Some Considerations on the Fisher Information in Nonlinear Mixed Effects Models

T. Mielke, R. Schwabe

Abstract The inverse of the Fisher Information Matrix is a lower bound for the covariance matrix of any unbiased estimator of the parameter vector and, given this, it is important for the construction of optimal designs. For normally distributed observation vectors with known variance, the Fisher Information can be easily constructed. For nonlinear mixed effects models, the problem of the missing closed-form solution of the likelihood function carries forward to the calculation of the Fisher Information matrix. The often used approximation of the Fisher Information by linearizing the model-function in the fixed effects is generally not reliable, as will be shown in this article.

1 Introduction

In population pharmacokinetic studies, the observations of different individuals are often assumed to follow one common function with small differences, which are generated by random individual parameters. One main interest in these studies lies in the estimation of the population parameters. Usually maximum likelihood estimation is desirable as fewer observations per individual are needed to estimate the population parameters, than for a two-stage procedure. The occurring models are nonlinear in the random parameters and with this the likelihood generally cannot be described in a explicit form. Numerical procedures, such as described by Davidian and Giltinan (1995) or Pinheiro and Bates (2000), are used to approximately solve the maximum likelihood problem. Knowledge of the Fisher Information is of interest for designing the experiment. A well known approach to approximating the Fisher Information is to linearize the regression function and to assume the linearized model to be normally distributed.

T. Mielke , R. Schwabe
Otto-von-Guericke University Magdeburg, e-mail: tobias.mielke@ovgu.de
This work was supported by the BMBF grant SKAVOE 03SCPAB3

In this article we outline some problems occurring, when using this approximated Fisher Information. The second section briefly describes estimation in nonlinear models and asymptotic distributions of estimators. Continuing from the ideas of nonlinear regression and of linear mixed effects models, we describe in section 3 the first-order linearization, which is used to approximate the Fisher Information. In section 4 a problem with this approximation is illustrated through a simple example.

2 Non-linear models

In our considered model, the observation $Y(x_i)$, taken at known experimental settings x_i in a design region X is modeled by

$$Y(x_i) = \eta(x_i,\beta) + \varepsilon_i, \text{with } E(\varepsilon_i) = 0 \text{ and } Var(\varepsilon_i) = \sigma^2.$$

The real valued regression function η is assumed to be nonlinear in the unknown parameter vector $\beta \in \mathbb{R}^p$. To avoid difficulties we assume that η is continuous in x_i and differentiable in β. For an unknown error-distribution, a standard approach for the estimation of the vector β would be the use of least squares techniques.

Let $\xi = (x_1,..,x_k)$ be a concrete design and denote

$$F_\beta(\xi) := (\frac{\partial \eta(x_1,\beta)}{\partial \beta},...,\frac{\partial \eta(x_k,\beta)}{\partial \beta})^T.$$

For a vector β_0 near to the true parameter vector β, the nonlinear model can be approximated by a linear model:

$$Y \approx \eta(\xi,\beta_0) + F_{\beta_0}(\xi)(\beta - \beta_0) + \varepsilon,$$

with vectors $Y = (Y(x_1),...,Y(x_k))^T$, $\eta(\xi,\beta_0) = (\eta(x_1,\beta_0),...,\eta(x_k,\beta_0))^T$ and $\varepsilon = (\varepsilon_1,...,\varepsilon_k)^T$. Under the assumption of a negligible linearization error, estimation of β in the approximated model

$$Y_{\beta_0} = F_{\beta_0}(\xi)\beta + \varepsilon, \text{ where } Y_{\beta_0} := Y - \eta(\xi,\beta_0) + F_{\beta_0}(\xi)\beta_0,$$

leads to an updated guess for the true parameter vector β. For β_0 close enough to β, this procedure leads to an estimate approximately fulfilling the estimating equation

$$F_\beta(\xi)^T(y - \eta(\xi,\beta)) = 0$$

which is fulfilled for the ordinary least squares estimator $\hat{\beta}_{OLS}$. For homoscedastic errors ε_i, n replications of the design ξ and under appropriate regularity conditions, the ordinary least squares estimator $\hat{\beta}_{OLS}$ is asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \to N(0, \sigma^2(F_\beta(\xi)^T F_\beta(\xi))^{-1}) \text{ as } n \to \infty$$

and has for normally distributed homoscedastic errors ε_i the nice property of coinciding with the *ML*-estimator.

