A Conspectus on Estimating Function Theory and its Applicability to Recurrent Modeling Issues in Forest Biometry

Oliver Schabenberger and Timothy G. Gregoire

Schabenberger, O. & Gregoire, T.G. 1995. A conspectus on Estimating Function theory and its applicability to recurrent modeling issues in forest biometry. Silva Fennica 29(1): 49–70.

Much of forestry data is characterized by a longitudinal or repeated measures structure where multiple observations taken on some units of interest are correlated. Such dependencies are often ignored in favor of an apparently simpler analysis at the cost of invalid inferences. The last decade has brought to light many new statistical techniques that enable one to successfully deal with dependent observations. Although apparently distinct at first, the theory of Estimating Functions provides a natural extension of classical estimation that encompasses many ot these new approaches. This contribution introduces Estimating Function Theory as a principle with potential for unification and presents examples covering a variety of modeling issues to demonstrate its applicability.

Keywords longitudinal data, mixed models, generalized linear models, optimality.
Authors' address Virginia Polytechnic Institute and State University, Department of Forestry, Section Forest Biometrics, Blacksburg, VA 24061-0324, USA **Fax** to *Schabenberger* +1 703 231 3698 **E-mail** oschab@vt.edu
Accepted January 30, 1995

1 Introduction

Estimation in statistical models is governed by principles, such as the least squares principle, the maximum likelihood principle, Bayes principle, etc. In fitting a stipulated model to data, we explicitly or implicitly employ one of these principles together with an appropriate technique to meet a specific criterion. Computing the coefficients for a regular linear regression model with a statistical package, for example, typically invokes the ordinary least squares principle. The criterion is the minimum sum of squared residuals which is achieved by finding the roots of the derivatives of the sum of squares in closed form.

The choice of an estimation principle is not

always governed by convenience. We want also estimates that are *good* in some sense. That is, to select a principle among a suite of candidates one may formulate minimal properties the estimator must possess and properties that guide final selection. The reason for the necessity of such an approach is that in any situation no single best estimator emerges. Whatever the criterion that governs selection, it should be emphasized that the understanding of optimality depends on it and needs to be defined. There is no optimal method or estimator per se. Optimality can be achieved only when conditions such as unbiasedness, minimum variance, minimum mean square error, etc. are imposed.

The coincidence of classical estimation principles in quite different settings should prompt a curiosity to find their common denominator. The large number of alternative methods provides impetus to seek a more unifying approach to statistical estimation. This contribution introduces Estimating Function (EF) theory to the forestry literature as such an approach with potential to deal successfully with many typical modeling problems. Special emphasis will be placed on modeling serially or spatially correlated data, which arise often in forestry. Modeling of both continuous and categorical responses will be treated. Section 2 highlights the important concepts behind Estimating Function Theory. Necessary theoretical developments are deferred to the Appendix. In Section 3 an optimal function is introduced that can serve in many situations as an efficient means to statistical modeling. Section 4 discusses several examples using forestry data and depicts the results from a simulation study.

2 Estimating Function Theory

Godambe (1960) published a fundamental article about what he termed "the optimality of regular maximum likelihood". It is here that the idea of estimating-function-based inferences originated. To be specific let $\mathbf{X}$ denote a sample from a distribution $p(\mathbf{x}, \theta)$, that depends on a scalar parameter θ. The vector parameter case is discussed in Section 3. We are interested to find an estimator $\hat{\theta}$ of θ that is a function of the data. All classical estimation principles solve an equation of the form

$$g(\mathbf{x}, \theta) = 0 \qquad [1]$$

for θ. If the root of [1] exists, it serves as the estimator $\hat{\theta} = T(\mathbf{x})$ of θ. For example, the ordinary least squares principle invokes

$$g(\mathbf{x},\theta) = \sum_i \left(x_i - E(x_i)\right)\partial E(x_i) / \partial\theta = 0.$$

While classical analysis focuses on evaluating the properties of $\hat{\theta} = T(\mathbf{x})$, Estimating Function theory focuses on properties of the function $g(\mathbf{x}, \theta)$. Among all possible $g(\mathbf{x}, \theta)$, which are functions of data and the parameters, one selects that which is optimal in some sense.

This appears to be a natural extension of classical estimation where one searches for the estimator that is in some sense optimal. Not surprisingly, just as there is typically no single best estimator, neither does there appear to be a uniformly best estimating function. However, since all classical estimating equations are of the form [1], it seems reasonable to restrict the search to the class $\mathfrak{S}$ of unbiased EF's, which are defined as having zero expectation regardless of the value of θ, i.e. $E(g(\mathbf{x}, \theta)) = 0\ \forall\ \theta$. It should be noted that unbiasedness of $g(\mathbf{x}, \theta)$ does not necessarily mean unbiasedness of $\hat{\theta}$, the implied estimator of θ. We restrict this discussion to EF's that are linear in the observables $\mathbf{X}$, but can be non-linear in θ.

Godambe (1960) defined $g(\mathbf{x}, \theta)$, as an optimal estimating function (OEF), provided that

$$\frac{E\left(g(\mathbf{x},\theta)^2\right)}{\left\{E\left(\frac{\partial g(\mathbf{x},\theta)}{\partial\theta}\right)\right\}^2} \le \frac{E\left(g^*(\mathbf{x},\theta)^2\right)}{\left\{E\left(\frac{\partial g^*(\mathbf{x},\theta)}{\partial\theta}\right)\right\}^2} \qquad (2)$$

for all θ, where $g^*(\mathbf{x}, \theta)$ is any other unbiased EF in $\mathfrak{S}$. Godambe and Kale (1991) term $g_s(\mathbf{x}, \theta) = g(\mathbf{x}, \theta) / E(\partial g(\mathbf{x}, \theta) / \partial\theta)$ a standardized estimating function (cf. Godambe and Heyde 1987, Godambe and Thompson 1989). [2] is thus equivalent to stating that an OEF is the standardized EF with smallest variance in $\mathfrak{S}$. This latter interpretation has appeal for several reasons.

(i) If $g_s(\mathbf{x}, \theta)$ is linear in θ, $\text{var}(g_s(\mathbf{x}, \theta)) = \text{var}(\hat{\theta})$, where $\hat{\theta}$ is the estimator obtained by solving $g_s(\mathbf{x}, \theta) = E(g_s(\mathbf{x}, \theta)) = 0$. Choosing an OEF thus yields a minimum variance unbiased estimator (MVUE).

(ii) If $g_s(\mathbf{x}, \theta)$ is non-linear in θ, an iterative algorithm may be necessary to estimate θ. Solving $g(\mathbf{x}, \theta) = 0$ via Fisher scoring applies a directional increment

$$\delta\hat{\theta} = -E\left\{\frac{\partial g(\mathbf{x},\theta)}{\partial\theta}\right\}^{-1} g(\mathbf{x},\theta)$$

to current solutions. Hence the estimate at iteration number $u + 1$ is

$$\hat{\theta}_{u+1} = \hat{\theta}_u + \delta\hat{\theta} = \hat{\theta}_u + g_s(\mathbf{x}, \hat{\theta}_u).$$

The directional increment is a standardized EF evaluated at the current solution. Choosing the $g_s(\mathbf{x}, \theta)$ with minimal dispersion results in fast convergence, since the correction term will be small on average.

(iii) Godambe (1960) showed that the likelihood score function is an OEF in the sense of smallest standardized variance in the class $\mathfrak{S}$. This corresponds to highest information, since the variance of the score function is Fisher's information number. Godambe and Heyde (1987) show that selection of the minimally dispersed $g_s(\mathbf{x}, \theta)$ is equivalent to maximizing the closeness to the likelihood score function, which may be unknown. This allows one to find an efficient means for estimation in cases where maximum likelihood estimation is infeasible or impossible with minimal loss of information about the parameter θ.

Succinctly stated, when estimation is based on an OEF, the properties of the estimator are of minor importance, because the OEF utilizes the maximum of information in the data about the parameter. It is remarkable, that the OEF's in many situations correspond to well-known classical estimation equations, mostly least squares normal equations and likelihood score equations. The advantage of EF theory though, is not to keep the classical principles distinct, but to search for the OEF in any situation, regardless of whether the results correspond to classical estimators.

The greater importance and contribution of EF theory to the recent developments in statistical estimation is that it provides consistent estimates in situations where classical methods do not. For example let $y_i = f_i(\alpha) + \varepsilon_i$, where $E(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma^2(\alpha)$. In this location model, variances and means are functionally related through the parameter α. It is well-known that least squares estimates obtained under these circumstances are biased and inconsistent. An optimal estimating function yielding consistent estimates, exists however (Godambe and Kale 1991).

In other situations, one may want to employ the maximum likelihood principle, but the joint distribution of the responses may be inscrutable or impossible to evaluate. An OEF that does not require a specification of the entire distribution, hence the likelihood, may provide a viable alternative means of estimation.

3 A General Estimating Function

For many who are well trained in classical methods of modeling and analysis, two related aspects of EF theory are difficult to appreciate at first glance. One, the need to identify an OEF in any given situation and two, that we are mainly concerned with the first two moments of $g(\mathbf{x}, \theta)$, rather than the moments of the estimator itself. In this section an EF is presented which is optimal in many classical and special situations, reduces to classical estimators in many settings and has interesting properties that transpire into properties of the obtained estimator. It is believed that the acceptability of EF theory depends on the existence of such a function.

To fix ideas, let $\mathbf{Y}_i = \left[Y_{i1}, \ldots, Y_{in_i}\right]$, $i = 1,\ldots,K$ be the vector of observed responses for the ith subject. No restriction is placed on n_i, the number of measurements made for subject i. In the case $n_i = 1, \forall\ i$, we have a classical (unrepeated) data structure, if $n_i = n > 1, \forall\ i$ this is a balanced repeated measurement structure. It is assumed throughout this contribution that the $\mathbf{Y}_i$'s are uncorrelated, but repeated measures on the same subject (the Y_{ij}'s) are dependent. The spacing of the Y_{ij} will often be associated with some meter

of time or space. But this is not required. In modeling the responses we postulate a location model, i.e. a statistical description for the expectation of $\mathbf{Y}_i$, with an additive error:

$$\mathbf{Y}_i = \eta(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) + \boldsymbol{\varepsilon}_i. \quad [3]$$

In [3], $\boldsymbol{\beta}$ is a $(p*1)$ vector of fixed effects, $\mathbf{b}_i$ is a $(q*1)$ vector of random variables with $E(\mathbf{b}_i) = \mathbf{0}$, $\text{var}(\mathbf{b}_i) = \mathbf{B}$. $\boldsymbol{\varepsilon}_i$ are random errors with $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$, $\text{var}(\boldsymbol{\varepsilon}_i) = \mathbf{R}_i$. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices associated with $\boldsymbol{\beta}$ and $\mathbf{b}_i$, respectively. If $\eta(\bullet)$ is the identity function and $\mathbf{Z}_i = \mathbf{0}$, [3] is a linear regression model $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ for the ith subject. If $\mathbf{Z}_i \neq \mathbf{0}$, $\mathbf{b}_i \neq \mathbf{0}$, a linear mixed model of the Laird-Ware form (Laird and Ware 1982) emerges as $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$. Gregoire, Schabenberger and Barrett (1995) analyzed models of this form. If $\eta(\bullet)$ is an invertible, non-linear, monotonic transform, [3] is a Generalized Linear Model (GLM) if the distribution of $\boldsymbol{\varepsilon}_i$ is in the exponential family of distributions (Nelder and Wedderburn 1972, Bickel and Doksum 1977, McCullagh and Nelder 1989). The inverses, $\eta^{-1}(\bullet)$, are known as link functions. This construction allows for categorical response variables, i.e. binary, nominal, ordinal responses and count data.

