written as :

$$y_j = \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u} + \epsilon_j = \mu_j + \epsilon_j \qquad [2]$$

where $\mathbf{u}$ includes both sire and maternal grandsire effects, and $\mathbf{z}_j$ is an incidence vector with elements appropriately defined to take into account presence or absence of the effect. As usual (GIANOLA & FOULLEY, 1983) :

$$\mathrm{E}(y_j \mid \boldsymbol{\beta},\mathbf{u}) = \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u} \qquad [3]$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{G} \qquad [4]$$

and now

$$\epsilon_j \sim \mathrm{N\,I\,I\,D}(0,\sigma_j^2) \qquad [5]$$

where $\sigma_j^2 = 1$, $1 + \sigma_s^2$, or $1 + \frac{1}{4}\sigma_s^2$, depending on the situation.

Let m be the number of categories as described by GIANOLA & FOULLEY (GF, 1983) and HARVILLE & MEE (1984). The conditional probability that observation j is in category k, given μ_j, can be written as :

$$P_{jk} = \Phi[(t_k - \mu_j)/\sigma_j] - \Phi[(t_{k-1} - \mu_j)/\sigma_j] \qquad [6]$$

where $t_1 < t_2 < \ldots < t_{m-1}$ is a set of fixed thresholds which partition the real line into m mutually exclusive and exhaustive intervals. The log posterior density function of $\boldsymbol{\theta}' = (\mathbf{t}',\boldsymbol{\beta}',\mathbf{u}')$, with $\mathbf{t}$ being the vector of thresholds is :

$$\mathrm{L}(\boldsymbol{\theta}) = \sum_{j=1}^{s}\sum_{k=1}^{m} n_{jk}\ln(P_{jk}) - \frac{1}{2}\,\mathbf{u}'\mathbf{G}^{-1}\mathbf{u} + \text{constant}. \qquad [7]$$

where s is as in GF.

This function is then maximized with respect to $\boldsymbol{\theta}$ using Fisher's scoring algorithm :

$$\left[-\mathrm{E}\left(\frac{\partial^2\mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right)\right]_{\boldsymbol{\theta}\,=\,\boldsymbol{\theta}^{[i-1]}} \bullet \quad \boldsymbol{\Delta}^{[i]} = \left[\frac{\partial\mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}\,=\,\boldsymbol{\theta}^{[i-1]}} \qquad [8]$$

where $[i]$ is round number and $\boldsymbol{\Delta}^{[i]} = \boldsymbol{\theta}^{[i]} - \boldsymbol{\theta}^{[i-1]}$. Let $\boldsymbol{\alpha}_j = \boldsymbol{\theta}/\sigma_j$, and note that P_{jk} in [6] is as in GF, but allows for heterogenous variance. Then :

$$\begin{aligned}\frac{\partial\mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= \sum_{j=1}^{s}\sum_{k=1}^{m}\left(\frac{n_{jk}}{\mathrm{P}_{jk}}\frac{\partial P_{jk}}{\partial\boldsymbol{\alpha}_j}\right)\frac{\partial\boldsymbol{\alpha}_j}{\partial\boldsymbol{\theta}} - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}) \\ &= \sum_{j=1}^{s}\sum_{k=1}^{m}\sigma_j^{-1}\left(\frac{n_{jk}}{P_{jk}}\frac{\partial\mathrm{P}_{jk}}{\partial\boldsymbol{\alpha}_j}\right) - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{u}'\mathbf{G}^{-1}\mathbf{u})\end{aligned} \qquad [9]$$

This vector is exactly as in GF except for two aspects : (1) the scalar σ_j^{-1} appears, and (2) P_{jk} is evaluated as in [6], as opposed to taking $\sigma_j = 1$ for all observations. Thus :

$$\frac{\partial\mathrm{L}}{\partial\boldsymbol{\theta}} = \begin{bmatrix}\mathbf{p}^* \\ \mathbf{X}'\mathbf{v}^* \\ \mathbf{Z}'\mathbf{v}^* - \mathbf{G}^{-1}\mathbf{u}\end{bmatrix} \qquad [10]$$

where $\mathbf{p}^*$ and $\mathbf{v}^*$ are similar to $\mathbf{p}$ and $\mathbf{v}$ in GF :

