FACTA UNIVERSITATIS
Series: **Economics and Organization** Vol. 2, N^{o} 1, 2003, pp. 59 - 64

HENDERSON'S APPROACH TO VARIANCE COMPONENTS ESTIMATION FOR UNBALANCED DATA

UDC 519.233.4

Vera Djordjević, Vinko Lepojević

Faculty of Economics, University of Niš, 18000 Niš, Serbia and Montenegro

Abstract. *The analysis of variance is a frequently used technique of data analysing and probably the most used statistical method in 20th century. However, there are a lot of ambiguities in the interpretation of particular aspects. The aim of this text is an explanation of one of the fundamental parts of this method - the variance components estimation. Namely, by calculating variability in observations we can divide their variance into parts which correspond to the factors we examine, and that means that the subject of study are the variance components. This text is a presentation of Henderson's method for variance components estimation in models of analysis of variance.*

1. INTRODUCTION

Estimating variance components from unbalanced data is not as straightforward as that obtained from balanced data. This is so for two reasons. First, several methods of estimation are available (most of which are reduced to the analysis of variance method for balanced data), but not one of them has yet been clearly established as superior to the others. Second, all the methods involve relatively cumbersome algebra; discussion of unbalanced data can therefore easily deteriorate into a welter of symbols, a situation we do our best (perhaps not successfully) to minimize here.

It is probably safe to describe the Henderson (1953) paper as the foundation paper dealing with variance component estimation from unbalanced data. The methods there described have, accordingly, often been referred to as Henderson's methods 1,2 and 3. As described in Searle (1968, 1971, 1992), Method 1 is simply an analogue of the analysis of variance method used with balanced data; Method 2 is designed to correct a deficiency of Method 1 that arises with mixed models; and Method 3 is based on the method of fitting constants so often used in fixed effects models. Prior to the publication of these methods Windsor and Clark (1940) had utilized the analysis of variance method in studying variation in the catch of plankton nets, Eisenhart (1947) had clearly specified distinctions between fixed, random, and mixed models, and Crump (1946, 1947, 1951) had established

Received December 23, 2003

sampling variances of the variance component estimators in the one-way classification. Henderson (1953) however, greatly extended the estimation procedures, especially in describing three different methods and in indicating their use in multi-way classifications. Since then, a number of developments have been made. Variances of estimated components have been considered by Tukey (1957a), Searle (1956, 1958, 1961a, 1992), Khuri, A.I., and Sahai, H. (1985) Hocking (1996); defects in Henderson's Method 2 have been demonstrated by Searle (1968,1992), and difficulties with the mixed model have been discussed by Searle and Henderson (1961), Cunningham and Henderson (1968), and Thompson (1969) Christensen, R. (1996); and other methods of estimation have been developed: maximum likelihood by Hartley and Rao (1967) and large sample variances there from by Searle (1967,1992), and Sahai and Ageel (2001) .

Not all of these developments have been applied to all of even the most straightforward applications and some of them are more specialized than others.

2. General quadratic forms

All currently available methods for estimating variance components from unbalanced data use, in one way or another, quadratic forms of the observations. Before describing the methods we therefore outline properties of quadratic forms of observations coming from a general linear model. This is taken as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of observations, $\mathbf{X}$ is a matrix of known values, $\boldsymbol{\beta}$ is a vector of parameters (including both fixed and random effects) and $\mathbf{e}$ is a vector of the customary error terms. The vector of means and the variance-covariance matrix are taken respectively as:

$$E(\mathbf{y}) = \boldsymbol{\mu} \text{ and } \mathbf{V} = \text{var}(\mathbf{y}) = E(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'.$$

Expected values

The expected value under the above- mentioned model of the quadratic form $\mathbf{y'Qy}$ is:

$$E(\mathbf{y'Qy}) = \text{tr}\,(\mathbf{QV}) + \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu} \tag{1}$$

where 'tr' represents the trace operation on matrix, that of summing its diagonal elements. This result is the basis of most methods of estimating variance components from unbalanced data. The general methodology is to obtain expected values of quadratic forms from (1) and to equate them to their observed values; i.e. to equate E ($\mathbf{y'Qy}$) to the observed $\mathbf{y'Qy}$. This is exactly what is done with mean squares (which are quadratic forms of the observations) in the analysis of variance method for balanced data. But, whereas with balanced data there is ' obviously' only one set of quadratic forms to use (the analysis of variance mean squares) and they lead to estimators that have some optimal properties, there are many sets of quadratics that can be used for unbalanced data. However, most of such sets lead to estimators that have few optimal properties and no particular set of quadratics has yet been established as more optimal than any other set.