In the cases where $\eta(\bullet)$ is either the identity or non-linear monotonic, the matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ are of size (n_i*p) and (n_i*q), respectively, and $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ is called the linear predictor ξ_i. For continuous responses $\mathbf{Y}_i$, one oftentimes models more general non-linear functions than link functions. A typical example are growth models. In this case we allow $\eta(\bullet)$ to be an arbitrary non-linear function and conformity between $\mathbf{X}_i$ and $\boldsymbol{\beta}$ on one hand, and $\mathbf{Z}_i$ and $\mathbf{b}_i$ on the other hand is not required. To indicate general $\eta(\bullet)$ we write

$$\mathbf{Y}_i = \eta(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{b}_i) + \boldsymbol{\varepsilon}_i. \quad [4]$$

Models [3] and [4] have a very broad scope. Linear and non-linear regression, mixed linear (Laird and Ware 1982) and mixed non-linear models (Lindstrom and Bates 1990, Wolfinger 1993), as well as Generalized Linear (McCullagh and Nelder 1989) and Generalized Mixed Linear Models (Breslow and Clayton 1993) are contained in it, all of these either with or without repeated measurements.

3.1 Known Variance-Covariance Structure

To find an optimal estimating function for [3], [4] we initially focus on the fixed models only, i.e. $\mathbf{b}_i = \mathbf{0}$, where the elements in $\mathbf{Y}_i$ have known dispersion $\mathbf{V}_i$. Stack all matrices and vectors to dispose of the subscript i for the time being. That is we let

$$\mathbf{Y} = \left[\mathbf{Y}_1', \ldots, \mathbf{Y}_K'\right]', \text{var}(\mathbf{Y}) = \mathbf{V}, N = \textstyle\sum_i n_i.$$

Under these and some regularity conditions (Appendix A1), the optimal EF in the class $\mathfrak{S}$ of unbiased EF's linear in $\mathbf{Y}_i$ is

$$U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - E(\mathbf{Y})), \quad [5]$$

where $\mathbf{D} = \partial E(\mathbf{Y}) / \partial\boldsymbol{\beta}$ is of dimension $(N*p)$. A proof is outlined in Appendix A2. But what does this result imply about the estimator of $\boldsymbol{\beta}$ obtained from solving $U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{0}$? To see this we first need a device to solve $U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{0}$. Since $\eta(\bullet)$ may be non-linear in the parameters, an iterative procedure such as the Newton-Raphson algorithm with Fisher scoring is needed. Starting from an initial guess $\hat{\boldsymbol{\beta}}_0$, compute directional increments

$$-E\left(\frac{\partial U(\boldsymbol{\beta},\mathbf{y})}{\partial\boldsymbol{\beta}}\right)U(\boldsymbol{\beta},\mathbf{y}) = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)\left(\mathbf{D}'\mathbf{V}^{-1}\left((\mathbf{y} - E(\mathbf{Y}))\right)\right)$$

and then compute

$$\hat{\boldsymbol{\beta}}_{u+1} = \hat{\boldsymbol{\beta}}_u + \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\left(\mathbf{D}'\mathbf{V}^{-1}\left(\mathbf{y} - \hat{E}(\mathbf{Y})\right)\right). \quad [6]$$

For computational purposes [6] can be re-expressed as

$$\hat{\boldsymbol{\beta}}_{u+1} = \hat{\boldsymbol{\beta}}_u + \left(\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \hat{\mu}_i\right).$$

At convergence the asymptotic moments of $\hat{\boldsymbol{\beta}}$ are $\hat{\boldsymbol{\beta}} \sim (\boldsymbol{\beta}, (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}))$, i.e. $\hat{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator, with known variance-covariance matrix. Whenever [6] has a closed

form, non-iterative solution, this result holds for any sample size. Zeger and Liang (1986) furthermore showed that the distribution of $\hat{\boldsymbol{\beta}}$ is asymptotically Gaussian.

A few examples will show that [5] and [6] together lead to classical estimates in many situations:

1) Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_N)$. Then [5] becomes $U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, and [6] reduces to

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_0 + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0\right)\\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y},\end{aligned}$$

the familiar OLS estimates.

2) Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{V})$. Then $U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}(\boldsymbol{\beta})$ and

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_0 + \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0\right)\\ &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.\end{aligned}$$

As in 1) the solution exists in closed form, does not depend on the starting value $\hat{\boldsymbol{\beta}}_0$ and is the classical generalized least squares estimator.

3) Let $\mathbf{y} = \eta(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (0, \sigma^2\mathbf{I}_N)$ where $\eta(\boldsymbol{\beta})$ is an arbitrary function, non-linear in $\boldsymbol{\beta}$. [5] becomes $U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{D}'(\mathbf{y} - \eta(\boldsymbol{\beta}))$ and the estimates are calculated as

$$\hat{\boldsymbol{\beta}}_{u+1} = \hat{\boldsymbol{\beta}}_u + \left(\mathbf{D}'\mathbf{D}\right)^{-1}\left(\mathbf{D}'\left(\mathbf{y} - \eta\left(\hat{\boldsymbol{\beta}}_u\right)\right)\right).$$

This is non-linear least squares using a Gauss-Newton algorithm.

4) Let $Y_{ij} = \eta(\mathbf{x}_{ij}'\boldsymbol{\beta}) + \varepsilon_{ij}$, $\varepsilon_{ij} \sim (0, h(E(Y_{ij})))$, where $\eta()$ is an inverse link function and $h()$ is a variance function, depending on the means $E(Y_{ij})$. This is a generalized linear model. If the distribution of ε_{ij} belongs to the exponential family of distributions with canonical link $\eta^{-1}()$ (McCullagh and Nelder 1989), the likelihood for the ijth observation can be written as

$$l(\boldsymbol{\beta}, y_{ij}) = \frac{\partial E(Y_{ij})}{\partial\boldsymbol{\beta}}\frac{1}{\text{var}(Y_{ij})}\left(y_{ij} - E(Y_{ij})\right).$$

This is just [5] for a single observation and under the assumption that the Y_{ij}'s are uncorrelated, the likelihood for $\mathbf{Y}_i$ is exactly [5] and the OEF produces the maximum likelihood estimates. The Fisher scoring algorithm is then equivalent to iteratively reweighted least squares (cf. McCullagh and Nelder 1989).

We have seen that the estimating function [5] appears reasonable in many cases, for one reason, since it coincides with many classical estimates. The quality of the classical estimates is well established. Under the conditions of 1), $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate. If we add normality to 1), we furthermore obtain the property of being a uniformly minimum variance unbiased estimate (UMVUE). The properties of $\hat{\boldsymbol{\beta}}$ under conditions 3) depend among other things on the size of the sample, since we are dealing with an iterative solution and asymptotic results.

From the standpoint of Estimating Function theory, [5] has been found to be the optimal function in its (restricted) class $\mathfrak{S}$, and it can be shown (see Appendix A3) that estimators derived from [5] are asymptotically minimally dispersed. This result can be viewed as an extension of the Gauss-Markov Theorem to Estimating Function theory. For extensions of other important theorems in statistical inference, such as the Cramér-Rao lower bound or the Rao-Blackwell theorem see Bhapkar (1991).

Another important property of the estimating function [5] is its invariance with respect to linear transformations, a property shared with maximum likelihood estimates, but not, for example with UMVU estimates. To see this let $\mathbf{Z} = \mathbf{GY}$ where $\mathbf{G}$ is a conformable non-singular matrix of real values (constants). Then $\text{var}(\mathbf{Z}) = \mathbf{GVG}'$, $E(Z) = \mathbf{G}E(\mathbf{Y})$, $\partial\mathbf{E}(\mathbf{Z}) / \partial\boldsymbol{\beta} = \mathbf{GD}$. The optimal EF for $\boldsymbol{\beta}$ is still $U(\boldsymbol{\beta},\mathbf{z}) = \mathbf{D}'\mathbf{G}'\mathbf{G}'^{-1}\mathbf{V}^{-1}\mathbf{G}^{-1}\mathbf{G}(\mathbf{y} - E(\mathbf{Y})) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - E(\mathbf{Y})) = U(\boldsymbol{\beta},\mathbf{y})$.

3.2 Variance-Covariance Structure Subject to Estimation

In this section we examine [5] in situations when the weight matrix $\mathbf{V}$ is not known completely a priori and subject to estimation. Only the first

two moments of the distribution of $\mathbf{Y}$ must be known to invoke [5], the first moment, $E(\mathbf{Y})$ is the objective of modeling and var($\mathbf{Y}$) is a feature of both the data and the model. However, a complete specification of the distribution is not necessary and only first derivatives of $E(\mathbf{Y})$ are required. In fact, all that is necessary to yield the asymptotic properties at this point is that $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$, $\forall\ i$, that is, the model is correctly specified.

If one were to use [5] in models with correlated observations, $\mathbf{V}$ must be known. In the case of unequal variances or correlated observations, $\mathbf{V}$ is typically at least partially unknown. For example, with repeated binary responses where $\mathbf{Y}_i$ is a $(n_i * 1)$ vector of zeros and ones, the variances of the Y_{ij} are completely determined by the binary nature of the response: $\text{var}(Y_{ij}) = E(Y_{ij})(1 - E(Y_{ij}))$. However, the off-diagonal elements in $\mathbf{V}$ are non-zero and unknown a priori. The standard approach is to make $\mathbf{V}$ a function of parameters in θ, and to replace the weight matrix during iterations by $\hat{\mathbf{V}} = \mathbf{V}(\hat{\theta})$. This causes the coefficient estimates to depend on $\hat{\theta}$ also and one proceeds with a two-step procedure: after estimating $\hat{\boldsymbol{\beta}}(\hat{\theta})$ new estimates are obtained for θ. These are used to obtain next estimates for $\boldsymbol{\beta}$. The scheme is continued until convergence.

Two different approaches exist to introduce the parameters θ into the model. One way is by using random terms $\mathbf{b}_i$ in the predictor. In this case θ comprises the unique elements in var($\mathbf{b}_i$) = $\mathbf{B}(\theta)$. For example consider a linear mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \mathbf{R}_i = \sigma^2\mathbf{I}).$$

The marginal variance var($\mathbf{Y}$)$_i$ becomes var($\mathbf{Y}$)$_i$ = $\mathbf{V}_i = \mathbf{Z}_i\mathbf{B}(\theta)\mathbf{Z}_i' + \sigma^2\mathbf{I}$. Having estimates $\hat{\theta}$ and $\hat{\sigma}^2$ available, one can estimate $\hat{\mathbf{V}}_i$. Similarly, the random terms enter non-linearly into the model.

