

READER REACTION

Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms

H.-P. Piepho

Institut für Nutzpflanzenkunde, Universität Kassel,
D-37213 Witzenhausen, Germany

SUMMARY

This note considers an REML procedure for mixed models with multiplicative terms that was recently suggested by Gogel, Cullis, and Verbyla (1995, *Biometrics* 51, 744-749). It discusses an extension to certain heteroscedastic mixed models, which are of interest for plant breeders.

1. Introduction

Genotype-environment ($G \times E$) data from plant cultivar trials are frequently analyzed by a two-way linear model of the form

$$y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon_{ij} \quad (i = 1, \dots, I; j = 1, \dots, J), \quad (1.1)$$

where y_{ij} = mean yield of i th genotype in j th environment, μ = general mean, g_i = effect of i th genotype, e_j = effect of j th environment, $(ge)_{ij}$ = interaction of i th genotype and j th environment, ε_{ij} = random error associated with means y_{ij} , assumed to be distributed as $N(0, \sigma^2)$. This model is appropriate for the analysis of means from equally replicated data with homoscedastic errors. In an analysis of $G \times E$ data, it may be useful to fit a more specific model describing the interaction. The most common of such models is the regression model first suggested by Yates and Cochran (1938), which was further elaborated by Finlay and Wilkinson (1963) and Eberhart and Russell (1966). It may be written as

$$y_{ij} = \mu + g_i + e_j + a_i e_j + \varepsilon_{ij}, \quad (1.2)$$

where a_i (subject to the estimability constraint $\sum a_i = 0$) is a regression coefficient corresponding to the i th genotype. Thus, the interaction is modeled by a regression on the environmental effect. Note that if the regression does not explain all interaction, the residual ε_{ij} comprises both experimental error and components of interaction not explained by the regression. It is often assumed that both genotypes and environments are fixed factors. However, in some cases it may be reasonable to regard testing environments as a random sample from a population of environments. Gogel, Cullis, and Verbyla (1995) have suggested an REML procedure to estimate the Finlay-Wilkinson (F-W) model in (1.2) under this assumption. This procedure is expected to prove very useful in practical applications, particularly because unbalanced data can be handled.

The F-W model is somewhat restrictive in that the independent variable of the regression is taken to be the environmental effect. It is often found in practice that this model captures only a small part of the genotype-environment interaction (Crossa, 1990). A more general and flexible model is given by

$$y_{ij} = \mu + g_i + e_j + \theta \alpha_i \beta_j + \varepsilon_{ij} \quad (1.3)$$

Key words: Factor analytic covariance structure; Finlay-Wilkinson regression; REML; Stability variance.

(Williams, 1952; Mandel, 1971), where α_i and β_j are scores for genotype i and environment j , respectively, and θ is a scalar constant. Extension to more than one multiplicative term is straightforward. These models have become popular in agricultural and biological applications as biadditive models (Denis and Gower, 1994) and as additive main effects multiplicative interaction (AMMI) models (Gauch, 1988). So far they have been mainly used in a fixed model framework.

Another generalization of the F-W model is to drop the assumption of homogeneous residual variance and assign a separate variance for every genotype (Shukla, 1972), i.e.,

$$\text{var}(\varepsilon_{ij}) = \sigma_i^2 \quad \text{for every } i. \quad (1.4)$$

The variance σ_i^2 may be regarded as a measure of stability (stability variance) of the i th genotype. Such models are often used to analyze cultivar trial data (Kang and Gauch, 1996; Piepho, 1995).

The purpose of this paper is to show how procedures like those suggested by Oman (1991) and Gogel et al. (1995) can be used to fit mixed models with multiplicative terms that entail the generalizations sketched above. Emphasis will be on modeling the covariance structure. All computations have been performed using standard statistical software (SAS).

