Optimal Design and Inference for Correlated Bernoulli Variables using a Simplified Cox Model

Daniel Bruce

Stockholm
University

Doctoral Dissertation
Department of Statistics
Stockholm University
SE-106 91 Stockholm
Sweden

Abstract

This thesis proposes a simplification of the model for dependent Bernoulli variables presented in Cox and Snell (1989). The simplified model, referred to as the simplified Cox model, is developed for identically distributed and dependent Bernoulli variables.

Properties of the model are presented, including expressions for the log-likelihood function and the Fisher information. The special case of a bivariate symmetric model is studied in detail. For this particular model, it is found that the number of design points in a locally D-optimal design is determined by the log-odds ratio between the variables. Under mutual independence, both a general expression for the restrictions of the parameters and an analytical expression for locally D-optimal designs are derived.

Focusing on the bivariate case, score tests and likelihood ratio tests are derived to test for independence. Numerical illustrations of these test statistics are presented in three examples. In connection to testing for independence, an E-optimal design for maximizing the local asymptotic power of the score test is proposed.

The simplified Cox model is applied to a dental data. Based on the estimates of the model, optimal designs are derived. The analysis shows that these optimal designs yield considerably more precise parameter estimates compared to the original design. The original design is also compared against the E-optimal design with respect to the power of the score test. For most alternative hypotheses the E-optimal design provides a larger power compared to the original design.

Key words: Cox binary model, correlated binary data, log-odds ratio, D-optimality, E-optimality, power maximization, efficiency

ISBN 978-91-7155-657-8

Printed in Sweden by ABA Copy 2008
Distributor: Department of Statistics, Stockholm University

Acknowledgements

During the work with this thesis I have received valuable support from a number of people.

Foremost I thank my supervisor Professor Hans Nyquist. Your guidance, support, and ideas have been invaluable for the progress of this thesis. Thank you for showing such a genuine interest not only in my research problems but also in other important issues such as skiing and training in general.

A large group of people whom I am deeply grateful to are all my friends and colleagues at the Department of Statistics, Stockholm University. You have shown great friendship and enthusiasm during these years. In particular, I would like to thank Ellinor Fackle Fornius, Jessica Franzén, and Mattias Torkelsson. You were my fellow doctoral students from the first day at the Department. A special thanks to Ellinor Fackle Fornius for exchanging ideas on SWP, MATLAB, optimal designs, running, and other issues. I also thank doctoral students, former and present, Jonas Larsson, Pär Stockhammar, Bertil Wegmann, Johan Koskinen, Michael Carlson, Christian Tallberg, Jan Hagberg, Jörgen Säve-Söderbergh, Linda Wänström, Nicklas Petersson, Termeh Shafie, and Marek Balog.

I am particular grateful to all my friends and family of varying description, notably Ardhiel, Cella, Hanna, syster Tomas, Hedvig and the family Vestergren. A special thanks to the other five members of ”bommel” Björn (Stubben), Johan (Poss), Tomas (don Tommaso), Thomas (Peaceman), and Daniel (Setta).

My true gratitude goes to my mother Eva and my sister Nina for always supporting and encouraging me.

Karin, you are a very special girl to me. You have given me love, happiness, and confidence. I think of a song with the lyrics: *She's just the girl for me and I want all the world to see we've met.*

Stockholm, April 2008 Daniel Bruce

Contents

Appendices

Chapter 1

Introduction

This thesis considers a model for dependent Bernoulli variables. The starting point is that independent observations are made on a cluster or batch of Bernoulli variables. In general the Bernoulli variables within a batch are assumed to be dependent. Although the size of the batch can be large in some applications, models for pairs of dependent Bernoulli variables are studied in more detail. The key property in the models is that the univariate marginal probabilities are identical for all Bernoulli variables within a batch. Note that this does not impose Bernoulli variables from different batches to have the same marginal probabilities. An advantage with the model, compared to other more general models for dependent Bernoulli variables, is that the expressions for the likelihood as well as the Fisher information matrix are relatively uncomplicated. Consequently, parameter estimators are obtained quite readily even for a model with many parameters to estimate. Examples with identically distributed dependent Bernoulli variables exist in several sciences including both observational and experimental studies.

One example of an application is the analysis of visual impairment data. The probability for visual impairment on the left eye is assumed to be equal to the probability of visual impairment on the right eye, for a particular individual. There is also a dependence between the eyes. To further improve the analysis, explanatory variables (covariates) such as intraocular pressure or age could be incorporated. These kinds of data have been studied by Rosner (1984), Tielsch et al. (1991), and Liang et al. (1992).

Another example is an experiment where fry (of fish) are studied. The objective of the experiment is to determine how different types of food (treatments) affect some property of the fish. Fry that are assigned to a certain treatment are therefore kept isolated. The fry can be further specified by other covariates. It is clear that the responses from the fry that are kept in the same isolation box are dependent. The model with identical marginal probabilities can be applied if the fry within a treatment are homogeneous.

In a third example, groups of plants grow in common soil. Different batches of plants are then exposed to different amounts of some fertilizer. Since the plants share soil, the condition of each plant is dependent of the other plants within the same batch. If the response variable is binary or coded as binary, the model applies to this kind of experiment. In this kind of applications it is natural to include a distance covariate for the spatial dependence between plants within a batch.

Mandel et al. (1982) describe a clinical trial using ear data. Briefly, this study is a double-blind randomized clinical trial in which two antibiotics are compared. The subjects of the experiment are children, divided into age groups, with acute otitis media in both ears. After 14 days of treatment the number of cured ears were recorded for each child. The response from each ear is a Bernoulli variable, ”cured” or ”not cured”. Moreover, it is reasonable to assume that the responses of the left and right ears are dependent.

Andrews and Herzberg (1985) present a clinical trial using dental data. In the experiment rats are randomly assigned to different diets to see if the cariogenic effect can be reduced. At the end of the experiment, occlusal surfaces in each rat were examined with two possible responses: ”caries” and ”no caries”. The responses from the occlusal surfaces within a certain rat are dependent and assumed to be identically distributed. This example is considered in greater detail in Chapter 9.

A last example involves a company that produces a certain product. The company wants to evaluate the test procedure for quality control of the product. Employees therefore perform repeated measurements on a sample of products to investigate whether the product pass the quality control test or not. The company wants to investigate to what degree the results in the different measurements are equal. The probability that the product pass the test is assumed to be constant for different measurements. Hence, a model for identically distributed dependent Bernoulli variables is suitable for this situation.

More applications with dependent Bernoulli variables exist in biology and in the medical sciences. Some examples given in Zucker and Wittes (1992) include, development of tumor in animals within a litter, presence of arthritic pain in different joints, and occurrence of plaque progression in each of several vein grafts in patients with prior coronary artery bypass surgery.

The above examples with fry allocated to different treatments, plants growing in common soil, children with otitis media, and cariogenic effect of diets are examples of planned experiments. In a planned experiment the design of the experiment needs to be determined. The design of an experiment includes choosing the treatments and choosing the corresponding number of observations to be allocated to each treatment. The design is important since all analysis is based on the design. A design that optimizes some inferential property of the model, according to some criterion, is referred to as an optimal design.

The main aim of this thesis is to present a model for k identically distributed and possibly dependent Bernoulli variables, called the simplified Cox model. Different properties of the model are explored, including the loglikelihood function and the Fisher information. When exploring the model, suggestions for relevant generalizations of the model are made. The extensions include a covariate for the distance between observed subjects within a batch and a generalization from binary data to polytomous data.Furthermore the aim is to derive analytical and numerical results for locally D-optimal designs.

The simplified Cox model is a special case of the more general Cox model, given in Cox and Snell (1989). When it can be assumed that the Bernoulli variables are identically distributed, the Cox model has an unnecessary complex structure with too many response categories, compared to the simplified Cox model. This follows since under the assumption of identical marginal distributions, the models are the same. Hence, the simplified Cox model is preferable since it is parsimonious compared to the Cox model.

The thesis also addresses test procedures for tests of mutual independence between the variables. The aim is further to propose and motivate optimal designs for maximizing the power of these tests. A numerical example is included to illustrate the optimal properties of one of the tests. When the Bernoulli variables are mutually independent, the simplified Cox model is estimated with just two parameters. This simple structure makes it important to test if mutual independence can be assumed. Therefore, parameter restrictions for mutual independence are derived.

The thesis is organized as follows. In Chapter 2 a brief overview over different models for bivariate Bernoulli variables is given. The simplified Cox model is presented in Chapter 3. Expressions for the likelihood function, the score function, and the Fisher information are derived. An alternative model incorporating a distance covariate as well as a generalization of the model to polytomous data are outlined as well.

Chapter 4 contains an introduction to the concept of optimal designs. Short summaries of the different techniques, locally optimal designs, sequential optimal designs, optimum in average designs, and minimax designs are given. The different design criteria used throughout the thesis are illustrated by examples. A symmetric bivariate model is outlined in Chapter 5. Different symmetry properties for the probability distribution are given together with examples of D-optimal designs and some general results on D-optimal designs. In Chapter 6 the model for mutually independent variables is explored. Analytical expressions for parameter restrictions as well as for locally D-optimal designs are obtained in two theorems. In the case of paired data, properties of the model are examined in more detail.

Likelihood ratio tests and score tests in a test for independence between the Bernoulli variables are discussed in Chapter 7. Using examples, test procedures where covariates are incorporated and test procedures without covariates are both illustrated. In Chapter 8, an expression for an optimal design that maximizes the local asymptotic power of the score test is derived. For a particular example the performance of the design in small samples is examined in a simulation experiment. The robustness of the optimal design is also examined. Chapter 9 gives an example of the simplified Cox model including estimation of the model and a test for independence. The original design is compared against both a locally D-optimal design and a locally E-optimal design with respect to precision in the parameter estimates. A short comparison between the original design and the E-optimal design with respect to the power of the score test is performed. Finally, Chapter 10 discusses assumptions made when using the simplified Cox model. Additionally, suggestions for further research using the simplified Cox model are given.

Chapter 2

Overview of Models for Bivariate Bernoulli Variables

The aim of this chapter is to present different models for dependent Bernoulli variables. This overview is by no mean comprehensive in the sense that it treats all families of models. The models presented are given in the bivariate case and includes just one covariate. This is because it is easier to get an overview of the model when the number of parameters and the number of response categories are limited. Nevertheless, several of the models can be generalized to an arbitrary number of variables as well as response categories.

Let S_1 and S_2 denote two possibly dependent Bernoulli variables. Moreover, let x be a covariate associated to the distribution of S_1 and S_2. Several ways of modelling the joint distribution of S_1 and S_2 as a function of x has been proposed. A summary of different approaches was given already in Cox (1972). Bonney (1987) presented general loglinear multivariate logistic models for an arbitrary number of dependent binary variables. Using the unsaturated model given by Bonney, let η_1 and η_2 be

$$
\begin{aligned}
\eta_1 &= \ln \frac{P\left(S_1 = 1 \mid x\right)}{P\left(S_1 = 0 \mid x\right)} = \alpha + \beta x \\
\eta_2 &= \ln \frac{P\left(S_2 = 1 \mid S_1, x\right)}{P\left(S_2 = 0 \mid S_1, x\right)} = \alpha + \gamma Z + \beta x,
\end{aligned}
$$

where

$$Z = 2S_1 - 1.$$

If S_1 and S_2 are independent then $\gamma = 0$. Based on

$$\pi_{S_1S_2}(x) = \prod_{i=1}^{2} \frac{e^{\eta_i S_i}}{1+e^{\eta_i}},$$

the probability of the four possible outcomes of (S_1, S_2), $(1,1)$, $(1,0)$, $(0,1)$, and $(0,0)$ are

$$\begin{aligned}
\pi_{(S_1=1,S_2=1)}(x) &= \pi_{11}(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \frac{e^{\alpha+\gamma+\beta x}}{1+e^{\alpha+\gamma+\beta x}} \\
\pi_{10}(x) &= \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \frac{1}{1+e^{\alpha+\gamma+\beta x}} \\
\pi_{01}(x) &= \frac{1}{1+e^{\alpha+\beta x}} \frac{e^{\alpha-\gamma+\beta x}}{1+e^{\alpha-\gamma+\beta x}} \\
\pi_{00}(x) &= \frac{1}{1+e^{\alpha+\beta x}} \frac{1}{1+e^{\alpha-\gamma+\beta x}},
\end{aligned}$$

respectively. Thus, the bivariate probability distribution of (S_1, S_2) can be expressed as products of ordinary logistic functions. Therefore the loglikelihood function and the information matrix can be obtained quite readily. The saturated model allows for different intercepts α and different slopes β in the linear predictors η_1 and η_2.

Murtaugh and Fisher (1990) utilized the bivariate logistic cumulative distribution function (cdf) given by Gumbel (1961). Define

$$\begin{aligned}
\eta_1 &= \alpha_1 + \beta_1 x \\
\eta_2 &= \alpha_2 + \beta_2 x.
\end{aligned}$$

The bivariate Gumbel distribution has cdf

$$F_{U,V}(u,v) = \frac{1}{1+e^{-u}} \frac{1}{1+e^{-v}} \left\{ 1 + \frac{\gamma e^{-u-v}}{(1+e^{-u})(1+e^{-v})} \right\},$$

and U and V are considered as continuous latent variables. The binary variables S_1 and S_2 are assumed to be indicators, indicating whether U and V exceeds a certain threshold:

$$\begin{aligned}
S_1 = 1 &\text{ iff } U \leq \eta_1 \\
S_2 = 1 &\text{ iff } V \leq \eta_2.
\end{aligned}$$

The parameter γ incorporates the possible dependence between S_1 and S_2 in the model. Using $F_{U,V}(u,v)$, the probabilities

$$
\begin{aligned}
\pi_{11}(x) &= \frac{1}{1+e^{-\eta_1}}\frac{1}{1+e^{-\eta_2}}+\frac{\gamma e^{-\eta_1-\eta_2}}{\left(1+e^{-\eta_1}\right)^2\left(1+e^{-\eta_2}\right)^2} && \text{(2.1a)}\\
\pi_{10}(x) &= \frac{1}{1+e^{-\eta_1}}-\frac{1}{1+e^{-\eta_1}}\frac{1}{1+e^{-\eta_2}}-\frac{\gamma e^{-\eta_1-\eta_2}}{\left(1+e^{-\eta_1}\right)^2\left(1+e^{-\eta_2}\right)^2} && \text{(2.1b)}\\
\pi_{01}(x) &= \frac{1}{1+e^{-\eta_2}}-\frac{1}{1+e^{-\eta_1}}\frac{1}{1+e^{-\eta_2}}-\frac{\gamma e^{-\eta_1-\eta_2}}{\left(1+e^{-\eta_1}\right)^2\left(1+e^{-\eta_2}\right)^2} && \text{(2.1c)}\\
\pi_{00}(x) &= 1-\frac{1}{1+e^{-\eta_1}}-\frac{1}{1+e^{-\eta_2}}+\frac{1}{1+e^{-\eta_1}}\frac{1}{1+e^{-\eta_2}} && \text{(2.1d)}\\
&\quad +\frac{\gamma e^{-\eta_1-\eta_2}}{\left(1+e^{-\eta_1}\right)^2\left(1+e^{-\eta_2}\right)^2}
\end{aligned}
$$

are obtained. It follows directly that S_1 and S_2 are independent if and only if $\gamma = 0$. As Murtaugh and Fisher (1990) point out, the marginal probabilities of S_1 and S_2 are logistic in η_1 and η_2, respectively. The likelihood function follows directly from (2.1a), (2.1b), (2.1c), and (2.1d). Maximum likelihood estimation of $(\alpha_1, \beta_1, \alpha_2, \beta_2, \gamma)$ are conducted by numerical maximization of the likelihood function. Heise and Myers (1996) and Dragalin and Fedorov (2006) also used the Gumbel model in bivariate logistic regression models.

Murtaugh and Fisher (1990) and Dragalin and Fedorov (2006) also used the Cox bivariate binary model, given in Cox and Snell (1989), to model dependent binary variables. This model treats each possible outcome as a separate response category. In the bivariate case, the model has four response categories. The corresponding probability of each response category is

$$
\begin{aligned}
\pi_{11}(x) &= \frac{e^{\eta_{11}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}\\
\pi_{10}(x) &= \frac{e^{\eta_{10}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}\\
\pi_{01}(x) &= \frac{e^{\eta_{01}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}\\
\pi_{00}(x) &= \frac{1}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}},
\end{aligned}
$$

where

$$\begin{aligned}\eta_{11} &= \alpha_{11} + \beta_{11} x \\ \eta_{10} &= \alpha_{10} + \beta_{10} x \\ \eta_{01} &= \alpha_{01} + \beta_{01} x.\end{aligned}$$

The model can be written in a more compact form as

$$\pi_{ij}(x) = \frac{e^{i(1-j)\eta_{10}+j(1-i)\eta_{01}+ij\eta_{11}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}} \quad i,j = 0,1.$$

The marginal probabilities of S_1 and S_2 are not logistic in η_{11}, η_{10}, and η_{01}. Instead it is the conditional probabilities for one of the variables given the other variable that are logistic in η_{11}, η_{10}, and η_{01}, Murtaugh and Fisher (1990). Throughout the thesis, the Cox binary model will be referred to as the Cox model. This simple illustration of the Cox model can be directly generalized to model k Bernoulli variables. Hirji (1994) suggested a similar model. He extended the model to include subject specific covariates.

In some applications it is reasonable or natural to assume that the Bernoulli variables have the same univariate marginal distribution. Under such an assumption it is irrelevant whether the event $(1,0)$ or the event $(0,1)$ occurred. By imposing such a restriction on the Cox model, the joint probability function for (S_1, S_2) becomes as shown in Table 2.1. Note that $\pi_{10} = \pi_{01}$ and that the linear predictors η_{10} and η_{01} have been replaced by one new linear predictor η_1. This is because the only information incorporated in the model is the number of "success". In the bivariate case the restriction of identical marginal distributions implies that S_1 and S_2 are exchangeable.

The situation outlined in Table 2.1 describes a special case of the Cox model which is central throughout this thesis. When all the Bernoulli variables have the same marginal distribution the Cox model is therefore denoted the simplified Cox model.

In the Cox model there are four possible outcomes $(1,1)$, $(1,0)$, $(0,1)$, and $(0,0)$. Under the simplified Cox model, the two outcomes $(1,0)$ and $(0,1)$ are merged into one outcome. Thus it is sufficient to model the joint probability of (S_1, S_2) through the random variable $S = S_1 + S_2$,

		S_2		
		0	1	
S_1	0	$\pi_{00} = \dfrac{1}{1+e^{\eta_1}+e^{\eta_2}}$	$\pi_{01} = \dfrac{e^{\eta_1}/2}{1+e^{\eta_1}+e^{\eta_2}}$	$1-\pi. = \dfrac{1+e^{\eta_1}/2}{1+e^{\eta_1}+e^{\eta_2}}$
	1	$\pi_{10} = \dfrac{e^{\eta_1}/2}{1+e^{\eta_1}+e^{\eta_2}}$	$\pi_{11} = \dfrac{e^{\eta_2}}{1+e^{\eta_1}+e^{\eta_2}}$	$\pi. = \dfrac{e^{\eta_1}/2+e^{\eta_2}}{1+e^{\eta_1}+e^{\eta_2}}$
		$1-\pi. = \dfrac{1+e^{\eta_1}/2}{1+e^{\eta_1}+e^{\eta_2}}$	$\pi. = \dfrac{e^{\eta_1}/2+e^{\eta_2}}{1+e^{\eta_1}+e^{\eta_2}}$	1

Table 2.1: Joint probability function for (S_1, S_2) and marginal distributions for S_1 and S_2.

where $s = 0, 1$, or 2. Under a simplified Cox model the joint probability of (S_1, S_2) is described by the probabilities

$$\begin{aligned}
\pi_{(S=2)}(x) &= \pi_2(x) = \frac{e^{\eta_2}}{1+e^{\eta_1}+e^{\eta_2}} \\
\pi_1(x) &= \frac{e^{\eta_1}}{1+e^{\eta_1}+e^{\eta_2}} \\
\pi_0(x) &= \frac{1}{1+e^{\eta_1}+e^{\eta_2}},
\end{aligned}$$

where

$$\begin{aligned}
\eta_1 &= \alpha_1 + \beta_1 x \\
\eta_2 &= \alpha_2 + \beta_2 x.
\end{aligned}$$

As for the Cox model, the simplified Cox model can in the bivariate case be defined in a more compact form as

$$\pi_{i+j}(x) = \frac{e^{|i-j|\eta_1 + ij\eta_2}}{1+e^{\eta_1}+e^{\eta_2}} \quad i, j = 0, 1.$$

The conditional probabilities for S_1 given S_2 and vice versa are logistic in η_1 and η_2 assuming the covariate x is held constant. For example,

$$P(S_1 = 1 \mid S_2 = 0) = \frac{e^{\eta_1}}{2+e^{\eta_1}},$$

is logistic in η_1.

In the sequel, the probabilities $\pi_0(x)$, $\pi_1(x)$, and $\pi_2(x)$ will be written in the shorter form π_0, π_1, and π_2 although, they are usually governed by a covariate. Since the simplified Cox model is a special case of the Cox model, the two models have a similar structure. Assuming a bivariate model and that $\pi_{10} = \pi_{01}$, the Cox model reduces to the simplified Cox model if

$$\pi_{10} + \pi_{01} = \pi_1$$

This is equivalent to imposing the restrictions

$$\eta_{10} = \eta_{01} = \eta_1 - \ln 2,$$

on the bivariate Cox model.

Using the expression for the linear predictors above, the following connection between the parameters in the models appear

$$\begin{cases} \alpha_{10} = \alpha_{01} = \alpha_1 - \ln 2 \\ \alpha_{11} = \alpha_2 \\ \beta_{10} = \beta_{01} = \beta_1 \\ \beta_{11} = \beta_2. \end{cases}$$

Thus the simplified Cox model is a special case of the original Cox model. The advantage with the simplified Cox model compared to the Cox model is that the number of response categories as well as the number of parameters is considerably reduced. In the bivariate case the number of response categories is reduced from four to three as shown above. Moreover the number of parameters is reduced from six to four. For a general model with k variables the number of response categories is reduced from 2^k to $k+1$ and the number of parameters is reduced from $2\left(2^k - 1\right)$ to $2k$.

This paragraph considers the general simplified Cox model with k variables. Due to the assumption that $S_1, S_2, \ldots, S_k$ are treated as dependent and identically distributed variables under the simplified Cox model, the correlation structure of the simplified Cox model needs some additional comments. Within an observation on $(S_1, S_2, \ldots, S_k)$, the correlation between any pair (S_i, S_j) has the same value for any $i \neq j$. In addition, any two variables from different observations on $(S_1, S_2, \ldots, S_k)$ are assumed to be independent. In Chapter 3 the simplified Cox model is outlined including a discussion on the correlation structure.

Consider again the bivariate case and a model for S_1 and S_2. In another class of models, $S = S_1 + S_2$ is assumed to follow a beta-binomial distribution. Let $\pi.$ denote the marginal probability of "success", i.e.

$$P(S_1 = 1) = P(S_2 = 1) = \pi.$$

Further, assume that $S = S_1 + S_2$ given $\pi.$ follows a binomial distribution,

$$P(S = s \mid \pi.) = \tbinom{2}{s}\pi.^s (1 - \pi.)^{2-s}, \quad s = 0, 1, 2.$$

The probability of "success" may vary between different batches or pairs, not only because different batches are assigned to different treatments but also because different batches may have different correlation structure. Skellam (1948) suggested that the probability of "success" should be described by a beta distribution. Williams (1975) used this to obtain a beta-binomial distribution for the probability distribution of S. By letting $\pi.$ be beta distributed with the parameters α and β, the probability distribution of S becomes beta-binomial distributed with

$$P(S = s) = \binom{2}{s}\frac{B(s + \alpha, 2 - s + \beta)}{B(\alpha, \beta)} \quad s = 0, 1, 2,$$

where $B(\alpha, \beta)$ is the beta function with the parameters α and β. In addition, different treatments usually have different α and β. The beta-binomial distribution is obtained by assuming that the Bernoulli variables are from an infinite sequence of exchangeable Bernoulli variables, George and Bowman (1995a). This requirement imposes the correlation between the Bernoulli variables to be positive.

The beta-binomial model was extended by Rosner (1984) and Prentice (1986). Rosner (1984) worked specifically with the bivariate case, incorporating covariates via a polychotomous logistic regression model. Prentice (1986) suggested an extended beta-binomial model, allowing the correlation within a batch to be negative. Zucker and Wittes (1992) compared the beta-binomial model with a model denoted Markov-like susceptibility model which is another conditional binomial model. As for the simplified Cox model, the beta-binomial model allows the correlation between (S_1, S_2) to vary across treatments.

In a series of papers, George and Bowman and George and Kodell respectively, proposed and discussed a model which, at least in terms of estimation of $P\left(S_1=s_1,S_2=s_2\right)$, is similar to the simplified Cox model, George and Bowman (1995a,b); George and Kodell (1996). Let S_1 and S_2 be exchangeable and let

$$\begin{cases} \lambda_2 = P\left(S_1=1,S_2=1\right) \\ \lambda_1 = P\left(S_1=1\right) \\ \lambda_0 = 1. \end{cases}$$

George and Bowman (1995a) showed that

$$\pi_s = \tbinom{2}{s}\sum_{j=0}^{2-s}\left(-1\right)^j\tbinom{n-s}{j}\lambda_{s+j}, \quad s=0,1,2,$$

which yields

$$\begin{cases} \pi_2 = \lambda_2 \\ \pi_1 = 2\left(\lambda_1-\lambda_2\right) \\ \pi_0 = 1-2\lambda_1+\lambda_2. \end{cases}$$

When S_1 and S_2 are independent

$$\lambda_2 = \lambda_1^2,$$

so that S has a binomial distribution with parameters $n=2$ and λ_1. The maximum likelihood estimators under the restriction of independence and the unrestricted estimators are derived in George and Kodell (1996). These estimators are equivalent to the corresponding estimators for the simplified Cox model derived in Chapter 7.

In George and Bowman (1995a), π_s is linked to a covariate using the logistic function,

$$\lambda_s\left(\alpha,\beta\right) = \frac{2}{1+\exp\left\{\left(\alpha+\beta x\right)\ln\left(s+1\right)\right\}}.$$

They compared estimates of this model with the beta-binomial model discussed above and with a third estimating procedure. The third procedure estimates their proposed model using a quasi-likelihood technique where a generalized estimating equations procedure is used. Generalized estimating equations for correlated binary data are treated in, e.g. Prentice (1988) and Liang et al. (1992).

A different class of models are the loglinear models. A loglinear model for two Bernoulli variables is defined by

$$\ln N\pi_{ij} = \lambda + \alpha_i + \beta_j + (\alpha\beta)_{ij}\,, \quad i,j = 0,1,$$

where $N\pi_{ij}$ is the expected frequency under the current model. The model is analogous to a model for analysis of variance. To model the probabilities π_{11}, π_{10}, π_{01}, and π_{00} a four factor model is required. Agresti (2002) points out that loglinear models focus on association and interaction in the joint distribution of categorical response variables. Logit models are preferable if a single categorical response variable depends on explanatory variables. This thesis focuses on the latter situation where the probability of the different outcomes of S depend on an explanatory variable, x. Loglinear models are presented in e.g. Bishop et al. (1975) , Christensen (1997), and Agresti (2002).

Another type of model utilizes the odds ratio as a measure of the dependence between S_1 and S_2. This type of model is based on the cross-ratio model, see e.g. Dale (1986), Palmgren (1991), Le Cassie and Van Houwelingen (1994), and Appelgren (2004) used this model for bivariate binary responses. Let $\pi_{1\cdot} = \pi_{11} + \pi_{10}$ and $\pi_{\cdot 1} = \pi_{11} + \pi_{01}$ denote the marginal probabilities $P(S_1 = 1)$ and $P(S_2 = 1)$, respectively. Moreover, let Ω denote the odds ratio between S_1 and S_2, defined as

$$\Omega = \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}.$$

Using the expression from Palmgren (1991)

$$\pi_{11} = \left\{ \begin{array}{cl} \frac{1}{2}(\Omega - 1)^{-1}\left\{a - \sqrt{a^2 + b}\right\} & \text{if } \Omega \neq 1 \\ \pi_{1\cdot}\pi_{\cdot 1} & \text{if } \Omega = 1 \end{array} \right. ,$$

where

$$\begin{aligned} a &= 1 + (\pi_{1\cdot} + \pi_{\cdot 1})(\Omega - 1) \\ b &= -4\Omega(\Omega - 1)\pi_{1\cdot}\pi_{\cdot 1}. \end{aligned}$$

The other probabilities π_{10}, π_{01}, and π_{00} follow from the marginal probabilities $\pi_{1\cdot}$ and $\pi_{\cdot 1}$. These probabilities can be associated with covariates using the bivariate logistic regression model given by McCullagh and

Nelder (1989). One example is obtained if

$$\begin{aligned}
\ln \frac{\pi_{1\cdot}}{1-\pi_{1\cdot}} &= \eta_1 = \alpha_1 + \beta_1 x \\
\ln \frac{\pi_{\cdot 1}}{1-\pi_{\cdot 1}} &= \eta_2 = \alpha_2 + \beta_2 x \\
\ln \Omega &= \eta_{12} = \alpha_{12} + \beta_{12} x.
\end{aligned}$$

In this model S_1 and S_2 are independent if and only if $\ln \Omega = 0$.

In some situations the data have a hierarchical structure. A class of models that utilizes this is multilevel models. A multilevel model has several levels, in which different factors enter at different levels. In Agresti (2002), an example with students writing a battery of exams is given. For each exam the response is binary, the student can either pass or fail. Suppose that a model is to be set up in order to estimate the probability that a student passes an exam. In this model other factors not necessary related to the student might be of interest. Such a factor could be the student's school. In the example, the exam is a level 1 factor and student is a level 2 factor, accounting for the variability among students in ability. School is a factor at level 3, accounting for factors such as percapita expenditure in the school's budget. Multilevel models in general are treated in e.g. Goldstein (2003).

The simplified version of the Cox model can be represented in terms of a model often referred to as multinomial logistic model. Models for multinomial responses can be categorized depending on the type of data. Zocchi and Atkinson (1999) argued that there are different models for nominal, ordinal and hierarchical data. Agresti (2002) divided the models in a similar way. Models for nominal data have been explored by Fahrmeir and Tutz (2001), Agresti (2002), and Puu (2003). This kind of models is sometimes called simple multinomial logit models. When there is an ordering between the outcomes of a response, several models exist. Zocchi and Atkinson (1999), Fahrmeir and Tutz (2001), Agresti (2002), and Dobson (2002) have presented some models, examples include the cumulative logit model, the proportional odds model and the continuation-ratio logit model. The continuation-ratio logit model is further explored in Fan (1999) and Fan and Chaloner (2004). Another model which resembles the continuation-ratio logit model is the contingent response model, discussed in Rabie (2004). The models for ordered responses are especially useful for efficacy-toxicity responses where a natural order among the different responses exist.

All the models above use the same link function, the logit link. Other link functions such as probit link and complementary log-log link are discussed in Fahrmeir and Tutz (2001), Agresti (2002), and Dobson (2002).

Chapter 3

The Simplified Cox Model

In the last chapter the simplified Cox model was briefly introduced. In this chapter the simplified Cox model is outlined in more detail. The main part of Section 3.1, Section 3.2, and Section 3.3 is presented in Bruce (2008).

