journal homepage: www.intl.elsevierhealth.com/journals/cmpb

ODMixed: A tool to obtain optimal designs for heterogeneous longitudinal studies with dropout

*Shirley Ortega**, *Frans E.S. Tan*, *Martijn P.F. Berger*

University of Maastricht, Department of Methodology and Statistics, P.O. Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 5 March 2009
Received in revised form
2 April 2010
Accepted 14 April 2010

Keywords:
D-optimal designs
Dropout
Heterogeneous autocorrelation
Linear mixed models
Longitudinal data
Relative efficiency

ABSTRACT

ODMixed is a computer program to obtain optimal designs for linear mixed models of longitudinal studies. These designs account for heterogeneous correlated errors and for data with dropout. Designs are compared by using relative efficiencies, e.g., between a D-optimal design for homogeneous data and another for heterogeneous data or between a D-optimal design for complete data against another that optimizes designs when data is missing at random. Two examples are worked out to illustrate how researchers could use this computer program to profit of optimal design theory at the planning stage of longitudinal studies.

1. Introduction

At the planning stage of longitudinal studies, the allocation of the resources (time, subjects and/or money) is a critical issue. To collect longitudinal data, it is common practice to use equally spaced designs since these designs are model-free. However, these designs can be suboptimal when compared with their optimal counterparts [1] or when data is missing [2]. Both studies showed that at the planning stage of longitudinal data, researchers may profit, substantially, from optimal design theory.

Optimal design of experiments give the lowest estimators variance such that the estimators have high precision. However, these designs are model-dependent, i.e., the researcher must have prior knowledge about the underlying model fitting the longitudinal data. These data are usually correlated, can have heterogeneous variances and/or can be affected by dropout. The most suitable model to fit data with correlated and unbalanced data structures is the linear mixed models [3].

In this paper, we present a computer program ODMixed that computes optimal designs for complete data, for data suffering from dropout and for data having heterogeneous error structures. ODMixed is made in Matlab. Matlab is chosen because of its flexible plotting capabilities, robust optimization algorithms and steadily growing number of toolboxes, not to mention the fact that it is been steadily introduced in biomedical and health sciences applications.

To our knowledge this is the first program that computes optimal designs for heterogeneous longitudinal data and data missing at random. Notice that the optimal designs with complete data and homogenous error structure can be matched with those obtained for one-cohort using the Program for Optimal design of Longitudinal Studies (POLS) which is an interactive program implemented in Matlab that allows to compute D-optimal designs for different polynomial models

* *Corresponding author*. Tel.: +31 070 3352072.
E-mail addresses: shortegaazurduy@yahoo.com, sota@cbs.nl (S. Ortega).

doi:10.1016/j.cmpb.2010.04.004

with mixed effects for a set of different number of cohorts [4].

This paper is organized as follows. In Section 2, the design problem for the data with dropout and heterogeneous error variance structure will be briefly discussed and the relative efficiency will be introduced as a measure to compare two or more designs. In Section 3, the graphical user interface will be introduced and applied to two motivating examples. Finally, some conclusions and recommendations are provided in Section 4.

2. Longitudinal data and linear mixed models

In longitudinal studies, the conditions for continuous responses $\mathbf{y}_i$ of n subjects must be planned well ahead the collection of the data. A typical model for longitudinal data is the linear mixed models. It is defined as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{1}$$

where $\mathbf{y}_i$ is the $q \times 1$ outcome vector of subject i, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the $q \times p$ and $q \times k$ design matrices for fixed and random parameter vectors $\boldsymbol{\beta}$ and $\mathbf{b}_i$, respectively, and $\mathbf{e}_i$ is the $q \times 1$ error matrix. The indices $i = 1, \ldots, n$ and $j = 1, \ldots, q$ denote each subject participating in the biomedical or social longitudinal study and each time point (date) in the planning scheme of the longitudinal study, respectively.

For the purpose of the program ODMIXED, the model (1) is rewritten as follows

$$\mathbf{y}_i = f(\mathbf{X}_i; \boldsymbol{\beta}) + g(\mathbf{Z}_i|\mathbf{b}_i) + \mathbf{e}_i, \tag{2}$$

where f is a polynomial function representing the time-connected fixed-effects of up to the third degree, i.e., $p = 4$ parameters; g is a polynomial function representing the also time-connected random-effects of up to the first degree, i.e., $k = 2$, and $\mathbf{e}_i$ is the random error.