2. The Mixed Model Perspective

First consider the mixed model version of the F-W model as considered by Gogel et al. (1995) in an analysis of wheat yield data by Digby (1979). The model may be rewritten as

$$y_{ij} = \mu + g_i + (a_i + 1)e_j + \varepsilon_{ij} = \mu + g_i + b_i e_j + \varepsilon_{ij}, \quad (2.1)$$

where $b_i = 1 + a_i$. It is assumed that the random terms e_j and ε_{ij} are independently normally distributed with zero mean and variances σ_e^2 and σ^2 , respectively. Note that this model does not facilitate explicit estimation of an error variance component because ε_{ij} subsumes both experimental error and unexplained interaction. This is acceptable when errors are homoscedastic and data are equireplicated. The vector of genotype means in the j th environment, $\mathbf{y}_j = (y_{1j}, \dots, y_{Ij})'$, is multivariate normal with variance-covariance matrix

$$\text{var}(\mathbf{y}_j) = \lambda \lambda' + \mathbf{D}, \quad (2.2)$$

where $\mathbf{D} = \sigma^2 \mathbf{I}_I$ and λ is an I -row vector with elements $\lambda_i = (1 + a_i)\sigma_e = b_i \sigma_e$. The covariance structure expressed in terms of b_i and σ_e is overparameterized, so the restriction $\bar{b} = 1$ must be imposed to ensure estimability. Because of this overparameterization it is more feasible to first estimate λ_i and then use the relations

$$b_i = \lambda_i / \bar{\lambda}, \quad \sigma_e^2 = (\bar{\lambda})^2 \quad (2.3)$$

(Oman, 1991). The parameterization in (2.2) is essentially the factor-analytic covariance structure suggested by Jennrich and Schluchter (1986), who showed how to obtain ML estimates of the variance-covariance parameters. Gogel et al. (1995) used a scoring algorithm (the average information algorithm of Gilmour, Thompson, and Cullis, 1995) to obtain REML estimates of λ_i and σ^2 . Alternatively, a Newton-Raphson algorithm may be used (Wolfinger, Tobias, and Sall, 1994). Statistical packages like SAS and BMDP have facilities to fit factor-analytic models as in (2.2) by likelihood-based methods.

Now consider the multiplicative model in (1.3). If environments are considered random, main effects e_j and scores β_j may be taken to be independently normally distributed with zero mean and variances σ_e^2 and σ_β^2 , respectively. Again, the covariance structure is overparameterized, so restrictions need to be imposed, for example $\theta = \bar{\alpha} = 1$, similar to the Finlay-Wilkinson regression. The variance-covariance structure may be written as

$$\text{var}(\mathbf{y}_j) = \sigma_e^2 \mathbf{J}_I + \lambda \lambda' + \mathbf{D}, \quad (2.4)$$

where \mathbf{J}_I is an $I \times I$ matrix of ones everywhere and $\mathbf{D} = \sigma^2 \mathbf{I}_I$. Now, the row-vector λ has elements $\lambda_i = \theta \alpha_i \sigma_\beta = \alpha_i \sigma_\beta$. This model may be estimated by ML or REML methods analogous to those cited above. In the following discussion, model (1.3) with the covariance structure in (2.4) will be referred to as Mandel's model (though it is emphasized that Mandel originally considered both factors of his two-way model as fixed). Extension to more than one multiplicative term is possible. If there are R multiplicative terms, λ may be replaced by an $I \times R$ matrix Λ of unknown coefficients. If $R > 1$, estimability is attained by constraining $R(R-1)/2$ of these coefficients to zero (Jennrich and Schluchter, 1986).

Standard errors for the estimates of λ_i and σ^2 in (2.2) and of σ_e^2 , λ_i , and σ^2 in (2.4) can be computed from the inverse of the Fisher information matrix or from the inverse of the empirical information matrix (Jennrich and Schluchter, 1986). The parameters σ_e^2 and b_i of the F-W model (collected in a vector η , say) or of σ_β^2 and α_i in Mandel's model may be expressed as functions of $\lambda = (\lambda_1, \dots, \lambda_I)'$. By a result on asymptotically normal random variables given in Seber (1984, p. 532), we have

$$\text{var}(\hat{\eta}) \approx \mathbf{F} \text{var}(\hat{\lambda}) \mathbf{F}', \tag{2.5}$$

where \mathbf{F} is the matrix of partial derivatives of η with respect to λ , evaluated at the current estimates. This approach has been used by Oman (1991) to fit similar models (Oman, personal communication). Gogel et al. (1995) apparently used a similar method.