$$\mathbf{p}^* = \{p_k\} = \left\{\sum_{j=1}^{s}\left[\frac{n_{jk}}{P_{jk}} - \frac{n_{j(k+1)}}{P_{j(k+1)}}\right]\phi\left(\frac{t_k - \mu_j}{\sigma_j}\right)\sigma_j^{-1}\right\}\ ;\ k = 1,\ \ldots,\ m-1$$

and

$$\mathbf{v}^* = \{v_j\} = \left\{\sum_{k=1}^{m}\left[n_{jk}\frac{\phi\left(\frac{t_{k-1} - u_j}{\sigma_j}\right) - \phi\left(\frac{t_k - u_j}{\sigma_j}\right)}{P_{jk}}\right]\sigma_j^{-1}\right\}\ ;\ j = 1,\ \ldots,\ s.$$

Similarly, the second derivatives of $\mathrm{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ can be written as :

$$\frac{\partial^2\mathrm{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} = \sum_{j=1}^{s}\sum_{k=1}^{m} n_{jk}\,\sigma_j^{-2}\,\frac{\partial^2\ln(P_{jk})}{\partial\boldsymbol{\alpha}_j\,\partial\boldsymbol{\alpha}'_j} - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}(\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}). \qquad [11]$$

Again, this matrix is as in GF except for the factor σ_j^{-2} and with P_{jk} calculated as in [6]. Hence, after taking expectations in Fisher's scoring :

$$-E\,\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} = \begin{bmatrix} \mathbf{T}^* & \mathbf{L}^{*\prime}\mathbf{X} & \mathbf{L}^{*\prime}\mathbf{Z} \\ \mathbf{X}'\mathbf{L}^* & \mathbf{X}'\mathbf{W}^*\mathbf{X} & \mathbf{X}'\mathbf{W}^*\mathbf{Z} \\ \mathbf{Z}'\mathbf{L}^* & \mathbf{Z}'\mathbf{W}^*\mathbf{X} & \mathbf{Z}'\mathbf{W}^*\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \qquad [12]$$

where each element of $\mathbf{T}^*$, $\mathbf{L}^*$, and $\mathbf{W}^*$ is evaluated as in GF with the following modifications : (1) replace $\phi\,(t_k - \mu_j)$ by $\phi\,[(t_k - \mu_j)/\sigma_j]$, (2) calculate P_{jk} as in [6], (3) multiply each elementary term (the « contribution » of each row in the contingency table) by σ_j^{-2}. Using [10] and [12], iteration proceeds with [8].

From a computational viewpoint, it is useful to observe that [8] is usually built summing « contributions » from each observation or each row in the contingency table. Let $\mathbf{C}_j^{[i-1]}$ and $\mathbf{r}_j^{[i-1]}$ be the « contributions » of the row j in round $i - 1$ to the coefficient matrix and the right-hand sides, respectively. The modified system of equations is then :

$$\left\{\sum_j \sigma_j^{-2}\,\mathbf{C}_j^{[i-1]} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{G}^{-1} \end{bmatrix}\right\} \begin{bmatrix} \Delta\mathbf{t}^{[i]} \\ \Delta\boldsymbol{\beta}^{[i]} \\ \Delta\mathbf{u}^{[i]} \end{bmatrix} = \sum_j \sigma_j^{-1}\,\mathbf{r}_j^{[i-1]} - \begin{bmatrix} 0 \\ 0 \\ \mathbf{G}^{-1}\mathbf{u}^{[i-1]} \end{bmatrix}. \qquad [13]$$

III. Numerical example

A hypothetical example involving two unrelated sires from the same population, appearing also as maternal grandsires, was considered. It was assumed that the offspring of these sires were recorded in the same testing environment. The response was binary and the 15 observations available are as shown below :

Row	Sire	Maternal grandsire	Responses in category	
			1	2
1	1	2	2	4
2	2	—	1	1
3	—	1	4	3

Because of the assumptions, fixed effects need not be considered, and the model for the underlying variable is :

$$y_{ijk} = s_i + \frac{1}{2}\,s_j + e_{ijk}.$$

Above, s_i and $\frac{1}{2}\,s_j$ are the random effects of sire i and maternal grandsire j, respectively. Under additive inheritance, $\sigma_s^2 = \sigma_a^2/4$, where σ_a^2 is additive genetic variance.