Result (1) applies no matter what form of the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ + e is used: $\boldsymbol{\beta}$ always includes all the effects in the model, be they fixed or random or a mixture of both. In most situations we assume that E(**e**) = **0**, so that var (**e**) is E (**ee**') $=\sigma_e^2$ **I**. In addition, when $\boldsymbol{\beta}$ is a vector of fixed effects, E($\boldsymbol{\beta}$**e**') = $\boldsymbol{\beta}$E(**e**') = **0**; and when β includes elements that are random effects they are assumed to have zero mean and zero covariance with the elements in **e**; thus at all times we take E($\boldsymbol{\beta}$**e**') = E(**e**$\boldsymbol{\beta}$') = **0**.

In a fixed effects model $\boldsymbol{\beta}$ is a vector of fixed effects, E(**y**) = **X**$\boldsymbol{\beta}$ and **V** = var (**y**) = var (**e**) = $\sigma_e^2 I_N$, where are N observations, i.e., **y** is N X 1. Then (1) becomes:

$$E(\mathbf{y'Qy}) = \boldsymbol{\beta}'\mathbf{XQX}\boldsymbol{\beta} + 2\sigma_e^2 tr\,(\mathbf{Q})$$

In a mixed model $\boldsymbol{\beta}$' can be partitioned as

$$\beta' = (\beta_1'\beta_A'\beta_B'...\beta_K'),$$

where $\boldsymbol{\beta}_1$ contains all the fixed effects of the model (including the mean) and where the other $\boldsymbol{\beta}$'s each represent the set of random effects for the factors **A, B, C, ..., K**, these random effects having zero means and zero co variances with the effects of any other set. (Although, only single subscripts are used, interaction effects and/or nested-factor effects are not excluded by this notation. They are considered merely as factors, each identified by a single subscript rather than the letters of the appropriate main effects; for example, **AB**- interaction effects might be in the vector labeled $\boldsymbol{\beta}_F$).

3. THE ANALYSIS OF VARIANCE METHOD (HENDERSON'S METHOD 1)

The analysis of variance method of estimating variance components is essentially the only method in use for balanced data. It is also the most frequently used method with unbalanced data, although its application is not as straightforward and its deficiencies are more pronounced. Nevertheless, it is likely to continue as an oft-used method and so considerable attention is devoted to it here. With balanced data the method consists of equating mean squares to their expected values. Essentially the same procedure is used with unbalanced data, as is now shown in terms of an example, the two-way crossed classification with interaction.

The model for *a* levels of an **A**-factor crossed with *b* levels of a **B**-factor is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad (2)$$

where y_{ijk} is the *k*-th observation (for *k* = 1,2, ,n_{ij}) in the *i*-th level of the **A**-factor and the *j*-th level of the **B**-factor, where *i* = 1,2,…., *a* and *j* = 1,2,….,*b*. Thus n_{ij} is the number of observations in the (*i,j*) cell- the *i*-th level of **A** and the *j*-th level of **B**. Since not all of the cells may contain observations we let *s* represent the number that do; i.e., *s* is the number of n_{ij}'s for which $n_{ij} > 0$. Thus *ab-s* is the number of cells containing no data (n_{ij} = 0). In (43), μ is a general mean, α_i is the effect due to the *i*-th level of the **A**-factor, $\boldsymbol{\beta}_j$ is the effect due to the *j*-th level of the **B**-factor, γ_{ij} is the interaction effect and e_{ijk} is the customary error term.

4. ADJUSTING FOR BIAS IN MIXED MODELS (HENDERSON METHOD 2)

Mixed models involve dual estimation problems - estimating both fixed effects and variance components. For the moment attention is directed just to estimating the variance components. In some situations this is exactly what might be done in practice; with genetic data, for example, time or year effects might be considered fixed and of little interest compared to the genetic variance components. On the other hand, time trends may be of very real interest in some data, in which case their estimation together with that of the variance components would be considered simultaneously. This dual estimation problem is considered subsequently.

