ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN – POLONIA
VOL. XXVII, 8 SECTIO A 1973

Katedra Zastosowań Matematyki, Akademia Rolnicza, Lublin

HENRYK MIKOS

Variance Component Estimation in the Unbalanced N-way Nested Classification

Estymacja komponentów wariancyjnych w niezrównoważonej N-krotnej klasyfikacji hierarchicznej

Оценка компонент дисперсии по несбалансированным данным N-факторной иерархической классификации

Introduction. Estimates of the variance components for the unbalanced nested classification obtained with the use of analytical methods are given in the papers of Gates and Shiue [3], Gower [5], Oktaba [11], Ahrens [1] and Gaylor and Hartwell [4]. Matrix methods for obtaining the variance components estimates are shown in the papers of Searle [13] and Mahamunulu [9]. This paper gives the estimates of variance components for the unbalanced N-way nested classification obtained with the use of the properties of linear spaces.

Model and analysis of variance. The linear model for an observation $y_{i_1i_2\ldots i_{N+1}}$ is taken as

$$(1) \qquad y_{i_1i_2\ldots i_{N+1}} = \mu + a^1_{i_1} + a^2_{i_1i_2} + \ldots + a^N_{i_1i_2\ldots i_N} + e_{i_1i_2\ldots i_{N+1}}$$

where μ is the general mean, $a^1_{i_1}$ is the effect due to the i_1-th first stage class $A^1_{i_1}$, $a^2_{i_1i_2}$ is the effect due to the i_2-th second stage class $A^2_{i_1i_2}$ within $A^1_{i_1}, \ldots, a^N_{i_1i_2\ldots i_N}$ is the effect due to i_N-th N-th stage class $A^N_{i_1i_2\ldots i_N}$ within $A^{N-1}_{i_1i_2\ldots i_{N-1}}$ and $e_{i_1i_2\ldots i_{N+1}}$ is the residual error of the observation $y_{i_1i_2\ldots i_{N+1}}$. We assume that the number of the first stage classes $A^1_{i_1}$ is a^1 so that $i_1 = 1, 2, \ldots, a^1$. Within each $A^1_{i_1}$-class there are $a^2_{i_1}$ A^2-classes so that $i_2 = 1, 2, \ldots, a^2_{i_1}$. Furthemore, within each $A^{p-1}_{i_1i_2\ldots i_{p-1}}$ class ($p = 2, 3, \ldots, N$) there are $a^h_{i_1i_2\ldots i_{p-1}}$ A^p-classes so that $i_p = 1, 2, \ldots, a^p_{i_1i_2\ldots i_{p-1}}$. The number of observations in the N-th stage class $A^N_{i_1i_2\ldots i_N}$ is $n_{i_1i_2\ldots i_N}$ where $n_{i_1i_2\ldots i_N} > 0$. All terms of the model (except μ) are assumed to be independent and normally distributed random variables with zero means and variances $\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_N$ and σ^2_e, respectively. These are the variance components which are to be estimated.

Let $\mathbf{y}$ be the column vector with the elements $y_{i_1 i_2 \dots i_{N+1}}$ $\boldsymbol{\alpha}^p$ — the column vector with elements $a^p_{i_1 i_2 \dots i_p}$ and $\mathbf{e}$ — the column vector the elements of which are the residual errors $e_{i_1 i_2 \dots i_{N+1}}$. Let furthermore $X_p (p = 1, 2, \dots, N)$ denote the $n \times a^p_.$ matrix, where

$$n = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} n_{i_1 i_2 \dots i_N},$$

(2)

$$a^p_. = \sum_{i_1} \sum_{i_2} \dots \sum_{i_{p-1}} a^p_{i_1 i_2 \dots i_{p-1}},$$

in which the element of the q-th row and $(i_1, i_2, \dots, i_p)$-th column is either zero or one; one- if the q-th observation is in the $A^p_{i_1 i_2 \dots i_p}$ -th class of the classification A^p, in other cases — zero.

