Ann. Génét. Sél. anim., 1981, **13** (2), 181-188

NOTE

La régression parent-descendant dans le cas d'un nombre variable de mesures par parent

J.-L. FOULLEY

I.N.R.A., Station de Génétique quantitative et appliquée,
Centre de Recherches zootechniques
F 78350 Jouy-en-Josas (France)

Résumé

Cet article généralise la méthode donnée par OLLIVIER (1974) pour estimer la régression parent-descendant au cas où plusieurs mesures du même caractère, en nombre variable, sont prises par parent. La méthode proposée repose sur un changement de variable parentale. Une application numérique est présentée.

Les premiers, KEMPTHORNE & TANDON (1953) ont proposé une méthode d'estimation de la régression (β) du descendant sur le parent qui tient compte d'un nombre variable de descendants par parent. OLLIVIER (1974) a généralisé cette méthode dans le cas où il existe deux types de relation de parenté (pleins frères et demi-frères par exemple) chez les descendants. Ce modèle peut également être appliqué à la situation simple d'une seule relation de parenté, mais avec répétition en nombre variable des mesures effectuées sur les descendants. Par contre l'utilisation de plusieurs mesures par parent nécessite un aménagement de la méthode d'OLLIVIER dont cette note fait l'objet. En effet, il faut tenir compte de l'information inégale existant chez les parents qui se traduit notamment par le fait que la régression du descendant sur la moyenne du parent dépend du nombre de mesures faites sur celui-ci.

En vue de l'application de la méthode d'OLLIVIER à cette situation, on est alors conduit :

1) à définir une variable parentale z_i telle que la régression d'une performance d'un descendant (i) sur z_i soit égale à β ;

2) à apporter en conséquence quelques modifications à l'estimateur d'OLLIVIER eu égard à l'hétérogénéité des variances et covariances résiduelles.

Changement de variable parentale

En reprenant les notations d'OLLIVIER, le modèle général s'écrit sous la forme :

$$y_{ijk} = \mu_y + \beta z_i + e_{ijk}$$

où y_{ijk} est la variable dépendante d'espérance μ_y.

z_i est une variable indépendante qui caractérise le parent i composée *a priori* ainsi

$$z_i = \sum_{ul} \overset{(i)}{\alpha}_{ul} x_{ul}$$

$u = 1, 2, ..., i, ..., p$ est l'indice du parent.
$l = 1, 2, ..., q_i$ est l'indice de répétition chez le parent.
j est l'indice du sous-groupe chez les descendants.
$k = 1, 2, ..., n_{ij}$ est l'indice de l'observation intra sous groupe ij.

Les coefficients $\overset{(i)}{\alpha}_{ul}$ recherchés pour le parent i qui pondèrent l'ensemble des performances parentales découlent des conditions que l'on pose sur z_i à savoir :

- $E(z_i) = 0$ soit $\sum_{ul} \overset{(i)}{\alpha}_{ul} = 0$ (2)
- var z_i maximum (3)
- $\text{Cov}(y_{ijk}, z_i) = \beta \text{ var } z_i$, β étant le coefficient de régression de y_{ijk} en x_{il}, ce qui implique :

$$\text{var } z_i = \sum_{l} \overset{(i)}{\alpha}_{il} \text{ var } x_{il} \qquad (4)$$

La détermination de ces coefficients revient à maximiser la fonction H_i suivante :

$$H_i = \text{var } z_i - \Pi \left[\text{var } z_i - (\overset{2}{\sigma}_a + \overset{2}{\sigma}_b) \sum_{l} \overset{(i)}{\alpha}_{il} \right] - \Phi \sum_{ul} \overset{(i)}{\alpha}_{ul}$$

par rapport aux paramètres $\overset{(i)}{\alpha}_{ul}$, Π et Φ (multiplicateurs de Lagrange)

$\overset{2}{\sigma}_a$ et $\overset{2}{\sigma}_b$ sont les composantes de la variance entre parents et intraparents.

