Improving parameter estimates in generalized linear mixed models

Summary

The aim of this thesis is to find methods to improve the estimation process of the regression parameters for Generalized Linear Mixed Models. The proposed methods concern the optimization of the choice of the design and the detection of influential data. The methods are discussed for the Linear Mixed Model and the Generalized Linear Mixed Model. The optimality criteria are discussed in chapter 1, 2, 3 and 6, while the influence measures are discussed in chapter 1, 4, 5 and 6.

Chapter 1 starts with the introduction of the Linear Mixed Model and the Generalized Linear Mixed Model. Thereafter, the most frequently used optimality criteria are discussed. One of these criteria is the D-optimality criterion, which is used to find a design for which the volume of the confidence ellipsoid is minimized. The preference for the D-optimality criterion is based on its invariance with respect to the coding of the independent variables.

Which design is D-optimal depends on the values of the model parameters in the Generalized Linear Mixed Model. Because these values are unknown in practice, the sequential design approach, the Bayesian design approach and the maximin approach can be used. In literature, the maximin approach is used to find a design for which the smallest efficiency over the model parameter space is maximized. The choice to let the maximin approach be based on the relative efficiency is explained in this chapter.

The chapter proceeds with the introduction of Cook's Distance and local influence measures to detect influential data. Cook's Distance assesses the change in the estimates, due to case deletion. The local influence measures assess the change in the estimates, due to infinitesimally perturbations of the data set. It is proposed to detect the influential cases by comparing the scores of the cases on the influence measures with each other. The importance of the detection of influential observations in addition to the detection of influential subjects is illustrated.

Chapter 1 ends by explaining the relationship between the D-optimality criterion and the expected/potential influence of observations for the linear regression model with

uncorrelated measurement errors. It is shown that each design point of the D-optimal design has the same expected/potential influence and that there exists no design point with larger expected/potential influence.

Chapter 2 starts with the presentation of D-optimal designs for the first and second degree polynomial model with random intercept, random slope and AR(1) correlated measurement errors. It shows that the solution of the D-optimality criterion depends on the unknown values of the model parameters. As a consequence, the chapter proceeds with the search of designs which are highly efficient for all likely values of the model parameters. Large classes of highly efficient symmetric designs are presented. The evaluation of the relative efficiency leads to the maximin criterion based on the *relative efficiency*.

The optimality criteria discussed in literature assume that *all* independent variables are experimentally controlled. This assumption is often violated in practice. For example, consider the project 'teeth-brushing on the elementary school', which is discussed in more detail in chapter 6. For this project, only the marginal proportion treated schools is experimentally controlled, where the marginals are taken over the joint distribution of the baseline value, the number of pupils within the schools and the social-economic status of the pupils. Because only the marginal proportion treated schools can be experimentally controlled, the existing optimality criteria are inadequate for this situation.

To deal with the previously described situation, the sample design is factorized in chapter 3 in a part which is experimentally controlled and a part which is not. The experimentally controlled part is called the experimental design and corresponds in this chapter with a marginal distribution (cf. marginal proportion treated schools). The other part of the factorization corresponds with a conditional distribution (cf. conditional joint distribution of the baseline value, the number of pupils within the schools and the social-economic status of the pupils). The experimental design should be optimized, taking this unknown conditional distribution into account. Because the conditional distribution is unknown, a maximin criterion is proposed which searches for an experimental design that maximizes the smallest relative efficiency over all likely values of the model

parameters and all likely conditional distributions. It is shown that under certain conditions, the maximin experimental design is balanced in the discrete variables. The proposed maximin criterion is illustrated using a Logistic Mixed Model.

The second method to improve the estimation process is the detection of influential subjects and observations. For the Linear Mixed Model, Cook's Distance is defined marginally over the random effects. Unfortunately, this measure may fail to detect or may incorrectly detect influential observations, due to the random effect variances and covariances. In chapter 4, Cook's Distance is defined conditionally on the subjects in the sample. The conditionally defined Cook's Distance can be partitioned into two influence measures, one measuring the effect on the estimated average profile and one measuring the effect on the estimated subject-specific deviations. The conditionally defined Cook's Distance works better than the marginally defined Cook's Distance.