Now we can express the linear model (1) as follows:

$$\mathbf{y} = \mathbf{J}_n \mu + \sum_{p=1}^{N} X_p \alpha^p + \mathbf{e} \tag{3}$$

where $\mathbf{J}'_n = [1, 1, \dots, 1]$. It is easy to see, that

$$R[J_n] \subset R[X_1] \subset R[X_2] \subset \dots \subset R[X_N] \tag{4}$$

where $R[X]$ denotes the range space of the matrix X, i.e. the space spanned by the column vectors of $\mathbf{X}$. It follows from this that each observation which belongs to the class $A^p_{i_1 i_2 \dots i_p}$ of the classification A^p belongs also to the class $A^{p-1}_{i_1 i_2 \dots i_{p-1}}$ of the classification A^{p-1}

It is easy to see that the random vectors $\alpha^p (p = 1, 2, \dots, N)$ are distributed as $N[\mathbf{0}, \sigma^2_p I_{a^p_.}]$ and the random vector $\mathbf{e}$ is distributed as $N[\mathbf{0}, \sigma^2_e I_n]$. Thus we have the following covariance matrix of the vector $\mathbf{y}$ /c.f. [12]/

$$\Sigma_y = \sum_{p=1}^{N} X_p X'_p \sigma^2_p + \sigma^2_e I_n. \tag{5}$$

In the customary analysis of variance there are the following sums of squares /c.f. [10]/

$$\begin{cases} SS_1 = \sum_{i_1} \sum_{i_2} \dots \sum_{i_{N+1}} (\bar{y}^1_{i_1} - \bar{y})^2, \\ SS_t = \sum_{i_1} \sum_{i_2} \dots \sum_{i_{N+1}} (\bar{y}^t_{i_1 i_2 \dots i_t} - \bar{y}^{t-1}_{i_1 i_2 \dots i_{t-1}})^2 \; (t = 2, 3, \dots, N) \\ SS_e = \sum_{i_1} \sum_{i_2} \dots \sum_{i_{N+1}} (y_{i_1 i_2 \dots i_{N+1}} - \bar{y}^N_{i_1 i_2 \dots i_N})^2 \end{cases} \tag{6}$$

and

where

$$\bar{y}^t_{i_1 i_2 \ldots i_t} = (n^t_{i_1 i_2 \ldots i_t})^{-1} \sum_{i_{t+1}} \sum_{i_{t+2}} \cdots \sum_{i_{N+1}} y_{i_1 i_2 \ldots i_{N+1}}, \quad (t = 1, 2, \ldots, N)$$

(8)
$$\bar{y} = \frac{1}{n} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_{N+1}} y_{i_1 i_2 \ldots i_{N+1}}$$

and

(9)
$$n^t_{i_1 i_2 \ldots i_t} = \sum_{i_{t+1}} \sum_{i_{t+2}} \ldots \sum_{i_N} n_{i_1 i_2 \ldots i_N}.$$

It can be shown that the sums of squares (6) can be written in the following form

(10)
$$SS_t = y'(P[X_t] - P[X_{t-1}])y \quad (t = 1, 2, \ldots, N),$$
$$SS_e = y'(I - P[X_N])y$$

where, if $t = 1$, X_0 should be replaced by $\mathbf{J}_n$ and the term $P[X_t]$ denotes the orthogonal projection operator to the range space of the matrix X_t/c.f. [14]/. It is worth noticing that from the relationship (4) follows immediately that

(11)
$$(I - P[X_N])(P[X_t] - P[X_{t-1}]) = \mathbf{0} \text{ for } t = 1, 2, \ldots, N$$

and

(12)
$$(P[X_t] - P[X_{t-1}])(P[X_r] - P[X_{r-1}]) = \mathbf{0} \text{ for } r \neq t \text{ and } r, t = 1, \ldots, N.$$

Estimation of variance components. To obtain the unbiased estimators of the variance components $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2, \sigma_e^2$, we must have the expected values of the sums of squares (10). According to the formula 2.1.24 in [1] we get

(13)
$$E[SS_e] = \sum_{p=1}^{N} \mathrm{tr}[(I - P[X_N]) X_p X_p' \sigma_p^2] + \mathrm{tr}[(I - P[X_N])\sigma_e^2],$$
$$E[SS_t] = \sum_{p=1}^{N} \mathrm{tr}[(P[X_t] - P[X_{t-1}]) X_p X_p' \sigma_p^2] + \mathrm{tr}[(P[X_t] - P[X_{t-1}])\sigma_e^2] \quad (t = 1, 2, \ldots, N)$$

where $\mathrm{tr}[X]$ denotes the trace of the matrix X.

