A Simple Argument Showing How to Derive Restricted Maximum Likelihood

J. L. FOULLEY[1]
National Institute for Agricultural Research
Quantitative and Applied Genetics Unit
78352 Jouy-en-Josas Cedex, France

ABSTRACT

This paper presents a pedagogical argument to explain to quantitative geneticists how REML can be derived from maximum likelihood for estimation of variance components. The argument is first developed for N independent normal observations with mean μ and variance σ^2 and is afterward extended to a general linear mixed model structure, $\mathbf{y} \sim N(\mathbf{X}\beta,\mathbf{V})$. The argument is taken from expectation-maximization theory and consists of replacement of a quadratic in μ or β by its conditional expectation given the observed data and the variance components.

(**Key words**: variance components, mixed model, restricted maximum likelihood, expectation-maximization)

Abbreviation key: **EM** = expectation-maximization, **MINQUE** = minimum norm quadratic unbiased estimation, **ML** = maximum likelihood.

INTRODUCTION

Restricted maximum likelihood, introduced by Patterson and Thompson (15), has become the method of choice in animal breeding for estimation of variance components of Gaussian linear mixed models because of the development of computer technology and the availability of simple and efficient algorithms based on Henderson's (9) mixed model equations [see reviews (14) and (16)].

On theoretical grounds, justifications given in review articles or textbooks (19, 20, 21) rely mainly on the fact that degrees of freedom associated with estimation of fixed effects are not taken into account with maximum likelihood (**ML**). Then, REML estimators are obtained from maximization of only the part of the likelihood that does not depend on the fixed effects (15) or, equivalently, the likelihood of "error contrasts" free of fixed effects (8). Derivation of the maximum for variance components of such functions is technically complex (18). Furthermore, the arguments used to justify REML are not completely transparent. In particular, it is difficult to understand why the maximization of the likelihood can be restricted to the part that is free of the fixed effects and the remainder can be eliminated without loss of information. In other words, according to Kalbfleisch and Sprott (11), this last part of the likelihood "contains no information about the parameter of interest (variance components) in the absence of knowledge of nuisance parameters (fixed effects)". This property is related to the classical concepts of marginal likelihood and generalized sufficiency (1, 10) and involves for its proof complex mathematical arguments (e.g., homomorphic transformation groups) that are far beyond the training and interest of most geneticists. Even well-known statisticians remain seemingly doubtful about the value of such proofs; for instance, McCullagh and Nelder (13) have said "In this example, there appears to be no loss of information on θ (here 'variance components') by using $\mathbf{R}$ ('contrasts') in place of $\mathbf{Y}$ ('data'), though it is difficult to give a totally satisfactory justification of this claim" (p. 247).

However, Harville (7) showed that REML can be viewed as the mode of the marginal posterior distribution of variance components with flat priors on fixed effects and variance components. Although coherent and pertinent, this Bayesian interpretation of REML is seldom discussed in the animal breeding literature despite some recent advocacy for Bayesian methods (5).

Received November 9, 1992.
Accepted March 19, 1993.
[1]Group of Statistical Genetics.

The purpose of this paper is mainly to provide quantitative geneticists and animal breeders with a simple argument showing how REML can be introduced starting from ML theory to correct some of its potential drawbacks. The reasoning is presented first in the case of N independent observations from a normal distribution with mean μ and variance σ^2 and later extended to a general linear mixed model structure.

MATERIALS AND METHODS

Theory

N Independent Observations. Let $\mathbf{y} = \{y_i\}$ be the $(N \times 1)$ vector of observations, $y_i \sim NIID(\mu,\sigma^2)$, assumed normally, independently, and identically distributed with mean μ and variance σ^2. The density of the data vector is

$$p(\mathbf{y}|\mu,\sigma^2) = \prod_{i=1}^{N} p(y_i|\mu,\sigma^2)$$

$$= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left[-\sum_{i=1}^{N} \frac{(y_i - \mu)^2}{2\sigma^2} \right], \quad [1]$$

and the log-likelihood can be written as

$$L(\mu,\sigma^2;\mathbf{y}) = -\left[N(\ln 2\pi + \ln\sigma^2) + \sigma^{-2} \sum_{i=1}^{N} (y_i - \mu)^2 \right]/2. \quad [2]$$