Another way to accommodate correlated observations is by stipulating a possible correlation pattern and estimating the associated correlation parameters. In this approach a so-called *working covariance matrix* replaces $\mathbf{V}_i$. This is profoundly different from using random effects. By simply estimating a *working* correlation pattern one is entirely free to stipulate the correlation structure. Estimation of this pattern focuses on $\mathbf{R}_i$ only. In the above linear mixed model however, as soon as $\mathbf{Z}_i$ is determined, the marginal variance is completely specified and governed by the shared subject effects $\mathbf{b}_i$. In accordance with Zeger et al. (1988) we term approaches based on random effects, subject-specific (SS) and those based on working assumptions, population-averaged (PA). The distinction between these approaches is important with respect to the interpretation of the coefficient estimates. In a PA model, the coefficients describe changes in the population averaged response with changes in the covariates. In a SS model they describe changes in a subject's response. It should also be noted, that for linear models this distinction is of minor importance. In the above example, if $\mathbf{Z}_i = \mathbf{X}_i$, $\boldsymbol{\beta}$ would be considered the population averaged coefficients, $\boldsymbol{\beta} + \mathbf{b}_i$ the subject specific coefficients. Linear models thus yield both interpretations naturally. If one omits random terms and models a *working* structure directly, however, only PA coefficients can be obtained.

Both approaches of modeling correlations can be combined. Gregoire, Schabenberger, and Barrett (1995) use direct modeling of $\mathbf{R}_i$ based on continuous autoregressive processes with random terms in a linear mixed model.

Zeger and Liang (1986) and Liang and Zeger (1986) proved that if $\mathbf{V}_i$ is replaced by a consistent estimator $\hat{\mathbf{V}}_i$, say, the asymptotic normality, unbiasedness and efficiency of the EF estimator of $\boldsymbol{\beta}$ is retained. The covariance estimator of $\hat{\boldsymbol{\beta}}$ is obtained as $\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{D}'\hat{\mathbf{V}}^{-1}\mathbf{D})^{-1}$ substituting the estimated covariance parameters for θ.

Estimating function [5] does not require specification of more than the variances and the means of the responses. In contrast to likelihood-based estimation, distributional assumptions are not needed. Oftentimes, fully parametric estimation of θ is computationally more cumbersome than necessary (Gregoire and Schabenberger 1994, Schabenberger 1995b). It is important to estimate θ consistently to obtain consistent estimates of $\mathbf{V}_i$. A simple, consistent estimator may prove sufficient. The basic scheme of [5] and [6] remains the same, with $\hat{\mathbf{V}}_i$ replacing $\mathbf{V}_i$, i.e.

$$U(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^{K} \mathbf{D}_i\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - E(\mathbf{Y}_i)\right),$$

$$\hat{\boldsymbol{\beta}}_{u+1} = \hat{\boldsymbol{\beta}} + \left(\sum_{i=1}^{K} \mathbf{D}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{D}_i\right)^{-1} \sum_{i=1}^{K} \mathbf{D}_i'\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - \hat{\mu}_i\right).$$

In a mixed model, we use the model structure itself to estimate θ. When the mixed model is linear, we directly obtain var($\mathbf{Y}_i$) = $\mathbf{V}_i$ = $\mathbf{Z}_i\mathbf{B}\mathbf{Z}_i' + \mathbf{R}_i$ and $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$. If $\eta(\bullet)$ is anything other than the identity function we have to find the marginal expectation and variance-covariance matrix from the conditional moments

$$E(\mathbf{Y}_i \mid \mathbf{b}_i) = \eta(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{b}_i)$$
$$\mathbf{V}_i = \text{var}(E(\mathbf{Y}_i \mid \mathbf{b}_i)) + E(\text{var}(\mathbf{Y}_i \mid \mathbf{b}_i)).$$

How this is accomplished differs for Generalized Linear Models and arbitrary non-linear functions. For GLM's the marginal mean can be calculated or approximated from the conditional ones by use of attenuation correction factors. They are described for the most frequently entertained link functions (log, logit, probit) in Zeger et al. (1988). For the ijth observation they typically apply a multiplicative offset to $\mathbf{x}_{ij}'\boldsymbol{\beta}$. The marginal variance is obtained by using Taylor series expansions around $\mathbf{b}_i = \mathbf{0}$ in the two conditional pieces above. The resulting variance can be written as

$$\text{var}(\mathbf{Y}_i) \doteq \mathbf{L}_i\mathbf{Z}_i\mathbf{B}\mathbf{Z}_i'\mathbf{L}_i + \mathbf{R}_i$$

where $\mathbf{L}_i$ is a diagonal matrix with elements $\partial\eta(\mathbf{x}_{ij}'\boldsymbol{\beta}) / \partial\mathbf{x}_{ij}'\boldsymbol{\beta}$. We can also write

$$\text{var}(\mathbf{Y}_i) \doteq \tilde{\mathbf{Z}}_i\mathbf{B}\tilde{\mathbf{Z}}_i' + \mathbf{R}_i \qquad [7]$$

where $\tilde{\mathbf{Z}}_i$ is $\partial\eta(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) / \partial\mathbf{b}_{i|\mathbf{b}_i=\mathbf{0}}$ is called a local design matrix with respect to the random terms.

Schabenberger (1995b) has adapted Zeger et al.'s approach for Generalized Mixed Linear Models to accommodate arbitrary $\eta(\bullet)$. The marginal variance is derived along the same lines, but the marginal mean is also developed through Taylor series around $\mathbf{b}_i = \mathbf{0}$. Consequently, $\text{E}(\mathbf{Y}_i) \doteq \eta(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{0})$. $\mathbf{D}$ is replaced by another localized design matrix, based on this Taylor series, i.e. $\mathbf{D}_i = \partial\eta(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{0}) / \partial\boldsymbol{\beta} = \tilde{\mathbf{X}}_i$. $\tilde{\mathbf{X}}_i$ is the same derivative matrix one would use in fitting a non-linear model by least squares without any random terms. After iterating for $\boldsymbol{\beta}$ once an estimate of $\mathbf{V}_i$ is needed. For simplicity, assume that there are no additional parameters in $\mathbf{R}_i$ apart from a possible scalar σ^2. For continuous responses $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$, for categorical responses, $\mathbf{R}_i$ is a diagonal matrix with the variance functions $h(E(Y_{ij}))$ on its diagonal. Then all covariance parameters of interest are the unique elements of $\mathbf{B}$. From [7] one obtains after solving for $\mathbf{B}$:

$$\mathbf{B} = \left(\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_i\right)^{-1}\tilde{\mathbf{Z}}_i'(\mathbf{V}_i - \mathbf{R}_i)\tilde{\mathbf{Z}}_i\left(\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_i\right)^{-1}$$

suggesting the simple and consistent moment estimator

$$\hat{\mathbf{B}} = \frac{1}{K}\sum_{i=1}^{K}\left(\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_i\right)^{-1}\tilde{\mathbf{Z}}\left(\left(\mathbf{y}_i - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}\right)\left(\mathbf{y}_i - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}\right)' - \mathbf{R}_i\right) \qquad [8]$$
$$\tilde{\mathbf{Z}}_i\left(\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_i\right)^{-1}.$$

At convergence of the algorithm, the Best Linear Unbiased Predictors (BLUP) for the random effects can be computed as (c.f. Laird and Ware 1982, Breslow and Clayton 1993)

$$\hat{\mathbf{b}}_i = \hat{\mathbf{B}}\tilde{\mathbf{Z}}_i'\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}\right).$$

Population averaged approaches require the modeler to stipulate an appropriate structure for the variance-covariance matrix $\mathbf{R}_i$ and to provide consistent estimates for its parameters without resorting to distributional assumptions. Since one typically specifies $\mathbf{R}_i$ in terms of correlation parameters we put $\mathbf{V}_i(\theta) = \mathbf{A}_i^{1/2}\mathbf{C}_i(\theta)\mathbf{A}_i^{1/2}$ where $\mathbf{C}_i(\theta)$ is the correlation matrix for the ith subject and $\mathbf{A}_i^{1/2}$ is a diagonal matrix with the standard deviations of the Y_{ij}'s on the diagonal. Typical structures for $\mathbf{C}_i(\theta)$ are

1) compound symmetry: $\mathbf{C}_i(\theta) = \begin{bmatrix} 1 & \theta & \ldots & \theta \\ \theta & 1 & \ldots & \theta \\ \theta & \theta & 1 & \ldots \\ \theta & \ldots & \theta & 1 \end{bmatrix}$

where it is assumed that all pairs of observations are equicorrelated,

2) 1-dependence: $\mathbf{C}_i(\theta) = \begin{bmatrix} 1 & \theta & 0 & 0 \\ \theta & 1 & \theta & 0 \\ 0 & \theta & 1 & \theta \\ 0 & \ldots & \theta & 1 \end{bmatrix}$

3) k-dependence: $\mathbf{C}_i(\theta) = \begin{bmatrix} 1 & \theta_1 & \ldots & \theta_k & 0 \\ \theta_1 & 1 & \theta_1 & \ldots & \theta_k \\ \theta_2 & \theta_1 & 1 & \theta_1 & \\ \vdots & & & \ddots & \theta_1 \\ 0 & \theta_k & \ldots & \theta_1 & 1 \end{bmatrix}$

Independence is a special case of these structures where $\theta = 0$.

Compound symmetry is a structure that does not imply any ordering among the measurements within a subject. For example, in ophthalmology, where measurements on the two eyes of an individual are taken, there is no prevalence of one eye over the other. In forestry, however, there usually is an implied ordering of the repeated measurements in time or in space. For such data one of the k-dependence structures may seem more appropriate. They imply that measurements are correlated as long as they are less than k units of remeasurements apart. If plots are revisited in 3 year intervals and the dominant heights recorded, a 3-dependence correlation structure implies that measurements are correlated unless they are more than 9 years apart. The compound and dependence structure can be combined. Such a correlation matrix would estimate a common correlation coefficient θ and set all correlations > k to zero. Obviously there is much latitude and freedom in experimenting with different correlation structures. If prior information about the likely structure is available, it should of course be utilized, since closeness of $\mathbf{C}_i(\hat{\theta})$to the actual $\mathbf{C}_i$ increases the efficiency of $\hat{\boldsymbol{\beta}}$.

To estimate θ in a PA model we utilize residuals and their cross-products as building blocks. In general we write the standardized residual for the ijth observation as

$$r_{ij} = \frac{\left(y_{ij} - \hat{\eta}_{ij}\right)}{\hat{\text{var}}\left(Y_{ij}\right)^{1/2}}$$

where $\hat{\text{var}}(Y_{ij})$ is typically $\hat{\sigma}^2$ for a continuous response and $h(\hat{E}(y_{ij}))$ for a categorical response. The correlation parameters are then functions of $r_{ijj'} = r_{ij}r_{ij'}$ which will be averaged over specific subsets of the sample to obtain consistent estimates of all elements in θ. For the k-dependence structure we can estimate

$$\theta_j = \sum_{i=1}^{K} r_{ij}r_{i(j+1)} / (K - p)$$

and combine these estimates to obtain the 1-dependence estimates

$$\hat{\theta} = \sum_{j=1}^{k-1} \theta_j / (k-1).$$

Under a compound symmetry structure we can average over all individuals and residual combinations as

$$\hat{\theta} = \sum_{i=1}^{K} \sum_{j>j'} \frac{r_{ij}r_{ij'}}{-p + \sum_{i=1}^{K} \frac{1}{2} n_i(n_i - 1)}.$$

The estimates $\hat{\theta}$ are not difficult to compute and the iterative algorithm is simpler than for the mixed models, since no Taylor series expansions of any kind or attenuation corrections are required.