As in ordinary least squares estimation each deviation receives equal weights, this method may be inefficient for heteroscedastic observation errors. Weighted least squares estimators for known variance structures and generalized least squares methods for unknown variance structures take the heteroscedasticities into account. With a variance matrix $V_\beta(\xi) = diag(\sigma^2(x_1,\beta),...,\sigma^2(x_k,\beta))$ depending on the experimental settings and the parameter vector β, one might transform the original model into a homoscedastic error model. Starting from a prior guess β_0 of the parameter vector β and a prior guess $V_{\beta_0}(\xi)$ of $V_\beta(\xi)$, new iterates for estimating β and $V_\beta(\xi)$ in the transformed model can be deduced. For variance functions $\sigma^2(x,\beta)$, which are known up to the vector β, independent errors and with some regularity conditions, the *GLS* estimator $\hat{\beta}_{GLS}$ is asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) \to N(0, (F_\beta(\xi)^T V_\beta(\xi)^{-1} F_\beta(\xi))^{-1}) \text{ as } n \to \infty.$$

Contrary to the homoscedastic case with normally distributed errors, the *ML*-estimator and *GLS*-estimator generally do not coincide for heteroscedastic errors. For nonlinear regression functions with normally distributed heteroscedastic errors and a known variance matrix $V_\beta(\xi)$, additional information can be drawn from the variance of the observations, such that the Fisher information for β results in

$$\begin{aligned} M_\beta(\xi) &= E(\frac{\partial \ln(f_Y(y,\beta))}{\partial \beta}^T \frac{\partial \ln(f_Y(y,\beta))}{\partial \beta}) \\ &= F_\beta^T(\xi) V_\beta(\xi)^{-1} F_\beta(\xi) + \frac{1}{2}\tilde{S}(\xi), \end{aligned}$$

where $f_Y(y,\beta)$ is the likelihood function for the model of Y and $\tilde{S}(\xi)$ is a matrix with

$$\tilde{S}(\xi)_{ij} = Tr(\frac{\partial V_\beta(\xi)}{\partial \beta_i} V_\beta(\xi)^{-1} \frac{\partial V_\beta(\xi)}{\partial \beta_j} V_\beta(\xi)^{-1}), \; i \text{ and } j = 1,...,p.$$

The *ML*-estimator in the normal model with heteroscedastic errors is known to be asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_{ML} - \beta) \to N(0, M_\beta(\xi)^{-1}) \text{ as } n \to \infty$$

and with this it follows that for normally distributed heteroscedastic errors the *ML*-estimator is asymptotically more efficient than the *GLS*-estimator.

As Davidian and Giltinan (1995) point out, the *ML*-estimator under assumed normality loses the advantage of efficiency very quickly in the case of nonnormal data and is highly sensitive to outlying observations, while the *GLS*-estimator is more robust. Moreover, misspecified variance functions $\sigma^2(x,\beta)$ may result in bi-

ased *ML*-estimates, whereas *GLS*-estimates are not this sensitive to variance function misspecification.

Nonlinear models with heteroscedastic and homoscedastic errors have in common that the variance of the estimator $\hat{\beta}$ depends on the parameter vector β itself, such that optimal designs for these models are in general just locally optimal.

3 Mixed-Effects Models

Consider that the j-th observation of individual i with experimental settings $x_{ij} \in X$ is described by

$$Y(x_{ij}) = \eta(x_{ij}, \beta_i) + \varepsilon_{ij}.$$

The individual parameter vector β_i is assumed to be a random vector with mean β and some covariance matrix D. The observation error is assumed to have zero mean and a known constant variance σ^2. Observation errors and individual parameter vector are considered to be independent of each other.