It may also be useful to consider modifications of the residual variance matrix \mathbf{D} , which so far has been set equal to $\sigma^2 \mathbf{I}_I$, as for of the F-W model. For example, all diagonal elements may be different, which leads to models similar to those given by Shukla (1972). This structure for \mathbf{D} can also be introduced in equation (2.4) to extend Mandel's model. The diagonal elements may be taken as a measure of stability (stability variance) of the corresponding cultivars. Note that the

Table 1
REML estimates (asymptotic standard errors)^a for different multiplicative models (data by Digby, 1979)

	F-W		Shukla ^c		Mandel ^d
$\text{var}(y_j)^b$	$\sigma_e^2 \lambda \lambda' + \sigma^2 \mathbf{I}_I$		$\sigma_e^2 \lambda \lambda'$		$\sigma_e^2 \mathbf{J}_I + \sigma_\beta^2 \lambda \lambda' + \sigma^2 \mathbf{I}_I$
			+ $\text{diag}(\sigma_1^2, \dots, \sigma_I^2)$		
-2LL ^e	129.4734		112.7636		114.4961
σ_e^2	1.0580 (0.3781)	σ_e^2	1.0682 (0.3814)	σ_e^2	0.5813 (0.2420)
				σ_β^2	0.4808 (0.3394)
b_1	0.9557 (0.0648)	b_1	0.9426 (0.0652)	α_1	0.8844 (0.0960)
b_2	0.7414 (0.0642)	b_2	0.7312 (0.0633)	α_2	0.6219 (0.1381)
b_3	1.0878 (0.0792)	b_3	1.0998 (0.0901)	α_3	1.3014 (0.2060)
b_4	1.0299 (0.0789)	b_4	1.0470 (0.0443)	α_4	1.0704 (0.1461)
b_5	1.1457 (0.0654)	b_5	1.1472 (0.0618)	α_5	1.3580 (0.1442)
b_6	0.8798 (0.0663)	b_6	0.8574 (0.1004)	α_6	0.5575 (0.1744)
b_7	1.0922 (0.0663)	b_7	1.0958 (0.0418)	α_7	1.1663 (0.1039)
b_8	0.9152 (0.1117)	b_8	0.9066 (0.0936)	α_8	0.8214 (0.1758)
b_9	1.2005 (0.0669)	b_9	1.2018 (0.0579)	α_9	1.3506 (0.1392)
b_{10}	0.9517 (0.0790)	b_{10}	0.9699 (0.0775)	α_{10}	0.8683 (0.1763)
σ^2	0.0737 (0.0100)	σ_1^2	0.0765 (0.0301)	σ^2	0.0517 (0.0077)
		σ_2^2	0.0727 (0.0275)		
		σ_3^2	0.0980 (0.0523)		
		σ_4^2	0.0150 (0.0117)		
		σ_5^2	0.0660 (0.0285)		
		σ_6^2	0.1880 (0.0714)		
		σ_7^2	0.0211 (0.0137)		
		σ_8^2	0.0516 (0.0333)		
		σ_9^2	0.0518 (0.0239)		
		σ_{10}^2	0.0687 (0.0356)		

^a Standard errors for b_i and σ_e^2 of the F-W model and of Shukla's model, as well as for α_i and σ_β^2 of Mandel's model, were obtained from the inverse of the observed information matrix of original parameters by the approximation in (2.5).

^b $\lambda = (\lambda_1, \dots, \lambda_I)'$.

^c Same as F-W model, except that $\text{var}(\epsilon_{ij}) = \sigma_i^2$.

