Optimal Designs For Mixed Effects Poisson Regression Models

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)

von Mehrdad Niaparast
geb. am 06.07.1971 in Behbahan, Iran

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Rainer Schwabe
Dr. Dave Woods

eingereicht am: 30.10.2009
Verteidigung am: 29.01.2010

Acknowledgement

This PhD thesis would not have been possible without the help of many people, whom I would like to thank.

First of all I would like to thank and acknowledge my supervisor Prof. Dr. Rainer Schwabe for suggesting the subject of this study, his help and guidance. His guidance and encouragement were greatly appreciated. I thank him for the many hours of discussion and advice which helped me immensely in completing this work.

I would like to express my gratitude to my wife, Leila Behbood, for her encouragement and support throughout the study period and also my daughters, Sheyda and Dorsa, for their love and encouraging priceless smile. I deeply acknowledge all my colleagues in our Institute for creating a friendly working environment; I had unforgettable time with them. I am also grateful to Kerstin Altenkirch for her kind help.

I am much obliged to the Ministry of Science, Research & Technology, Iran for the financial support of my PhD program.

Summary

Optimal experimental designs for models with random effects have received increasing attention in recent years. Binary data models, especially logistic, form the main part of the presented research.
The main goal of this thesis is to develop optimal experimental designs for the Poisson regression models with random intercept and random slope.
An introduction will be presented about fundamental concepts including linear models, generalized linear models, linear mixed models and generalized linear mixed models. Complications in the design process arise with the use of random effects, i.e. when some model parameters are allowed to vary randomly between subjects. In fact the Fisher information matrix can not be written down in closed form for generalized linear mixed models due to the random effects. Therefore we apply a different estimating method to derive an approximating information matrix. This method is called the quasi-likelihood method and the information matrix based on this method is the quasi-information matrix. Some properties of the quasi-score function are studied as a special case of the estimating function.
A simulated example shows that the quasi-likelihood estimations are close to the MLE of the unknown model parameters, especially when the variance of random effects is small.
Using the quasi-likelihood method, the quasi-information matrices are obtained for different Poisson models.
Convex design theory for ordinary linear models could not be extended to the proposed models due to the fact that the quasi-information matrices are not additive because of the existence of random effects in the models. We obtain some new theorems that allow us to apply convex design theory to our models. Besides this, equivalence theorems, similar to the ones known for ordinary linear models, are derived for our situations.
The best experimental settings to do an experiment are usually selected via a real-valued function of the respective information matrix. In this work, we derive different representation of these functions based on the quasi-information matrices.
Some examples from the models are presented to illustrate proposes. This thesis is closed with a discussion of future work.

Zusammenfassung

Die Bestimmung optimaler Versuchspläne für Modelle mit zufälligen Effekten erfreut sich in den letzten Jahren wachsenden Aufmerksamkeit in der Literatur. Modelle mit binären Daten, speziell logistischer Form, bilden den Hauptteil dieser Arbeiten .
Das Ziel der vorliegenden Arbeit ist die Herleitung optimaler Versuchspläne für das Poisson-Regressions-Modell mit zufälligem Achsenabschnitt bzw. mit zufälliger Steigung. Es wird eine Einführung in grundlegende Konzepte gegeben, die lineare Modelle, verallgemeinerte lineare Modelle, lineare gemischte Modelle und verallgemeinerte lineare gemischte Modelle umschließen. Die Einführung zufälliger Effekte zur Modellierung individueller Parameter verkompliziert die Bestimmung optimaler Designs. Für verallgemeinerte lineare gemischte Modelle lässt sich auf Grund der zufälligen Effekte keine geschlossene Form der zugehörige Fisher-Information herleiten. Deswegen wenden wir eine andere Schätzmethode an und approximieren die zugehörige Informationsmatrix. Diese Methode wird Quasi-Likelihood-Methode genannt, und die aus dieser Methode resultierende Informationsmatrix wird als Quasi-Informationsmatrix bezeichnet. Einige Eigenschaften der Quasi-Score-Funktion als Spezialfall der Schätzfunktion werden hier untersucht.
Ein simuliertes Beispiel zeigt, dass sich Quasi-Likelihood- und Maximum-Likelihood-Schätzungen der unbekannten Parameter nicht stark unterscheiden, speziell wenn die Varianz der zufälligen Effekte klein ist. Mit der Quasi-Likelihood-Methode können die Quasi-Informationsmatrizen für verschiedene Poisson-Modelle hergeleitet werden.
Bisher konnte die konvexe Design-Theorie für gewöhnliche lineare Modelle nicht auf die vorgestellten Modelle erweitert werden, da die Quasi-Informationsmatrizen auf Grund des Vorliegens der zufälligen Effekte nicht additiv sind. Wir können jedochneue Theoreme herleiten, die uns erlauben, die konvexe Design-Theorie auf unsere Modelle anzuwenden. Des Weiteren werden Äquivalenz-Theoreme für die betrachteten Modelle bewiesen.
Die optimalen Versuchspläne werden üblicherweise mit Hilfe einer reellwertigen Funktion der betreffenden Informationsmatrix bestimmt. In dieser Arbeit leiten wir eine auf der Quasi-Informationsmatrix basierte Form dieser Kriterien her.
Einige Beispiele der Modelle werden vorgestellt, um das Vorhaben zu illustrieren. Die Arbeit schließt mit einer Diskussion über mögliche zukünftige Entwicklungen auf dem bearbeiteten Gebiet.

Contents

List of Figures

List of Tables

1 Introduction

Optimal design of experiments goes back at least as far as Smith (1918), who defined the objective of minimizing the worst-case prediction error on the construction of univariate polynomial models up to the sixth degree. In that paper, the G-optimality criterion (a term which was introduced by Kiefer and Wolfowitz (1959)) was first used. Kiefer and Wolfowitz (1959) gave the name D-optimality to the criterion introduced by Wald (1943), which is based on the parameters. Kiefer and Wolfowitz (1960) also related D-optimality and G-optimality. In fact, the contributions in this area are relatively sparse until the fifties and sixties, before Kiefer and Wolfowitz (1959) published a paper on this topic. The history behind optimal designs can be found in Atkinson and Bailey (2001) and Kiefer et al. (1985).

A review of literature in the optimal design of experiments (Fedorov (1972), Silvey (1980), Pukelsheim (1993)) indicates that the optimality for linear models was the main part of the research on this topic, while generalized linear and non-linear models are often more applicable to real data. Certain complication arise due to the fact that the information matrix based on the likelihood function depends on the parameter values, hence we encounter to a dual problem: parameter estimating needs to design an experiment and experimental design needs the parameter values. Box and Lucas (1959) investigated locally D-optimal design for non-linear models. They suggested using of some initial guess of the values of the parameters.

Despite considerable work on optimal design for binary data models, especially for the logistic regression model (Myers et al. (1994), Sitter (1992)), there are few researches on the optimal design for the Poisson regression model. Recently, Wang et al. (2006) and Russell et al. (2009) have done an extensive work on the Poisson regression model.

Linear and non-linear mixed models and generalized linear mixed models (McCulloch and Searle (2001)) consider random effects beside fixed effects. Despite wide theoretical work on the analysis of these models, very little research has focused on this topic in optimal experimental designs.

Mentré et al. (1997) propose the linearization approach to non-linear mixed model. Gladitz and Pilz (1982) considered a Bayesian framework for individual prediction in the random coefficient regression models. Fedorov and Hackl (1997) Liski et al. (2002) and Schmelter (2006) gave some results for linear mixed model.

A common characteristic in the above literature on optimal design is to apply the likelihood method for estimating the parameter. Due to the random effects in generalized linear mixed models, we can not obtain a closed form for the likelihood function and hence the information matrix for the parameters. Waterhouse (2005) has done a numerically study on optimal designs for the generalized linear mixed models and non-linear mixed

models.
The quasi-likelihood function is as an alternative approach to the likelihood function. To define a quasi-likelihood function we need only to specify a relation between the mean and the variance. Using this method, in present work, we address a new approach to optimal designs.
This thesis is organized of follows.
An introduction to linear models, generalized linear models, linear mixed models and generalized linear mixed models is presented in chapter 2. Furthermore the differences between these models are discussed. In chapter 3, the quasi-score function as a special case of estimating function and the quasi-likelihood estimator are described. We also obtained many properties of quasi-score function and quasi-likelihood function. In chapter 4, we define the considered models under the names Poisson regression model with random intercept and random slope as special cases of generalized linear mixed models. The relations between mean and variance in these models are obtained in this chapter. We obtain quasi-information matrix for these models. The main parts of our work are in the last two chapters where, in chapter 5, we obtain new statements to develop convex design theory to the considered models. In chapter 6, we apply the theoretical parts of our study to find some locally D-optimal designs for our models. We also do some curiosities in optimal design for our models in this chapter. We will close this work with a short discussion and some suggestions for future studies in the last chapter, chapter 7.

2 Generalized Linear Mixed Models: A review

2.1 Introduction

Generalized Linear Mixed Models (GLMMs) are a useful extension of Linear Mixed Models (LMMs) and Generalized Linear Models (GLMs) for assessing additional components of variability due to latent random effects. In other words GLMMs are an extension of GLMs by the inclusion of random effects. The essence of this extension is two-fold: First, data are not necessarily assumed to be normally distributed, and second, that the mean is not necessarily taken as a linear combination of parameters but that some function of the mean is. These models provide a general framework which includes a wide range of models, in addition to Linear Models (LMs), GLMs and LMMs, like Poisson regression models with random effects and logistic model with random effects. Because of this generality and because of the availability computer capacity, the use of such models has increased dramatically in the last two decades. McCulloch and Searle (2001) have provided a comprehensive summary of the development that took place in the last century. Also more details can be found in Jiang (2007) on these topics.
The maximum likelihood estimator (MLE) is used to make inference about the unknown parameters. Obtaining MLEs involves tremendous analytical and computational difficulties due to integrated likelihoods. There are two different approaches for finding MLEs. The first approach emphasizes on the numerical techniques to find a solution for ML equations (e.g. Pinheiro J.C. and Bates D.M. (1995), Booth and Hobert (1999)). The second approach which is based on alternative method for MLEs, include Quasi-Likelihood (McCullagh and Nelder (1998)), Penalized Quasi-Likelihood (PQL) (Breslow and Glayton (1993)), Generalized Estimating Equations (GEE) (Liang and zeger (1986)), Conditional second-order Generalized Estimating Equation (CGEE2) (Vonesh et al. (2002)), among others. When the number of observations is large, most of the alternatives of the MLEs work well (see e.g. Sinha (2004) and Nie (2007)).
The aim of this chapter is to provide some basic information and fundamental concepts that we need to use of GLMs and GLMMs.
This chapter is organized as follow: In the next two sections, we recall GLMs and LMMs briefly. In section 4 we describe GLMMs and a method to estimate the unknown parameters.

2.2 Generalized Linear Models

One of the flexible tools for statistical inference is Generalized Linear Models which is formulated by Nelder and Wedderburn (1972) to unify various statistical models, including linear regression, logistic regression, probit regression and Poisson regression, under one framework. This unification helps us to estimate the parameters of models under the same algorithm.
The linear Model (LM) for a response Y has the following form

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon \tag{2.1}$$

where $\mathbf{x}$ is a vector of known explanatory variable, $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of fixed effects and uncorrelated ϵ is the error term,which follows a Normal distribution with mean μ and variance σ^2. Note that we can replace $\mathbf{x}^T$ by $\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))$, then

$$Y = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + \epsilon \tag{2.2}$$

where $f_i(\mathbf{x})$ can be an arbitrary function of $\mathbf{x}$.
In general, we can write this model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad E(\boldsymbol{\epsilon}) = \mathbf{0}\,\text{and}\,Var(\boldsymbol{\epsilon}) = \sigma^2 I \tag{2.3}$$

or, corresponding to (2.2)

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.4}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is the data vector which summarizes the whole observations. $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ (or $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1) \ldots \mathbf{f}(\mathbf{x}_n))^T$) is known as the design matrix.
If $(\mathbf{F}^T\mathbf{F})$ is regular then the maximum likelihood estimator, say $\hat{\boldsymbol{\beta}}$, for the parameter vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{Y} \tag{2.5}$$

which coincides with the best linear unbiased estimator(BLUE) of the parameter vector $\boldsymbol{\beta}$.
The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is then

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{F}^T\mathbf{F})^{-1} \tag{2.6}$$

that is free of $Y_i s$ and parameters.
Note that if $\mathbf{F}$ is rank-deficient, either because of intrinsic aliasing among factors or for some other reasons, then $(\mathbf{F}^T\mathbf{F})$ is singular. In this case, there is no unique estimator for $\boldsymbol{\beta}$. But if $\gamma(\boldsymbol{\beta}) = L_\gamma\boldsymbol{\beta}$ is any estimable among $\boldsymbol{\beta}$s, i.e. $\gamma(\boldsymbol{\beta}) = L_\gamma\boldsymbol{\beta}$ is identifiable, then the best linear unbiased estimator of $\gamma(\boldsymbol{\beta})$ is

$$\widehat{\gamma(\boldsymbol{\beta})} = L_\gamma(\mathbf{F}^T\mathbf{F})^{-}\mathbf{F}^T\mathbf{Y} \tag{2.7}$$

where $(\mathbf{F}^T\mathbf{F})^-$ is a g-inverse for $(\mathbf{F}^T\mathbf{F})$.(see e.g. Schott 1997 sec.5.8). The variance-covariance matrix of $\widehat{\gamma(\boldsymbol{\beta})}$ is then

$$Var(\widehat{\gamma(\boldsymbol{\beta})}) = \sigma^2 L_\gamma (\mathbf{F}^T\mathbf{F})^- L_\gamma^T \tag{2.8}$$

which also does not depend on $\boldsymbol{\beta}$.

GLMs seek to extend the domain of applicability of LM by relaxing of the normal assumption or, more generally, of the assumption of additive error, i.e. we denote $f_{\mathbf{Y}}(\mathbf{y}) = f_{\boldsymbol{\epsilon}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ instead of $f_{\mathbf{Y}}(\mathbf{y}) = f(\mathbf{y}; \mathbf{X}\boldsymbol{\beta})$.
If $Y_1, \ldots, Y_n$ be the random sample of Y the two following statements are sufficient to define GLMs:

1. **Random component**: Y_i is a member of an exponential family (Jørgensen 1987) of the form

$$f(y_i; \gamma_i, \phi) = \exp\left\{\frac{y_i\gamma_i - b(\gamma_i)}{a(\phi)} + c(y_i; \phi)\right\} \tag{2.9}$$

   where $a(.)$, $b(.)$ and $c(.;.)$ are known functions. γ_i is the canonical parameter of the distribution of Y_i and ϕ is a nuisance parameter which may be known or not.

2. **Systematic component**: The expectation of Y_i, denoted as μ_i is related to $\mathbf{x}_i$ through a known monotone function h, i.e.,

$$h(\mu_i(\boldsymbol{\beta})) = \mathbf{x}_i^T\boldsymbol{\beta} \quad \Leftrightarrow \quad \mu_i(\boldsymbol{\beta}) = h^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}) \tag{2.10}$$

   Since h links together the mean of Y_i and linear form of predictors, h is called the link function. Also γ_i would be some known function of μ_i, i.e. $\gamma_i = \gamma(\mu_i)$. It is clear that γ_i in (2.9) is a function of $\boldsymbol{\beta}$ through $\mu_i(\boldsymbol{\beta})$. For simplicity we suppress the arguments of $\gamma(\mu_i(\boldsymbol{\beta}))$ and $\mu_i(\boldsymbol{\beta})$ except for the cases which would lead to ambiguities.

For the sake of clarity, we list some special cases.

- **LM**. The corresponding link function is $h(\mu) = \mu$, the "identity" link function, $\phi = \sigma^2$, $a(\phi) = \phi$ and $\gamma_i = \mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$.
- **Logistic Model**. This corresponds to the case that Y_i's have Bernoulli distribution with canonical parameter $\gamma_i = \log(\frac{\mu_i}{1-\mu_i})$. $h(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$ is the "logit" link and $a(\phi) = 1$, i.e. ϕ can be neglected.
- **Poisson Model**. The distribution of the response Y_i is Poisson with mean parameter μ_i. $h(\mu_i) = \log\mu$ is the "log" link and $a(\phi) = 1$

Remark 2.1. *When the link function $h(\mu)$ is the same as the canonical parameter $\gamma(\mu)$, we call $h(\mu)$ the canonical link function.*

2.2.1 Some Properties

With regard to the exponential distribution family the log-likelihood function for (2.9) is

$$l = l(\boldsymbol{\gamma}, \phi; \mathbf{y}) = \sum_{i=1}^{n} (y_i \gamma_i - b(\gamma_i))/a(\phi) + \sum_{i=1}^{n} c(y_i; \phi) \tag{2.11}$$

The following elementary properties of log-likelihood function, under suitable regularity conditions which are fulfilled for (2.11)

$$E(\frac{\partial l(\boldsymbol{\gamma}, \phi; \mathbf{Y})}{\partial \gamma_i}) = 0 \tag{2.12}$$

$$Var(\frac{\partial l(\boldsymbol{\gamma}, \phi; \mathbf{Y})}{\partial \gamma_i}) = -E(\frac{\partial^2 l}{\partial \gamma_i^2}) \tag{2.13}$$

lead us to the basic concepts in GLMs ,

$$\frac{\partial l(\boldsymbol{\gamma}, \phi; \mathbf{y})}{\partial \gamma_i} = (y_i - b'(\gamma_i))/a(\phi) \Rightarrow E(Y_i) = \mu_i = b'(\gamma_i) \tag{2.14}$$

$$\begin{aligned} Var(\frac{\partial l(\boldsymbol{\gamma}, \phi; \mathbf{Y})}{\partial \gamma_i}) &= Var((Y_i - b'(\gamma_i))/a(\phi)) = a^{-2}(\phi) Var(Y_i) \\ \text{and } -E(\frac{\partial^2 l(\boldsymbol{\gamma}, \phi; \mathbf{Y})}{\partial \gamma_i^2}) &= a^{-1}(\phi) b''(\gamma_i) \end{aligned}$$

$$\Rightarrow Var(Y_i) = a(\phi) b''(\gamma_i) = a(\phi) v(\mu_i) \tag{2.15}$$

where $b'(.)$ and $b''(.)$ indicate the first and the second derivative of $b(.)$.The last expression for the variance of Y_i describes the rational behind the phrase "dispersion parameter" for ϕ. Since $v(\mu_i) = b''(\gamma_i) = b''(\gamma(\mu_i))$ indicates how the variance of Y_i is related to the mean of Y_i, and it is called the variance function.

Remark 2.2. *Since in all model that we apply here $a(\phi) = 1$, we ignore this in the remainder of this chapter.*

Remark 2.3. *Regarding to the relationship between γ_i and $\boldsymbol{\beta}$ through the relations* (2.10), (2.14) *and* (2.15), *we can represent $f(y_i; \gamma_i)$ in* (2.9) *as $f(y_i; \boldsymbol{\beta})$.*

In addition to above equations, a straightforward algebra leads to

$$\frac{\partial \gamma_i}{\partial \mu_i} = \frac{1}{v(\mu_i)} \quad \text{and} \quad \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{1}{h'(\mu_i)} \cdot \mathbf{x}_i \tag{2.16}$$

And if we have the canonical link function

$$h'(\mu_i) = \frac{1}{v(\mu_i)} \text{ and hence } \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} = \mathbf{x}_i \tag{2.17}$$

Thus, in the general case, the maximum likelihood equations for $\boldsymbol{\beta}$ are given as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{v(\mu_i)h'(\mu_i)}\mathbf{x}_i \tag{2.18}$$

We can represent the above equation in matrix notation as follow

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \tag{2.19}$$

where $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta})$ is a diagonal matrix with the diagonal entries $[v(\mu_i)h'(\mu_i)]^{-1}$ and depends on $\boldsymbol{\beta}$ via $\mu_i = \mu_i(\boldsymbol{\beta})$. Because μ is a function of $\boldsymbol{\beta}$, $\frac{\partial l}{\partial \boldsymbol{\beta}}$ is also a function of $\boldsymbol{\beta}$ through μ and $\mathbf{W}$. If we consider the canonical link function, then $\mathbf{W} = \mathbf{I}$ where $\mathbf{I}$ is identity matrix. Sometimes we can maximize this analytically and find an exact solution for the MLE of $\boldsymbol{\beta}$, says $\hat{\boldsymbol{\beta}}$, but the Normal GLM is the only common case where this is possible. Typically, we must use numerical optimization. By applying the Fisher scoring method,which is a method to solve maximum likelihood equations numerically(McCulloch and Searle (2001)), McCullagh and Nelder (1998) show that the optimization is equivalent to Iterative Weighted Least Squares(IWLS) based on a working variable. Many softwares provide program to find estimations of the parameters (e.g. Faraway 2006).
The second derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X} + \mathbf{X}^T(\frac{\partial}{\partial \boldsymbol{\beta}^T}\mathbf{W})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$$

then,

$$-E(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

which is the information matrix for $\boldsymbol{\beta}$. Under the regularity conditions, the asymptotic variance- covariance matrix of $\hat{\boldsymbol{\beta}}$ equals to the inverse of the information matrix (Pourahmadi 2002)

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \tag{2.20}$$

which depends on the unknown parameter vector through $\mathbf{W}$.

2.3 Linear Mixed Models

A natural question to raise is: what happens if one ignores the random effects or dependency of observations? Before we describe LMMs we answer these questions (or sometimes one question) with a simple example.

Example 2.1. *Suppose that Y_{ij} is the jth observation from the ith subject, and*

$$Y_{ij} = \mu + a_i + \epsilon_{ij} \quad i = 1, \dots m \quad \text{and} \quad j = 1, \dots, n \tag{2.21}$$

where $a_i \sim N(0, \sigma_a^2)$, ϵ_{ij}s are uncorrelated error terms for all i and j which are normally distributed with mean 0 and variance σ^2. $Cov(a_i, a_{i'}) = 0$ for all $i \neq i'$ and $Cov(a_i, \epsilon_{i'j}) = 0$ for all i, i' and j. In other words, we consider a balanced design with m subjects and n observations per subject.

$$Cov(Y_{ij}, Y_{ij'}) = \sigma_a^2 \Rightarrow corr(Y_{ij}, Y_{ij'}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} = \frac{1}{1+d}$$

where $d = \frac{\sigma^2}{\sigma_a^2}$. It is easy to indicate that $corr(Y_{ij}, Y_{ij'})$ is a decreasing function of d and hence increasing function of σ_a^2.
The focus here is on the inference for μ when we ignore the fact that the random effect exists. Let $\hat{\mu}$ be the maximum likelihood estimator of μ, it is easy to see that

$$\hat{\mu} = \bar{Y}.. \quad \text{and} \quad V_1 = Var(\hat{\mu}) = \frac{\sigma^2 + n\sigma_a^2}{nm}$$

Now, we ignore a_i, the random effect of the model, i.e. $Y_{ij} = \mu + \epsilon_{ij}$. We have

$$\hat{\mu} = \bar{Y}.. \quad \text{and} \quad V_2 = Var(\hat{\mu}) = \frac{\sigma^2}{nm}$$

Thus the ratio of V_2/V_1 is $d/(d+n)$
Figure 2.1 shows plot of the ratio V_2/V_1 versus d for three different n.

The main result in Figure 2.1 is that how can the ignoring of random effect influence on the variance for simple model (2.21). The message is rather clear. The discrepancy between V_1 and V_2 decreases with the rising in d and hence decreasing in $corr(Y_{ij}, Y_{ij'})$. Under the random effect ignoring, the confidence interval based on V_2 is narrower than should be. Therefore increasing in d causes that the confidence intervals based on V_1 and V_2 are closed.
Different from LMs and GLMs, which consider independence of data and involve only fixed effects, LMMs can be applied to model correlation between observations and consider random effects in addition to fixed effects. In fact, the use of random effects reflects the

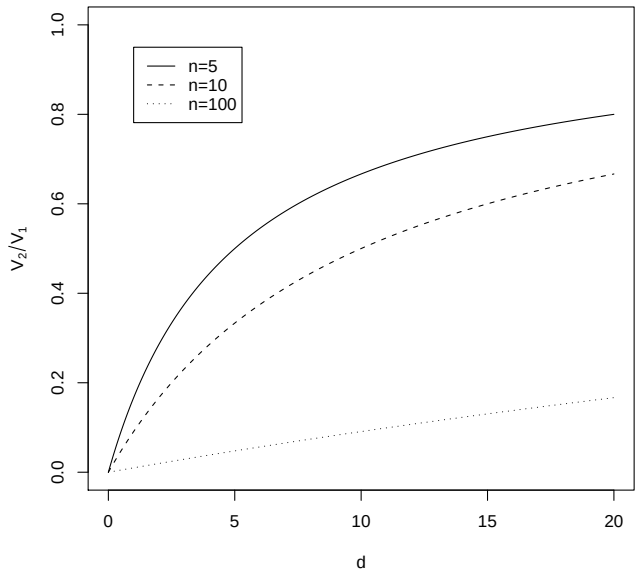


Figure 2.1: The plot of V_2/V_1 against d for $n = 5$, $n = 10$ and $n = 100$.

belief that there is heterogeneity in the subjects for a subset of the regression coefficients in $\boldsymbol{\beta}$. Since Laird and Ware (1982), mixed models have become popular and widely used tool for modeling repeated measurements in the framework of normal regression models. Before describing GLMMs, we recall the main principles of Linear Mixed Models.
Assume that n_i denote the number of observations for the ith subject provided that $n = \sum_{i=1}^{m} n_i$ be the number of the whole data at hand. The general LMM is given by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{2.22}$$

where $\mathbf{Y}_i$ is a $n_i \times 1$ vector of observation for the ith subject ($i = 1, \ldots, m$), $\boldsymbol{\beta}$ is $p \times 1$ vector of unknown fixed effects parameters and $\mathbf{b}_i$ is the $q \times 1$ vector of the random effects parameters for the ith subject in the data set which is often supposed to follow a $N_q(\mathbf{0}, \mathbf{D})$. $\mathbf{X}_i$, of order $n_i \times p$, and $\mathbf{Z}_i$, of order $n_i \times q$, are design matrices for the fixed and the random effects respectively. $\boldsymbol{\epsilon}_i$ contains the error terms for subject i and we suppose that it is normally distributed with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{R}_i$. In most application we suppose $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$. We assume that the different subjects are independent and $\mathbf{b}_i$ and $\boldsymbol{\epsilon}$ are supposed to be independent. Since

$$\boldsymbol{\mu}_i = E(\mathbf{Y}_i) = E(E(\mathbf{Y}_i \mid \mathbf{b}_i)) = \mathbf{X}_i\boldsymbol{\beta} + E(\mathbf{Z}_i\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} \tag{2.23}$$

$$\mathbf{V}_i = Var(\mathbf{Y}_i) = Var(E(\mathbf{Y}_i \mid \mathbf{b}_i)) + E(Var(\mathbf{Y}_i \mid \mathbf{b}_i)) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i \tag{2.24}$$

then $\mathbf{Y}_i \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i)$. So we can find log-likelihood function under the conditions of a known variance-covariance matrix or unknown variance-covariance matrix and then numerical methods leads us to obtain the maximum likelihood estimator.
A useful brief on this topic can be found in chapter 6 of McCulloch and Searle (2001) and more details in chapter 1 and chapter 2 of Jiang (2007).

2.4 Generalized Linear Mixed Models

After a short discussion about GLMs and LMMs, we are ready to introduce GLMMs as a generalized version of GLMs and LMMs.
Suppose that, given a vector of random effects $\mathbf{b}_i$, the response $Y_{ij}\,(i = 1, \ldots, m\,,\, j = 1, \ldots, n_i)$ corresponding to the jth observation of the ith subject are conditionally independent and the conditional density function of Y_{ij} given $\mathbf{b}_i$ is

$$f(y_{ij} \mid \mathbf{b}_i) = \exp\{\frac{y_{ij}\gamma_{ij}^{(\mathbf{b}_i)} - b(\gamma_{ij}^{(\mathbf{b}_i)})}{a(\phi)} + c(y_{ij};\phi)\} \quad \begin{cases} i = 1, \ldots, m \\ j = 1, \ldots, n_i \end{cases} \tag{2.25}$$

where, as in GLMs, $b(.)$ $a(.)$ and $c(.;.)$ are known function. In other words we assume that the conditional distribution of Y_{ij} given $\mathbf{b}_i$ is a member of an exponential family. It is also assumed that the vector $\mathbf{b}_i$ is normally distributed with mean $\mathbf{0}$ and variance-covariance $\mathbf{D}$, where $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ depends on a vector $\boldsymbol{\alpha}$ of unknown variance components.
Regarding to remark (2.2), we consider the case with known ϕ.
Since the conditional distribution is a member of an exponential family, properties of GLMs which are obtained based on this family are satisfied for conditional distributions in GLMMs. For example,

$$\mu_{ij}^{(\mathbf{b}_i)} = E(Y_{ij} \mid \mathbf{b}_i) = b'(\gamma_{ij}^{(\mathbf{b}_i)}) \tag{2.26}$$

$$Var(Y_{ij} \mid \mathbf{b}_i) = a(\phi)b''(\gamma_{ij}^{(\mathbf{b}_i)}) = a(\phi) \cdot v(\mu_{ij}^{(\mathbf{b}_i)}) \tag{2.27}$$

Similar to GLMs, the conditional mean is related to a linear predictor

$$\eta_{ij}^{(\mathbf{b}_i)} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i \tag{2.28}$$

by a link function h, $h(\mu_{ij}^{(\mathbf{b}_i)}) = \eta_{ij}^{(\mathbf{b}_i)}$. Also $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are p-dimensional and q-dimensional vectors of known covariate values. $\boldsymbol{\beta}$ and $\mathbf{b}_i$ have the same definition as in LMMs.

Remark 2.4. *As we defined in remark* (2.1)*,if* $\gamma_{ij}^{(\mathbf{b}_i)} = h(\mu_{ij}^{(\mathbf{b}_i)})$ *then we say* $h(\mu_{ij}^{(\mathbf{b}_i)})$ *is a canonical link function. In most application of GLMMs, this type of link function is considered.*

The marginal mean of Y_{ij} can be derived:

$$\mu_{ij} = E(Y_{ij}) = E(E(Y_{ij} \mid \mathbf{b}_i)) = E(\mu_{ij}^{(\mathbf{b}_i)}) = E(h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i)) \tag{2.29}$$

Because of the nonlinear function $h^{-1}(.)$, more specification, in general, is not possible. The marginal variance can be also obtained as

$$\begin{aligned} Var(Y_{ij}) &= Var(E(Y_{ij} \mid \mathbf{b}_i)) + E(Var(Y_{ij} \mid \mathbf{b}_i)) \\ &= Var(h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i)) + E(a(\phi) \cdot v(h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i)) \end{aligned} \tag{2.30}$$

which again can not be more simplified without making further assumptions on the form of $h^{-1}(.)$ and/or the conditional distribution of Y_{ij}.
By the same way, we have

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \mathrm{Cov}(h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i), h^{-1}(\mathbf{x}_{ik}^T\boldsymbol{\beta} + \mathbf{z}_{ik}^T\mathbf{b}_i)). \tag{2.31}$$

It is often assumed that the different subjects are independent, So

$$\mathrm{Cov}(Y_{ij}, Y_{i'k}) = 0 \quad \text{for all } i \neq i'. \tag{2.32}$$

Example 2.2. *Let Y_{ij} denotes the jth observation for the ith subject or individual and the Y_{ij} given $\mathbf{b}_i$ are independent following a Poisson distribution, i.e.*

$$f(y_{ij} \mid \mathbf{b}_i) = \frac{(\mu_{ij}^{(\mathbf{b}_i)})^{y_{ij}} e^{-\mu_{ij}^{(\mathbf{b}_i)}}}{y_{ij}!} \; i = 1, \ldots, m;\; j = 1, \ldots, n$$

for non-negative integer y_{ij}. We consider canonical link function, i.e. $\log(\mu_{ij}^{(\mathbf{b}_i)}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i$. Also we suppose $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$. So the mean and the variance are

$$\mu_{ij} = E(Y_{ij}) = E(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i}) = e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} \cdot E(e^{\mathbf{z}_{ij}^T\mathbf{b}_i}) = e^{\mathbf{x}_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}_{ij}^T\mathbf{D}\mathbf{z}_{ij}}$$

$$\begin{aligned} Var(Y_{ij}) &= Var(E(Y_{ij} \mid \mathbf{b}_i)) + E(Var(Y_{ij} \mid \mathbf{b}_i)) = Var(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i}) + E(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i}) \\ &= E(E(Y_{ij} \mid \mathbf{b}_i))^2 - E^2(E(Y_{ij} \mid \mathbf{b}_i)) + E(Var(Y_{ij} \mid \mathbf{b}_i)) \\ &= \mu_{ij}^2(e^{\mathbf{z}_{ij}^T\mathbf{D}\mathbf{z}_{ij}} - 1) + \mu_{ij} \end{aligned}$$

Since the term in parentheses on the right side of the last expression is always positive, the variance of Y_{ij} is greater than the mean of Y_{ij}. This fact, compared to the properties of the Poisson distribution, where we have that the mean and the variance are equal, is called over dispersion. For different subjects the covariances of responses are zero and regarding to (2.31) *for the same subject*

$$Cov(Y_{ij}, Y_{ik}) = Cov(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i}, e^{\mathbf{x}_{ik}^T\boldsymbol{\beta} + \mathbf{z}_{ik}^T\mathbf{b}_i}) + E(Cov(Y_{ij}, Y_{ik} \mid \mathbf{b}_i))$$

with regard to the definition of the model $Cov(Y_{ij}, Y_{ik} \mid \mathbf{b}_i) = 0$ for $j \neq k$ then,

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= e^{\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{x}_{ik}^T\boldsymbol{\beta}}[E(e^{(\mathbf{z}_{ij}^T + \mathbf{z}_{ik}^T)\mathbf{b}_i}) - E(e^{\mathbf{z}_{ij}^T\mathbf{b}_i})E(e^{\mathbf{z}_{ik}^T\mathbf{b}_i})] \\ &= \mu_{ij}\mu_{ik}(e^{\mathbf{z}_{ij}^T\mathbf{D}\mathbf{z}_{ik}} - 1) \end{aligned}$$

Remark 2.5. *If* $\mathbf{Y}$ *is the vector of all responses, then we can represent the conditional density function of* $\mathbf{Y}$ *given* $\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_m \end{pmatrix}$ *(as a member of the exponential family) in the matrix form as following,*

$$f(\mathbf{y} \mid \mathbf{b}; \boldsymbol{\beta}, \phi) = \exp\{(\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})/a(\phi) + \mathbf{1}^T c(\mathbf{y}; \phi)\} \tag{2.33}$$

where $\boldsymbol{\beta}$ *and* $\mathbf{X}$ *are the vector of fixed effects and the corresponding design matrix for whole data respectively.* $\mathbf{b}$ *is the vector of all random effects and* $\mathbf{Z}$ *is a corresponding block diagonal design matrix with elements* $\mathbf{Z}_i (i = 1, \ldots, m)$ *and also* $\mathbf{Z}\mathbf{b} = \sum_{i=1}^{m} \mathbf{Z}_i \mathbf{b}_i$ *with* $\mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i})^T$. $\mathbf{1}$ *is a vector with all entries one of appropriate length. Also we note that for a general vector* $\mathbf{u} = (u_1, \ldots, u_r)^T$, $a(\mathbf{u})$ *denotes the vector* $(a(u_1), \ldots, a(u_r))^T$.