In chapter 5, we considered likelihood-based influence measures. To accelerate the detection of influential data, these measures are approximated by Taylor expansions, which are called local influence measures. In literature, the local influence measures are defined at subject-level and are only discussed for the Linear Mixed Model. We extended these measures not only to the Generalized Linear Mixed Model, but also to the observation-level. It is shown that the subject-oriented influence measure is a special case of the proposed observation-oriented influence measure. An illustration of a two-treatment multiple period cross-over trial demonstrates the practical importance of the detection of influential observations in addition to the detection of influential subjects.

In chapter 6, the optimal design criterion which is proposed in Chapter 3 and the local influence measures which are proposed in chapter 5 are illustrated based on the project 'teeth-brushing on the elementary school'. It appeared that the balanced design is highly efficient, irrespective of the uncontrolled conditional distribution and model parameter vector at hand. The detection of influential data leaded to an assessment of the stability of the parameter estimates, a change in the evaluated model and the suggestion to sample more schools for the validation of a random effect for the social-economic status.

Dutch Summary

In dit proefschrift worden methoden voorgesteld, die tot kwalitatief betere schatters van de regressieparameters leiden. De methoden zijn gericht op de optimalisatie van de keuze van het design en de detectie van invloedrijke data en worden besproken voor het lineaire mixed effect model en het gegeneraliseerde lineaire mixed effect model. De optimalisatiecriteria zijn behandeld in hoofdstuk 1, 2, 3 en 6. De invloedsmaten zijn terug te vinden in hoofdstuk 1, 4, 5 en 6. Hieronder volgt een samenvatting van elk hoofdstuk.

Hoofdstuk 1 begint met de introductie van het lineaire mixed effect model en het gegeneraliseerde lineaire mixed effect model. Het hoofdstuk vervolgt met de bespreking van de meest gebruikte optimalisatiecriteria. Een van deze criteria is het D-optimalisatiecriterium. Dit criterium wordt gebruikt voor het bepalen van een design dat het volume van de betrouwbaarheidsellipsoïde minimaliseert. De voorkeur voor dit criterium is gebaseerd op zijn invariantie met betrekking tot de codering van de onafhankelijke variabelen.

Welk design D-optimaal is hangt af van de waarden van de modelparameters. Helaas zijn deze waarden in de praktijk onbekend. Om toch tot een goede keuze van het design te kunnen komen, zijn de sequentiële design benadering, de Bayesiaanse design benadering en de maximin benadering geïntroduceerd. Tot nu toe is de maximin benadering gebruikt voor het bepalen van een design dat de kleinste efficiëntie over de modelparameter ruimte maximaliseert. De keuze om de maximin benadering te baseren op de relatieve efficiëntie is in hoofdstuk 1 gemotiveerd.

Het hoofdstuk gaat verder met de introductie van Cook's Distance en lokale invloedsmaten voor het detecteren van invloedrijke data. Cook's Distance meet de verandering in de regressieparameters door het wegnemen van data. De lokale invloedsmaten meten de verandering in de schatters door infinitesimale verstoringen van de data set. De invloedrijke cases worden gedetecteerd door het onderling vergelijken van de scores van de cases op de invloedsmaten. De toegevoegde waarde van de detectie

van invloedrijke observaties is aangetoond.

Hoofdstuk 1 eindigt met de relatie tussen het D-optimalisatiecriterium en de verwachte/potentiële invloed van observaties voor het lineaire regressie model met ongecorreleerde meetfouten. Het blijkt dat ieder design punt van het D-optimaal design dezelfde verwachte/potentiële invloed heeft en dat geen enkel ander design punt een nog grotere verwachte/potentiële invloed heeft.

Hoofdstuk 2 start met de presentatie van D-optimale designs voor de eerstegraads en tweedegraads polynomen met stochastisch intercept, stochastische helling en AR(1) gecorreleerde meetfouten. Uit deze presentatie blijkt dat de oplossing van het D-optimalisatiecriterium afhangt van de (onbekende) waarden van de modelparameters. Als een gevolg hiervan wordt overgegaan tot het zoeken van designs die voor alle aannemelijke waarden van de modelparameters hoog efficiënt zijn. De evaluatie van de relatieve efficiëntie over de gehele modelparameter ruimte leidt tot de definitie van de maximin benadering op basis van de relatieve efficiëntie.

