D-optimal designs for prospective cohort studies

DOI:
10.26481/dis.20081008ft

Summary and conclusion

Generalized Linear Mixed Models (GLMM) have been used quite effectively to model longitudinal data with discrete as well as continuous outcome variable. The choice of a design for GLMM is a very important task in the development and building of an adequate model. In addition to an appropriate method of analysis, a carefully designed study will safeguard the validity of the final conclusion of longitudinal studies. Optimal planning of research designs may also save money and time by specifying the optimal number of cohorts, time points, and optimal allocation of measurement points per subject with limited cost.

The objectives of this thesis are:

- to find the optimal number of cohorts,
- to identify the optimal number of repeated measurements per subject, and
- to determine the optimal allocations of time points within a study period for longitudinal studies with the constraint of the study cost (budget).

Chapter 1 is introductory. The GLMM and its two specific models namely the linear and logistic mixed models for continuous and dichotomous outcome variables respectively are described in this chapter. Designs for linear models are studied in chapters 2, 3 and 4 while designs for logistic models are studied in chapters 5 and 6. The D-optimality criterion is used in this thesis and is discussed in chapter 1. This criterion is popular because of its interpretation in terms of the volume of a confidence ellipsoid of the parameter estimators and its nice property that the D-optimal designs do not depend on the chosen time interval. The cost parameters encountered in longitudinal studies are introduced and discussed in the first chapter and implemented in optimization of designs in chapters 2, 3 and 4. The relative efficiency that helps to compare different designs is also discussed in the first chapter.

In chapter 2, we studied and identified optimal number of time points and allocations for longitudinal studies with fixed costs. Linear mixed models with polynomials of degree 1, 2 and 3 are considered. A numerical study shows that the optimal design results for a linear mixed model with cost constraints are almost similar to the results in literature for fixed effects linear models. It is found that the optimal number of time points is equal to the number of regression parameters in the mixed model except when the ratio of costs of recruiting a subject to the costs of additional measurement per subject is larger than about 20 and the autocorrelation coefficient is close to 0. The optimal allocations are almost at equi-distance time points except when the autocorrelation coefficient between successive measurements within a subject is less than 0.1. In most practical cases, correlations between successive measurements per subject in longitudinal studies are expected to be greater or equal to 0.1. Hence, in most practical cases, the optimal number of time points per subject is equal to the number of regression parameters and the optimal allocations are approximately equi-distant. From the numerical study, the main factors that affect the optimal design are the number of regression parameters in the model, the ratio between the cost of recruiting a new subject and the cost of an additional observation per subject, and the autocorrelation coefficient between repeated measurements per subject.

Depending on the number of cohorts (groups) over time, the design for longitudinal studies may be a *purely longitudinal* design, with only one cohort measured over the complete time interval, a *mixed longitudinal* design, with several cohorts measured over a shorter time interval for each cohort and a *cross-sectional* design, where groups of subjects are measured at a single time point for each group (Rao and Rao, 1966; Woolson et al., 1978). The purely longitudinal cohort design is studied in Chapter 2.

In chapter 3, the extension to other cohort designs is made by considering more than one cohort of subjects over time, i.e. *purely longitudinal*, *mixed longitudinal* and *cross-sectional* cohort designs are considered in chapter 3. The general finding shows that the optimal number of time points is a function of the number of cohorts and the number of regression parameters in the model. It is found that the optimal number of time points is the sum of the number of cohorts and the number of regression parameters in the model minus 1. Comparison of the three cohort designs (*Purely longitudinal*, *mixed longitudinal* and *cross-sectional* cohort designs) shows that the purely longitudinal cohort design with one cohort of subjects over the study period is the most efficient design. Thus, unless there is some biological or practical reason to consider, it is recommended to follow one group of subjects over the study period and take their measurements at the optimal allocation points which in most cases are approximately the same as the equi-distance time points.

In chapter 4, an interactive user-friendly computer program is introduced and described. This computer program is written in matlab to implement the basic ideas discussed in chapters 2 and 3. The program helps practitioners to obtain optimal cohort design with an optimal number of repeated measurements and optimal allocations of time points for longitudinal studies without going in to the details of optimal design theory. It also helps to computes the loss in efficiency of any alternative design compared to the optimal design.

In chapter 5, designs for fixed effects logistic models with categorical independent variables are considered. It is known that designs for non linear models including logistic models depend on the model parameters that are not known before data collection. We proposed the use of what we call r-design in chapter 5 to go around this problem. The r-design assigns an equal weight to design points that give more information in estimating the regression parameters and ignores those design points which contain less information. This design is D-optimal on a restricted design region. The r- design can be constructed in practice by using the approximate rank order of the response variances at the design points. The design points are finite when all the independent variables are categorical. A numerical study shows that the r-designs are highly efficient in most cases as compared to the usual balanced designs that assign equal weight to all design points.

In chapter 6, designs for logistic mixed effects models for binary longitudinal data are considered. Locally D-optimal designs are studied and they are given as a function of the regression parameters. Maximin D-optimal designs are considered to solve the problem of parameter dependency of designs for logistic mixed models. Maximin D-optimal designs at a given interval of parameter values are computed by means of a maximin relative efficiency measure. The performance of the maximin D-optimal designs in terms of the maximin efficiency is high for a range of parameter values that is common in practice. Further, maximin designs with different numbers of time points are compared. The results show that the optimal number of repeated measurements depends on the number of regression parameters in the model. The design locations generally shift to the left as compared to the design locations for linear mixed effects models known in the literature. A user-friendly computer program that helps to obtain locally and maximin D-optimal designs for logistic mixed models is also described in chapter 6.