2.4.1 Maximum Likelihood Estimation

If $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ be the vector of observations for the ith subject, the likelihood function for subject i becomes

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi; \mathbf{y}_i) &= f(\mathbf{y}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \int f(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\beta}, \phi) f(\mathbf{b}_i; \boldsymbol{\alpha}) d\mathbf{b}_i \\ &= \int \prod_{j=1}^{n_i} f(y_{ij} \mid \mathbf{b}_i; \boldsymbol{\beta}, \phi) f(\mathbf{b}_i; \boldsymbol{\alpha}) d\mathbf{b}_i \end{aligned} \tag{2.34}$$

The overall likelihood function for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ is obtained as,

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi; \mathbf{y}) = \prod_{i=1}^{m} L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi; \mathbf{y}_i) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij} \mid \mathbf{b}_i; \boldsymbol{\beta}, \phi) f(\mathbf{b}_i; \boldsymbol{\alpha}) d\mathbf{b}_i \tag{2.35}$$

Note that $f(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\beta}, \phi)$ is the same as $f(\mathbf{y}_i \mid \mathbf{b}_i)$. Unlike linear models, the likelihood function under a GLMM typically does not have an explicit expression. Following Wand (2007) we consider the form (2.33) for the conditional density function of $\mathbf{Y}$ given $\mathbf{b}$. Also we assume that $a(\phi) = 1$. The derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \mathbf{y})}{\partial \boldsymbol{\beta}} = \frac{\partial \log L(\boldsymbol{\beta}, \boldsymbol{\alpha}; \mathbf{y})}{\partial \boldsymbol{\beta}} = \frac{\partial \log f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}}$$

where

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \int f(\mathbf{y} \mid \mathbf{b}; \boldsymbol{\beta}) f(\mathbf{b}; \boldsymbol{\alpha}) d\mathbf{b} \\ &= \int \exp\{[\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})] + \mathbf{1}^T c(\mathbf{y})\} f(\mathbf{b}; \boldsymbol{\alpha}) d\mathbf{b} \end{aligned}$$

thus

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha};\mathbf{y})}{\partial \boldsymbol{\beta}} &= \frac{\frac{\partial f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}}}{f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\alpha})} = \frac{\int f(\mathbf{y},\mathbf{b};\boldsymbol{\beta},\boldsymbol{\alpha})\{[\mathbf{y}-b'(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b})]^T\mathbf{X}\}d\mathbf{b}}{f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\alpha})} \\ &= \int \frac{f(\mathbf{y},\mathbf{b};\boldsymbol{\beta},\boldsymbol{\alpha})}{f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\alpha})}\{[\mathbf{y}-b'(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b})]^T\mathbf{X}\}d\mathbf{b} \\ &= \int \{[\mathbf{y}-b'(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b})]^T\mathbf{X}\}f(\mathbf{b}\mid\mathbf{y};\boldsymbol{\beta},\boldsymbol{\alpha})d\mathbf{b} \\ &= (\mathbf{y}-E(\boldsymbol{\mu}^{(\mathbf{b})}\mid\mathbf{y}))\mathbf{X} \qquad (2.36)\end{aligned}$$

By the same way, the derivative of the log-likelihood with respect to $\boldsymbol{\alpha}$ is

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha};\mathbf{y})}{\partial \boldsymbol{\alpha}} = \int \frac{\partial \log f(\mathbf{b};\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} f(\mathbf{b}\mid\mathbf{y})d\mathbf{b}$$

Since in this thesis, we suppose that the parameters of random effects are known we ignore more detail about the MLE of $\boldsymbol{\alpha}$.
It is easy to see that (2.36) can not be further simplified, so we, in the general case, can not find a closed form for the estimators of $\boldsymbol{\beta}$. In most cases, due to the multidimensional integral involved in (2.36), such an integral is difficult to evaluate even numerically.

3 Quasi-likelihood

3.1 Introduction

In applications of a statistical model, in a few cases, we know the specific distribution of the random variable which has generated the data. Also, in many cases, in spite of a known density function of the data, due to difficulty in integrating we can not find a closed form for the Likelihood function. This case occurs especially in the models with random effects. We may, however, be able to specify some characteristics of data like cumulants.
Historically, two particular views are considered in parameter estimation theory. The first one, introduced by Fisher, is the maximum likelihood (ML) method which is dependent on the entire form of the underlying distribution. The second view, introduced by Gauss and Legendre, is the least squares (LS) method that works based on minimizing the sum of squared errors.
These approaches have been unified under the general description of Quasi-Likelihood (QL) via estimating functions. Earlier works on the QL method was made available by Wedderburn (1974) who relaxed the distributional assumption through the specification of a variance function in Generalized Linear Models where the distribution of observations is from the exponential family. This suggestion raises the question of what happens when the true underlying distribution is not from the exponential family? This question is answered by relating the Quasi-Likelihood to estimating functions. Some properties of QL estimates, irrespective of underlying distribution, via estimating function have been considered in Godambe and Heyde (1987) and Heyde (1997). A brief review on QL through estimating functions has been prepared in Desmond (1997). McCullagh (1983) provided some asymptotic properties of QL estimators. An extensive work can be found in McCullagh and Nelder (1998)(ch.9) on this topic.
The purpose of this chapter is to show how inference can be drawn from experiments in which there is insufficient information to construct the likelihood function explicitly.
In the next section, we introduce estimating functions and the Optimal Estimating Function to estimate parameters. In section 3, we first introduce the Quasi-Likelihood method to estimate and then use the relation between QL and estimating functions, we extend the properties of estimating functions to QL method.

3.2 Estimating Functions

Estimating functions, the term may have been coined by Kimball (1946), provide a general framework for finding estimators and studying their properties in many different kinds of statistical models. The estimating function approach has turned out to be very useful in obtaining estimators where the likelihood function is usually not explicitly known.

Def 3.1. *An estimating function g_n for μ is a function of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as well as the parameter of interest μ, i.e. $g_n(\mathbf{Y}; \mu)$.*

We get an estimator by solving $g_n(\mathbf{Y}; \mu) = 0$. That is, $\hat{\mu}$ is an estimator based on $\mathbf{y}$ and g_n, if $g_n(\mathbf{y}, \hat{\mu}) = 0$. There might be more than one solution or no solution at all.

Def 3.2. *$g_n(\mathbf{Y}; \mu)$ is an unbiased estimating function if $E_\mu(g_n(\mathbf{Y}; \mu)) = 0$ for all μ.*

Example 3.1. *As an immediate consequence, under the regularity conditions the derivative of log-likelihood function is an unbiased estimating function as well as least square.*

It is pointed out that not all estimates covered by the estimating function method need be unbiased, while the corresponding estimating function is unbiased. Higher order moments of $g_n(\mathbf{Y}; \mu)$ might be dependent on the parameter, so that the estimating function need not be a pivotal function.
Note that we consider unbiased and square integrable version of the estimating function in this text as needed. To clarify the basic concepts, we first consider the simplest case with random independent variables and a scalar parameter μ in the following simple example.

Example 3.2. *Let $Y_1, \ldots, Y_n$ be independent random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$. Consider the estimating function $g_n(\mathbf{Y}; \mu) = \sum_{i=1}^{n} b_i(Y_i - \mu)$, then $\hat{\mu}_b = \frac{\sum_{i=1}^{n} b_i Y_i}{\sum_{i=1}^{n} b_i}$ with $\sum_{i=1}^{n} b_i \neq 0$ is an unbiased estimator which is the solution of $g_n(\mathbf{Y}; \mu) = 0$.*

For simplicity we denote $g_n(\mathbf{Y}; \mu)$ by g_n whenever this doesn't lead to ambiguities.

It is easy to see that g_n and $\dot{g_n} = kg_n$ (k is a constant) result in the same estimator $\hat{\mu}_b$. But $Var(\dot{g_n}) = k^2 Var(g_n)$ can however be made arbitrarily small, thus two estimating functions are not comparable based on the variance of the estimating functions. One remedy is to use some standardized versions of estimating functions. One possible standardization is to define it as (Heyde 1997)

$$g_n^{(s)}(\mathbf{Y}; \mu) = (-E_\mu(g_n'))(E_\mu(g_n^2))^{-1} g_n(\mathbf{Y}; \mu) \tag{3.1}$$

where g_n' is the derivative of $g_n(\mathbf{y}, \mu)$ with respect to μ. It produces the same estimator of μ for different k and $Var(g_n^{(s)}) = Var(kg_n)^{(s)}$.
This standardization helps us to find a measure to compare different estimating functions in some sense. In addition to the above property of standardization, the following advantage of a standard version of g_n is important (Heyde 1997 and Godambe 1991).

- One measure that we want to minimize, is $Var(g_n) = E(g_n^2)$. On the other hand we would like that $g_n(\mathbf{Y}; \mu)$ be sensitive to small varies in μ where μ is true value. That is we want $g_n(\mathbf{Y}; \mu + \delta\mu) - g_n(\mathbf{Y}; \mu)$ with $\delta > 0$, to differ as much as possible from 0. These two statements are achieved by maximizing $Var(g_n^{(s)}) = \frac{(E(g_n'))^2}{E(g_n^2)}$.

The following property is also true in example (3.2).

- Under the Feller condition (See e.g. Bauer (1996), page 235) which guarantees that the ${b_i}^2$ are small compared to their sum, $\sum_{i=1}^{n} b_i^2$, in the sense that for given $\epsilon > 0$, $\frac{b_i}{\sqrt{\sum_{i=1}^{n} b_i^2}} < \epsilon$ for $i = 1, \ldots, n$ when n is sufficiently large , using the Lindeberg-Feller Central Limit theorem (Sec. 2.8 in Van der Vaart (1998))

$$\sum_{i=1}^{n} b_i(Y_i - \mu)/(\sigma^2 \sum_{i=1}^{n} b_i^2)^{\frac{1}{2}} \xrightarrow{D} N(0, 1)$$

  Hence with regard to estimator of μ, $\hat{\mu}_b = \frac{\sum_{i=1}^{n} b_i Y_i}{\sum_{i=1}^{n} b_i}$, we have

$$\hat{\mu} - \mu \xrightarrow{D} N(0, Var^{-1}(g_n^{(s)})) = \frac{E(g_n^2)}{(E(g_n'))^2})$$

  where $Var^{-1}(g_n^{(s)}) = \frac{E(g_n^2)}{(E(g_n'))^2}$. The length of confidence interval for μ is proportional to the inverse of the $Var(g_n^{(s)})$. Thus maximizing of $Var(g_n^{(s)})$ coincides with small confidence interval.

The following theorem states an important property of estimating function which is called invariance property.

Theorem 3.2.1 (Godambe(1991))**.** *The estimating function is invariant under one to one transformation of the parameter μ.*

Proof. We have to show that if $g_n(\mathbf{Y}; \mu)$ is an estimating function and $\hat{\mu}$ is an estimate for $\mu \in \Theta$, then under the one-to-one transformation $\varphi = \alpha(\mu)$ $(\varphi : \Theta \longrightarrow \Lambda)$, $\hat{\varphi} = \alpha(\hat{\mu})$ is an estimate for φ.
If $\hat{\mu}$ is an estimator based on $g_n(\mathbf{Y}; \mu)$ then

$$g_n(\mathbf{Y}, \hat{\theta}) = 0 \Longrightarrow \exists\, \hat{\varphi},\ g_n(\mathbf{Y}, \alpha^{-1}(\hat{\varphi})) = \dot{g}_n(\mathbf{Y}, \hat{\varphi}) = 0$$

then $\hat{\varphi}$ is an estimator based on $\dot{g}_n(\mathbf{Y}, \varphi)$. □

Although this property doesn't hold for unbiased minimum variance estimators, it is well known that the invariance property is enjoyed by the maximum likelihood estimators (Godambe and Thompson (1978)).

Remark 3.1. *b_i's might depend on μ_i which differ from one individual to another. And also μ_i may be link to a linear or non-linear combination of p-vector unknown parameter, $\boldsymbol{\beta}$, through a known function, where $\boldsymbol{\beta}$ is taking values in an open subset Θ of p-dimensional Euclidian space, $\mathbb{R}^p$.*

Now we concentrate on a more general class of unbiased and square integrable estimating functions $G = \{\mathbf{g}(Y_1, \ldots, Y_n; \boldsymbol{\beta}) : E(\mathbf{g}(Y_1, \ldots, Y_n; \boldsymbol{\beta})) = \mathbf{0}, \boldsymbol{\beta} \in \Theta\}$, where $\mathbf{g}_n = \mathbf{g}(Y_1, \ldots, Y_n; \boldsymbol{\beta})$ is a $p \times 1$ vector with element $g_{i(n)}$.

Regarding to the class $\mathbf{G}$, the standardized version of $\mathbf{g}_n$ can be represent as

$$\mathbf{g}_n^{(s)} = -E(\mathbf{g}_n')^T (E(\mathbf{g}_n \mathbf{g}_n^T))^{-1} \mathbf{g}_n \tag{3.2}$$

which is a generalization of Fisher information. In this expression the components of $\mathbf{g}_n'$, of order $p \times p$, are $g'_{ir(n)} = \frac{\partial g_{i(n)}}{\partial \beta_r}$

Def 3.3. *Consider the class of unbiased estimating function G. A member of G, $\mathbf{g}_n^*$ is F-optimal (That is finite sample optimality (Desmond (1991))) within this class, if it maximizes*

$$E(\mathbf{g}_n^{(s)} \mathbf{g}_n^{(s)^T}) = E(\mathbf{g}')^T (E(\mathbf{g}_n \mathbf{g}_n^T))^{-1} E(\mathbf{g}_n') \tag{3.3}$$

i.e. $Var(\mathbf{g}_n^{*^{(s)}}) \geq Var(\mathbf{g}_n^{(s)})$ for all $\mathbf{g}_n \in \mathbf{G}$ uniformly in $\boldsymbol{\beta} \in \Theta$.

It is clear from the nature of the above definition that finding the F-optimality based on the comparison, in the case with more than one-dimensional in vector of parameters, between two matrices is difficult and sometimes is impossible. By results in Heyde (1997) $\mathbf{g}^*$ is F-optimal if

$$E(\mathbf{g}_n^{*(s)} \mathbf{g}_n^{(s)^T}) = E(\mathbf{g}_n^{(s)} \mathbf{g}_n^{*(s)^T}) = E(\mathbf{g}_n^{(s)} \mathbf{g}_n^{(s)^T}) \text{ for all } \mathbf{g}_n^{(s)} \tag{3.4}$$

or equivalently

$$(E(\mathbf{g}_n'))^{-1} E(\mathbf{g}_n \mathbf{g}_n^{*T}) \tag{3.5}$$

is a constant matrix for all $\mathbf{g}_n \in G$. In the practical point of view, these conditions are simpler than the condition in the definition of F-optimality.

Note that under the condition of existence of an F-optimal estimating function, we may compare matrices by some real functions. The following theorem (Heyde 1997) states that.

Theorem 3.2.2. *Suppose* $\mathbf{G}$ *is a set of estimating functions for which an F-optimal estimating function,* $\mathbf{g}_n^*$*, exists. "*$\mathbf{g}_n^*$ *is F-optimal" is equivalent to either of the three following statements:*

1. *Trace optimality:* $tr(E(\mathbf{g}_n^*\mathbf{g}_n^{*^T})) \geq tr(E(\mathbf{g}_n\mathbf{g}_n^T))$ *for all* $\mathbf{g}_n$
2. *Determinant optimality:* $\det(E(\mathbf{g}_n^*\mathbf{g}_n^{*^T})) \geq \det(E(\mathbf{g}_n\mathbf{g}_n^T))$ *for all* $\mathbf{g}_n$
3. *Smallest eigenvalue optimality:* $\lambda_{min}(E(\mathbf{g}_n^*\mathbf{g}_n^{*^T})) \geq \lambda_{min}(E(\mathbf{g}_n\mathbf{g}_n^T))$ *for all* $\mathbf{g}_n$

where tr, det and λ_{min} denote trace, determinant and minimum of eigenvalues respectively.

Proof. The proof can be found in Heyde (1997)(pages 19-21).

3.3 Quasi-Likelihood

Wedderburn (1974) observed that, from a computational point of view, the only two assumptions of GLM necessary to fit the model were a specification of the mean (in term of the regression parameters) and the relationship between the mean and the variance. This led him to replace the full distributional assumption about the random component in the model by a much weaker assumption of mean-variance relationship alone.
Suppose the $n \times 1$ random variable $\mathbf{Y}$ has mean $\boldsymbol{\mu}(\boldsymbol{\beta})$ and variance-covariance matrix $a(\phi)\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$ where $\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$ is the variance function . Both are known functions of the p-dimensional parameter vector $\boldsymbol{\beta}$ and $\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$ is a positive definite matrix. Also ϕ is the nuisance parameter. Nuisance parameters, by definition, are of little intrinsic interest to investigators, yet necessary to fully specify the random mechanism of the data.

Remark 3.2. *As we mentioned in previous chapter, we suppose that* ϕ *is known then without loss of generality we assume that* $a(\phi) = 1$ *if this doesn't lead to ambiguity.*

The p-dimensional vector $U^{(n)}(\boldsymbol{\beta})$ with the entries $U_r^{(n)}(\boldsymbol{\beta})\, i = 1, \ldots, p$ is the quasi-score function, which considers as a function of $\boldsymbol{\beta}$, and it is given by the system of partial differential equations

$$U^{(n)}(\boldsymbol{\beta}) = \frac{\partial ql(\boldsymbol{\beta}; \mathbf{y})}{\partial \boldsymbol{\beta}} = \frac{\partial \log(Ql(\boldsymbol{\beta}; \mathbf{y}))}{\partial \boldsymbol{\beta}} = \mathbf{D}^T\mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}))(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \tag{3.6}$$

where $\mathbf{D}^T = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. In other words, $ql(\boldsymbol{\beta}; \mathbf{y})$ as a solution for the above differential equation (if it exists) is called the log Quasi-likelihood function.
$\hat{\boldsymbol{\beta}}^{(n)}$ is maximum QL estimator of $\boldsymbol{\beta}$ if $U^{(n)}(\hat{\boldsymbol{\beta}}^{(n)}) = \mathbf{0}$.Unfortunately, we can not usually find a closed form expression to obtain the root of (3.6), although root of that can be found through out numerical methods. From the point of view of the statistical inference, fortunately, we can study the properties of $\hat{\boldsymbol{\beta}}^{(n)}$ indirectly by studying the Quasi-score function.

Remark 3.3. *The expression* (3.6)*, as we will indicate later, possesses the properties of the derivative of the log-likelihood .*

Remark 3.4. *For a single observation* $ql(\boldsymbol{\beta}; y)$ *can be defined as an integral* $\int_y^{\mu(\boldsymbol{\beta})} \frac{y-t}{V(t)} dt$*, if the integral exists.*

When the certain mean-variance relationship are specified, the QL function sometimes turn out to be a recognizable likelihood function. Wedderburn (1974) indicated that the QL is unified with the likelihood function if the family of distributions is from an one-parameter exponential family. The efficiency of the QL estimator of $\boldsymbol{\beta}$ relative to the ML estimator of $\boldsymbol{\beta}$ has been considered by Firth (1987) when the true probability mechanism of the random variables doesn't follow an exponential family distribution. Several different types of departure from the exponential dispersion family were established and high efficiency of QL estimator of $\boldsymbol{\beta}$ is maintained when the departure of f, the true density function, from the exponential dispersion family is only modest.
In the simple case, we suppose that the components of the response vector $\mathbf{Y}$ are independent. Thus the matrix $\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$ is diagonal

$$\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta})) = diag\{V_1(\boldsymbol{\mu}(\boldsymbol{\beta})), \ldots, V_n(\boldsymbol{\mu}(\boldsymbol{\beta}))\}$$

where $Var(Y_i) = V_i(\boldsymbol{\mu}(\boldsymbol{\beta}))$. Also, in most applications, one can assume that $V_i(\boldsymbol{\mu}(\boldsymbol{\beta}))$ depends on the ith component of $\boldsymbol{\mu}(\boldsymbol{\beta})$, i.e. $\mu_i(\boldsymbol{\beta})$. Thus

$$\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta})) = diag\{V_1(\mu_1(\boldsymbol{\beta})), \ldots, V_n(\mu_n(\boldsymbol{\beta}))\}$$

An unnecessary assumption is to suppose that the functions $V_1(.), \ldots, V_n(.)$ are identical, through their arguments, and hence their values, are different. The condition is satisfied in the majority of the applied models.
Thus under the above assumptions, we can represent $U^{(n)}(\boldsymbol{\beta})$ in the form of the gradient vector of the QL function as following

$$U^{(n)}(\boldsymbol{\beta}) = \mathbf{D}^T \boldsymbol{\Delta} \tag{3.7}$$

where $\boldsymbol{\Delta}$ is a $n \times 1$ vector with entries $\frac{y_i - \mu_i(\boldsymbol{\beta})}{V(\mu_i(\boldsymbol{\beta}))}$, $i = 1, \ldots, n$.

There are many situations where the dependence relationship among the data is so strong such that we can not ignore them. Repeated measurements observations, for example longitudinal data, can be recognized as dependent observation. Liang and Zeger (1986) proposed the generalized estimating equations (GEE) approach as an application of the QL approach to longitudinal data analysis. They considered the case that $Var(\mathbf{Y}) = \mathbf{V}(a(\phi), \boldsymbol{\rho})$, where $\boldsymbol{\rho}$ is a vector of autoregressive coefficients. Here and in the following we consider the case that we have $Var(\mathbf{Y}) = \mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$.In this case, the Quasi-score function (3.6) has the following properties (McCullagh and Nelder (1998))

1. $E(U_r^{(n)}(\boldsymbol{\beta})) = \mathbf{0}$

2. $Var(U^{(n)}(\boldsymbol{\beta})) = \mathbf{D}^T\mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}))\mathbf{D}/\phi^2 = M_{\boldsymbol{\beta},n}$

where $M_{\boldsymbol{\beta},n} = -E(\frac{\partial U^{(n)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})$.
If the covariance matrix of $\mathbf{Y}$, or $\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$, is not of full rank then any generalized inverse can be used, providing that this inverse, say $\mathbf{V}^{-}(\boldsymbol{\mu}(\boldsymbol{\beta}))$, ensures that $Var(U^{(n)}(\boldsymbol{\beta})) = \mathbf{D}^T\mathbf{V}^{-}(\boldsymbol{\mu}(\boldsymbol{\beta}))\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))\mathbf{V}^{-}(\boldsymbol{\mu}(\boldsymbol{\beta}))\mathbf{D}$ is equal to $M_{\boldsymbol{\beta},n}$.(McCullagh 1991)
The above properties of $U^{(n)}(\boldsymbol{\beta})$ are in accordance with these of a score function based on the likelihood function. There is a curious limitation in the equivalence between the Likelihood function and the Quasi-likelihood function which pointed the following remark.

Remark 3.5. *There is an elegant point which is not clear for inference in the case of quasi-likelihood. An important property, which is satisfied by the score function, is that the derivative matrix of the score function with respect to $\boldsymbol{\beta}$ is symmetric. According to McCullagh and Nelder (1998), $U^{(n)}(\boldsymbol{\beta})$ is the gradient vector of a log quasi-likelihood if and only if the derivative matrix of $U^{(n)}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is symmetric. For more details see 9.3.2 in McCullagh and Nelder (1998).*

Remark 3.6. *The above properties refer to the derivative with respect to $\boldsymbol{\beta}$ and not with respect to ϕ. In other words, the statistical properties of the QL function, in terms of the quasi-score function, are similar to those of ordinary likelihood functions except that the nuisance parameter, ϕ, when it is unknown, is treated separately from $\boldsymbol{\beta}$. So the nuisance parameter could have non-trivial impact on estimating $\boldsymbol{\beta}$ as well. Nelder and Pregibon (1987) extended the quasi-likelihood to the case in the presence of a nuisance parameter. In our text we consider only a known nuisance parameter.*

Under some weak conditions on the third derivative of the link function and assuming that $n^{-1}M_{\boldsymbol{\beta},n}$ has a positive definite limit and also that the third moments of $\mathbf{Y}$ are finite, following McCullagh (1983), we have the following results:

1. $U^{(n)}(\boldsymbol{\beta}) = O_p(n)$
2. $n^{-\frac{1}{2}}U^{(n)}(\boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, M_{\boldsymbol{\beta},n}/n) \quad \text{as } n \to \infty$
3. $I_{\boldsymbol{\beta}}^{(n)} = -\frac{\partial}{\partial \boldsymbol{\beta}}(U^{(n)}(\boldsymbol{\beta})) = O_p(n)$
4. $I_{\boldsymbol{\beta}}^{(n)} - M_{\boldsymbol{\beta},n} = O_p(n^{1/2})$

Furthermore, the maximum Quasi-likelihood estimator, $\hat{\boldsymbol{\beta}}^{(n)}$, is consistent for $\boldsymbol{\beta}$, i.e.,

$$\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta} = O_p(n^{-1}) \tag{3.8}$$

Also the distribution of $\hat{\boldsymbol{\beta}}^{(n)}$ is asymptotic Normal as following

$$n^{-1/2}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, ni_{\boldsymbol{\beta}}^{-1}) \quad \text{as } n \to \infty \tag{3.9}$$

That is $\hat{\boldsymbol{\beta}}^{(n)}$ is unbiased for $\boldsymbol{\beta}$ asymptotically and the covariance matrix of $\hat{\boldsymbol{\beta}}^{(n)}$ is $i^{-1}_{\boldsymbol{\beta},n}$, given that the eigenvalues λ of $M_{\boldsymbol{\beta},n}$ be enough large for all $\boldsymbol{\beta}$.
We consider the class of the linear unbiased estimating functions, $\mathbb{H}$, which is defined by components

$$\mathbf{g}_n = \mathbf{H}^T(\boldsymbol{\beta})(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \tag{3.10}$$

where $\mathbf{H}(\boldsymbol{\beta})$, an $n \times p$ matrix, may be dependent on unknown parameter, $\boldsymbol{\beta}$, but it doesn't depend on $\mathbf{Y}$. As we see from the definition of $\mathbf{g}_n$, "linear" for $\mathbf{g}_n$ is "linear in $(Y_i - \mu_i(\boldsymbol{\beta}))$" $(i = 1, \ldots, n)$.
Regarding to the definition of Quasi-score function, we have

$$\left.\begin{array}{c} E(\mathbf{g}_n') = -\mathbf{H}^T(\boldsymbol{\beta})\frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{H}^T(\boldsymbol{\beta})\mathbf{D} \\ \text{and} \\ E(\mathbf{g}_n \mathbf{U}^{(n)^T}) = E(\mathbf{H}^T(\boldsymbol{\beta})(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T \mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}))\mathbf{D}) \end{array}\right\} \Rightarrow (E(\mathbf{g}_n'))^{-1} E(\mathbf{g}_n \mathbf{U}^{(n)^T}) = -I$$

is a constant matrix. Thus according to (3.5) the Quasi-score function is a F-optimal estimating function in the linear estimating function class $\mathbb{H}$.
In addition to the properties which have been obtained in the previous section, in the reminder of this section we describe some other properties which are satisfied for the Quasi-likelihood function as a F-optimal estimating function.
If $f_{\mathbf{y}}(\boldsymbol{\beta})$ is the density function of $\mathbf{Y}$, then we can write, in the general case, $\mathbf{U}^{(n)}(\boldsymbol{\beta}) = f_{\mathbf{y}}^{-1}(\boldsymbol{\beta}) f_{\mathbf{y}}'(\boldsymbol{\beta})$ for the score function. Also we assume that $\mathbf{U}^{(n)}(\boldsymbol{\beta})$ is almost surely differentiable with respect to the components of $\boldsymbol{\beta}$.
Now we suppose that $\mathbf{g}_n = (g_{1,n}, \ldots, g_{p,n})$, $\mathbf{U}^{(n)} = (U_1^{(n)}, \ldots, U_p^{(n)})$ and $\mathbf{H}(\boldsymbol{\beta}) = (\mathbf{h}_1(\boldsymbol{\beta}), \ldots, \mathbf{h}_p(\boldsymbol{\beta}))$ under the regularity conditions,

$$E(\mathbf{g}_n \mathbf{U}^{(n)^T}) = (E(g_{i,n} U_j^{(n)}))_{i,j=1}^p \text{ and,}$$

$$\begin{aligned} E(g_{i,n} U_j^{(n)}) &= \int \mathbf{h}_i^T(\boldsymbol{\beta})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))(\frac{\partial}{\partial \beta_j} f_{\mathbf{y}}(\boldsymbol{\beta}))\frac{1}{f_{\mathbf{y}}(\boldsymbol{\beta})} f_{\mathbf{y}}(\boldsymbol{\beta}) d\mathbf{y} \\ &= \int \mathbf{h}_i^T(\boldsymbol{\beta}) \mathbf{y} \frac{\partial}{\partial \beta_j} f_{\mathbf{y}}(\boldsymbol{\beta}) d\mathbf{y} - \mathbf{h}_i^T(\boldsymbol{\beta}) \boldsymbol{\mu}(\boldsymbol{\beta}) \frac{\partial}{\partial \beta_j} \int f_{\mathbf{y}}(\boldsymbol{\beta}) d\mathbf{y} \\ &= \mathbf{h}_i^T(\boldsymbol{\beta}) \frac{\partial}{\partial \beta_j} \boldsymbol{\mu}(\boldsymbol{\beta}) \end{aligned}$$

$$\Rightarrow E(\mathbf{g}\mathbf{U}^{(n)^T}) = H^T \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}) = -E(\mathbf{g}') \tag{3.11}$$

Now we consider the vector correlation which measures the association between $\mathbf{g}$ and $\mathbf{U}^{(n)}$, and which is defined by

$$\rho^2_{\mathbf{g},\mathbf{U}^{(n)}} = \frac{(\det(E(\mathbf{g}\mathbf{U}^{(n)^T})))^2}{\det(E(\mathbf{g}\mathbf{g}^T)) \det(E(\mathbf{U}^{(n)}\mathbf{U}^{(n)^T}))} \tag{3.12}$$

(Hotelling (1936)). Regarding to (3.11), the maximization of $\rho^2_{\mathbf{g},\mathbf{U}^{(n)}}$ is equivalent to the maximizing

$$(\det(E(\mathbf{g}')))^2/\det(E(\mathbf{g}\mathbf{g}^T))$$

which is equivalent to maximize

$$\det(E(\mathbf{g}')^T(E(\mathbf{g}\mathbf{g}^T))^{-1}E(\mathbf{g}'))$$

That is, the maximum of the correlation between estimating function in the class $\mathbb{H}$ and true score function is achieved when the estimating function is F-optimal or the Quasi-score function, if it exists.
The next property is that the Quasi-score function as a F-optimal estimating function is unique, in the sense that the standardized version of optimal estimating function is unique.
Since, even under one to one transformation, unbiased estimators is not invariant thus we can not extend invariance property to optimal estimating function. But Suppose that $\mathbf{Y}_L = \mathbf{L}\mathbf{Y}$, $\boldsymbol{\mu}_\mathbf{L}(\boldsymbol{\beta}) = \mathbf{L}\boldsymbol{\mu}(\boldsymbol{\beta})$ and $\mathbf{V}_\mathbf{L}(\boldsymbol{\beta}) = \mathbf{L}\mathbf{V}(\boldsymbol{\beta})\mathbf{L}$ where $\mathbf{L}$ is a nonsingular matrix of order n. The Quasi-score function based on $\mathbf{Y}_L$

$$\mathbf{U}_\mathbf{L}(\boldsymbol{\beta}) = (\frac{\partial\mu_\mathbf{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}})^T\mathbf{V}_\mathbf{L}^{-1}(\boldsymbol{\beta})(\mathbf{Y}_L - \boldsymbol{\mu}_\mathbf{L}(\boldsymbol{\beta})) = \mathbf{D}^T\mathbf{V}^{-1}(\boldsymbol{\beta})(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{U}(\boldsymbol{\beta})$$

is the same as the Quasi-score function based on $\mathbf{Y}$, $\mathbf{U}^{(n)}(\boldsymbol{\beta})$. This condition is weaker than the corresponding condition for Score function.

3.3.1 A simulated example

To explore the adequacy of QL estimations a simulation study was performed. Data was generated according to the following model

$$Y_{ijk} \mid b_i \sim ind.P(\mu_{ij}^{(b_i)})\; i = 1, \ldots, m;\; j = 1, \ldots, t,\; k = 1, \ldots s \tag{3.13}$$

where $\mu_{ij}^{(b_i)} = e^{b_i+\beta_0+\beta_1 x_j}$ is the conditional expectation of Y_{ijk} given b_i.
As we described in previous chapter, this model is a special case of GLMMs, where $\mathbf{z}_{ij} = 1$ and b_i is one-dimensional for all i and j, and $\mathbf{x}_{ij}^T = (1, x_j)$ for all i. Data are generated for the following settings $m = 30, t = 2, s = 2, 12, 24$, $\boldsymbol{\beta} = (3, -2)$ and $x_j = 0, 3$ for different σ^2 ($\sigma^2 = 0$, $\sigma^2 = 2$ and $\sigma^2 = 6$). The results including the mean and standard deviation are obtained in Table (3.1).
Note that due to the random effects generating in simulated data, the results might be not fixed for different tries but they must be close to these results. For this simulation I used 300 repetitions.