3.1 The Model

Let $S_1, S_2, \dots, S_k$ denote k identically distributed Bernoulli variables. Let

$$S = \sum_{i=1}^{k} S_i,$$

and

$$P\left(S = s\right) = \pi_s \quad s = 0, 1, \dots, k.$$

A model for S can be viewed as a multivariate generalized linear model (MGLM). In a MGLM the response variable, the linear predictor, and the link function are vector-valued functions, see Fahrmeir and Tutz (2001). The response vector is

$$Y = \begin{pmatrix} Y_1 & Y_2 & \dots & Y_k \end{pmatrix}^T,$$

where

$$Y_s = \begin{cases} 1, \text{ if } S = s \\ 0, \text{ otherwise} \end{cases} \qquad s = 1, 2, \dots, k.$$

Hence, the expected value of Y is

$$\mu = E\left[\begin{pmatrix} Y_1 & Y_2 & \dots & Y_k \end{pmatrix}^T\right] = \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_k \end{pmatrix}^T.$$

In a multivariate logit model, one of the response categories is chosen to be a reference category. Because of the way Y is defined, the event $S = 0$ is chosen to be reference category. Given the reference category, the logit link function $g(\pi_1, \pi_2, ..., \pi_k)$ is

$$g(\pi_1, \pi_2, \dots, \pi_k)^T = \begin{pmatrix} \ln\frac{\pi_1}{\pi_0} & \ln\frac{\pi_2}{\pi_0} & \dots & \ln\frac{\pi_k}{\pi_0} \end{pmatrix}^T = \eta,$$

where η is the linear predictor. With just one covariate, η is

$$\eta = \begin{pmatrix} \eta_1 & \eta_2 & \dots & \eta_k \end{pmatrix}^T = \begin{pmatrix} \alpha_1 + \beta_1 x & \alpha_2 + \beta_2 x & \dots & \alpha_k + \beta_k x \end{pmatrix}^T = \mathbf{x}\theta,$$

where

$$\mathbf{x} = \begin{pmatrix} 1 & 0 & \dots & 0 & x & 0 & \dots & 0 \\ 0 & \ddots & & \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & x \end{pmatrix}$$

and

$$\theta = \begin{pmatrix} \alpha_1 & \dots & \alpha_k & \beta_1 & \dots & \beta_k \end{pmatrix}^T.$$

Note that $\eta_0 = 0$ by definition.

$\mathbf{x}$ is a $(k \times 2k)$ matrix and θ is a size $2k$ vector. The probabilities $\pi_0, \pi_1, \dots, \pi_k$ as functions of x are

$$\pi_s = \frac{e^{\eta_s}}{\sum_{i=0}^{k} e^{\eta_i}} \quad s = 0, 1, \dots, k. \tag{3.1}$$

No simple and direct interpretation of the parameters exist. The parameters α_s and β_s in η_s are interpreted from the expression $\eta_s = \ln\frac{\pi_s}{\pi_0}$, $s = 1, \dots, k$. Thus it is difficult to interpret how different parameters affect the joint probability distribution of $S_1, S_2, \dots, S_k$.

When there are only two dependent variables (S_1, S_2), the (marginal) odds ratio can be given by just one expression.

Property 3.1 *The odds ratio for $S_1 = 1$ is $4e^{\eta_2 - 2\eta_1}$.*

Proof. Denote the odds ratio for $S_1 = 1 \mid S_2 = 1$ relative to $S_1 = 1 \mid S_2 = 0$ by Ω.

$$\Omega = \frac{\frac{\pi_{11}}{\pi_{01}}}{\frac{\pi_{10}}{\pi_{00}}} = \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = \frac{\frac{1}{1+e^{\eta_1}+e^{\eta_2}} \frac{e^{\eta_2}}{1+e^{\eta_1}+e^{\eta_2}}}{\frac{e^{\eta_1}}{2(1+e^{\eta_1}+e^{\eta_2})} \frac{e^{\eta_1}}{2(1+e^{\eta_1}+e^{\eta_2})}} = 4e^{\eta_2 - 2\eta_1}.$$

■

Hence the log-odds ratio is

$$\ln \Omega = \ln 4 + \alpha_2 - 2\alpha_1 + x\left(\beta_2 - 2\beta_1\right). \tag{3.2}$$

In general the log-odds ratio depends on the value of x. A model for (S_1, S_2) contains the four parameters,

$$\theta^T = \begin{pmatrix} \alpha_1 & \alpha_2 & \beta_1 & \beta_2 \end{pmatrix}.$$

The parameters can be interpreted using $\ln \frac{\pi_1}{\pi_0}$ and $\ln \frac{\pi_2}{\pi_0}$ as described above. Another way of interpreting the parameters is to use the expression for $\ln \Omega$. For example, the effect of the covariate on $\ln \Omega$ is controlled by β_1 and β_2.

To see how the probability distribution of S changes with x, four plots with different parameter values are shown in Figure 3.1. Although the plots differ a lot, they share some general properties. Since β_1 and β_2 are larger than zero, π_0 decreases with x and π_2 increases with x.

3.2 The Simplified Cox Model viewed as a Multidimensional Table

The joint probability distribution of $S_1, S_2, \ldots, S_k$ can also be viewed as a $k-$dimensional $2 \times 2 \times \ldots \times 2$ table. Each cell in this table represents a unique sequence of $S_1, S_2, \ldots, S_k$, where $S_i = \{0 \text{ or } 1\}$, $i = 1, 2, \ldots, k$. Moreover there are $\binom{k}{s}$ cells in which $\sum_{i=1}^{k} S_i = s$. This follows from the fact that there are $\binom{k}{s}$ ways of observing s "successes" among a total of k variables. The restrictions of the simplified Cox model state that $S_1, S_2, \ldots, S_k$ are identically distributed so that every outcome resulting in s "successes" among k variables have the same probability. Since π_s

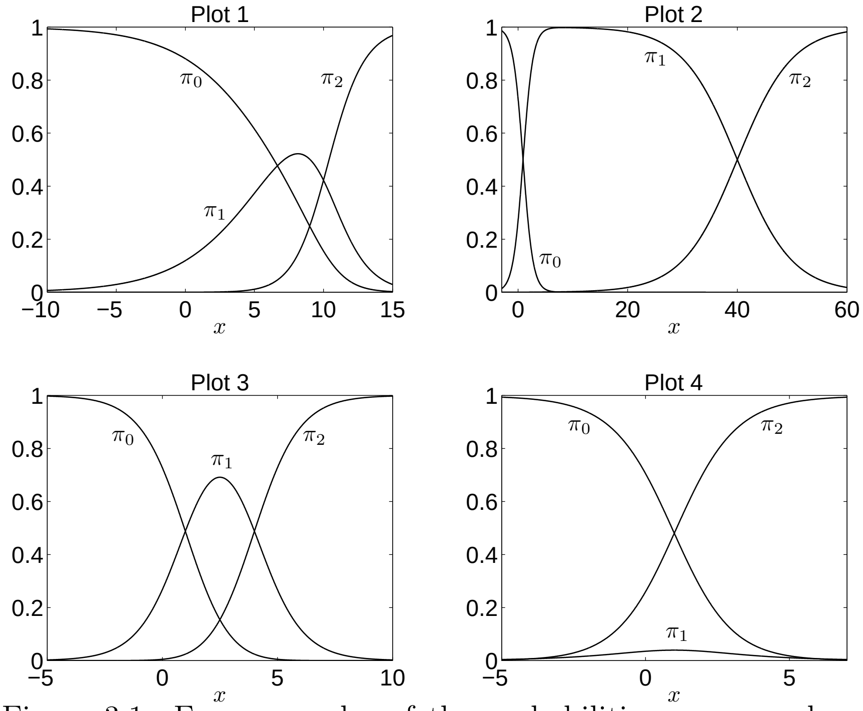


Figure 3.1: Four examples of the probabilities π_0, π_1, and π_2 as functions of x. The parameters are $\theta_1^T = (-2, -9, 0.3, 1), \theta_2^T = (-1, -9, 1.1, 1.3), \theta_3^T = (-1, -5, 1, 2)$, and $\theta_4^T = (-3, -1, 0.5, 1)$, respectively.

denotes the probability of obtaining s "successes", this imposes all cell probabilities where $\sum_{i=1}^{k} S_i = s$ to be equal to

$$\frac{\pi_s}{\binom{k}{s}}.$$

From the $2 \times 2 \times \ldots \times 2$ table, local tables can be formed. A local table for any two of the variables $S_1, S_2, \ldots, S_k$ is a 2×2 subset from the $2 \times 2 \times \ldots \times 2$ table. The local table is formed so that the outcomes of the other $k - 2$ variables are the same in all four cells. Since $S_1, S_2, \ldots, S_k$ are identically distributed there are only $k - 1$ unique local tables. As an example, consider the case with $k = 3$. The $2 \times 2 \times 2$ table for (S_1, S_2, S_3) is given in Figure 3.2. The left and right box shows the local table for (S_1, S_2) when $S_3 = 0$ and the local table for (S_1, S_2) when

$S_3 = 1$, respectively. Note that a local table is not formed by conditioning on the other $k-2$ variables.

$S_3 = 0$

	$S_2 = 0$	$S_2 = 1$
$S_1 = 0$	π_0	$\frac{\pi_1}{3}$
$S_1 = 1$	$\frac{\pi_1}{3}$	$\frac{\pi_2}{3}$

$S_3 = 1$

	$S_2 = 0$	$S_2 = 1$
$S_1 = 0$	$\frac{\pi_1}{3}$	$\frac{\pi_2}{3}$
$S_1 = 1$	$\frac{\pi_2}{3}$	π_3

Figure 3.2: The joint probabilities for (S_1, S_2, S_3) viewed as a $2 \times 2 \times 2$ table. The two local tables for (S_1, S_2) when $S_3 = 0$ and $S_3 = 1$ are enclosed by boxes.

An arbitrary local table is given by

	$S_j = 0$	$S_j = 1$
$S_i = 0$	$\frac{\pi_{s-2}}{\binom{k}{s-2}}$	$\frac{\pi_{s-1}}{\binom{k}{s-1}}$
$S_i = 1$	$\frac{\pi_{s-1}}{\binom{k}{s-1}}$	$\frac{\pi_s}{\binom{k}{s}}$

,

where s is determined by the value of the $k-2$ variables that are held constant, $\sum_{i=1}^{k-2} s_i = s-2$. The possible values for s are therefore $2, 3, \ldots, k$. The local log-odds ratio for an arbitrary local table is

$$\ln \Omega_s = \ln \left\{ \frac{\frac{\pi_s}{\binom{k}{s}} \frac{\pi_{s-2}}{\binom{k}{s-2}}}{\frac{\pi_{s-1}^2}{\binom{k}{s-1}^2}} \right\} = \ln \left\{ \frac{\pi_s \pi_{s-2}}{\pi_{s-1}^2} \frac{s\,(k-s+2)}{(s-1)\,(k-s+1)} \right\}. \tag{3.3}$$

In Agresti (2002), different types of independence for multidimensional tables are compared. As Agresti points out, mutual independence is the strongest type of independence. When $S_1, S_2, \ldots, S_k$ are mutually independent, every cell probability is equal to the product of its respective marginal probabilities. Let $\pi_.$ denote the univariate marginal probability of "success" for $S_1, S_2, \ldots, S_k$. Under mutual independence an arbitrary cell probability is

$$\frac{\pi_s}{\binom{k}{s}} = \pi_.^s \left(1 - \pi_.\right)^{k-s}.$$

Moreover, $\ln\Omega_s$ is then equal to

$$\begin{aligned}\ln\Omega_s &= \ln\left\{\frac{\pi_\cdot^s\left(1-\pi_\cdot\right)^{k-s}\pi_\cdot^{s-2}\left(1-\pi_\cdot\right)^{k-s+2}}{\pi_\cdot^{2(s-1)}\left(1-\pi_\cdot\right)^{2(k-s+1)}}\right\}\\ &= \ln\left\{\frac{\pi_\cdot^{2(s-1)}\left(1-\pi_\cdot\right)^{2(k-s+1)}}{\pi_\cdot^{2(s-1)}\left(1-\pi_\cdot\right)^{2(k-s+1)}}\right\}\\ &= 0.\end{aligned}$$

Hence, mutual independence among $S_1, S_2, \ldots, S_k$ implies that $\ln\Omega_s = 0$ for all local tables.

3.3 Likelihood and Fisher Information

The loglikelihood function and the Fisher information for the simplified Cox model are similar as compared to the Cox model. For the Cox model the loglikelihood function and the Fisher information matrix are outlined in Dragalin and Fedorov (2006).

The probability function of a single observation on Y under the simplified Cox model is

$$P\left(Y=y;\theta\right)=\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_k^{y_k}\left(1-\pi_1-\pi_2-\ldots-\pi_k\right)^{(1-y_1)(1-y_2)\ldots(1-y_k)}.$$

From the expression for the probability distribution of Y it follows that the distribution of Y is an exponential family. Assuming that the sample consists of N independent observations on $(S_1, S_2, \ldots, S_k)$, the likelihood function for a whole sample is

$$L\left(\theta;\mathbf{y}\right)=\prod_{i=1}^{N}\left\{\pi_{1i}^{y_{1i}}\pi_{2i}^{y_{2i}}\ldots\pi_{ki}^{y_{ki}}\left(1-\pi_{1i}-\pi_{2i}-\ldots-\pi_{ki}\right)^{(1-y_{1i})(1-y_{2i})\ldots(1-y_{ki})}\right\},$$

where $\mathbf{y}$ is a matrix with the responses from N observations on $y_0, y_1, \ldots, y_k$. The loglikelihood function is

$$l\left(\theta;\mathbf{y}\right)=\sum_{i=1}^{N}\left\{y_{1i}\eta_{1i}+\ldots+y_{ki}\eta_{ki}-\ln\left(1+e^{\eta_{1i}}+\ldots+e^{\eta_{ki}}\right)\right\}.$$

The score function of a single observation can be derived using the chain rule,

$$u\left(\theta\right)=\left(\frac{\partial\eta}{\partial\theta}\right)^{T}\left(\frac{\partial\pi}{\partial\eta}\right)^{T}\left(\frac{\partial l}{\partial\pi}\right)^{T}.$$

The derivatives are given by

$$\left(\frac{\partial\eta}{\partial\theta}\right)^{T}=\mathbf{x}^{T}=\begin{pmatrix}\frac{\partial\eta_1}{\partial\alpha_1} & \frac{\partial\eta_1}{\partial\alpha_2} & \cdots & \frac{\partial\eta_1}{\partial\alpha_k} & \frac{\partial\eta_1}{\partial\beta_1} & \frac{\partial\eta_1}{\partial\beta_2} & \cdots & \frac{\partial\eta_1}{\partial\beta_k}\\ \frac{\partial\eta_2}{\partial\alpha_1} & \ddots & & \frac{\partial\eta_2}{\partial\alpha_k} & \frac{\partial\eta_2}{\partial\beta_1} & \ddots & & \frac{\partial\eta_2}{\partial\beta_k}\\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots\\ \frac{\partial\eta_k}{\partial\alpha_1} & \frac{\partial\eta_k}{\partial\alpha_2} & \cdots & \frac{\partial\eta_k}{\partial\alpha_k} & \frac{\partial\eta_k}{\partial\beta_1} & \frac{\partial\eta_k}{\partial\beta_2} & \cdots & \frac{\partial\eta_k}{\partial\beta_k}\end{pmatrix}^{T},$$

$$\begin{aligned}\left(\frac{\partial\pi}{\partial\eta}\right)^{T} &= D=\begin{pmatrix}\frac{\partial\pi_1}{\partial\eta_1} & \frac{\partial\pi_2}{\partial\eta_1} & \cdots & \frac{\partial\pi_k}{\partial\eta_1}\\ \frac{\partial\pi_1}{\partial\eta_2} & \frac{\partial\pi_2}{\partial\eta_2} & & \frac{\partial\pi_k}{\partial\eta_2}\\ \vdots & & \ddots & \vdots\\ \frac{\partial\pi_1}{\partial\eta_k} & \frac{\partial\pi_2}{\partial\eta_k} & \cdots & \frac{\partial\pi_k}{\partial\eta_k}\end{pmatrix}\\ &= \begin{pmatrix}\pi_1\left(1-\pi_1\right) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_k\\ -\pi_1\pi_2 & \pi_2\left(1-\pi_2\right) & & -\pi_2\pi_k\\ \vdots & & \ddots & \\ -\pi_1\pi_k & -\pi_2\pi_k & & \pi_k\left(1-\pi_k\right)\end{pmatrix},\end{aligned}$$

and

$$\left(\frac{\partial l}{\partial\pi}\right)^{T}=\begin{pmatrix}\frac{y_1}{\pi_1}-\frac{(1-y_1)(1-y_2)\ldots(1-y_k)}{(1-\pi_1-\pi_2-\ldots-\pi_k)}\\ \vdots\\ \frac{y_k}{\pi_k}-\frac{(1-y_1)(1-y_2)\ldots(1-y_k)}{(1-\pi_1-\pi_2-\ldots-\pi_k)}\end{pmatrix}.$$

The matrix D is symmetric. Moreover D is equal to $\mathrm{Var}(Y)$. Using the fact that

$$D\left(\frac{\partial l}{\partial\pi}\right)^{T}=\left(y-\mu\right)$$

yields the score function for a whole sample

$$u_{\cdot}\left(\theta\right)=\begin{pmatrix}u_{\alpha_1\cdot}\left(\theta\right)\\ \vdots\\ u_{\alpha_k\cdot}\left(\theta\right)\\ u_{\beta_1\cdot}\left(\theta\right)\\ \vdots\\ u_{\beta_k\cdot}\left(\theta\right)\end{pmatrix}=\sum_{i=1}^{N}\mathbf{x}_i^{T}\left(y_i-\mu_i\right)=\sum_{i=1}^{N}\begin{pmatrix}\left(y_{1i}-\pi_{1i}\right)\\ \vdots\\ \left(y_{ki}-\pi_{ki}\right)\\ x_i\left(y_{1i}-\pi_{1i}\right)\\ \vdots\\ x_i\left(y_{ki}-\pi_{ki}\right)\end{pmatrix}.$$

The Fisher information matrix for a single observation is denoted $I(\theta, x)$ to stress that it depends on x. The Fisher information matrix is derived using the score function.

$$\begin{aligned} I(\theta, x) &= E\left[u(\theta)\, u^T(\theta)\right] \\ &= E\left[\mathbf{x}^T (y-\mu)(y-\mu)^T \mathbf{x}\right] \\ &= \mathbf{x}^T D \mathbf{x} \end{aligned}$$

$$= \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_k & x\pi_1(1-\pi_1) & -x\pi_1\pi_2 & \dots & -x\pi_1\pi_k \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & & \vdots & -x\pi_1\pi_2 & x\pi_2(1-\pi_2) & & \vdots \\ \vdots & & \ddots & -\pi_{k-1}\pi_k & \vdots & & \ddots & -x\pi_{k-1}\pi_k \\ -\pi_1\pi_k & -\pi_2\pi_k & \dots & \pi_k(1-\pi_k) & -x\pi_1\pi_k & -x\pi_2\pi_k & \dots & x\pi_k(1-\pi_k) \\ x\pi_1(1-\pi_1) & -x\pi_1\pi_2 & \dots & -x\pi_1\pi_k & x^2\pi_1(1-\pi_1) & -x^2\pi_1\pi_2 & \dots & -x^2\pi_1\pi_k \\ -x\pi_1\pi_2 & x\pi_2(1-\pi_2) & & \vdots & -x^2\pi_1\pi_2 & x^2\pi_2(1-\pi_2) & & \vdots \\ \vdots & & \ddots & -x\pi_{k-1}\pi_k & \vdots & & \ddots & -x^2\pi_{k-1}\pi_k \\ -x\pi_1\pi_k & -x\pi_2\pi_k & \dots & x\pi_k(1-\pi_k) & -x^2\pi_1\pi_k & -x^2\pi_2\pi_k & \dots & x^2\pi_k(1-\pi_k) \end{pmatrix}$$

$$= \mathbf{x}^* \otimes D,$$

where

$$\mathbf{x}^* = \begin{pmatrix} 1 & x \\ x & x^2 \end{pmatrix}.$$

Although the likelihood function for the simplified Cox model can be expressed explicitly with a relatively simple expression, it is sometimes problematic to numerically obtain maximum likelihood estimates. If the data are such that all observations on Y fall in the same response category, no estimates can be obtained. As an example in the bivariate case, this means e.g. that $S = 2$ for all $i = 1, 2, \dots, N$, observing only pairs of "successes". Then no information about the relationship between θ, and (π_0, π_1, π_2) is provided, and therefore no maximum likelihood estimates can be obtained. In general the estimation procedure is sensitive against situations where the number of observations falling into a particular response category is close to zero.

3.4 A Distance Covariate

In the Introduction, an experiment with plants growing in common soil was briefly described. Assume that the plants grow in pairs, so that independent observations are made on (S_1, S_2). In this situation it is realistic to include both a covariate for a fertilizer (x) and a covariate for the distance between the plants (z). Alternatively, z can also represent differences in time.

For a bivariate simplified Cox model, the dependence between (S_1, S_2) is described by the log-odds ratio, $\ln \Omega$. When a distance covariate is included, it is natural to let $\ln \Omega$ depend on z. For example, if the distance between two plants is small, the dependence between them is strong. In general the dependence between the variables is a function of both the treatment x and the distance z. The linear predictors are then given by

$$\begin{aligned} \eta_1 &= \alpha_1 + \beta_1 x \\ \eta_2 &= \alpha_2 + \beta_2 x + h\left(z, \gamma\right), \end{aligned}$$

where $h\left(z, \gamma\right)$ is a function of z and the parameter γ. Usually $h\left(z, \gamma\right)$ is a decreasing function in z. As an example let $\alpha_2 = 2\alpha_1 - \ln 4$, $\beta_2 = 2\beta_1$, and $h\left(z, \gamma\right) = \frac{\gamma}{z}$, then the log-odds ratio does not depend on x and the log-odds ratio becomes

$$\ln \Omega = \frac{\gamma}{z}.$$

In this setup the dependence between S_1 and S_2 is a function of the parameter γ and z, $z \neq 0$. In particular, note that $\gamma = 0$ is equivalent to independence between S_1 and S_2. Moreover, assuming that $\gamma > 0$ and $z > 0$ the dependence between the variables decreases as z gets larger.

In order to see how the probabilities π_0, π_1, and π_2 change with z, let $\alpha_1 = 0$, $\beta_1 = 0.5$, and $x = \ln 4$ so that π_0, π_1, and π_2 are functions of z only. Two examples using $h\left(z, \gamma\right) = \frac{\gamma}{z}$ but with different values of γ are given in Figure 3.3. In the upper plot, $\gamma = -3$ and hence the correlation between S_1 and S_2 is negative. A large negative correlation yields a large value on π_1, as demonstrated in the plot. Given that S_1 is a "success", the probability that S_2 is a "failure" is high and the other way around. Note that both plots use a logarithmic scale on the x-axis.

A different situation is described in the lower plot where $\gamma = 3$. For values of z close to zero, there is a large positive correlation between S_1 and S_2. Given that S_1 is a "success", the probability that S_2 is also a "success" is high and vice versa. Therefore π_2 is close to one for z close to zero. When z gets larger the correlation between S_1 and S_2 tends to zero regardless of γ. In other words, a large value on z has little influence on the dependence between S_1 and S_2. In both the upper and lower plot, the right y-axis shows $\ln\Omega$ as a function of z. The lower plot shows that a large positive $\ln\Omega$ has a strong influence on π_0, π_1, and π_2. When $\ln\Omega$ is large, the probabilities π_0, π_1, and π_2 are constrained to be around zero, zero, and one, respectively.

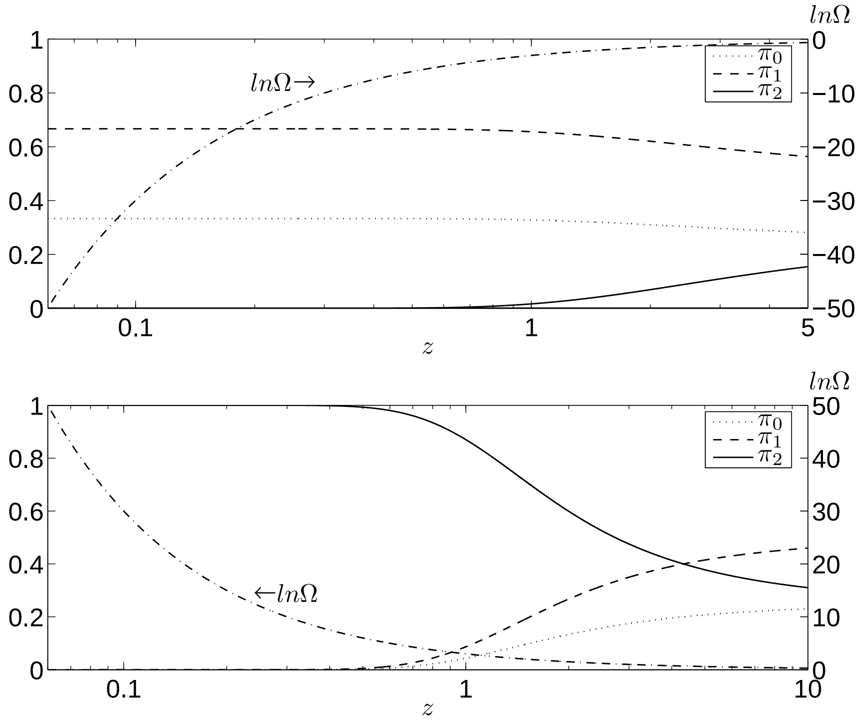


Figure 3.3: The probabilities π_0, π_1, and π_2 as functions of z for two different values of γ. For both plots, $\alpha_1 = 0$, $\beta_1 = 0.5$, and $x = \ln 4$, respectively. In the upper plot $\gamma = -3$ and in the lower plot $\gamma = 3$. In both plots, the x-axis has a logarithmic scale.The right y-axis shows the log-odds ratio, $ln\Omega = \frac{\gamma}{z}$.

3.5 Extension to the Polytomous Case

In this section the simplified Cox model is generalized to incorporate polytomous variables, $S_1, S_2, \ldots, S_k$. As an example, consider the experiment with cariogenic effect of different diets introduced in Chapter 1. For this experiment it is reasonable that each occlusal surface, S_i, has three possible responses, ”no caries”, ”caries in enamel”, and ”caries in dentin”. A polytomous model has the same structure as the one for binary data. Mainly it is just the dimensions of the different components such as μ, η, and I that increase. To limit the complexity of the notation, only the case with three response categories is presented here. An extension to arbitrarily many response categories follows the same structure.

Let $S_1, S_2, \ldots, S_k$ be k identically distributed and possibly dependent variables. Furthermore, let the probability for the different outcomes of S_i be

$$\pi_{j\cdot} = P\left(S_i = j\right) \quad j = 0, 1, 2 \;\text{ and }\; i = 1, 2, \ldots, k,$$

so that each S_i has three different response categories. Note that S_i is assumed to be a nominal variable without any ordering among the outcomes. In some applications, such as the experiment with cariogenic effects, it is more convenient to use a model that incorporates the ordering between the outcomes of responses. Some models for ordinal data are mentioned in Chapter 2. Henceforth, all properties of the model are derived for a particular observation on $(S_1, S_2, \ldots, S_k)$. Let Y denote the response matrix with elements

$$Y_{ij} = \begin{cases} 1, \text{ if } S_i = j \\ 0, \text{ otherwise} \end{cases} \quad j = 1, 2 \;\text{ and }\; i = 1, 2, \ldots, k.$$

The number of outcomes in the respective category is defined by

$$Y_j = \sum_{i=1}^{k} Y_{ij} \quad j = 1, 2,$$

so that

$$Y_1 + Y_2 \leq k.$$

In this model, the probability of the different outcomes of (Y_1, Y_2) are associated with a covariate using a multinomial logit model. The vector of probabilities for the outcome of (Y_1, Y_2) is

$$\pi_{10} \quad \pi_{20} \quad \ldots \quad \pi_{k0} \quad \pi_{01} \quad \pi_{11} \quad \ldots \quad \pi_{k-1,1} \quad \ldots \quad \pi_{0k} \ ,$$

where $\pi_{y_1 y_2}$ is the element of the vector corresponding to $P\left(Y_1 = y_1, Y_2 = y_2\right)$. Using the outcome $(0,0)$ as a reference category, the linear predictor becomes,

$$\eta = \ln \left[\frac{1}{\pi_{00}} \begin{pmatrix} \pi_{00} \\ \pi_{10} \\ \pi_{20} \\ \vdots \\ \pi_{k0} \\ \pi_{01} \\ \pi_{11} \\ \vdots \\ \pi_{k-1,1} \\ \pi_{02} \\ \vdots \\ \pi_{0k} \end{pmatrix} \right] = \begin{pmatrix} \eta_{00} \\ \eta_{10} \\ \eta_{20} \\ \vdots \\ \eta_{k0} \\ \eta_{01} \\ \eta_{11} \\ \vdots \\ \eta_{k-1,1} \\ \eta_{02} \\ \vdots \\ \eta_{0,k} \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha_{10} + \beta_{10}x \\ \alpha_{20} + \beta_{20}x \\ \vdots \\ \alpha_{k0} + \beta_{k0}x \\ \alpha_{01} + \beta_{01}x \\ \alpha_{11} + \beta_{11}x \\ \vdots \\ \alpha_{k-1,1} + \beta_{k-1,1}x \\ \alpha_{02} + \beta_{02}x \\ \vdots \\ \alpha_{0k} + \beta_{0k}x \end{pmatrix}.$$

There are $\binom{k+2}{2}$ response categories and accordingly $2\left\{\binom{k+2}{2} - 1\right\}$ parameters in the model. The vector of parameters is

$$\theta^T = \begin{pmatrix} \alpha_{10} & \alpha_{20} & \ldots & \alpha_{0k} & \beta_{10} & \beta_{20} & \ldots & \beta_{0k} \end{pmatrix}.$$

The probabilities π as functions of x are

$$\pi_{y_1 y_2} = \frac{e^{\eta_{y_1 y_2}}}{\sum_{\substack{i,j \geq 0 \\ i+j \leq k}} e^{\eta_{ij}}} \quad y_1, y_2 \geq 0 \text{ and } y_1 + y_2 \leq k. \tag{3.4}$$

As for the model with binary data, the joint probability distribution of $S_1, S_2, \ldots, S_k$ can also be viewed as a $k-$dimensional table. The table has dimensions $3 \times 3 \times \ldots \times 3$ where each cell corresponds to a unique sequence of $S_1, S_2, \ldots, S_k$, where $S_i = \{0, 1, 2\}$, $i = 1, 2, \ldots, k$. The cell probability for all cells with $(Y_1 = y_1, Y_2 = y_2)$ is equal to

$$\frac{\pi_{y_1 y_2}}{\begin{pmatrix} & k & \\ y_1 & & y_2 \end{pmatrix}}.$$

The denominator is the multinomial coefficient defined as

$$\binom{k}{y_1 \quad y_2} = \frac{k!}{y_1!y_2!\left(k - y_1 - y_2\right)!}.$$

In analogy to the case with binary outcomes, the $k-$dimensional table can be divided into local tables. In the case of three outcomes, three different local tables exist. Note that interest is only in the symmetrical local tables. Thus, out of the possible nine local tables only the three cases, $\left(S_i = 1, S_j = 0\right)$, $\left(S_i = 2, S_j = 0\right)$, and $\left(S_i = 2, S_j = 1\right)$ for $i, j = 1, 2, \ldots, k$ and $i \neq j$ are considered. Let S^* be the set of the other variables, $S^* = \left\{S_1, S_2, \ldots, S_k \backslash \left(S_i, S_j\right)\right\}$. Then, the three different local odds ratios for S_i and S_j with $S^* = s^*$ are

$$\ln \Omega_{S_i=1,S_j=0;s^*} = \ln \left[\frac{\frac{\pi_{y_1 y_2}}{\binom{k}{y_1 \quad y_2}} \frac{\pi_{y_1-2 y_2}}{\binom{k}{y_1 - 2 \quad y_2}}}{\left\{ \frac{\pi_{y_1-1 y_2}}{\binom{k}{y_1 - 1 \quad y_2}} \right\}^2} \right], \qquad (3.5)$$
$$y_1 \geq 2, y_2 \geq 0 \text{ and } y_1 + y_2 \leq k$$

$$\ln \Omega_{S_i=2,S_j=0;s^*} = \ln \left[\frac{\frac{\pi_{y_1 y_2}}{\binom{k}{y_1 \quad y_2}} \frac{\pi_{y_1 y_2-2}}{\binom{k}{y_1 \quad y_2 - 2}}}{\left\{ \frac{\pi_{y_1 y_2-1}}{\binom{k}{y_1 \quad y_2 - 1}} \right\}^2} \right], \qquad (3.6)$$
$$y_1 \geq 2, y_2 \geq 2 \text{ and } y_1 + y_2 \leq k$$

$$\ln \Omega_{S_i=2,S_j=1;s^*} = \ln \left[\frac{\frac{\pi_{y_1 y_2-2}}{\binom{k}{y_1 \quad y_2 - 2}} \frac{\pi_{y_1-2 y_2}}{\binom{k}{y_1 - 2 \quad y_2}}}{\left\{ \frac{\pi_{y_1-1 y_2-1}}{\binom{k}{y_1 - 1 \quad y_2 - 1}} \right\}^2} \right], \qquad (3.7)$$
$$y_1, y_2 \geq 2 \text{ and } y_1 + y_2 \leq k + 2.$$

Expressions for the likelihood function, the score function, and the Fisher information matrix follow straight forwardly from the corresponding expressions for the model with binary outcomes.