4 Examples

Several examples and a small scale simulation study are presented in this section to demonstrate the broad scope of the Estimating Function concept and especially of function [5]. The examples are concerned with forestry data in which repeated measurements of some sort are correlated. The unifying appearance of estimating function [5] allows one to program the estimation algorithm with any high-level matrix programming language, since at the present time no commercially available statistics package supports [5]. The analysis of the following examples has been facilitated by a program written in GAUSS[1] which is available from the senior author upon request.

4.1 Non-Linear Mixed Models, Continuous Response

Equations to predict the accumulated bole volume to any upper diameter of a tree are important devices to assess merchantable volume. In the past, it has been common practice to fit separate volume functions depending on the upper diameter of interest. This is a wasteful approach, since what is considered merchantable in a specific sector is non-merchantable timber in another. Furthermore, such merchantability limits are constantly changing over time. Hence attempts have been made to include upper diameters in volume equations to enable predictions to any arbitrary upper diameter. The resulting models are typically of the form

$$V_{d_u} = V_0(\boldsymbol{\beta}_1)R(\boldsymbol{\beta}_2)$$

where V_0 is a total volume function depending on covariates like diameter at breast height D, total tree height H, etc. and parameters in $\boldsymbol{\beta}_1$. $R(\boldsymbol{\beta}_2)$ depends on the upper diameter d_u and adjusts $V_0(\boldsymbol{\beta}_1)$ downward depending on values of d_u. Amateis and Burkhart (1987) used $V_o(\boldsymbol{\beta}_1) = \beta_0 + \beta_1 D^2 H$ together with

$$R(\beta_2) = 1 + \beta_2 d_u^{\beta_3} / D^{\beta_4}. \qquad [9]$$

Even if both $V_0(\boldsymbol{\beta}_1)$ and $R(\boldsymbol{\beta}_2)$ are linear, V_{d_u} is a non-linear function of the parameters. Non-linear estimating techniques are required to fit utilization models of this type. The correction terms $R(\boldsymbol{\beta}_2)$ have to obey some simple and a few less obvious constraints to be useful. Obviously $R(\boldsymbol{\beta}_2)$ has to be bounded by 1 to guarantee $V_o = V_0(\boldsymbol{\beta}_1)$ at the tree tip. Plotting empirical cumulative volume functions on a relative scale (see Fig. 1) one can generally depict a sigmoid shape reminiscent of empirical cumulative distribution functions. The correction term $R(\boldsymbol{\beta}_2)$ thus has to have an interior inflection point. A less obvious constraint is that transcendent forms in the derivatives $\partial V_d / \partial \beta$ should be avoided. The reason herefore is simply that in the course of a non-linear fit derivatives have to be evaluated at the current solutions, which can be negative.

A correction term that meets these criteria, depends on only two parameters, and is easy to fit, since parameters are only involved as powers of e, is

$$R(\boldsymbol{\beta}_2) = \exp\left\{-\beta_2 t * \exp\{\beta_3 t\}\right\} \quad \text{where } t = \frac{d_u}{D}.$$

This correction term together with a total volume function $V_o(\boldsymbol{\beta}_1) = \beta_0 + \beta_1 D^2 H$ is used here to model upper diameter volume of sweet gum (*Liquidambar styraciflua* L.). 39 sweet gum trees were felled and measured in the East Texas region. Tree outside-bark diameters and section-wise volume were obtained at 3 foot intervals along the boles. The total number of cumulative bole volumes was 1058. On average there were 27 measurements per tree. While it can be assumed that trees are independent, the measurements on any single tree are of course correlated. The goal of the study was not only to develop and test a new correction term $R(\boldsymbol{\beta}_2)$ and to account for serial correlation, but also to model each tree's volume profile most accurately. This calls for random terms so that the BLUPs can be used to individualize the predictions. The actual model fitted was

$$V_{d_u ij} = \left(\left\{\beta_0 + b_i\right\} + \beta_1 \frac{D_i^2 H_i}{1000}\right) \exp\left\{-\frac{\beta_3 t_{ij}}{1000} * \exp\left\{\beta_4 t_{ij}\right\}\right\} \qquad [10]$$

with a random intercept. Note that this construction makes the variances of V_{d_u} a function of the upper bole diameters which are likely to be a sensible proxy for the correlations, i.e. as two diameters become distant, it is likely that the correlations will decrease, compared to directly adjacent measurements. Although a random intercept is used, the multiplicative nature of the model makes the marginal variance of the responses a function of the meta meter of correlations.

Nonlinear mixed models such as [10] can be

[1] Gauss is a trademark of Aptech Systems, Inc., 23804 S. E. Kent-Kangley Road, Maple Valley, Washington, U.S.A.

Table 1. Results from fitting non-linear mixed model to sweet gum data. Asymptotic standard errors in parentheses.

Statistic	Non-linear least squares	Mixed model
	OEF under independence	OEF with random intercept
$\hat{\beta}_0$	3.31113	4.88284
	(0.7129)	(2.5552)
$\hat{\beta}_1$	2.11338	2.1024
	(0.0210)	(0.0655)
$\hat{\beta}_2$	15.7994	5.51959
	(3.0605)	(0.7022)
$\hat{\beta}_3$	6.03415	6.35841
	(0.2661)	(0.1525)
$\hat{\sigma}^2$	126.98	41.0306
$\hat{\mathbf{B}}$		113.0545

fit by approximate likelihood techniques based on linearizations. This requires specification of the distributions for $\mathbf{b}_i$ and the model errors, which for convenience are typically presumed Gaussian. A classical algorithm to accomplish this task was presented by Lindstrom and Bates (1988). Schabenberger (1995b) demonstrated a more efficient implementation based on theoretical results by Wolfinger (1993). However, the computational demands of likelihood methods for non-linear models are still considerable and compare poorly to the speed of the estimating function based algorithm, which does not involve distributional assumptions.

Table 1 displays results from fitting model [10] to the data using the OEF [5] and the mixed model algorithm outlined in Section 3.2.2.

The estimates obtained from a non-linear least squares fit under the incorrect assumption of uncorrelated observations are profoundly different from the estimating function results. Of special interest is the reduction in residual variation $\hat{\sigma}^2$. A considerable amount of variation has been explained by inclusion of the random intercept. The residual sum of squares in the mixed model is only 43,246 compared to 133,837 in the least squares fit. At the cost of only one additional parameter, $\text{var}(b_i) = \mathbf{B}$, variation was distributed in across and within subject sources. Making the slopes or terms in $R(\boldsymbol{\beta}_2)$ random showed no improvement over model [10].

Motivations for inclusion of the random term was (i) to accurately attribute variation to separate sources, and (ii) the desire to individualize the fit. Fig. 1 depicts the observed and fitted volume profiles for four trees. The sigmoidal shape of the empirical profiles is apparent. The solid line denotes predictions using only the fixed effects of model [10]. Apparently, for trees 5, 7, and 8, there is a severe overprediction of cumulative volume. Tree 6 shows underpredictions based on only the fixed effects. The reason herefore is that the fixed effects alone target the average behavior in the population. Whether predictions using only $\hat{\boldsymbol{\beta}}$ provide a good description of a tree's profile depends on how close the individual profile mimics the population average. The dashed lines in Fig. 1 include the BLUPs in the linear predictor. The improvement over a prediction based on $\hat{\boldsymbol{\beta}}$ only is remarkable.

4.2 Linear Mixed Models, Continuous Response

One motivation for entertaining semi-parametric estimating functions instead of complete likelihood analysis for non-linear mixed models is to reduce computational effort. While providing excellent answers, that have been found to be hardly distinguishable from the likelihood results (cf., Gregoire and Schabenberger 1994, 1995) the computations are much less involved. Fitting model [10] to the sweet gum data set with 1058 observations required a mere 12 seconds on a 486/66 PC which includes all aspects of inference and calculation of predictions. The full likelihood implementation can take up to several minutes. When mixed models are linear, this advantage vanishes. Starting with Release 6.07 of SAS/STAT[2], PROC MIXED has been available that provides normal theory inference for mixed linear models efficiently. Estimating functions, however, do not lose their comparative merits. Normal theory inference depends on distributional assumptions, which oftentimes are

[2] SAS is a registered trademark of SAS Institute Inc., Cary, North Carolina, U.S.A.

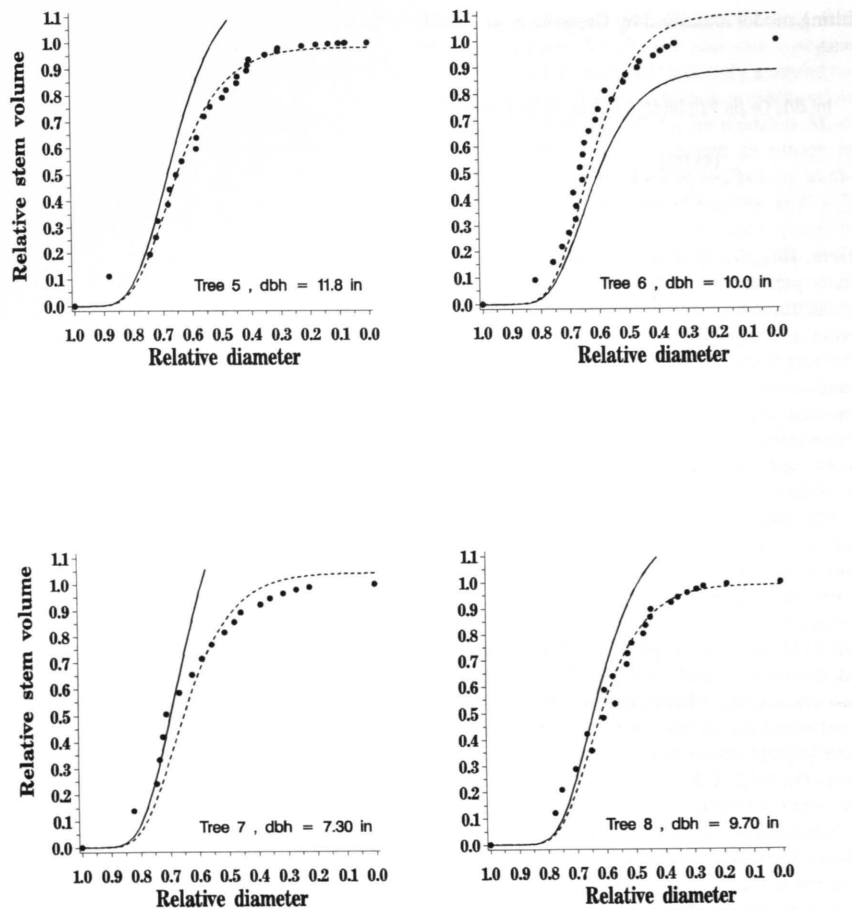

Fig. 1. Upper diameter volume model with random intercept for sweet gum data fitted via optimal estimating function. Observed values denoted with bullets, fixed effects predictions with solid line, and predictions based on BLUPs with dashed line.

wrong. Much of mixed linear model work today assumes $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{B})$, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$ simply because software is available that supports this assumption. Gregoire, Schabenberger, and Barrett (1995) present mixed model diagnostics that help to address the question of normality. The validity of distributional assumptions will, however, always be subject to question.