For linear regression functions, the assumption of normally distributed individual parameter vectors and observation errors carries forward to the marginal distribution of the observation vector. Let $Y = (Y_1^T, ..., Y_N^T)$ describe the vector of all observations, where $Y_i = (Y(x_{i1}), ..., Y(x_{im_i}))^T$ is the observation vector of the i-th individual with a concrete design ξ_i. As the regression function is linear in the parameter vector β, the design matrix $F(\xi_i) = F_\beta(\xi_i)$ defined in the previous section does not depend on the parameter vector. For the observation vector Y_i we obtain

$$Y_i \sim N(F(\xi_i)\beta, (F(\xi_i)DF(\xi_i)^T + \sigma^2 I_{m_i})).$$

Often primary interest lies in estimating the mean parameter vector β. With

$$\begin{aligned} F &:= (F(\xi_1)^T, ..., F(\xi_N)^T)^T, \\ G &:= diag(F(\xi_1), ..., F(\xi_N)), \\ V(\xi_i) &:= (F(\xi_i)DF(\xi_i)^T + \sigma^2 I_{m_i}) \text{ and} \\ V &:= diag(V(\xi_1), ..., V(\xi_N)), \end{aligned}$$

the model of all observations is described by

$$Y = F\beta + Gb + \varepsilon, \text{ where } b = ((\beta_1 - \beta)^T, ..., (\beta_N - \beta)^T)^T \text{ and } \varepsilon = (\varepsilon_1^T, ..., \varepsilon_N^T)^T.$$

It readily follows that $Y \sim N(F\beta, V)$ and that *ML*- and *GLS*-estimator in the case of a known matrix D and σ^2 coincide:

$$\hat{\beta}_{ML} = \hat{\beta}_{GLS} = (F^T V^{-1} F)^{-1} F^T V^{-1} Y \text{ and } cov(\hat{\beta}_{ML}) = (F^T V^{-1} F)^{-1}.$$

Note that for a variance matrix V_β depending on the parameter vector β, *ML*- and *GLS*-estimator generally do not coincide, as the Fisher Information is then of a similar form as for nonlinear regression functions with normally distributed heteroscedastic errors (Atkinson and Cook (1995)). For nonlinear mixed effects models, these results usually cannot be observed, as the observations will generally not be normally distributed. For estimating the population parameters in nonlinear mixed effects models, two-stage procedures might be helpful. In a first step individual parameter vectors should be estimated and based on these estimates the population parameter vector might be estimated. However, often reliable individual estimates cannot be obtained for the subjects. Maximum likelihood estimation on the marginal model of the observations would then be an alternative approach to obtain reliable estimates of the population parameter vector. Due to the nonlinearity of the regression function in the random parameters, a closed-form description of the likelihood of the observations y is in general nonexistent. Different numerical approaches are used to make the optimization of the likelihood a tractable problem.

If we have prior knowledge in the form of a first guess β_0 of the true population mean β, then linearization of the model around β_0 leads on the individual level by

$$\begin{aligned} Y_i &= \eta(\xi_i,\beta_i)+\varepsilon_i \\ &\approx \eta(\xi_i,\beta_0)+F_{\beta_0}(\xi_i)(\beta_i-\beta_0)+\varepsilon_i \\ &= \eta(\xi_i,\beta_0)+F_{\beta_0}(\xi_i)(\beta-\beta_0)+F_{\beta_0}(\xi_i)(\beta_i-\beta)+\varepsilon_i \end{aligned}$$

to a linear mixed effects model. With the earlier assumptions of the normal distribution for the parameter and the error vector and with the assumption that the approximating model is almost exact, one obtains

$$Y_{i,\beta_0} = F_{\beta_0}(\xi_i)\beta + F_{\beta_0}(\xi_i)(\beta_i-\beta)+\varepsilon_i, \text{ with } Y_{i,\beta_0} := Y_i - \eta(\xi_i,\beta_0)+F_{\beta_0}(\xi_i)\beta_0.$$

As a consequence it is assumed

$$Y_{i,\beta_0} \sim N(F_{\beta_0}(\xi_i)\beta, V_{\beta_0}(\xi_i)), \text{ where } V_{\beta_0}(\xi_i) := F_{\beta_0}(\xi_i)DF_{\beta_0}(\xi_i)^T + \sigma^2 I_{m_i}$$

and with this

$$\hat{\beta} = (\sum_{i=1}^{N} F_{\beta_0}(\xi_i)^T V_{\beta_0}(\xi_i)^{-1} F_{\beta_0}(\xi_i))^{-1} \sum_{i=1}^{N} F_{\beta_0}(\xi_i) V_{\beta_0}(\xi_i)^{-1} Y_{i,\beta_0}$$

is the *ML*-Estimator for β in the linearized model around β_0 and might be used as starting point for a next iteration.