Chapter 4

Optimal Designs

4.1 Introduction

In an ideal experimental study, explanatory variables or covariates perfectly govern the value of the response variables of the experiment. In practice though, other unobservable factors also influence the outcome of the experiment. To minimize the influence of these factors, more covariates can be added to the model. Usually some of the techniques matching, blocking, balancing, blinding, or double blinding are also employed. What hopefully remains of the unobservable variation is then just small random errors.

Assuming that the experimenter has chosen one explanatory variable, the design of the experiment needs to be established. By determining which levels of the explanatory variable to use and the corresponding number of observations to be allocated to each level the design of the experiment is established. A design is therefore usually written in the form

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \ldots & x_n \\ N_1 & N_2 & \ldots & N_n \end{array} \right\},$$

where $x_1, x_2, \ldots, x_n$ are the chosen levels, called design points. N_1, $N_2, \ldots, N_n$ are the corresponding number of observations to be taken at the different design points, $\sum_{i=1}^{n} N_i = N$. For the design to be realizable all N_i need to be integers. Such a design is referred to as an exact design. In practice, when deriving optimal designs the restriction that N_i is an integer is often relaxed yielding a continuous design. The reason why continuous designs are preferable is that continuous optimization problems are usually mathematically and numerically less cumbersome to work with than discrete optimization problems. For a continuous design it is more convenient to denote the design as

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_n \\ w_1 & w_2 & \dots & w_n \end{array} \right\},$$

where $w_1, w_2, \dots, w_n$ are design weights, satisfying

$$w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1.$$

The design weights determine the proportion of observations to be taken at the different design points. Only continuous designs are considered in this thesis. In practice, these designs will only be approximate optimal designs, since the design weights of the continuous design have to be rounded in order for the design to be realizable. Consequently, the exact optimal design for the same sample size may differ considerably, making the efficiency of the rounded design lower.

4.2 Optimality Criteria

A design is optimal according to a specific criterion if it minimizes the corresponding criterion function, ψ. Formally, the design ξ^* is $\psi-$optimal if

$$\xi^* = \arg\min_{\xi \in \Xi} \psi\left(\theta, \xi\right),$$

where Ξ is the set of all possible designs. The choice of criterion function is controlled by the objectives of the experiment. These objectives are usually connected to the precision in the parameter estimators. Given regularity conditions, see e.g. Casella and Berger (2002), the covariance matrix of the maximum likelihood estimator is asymptotically equal to the inverse of the Fisher information matrix for the whole sample. Therefore, many optimality criteria optimize some function of the Fisher information matrix. Two of these criteria, D-optimality and E-optimality, are outlined in the two coming sections. A more comprehensive description of different optimality criteria are given in Atkinson and Donev (1992). The Fisher information matrix is positive semi-definite, symmetric, and additive (for independent observations), Fedorov and Hackl

(1997). When deriving optimal designs in this thesis, the cost for taking an observation is not incorporated in the criterion function. If the criterion function would consider the cost of different observations, the information matrix at each design point would be weighted with the cost of taking an observation at that design point.

In order to stress that the information matrix is a sum of independent observations from the design ξ, the information matrix is denoted

$$I_{\cdot}\left(\theta,\xi\right).$$

Optimal design theory uses the standardized information matrix, denoted

$$M\left(\theta,\xi\right)=\frac{I_{\cdot}\left(\theta,\xi\right)}{N},$$

rather than the Fisher information matrix.

Let

$$\overline{\xi}_x=\left\{\begin{array}{c} x \\ 1 \end{array}\right\}$$

denote the design which puts unit mass at the point x. The directional derivative of $\psi\left\{M\left(\theta,\xi\right)\right\}$ in the direction of $\overline{\xi}$ is

$$\phi\left(\theta,x,\xi\right)=\lim_{\alpha\longrightarrow 0+}\frac{1}{\alpha}\left[\psi\left\{\left(1-\alpha\right)M\left(\theta,\xi\right)+\alpha M\left(\theta,\overline{\xi}_x\right)\right\}-\psi\left\{M\left(\theta,\xi\right)\right\}\right],$$

see Atkinson and Donev (1992). The General Equivalence Theorem, (Kiefer, 1959; Kiefer and Wolfowitz, 1960), states the equivalence of the following three conditions for ξ^* to be $\psi-$optimal.

1. $\xi^*=\arg\min\limits_{\xi\in\Xi}\psi\left(\theta,\xi\right)$
2. $\min\limits_{x\epsilon\mathfrak{X}}\phi\left(\theta,x,\xi^*\right)\geqslant 0$
3. $\phi\left(\theta,x,\xi^*\right)$ attains its minimum at all the design points.

$\mathfrak{X}$ is called design region and specifies the possible values of x. Using these statements, The General Equivalence Theorem is used to verify that a design really is optimal.

4.2.1 Optimal Designs for Nonlinear Models

For nonlinear models, optimal designs generally depend on the true and unknown values of the parameters. Typically, $M(\theta, \xi)$ depends on the parameter vector θ as well as the design ξ. So in order to get optimal parameter estimates, it is required that the true values of the parameters are known. To handle this dilemma at least four strategies have been proposed.

Locally Optimal Designs

By treating guessed parameter values as the true parameter values, a locally optimal design can be derived. A guess can, e.g. be based on prior knowledge. Thus, a locally optimal design is optimal only in case the true value on the parameter vector equals the particular value chosen when determining the design. If the true value of the parameter vector is different from that chosen for determining the design, there is no guarantee that the design has any favorable properties.

Sequential Optimal Designs

This method can be described as an iterative procedure where locally optimal designs and parameter estimates are obtained in each step. From an initial design a subexperiment is conducted yielding estimates of the parameters.

Next a weighted information matrix is constructed using these parameter estimates. This weighted information matrix is the weighted sum of the information from the previous steps and the information obtained in the current step. By applying this weighted information matrix in the criterion function, a new design is derived. Then a new subexperiment based on the new design yields new estimates of the parameters, and so on. An advantage with a sequential design is that a poor initial guess of the parameter values, can be corrected as more information about the parameters are obtained from the following subexperiments. The major drawback is that it could be time consuming to obtain an optimal sequential design. Therefore, it is not feasible to use sequential designs in some applications such as afforestation experiments. An overview of references on sequential designs is given in Wang (2002).

Optimum in Average Designs

Assume that there is a prior distribution for the parameters denoted $\vartheta(\theta)$. This distribution reflects the experimenter's belief in different parameter values. A function for deriving an optimum in average design is then obtained by weighting different values of the criterion function using $\vartheta(\theta)$,

$$B(\vartheta, \xi) = \int \psi(\theta, \xi) \, d\vartheta(\theta).$$

A design, ξ^{ϑ}, is optimum in average with respect to the prior distribution $\vartheta(\theta)$ if

$$B(\vartheta, \xi^{\pi}) = \inf_{\xi \in \Xi} B(\vartheta, \xi).$$

For continuous prior distributions, the evaluation of the integral above is a potential problem. Atkinson and Haines (1996) suggest some approaches to solve this problem, such as discretizing the prior and then work with a weighted sum instead. Due to the incorporated prior distribution, optimum in average designs are generally more robust than locally D-optimal designs. Optimal in average designs in general are treated in Chaloner and Larntz (1989), Fedorov and Hackl (1997), Pettersson (2001), and Pettersson and Nyquist (2003).

Minimax Designs

In a minimax approach it is believed that θ belongs to a subset $\Theta_0 \subset \Theta$ of the parameter space Θ. For each design, $\xi \in \Xi$, the maximum value of the criterion function, $\max\limits_{\theta \in \Theta_0} \psi(\theta, \xi)$, can then be derived. ξ^M is a minimax design if the maximum of the criterion function, the maximum taken over Θ_0, is minimized for ξ^M. Thus, a minimax design ξ^M satisfies

$$\max_{\theta \in \Theta_0} \psi\left(\theta, \xi^M\right) = \min_{\xi \in \Xi} \max_{\theta \in \Theta_0} \psi(\theta, \xi).$$

Minimax designs also have a robustness property. Compared to a locally optimal design, a minimax design can not be too bad as long as $\theta \in \Theta_0$, i.e. Θ_0 is large enough, Häggström (2000). In practice though, it is mathematically and numerically difficult to derive minimax designs, (Fedorov and Hackl, 1997; Häggström, 2000).

Optimal designs for nonlinear models in general are treated in e.g. Silvey (1980), Atkinson and Donev (1992), Atkinson and Haines (1996), and

Fedorov and Hackl (1997). A short description of the two criteria used in this thesis is given below.

4.2.2 D-optimality

The criterion function for D-optimality is

$$\psi \left\{ M\left(\theta,\xi\right)\right\} = \ln \left| M^{-1}\left(\theta,\xi\right)\right|,$$

where $|A|$ stands for the determinant of A. A majority of the articles mentioned above have considered the D-optimality criterion. This criterion is appealing for at least two reasons. The determinant of the inverse of the standardized information matrix is proportional to the generalized volume of the confidence ellipsoid of the parameters. Hence, a smaller value of the criterion function leads to greater precision in the parameter estimates. Moreover, if no particular subset of the parameters or linear combination of the parameters is of interest the D-optimality criterion is suitable.

Silvey (1980) showed that the directional derivative of the criterion function for a D-optimal design is

$$\phi\left(\theta,x,\xi\right) = p - d(x,\xi),$$

where p is the number of parameters in the model and $d(x,\xi)$ denotes the standardized variance of the predicted response,

$$d(x,\xi) = tr\left\{ M\left(\theta,\xi\right)^{-1} M\left(\theta,\overline{\xi}_x\right)\right\} \quad \forall x \epsilon \mathfrak{X}.$$

Although the standardized variance of the predicted response depends on θ in general, it is denoted $d(x,\xi)$ throughout this thesis. The above expression for $\phi\left(\theta,x,\xi\right)$ is very useful when inserted in the General Equivalence Theorem. It then follows directly that a design, ξ^*, that satisfies

$$d(x,\xi^*) \leqslant p \quad \forall x \epsilon \mathfrak{X}$$

is D-optimal. Furthermore, the General Equivalence Theorem states that $d(x,\xi^*) = p$ at the design points. These two conditions on $d(x,\xi)$ are then used when deriving a D-optimal design and to verify if a design is D-optimal or not.

Example 4.1

To illustrate a D-optimal design, Figure 4.1 shows π_0, π_1, and π_2 for a bivariate example of the simplified Cox model. The parameters are $\alpha_1 = -3$, $\alpha_2 = -12$, $\beta_1 = 0.7$, and $\beta_2 = 1.3$. In Figure 4.1, $d(x, \xi^*)$ for the locally D-optimal design

$$\xi^* = \left\{ \begin{array}{cccc} 2.4661 & 7.1641 & 11.9644 & 17.1245 \\ 0.2472 & 0.2908 & 0.2205 & 0.2415 \end{array} \right\}$$

is also included. Note that $d(x, \xi^*) = 4 = p$ at the design points. Since this result is in line with the General Equivalence Theorem it follows that ξ^* is a locally D-optimal design.

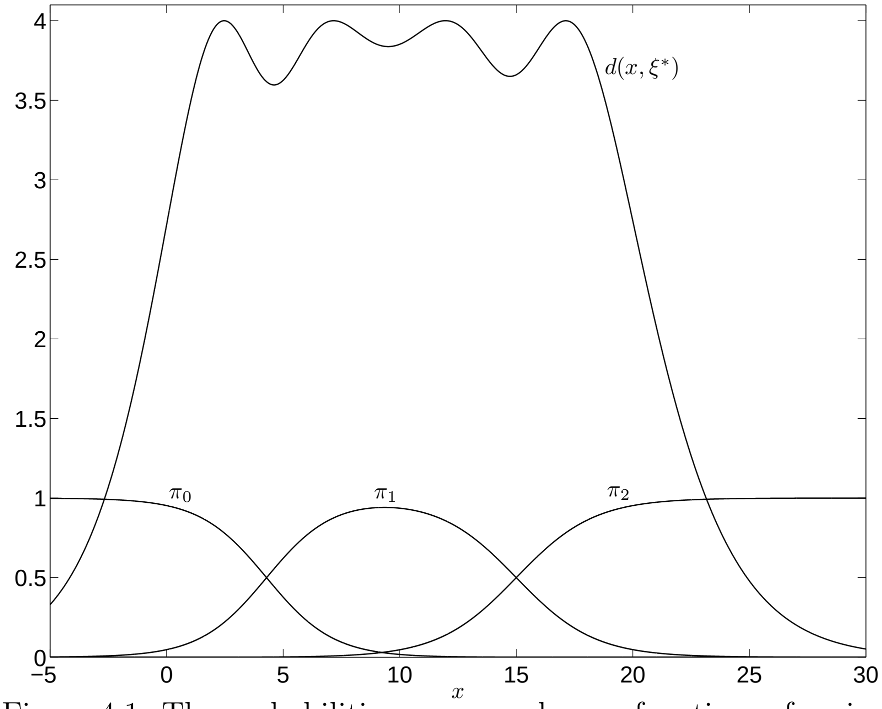


Figure 4.1: The probabilities π_0, π_1, and π_2 as functions of x given $\alpha_1 = -3, \alpha_2 = -12, \beta_1 = 0.7$ and $\beta_2 = 1.3$. The figure also shows the standardized variance of the predicted response for a D-optimal design $\xi^*, d(x, \xi^*)$.

4.2.3 E-optimality

E-optimal designs minimize the variance of the worst estimated linear contrast, $a^T\theta$. An E-optimal design, ξ_E, is defined by

$$\xi_E = \arg\min_{\xi\in\Xi}\max_{i=1,\ldots,p}\frac{1}{\lambda_i},$$

where $\frac{1}{\lambda_i}$ is an eigenvalue to $M\left(\theta,\xi\right)^{-1}$. The E-optimal design is interpreted as the design that minimizes the length of the long axis of the confidence ellipsoid of the parameters, Pettersson and Nyquist (2003).

The directional derivative of the criterion function for an E-optimal design is

$$\phi\left(\theta,x,\xi\right) = \lambda_{\min} - v^T M\left(\theta,\overline{\xi}_x\right)v, \tag{4.1}$$

Atkinson and Donev (1992). $\lambda_{\min}$ is the smallest eigenvalue of $M\left(\theta,\xi\right)$ and v^T is the corresponding eigenvector. The expression in (4.1) can be used to verify that a design is E-optimal.

Example 4.2

Assume that the parameters are the same as in Example 4.1 above. With these parameter values a locally E-optimal design, denoted ξ_E, is a $3-$point design with unequal design weights,

$$\xi_E = \left\{\begin{array}{ccc} 1.8657 & 10.7990 & 18.9416 \\ 0.2272 & 0.5134 & 0.2594 \end{array}\right\}.$$

When compared, ξ_E and the locally D-optimal design in Example 4.1, ξ^*, are very similar. The main difference is that two of the design points in ξ^*, at 7.1641 and at 11.9644, are merged into one design point in ξ_E, at 10.7990. The directional derivative for ξ_E is given in Figure 4.2. Figure 4.2 shows that $\phi\left(\theta,x,\xi_E\right)$ achieves its minima at the design points and that the minimum value is equal to zero. According to the General Equivalence Theorem, ξ_E is therefore E-optimal.

The D-optimality criterion and the E-optimality criterion have in common that their respective criterion function is a function of the standardized information matrix, and consequently also a function of the unknown parameters θ. Therefore, the derived optimal designs in this

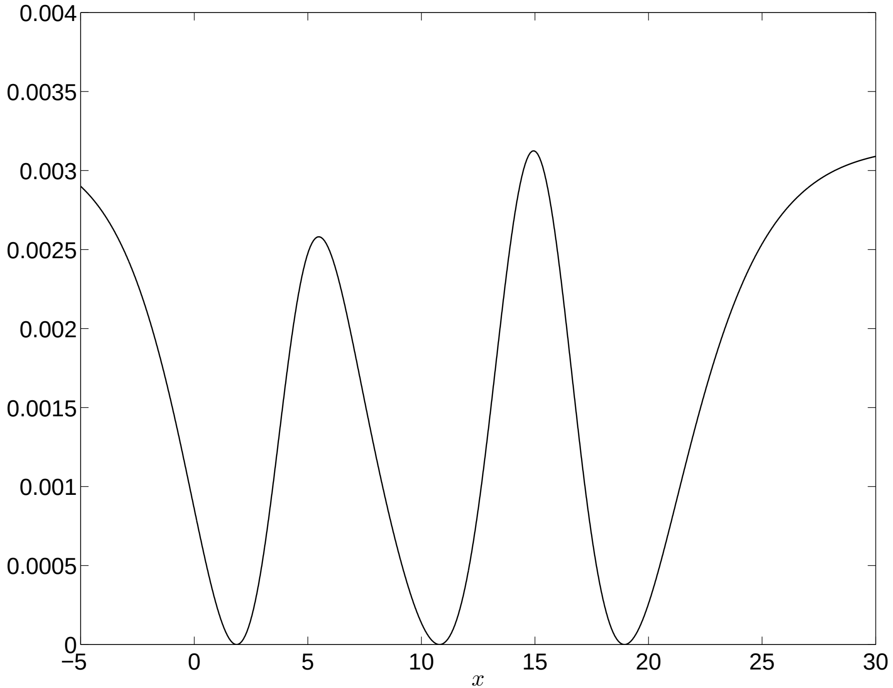


Figure 4.2: Directional derivative function, $\phi(x, \xi_E)$, for the E-optimal design ξ_E. The parameters are $\alpha_1 = -3, \alpha_2 = -12, \beta_1 = 0.7$, and $\beta_2 = 1.3$, respectively.

thesis are locally optimal designs. Furthermore, there is in general no closed form formula that defines an optimal design, it must be numerically determined using some routines for function optimization. In this thesis routines in Mathcad and MATLAB have been used.

As outlined previously, the optimal designs are derived based on the asymptotic covariance matrix of the maximum likelihood estimator $\widehat{\theta}$. In practice, the performance of a particular design depends on how well the asymptotic sampling distribution resembles the actual sampling distribution used in the experiment.

4.3 Concluding Remarks

This section gives a short and by no means complete overview of references concerning optimal designs for dependent Bernoulli variables.

There is a large amount of articles in which authors address a dose-finding experiment in clinical trials, where the binary responses efficacy (yes/no) and toxicity (yes/no) have a joint distribution. Heise and Myers (1996) derived locally D-optimal designs for this situation using the Gumbel model described above. The locally D-optimal designs are derived for different values on the parameters. The results show that the design points are often symmetrically allocated about some ratio of the parameters. They also studied locally Q-optimal designs. The Q-optimal designs minimize the predicted variance of the response (”efficacy”,”no toxicity”). Dragalin and Fedorov (2006) presented locally D-optimal designs based on Cox bivariate binary model. They include a penalty function in the criterion function in order to avoid situations where the covariate attains unethical or impractical values.

Other authors have used the trinomial model with response categories ”no response”, ”efficacy”, and ”adverse reaction”. Puu (2003) considered locally D- and D$_A$-optimal designs for a multinomial logit model. Zocchi and Atkinson (1999) derived D-optimal in average designs and compare them with locally D-optimal designs for a trinomial model.

Appelgren (2004) derived locally D-optimal designs for the bivariate logistic regression model, (McCullagh and Nelder, 1989). He studied models with independent margins as well as models with dependent margins. Results show that the parameters for the margins are most important for the location of the design points. The locally D-optimal designs have two, three, or four design points.

Fan (1999) and Fan and Chaloner (2004) considered locally D-optimal designs, D-optimal in average designs, and locally c-optimal designs for a continuation-ratio logit model. They derived analytical expressions for a design, which is referred to as limiting locally D-optimal design. The design is not optimal but it tends to be optimal when a certain difference between the parameters tends to infinity. The limiting locally D-optimal design proves to be useful in that an analytical expression can be found when the ordinary locally optimal design has to be determined numerically. Rabie (2004) also worked with locally D-optimal designs, locally c-optimal designs, and limiting locally D-optimal designs but for the contingent response model.

All the above authors have addressed a situation where the N observations are assumed to be independent. As an example, in a model for trinomial responses, the outcomes of a response is dependent but the responses from different experimental units are assumed to be independent. Müller and Pázman have in a series of articles developed theory for optimal designs in a model with correlated observations, Müller and Pázman (1998, 1999, 2001, 2003). The main obstacle when deriving an optimal design for correlated observations is the non-additivity, and consequently, the non-differentiability of the information matrix. Müller and Pázman handle this by deriving a differentiable approximation of the information matrix. Müller and Pázman primarily worked with a linear model, particularly useful for spatial data in situations where observations cannot be replicated.

Chapter 5

Bivariate Symmetric Model

For a bivariate simplified Cox model, $\ln \Omega$ was given in (3.2). As mentioned above, the dependence between the variables is described by $\ln \Omega$. The restriction $\beta_2 = 2\beta_1$ implies that the log-odds ratio between the variables is constant and does not depend on x. Hence under the restriction $\beta_2 = 2\beta_1$, the dependence between the variables is the same for all values on x. Furthermore, $\pi_0 \left(x_0 - d\right) = \pi_2 \left(x_0 + d\right)$ for all values on the constant d, where x_0 is defined as

$$x_0 = \arg \max_{x \in \mathfrak{X}} \pi_1$$

The last property is referred to as a symmetry property in this thesis. A model under the above restriction is studied in this chapter using bivariate data, (S_1, S_2). Note that although $\ln \Omega$ does not depend on x, S_1 and S_2 may still be dependent.

5.1 Some Properties

The model is a particular case of the simplified Cox model, but since $\beta_2 = 2\beta_1$, some properties are more specific. Compared with the simplified Cox model for two variables, the response variable and the link function are left unchanged. The linear predictor changes to

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 + \beta_1 x \\ \alpha_2 + 2\beta_1 x \end{pmatrix} = \mathbf{x}\theta,$$

where

$$\mathbf{x} = \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & 2x \end{pmatrix}$$

and

$$\theta^T = \begin{pmatrix} \alpha_1 & \alpha_2 & \beta_1 \end{pmatrix}.$$

Based on the expression derived for the simplified Cox model, the log-odds ratio is

$$\ln\Omega = \ln 4 + \alpha_2 - 2\alpha_1. \tag{5.1}$$

Note that $\ln\Omega$ depends neither on β_1 nor on x. Moreover, using the expression for $\ln\Omega$ the model from Section 3.4 can easily be obtained. By setting $\alpha_2 = 2\alpha_1 - \ln 4$ and adding the term $\frac{\gamma}{z}$ to η_2, $\ln\Omega$ becomes a function of $\frac{\gamma}{z}$ only. Accordingly, the model with a distance covariate is obtained.

The probability distribution for S depends on the parameters α_1, α_2, and β_1 and the covariate x. In Figure 5.1, π_0, π_1, and π_2 are plotted for four different combinations of α_1, α_2, and β_1.

The maximum value of π_1 decreases as $\ln\Omega$ increases. In Plot 4 with $\ln\Omega_4 = 20.39$, π_1 is not possible to see since it is so close to zero. Since $\beta_1 > 0$, π_0 decreases with x and π_2 increases with x.

Property 5.1

$$x_0 = \frac{-\alpha_2}{2\beta_1}$$

Proof.

$$\frac{d\pi_1(x)}{dx} = \frac{\beta_1 e^{\eta_1} - \beta_1 e^{\eta_1+\eta_2}}{\left(1 + e^{\eta_1} + e^{\eta_2}\right)^2}$$

Equating to zero yields

$$x_0 = \frac{-\alpha_2}{2\beta_1}.$$

By applying standard calculus technique one can show that x_0 is a global maximum of π_1.

■

The term x_0 is important in obtaining D-optimal designs and to show the symmetry properties for this model.

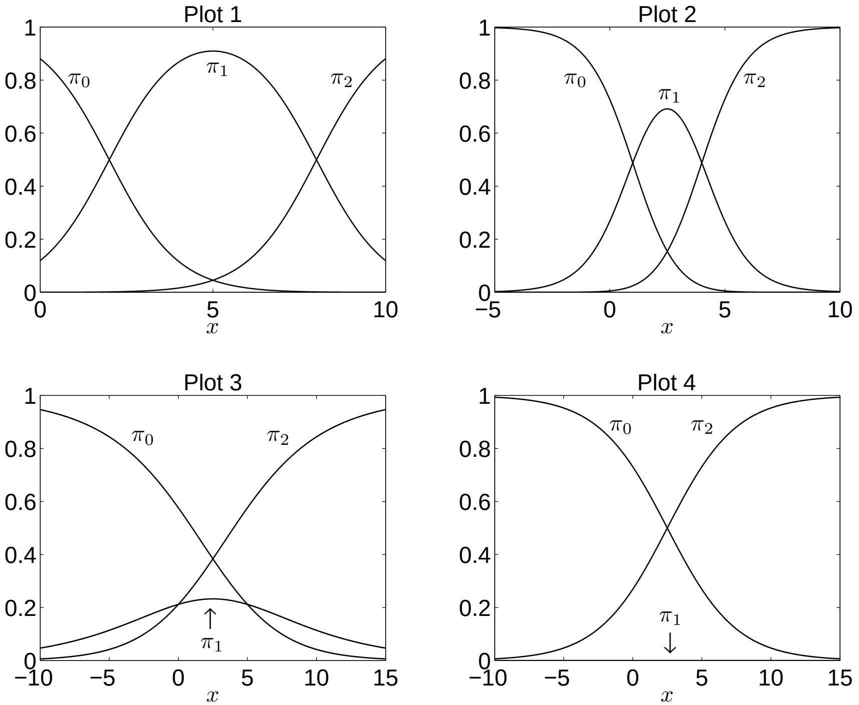


Figure 5.1: π_0, π_1, and π_2 as functions of x. The parameters are $\theta_1^T = (-2, -10, 1), \theta_2^T = (-1, -5, 1), \theta_3^T = (-1, -1, 0.2)$, and $\theta_4^T = (-10, -1, 0.2)$, respectively. The corresponding log-odds ratios are $\ln\Omega_1 = -4.61$, $\ln\Omega_2 = -1.61$, $\ln\Omega_3 = 2.39$, and $\ln\Omega_4 = 20.39$, respectively.

Property 5.2

$$\pi_1(x_0) = \frac{1}{1 + \sqrt{\Omega}}$$

Proof.

$$\pi_1(x_0) = \frac{e^{\alpha_1+\beta_1 x_0}}{1 + e^{\alpha_1+\beta_1 x_0} + e^{\alpha_2+2\beta_1 x_0}} = \frac{e^{\alpha_1-\frac{\alpha_2}{2}}}{2 + e^{\alpha_1-\frac{\alpha_2}{2}}} = \frac{1}{1+\sqrt{\Omega}}$$

■

The value of $\pi_1(x_0)$ depends only on the odds ratio Ω. Thus a very large $\ln\Omega$ gives a very small $\pi_1(x_0)$ and vice versa.

The following property shows that π_0 and π_2 are symmetric around x_0 in the sense that $\pi_0(x_0 - d) = \pi_2(x_0 + d)$.

Property 5.3

$$\pi_0\left(x_0-d\right)=\pi_2\left(x_0+d\right)\quad \textit{for all } d$$

Proof.

$$\pi_0\left(x_0-d\right)=\frac{1}{1+e^{\alpha_1+\beta_1(x_0-d)}+e^{\alpha_2+2\beta_1(x_0-d)}}=\frac{1}{1+e^{\alpha_1-\frac{\alpha_2}{2}-d\beta_1}+e^{-2d\beta_1}}$$

$$\pi_2\left(x_0+d\right)=\frac{e^{\alpha_2+2\beta_1(x_0+d)}}{1+e^{\alpha_1+\beta_1(x_0+d)}+e^{\alpha_2+2\beta_1(x_0+d)}}=\frac{1}{1+e^{\alpha_1-\frac{\alpha_2}{2}-d\beta_1}+e^{-2d\beta_1}}$$

Hence π_0 and π_2 are symmetric around x_0. ■

For the current model S_1 and S_2 are not independent in general. Expressions for the covariance and the correlation between S_1 and S_2 are derived below.

Property 5.4

$$Cov\left(S_1,S_2\right)=\frac{4e^{\eta_2}-e^{2\eta_1}}{4\left(1+e^{\eta_1}+e^{\eta_2}\right)^2}$$

Proof.

$$\begin{aligned}Cov\left(S_1,S_2\right) &= \frac{e^{\eta_2}}{1+e^{\eta_1}+e^{\eta_2}}-\pi_{1.}\pi_{.1}\\ &= \frac{4e^{\eta_2}-e^{2\eta_1}}{4\left(1+e^{\eta_1}+e^{\eta_2}\right)^2},\end{aligned}$$

since $\beta_2=2\beta_1$. ■

Property 5.5

$$Corr\left(S_1,S_2\right)=\frac{4e^{\alpha_2}-e^{2\alpha_1}}{\left(2+e^{\alpha_1+\beta_1 x}\right)\left(2e^{\alpha_2}+e^{\alpha_1-\beta_1 x}\right)}$$

Proof.

$$\begin{aligned}Corr\left(S_1,S_2\right) &= \frac{Cov\left(S_1,S_2\right)}{\sqrt{\pi_{1.}\pi_{0.}}\sqrt{\pi_{.1}\pi_{.0}}}\\ &= \frac{\frac{4e^{\eta_2}-e^{2\eta_1}}{4(1+e^{\eta_1}+e^{\eta_2})^2}}{\frac{(2+e^{\eta_1})(2e^{\eta_2}+e^{\eta_1})}{4(1+e^{\eta_1}+e^{\eta_2})^2}}\\ &= \frac{4e^{\alpha_2}-e^{2\alpha_1}}{\left(2+e^{\alpha_1+\beta_1 x}\right)\left(2e^{\alpha_2}+e^{\alpha_1-\beta_1 x}\right)},\end{aligned}$$

since $\beta_2 = 2\beta_1$.

■

The variance of S is given by

Property 5.6

$$Var\left(S\right) = \frac{4e^{\eta_2} + e^{\eta_1}\left(1 + e^{\eta_2}\right)}{\left(1 + e^{\eta_1} + e^{\eta_2}\right)^2}.$$

Proof.

$$\begin{aligned} Var\left(S\right) &= \pi_{1\cdot}\pi_{0\cdot} + \pi_{\cdot 1}\pi_{\cdot 0} + 2\frac{4e^{\eta_2} - e^{2\eta_1}}{4\left(1 + e^{\eta_1} + e^{\eta_2}\right)^2} \\ &= \frac{4e^{\eta_2} + e^{\eta_1}\left(1 + e^{\eta_2}\right)}{\left(1 + e^{\eta_1} + e^{\eta_2}\right)^2} \end{aligned}$$

■

Consider the following expression for the covariance between S_1 and S_2,

$$Cov\left(S_1, S_2\right) = \sum_{s_1=0,1}\sum_{s_2=0,1}\left(s_1 - \pi_{\cdot}\right)\left(s_2 - \pi_{\cdot}\right)P\left(S_1 = s_1, S_2 = s_2\right).$$

It was stated previously that π_0 is a decreasing function in x and π_2 is an increasing function in x if $\beta_1 > 0$. For very small x, $P\left(S_1 = s_1, S_2 = s_2\right)$ is close to zero for all s_1, s_2 except for $s_1 = s_2 = 0$. Since $\pi_{\cdot}$ is close to zero for very small x, $Cov\left(S_1, S_2\right)$ is close to zero for very small x. In a similar way $Cov\left(S_1, S_2\right)$ is close to zero for very large x. Hence, the covariance tends to zero when x tends to minus or plus infinity.

Figure 5.2 includes four plots of the correlation between S_1 and S_2 for the same parameter values as in Figure 5.1. In Plot 1 and Plot 2 the parameter values generate a negative $\ln\Omega$, hence the correlation is also negative. Plot 4 is based on parameter values that generate a large $\ln\Omega$. The correlation is therefore close to one for values of x in the interval -30 to 30. In all plots the correlation tends to zero for very large and for very small values on x.

As mentioned previously, the simplified Cox model in general allows the correlation between S_1 and S_2 to vary across treatments. This is true also for the symmetric model where the correlation varies symmetrically around x_0 as Figure 5.2 illustrates and the following property shows.