The analysis of variance method for mixed model leads, with unbalanced data, to biased estimators of variance components. The method known as Method 2 in Henderson (1953) is designed to correct this deficiency. It uses the data first to estimate fixed effects of the model and then, using these estimators to adjust the data, variance components are estimated from the adjusted data, by the analysis of variance method. The whole procedure is designed so that the resulting variance component estimators are not biased by the presence of the fixed effects in the model, as they are with analysis of variance estimators derived from the basic data. So far as the criterion of unbiased ness is concerned, this is certainly achieved by Method 2. But the general method of analyzing data adjusted accordant to some estimator of the fixed effects is open to criticism on other grounds: it cannot be uniquely defined, and a simplified form of it, of which Henderson's Method 2 is a special case, cannot be used whenever the model includes interactions between the fixed effects and the random effects.

The general approach of Method 2 can be considered in terms of the model:

$$y = \mu 1 + X_f \beta_f + X_r \beta_r + e$$

where all fixed effects other than μ are represented by $\boldsymbol{\beta_f}$, and all random effects by $\boldsymbol{\beta_r}$. As usual $E(\boldsymbol{\beta}_r) = \mathbf{0}$ and so $E(\beta_r \beta'_r) = \mathbf{V}(\beta_r)$, the variance-covariance matrix of the random effects. The general effect of correcting the data vector y accordant to an estimator of the fixed effects β_f is to suppose that such an estimator is $\tilde{\beta}_f = Ly$ for some matrix $\boldsymbol{L}$ so that the vector of corrected data is $z = yX_f \tilde{\beta}_f$. It can then be shown (Searle 1992) that the model for z contains no terms in β_f provided $\boldsymbol{L}$ is a generalized inverse of $\boldsymbol{X_f}$. Under this condition the analysis of variance method applied to $\mathbf{y}$ - $X_f \tilde{\beta}_f$. will yield unbiased estimators of the variance components. However, the fact that $\mathbf{L}$ has only to be a generalized inverse of $\boldsymbol{X_f}$ indicates the arbitrariness of the method. This lack of specificity means the method is not uniquely defined and hence is impractical.

The model for $\mathbf{z}$ = $\mathbf{y}$ - $X_f \tilde{\beta}_f$. just described contains no term in $\boldsymbol{\beta_f}$. An additional restriction would be for the model to have the same term $\boldsymbol{X_r} \boldsymbol{\beta_r}$ as does the model for $\mathbf{y}$, as wall as a mean term $\mu_1 1$ where μ_1 is not necessarily equal to μ.

5. THE FITTING CONSTANTS METHOD (HENDERSON'S METHOD 3)

The third method described by Henderson is based on the method of fitting constants traditionally used in fixed effects models. It uses reductions in sums of squares due to fitting different subgroups of factors in the model, using them in exactly the same manner as the S's are used in the analysis of variance method, namely estimating the variance components by equating each computed reduction to its expected value.

In an example of the two-way classification random model with interaction, the four quadratics equated to their expected values in the analysis of variance method are S_A, S_B, S_{AB} and SSE shown as functions of T's in equations. The fitting constants method also uses four quadratics, derived from fitting the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

and three sub-models:

$$y_{ijk} = \mu + e_{ijk}\,,\ y_{ijk} = \mu + \alpha_i + e_{ijk}\,,\ y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

In fitting these models the total sum of squares is $T_0 = \Sigma\Sigma\Sigma\ y^2_{ijk}$. With each model there is a reduction in sum of squares, $\mathbf{y'X(X'X)^{-}X'y}$, due to fitting the model. These reductions can be donated by:

$$R(\mu, A, B, AB)\,,\ R(\mu)\,,\ R(\mu, A)\,,\quad R(\mu, A, B)$$

respectively, where the letters in parentheses indicate the factors fitted in the respective models. (In this notation AB represents, as usual, A-by-B interaction). By way of example, the last of the above models fits μ, A-, and B- factors, and so the reduction is symbolized as R(μ,A,B). Writing the model as $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where $\mathbf{b}$ is the vector containing μ, the α's and β's we have R(μ,A,B) = $\mathbf{y'X(X'X)^{-}X'y}$ for that $\mathbf{X}$. The fitting constants method of estimating variance components uses T_0 and these R( ) - reductions by equating certain differences among them to their expected values, so setting up equations in the σ^2's whose solutions are the estimators.