Let

$$\bar{y} = \left(\sum_{i=1}^{N} y_i \right)/N,$$

and

$$s^2 = \left[\sum_{i=1}^{N} (y_i - \bar{y})^2 \right]/N$$

designate the sample mean and variance, respectively, so

$$\sum_{i=1}^{N} (y_i - \mu)^2$$

can be expressed as

$$\sum_{i=1}^{N} (y_i - \mu)^2 = Ns^2 + N(\bar{y} - \mu)^2. \quad [3]$$

Hence,

$$L(\mu,\sigma^2;\mathbf{y}) = -\frac{N}{2}\left[\ln 2\pi + \ln\sigma^2 + \frac{s^2 + (\bar{y} - \mu)^2}{\sigma^2} \right]. \quad [4]$$

Differentiating for μ and σ^2, respectively, leads to

$$\partial L(\mu,\sigma^2;\mathbf{y})/\partial\mu = N(\bar{y} - \mu)/\sigma^2, \quad [5a]$$

$$\partial L(\mu,\sigma^2;\mathbf{y})/\partial\sigma^2 = -(N/2\sigma^2)\left[1 - \frac{s^2 + (\bar{y} - \mu)^2}{\sigma^2} \right], \quad [5b]$$

and equating to zero gives, for $N \geq 2$, the usual ML estimates of μ and σ^2, i.e.,

$$\hat{\mu} = \bar{y}, \quad [6a]$$

$$\hat{\sigma}^2_{ML} = s^2. \quad [6b]$$

Had μ been known, the ML estimator of σ^2 would have been obtained when solving for σ^2, the expression in [5b] set to zero, i.e.,

$$\hat{\sigma}^2 = s^2 + (\bar{y} - \mu)^2, \quad [7]$$

which is always equal to or larger than s^2, the ML estimator.

In fact, μ is unknown, and the contribution to $(\bar{y} - \mu)^2$ to [7] is evaluated with ML by assuming that μ equals its ML estimate $\bar{y}$ (i.e., by a nil contribution). An alternative procedure is to use ideas behind the expectation-maximization (EM) theory (2, 12) and its Bayesian extension to evaluate more properly the contribution of $(\bar{y} - \mu)^2$ to the estimator of σ^2.

If μ belongs to the complete data (2), σ^2 is unbiasedly estimated from its sufficient statis-

tic $\sum_{i=1}^{N} (y_i - \mu)^2/N$, which, because μ is missing, is replaced by its conditional expectation $E[N^{-1} \sum_{i=1}^{N} (y_i - \mu)^2|y,\sigma^2]$, given the observed data by $s^2 + E[(\bar{y} - \mu)^2|y,\sigma^2]$ (Equation [3]). In practice, the EM algorithm proceeds as follows. For the expectation (E) step, at round [n], the unknown quantity $(\bar{y} - \mu)^2$ is replaced by its conditional expectation $E[(\bar{y} - \mu)^2|y,\sigma^{2[n]}]$ given the data y and the current value $\sigma^{2[n]}$ of the variance; this expectation is taken with respect to the distribution of μ to take into account uncertainty for that parameter. For the maximization (M) step, compute the next value $\sigma^{2[n+1]}$ from Equation [7], i.e.,

$$\sigma^{2[n+1]} = s^2 + E[(\bar{y} - \mu)^2|y,\sigma^{2[n]}]. \quad [8]$$

Now, $\frac{\bar{y} - \mu}{\sigma/\sqrt{N}} \sim N(0,1)$ can be interpreted either 1) classically as $\bar{y}|\mu,\sigma^2 \sim N(\mu,\sigma^2/N)$ or 2) from a noninformative Bayesian (or fiducial) method as, $\mu|\bar{y},\sigma^2 \sim N(\bar{y},\sigma^2/N)$. Using the latter interpretation,

$$E[(\bar{y} - \mu)^2|y,\sigma^2] = \mathrm{Var}(\mu|\bar{y},\sigma^2) = \sigma^2/N, \quad [9]$$