Table 3.1: QL estimation of parameters (and standard deviation) of the model (3.13) based on simulated data

		$l=2$	$l=12$	$l=24$
$\sigma^2=0$	$\hat{\beta}_0$	3.0042(0.0091)	2.9995(0.0112)	3.0011(0.0900)
	$\hat{\beta}_1$	-1.9981(0.0562)	-2.0049(0.0771)	-2.0078(0.0590)
$\sigma^2=2$	$\hat{\beta}_0$	3.0369(0.3799)	3.0466(0.3743)	2.9728(0.3867)
	$\hat{\beta}_1$	-2.0082(0.0876)	-2.0036(0.0380)	-2.0007(0.0262)
$\sigma^2=2$	$\hat{\beta}_0$	3.4073(1.0603)	3.2619(1.0433)	3.2979(1.0432)
	$\hat{\beta}_1$	-2.0011(0.0022)	-1.9999(0.0032)	-1.9998(0.0023)

Note that $\sigma^2=0$ corresponds to the model without random effect, i.e. $Y_{ijk} \sim ind.P(\mu_{ij})$ where $\mu_{ij} = e^{\beta_0+\beta_1 x_j}$ for all i. The results indicate that when σ^2 increases the accuracy of estimation of parameters will be less. Also when s rises the estimation of parameters will be closer to the true values of the parameters.

3.4 Penalized Quasi-Likelihood Estimator

As we saw in the previous sections, in the QL method random effects are not involved with unknown parameters. Approximation of the likelihood function is an alternative method to find an estimation of the parameters in the GLMMs. Several of approximations have been proposed in the literature. Among them penalized quasi-likelihood by Breslow and Glayton (1993) is the most popular for GLMMs. It approximates the integral involved in the likelihood function using the well-known Laplace approximation.

The integrated quasi-likelihood based on the observations of ith subject corresponding to (2.34) is given by (Breslow and Glayton (1993))

$$IQL = (2\pi)^{\frac{q}{2}}(\det(\mathbf{D}))^{-\frac{1}{2}} \int \exp[\sum_{j=1}^{n_i} ql(\mu_{ij}^{(\mathbf{b}_i)}; y_{ij}) - \frac{1}{2}\mathbf{b}_i^T\mathbf{D}^{-1}\mathbf{b}_i]d\mathbf{b}_i \tag{3.14}$$

We can represent the above expression as

$$IQL_i = c(det(\mathbf{D}))^{-\frac{1}{2}} \int e^{-k(\mathbf{b})}d\mathbf{b}$$

where $k(\mathbf{b}) = -\sum_{j=1}^{n_i} ql(\mu_{ij}^{(\mathbf{b}_i)}; y_{ij}) + \frac{1}{2}\mathbf{b}_i^T\mathbf{D}^{-1}\mathbf{b}_i$. Using Laplace approximation (Barndorff-Nielsen and Cox (1989) sec 3.3),

$$\log(IQL_i) \approx -\frac{1}{2}\log(det(\mathbf{D})) - \frac{1}{2}\log(k''(\tilde{\mathbf{b}})) - k(\tilde{\mathbf{b}})$$

where $\tilde{\mathbf{b}}$ is the root of $k'(\mathbf{b}) = 0$. Thus following Breslow and Glayton (1993), $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}})$ which maximize jointly

$$PQL_i(\boldsymbol{\beta}, \mathbf{b}) = \sum_{j=1}^{n_i} ql(\mu_{ij}^{(\mathbf{b}_i)}; y_{ij}) - \frac{1}{2}\mathbf{b}_i^T \mathbf{D}^{-1}\mathbf{b}_i \tag{3.15}$$

are called penalized quasi-likelihood(PQL) estimations of $\boldsymbol{\beta}$ and $\mathbf{b}$ based on the observations of subject i. $PQL_i(\boldsymbol{\beta}, \mathbf{b})$ is individual PQL function which has been penalized the QL function by the term $\frac{1}{2}\mathbf{b}_i^T \mathbf{D}^{-1}\mathbf{b}_i$. And for whole observations, the PQL function is

$$PQL(\boldsymbol{\beta}, \mathbf{b}) = \sum_{i=1}^{m} PQL_i(\boldsymbol{\beta}, \mathbf{b}) \tag{3.16}$$

The PQL estimators of the parameters are biased. Lin and Breslow (1996) consider some correction for the bias of the estimators.

4 Poisson Regression Models with Random Intercept and Random Slope

In the previous chapters, we considered a general framework for our models. there were many problems to achieve a closed form for variance-covariance structure and hence for information matrix due to the unspecified models.
In this chapter using the two previous chapters we apply quasi-likelihood in some special cases of GLMMs which are well-known as the Poisson Regression Model with Random Intercept and the Poisson Regression Model with Random Slope.
In the next section we consider a Poisson regression with random intercept which is divided in detail in two models: simple Poisson regression model with random intercept and quadratic Poisson regression model with random intercept. After that in section 3, we discuss the simple Poisson regression with random slope. We will find the variance-covariance structure of the estimator of the fixed effect parameters to be similar to the Poisson regression model with random intercept.

4.1 Poisson Regression Model with Random Intercept

We consider the Poisson regression model with random intercept which can be written as:

$$Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \begin{cases} i = 1, \ldots, s & \sum_{j=1}^{t_i} m_{ij} = m_i \\ j = 1, \ldots, t_i & \\ k = 1, \ldots, m_{ij} & n = \sum_{i=1}^{s} m_i \end{cases} \tag{4.1}$$

i.e. $f(y_{ijk} \mid b_i) = \frac{(\mu_{ij}(b_i))^{y_{ijk}} e^{\mu_{ij}(b_i)}}{y_{ijk}!}$
where $\mu_{ij}(b_i) = exp(b_i + \mathbf{f}^T(x_{ij})\boldsymbol{\beta})$ is specified by the canonical link function. The deviation b_i is assumed to be normal distributed with mean 0 and known variance σ^2. The random intercepts are uncorrelated for different individuals, i.e., $cov(b_i, b_{i'}) = 0$ for all $i \neq i'$ Here, Y_{ijk} stands for the kth replication for the individual i at the experimental setting x_{ij} from the experimental region τ. Also we suppose that m_{ij} denotes the number of replications of individual i at the jth level of x. The vector of known regression functions $\mathbf{f} = (1, f_1, \ldots, f_{p-1})^T$ is the same for all individuals. The p-dimensional vector of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ associated with the mean response curve is unknown and it is also the same for all individual. The random variable b_i is the individual deviation from the overall population intercept β_0, i.e. $\beta_0 + b_i$ is the random intercept

which varies and depends on the different individuals. The deviation b_i is assumed to be normally distributed with mean 0 and known variance σ^2. The random intercepts are uncorrelated for different individuals, i.e., $cov(b_i, b_{i'}) = 0$ for all $i \neq i'$. Note that $Var(Y_{ijk} \mid b_i) = E(Y_{ijk} \mid b_i) = \mu_{ij}(b_i)$.
As we have seen in the example (2.2) this model is a special case of GLMMs with $\mathbf{z}_{ij} = 1$
With regard to section(2.4), if σ^2 is known then the likelihood function for $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{s} \int \prod_{j=1}^{t} \prod_{k=1}^{m_{ij}} \frac{e^{\mu_{ij}(b_i)}(\mu_{ij}(b_i))^{y_{ijk}}}{y_{ijk}!} . \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} b_i^2} db_i$$

so we can not find a closed form for the ML estimator of $\boldsymbol{\beta}$ and the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can not be obtained.
In many applications including optimal design we need to have the variance of estimator of parameters instead of seeking for the exact estimator of $\boldsymbol{\beta}$. These facts lead us to use of the quasi likelihood as an approximate method (sec (3.3)) to estimate parameters; with regard to (3.6) the quasi-score function is

$$U^{(n)}(\boldsymbol{\beta}) = \mathbf{D}^T \mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}))(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \tag{4.2}$$

and the information matrix for $\hat{\boldsymbol{\beta}}$ is

$$M_{\boldsymbol{\beta},n} = \mathbf{D}^T \mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}))\mathbf{D} \tag{4.3}$$

where $\mathbf{Y}$ is the vector of responses, $\boldsymbol{\mu}(\boldsymbol{\beta}) = E(\mathbf{Y})$ is expectation of $\mathbf{Y}$ and $Var(\mathbf{Y}) = \mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$ where $\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$ is the variance function and indicates the relation between the mean and variance of $\mathbf{Y}$. $\mathbf{D} = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \mathbf{D}_1 \\ \vdots \\ \mathbf{D}_s \end{pmatrix}$ and $\mathbf{D}_i = \begin{pmatrix} \left.\begin{matrix} \boldsymbol{\mu}_{i1}\mathbf{f}^T(x_{i1}) \\ \vdots \\ \boldsymbol{\mu}_{i1}\mathbf{f}^T(x_{i1}) \end{matrix}\right\}_{m_{i1}\times p} \\ \vdots \\ \left.\begin{matrix} \boldsymbol{\mu}_{it}\mathbf{f}^T(x_{it_i}) \\ \vdots \\ \boldsymbol{\mu}_{it}\mathbf{f}^T(x_{it_i}) \end{matrix}\right\}_{m_{it_i}\times p} \end{pmatrix}_{m_i \times p}$

for $i = 1, \ldots, s$.
We first need to obtain the variance-covariance structure of $\mathbf{Y}$ to find the information matrix for $\boldsymbol{\beta}$ i.e. $M_{\boldsymbol{\beta},n}$. For sake of simplicity we will denote $M_{\boldsymbol{\beta},n}$ by $M_{\boldsymbol{\beta}}$.
As we have seen in example (2.2),

$$Var(Y_{ijk}) = Var(E(Y_{ijk} \mid \mathbf{b}_i)) + E(Var(Y_{ijk} \mid \mathbf{b}_i)) = \mu_{ij}^2(e^{\sigma^2} - 1) + \mu_{ij} = V(\mu_{ij}) \tag{4.4}$$

where $\mu_{ij} = \mu_{ij}(\boldsymbol{\beta}) = E(Y_{ijk}) = E(E(Y_{ijk} \mid b_i)) = e^{\mathbf{f}^T(x_{ij})\boldsymbol{\beta} + \frac{1}{2}\sigma^2}$ is a function of $\boldsymbol{\beta}$. For the sake of simplicity we suppress the argument $\boldsymbol{\beta}$ in $\mu_{ij}(\boldsymbol{\beta})$.

For different subjects the covariance will be zero, i.e.

$$\text{Cov}(Y_{ijk}, Y_{i'j'k'}) = 0 \text{ for all } i \neq i' \text{ and all } j, j', k, k' \tag{4.5}$$

Also,

$$\begin{aligned} \text{Cov}(Y_{ijk}, Y_{ij'k'}) &= \text{Cov}(E(Y_{ijk} \mid b_i), E(Y_{ij'k'} \mid b_i)) + E(\text{Cov}(Y_{ijk}, Y_{ij'k'} \mid b_i)) \\ &= \mu_{ij}\mu_{ij'}(e^{\sigma^2} - 1) \text{ for all } (j,k) \neq (j',k') \end{aligned} \tag{4.6}$$

Let $\mathbf{Y}_i = (\mathbf{Y}_{i1}^T, \ldots, \mathbf{Y}_{it_i}^T)^T$ be the $m_i \times 1$ vector of measurements on the ith individual, where $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijm_{ij}})^T (j = 1, \ldots, t_i)$ is the $m_{ij} \times 1$ vector of replications of individual i at the jth level of x. Then by applying some matrix algebra, $\mathbf{V}_i = Var(Y_i) = \mathbf{V}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))$ is

$$\mathbf{V}_i = \begin{pmatrix} \mathbf{V}_i^{(11)} & \mathbf{V}_i^{(12)} & \ldots & \mathbf{V}_i^{1t} \\ \mathbf{V}_i^{(21)} & \mathbf{V}_i^{(22)} & \ldots & \mathbf{V}_i^{2t} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_i^{(t1)} & \mathbf{V}_i^{(t2)} & \ldots & \mathbf{V}_i^{tt} \end{pmatrix} \tag{4.7}$$

where $\mathbf{V}_i^{(jk)} = \text{Cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{ik})$ for all j and k and it will be

$$\begin{aligned} \mathbf{V}_i^{(kk)} &= \begin{pmatrix} (e^{\sigma^2} - 1)\mu_{ik}^2 + \mu_{ik} & (e^{\sigma^2} - 1)\mu_{ik}^2 & \ldots & (e^{\sigma^2} - 1)\mu_{ik}^2 \\ (e^{\sigma^2} - 1)\mu_{ik}^2 & (e^{\sigma^2} - 1)\mu_{ik}^2 + \mu_{ik} & \ldots & (e^{\sigma^2} - 1)\mu_{ik}^2 \\ \vdots & \vdots & \ddots & \vdots \\ (e^{\sigma^2} - 1)\mu_{ik}^2 & (e^{\sigma^2} - 1)\mu_{ik}^2 & \ldots & (e^{\sigma^2} - 1)\mu_{ik}^2 + \mu_{ik} \end{pmatrix}_{m_{ik} \times m_{ik}} \\ &= \mu_{ik} I_{m_{ik}} + (e^{\sigma^2} - 1)\mu_{ik}^2 J_{m_{ik} \times m_{ik}} \end{aligned} \tag{4.8}$$

and

$$\mathbf{V}_i^{(jk)} = \begin{pmatrix} (e^{\sigma^2} - 1)\mu_{ij}\mu_{ik} & \ldots & (e^{\sigma^2} - 1)\mu_{ij}\mu_{ik} \\ \vdots & \ddots & \vdots \\ (e^{\sigma^2} - 1)\mu_{ij}\mu_{ik} & \ldots & (e^{\sigma^2} - 1)\mu_{ij}\mu_{ik} \end{pmatrix}_{m_{ij} \times m_{ik}} = (e^{\sigma^2} - 1)\mu_{ij}\mu_{ik} J_{m_{ij} \times m_{ik}} \tag{4.9}$$

where $j \neq k$ and $J_{m \times n}$ is a matrix of $m \times n$ order in which all entries are equal to 1.

$$\begin{aligned} \mathbf{V}_i = Var(\mathbf{Y}_i) &= \begin{pmatrix} \mu_{i1} I_{m_{i1}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mu_{i2} I_{m_{i2}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mu_{it} I_{m_{it_i}} \end{pmatrix} \\ &\quad + (e^{\sigma^2} - 1) \begin{pmatrix} \mu_{i1}\mathbf{1}_{m_{i1}} \\ \mu_{i2}\mathbf{1}_{m_{i2}} \\ \vdots \\ \mu_{it_i}\mathbf{1}_{m_{it_i}} \end{pmatrix} \begin{pmatrix} \mu_{i1}\mathbf{1}_{m_{i1}}^T & \mu_{i2}\mathbf{1}_{m_{i2}}^T & \cdots & \mu_{it_i}\mathbf{1}_{m_{it_i}}^T \end{pmatrix} \\ &= \dot{\mathbf{A}}_i + \dot{\mathbf{a}}_i \dot{\mathbf{a}}_i^T \end{aligned} \tag{4.10}$$

where, $\dot{\mathbf{a}}_i^T = \sqrt{e^{\sigma^2}-1}\begin{pmatrix} \mu_{i1}\mathbf{1}_{m_{i1}}^T & \cdots & \mu_{it_i}\mathbf{1}_{m_{it_i}}^T \end{pmatrix}$ and $\dot{\mathbf{A}}_i = diag\{\mu_{i1}I_{m_{i1}}, \cdots, \mu_{it_i}I_{m_{it_i}}\}$. Here and throughout I_ν denotes the $\nu \times \nu$ identity matrix and $\mathbf{1}_\nu$ is an $\nu \times 1$ vector with all entries equal to 1.
We suppose that $\mathbf{Y}^T = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_s^T)$ is the vector of the whole observation. Independence of different individuals leads to,

$$\mathbf{V} = Var(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V}_s \end{pmatrix} = \mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$$

according to (4.3) the quasi-information matrix of $\boldsymbol{\beta}$, $M_{\boldsymbol{\beta}}$, is

$$M_{\boldsymbol{\beta}} = \sum_{i=1}^{s} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i = \sum_{i=1}^{s} M_{\boldsymbol{\beta}}^i \tag{4.11}$$

where $M_{\boldsymbol{\beta}}^i = \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i$ is the individual quasi-information matrix for $\boldsymbol{\beta}$ associated with the observations $\mathbf{Y}_i$ of a single individual i. By replacing $\mathbf{V}_i$ (Eq. (4.10)) in to the expression of $M_{\boldsymbol{\beta}}^i$ we can represent $M_{\boldsymbol{\beta}}^i$ as following

$$M_{\boldsymbol{\beta}}^i = \mathbf{D}_i^T (\dot{\mathbf{A}}_i + \dot{\mathbf{a}}_i \dot{\mathbf{a}}_i^T)^{-1} \mathbf{D}_i \tag{4.12}$$

From (4.12), the information matrix is strongly dependent on the parameters.
If we define $\dot{\mathbf{F}}_i = \begin{pmatrix} \left.\begin{matrix} \mathbf{f}^T(x_{i1}) \\ \vdots \\ \mathbf{f}^T(x_{i1}) \end{matrix}\right\}_{m_{i1}\times p} \\ \vdots \\ \left.\begin{matrix} \mathbf{f}^T(x_{it_i}) \\ \vdots \\ \mathbf{f}^T(x_{it_i}) \end{matrix}\right\}_{m_{it_i}\times p} \end{pmatrix}_{m_i \times p}$ then $\mathbf{D}_i = \dot{\mathbf{A}}_i \dot{\mathbf{F}}_i$ and consequently we obtain the following lemma .

Lemma 4.1.1. *The individual information matrix* (4.12) *can be represented as*

$$M_{\boldsymbol{\beta}}^i = \dot{\mathbf{F}}_i^T (\dot{\mathbf{A}}_i^{-1} + (e^{\sigma^2}-1)\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T)^{-1} \dot{\mathbf{F}}_i \tag{4.13}$$

Proof.

$$M_{\boldsymbol{\beta}}^i = \mathbf{D}_i^T (\dot{\mathbf{A}}_i + \dot{\mathbf{a}}_i \dot{\mathbf{a}}_i^T)^{-1} \mathbf{D}_i = \mathbf{D}_i^T \dot{\mathbf{A}}_i^{-1} (\dot{\mathbf{A}}_i^{-1} + \dot{\mathbf{A}}_i^{-1} \dot{\mathbf{a}}_i \dot{\mathbf{a}}_i^T \dot{\mathbf{A}}_i^{-1})^{-1} \dot{\mathbf{A}}_i^{-1} \mathbf{D}_i$$

Since $\mathbf{D}_i^T \dot{\mathbf{A}}_i^{-1} = \dot{\mathbf{F}}_i^T$ and $\dot{\mathbf{A}}_i^{-1} \dot{\mathbf{a}}_i = \sqrt{e^{\sigma^2}-1}\mathbf{1}_{m_i}$, the result follows. □

Define $\mathbf{F}_i = \begin{pmatrix} \mathbf{f}^T(x_{i1}) \\ \vdots \\ \mathbf{f}^T(x_{it_i}) \end{pmatrix}_{t_i \times p}$ the row individual design matrix neglecting the number of replications. Then the information matrix can be simplified.

Lemma 4.1.2. *(Niaparast 2009) The individual information matrix can be represented as*

$$M^i_{\boldsymbol{\beta}} = \mathbf{F}_i^T(\mathbf{A}_i^{-1} + (e^{\sigma^2}-1)\mathbf{1}_{t_i}\mathbf{1}_{t_i}^T)^{-1}\mathbf{F}_i = \mathbf{F}_i^T(\mathbf{A}_i - \frac{(e^{\sigma^2}-1)\mathbf{A}_i\mathbf{1}_{t_i}\mathbf{1}_{t_i}^T\mathbf{A}_i}{1+(e^{\sigma^2}-1)\mathbf{1}_{t_i}^T\mathbf{A}_i\mathbf{1}_{t_i}})\mathbf{F}_i \quad (4.14)$$

with $\mathbf{A}_i = \begin{pmatrix} m_{i1}\mu_{i1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & m_{it_i}\mu_{it_i} \end{pmatrix}$

Proof. Because of (Schott (1997), Corollary 1.7.2)

$$M^i_{\boldsymbol{\beta}} = \dot{\mathbf{F}}_i^T(\dot{\mathbf{A}}_i^{-1} + (e^{\sigma^2}-1)\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T)^{-1}\dot{\mathbf{F}}_i = \dot{\mathbf{F}}_i^T(\dot{\mathbf{A}}_i - \frac{(e^{\sigma^2}-1)\dot{\mathbf{A}}_i\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\dot{\mathbf{A}}_i}{1+(e^{\sigma^2}-1)\mathbf{1}_{m_i}^T\dot{\mathbf{A}}_i\mathbf{1}_{m_i}})\dot{\mathbf{F}}_i$$

Since $\dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i\dot{\mathbf{F}}_i = \mathbf{F}_i^T\mathbf{A}_i\mathbf{F}_i$, $\dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i\mathbf{1}_{m_i} = \mathbf{F}_i^T\mathbf{A}_i\mathbf{1}_{t_i}$ and $\mathbf{1}_{m_i}^T\dot{\mathbf{A}}_i\mathbf{1}_{m_i} = \sum_{j=1}^{t_i} m_{ij}\mu_{ij} = \mathbf{1}_{t_i}^T\mathbf{A}_i\mathbf{1}_{t_i}$, we obtain

$$M^i_{\boldsymbol{\beta}} = \mathbf{F}_i^T\mathbf{A}_i\mathbf{F}_i - \frac{(e^{\sigma^2}-1)(\mathbf{F}_i^T\mathbf{A}_i\mathbf{1}_{t_i})(\mathbf{1}_{t_i}^T\mathbf{A}_i\mathbf{F}_i)}{1+(e^{\sigma^2}-1)\mathbf{1}_{t_i}^T\mathbf{A}_i\mathbf{1}_{t_i}}$$

and the representation follows □

Lemma 4.1.3. *The above representation of information matrix can be simplified as*

$$M^i_{\boldsymbol{\beta}} = ((\mathbf{F}_i^T\mathbf{A}_i\mathbf{F}_i)^{-1} + \mathbf{U})^{-1} \quad (4.15)$$

where $\mathbf{U} = \begin{pmatrix} e^{\sigma^2}-1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$ *is a matrix of order* $p \times p$.

Proof. Since $(\sqrt{e^{\sigma^2}-1})\mathbf{1}_{t_i} = \mathbf{F}_i\mathbf{u}_p$ with $\mathbf{u}_p^T = \sqrt{e^{\sigma^2}-1}\begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}$, we have

$$M^i_{\boldsymbol{\beta}} = \mathbf{F}_i^T(\mathbf{A}_i^{-1} + (\mathbf{F}_i\mathbf{u}_p)(\mathbf{F}_i\mathbf{u}_p)^T)^{-1}\mathbf{F}_i = \mathbf{F}_i^T(\mathbf{A}_i^{-1} + \mathbf{F}_i^T\mathbf{U}\mathbf{F}_i)^{-1}\mathbf{F}_i$$

where $\mathbf{U} = \mathbf{u}_p\mathbf{u}_p^T$. With regard to lemma 1 in Schmelter (2007) the claim follows. □

Eq.(4.14) can be represented as,

$$M^i_{\boldsymbol{\beta}} = e^{\frac{1}{2}\sigma^2}(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{F}_i - \frac{e^{\frac{1}{2}\sigma^2}(e^{\sigma^2}-1)}{1+e^{\frac{1}{2}\sigma^2}(e^{\sigma^2}-1)\sum_{j=1}^{t_i} m_{ij}\check{\mu}_{ij}}(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})^T)$$

where $\check{\mu}_{ij} = e^{-\frac{1}{2}\sigma^2}\mu_{ij}$, $\check{\mathbf{A}}_i = e^{-\frac{1}{2}\sigma^2}\mathbf{A}_i = diag\{m_{i1}\check{\mu}_{i1}, \ldots m_{it_i}\check{\mu}_{it_i}\}$ and, hence, $\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{F}_i$ is the information matrix for the model without random intercept, i.e.

$$\check{Y}_{ijk} \overset{ind}{\sim} P(\check{\mu}_{ij}) \text{ with } \check{\mu}_{ij} = \exp(\mathbf{f}^T(x_{ij})\boldsymbol{\beta})) \qquad \begin{cases} i = 1, \ldots, s \\ j = 1, \ldots, t_i \\ k = 1, \ldots, m_{ij} \end{cases}$$

If $\check{\mu}_i$ is the mean of the individual's means for individual i in different points of experimental setting, i.e., $\check{\mu}_i = \frac{1}{m_i}\sum_{j=1}^{t_i} m_{ij}\check{\mu}_{ij}$, then

$$\begin{aligned} M^i_{\boldsymbol{\beta}} &= e^{\frac{1}{2}\sigma^2}(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{F}_i - c(\sigma^2, m_i, \check{\mu}_i)(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})^T) \\ &= e^{\frac{1}{2}\sigma^2}((1 - c(\sigma^2, m_i, \check{\mu}_i))\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{F}_i + c(\sigma^2, m_i, \check{\mu}_i)(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{F}_i - (\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})^T) \end{aligned} \tag{4.16}$$

where $c(\sigma^2, m_i, \check{\mu}_i) = \frac{e^{\frac{1}{2}\sigma^2}(e^{\sigma^2}-1)}{1+e^{\frac{1}{2}\sigma^2}(e^{\sigma^2}-1)m_i\check{\mu}_i}$. Using some simple algebra it is easy to see that $c(\sigma^2, m_i, \check{\mu}_i)$ is an increasing function of σ^2 for all m_i and $\check{\mu}_i$ and it takes a number in $[0, \frac{1}{m_i\check{\mu}_i})$. $\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{F}_i - (\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})(\mathbf{F}_i^T \check{\mathbf{A}}_i \mathbf{1}_{t_i})^T$ is the information matrix for the parameter vector $(\beta_1, \ldots, \beta_{p-1})$ in the model with individual fixed effects. As we have indicated, (4.16) concludes a linear combination of two information matrix, the fact that is in contrary to the result in Schwabe and Schmelter (2008) which is considered a convex combination of these two information matrices.

Now we consider two special cases of the Poisson regression model with random intercept which are called simple Poisson regression with random intercept and quadratic Poisson regression model with random intercept. In the first model we suppose that $\mathbf{f}(x_{ij}) = (1, x_{ij})$,i.e.

$$Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \text{ where } \mu_{ij}(b_i)) = \exp(b_i + \beta_0 + \beta_1 x_{ij}) \tag{4.17}$$

In the spaghetti plot of Figure 4.1, we show an aspect of conditional individual mean response lines, $\exp(\beta_0 + b_i x)$, and marginal mean response line, $\exp(\beta_0 + \beta_1 x + \frac{1}{2}\sigma^2)$, to get a feeling for the patterns.

It is easy to see that $\mathbf{F}_i$ will be $\begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{it_i} \end{pmatrix}$ in this model. Thus after some matrix

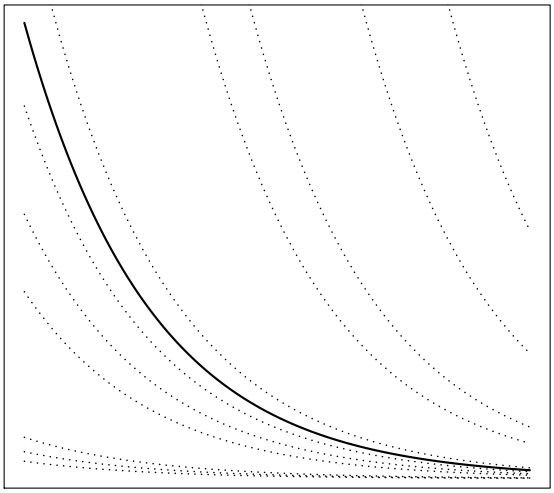


Figure 4.1: Conditional (dashed lines) and population mean (solid line) for the simple Poisson regression model with random intercept.

algebra we have,

$$
\begin{aligned}
M^i_{\boldsymbol{\beta}} &= \begin{pmatrix} \sum_{j=1}^{t_i} m_{ij}\mu_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij} \\ \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2 \end{pmatrix} \\
&- \frac{e^{\sigma^2}-1}{1+(e^{\sigma^2}-1)\sum_{j=1}^{t_i} m_{ij}\mu_{ij}} \begin{pmatrix} (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})^2 & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}) \\ (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}) & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij})^2 \end{pmatrix}
\end{aligned}
\tag{4.18}
$$

In the second model, the quadratic Poisson regression model with random intercept, we have $\mathbf{f}(x_{ij}) = (1\, x_{ij}\, x_{ij}^2)$, an

$$
Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \text{ where } \mu_{ij}(b_i)) = \exp(b_i + \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2) \tag{4.19}
$$

In the following plot (Figure 4.2) a general pattern of the behavior of the conditional and unconditional means of the response is given.

The quasi-information matrix for $\boldsymbol{\beta}$, based on individual i, can be represented as

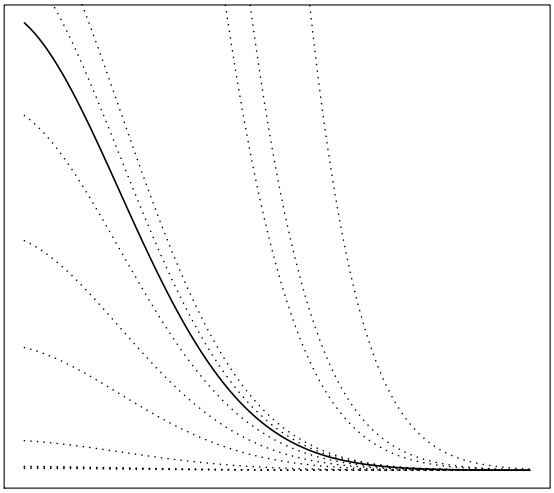


Figure 4.2: Conditional (dashed lines) and population mean (solid line) for the quadratic Poisson regression model with random intercept.

$$
\begin{aligned}
M^i_{\boldsymbol{\beta}} = &\begin{pmatrix} 1 & x_{i1} & x_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{it_i} & x_{it_i}^2 \end{pmatrix}^T \begin{pmatrix} m_{i1}\mu_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & m_{it_i}\mu_{it_i} \end{pmatrix} \begin{pmatrix} 1 & x_{i1} & x_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{it_i} & x_{it_i}^2 \end{pmatrix} \\
&- \frac{e^{\sigma^2}-1}{1+(e^{\sigma^2}-1)\sum_{j=1}^{t_i} m_{ij}\mu_{ij}} \begin{pmatrix} \sum_{j=1}^{t_i} m_{ij}\mu_{ij} \\ \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij} \\ \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2 \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{t_i} m_{ij}\mu_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2 \end{pmatrix} \\
= &\begin{pmatrix} \sum_{j=1}^{t_i} m_{ij}\mu_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2 \\ \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij} & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2 & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^3 \\ \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2 & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^3 & \sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^4 \end{pmatrix} - \frac{e^{\sigma^2}-1}{1+(e^{\sigma^2}-1)\sum_{j=1}^{t_i} m_{ij}\mu_{ij}} \\
&\begin{pmatrix} (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})^2 & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}) & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2) \\ (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}) & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij})^2 & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2) \\ (\sum_{j=1}^{t_i} m_{ij}\mu_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2) & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij})(\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2) & (\sum_{j=1}^{t_i} m_{ij}\mu_{ij}x_{ij}^2)^2 \end{pmatrix}
\end{aligned}
\tag{4.20}
$$

4.2 Poisson Regression Model with Random Slope

An assumption that might sometimes be in challenge is that the effect of the explanatory variable is constant across the different subjects. On the contrary of a random intercept model, random slope model allows the explanatory variable to have a different effect for each individual.
The Poisson regression model with random slope can be written as,

$$Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \begin{cases} i = 1, \ldots, s & \sum_{j=1}^{t_i} m_{ij} = m_i \\ j = 1, \ldots, t_i & \\ k = 1, \ldots, m_{ij} & n = \sum_{i=1}^{s} m_i \end{cases} \tag{4.21}$$

where $\mu_{ij}(b_i) = \exp(\beta_0 + b_i x_{ij})$, and b_i is normally distributed with mean β_1 and variance σ^2.
Similar to the model with random intercept, we have

$$\mu_{ij} = E(Y_{ijk}) = e^{\beta_0 + \beta_1 x_{ij} + \frac{1}{2}\sigma^2 x_{ij}^2} \text{ for all } k \tag{4.22}$$

A good comparison of the conditional mean and population mean is prepared in the following plot (Figure 4.2)

$$Var(Y_{ijk}) = \mu_{ij}^2 (e^{\sigma^2 x_{ij}^2} - 1) + \mu_{ij} \text{ for all } k \tag{4.23}$$

$$\text{Cov}(Y_{ijk}, Y_{ij'k'}) = \mu_{ij}\mu_{ij'}(e^{\sigma^2 x_{ij} x_{ij'}} - 1) \text{ for all } (j,k) \neq (j',k') \tag{4.24}$$

For different individuals ,

$$\text{Cov}(Y_{ijk}, Y_{i'j'k'}) = 0 \text{ for all } i \neq i' \text{ and all } j, j', k, k' \tag{4.25}$$

Consider the same notation as in the previous section , then we have

$$\mathbf{V}_i^{(kk)} = \mu_{ik} I_{m_{ik}} + (e^{\sigma^2 x_{ik}^2} - 1)\mu_{ik}^2 J_{m_{ik} \times m_{ik}}$$

and also,

$$\mathbf{V}_i^{(jk)} = (e^{\sigma^2 x_{ij} x_{ik}} - 1)\mu_{ij}\mu_{ik} J_{m_{ij} \times m_{ik}} \text{ for } j \neq k$$

where $\mathbf{V}_i^{(kk)}$ and $\mathbf{V}_i^{(jk)}$ are the same notation as in Model with random intercept. Therefore the variance-covariance matrix of the $\mathbf{Y}_i$, the vector of the ith individual observations, is

$$\mathbf{V}_i \;=\; Var(\mathbf{Y}_i) = \dot{\mathbf{A}}_i + \dot{\mathbf{B}}_{ii} \tag{4.26}$$

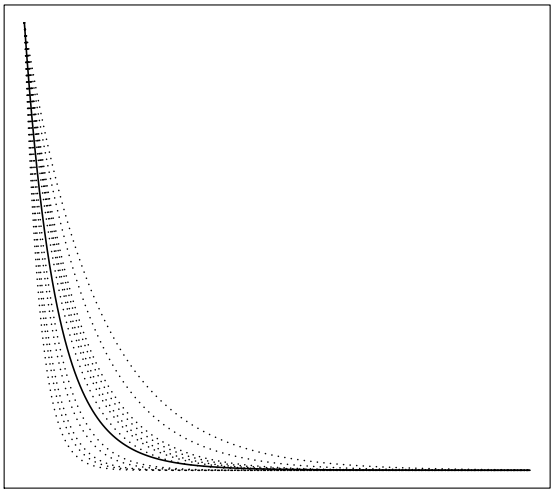


Figure 4.3: Conditional (dashed lines) and population mean (solid line) for the Poisson regression model with random slope.

where $\dot{\mathbf{A}}_i$ is the same as the definition of $\dot{\mathbf{A}}_i$ in the model with random intercept (4.10), i.e.,

$$\dot{\mathbf{A}}_i = diag\{\mu_{i1} I_{m_{i1}}, \cdots, \mu_{it_i} I_{m_{it_i}}\} \tag{4.27}$$

and $\dot{\mathbf{B}}_{ii} = (\mathbf{B}_{(ii)jk})_{j,k=1}^{m_i}$ is a $m_i \times m_i$ block matrix, where $\mathbf{B}_{(ii)jk} = (e^{\sigma^2 x_{ij} x_{ik}} - 1)\mu_{ij}\mu_{ik} J_{m_{ij} \times m_{ik}}$.