Le système d'équations s'écrit :

$$\left.\begin{array}{ll} \bullet \quad \dfrac{\partial H_i}{\partial \overset{(i)}{\alpha}_{ul}} = 0 & (\Sigma\, q_i \text{ équations}) \\ \bullet \quad \sum_{ul} \overset{(i)}{\alpha}_{ul} = 0 & \text{(condition 2)} \\ \bullet \quad \overset{2}{\sigma}_a \sum_{u} \left(\sum_{l} \overset{(i)}{\alpha}_{ul} \right)^2 + \overset{2}{\sigma}_b \sum_{ul} \overset{(i)2}{\alpha}_{ul} = (\overset{2}{\sigma}_a + \overset{2}{\sigma}_b) \sum_{l} \overset{(i)}{\alpha}_{il} & \text{(condition 4)} \end{array}\right\} \quad (5)$$

Or :

$$\frac{\partial H_i}{\partial \overset{(i)}{\alpha_{il}}} = \overset{2}{\sigma_a} \sum_l \overset{(i)}{\alpha_{il}} + \overset{2}{\sigma_b}\, \overset{(i)}{\alpha_{il}} = \frac{\Phi - \Pi\,(\overset{2}{\sigma_a} + \overset{2}{\sigma_b})}{2\,(1 - \Pi)}$$

$$\frac{\partial H_i}{\partial \overset{(i)}{\alpha_{i'l}}} = \overset{2}{\sigma_a} \sum_l \overset{(i)}{\alpha_{i'l}} + \overset{2}{\sigma_b}\, \overset{(i)}{\alpha_{i'l}} = \frac{\Phi}{2(1 - \Pi)} \quad \text{pour } i' \neq i$$

En posant :

$$\Pi' = -\frac{\Pi\,(\overset{2}{\sigma_a} + \overset{2}{\sigma_b})}{2\,(1 - \Pi)}\,; \quad \Phi' = -\frac{\Phi}{2\,(1 - \Pi)}\,; \quad l_i = \frac{1}{q_i\,\overset{2}{\sigma_a} + \overset{2}{\sigma_b}}$$

le système (5) s'écrit :

$$\left.\begin{aligned} &\overset{(i)}{\alpha_{il}} = \overset{(i)}{\alpha_i} = (\Pi' + \Phi')\, l_i \\ &\overset{(i)}{\alpha_{i'l}} = \overset{(i)}{\alpha_{i'}} = \Phi' l_{i'} \text{ pour } i' \neq i \\ &(2) \\ &(4) \end{aligned}\right\} \quad (6)$$

La condition (2) équivaut à $\Pi' q_i l_i + \Phi' \sum_i q_i l_i = 0$.

En exprimant ainsi Π' en fonction de Φ' et en remplaçant dans la condition (4) il vient :

$$\Phi' = -\frac{(\sum_u k_u - k_i)\,(\overset{2}{\sigma_a} + \overset{2}{\sigma_b})^2}{\overset{2}{\sigma_a}\left[(\sum_u k_u - k_i)^2 + \sum_{u \neq i} \overset{2}{k_u}\right] + \overset{2}{\sigma_b}\left[\frac{1}{q_i}(\sum_u k_u - k_i)^2 + \sum_{u \neq i} \frac{k_u^2}{q_u}\right]} \quad (7)$$

avec :

$$k_u = l_u q_u\,(\overset{2}{\sigma_a} + \overset{2}{\sigma_b})$$

Le dénominateur de (7) se met sous la forme :

$$\frac{\overset{2}{\sigma_a} + \overset{2}{\sigma_b}}{k_i} \sum_u k_u\,(\sum_u k_u - k_i)$$

d'où :

$$\left.\begin{aligned} \Phi' &= -\frac{k_i}{\sum_u k_u}\,(\overset{2}{\sigma_a} + \overset{2}{\sigma_b}) \\ \Pi' + \Phi' &= \frac{\sum_u k_u - k_i}{\sum_u k_u}\,(\overset{2}{\sigma_a} + \overset{2}{\sigma_b}) \end{aligned}\right\} \quad (8)$$