Gregoire et al. (1994) analyzed a data set of eastern white pine (*Pinus strobus* L.) to develop a model for log(basal area) that appeared to support the normality assumption reasonably well. On 59 plots located in New Hampshire, the basal area along with several plot level covariates were measured. Remeasurements varied between 3 and 23 years, the number of remeasurements per plot ranged from 2 to 6 between 1960 and 1991. The data set comprises 268 observations. The best

fitting model identified by Gregoire et al. (1994) was

$$\ln(BA_{ij}) = \beta_0 + \beta_1 \ln(H_{ij}) + \beta_2 \ln(N_{ij}) + \beta_3\left(\frac{\ln N_{ij}}{A_{ij}}\right) + \beta_4 \frac{(\ln H_{ij})}{A_{ij}} + b_i \frac{1}{A_{ij}}.$$

Here, BA_{ij}, N_{ij}, and H_{ij} denote basal area (m^2), trees per hectare and height (m) at the jth remeasurement of plot i. That is four fixed effects were accompanied by a random coefficient in the reciprocal of the ijth age (A_{ij}). Although the random term b_i is a characteristic of the ith plot, the marginal variance-covariance matrix for the measurements of the ith plot are a function of the ages, since the rows of the $\mathbf{Z}_i$ matrix vary within a subject.

The model for $\ln(BA_{ij})$ was fit in three different ways. Using the OEF without accounting for the serial correlation in any way leads to ordinary least squares estimates one would obtain from any statistical regression package. PROC MIXED was used to provide a fully parametric fit, here a restricted maximum likelihood analysis was chosen. Finally the model with random coefficient was fit employing the estimating function [5] and the moment estimator $\hat{\mathbf{B}}$ described in Section 3.2.2. Results of these three analyses are summarized in Table 2.

The OLS results are again much different from the mixed model analyses. The estimate of residual variation $\hat{\sigma}^2$ under OLS is order of magnitude larger. This sheds some light on how much variation has been explained by the inclusion of the random coefficient. The agreement between the REML and the OEF/Mixed $\hat{\beta}_k$ estimates is very satisfying. The minute differences are caused entirely by the differences in the $\hat{\mathbf{B}}$ estimates, since the normal theory scoring algorithm for $\boldsymbol{\beta}$ can actually be written in the form of [5]. If the covariance parameters $\mathbf{B}$ and σ^2 would have been estimated by maximum likelihood instead of restricted maximum likelihood the agreement between OEF and parametric analysis would be even greater, since ML estimates of covariance parameters are slightly downward biased (Searle et al. 1992). The ML estimates of $\mathbf{B}$ and σ^2 are 135.209 and 0.0019, respectively, showing perfect agreement with the OEF estimates for the mixed model.

Table 2. Results from fitting linear mixed model to white pine data. Standard errors in parentheses.

Statistic	OLS OEF under independence	Mixed model OEF with random coefficient	Mixed model Fully parametric REML
$\hat{\beta}_0$	–0.8086	–0.4175	–0.4197
	(0.3091)	(0.3426)	(0.3459)
$\hat{\beta}_1$, $\ln H_{ij}$	0.8815	0.5777	0.5778
	(0.0907)	(0.0624)	(0.0629)
$\hat{\beta}_2$, $\ln N_{ij}$	0.2837	0.4983	0.4986
	(0.0272)	(0.0314)	(0.0317)
$\hat{\beta}_3$, $(\ln N_{ij}) / A_{ij}$	–0.6903	4.3778	4.3824
	(2.8687)	(0.7801)	(0.7872)
$\hat{\beta}_4$, $(\ln H_{ij}) / A_{ij}$	2.59138	–24.4023	–24.422
	(2.8687)	(2.4903)	(2.5133)
$\hat{\sigma}^2$	0.01762	0.00190	0.00193
$\hat{\mathbf{B}}$		135.961	139.084

However, the results for the mixed model using the estimating function approach have been obtained without any distributional assumptions. Fig. 2 displays the predicted values for the 59 plots obtained from the REML and the OEF fit, utilizing the BLUPs. The solid lines in Fig. 2 connect the REML predictions, the circles denote predicted values from the OEF. For all practical purposes, the two are indistinguishable and it can be asked, what has been gained by making a rather strong assumption about the distribution of the $\mathbf{b}_i$'s and the $\boldsymbol{\varepsilon}_i$'s.

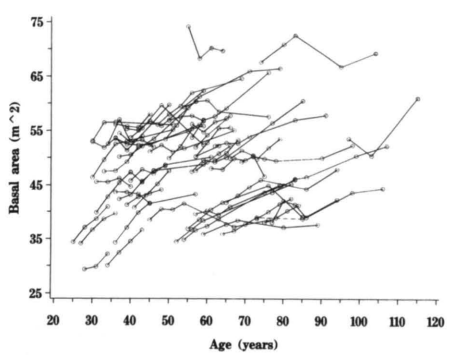

Fig. 2. White pine data. Mixed linear model with random coefficient. Predicted values from REML fit denoted by lines, values obtained from OEF denoted by circles.

4.3 Categorical Response

Modeling categorical responses has attracted much interest in the last years. One reason for this is it allows one to model and predict probabilities and/or counts, which are important quantities in many areas of scientific research. Another reason is that researchers have become more aware of the necessity to use appropriate tools for non-continuous responses and have resorted to generalized linear models, that are of specific importance in categorical data analysis. Commercially available software such as PROC LOGISTIC of SAS/STAT for example has contributed much to the recently developed appreciation of logistic and probit analyses. In this section, cumulative link models, a special type of categorical response models, are used to model multinomial responses. They are important because they allow ordering in more than just 2 response categories but reduce to classical models for binary response when only two possible outcomes are observed.

Assume a random experiment where we observe a categorical response C which can fall into one and only one of P mutually exclusive categories. If $P > 2$ we assume tacitly that the categories are ordered and not merely nominally scaled. To develop a model that allows one to predict the outcome of such a random experiment we recode C initially as

$$\begin{aligned} Y_p &= 1 \text{ if } C \text{ is in category } p \\ &= 0 \text{ otherwise.} \end{aligned}$$

By construction, if we find a model for the expectation of Y_p, we will model the probability that C falls in category p, since

$$E(Y_p) = \Pr(Y_p = 1) + 0 * \Pr(Y_p \neq 1) = \Pr(C = p).$$

Since the categories are exclusive, there is a linear constraint $\sum_{p=1}^{P} Y_p = 1$ and one category can be omitted. Hence, if $P = 2$, only a model for $E(Y)$ is needed, if $P > 2$, this is a multivariate setting, where $E(Y_1),\ldots,E(Y_{P-1})$ are modeled. Models in the first group are known as binary or dichotomous response models and belong to the family of Generalized Linear Models. If $P > 2$, we can use similar constructs, but are required to employ what is known as composite link functions. One particularly useful class of composite links is the family of cumulative links (McCullagh and Nelder 1980, Agresti 1990). Instead of modeling $\Pr(C = p)$ one models $\Pr(C \leq p)$. The link function itself is often chosen freely, and commonly employed functions include the log-link $\eta^{-1}(t) = \log(t)$, logit link $\eta^{-1}(t) = \log(t/1 - t)$, log-log link $\eta^{-1}(t) = -\log(-\log(t))$, and the probit link corresponding to the inverse cumulative standard normal distribution. For binary and multinomial responses, the logit link has enjoyed particularly frequent usage because it leads to interesting interpretation of the model coefficients in terms of odds and because it is the canonical link for binomial and multinomial distributions. Details can be found in McCullagh and Nelder (1989), Agresti (1990), and Schabenberger (1995a).

Schabenberger (1995a) introduced a particularly simple cumulative logit model into the forestry literature. It is known as McCullagh's proportional odds model (McCullagh 1980) and can be motivated in the following fashion: Assume one observes only the category in which the ith subject responds, but recognizes a putative, underlying continuous scale Z, say, which is divided into intervals by the categories of C. An example is the observation of tree breast height diameters in DBH classes[3]. Presume a linear model holds on the continuous scale

$$Z_i = \boldsymbol{\varepsilon}_i - \mathbf{x}_i'\boldsymbol{\beta}$$

where ε_i follows a distribution law with cdf $\eta(\bullet)$. This construction is possible since the link functions above are the inverse functions of cumula-

[3] It is important, however, that the existence of Z is only a device to motivate the development of the following model, it is not required for its existence or validity.

tive distribution functions. The cut-off points on the scale of Z that divide its range into categories are denoted α_j, where $\alpha_0 = -\infty$, $\alpha_P = \infty$. There are only $P-1$ unknown cut-off parameters. Then

$$\Pr(C_i \le j|\mathbf{x}_i) = \Pr(Z_i \le \alpha_j|\mathbf{x}_i) = \Pr(\varepsilon_i - \mathbf{x}_i'\boldsymbol{\beta} \le \alpha_j|\mathbf{x}_i) = \Pr(\varepsilon_i \le \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}) = \eta(\alpha_j + \mathbf{x}_i'\boldsymbol{\beta}). \quad [11]$$

The negative sign in the model for Z_i is a convention, to ensure that the linear predictor in [11] is in standard form. Model [11] states that the probability to be in any category given $\mathbf{x}_i$ depends on the cut-off points only. This is also known as the parallel lines or proportionality assumption (McCullagh 1980, Anderson 1984, Schabenberger 1995a). Although this appears restrictive at first, it allows a very parsimonious description of a multivariate response and ensures strict stochastic ordering, since $\alpha_j < \alpha_{j+1}$, $\forall\, i$.

Applying a logit link $\eta^{-1}(t) = \log(t / (1-t))$ to [11] yields

$$\log\left\{\frac{\Pr(C_i \le j|\mathbf{x}_i)}{\Pr(C_i > j|\mathbf{x}_i)}\right\} = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta} \quad [12]$$

which is known as the proportional odds model. To express this in terms of the Y_i it is useful to redefine Y_i as a cumulative indicator, $Y_{ip} = 1$ if C_i is in category p or less, 0 otherwise. Whether cumulative or direct indicators are used does not make a difference for a binary response, but simplifies the analysis when $P > 2$. [12] can be rewritten as

$$\log\left\{\frac{E(Y_{ip})}{1-E(Y_{ip})}\right\} = \alpha_p + \mathbf{x}_i'\boldsymbol{\beta}$$

or

$$E(Y_{ip}) = \eta(\alpha_p + \mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp\{\alpha_p + \mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\alpha_p + \mathbf{x}_i'\boldsymbol{\beta}\}}.$$

If $P = 2$ we of course have only 1 cut-off point that serves as intercept and the cumulative logit model reduces to a logistic regression model. The potential uses for the proportional odds model in forestry are plentiful, ranging from mortality models to the classification of disease propensities, stand density assessments, etc. Schabenberger (1995a) lists other potential applications and provides examples. The potential of [12] to serve as a classification tool has been examined by Greenwood and Farewell (1988). Regardless of P, model [12] can be fit with the optimal estimating function [5]. Assuming independence among the responses, [5] yields maximum likelihood estimates. For $P > 2$ the estimating algorithm [6] is a multivariate extension of iteratively reweighted least squares (McCullagh and Nelder 1989, Seber and Wild 1989). The independence assumption often will be not tenable however. If categorical (binary, ordered) responses are obtained repeatedly for any subject, serial or spatial correlation is introduced which has to be accounted for. Since maximum likelihood is to date the most important principle for estimation in GLM's it seemed natural to extend the GLM to allow for random terms in the linear predictor. Maximum likelihood analyses are difficult in this context, however, because one has to commence estimation from the marginal distribution of the observables, Y_{ij}, which requires integration over the random effects distribution. The non-linearity of the link function makes this difficult, because the integrals do not exist in closed form in most instances. One can employ a quadrature method to evaluate the integrals numerically (Jansen 1990, Longford 1993, Hedeker and Gibbons 1994) or approximate the marginal likelihood by response surface methods (Longford 1993), but both approaches are rather computationally intensive. The estimating function [5] provides a simpler and less cumbersome means of estimation.