Under known σ^2, the quasi-information matrix of $\boldsymbol{\beta}$, $M^i_{\boldsymbol{\beta}}$, can be obtained by replacing (4.27) in (4.3),

$$M^i_{\boldsymbol{\beta}} = \mathbf{D}_i^T (\dot{\mathbf{A}}_i + \dot{\mathbf{B}}_{ii})^{-1} \mathbf{D}_i \tag{4.28}$$

Lemma 4.2.1. *Suppose that* $\mathbf{H} = \begin{pmatrix} e^{\sigma^2 x_1^2} - 1 & \dots & e^{\sigma^2 x_1 x_t} - 1 \\ \vdots & \ddots & \vdots \\ e^{\sigma^2 x_1 x_t} - 1 & \dots & e^{\sigma^2 x_t^2} - 1 \end{pmatrix}$, *then* H *is positive semi-definite.*

Proof. Consider the following model

$$Y_i \mid b \sim P(\mu_i(b)), i = 1, \dots, t \text{ and } \mu_i(b) = e^{\beta_0 + b x_i}$$

Let $\widetilde{\widetilde{\mathbf{Y}}} = (Y_1, \ldots, Y_t)^T$ be the $t \times 1$ vector of observations, we have

$$Var(E(\widetilde{\widetilde{\mathbf{Y}}} \mid b)) = \begin{pmatrix} (e^{\sigma^2 x_1^2} - 1)\mu_1^2 & \ldots & (e^{\sigma^2 x_1 x_t} - 1)\mu_1\mu_t \\ \vdots & \ddots & \vdots \\ (e^{\sigma^2 x_1 x_t} - 1)\mu_1\mu_t & \ldots & (e^{\sigma^2 x_t^2} - 1)\mu_t^2 \end{pmatrix}$$

where $\mu_j = e^{\beta_0 + \beta_1 x_j + \frac{1}{2}\sigma^2 x_j^2}$. We define $\mathbf{Q} = diag\{\mu_1, \ldots \mu_t\}$ as a diagonal matrix, thus

$$\mathbf{H} = \mathbf{Q}^{-1} Var(E(\widetilde{\widetilde{\mathbf{Y}}} \mid b))\mathbf{Q}^{-1}$$

Since $Var(E(\widetilde{\widetilde{\mathbf{Y}}} \mid b))$ is positive semi definite (PSD) matrices and $\mathbf{Q} = \mathbf{Q}^T$, $\mathbf{H}$ is PSD. □

The following lemma indicates a simplified version of the quasi-information matrix for the model with random slope.

Lemma 4.2.2. *The quasi-information matrix $M^i_{\boldsymbol{\beta}}$ in (4.28) can be represented as*

$$M^i_{\boldsymbol{\beta}} = \mathbf{F}_i^T(\mathbf{A}_i^{-1} + \mathbf{B}_{ii})^{-1}\mathbf{F}_i \tag{4.29}$$

where $A_i = diag\{m_{i1}\mu_{i1}, m_{i2}\mu_{i2}, \ldots, m_{it_i}\mu_{it_i}\}$ is the same as A_i in the model with random intercept, $\mathbf{F}_i = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_{i1} & x_{i2} & \ldots & x_{it_i} \end{pmatrix}^T$ is the design matrix of the model, and

$$\mathbf{B}_{ii} = \begin{pmatrix} e^{\sigma^2 x_{i1}^2} - 1 & e^{\sigma^2 x_{i1}x_{i2}} - 1 & \ldots & e^{\sigma^2 x_{i1}x_{it_i}} - 1 \\ e^{\sigma^2 x_{i1}x_{i2}} - 1 & e^{\sigma^2 x_{i2}^2} - 1 & \ldots & e^{\sigma^2 x_{i2}x_{it_i}} - 1 \\ \vdots & \vdots & \ddots & \vdots \\ e^{\sigma^2 x_{it_i}x_{i1}} - 1 & e^{\sigma^2 x_{it_i}x_{i2}} - 1 & \ldots & e^{\sigma^2 x_{it_i}^2} - 1 \end{pmatrix} \tag{4.30}$$

Proof. Consider $M^i_{\boldsymbol{\beta}}$ in (4.28) it is easy to see that under known σ^2, $\mathbf{D}_i^T = \dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i$, then

$$\mathbf{D}_i^T(\dot{\mathbf{A}}_i + \dot{\mathbf{B}}_{ii})^{-1}\mathbf{D}_i = \dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i(\dot{\mathbf{A}}_i + \dot{\mathbf{B}}_{ii})^{-1}\dot{\mathbf{A}}_i\dot{\mathbf{F}}_i = \dot{\mathbf{F}}_i^T(\dot{\mathbf{A}}_i^{-1} + \dot{\mathbf{A}}_i^{-1}\dot{\mathbf{B}}_{ii}\dot{\mathbf{A}}_i^{-1})^{-1}\dot{\mathbf{F}}_i$$

this equality holds because $\dot{\mathbf{A}}_i$ is a diagonal matrix with non-zero diagonal entries and, hence, it is symmetric and invertible.

$$\begin{aligned} &\dot{\mathbf{A}}_i^{-1}\dot{\mathbf{B}}_{ii}\dot{\mathbf{A}}_i^{-1} \\ &= \begin{pmatrix} (e^{\sigma^2 x_{i1}^2} - 1)J_{m_{i1}\times m_{i1}} & (e^{\sigma^2 x_{i1}x_{i2}} - 1)J_{m_{i1}\times m_{i2}} & \ldots & (e^{\sigma^2 x_{i1}x_{it_i}} - 1)J_{m_{i1}\times m_{it_i}} \\ (e^{\sigma^2 x_{i1}x_{i2}} - 1)J_{m_{i2}\times m_{i1}} & (e^{\sigma^2 x_{i2}^2} - 1)J_{m_{i2}\times m_{i2}} & \ldots & (e^{\sigma^2 x_{i2}x_{it_i}} - 1)J_{m_{i2}\times m_{it_i}} \\ \vdots & \vdots & \ddots & \vdots \\ (e^{\sigma^2 x_{it_i}x_{i1}} - 1)J_{m_{it_i}\times m_{i1}} & (e^{\sigma^2 x_{it_i}x_{i2}} - 1)J_{m_{it_i}\times m_{i2}} & \ldots & (e^{\sigma^2 x_{it_i}^2} - 1)J_{m_{it_i}\times m_{it_i}} \end{pmatrix} \\ &= \mathbf{C}_i\mathbf{B}_{ii}\mathbf{C}_i^T \end{aligned}$$

where $\mathbf{C}_i = diag\{\mathbf{1}_{m_{i1}}, \mathbf{1}_{m_{i2}}, \ldots, \mathbf{1}_{m_{it_i}}\}$ which is a block diagonal matrix of order $m_i \times t_i$. using 4.2.1, Under conditions of unequal and non-zero measurements for x_{ij}, regarding to the Lemma 4.2.1 $\mathbf{B}_{ii}$ is invertible. Applying the theorem 1.7 of Schott (1997), we have

$$\begin{aligned} M^i_{\boldsymbol{\beta}} &= \dot{\mathbf{F}}_i^T(\dot{\mathbf{A}}_i^{-1} + \mathbf{C}_i\mathbf{B}_{ii}\mathbf{C}_i^T)^{-1}\dot{\mathbf{F}}_i \\ &= \dot{\mathbf{F}}_i^T[\dot{\mathbf{A}}_i - \dot{\mathbf{A}}_i\mathbf{C}_i(\mathbf{B}_{ii}^{-1} + \mathbf{C}_i^T\dot{\mathbf{A}}_i\mathbf{C}_i)^{-1}\mathbf{C}_i^T\dot{\mathbf{A}}_i]\dot{\mathbf{F}}_i \\ &= \dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i\dot{\mathbf{F}}_i - \dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i\mathbf{C}_i(\mathbf{B}_{ii}^{-1} + \mathbf{C}_i^T\dot{\mathbf{A}}_i\mathbf{C}_i)^{-1}\mathbf{C}_i^T\dot{\mathbf{A}}_i\dot{\mathbf{F}}_i \end{aligned}$$

Since $\dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i\dot{\mathbf{F}}_i = \mathbf{F}_i^T\mathbf{A}_i\mathbf{F}_i$, $\dot{\mathbf{F}}_i^T\dot{\mathbf{A}}_i\mathbf{C}_i = \mathbf{F}_i^T\mathbf{A}_i$ and $\mathbf{C}_i^T\dot{\mathbf{A}}_i\mathbf{C}_i = \mathbf{A}_i$, thus

$$\begin{aligned} & M^i_{\boldsymbol{\beta}} = \mathbf{F}_i^T[\mathbf{A}_i - \mathbf{A}_i(\mathbf{A}_i + \mathbf{B}_{ii}^{-1})^{-1}\mathbf{A}_i]^{-1}\mathbf{F}_i \\ \Rightarrow \quad & M^i_{\boldsymbol{\beta}} = \mathbf{F}_i^T(\mathbf{A}_i^{-1} + \mathbf{B}_{ii})^{-1}\mathbf{F}_i \end{aligned} \tag{4.31}$$

where the last expression holds because of (Schott(1997), Corollary 1.7.1) □

Remark 4.1. *In the above proof an essential assumption was $x_{ij} \neq 0$. In case of $x_{ij} = 0$ which will be typically occurred to optimal designs study, we replace x_{ij} by $x_{ij} + \delta$ and hence $\mathbf{B}_{ii}$ by $\mathbf{B}^{\delta}_{ii}$. Since $\lim_{\delta \to 0} \mathbf{B}^{\delta}_{ii} = \mathbf{B}_{ii}$ we can follow the same line in the proof as the above proof and thus result follows.*

As we will see in next chapters, the analysis of this model based on the quasi-information matrix differs from the model with random intercept. This discrepancy comes from the dependency of $\mathbf{B}_{ii}$ on the support points of the experimental setting.

5 Optimal Designs

5.1 Introduction

This chapter presets a compact review of the topics of optimal designs of experiments and locally optimal design of experiments. A short introduction to the optimal design terminology, which we need to introduce the optimal design theory and its applications, starts the review. Despite wide theoretical work on optimal designs for linear models (see e.g. Fedorov (1972), Silvey (1980) and many others) and optimal linear mixed models(see e.g. Liski et al. 2002, Fedorov and Hackl 1997, Schmelter 2007 and others) there are only a few results on the case of Generalized Linear Mixed Models. With regard to the structure of these models (see sec. 2.4), deriving analytical result is difficult.
In the next section, we introduce some arbitrary concepts and definitions of optimal design. After that, in the third and the fourth sections, a review of convex design theory for two popular models including linear models and linear mixed models will be prepared. In the last two sections some new results on the convex design theory are presented for two special models which are the Poisson regression model with random intercept and the Poisson regression model with random slope separately.

5.2 Basic Concepts

Consider the linear model,

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

As we indicated in (2.6) and (2.8) the variance-covariance matrices of $\hat{\boldsymbol{\beta}}$ and $\widehat{\gamma(\boldsymbol{\beta})}$ are

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{F}^T\mathbf{F})^{-1} \\ Var(\gamma(\hat{\boldsymbol{\beta}})) &= \sigma^2 L_\gamma(\mathbf{F}^T\mathbf{F})^{-1}L_\gamma^T \end{aligned}$$

respectively. They depend heavily on the ***experimental setting*** $\{x_1, \ldots, x_n\}$ via $\mathbf{F}$. The settings x_i may be chosen from $\tau = [a, b]$ which is called the ***experimental domain***. Most literature on optimal designs focuses on the specific symmetric domain $[-1, 1]$ or on the specific asymmetric domain $[0, 1]$. It is easy to see that we can replace $\tau = [a, b]$ by a symmetric standardized experimental domain $\tau = [-1, 1]$(e.g. Atkinson et al. (2007) page 18) or an asymmetric standardized experimental domain $\tau = [0, 1]$ (e.g. Liski et al. (2002) page 38), although the relations between optimality study in different domains are not clear generally.

Note that if we have replications in the experimental settings,i.e. $\{n_1, \ldots, n_m\}$ is the number of replications corresponding to $\{x_1, \ldots, x_m\}$ with the total number of observations $n = \sum_{i=1}^{m} n_i$ and $x_i \in \tau$, then we can represent the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\widehat{\gamma(\boldsymbol{\beta})}$

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= \frac{1}{n}\sigma^2(\mathbf{F}^T\mathbf{W}\mathbf{F})^{-1} \\ Var(\gamma(\hat{\boldsymbol{\beta}})) &= \frac{1}{n}\sigma^2 L_\gamma(\mathbf{F}^T\mathbf{W}\mathbf{F})^{-}L_\gamma^T \end{aligned} \tag{5.1}$$

respectively. In the above expressions $\mathbf{W}$ is a diagonal matrix with entries $\frac{n_i}{n}$, $i = 1, \ldots, m$, i.e. $\mathbf{W} = \begin{pmatrix} \frac{n_1}{n} & & \\ & \ddots & \\ & & \frac{n_m}{n} \end{pmatrix}$.

Remark 5.1. *Since the analytical inference based on $Var(\hat{\boldsymbol{\beta}})$ is the same as analytical inference based on $Var(\gamma(\hat{\boldsymbol{\beta}}))$, in continue we consider only $\hat{\boldsymbol{\beta}}$ and $Var(\hat{\boldsymbol{\beta}})$ except in special cases that we say.*

Because of the dependency on experimental settings, the experimenters might think how they can plan an experiment to make the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ as small as possible or alternatively to make its inverse as large as possible in some sense. The answer of this question is the main subject of ***optimal design***.
Consider an ***exact design*** d_n as following

$$d_n = \begin{Bmatrix} x_1 & \ldots & x_m \\ n_1 & \ldots & n_m \end{Bmatrix} \tag{5.2}$$

then,

$$Var_d(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sigma^2(\mathbf{F}^T\mathbf{W}\mathbf{F})^{-1} \tag{5.3}$$

We subscribe Var with d to lay emphasis on the design. For the sake of simplicity in writing we suppress the index of d_n.
Each exact design d_n can be considered as a discrete design measure over τ, ξ_n. If the design has trails at m distinct points in τ, we can write,

$$\xi_n = \begin{Bmatrix} x_1 & \ldots & x_m \\ p(x_1) & \ldots & p(x_m) \end{Bmatrix} \tag{5.4}$$

where $p(x_i)$ is the assigned relative frequencies for the point x_i $(i = 1, \ldots, m)$. Note that $np(x_i) = n_i$ and $\sum_{i=1}^{m} p(x_i) = 1$.

If σ^2 is known, then $Var_d(\hat{\boldsymbol{\beta}})$ depends on the x_i and the n_i. It seems to be reasonable that we try to minimize $\frac{1}{n}(\mathbf{F}^T\mathbf{W}\mathbf{F})^{-1}$ or equivalently to maximize $n(\mathbf{F}^T\mathbf{W}\mathbf{F})$ or $(\mathbf{F}^T\mathbf{W}\mathbf{F})$ with $\mathbf{W} = diag\{p_1, \ldots, p_m\}$ and $p_i = p(x_i)$, $i = 1, \ldots, m$. $n(\mathbf{F}^T\mathbf{W}\mathbf{F})$ and $(\mathbf{F}^T\mathbf{W}\mathbf{F})$ are well-known as the information matrix and the normalized information matrix respectively. Following Fedorov and Hackl (1997) we denote the information matrix and normalized information matrix with $\underline{M}_{\boldsymbol{\beta}}(\xi_n)$ and $M_{\boldsymbol{\beta}}(\xi_n)$ respectively.

The design $\xi_n^{(1)}$ dominates the design $\xi_n^{(2)}$ in the ***Loewner ordering sense***, denoted by $\xi_n^{(1)} \succ \xi_n^{(2)}$, if $M_{\boldsymbol{\beta}}(\xi_n^{(1)}) - M_{\boldsymbol{\beta}}(\xi_n^{(2)})$ is a non-negative definite matrix. We also denote $M_{\boldsymbol{\beta}}(\xi_n^{(1)}) \geq M_{\boldsymbol{\beta}}(\xi_n^{(2)})$ or $M_{\boldsymbol{\beta}}(\xi_n^{(1)}) - M_{\boldsymbol{\beta}}(\xi_n^{(2)}) \geq 0$ when $M_{\boldsymbol{\beta}}(\xi_n^{(1)}) - M_{\boldsymbol{\beta}}(\xi_n^{(2)})$ is non-negative definite.

The matrix inversion A^{-1} is an antitonic mapping from the open cone of positive definite matrices to itself with respect to the Loewner ordering sense , that means if $A \geq B$ then $A^{-1} \leq B^{-1}$(see Pukelsheim (1993), page 13).

Since the information matrix for the parameter $\boldsymbol{\beta}$, in general, equals (or approximately equals) the inverse of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, this property leads us to an equivalence in the comparison of variance-covariance matrices of two different designs and the comparison of their information matrices in the sense of Loewner ordering, i.e.

$$Var_{\xi_n^{(1)}}(\hat{\boldsymbol{\beta}}) \leq Var_{\xi_n^{(2)}}(\hat{\boldsymbol{\beta}}) \Leftrightarrow M_{\boldsymbol{\beta}}(\xi_n^{(1)}) \geq M_{\boldsymbol{\beta}}(\xi_n^{(2)}) \Leftrightarrow \xi_n^{(1)} \succ \xi_n^{(2)} \tag{5.5}$$

An immediate question arises: Can we find an exact design ξ_n^* which dominates any other design of experiment?

In the simple case, when $\boldsymbol{\beta}$ is scalar, to find an optimal design is straightforward but, in general, for the situation for more than one dimensional $\boldsymbol{\beta}$ there is no optimal design in the Loewner ordering sense.

A remedy way out of this situation is to compare the designs with respect to some design criterion function Φ.

In general Φ is an antitonic convex real-valued function on the $p \times p$ symmetric matrices which is called an optimality criterion Φ. We suppose without loss of generality that the best design according to Φ, the ***Φ-optimal design***, is the one which minimizes Φ. Thus ξ_n^* is Φ-optimal design if

$$\xi_n^* = arg \min_{\xi_n \in \Xi_n} \Phi(M_{\boldsymbol{\beta}}(\xi_n)) \tag{5.6}$$

where Ξ_n is the set of exact designs of the size n (see (5.4)).

Because of the discreteness in the exact design and consequently the set

$$\mathcal{M}_n = \{M_{\boldsymbol{\beta}}(\xi_n), \xi_n \in \Xi_n\} \tag{5.7}$$

finding a solution to minimize $\Phi(M_{\boldsymbol{\beta}}(\xi_n))$ over Ξ_n may be extremely difficult both analytically or computationally (see Pukelsheim (1993), sec. 4.7). The mathematical problem is avoided by considering ***approximate designs*** (continuous designs) which ignore the restriction that the number of trails at any design points must be integer. In fact

we allow that the weights at the experimental points lie in $[0,1]$ without the condition $np_i = n_i\,(i = 1,\ldots,m)$. If Ξ be the set of all probability measures over τ, ξ, then the Φ-optimal design might be not an exact design, the fact that the "approximate" word could be come up.
For the approximate design $\xi = \left\{ \begin{array}{ccc} x_1 & \ldots & x_m \\ p_1 & \ldots & p_m \end{array} \right\} \in \Xi$ with $\sum_{i=1}^{m} p_i = 1$ the information matrix is

$$M_{\boldsymbol{\beta}}(\xi) = \int_{\tau} \mathbf{f}(x)\mathbf{f}^T(x)\xi(dx) = n\sum_{i=1}^{m} p_i \mathbf{f}(x_i)\mathbf{f}^T(x_i) = n(\mathbf{F}_{\xi}^T \mathbf{W}\mathbf{F}_{\xi})$$

where $\mathbf{W}$ is the same as before with entries p_i.
There are many different optimality criteria which are indexed by some alphabetic letters. we review some of them in this text.
Note that we suppress indices from $\mathbf{F}_{\xi}$ and $M_{\boldsymbol{\beta}}(\xi)$ for simplicity in writing. The most popular criterion is based on the determinant of the information matrix $M_{\boldsymbol{\beta}}(\xi)$ which is defined as following

Def 5.1. *ξ^* is called a* **D-optimal** *design if*

$$\det(M(\xi^*)) \geq \det(M(\xi)) \quad \textit{for all } \xi \in \Xi \tag{5.8}$$

This is equivalent to maximizing the logarithm of the determinant of $M(\xi)$ over Ξ, i.e.

$$\xi^* = \arg\min_{\xi\in\Xi}[-\log(\det(M(\xi)))] \tag{5.9}$$

or $\Phi(M(\xi)) = -\log(\det(M(\xi)))$. Using the logarithm function has the advantage that it is convex on the space of information matrices $\mathcal{M}$, so that a local minimum will in fact be a global minimum.

In other words under the assumption of normal errors the confidence ellipsoid for $\boldsymbol{\beta}$ will be

$$\{\boldsymbol{\beta} :\ \frac{n}{\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T M(\xi)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq c\} = \{\boldsymbol{\beta} :\ (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T Var^{-1}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq c\} \tag{5.10}$$

where c is a constant which depends on the significance level. A common way is to make this ellipsoid as small as possible. A measure of the size of this ellipsoid is its volume which is proportional to the $(\det(M(\xi)))^{-\frac{1}{2}}$. Thus it is reasonable that we maximize $\det(M(\xi))$ or minimize $-\log(\det(M(\xi)))$ which coincides with the definition of D-optimal design.
If a certain linear combination of parameter vector $\boldsymbol{\beta}$, $A^T\boldsymbol{\beta}$, is of interest, then,

$$M_A(\xi) = (A^T M^{-1}(\xi)A)^{-1} \Rightarrow \Phi(M_A(\xi)) = \log(\det(A^T M^{-1}(\xi)A)) \tag{5.11}$$

is an optimal criterion, which is called the D_A***-criterion***. Here A is a $s \times m$ matrix of rank s. If $A = [I_s\,\mathbf{0}]$ then $A^T\boldsymbol{\beta}$ is the part of $\boldsymbol{\beta}$ containing the first s elements. This

optimal criterion is called the D_s**-** criterion (Kiefer and Wolfowitz (1959)).
Suppose that $\lambda_{min}(M(\xi))$ is the minimal eigenvalue of the matrix $M(\xi)$ then the length of the largest principle axis of the ellipsoid (5.10) is $1/\sqrt{\lambda_{min}(M)}$. So this is another optimal design which minimizes $\lambda_{min}(M(\xi))$.

Def 5.2. Φ *is called* E**-criterion** *(Ehrenfeld (1955)) if*

$$\Phi(M(\xi)) = \lambda_{min}(M(\xi)) \tag{5.12}$$

and, hence, ξ^* *will be* E*-optimal design if*

$$\xi^* = \max_{\xi \in \Xi} \lambda_{min}(M(\xi)) \tag{5.13}$$

The A-criterion is a criterion which is considered to minimize the total variance of parameter estimates or equivalently minimizing the average variance.

Def 5.3. ξ^* *is an* **A-optimal** *design (Chernoff (1953)) if it minimize* $tr(M^{-1}(\xi))$ *over* Ξ*, i.e.*

$$\Phi(M(\xi)) = tr(M^{-1}(\xi)) \tag{5.14}$$

If $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $M(\xi)$, then the A-optimality criterion will be ,

$$\Phi(M(\xi)) = \sum_{i=1}^{p} \lambda_i^{-1}. \tag{5.15}$$

There are more criteria to define an optimal design in different ways, which we will not treat in this note.
A wide class of optimal design criteria can be considered as

$$\Phi_k(M(\xi)) = (\frac{1}{p} \sum_{i=1}^{p} \lambda_i^{-k})^{\frac{1}{k}} \tag{5.16}$$

D, A and E-optimality criteria are known as special cases when $k \to 0$, 1 and $k \to \infty$ respectively.
All above criteria are relative to parameter space.
Smith(1918)was one of the first to state a criterion and obtain optimal design based on predictions of Y at x, $\hat{Y}_x$, for regression models. She proposed a criterion which is based on the minimization of the maximum variance of any predicted value over the experimental domain, i.e.

$$\min_{\xi \in \Xi} \max_{x \in \tau} Var(\hat{Y}_x). \tag{5.17}$$

With regard to (2.2), in the linear model,

$$Var(\hat{Y}_x) = \sigma^2 \mathbf{f}^T(x) M^{-1}(\xi) \mathbf{f}(x) \tag{5.18}$$

Def 5.4. *A design ξ^* is* **G-optimal** *if*

$$\xi^* = \arg \min_{\xi \in \Xi} \max_{x \in \tau} \mathbf{f}^T(x) M^{-1}(\xi) \mathbf{f}(x) \tag{5.19}$$

In other words, ξ^* is G-optimal if

$$\max_{x \in \tau} \mathbf{f}^T(x) M^{-1}(\xi^*) \mathbf{f}(x) \leq \max_{x \in \tau} \mathbf{f}^T(x) M^{-1}(\xi) \mathbf{f}(x) \tag{5.20}$$

for all $\xi \in \Xi$.

Since D and G-optimal criteria are the criteria receiving the most attention in applied research and strong connection between D and G-optimality, we will more focus on these criteria.
Two important questions which arise here, how would one shows that a specific design, ξ^*, is the best one? And second, how does D-optimality as a parameter estimation criterion relate to G-optimality as a response estimation criterion? The answers will be through ***convex design theory*** and an ***equivalence theorem***.

5.3 Convex Design Theory

Under the following mild assumptions,

1. τ is compact set.
2. $f_i(.)\,(i = 1, \ldots p)$ are continuous functions.

which guarantee the existence a Φ-optimal design, ***Carathéodory's theorem*** concludes that every element of the design space $\mathcal{M}$ can be expressed as a convex combination of no more than $\frac{p(p+1)}{2} + 1$ elements of the form $\mathbf{f}(x)\mathbf{f}^T(x)$. Moreover as we further see if $\Phi(M(\xi))$ is monotone, then a boundary point of $\mathcal{M}$ minimizes $\Phi(M(\xi))$ and, hence, an optimal design with at most $\frac{p(p+1)}{2}$ support points can be found (see Silvey (1980), Appendix 2).
All optimal criteria have the following properties,

1. Monotonicity: If $M(\xi_1) \leq M(\xi_2)$ (Loewner ordering sense), then $\Phi(M(\xi_1)) \geq \Phi(M(\xi_2))$.
   this property ensures that the minimum of $\Phi(M(\xi))$ occurs at a boundary point.
2. Convexity:$\Phi(M((1-\alpha)\xi_1 + \alpha\xi_2)) \leq (1-\alpha)\Phi(M(\xi_1)) + \alpha\Phi(M(\xi_2))$ for $\alpha \in [0, 1]$.
   this property guarantees that a local minimum will in fact be a global minimum.

Now we are ready to find an answer for the question which was stated to start this section, "how would one shows that a specific design, ξ^*, is the best there is?", by means of the Fréchet directional derivative.

Def 5.5. *For any M_1 and $M_2 \in \mathcal{M}$, the* **Fréchet directional derivative** *of $\Phi(.)$ at M_1 in the direction of M_2 is defined as,*

$$F_\Phi(M_1, M_2) = \lim_{\alpha \to 0^+} \frac{1}{\alpha}[\Phi((1-\alpha)M_1 + \alpha M_2) - \Phi(M_1)] \tag{5.21}$$

or equivalently ,

$$F_\Phi(M_1, M_2) = \frac{d}{d\alpha}\Phi((1-\alpha)M_1 + \alpha M_2) \mid_{\alpha=0^+} \tag{5.22}$$

Note that, since that $M((1-\alpha)\xi_1 + \alpha\xi_2) = (1-\alpha)M(\xi_1) + \alpha M(\xi_2)$ we replaced $M((1-\alpha)\xi_1 + \alpha\xi_2)$ by $(1-\alpha)M(\xi_1) + \alpha M(\xi_2)$in (5.21) and (5.22).

Remark 5.2. *Note that differentiability of Φ at M_1 is a necessary condition. To check if Φ is Fréchet differentiable at M_1 when Φ is convex and $\phi(M_1)$ is finite, a sufficient and necessary condition stated in Silvey (1980), Appendix 3, can be used, i.e., one can check whether the Gâteaux derivative is linear in its second argument, in other words,*

$$G_\Phi(M_1, \sum a_i M_i) = \sum a_i G_\Phi(M_1, M_i)$$

where a_i is a real number with$\sum a_i = 1$ and

$$G_\Phi(M_1, M_i) = \lim_{\alpha \to 0^+} \frac{1}{\alpha}[\Phi(M_1 + \alpha M_2) - \Phi(M_1)]$$

Or equivalently, with regard to Rockafeller (1972) a necessary and sufficient condition for Φ to be Fréchet differentiable is

$$F_\Phi(M_1, \sum a_i M_i) = \sum a_i F_\Phi(M_1, M_i) \tag{5.23}$$

Note that linearity in the second argument of the Gâteaux derivative does not guarantee Fréchet differentiability in general, but since that the optimal design is unique it guarantees existence of Fréchet derivative (For more detail see P.113-115 of Wayne and Varberg(1973)). After this definition, we are ready to obtain the equivalence theorem as a consequence of the Fréchet directional derivative. Silvey(1980) has proved the following theorem which is known as the general ***equivalence theorem*** in the literature.

Theorem 5.3.1. *If $\boldsymbol{\beta}$ is the parameter vector and $\Phi(.)$ is convex on $\mathcal{M}$, the set of design information matrices, and differentiable at $M(\xi^*)$, then the following statements are equivalent*

1. *The measure ξ^* is Φ- optimal*
2. *The Fréchet derivative $F_\Phi(M(\xi^*), \mathbf{f}^T(x)\mathbf{f}^T(x)) \geq 0$ for all $x \in \tau$*

3. The following equality holds,

$$\max_{x\in\tau} F_\Phi(M(\xi^*), \mathbf{f}(x)\mathbf{f}^T(x)) = \min_{\xi\in\Xi}\max_{x\in\tau} F_\Phi(M(\xi), \mathbf{f}(x)\mathbf{f}^T(x)) \tag{5.24}$$

If $\Phi(M(\xi)) = -\log(\det(M(\xi)))$, the last statement expresses the equivalence between D-optimality designs and G-optimality designs. In other words the equivalence theorem says that these two design criteria are identical when the design is expressed as a measure on τ. A compact discussion can be found in Silvey (1980) and many theorems have been provided there.

Remark 5.3. *The condition $F_\Phi(M(\xi^*), \mathbf{f}(x)\mathbf{f}^T(x)) \geq 0$ in the above theorem can often be transformed to the form $\phi(x, \xi^*) \leq C(M(\xi^*))$ where $\phi(x, \xi^*)$ is usually called the* **sensitivity function***, as it shows us how moving some small measure from the support set ξ^* into the direction of x influences the optimality criterion Φ. $C(M(\xi^*))$ is a function of $M(\xi^*)$. For instance if we consider D-optimality in the ordinary linear model, the sensitivity function will be*

$$\phi(x, \xi^*) = \mathbf{f}^T(x)M^{-1}(\xi^*)\mathbf{f}(x)$$

and $C(M(\xi^*)) = p$ where p is the number of parameters. A summarized table of sensitivity functions and $C(M(\xi^*))$ for different optimality criteria in the ordinary linear models has been prepared in table 2.1 in Fedorov and Hackl (1997).