A second possible approach would be the linearization of the function η around the unknown expected value of β_i as described by Davidian and Giltinan (1995):

$$\begin{aligned} Y_i &= \eta(\xi_i,\beta_i)+\varepsilon_i \\ &\approx \eta(\xi_i,\beta)+F_\beta(\xi_i)(\beta_i-\beta)+\varepsilon_i. \end{aligned}$$

Assuming the linearization error as negligible and with a covariance matrix $V_\beta(\xi_i) = F_\beta(\xi_i)DF_\beta(\xi_i)^T + \sigma^2 I_{m_i}$ depending on β, the marginal model results in

$$Y_{i,\beta} \sim N(\eta(\xi_i,\beta), V_\beta(\xi_i)).$$

For the estimation of β one might in this case recall the parameter estimation in nonlinear heteroscedastic models with normal observation errors.

This first-order linearization around the expected value of the individual parameter vector is often used to approximate the true nonlinear mixed effects model. In their description of the first-order linearization in nonlinear mixed effects models, Davidian and Giltinan (1995) point out that the observation vector Y is taken in this method as approximately normally distributed with the moments

$$E(Y_i) \approx \eta(\xi_i,\beta) \text{ and } cov(Y_i) \approx V_\beta(\xi_i).$$

This has the drawback that if the inter-individual variation is substantial, then the linearized model may lead to biased imprecise estimation of the fixed parameters. In fact, the linearization around the population parameter vector might misleadingly suggest generating some information, as can be seen in the following example.

4 Example

Assume the observations of an experiment to follow some quadratic model. The measurements in the experimental settings $x \in X$ are considered to be exact:

$$Y_i(x) = \beta_{1,i} + \beta_{2,i}x + \beta_{3,i}x^2 =: f(x)^T\beta_i,\ \varepsilon_i = 0$$

and the individual parameter vector β_i is assumed to be normally distributed with mean vector β and a positive definite covariance matrix D. For simplicity assume that each individual is observed under 3 different experimental settings $x_{ij} \in X$. With these assumptions it follows that

$$Y_i \sim N(F_i\beta, V_i), \text{ where } F_i = (f(x_{i1}), f(x_{i2}), f(x_{i3}))^T \text{ and } V_i = F_iDF_i^T.$$

For the population model with N individuals follows:

$$Y \sim N(F\beta, V), \text{ where } F = (F_1^T, ..., F_N^T)^T \text{ and } V = diag(V_1, ..., V_N).$$

In this model *ML*-estimator and least squares estimator coincide:

$$\hat{\beta} = (F^TV^{-1}F)^{-1}F^TV^{-1}Y \text{ with the covariance } cov(\hat{\beta}) = (F^TV^{-1}F)^{-1} = \frac{1}{N}D.$$

This is obvious, since we obtain for each individual the true parameter vector and with this:

$$\hat{\beta} \sim N(\beta, \frac{1}{N}D).$$

In a next step consider the lognormal model:

$$Y_i(x) = \eta(x, \beta_i) = exp(\beta_{1,i} + \beta_{2,i}x + \beta_{3,i}x^2)$$

with the same assumptions as in the former example. Notice that the regression function is no longer linear in the parameters. For the *ML*-estimate under the implied lognormal model, with the design matrix F_i and variance V_i as before, we obtain

$$\begin{aligned}\hat{\beta}_{ML} &= (\sum_{i=1}^{N} F_i^T V_i^{-1} F_i)^{-1} \sum_{i=1}^{N} F_i^T V_i^{-1} ln(Y_i) \\ &= (ND)^{-1} D \sum_{i=1}^{N} F_i^{-1} ln(Y_i) = \frac{1}{N}\sum_{i=1}^{N} F_i^{-1} F_i \beta_i \\ &= \frac{1}{N}\sum_{i=1}^{N} \beta_i \sim N(\beta, \frac{1}{N}D).\end{aligned}$$