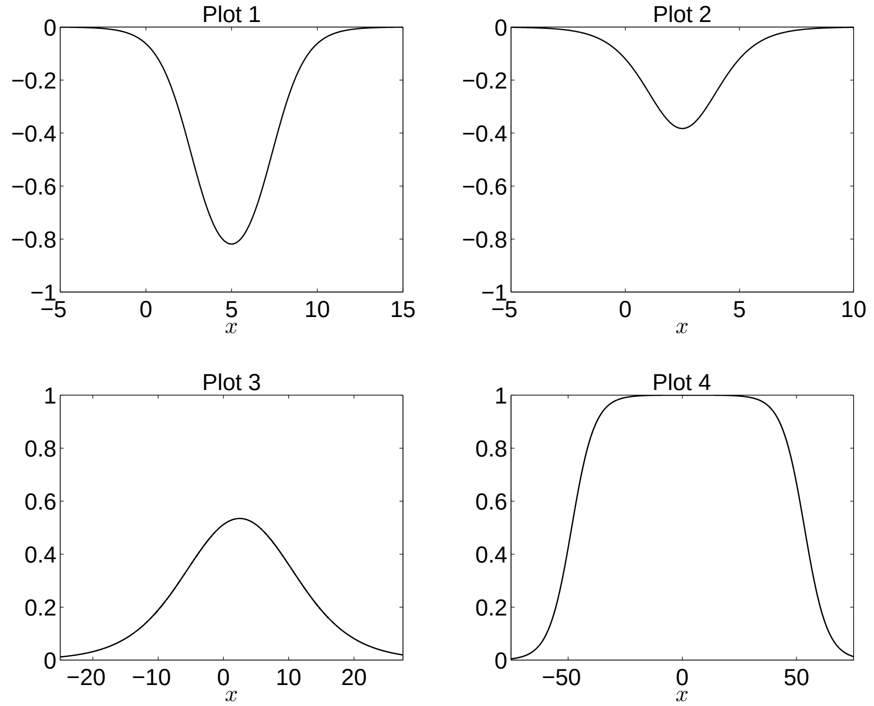


Figure 5.2: Correlation between S_1 and S_2 for different sets of parameter values. The parameters are $\theta_1^T = (-2, -10, 1), \theta_2^T = (-1, -5, 1), \theta_3^T = (-1, -1, 0.2)$, and $\theta_4^T = (-10, -1, 0.2)$, respectively. The corresponding log-odds ratios are $\ln\Omega_1 = -4.61$, $\ln\Omega_2 = -1.61$, $\ln\Omega_3 = 2.39$, and $\ln\Omega_4 = 20.39$, respectively.

Property 5.7 *$Corr\,(S_1, S_2)$ has a global minimum or maximum at $x = x_0$.*

Proof.

$$\frac{d}{dx}Corr(S_1,S_2)=\frac{\beta_1\left(e^{2\alpha_1}-4e^{\alpha_2}\right)\left\{e^{\alpha_1-\beta_1 x}\left(2+e^{\alpha_1+\beta_1 x}\right)-e^{\alpha_1+\beta_1 x}\left(2e^{\alpha_2}+e^{\alpha_1-\beta_1 x}\right)\right\}}{\left(2+e^{\alpha_1+\beta_1 x}\right)^2\left(2e^{\alpha_2}+e^{\alpha_1-\beta_1 x}\right)^2}$$

Equating to zero yields

$$x = \frac{-\alpha_2}{2\beta_1} = x_0.$$

By applying standard calculus technique one can show that x_0 is a global minimum or maximum of $Corr\,(S_1, S_2)$.

■

The conditional probability that $S_1 = 1$ given that $S_2 = 1$ is derived below.

Property 5.8

$$P\left(S_1 = 1 \mid S_2 = 1\right) = \frac{2}{2 + e^{\alpha_1 - \alpha_2 - \beta_1 x}}$$

Proof.

$$\begin{aligned} P\left(S_1 = 1 \mid S_2 = 1\right) &= \frac{P\left(S_1 = 1, S_2 = 1\right)}{P\left(S_2 = 1\right)} \\ &= \frac{\frac{e^{\eta_2}}{1 + e^{\eta_1} + e^{\eta_2}}}{\frac{2e^{\eta_2} + e^{\eta_1}}{2\left(1 + e^{\eta_1} + e^{\eta_2}\right)}} \\ &= \frac{2}{2 + e^{\alpha_1 - \alpha_2 - \beta_1 x}} \end{aligned}$$

■

If β_1 is positive the conditional probability that $S_1 = 1$ given that $S_2 = 1$ tends to one when x tends to infinity and to zero when x tends to minus infinity. Figure 5.3 presents the conditional probability, $P\left(S_1 = 1 \mid S_2 = 1\right)$, for the same parameter values as in Figure 5.1 and Figure 5.2.

Assuming that β_1 is positive, $P(S = 2)$ is an increasing function in x and consequently $P\left(S_1 = 1 \mid S_2 = 1\right)$ increases with x. This property is illustrated in Figure 5.3. The conditional probability when $\theta = \theta_4$ is close to one for so small values on x as -30. This is explained by a strong dependence between S_1 and S_2. When $\theta = \theta_1$ and $\theta = \theta_2$, β_1 is larger compared to when $\theta = \theta_3$ and $\theta = \theta_4$. Therefore the conditional probability increases more rapidly for θ_1 and θ_2.

5.2 Likelihood and Fisher Information

Since the distribution of the response variable for the simplified Cox model belongs to the exponential family, it follows immediately that also the distribution of the response variable under the symmetric model does so. The likelihood function is basically the same likelihood function as for the simplified Cox model,

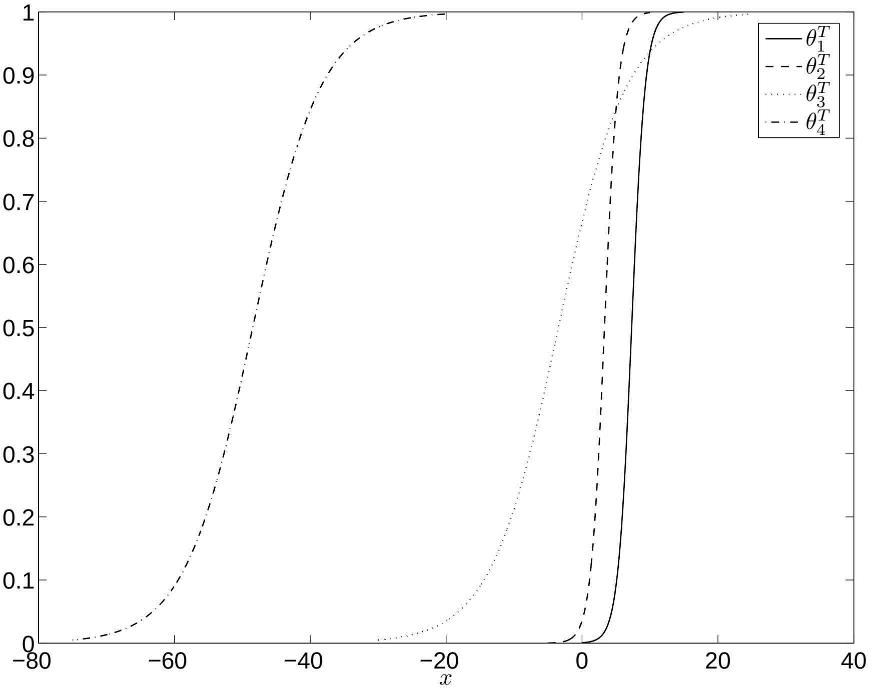


Figure 5.3: $\mathrm{P}(S_1 = 1 \mid S_2 = 1)$ for different sets of parameter values. The parameters are $\theta_1^T = (-2, -10, 1), \theta_2^T = (-1, -5, 1), \theta_3^T = (-1, -1, 0.2)$, and $\theta_4^T = (-10, -1, 0.2)$, respectively. The corresponding log-odds ratios are $\ln\Omega_1 = -4.61$, $\ln\Omega_2 = -1.61$, $\ln\Omega_3 = 2.39$, and $\ln\Omega_4 = 20.39$, respectively.

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{N} \left\{ \pi_{1i}^{y_{1i}} \pi_{2i}^{y_{2i}} \left(1 - \pi_{1i} - \pi_{2i}\right)^{(1-y_{1i})(1-y_{2i})} \right\}.$$

The loglikelihood function is

$$l(\theta; \mathbf{y}) = \sum_{i=1}^{N} \left(y_{1i}\eta_{1i} + y_{2i}\eta_{2i} - \ln\left(1 + e^{\eta_{1i}} + e^{\eta_{2i}}\right) \right).$$

Also the score function is similar to the score function under the simplified Cox model. The only difference is the matrix $\mathbf{x}$.

$$\begin{pmatrix} u_{\alpha_1\cdot}(\theta) \\ u_{\alpha_2\cdot}(\theta) \\ u_{\beta_1\cdot}(\theta) \end{pmatrix} = \sum_{i=1}^{N} \begin{pmatrix} (y_{1i} - \pi_{1i}) \\ (y_{2i} - \pi_{2i}) \\ x_i\left(y_{1i} - \pi_{1i} + 2y_{2i} - \pi_{2i}\right) \end{pmatrix} = \sum_{i=1}^{N} \mathbf{x}_i^T \left(y_i - \mu_i\right)$$

The Fisher information of a single observation is partly the same compared with the simplified Cox model.

$$I(\theta, x) = \mathbf{x}^T D \mathbf{x} =$$

$$\begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & x(\pi_1(1-\pi_1) - 2\pi_1\pi_2) \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & x(2\pi_2(1-\pi_2) - \pi_1\pi_2) \\ x(\pi_1(1-\pi_1) - 2\pi_1\pi_2) & x(2\pi_2(1-\pi_2) - \pi_1\pi_2) & x^2(\pi_1(1-\pi_1) - 4\pi_1\pi_2 + 4\pi_2(1-\pi_2)) \end{pmatrix} \tag{5.2}$$

5.3 Locally D-optimal Designs

The example of a D-optimal design in Section 4.2.2 had no symmetry properties in the design. For instance, no design weights were equal. For the symmetric model in this chapter, the number of design points and the design weights change with different parameter values. Nevertheless, D-optimal designs have some symmetry properties under this model.

Example 5.1

Four D-optimal designs are presented in Table 5.1.

#	α_1	α_2	β_1	$\ln\Omega$	x_0	D-optimal design
1	-2	-10	1	-4.61	5	$\left\{\begin{matrix} 1.176 & 3.680 & 6.320 & 8.824 \\ 0.269 & 0.231 & 0.231 & 0.269 \end{matrix}\right\}$
2	-1	-5	1	-1.61	2.5	$\left\{\begin{matrix} 0.132 & 2.5 & 4.868 \\ 0.255 & 0.49 & 0.255 \end{matrix}\right\}$
3	-1	-1	0.2	2.39	2.5	$\left\{\begin{matrix} -1.487 & 6.487 \\ 0.5 & 0.5 \end{matrix}\right\}$
4	-10	-1	0.2	20.39	2.5	$\left\{\begin{matrix} -0.889 & 5.889 \\ 0.5 & 0.5 \end{matrix}\right\}$

Table 5.1: D-optimal design for different sets of parameter values. The parameter values are the same as in Figure 5.1, Figure 5.2, and Figure 5.3.

Although the number of design points are different for the designs, there are some general properties. The design points are placed symmetrically around x_0. In a plot of $d(x, \xi^*)$, the function should have maximum points at the design points. $d\left(x,\xi^*\right)$ at these maximum points should also be equal to 3 (the number of parameters in the model). These properties are illustrated in Figure 5.4 where it is also possible to see x_0 as a local minimum or maximum point.

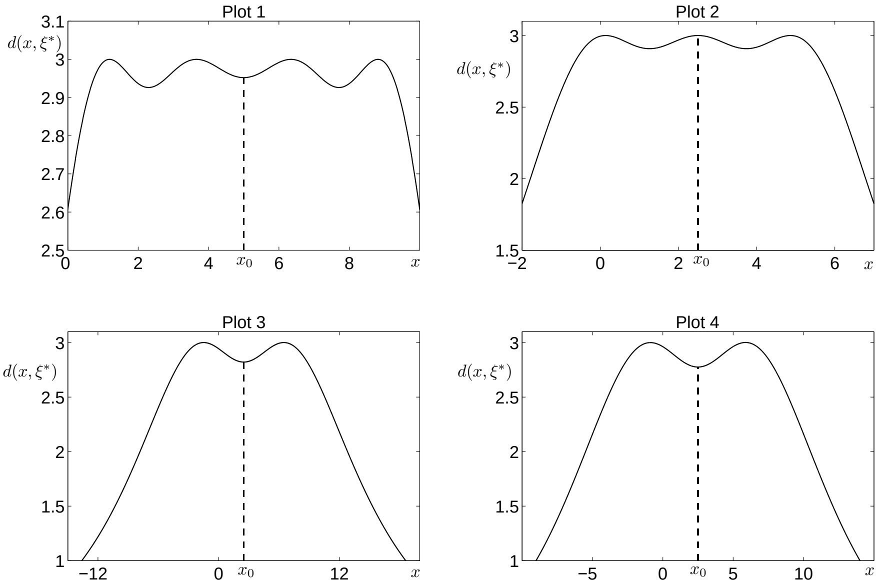


Figure 5.4: $d(x,\xi^*)$ for the different sets of parameter values described in Table 5.1.

The results of the example are in line with Fan (1999) and Fan and Chaloner (2004). They also found optimal designs with two, three or four design points although they had a different model. Fan (1999) argues that the differences in the number of design points and design weights can be explained by certain differences and ratios between parameters. The differences in the number of design points can also be explained using $\ln\Omega$. Several plots have shown that the value of $\ln\Omega$ determines the number of design points in the D-optimal design. For example, when $\ln\Omega$ is large it is sufficient with two design points. In other words it is sufficient to gather information about the model using only two design points. When $\ln\Omega$ decreases, $\pi_1\left(x_0\right)$ increases and a design with two design points is no longer optimal. The design points giving the most information about π_0 do not give any information about π_2 and the other way around. Therefore the optimal design contains three or more design points.

In order to derive some more results on D-optimal designs the following two properties of the determinant of a matrix are needed.

Property 5.9 *If the square matrix B is formed from the square matrix A by multiplying all of the elements of one row or one column of A by the same scalar k, then*

$$|B| = k\,|A|\,.$$

Proof. The reader is referred to Harville (1997) for a proof. ■

Property 5.10 *Let B be the matrix formed from the matrix A by adding, to any row or column of A, scalar multiples of one or more of the other rows or columns. Then,*

$$|B| = |A|\,.$$

Proof. The reader is referred to Harville (1997) for a proof. ■

Define $\theta_g^T = \begin{pmatrix} g & 0 & 1 \end{pmatrix}$ where $g = \alpha_1 - \frac{1}{2}\alpha_2$.

Lemma 5.1 *Let the design*

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_n \\ w_1 & w_2 & \dots & w_n \end{array} \right\}$$

and the design

$$\xi_{\alpha_2,\beta_1} = \left\{ \begin{array}{cccc} \frac{x_1-\frac{\alpha_2}{2}}{\beta_1} & \frac{x_2-\frac{\alpha_2}{2}}{\beta_1} & \dots & \frac{x_n-\frac{\alpha_2}{2}}{\beta_1} \\ w_1 & w_2 & \dots & w_n \end{array} \right\},$$

$\beta_1 \neq 0$. *Then* $\left|M\left(\theta, \xi_{\alpha_2,\beta_1}\right)\right| = \frac{1}{\beta_1^2}\left|M\left(\theta_g, \xi\right)\right|.$

Proof. From (5.2) the information matrix for the design ξ and the parameter vector θ is

$$M\left(\theta,\xi\right) = \begin{pmatrix} M_a\left(\theta,\xi\right) & M_{ab}\left(\theta,\xi\right) & M_{ac}\left(\theta,\xi\right) \\ M_{ab}\left(\theta,\xi\right) & M_b\left(\theta,\xi\right) & M_{bc}\left(\theta,\xi\right) \\ M_{ac}\left(\theta,\xi\right) & M_{bc}\left(\theta,\xi\right) & M_c\left(\theta,\xi\right) \end{pmatrix}$$

where

$$\begin{aligned}
M_a(\theta,\xi) &= \sum_{i=1}^{n} w_i \pi_{1i}(1-\pi_{1i}) \\
M_{ab}(\theta,\xi) &= -\sum_{i=1}^{n} w_i \pi_{1i}\pi_{2i} \\
M_{ac}(\theta,\xi) &= \sum_{i=1}^{n} w_i x_i \left\{\pi_{1i}(1-\pi_{1i}) - 2\pi_{1i}\pi_{2i}\right\} \\
M_b(\theta,\xi) &= \sum_{i=1}^{n} w_i \pi_{2i}(1-\pi_{2i}) \\
M_{bc}(\theta,\xi) &= \sum_{i=1}^{n} w_i x_i \left\{2\pi_{2i}(1-\pi_{2i}) - \pi_{1i}\pi_{2i}\right\} \\
M_c(\theta,\xi) &= \sum_{i=1}^{n} w_i x_i^2 \left\{\pi_{1i}(1-\pi_{1i}) - 4\pi_{1i}\pi_{2i} + 4\pi_{2i}(1-\pi_{2i})\right\}.
\end{aligned}$$

Consider now the transformation $\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}$, $i=1,\ldots,n$, of the design points of ξ. Under this transformation, the expressions for $\pi_{1i}(1-\pi_{1i})$, $\pi_{1i}\pi_{2i}$, and $\pi_{2i}(1-\pi_{2i})$ are found to be

$$\begin{aligned}
\pi_{1i}(1-\pi_{1i}) &= \frac{e^{\alpha_1+\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\left(1+e^{\alpha_2+2\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\right)}{\left(1+e^{\alpha_1+\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}+e^{\alpha_2+2\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\right)^2} \\
&= \frac{e^{x_i+g}\left(1+e^{2x_i}\right)}{\left(1+e^{x_i+g}+e^{2x_i}\right)^2},
\end{aligned}$$

$$\begin{aligned}
\pi_{1i}\pi_{2i} &= \frac{e^{\alpha_1+\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}e^{\alpha_2+2\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}}{\left(1+e^{\alpha_1+\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}+e^{\alpha_2+2\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\right)^2} \\
&= \frac{e^{x_i+g}e^{2x_i}}{\left(1+e^{x_i+g}+e^{2x_i}\right)^2},
\end{aligned}$$

$$\begin{aligned}
\pi_{2i}\left(1-\pi_{2i}\right) &= \frac{e^{\alpha_2+2\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\left(1+e^{\alpha_1+\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\right)}{\left(1+e^{\alpha_1+\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}+e^{\alpha_2+2\beta_1\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}}\right)^2} \\
&= \frac{e^{2x_i}\left(1+e^{x_i+g}\right)}{\left(1+e^{x_i+g}+e^{2x_i}\right)^2},
\end{aligned}$$

respectively. In order to obtain a more compact expression for $M\left(\theta,\xi_{\alpha_2,\beta_1}\right)$, let

$$a^*\left(\theta,x_i\right) = \frac{e^{x_i+g}\left(1+e^{2x_i}\right)}{\left(1+e^{x_i+g}+e^{2x_i}\right)^2} \tag{5.3a}$$

$$b^*\left(\theta,x_i\right) = -\frac{e^{x_i+g}e^{2x_i}}{\left(1+e^{x_i+g}+e^{2x_i}\right)^2} \tag{5.3b}$$

$$c^*\left(\theta,x_i\right) = \frac{e^{2x_i}\left(1+e^{x_i+g}\right)}{\left(1+e^{x_i+g}+e^{2x_i}\right)^2}. \tag{5.3c}$$

All elements in $M\left(\theta,\xi_{\alpha_2,\beta_1}\right)$ are then expressed as combinations of (5.3a), (5.3b), and (5.3c).

$$M_a\left(\theta,\xi_{\alpha_2,\beta_1}\right) = \sum_{i=1}^{n} w_i\pi_{1i}\left(1-\pi_{1i}\right) = \sum_{i=1}^{n} w_i a^*\left(\theta,x_i\right)$$

$$M_{ab}\left(\theta,\xi_{\alpha_2,\beta_1}\right) = -\sum_{i=1}^{n} w_i\pi_{1i}\pi_{2i} = \sum_{i=1}^{n} w_i b^*\left(\theta,x_i\right)$$

$$M_b\left(\theta,\xi_{\alpha_2,\beta_1}\right) = \sum_{i=1}^{n} w_i\pi_{2i}\left(1-\pi_{2i}\right) = \sum_{i=1}^{n} w_i c^*\left(\theta,x_i\right)$$

$$\begin{aligned}
& M_{ac}\left(\theta,\xi_{\alpha_2,\beta_1}\right) \\
&= \sum_{i=1}^{n} w_i\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}\left\{\pi_{1i}\left(1-\pi_{1i}\right)-2\pi_{1i}\pi_{2i}\right\} \\
&= \sum_{i=1}^{n}\frac{w_i}{\beta_1}\left[x_i\left\{a^*\left(\theta,x_i\right)+2b^*\left(\theta,x_i\right)\right\}-\frac{\alpha_2}{2}\left\{a^*\left(\theta,x_i\right)+2b^*\left(\theta,x_i\right)\right\}\right]
\end{aligned}$$

$$\begin{aligned}
&M_{bc}\left(\theta,\xi_{\alpha_2,\beta_1}\right)\\
=&\sum_{i=1}^{n} w_i \frac{x_i-\frac{\alpha_2}{2}}{\beta_1}\left\{2\pi_{2i}\left(1-\pi_{2i}\right)-\pi_{1i}\pi_{2i}\right\}\\
=&\sum_{i=1}^{n} \frac{w_i}{\beta_1}\left[x_i\left\{2c^*\left(\theta,x_i\right)+b^*\left(\theta,x_i\right)\right\}-\frac{\alpha_2}{2}\left\{2c^*\left(\theta,x_i\right)+b^*\left(\theta,x_i\right)\right\}\right]
\end{aligned}$$

$$\begin{aligned}
&M_{c}\left(\theta,\xi_{\alpha_2,\beta_1}\right)\\
=&\sum_{i=1}^{n} w_i \left(\frac{x_i-\frac{\alpha_2}{2}}{\beta_1}\right)^2\left\{\pi_{1i}\left(1-\pi_{1i}\right)-4\pi_{1i}\pi_{2i}+4\pi_{2i}\left(1-\pi_{2i}\right)\right\}\\
=&\sum_{i=1}^{n} \frac{w_i}{\beta_1^2}\left\{x_i^2-x_i\alpha_2+\left(\frac{\alpha_2}{2}\right)^2\right\}\left\{a^*\left(\theta,x_i\right)+4b^*\left(\theta,x_i\right)+4c^*\left(\theta,x_i\right)\right\}
\end{aligned}$$

In addition, let

$$\begin{aligned}
a &= \sum_{i=1}^{n} w_i a^*\left(\theta,x_i\right)\\
b &= \sum_{i=1}^{n} w_i b^*\left(\theta,x_i\right)\\
c &= \sum_{i=1}^{n} w_i x_i\left\{a^*\left(\theta,x_i\right)+2b^*\left(\theta,x_i\right)\right\}\\
d &= \sum_{i=1}^{n} w_i c^*\left(\theta,x_i\right)\\
e &= \sum_{i=1}^{n} w_i x_i\left\{2c^*\left(\theta,x_i\right)+b^*\left(\theta,x_i\right)\right\}\\
f &= \sum_{i=1}^{n} w_i x_i^2\left\{a^*\left(\theta,x_i\right)+4b^*\left(\theta,x_i\right)+4c^*\left(\theta,x_i\right)\right\},
\end{aligned}$$

so that the information matrix can be written as

$$M\left(\theta,\xi_{\alpha_2,\beta_1}\right)=\begin{pmatrix} a & b & \frac{c-\frac{\alpha_2}{2}(a+2b)}{\beta_1}\\ b & d & \frac{e-\frac{\alpha_2}{2}(b+2d)}{\beta_1}\\ \frac{c-\frac{\alpha_2}{2}(a+2b)}{\beta_1} & \frac{e-\frac{\alpha_2}{2}(b+2d)}{\beta_1} & \frac{f-\alpha_2(c+2e)+\left(\frac{\alpha_2}{2}\right)^2(a+4b+4d)}{\beta_1^2}\end{pmatrix}.$$

Using Property 5.9, $|M(\theta, \xi_{\alpha_2,\beta_1})|$ can be written as

$$\frac{1}{\beta_1^2}\begin{vmatrix} a & b & c-\frac{\alpha_2}{2}(a+2b) \\ b & d & e-\frac{\alpha_2}{2}(b+2d) \\ c-\frac{\alpha_2}{2}(a+2b) & e-\frac{\alpha_2}{2}(b+2d) & f-\alpha_2(c+2e)+\left(\frac{\alpha_2}{2}\right)^2(a+4b+4d) \end{vmatrix}$$

By applying Property 5.10 to the expression of $|M(\theta, \xi_{\alpha_2,\beta_1})|$ it is found that

$$|M(\theta, \xi_{\alpha_2,\beta_1})| = \frac{1}{\beta_1^2}\begin{vmatrix} a & b & c \\ b & d & e \\ c & e & f \end{vmatrix}.$$

Next assume that the design is ξ and the parameter vector is θ_g. Then it is readily found that the expressions for $\pi_{1i}(1-\pi_{1i})$, $\pi_{1i}\pi_{2i}$, and $\pi_{2i}(1-\pi_{2i})$ are the same as in the case with ξ_{α_2,β_1} and θ. By applying this result to all elements in $M(\theta_g, \xi)$,

$$M(\theta_g, \xi) = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}.$$

Hence, $|M(\theta, \xi_{\alpha_2,\beta_1})| = \frac{1}{\beta_1^2}|M(\theta_g, \xi)|$. ■

Theorem 5.1 *If*

$$\xi^* = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_n \\ w_1 & w_2 & \dots & w_n \end{array} \right\}$$

is locally D-optimal for the parameter vector θ_g then

$$\xi^*_{\alpha_2,\beta_1} = \left\{ \begin{array}{cccc} \frac{x_1-\frac{\alpha_2}{2}}{\beta_1} & \frac{x_2-\frac{\alpha_2}{2}}{\beta_1} & \dots & \frac{x_n-\frac{\alpha_2}{2}}{\beta_1} \\ w_1 & w_2 & \dots & w_n \end{array} \right\}$$

is locally D-optimal for the parameter vector θ.

Proof. The design $\xi^*_{\alpha_2,\beta_1}$ is obtained from the design ξ^* by one-to-one transformation of the design points $x_i \longrightarrow \frac{x_i-\frac{\alpha_2}{2}}{\beta_1}$, $i = 1, \dots, n$. Maximizing $\ln|M(\theta, \xi)|$ is the same as maximizing $|M(\theta, \xi)|$ over all possible designs. From Lemma 5.1 $|M(\theta, \xi_{\alpha_2,\beta_1})| = \frac{1}{\beta_1^2}|M(\theta_g, \xi)|$. Thus, if ξ^* is locally D-optimal for the parameter vector θ_g, it follows that $\xi^*_{\alpha_2,\beta_1}$ is the locally D-optimal design for the parameter vector θ by the one-to-one property of $\xi^* \longrightarrow \xi^*_{\alpha_2,\beta_1}$. ■

By using Lemma 5.1, Theorem 5.1 states that once a locally D-optimal design has been found for the parameter vector θ_g, a locally D-optimal design has also been found for the parameter vector θ.

Theorem 5.2 *Under a bivariate symmetric simplified Cox model, the number of design points in a locally D-optimal design is determined by the log-odds ratio between the variables.*

Proof. Let ξ^* be the locally D-optimal design for the parameter vector θ. Lemma 5.1 showed that for every setup of θ and ξ^*, there is a corresponding setup with θ_g and ξ, where ξ and ξ^* have equally many design points. In addition, the value of θ_g is completely determined by $g = \alpha_1 - \frac{1}{2}\alpha_2$, so that the number of design points in ξ^* is determined by g. From (5.1) it follows that $\ln\Omega$ is equal to $\ln 4 - 2g$, making $\ln\Omega$ a function of g and a constant. Hence, $\ln\Omega$ determines the number of design points in a locally D-optimal design. ■

Note that given a certain $\ln\Omega$, Theorem 5.2 does not provide the optimal number of design points in a locally D-optimal design. However, the relationship between $\ln\Omega$ and the number of design points can be found numerically as illustrated in Figure 5.5. The result has been derived by determining the locally D-optimal design for a number of different values on $\ln\Omega$. The parameter β_1 was held constant during the examination because Theorem 5.2 implies that the value of β_1 does not affect the number of design points. Depending on $\ln\Omega$, a $2-$point, $3-$point, or $4-$point design is then derived. The optimality of the proposed design is checked by calculating max $d(x,\xi)$.

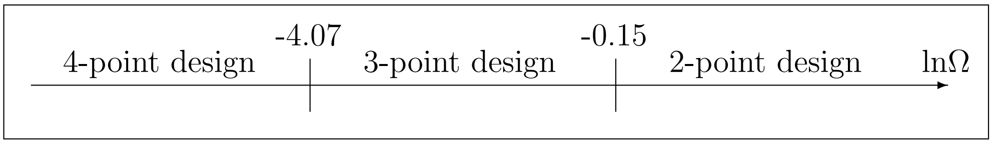


Figure 5.5: Number of design points given the log-odds ratio.

5.3.1 Locally D-optimal Designs when the Log-odds Ratio is Large

For symmetric simplified Cox models with a large $\ln \Omega$ it was shown numerically that a D-optimal design has two design points. In this section, the particular case when $\alpha_2 = 0$ and $\beta_1 = 1$ is examined. All other cases can be obtained by using Theorem 5.1. Given $\alpha_2 = 0$ and $\beta_1 = 1,\ x_0 = 0$. Assume further that the design points are placed symmetrically around x_0 with equal design weights and that the design points are denoted by $-c$ and c. Then the proposed design is given by

$$\xi = \left\{ \begin{array}{cc} -c & c \\ 0.5 & 0.5 \end{array} \right\}.$$

The restriction on α_2 and β_1 simplifies the calculations considerable. The standardized information matrix for ξ is

$$M\left(\alpha_1, c\right) = \frac{1}{2}\left(I(\alpha_1, -c) + I\left(\alpha_1, c\right)\right).$$

The determinant of M is

$$\left|M\left(\alpha_1, c\right)\right| = \frac{c^2 e^{\alpha_1 - 6c}\left(e^{\alpha_1} + e^{\alpha_1 + 2c} + 4e^c\right)}{\left(1 + e^{\alpha_1 - c} + e^{-2c}\right)^5}.$$

Using this expression, the derivative of the determinant of M with respect to c is

$$\frac{d\left|M\left(\alpha_1, c\right)\right|}{dc} = \frac{1}{\left(1 + e^{\alpha_1 - c} + e^{-2c}\right)^6}\{ce^{\alpha_1 - 4c}$$
$$[2\left(1 + e^{\alpha_1 - c} + e^{-2c}\right)\left\{e^{-2c}\left(e^{\alpha_1} + e^{\alpha_1 + 2c} + 4e^c\right) - c\left(3e^{\alpha_1 - 2c} + 2e^{\alpha_1} + 10e^{-c}\right)\right\}$$
$$+ 5ce^{-3c}\left(e^{\alpha_1} + 2e^{-c}\right)\left(e^{\alpha_1} + e^{\alpha_1 + 2c} + 4e^c\right)]\}.$$

Setting

$$\frac{d\left|M\left(\alpha_1, c\right)\right|}{dc} = 0$$

yields,

$$2\left(1 + e^{\alpha_1 - c} + e^{-2c}\right)\left\{e^{-2c}\left(e^{\alpha_1} + e^{\alpha_1 + 2c} + 4e^c\right) - c\left(3e^{\alpha_1 - 2c} + 2e^{\alpha_1} + 10e^{-c}\right)\right\}$$
$$+5ce^{-3c}\left(e^{\alpha_1} + 2e^{-c}\right)\left(e^{\alpha_1} + e^{\alpha_1 + 2c} + 4e^c\right) = 0. \qquad (5.4)$$

Based on the numerical solution of (5.4), Figure 5.6 shows the value of c as a function of $\ln\Omega$. For the same c and the same log-odds ratio

$$\max_{x\in\mathfrak{X}} d\left(x,\xi\left(c\right)\right)$$

is plotted. $\max d\left(x,\xi\left(c\right)\right)$ is used since it is a direct way to verify if a design is D-optimal or not. In the plot over c note that when $\ln\Omega$ becomes smaller c increases. When $\ln\Omega$ is less than -0.15 a $2-$point design is no longer optimal. That is why max $d\left(x,\xi\left(c\right)\right)$ is larger than three as soon as $\ln\Omega$ is less than -0.15. This result is in line with the result shown in Figure 5.5.