5.4 Convex Design Theory for Linear Mixed Models

The structure of the last section was built on the basis of the fixed effects linear model. In those models the main characteristic of models was independence of observations.
In many cases of application, where the individuals can be observed more than once or the repeated measurements are available, this condition is not true. Fedorov and Hackl (1997), Liski et al. (2002), Luoma (2000) have extensively provided some theoretical results in optimal designs for these models. Liu (2006), and Schmelter (2007a, 2007b and 2007c) have recently done some extensive works on this topic.
Consider again the linear mixed model (2.22),

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{5.25}$$

We focus on a special case of LMM, where $\mathbf{Z}_i = \mathbf{X}_i$, i.e.,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{5.26}$$

This model is called Random Coefficient Regression (RCR) model. $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ is the vector of all observations for individual i, Y_{ij} $(j = 1, \ldots, m_i)$ is observed under the following experimental design

$$\xi_i = \left\{ \begin{array}{ccc} x_{i1} & \ldots & x_{it_i} \\ p_{i1} & \ldots & p_{it_i} \end{array} \right\}$$

where $\sum_{i=1}^{t_i} p_{ij} = 1$. We replace $\mathbf{X}_i$ by $\mathbf{F}_i$ for unifying notation and generality, where F_i has the same definition as F_i in the linear model. Thus we restrict ourselves to the following model,

$$\mathbf{Y}_i = \mathbf{F}_i\boldsymbol{\beta} + \mathbf{F}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{5.27}$$

where $\mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{U})$. We suppose that $\mathbf{U}$ is known. $\boldsymbol{\epsilon}_i$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$.
Regarding to (2.24), the covariance matrix of the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ corresponding to the individual i, is

$$Var(\hat{\boldsymbol{\beta}}_i) = \sigma^2(\mathbf{F}_i^T(\mathbf{F}_i\mathbf{U}\mathbf{F}_i^T + m_i\mathbf{W}_i)^{-1}\mathbf{F}_i)^{-1}$$

where $\mathbf{W}_i$ is the diagonal matrix with the weights p_{ij} as diagonal elements. If σ^2 is known, then the ith individual information matrix, $\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i)$ is defined as:

$$\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i) = \mathbf{F}_i^T(\mathbf{F}_i\mathbf{U}\mathbf{F}_i^T + m_i\mathbf{W}_i)^{-1}\mathbf{F}_i \tag{5.28}$$

So

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\sum_{i=1}^{s}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{U}\mathbf{F}_i^T + m_i\mathbf{W}_i)^{-1}\mathbf{F}_i)^{-1} \Rightarrow \mathfrak{M}_{\boldsymbol{\beta}} = \sum_{i=1}^{s}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{U}\mathbf{F}_i^T + m_i\mathbf{W}_i)^{-1}\mathbf{F}_i$$

where $\mathfrak{M}_{\boldsymbol{\beta}}$ is the information matrix for the whole sample population which depends heavily on the experimental setting. The last expressions hold because of independence of individuals. For more details see Schmelter (2007a).

Lemma 5.4.1. *(cf. Schmelter (2007a) Lemma1)* $\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i)$ *in* (5.28) *can be represented as*

$$\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i) = ((m_i\mathbf{F}_i^T\mathbf{W}_i\mathbf{F}_i)^{-1} + \mathbf{U})^{-1} = (\underline{M}^{-1}(\xi_i) + \mathbf{U})^{-1}$$

where $\underline{M}(\xi_i) = m_i\mathbf{F}_i^T\mathbf{W}_i\mathbf{F}_i$ *is the non-normalized information matrix of* ξ_i *in the corresponding fixed effect model.*

The same result can be found in Liski et al.(2002). This above representation of $\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i)$ clearly separates the effects of the random part and the fixed part on the information matrix. We can also obtain thus any information matrix dominance in the fixed effects models will also carry through for random coefficient regression models.
The main property of the information matrix of the ordinary linear model is that the information matrix of the convex combination of two designs is the convex combination of the information matrix of the two designs, a property which does not carry over through random coefficient regression model, i.e.

$$\mathfrak{M}_{\boldsymbol{\beta}}(\alpha\xi_1 + (1-\alpha)\xi_2) \neq \alpha\mathfrak{M}_{\boldsymbol{\beta}}(\xi_1) + (1-\alpha)\mathfrak{M}_{\boldsymbol{\beta}}(\xi_2)$$

so we cannot directly apply convex design theory and find an equivalence theorem for RCR models and hence we have to reconstruct the concepts and theorem for this case.
Under mild assumptions, lemma 8.5 in Schmelter (2007c) allows us to directly apply convex design theory as described in the last section to the ordinary linear model. For instance , if

$$\Phi[(\underline{M}^{-1}(\xi) + \mathbf{U})^{-1}] = \log(\det(\underline{M}^{-1}(\xi) + \mathbf{U}))$$

i.e. the optimality criterion is the D-criterion. The second statement in Theorem 5.3.1 leads us to,

$$m\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}_{\boldsymbol{\beta}}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x) \leq tr[\mathfrak{M}_{\boldsymbol{\beta}}(\xi^*)\underline{M}^{-1}(\xi^*)] \text{ for all } x \in \tau \tag{5.29}$$

as a necessary and sufficient condition for ξ^* to be optimal for the estimation of $\boldsymbol{\beta}$.
In general, for different individuals we may use different designs for data collection, we introduce population design ζ:

$$\zeta = \left\{ \begin{array}{ccc} \xi_1 & \dots & \xi_m \\ q_1 & \dots & q_m \end{array} \right\} \tag{5.30}$$

with $\sum_{i=1}^{m} q_i = 1$, where the ξ_i, $i = 1, \dots, m$ are individual designs. The following theorem expresses the relation between individual optimal design and population optimal designs.

Theorem 5.4.1. *(cf. Schmelter (2007a) Theorem 1) Optimal designs can be found among those which are uniform across the individuals, i.e. if ξ_i^* is Φ-optimal for the individual design then we can observe all individual under this experimental design.*

The proof can be found in Schmelter (2007a).
An immediately conclusion of this theorem is that we can ignore index i in $\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i)$, or in other words

$$\mathfrak{M}_{\boldsymbol{\beta}} = s\mathfrak{M}_{\boldsymbol{\beta}}(\xi_1)$$

5.5 Locally Optimal Designs

A further common property of generalized linear models and generalized linear mixed models is that the information matrix depends on the unknown parameters and hence the optimum design will also depend upon the value of $\boldsymbol{\beta}$. It poses a dual problem: to find the optimal design we must know the parameters in advance and to get knowledge about the parameters we need the experimental design to perform the experiment. A simple approach to this problem is to look for locally optimal designs (the term introduced by Chernoff (1953)) which are based on an initial guess of the parameters and then find optimal designs which are optimal with respect to this initial guess. This initial guess

used in locally optimal design might come from previous experimentation, or from a pilot experiment conducted particularly for this purpose, or merely a guess. We shall call this parameter guess the initial estimate or initial guess no matter how it is obtained.
Locally optimal designs will often apply in the next chapter where we attempt to find optimal designs for Poisson regression models with random intercept and a Poisson regression model with random slope.

5.6 Convex Design Theory for Poisson Regression Models with Random Intercept

We consider again the individual information matrix of a Poisson regression model with random intercept (4.15),

$$\mathfrak{M}_{\boldsymbol{\beta}}(\xi) = (\underline{M}_{\boldsymbol{\beta}}^{-1}(\xi) + \mathbf{U})^{-1}$$

We ignore the index i and then $\boldsymbol{\beta}$ in this expression just for simplicity. This representation is very similar to the information matrix of a RCR model which has briefly been described in the previous section. Here we have

$$\mathbf{U} = \begin{pmatrix} e^{\sigma^2} - 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

In this section we try to extend convex design theory and the equivalence theorem for the RCR model to the Poisson regression model with random intercept.
The following lemma is basically needed to find later results.

Lemma 5.6.1. *The following inequality holds for every ξ_1 and ξ_2*

$$\mathfrak{M}((1-\alpha)\xi_1 + \alpha\xi_2) \geq (1-\alpha)\mathfrak{M}(\xi_1) + \alpha\mathfrak{M}(\xi_2)$$

with respect to the Loewner partial ordering of symmetric non-negative definite matrices, and where

$$\mathfrak{M}(\xi) = (\underline{M}^{-1}(\xi) + \mathbf{U})^{-1}$$

Proof. The matrix $\mathbf{U}$ is not invertible, so we use some regularization

$$(((1-\alpha)\underline{M}(\xi_1) + \alpha\underline{M}(\xi_2))^{-1} + \mathbf{U})^{-1} = \lim_{\delta \to 0^+} (((1-\alpha)\underline{M}(\xi_1) + \alpha\underline{M}(\xi_2))^{-1} + (\mathbf{U} + \delta I))^{-1}$$

$$= \lim_{\delta \to 0^+} [(\mathbf{U} + \delta I)^{-1} - (\mathbf{U} + \delta I)^{-1}((1-\alpha)\underline{M}(\xi_1) + \alpha\underline{M}(\xi_2) + (\mathbf{U} + \delta I)^{-1})^{-1}(\mathbf{U} + \delta I)^{-1}]$$

$$\geq \lim_{\delta\to 0^+}[(\mathbf{U}+\delta I)^{-1} - (\mathbf{U}+\delta I)^{-1}((1-\alpha)(\underline{M}(\xi_1)+(\mathbf{U}+\delta I)^{-1})^{-1} + \alpha(\underline{M}(\xi_2)+(\mathbf{U}+\delta I)^{-1})^{-1})(\mathbf{U}+\delta I)^{-1}]$$
$$= \lim_{\delta\to 0^+}[(1-\alpha)(\underline{M}^{-1}(\xi_1)+(\mathbf{U}+\delta I))^{-1} + \alpha(\underline{M}^{-1}(\xi_2)+(\mathbf{U}+\delta I))^{-1}]$$
$$= (1-\alpha)(\underline{M}^{-1}(\xi_1)+\mathbf{U})^{-1} + \alpha(\underline{M}^{-1}(\xi_2)+\mathbf{U})^{-1}$$

and the representation immediately follows. The inequality holds because of $(\alpha A + (1-\alpha)B)^{-1} \leq \alpha A^{-1} + (1-\alpha)B^{-1}$ for invertible matrices A and B (see ,for example, Fedorov and Hackl (1997) p.107). □

This outcome does not coincide with the result in ordinary linear models. The following lemma is an immediate consequence of the above lemma and helps us to relate convex design theory for our model and convex design theory for the ordinary linear model.

Lemma 5.6.2. *Suppose that $\mathcal{M}$ is the set of non-negative definite (N.N.D.) Matrices of order $k\times k$ and $\Phi : \mathcal{M} \to (-\infty, \infty]$ is an optimality criterion with regularity assumptions, i.e. convexity, monotonicity. If we define $\Psi : \Xi \to (-\infty, \infty]$ with $\Psi(\xi) = \Phi(\mathfrak{M}(\xi)) = \Phi[(M^{-1}(\xi)+D)^{-1}]$, then Ψ has the same properties as Φ, i.e., $\Psi(\xi)$ is also monotone, convex, where Ξ is the set of all probability measures on τ*

.

Proof. Let ξ_1 and ξ_2 two designs in Ξ
Monotonicity property:

$$\begin{aligned}\xi_1 \succ \xi_2 &\Leftrightarrow \mathfrak{M}(\xi_1) \geq \mathfrak{M}(\xi_2)\\ &\Leftrightarrow \Phi(\mathfrak{M}(\xi_1)) \leq \Phi(\mathfrak{M}(\xi_2)) \Leftrightarrow \Psi(\xi_1) \leq \Psi(\xi_2)\end{aligned}$$

convexity property:

$$\begin{aligned}\Psi((1-\alpha)\xi_1+\alpha\xi_2) &= \Phi(\mathfrak{M}((1-\alpha)\xi_1+\alpha\xi_2))\\ &\leq \Phi((1-\alpha)\mathfrak{M}(\xi_1)+\alpha\mathfrak{M}(\xi_2))\\ &\leq (1-\alpha)\Phi(\mathfrak{M}(\xi_1))+\alpha\Phi(\mathfrak{M}(\xi_2))\\ &= (1-\alpha)\Psi(\xi_1)+\alpha\Psi(\xi_2)\end{aligned}$$

The first inequality is because of the antitonicity of Φ and Lemma 5.6.1 and the second inequality is due to the convexity of Φ. □

A similar proof can be found in Schmelter(2007c).
This lemma guarantees the applicability of convex design theory of the ordinary linear model to our model, the Poisson regression model with random intercept.
If we consider the D-criterion

$$\Psi(\xi) = \log(\det(\underline{M}^{-1}(\xi)+\mathbf{U})) \tag{5.31}$$

then the following theorem gives a sufficient and necessary condition that ξ^* is D-optimal.

Theorem 5.6.1. *ξ^* is D-optimal for a Poisson regression model with random intercept if and only if*

$$m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x) \leq tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)]\,\forall x \in \tau$$
$$\Leftrightarrow \mu(x)\mathbf{f}^T(x)M^{-1}(\xi^*)[M^{-1}(\xi^*)+m\mathbf{U}]^{-1}M^{-1}(\xi^*)\mathbf{f}(x) \leq p-1+\frac{m}{m+\mathbf{u}_p^T M(\xi^*)\mathbf{u}_p}\,\forall x \in \tau$$

where $M(\xi^*)$ and $\underline{M}(\xi^*)$ are the normalized and non-normalized information matrices for the model without random effect respectively and $\mu(x) = e^{\beta_0+\beta_1 x+\frac{1}{2}\sigma^2}$

Proof. First of all we need to check the Fréchet differentiability of Ψ defined in (5.31). The Gâteaux derivative of the criterion function Ψ is

$$\begin{aligned}
G_\Psi(\xi,\xi') &= \lim_{\alpha\to 0^+}\frac{1}{\alpha}[\log(\det(\underline{M}(\xi)+\alpha\underline{M}(\xi'))^{-1}+\mathbf{U})-\log(\det(\underline{M}^{-1}(\xi)+\mathbf{U}))]\\
&= \lim_{\alpha\to 0^+}\frac{1}{\alpha}[\log\frac{\det(I+\mathbf{U}(\underline{M}(\xi)+\alpha\underline{M}(\xi')))}{\det(\underline{M}(\xi)+\alpha\underline{M}(\xi'))}-\log\frac{\det(I+\mathbf{U}\underline{M}(\xi))}{\det(\underline{M}(\xi))}]\\
&= \lim_{\alpha\to 0^+}\frac{1}{\alpha}[\log\frac{\det(I+\mathbf{U}(\underline{M}(\xi)+\alpha\underline{M}(\xi')))\det(\underline{M}(\xi))}{\det(\underline{M}(\xi)+\alpha\underline{M}(\xi'))\det(I+\mathbf{U}\underline{M}(\xi))}]\\
&= \lim_{\alpha\to 0^+}\frac{1}{\alpha}[\log(\det(I+\alpha\mathbf{U}\underline{M}(\xi')(I+\mathbf{U}\underline{M}(\xi))^{-1}))-\log(\det(I+\alpha\underline{M}(\xi')\underline{M}^{-1}(\xi)))]\\
&= \lim_{\alpha\to 0^+}\frac{1}{\alpha}[(1+\alpha tr(\mathbf{U}\underline{M}(\xi')(I+\mathbf{U}\underline{M}(\xi))^{-1}))-(1+\alpha tr(\underline{M}(\xi')\underline{M}^{-1}(\xi')))]\\
&= tr(\mathbf{U}\underline{M}(\xi')(I+\mathbf{U}\underline{M}(\xi))^{-1})-tr(\underline{M}(\xi')\underline{M}^{-1}(\xi))\\
&= tr((\underline{M}^{-1}(\xi)+\mathbf{U})^{-1}\mathbf{U}-I)(\underline{M}(\xi')\underline{M}^{-1}(\xi))
\end{aligned}$$

let $\xi' = \sum_{i=1}^{r} a_i\xi_i$ with $\sum_{i=1}^{r} a_i = 1$ therefore

$$\begin{aligned}
G_\Psi(\xi_1,\sum_{i=1}^{r}a_i\xi_i) &= tr((\underline{M}^{-1}(\xi_1)+\mathbf{U})^{-1}\mathbf{U}-I)((\sum_{i=1}^{r}a_i\underline{M}(\xi_i))\underline{M}^{-1}(\xi_1))\\
&= \sum_{i=1}^{r}a_i tr((\underline{M}^{-1}(\xi_1)+\mathbf{U})^{-1}\mathbf{U}-I)\underline{M}(\xi_i)\underline{M}^{-1}(\xi_1)\\
&= \sum_{i=1}^{r}a_i G_\Psi(\xi_1,\xi_i)
\end{aligned}$$

so Ψ is Fréchet differentiable, and by applying the rules for matrix/vector differential calculus (see, e.g. Magnus and Neudecker (1988) or Wand (2002)) the Fréchet derivative

which is defined in (5.21) or equivalently in (5.22), will be

$$\begin{aligned}
F_\Psi(\xi^*,\xi) &= \frac{d}{d\alpha}\psi((1-\alpha)\xi^* + \alpha\xi)\mid_{\alpha=0^+} \\
&= \frac{d}{d\alpha}\log\{\det[((1-\alpha)\underline{M}(\xi^*) + \alpha\underline{M}(\xi))^{-1} + \mathbf{U}])\}\mid_{\alpha=0^+} \\
&= tr\{[((1-\alpha)\underline{M}(\xi^*) + \alpha\underline{M}(\xi))^{-1} + \mathbf{U}]^{-1}\frac{d}{d\alpha}[((1-\alpha)\underline{M}(\xi^*) + \alpha\underline{M}(\xi))^{-1} + \mathbf{U}]\}\mid_{\alpha=0^+} \\
&= -tr\{(\underline{M}^{-1}(\xi^*) + \mathbf{U})^{-1}((1-\alpha)\underline{M}(\xi^*) + \alpha\underline{M}(\xi))^{-1}(\underline{M}(\xi) \\
&\qquad -\underline{M}(\xi^*))((1-\alpha)\underline{M}(\xi^*) + \alpha\underline{M}(\xi))^{-1}\} \\
&= tr\{(\underline{M}^{-1}(\xi^*) + \mathbf{U})^{-1}\underline{M}^{-1}(\xi^*)(\underline{M}(\xi^*) - \underline{M}(\xi))\underline{M}^{-1}(\xi^*)\}
\end{aligned}$$

where $\underline{M}(\xi) = m\mu(x)\mathbf{f}(x)\mathbf{f}^T(x)$ is the information matrix for a single point design ξ. With regard to the theorem (5.3.1) ξ^* is D-optimal if

$$\begin{aligned}
&F_\Psi(\xi^*,\xi) \geq 0 \text{ for all } \xi \\
&\quad \Leftrightarrow m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x) \leq tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)]
\end{aligned}$$

the right expression of the above inequality can be simplified as

$$\begin{aligned}
tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)] &= tr[(\underline{M}^{-1}(\xi^*) + \mathbf{U})^{-1}\underline{M}^{-1}(\xi^*)] \\
&= tr[(\underline{M}(\xi^*) - \frac{\underline{M}(\xi^*)\mathbf{u}_p\mathbf{u}_p^T\underline{M}(\xi^*)}{1 + \mathbf{u}_p^T\underline{M}(\xi^*)\mathbf{u}_p})\underline{M}^{-1}(\xi^*)] \\
&= p - \frac{tr\underline{M}(\xi^*)\mathbf{u}_p\mathbf{u}_p^T}{1 + \mathbf{u}_p^T\underline{M}(\xi^*)\mathbf{u}_p} = p - \frac{\mathbf{u}_p^t\underline{M}(\xi^*)\mathbf{u}_p}{1 + \mathbf{u}_p^T\underline{M}(\xi^*)\mathbf{u}_p} \\
&= \frac{p + (p-1)\mathbf{u}_p^T\underline{M}(\xi^*)\mathbf{u}_p}{1 + \mathbf{u}_p^T\underline{M}(\xi^*)\mathbf{u}_p} = (p-1) + \frac{1}{1 + \mathbf{u}_p^T\underline{M}(\xi^*)\mathbf{u}_p}
\end{aligned}$$

since that $\underline{M}(\xi^*) = \frac{1}{m}M(\xi^*)$ thus the representation follows. □

Note that, by the same way as in the linear mixed models, we can indicate that the population design is optimal when all individuals are observed under the same individual optimal design.

5.7 Convex Design Theory for the Poisson Regression Model with Random Slope

We consider the individual design as follow:

$$\xi_i = \left\{ \begin{array}{ccc} x_{i1} & \dots & x_{it_i} \\ p_{i1} & \dots & p_{it_i} \end{array} \right\}.$$

Lemma 5.7.1. *For the above design the quasi-information matrix of the Poisson regression model with random slope is*

$$\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i) = \mathbf{F}_i^T(\mathbf{A}_i^{-1} + \mathbf{B}_{ii})^{-1}\mathbf{F}_i \tag{5.32}$$

where $\mathbf{A}_i = m_i diag\{p_{i1}\mu_{i1}, p_{i2}\mu_{i2}, \ldots, p_{it_i}\mu_{it_i}\}$, $\mathbf{F}_i = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_{i1} & x_{i2} & \ldots & x_{it_i} \end{pmatrix}^T$ *and*

$$\mathbf{B}_{ii} = \begin{pmatrix} e^{\sigma^2 x_{i1}^2} - 1 & e^{\sigma^2 x_{i1}x_{i2}} - 1 & \ldots & e^{\sigma^2 x_{i1}x_{it_i}} - 1 \\ e^{\sigma^2 x_{i1}x_{i2}} - 1 & e^{\sigma^2 x_{i2}^2} - 1 & \ldots & e^{\sigma^2 x_{i2}x_{it_i}} - 1 \\ \vdots & \vdots & \ddots & \vdots \\ e^{\sigma^2 x_{it_i}x_{i1}} - 1 & e^{\sigma^2 x_{it_i}x_{i2}} - 1 & \ldots & e^{\sigma^2 x_{it_i}^2} - 1 \end{pmatrix}. \tag{5.33}$$

Proof. Regarding to the Lemma 4.2.2, the proof is immediately obtained . □

In contrast to the previous model, we can not separate the random effect part and fixed effect part. This fact causes that all theorems and lemmas which are constructed based on these two parts cannot be used. So we build a new view to optimal design for this model.
The following lemma will be needed in the reminder of this section.

Lemma 5.7.2. *If* ξ_1 *and* ξ_2 *be two experimental designs in* $\in \Xi$ *then*

$$\mathfrak{M}_{\boldsymbol{\beta}}((1-\alpha)\xi_1 + \alpha\xi_2) \geq (1-\alpha)\mathfrak{M}_{\boldsymbol{\beta}}(\xi_1) + \alpha\mathfrak{M}_{\boldsymbol{\beta}}(\xi_2)$$

where $\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i)$ $(i = 1, 2)$ *is the quasi-information matrix for* $\boldsymbol{\beta}$ *based on individual design* ξ_i*. For simplicity we suppress the index* $\boldsymbol{\beta}$*.*

Proof.

$$\begin{aligned} \mathfrak{M}_{\boldsymbol{\beta}}((1-\alpha)\xi_1 + \alpha\xi_2) &= \mathbf{F}^T \left(\begin{pmatrix} (1-\alpha)\mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{A}_2 \end{pmatrix}^{-1} + \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \right)^{-1} \mathbf{F} \\ &= \mathbf{F}^T \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1} + \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \alpha^{-1}\mathbf{A}_2^{-1} + \mathbf{B}_{22} \end{pmatrix}^{-1} \mathbf{F} \end{aligned}$$

where $\mathbf{F} = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}$ with $\mathbf{F}_1 = \begin{pmatrix} 1 & \ldots & 1 \\ x_{11} & \ldots & x_{1t_1} \end{pmatrix}^T$ and $\mathbf{F}_2 = \begin{pmatrix} 1 & \ldots & 1 \\ x_{21} & \ldots & x_{2t_2} \end{pmatrix}^T$ as the design matrices corresponding to ξ_1 and ξ_2 respectively. We also have

$$\mathbf{B}_{12} = \begin{pmatrix} e^{\sigma^2 x_{11}x_{12}} - 1 & \ldots & e^{\sigma^2 x_{11}x_{1t_2}} - 1 \\ \vdots & \ldots & \vdots \\ e^{\sigma^2 x_{1t_1}x_{21}} - 1 & \ldots & e^{\sigma^2 x_{1t_1}x_{2t_2}} - 1 \end{pmatrix} = \mathbf{B}_{21}^T$$

Due to the lemma(4.2.1), $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$ is PSD and hence all leading principle minors are non-negative. On the other hand $0 < \alpha < 1$ and hence the constants $\frac{\alpha}{1-\alpha}$ and $\frac{1-\alpha}{\alpha}$ are positive and inverse to each other. Therefore all leading principle minors of $\begin{pmatrix} (\frac{\alpha}{1-\alpha})\mathbf{B}_{11} & -\mathbf{B}_{12} \\ -\mathbf{B}_{21} & (\frac{1-\alpha}{\alpha})\mathbf{B}_{22} \end{pmatrix} = \mathbf{C}\mathbf{B}\mathbf{C}^T$ are positive semi-definite with $\mathbf{C} = \begin{pmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\sqrt{\frac{1-\alpha}{\alpha}}\mathbf{I} \end{pmatrix}$.

$$\begin{pmatrix} (\frac{\alpha}{1-\alpha})\mathbf{B}_{11} & -\mathbf{B}_{12} \\ -\mathbf{B}_{21} & (\frac{1-\alpha}{\alpha})\mathbf{B}_{22} \end{pmatrix} = \begin{pmatrix} (\frac{1}{1-\alpha}-1)\mathbf{B}_{11} & -\mathbf{B}_{12} \\ -\mathbf{B}_{21} & (\frac{1}{\alpha}-1)\mathbf{B}_{22} \end{pmatrix} \geq 0$$

$$\Leftrightarrow \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+\mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \alpha^{-1}\mathbf{A}_2^{-1}+\mathbf{B}_{22} \end{pmatrix} - \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+(1-\alpha)^{-1}\mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \alpha^{-1}\mathbf{A}_2^{-1}+\alpha^{-1}\mathbf{B}_{22} \end{pmatrix} \leq 0$$

$$\Leftrightarrow \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+\mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \alpha^{-1}\mathbf{A}_2^{-1}+\mathbf{B}_{22} \end{pmatrix} \leq \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+(1-\alpha)^{-1}\mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \alpha^{-1}\mathbf{A}_2^{-1}+\alpha^{-1}\mathbf{B}_{22} \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+\mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \alpha^{-1}\mathbf{A}_2^{-1}+\mathbf{B}_{22} \end{pmatrix}^{-1} \geq \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+(1-\alpha)^{-1}\mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \alpha^{-1}\mathbf{A}_2^{-1}+\alpha^{-1}\mathbf{B}_{22} \end{pmatrix}^{-1}$$

the last inequality is because of the antitonic property of the matrix inversion (Pukelsheim(1993), page 13). So

$$\mathbf{F}^T \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+\mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \alpha^{-1}\mathbf{A}_2^{-1}+\mathbf{B}_{22} \end{pmatrix}^{-1} \mathbf{F}$$
$$\geq \mathbf{F}^T \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+(1-\alpha)^{-1}\mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \alpha^{-1}\mathbf{A}_2^{-1}+\alpha^{-1}\mathbf{B}_{22} \end{pmatrix}^{-1} \mathbf{F}$$

$$\Leftrightarrow \mathbf{F}^T \begin{pmatrix} (1-\alpha)^{-1}\mathbf{A}_1^{-1}+\mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \alpha^{-1}\mathbf{A}_2^{-1}+\mathbf{B}_{22} \end{pmatrix}^{-1} \mathbf{F}$$
$$\geq \mathbf{F}_1^T((1-\alpha)^{-1}\mathbf{A}_1^{-1}+(1-\alpha)^{-1}\mathbf{B}_{11}^{-1})^{-1}\mathbf{F}_1 + \mathbf{F}_2^T(\alpha^{-1}\mathbf{A}_2^{-1}+\alpha^{-1}\mathbf{B}_{22})^{-1}\mathbf{F}_2$$

So the claim follows. □

Theorem 5.7.1. *If $\Psi : \Xi \to (-\infty, \infty]$ with $\Psi(\xi) = \Phi(\mathfrak{M}(\xi))$ and ξ_1 and ξ_2 be two designs in Ξ, the following statements hold*

1. *Ψ is antitonic, that is if ξ_1 and ξ_2 be two designs in Ξ and $\xi_1 \succ \xi_2$ then $\Psi(\xi_1) \leq \Psi(\xi_2)$.*

2. *Ψ is convex, that is if ξ_1 and ξ_2 be two designs in Ξ, then*

$$\Psi((1-\alpha)\xi_1 + \alpha\xi_2) \leq (1-\alpha)\Psi(\xi_1) + \alpha\Psi(\xi_2)$$

Proof.

1. $\xi_1 \succ \xi_2 \Rightarrow \mathfrak{M}(\xi_1) \geq \mathfrak{M}(\xi_2) \Rightarrow \Phi(\mathfrak{M}(\xi_1)) \leq \Phi(\mathfrak{M}(\xi_2)) \Rightarrow \Psi(\xi_1) \leq \Psi(\xi_2)$
2. $\Psi((1-\alpha)\xi_1 + \alpha\xi_2) = \Phi(\mathfrak{M}((1-\alpha)\xi_1 + \alpha\xi_2))$
   $\leq \Phi((1-\alpha)\mathfrak{M}(\xi_1) + \alpha\mathfrak{M}(\xi_2)) \leq (1-\alpha)\Phi(\mathfrak{M}(\xi_1)) + \alpha\Phi(\mathfrak{M}(\xi_2))$
   $= (1-\alpha)\Psi(\xi_1) + \alpha\Psi(\xi_2)$

The second statement holds because of lemma (5.7.2) and antitonicity of Φ and the last inequality in there is true due to convexity of Φ. □

This theorem allows the practitioner to verify that a given design is globally optimal. As we have seen the most popular optimal criterion is the D-optimality criterion. We are now ready to obtain a new version of the equivalence theorem to check D-optimality of a design in this case.
Let $\mathbf{F}_0, \mathbf{F}_1, \ldots, \mathbf{F}_k$ be the design matrices and $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_k$ are diagonal matrices with elements $m_{ij}\mu_{ij}$ and $\mu_{ij} = e^{\beta_0+\beta_1 x_{ij}+\frac{1}{2}\sigma^2 x_{ij}^2}$ $(j = 1, \ldots, t_i)$ corresponding to the designs $\xi, \xi_1, \ldots, \xi_k$. The following lemma is necessary to prove Fréchet differentiability of D-criterion.

Lemma 5.7.3. *Suppose that* $f(\epsilon) = \det[\mathfrak{M}((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i)]$. $f(\epsilon)$ *is continuous, concave and differentiable at any right hand neighborhood of* $\epsilon = 0$.