Comme, d'après (6) la variable z_i peut s'écrire :

$$z_i = \sum_u \overset{(i)}{\alpha_u} \; q_u \, x_{u\cdot}$$

$$\text{avec} \quad x_{u\cdot} = \sum_u x_{ul}/q_u$$

En reportant (8) dans (6), on en déduit les valeurs des pondérations :

- pour $u = i$

$$\overset{(i)}{\alpha_i} \; q_i = \frac{\sum_u k_u - k_i}{\sum_u k_u} k_i$$

- pour $u = i' \neq i$

$$\overset{(i)}{\alpha_{i'}} \; q_{i'} = -\frac{k_i k_{i'}}{\sum_u k_u}$$

Ainsi, z_i s'exprime sous la forme :

$$z_i = k_i (x_{i\cdot} - \tilde{x}) \qquad (9)$$

avec :

$$x_{i\cdot} = \sum_l x_{il}/q_i$$

$$\tilde{x} = (\sum_i k_i x_{i\cdot})/\sum_i k_i$$

$k_i = \dfrac{q_i}{1 + (q_i - 1)\, r}$, r étant égal à la répétabilité $\dfrac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2}$.

Dans ces conditions, $\text{var}\, z_i = h_i \, \text{var}\, x$

avec :

$$h_i = \frac{\sum_u k_u - k_i}{\sum_u k_u} k_i$$

Estimation du coefficient β

Avec la définition de la variable indépendante établie en (9), la matrice des variances et covariances des résidus e_{ijk} dépend de i. On peut aisément adapter la méthode proposée par OLLIVIER à cette situation. Pour ce faire, nous poserons en reprenant ses notations :

- $\text{var } e_{ijk} = \overset{2}{\sigma}_i$
- $\overset{2}{\sigma}_i = \sigma^2 / v_i$ où $\sigma^2 = \text{var } y\,(1 - \beta^2\tau)$
- $v_i = \dfrac{1 - \beta^2\tau}{1 - h_i\beta^2\tau}$
- $\tau = \text{var } x / \text{var } y$
- $\varrho_i = \dfrac{\text{Cov}\,(e_{ijk}, e_{ij'm})}{\sqrt{\text{var } e_{ijk}\ \text{var } e_{ij'm}}} = \dfrac{\varrho_1 - h_i\beta^2\tau}{1 - h_i\beta^2\tau}$ pour $j \neq j'$
- $\varrho'_i = \dfrac{\text{Cov}\,(e_{ijk}, e_{ijk'})}{\sqrt{\text{var } e_{ijk}\ \text{var } e_{ijk'}}} = \dfrac{\varrho_2 - h_i\beta^2\tau}{1 - h_i\beta^2\tau}$ pour $k \neq k'$

où ϱ_1 et ϱ_2 sont tels que :

$$\frac{\text{Cov}\,(y_{ijk}, y_{ij'k'})}{\sqrt{\text{var } y_{ijk}\ \text{var } y_{ij'k'}}} = \begin{cases} \varrho_1 \text{ si } j \neq j' \\ \varrho_2 \text{ si } j = j' \text{ et } k \neq k' \end{cases}$$

L'estimateur linéaire L de β que l'on cherche, soit $L = \sum_{ijk} \lambda_{ijk}\, y_{ijk}$ est choisi tel que :

- $E(L) = \beta$ soit $\sum_{ijk} \lambda_{ijk} = 0$ et $\sum_{ijk} \lambda_{ijk}\, z_i = 1$
- var L minimum par rapport aux λ_{ijk}

Les λ_{ijk} sont obtenus en minimisant la quantité.

$$Q = \text{var } L - 2\,\Pi'' \sum_{ijk} \lambda_{ijk} - 2\,\Phi'' \left(\sum_{ijk} \lambda_{ijk}\, z_i - 1\right)$$

où :

$$\text{var } L = \sigma^2 \left[\sum_{ijk} \frac{1 - \varrho'_i}{v_i} \lambda^2_{ijk} + \sum_{ij} \frac{\varrho'_i}{v_i} \left(\sum_k \lambda_{ijk}\right)^2 + \sum_i \frac{\varrho_i}{v_i} \sum_{j \neq j'} \sum_{km} \lambda_{ijk}\, \lambda_{ij'm} \right]$$

Π'' et Φ'' sont des multiplicateurs de Lagrange.