De in de literatuur behandelde optimalisatiecriteria veronderstellen dat *alle* onafhankelijke variabelen experimenteel gecontroleerd worden. Deze veronderstelling wordt in de praktijk vaak geschonden. Neem bijvoorbeeld het project 'tandenpoetsen op de basisschool', dat uitgebreid behandeld wordt in hoofdstuk 6. In dit project wordt alleen de marginale proportie behandelde scholen experimenteel gecontroleerd, waar de marginalen genomen worden over de verdeling van de beginmeting, het aantal leerlingen per school en de sociaal-economische status van de leerlingen. Omdat alleen de marginale proportie behandelde scholen experimenteel gecontroleerd wordt, zijn de bestaande optimalisatiecriteria voor deze situatie ontoereikend.

Een maximin criterium voor de hiervoor beschreven situatie wordt in hoofdstuk 3 afgeleid door het ontbinden van het steekproef design in dat deel dat experimenteel gecontroleerd wordt en dat deel dat niet experimenteel gecontroleerd wordt. Het experimenteel gecontroleerde gedeelte wordt experimenteel design genoemd en correspondeert in dit hoofdstuk met een marginale verdeling (marginale proportie behandelde scholen). Het andere deel correspondeert met een conditionele verdeling

(conditionele verdeling van de basismeting, het aantal leerlingen per school en de sociaal-economische status van de leerlingen). De optimalisatie van het experimenteel gecontroleerde deel is nu van belang, terwijl er rekening moet worden gehouden met het ongecontroleerde deel. Omdat de conditionele verdeling onbekend is, wordt een maximin benadering voorgesteld, die zoekt naar een experimenteel design waarvoor de kleinste relatieve efficiëntie over alle aannemelijke waarden voor de modelparameters en alle aannemelijke conditionele verdelingen het grootst is. Onder bepaalde omstandigheden blijkt het maximin experimentele design gebalanceerd te zijn in de discrete variabelen. De voorgestelde maximin benadering wordt aan de hand van een logistisch mixed model geïllustreerd.

In hoofdstuk 4 worden twee generalisaties van Cook's Distance voor het lineaire mixed model besproken. Cook's Distance is in het verleden naar het lineaire mixed effect model uitgebreid. Dit werd gedaan via marginalisatie over de stochastische effecten. Helaas detecteert de zo verkregen invloedsmaat de invloedrijke observaties niet altijd en worden soms zelfs de verkeerde observaties gedetecteerd. Dit is voornamelijk toe te wijzen aan de varianties en covarianties van de stochastische effecten. In hoofdstuk 4 wordt Cook's Distance daarom geconditioneerd op de subjecten. De conditioneel gedefinieerde Cook's Distance kan worden ontbonden in twee invloedsmaten, waarbij de ene het effect op het geschatte algemene profiel meet en de andere het effect op de geschatte subject-specifieke afwijkingen. De conditioneel gedefinieerde Cook's Distance werkt beter dan de marginaal gedefinieerde Cook's Distance.

In hoofdstuk 5 behandelen we likelihood-gebaseerde invloedsmaten. Om de detectie van invloedrijke data te versnellen worden de invloedsmaten benaderd door Taylor expansies. Deze Taylor expansies worden lokale invloedsmaten genoemd. In de literatuur worden de lokale invloedsmaten op subject-niveau gedefinieerd en worden zij alleen besproken voor het lineaire mixed effect model. In hoofdstuk 5 worden deze maten uitgebreid tot de gegeneraliseerde lineaire mixed effect modellen en tot het observatie-niveau. De subject-georiënteerde invloedsmaten blijken speciale gevallen van de observatie-georiënteerde invloedsmaten te zijn. De toegevoegde waarde van de

observatie-georiënteerde invloedsmaten wordt aan de hand van een twee-behandelingen meervoudige periode kruisproef studie aangetoond.

In hoofdstuk 6 worden het optimalisatiecriterium van hoofdstuk 3 en de lokale invloedsmaten van hoofdstuk 5 aan de hand van het project 'tandenpoetsen op de basisschool' geïllustreerd. Het blijkt dat het gebalanceerde design hoog efficiënt is, onafhankelijk van de conditionele verdeling en de waarden van de modelparameters. De detectie van invloedrijke data leidt tot een beoordeling van de stabiliteit van de parameterschatters, een verandering in het geëvalueerde model en de suggestie om meer scholen op te nemen in de steekproef voor de validatie van een stochastische helling voor de sociaal-economische status.