Proof. Suppose that $\mathbf{F}^T = \begin{pmatrix} \mathbf{F}_0^T & \mathbf{F}_1^T & \ldots & \mathbf{F}_k^T \end{pmatrix} = \begin{pmatrix} \mathbf{F}_0^T & \widetilde{\mathbf{F}}^T \end{pmatrix}$ is the design matrix of the convex combination of ξ and $(1-\epsilon)\xi+\epsilon\sum_{i=1}^{k} a_i\xi_i$ and $\mathbf{A}(\epsilon) = diag\{(1-\epsilon)\mathbf{A}_0, \epsilon a_1\mathbf{A}_1, \ldots, \epsilon a_k\mathbf{A}_k\} = \begin{pmatrix} \mathbf{A}_0(\epsilon) & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{A}}(\epsilon) \end{pmatrix}$ is a block diagonal matrix where $\mathbf{A}_0(\epsilon) = (1-\epsilon)\mathbf{A}_0$ and $\widetilde{\mathbf{A}}(\epsilon) = \epsilon \cdot diag\{a_1\mathbf{A}_1, \ldots, a_k\mathbf{A}_k\}$. We also suppose that $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{00} & \widetilde{\mathbf{B}}_{01} \\ \widetilde{\mathbf{B}}_{10} & \widetilde{\mathbf{B}}_{11} \end{pmatrix}$ where $\widetilde{\mathbf{B}}_{01} = \widetilde{\mathbf{B}}_{10}^T = \begin{pmatrix} \mathbf{B}_{01} & \ldots & \mathbf{B}_{0k} \end{pmatrix}$ and $\widetilde{\mathbf{B}}_{11} = (\mathbf{B}_{hj})_{h,j=1}^{k}$ are also appropriate block matrices with $\mathbf{B}_{hj} = (e^{\sigma^2 x_{hl}x_{jl'}} - 1)_{l,l'=1}^{t_h,t_j}$ for h and $j = 0, 1, \ldots, k$.
-Continuity of $f(\epsilon)$: Let $\mathbf{B}$ be invertible. We have $f(\epsilon) = \det[\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F}]$ for $\epsilon > 0$ and $f(0) = \mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0)$, so $f(\epsilon)$ is a continuous function of $\epsilon > 0$ and

$$\begin{aligned}
\lim_{\epsilon\to 0} f(\epsilon) &= \lim_{\epsilon\to 0} \det[\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F}] = \lim_{\epsilon\to 0} \det[\mathbf{F}^T(\mathbf{A}(\epsilon) - \mathbf{A}(\epsilon)(\mathbf{A}(\epsilon) + \mathbf{B}^{-1})^{-1}\mathbf{A}(\epsilon))\mathbf{F}] \\
&= \lim_{\epsilon\to 0} \det\left[\mathbf{F}^T\left(\mathbf{A}(\epsilon) - \mathbf{A}(\epsilon)\begin{pmatrix} \mathbf{A}_0(\epsilon) + \mathbf{B}^{00} & \widetilde{\mathbf{B}}^{01} \\ \widetilde{\mathbf{B}}^{10} & \widetilde{\mathbf{A}}(\epsilon) + \widetilde{\mathbf{B}}^{11} \end{pmatrix}^{-1}\mathbf{A}(\epsilon)\right)\mathbf{F}\right] \\
&= \det\left[\mathbf{F}^T\left(\begin{pmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{A}_0 + \mathbf{B}^{00} & \widetilde{\mathbf{B}}^{01} \\ \widetilde{\mathbf{B}}^{10} & \widetilde{\mathbf{B}}^{11} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right)\mathbf{F}\right]
\end{aligned}$$

where $\begin{pmatrix} \mathbf{B}^{00} & \widetilde{\mathbf{B}}^{01} \\ \widetilde{\mathbf{B}}^{10} & \widetilde{\mathbf{B}}^{11} \end{pmatrix} = \mathbf{B}^{-1}$. If we define that $\begin{pmatrix} \mathbf{A}_0(\epsilon) + \mathbf{B}^{00} & \widetilde{\mathbf{B}}^{01} \\ \widetilde{\mathbf{B}}^{10} & \widetilde{\mathbf{A}}(\epsilon) + \widetilde{\mathbf{B}}^{11} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{T}_{00} & \widetilde{\mathbf{T}}_{01} \\ \widetilde{\mathbf{T}}_{10} & \widetilde{\mathbf{T}}_{11} \end{pmatrix}$ with $\mathbf{T}_{00} = (\mathbf{A}_0 + \mathbf{B}^{00} - \widetilde{\mathbf{B}}^{01}(\widetilde{\mathbf{B}}^{11})^{-1}\widetilde{\mathbf{B}}^{10})^{-1} = (\mathbf{A}_0 + \mathbf{B}_{00}^{-1})^{-1}$, we have

$$\begin{aligned} \lim_{\epsilon \to 0} f(\epsilon) &= \det\left[\mathbf{F}^T \left(\begin{pmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{A}_0\mathbf{T}_{00}\mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right) \mathbf{F}\right] \\ &= \det[\mathbf{F}_0^T(\mathbf{A}_0 - \mathbf{A}_0(\mathbf{A}_0 + \mathbf{B}_{00}^{-1})^{-1}\mathbf{A}_0)\mathbf{F}_0] = \det(\mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0) \end{aligned}$$

- Concavity of $f(\epsilon)$:

$$\begin{aligned} f((1-\beta)\epsilon + \beta\epsilon') &= \det(\mathfrak{M}((1 - (1-\beta)\epsilon + \beta\epsilon')\xi + ((1-\beta)\epsilon + \beta\epsilon')\sum_{i=1}^{k} a_i\xi_i)) \\ &= \det(\mathfrak{M}((1-\beta)((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i) + \beta((1-\epsilon')\xi + \epsilon'\sum_{i=1}^{k} a_i\xi_i)) \\ &\geq (1-\beta)f(\epsilon) + \beta f(\epsilon) \end{aligned}$$

- Differentiability of $f(\epsilon)$:
Since that $\mathbf{A}(\epsilon)$ is invertible for $\epsilon > 0$, then $f(\epsilon) = \det[\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F}]$ is differentiable at any right hand neighborhood of $\epsilon = 0$ and

$$\begin{aligned} f'(\epsilon) &= \det\left[\mathfrak{M}((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i)\right] \\ &\quad \cdot tr\left\{\left[\mathfrak{M}((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i)\right]^{-1} \frac{d}{d\epsilon}\left[\mathfrak{M}((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i)\right]\right\} \\ &= \det\left[\mathfrak{M}((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i)\right] \\ &\quad \cdot tr\{\left[\mathfrak{M}((1-\epsilon)\xi + \epsilon\sum_{i=1}^{k} a_i\xi_i)\right]^{-1} \frac{d}{d\epsilon}(\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F})\} \end{aligned} \tag{5.34}$$

The derivative inside tr in (5.34) will be

$$\begin{aligned} \frac{d}{d\epsilon}\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F} &= \mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}(\frac{d}{d\epsilon}(\mathbf{A}^{-1}(\epsilon)))(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F} \\ &= \mathbf{F}^T(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{A}^{-1}(\epsilon)(\frac{d}{d\epsilon}\mathbf{A}(\epsilon))\mathbf{A}^{-1}(\epsilon)(\mathbf{A}^{-1}(\epsilon) + \mathbf{B})^{-1}\mathbf{F} \end{aligned} \tag{5.35}$$

where we have

$$\mathbf{A}^{-1}(\epsilon)(\frac{d}{d\epsilon}\mathbf{A}(\epsilon))\mathbf{A}^{-1}(\epsilon) = \begin{pmatrix} -(1-\epsilon)^{-2}\mathbf{A}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \epsilon^{-2}\widetilde{\mathbf{A}}^{-1} \end{pmatrix} \tag{5.36}$$

Suppose that

$$(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1} = \begin{pmatrix} \mathbf{A}_0^{-1}(\epsilon)+\mathbf{B}_{00} & \widetilde{\mathbf{B}}_{01} \\ \widetilde{\mathbf{B}}_{10} & \widetilde{\mathbf{A}}^{-1}(\epsilon)+\widetilde{\mathbf{B}}_{11} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{K}_{11}(\epsilon) & \mathbf{K}_{12}(\epsilon) \\ \mathbf{K}_{21}(\epsilon) & \mathbf{K}_{22}(\epsilon) \end{pmatrix} \tag{5.37}$$

where

$$\begin{aligned}
\mathbf{K}_{11}(\epsilon) &= (\mathbf{A}_0^{-1}(\epsilon)+\mathbf{B}_{00}-\widetilde{\mathbf{B}}_{01}(\widetilde{\mathbf{A}}^{-1}(\epsilon)+\widetilde{\mathbf{B}}_{11})^{-1}\widetilde{\mathbf{B}}_{10})^{-1} \text{ of order } t_0 \times t_0 \\
\mathbf{K}_{22}(\epsilon) &= (\widetilde{\mathbf{A}}^{-1}(\epsilon)+\widetilde{\mathbf{B}}_{11}-\widetilde{\mathbf{B}}_{10}(\mathbf{A}_0^{-1}(\epsilon)+\mathbf{B}_{00})^{-1}\widetilde{\mathbf{B}}_{01})^{-1} \text{ of order } (\sum_{j=1}^{k} t_j) \times (\sum_{j=1}^{k} t_j) \\
\mathbf{K}_{12}(\epsilon) &= -(\mathbf{A}_0^{-1}(\epsilon)+\mathbf{B}_{00})^{-1}\widetilde{\mathbf{B}}_{01}\mathbf{K}_{22}(\epsilon) \text{ of order } t_0 \times (\sum_{j=1}^{k} t_j) \\
\mathbf{K}_{21}(\epsilon) &= -(\widetilde{\mathbf{A}}^{-1}(\epsilon)+\widetilde{\mathbf{B}}_{11})^{-1}\widetilde{\mathbf{B}}_{10}\mathbf{K}_{11}(\epsilon) \text{ of order } (\sum_{j=1}^{k} t_j) \times t_0
\end{aligned}$$

We take (5.36) and (5.37) in (5.35),

$$\begin{aligned}
&\frac{d}{d\epsilon}\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F} \\
&\quad = \mathbf{F}^T \begin{pmatrix} \mathbf{K}_{11}(\epsilon) & \mathbf{K}_{12}(\epsilon) \\ \mathbf{K}_{21}(\epsilon) & \mathbf{K}_{22}(\epsilon) \end{pmatrix} \begin{pmatrix} -(1-\epsilon)^{-2}\mathbf{A}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \epsilon^{-2}\widetilde{\mathbf{A}}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{K}_{11}(\epsilon) & \mathbf{K}_{12}(\epsilon) \\ \mathbf{K}_{21}(\epsilon) & \mathbf{K}_{22}(\epsilon) \end{pmatrix} \mathbf{F} \\
&\quad = \mathbf{F}^T \begin{pmatrix} -(1-\epsilon)^{-2}\mathbf{K}_{11}(\epsilon)\mathbf{A}_0^{-1}\mathbf{K}_{11}(\epsilon)+\epsilon^{-2}\mathbf{K}_{12}(\epsilon)\widetilde{\mathbf{A}}^{-1}\mathbf{K}_{21}(\epsilon) & \\ -(1-\epsilon)^{-2}\mathbf{K}_{21}(\epsilon)\mathbf{A}_0^{-1}\mathbf{K}_{11}(\epsilon)+\epsilon^{-2}\mathbf{K}_{22}(\epsilon)\widetilde{\mathbf{A}}^{-1}\mathbf{K}_{21}(\epsilon) & \\ & -(1-\epsilon)^{-2}\mathbf{K}_{11}(\epsilon)\mathbf{A}_0^{-1}\mathbf{K}_{12}(\epsilon)+\epsilon^{-2}\mathbf{K}_{12}(\epsilon)\widetilde{\mathbf{A}}^{-1}\mathbf{K}_{22}(\epsilon) \\ & -(1-\epsilon)^{-2}\mathbf{K}_{21}(\epsilon)\mathbf{A}_0^{-1}\mathbf{K}_{12}(\epsilon)+\epsilon^{-2}\mathbf{K}_{22}(\epsilon)\widetilde{\mathbf{A}}^{-1}\mathbf{K}_{22}(\epsilon) \end{pmatrix} \mathbf{F}
\end{aligned}$$

Since that,

$$\begin{aligned}
\lim_{\epsilon\to 0^+} \mathbf{K}_{11}(\epsilon) &= (\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1} \\
\lim_{\epsilon\to 0^+} \epsilon^{-1}\mathbf{K}_{22}(\epsilon) &= \lim_{\epsilon\to 0^+} \epsilon^{-1}(\widetilde{\mathbf{A}}^{-1}(\epsilon)+\widetilde{\mathbf{B}}_{11})^{-1} = \widetilde{\mathbf{A}} \\
\lim_{\epsilon\to 0^+} \mathbf{K}_{12}^T(\epsilon) = \lim_{\epsilon\to 0^+} \mathbf{K}_{21}(\epsilon) &= \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}_{10}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}
\end{aligned}$$

so we have

$$\frac{d}{d\epsilon}\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F}\mid_{\epsilon=0^+}$$
$$= \mathbf{F}^T\begin{pmatrix} -(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\mathbf{A}_0^{-1}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}+(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\widetilde{\mathbf{B}}_{01}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}_{10}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1} & \\ -\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}_{10}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1} & \\ & -(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\widetilde{\mathbf{B}}_{12}\widetilde{\mathbf{A}} \\ & \widetilde{\mathbf{A}} \end{pmatrix}\mathbf{F}$$
$$= -\mathbf{F}_0^T(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\mathbf{A}_0^{-1}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\mathbf{F}_0+\mathbf{F}_0^T(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\widetilde{\mathbf{B}}_{01}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}_{10}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\mathbf{F}_0$$
$$-\widetilde{\mathbf{F}}^T\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}_{10}(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\mathbf{F}_0-\mathbf{F}_0^T(\mathbf{A}_0^{-1}+\mathbf{B}_{00})^{-1}\widetilde{\mathbf{B}}_{01}\widetilde{\mathbf{A}}\widetilde{\mathbf{F}}+\widetilde{\mathbf{F}}^T\widetilde{\mathbf{A}}\widetilde{\mathbf{F}}$$

We denote the last term in the above with T in the reminder to avoid a long expression, so

$$f'(\epsilon)\mid_{\epsilon=0^+}=\det[\mathfrak{M}(\xi)]\cdot tr\{\mathfrak{M}^{-1}(\xi)\cdot T\} \tag{5.38}$$

Regarding to concavity of $f(\epsilon)$, $f'(0^+)=f'(\epsilon)\mid_{\epsilon=0^+}$. Note that if $\mathbf{B}$ is singular, then we consider $\mathbf{B}=\lim_{\gamma\to 0}(\mathbf{B}+\gamma\mathbf{I})$ and we follow the same way as the above case. □

If we consider $\Psi(\xi)=\Phi(\mathfrak{M}(\xi))=-\log(\det(\mathfrak{M}(\xi)))$ then the following theorem guarantees the Fréchet differentiability of $\Psi(\xi)$ at ξ_1 in the direction of ξ_2.

Theorem 5.7.2. *If $\Psi(\xi)=\Phi(\mathfrak{M}(\xi))=-\log(\det(\mathfrak{M}(\xi)))$ is the D-optimality criterion, then $F_\Psi(\xi,\xi')$ is linear in the second argument. In other words, if $\xi'=\sum_{i=1}^{k}a_i\xi_i$*

$$F_\Psi(\xi,\xi')=\sum_{i=1}^{k}a_iF_\Psi(\xi,\xi_i)$$

where $F_\Psi(\xi,\xi')$ is the Fréchet derivative of $\Psi(\xi)$ at ξ in the direction of ξ'.

Proof.

$$\begin{aligned} F_\Psi(\xi,\xi') &= \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}[\Phi(\mathfrak{M}((1-\epsilon)\xi+\epsilon\xi'))-\Phi(\mathfrak{M}(\xi))] \\ &= \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\{\log(\det(\mathfrak{M}((1-\epsilon)\xi+\epsilon\xi')))-\log(\det(\mathfrak{M}(\xi)))\} \\ &= \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\{\log(\det(\mathfrak{M}((1-\epsilon)\xi+\epsilon\xi')(\mathfrak{M}(\xi))^{-1}))\} \\ &= \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\{\log(g(\epsilon))\end{aligned}$$

where

$$g(\epsilon)=\det(\mathfrak{M}((1-\epsilon)\xi+\epsilon\xi'))\det(\mathfrak{M}(\xi))^{-1}$$

We can represent the above expression as $g(\epsilon) = c \cdot f(\epsilon)$, where $c = \det(\mathfrak{M}(\xi))^{-1}$ is a constant and $f(\epsilon)$ has the same definition as in Lemma 5.7.3. where $\mathbf{F}^T = \begin{pmatrix} \mathbf{F}_0^T & \mathbf{F}_1^T & \dots & \mathbf{F}_k^T \end{pmatrix} = \begin{pmatrix} \mathbf{F}_0^T & \widetilde{\mathbf{F}}^T \end{pmatrix}$, $\mathbf{A}(\epsilon) = \begin{pmatrix} \mathbf{A}_0(\epsilon) & \\ & \widetilde{\mathbf{A}}(\epsilon) \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{00} & \widetilde{\mathbf{B}}_{01} \\ \widetilde{\mathbf{B}}_{10} & \widetilde{\mathbf{B}}_{11} \end{pmatrix}$ are the same matrices as before in Lemma 5.7.3.

With regard to the Lemma 5.7.3, $g(\epsilon)$ is continuous, concave and differentiable at the right hand neighborhood of $\epsilon = 0$ ($\epsilon = 0^+$), and hence the Taylor expansion of $g(\epsilon)$ around $\epsilon = 0^+$ is

$$\begin{aligned} g(\epsilon) &= \lim_{\epsilon \to 0^+} c \cdot f(\epsilon) + (\lim_{\epsilon \to 0^+} c \cdot \frac{d}{d\epsilon} f(\epsilon))\epsilon + o(\epsilon) \qquad (5.39) \\ &= 1 + \epsilon \cdot tr(c \cdot T) + o(\epsilon) \qquad (5.40) \end{aligned}$$

The last expression holds because of $\lim_{\epsilon \to 0^+} g(\epsilon) = 1$.

$$\begin{aligned} F_\Psi(\xi, \sum_{i=1}^k a_i \xi_i) &= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \log(1 + \epsilon \cdot tr(c \cdot T) + o(\epsilon)) \\ &= \lim_{\epsilon \to 0^+} \frac{1}{\epsilon}(T\epsilon) = tr(c \cdot T) \end{aligned}$$

The last equality holds because of $\log(1 + t) = t + o(t)$.

$$\begin{aligned} F_\Psi(\xi, \sum_{i=1}^k a_i \xi_i) &= tr\{[-\mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{A}_0^{-1}(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0 \\ &\quad + \mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1} \sum_{i=1}^k a_i \mathbf{B}_{0i}\mathbf{A}_i\mathbf{B}_{i0}(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0 \\ &\quad - \sum_{i=1}^k a_i \mathbf{F}_i^T \mathbf{A}_i \mathbf{B}_{i0}(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0 \\ &\quad - \mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1} \sum_{i=1}^k a_i \mathbf{B}_{0i}\mathbf{A}_i\mathbf{F}_i + \sum_{i=1}^k a_i \mathbf{F}_i^T \mathbf{A}_i \mathbf{F}_i] \cdot (\mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0)\} \\ &= \sum_{i=1}^k a_i tr\{[-\mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{A}_0^{-1}(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0 \\ &\quad + \mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{B}_{0i}\mathbf{A}_i\mathbf{B}_{i0}(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0 \\ &\quad - \mathbf{F}_i^T \mathbf{A}_i \mathbf{B}_{i0}(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0 \\ &\quad - \mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{B}_{0i}\mathbf{A}_i\mathbf{F}_i + \mathbf{F}_i^T \mathbf{A}_i \mathbf{F}_i] \cdot (\mathbf{F}_0^T(\mathbf{A}_0^{-1} + \mathbf{B}_{00})^{-1}\mathbf{F}_0)\} \\ &= \sum_{i=2}^k a_i F_\Psi(\xi, \xi_i) \end{aligned}$$

□

We are now ready to give an equivalence theorem for the D-criterion for the model with random slope. The following theorem does not give us a closed form of equivalence theorem to check whether a design is D-optimal or not, but it prepares an expression to check it computationally.

Theorem 5.7.3. *The design* $\xi_* = \left\{ \begin{array}{ccc} x_1^* & \dots & x_t^* \\ p_1^* & \dots & p_t^* \end{array} \right\}$ *is D-optimal for the Poisson regression model with random slope if and only if*

$$tr\{m\mu_x \mathbf{f}(x)\mathfrak{M}^{-1}(\xi_*)\mathbf{f}^T(x) - \mathfrak{M}^{-1}(\xi_*)[\mathbf{F}_*^T(\mathbf{A}_*^{-1}+\mathbf{B}_{**})^{-1}(\mathbf{I}+\mathbf{B}_{**}\mathbf{A}_*)^{-1}\mathbf{F}_* - \mathbf{F}_*^T(\mathbf{A}_*^{-1}+\mathbf{B}_{**})^{-1}\mathbf{B}_{*x}\mathbf{B}_{x*}(\mathbf{A}_*^{-1}+\mathbf{B}_{**})^{-1} + m\mu_x\mathbf{f}(x)\mathbf{B}_{x*}(\mathbf{A}_*^{-1}+\mathbf{B}_{**})^{-1} + m\mu_x\mathbf{F}_*^T(\mathbf{A}_*^{-1}+\mathbf{B}_{**})^{-1}\mathbf{B}_{*x}\mathbf{f}^T(x)]\} \leq 0 \text{ for all } x \in \tau$$

where $\mathbf{F}_*$ *is the design matrix of* ξ_*, $\mathbf{A}_* = m \cdot diag\{p_1^*\mu_1^*, \dots, p_t^*\mu_t^*\}$ *with* $\mu_i^* = \exp(\beta_0 + \beta_1 x_i^* + \frac{1}{2}\sigma^2 x_i^{*^2})$, $\mu_x = \exp(\beta_0 + \beta_1 x + \frac{1}{2}\sigma^2 x^2)$, $\mathbf{B}_{x*} = \mathbf{B}_{*x}^T = \left(\begin{array}{ccc} e^{\sigma^2 x_1^* x} - 1 & \dots & e^{\sigma^2 x_t^* x} - 1 \end{array} \right)$ *and* $\mathbf{B}_{**} = \left(e^{\sigma^2 x_i^* x_j^*} - 1 \right)_{i,j}^t$ *of order* $t \times t$.

Proof. Consider the D-criterion, $\Psi(\xi) = -\log(\det[\mathfrak{M}(\xi)])$ then with regard to theorem 5.7.2 $\Psi(\xi)$ is Fréchet differentiable. So

$$F_\Psi(\xi_*, \xi) = \frac{d}{d\epsilon}\Phi(\mathfrak{M}((1-\epsilon)\xi_* + \epsilon\xi)) \mid_{\epsilon=0}$$

Since $\Phi(\mathfrak{M}((1-\epsilon)\xi_* + \epsilon\xi)) = -\log(\det(\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F}))$, thus

$$\begin{aligned} \frac{d}{d\epsilon}\Phi(\mathfrak{M}((1-\epsilon)\xi_* + \epsilon\xi)) &= -tr[(\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F})^{-1}\frac{d}{d\epsilon}(\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F})] \\ &= -tr[(\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F})^{-1}\{\mathbf{F}^T(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{A}^{-1}(\epsilon)\mathbf{A}\mathbf{A}^{-1}(\epsilon)(\mathbf{A}^{-1}(\epsilon)+\mathbf{B})^{-1}\mathbf{F}\}] \end{aligned}$$

Note that $\mathbf{A}(\epsilon) = \begin{pmatrix} (1-\epsilon)\mathbf{A}_* & \mathbf{0} \\ \mathbf{0} & \epsilon\mathbf{A}_1 \end{pmatrix}$ and $\mathbf{A} = \begin{pmatrix} -\mathbf{A}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix}$, where $\mathbf{A}_*$ and $\mathbf{A}_1$ are diagonal matrices with the weighted response means $(m_{ij}\mu_{ij})$ as diagonal elements corresponding to the ξ_* and the ξ respectively. The first equality holds according to the derivative of the determinant (Schott (1997), Corollary 8.1.1) and the second equality holds because of theorem 8.2 in Schott(1997). After some matrix algebra similar to the proof of theorem (5.7.2) the result follows. □

5.8 G-Optimality

We considered the D-criterion as a measure to compare different designs all over the previous sections. The demand for G-optimality is arising as well as applying D-optimality.

Unfortunately, due to the nonlinear structure of the prediction of the response variable in the two mentioned classes (Poisson regression with random intercept and Poisson regression with random slope), the variance of the predictor can not be obtained by the linear model techniques. On the other hand, we have obtained that the quasi likelihood estimators are asymptotically normal distributed (see section (3.3)). These two reasons are sufficient to suggest an asymptotic method to find the variance of the predictor of Y_x, where Y_x is the response variable at x and we eliminated index in response and experimental point just for simplicity in writing. This asymptotic method is called the ***delta method*** (or δ-method) (Rao (1973), page 388 or Casella and Berger (2002)).
Let $\hat{\boldsymbol{\beta}}^{(n)}$ be the quasi-likelihood estimator of $\boldsymbol{\beta}$, then with regard to (3.9), $\hat{\boldsymbol{\beta}}^{(n)}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ equals to the inverse of quasi-information matrix of $\boldsymbol{\beta}$. According to the δ- method theorem if g(.) be any function with continuous first partial derivative then

$$g(\hat{\boldsymbol{\beta}}^{(n)}) - g(\boldsymbol{\beta}) \xrightarrow{D} N(0, \nabla_g^T \boldsymbol{\Sigma} \nabla_g) \tag{5.41}$$

where $\nabla_g^T = (\frac{dg(\boldsymbol{\beta})}{d\beta_1}, \ldots, \frac{dg(\boldsymbol{\beta})}{d\beta_p})$. In other words,

$$\text{asymptotic } Var(g(\hat{\boldsymbol{\beta}}^{(n)})) = \nabla_g^T \boldsymbol{\Sigma} \nabla_g. \tag{5.42}$$

We are now ready to find the variance of the response predictor for two proposed classes separately. We consider again Poisson regression model with random intercept (4.1), so

$$\hat{Y}_x = e^{\mathbf{f}(x)^T \hat{\boldsymbol{\beta}} + \frac{1}{2}\sigma^2}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$. Using the (5.42), let $g(\hat{\boldsymbol{\beta}}) = e^{\mathbf{f}(x)^T \hat{\boldsymbol{\beta}} + \frac{1}{2}\sigma^2}$, we have

$$\text{as. } Var(\hat{Y}_x) = \mu_x^2 \mathbf{f}^T(x) \mathfrak{M}_{\boldsymbol{\beta}}^{-1}(\xi) \mathbf{f}(x) \text{ and } \mu_x = e^{\beta_0 + \beta_1 x + \frac{1}{2}\sigma^2}$$

For simplicity we omit the superscript (n). So, ξ^* is G-optimality design if

$$\xi^* = \arg \min_{\xi_I \in \Xi} \max_{x \in \tau} \mu_x^2 \mathbf{f}^T(x) \mathfrak{M}_{\boldsymbol{\beta}}^{-1}(\xi_I) \mathbf{f}(x)$$

where ξ_I is the experimental design to estimate the parameters of the Poisson regression model with random intercept model.
For the Poisson regression model with random slope (4.21)

$$\hat{Y}_x = e^{\hat{\beta}_0 + \hat{\beta}_1 x + \frac{1}{2}\sigma^2 x^2}$$

If we consider $g(\hat{\boldsymbol{\beta}}) = e^{\hat{\beta}_0 + \hat{\beta}_1 x + \frac{1}{2}\sigma^2 x^2}$ then using the (5.42), we have

$$\text{as. } Var(\hat{Y}_x) = \mu_x^2 (1 \quad x) \mathfrak{M}_{\boldsymbol{\beta}}^{-1}(\xi) \begin{pmatrix} 1 \\ x \end{pmatrix} \text{ and } \mu_x = E(Y_X) = e^{\beta_0 + \beta_1 x + \frac{1}{2}\sigma^2 x^2}$$

So, ξ^* is G-optimality design if

$$\xi^* = \arg \min_{\xi_S \in \Xi} \max_{x \in \tau} \mu_x^2 (1 \quad x) \mathfrak{M}_{\boldsymbol{\beta}}^{-1}(\xi_S) \begin{pmatrix} 1 \\ x \end{pmatrix}$$

Here ξ_S stands for the experimental design for the Poisson regression with random slope model.
From these expressions for as. $Var(\hat{Y}_x)$, it becomes evident that even for simple Poisson regression with random intercept, i.e. model (4.17), we can not obtain a closed form for $Var(\hat{Y}_x)$.
In the next chapter we will consider some numerically results on this subject.

6 Some Results

6.1 Introduction

The main purpose of this chapter is to present some applications of the theory in the last two chapters, in other words we would like to obtain some locally optimal designs. The most popular criterion, the D-criterion, is considered and we obtain the locally D-optimal designs for some selected values of parameters.
In the next two next sections, we consider once again the simple and quadratic Poisson regression models with the random intercept. The locally D-optimality will be obtained in these cases. The structure of these two models are different from the model with random slope. In section 6.4 the locally D-optimal design will be described for the Poisson regression model with random slope. The relation to the local G-optimality will be discussed throughout these sections.

6.2 Locally D-optimal Design for Simple Poisson Regression with Random Intercept

In a recent manuscript Yanping et al. (2006) obtain the locally D-optimal design in a Poisson regression model without random effects. In this section we consider a simple Poisson regression model with random intercept

$$Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \text{ where } \mu_{ij}(b_i) = \exp(b_i + \beta_0 + \beta_1 x_j) \tag{6.1}$$

to see the effect of blocks and, hence, of intra-individual correlation. We ignore the index i in the settings, i.e. $x_{ij} = x_j$, because of the fact that the optimal designs can be found among those which are uniform across the individuals.
We want to find locally D-optimal designs to estimate $\boldsymbol{\beta}$, which maximize the determinant of the information matrix.
In most applications of this model, like bioscience, pharmacokinetics etc., the design region is the non-negative real line or a subset of that. In other words $\tau = [h, \infty)$ is considered as an unbounded subset and $\tau = [h, g]$ is considered as a bounded subset, where $h \geq 0$ and $g > h$ determine the bounds for the design region, which has to be defined by experimenter. In this text we consider only the cases with positive design regions. The expectation $\mu_j = \mu(x_j) = e^{\frac{1}{2}\sigma^2 + \beta_0 + \beta_1 x_j}$ of Y_{ijk}, is a monotone function of x_j. We consider the special case, where $\mu(x_j)$ is considered as a decreasing function of x_j, i.e., $\beta_1 < 0$. Therefore the maximum and the minimum of the mean response are attained at the lower and

upper bound of the design region respectively. Let $\mu(h) = e^{\beta_0+\beta_1 h+\frac{1}{2}\sigma^2}$ be the expectation of Y_{ijk} at h, the canonical standardized mean $\tilde{\mu}_j = \tilde{\mu}(x_j) = \frac{\mu_j}{\mu(h)} = e^{\beta_1(x_j-h)}$, will always lie in $(0, 1]$ and $[\tilde{\mu}(g), 1]$ respectively corresponding to the design regions $[h, \infty)$ and $[h, g]$.

We outline the following lemma which allows us to restrict ourselves to designs with only two different settings x_1 and x_2.

Lemma 6.2.1. *(Niaparast (2009)) For the model* (6.1)*, the D-optimal design* $\zeta^* = \left\{ \begin{array}{c} \xi^* \\ 1 \end{array} \right\}$ *has exactly two different support points* x_1^* *and* x_2^**, i.e.,* $\xi^* = \left\{ \begin{array}{cc} x_1^* & x_2^* \\ p_1^* & p_2^* \end{array} \right\}$ *where* $p_1^*, p_2^* > 0$.

Proof. As we have seen in theorem (5.6.1) a necessary and sufficient condition for ξ^* to be D-optimal for our model is

$$m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x) \leq tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)] \quad \forall x \in \tau$$

We represent the above inequality in the following form

$$\mu(x) \left(\begin{array}{cc} 1 & x \end{array} \right) \left(\begin{array}{cc} a & b \\ b & c \end{array} \right) \left(\begin{array}{c} 1 \\ x \end{array} \right) \leq d \tag{6.2}$$

with appropriate constant a, b, c, d. Therefore the inequality (6.2) is equivalent to

$$\mu(x)(cx^2 + 2bx + a) \leq d \Longleftrightarrow ax^2 + bx + c \leq \frac{d}{\mu(x)}$$

Suppose that the D-optimal design has at least three support points $z_1 < z_2 < z_3$, i.e. $h(z_i) = k(z_i)$ $(i = 1, 2, 3)$ where $h(x) = \frac{1}{d}(cx^2+2bx+a)$ and $k(x) = \frac{1}{\mu(x)}$. Applying Rolle's theorem, there are z_1', z_2' such that $z_1 < z_1' < z_2 < z_2' < z_3$ and $h'(z_i') = k'(z_i')$ $(i = 1, 2)$. Because $h(z) \leq k(z)$ for all z, the points $z_i (i = 1, 2, 3)$ are local extrema, then $h'(z_2) = k'(z_2)$. By Rolle's theorem again for the function $h'(z)-k'(z)$, we receive points z_1'', z_2'', such that $z_1' < z_1'' < z_2 < z_2'' < z_2'$ and $h''(z_i'') = k''(z_i'')(i = 1, 2)$. Because $h''(z) = 2\frac{c}{d}$ for all z, we have $k''(z_i'') = 2\frac{c}{d}$ $(i = 1, 2)$. Since that in the model (model (6.1))$k(x) = e^{-(\beta_0+\beta_1 x+\frac{1}{2}\sigma^2)}$ and hence $k''(x) = \beta_1^2 e^{-(\beta_0+\beta_1 x+\frac{1}{2}\sigma^2)}$ and $k''(x) = r$ (for all r) has at most one root, thus it is in contrary to result that $k''(x) = 2\frac{c}{d}$ has two roots and the claim follows. Similar idea can be found in Biedermann et al. (2006). □

Using this lemma, we can restrict ourselves to the case $t = 2$. In other words, this lemma reduce our search to the the experimental setting with two points.

Lemma 6.2.2. *Consider the model* (6.1)*. In terms of the canonical standardized mean, the D-criterion for a design* $\xi = \left\{ \begin{array}{cc} x_1 & x_2 \\ p_1 & p_2 \end{array} \right\}$ *to estimate* $\boldsymbol{\beta}$ *depends on the parameters only through* $\gamma(m, \beta_0(h), \sigma^2) = me^{\beta_0(h)+\frac{1}{2}\sigma^2}(e^{\sigma^2} - 1)$ *as follows.*

$$\det(\mathfrak{M}(\xi)) \propto \frac{p_1(1-p_1)\tilde{\mu}_1\tilde{\mu}_2(\ln(\tilde{\mu}_1)-\ln(\tilde{\mu}_2))^2}{1+\gamma(m,\beta_0(h),\sigma^2)(p_1\tilde{\mu}_1+p_2\tilde{\mu}_2)} \tag{6.3}$$

where $\beta_0(h) = \beta_0 + \beta_1 h$ *is the mean response at the lower bound* $x = h$ *of the design region.*

Proof. Let $m_i = mp_i$ $(i = 1, 2)$, from (4.14) the information matrix for model (6.1) is

$$\begin{aligned}
\mathfrak{m}(\xi) &= \begin{pmatrix} 1 & 1 \\ x_1 & x_2 \end{pmatrix} \left(\begin{pmatrix} m_1\mu_1 & 0 \\ 0 & m_2\mu_2 \end{pmatrix}^{-1} + (e^{\sigma^2}-1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \\
&= \begin{pmatrix} \sum_{j=1}^{2} m_j\mu_j & \sum_{j=1}^{2} m_j\mu_j x_j \\ \sum_{j=1}^{2} m_j\mu_j x_j & \sum_{j=1}^{2} m_j\mu_j x_j^2 \end{pmatrix} \\
&\quad - \frac{(e^{\sigma^2}-1)}{1+(e^{\sigma^2}-1)\sum_{j=1}^{2} m_j\mu_j} \begin{pmatrix} \sum_{j=1}^{2} m_j\mu_j \\ \sum_{j=1}^{2} m_j\mu_j x_j \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{2} m_j\mu_j & \sum_{j=1}^{2} m_j\mu_j x_j \end{pmatrix} \\
&= \frac{1}{1+(e^{\sigma^2}-1)(m_1\mu_1+m_2\mu_2)} \\
&\quad \times \begin{pmatrix} m_1\mu_1+m_2\mu_2 & m_1\mu_1x_1+m_2\mu_2x_2 \\ m_1\mu_1x_1+m_2\mu_2x_2 & m_1\mu_1x_1^2+m_2\mu_2x_2^2+(e^{\sigma^2}-1)m_1\mu_1m_2\mu_2(x_1-x_2)^2 \end{pmatrix}
\end{aligned}$$

A straightforward calculation leads to

$$\det(\mathfrak{m}(\xi)) = \frac{m_1\mu_1m_2\mu_2(x_1-x_2)^2}{1+(e^{\sigma^2}-1)\sum_{j=1}^{2} m_j\mu_j} = \frac{m^2\mu_h^2p_1p_2\tilde{\mu}_1\tilde{\mu}_2(x_1-x_2)^2}{1+(e^{\sigma^2}-1)m\mu(h)\sum_{j=1}^{2} p_j\tilde{\mu}_j}$$

As $(x_1 - x_2) = \frac{\log\tilde{\mu}_1 - \log\tilde{\mu}_2}{\beta_1}$, the representation follows □

Theorem 6.2.1. *Let* $\tau = [h, \infty)$. *If* $\tilde{\mu}_1^*$, $\tilde{\mu}_2^*$, p_1^*, $p_2^* = 1 - p_1^*$ *maximize the expression* (6.3), *the D-optimal design for model* (6.1) *is given by* $\xi^* = \begin{Bmatrix} x_1^* & x_2^* \\ p_1^* & p_2^* \end{Bmatrix}$. *where* $x_j^* = \frac{1}{\beta_1}\log\tilde{\mu}_j^* + h$, *and* $\tilde{\mu}_1^*$, $\tilde{\mu}_2^*$, p_1^*, $p_2^* = 1 - p_1^*$ *maximize the expression* (6.3).