^d Model given by (1.3) and (2.4).

^e LL = log likelihood.

residual term ϵ_{ij} includes both experimental error and interaction not explained by the regression. Shukla (1972) assumes that errors are homoscedastic and the number of replications is the same in each environment, such that the analysis can be based on means y_{ij} and heterogeneity in σ_i^2 indicates heterogeneity in interaction variances. When replicate observations are available, the assumption of homogeneity of error variances can be tested (Piepho, 1995). In case of heteroscedastic errors and/or unequal replication, models for replicate data can be used, such as those considered by Oman (1991) and Gogel et al. (1995) for a tomato dataset, with a possible extension to allow for heteroscedasticity in the interactions.

3. Numerical Example

I will use an incomplete data set by Digby (1979), which was also employed by Gogel et al. (1995) to demonstrate their REML procedure for the Finlay–Wilkinson (F–W) model. The data consist of the yields of 10 spring wheat varieties as tested in 17 statutory trials in the U.K. in 1976. Some varieties were not tested in all locations, so that the two-way table of yields is incomplete. All computations were performed using the procedure MIXED of the SAS system (SAS, 1992, 1994), which has an option for the factor-analytic covariance structure. The REML method was used throughout. Standard errors of the F–W parameters were obtained from the asymptotic variance–covariance matrix of estimates of λ_i given by the approximation (2.5). Results for the F–W model are presented in Table 1. Parameter estimates coincide with those of Gogel et al. (1995), but standard errors differ slightly because I have employed the observed information matrix, while Gogel et al. (1995) used the expected information matrix to obtain the asymptotic variance–covariance matrix of parameters. In addition, two models were fitted: Shukla's model, which generalizes the F–W model in that \mathbf{D} is taken to be a diagonal matrix of stability variances, and Mandel's model (see results in Table 1). The F–W model and Shukla's model are nested, so that a likelihood ratio test (LRT) is appropriate to check adequacy of the more general model. The drop in minus two times the log likelihood ($-2LL$) is 16.71, with an associated p -value of 0.053 if compared to a chi-squared distribution with 9 d.f. This result gives some indication that Shukla's heteroscedastic model may be worthwhile, although the standard errors of the stability variance estimates are quite large. While Mandel's model leads to a similar drop in $-2LL$, it is more parsimonious (only one additional parameter (θ) compared to F–W) and is therefore preferable to Shukla's model. Finally, Mandel's model was fitted with heteroscedastic residual variances, i.e., with $\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_I^2)$ (results not shown). The drop in $-2LL$ compared to Mandel's model with $\mathbf{D} = \sigma^2 \mathbf{I}_I$ was 8.8056, which was not significant ($p = 0.4554$) compared to a chi-squared distribution with 9 d.f. Thus, the multiplicative term in Mandel's model seems to account for most of the heteroscedasticity in the data.

Interestingly, the sensitivities b_i of the F–W model and the coefficients α_i of Mandel's model show very similar rankings. In the present example the advantage of Mandel's model is the better fit compared to the F–W model, at the cost of just one extra parameter. Note, however, that the standard errors of estimates of b_i under both the F–W model and the Shukla model are relatively smaller than those of the estimates of α_i , so that the final decision as to the best model is ambiguous.

Table 2
REML estimates (asymptotic standard errors) of λ_i for
different multiplicative models (data by Digby, 1979)

	F–W	Shukla ^a	Mandel ^b
λ_1	0.9830 (0.1867)	0.9742 (0.1857)	0.6132 (0.2454)
λ_2	0.7626 (0.1511)	0.7564 (0.1499)	0.4312 (0.2266)
λ_3	1.1189 (0.2165)	1.1367 (0.2254)	0.9024 (0.2812)
λ_4	1.0594 (0.2067)	1.0821 (0.1965)	0.7422 (0.2716)
λ_5	1.1784 (0.2192)	1.1857 (0.2193)	0.9416 (0.2732)
λ_6	0.9050 (0.1747)	0.8862 (0.1921)	0.3865 (0.2389)
λ_7	1.1234 (0.2105)	1.1326 (0.2037)	0.8087 (0.2633)
λ_8	0.9414 (0.2082)	0.9370 (0.1960)	0.5695 (0.2647)
λ_9	1.2348 (0.2293)	1.2421 (0.2274)	0.9365 (0.2753)
λ_{10}	0.9789 (0.1938)	1.0024 (0.1971)	0.6021 (0.2730)

^a Same as F–W model, except that $\text{var}(\epsilon_{ij}) = \sigma_i^2$.