On obtient alors par une démonstration analogue à celle d'OLLIVIER :

$$\hat{\beta} = \frac{\sum_{ij} w_{ij}\, y_{ij\cdot}\, (z_i - \tilde{z})}{\sum_i w_{io}\, (z_i - \tilde{z})^2} \qquad (10)$$

avec :

$$w_{ij} = \frac{v_i\, N_{ij}}{1 + \varrho_i\, N_{io}}$$

$$N_{ij} = \frac{n_{ij}}{1 + \varrho'_i (n_{ij} - 1) - \varrho_i n_{ij}}$$

$$w_{io} = \sum_j w_{ij} \; ; \; N_{io} = \sum_j N_{ij}$$

$$y_{ij\cdot} = \sum_k y_{ijk}/n_{ij} \; ; \; \tilde{z} = (\sum_i w_{io} z_i)/\sum_i w_{io}$$

La variance d'échantillonnage de cet estimateur est donnée par :

$$\text{var}\,\hat{\beta} = \frac{\sigma^2}{\left[\sum_i w_{io} (z_i - \tilde{z})^2\right]} \sum_i \frac{(z_i - \tilde{z})^2}{v_i} \left(\sum_j \frac{w_{ij}^2}{N_{ij}} + \varrho_i w_{io}^2 \right) \quad (11)$$

qui se réduit à $\sigma^2/\sum_i w_{io} (z_i - \tilde{z})^2$ lorsque les coefficients ϱ_i et ϱ'_i utilisés dans le calcul des pondérations w_{ij} sont les valeurs vraies des paramètres. OLLIVIER (1974) propose comme estimation de la variance d'échantillonnage définie en (11) celle obtenue en remplaçant σ^2, ϱ_i et N_{ij} par les valeurs correspondantes utilisées lors de l'estimation (10) de β.

Application numérique

Elle concerne 37 et 39 teneurs en γ globulines du colostrum bovin prises sur 23 mères et 32 filles respectivement de 1972 à 1975. Ces données ont été présentées et analysées en détail par ailleurs par DARDILLAT *et al.* (1978).

L'échantillon utilisé ici se présentait schématiquement de la façon suivante :

	Nombre de filles par mère			Nombre de mesures par individu			
	1	2	3	1	2	3	4
Mères	15	7	1	13	7	2	1
Filles				28	1	3	

Une première analyse par moindres carrés a été effectuée sur un fichier de 135 mesures provenant de 62 vaches non apparentées suivant un modèle factoriel « année », « individu ». Des facteurs de correction de l'effet « année » ont été alors obtenus ainsi qu'une estimation du coefficient de répétabilité $\hat{r} = 0{,}40$ des teneurs ajustées pour l'année.

Ici ϱ_2 étant égal *a priori* à la répétabilité des mesures et ϱ_1 au quart de l'héritabilité (donc au maximum au quart de r), les valeurs introduites pour ces

paramètres ont été de $\varrho_2 = 0{,}40$ et $\varrho_1 = 0{,}10$. Les variances des mesures chez les mères et les filles étant très voisines et non significativement différentes, les calculs ont été faits avec $\tau = 1$.

Dans ces conditions, on a obtenu $\widehat{\beta} = 0{,}217$ et $\widehat{\sigma\beta} = 0{,}141$. Le même calcul, conduit sans prendre en compte l'hétérogénéité des variances et covariances résiduelles, à aboutir à $\widehat{\beta} = 0{,}214$ et $\widehat{\sigma\beta} = 0{,}144$. Cette simplification affecte ainsi très peu les estimations, relatives, il est vrai, à un échantillon très limité et relativement peu déséquilibré.