Proof. the proof is immediately obtained using the lemmas 6.2.1 and 6.2.2. □

According to Theorem 6.2.1, numerical methods can be used to maximize this criterion in order to find D-optimal designs. The D-optimal design for some representative values

Table 6.1: D-optimal designs for model (6.1)

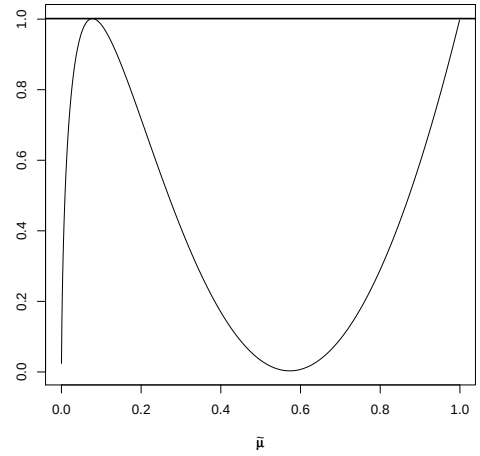
$\gamma(m,\beta_0(h),\sigma^2)$	p_1^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\gamma(m,\beta_0(h),\sigma^2)$	p_1^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$
0	0.500	0.1353	1	30	0.746	0.0864	1
0.5	0.543	0.1279	1	60	0.762	0.0825	1
2	0.609	0.1155	1	100	0.769	0.0807	1
5	0.665	0.1044	1	1000	0.781	0.0779	1
10	0.703	0.0962	1	10000	0.782	0.0776	1

of $\gamma(m,\beta_0(h),\sigma^2)$ are listed in Table 6.1 .

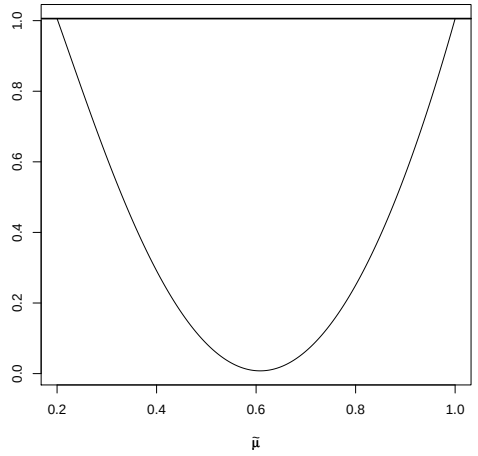
Note that $\tilde{\mu}_2^* = 1$ corresponds to $x_2^* = h$ which means that one point in the optimal setting will be at the lower bound of the design region. For the Fixed Effect Simple Poisson Regression models, i.e. $\sigma^2 = 0$, the constant $\gamma(m,\beta_0(h),\sigma^2)$ is equal to zero for all m and $\beta_0(h)$. In this case we find that the D-optimal design is the design with 50% to the experimental runs at $x_1^* = \frac{1}{\beta_1}\ln(0.1353) + h$ for all $\beta_1 < 0$ as long as $x_1^* \leq g$ and the remaining at $x_2^* = h$, in accordance with the results in Yanping et al. (2006).
$\gamma(m,\beta_0(h),\sigma^2)$ is an increasing function of σ^2 and it can also be easily seen that the intra-individual correlation , $corr_{\sigma^2}(Y_{ijk}, Y_{ij'k'})$, is an increasing function of σ^2. So for fixed m and $\beta_0(h)$ increasing in $corr_{\sigma^2}(Y_{ijk}, Y_{ij'k'})$ is equivalent to increasing in $\gamma(m,\beta_0(h),\sigma^2)$. From Table 6.1 it can be seen that a larger value of σ^2 and, hence, of the intra-individual correlation decreases the proportion of observations at $x_2^* = h$. When σ^2 tends to infinity, 78% of the experiments should be run at $x_1^* = \frac{1}{\beta_1}\ln(0.0776) + h$ and 22% at $x_2^* = h$. These results coincide with the results to find the D-optimal design for estimation of the slope in the corresponding model with fixed block effects (Minkin (1993)).
Theoretically, x_i can tend to ∞ and hence $\tilde{\mu}_i$ can be as small as 0. But this is rarely occurs in practice. In real experiment, a design point with very small $\tilde{\mu}$ is not reasonable, so it is more practical to consider design on a restricted (bounded) region. In the case of a restricted region $\tau = [h, g]$, the D-optimal design is the same as the D-optimal in the unrestricted case, if $\tilde{\mu}(g) \leq \tilde{\mu}_1^*$. Otherwise, $\tilde{\mu}(g)$ and μ_2^*, i.e., $x_1^* = g$ and $x_2^* = h$ will be optimal values but with different weights p_1^* and p_2^*.
We have evaluated the sensitivity function,

$$\phi(x,\xi^*) = m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x)$$

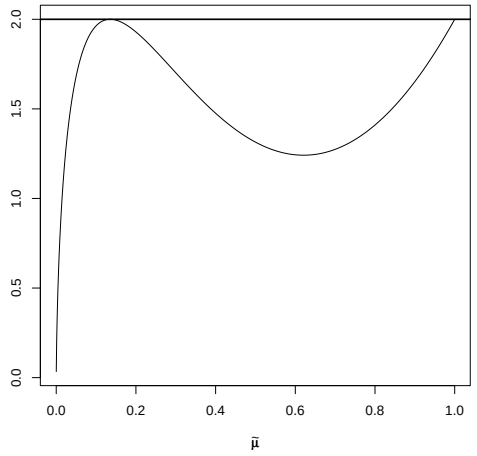
over the experimental domain for this model. We have found that this expression achieves the maximum of $tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)]$ at the experimental setting points of the locally optimal design for some special cases.
Figure 6.1 shows the locally D-optimal design for the first order Poisson regression model with random intercept when $\gamma(m,\beta_0(h),\sigma^2) = 1000$ for two special cases: (i) unrestricted region or $\tilde{\mu} \in (0, 1]$ and (ii) restricted region or $\tilde{\mu} \in (0.2, 1]$ corresponding to these cases for the model without random intercept, i.e. $\gamma(m,\beta_0(h),\sigma^2) = 0$ respectively. For the first two cases, figures 6.1(a) and 6.1(b), the sensitivity function attains the maximum



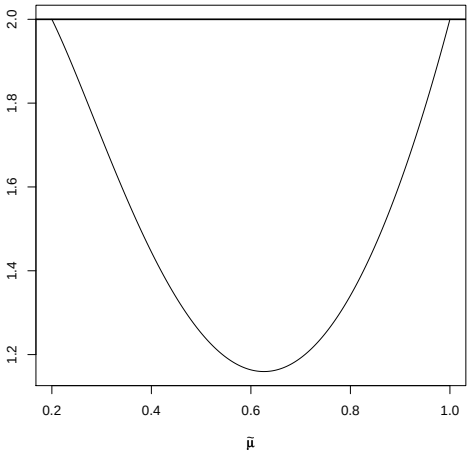
(a) unrestricted region ; $\gamma(m, \beta_0(h), \sigma^2) = 1000$



(b) restricted region ;$\gamma(m, \beta_0(h), \sigma^2) = 1000$



(c) unrestricted region ; $\gamma(m, \beta_0(h), \sigma^2) = 0$



(d) restricted region ;$\gamma(m, \beta_0(h), \sigma^2) = 0$

Figure 6.1: Locally D-optimal designs for the model (6.1): (a) $\tilde{\mu} \in (0, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 1000$; (b) $\tilde{\mu} \in [0.2, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 1000$; (c): $\tilde{\mu} \in (0, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 0$ (without random intercept); (d)$\tilde{\mu} \in [0.2, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 0$ (without random intercept).

$tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)] = 1.001$ at $\tilde{\mu}_1^* = 0.0779$ and $\tilde{\mu}_2^* = 1$ whereas for the cases without random effects, the maximum of sensitivity function, $\phi(x, \xi^*) = m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathbf{f}(x)$ is 2, the number of parameters.
The coincidence between G- and D- optimality is an immediate consequent of the Kiefer-Wolfowitz (1959) equivalence theorem for the fixed effects linear models. Schwabe and Schmelter (2008) have considered a linear model with random intercept which is an special case of L.M.M. They have indicated that the equivalence of D- and G-optimality is not retained even for a simple linear model with random intercept. The similar result holds for the simple linear model with random slope which has been indicated in Schmelter et al. (2007).
If we confine the search for G-optimal designs to those with observation at x_1^* and x_2^* corresponding to μ_1^* and μ_2^* respectively, then candidates for an G-optimal design are characterized by p_1, the proportion of observations at μ_1^*. The quasi-information is evaluated at μ_1^* and μ_2^*. $\xi_G^* = \left\{ \begin{array}{cc} x_1^* & x_2^* \\ p_1^{'*} & p_2^{'*} \end{array} \right\}$ is G-optimal if

$$\xi_G^* = \arg \min_{\xi \in \Xi_G} \max_x Var(\hat{Y}_x)$$

where Ξ_G is the set of experimental design ξ_G such that $\xi_G = \left\{ \begin{array}{cc} x_1^* & x_2^* \\ p_1' & p_2' \end{array} \right\}$. There is no analytical solution to find ξ_G^*. Figure 6.2 gives a general view how the proportion varies in terms of σ^2: For $\sigma^2 = 0$, $p_1^* = p_1^{'*}$ which is in accordance with the celebrated equivalence theorem for fixed effects models. If $\sigma^2 \geq 0$ then p_1^* and $p_1^{'*}$ go far away as σ^2 is increasing.
As we have described before, in real experimental design, it is more reasonable to consider the bounded regions, i.e. $\tilde{\mu}_j \in [\tilde{\mu}(g), 1]$. Experimenters might consider a standard two points design which consist of the two endpoints, i.e. $\tilde{\mu}_1 = \tilde{\mu}(g)$ and $\tilde{\mu}_2 = 1$ with equal allocation, i.e. $p_1 = p_2 = \frac{1}{2}$. We define the D-efficiency as

$$\text{D-efficiency}(\xi) = \left(\frac{\det(\mathfrak{M}_{\boldsymbol{\beta}}(\xi))}{\det(\mathfrak{M}_{\boldsymbol{\beta}}(\xi^*))} \right)^{\frac{1}{p}}$$

where ξ^* is the D-optimal design in the corresponding model.
The results in Table 6.3 indicate that, for the case with $\tilde{\mu}(g) = 0.01$ which is nearly unrestricted case, the standard two points design $\xi_0 = \left(\begin{array}{cc} h & g \\ 0.5 & 0.5 \end{array} \right)$ is not robust to the values of $\gamma(m, \beta_0(h), \sigma^2)$ whereas it is robust for the case with $\tilde{\mu}(g) = 0.2$. In fact in the latter case the experimental design points of the standard design and of the optimal design coincide while they only have different proportions.
Also we present a graphical view (Figure 6.3) for the D-efficiency of the standard design ξ_0 for different values of $\gamma(m, \beta_0(h), \sigma^2)$.

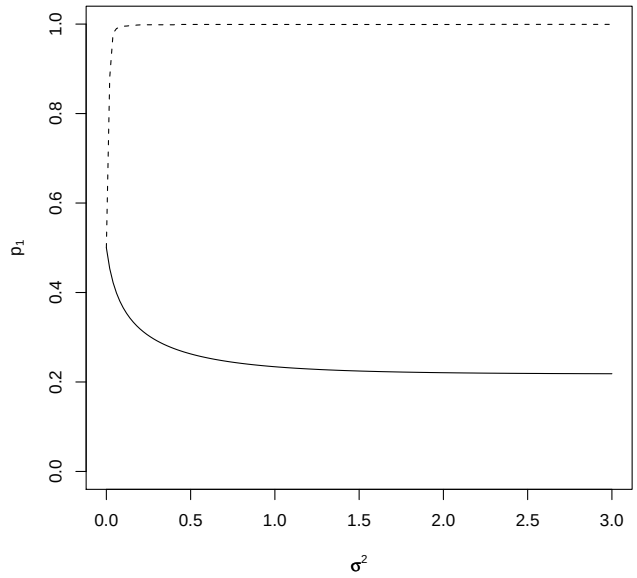


Figure 6.2: Optimal proportions of observations at x_1^* versus σ^2. D-optimal proportion (solid line) and G-optimal proportion (dashed line).

Table 6.2: D-efficiency of the standard design for the simple Poisson regression model with random intercept

$\gamma(m, \beta_0(h), \sigma^2)$	D-efficiency ($\tilde{\mu}(g) = 0.01$)	D-efficiency ($\tilde{\mu}(g) = 0.2$)
0	0.6259	1.0000
0.5	0.6311	0.9970
2	0.6291	0.9827
5	0.6187	0.9659
10	0.6081	0.9541

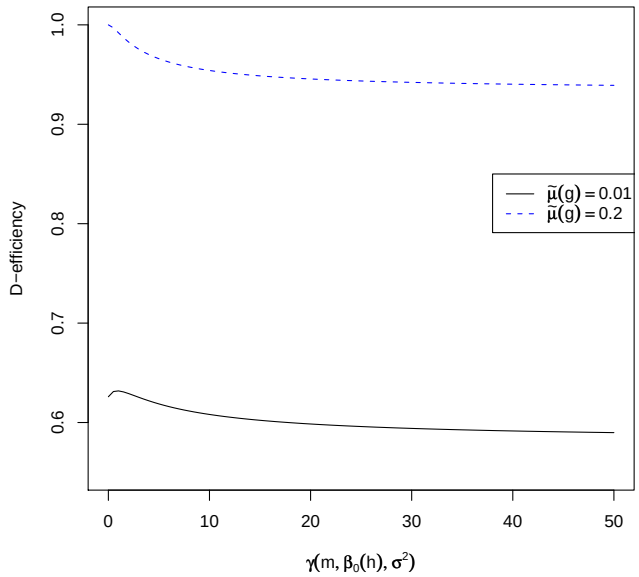


Figure 6.3: D-efficiency of the standard design ξ_0 in dependence on γ.

6.3 Locally D-optimal Designs for the Quadratic Poisson Regression Model with Random Intercept

In the previous section we focused on the first order model of Poisson regression as defined in (6.1). However the effect of explanatory variables sometimes are stronger than that which the model (6.1) describe as the relation between the explanatory variable and response variable. Thus the quadratic model as we indicated in (4.19) may be suitable. We consider again the quadratic Poisson regression model with random intercept

$$Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \text{ where } \mu_{ij}(b_i)) = \exp(b_i + \beta_0 + \beta_1 x_j + \beta_2 x_j^2). \tag{6.4}$$

We assume here that the designs used for different individuals are the same and hence we do not consider index i in the experimental settings. We suppose that the expectation of Y_{ijk}, $\mu_j = \mu(x_j) = e^{\frac{1}{2}\sigma^2+\beta_0+\beta_1 x_j+\beta_1 x_j^2}$, is a monotone function on the design region, an assumption in accordance with the most applications in real experiments. Without loss of generality, we consider the function $\mu(x_j)$ to be decreasing.
We also assume that the design region is restricted to the non-negative real numbers or a subset of that as in the first order model with random intercept.

Remark 6.1. *All design regions are considered in this text are bounded below. The restricted design regions are denoted by $[h, g]$ and the unrestricted design region by $[h, \infty)$.*

Let $\tilde{\mu}_j = \tilde{\mu}(x_j) = \frac{\mu_j}{\mu(h)} = e^{\beta_1(x_j-h)+\beta_2(x_j^2-h^2)}$ be the canonical standardized mean at x_j, where $x_j \in [h, g]$. For sake of simplicity, we suppose $h = 0$. As we have seen in the first order model , $\beta_1 \leq 0$ guarantees that $\mu(x_j)$ is a decreasing function of x_j. From a descriptive point of view, this assumption should carry over to quadratic model. Also, β_2 should not be as much large positive number so that the quadratic model is capable to describe a stronger effect than the first order model. Mathematically, the signs of β_0 and β_1 can be derived from the assumption that $\tilde{\mu}(x_j)$ is a decreasing function of x_j.
We reconsider

$$\tilde{\mu}(x) = e^{\beta_1 x+\beta_2 x^2} \text{ or equivalently } \log(\mu(x)) = \beta_1 x + \beta_2 x^2$$

This function is a decreasing function of x if

$$\frac{d}{dx}\log(\tilde{\mu}(x)) = \beta_1 + 2\beta_2 x \leq 0 \text{ for all } x \in [0, g]$$

Since that $\beta_1 + 2\beta_2 x$ must be non-positive for all x, thus β_1 will be non-positive number. Despite of the non-positive constant sign of β_1 the sign of β_2 is not constant. If $\beta_2 \leq 0$, then $\tilde{\mu(x)}$ is a decreasing function and for $\beta_2 \geq 0$, under the condition $\frac{-\beta_1}{2\beta_2} \leq g$, $\tilde{\mu(x)}$ is still a decreasing function of x.

Lemma 6.3.1. *For the model* (6.4) *the D-optimal design* $\zeta^* = \left\{ \begin{array}{c} \xi^* \\ 1 \end{array} \right\}$ *has exactly three different support points* x_1^*, x_2^* *and* x_3^*, *i.e.,* $\xi^* = \left\{ \begin{array}{ccc} x_1^* & x_2^* & x_3^* \\ p_1^* & p_2^* & p_3^* \end{array} \right\}$ *where* p_1^*, p_2^* *and* $p_3^* > 0$.

Proof. We consider $\beta_2 < 0$. According to the Theorem 5.6.1 ξ^* is the D-optimal design for model (6.4) if and only if

$$m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x) \leq tr[\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)] \quad \forall x \in \tau.$$

We can represent the above inequality as:

$$\begin{pmatrix} 1 & x & x^2 \end{pmatrix} \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} \leq \frac{1}{\mu(x)} \tag{6.5}$$

for some constant a, b, c, d, e, f. Therefore the left side of (6.5) can be represented as a forth order degree polynomial and hence the above inequality can be written as:

$$h(x) \leq k(x) \text{ with } h(x) = a_1x^4 + a_2x^3 + a_3x^2 + a_4x^+a_5$$

where a_i $(i = 1, \ldots, 5)$ are some appropriate constants. We suppose that $h(x) - k(x) = 0$, where $k(x) = \frac{1}{\mu(x)}$, has at least four roots $z_1 < z_2 < z_3 < z_4$, i.e. $h(z_i) = k(z_i)$for $i = 1, 2, 3, 4$. Regarding to Rolle's theorem we have z_1', z_2' and z_3' such that $z_1 < z_1' < z_2 < z_2' < z_3 < z_3' < z_4$ and $h'(z_i') = k'(z_i')$ $(i = 1, 2, 3)$. Since that $h(z) \leq k(z)$ thus z_i $(i = 2, 3, 4)$

are tangent points, i.e. $h'(z_i) = k'(z_i)$ for $i = 2, 3, 4$. Applying Rolle's theorem again, we have z_i'' $i = 1, \dots 5$, such that $z_1' < z_1'' < z_2 < z_2'' < z_2' < z_3'' < z_3' < z_4'' < z_3 < z_5'' < z_4$. We apply the Rolle's theorem furthermore three times. We find $z_1^{(5)} < z_2^{(5)}$ such that $h^{(5)}(z_i^{(5)}) = k^{(5)}(z_i^{(5)})$ for $i = 1, 2$. Because of $h^{(5)}(z_i^{(5)}) = 0$, then we have $k^{(5)}(x) = 0$ has two roots. Since that in (6.4) $k(x) = e^{-(\beta_1 x + \beta_2 x^2)}$ and hence,

$$k^{(5)} = k(x)(\beta_1 + 2\beta_2 x)[-60\beta_2^2 + 20\beta_2(\beta_1 + 2\beta_2 x)^2 - (\beta_1 + 2\beta_2 x)^4]$$

since that $\beta_1 + 2\beta_2 x \leq 0$ then $k^{(5)} = 0$ has one root, thus It is in contrary to result that $k^{(5)}(x) = 0$ has two roots and the claim follows. □

Remark 6.2. *when $\beta_2 > 0$, then the value of β_2 depends on β_1, so we can not find a general proof for this case. We can find D-optimal designs in the three points design class and check the results via equivalence theorem.*

Using this lemma, we consider designs with three support points. For sake of simplicity, we consider the case where $h = 0$. The following lemma is immediately concluded from the above lemma.

Lemma 6.3.2. *Consider the model* (6.4). *The determinant of the information matrix for $\boldsymbol{\beta}$ is as follows for $\xi = \begin{pmatrix} x_1 & x_2 & x_3 \\ p_1 & p_2 & p_3 \end{pmatrix}$*

$$\det(\mathfrak{M}(\xi)) = \frac{1}{1 + \gamma(m, \beta_0, \sigma^2)(\sum_{j=1}^{3} p_j \tilde{\mu}_j)} \cdot p_1 p_2 (1 - p_1 - p_2)\tilde{\mu}_1 \tilde{\mu}_2 \tilde{\mu}_3 \{x_1^2(x_2 - x_3) - x_2^2(x_1 - x_3) + x_3^2(x_1 - x_2)\}^2 \quad (6.6)$$

where $\gamma(m, \beta_0, \sigma^2) = m\mu_0(e^{\sigma^2} - 1)$, $\mu_0 = e^{\beta_0 + \frac{1}{2}\sigma^2}$ and $\tilde{\mu}_j = \frac{\mu_j}{\mu_0}$ for $j = 1, 2, 3$.

Proof. Let $m_j = mp_j$. From (4.20), for the design $\xi = \begin{pmatrix} x_1 & x_2 & x_3 \\ p_1 & p_2 & p_3 \end{pmatrix}$ for model (6.4), the information matrix is given as follows

$$\mathfrak{M}_{\boldsymbol{\beta}}(\xi) = \begin{pmatrix} \sum_{j=1}^{3} m_j\mu_j & \sum_{j=1}^{3} m_j\mu_j x_j & \sum_{j=1}^{3} m_j\mu_j x_j^2 \\ \sum_{j=1}^{3} m_j\mu_j x_j & \sum_{j=1}^{3} m_j\mu_j x_j^2 & \sum_{j=1}^{3} m_j\mu_j x_j^3 \\ \sum_{j=1}^{3} m_j\mu_j x_j^2 & \sum_{j=1}^{3} m_j\mu_j x_j^3 & \sum_{j=1}^{3} m_j\mu_j x_j^4 \end{pmatrix} - \frac{e^{\sigma^2} - 1}{1 + (e^{\sigma^2} - 1)\sum_{j=1}^{3} m_j\mu_j}$$

$$\begin{pmatrix} (\sum_{j=1}^{3} m_j\mu_j)^2 & (\sum_{j=1}^{3} m_j\mu_j)(\sum_{j=1}^{3} m_j\mu_j x_j) & (\sum_{j=1}^{3} m_j\mu_j)(\sum_{j=1}^{3} m_j\mu_j x_j^2) \\ (\sum_{j=1}^{3} m_j\mu_j)(\sum_{j=1}^{3} m_j\mu_j x_j) & (\sum_{j=1}^{3} m_j\mu_j x_j)^2 & (\sum_{j=1}^{3} m_j\mu_j x_j)(\sum_{j=1}^{3} m_j\mu_j x_j) \\ (\sum_{j=1}^{3} m_j\mu_j)(\sum_{j=1}^{3} m_j\mu_j x_j^2) & (\sum_{j=1}^{3} m_j\mu_j x_j)(\sum_{j=1}^{3} m_j\mu_j x_j^2) & (\sum_{j=1}^{3} m_j\mu_j x_j^2)^2 \end{pmatrix}$$

The claim is concluded after straightforward calculation. □

Theorem 6.3.1. *Let* $\tau = [0, \infty)$. $\xi^* = \begin{Bmatrix} x_1^* & x_2^* & x_3^* \\ p_1^* & p_2^* & p_3^* \end{Bmatrix}$ *is D-optimal design for model* (6.4) *if it maximizes* (6.6)

Proof. According to the Lemma 6.3.1, the optimal design for model (6.4) is a saturated design like most D-optimal designs. From Lemma 6.3.2 for a saturated design for a quadratic Poisson regression model with random intercept, (6.6) gives the D-criterion for this model. So the D-optimal design can be obtained though maximizing (6.6). Note that

$$\tilde{\mu}_j^* = e^{\beta_1 x_j^* + \beta_2 x_j^{*^2}} \Rightarrow \beta_2 x_j^{*^2} + \beta_1 x_j^* - \log \tilde{\mu}_j^* = 0$$

is a quadratic equation in x_j^*. Therefore

$$x_j^* = \frac{-\beta_1 \pm \sqrt{\beta_1^2 + 4\beta_2 \log \tilde{\mu}_j^*}}{4\beta_2}$$

is divided to two separated cases corresponding to the sign of β_2: i) $x_j^* = \frac{1}{\sqrt{2\beta_2}}\{\sqrt{r} - \sqrt{r + 4\log\tilde{\mu}_j}\}$ if $\beta_2 > 0$ and ii) $x_j^* = \frac{1}{\sqrt{-2\beta_2}}\{\sqrt{-r} + \sqrt{-r - 4\log\tilde{\mu}_j}\}$ if $\beta_2 < 0$, where $r = \frac{\beta_1^2}{2\beta_2}$. □

Note that $\gamma(m, \beta_0, \sigma^2) = m\mu_0(e^{\sigma^2} - 1)$ then the criterion (6.6), in terms of the canonical standardized mean, depends on the parameters through $\gamma(m, \beta_0, \sigma^2)$ and r.
With regard to Theorem 6.3.1, numerical methods can be used to maximize $\det(\mathfrak{M}(\xi))$ or to minimize $-\log(\det(\mathfrak{M}(\xi)))$ in order to find D-optimal designs. The D-optimal design for some representative values of $\gamma(m, \beta_0, \sigma^2)$ and r are listed in Tables 6.3 and 6.4.

Table 6.3: D-optimal designs for model (6.4), $r > 0$

	$r = 20$					$r = 50$				
$\gamma(m, \beta_0, \sigma^2)$	p_1^*	p_2^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\tilde{\mu}_3^*$	p_1^*	p_2^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\tilde{\mu}_3^*$
0	0.333	0.333	0.0067	0.1806	1	0.333	0.333	0.0046	0.2425	1
10	0.439	0.360	0.0067	0.1434	1	0.439	0.358	0.0039	0.1959	1
100	0.455	0.400	0.0067	0.1229	1	0.485	0.358	0.0035	0.1745	1
1000	0.493	0.373	0.0067	0.1218	1	0.495	0.356	0.0034	0.1703	1
10000	0.587	0.302	0.0067	0.1218	1	0.496	0.355	0.0034	0.1699	1

According to these tables $x_3^* = 0$ is one point in experimental setting which is corresponding to $\mu_3^* = 1$.
$\gamma(m, \beta_0, \sigma^2) = 0$ is corresponding to $\sigma^2 = 0$ for all m and β_0. The results for $\gamma(m, \beta_0, \sigma^2) = 0$ coincide with the results in Yanping et al. (2006).
For the restricted region, i.e. $\tau = [h, g]$, if $\tilde{\mu}(g) \leq \tilde{\mu}_1^*$ the D-optimal design is equal to the

Table 6.4: D-optimal designs for model (6.4), $r < 0$

	$r = -20$					$r = -50$				
$\gamma(m, \beta_0, \sigma^2)$	p_1^*	p_2^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\tilde{\mu}_3^*$	p_1^*	p_2^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\tilde{\mu}_3^*$
0	0.333	0.333	0.0149	0.3294	1	0.333	0.333	0.0117	0.3050	1
10	0.440	0.344	0.0128	0.2801	1	0.439	0.349	0.0101	0.2559	1
100	0.480	0.339	0.0120	0.2620	1	0.483	0.343	0.0093	0.2367	1
1000	0.489	0.337	0.0118	0.2584	1	0.490	0.342	0.0092	0.2332	1
10000	0.489	0.337	0.0118	0.2585	1	0.491	0.341	0.0091	0.2327	1

Table 6.5: D-optimal designs for model (6.4), $r < 0$ (restricted region $\tilde{\mu}(g) = 0.1$)

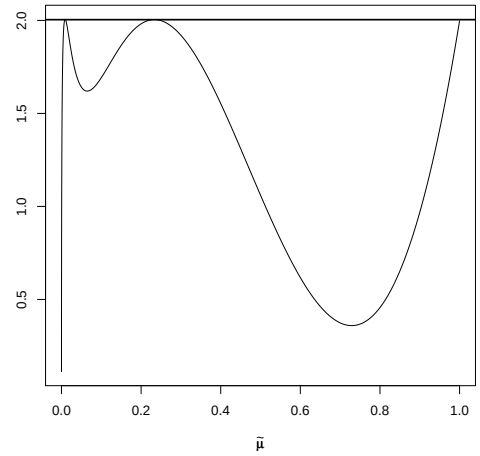
	$r = -20$					$r = -50$				
$\gamma(m, \beta_0, \sigma^2)$	p_1^*	p_2^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\tilde{\mu}_3^*$	p_1^*	p_2^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	$\tilde{\mu}_3^*$
0	0.333	0.333	0.1	0.4500	1	0.333	0.333	0.1	0.4397	1
10	0.421	0.334	0.1	0.4188	1	0.417	0.337	0.1	0.4077	1
100	0.442	0.332	0.1	0.4117	1	0.442	0.335	0.1	0.4023	1
1000	0.446	0.331	0.1	0.4109	1	0.438	0.339	0.1	0.4000	1
10000	0.446	0.330	0.1	0.4109	1	0.443	0.335	0.1	0.3998	1

D-optimal for the unrestricted case. Otherwise the second point of experimental setting will be $x_1^* = g$. The results for some special values for $\gamma(m, \beta_0, \sigma^2)$ and r are listed in Table 6.5.

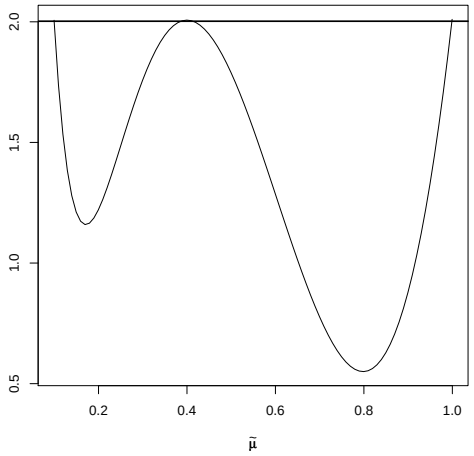
Based on Tables 6.3, 6.4 and 6.5 the D-optimal designs for the restricted design region are completely different from the D-optimal designs for the unrestricted cases, in defiance of the similar trends in all cases. For instance when the intra-individual correlation is increasing then the proportion of observations at $x_3^* = 0$ decreases.

Similar to the previous section, we need to check the locally D-optimality for the quadratic Poisson regression model with random intercept. Regarding to Theorem 5.6.1, we have evaluated sensitivity function, $\phi(x, \xi^*) = m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x)$, over the experimental domain for the model (6.4). The results confirmed the locally D-optimal designs for special cases which have been obtained in Tables 6.3, 6.4 and 6.5. For instance, we illustrate these evaluation for some representative values of $\gamma(m, \beta_0, \sigma^2)$ and r. Four cases have been considered in Figure 6.4: restricted design domain (or $\tilde{\mu} \in [0.1, 1]$) for two different values $\gamma(m, \beta_0, \sigma^2) = 1000$ and $\gamma(m, \beta_0, \sigma^2) = 0$ and unrestricted design region (or $\tilde{\mu} \in (0, 1]$) for the same $\gamma(m, \beta_0, \sigma^2)$ as the restricted design region. Note that in these four cases we consider $r = -50$.

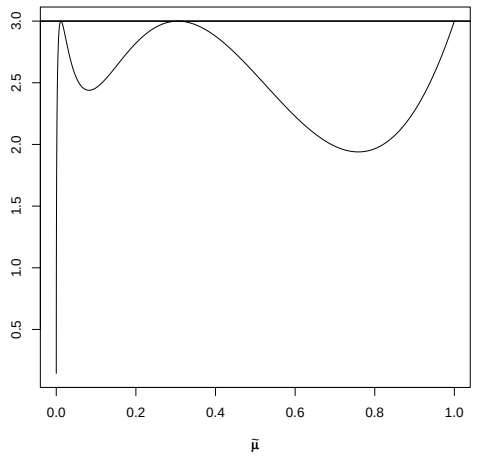
In the restricted case of design region, Figure 6.4(b) and Figure 6.4(d), two of the maxima points of sensitivity function $\phi(x, \xi^*)$ are on the border points of the design region and the remainder can be observed at the corresponding points to $\tilde{\mu}_2^* = 0.4000$ if $\gamma(m, \beta_0, \sigma^2) = 1000$ or at $\tilde{\mu}_2^* = 0.4397$ if $\gamma(m, \beta_0, \sigma^2) = 0$, respectively.



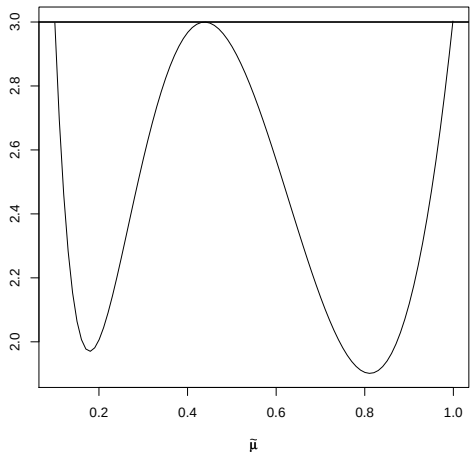
(a) unrestricted region ; $\gamma(m, \beta_0, \sigma^2) = 1000$



(b) restricted region ;$\gamma(m, \beta_0, \sigma^2) = 1000$



(c) unrestricted region ; $\gamma(m, \beta_0, \sigma^2) = 0$



(d) restricted region ;$\gamma(m, \beta_0, \sigma^2) = 0$

Figure 6.4: Locally D-optimality for model (6.4) (a) $\tilde{\mu} \in (0, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 1000$; (b) $\tilde{\mu} \in [0.1, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 1000$; (c) $\tilde{\mu} \in (0, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 0$ (without random intercept); (d) $\tilde{\mu} \in [0.1, 1]$ and $\gamma(m, \beta_0(h), \sigma^2) = 0$ (without random intercept).