so that Equation [8] reduces to

$$\sigma^{2[n+1]} = s^2 + (\sigma^{2[n]}/N), \quad [10]$$

which, at the limit for $\hat{\sigma}^2 = \sigma^{2[n+1]} = \sigma^{2[n]}$, leads to the usual REML unbiased estimator of σ^2:

$$\hat{\sigma}^2 = Ns^2/(N - 1). \quad [11]$$

Equations [8] and [10] clearly illustrate the difference between ML and REML and show the bias removed by REML. For ML, the contribution of $(\bar{y} - \mu)^2$ to the estimator is obtained by pretending that μ equals its ML estimate $\bar{y}$ (i.e., by a nil contribution), whereas, for REML, this quadratic form is replaced by its conditional expectation given the data observed.

Linear Mixed Model Structure. The argument developed for this simple situation can be extended to a general mixed linear structure under the normality assumption, i.e., with

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta},\mathbf{V}), \quad [12]$$

where $\boldsymbol{\beta}$ is a (p × 1) vector of fixed effects, $\mathbf{X}$ is the corresponding (N × p) incidence matrix (assumed to have full column rank for simplicity), and $\mathbf{V} = \sum_{k=1}^{K} \mathbf{V}_k\sigma_k^2$, is an (N × N) variance covariance matrix linear for the K variance components (σ_k^2; k = 1, 2, . . ., K) where $\mathbf{V}_k$ is an (N × N) matrix of known elements.

With $\sigma^2 = \{\sigma_k^2\}$, the log-likelihood is given by

$$L(\boldsymbol{\beta},\sigma^2;\mathbf{y}) = \text{const} - \frac{1}{2} \ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad [13]$$

Given that $\boldsymbol{\beta}$ is known, the ML of σ^2 is the solution of the following system (for k = 1, 2, . . ., K)

$$\frac{\partial}{\partial\sigma_k^2} L(\boldsymbol{\beta},\sigma^2;\mathbf{y}) = -\frac{1}{2}\frac{\partial}{\partial\sigma_k^2} \ln|\mathbf{V}| - \frac{1}{2}\frac{\partial}{\partial\sigma_k^2}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] = 0. \quad [14]$$

The algebraic identity in Equation [3] can be extended to the quadratic form $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ (6, 24) as follows:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \quad [15]$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is the generalized least squares estimator of $\boldsymbol{\beta}$. Using Equation [15] in Equation [14] leads to

$$\frac{\partial}{\partial\sigma_k^2} L(\boldsymbol{\beta},\sigma^2;\mathbf{y}) = -\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{V}_k) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = 0. \quad [16]$$

The ML solutions for σ^2 are obtained from Equation [16] by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, thus setting the last quadratic form to zero. As previously, this quadratic can alternatively be replaced by its conditional expectation given y and V.

Because the generalized least squares estimator of $\boldsymbol{\beta}$ is normally distributed with mean $\boldsymbol{\beta}$ and variance $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ or alternatively $\boldsymbol{\beta}|\mathbf{y},\mathbf{V} \sim N[\hat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}]$ (assuming a "flat" prior distribution on $\boldsymbol{\beta}$), the expectation of this quadratic form with respect to that distribution is equal to $\frac{1}{2}\text{tr}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}_k\ \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}]$. When this expression is entered into Equation [16], $\mathbf{V}_k$ is factored, and the P matrix of Searle et al. (21) (i.e., $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$) is introduced, the following system is obtained:

$$-\frac{1}{2}\text{tr}(\mathbf{PV}_k) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \qquad [17a]$$

or

$$\text{tr}(\mathbf{PV}_k) = \mathbf{y}'\mathbf{PV}_k\mathbf{Py}. \qquad [17b]$$

These REML equations are exactly those derived classically [see, e.g., Searle, (18), formulas 5.14 and 5.15, and Searle et al. (21), formula 89, page 251]. More detail about an explicit description of the expectation and maximization steps is given in the appendix.

DISCUSSION

The EM algorithm has been widely used for calculation of ML or REML estimates of variance components with linear models (12, 14, 16). As shown in this paper, this procedure is also helpful to explain the difference between ML and REML. The only mathematical operation required to go from ML to REML reduces to taking the expectation of a quadratic form. Moreover, this derivation shows that, although REML is generally biased because of parameter space constraints (20), it nevertheless removes some bias from ML, as already indicated (22).