For an ordered response with P categories, observed for subjects $i = 1,...,K$ at time points $j = 1,...,n_i$, one records C_{ij} and defines the cumulative indicators as

$$Y_{ijp} = 1 \text{ if } C_{ij} \le p$$
$$= 0 \text{ otherwise.}$$

The model for the mean response using a logit link function is

$$E(Y_{ijp}) = \mu_{ijp} = \eta(\alpha_p + \mathbf{x}_{ij}'\boldsymbol{\beta}) = \frac{\exp\{\alpha_p + \mathbf{x}_{ij}'\boldsymbol{\beta}\}}{1+\exp\{\alpha_p + \mathbf{x}_{ij}'\boldsymbol{\beta}\}}. \quad [13]$$

The responses are stacked as $\mathbf{Y}_{ij} = [Y_{ij1},...,Y_{ijP-1}]$, $\mathbf{Y}_i = [\mathbf{Y}_{i1}',...,\mathbf{Y}_{in_i}']'$, the means accordingly as $\mu_{ij} = [\mu_{ij1},...,\mu_{ijP-1}]$, $\mu_i = [\mu_{i1}',...,\mu_{in_i}']'$. Under the assumption of uncorrelatedness across subjects [5] can be rewritten as

$$U(\alpha_1,...,\alpha_{P-1},\boldsymbol{\beta};\mathbf{y}_i) = \sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - E(\mathbf{Y}_i)).$$

The elements of $\mathbf{D}_i$ are given by the derivatives of [13] with respect to the cut-off parameters and $\boldsymbol{\beta}$. If the observations within a subject were uncorrelated $\mathbf{V}_i$ would be block-diagonal with elements

$$\text{cov}(Y_{ijp}, Y_{ijp'}) = \mu_{ijp}(1-\mu_{ijp'}),\ p \le p' \quad [14]$$

where each block is of size $(P-1)*(P-1)$. These are the nominal variances and covariances of a cumulative multinomial random variable. In a mixed model where

$$E(Y_{ijp}|\mathbf{b}_i) = \mu_{ijp} = \eta(\alpha_p + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i),\ \mathbf{b}_i \sim (0,\mathbf{B})$$

$\mathbf{V}_i$ is given by [7] where $\mathbf{R}_i$ is the block-diagonal matrix with elements [14].

The application of a mixed POM for a repeatedly measured ordered response with four categories is illustrated with the following example.

The East Texas Pine Plantation Research Project (ETPPRP) is a long-term research program designed to investigate factors that affect management of loblolly pine (*Pinus taeda* L.) and slash pine (*Pinus elliottii* Engelm.) plantations in East Texas. We focus here on the slash pine data. Plots were visited in three year intervals and among other characteristics the status of each tree with respect to fusiform rust (*Cronartium quercuum* (Berk.) Miyabe ex Shirai f. sp. *fusiforme*) was recorded in 4 categories: *healthy*, *branch infected*, *stem infected*, and *dead*. For ease of identification the scores {0, 1, 2, 3} are assigned to the four categories in the analysis. Owing to the large number of plots and their geographical dispersion, only one third of all plots were visited in any year, creating a rotating panel design with three measurement waves. One of the research questions of interest is, whether it is possible to predict fusiform rust disease probabilities from classical inventory data such as site index, tree dbh, quadratic mean diameter, age, etc. and at the same time account for the serial correlation obviously present. Such models can be developed either on the plot, or the tree level. We focus on plot models in this example.

While the development of the rust disease is well understood (c.f. Powers et al. 1981), it is still a matter of investigation, if, and how well one can predict rust association with inventory data (Borders and Bailey 1986). Fitting only plot-level covariates that vary across time points (trees per acre, stand height, for example) or remain constant for a plot (site index, slope of terrain, etc.) one can make use of the fact that a number of trees on a plot share the same response and covariate pattern. This simplifies the analysis since the plots can serve as subjects and the outcome vector can be represented by a vector of counts representing the number of trees on a plot in any of the four states.

The four categories are clearly ordered, hence only models that preserve this ordering are meaningful, the POM is a convenient choice. The correlations between multiple measurements of the same plot are taken care of by means of random effects. It is reasonable to assume that the correlations between measurements on the same plot will decrease over time. In this case, it is advisable to chose the $\mathbf{Z}_i$ matrix in the linear predictor such that its rows are not constant, but vary over time. Covariates like age, stand height, basal area, etc. are particularly suited since they convey plot-specific information and are an actual or biological measure of time.

From a large suite of covariates the following have been identified as providing an adequate description of the fusiform rust disease category probabilities:

$\mathbf{x}_{ij} = [SL_i, A_{ij}, TPA_{ij}, SI_i, SH_{ij}, I_i]$

$\mathbf{z}_{ij} = \left[\frac{A_{ij}}{10}\right].$

The covariates denote:

SL_i	Slope of the plot area in %
A_{ij}	Age of plot i at measurement j
TPA_{ij}	Trees per acre
SI_i	Site index in feet
SH_{ij}	Stand height in feet
I_i	Indicator, 1 if initial age ≤ 5

The covariate in the random effects design matrix has been scaled to increase var($\mathbf{b}_i$). This is advisable in some cases to remove $\mathbf{B}$ from the boundary point 0. Note that SL_i and SI_i do not vary within a plot. The inclusion of I_i was motivated by experience. The fusiform rust disease process oftentimes develops differently and more strongly for trees that were infected at young ages. Wells and Dinus (1978) found age 5 as a meaningful point to subgroup the disease development. The covariate information in this model is typical inventory data. Table 3 depicts the results of fitting the model to the slash pine data. Over interpretation of the coefficients should be avoided because the study is not a designed experiment and the covariates do not *explain* fusiform rust incidence. They have simply been identified as a suite of well fitting descriptors.

The first three rows in Table 3 display the estimates of the cut-off points α_p. They are usually considered incidental parameters and no significance tests are attached to them (Schabenberger 1995a). It should be noted however, that the estimates are increasing $\hat{\alpha}_p < \hat{\alpha}_{p+1}$ as is necessary to preserve ordering. The closeness of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ is indicative of generally small probabilities to observe category 1 (branch infection) compared to the other categories.

The first column of Table 3 corresponds to a MLE fit ignoring correlations. In the second column a population-averaged model with 1-dependent correlation structure has been fitted. The weighted sums of squares as a prediction oriented goodness-of-fit criterion (Arabatzis 1990) decreased by 25 %. The results of the mixed model fit show another decrease in WSS over the 1-dependent structure. The coefficient estimates

Table 3. Results from fitting proportional odds model to slash pine data. WSS denotes weighted sums of squares, a predicted oriented goodness-of-fit criterion. Standard errors in parentheses. Cut-off points α_p are coded: 0 = healthy, 1 = branch infected, ...

Statistic		OEF under independence	OEF under 1-dependence	Mixed model OEF with random coeff.
$\hat{\alpha}_0$		2.01987	1.41983	2.51490
$\hat{\alpha}_1$		2.24311	1.65551	2.74971
$\hat{\alpha}_2$		5.36329	4.86911	6.01988
$\hat{\beta}_1$	SL_{ij}	0.01321	0.02459	0.01259
		(0.0048)	(0.0063)	(0.0092)
$\hat{\beta}_2$	A_{ij}	–0.02768	–0.01452	–0.00660
		(0.0121)	(0.0123)	(0.0177)
$\hat{\beta}_3$	TPA_{ij}	0.00164	0.00174	0.00177
		(0.00007)	(0.00009)	(0.0001)
$\hat{\beta}_4$	SI_{ij}	–0.02408	–0.01729	–0.03323
		(0.0023)	(0.0026)	(0.0033)
$\hat{\beta}_5$	SH_{ij}	–0.02530	–0.02473	–0.02841
		(0.0037)	(0.0036)	(0.0048)
$\hat{\beta}_6$	I_{ij}	0.81998	0.69550	0.91371
		(0.0580)	(0.0353)	(0.0606)
WSS		5644.1	4233.82	4108.68
$\hat{\mathbf{B}}$				0.23509

are close to those for the MLE fit in column 1, except for the fact that the coefficient for A_{ij} has changed due to the involvement of the same variable in the random part of the model. The standard error estimates under the mixed model specification are larger than the incorrect standard error estimates in the first column. Ignoring the correlations lends too much confidence to the precision of the coefficient estimates.

To calculate the category probabilities the coefficients are used in the following fashion. Assume the ijth measurement of a plot provides the following information:

SL_i: 10 %, A_{ij}: 9 years, TPA_{ij}: 270, SI_i: 65 ft, SH_{ij}: 30 ft, I_i: 0

Using the estimates in the right hand column of Table 3, the three linear predictors are calculated as

$\hat{\xi}_0 = 2.51490 + (-2.46576) = 0.04914$

$\hat{\xi}_1 = 2.74971 + (-2.46576) = 0.28394$

$\hat{\xi}_2 = 6.01988 + (-2.46576) = 3.55412$

and the cumulative probabilities are

$$\hat{\Pr}(C_{ij} \le 0) = \frac{\exp\{\hat{\xi}_0\}}{1+\exp\{\hat{\xi}_0\}} = 0.51228$$

$$\hat{\Pr}(C_{ij} \le 1) = \frac{\exp\{\hat{\xi}_1\}}{1+\exp\{\hat{\xi}_1\}} = 0.57051$$

$$\hat{\Pr}(C_{ij} \le 2) = \frac{\exp\{\hat{\xi}_2\}}{1+\exp\{\hat{\xi}_2\}} = 0.97219$$

The probabilities to fall in any category are obtained by subtraction as

$\hat{\Pr}(C_{ij} = 0) = 0.51228$
$\hat{\Pr}(C_{ij} = 1) = 0.57051 - 0.51228 = 0.05823$
$\hat{\Pr}(C_{ij} = 2) = 0.97219 - 0.57051 = 0.40168.$

Consequently, on a plot with the above covariates there is a 51 % chance, that a randomly selected tree will be healthy, a 6 % chance of having a branch infection, a 40 % chance of being stem infected and about 3 % of the trees will have died because of fusiform rust. It is believed that such analyses are highly meaningful to determine the risks and liabilities associated with pine management in the southern forests of the United States.