Table 6.6: D-efficiency of the standard design for the Quadratic Poisson Regression Model with Random Intercept, $r < 0$ (restricted region $\mu(g) = 0.1$)

	D-efficiency			
$\gamma(m, \beta_0(0), \sigma^2)$	$r = -10$	$r = -20$	$r = -30$	$r = -50$
0	0.9200	0.9148	0.9113	0.9070
10	0.8609	0.8490	0.8424	0.8353
100	0.8367	0.8222	0.8144	0.8063
1000	0.8019	0.8181	0.8101	0.8019
10000	0.8006	0.8179	0.8100	0.8018

In the case of an unrestricted design region, Figure 6.4(a) and Figure 6.4(c), if $\gamma(m, \beta_0, \sigma^2) = 1000$, $m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x)$ gets its maximum at $\tilde{\mu}_1^* = 0.0092$, $\tilde{\mu}_2^* = 0.2332$ and $\tilde{\mu}_3^* = 1$ and if $\gamma(m, \beta_0, \sigma^2) = 0$ (model without random effects),

$$m\mu(x)\mathbf{f}^T(x)\underline{M}^{-1}(\xi^*)\mathfrak{M}(\xi^*)\underline{M}^{-1}(\xi^*)\mathbf{f}(x)$$

attains its maximum at $\tilde{\mu}_1^* = 0.0117$, $\tilde{\mu}_2^* = 0.3050$ and $\tilde{\mu}_3^* = 1$.
A result in the previous section was that the D-optimal designs do not coincide with the G-optimal designs.
If we consider $\xi_G^* = \left\{ \begin{array}{ccc} x_1^* & x_2^* & x_3^* \\ p_1^{'*} & p_2^{'*} & p_3^{'*} \end{array} \right\}$ as a G-optimal design for model (6.4), i.e.

$$\xi_G^* = \arg \min_{\xi \in \Xi_G} \max_x Var(\hat{Y}_x)$$

where $Var(\hat{Y}_x)$ is the variance of predictor of Y at the point x and can be obtained through δ-method (see section 5.8). Ξ_G is the set of experimental design such that $\xi_G^* = \left\{ \begin{array}{ccc} x_1^* & x_2^* & x_3^* \\ p_1' & p_2' & p_3' \end{array} \right\}$ and p_1', p_2', p_3' are the allocated proportions to x_1^*, x_2^*, x_3^* respectively with $p_1' + p_2' + p_3' = 1$. There is no analytical solution to find G-optimal design for the quadratic Poisson regression model with random intercept. As in the previous section, we have numerically that the D-optimality and G-optimality do not coincide. For instance, we plot the proportion of the observations at x_1^* from the D-optimal design and from the G-optimal design versus σ^2 (Figure 6.5). p_1^* and $p_1^{'*}$ go far away when σ^2 increases. We consider a standard three point design which includes the two endpoints and the middle point. We also suppose that the corresponding proportions are equal, i.e. $p_1 = p_2 = p_3 = \frac{1}{3}$. Table 6.6 indicates that this standard three point design is robust as long as $\tilde{\mu}(g)$ is not too small, i.e. if g is not too large. The results also show that the efficiency decreases when r is increasing. The same relation can be seen for $\gamma(m, \beta_0(0), \sigma^2)$.
Figure 6.6 indicates the above results for D-efficiency of the standard design graphically.

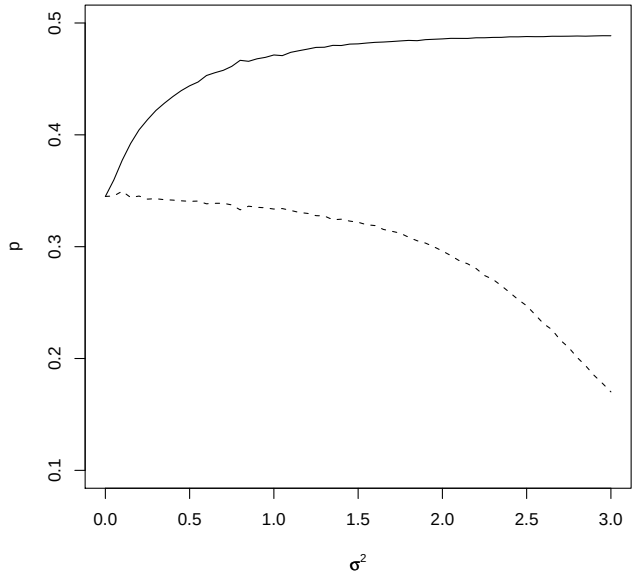


Figure 6.5: D and G-optimal proportions of observations at x_1^* versus σ^2 (for model (6.4)). D-optimal proportions (solid line) and G-optimal proportion (dashed line).

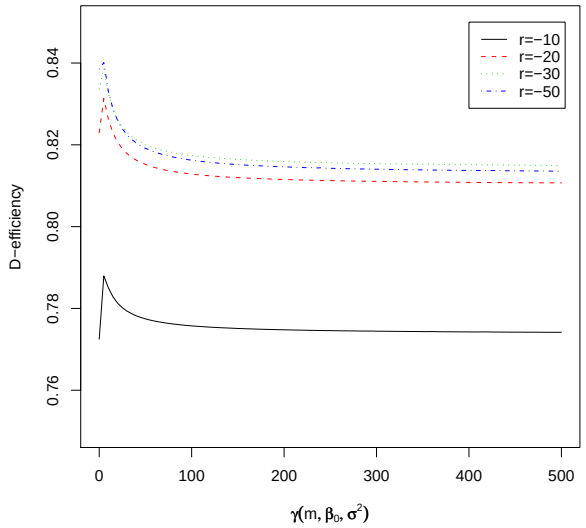


Figure 6.6: D-efficiency of the standard design in model (6.4) versus σ^2.

6.4 Locally D-optimal Designs for the Poisson Regression Model With Random Slope

The last model, which is considered in this section, is the Poisson regression model with random slope. In the previous chapter we described some properties of this model theoretically. Some examples of locally D-optimal designs are obtained in the sequel of this section.
In general, for different individuals, we may use different designs for data collection. In this section we confine ourselves to the case where identical designs are used for different individuals. So this model will be as follows

$$Y_{ijk} \mid b_i \overset{ind}{\sim} P(\mu_{ij}(b_i)) \begin{cases} i = 1, \dots, s & \sum_{j=1}^{t} m_j = m \\ j = 1, \dots, t & \\ k = 1, \dots, m_j & n = sm \end{cases} \tag{6.7}$$

where $\mu_{ij}(b_i) = \exp(\beta_0 + b_i x_j)$ with b_i is normally distributed with mean β_1 and variance $\sigma^2$and $\mathrm{Cov}(b_i, b_{i'}) = 0\,(i \neq i')$. Here, the aim is to find D-optimal design for estimating β_0 and β_1. For the same reason as previous sections, we suppose that the design regions are non-negative real number. So $\mu_j = \exp(\frac{1}{2}\sigma^2 x_j^2 + \beta_0 + \beta_1 x_j)$ lies in $(0, \mu_0]$, where $\mu_0 = e^{\beta_0}$ stands for the mean response at $x = 0$. If μ_j is decreasing , $\tilde{\mu}_j = \frac{\mu_j}{\mu_0}$ lies in $(0, 1]$.

Remark 6.3. *We can consider the general case for experimental design domain where $x \in [h, g]$ and hence $\mu_j \in [\mu_g, \mu_h]$. For simplicity we consider $x \in [0, g]$, where g can tend to ∞.*

The following theorem gives us a criterion for D-optimal designs in the Poisson regression model with random slope.

Theorem 6.4.1. *Consider the model* (6.7)*. If ξ is any design with two points in the experimental setting, i.e. $\xi = \begin{Bmatrix} x_1 & x_2 \\ p_1 & 1-p_1 \end{Bmatrix}$ then the determinant of the quasi-information matrix for $\boldsymbol{\beta}$ is as follows*

$$\det(\mathfrak{M}_{\boldsymbol{\beta}}(\xi)) = \frac{m^2\mu_0^2 p_1(1-p_1)\tilde{\mu}_1\tilde{\mu}_2(x_1-x_2)^2}{1 + mp_1\mu_0\tilde{\mu}_1(e^{\sigma^2 x_1^2} - 1) + m(1-p_1)\mu_0\tilde{\mu}_2(e^{\sigma^2 x_2^2} - 1) + m^2 p_1(1-p_1)\mu_0^2\tilde{\mu}_1\tilde{\mu}_2(e^{\sigma^2 x_1 x_2} - 1)} \tag{6.8}$$

Proof. with regard to (5.32)

$$\mathfrak{M}_{\boldsymbol{\beta}}(\xi) = \mathbf{F}^T(\mathbf{A}^{-1} + \mathbf{B})^{-1}\mathbf{F}$$
$$= \begin{pmatrix} 1 & 1 \\ x_1 & x_2 \end{pmatrix} \left(\left(m\mu_0 \begin{pmatrix} p_1\mu_1 & 0 \\ 0 & (1-p_1)\mu_2 \end{pmatrix} \right)^{-1} + \begin{pmatrix} e^{\sigma^2 x_1^2} - 1 & e^{\sigma^2 x_1 x_2} - 1 \\ e^{\sigma^2 x_1 x_2} - 1 & e^{\sigma^2 x_2^2} - 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}$$

After using some matrix algebra and a straightforward calculation the result follows. □

Using the numerical methods, we maximize this criterion (6.8)(or minimize $-\log(\mathfrak{M}_{\boldsymbol{\beta}}(\xi))$) for some representative values of β_0, β_1, σ^2 and m. The results are listed in Table 6.7.

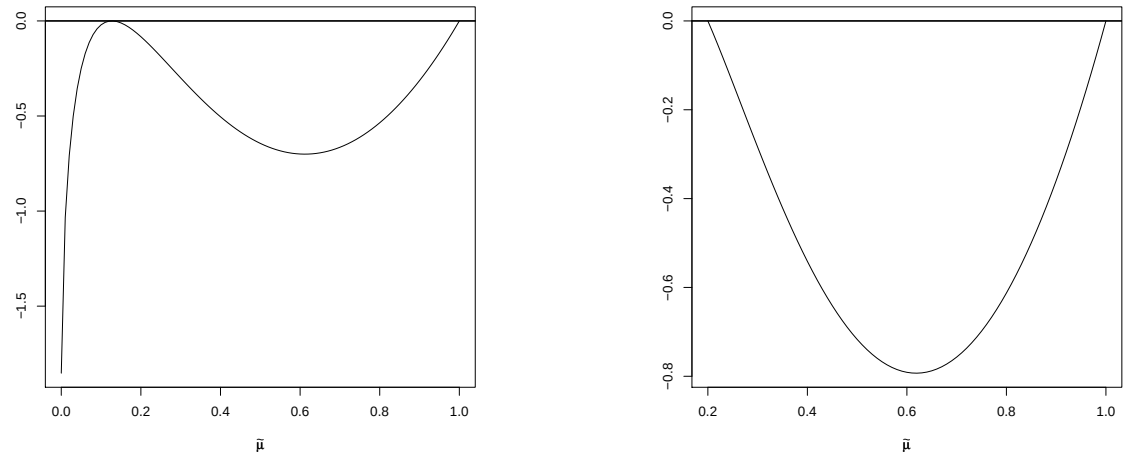
Note that $\tilde{\mu}_2^* = 1$ corresponds to $x_2^* = 0$. If $\sigma^2 = 0$, i.e. the model without random

Table 6.7: D-optimal designs for model (6.7)

	$m = 100$, $\beta_0 = -2$ and $\beta_1 = -5$			$m = 200$, $\beta_0 = -2$ and $\beta_1 = -5$		
σ^2	p_1^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$	p_1^*	$\tilde{\mu}_1^*$	$\tilde{\mu}_2^*$
0	0.500	0.1353	1	0.500	0.1353	1
0.5	0.482	0.1305	1	0.465	0.1313	1
1	0.462	0.1272	1	0.434	0.1308	1
2	0.422	0.1296	1	0.379	0.1441	1
3	0.384	0.1487	1	0.338	0.1743	1
4	0.354	0.1798	1	0.309	0.2114	1
5	0.331	0.2147	1	0.289	0.2487	1

effects, the results are in accordance with the results in Table 6.1.

From Table 6.7, for fixed m, β_0 and β_1 the proportion of observations at x_1^*, p_1^*, decreases when σ^2 is increasing. The same trend can be seen when m increases. On this point of view the results in Table 6.7 is in contrast with the results in Table 6.1, where p_1^* increases when σ^2 or m are rising.



(a) unrestricted region

(b) restricted region

Figure 6.7: Locally D-optimality for the model (6.7) (a) unrestricted domain, where $m = 100$, $\beta_0 = -2$, $\beta_1 = -5$ and $\sigma^2 = 1$; (b) restricted domain $\tilde{\mu} \in [0.2, 1]$, where $m = 100$, $\beta_0 = -2$, $\beta_1 = -5$ and $\sigma^2 = 1$;

Theorem 5.7.3 helps us to check the optimality of ξ^* for different cases. We obtain lo-

cally D-optimal designs which are showed in Table 6.7 and also for the restricted domain cases. For instance we plot two different cases corresponding to the different values of the parameters (Figure 6.7). For $m = 100$, $\beta_0 = -2$, $\beta_1 = -5$ and $\sigma^2 = 1$, in the unrestricted case, Figure 6.7(a) shows that the sensitivity function, $\phi(x, \xi^*)$ achieves the maximum zero at $\tilde{\mu}_1 = 0.1272$ and $\tilde{\mu}_2 = 1$. In the restricted domain case, the Fréchet derivative is zero at $\tilde{\mu}_1 = 0.2$ and $\tilde{\mu}_2 = 1$.
For some fixed values of β_0, β_1, m and σ^2 ($\sigma^2 > 0$), the numerical results show that the D-optimal designs do not coincide with the corresponding G-optimal design.
Figure 6.8 shows that the design is robust when the design points of Standard design and optimal design are the same.

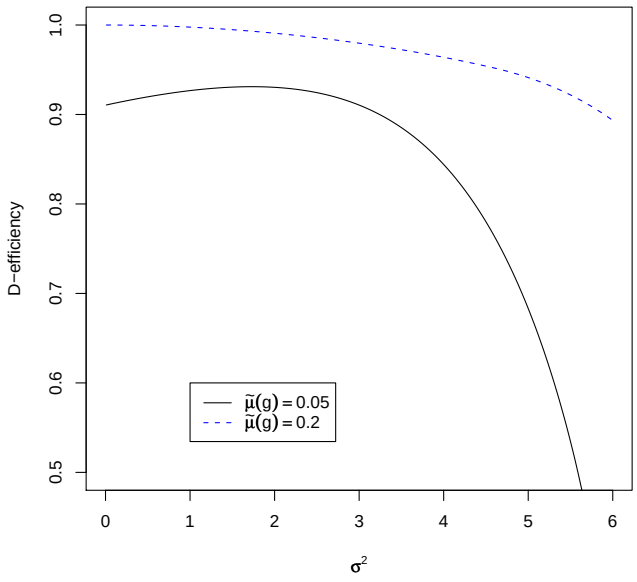


Figure 6.8: D-efficiency of the design in model (6.7) versus σ^2.

7 Discussion and Future Research

The main purpose of this research was to develop efficient and robust experimental design for the Poisson regression model with random effects. This model is a special case of a broad class, generalized linear mixed models. These random effects caused some theoretical problems with the likelihood method and in finding the information matrix. At least two ways could be considered to overcome of these limitations. First one is to use numerical methods for obtaining locally optimal designs directly. The second way is to apply some approximate method to find a closed form for information matrices and hence maximizing an appropriate optimality criterion. The quasi-likelihood method is an approximate method which is applied in our research. In the Poisson regression model with random effects, the quasi-likelihood method encountered us to new theoretical challenges with convex design theory. We obtained some solutions to remedy these problems in our models separately.
The best experimental settings, as a main goal of the optimal design studies, have been considered in chapter 6 and we found locally D-optimal designs for different models. We have made two important assumptions there: First we supposed that experimental design domains are non-negative subsets of the real line and the second assumption was that increasing the design variable causes a decrease in the mean response. These assumptions seem to be justified in many applications including industrial studies, biosciences etc.
A possible design, which might be considered for a linear model, is a saturated standard design. The numerical results indicated that these designs could not be suggested for our models due to low efficiency for these saturated standard designs when experimental design regions are large.
This research could potentially be extended in several directions, which are listed as follows:

1. Tekle et al. (2008) have done a numerical study in optimal designs for logistic mixed effects models for a binary longitudinal response as a special case of generalized linear mixed models. They studied locally optimal designs for these models under a different method of estimation. So a numerical comparison between locally optimal designs based on the quasi-likelihood method and their likelihood method might be interesting.

2. We considered models with one explanatory variable. An extension to higher dimensions of the explanatory variables is a topic for further studies. A recent work in the Poisson regression model with fixed effects has been published by Russell et al. (2009).

3. Although some numerical results has been presented for different values of σ^2 (variance of the random effect), we only considered random effects with known variance in the theoretical part of this thesis. One could work on the case with random effects with unknown variance numerically and analytically.

4. A main assumption was that the random effects are normally distributed. Selecting a different distribution, a deviation of this condition, could be a subject for further investigation

5. Bayesian or other methods can be considered to overcome the dependence of design performance on parameter values. For example, we can see Woods et al.(2006) and Gotwalt et al.(2009).

Abbreviation

$\xrightarrow{D}$	asymptotic distribution, 20
$()^-$	g-inverse, 3
∇_g^T	gradient vector, 58
$\mathbf{1}_t$	t-dimensional vector containing only ones, 28
$\mathbf{A}_i$	$diag\{m_{i1}\mu_{i1}, \ldots, m_{it_i}\mu_{it_i}\}$, 29
$\dot{\mathbf{A}}_i$	$diag\{\mu_{i1}I_{m_{i1}}, \cdots, \mu_{it_i}I_{m_{it_i}}\}$, 28
$\dot{\mathbf{a}}_i^T$	$\sqrt{e^{\sigma^2}-1}\left(\begin{array}{ccc} \mu_{i1}\mathbf{1}_{m_{i1}}^T & \cdots & \mu_{it_i}\mathbf{1}_{m_{it_i}}^T \end{array}\right)$, 28
$\mathbf{b}_i$	vector of the random effects parameters for the ith subject, 9
Cov	covariance matrix, 4
$\mathbf{D}$	partial derivative of $\boldsymbol{\mu}(\boldsymbol{\beta})$, 19
d_n	exact design of size n, 37
det()	determinant of a matrix, 18
$diag\{\}$	diagonal matrix, 19
$E()$	expectation, 4
$e^{()}$	exponential function, 22
exp()	exponential function, 5
$\mathbf{F}$	design matrix, 4
$\mathbf{F}_i$	row individual design matrix neglecting the number of replications , 29
$\dot{\mathbf{F}}_i$	design matrix for individual i, 28
F_Φ	Fréchet directional derivative, 42
GLM	generalized linear models, 3
GLMM	Generalized Linear Mixed Model, 3
G_Φ	Gâteaux derivative, 42
$\mathbf{g}_n$	estimating function for p dimensional vector of parameters, 17
$\mathbf{g}()$	estimating function for p dimensional vector of parameters, 17
$\mathbf{g}_n^*$	F-optimal estimating function, 17

g	upper bound of experimental domain, 59
$g_n = g_n()$	estimating function, 15
$g_n^{(s)}$	standardized versions of estimating function, 15
$g_n^{(s)}()$	standardized versions of estimating function, 15
h	lower bound of experimental domain, 60
$h()$	link function, 5
$\mathbf{I}$	identity matrix, 7
$I_{\boldsymbol{\beta}}^{(n)}$	minus derivative of score function w.r.to $\boldsymbol{\beta}$, 20
$\mathbf{J}_{m\times n}$	a matrix of $m \times n$ order containing only ones, 27
LM	linear model, 3
LMM	linear mixed model, 3
LS	least square, 14
$l = l()$	log-likelihood function, 6
log()	natural logarithm, 5
ML	maximum likelihood, 14
$M_{\boldsymbol{\beta},n}$	quasi-information matrix, 20
$M_{\boldsymbol{\beta}}^i$	quasi-information matrix for $\boldsymbol{\beta}$ corresponding to the individual i, 28
$\mathcal{M}_n$	set of the information matrices of exact designs d_n, 38
$\mathfrak{M}_{\boldsymbol{\beta}}(\xi_i)$	ith individual information matrix for $\boldsymbol{\beta}$, 44
m_i	total number of observations for individual i , 25
m_{ij}	number of replication for the individual i at x_{ij} , 25
$P()$	Poisson distribution , 22
PQL	penalized quasi-likelihood , 24
QL	quasi-likelihood , 14
$ql()$	log-QL function , 19
t_i	number of points at experimental setting for individual i , 25
$tr()$	trace of a matrix, 18
$U^{(n)}(\boldsymbol{\beta})$	quasi-score function for $\boldsymbol{\beta}$, 18
$U_r^{(n)}(\boldsymbol{\beta})$	rth element of $U^{(n)}(\boldsymbol{\beta})$, 18
$\mathbf{V}$	variance of $\mathbf{Y}$, 28

$\mathbf{V}_i$	variance of Y_i, 27
$\mathbf{V}_i^{(jk)}$	$\mathrm{Cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{ik})$, 27
$\mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}))$	$Var(\mathbf{Y})$, 18
$Var()$	variance, 4
$v()$	variance function, 6
$\mathbf{X}$	design matrix, 4
x_{ij}	jth level of explanatory variable for ith individual , 25
$\mathbf{Y}$	vector of the whole observations, 4
$\mathbf{Y}_i$	vector of all responses for individual i, 27
$\mathbf{Y}_{ij}$	vector of replication for individual i at x_{ij} , 27
Y_{ij}	jth observation of the ith subject (individual) , 10
$\mathbf{Z}_i$	design matrix for the random effects, 9
$\boldsymbol{\beta}$	unknown vector of fixed effects, 4
$\hat{\boldsymbol{\beta}}^{(n)}$	maximum QL estimator of $\boldsymbol{\beta}$, 19
γ_i	canonical parameter of the distribution of Y_i , 5
ϵ	error term, 4
ζ	population design, 45
λ_{min}	minimum of eigenvalues of a matrix, 18
$\boldsymbol{\mu}(\boldsymbol{\beta})$	expectation of $\mathbf{Y}$, 18
$\boldsymbol{\mu}_i(\boldsymbol{\beta})$	expectation of Y_i, 27
$\mu_{ij}^{(\mathbf{b}_i)}$	conditional mean of Y_{ij} or Y_{ijk}, 10
$\tilde{\mu}()$	canonical standardized mean evaluated at g , 61
Ξ	set of all probability measures over τ , 39
Ξ_n	set of exact designs of size n, 38
ξ_i	ith individual design, 43
ξ_n	discrete design measure corresponding to d_n , 37
ξ^*	optimal design, 38
σ^2	variance of the error term, ϵ, 4
σ^2	variance of the random effect, 25
τ	experimental (design) domain (region), 36

Φ optimality criterion, 38

ϕ nuisance parameter, 5

$\phi()$ sensitivity function, 43

Ψ optimality criterion, 47

Bibliography

[1] Atkinson A.C., Donev A.N. and Tobias R.D. (2007), "Optimum Experimental Designs, with SAS", Oxford University Press, New york.

[2] Atkinson A.C. and Bailey R.A. (2001), "One Hundred Years of the Design of Experiments on and off the pages Biometika", Biometrika, **88**, 53-97.

[3] Bauer H. (1996), "Probability Theory", Walter de Gruyter, Berlin.

[4] Biedermann S., Dette H. and Zhu W., (2006), "Optimal designs for Dose-response Models With Restricted Design Spaces", Journal of the American Statistical Association, **101**,747-759.

[5] Box G.E.P. and Lucas H.L. (1959)," Design of Experiments in Non-Linear Situations", Biometrika, **46**, 7790.

[6] Breslow N.E. and Glayton D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models", Journal of the American Statistical Association, **88**, 9-25.

[7] Booth J.G. and Hobert J.P. (1999), "Maximizing Generalized Linear Mixed Models with an Automated Monte Carlo EM Algorithm", Journal of the Royal Statistical Society B, Methodological **61**, 265-285.

[8] Casella G. and Berger, R.L. (2002), "Statistical Inference", Duxbury Press, second edition.

[9] Chernoff H. (1953), "Locally Optimal Designs for Estimating Parameters", The Annals of Mathematical Statistics,**26**, 586-602.

[10] Desmond A.F. (1997), "Optimal Estimating Function, Quasi-Likelihood and Statistical Modeling",Journal of Statistical Planning and Inference, **60**, 77-121.

[11] Ehrenfeld, S. (1955), "On the Efficiency of Experimental Designs", **26**, 247-255.

[12] Faraway J.J. (2006), " Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models", Chapmann and Hall, London.

[13] Fedorov V.V. (1972), "Theory of Optimal Experiments", Academic Press, New York.

[14] Fedorov V.V., Hackl P., (1997), "Model-Oriented Design of Experiments", Volume 125 of Lecture Notes in Statistics. Springer, New York.

[15] Firth D. (1987), "On the Efficiency of Quasi-Likelihood Estimation", Biometrika, **74**, 233-245.

[16] Gladitz J. and Pliz J. (1982), "Construction of optimal designs in random coefficient regression models", Statistics: A Journal of Theoretical and Applied Statistics, **13**, 371-385.

[17] Godambe V.P. (1991), "Estimating Functions", Oxford University Press, Oxford.

[18] Godambe V.P. and Thompson M.E. (1978), "Some Aspects of the Theory of Estimating Equations", Journal of Statistical Planning and Inference, **2**, 95-104.

[19] Godambe V.P. and Heyde C.C. (1987), "Quasi-Likelihood and Optimal Estimation", International Statistical Review, **78**, 231-244.

[20] Gotwalt C.M., Jones B.A. and Steinberg D.M., (2009), "Fast Computation of Designs robustto Parameters Uncertainty for Nonlinear Settings", Technometrics, **51**, 88-95.

[21] Graybill F. A.(1976), "Theory and Application of the Linear Model", Duxbury, Mass.

[22] Heyde C. C. (1997), "Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation", Springer, New York.

[23] Hotelling H.(1936), "Relations Between Two Sets of Variables", Biometrika, **28**, 321-377.

[24] Jørgensen B. (1987), "Exponential Dispersion Models (with discussion)", Journal of the Royal Statistical Society B, **49**, 127-162.

[25] Jiang J. (2007), "Linear and Generalized Linear Mixed Models and their applications", Springer, New York.

[26] Kiefer J. and Wolfowitz J. (1959), "Optimum Designs in Regression Problems", The Annals of Mathematical Statistics, **30**, 271-294.

[27] Kiefer J. and Wolfowitz J. (1960), "The Equivalence of two Extremum Problem", Canadian Journal of Mathematics, **12**, 363-366.

[28] Kiefer J. , Brown L.D. and Olkin I. (1985), "Collected Papers Volume 3: Design of Experiments", Springer, New York.

[29] Kimball B. F. (1946), "Sufficient statistical estimation functions for the parameters of the distribution of maximum values", The Annals of Mathematical Statistics, **17**, 299-309.

[30] Laird N.M. and Ware J.M.(1982), "Random Effects Models for Longitudinal Data ", Biometrics, **38**, 963-974.

[31] Liang K.Y. and Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models", Biometrika, **73**, 13-22.

[32] Lin X. and Breslow N.E. (1996), "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion", Journal of the American Statistical Association, **91**, 1007-1016.

[33] Liski E.P., Mandal N.K., Shah K.R. and Sinha B.K. (2002), "Topics in Optimal Design, Volume 163 of Lecture Note in Statistics" Springer, New York.

[34] Magnus J.R. and Neudecker H. (1988), "Matrix Differential Calculus with Application in Statistics and Economics", Wiley, Chichester.

[35] McCullagh P.(1983), "Quasi-Likelihood Functions",The Annals of Statistics,**11**, 59-67.

[36] McCullagh P.(1991), In Hinkley D.V., Reid N. and Snell E.J.,Eds., "Quasi-Likelihood and Estimating Functions In Statistical Theory and Modeling", Chapman and Hall, London Ch.11.

[37] McCullagh P. and Nelder J. A.(1998), "Generalized Linear Models",(Reprinted), Chapman and Hall, London.

[38] McCulloch C.E. and Searle S.R. (2001), "Generalized, Linear, and Mixed Models", Wiley, New York.

[39] Minkin, S., 1993. Experimental Designs for Clonogenic Assay in Chemotherapy. Journal of the American Statistical Association **88**, 410-420.

[40] Myers W.R., Myers R.H. and Carter W.H. Jr. (1994), " Some alphabetic Optimal Designs for the Logistic Regression Models", Journal of Statistical Planning and Inference, **42**, 57-77.

[41] Nelder J.A. and Pregibon D. (1987), "An Extended Quasi-Likelihood Function", Biometrika, **74**, 221-32.

[42] Nelder J.A. and Wedderburn R.W.M. (1972), "Generalized Linear Models", Journal of the Royal Statistical Society A, **135**, 370-384.

[43] Niaparast M. (2009a), "On optimal Design for a Poisson Regression Model with Random Intercept ", Statistics and Probability Letters, **79**, 741-747.

[44] Niaparast M. (2009b), "On the Locally D-optimality for the Poisson Regression Model with Random Block Effect", Proceeding of ENBIS9, Gothenburg, Sweden.

[45] Nie L. (2007), "Convergence Rate of MLE in Generalized Linear and Non-Linear Mixed Effects Models: Theory and Applications", Journal of Statistical Planning and Inference, **137**, 1787-1804.

[46] Ouwens J.N.M. Mario, Tan E.S. Frans and Berger P.F. Martijn, (2006), "A maximin criterion for the logistic random intercept model with covariates", Journal of Statistical Planning and Inference, **136**, 962-981.

[47] Pinheiro J.C. and Bates D.M. (1995), "Approximation to the log-likelihood function in nonlinear mixed effects models", Journal of Computational and Graphical Statistics, **4**, 12-35.

[48] Pourahmadi M.(2000), "Maximum Likelihood Estimation of Generalized Linear Models for Multivariate Normal Covariance Matrix", Biometrika, **87**, 425-435.

[49] Pukelsheim F. (1993), "Optimal Design of Experiments", Wiley, New York.

[50] Liu Qing (2006), " Optimal Experimental Designs for Hyperparameter Estimation in Hierarchical Linear Models", PhD thesis, The Ohio State University,USA.

[51] Rao C.R. (1973), "Linear Statistical Inference and its Applications", Wiley, New York, second edition.

[52] Rockafellar R.T. (1972), "Convex Analysis", Princeton University Press, Princeton.

[53] Russell K.G., Woods D.C., Lewis S.M. and Eccleston (2009), "D-Optimal Designs for Poisson Regression Models", Statistica Sinica, **19**, 721-730.

[54] Schmelter T. (2007a), "The Optimality of Single-group Designs for Certain Mixed Models", Metrika, **65**, 183-193.

[55] Schmelter T. (2007b), " Considerations on Group-Wise Identical Designs for Linear Mixed Models", Journal of Statistical Planning and Inference, **137**, 4003-4010.

[56] Schmelter T. (2007c), "Experimental Design for Mixed Models with Application to Population Pharmacokinetic Studies" PhD thesis, Magdeburg University, Germany.

[57] Schmelter T., Benda N., Schwabe R. (2007), "Some Curiosities in Optimal Designs for Random Slopes", mODa 8 - Advances in Model-Oriented Design and Analysis (Proceedings of the 8th International Workshop in Model-Oriented Design and Analysis), 4-8.

[58] Schott J.R. (1997), "Matrix Analysis for Statistics", Wiley, New York.

[59] Schwabe R. and Schmelter T. (2008), "On Optimal Designs in Random Intercept Models", Tatra Mountains Mathematical Publications, **39**, 145-153.

[60] Silvey S.D. (1980), " Optimal Design", Chapman Hall, London.

[61] Sinha S. (2004) Robust Analysis in Generalized Linear Mixed Models" Journal of the American Statistical Association, **99**, 451-460.

[62] Sitter R.R. (1992), "Robust Designs for Binary Data", Biometrics, **48**, 1145-1155.

[63] Smith K. (1918), "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and Its Constants and the Guidance They Give Towards a proper Choice of the Distribution of Observations", Biometrika, **12**, 1-85.

[64] Tekle F.B., Tan F.E.S. and Berger M.P.F (2008), " Maximin D-optimal Designs for Binary Longitudinal Responses", Computational Statistics and Data Analysis, **52**, 5253-5262.

[65] Van der Vaart A.W. (1998), "Asymptotic Statistics", Cambridge University Press.

[66] Vonesh E.F., Wang H., Nie L. and Majumdar D. (2002), "Conditional Second-Order Generalized Estimating Equations for Generalized Linear and Non-Linear Mixed Effects Models", Journal of the American Statistical Association, **96**, 282-291.

[67] Wald A. (1943), "On the Efficient Design of Statistical Investigations", Annals of Mathematical Statistics, **14**, 134140.

[68] Wand M.P. (2002), "Vector Differential Calculus in Statistics", The American Statistician, **56**, 55-62.

[69] Wand M.P. (2007), "Fisher Information for Generalised Linear Mixed Models", Journal of Multivariate Analysis, **98**, 1412-1416.

[70] Wang Y., Myers R.H., Smith E.P. and Ye K. (2006), "D-optimal Designs for Poisson Regression Models", Journal of Statistical Planning and Inference, **136**, 2831-2845.

[71] Waterhouse T.H. (2005), "Optimal Experimental Design for Nonlinear and Generalised Linear Models", PhD thesis, University of Queensland, Australia.

[72] Wedderburn R.W.M. (1974), "Quasi-likelihood functions, Generalized Linear Models and the Gauss-Newton Method", Biometrika, **61**, 439-447.

[73] Woods D.C., Lewis S.M., Eccleston J.A. and Russell K.G. (2006), "Designs for Generalized Linear Models with Several Variables and Model Uncertainty", Technomtrics, **48**, 284-292.