Although this procedure is pedagogically self-contained, it can be fully justified more theoretically. Indeed, it can be shown (3, 4) that

$$\frac{\partial}{\partial\sigma_k^2}\ln\ p(\mathbf{y}|\sigma^2) = E_c\left[\frac{\partial}{\partial\sigma_k^2}\ln\ p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\right], \qquad [18]$$

where $p(\mathbf{y}|\sigma^2)$ is the marginal density of y after $\boldsymbol{\beta}$ is integrated out (assuming a flat prior distribution), and E_c indicates a conditional expectation taken with respect to the posterior density of $\boldsymbol{\beta}$ given σ^2.

The argument developed in this paper consists of computing the right-hand side of Equation [18] and equating it to zero. This procedure is equivalent, because of this identity, to setting to zero the derivative of the log-marginal likelihood with respect to σ^2, i.e., to compute the maximum marginal likelihood of σ^2 (or the mode of its posterior density with a flat prior on σ^2). In that respect, our argument can be considered to be Bayesian as in Harville (7), but without recourse to a direct integration of the fixed effects. Other arguments for introduction and justification of REML can be used, such as the equivalence with iterative minimum norm quadratic unbiased estimation (**MINQUE**) (21) or the formal identity of EM formulas with those of the pseudoexpectation approach (17, 23) applied to MINQUE quadratics.

ACKNOWLEDGMENTS

This paper draws widely from material presented during a course on variance component estimation taught by the author in Paris XI University for the "Diplome d'Etudes Approfondies" in Quantitative and Population Genetics. Thanks are expressed to J. Genermont and E. Verrier for having scheduled that course and to the students for their questions and comments on those topics. The author is also indebted to J. J. Colleau, V. Ducrocq, R. L. Quaas, M. San Cristobal, and two anonymous referees for their critical reading of the manuscript.

REFERENCES

1 Dawid, A. P. 1980. A Bayesian look at nuisance parameters. Page 167 *in* Bayesian Statistics. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A.F.M. Smith, ed. Univ. Press, Valencia, Spain.

2 Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM

algorithm. J. R. Stat. Soc. B 39:1.
3 Foulley, J. L., S. Im, D. Gianola, and I. Hoeschele. 1987. Empirical Bayes estimation of parameters for n polygenic binary traits. Genet. Sel. Evol. 19:197.
4 Foulley, J. L., M. San Cristobal, D. Gianola, and S. Im. 1992. Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. Comput. Stat. Data Anal. 13:291.
5 Gianola, D., and R. L. Fernando. 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63: 217.
6 Gianola, D., J. L. Foulley, and R. L. Fernando. 1986. Prediction of breeding values when variances are not known. Genet. Sel. Evol. 18:485.
7 Harville, D. A. 1974. Bayesian inference for variance components using only error contrasts. Biometrika 61: 383.
8 Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72:320.
9 Henderson, C. R. 1984. Applications of linear models in animal breeding. Univ. Guelph, Guelph, ON, Canada.
10 Kalbfleisch, J. D. 1986. Pseudo-likelihood. Page 324 *in* Encyclopedia of Statistical Science. Vol. 7. S. Kotz and N. L. Johnson, ed. John Wiley & Sons, New York, NY.
11 Kalbfleisch, J. D., and D. A. Sprott. 1970. Application of the likelihood methods to models involving large numbers of parameters (with discussion). J. R. Stat. Soc. B 32:175.
12 Little, R.J.A., and D. B. Rubin. 1987. Statistical Analysis with Missing Data. John Wiley & Sons, New York, NY.
13 McCullagh, P., and J. Nelder. 1989. Generalized Linear Models. 2nd ed. Chapman and Hall, London, Engl.
14 Meyer, K. 1990. Present status of knowledge about statistical procedures and algorithms to estimate variance and covariance components. Proc. 4th World Congr. Genet. Appl. Livest. Prod., Edinburgh, Scotland XIII:407.
15 Patterson, H. D., and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. Biometrika 58:545.
16 Quaas, R. L. 1992. REML NoteBook. Dep. Anim. Sci., Cornell Univ., Ithaca, NY.
17 Schaeffer, L. R. 1986. Pseudoexpectation approach to variance component estimation. J. Dairy Sci. 69:2884.
18 Searle, S. R. 1979. Notes on variance component estimation. A detailed account of maximum likelihood and kindred methodology. Paper BU-673-M, Cornell Univ., Ithaca, NY.
19 Searle, S. R. 1987. Linear Models for Unbalanced Data. John Wiley & Sons, New York, NY.
20 Searle, S. R. 1989. Variance components—some history and a summary account of estimation methods. J. Anim. Breed. Genet. 106:1.
21 Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance Components. John Wiley & Sons, New York, NY.
22 Thompson, R. 1989. REML. Biometric Bull. 6:4.
23 Van Raden, P. M., and Y. C. Young. 1988. A general purpose approximation to restricted maximum likelihood: the tilde-hat approach. J. Dairy Sci. 71:187.
24 Zellner, A. 1971. An Introduction to Bayesian Inference in Econometrics. John Wiley & Sons, New York, NY.