Let

$$k_{rs} = \mathrm{tr}[P[X_r] X_s X_s'], \quad (r \leqslant s;\ r, s = 0, 1, 2, \ldots, N).$$

Then, in regard to the range space of the matrix X_p and the range space of the matrix $P[X_t]-P[X_{t-1}]$ are orthogonal if $p<t$, we have

$$(14)\quad \begin{aligned} E[SS_e] &= (n-a_{\cdot}^{N})\sigma_e^2, \\ E[SS_t] &= (a_{\cdot}^{t}-a_{\cdot}^{t-1})\sigma_e^2+\sum_{p=t}^{N}(k_{tp}-k_{t-1,p})\sigma_p^2, \quad (t=1,2,\ldots,N) \end{aligned}$$

Now we will find the coefficients k_{rs}. If $r=s=p(p=1,2,\ldots,N)$ we get

$$(15)\quad k_{pp}=\mathrm{tr}[P[X_p]X_pX_p']=\mathrm{tr}[X_pX_p']=\mathrm{tr}[X_p'X_p]=\sum_{i_1}\ldots\sum_{i_p}n_{i_1\ldots i_p}^{p}=n.$$

Similarly

$$(16)\quad k_{Op}=\mathrm{tr}[P[J_n]X_pX_p']=\frac{1}{n}\mathrm{tr}[J_nJ_n'X_pX_p']=\frac{1}{n}\mathrm{tr}[(J_n'X_p)'J_n'X_p]$$

$$=\frac{1}{n}\sum_{i_1}\sum_{i_2}\ldots\sum_{i_p}(n_{i_1i_2\ldots i_p}^{p})^2.$$

In an analogous way we can obtain

$$(17)\quad k_{pr}=\sum_{i_1}\sum_{i_2}\ldots\sum_{i_r}\frac{(n_{i_1i_2\ldots i_r}^{r})^2}{n_{i_1i_2\ldots i_p}^{p}} \quad (p<r;\ p,r=1,2,\ldots,N).$$

Hendorson's first method (c.f. [8]) for estimating the components variance is to equate each of the sums of squares $SS_1, SS_2, \ldots, SS_N, SS_e$ to its expected value. Denoting the resulting estimates as $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \sigma_N^2$, and $\hat{\sigma}_e^2$ the equations for obtaining them are

$$(18)\quad \begin{aligned} SS_t &= \sum_{p=t}^{N}(k_{tp}-k_{t-1,p})\hat{\sigma}_p^2+(a_{\cdot}^{t}-a_{\cdot}^{t-1})\hat{\sigma}_e^2 \quad (t=1,2,\ldots,N), \\ SS_e &= (n-a_{\cdot}^{N})\hat{\sigma}_e^2. \end{aligned}$$

The equations (18) can be written in the following matrix form

$$(19)\quad \mathbf{S}=K\hat{\boldsymbol{\sigma}}^2$$

where $\mathbf{S}=[SS_1, SS_2, \ldots, SS_N, SS_e]'$, $\hat{\boldsymbol{\sigma}}^2=[\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_N^2, \hat{\sigma}_e^2]'$ and K is the triangular matrix of k's. Since all diagonall elements of the matrix K are not equal zero, the matrix K is nonsingular. Hence the following unique solution of the equation (19) exists

$$\hat{\boldsymbol{\sigma}}^2=K^{-1}\mathbf{S}.$$

The sampling covariance matrix of the vector $\hat{\boldsymbol{\sigma}}^2$ can be found for the unbalanced data by the method of Ahrens (c.f. [2]).

Balanced data. When all the $n_{i_1 i_2 \ldots i_N}$ are equal, say m, and when all the $a^p_{i_1 i_2 \ldots i_{p-1}}$ are equal, say a^p ($p = 1, 2, \ldots, N$), i.e. when the data are balanced, we can explicitly obtain the estimates of variance components as well as the sampling variances of their estimates. In this case for $p < r (p, r = 0, 1, 2, \ldots, N)$

$$k_{pr} = a^0 a^1 a^2 \ldots a^p a^{r+1} a^{r+2} \ldots a^{N+1} \tag{20}$$

where to simplify the notation it is taken $1 = a^0$, $m = a^{N+1}$.
Now the equations (18) can be expressed as

$$\begin{gathered} MS_t = \hat{\sigma}_e^2 + \sum_{p=t}^{N} a^{p+1} a^{p+2} \ldots a^{N+1} \hat{\sigma}_p^2 \quad (t = 1, 2, \ldots, N), \\ MS_e = \hat{\sigma}_e^2 \end{gathered} \tag{21}$$

where $MS_e = (1/f_e) SS_e$, $MS_t = (1/f_t) SS_t$ are the mean squares due to the error and the t-th classification, respectively. The notations f_e, f_t are used for the degrees of freedom due to the error and the t-th classification, respectively. It can be shown that

$$\begin{gathered} f_e = a^1 a^2 \ldots a^N (m-1), \\ f_t = a^0 a^1 a^2 \ldots a^{t-1} (a^t - 1) \quad (t = 1, 2, \ldots, N). \end{gathered} \tag{22}$$

We can readily see that

$$MS_t = a^{t+1} a^{t+2} \ldots a^{N+1} \hat{\sigma}_t^2 + MS_{t+1}$$

and hence

$$\hat{\sigma}_t^2 = \frac{1}{a^{t+1} a^{t+2} \ldots a^{N+1}} (MS_t - MS_{t+1}) \quad (t = 1, 2, \ldots, N) \tag{23}$$

where, if $t = N$, MS_{N+1} should be replaced by MS_e.