6. CONCLUSION

Henderson's approach to variance components estimation for unbalanced data was the foundation paper for other statisticians in attempt to solve this problem. The methods there described have, accordingly, often been referred to as Henderson's methods 1,2 and 3. Method 1 is simply an analogue of the analysis of variance method used with balanced data; Method 2 is designed to correct a deficiency of Method 1 that arises with mixed models; and Method 3 is based on the method of fitting constants so often used in fixed effects models.

REFERENCES

1. Christensen, R. 1996. Analysis of Variance, Design and Regression. London: Chapman&Hall.
2. Eisenhart, C. 1947 "The assumptions underling the analysis of variance." *Biometrics* 3: 1-21.
3. Hartley, H.O. 1967. "Expectations variances and covariance's of ANOVA mean squares by "synthesis." *Biometrics* 25: 105-114.
4. Hartley, H.O. and Rao, J.N.K. 1967. "Maximum likelihood estimation for the mixed analysis of variance models." *Biometrics* 23: 93-108.
5. Hartley, H.O. and Searle, R.R. 1969. "A discontinuity in mixed model analysis ." *Biometrics* 25: 573-576
6. Henderson, C.R. 1953 "Estimation of variance and covariance components." *Biometrics* 9: 226-252.
7. Hoaglin, D.C., Mosteller, F., and Tukey, J.W. 1991. *Fundamentals of Exploratory Analysis of Variance*. New York; Wiley.
8. Hockihg, R.R., Green, J.W. and Bremer, R.H. 1989. "Variance component estimation with model-based diagnostics." *Tehnomestrics* 31: 227-239.
9. Hocking, R.R. 1996. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. New York: Wiley.
10. Kempthorne, O. 1975. "Fixed and mixed models in the analysis of variance." *Biometrics* 31:473-486.
11. Khuri, A.I., and Sahai, H. 1985. "Variance components analysis: Selective literature review." *International Statistics* 11: 279-300.
12. Miller, R.G. 1997. *Beyond ANOVA*. Chapman&Hall
13. Sahai, H. 1979. "A Bibliography on variance components." *International Statistical Reviews* 47: 177-222.
14. Sahai H., Ageel M., 2001. "*The analysis of variance*", Birkhauser,
15. Sheffe, H. 1959. *The analysis of Variance*. New York: Wiley.
16. Searle, S.R. 1969. "Another look at Hendersons methods of estimating variance components." *Biometrics* 24: .749-788.
17. Searle, S.R. 1971. "Topics in variance component estimation." *Biometrics* 27: 1-76.
18. Searle, S.R., Casella, G., and McCulloch, C.E. 1992. *Variance Components*. New York: Wiley.
19. Thompson, W.A.Jr., and Moore, J.R. 1963. Non-negative estimates of variance components. *Tehnometrics* 5: 441-450.

HENDERSONOV PRISTUP OCENI VARIJANSNIH KOMPONENTI ZA NEURAVNOTEŽENE PODATKE

Vera Djordjević, Vinko Lepojević

Analiza varijanse je obilato korišćena tehnika analize podataka i verovatno najčešće korišćeni statistički metod u dvadesetom veku. No, i pored toga, još uvek postoje nejasnoće u tumačenju pojedinih njenih aspekata. Ovaj rad ima za cilj osvetljavanje jednog od fundamentalnih delova ovog metoda - ocenu varijansnih komponenti. Naime, u cilju izračunavanja varijabiliteta u posmatranjima možemo razložiti njihovu varijansu na delove koji odgovaraju faktorima koje proučavamo, a to u stvari znači da predmet proučavanja predstavljaju varijansne komponente. U radu je prezentovan kritički pristup Henderson-ov metode ocene varijansnih komponenti u modelima analize varijanse.