APPENDIX

Foulley et al. (3) have shown that $E_c\left[\frac{\partial}{\partial\sigma_k^2}\ln p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\sigma}^2)\right]$ can be evaluated iteratively using

$$\frac{\partial}{\partial\sigma_k^2}E_c^{[n]}[\ln\ p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\sigma}^2)], \qquad [A1]$$

where $E_c^{[t]}[.]$ indicates that the expectation of the quantity within brackets (viewed as a function of $\boldsymbol{\beta}$) has to be taken with respect to the posterior density of $\boldsymbol{\beta}$, given that $\boldsymbol{\sigma}^2$ equals its value $\boldsymbol{\sigma}^{2[n]}$ at the last iteration [n].

Actually this formula defines the expectation and maximization steps of the EM algorithm for computing the REML estimates of $\boldsymbol{\sigma}^2$, using $\ln p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\sigma}^2)$ as the sufficient statistics for $\boldsymbol{\sigma}^2$ for the complete data (2).

The expectation step consists of evaluating $E_c^{[t]}[\ln p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\sigma}^2)]$, i.e., of computing

$$Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2[n]}) = -\frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[n]})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[n]}) - \frac{1}{2}\mathrm{tr}\{\mathbf{V}^{-1}\mathbf{X}[\mathbf{X}'(\mathbf{V}^{[n]})^{-1}\mathbf{X}]^{-1}\mathbf{X}'\}, \qquad [A2]$$

where, as in Dempster et al. (2), $Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2[n]}) = E_c^{[n]}[\ln p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\sigma}^2)]$, and $\boldsymbol{\beta}^{[n]}$ is solution of $\mathbf{X}'(\mathbf{V}^{[n]})^{-1}\mathbf{X}\boldsymbol{\beta}^{[n]} = \mathbf{X}'(\mathbf{V}^{[n]})^{-1}\mathbf{y}$, and $\mathbf{V}^{[n]} = \mathbf{V}(\boldsymbol{\sigma}^{2[n]})$.

The maximization step consists of maximizing $Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2[n]})$, by solving the equation $\frac{\partial}{\partial\sigma_k^2}Q(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2[n]}) = 0$ for $\boldsymbol{\sigma}^2$. This solution is achieved by selection of $\boldsymbol{\sigma}^{2[n+1]}$ such that

$$-\frac{1}{2}\mathrm{tr}[(\mathbf{V}^{[n+1]})^{-1}\mathbf{V}_k] + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[n]})'(\mathbf{V}^{[n+1]})^{-1}\mathbf{V}_k(\mathbf{V}^{[n+1]})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{[n]}) + \frac{1}{2}\mathrm{tr}\{(\mathbf{V}^{[n+1]})^{-1}\mathbf{V}_k(\mathbf{V}^{[n+1]})^{-1}\mathbf{X}[\mathbf{X}'(\mathbf{V}^{[n]})^{-1}\mathbf{X}]^{-1}\mathbf{X}'\} = 0. \qquad [A3]$$

At convergence for $\hat{\mathbf{V}} = \mathbf{V}^{[n]} = \mathbf{V}^{[n+1]}$, Equation [A3] reduces to Equation [17a].