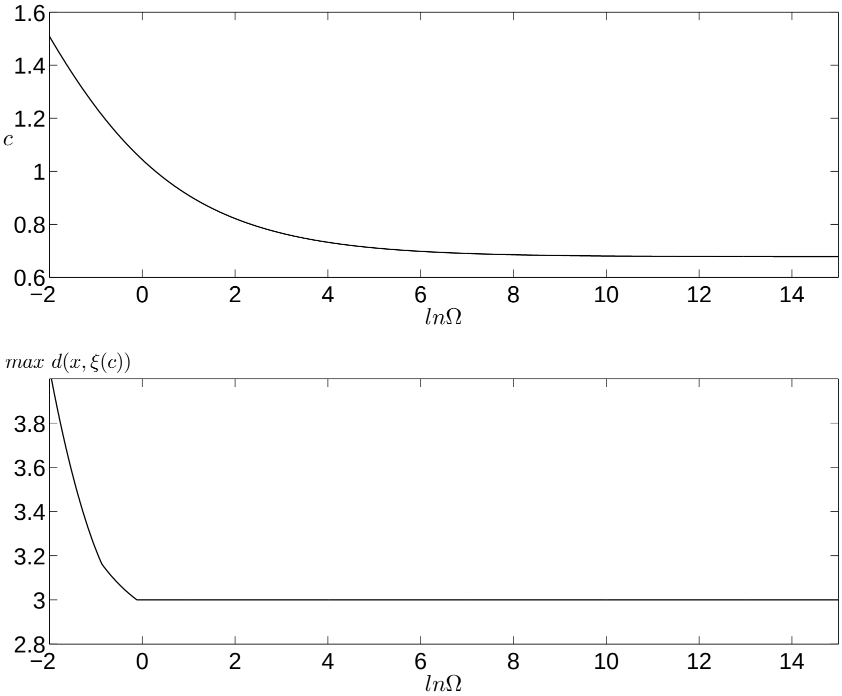


Figure 5.6: The value of c that maximizes $\mid M(\alpha_1,\xi)\mid$ for different values of $ln\Omega$, (upper plot). Maximum of $d(x,\xi(c))$ as a function of both c and $ln\Omega$, (lower plot).

Consider again the equation in (5.4). If $\alpha_2=0$ then

$$\ln\Omega\to\infty \ \text{ when } \ \alpha_1\to-\infty.$$

Let $\alpha_1 \rightarrow -\infty$ and it follows that

$$\begin{aligned} c &= \frac{2\left(1+e^{-2c}\right)}{5\left(1-e^{-2c}\right)}. \\ c &\approx 0.6778 \end{aligned}$$

In Figure 5.6 c is constant for a $\ln \Omega$ of approximately 10 and larger. This value of c is around 0.6778. One of the examples shown previously had a $\ln \Omega$ of around 20. It is possible to verify that c is around 0.6778 in that example[1].

The proposed design with $c = 0.6778$ can be evaluated using the D-efficiency, given in e.g. Atkinson and Donev (1992). The D-efficiency is defined as

$$D_{eff} = \left(\frac{\left|M\left(\theta, \xi\left(c\right)\right)\right|}{\left|M\left(\theta, \xi^{*}\right)\right|}\right)^{\frac{1}{p}},$$

where p is the number of parameters in the model. Figure 5.7 presents the D-efficiency for the design with $c = 0.6778$ given different parameter values (different $\ln \Omega$). The D-efficiency was derived by first choosing a vector of values for α_1. Since $\alpha_2 = 0$ this vector corresponds to a vector of values for $\ln \Omega$. For the proposed design with $c = 0.6778$ and for each value of $\ln \Omega$ the determinant of the information matrix, $\left|M\left(\theta, \xi\left(c\right)\right)\right|$, is derived. Given the same parameter vector the locally D-optimal design, ξ^{*}, is found numerically. Finally the D-efficiency is calculated according to the expression above.

For parameter values with $\ln \Omega$ larger than five the proposed design is very close to being optimal. When $\ln \Omega = 0$ the D-efficiency is around 0.9136. This means that the design is quite efficient even when the variables are independent and another model is preferable. When $\ln \Omega$ is less than -0.15 a $2-$point design is no longer optimal and the D-efficiency decreases rapidly.

[1] In the example $a_1 = -10$ and $a_2 = -1$. This gives a log-odds ratio, $ln\Omega = ln4 + 19$. The symmetry point $x_0 = \frac{1}{0.4} = 2.5$. The design points are located $5.889 - 2.5 = 3.389$ from x_0. So in this example $c = 3.389\beta_1 = 0.6778$.

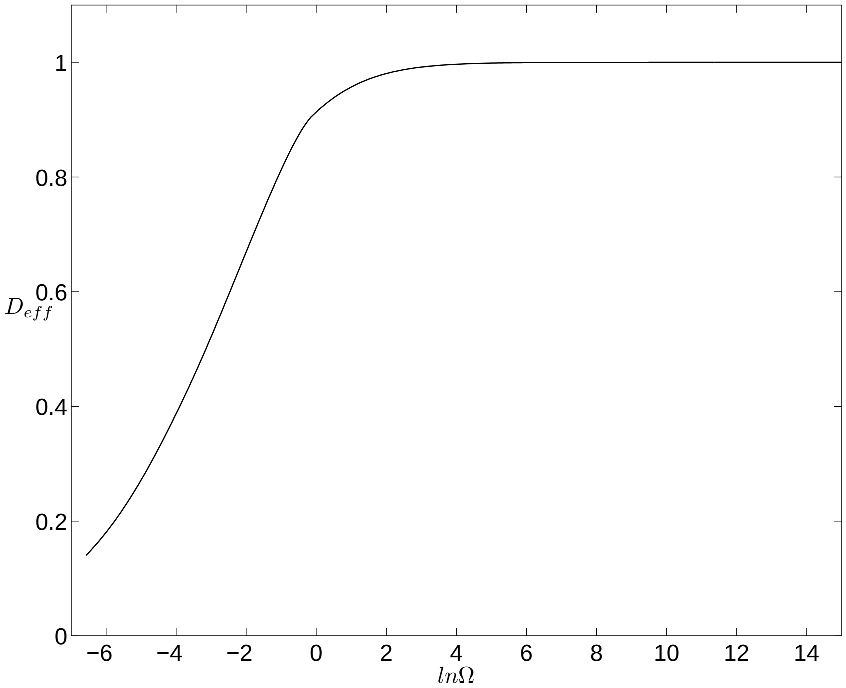


Figure 5.7: D-efficiency for the proposed design with $c = 0.6778$ for different parameter values.

5.3.2 Locally D-optimal Designs when the Log-odds Ratio is Large Negative

For parameter values that yield a large negative $\ln \Omega$, a 4−point design is optimal. The particular case when $\alpha_2 = 0$ and $\beta_1 = 1$ is examined below. As in the last section, all other cases can be obtained by applying Theorem 5.1. A plot of the probabilities π_0, π_1, and π_2 as functions of x for $\theta^T = \begin{pmatrix} 20 & 0 & 1 \end{pmatrix}$ is given in Figure 5.8.

Based on Figure 5.8 it is reasonable to believe that when $\ln \Omega$ is very large negative, the optimal design points are located around two symmetry points. These symmetry points are where π_0 equals π_1 and where π_1 equals π_2. When π_0 equals π_1, $\eta_1 = 0$ and hence

$$\alpha_1 + x = 0 \quad \text{so that} \quad x = -a_1.$$

In the same way, when π_1 equals π_2, $\eta_1 = \eta_2$ and hence

$$\alpha_1 + x = 2x \quad \text{so that} \quad x = a_1.$$

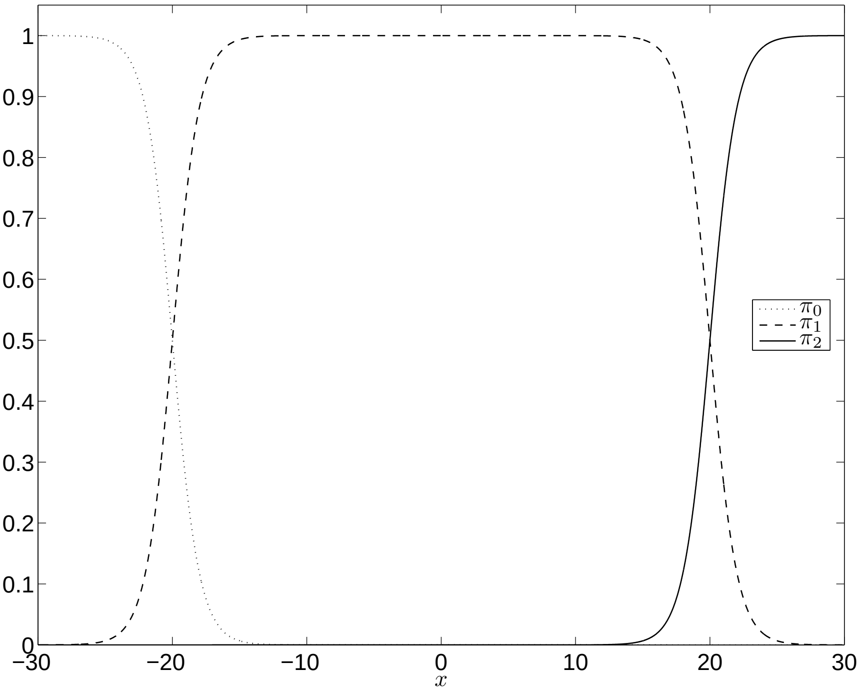


Figure 5.8: An example of the probabilities π_0, π_1, and π_2 as functions of x when the log-odds ratio is large negative. The parameters are $\theta^T = (20, 0, 1)$ and the log-odds ratio is $\ln\Omega = -38.61$.

The proposed 4−point design is then

$$\xi = \left\{ \begin{array}{cccc} -\alpha_1 - c & -\alpha_1 + c & \alpha_1 - c & \alpha_1 + c \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}. \tag{5.5}$$

The standardized information matrix based on ξ is then

$$M\left(\alpha_1, c\right) = \frac{1}{4}\left\{I(-\alpha_1, -c) + I\left(-\alpha_1, c\right) + I(\alpha_1, -c) + I\left(\alpha_1, c\right)\right\}.$$

This 4−point design, ξ, has a more complex expression for the determinant of M compared with the 2−point design in the last section. Therefore, the value of c that maximizes $|M|$ can only be found numerically. In Figure 5.9 the value of c that maximizes $|M\left(\alpha_1, c\right)|$ is plotted against $\ln\Omega$.

When $\ln\Omega$ is less than -10, $c \approx 1.2229$. With similar constraints on the parameters and for a similar situation as in Figure 5.8, Fan (1999)

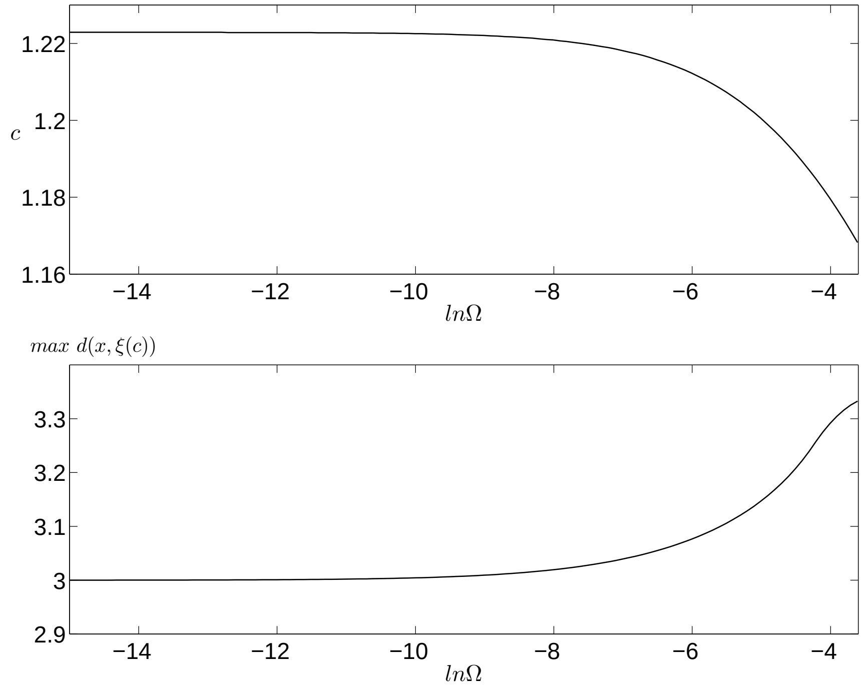


Figure 5.9: The value of c that maximizes $\mid M(\alpha_1, \xi) \mid$ for different values of $ln\Omega$, (upper plot). Maximum of $d(x, \xi(c))$ as a function of both c and $ln\Omega$, (lower plot).

derived a limiting locally optimal design. The limiting locally D-optimal design,

$$\xi_{\text{lim}} = \left\{ \begin{array}{cccc} -1.223 & +1.223 & \tau - 1.223 & \tau + 1.223 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\},$$

is similar to the 4-point design in (5.5). Both designs are 4-points designs with equal design weights. Moreover, a constant, $c \approx 1.223$ partly determines the allocation of the design points. Here, τ is a function of the parameters in the continuation-ratio logit model.

The proposed design in (5.5) with $c = 1.2229$ is an approximation of the locally D-optimal design. The approximation is evaluated by calculating the D-efficiency for different parameter values in the same manner as in the last section. The D-efficiency for different parameter values are shown in Figure 5.10.

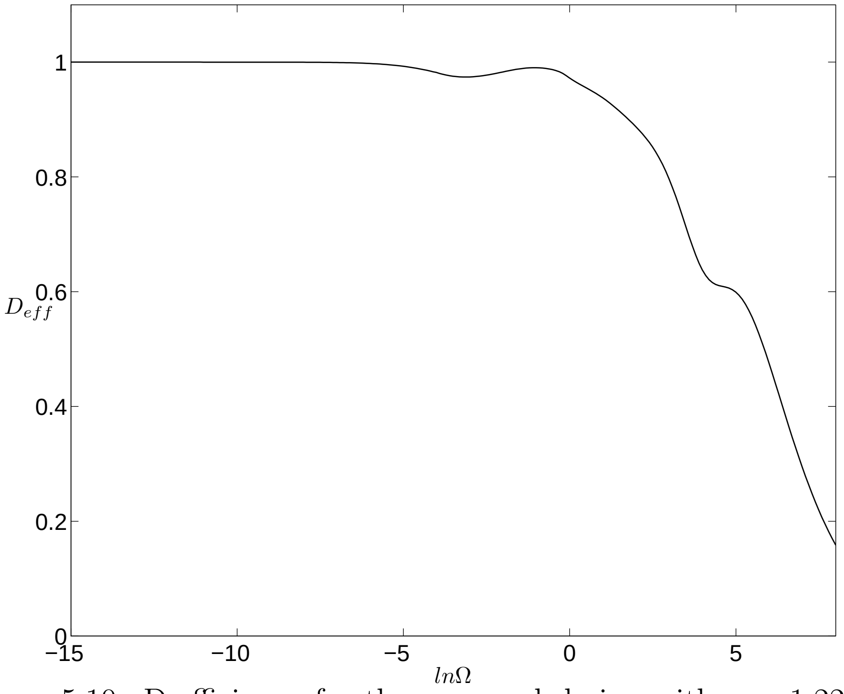


Figure 5.10: D-efficiency for the proposed design with $c = 1.229$ given different parameter values.

If $\ln \Omega$ is approximately -5 or smaller $c = 1.2229$ yields a very efficient design. A $3-$point design is D-optimal when $\ln \Omega$ is between -4.07 and -0.15. Nevertheless, the D-efficiency for the proposed $4-$point design is still very large when $\ln \Omega$ is between -4.07 and -0.15. The explanation is that two design points are almost equal for the proposed $4-$point design. Hence, the proposed design has almost the same D-efficiency as a D-optimal $3-$point design. For positive log-odds ratios the D-efficiency decreases fast.

Chapter 6

Mutual Independence

6.1 The Model

As mentioned in Chapter 1, the simplified Cox model has a much simpler structure when the Bernoulli variables are mutually independent. In fact when $S_1, S_2, \ldots, S_k$ are mutually independent, $S \sim Bin\,(k, \pi.)$ with

$$P\,(S = s) = \binom{k}{s} \pi_.^s \,(1 - \pi.)^{k-s} \qquad s = 0, 1, \ldots, k$$

and

$$\pi. = \frac{e^{\eta_1}}{k + e^{\eta_1}},$$

where $\pi.$ is the common probability for observing a "success". The parameters, α_1 and β_1, are interpreted from the expression for the log-odds of "success",

$$\ln \frac{\pi.}{1 - \pi.} = \eta_1 - \ln k = \alpha_1 - \ln k + \beta_1 x.$$

Under an ordinary binary logistic regression model for independent variables, the distribution for S is $S \sim Bin\,(k, \pi.)$ with

$$\pi. = \frac{e^{\eta}}{1 + e^{\eta}}.$$

The log-odds of "success" is then equal to η,

$$\ln \frac{\pi.}{1 - \pi.} = \eta = \alpha + \beta x.$$

Hence, the parameterization of the simplified Cox model under mutual independence is not the same as for the ordinary logistic model. In particular, α in the ordinary logistic model corresponds to $\alpha_1 - \ln k$ in the simplified Cox model.

6.2 Parameter Restrictions under Mutual Independence

Theorem 6.1 states necessary and sufficient conditions on the parameters to yield mutual independence. The theorem is used in Chapter 7, where a model under these restrictions is tested against a model without restrictions. Theorem 6.1 is presented in Bruce (2008).

Theorem 6.1 *$S_1, S_2, \ldots, S_k$ are mutually independent if and only if*

$$\eta_s = s\left(\eta_1 - \ln k\right) + \ln \binom{k}{s} \qquad s = 0, 1, \ldots, k. \tag{6.1}$$

Proof. The proof is divided into two parts where the first part shows that the parameter restrictions in (6.1) implies that $S_1, S_2, \ldots, S_k$ are mutually independent. The second part shows that mutual independence implies the parameter restrictions in (6.1).

Under a simplified Cox model π_s can be expressed as in (3.1). Assume now that the parameters are determined by (6.1) so that

$$\begin{aligned}
\eta_0 &= 0 \\
\eta_1 &= \eta_1 = \alpha_1 + \beta_1 x \\
\eta_2 &= 2\left(\alpha_1 + \beta_1 x - \ln k\right) + \ln \tbinom{k}{2} \\
&\vdots \\
\eta_k &= k\left(\alpha_1 + \beta_1 x - \ln k\right).
\end{aligned}$$

Then

$$e^{\eta_s} = e^{s(\eta_1 - \ln k) + \ln \binom{k}{s}} = \binom{k}{s}\left(\frac{e^{\eta_1}}{k}\right)^s \qquad s = 0, 1, \ldots, k$$

and

$$\sum_{s=0}^{k} e^{\eta_s} = \sum_{s=0}^{k} \binom{k}{s} \left(\frac{e^{\eta_1}}{k}\right)^s = \left(1 + \frac{e^{\eta_1}}{k}\right)^k$$

which follows from the Binomial Theorem. This yields

$$\pi_s = \frac{\binom{k}{s}\left(\frac{e^{\eta_1}}{k}\right)^s}{\left(1+\frac{e^{\eta_1}}{k}\right)^k} \qquad s = 0, 1, \ldots, k. \tag{6.2}$$

From (6.2) $\pi.$ and $(1 - \pi.)$ can be identified,

$$\begin{aligned} \pi_s &= \binom{k}{s} \frac{\left(e^{\eta_1}\right)^s}{\left(k + e^{\eta_1}\right)^k} k^{k-s} \\ &= \binom{k}{s} \underbrace{\left(\frac{e^{\eta_1}}{k + e^{\eta_1}}\right)^s}_{=\pi.} \underbrace{\left(\frac{k}{k + e^{\eta_1}}\right)^{k-s}}_{=(1-\pi.)} \\ &= \tbinom{k}{s} \pi.^s \left(1 - \pi.\right)^{k-s} \qquad s = 0, 1, \ldots, k. \end{aligned}$$

Hence, the sum $S = S_1 + S_2 + \ldots + S_k$ has a binomial distribution $Bin\,(k, \pi.)$ with $\pi. = \frac{e^{\eta_1}}{k+e^{\eta_1}}$, which is possible only if $S_1, S_2, \ldots, S_k$ are mutually independent.

Next assume that $S_1, S_2, \ldots, S_k$ are mutually independent. The parameter restrictions in (6.1) follow immediately for $s = 0, 1$. From (3.3) it follows that

$$\ln \Omega_s = \eta_s - 2\eta_{s-1} + \eta_{s-2} - \ln \tbinom{k}{s} + 2 \ln \tbinom{k}{s-1} - \ln \tbinom{k}{s-2}.$$

Since mutual independence implies that $\ln \Omega_s = 0$,

$$\eta_s = 2\eta_{s-1} - \eta_{s-2} + \ln \tbinom{k}{s} - 2 \ln \tbinom{k}{s-1} + \ln \tbinom{k}{s-2}. \tag{6.3}$$

Assume now that the restrictions (6.1) are true for $s - 1$ and s, so that

$$\eta_{s-1} = (s - 1)\left(\eta_1 - \ln k\right) + \ln \tbinom{k}{s-1}$$

and

$$\eta_s = s\left(\eta_1 - \ln k\right) + \ln \tbinom{k}{s}.$$

From (6.3) η_{s+1} is then given by

$$\begin{aligned}\eta_{s+1} &= 2\left\{s\left(\eta_1-\ln k\right)+\ln\tbinom{k}{s}\right\}-\left\{(s-1)\left(\eta_1-\ln k\right)+\ln\tbinom{k}{s-1}\right\}\\ &\quad +\ln\tbinom{k}{s+1}-2\ln\tbinom{k}{s}+\ln\tbinom{k}{s-1}\\ &= \eta_1\left(2s-s+1\right)-\ln k\left(2s-s+1\right)+\ln\tbinom{k}{s+1}\\ &= (s+1)\left(\eta_1-\ln k\right)+\ln\tbinom{k}{s+1},\end{aligned}$$

i.e. the restriction (6.1) is true also for $s+1$.
Furthermore, since (6.3) is true for all values of s, i.e. $s=2,3,\ldots,k$, the parameter restrictions in (6.1) follow by induction. ■

In Section 3.5 an extension to polytomous data was presented. Let $S_1,S_2,\ldots,S_k$ have three response categories so that Y_j is the number of outcomes in the respective response category, $j=0,1,2$. When $S_1,S_2,\ldots,S_k$ are mutually independent, (Y_1,Y_2) are multinomial distributed with

$$\begin{aligned}P\left(Y_1=y_1,Y_2=y_2\right) &= \begin{pmatrix} & k & \\ y_1 & & y_2\end{pmatrix}\pi_{1\cdot}^{y_1}\pi_{2\cdot}^{y_2}\left(1-\pi_{1\cdot}-\pi_{2\cdot}\right)^{(k-y_1-y_2)},\\ & \quad y_1,y_2\geq 0 \text{ and } y_1+y_2\leq k.\end{aligned}$$

As for the case with binary data, parameter restrictions for mutual independence are expressed analytically by the following theorem.

Theorem 6.2 *Let $S_1,S_2,\ldots,S_k$ be k identically distributed random variables each having three response categories. $S_1,S_2,\ldots,S_k$ are mutually independent if and only if*

$$\begin{aligned}\eta_{y_1y_2} &= y_1\left(\eta_{10}-\ln k\right)+y_2\left(\eta_{01}-\ln k\right)+\ln\begin{pmatrix} & k & \\ y_1 & & y_2\end{pmatrix}, \qquad (6.4)\\ & \quad y_1,y_2\geq 0 \ \textit{and}\ y_1+y_2\leq k.\end{aligned}$$

The proof is given in Appendix A.

For the case with arbitrary many response categories, corresponding parameter restrictions are obtained analogously.

6.3 D-optimal Designs

For the simplified Cox model, it is difficult to analytically derive locally D-optimal designs in general. This is usually due to a complex expression for the standardized information matrix. Typically the determinant of the standardized information matrix is a function, which can only be maximized numerically. In Section 5.3, expressions for locally D-optimal designs for a bivariate symmetric model were derived. However, this was done for a simplified model (symmetric) and for only two dependent variables. Still, assumptions had to be made to obtain any results at all.

It was shown above that under mutual independence, the distribution of the response variable is greatly simplified. In addition, the assumption about mutual independence yields a less complex expression for the standardized information matrix. Thus, it is possible to formulate a theorem for locally D-optimal designs under these restrictions.

Theorem 6.3 *Let $S_1, S_2, \ldots, S_k$ be mutually independent. The locally D-optimal design is then*

$$\xi^* = \left\{ \begin{array}{cc} \frac{-\alpha_k}{\beta_k} - \frac{c}{\beta_1} & \frac{-\alpha_k}{\beta_k} + \frac{c}{\beta_1} \\ 0.5 & 0.5 \end{array} \right\},$$

where c is the solution to the equation

$$c = \frac{e^c + 1}{e^c - 1}.$$

$$c \approx 1.5434$$

Proof. By applying the parameter restrictions for mutual independence given in Theorem 6.1 the design can be rewritten as

$$\xi^* = \left\{ \begin{array}{cc} \frac{\ln k - \alpha_1 - c}{\beta_1} & \frac{\ln k - \alpha_1 + c}{\beta_1} \\ 0.5 & 0.5 \end{array} \right\}. \tag{6.5}$$

Furthermore, let $\alpha_1 - \ln k$ correspond to α and let β_1 correspond to β in the ordinary logistic model defined previously. For the ordinary logistic model it is an established result that the locally D-optimal design is

$$\xi = \left\{ \begin{array}{cc} \frac{-\alpha - c}{\beta} & \frac{-\alpha + c}{\beta} \\ 0.5 & 0.5 \end{array} \right\},$$

see e.g. Atkinson et al. (2007). Thus, it follows directly that the proposed design, ξ^*, is the locally D-optimal design under mutual independence.

■

The difference between the design ξ^* and the D-optimal design under the ordinary logistic model, is explained by different parameterization of the models as shown above. However, using corresponding parameterizations both designs allocate observations at the same levels of $\pi.$,

$$\pi.\left(\frac{-\alpha - c}{\beta}\right) = \pi.\left(\frac{\ln k - \alpha_1 - c}{\beta_1}\right) = 0.1760$$

and

$$\pi.\left(\frac{-\alpha + c}{\beta}\right) = \pi.\left(\frac{\ln k - \alpha_1 + c}{\beta_1}\right) = 0.8240.$$

6.4 Paired Data

When there are only two variables, S_1 and S_2, the properties of the model can be further explored. Assume that S_1 and S_2 are independent for all x. The parameter restrictions under independence then follow from Theorem 6.1. These restrictions can also be derived by equating the log-odds ratio between S_1 and S_2 to zero, yielding

- $\alpha_2 = 2\alpha_1 - \ln 4$
- $\beta_2 = 2\beta_1$.

Because S_1 and S_2 are independent the distribution for S is simply

$$S \sim Bin\left(2, \pi.\right).$$

In this model S is the response variable, and using the restrictions for the model

$$\pi. = \frac{e^{\eta_1}}{2 + e^{\eta_1}}.$$

The variance of S is

$$Var(S) = 2\pi.\left(1 - \pi.\right).$$

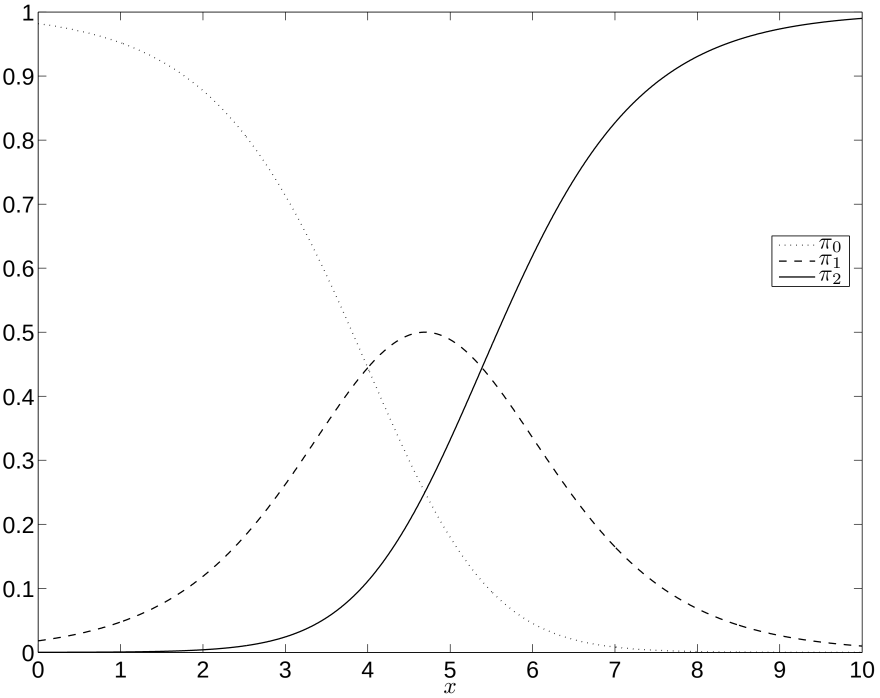


Figure 6.1: The probabilities π_0, π_1, and π_2 as functions of x in a model for independent pairs of responses and with parameters $\alpha_1 = -4$ and $\beta_1 = 1$.

Figure 6.1 presents an example of the probabilities π_0, π_1, and π_2 as a function of x.

In the remaining part of this chapter some further properties of the model for independent pairs of response variables are derived. As for the bivariate symmetric model in Chapter 5, $x_0 = \arg\max_{x \in \mathfrak{X}} \pi_1$ is important when deriving symmetry properties for this model.

Property 6.1

$$x_0 = \frac{\ln 2 - \alpha_1}{\beta_1}$$

Proof. Differentiate $\pi_1(x)$ with respect to x.

$$\frac{d\pi_1(x)}{dx} = \frac{8\beta_1 e^{\eta_1} - 4\beta_1 e^{2\eta_1}}{(2 + e^{\eta_1})^3}$$

Equating to zero yields

$$2 = e^{\eta_1}$$

and hence

$$x_0 = \frac{\ln 2 - \alpha_1}{\beta_1}.$$

By applying standard calculus technique it is easy to verify that x_0 is a global maximum of $\pi_1(x)$. ■

Property 6.2

$$\pi_1(x_0) = \frac{1}{2}$$

Proof.

$$\pi_1(x_0) = \frac{4e^{\eta_1}}{(2 + e^{\eta_1})^2} = \frac{1}{2}$$

■

Hence, in this model the maximum value of π_1 is always equal to $\frac{1}{2}$. This result is in line with the fact that $\pi_1 \leq \frac{1}{2}$ when $S \sim Bin(2, \pi.)$ regardless of π.

Analogously to the bivariate symmetric model in Chapter 5, π_0 and π_2 are symmetrical around x_0 in the sense that $\pi_0(x_0 - d) = \pi_2(x_0 + d)$.

Property 6.3

$$\pi_0(x_0 - d) = \pi_2(x_0 + d) \quad \textit{for all } d$$

Proof.

$$\pi_0(x_0 - d) = \frac{4}{(2 + e^{\alpha_1 + \beta_1(x_0 - d)})^2} = \frac{1}{(1 + e^{-d\beta_1})^2}$$

$$\pi_2(x_0 + d) = \frac{e^{2(\alpha_1 + \beta_1(x_0 + d))}}{(2 + e^{\alpha_1 + \beta_1(x_0 + d)})^2} = \frac{1}{(1 + e^{-d\beta_1})^2}$$

Hence $\pi_0(x_0 - d) = \pi_2(x_0 + d)$. ■

Since S has a binomial distribution, the distribution of S is an exponential family. The likelihood is well determined

$$L(\alpha_1, \beta_1; \mathbf{s}) = \prod_{i=1}^{N} \binom{2}{s_i} \pi_i^{s_i} (1 - \pi_i)^{2 - s_i},$$

$$\ln L\left(\alpha_1, \beta_1; \mathbf{s}\right) = \sum_{i=1}^{N} \left\{ \ln \binom{2}{s_i} + s_i \ln \left(\frac{\pi_i}{\left(1-\pi_i\right)} \right) + 2 \ln \left(1-\pi_i\right) \right\}.$$

The score function is determined by applying the chain rule,

$$\begin{pmatrix} u_{\alpha_1}.\left(\theta\right) \\ u_{\beta_1}.\left(\theta\right) \end{pmatrix} = \sum_{i=1}^{N} \begin{pmatrix} \left(s_i - 2\pi_i\right) \\ x_i\left(s_i - 2\pi_i\right) \end{pmatrix}.$$

The Fisher information is

$$\begin{aligned} I.\left(\theta, x\right) &= E\left[u\left(\theta\right) u^T\left(\theta\right)\right] \\ &= \sum_{i=1}^{N} 2\pi_i\left(1-\pi_i\right) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}. \end{aligned}$$

Locally D-optimal designs for two independent Bernoulli variables are easily found, see Kalish and Rosenberger (1978) for an early reference.