How well the model predicts overall has been assessed through a second, independent evaluation data set that arises, since each plot was divided into two subplots. One was used for development of the model, the other can be used for evaluation. The mixed model provided the best fit to the data. Since this is an independent data set, the BLUPs have not been used. The proportions of observed and predicted responses are

Category	0	1	2	3
Observed	0.5415	0.0516	0.3711	0.0358
Predicted	0.5629	0.0504	0.3589	0.0278

showing good agreement.

4.4 Monte Carlo Results

To gain additional insight about the performance of the estimating function approach compared to maximum likelihood, and to highlight the liabilitities of analyzing data incorrectly under the assumption of independent observations, a small scale simulation study was conducted. Observations were generated randomly for a Coile-Schumacher height equation (Clutter et al. 1983), where the intercepts vary between subjects. The model can be written on the logarithmic scale as

$$\ln(H_{ij}) = \beta_0 + \beta_1 \frac{1}{A_{ij}} + b_i + \varepsilon_{ij} \qquad [15]$$

where H_{ij} is height of the ith tree at the jth remeasurement ($j = 1,...,n_i$), A_{ij} is age in years associated with the ijth measurement, b_i is a random subject effect distributed according to $b_i \sim N(0, \tau^2)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The random disturbances ε_{ij}, $\varepsilon_{ij'}$ were chosen independently of each other and independently of the b_i's from their respective distributions. The fixed effects coefficients were chosen as $\beta_0 = 4$, $\beta_1 = -15$. The error structure in [15] implies that var($\mathbf{H}_i$) = $\tau^2\mathbf{J}_{n_i} + \sigma^2\mathbf{I}_{n_i}$, and that two observations from the same subject have a constant correlation of $\rho = \tau^2 / (\tau^2 + \sigma^2)$. It is for this reason, that the error structure in [15] has been termed the compound symmetry or exchangeable correlation structure (Longford 1993). The values of τ^2 and σ^2 were varied in the simulations, in order to examine the effect of differing degrees of the dependency among elementary observations. For each run, data were generated according to [15], and $\boldsymbol{\beta} = [\beta_0, \beta_1]'$ was estimated by ordinary least squares, ignoring the correlation among observations of the same subject; restricted maximum likelihood; and the OEF [5] combined with a moment estimator for τ^2. The full likelihood implementation is described for example in Gregoire, Schabenberger, and Barrett (1995). The number of repeated measurements per subject was also selected randomly between 1 and $n_{max} = 10$ to impose an unbalanced data structure. This process was repeated 50 times for

Table 4. Results from Monte-Carlo study of Coile-Schumacher height growth data. K = 30 subjects, n_{max} = 10 $\hat{se}(\hat{\beta}_k)$ denotes the estimated, $se(\hat{\beta}_k)$ the exact standard errors of the kth coefficient.

	$\sigma^2 = 0.1, \tau^2 = 0.01$ $\rho = 0.091$			$\sigma^2 = 0.01, \tau^2 = 0.01$ $\rho = 0.5$			$\sigma^2 = 0.005, \tau^2 = 0.05$ $\rho = 0.909$		
Statistic	OLS	OEF	REML	OLS	OEF	REML	OLS	OEF	REML
$\hat{\beta}_0$	3.9895	3.9883	3.9909	3.9924	3.9930	3.9930	3.9853	3.9841	3.9841
$\hat{se}(\hat{\beta}_0)$	0.0305	0.0366	0.0370	0.0129	0.0213	0.0219	0.0212	0.0401	0.0416
$se(\hat{\beta}_0)$	0.0376	0.0370	0.0370	0.0249	0.0221	0.0221	0.0521	0.0421	0.0421
$\hat{\beta}_1$	–15.04	–15.11	–15.08	–15.05	–15.06	–15.06	–15.13	–15.04	–15.03
$\hat{se}(\hat{\beta}_1)$	0.6639	0.6967	0.7228	0.2798	0.3044	0.3061	0.4592	0.2554	0.2543
$se(\hat{\beta}_1)$	0.7651	0.7439	0.7439	0.4575	0.3149	0.3149	0.9256	0.2598	0.2598
$\hat{\sigma}^2$	0.1069	0.0969	0.0959	0.0192	0.0093	0.0097	0.0518	0.0048	0.0479
$\hat{\tau}_2$		0.0115	0.0115		0.0093	0.0100		0.0464	0.0500

each setting of ρ and the results averaged. Table 4 lists these averages for 50 repetitions from selected runs.

Restricted maximum likelihood has been chosen over maximum likelihood for estimating the covariance parameters τ^2 and σ^2 because REML estimators exhibit less bias in unbalanced data sets than ML estimators. The correlation increases in Table 4 from the left to the right panel. The agreement between the OEF and REML estimates is remarkable, regardless of the strength of the correlation. Not much information about the model parameters is lost in the OEF approach while at the same time only minimal assumptions are required. The standard errors $se(\hat{\beta}_k)$ for OEF and REML are identical, since both are asymptotic and typically rendered conditionally on the estimated covariance parameters.

The moment estimator [8] is unconstrained, i.e. its middle piece $(\mathbf{y}_i - \eta)(\mathbf{y}_i - \eta)' - \mathbf{R}_i$ can take on negative values and force $\hat{\mathbf{B}}$ to be negative. Table 4 shows that even small covariance parameters such as $\tau^2 = 0.01$ can be estimated without problems.

It is obvious that the estimates for β_0 and β_1 obtained under OLS remain unbiased, even if the error structure is not correctly specified. There is little difference between the OLS, OEF, and REML estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$. This is a well established result. However, it is also known, that not taking the error structure, i.e. correlations into account, affects the standard error estimates in a most devastating way. As the correlation increases, the OLS estimates become increasingly inefficient compared to either mixed model analysis, as the exact standard errors of the coefficients $se(\hat{\beta}_0)$ and $se(\hat{\beta}_1)$ depict. For $\rho = 0.909$ in the last columns of Table 4. OLS is 3.5 times less efficient than REML or OEF in estimating the slope β_1. For $\rho = 0.091$ there is, however, hardly a difference as one would expect.

In practical applications these standard errors are of course unknown, and inference is based on their estimates $\hat{se}(\hat{\beta}_0)$ and $\hat{se}(\hat{\beta}_1)$. With increasing ρ the error model assumed by OLS, i.e. $\text{var}(\mathbf{H}_i) = \sigma^2\mathbf{I}_{n_i}$ deviates more from the actual variance of the observations. As this deviation increases, the OLS estimates not only become less efficient, the estimates of their standard errors become increasingly negatively biased, thereby giving a false sense of precision. This bias and inconsistency of $\hat{se}(\hat{\beta}_k)_{OLS}$ affects all aspects of inference, from hypothesis tests, confidence intervals, to influential data diagnostics. Typically in forestry, an analyst usually knows or suspects when observations are correlated, although the strength and structure of this correlation may not be discernible.

5 Discussion and Conclusion

This paper is concerned with presenting and demonstrating a unified approach to statistical estimation to a forestry audience. The approach is semi-parametric in that it is entirely based on only mean and variance specifications. This has the advantage of requiring minimal assumptions and leads to very efficient and well behaved estimating algorithms. However, the cost involved is the lack of a parametric basis for model inference. Much depends on asymptotic results for the model at hand. Since no likelihood function is specified, likelihood ratio tests are not possible to discriminate between competing models. Different criteria have to be used. In this paper focus was on the predictive capabilities of the models since many models in forestry are developed to serve as predictive tools. Goodness-of-fit statistics and model diagnostics for semi-parametric estimation procedures is currently an area of active research. In a sense, one faces the same dilemma as in regular linear regression with the least squares principle. To perform inference under the model one tacitly makes a distributional assumption, commonly normality, after estimates of the model coefficients have been obtained. One is willing to accept a rather strong assumption after estimation has taken place in order to establish a basis for inference.

We distance ourselves from doing so in the estimation function approach. Permitting minimal assumptions is a procedural merit. These assumptions are typically rather weak compared to distributional requirements. Data analysts will have more confidence in selecting a mean model than in specifying a distribution for the random quantities in the model. Modifying the set of assumptions after estimation usually makes the entire analysis more vulnerable, since added assumptions are typically more stringent than the initial ones. Research in the area of inference, influence and outlier diagnostics, and model discrimination has to provide the proper tools to give answers based on only the minimal assumptions involved in the estimation step.

The estimation function used in this paper resembles a quasi-likelihood function (McCullagh 1983, McCullagh and Nelder 1989, Candy 1989, Nelder and Lee 1992). The two approaches differ in spirit, however. Quasi-likelihoods are based on an assumed mean model, a link function, and a functional mean-variance relationship. The idea is to find a function that behaves like a likelihood score function, i.e. has zero expectation and whose gradient contains information about dispersion. OEF's are found by selecting functions that are close to the likelihood score function. Since discussion was restricted to estimating functions in $\mathfrak{S}$, which have zero expectation by definition, it is not surprising that OEF [5] is the quasi-likelihood in some cases. If one broadens the class of estimating functions, for example by including EF's that are non-linear in $\mathbf{Y}$, the correspondence between estimating function and quasi-likelihood theory does not hold.

Acknowledgments

We are indebted to Dr. J. David Lenhart for providing the slash pine data used in the fusiform rust analysis and the white pine data used in the second example.

References

Agresti, A. 1990. Categorical data analysis. John Wiley and Sons, Inc., New York.

Amateis, R.L. & Burkhart, H.E. 1987. Cubic-foot volume equations for loblolly pine trees in cutover, site-prepared plantations. Southern Journal of Applied Forestry 11: 190–192.

Anderson, J.A. 1984. Regression and ordered categorical variables. Journal of the Royal Statistics Society (B) 46(1): 1–30.

Arabatzis, A.A. 1990. Qualitative response models theory and its application to forestry. Unpublished Ph.D. dissertation. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Bhapkar, V.P. 1991. Sufficiency, ancillarity, and information in estimating functions. In: Godambe, V.P. (ed.). Estimating functions. Clarendon, Oxford. p. 241–254.

Bickel, P.J. & Doksum, K.A. 1977. Mathematical statistics: Basic ideas and selected topics. Holden-Day, Oakland.

Borders, B.E. & Bailey, R.L. 1986. Fusiform rust prediction models for site-prepared slash and Loblolly pine plantations in the Southeast. Southern Journal of Applied Forestry 10: 145–151.

Breslow, N.E. & Clayton, D.G. 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88(421): 9–25.

Candy, S.G. 1989. Growth and yield models for Pinus radiata in Tasmania. New Zealand Journal of Forestry Science 19: 112–133.

Clutter, J.L., Fortson, J.C., Pienaar, L.V., Brister, G.H. & Bailey, R.L. 1983. Timber management: A quantitative approach. John Wiley and Sons, New York.

Godambe, V.P.1960. An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31: 1208–1211.

— & Heyde, C.C. 1987. Quasi-likelihood and optimal estimation. International Statistical Review 55: 213–244.

— & Kale, B.K. 1991. Estimating functions: an overview. In: Godambe, V.P. (ed.). Estimating functions. Clarendon, Oxford. p. 3–20.

— & Thompson, M.E. 1989. An extension of quasi-likelihood estimation. Journal of Statistical Planning and Inference 22: 137–152.

Greenwood, C & Farewell, V. 1988. A comparison of regression models for ordinal data in an analysis of transplanted-kidney function. The Canadian Journal of Statistics 16(4): 325–335.

Gregoire, T.G. 1987. Generalized error structure for forestry yield models. Forest Science 33: 423–444.