For balanced data the covariance matrix of the vector y can be expressed as

$$\Sigma_y = I_n \sigma_e^2 + \sum_{p=1}^{N} a^{p+1} a^{p+2} \ldots a^{N+1} \sigma_p^2 P[X_p] \tag{24}$$

for $P[X_p] = (a^{p+1} a^{p+2} \ldots a^{N+1})^{-1} X_p X_p'$ ($p = 1, 2, \ldots, N$).
Now we can prove the following theorem:

Theorem 1. *The projection operators* $I - P[X_N]$, $P[X_t] - P[X_{t-1}]$ ($t = 1, 2, \ldots, N$) *and the covariance matrix* Σ_y *satisfy the following conditions*

$$(I - P[X_N]) \Sigma_y = \varphi_e (I - P[X_N])$$

$$(P[X_t] - P[X_{t-1}] \Sigma_y = \varphi_t (P[X_t] - P[X_{t-1}])$$

where

$$\varphi_e = \sigma_e^2, \; \varphi_t = \sigma_e^2 + \sum_{p=t}^{N} a^{p+1} a^{p+2} \ldots a^{N+1} \sigma_p^2.$$

Proof. Since for $p<t$ the range space $R[X_p]$ and the range space $R\left]P[X_t]-P[X_{t-1}]\right[$ are orthogonal

(25) $$(P[X_t]-P[X_{t-1}])P[X_p] = \mathbf{0} \quad (p<t)$$

and since for $p \geqslant t$ $R\left[P[X_t]-P[X_{t-1}]\right] \subset R[X_p]$ we have (c.f § 76 Theorem 2 in [7]),

(26) $$(P[X_t]-P[X_{t-1}])P[X_p] = P[X_t]-P[X_{t-1}] \quad (p \geqslant t).$$

Thus for $t = 1, 2, \ldots, N$

$$(P[X_t]-P[X_{t-1}])\Sigma_y = \sum_{p=1}^{N} a^{p+1}a^{p+2}\ldots a^{N+1}\sigma_p^2(P[X_t]-P[X_{t-1}])P[X_p]+$$

$$+(P[X_t]-P[X_{t-1}])\sigma_e^2 = \sum_{p=1}^{N} a^{p+1}a^{p+2}\ldots a^{N+1}\sigma_p^2(P[X_t[-P[X^{t-1}])+$$
$$+(P[X_t]-P[X_{t-1}])\sigma_e^2 = \varphi_t(P[X_t]-P[X_{t-1}]).$$

The first condition follows immediately from (4) and (24).
The straightforward conclusion from Theorem 1 is the following theorem:

Theorem 2. *The quadratic forms* $\frac{1}{\varphi_1}SS_1, \frac{1}{\varphi_2}SS_2, \ldots, \frac{1}{\varphi_N}SS_N, \frac{1}{\varphi_e}SS_e$ *are independently distributed as* χ^2 *(chi-square) with degrees of freedom* $f_1, f_2, \ldots, f_N, f_e$*, respectively.*

Proof. From Theorem 1 we have that the matices $1/\varphi_t(P[X_t]-P[X_{t-1}])\Sigma\mathbf{y}(t = 1, 2, \ldots, N)$ and $1/\varphi_t(I-P[X_N]\Sigma\mathbf{y}$ are idempotent. The expectation of the vector $\mathbf{y}$ is the vector J_n which is orthogonal to each of the rang spaces $R\left[I-P[X_N]\right]$, $R\left[P[X_t]-P[X_{t-1}]\right]$ ($t = 1, 2, \ldots, N$). The application of the theorem 4.9 in [6] completes the first part of the proof. The independence of the quadratic forms follows immediately from (11), (12), Theorem 1 and theorem 4.21 in [6].