In the contingency table, there are three situations corresponding to each of the rows. The residual variances for these cases are :

row 1 : $11/16\ \sigma_a^2 + \sigma_e^2$

row 2 : $12/16\ \sigma_a^2 + \sigma_e^2$

row 3 : $15/16\ \sigma_a^2 + \sigma_e^2$

where σ_e^2 is environmental variance. Setting the residual variance corresponding to a sire model equal to 1 (row 2), and assuming a heritability (h^2) of 0.25, one obtains $\sigma_1^2 = 0.9833$, $\sigma_2^2 = 1$, and $\sigma_3^2 = 1.05$.

Equations [13], using null starting values for threshold t and sire transmitting abilities s_1 and s_2, are :

$$\left\{ 1.016 \begin{bmatrix} 3.820 & -3.820 & -1.910 \\ -3.820 & 3.820 & 1.910 \\ -1.910 & 1.910 & 0.955 \end{bmatrix} + 1 \begin{bmatrix} 1.273 & 0 & -1.273 \\ 0 & 0 & 0 \\ -1.273 & 0 & -1.273 \end{bmatrix} \right.$$

$$\left. + 0.952 \begin{bmatrix} 4.457 & -2.228 & 0 \\ -2.228 & 1.114 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 15 & 0 \\ 0 & 0 & 15 \end{bmatrix} \right\} \begin{bmatrix} t^{[1]} \\ s_1^{[1]} \\ s_2^{[1]} \end{bmatrix}$$

$$= 1.009 \begin{bmatrix} -1.596 \\ 1.596 \\ 0.798 \end{bmatrix} + 1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + 0.976 \begin{bmatrix} 0.798 \\ 0.399 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 15 \times 0 \\ 15 \times 0 \end{bmatrix}$$

and after summation become :

$$\begin{bmatrix} 9.042 & -6.007 & -3.216 \\ -6.007 & 19.946 & 1.942 \\ -3.216 & 1.942 & 17.245 \end{bmatrix} \begin{bmatrix} t^{[1]} \\ s_1^{[1]} \\ s_2^{[1]} \end{bmatrix} = \begin{bmatrix} -0.831 \\ 1.220 \\ 0.805 \end{bmatrix}$$

where $t^{[1]}$ and $s_i^{[1]}$ are the solution for t and s_i at round 1 ; the number 15 is the ratio of residual to sire variance corresponding to $h^2 = 0.25$. Collecting terms and solving yields :

$$\begin{bmatrix} t^{[1]} \\ s_1^{[1]} \\ s_2^{[1]} \end{bmatrix} = \begin{bmatrix} -0.0498 \\ 0.0430 \\ 0.0325 \end{bmatrix}.$$

The solutions stabilize to 4 digits after the decimal point at the second round of the scoring algorithm :

$$\begin{bmatrix} t^{[2]} \\ s_1^{[2]} \\ s_2^{[2]} \end{bmatrix} = \begin{bmatrix} -0.0500 \\ 0.0431 \\ 0.0326 \end{bmatrix}.$$

Received October 12, 1987.

Accepted February 26, 1988.

Acknowledgements

Support of the Illinois Agriculture Experiment Station and of US-Israel Binational Agricultural Research and Development Project No. US-805-84 is gratefully acknowledged.

References

EVERETT R.W., QUAAS R.L., MCCLINTOCK A.E., 1979. Daughter's maternal grandsires in sire evaluation. *J. Dairy Sci.*, **62**, 1304-1313.

GIANOLA D., FOULLEY J.L., 1983. Sire evaluation for ordered categorical data with a threshold model. *Génét. Sél. Evol.*, **15**, 201-224.

HARVILLE D.A., MEE R.W., 1984. A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, **40**, 393-408.

QUAAS R.L., EVERETT R.W., MCCLINTOCK A.C., 1979. Maternal grandsire model for dairy sire evaluation. *J. Dairy Sci.*, **62**, 1648-1654.

QUAAS R.L., POLLAK E.J., 1980. Mixed model methodology for farm and ranch beef cattle testing programs. *J. Anim. Sci.*, **51**, 1280-1287.