^b Model given by (1.3) and (2.4).

It is noted that interpretation does not change if estimates of λ_i are considered (see Table 2) because (up to a possible change in sign) the ranking of sensitivities b_i and α_i will be the same as that of the corresponding λ_i . In the spring wheat data, the magnitudes of b_i , α_i , and λ_i are rather similar, which is a coincidence due to the chosen unit of measurement. It is noteworthy that despite similar magnitude of parameter estimates the standard errors of b_i and α_i are considerably smaller than those of λ_i . This suggests that the approximation (2.5) may not be particularly satisfactory in the present case. Moreover, the discrepancy might indicate that the asymptotic standard errors of λ_i are not very reliable in samples of the size considered here ($J = 17$). It therefore seems worthwhile to investigate the small sample properties of REML estimates and standard errors by Monte Carlo simulation.

ACKNOWLEDGEMENT

I would like to thank a referee for a very careful review of the manuscript.

RÉSUMÉ

Cette note étudie une procédure REML pour des modèles mixtes avec des termes multiplicatifs, récemment suggérée par Gogel, Cullis et Vertbyla (1995, *Biometrics* 51, 744-749). Elle concerne une extension à certains modèles mixtes avec hétéroscédasticité qui présentent un intérêt pour les reproducteurs de plantes.

REFERENCES

- Crossa, J. (1990). Statistical analyses of multilocation trials. *Advances in Agronomy* 44, 55-85.
- Denis, J. B. and Gower, J. C. (1994). Biadditive models. *Biometrics* 50, 310-311.
- Digby, P. G. N. (1979). Modified joint regression analysis for incomplete variety \times environment data. *Journal of Agricultural Science, Cambridge* 93, 81-86.
- Eberhart, S. A. and Russell, W. A. (1966). Stability parameters for comparing varieties. *Crop Science* 6, 36-40.
- Finlay, K. W. and Wilkinson, G. N. (1963). The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research* 14, 742-754.
- Gauch, H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics* 44, 705-715.
- Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440-1450.
- Gogel, B. J., Cullis, B. R., and Verbyla, A. P. (1995). REML estimation of multiplicative effects in multi-environment variety trials. *Biometrics* 51, 744-749.
- Jennrich, R. L. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 42, 805-820.
- Kang, M. S. and Gauch, H. G. (eds). (1996). *Genotype-by-Environment Interaction*. Boca Raton, Florida: CRC Press.
- Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics* 13, 1-18.
- Oman, S. D. (1991). Multiplicative effects in mixed model analysis of variance. *Biometrika* 78, 729-739.
- Piepho, H.-P. (1995). Detecting, interpreting and handling heteroscedasticity in yield trial data. *Communications in Statistics B—Simulation and Computation* 24, 243-274.
- SAS Institute. (1992). *SAS/STAT software: Changes and enhancements*, Release 6.07. SAS Technical Report P-229, SAS Institute, Cary, North Carolina.
- SAS Institute. (1994). *SAS/STAT software: Changes and enhancements*, Release 6.10. SAS Institute, Cary, North Carolina.
- Seber, G. A. F. (1984). *Multivariate Observations*. New York: Wiley.
- Shukla, G. K. (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity* 29, 237-245.
- Williams, E. J. (1952). The interpretation of interactions in factorial experiments. *Biometrika* 39, 65-81.
- Wolfinger, R. D., Tobias, R. D., and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing* 15(6), 1294-1310.