By using Theorem 6.3, a locally D-optimal design is given by

$$\xi^* = \left\{ \begin{array}{cc} x_0 - \frac{c}{\beta_1} & x_0 + \frac{c}{\beta_1} \\ 0.5 & 0.5 \end{array} \right\},$$

where $c \approx 1.5434$.

Example 6.1

For the parameters $\alpha_1 = -4$ and $\beta_1 = 1$, a D-optimal design is given by

$$\xi^* = \left\{ \begin{array}{cc} 3.1497 & 6.2365 \\ 0.5 & 0.5 \end{array} \right\}.$$

In Figure 6.2, $d(x, \xi^*)$ is plotted. Note that $d(x, \xi^*)$ has two maximum points appearing at the two design points 3.1497 and 6.2365. If ξ^* is a D-optimal design the criterion function ψ has minimum value for this design. However, ξ^* is also optimal if the determinant of the standardized information matrix, $\left|M\left(\theta, \xi^*\right)\right|$, has maximum points at the design points. Figure 6.2 illustrates that this is true for the current example. In Figure 6.2, $\left|M\left(\theta, \xi\right)\right|$ for the design

$$\xi = \left\{ \begin{array}{cc} \ln 2 + 4 - c & \ln 2 + 4 + c \\ 0.5 & 0.5 \end{array} \right\}$$

is given. The maximum value of $|M(\theta, \xi)|$ is attained for $c \approx 1.5434$. Moreover, since the value of $|M(\theta, \xi)|$ is maximized at the design points of a D-optimal design, the D-optimal design points are approximately 3.150 and 6.236.

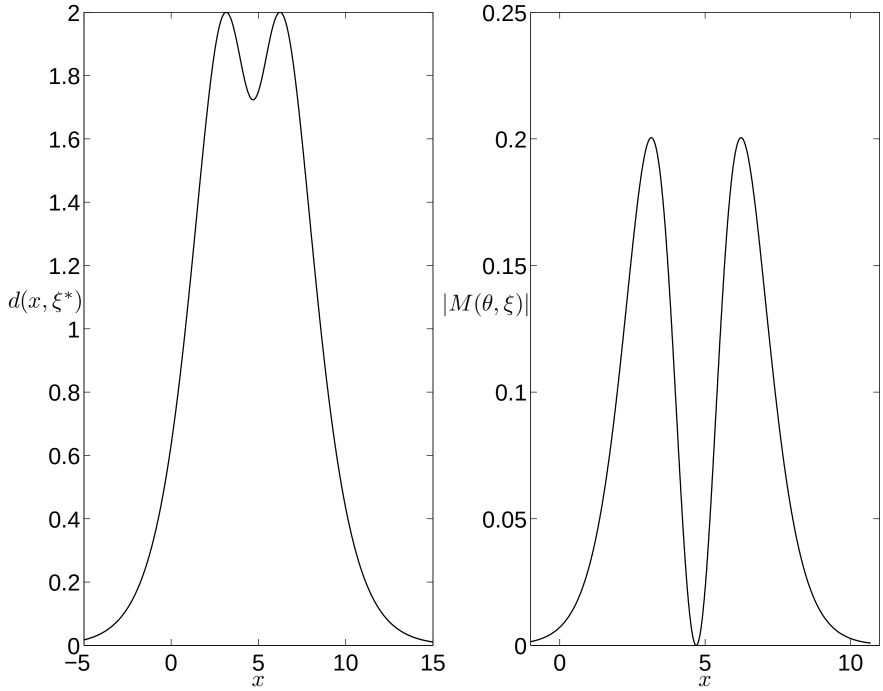


Figure 6.2: $d(x, \xi^*)$ for a model of independent pairs of responses and with parameters $\alpha_1 = -4$ and $\beta_1 = 1$. Determinant of the information matrix for different choices of design points given the same parameters.

Chapter 7

Testing for Independence

The previous chapter showed that under the restrictions for mutual independence, the model collapses to a simple form with only two parameters to estimate. Locally D-optimal designs are also easier to derive when the variables are mutually independent. Test procedures for testing the null hypothesis of independent variables are therefore of interest. By applying the parameter restrictions given in Theorem 6.1, the mutual independence model can be tested against a more general model. This chapter considers the score test and the likelihood ratio test for testing independence. If there is enough support for the independence model, the model can be estimated and evaluated using standard statistical software. Tests of independence for exchangeable binary data have been treated in e.g. George and Kodell (1996) and Kang and Park (2000). George and Kodell (1996) worked with the model presented in Chapter 2, see also George and Bowman (1995a,b). They also derived a test for heterogeneity between treatment groups and a test for a dose-related trend. Using the same model, Kang and Park (2000) derived an exact test for independence particularly useful in small samples. This chapter is based on Bruce and Nyquist (2007).

7.1 Models without Covariates

Assume a simplified Cox model with only two variables, S_1 and S_2. If no covariate is included in the model, the loglikelihood as a function of

π_0, π_1, and π_2 is

$$\ell\left(\pi;\mathbf{y}\right) = \sum_{i=1}^{N} y_{0i}\ln\pi_0 + y_{1i}\ln\pi_1 + y_{2i}\ln\pi_2, \tag{7.1}$$

where $\mathbf{y}$ is a matrix with the responses from N observations on $y_0,\ y_1$, and y_2. Hence,

$$\begin{aligned} y_{ji} \quad &= \quad \begin{cases} 1, \text{ if } S = S_1 + S_2 = j \text{ for observation } i. \\ 0, \text{ otherwise} \end{cases} \quad, \\ &j = 0,1,2 \text{ and } i = 1,2,\ldots,N. \end{aligned}$$

The test statistic for the score test is defined as

$$T_S = u^T_{.}\left(\widetilde{\pi}\right) I_{.}^{-1}\left(\widetilde{\pi}\right) u_{.}\left(\widetilde{\pi}\right), \tag{7.2}$$

where $\widetilde{\pi} = \left(\widetilde{\pi}_1, \widetilde{\pi}_2\right)^T$ is the estimated vector of probabilities under the null hypothesis. The hypothesis of independence implies the restrictions $\pi_0 = \left(1-\pi_{.}\right)^2$, $\pi_1 = 2\pi_{.}\left(1-\pi_{1.}\right)$, and $\pi_2 = \pi_{.}^2$, where $\pi_{.} = P\left(S_1 = 1\right) = P\left(S_2 = 1\right)$ is the marginal probability to observe a "success". Under the hypothesis of independence, the maximum likelihood estimator of $\pi_{.}$ is evidently

$$\widetilde{\pi}_{.} = \frac{1}{2N}\sum_{i=1}^{N}\left(y_{1i} + 2y_{2i}\right) = \frac{r_1 + 2r_2}{2N}, \tag{7.3}$$

where r_j is the number of observed pairs that result in $S = j,\ j = 0,1,2$. Hence, the estimator $\widetilde{\pi}_{.}$ equals the total number of observed "successes" divided by the number of observed variables. Maximum likelihood estimators of π_0, π_1, and π_2 are accordingly

$$\widetilde{\pi}_0 = \left(1-\widetilde{\pi}_{.}\right)^2, \widetilde{\pi}_1 = 2\widetilde{\pi}_{.}\left(1-\widetilde{\pi}_{.}\right), \text{ and } \widetilde{\pi}_2 = \widetilde{\pi}_{.}^2. \tag{7.4}$$

By deriving expressions for the scores and the information matrix from (7.1) and inserting these expressions in (7.2), the score test statistic becomes

$$T_S = \sum_{j=0}^{2}\frac{\left(r_j - N\widetilde{\pi}_j\right)^2}{N\widetilde{\pi}_j}. \tag{7.5}$$

The test statistic coincides with the χ^2-test statistic for testing the goodness of fit of a trinomial distribution with probabilities restricted as described above. Asymptotically as N tends to infinity, T_S has a χ^2 distribution with 1 degree of freedom, the approximation being good provided the expected frequencies, $N\widetilde{\pi}_j$, $j = 0, 1, 2$, are sufficiently large.

The test statistic for the likelihood ratio test is defined as

$$\begin{aligned} T_{LR} &= 2\left\{\ell\left(\widehat{\pi}; \mathbf{y}\right) - \ell\left(\widetilde{\pi}; \mathbf{y}\right)\right\} \\ &= 2\sum_{j=0}^{2} r_j \ln \frac{\widehat{\pi}_j}{\widetilde{\pi}_j}, \end{aligned}$$

where $\widehat{\pi}_j = \frac{r_j}{N}$ is the unrestricted maximum likelihood estimator of π_j, $j = 0, 1, 2$. This likelihood ratio test statistic is equivalent to the one derived in George and Kodell (1996).

This simple case generalizes straightforwardly to the case with several, say K, groups with N_k observations in each group. The distribution of the trinomial response vector in each group is here defined by the vector $\left(\pi_{0k}, \pi_{1k}, \pi_{2k}\right)^T$, $k = 1, 2, \ldots, K$, of probabilities. The test statistic for the score test now becomes

$$T_S = \sum_{k=1}^{K} \sum_{j=0}^{2} \frac{\left(r_{jk} - N_k\widetilde{\pi}_{jk}\right)^2}{N_k\widetilde{\pi}_{jk}}, \tag{7.6}$$

where r_{jk} is the observed frequency of category j, $j = 0, 1, 2$, in group k, $k = 1, 2, \ldots, K$,

$$\widetilde{\pi}_{0k} = \left(1 - \widetilde{\pi}_{\cdot k}\right)^2, \widetilde{\pi}_{1k} = 2\widetilde{\pi}_{\cdot k}\left(1 - \widetilde{\pi}_{\cdot k}\right), \widetilde{\pi}_{2k} = \widetilde{\pi}_{\cdot k}^2, \tag{7.7}$$

and

$$\widetilde{\pi}_{\cdot k} = \frac{r_{1k} + 2r_{2k}}{2N_k}. \tag{7.8}$$

Similarly, the test statistic for the likelihood ratio test becomes

$$T_{LR} = 2\sum_{k=1}^{K} \sum_{j=0}^{2} r_{jk} \ln \frac{\widehat{\pi}_{jk}}{\widetilde{\pi}_{jk}}. \tag{7.9}$$

where the unrestricted estimator is

$$\widehat{\pi}_{jk} = \frac{r_{jk}}{N_k}.$$

The test statistics T_S and T_{LR} are asymptotically equivalent and has a χ^2 distribution with K degrees of freedom, asymptotically. This asymptotic result relies on a fixed K and that $\min N_k$ tends to infinity, Fahrmeir and Tutz (2001). Here it is important that each $N_k\widetilde{\pi}_{jk}$ is sufficiently large for the approximation to be good.

7.2 Models with Covariates

A more structured model is obtained if the vector of probabilities π is governed by a vector of explanatory variables x. Assume that a simplified bivariate Cox model as described in Chapter 3 is used. The vector valued linear predictor, $\eta = (\eta_1, \eta_2)^T$ is then

$$\eta_j = \mathbf{x}_j^T \theta_j, \;\; j = 1, 2,$$

where $\mathbf{x}_j$ and θ_j are vectors of explanatory variables and associated parameters used for determining the probability π_j. Denoting the maximum likelihood estimator of the parameter vector θ under H_0 by $\widetilde{\theta}$, the score test statistic becomes

$$T_S = u_{\cdot}^T \left(\widetilde{\theta}\right) I_{\cdot}^{-1} \left(\widetilde{\theta}\right) u_{\cdot} \left(\widetilde{\theta}\right). \tag{7.10}$$

The test statistic can be calculated using the previously derived expressions for the score vector and the information matrix.

Denote the maximum likelihood estimator without restrictions by $\widehat{\theta}$. The likelihood ratio test statistic is then obtained as the difference of the loglikelihood function evaluated at $\widetilde{\theta}$ and at $\widehat{\theta}$:

$$T_{LR} = 2 \left\{ l\left(\widehat{\theta}; \mathbf{y}\right) - l\left(\widetilde{\theta}; \mathbf{y}\right) \right\}. \tag{7.11}$$

The expression for the likelihood under H_0 and the expression for the unrestricted likelihood have both been given above. The score test statistic and the likelihood ratio test statistic have the same χ^2 distribution asymptotically.

Suppose that data exist for K independent groups with N_k observations in each group. Then the two vectors of explanatory variables are identical and consist of dummy variables $\mathbf{x}_1 = \mathbf{x}_2 = (d_1, d_2, \ldots, d_K)^T$, where each d_k is either 1 or 0, indicating if an observation comes from response group k or not, $k = 1, 2, \ldots, K$. In this case the model reduces to the case with K response groups discussed above and the test statistics for independence are (7.6) and (7.9).

Example 7.1

The data for this example are taken from Liang et al. (1992). 5199 people are subject to a visual examination, measuring if the left eye and/or the right eye has a visual impairment or not. The outcome for each eye is binary, where " $+$ " indicates visual impairment and " $-$ " no visual impairment. Age is used as explanatory variable, see Table 7.1. In Table 7.1 there are, e.g. 3627 out of 3958 people in age $40-70$ that have no visual impairment.

Left	Right	Age: $40-70$	Age: 71+	Total
$-$	$-$	3627	913	4540
$+$	$-$	122	89	211
$-$	$+$	133	104	237
$+$	$+$	76	135	211
	Total	3958	1241	5199

Table 7.1: Joint distribution of visual impairment for both eyes, for the two age groups $40-70$ and over 70, respectively. Data are taken from Liang et al. (1992).

The probability that the left eye is visually impaired is assumed to be equal to the probability that the right eye is visually impaired. This assumption is reasonable since the risk of visual impairment (in percent) is similar for the left and the right eye in both age groups.

Let S_1 and S_2 be Bernoulli variables for visual impairment of the left eye and the right eye, respectively. The elements of the response vector $y_i = (y_{1i}, y_{2i})^T$, $i = 1, 2, \ldots, 5199$, are the corresponding indicator variables.

The vector of explanatory variables consists of the dummy variables d_1 and d_2 denoting the two age groups. The link function is therefore

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \beta_{11} d_1 + \beta_{12} d_2 \\ \beta_{21} d_1 + \beta_{22} d_2 \end{pmatrix}.$$

Suppose now that primary interest is in the possible dependence between S_1 and S_2. In this setup, S_1 and S_2 are independent if the parameter restrictions

$$\beta_{2j} = 2\beta_{1j},\ j = 1, 2$$

are satisfied. As stated previously the score test statistic is given by (7.6). The test statistic has a χ^2 distribution with 2 degrees of freedom, asymptotically. The observed test statistic for the data material in Table 7.1 becomes, using (7.6),

$$T_S \approx 751.22.$$

Hence the hypothesis of independence is rejected since the critical value on the 5% level is 5.991. The observed likelihood ratio test statistic is derived from (7.9). Since the value on the test statistic is

$$T_{LR} \approx 465.35,$$

the hypothesis about independence is rejected when using the likelihood ratio test as well.

Another model is used when the linear predictors consist of an intercept and a single explanatory variable, x, the same variable in both linear predictors, so that $\eta_j = \mathbf{x}^T \theta_j$, $\mathbf{x} = (1, x)^T$ and $\theta_j = (\alpha_j, \beta_j)^T$, $j = 1, 2$. In this model, the explanatory variable x may influence the "success" probabilities π_1 and π_2 differently. Here S_1 and S_2 are independent if $\beta_2 = 2\beta_1$ and $\alpha_2 = 2\alpha_1 - \ln 4$. Asymptotically, both T_S and T_{LR} are χ^2-distributed with 2 degrees of freedom. Values on the test statistics are now computed for two artificially created data materials.

Example 7.2

Data consists of 100 pairs of Bernoulli variables. Each pair is associated with a single covariate, x, ranging between zero and ten, see Figure 7.1.

Ignoring the covariate x, the observed frequencies for $S = 0, 1, 2$ are $49, 17$, and 34, respectively. By only looking at the observed frequencies, S_1 and S_2 seem to be dependent. The goodness of fit test given in (7.5) confirms this. The observed test statistic when the covariate is ignored is

$$\chi^2_{obs} \approx 42.54.$$

Clearly, the conclusion based only on this test would be that S_1 and S_2 are dependent. However, it is not sufficient to look at observed marginal frequencies only. When testing for independence, one has to study how the probabilities $\widehat{\pi}_0(x), \widehat{\pi}_1(x)$, and $\widehat{\pi}_2(x)$ change when taking account of the covariate, x. The relative low frequency of pairs where $S = 1$ is explained by the fact that many observations are taken at values of x where $\widehat{\pi}_1(x)$ is small.

The score test statistic and the likelihood ratio test statistic, given in (7.10) and (7.11), take covariates into account in the test procedures. The observed test statistics for the two tests become

$$T_S \approx 0.0340$$

and

$$T_{LR} \approx 0.0338,$$

respectively. Because the critical value at the 5% level is 5.991, the hypothesis of independence can not be rejected in either of the tests.

Another good indicator of the possible dependence between S_1 and S_2 is the estimated probabilities $\widehat{\pi}_0$, $\widehat{\pi}_1$, and $\widehat{\pi}_2$, given in Figure 7.1. The probabilities $\widehat{\pi}_0$, $\widehat{\pi}_1$, and $\widehat{\pi}_2$ as function of x, closely resembles the appearance of a distribution for independent Bernoulli variables, see Section 6.4. There are several characteristically properties in a model for independent data. Two of these properties are clearly shown in Figure 7.1. First the maximum value of $\widehat{\pi}_1$ is close to 0.5, and secondly $\widehat{\pi}_1$ is a symmetric function around $\arg\max_x \widehat{\pi}_1$. This example emphasizes the importance of including existing covariates in the analysis.

Example 7.3

The data in the third example have the same structure as the data in Example 7.2. Data consist of 100 pairs of Bernoulli variables, where each

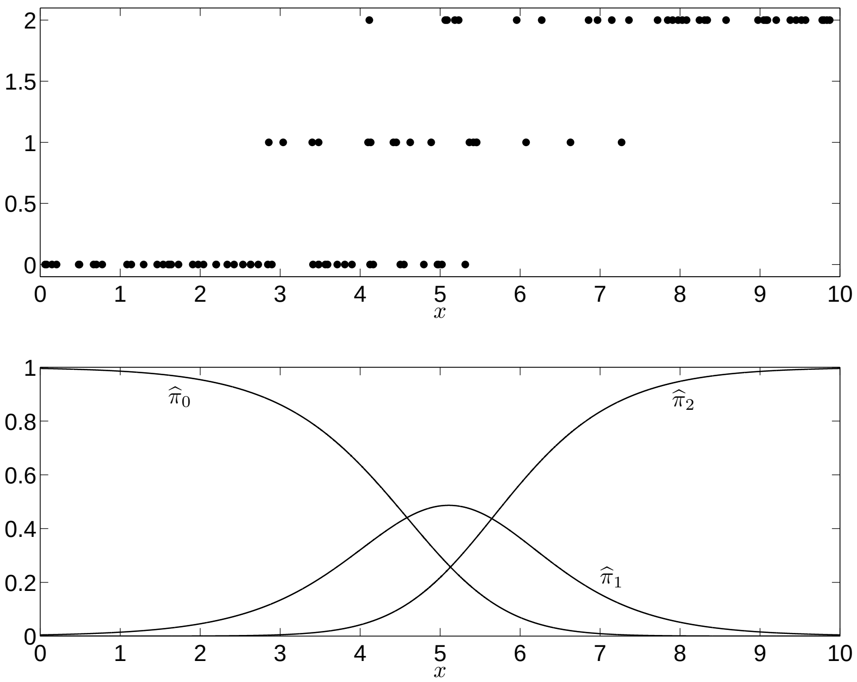


Figure 7.1: The 100 observations on S for Example 7.2, (upper plot). The estimated probabilities $\widehat{\pi}_0, \widehat{\pi}_1$, and $\widehat{\pi}_2$ as functions of x, (lower plot). The parameter values used are those obtained from the bivariate logit estimator $\widehat{\theta}$.

pair is associated with a single covariate. Thus, the same model can be fitted to this data material as to the previous data material. Figure 7.2 shows that the data in Example 7.3 resemble the data in Example 7.2. Nevertheless, the observed score test statistic and the observed likelihood ratio test statistic are given by

$$T_S \approx 6.3066$$

and

$$T_{LR} \approx 7.7791,$$

respectively. The hypothesis of independence is rejected in both tests because the observed test statistics exceed the critical value at the 5% level. Figure 7.2 presents the probability distribution of S based on the estimator of the unrestricted likelihood, $\widehat{\theta}$.

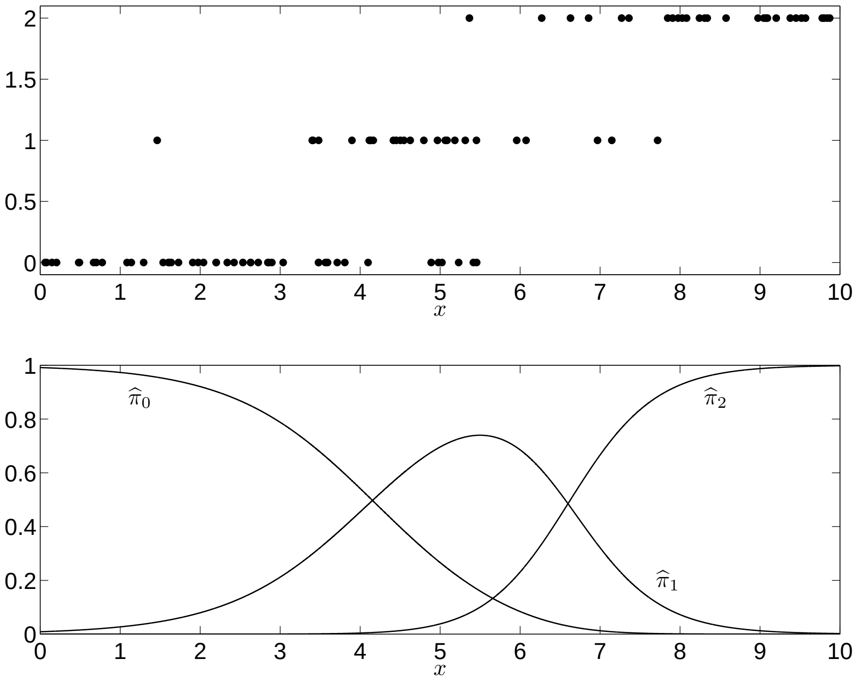


Figure 7.2: The 100 observations on S for Example 7.3 (upper plot). The lower plot shows the estimated probabilities $\widehat{\pi}_0, \widehat{\pi}_1$, and $\widehat{\pi}_2$ as functions of x.

The probability distribution does not share the properties of a probability distribution generated by independent Bernoulli variables. In particular, the maximum value of $\widehat{\pi}_1$ is relatively far from 0.5 and $\widehat{\pi}_1$ is not a symmetric function around $\arg\max_x \widehat{\pi}_1$.

Chapter 8

Optimal Design in a Test of Independence

In the previous chapter, examples showed that a covariate, which is important in explaining the probability distribution of S, should be included in a test for independence between S_1 and S_2. The examples illustrate how the response probabilities, π_0, π_1, and π_2, in some cases clearly depend on the covariate. In an experimental study the values of the covariate can be controlled. Therefore it is of interest to find values of the covariate so that properties of the test are optimized. In particular, different sets of values of the covariate generate different power of the test. In this framework, a favorable power function can be generated if the values of the covariate, i.e. the design, are chosen in an appropriate way. This chapter considers the problem of finding optimal designs such that the local asymptotic power of the score test in (7.6) is maximized.

8.1 Optimal Design

For determining an approximation to the power of the test at an alternative hypothesis close to H_0, let θ be the true value of the parameter vector and θ_0 the value under H_0. Let further $\delta = \sqrt{N}\left(\theta - \theta_0\right)$ be fixed so that θ converges to θ_0 as N tends to infinity. The first-order expansion of $\frac{1}{\sqrt{N}} u.\left(\theta_0\right)$ around θ is

$$\frac{1}{\sqrt{N}} u.\left(\theta_0\right) = \frac{1}{\sqrt{N}} u.\left(\theta\right) + \frac{1}{N} H \sqrt{N}\left(\theta_0 - \theta\right),$$

where H is the matrix of second-order derivatives of the loglikelihood function. Now, the first term in this expression converges to a normally distributed random variable with zero mean and variance $M(\theta_0)$, and the second term converges to $M(\theta_0)\,\delta$. Hence, the distribution of $\frac{1}{\sqrt{N}}u.(\theta_0)$ is approximately normal with expectation $M(\theta_0)\,\delta$ and variance $M(\theta_0)$, in large samples. This makes that the distribution of the score test statistic, T_S, can be approximated in large samples by a noncentral χ^2 distribution with 2 degrees of freedom and non-centrality parameter

$$\varphi = \delta^T M(\theta_0)\,\delta. \tag{8.1}$$

The asymptotic distribution of test statistics is treated in, e.g. Ferguson (1996). When evaluating the information matrix in φ, the true value θ can be used instead of θ_0 so that

$$\varphi = \delta^T M(\theta)\,\delta. \tag{8.2}$$

As Ferguson points out, (8.1) and (8.2) are asymptotically equal. In this chapter the performance of the score test statistic is examined by determining the power of the test in finite samples. The power of the test is found as the probability that T_S exceeds the critical value T_c given that θ is the true parameter vector. Therefore (8.2) will be used as non-centrality parameter throughout this chapter.

Obviously, the power depends on φ and is smallest in the direction δ in which $\delta^T M(\theta)\,\delta$ is minimized. The smallest possible value of φ is the smallest eigenvalue of $M(\theta)$ and δ is the eigenvector associated to the smallest eigenvalue. If an experiment is to be conducted in order to test H_0, it is reasonable to select a design that makes the power of the test as large as possible. Furthermore, the smallest power is in the direction of the eigenvector associated to the smallest eigenvalue of $M(\theta)$. If no direction is of particular interest, a design that maximizes the smallest eigenvalue of $M(\theta)$ is proposed. This design is recognized as an E-optimal design, see Section 4.2.3 for a description of E-optimal designs.

Unfortunately, the E-optimal design for maximizing the smallest local power depends on the unknown parameter vector, so only a locally optimal design can be determined. As an example, with $\alpha_1 = -2$ and

$\beta_1 = 1$, and accordingly $\alpha_2 = -4 - \ln 4 \approx -5.3863$ and $\beta_2 = 2$, the 3−point design

$$\xi_E = \left\{ \begin{array}{ccc} 0.1741 & 2.2049 & 5.7469 \\ 0.4414 & 0.3706 & 0.1880 \end{array} \right\} \tag{8.3}$$

is obtained. The probabilities π_0, π_1, and π_2 are estimated at each design point as in (7.7) and (7.8). The test statistic (7.6) is then used, treating the design points as different groups.

It should be noted that this design maximizes the smallest power. This means that there may be other deigns that yield a stronger power at some alternatives under H_1 but at some other parameter values under H_1 they yield a smaller power than the E-optimal design. On the other hand, there is no design that dominates the E-optimal design in that it provides a larger asymptotic power than that for the E-optimal design for all directions of the alternative hypothesis.

Section 8.2 and Section 8.3 illustrate how the performance of a locally optimal design can be investigated with respect to small samples and incorrect guesses of the parameter values. The results from the investigation can not be generalized to an arbitrary locally optimal design. It is merely one example on how to examine the performance of a proposed locally optimal design.

8.2 Small Sample Performances based on Simulation

In this section the performance of a locally optimal design in small samples is examined with respect to the power of the score test. Using simulation the power is calculated and then compared against the asymptotic power for the same sample size. Values of the parameters α_1 and β_1 are chosen arbitrary to $\alpha_1 = -2$ and $\beta_1 = 1$. The power of the score test depends on the alternative hypothesis. Values of α_2 and β_2 far away from their restrictions under H_0 yield in general a large power. In order to examine the power for different alternative hypothesis, α_2 and β_2 are varied over intervals of values. The sample sizes are chosen to be 100, 400, and 1000, respectively. Using simulation, a data set is created 5000

times for each value of α_2 and β_2. The score test statistic, based on the simulated data, is then calculated using (7.6), (7.7), and (7.8). The power is determined as the percentage of the score test statistics larger than the critical value T_c. The significance level is taken to be 5% in all studied cases.
The hypotheses when testing for independence are

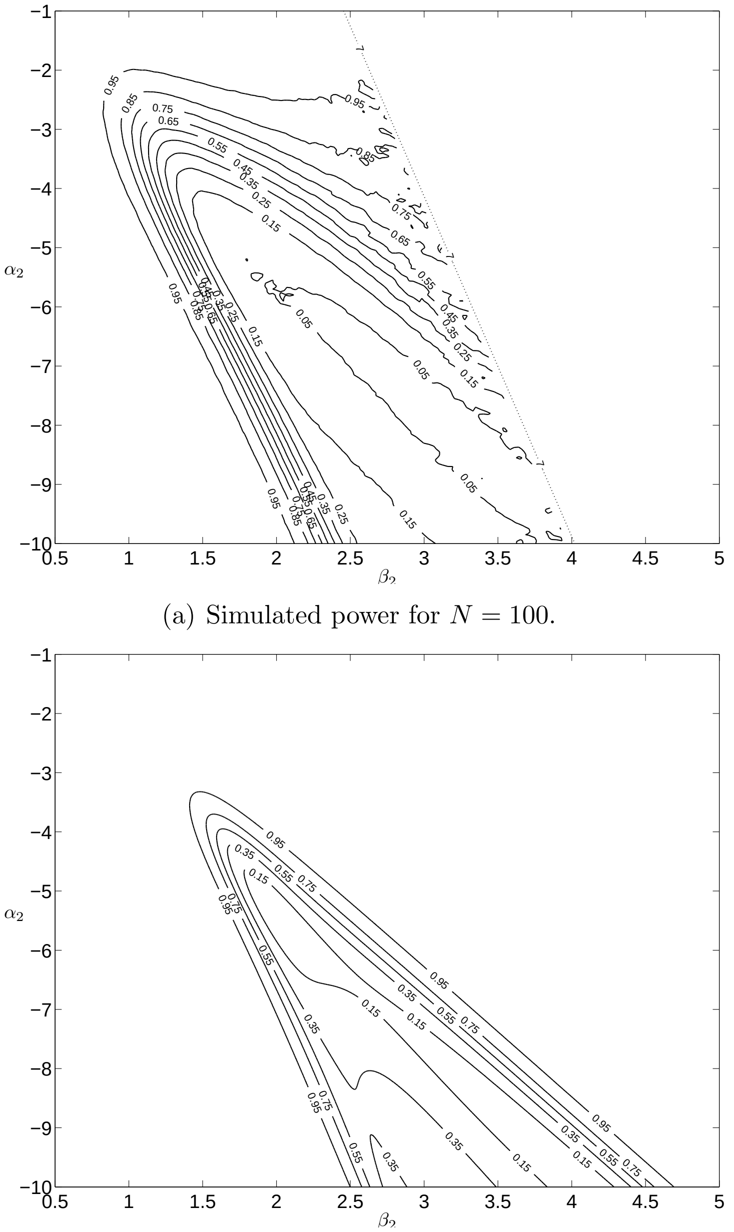
$$\begin{aligned} H_0 &: \begin{cases} \alpha_2 = 2\alpha_1 - \ln 4 \\ \beta_2 = 2\beta_1 \end{cases} \\ H_1 &: \alpha_2 \neq 2\alpha_1 - \ln 4 \quad \text{or} \quad \beta_2 \neq 2\beta_1. \end{aligned} \tag{8.4}$$

Using $(\alpha_1 = -2, \beta_1 = 1)$ a locally E-optimal design is given in (8.3). Note that the design depends only on α_1 and β_1. Contour plots of the power as a function of α_2 and β_2 for different sample sizes are given in Figure 8.1, Figure 8.2, and Figure 8.3. In each figure the simulated power function is compared against the asymptotic power function for the same sample size.