Ignoring the obvious distribution of Y_i and considering the linearization of the model around some vector β_0, we obtain for the linearized model:

$$Y_i(x) \approx \eta(x, \beta_0) + f_{\beta_0}(x)^T(\beta_i - \beta) + f_{\beta_0}(x)^T(\beta - \beta_0)$$

with

$$f_\beta(x) := (\frac{\partial \eta(x,\beta)}{\partial \beta_1}, \frac{\partial \eta(x,\beta)}{\partial \beta_2}, \frac{\partial \eta(x,\beta)}{\partial \beta_3})^T \text{ and } f_{\beta_0}(x) := f_\beta(x)|_{\beta=\beta_0}.$$

As the linearized design matrix $F_{i,\beta_0} = (f_{\beta_0}(x_{i1}), f_{\beta_0}(x_{i2}), f_{\beta_0}(x_{i3}))^T$ is for $\beta_0 \neq 0$ and 3 different experimental settings x_{ij} in X regular, it follows that the individual information is

$$F_{i,\beta_0}^T V_{i,\beta_0}^{-1} F_{i,\beta_0} = F_{i,\beta_0}^T (F_{i,\beta_0} D F_{i,\beta_0}^T)^{-1} F_{i,\beta_0} = D^{-1}.$$

The resulting observation vector $\tilde{Y}_{i,\beta_0} = Y_i - \eta(\xi_i, \beta_0) + F_{i,\beta_0}\beta_0$ and the assumption $\tilde{Y}_{i,\beta_0} \sim N(F_{i,\beta_0}\beta, V_{i,\beta_0})$ yield

$$\begin{aligned}\hat{\beta}_{GLS} &= (\sum_{i=1}^{N} F_{i,\beta_0}^T V_{i,\beta_0}^{-1} F_{i,\beta_0})^{-1} \sum_{i=1}^{N} F_{i,\beta_0}^T V_{i,\beta_0}^{-1} \tilde{Y}_{i,\beta_0} \\ &= (ND)^{-1} D \sum_{i=1}^{N} F_{i,\beta_0}^{-1} \tilde{Y}_{i,\beta_0} \sim N(\beta, \frac{1}{N}D).\end{aligned}$$

The linearized model around the expectation of the individual effects follows as described in the previous section:

$$Y_{i,\beta} \sim N(\eta(\xi_i,\beta), V_{i,\beta}) \text{ with } V_{i,\beta} = F_{i,\beta} D F_{i,\beta}^T.$$

The asymptotic variance of the *ML*-estimator in the heteroscedastic normal model leads according to section 2 to

$$cov(\hat{\beta}) = M_\beta^{-1} = (\sum_{i=1}^{N} F_{i,\beta}^T V_{i,\beta}^{-1} F_{i,\beta} + \frac{1}{2}\tilde{S}_{i,\beta})^{-1} = (ND^{-1} + \frac{1}{2}\sum_{i=1}^{N}\tilde{S}_{i,\beta})^{-1} < \frac{1}{N}D.$$

These differences in the information matrices are generally not negligible. Considering $D = I_3$ and $X = [-1,1]$, the determinant of the D-optimal information in the linearized model is more than 20000 times the determinant of the information of the *ML*-estimator in the lognormal model.

A consequence of these results for the linear mixed model would be, that the information might be improved by simply "nonlinearizing" the model and afterwards applying a Taylor expansion conditional on the unknown population parameter vector β and assuming that this approximation is exact. This would mean that information is generated by systematically misspecifying the model.

5 Discussion

The above example is a simple illustration of one main problem of the derivation of the Fisher Information using the first-order linearization. Misspecifications of the model may lead to wrong approximations of the Fisher Information and with this to wrong optimal designs. For nonlinear mixed effects models, the vector of observations will generally not be normally distributed, to such an extent that the approximation of the Fisher Information by linearization around the population mean is not reliable.

References

Atkinson, A. C. and R. Cook (1995). *d*-optimum designs for heteroscedastic linear models. *Journal of American Statistical Association 90*, 204–212.

Davidian, M. and D. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*, Volume 62 of *Monographs on Statistics and Applied Probability*. Chapman & Hall. London.

Pinheiro, J. and D. Bates (2000). *Mixed-Effects Models in S and S-Plus*. Statistics and Computing. Springer. New York.