The function g is conditional on $\mathbf{b}_i$, $\mathbf{b}_i$ represents how the ith subject deviates from the average population and it is assumed to be normally distributed with mean 0 and random covariance matrix $\text{Cov}(\mathbf{b}) = D$. D is a $k \times k$ matrix with elements d_{sl} where $s, l = 1, 2$. These elements denote the variances of the random intercepts, random slopes and their covariances.

The random error of each subject is represented by $\mathbf{e}_i$. This error is assumed to be normally distributed with mean 0 and covariance matrix $\text{Var}(\mathbf{e}_i) = \boldsymbol{\Sigma\Psi\Sigma}$ with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\sigma_1, \ldots, \sigma_q)$ diagonal allowing heterogeneity of the error. The within-subject correlation matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\rho)$ with ρ denoting the autocorrelation between the jth and j'th time points x_j and $x_{j'}$, respectively. Notice that the time points x_j are not necessarily equally spaced and that $\mathbf{b}_i$ and $\mathbf{e}_i$ are independent from each other.

Two cases are distinguished: errors $\mathbf{e}_i$ are homogeneous among different time points x_j, i.e., $\sigma_1 = \sigma_2 = \cdots = \sigma_q = \sigma$ and $\text{Var}(\mathbf{e}_i) = \sigma^2\boldsymbol{\Psi}$ or there is heterogeneity in the error $\text{Var}(\mathbf{e}_i) = \boldsymbol{\Sigma\Psi\Sigma}$, i.e., the diagonal elements of $\boldsymbol{\Sigma}$ are all unequal in value.

Notice that the subindex i have been omitted since the measurements of each subject i are performed at the same time-points.

2.1. *Optimal designs and response probability function*

The exact design ξ for the linear mixed model (1) is defined as:

$$\xi = \begin{Bmatrix} x_1 & x_2 & \ldots & x_j & \ldots & x_q \\ n_1 & n_2 & \ldots & n_j & \ldots & n_q \end{Bmatrix}, \quad \text{where } j = 1, \ldots, q, \tag{3}$$

At the first design point x_1 of the study, the probability of response (i.e., the chance that data from a subject is available at this time point) is $p(x_1) = 1$ and the number of subjects responding at x_1 is equal to n_1. At the jth design point with response probability $p(x_j)$, the expected number of subjects $n_j = n_1 p(x_j)$. Finally, at the last design point x_q, the number of subjects n_q represents the number of subjects completing the experiment. For $j = q$, n_q is the number of subjects responding at all q design points. We will assume that at least one subject is observed at this last point, i.e., $n_q \geq 1$. It is also assumed that $p(x_j)$ is a monotonically decreasing function with $p(x_1) \geq p(x_2) \geq \cdots \geq p(x_q)$ where $p(x_1) > p(x_q)$.

Notice that a response probability $p(x_j)$ at a design point x_j is complementary to a dropout probability $1 - p(x_j)$ at the same design point x_j and that an important assumption considered all along this paper is that the dropout occurs through a noninformative mechanism, i.e., responses are missing at random (MAR), see e.g., [9].

For complete data, the number of subjects responding to each x_j is assumed equal, i.e., $p(x_j) = 1$ for all j. But, if a dropout process arises, the number of subjects responding at x_j will (monotonically) decrease. So, the probability of obtaining data at design point x_j will depend upon each design point. The values of the design time points x_j and the response probability at those points are confined in a given design space $\mathcal{X}$ and probability range $0 < p(x_j) \leq 1$.

The computer program ODMIXED can handle two polynomial response functions, namely, a linear function $p_{lin}(x_j) = a_0 + a_1 x_j$ and a quadratic function $p_{quad}(x_j) = a_0 + a_1 x_j + a_2 x_j^2$. Further details are given in [2].

2.2. *Asymptotic covariance matrix*

We introduce the super-index $[j]$ to group subjects having j measurements with design matrices $\mathbf{X}^{[j]}$ for the fixed-effects and $\mathbf{Z}^{[j]}$ for the random-effects and connected to the response probability $p(x_j)$.

In general, if the number of subjects having all responses to all q design points is denoted by $m_q = n_q$ and the number of subjects with only $j < q$ responses is denoted by the difference $m_j = n_j - n_{j+1}$, then the number of subjects m_j having j responses is given as:

$$m_j = \begin{cases} n_q & \text{if } j = q, \\ n_j - n_{j+1} & \text{if } j < q. \end{cases} \tag{4}$$

The design matrices $\mathbf{X}^{[j]}$ and $\mathbf{Z}^{[j]}(j = 1, \ldots, q)$ have size $j \times p$ and $j \times k$, respectively. The linear mixed model (1) for data with dropout and heterogeneous error implies a marginal model