Discussion

La variable z_i a été déterminée de façon à satisfaire des propriétés intéressantes de la régression utilisée comme prédicteur de $y_{ijk} - \mu_y$. On a alors supposé β connu *a priori* et on ne s'est posé qu'en second lieu le problème de l'estimation de ce paramètre pour une variable z_i observable et définie selon cet objectif initial. Quant à l'intérêt de cette méthode de régression pondérée, il faut rappeler qu'il est d'autant plus net que les valeurs attribuées aux paramètres ϱ_i et ϱ'_i (donc aussi à τ, ϱ_1, ϱ_2, β et r sont plus proches des valeurs vraies. De plus, les formules présentées ici ne sont rigoureusement applicables que lorsqu'on utilise des valeurs *a priori* de ces paramètres. Dans ces conditions, l'estimation de β reste bien centrée, ce qui n'est plus vrai avec des méthodes itératives.

En fait, d'autres procédures auraient pu être envisagées. Une méthode intéressante consiste à utiliser l'ensemble de l'information disponible sur les parents et leurs descendants en vue d'estimer les paramètres auxquels on s'intéresse. En effet, on dispose d'informations permettant d'appréhender la variabilité génétique, non seulement à partir de la relation parent-descendant, mais aussi, à partir des relations entre les descendants eux-mêmes. De même, dans le cas du dispositif abordé ici, la répétabilité des mesures peut être appréciée à la fois sur les parents et sur leurs descendants. L'analyse fait alors appel à l'écriture de la matrice des variances et covariances des observations. Ici, cette matrice est paramétrable linéairement en fonction des composantes $\sigma^2 g$ (effets génétiques additifs), $\sigma^2 m$ (effets de dominance et de milieu permanent) et $\sigma^2 e$ (effets aléatoires de milieu) de sorte que

$$h^2 = \sigma^2 g/(\sigma^2 g + \sigma^2 m + \sigma^2 e) \text{ et } r = (\sigma^2 g + \sigma^2 m) / (\sigma^2 g + \sigma^2 m + \sigma^2 e).$$

Sous l'hypothèse de normalité des distributions, ces variances pourraient être alors estimées par maximum de vraisemblance suivant, par exemple, le procédé développé par THOMPSON (1977). On pourrait également utiliser des estimateurs quadratiques tels que ceux proposés par MATHERON *et al.* (1974) en vue de l'estimation des composantes de la variance en présence d'effets maternels ou ceux à norme minimum discutés notamment par CHEVALET (1976).

Reçu pour publication en mai 1981.

Remerciements

L'auteur tient à remercier L. OLLIVIER et J. RAZUNGLES pour leurs utiles critiques et suggestions à la lecture du manuscrit.

Summary

The parent progeny regression
in the case of a variable number of measurements per parent

This paper extends the method given by OLLIVIER (1974) for estimating the parent-offspring regression when different numbers of records on the parents are available. The method presented is based on a change of the parent variable. A numerical application is given.

Références bibliographiques

CHEVALET C., 1976. Estimation des composantes de la variance phénotypique dans une population consanguine. I - Elaboration du modèle. *Ann. Génét. Sél. anim.*, **8**, 181-206.

DARDILLAT J., TRILLAT G., LARVOR P., 1978. Colostrum immunoglobulin concentration in cows : relationship with their calf mortality and with the colostrum quality of their female offspring. *Ann. Rech. vét.*, **9**, 375-384.

KEMPTHORNE O., TANDON O.B., 1953. The estimation of heritability by regression of offspring on parent. *Biometrics*, **9**, 90-100.

MATHERON G., POUJARDIEU B., LEFORT G., 1974. Un modèle d'estimation des paramètres génétiques en présence d'effets génétiques directs et maternels chez le lapin. 1er Congr. mond. Génét. appl. Elev. Madrid, 7-11 octobre 1974, vol. III (Symposia), 447-454.

OLLIVIER L., 1974. La régression parent-descendant dans le cas de descendances subdivisées en familles de taille inégale. *Biometrics*, **30**, 59-66.

THOMPSON R., 1977. The estimation of heritability with unbalanced data. I. Observation available on parents and offspring. *Biometrics*, **33**, 485-495.