In general, the simulated power is smaller than the asymptotic power at an arbitrary alternative hypothesis. For a given power the contour line for the simulated power lies farther away from H_0 compared to the asymptotic power. The difference is larger for $N = 100$ than for $N = 400$ or $N = 1000$. Besides this the asymptotic power resembles the simulated power fairly well, at least for the larger sample sizes.

Two notes follow immediately from the results obtained, one note concerns the computation of the test statistic at certain parameter values, the other concerns the general shape of the power contours.

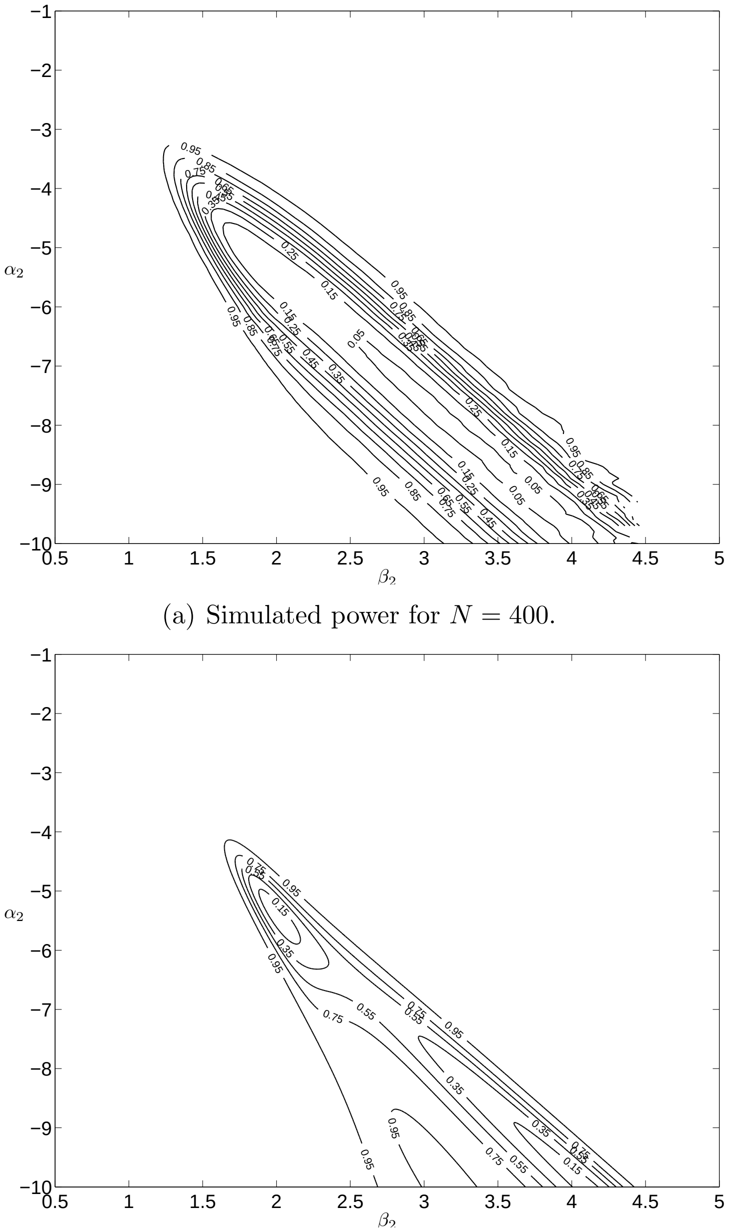
The first note may be explained in terms of the log-odds ratio between S_1 and S_2, $\ln \Omega$. The expression for $\ln \Omega$ was given in (3.2). Let $\ln \Omega$ for the particular value x_i be denoted by $\ln \Omega_{x=x_i}$. Since $\ln \Omega$ also depends on the parameter vector θ, different alternative hypotheses yield different values on $\ln \Omega$. In this particular example, large values on α_2 and β_2 generate a large $\ln \Omega$. When $\ln \Omega$ is large, the estimated marginal probability of observing a "success", $\widetilde{\pi}_{\cdot}$, is often equal to one. As a direct consequence, the score test statistic, T_S in (7.6), can not be computed. This results in computational problems when determining the power. Especially the design point at $x = 5.7469$ generates, for some values on α_2 and β_2, a $\ln \Omega$ that causes these computational problems. Because of these problems,



(a) Simulated power for $N = 100$.

(b) Asymptotic power for $N = 100$.

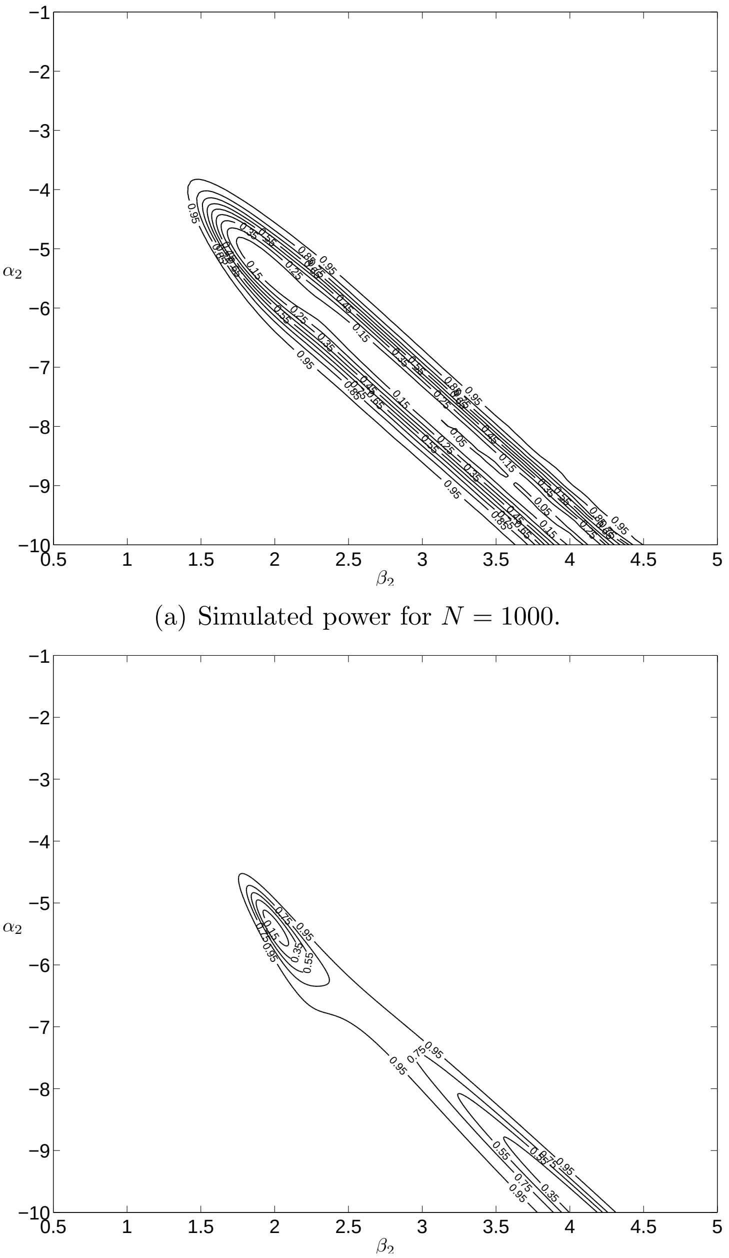
Figure 8.1: Contour plot of the simulated and the asymptotic power as a function of α_2 and β_2 for $N = 100$. The vertical dotted line in (a) is an approximate boundary showing for which (α_2, β_2) the computational problems are extensive.



(a) Simulated power for $N = 400$.

(b) Asymptotic power for $N = 400$.

Figure 8.2: Contour plot of the simulated and the asymptotic power as a function of α_2 and β_2 for $N = 400$.



(a) Simulated power for $N = 1000$.

(b) Asymptotic power for $N = 1000$.

Figure 8.3: Contour plot of the simulated and the asymptotic power as a function of α_2 and β_2 for $N = 1000$.

the test statistic in Figure 8.1(a) is not computed when $\ln\Omega_{x=5.7469}$ is larger than seven. A dotted line in Figure 8.1(a) indicates for which parameter values $\ln\Omega_{x=5.7469}$ is equal to seven. Although computational problems are more extensive in small samples, such as $N = 100$, these problems still appear when $N = 400$. In the lower right corner of the contour plot in Figure 8.2(a), the test statistic can not be computed for some parameter values. Due to the computational problems for some values of α_2 and β_2, the power is not based on 5000 replicates for all (α_2, β_2) in the plots.

Under H_0, $(\alpha_2 \approx -5.4, \beta_2 = 2)$, $\ln\Omega = 0$ since S_1 and S_2 are independent. Furthermore, the power of the test is close to 0.05 in this point. If the true values of α_2 and β_2 differ from H_0 in the direction $\alpha_2 \approx -2\beta_2$, the contour plots in all three figures show that the power is small. This area with low power starts in $(\alpha_2 \approx -5, \beta_2 \approx 2)$ and goes in the direction $\alpha_2 = k\beta_2$, where the value of k is between -2.5 and -2 depending on the sample size.

For the asymptotic power the result is explained by the fact that the appearance of the contour plots are determined by the non-centrality parameter φ and that φ in general is small in this direction. Note that φ gets larger as the sample size increases. Therefore, the asymptotic power in the direction $\alpha_2 \approx -2\beta_2$ in Figure 8.3(b) is larger compared to the asymptotic power in Figure 8.2(b) and in Figure 8.1(b).

To explain the result for the simulated power the calculation of the power needs to be studied in detail. In particular the probabilities π_{jk} in the alternative hypothesis have to be compared to the probabilities $\widetilde{\pi}_{jk}$ under the null hypothesis, $j = 0, 1, 2$ and $k = 1, 2, 3$. Given an arbitrary alternative hypothesis in the area with low power, Figure 8.4 compares π_{jk} with $\widetilde{\pi}_{jk}$ at the design points. The comparison shows that π_{jk} resembles $\widetilde{\pi}_{jk}$ very well at the design points, $x = 0.1741$ and $x = 2.2049$. In the third design point, at $x = 5.7469$, there is a small difference between π_2 and $\widetilde{\pi}_2$. Nevertheless, the difference is so small that it requires a very large sample to reject the hypothesis of independence. Hence, the estimated frequencies $N_k\widetilde{\pi}_{jk}$ in (7.6) are similar to the observed frequencies r_{jk} for $j = 0, 1, 2$ and $k = 1, 2, 3$. This results in low power, despite the fact that α_2 and β_2 are far from their restrictions under the null hypothesis.

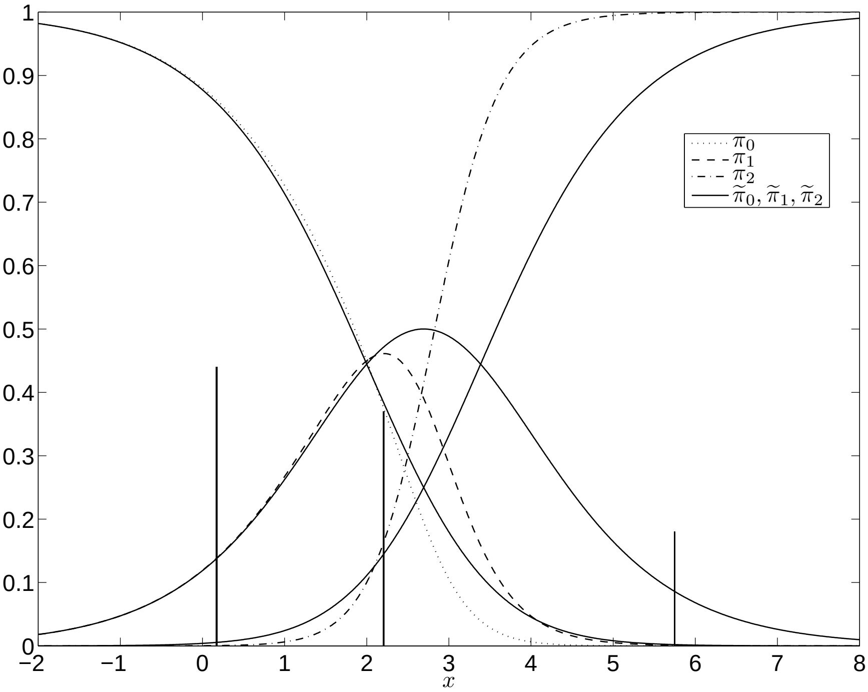


Figure 8.4: The probabilities π_0, π_1, π_2 for $\theta^T = (-2, -8, 1, 3.25)$ and the probabilities $\widetilde{\pi}_0, \widetilde{\pi}_1, \widetilde{\pi}_2$ under independence, i.e. for $\widetilde{\theta}^T = (-2, -5.4, 1, 2)$. The design points of the E-optimal design are shown as vertical lines where the height of each line corresponds to the size of the design weight.

The distribution of the score test statistic in small samples is not completely comparable to the asymptotic distribution of the score test statistic. An important difference is that the asymptotic power of the test is based only on the cumulative distribution function of the non-central χ^2 distribution. Hence, the described computational problems can never occur for the asymptotic power, since the non-centrality parameter φ is always computable. The expression for the test statistic in finite samples, on the other hand, is based on a finite number of observations where computational problems do occur.

8.3 Robustness of a Locally Optimal Design

As stated before, the optimal designs depend on the parameters α_1 and β_1. For that reason it is of great interest to study how well the optimal designs perform against incorrect guesses of the parameter values. Especially important is that the designs generate a good power function for the test they are supposed to optimize, regardless of incorrect guesses of the parameter values. A design is considered to be robust if the power of the test for fairly incorrect guesses of the parameter values is close to the power generated by the correct parameter values. Throughout the section, the design considered is E-optimal for $\alpha_1 = -2$ and $\beta_1 = 1$. Note that all power functions are based on the optimal design under consideration. Power is evaluated for a number of different alternative hypotheses, since the power function also depends on the alternative hypothesis. Because a complete robustness examination of the optimal design is extensive, only a sample of the resulting plots is shown here.

Consider the E-optimal design for testing both restrictions $\alpha_2 = 2\alpha_1 - \ln 4$ and $\beta_2 = 2\beta_1$. Assume that the alternative hypothesis is given by $\alpha_2 = 2\alpha_1 - \ln 4 + a$ and $\beta_2 = 2\beta_1 + b$ where a and b are constants. Since the asymptotic power, given α_1 and β_1, is an even function of a and b, only positive values of a and b are considered. The evaluation of the robustness utilizes the relative power for different (α_1, β_1) with respect to $(\alpha_1 = -2, \beta_1 = 1)$. Relative power is used because it gives a direct measure of the robustness.

Figure 8.5 shows the relative power as a function of both α_1 and β_1 given some values on a and b. Note that the relative power in all figures is equal to one in the point $(\alpha_1 = -2, \beta_1 = 1)$.

In the first three columns of contour plots the design is robust around the line $\alpha_1 + 6\beta_1 = d$, where d is between $[-0.5; 0.8]$ depending on the alternative hypothesis. The large power along this line is explained by large values on the non-centrality parameter, φ. All contour plots are parallel to this line, verifying that the design is robust for these parameter values. Some values of α_1 and β_1 generate a relative power larger than two, showing that the design is very efficient for these parameter values. On the other hand, the relative power decreases fast when the values on

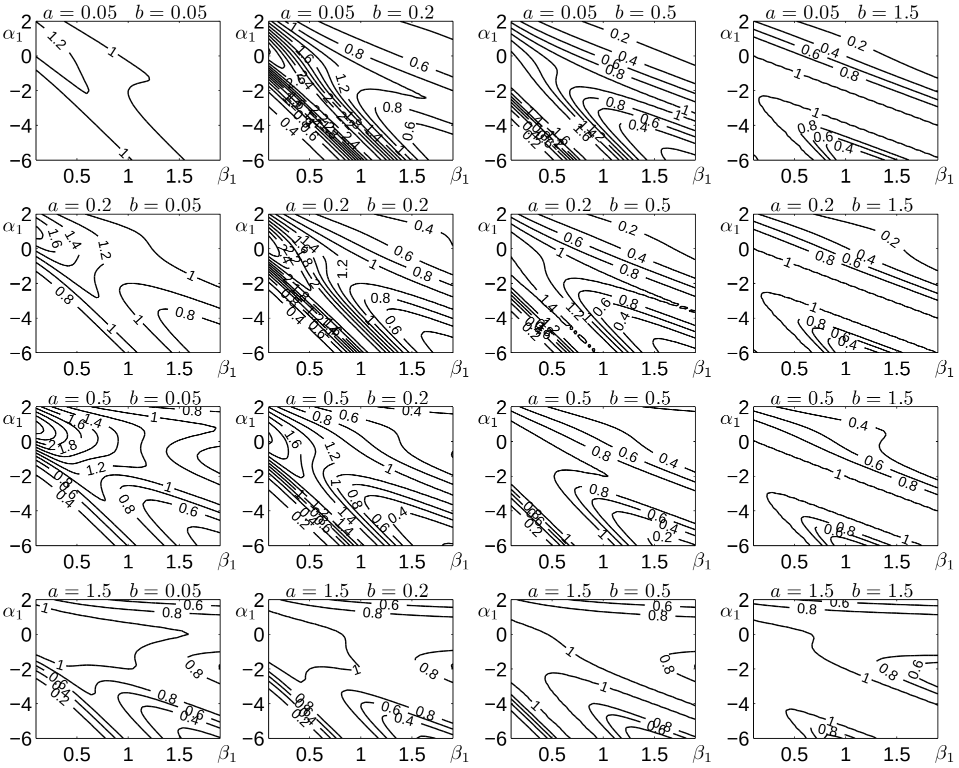
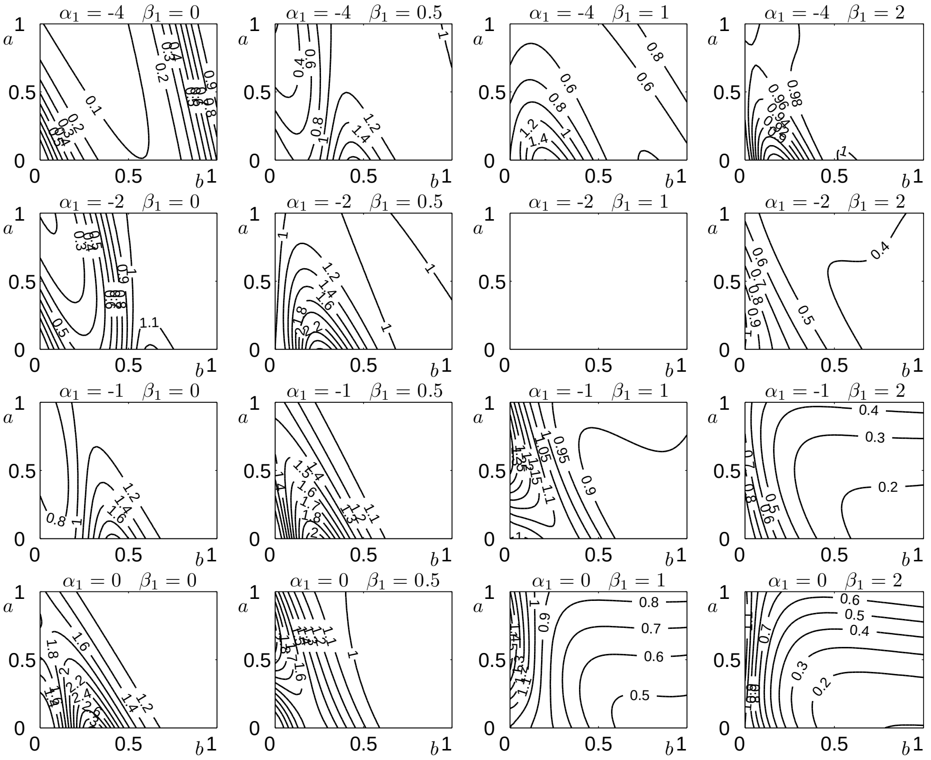


Figure 8.5: Contour plots of the relative power with respect to $\alpha_1 = -2$ and $\beta_1 = 1$. The different alternative hypotheses are given above each plot.

α_1 or β_1 change in the direction perpendicular to the line $\alpha_1 + 6\beta_1 = d$. The design is least efficient when $\alpha_1 + 6\beta_1 < d$. In the fourth and last column, where $b = 1.5$, the design is robust around the line $\alpha_1 + 2\beta_1 \approx 1$.

Figure 8.5 treats mainly how different (α_1, β_1) affects the robustness of the design. It is therefore hard based on those figures to conclude how the alternative hypothesis, i.e. a and b, affects the robustness.

Figure 8.6 illustrates in more detail how the alternative hypothesis affects the robustness. Figure 8.6 displays the relative power as a function of both a and b given some values on α_1 and β_1. By looking at the plots in row two and column three in Figure 8.6, the design is robust when just one of the guesses on α_1 and β_1 is incorrect. From Figure 8.6 it is clear that the relative power as a function of a and b changes a lot when both guesses on α_1 and β_1 are incorrect.



α_1 = -4 $\beta_1 = 0$ | α_1 = -4 $\beta_1 = 0.5$ | α_1 = -4 $\beta_1 = 1$ | α_1 = -4 $\beta_1 = 2$

α_1 = -2 $\beta_1 = 0$ | α_1 = -2 $\beta_1 = 0.5$ | α_1 = -2 $\beta_1 = 1$ | α_1 = -2 $\beta_1 = 2$

α_1 = -1 $\beta_1 = 0$ | α_1 = -1 $\beta_1 = 0.5$ | α_1 = -1 $\beta_1 = 1$ | α_1 = -1 $\beta_1 = 2$

α_1 = 0 $\beta_1 = 0$ | α_1 = 0 $\beta_1 = 0.5$ | α_1 = 0 $\beta_1 = 1$ | α_1 = 0 $\beta_1 = 2$

Figure 8.6: Contour plots of the relative power with respect to $\alpha_1 = -2$ and $\beta_1 = 1$ as a function of a and b. The different parameter values are given above each plot.

In summary, the locally optimal design is found to be fairly efficient against incorrect guesses of the parameter values when just one of the guesses on α_1 and β_1 are incorrect. Moreover when $b \leq 0.5$, the design is robust around the line $\alpha_1 + 6\beta_1 = d$ where d is between $[-0.5; 0.8]$ depending on the alternative hypothesis. However, outside this line large variations in the robustness occur for different alternative hypotheses when both guesses on α_1 and β_1 are incorrect.

Chapter 9

A Numerical Example: Cariogenic Effect of Diets

9.1 Description of the Example

Up to this point, the preceding chapters have outlined the simplified Cox model and discussed optimal designs and inferential aspects of the model. Some of the results obtained are demonstrated in this chapter using a data set. In the numerical example the model is estimated with and without restrictions of independence. Using these estimates the locally D-optimal design is derived. In addition, an E-optimal design, for maximizing the local power in a test for independence, is derived. The original design is compared to these locally optimal designs with respect to precision and power. The main part of Section 9.1 and Section 9.2 is presented in Bruce (2008).

The application, given in Andrews and Herzberg (1985), is a clinical trial using rats as experimental units. In the experiment 60 rats were randomly assigned to different doses of a certain substance. The purpose of the experiment was to see how this substance neutralizes caries in the teeth of the rats. During the feeding period all rats were therefore given a cariogenic control diet together with a dose of the substance that would hopefully neutralize the cariogenic effect. The used doses were

$0, 0.25, 0.5$, and 1 with equally many rats assigned to each dose[1]. At the end of the experiment, occlusal surfaces in each rat were examined for caries. The response variables are the first two occlusal surfaces, each binary coded as "caries" or "no caries".

The experiment is very briefly described in Andrews and Herzberg (1985). The briefness of the documentation is problematic since it is hard to interpret the choice of experimental design as well as other aspects of the experiment without a proper documentation. For example, is it difficult to determine if a high dose is feasible with respect to toxicity, ethical aspects, and costs. Unfortunately no further information on how the experiment was carried out has been found. Even a conversation with the authors of the book did not yield a full description of the experiment.

9.2 Model and Estimation

Let the response of the two occlusal surfaces be denoted by S_1 and S_2, respectively. Moreover, S_i equals one if caries is found and zero otherwise, $i = 1, 2$. Since there are two Bernoulli variables, the simplified Cox model for $S = S_1 + S_2$ has three response categories, $S = 0,\ S = 1$, and $S = 2$. The linear predictor is

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 + \beta_1 x \\ \alpha_2 + \beta_2 x \end{pmatrix},$$

where x denotes the used dose. Using the outcome $S = 0$ as reference category, the probabilities $\pi_0,\ \pi_1$, and π_2 become

$$\begin{aligned} \pi_0 &= \frac{1}{1 + e^{\eta_1} + e^{\eta_2}} \\ \pi_1 &= \frac{e^{\eta_1}}{1 + e^{\eta_1} + e^{\eta_2}} \\ \pi_2 &= \frac{e^{\eta_2}}{1 + e^{\eta_1} + e^{\eta_2}}. \end{aligned}$$

[1] During the experiment one rat from the group with dose 1 died. Consequently, the parameter estimates from the experiment are only based on 59 observations. Despite this, it is reasonable to conclude that the used design assigns equally many rats to each dose since this was the initial plan of the experiment.

In order to fit the simplified Cox model to the data set it must be assumed that S_1 and S_2 are identically distributed. If the distributions of S_1 and S_2 are different, a Cox model must be used instead. In the bivariate case, the Cox model has four outcomes with probabilities

$$\begin{aligned}
\pi_{00} &= \frac{1}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}\\
\pi_{10} &= \frac{e^{\eta_{10}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}\\
\pi_{01} &= \frac{e^{\eta_{01}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}\\
\pi_{11} &= \frac{e^{\eta_{11}}}{1+e^{\eta_{10}}+e^{\eta_{01}}+e^{\eta_{11}}}.
\end{aligned}$$

The linear predictor is

$$\eta = \begin{pmatrix}\eta_{10}\\ \eta_{01}\\ \eta_{11}\end{pmatrix} = \begin{pmatrix}\alpha_{10}+\beta_{10}x\\ \alpha_{01}+\beta_{01}x\\ \alpha_{11}+\beta_{11}x\end{pmatrix}.$$

To test if the Cox model can be reduced to the simplified Cox model, a likelihood ratio test is used. This test compares the unrestricted loglikelihood of the Cox model with the loglikelihood of the same model under the restrictions

$$\begin{cases}\alpha_{10}=\alpha_{01}\\ \beta_{10}=\beta_{01}.\end{cases}$$

The observed test statistic is very close to zero $(1.4211\cdot 10^{-14})$ indicating that S_1 and S_2 have the same distribution. The reason why the test statistic is so close to zero is that there are equally many "caries" for S_1 as for S_2 in the data.

Assuming that data are described by a simplified Cox model, the model can now be estimated. When estimating the model it is also of interest to test if the simplified Cox model can be reduced to a model for independent data. When S_1 and S_2 are independent, $S \sim Bin\,(2,\pi.)$ with

$$\pi. = \frac{e^{\eta_1}}{2+e^{\eta_1}}$$

and

$$\eta_1 = \alpha_1 + \beta_1 x.$$

Using Theorem 6.1, the hypotheses

$$\begin{aligned} H_0 &: \left\{ \begin{array}{l} \alpha_2 = 2\alpha_1 - \ln 4 \\ \beta_2 = 2\beta_1 \end{array} \right. \\ H_1 &: \alpha_2 \neq 2\alpha_1 - \ln 4 \quad \text{or} \quad \beta_2 \neq 2\beta_1 . \end{aligned}$$

for testing independence are obtained. The restrictions under H_0 impose the probabilities π_0, π_1, and π_2 to follow a binomial distribution for independent Bernoulli variables. The likelihood ratio test statistic is

$$T_{LR} = 2\left\{ l\left(\widehat{\theta}; \mathbf{y}\right) - l\left(\widetilde{\theta}; \mathbf{y}\right) \right\},$$

where $\mathbf{y}$ is the matrix of responses from the whole sample. $\widetilde{\theta}$ and $\widehat{\theta}$ are the maximum likelihood estimators under H_0 and H_1, respectively. The observed values of $\widetilde{\theta}$ and $\widehat{\theta}$ are found to be

$$\widetilde{\theta}^T = \left(\widetilde{\alpha}_1, \widetilde{\beta}_1\right) \approx (3.6138, -1.7628)$$

and

$$\widehat{\theta}^T = \left(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\beta}_1, \widehat{\beta}_2\right) \approx (-0.1068, 3.2843, -0.1824, -1.8952),$$

respectively. T_{LR} has a χ^2-distribution with 2 degrees of freedom, asymptotically. The observed test statistic is found to be

$$T_{LR} \approx 15.79.$$

Since the critical value on the 5% level is 5.991, the hypothesis of independence is rejected.

9.3 Locally D-optimal Design

Assuming that the parameter estimates $\widehat{\theta}^T = (-0.1068, 3.2843, -0.1824, -1.8952)$ are the true parameter values, i.e. $\theta = \widehat{\theta}$, a locally D-optimal design is derived. The locally D-optimal design is

$$\xi^* = \left\{ \begin{array}{ccc} 0.7893 & 2.4226 & 14.5712 \\ 0.3029 & 0.4515 & 0.2456 \end{array} \right\}.$$

Figure 9.1 presents the probabilities π_0, π_1, and π_2 as functions of the cariogenic neutralization substance under study, x. The design points for the D-optimal design can be seen as vertical lines where the height of each line corresponds to the size of the design weight. The design points of the original design only estimate π_0, π_1, and π_2 for a limited range of values of x, between 0 and 1. Figure 9.1 shows that the D-optimal design on the other hand provides much more information about π_0, π_1, and π_2 for larger values of x. The third design point of the D-optimal design, located at $x = 14.5712$, is far from any of the design points of the original design. The location of this third design point is explained by the fact that the D-optimal design is based on $\theta = (-0.1068, 3.2843, -0.1824, -1.8952)$. Given this θ, π_1 in Figure 9.1 decreases so slowly that a design point as far out as $x = 14.5712$ is needed.

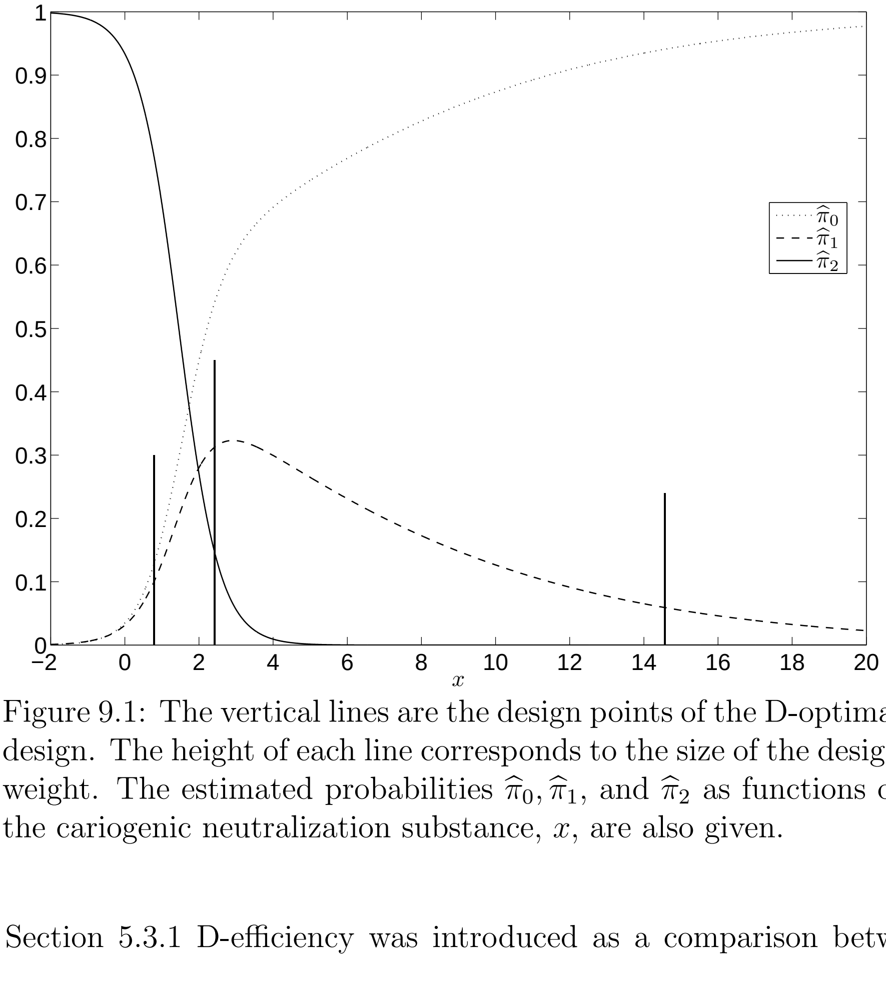


Figure 9.1: The vertical lines are the design points of the D-optimal design. The height of each line corresponds to the size of the design weight. The estimated probabilities $\widehat{\pi}_0, \widehat{\pi}_1$, and $\widehat{\pi}_2$ as functions of the cariogenic neutralization substance, x, are also given.