— & Schabenberger, O. 1994. Fitting bole-volume equations to spatially correlated within-tree data. In: Schwenke, J.R. & G.A. Milliken, G.A. (eds.). Proceedings of the 6th annual KSU conference on Applied Statistics in Agriculture, April 24–26, 1994, Manhattan, Kansas. In press.

— , Schabenberger, O. & Barrett, J.P. 1995. Modeling irregularly spaced, unbalanced, longitudinal data from permanent plot measurements. Canadian Journal of Forest Research 25. In press.

Hedeker, D.; Gibbons, R.D. 1994. A random-effects ordinal regression model for multilevel analysis. Biometrics 51. In press.

Hosmer, D.W.Jr. & Lemeshow, S. 1989. Applied logistic regression. John Wiley & Sons, Inc., New York.

Jansen, J. 1990. On the statistical analysis of ordinal data when extravariation is present. Applied Statistics 39(1): 75–84.

Jones, R.H. 1993. Longitudinal data with serial correlation: A state-space approach. Chapman & Hall, New York.

— & Ackerson, L.M. 1990. Serial correlation in unequally spaced longitudinal data. Biometrika 77: 721–731.

Laird, N.M; Ware, J.H. 1982. Random-effects models for longitudinal data. Biometrics 38: 963–974.

Liang, K.Y. 1992. Extensions of generalized linear models in the past twenty years. Overview and some biomedical applications. Proceedings of the XVIth International Biometric Conference, Hamilton, New Zealand, December 7–11, 1992.

— & Zeger, S.L. 1986. Longitudinal analysis usung generalized linear models. Biometrika 73(1): 13–22.

— , Zeger, S.L. & Qaqish, B. 1992. Multivariate regression analyses for categorical data. Journal of the Royal Statistics Society (B) 54(1): 3–40.

Lindstrom, M.J. & Bates, D.M. 1990. Nonlinear mixed effects models for repeated measures data. Biometrics 46: 673–687.

Longford, N.T. 1993. Random coefficient models. Clarendon, Oxford.

McCullagh, P. 1980. Regression models for ordinal data. Journal of the Royal Statistics Society (B) 42(2): 109–142.

— 1983. Quasi-likelihood functions. The Annals of Statistics 11(1): 59–67.

— & Nelder, J.A. 1989. Generalized linear models. 2. ed.. Chapman and Hall, London–New York.

Nelder, J.A. & Wedderburn, R.W.M. 1972. Generalized linear models. Journal of the Royal Statistics Society (A) 135: 370–384.

— & Lee, Y. 1992. Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. Journal of the Royal Statistics Society (B) 54(1): 273–284.

Powers, H.R. Jr., Schmidt, R.A.. & Snow, G.A. 1981. Current status and management of fusiform rust on southern pines. Ann. Rev. Phytopathology 19: 353–371.

Pregibon, D. 1981. Logistic regression diagnostics. Annals of Statistics 9: 705–724.

Rao, C.R. 1973. Linear statistical inference and its applications. 2. ed. John Wiley and Sons, New York.

Schabenberger, O. 1993. Loglinear models for square contingency tables. The validation of classifications. Paper presented at the 1993 Spring Statistical Meeting of the Biometric Society (ENAR), Philadelphia, March 21–24, 1993.

— 1995a. The use of ordinal response methodology in forestry. Forest Science 41. In press.

— 1995b. Nonlinear mixed effects growth models for repeated measures in ecology. Proceedings of the 1994 Joint Statistical Meetings, Toronto, August 1994. American Statistical Association, Section on Statistics and the Environment. In press.

Searle, S.R., Casella, G. & McCulloch, C.E. 1992. Variance components. John Wiley and Sons, New York.

Seber, G.A.F. & Wild, C.J. 1989. Nonlinear regression. John Wiley and Sons, New York.

Wells, O.O. & Dinus, R.J. 1987. Early infections as a predictor of mortality associated with fusiform rust of southern pines. Journal of Forestry 76: 8–12.

Wolfinger, R. 1993. Laplace's approximation for nonlinear mixed models. Biometrika 80(4): 791–795.

Zeger, S.L. & Liang, K.-Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42: 121–130.

— , Liang, K.-Y. & Albert, P.S. 1988. Models for longitudinal data: A generalized estimating equation approach. Biometrics 44: 1049–1060.

Total of 46 references

Appendix: Some Technical Details

A1. Regularity conditions

For the development of optimal EF's some regularity conditions are required.

Let $U(\mathbf{x}, \theta) = \partial \log\{p(\mathbf{x}, \theta)\} / \partial\theta$ denote the likelihood score function where the parameter θ belongs to a space Θ. We then require that

(i) Θ is an open interval on the real line,

(ii) first and second derivatives of the log likelihood with respect to θ exist,

(iii) differentiation and integration with respect to $p(\mathbf{x}, \theta)$ and $\log\{p(\mathbf{x}, \theta)\}$ are interchangeable.

(iv) The variance of the score function $U(\mathbf{x}, \theta)$ is finite for all θ.

These regularity conditions are known from theorems about the Cramér-Rao bound or the information inequality (Rao 1973, Bickel and Doksum 1977). These conditions are met for densities in the exponential family of distributions. It is furthermore required that $g(\mathbf{x}, \theta)$ is informative with respect to θ, i.e. $\partial g(\mathbf{x}, \theta)\} / \partial\theta$ exists and $\text{var}(g(\mathbf{x}, \theta) > 0 \; \forall \; \theta \in \Theta$. This is an obvious requirement, since a function like $\sum_i x_i = 0$ that does not depend on θ and implies no estimate.

A2. Optimal estimating function in $\mathfrak{S}$

In the class $\mathfrak{S}$ of unbiased EF's linear in $\mathbf{Y}$, natural candidates are of the form

$$U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{H}'(\mathbf{y} - E(\mathbf{Y})) \qquad [16]$$

where $(\mathbf{y} - E(\mathbf{Y}))$ ensures unbiasedness and linearity in $\mathbf{Y}$, $\mathbf{H}$ is a $(N*p)$ matrix that maps the expectation into the parameter space. $E(\mathbf{Y})$ depends on the $(p*1)$ parameter vector $\boldsymbol{\beta}$. To judge all candidates of form [16] against each other, we need an expression for the standardized variance of [16]. Following Bhapkar (1991) and Godambe and Heyde (1987) for the treatment of vector parameters, the standardized variance-covariance matrix has form

$$E\left(\frac{\partial U(\boldsymbol{\beta},\mathbf{y})}{\partial \boldsymbol{\beta}}\right)' E\left(U(\boldsymbol{\beta},\mathbf{y})U(\boldsymbol{\beta},\mathbf{y})'\right)^{-1} E\left(\frac{\partial U(\boldsymbol{\beta},\mathbf{y})}{\partial \boldsymbol{\beta}}\right).$$

Among two competing EF's $U_1(\boldsymbol{\beta}, \mathbf{y})$ and $U_2(\boldsymbol{\beta}, \mathbf{y})$, say, $U_1(\boldsymbol{\beta}, \mathbf{y})$ is preferable iff

$$E\left(\frac{\partial U_1}{\partial \boldsymbol{\beta}}\right)' E(U_1U_1')^{-1} E\left(\frac{\partial U_1}{\partial \boldsymbol{\beta}}\right) - E\left(\frac{\partial U_2}{\partial \boldsymbol{\beta}}\right)' E(U_2U_2')^{-1} E\left(\frac{\partial U_2}{\partial \boldsymbol{\beta}}\right) \qquad [17]$$

is non-negative definite. This would imply that $U_1(\boldsymbol{\beta}, \mathbf{y})$ contains more information about the parameter than $U_2(\boldsymbol{\beta}, \mathbf{y})$. A candidate against [16] is

$$U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{D}'\mathbf{V}(\mathbf{y} - E(\mathbf{Y})),$$

where $\mathbf{D} = \partial E(\mathbf{Y}) / \partial \boldsymbol{\beta}$ is of dimension $(N*p)$. But since $\partial U_2(\boldsymbol{\beta}, \mathbf{y}) / \partial \boldsymbol{\beta} = -\mathbf{H}'\mathbf{D}$, $E(U_2 U_2') = \mathbf{H}'\mathbf{V}\mathbf{H}$, $\partial U_1(\boldsymbol{\beta}, \mathbf{y}) / \partial \boldsymbol{\beta} = -\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}$, $E(U_1 U_1') = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}$, [16] reduces to

$$\mathbf{D}'(\mathbf{V}^{-1} - \mathbf{H}'(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}')\mathbf{D}$$

which is a residual covariance matrix, where the linear effect of $\mathbf{H}'\mathbf{y}$ was removed from $\mathbf{D}'\mathbf{V}^{-1}\mathbf{y}$, hence non-negative definiteness is guaranteed (McCullagh and Nelder 1989). Apparently, $U(\boldsymbol{\beta}, \mathbf{y})$ is the estimating function that improves all other members in $\mathfrak{S}$.

A3. Minimal dispersion of $\hat{\boldsymbol{\beta}}$

To see, how well the estimate $\hat{\boldsymbol{\beta}}$ performs compared to other candidates in $\mathfrak{S}$ we will give an asymptotic result since we have to allow for the case that solutions to $U(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{0}$ do not exist in closed form.

Let $U_2(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{H}'(\mathbf{y} - E(\mathbf{Y}))$ denote any other estimating function in $\mathfrak{S}$ and $\tilde{\boldsymbol{\beta}}$ the root of $U_2(\boldsymbol{\beta}, \mathbf{y}) = \mathbf{0}$. We want to know, whether $\mathrm{var}(\mathbf{a}'\tilde{\boldsymbol{\beta}}) \geq \mathrm{var}(\mathbf{a}'\hat{\boldsymbol{\beta}})$ which implies that $\mathrm{var}(\tilde{\boldsymbol{\beta}}) - \mathrm{var}(\hat{\boldsymbol{\beta}})$ is non-negative definite or $\mathrm{var}(\hat{\boldsymbol{\beta}})^{-1} - \mathrm{var}(\tilde{\boldsymbol{\beta}})^{-1}$ is non-negative definite. For $U_1(\boldsymbol{\beta}, \mathbf{y})$ we were led earlier to $\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}$. To obtain $\mathrm{var}(\tilde{\boldsymbol{\beta}})$ use a Taylor series approximation

$$U_2(\tilde{\boldsymbol{\beta}}, \mathbf{y}) \doteq U_2(\boldsymbol{\beta}, \mathbf{y}) + \mathbf{H}'\mathbf{D}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

which implies $\mathrm{var}(\tilde{\boldsymbol{\beta}}) \doteq (\mathbf{H}'\mathbf{D})^{-1}(\mathbf{H}'\mathbf{V}\mathbf{H})(\mathbf{D}'\mathbf{H})^{-1}$. Substituting these expressions into

$$\mathrm{var}(\hat{\boldsymbol{\beta}})^{-1} - \mathrm{var}(\tilde{\boldsymbol{\beta}})^{-1}$$

yields

$$\mathbf{D}'(\mathbf{V}^{-1} - \mathbf{H}'(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}')\mathbf{D}.$$

This is the same matrix found to be non-negative definite in A1. The optimality of the estimating function [5] translates directly into asymptotically minimally dispersed estimators.