The sampling variance of any quadratic form $\mathbf{y}'A\mathbf{y}$ of normally-distributed random variables represented by the vector y is $2\mathrm{tr}(\Sigma_y A)^2$ where Σ_y is the covariance matrix of $\mathbf{y}$. The well known formula, Theorem 1 and Theorem 2 will be applied to obtain the sampling variances of the estimates $\hat{\sigma}_e^2$ and $\hat{\sigma}_t^2$ ($t = 1, 2, \ldots, N$). First we will get the sampling variances of the mean squares MS_e and MS_t ($t = 1, 2, \ldots, N$).

$$\mathrm{var}(MS_e) = 2f_e^{-2}\mathrm{tr}\left[\varphi_e(I-P[X_N])\right]^2 = \frac{2\varphi_e^2}{f_e}$$

$$\mathrm{var}(MS_t) = 2f_t^{-2}\mathrm{tr}\left[\varphi_t(P[X_t]-P[X_{t-1}])\right]^2 = \frac{2\varphi_t^2}{f_t}.$$

Hence

$$\operatorname{var}(\hat{\sigma}_t^2) = (a^{t+1}a^{t+2}\dots a^{N+1})^{-2}\left(\frac{\varphi_t^2}{f_t}+\frac{\varphi_{t+1}^2}{f_{t+1}}\right)\ (t = 1, 2, \dots, N)$$

(27)

$$\operatorname{var}(\hat{\sigma}_e^2) = \frac{2\varphi_e^2}{f_e}.$$

On the basis of Theorem 2 we can say that the test function available to verify the hypothesis

$$H_t^0\colon \sigma_t^2 = 0 \text{ is } F_t = \frac{MS_t}{MS_{t+1}}\ (t = 1, 2, \dots, N)$$

where if $t = N$, MS_{N+1} should be replaced by MS_e. If H_t^0 is true, the test function F_t is distributed as $F(f_t, f_{t+1})$.

Acknowledgement. The author is indebted to Professor Dr Victor Oktaba for suggesting the subject of this paper, and for his advice during its preparation.

REFERENCES

[1] Ahrens, H., *Varianzanalyse,* Berlin 1967.

[2] Ahrens, H., *Standardfehler geschatzter Varianzkomponenten eines unbalanzierten Versuchsplanes in r-stufiger hierarchischer Klassifikation.* Monatsb. Deutsch. Akad. Wiss. Berlin 7(2), 1965.

[3] Gates, C.E. and Shiue, C., *The Analysis of Variance of the S-stage Hierarchical Classification.* Biometrics, 18 (1962), 529-536.

[4] Gaylor, D. W. and Hartwell, T. D., *Expected Mean Squares for Nested Classifications.* Biometrics. 25 (1969), 427-430.

[5] Gower, J. C., *Variance Component Estimation for Unbalanced Hierarchical Classifications.* Biometrics, 18 (1962), 537-542.

[6] Graybill, F.A., *An Introduction to Linear Statistical Models.* New York 1961.

[7] Halmos, P.R., *Finite-Dimensional Vector Spaces.* New York 1958.

[8] Henderson, C.R., *Estimation of Variance and Covariance Components.* Biometrics. 9 (1953) 226-252.

[9] Mahamunulu, D., M., *Sampling Variances of the Estimates of Variance Components in the Unbalanced 3-way Nested Classification.* Ann. Math. Statist., 34 (1963), 521-527.

[10] Mikos, H., *Orthonality in the N-way Nested Classification.* Ann. Univ. Mariae Curie-Skłodowska, Sect. A, 27 (1973), 55-63.

[11] Oktaba, W., *Nieortogonalne modele losowe klasyfikacji hierarchicznej.* Roczniki Nauk Rolniczych 82-B-3, 417-435.

[12] Oktaba, W., *Teoria układów eksperymentalnych. I Modele stałe,* PAN Wydz. V, Warszawa 1970.

[13] Searle, S.R., *Variance Components in the Unbalanced 2-way Nested Classification.* Ann. Math. Statist., 32 (1961), 1161-1166.

[14] Seber, G.A.F., *The Linear Hypothesis,* London 1966.

STRESZCZENIE

W pracy otrzymano nieobciążone estymatory komponentów wariancyjnych dla modelu losowego niezrównoważonej N-krotnej klasyfikacji hierarchicznej. Dla omówionego oddzielnie modelu z danymi zrównoważonymi otrzymano ponadto wariancje z próby uzyskanych estymatorów oraz testy istotności dla weryfikacji hipotez dotyczących parametrów modelu. Wszystkie wyniki uzyskano w oparciu o własności przestrzeni liniowych.

РЕЗЮМЕ

В этой работе получены несмещенные оценки компонент дисперсии по несбалансированным данным N-факторной иерархической классификации. Для отдельно обсуждаемой сбалансированной модели получены кроме того выборочные дисперсии оценок и критерии значимости для проверки гипотез об эффектах исследуемых факторов. Все результаты получены с использованием свойств линейных пространств.