In Section 5.3.1 D-efficiency was introduced as a comparison between designs with respect to the precision in the parameter estimator. Compared to the D-optimal design, the D-efficiency for the original design is

$$D_{eff} \approx 0.1981.$$

Thus, the D-optimal design needs only around 20 per cent as many observations as the original design to reach the same precision in the parameter estimator. Precision is here interpreted as the generalized volume of the confidence ellipsoid of the parameters.

9.4 Locally E-optimal Design

Suppose that interest is in deriving a design so that a test for independence is conducted. In Chapter 8 it was argued that a locally E-optimal design should be used in order to maximize the power of the test. Recall that the locally E-optimal design maximizes the smallest locally asymptotic power of the score test for testing the hypothesis of independence above. Assume that $\widetilde{\alpha}_1$ and $\widetilde{\beta}_1$ from the experiment are the true values of the parameters, i.e. $\theta = \widetilde{\theta}$, the locally E-optimal design, ξ_E, is derived. For $\theta^T = (\alpha_1, \beta_1) = (3.6138, -1.7628)$, ξ_E is a 3-point design with the following design points and design weights

$$\xi_E = \left\{ \begin{array}{ccc} 0.2406 & 1.3650 & 3.3924 \\ 0.2542 & 0.4721 & 0.2737 \end{array} \right\}.$$

Figure 9.2 shows the locations and the weights of the design points of ξ_E. The probabilities π_0, π_1, and π_2 as functions of x are also shown. As for the D-optimal design, the E-optimal design provides much more information about π_0, π_1, and π_2 for larger values of x compared to the original design. A difference between the D-optimal design and the E-optimal design is that there is no support in ξ_E for an extreme design point such as the third design point in ξ^*, at $x = 14.5712$. The lack of an extreme design point in ξ_E is explained by the appearance of π_0, π_1, and π_2 in Figure 9.2. For values of x larger than five, π_0 is very close to one. In this region there is no need for a design point since the design point would provide very little information on how π_0, π_1, and π_2 change as functions of x. Thus, the largest design point in ξ_E is located at $x = 3.3924$.

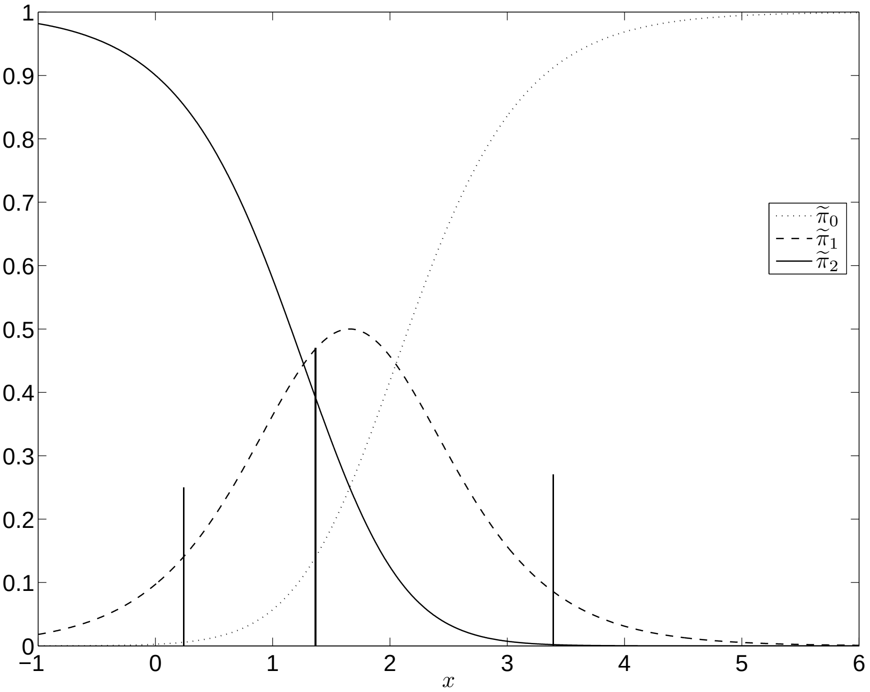


Figure 9.2: The estimated probabilities $\widetilde{\pi}_0, \widetilde{\pi}_1$, and $\widetilde{\pi}_2$ as functions of the cariogenic neutralization substance, x. The design points of the E-optimal design are shown as vertical lines.

To evaluate the precision in the parameter estimator based on the E-optimal design, D-efficiency is used once again. Compared to the E-optimal design, the D-efficiency for the original design is

$$D_{eff} \approx 0.2323.$$

Thus, the E-optimal design needs only around 23 per cent as many observations as the original design to reach the same precision in the parameter estimator. With respect to the D-optimal design, the E-optimal design has a D-efficiency of 0.8529.

Suppose that the designs are to be compared with respect to the smallest asymptotic power among all alternative hypotheses in a test for independence. The preferable measure is then E-efficiency, which compares the precision in the worst estimated contrast of the parameters. Let ζ denote the smallest eigenvalue of $M(\theta)$ based on the original design and let ζ_E

denote the smallest eigenvalue of $M(\theta)$ based on the E-optimal design, $\theta^T = (3.6138, -1.7628)$. The E-efficiency is defined as

$$E_{eff} = \left(\frac{\zeta}{\zeta_E}\right).$$

In the example, the E-efficiency is around 0.098. The E-efficiency can be interpreted in several ways. The most interesting interpretation is perhaps to compare the designs with respect to the number of observations required to reach a certain precision. In an experiment based on the original design, let s^2 and N denote the variance of the worst estimated contrast of the parameters and the total number of observations, respectively. If the experiment is based on the E-optimal design, let s_E^2 and N_E denote the variance of the worst estimated contrast of the parameters and the total number of observations, respectively. Furthermore, ζ and ζ_E are inversely proportional to

$$\zeta \propto \frac{N}{s^2} \quad \text{and} \quad \zeta_E \propto \frac{N_E}{s_E^2},$$

respectively. This follows from the fact that the length of the longest axis of the confidence ellipsoid is inversely proportional to the smallest eigenvalue of the information matrix, Atkinson and Donev (1992). Thus, given the same number of observations, the variance of the worst estimated contrast of the parameters for an experiment based on the E-optimal is 0.098 times the corresponding variance for an experiment based on the original design. Assume that

$$\frac{s^2}{N} = p, \tag{9.1}$$

where p is the desired precision. Given that $s_E^2 = 0.098s^2$ and that the precision is equal to p,

$$N_E = \frac{0.098s^2}{p}. \tag{9.2}$$

The two equations in (9.1) and (9.2) together imply that

$$N_E = 0.098 \cdot N.$$

Hence, the E-optimal design needs only around 10 per cent as many observations as the original design to reach the same precision in the worst estimated contrast of the parameters.

E-efficiency only compares designs with respect to the worst estimated contrast of the parameters. By comparing the designs with respect to power in several different alternative hypotheses, a more comprehensive comparison of the designs is obtained. The power for different alternative hypotheses, i.e. different α_2 and β_2, is calculated using simulated data as follows. For each alternative hypothesis a data set with a sample size of 1000 is created 5000 times. The score test statistic is then calculated using the formulas (7.6), (7.7), and (7.8) in Chapter 7. The power is given by the percentage of the score test statistics that are larger than the critical value at the significance level 5%.

In Figure 9.3, the relative power between the E-optimal design and the original design is given. For interpretation of the figure, consider an area where the relative power is around three. This means that the power of the E-optimal design is approximately three times larger than the power of the original design. Since $\alpha_1 = 3.6138$ and $\beta_1 = -1.7628$, the values for α_2 and β_2 under H_0 are $\alpha_2 \approx 5.84$ and $\beta_2 \approx -3.53$, respectively. Figure 9.3 shows that the power for the E-optimal design is about as high or higher compared to the original design for almost all alternative hypotheses. The only exception is the upper-left corner where the power for the original design is approximately twice as large compared to the E-optimal design. In the lower-right part, there are alternative hypotheses for which the power of the E-optimal design is as much as ten times larger than the power of the original design.

9.5 Concluding Remarks

The last two sections show that the original design is very inefficient in estimating the parameters compared to both the D-optimal design and the E-optimal design. The D-optimal design and the E-optimal design, only require around 20 per cent as many observations as the original design to reach the same precision in the parameter estimator. This implies that a substantial amount of money had been saved if the D-optimal design or the E-optimal design had been implemented instead of the original design.

It should be stated that this conclusion depends on the assumption that $\widehat{\theta}$ and $\widetilde{\theta}$ are true values for the D-optimal design and the E-optimal design,

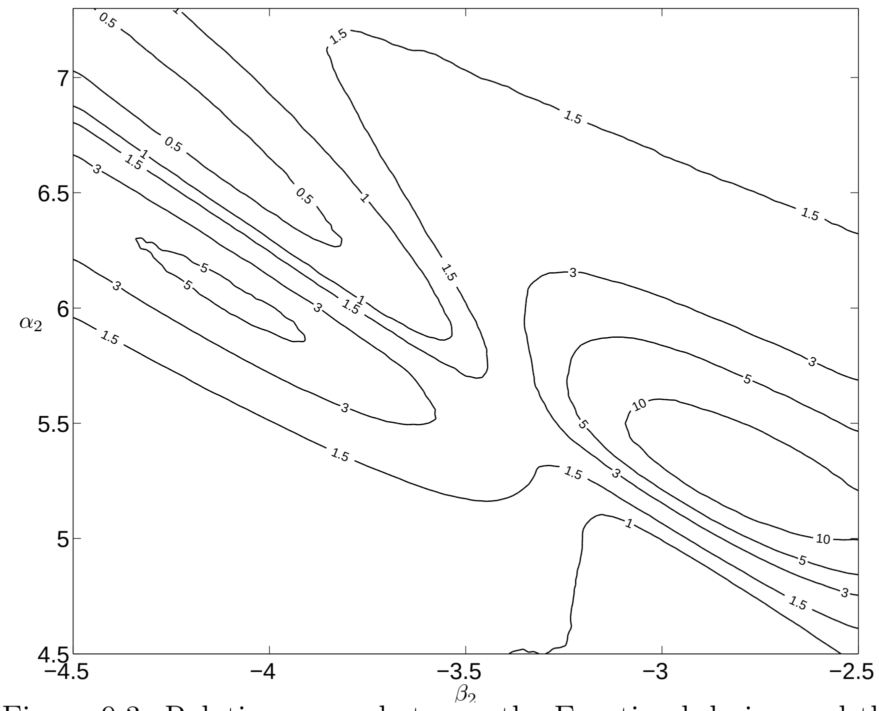


Figure 9.3: Relative power between the E-optimal design and the original design for different alternative hypotheses, (α_2, β_2). The power for both designs is calculated for the score test which tests if the variables are independent. The values for α_2 and β_2 under H_0 are $\alpha_2 \approx 5.84$ and $\beta_2 \approx -3.53$, respectively.

respectively. If these guesses on the parameter values differ considerably from the true parameter values, the favorable properties of the respective design might not hold. Moreover, the comparison is somewhat unfair in that the goals of the original experiment are not known. For example, it is unlikely that the purpose of the experiment was to maximize the power in a test for independence.

On the other hand, if $\widehat{\theta}$ and $\widetilde{\theta}$ are inefficient estimates it is quite possible that the ineffectiveness was caused by a poor original design. Another possible consequence of incorrect parameter values is the extreme design point at $x = 14.5712$ in the D-optimal design. If the original design had generated parameter estimates closer to the true values, it is possible that the corresponding locally D-optimal design would not include such an extreme design point. However, it is also possible that different parameter estimates generate an even more extreme design point than $x = 14.5712$.

When deriving the D-optimal and E-optimal designs, different guesses on the parameter values are used. This is due to different inferential goals of the two designs. The D-optimal design is based on $\widehat{\theta}$ since the objective is estimating the parameters with good precision, whereas the E-optimal design is based on $\widetilde{\theta}$ since the design is derived under the hypothesis of independence.

Chapter 10

Discussion and Suggestions for Further Research

A multinomial logit model for identically distributed Bernoulli variables is examined in this thesis. How well the model fit the data depends on several factors and assumptions.

The model is only applicable to data where identically distributed Bernoulli variables exist. For a similar model Agresti (2002) argued that data fit poorly if the marginal distributions of the Bernoulli variables differ substantially. Hence, it remains to investigate how poorly the model fits when the assumption is not valid.

Another critical assumption is that an observation on $S_1, S_2, \ldots, S_k$ is really composed of k Bernoulli variables. Estimation of the model can not be carried out when there is an item nonresponse on some of the variables $S_1, S_2, \ldots, S_k$. This restriction implies that all applications must have a fixed batch or litter size. The problem does not only occur in observational studies, but can also occur in experimental studies where the sizes of the batches are under control. George and Bowman (1995b) extended the model presented in Chapter 2 to incorporate data with random batch sizes. A problem with incorporating the batch size into the model is that the correlation between $(S_1, S_2, \ldots, S_k)$ is affected by the batch size. According to Williams (1987) this often leads to problems with overdispersion. In addition, the size of the batch may be affected by the treatment, (Williams, 1987; George and Bowman, 1995a).

All observations with a certain value on the covariates are assumed to be homogeneous in the sense that they have the same parameter values. A further development would be to include, e.g. random effect parameters in the linear predictor. These parameters would then account for variations among observations. Parameters for modelling the heterogeneity among individuals are included in the general model for dependent binary responses given in Agresti (1997).

A bivariate model where the dependence between the variables is controlled by a distance covariate is proposed in Section 3.4. This model is just briefly discussed and additional research is required. Examples of issues to study include interpretation of the model, parameter restrictions in a test for independence, and various optimal designs including optimal designs in two variables, both a distance covariate and a covariate for the dose.

In Chapter 8, an E-optimal design which maximizes the smallest locally asymptotic power of the score test for testing independence is proposed. The asymptotic power function of the test is compared with power functions for finite sample sizes using a small simulation experiment. The results indicate that the locally optimal designs perform well as long as $\ln \Omega$ is negative. A problem occurs however, for large values of $\ln \Omega$. The problem is related to the fact that almost all observations fall in the same response category. This affects the performance of the test statistic in small samples. If the Bernoulli variables are strongly correlated the value of the test statistic might not exist. On the other hand, other test procedures for testing independence in 2×2 contingency tables have the same problem when small expected cell frequencies appear, see (Haberman, 1988; Agresti, 2002).

The locally optimal design is fairly robust against incorrect parameter values. It should be stated that the investigation about robustness is not comprehensive since only one parameter setup is examined, i.e. $\alpha_1 = -2$ and $\beta_1 = 1$. To obtain a more general result on the robustness of an arbitrary design, a spectrum of values for α_1 and β_1 should be examined. The reason why it has not been done here is that such investigation would totally overwhelm the other results of the thesis.

An additional aspect is that the robustness evaluation relies on asymptotic results instead of results derived using simulated data. It was shown in Section 8.2 that the power function based on simulations does not resemble the asymptotic power function completely, at least in situations where the correlation between the variables is strong. Altogether, this shows that in the particular situation of interest, it is important to thoroughly examine the performance of the test statistic before the design of the experiment is determined.

In Chapter 9 an example with diets reducing the cariogenic effect is analyzed. When deriving optimal designs for this application as well as many other applications involving clinical trials, a design point is sometimes too extreme with respect to toxicity and ethical aspects. The locally D-optimal design derived in the cariogenic example has a design point around 15. This should be compared with the original design, where all design points were between 0 and 1. In this context the probability of toxicity is certainly too high for a design point around 15. This raises the issue of using some constraint optimization technique, Cook and Fedorov (1995), or a penalty function as in Dragalin and Fedorov (2006), to incorporate the restriction of a maximum tolerated dose in the model.

Another issue for further research is to derive algebraic forms for locally D-optimal designs. One approach to do this is to first derive analytical results which show how the information matrix is affected when the values of the parameters change. In this context, Fan (1999) and Puu (2003) derived analytical results for the determinant of the information matrix. Then, the results about the information matrix are used to derive general expressions for locally D-optimal designs. Although Fan (1999) and Puu (2003) worked with other models, it is of interest to derive similar results for the model in this thesis. The derivation for algebraic expressions is also an issue for the asymptotically optimal designs discussed in Section 5.3. The proposed designs have a setup which resemble the limiting locally optimal designs in Fan (1999). A future task is therefore to analogously derive closed forms for limiting optimal designs under the simplified Cox model.

In a clinical trial with the responses efficacy (yes/no) and toxicity (yes/no) the main objective may not be generally good precision in the parameter estimates. Instead, focus could be on finding conditions that maximize the probability for a particular outcome, say the response ("efficacy","no toxicity"). Optimal designs with this objective have been studied in e.g. Heise and Myers (1996), Fan (1999), and Rabie (2004). Derivations of similar designs are of interest for the Cox model and perhaps also the simplified Cox model.

Finally, another comment about the robustness of the various locally D-optimal designs derived in the thesis: According to Zocchi and Atkinson (1999) D-optimal designs for logistic models are dependent on the parameters. Changes in the parameter values result in different design points, different design weights and sometimes even different number of design points. It is therefore of interest to derive- optimal in average designs and sequential designs for the models studied in this thesis.

Bibliography

Agresti, A. (1997). “A Model for Repeated Measurements of a Multivariate Binary Response.” *J. Amer. Statist. Assoc.*, 92, 437, 315–321.

— (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics, 2nd ed. New York: Wiley-Interscience [John Wiley & Sons].

Andrews, D. F. and Herzberg, A. M. (1985). *Data : a Collection of Problems from many Fields for the Student and Research Worker*. Springer Series in Statistics. New York: Springer-Verlag.

Appelgren, J. (2004). “Locally D-optimal Designs for Bivariate Logistic Regression.” Department of Statistics, University of Umeå. Statistical Studies No. 32.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford science publications.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*, vol. 34 of *Oxford Statistical Science Series*. Oxford: Oxford University Press.

Atkinson, A. C. and Haines, L. M. (1996). “Designs for Nonlinear and Generalized Linear Models.” In *Design and Analysis of Experiments*, vol. 13 of *Handbook of Statist.*, 437–475. Amsterdam: North-Holland.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, Mass.-London. With the collaboration of Richard J. Light and Frederick Mosteller.

Bonney, G. E. (1987). “Logistic Regression for Dependent Binary Observations.” *Biometrics*, 43, 4, 951–973.

Bruce, D. (2008). "Some Properties for a Simplified Cox Binary Model." *Comm. Statist. Theory Methods*. To appear.

Bruce, D. and Nyquist, H. (2007). "Testing for Dependency of Bernoulli Variables." *International Journal of Statistical Sciences*, 6, 151–161.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. 2nd ed. Duxbury.

Chaloner, K. and Larntz, K. (1989). "Optimal Bayesian Design Applied to Logistic Regression Experiments." *Journal of Statistical Planning and Inference*, 21, 2, 191–208.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. Springer Texts in Statistics, 2nd ed. New York: Springer-Verlag.

Cook, D. and Fedorov, V. (1995). "Constrained Optimization of Experimental Design." *Statistics*, 26, 2, 129–178. With discussion and a rejoinder by the authors.

Cox, D. R. (1972). "The Analysis of Multivariate Binary Data." *Applied Statistics*, 21, 2, 113–120.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, vol. 32 of *Monographs on Statistics and Applied Probability*. 2nd ed. London: Chapman & Hall.

Dale, J. R. (1986). "Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses." *Biometrics*, 42, 4, 909–917.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC Texts in Statistical Science Series, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.

Dragalin, V. and Fedorov, V. (2006). "Adaptive Designs for Dose-finding based on Efficacy-Toxicity Response." *Journal of Statistical Planning and Inference*, 136, 6, 1800–1823.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer Series in Statistics, 2nd ed. New York: Springer-Verlag. With contributions by Wolfgang Hennevogl.

Fan, S. K. (1999). "Multivariate Optimal Designs." Ph.D. thesis, School of Statistics, University of Minnesota.

Fan, S. K. and Chaloner, K. (2004). "Optimal Designs and Limiting Optimal Designs for a Trinomial Response." *Journal of Statistical Planning and Inference*, 126, 1, 347–360.

Fedorov, V. V. and Hackl, P. (1997). *Model-Oriented Design of Experiments*, vol. 125 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Texts in Statistical Science Series. London: Chapman & Hall.

George, E. O. and Bowman, D. (1995a). "A Full Likelihood Procedure for Analysing Exchangeable Binary Data." *Biometrics*, 51, 2, 512–523.

— (1995b). "A Saturated Model for Analyzing Exchangeable Binary Data: Applications to Clinical and Developmental Toxicity Studies." *J. Amer. Statist. Assoc.*, 90, 431, 871–879.

George, E. O. and Kodell, R. L. (1996). "Tests of Independence, Treatment Heterogeneity, and Dose-Related Trend with Exchangeable Binary Data." *J. Amer. Statist. Assoc.*, 91, 436, 1602–1610.

Goldstein, H. (2003). *Multilevel Statistical Models*. 3rd ed. London: Arnold.

Gumbel, E. J. (1961). "Bivariate Logistic Distributions." *J. Amer. Statist. Assoc.*, 56, 335–349.

Haberman, S. J. (1988). "A Warning on the use of Chi-squared Statistics with Frequency Tables with Small Expected Cell Counts." *J. Amer. Statist. Assoc.*, 83, 402, 555–560.

Häggström, J. (2000). "The Minimax Approach to Optimum Design of Experiment." Ph.D. thesis, Department of Statistics, University of Umeå. Statistical Studies No. 24.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.

Heise, M. A. and Myers, R. H. (1996). "Optimal Designs for Bivariate Logistic Regression." *Biometrics*, 52, 2, 613–624.

Hirji, K. F. (1994). "Exact Analysis for Paired Binary Data." *Biometrics*, 50, 4, 964–974.

Kalish, L. and Rosenberger, J. (1978). "Optimal Designs for the Estimation of the Logistic Function." Tech. Rep. 33, Pennsylvania State University, University Park, Pennsylvania.

Kang, S.-H. and Park, S. M. (2000). "Exact Likelihood Ratio Test of Independence of Binary Responses within Clusters." *Comput. Statist. Data Anal.*, 33, 1, 15–23.

Kiefer, J. (1959). "Optimum Experimental Designs." *J. Roy. Statist. Soc. Ser. B*, 21, 272–319.

Kiefer, J. and Wolfowitz, J. (1960). "The Equivalence of two Extremum Problems." *Canad. J. Math.*, 12, 363–366.

Le Cassie, S. and Van Houwelingen, J. (1994). "Logistic Regression for Correlated Binary Data." *Applied Statistics*, 43, 1, 95–108.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). "Multivariate Regression Analyses for Categorical Data." *Journal of the Royal Statistical Society, Series B*, 54, 1, 3–40.

Mandel, E., Bluestone, C., Rockette, H., Blatter, M., Reisinger, K., Wucher, K., and Harper, J. (1982). "Duration of Effusion after Antibiotic Treatment for Acute Otitis Media: Comparison of Cefaclor and Amoxicillin." *Pediatric Infectious Diseases*, 1, 1, 310–316.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall.

Müller, W. G. and Pázman, A. (1998). "Design Measures and Approximate Information Matrices for Experiments without Replications." *J. Statist. Plann. Inference*, 71, 1-2, 349–362.

— (1999). "An Algorithm for the Computation of Optimum Designs under a given Covariance Structure." *Comput. Statist.*, 14, 2, 197–211.

— (2001). "Optimal Design of Experiments Subject to Correlated Errors." *Statist. Probab. Lett.*, 52, 1, 29–34.

— (2003). "Measures for Designs in Experiments with Correlated Errors." *Biometrika*, 90, 2, 423–434.

Murtaugh, P. A. and Fisher, L. D. (1990). "Bivarate Binary Models of Efficacy and Toxicity in Dose-Ranging Trials." *Comm. Statist. Theory Methods*, 19, 6, 2003–2020.

Palmgren, J. (1991). "Approaches to Modelling Bivariate Binary Responses; An Empirical Adventure." National Public Health Institute, Helsinki.

Pettersson, H. (2001). "Optimum in Average and Minimax Designs for Estimation of Generalized Linear Models." Ph.D. thesis, Department of Statistics, University of Umeå. Statistical Studies No. 26.

Pettersson, H. and Nyquist, H. (2003). "Computation of Optimum in Average Designs for Experiments with Finite Design Space." *Communications in Statistics. Simulation and Computation*, 32, 1, 205–221.

Prentice, R. L. (1986). "Binary Regression using an Extended Beta-Binomial Distribution, with Discussion of Correlation Induced by Covariate Measurement Errors." *J. Amer. Statist. Assoc.*, 81, 394, 321–327.

— (1988). "Correlated Binary Regression with Covariates Specific to each Binary Observation." *Biometrics*, 44, 4, 1033–1048.

Puu, M. (2003). "Optimum Experimental Designs for Generalized Linear Models with Multinomial Response." Department of Statistics, University of Umeå. Statistical Studies No. 31.

Rabie, H. (2004). "Optimal Designs for Dose-Finding in Contingent Response Models." Ph.D. thesis, University of Missouri - Columbia.

Rosner, B. (1984). "Power and Sample Size for a Collection of 2 × 2 Tables." *Biometrics*, 40, 4, 1025–1035.

Silvey, S. D. (1980). *Optimal Design*. London: Chapman & Hall.

Skellam, J. G. (1948). "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials." *J. Roy. Statist. Soc. Ser. B.*, 10, 2, 257–261.

Tielsch, J., Sommer, A., Katz, J., Quigley, H., and Ezrine, S. (1991). "Socioeconomic Status and Visual Impairment among Urban Americans." *Archives of Ophthalmology*, 109, 5, 637–641.

Wang, Y. (2002). "Optimal Experimental Design for the Poisson Regression Model in Toxicity Studies." Ph.D. thesis, Department of Statistics, Virginia Tech.

Williams, D. A. (1975). "The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity." *Biometrics*, 31, 4, 949–952.

— (1987). "Reader Reaction: Dose-response Models for Teratological Experiments." *Biometrics*, 43, 4, 1013–1016.

Zocchi, S. S. and Atkinson, A. C. (1999). "Optimum Experimental Designs for Multinomial Logistics Models." *Biometrics*, 55, 2, 437–444.

Zucker, D. M. and Wittes, J. (1992). "Testing the Effect of Treatment in Experiments with Correlated Binary Outcomes." *Biometrics*, 48, 3, 695–710.

Appendix A

Proof of Theorem 6.2

Like the case with binary data, the proof is divided into two parts where the first part shows that the parameter restrictions in (6.4) implies that $S_1, S_2, \ldots, S_k$ are mutually independent. The second part shows that mutual independence implies the parameter restrictions in (6.4).

Under a simplified Cox model, $\pi_{y_1y_2}$ can be expressed as in (3.4). Assume now that the parameters are determined by (6.4) so that

$$\begin{aligned}
\eta_{00} &= 0 \\
\eta_{10} &= \eta_{10} \\
\eta_{20} &= 2\eta_{10} - \ln k \\
&\vdots \\
\eta_{k0} &= k\left(\eta_{10} - \ln k\right) \\
\eta_{01} &= \eta_{01} \\
\eta_{11} &= \eta_{10} + \eta_{01} - \ln k + \ln\left(k-1\right) \\
&\vdots \\
\eta_{k-1,1} &= \left(k-1\right)\left(\eta_{10} - \ln k\right) + \eta_{01} \\
\eta_{02} &= 2\eta_{01} - \ln k \\
&\vdots \\
\eta_{0k} &= k\left(\eta_{01} - \ln k\right).
\end{aligned}$$

Then

$$\begin{aligned} e^{\eta_{y_1y_2}} &= e^{y_1(\eta_{10}-\ln k)+y_2(\eta_{01}-\ln k)+\ln \binom{k}{y_1\ \ y_2}} \\ &= \left(\frac{e^{\eta_{10}}}{k}\right)^{y_1}\left(\frac{e^{\eta_{01}}}{k}\right)^{y_2}\binom{k}{y_1\ \ y_2}. \end{aligned}$$

Moreover, the denominator of (3.4) is a multinomial series

$$\sum_{\substack{y_1,y_2\geq 0\\ y_1+y_2\leq k}} e^{\eta_{y_1y_2}} = \sum_{\substack{y_1,y_2\geq 0\\ y_1+y_2\leq k}} \binom{k}{y_1\ \ y_2}\left(\frac{e^{\eta_{10}}}{k}\right)^{y_1}\left(\frac{e^{\eta_{01}}}{k}\right)^{y_2} 1^{k-y_1-y_2}.$$

From the Multinomial Theorem

$$\sum_{\substack{y_1,y_2\geq 0\\ y_1+y_2\leq k}} e^{\eta_{y_1y_2}} = \left(1+\frac{e^{\eta_{10}}}{k}+\frac{e^{\eta_{01}}}{k}\right)^{k}.$$

Hence $\pi_{y_1y_2}$ can be written as

$$\pi_{y_1y_2} = \frac{e^{\eta_{y_1y_2}}}{\sum\limits_{\substack{i,j\geq 0\\ i+j\leq k}} e^{\eta_{ij}}} =$$

$$\binom{k}{y_1\ \ y_2}\underbrace{\left(\frac{\frac{e^{\eta_{10}}}{k}}{1+\frac{e^{\eta_{10}}}{k}+\frac{e^{\eta_{01}}}{k}}\right)^{y_1}}_{\pi_{1\cdot}^{y_1}}\underbrace{\left(\frac{\frac{e^{\eta_{01}}}{k}}{1+\frac{e^{\eta_{10}}}{k}+\frac{e^{\eta_{01}}}{k}}\right)^{y_2}}_{\pi_{2\cdot}^{y_2}}\underbrace{\left(\frac{1}{1+\frac{e^{\eta_{10}}}{k}+\frac{e^{\eta_{01}}}{k}}\right)^{k-y_1-y_2}}_{(1-\pi_{1\cdot}-\pi_{2\cdot})^{k-y_1-y_2}}.$$

Thus $S_1, S_2, \ldots, S_k$ are mutually independent and (Y_1, Y_2) are multinomial distributed with $(\pi_{1\cdot}, \pi_{2\cdot}, k)$.

Without any assumption on $S_1, S_2, \ldots, S_k$ it follows immediately that (6.4) is true for η_{00}, η_{10}, and η_{01}. Next assume that $S_1, S_2, \ldots, S_k$ are mutually independent. This implies that $\ln\Omega$ for all local tables (3.5), (3.6), and (3.7) are equal to zero for all possible values on (y_1, y_2). Assume now that

$$\eta_{y_1-1y_2} = (y_1-1)(\eta_{10}-\ln k) + y_2(\eta_{01}-\ln k) + \ln\binom{k}{y_1-1\ \ y_2}$$

and that

$$\eta_{y_1y_2} = y_1\,(\eta_{10} - \ln k) + y_2\,(\eta_{01} - \ln k) + \ln\binom{k}{y_1 \quad y_2}.$$

Equating the expression in (3.5) to zero yields,

$$\begin{aligned}
\eta_{y_1+1y_2} &= 2\left\{y_1\,(\eta_{10} - \ln k) + y_2\,(\eta_{01} - \ln k) + \ln\binom{k}{y_1 \quad y_2}\right\}\\
&\quad -\left\{(y_1-1)\,(\eta_{10} - \ln k) + y_2\,(\eta_{01} - \ln k) + \ln\binom{k}{y_1-1 \quad y_2}\right\}\\
&\quad +\ln\binom{k}{y_1+1 \quad y_2} - 2\ln\binom{k}{y_1 \quad y_2} + \ln\binom{k}{y_1-1 \quad y_2}\\
&= \{\eta_{10}-\ln k\}(2y_1-y_1+1)+\{\eta_{01}-\ln k\}(2y_2-y_2)+\ln\binom{k}{y_1+1 \quad y_2}\\
&= (y_1+1)\,(\eta_{10} - \ln k) + y_2\,(\eta_{01} - \ln k) + \ln\binom{k}{y_1+1 \quad y_2}
\end{aligned}$$

Next assume that (6.4) is true for $\eta_{y_1y_2-1}$ and $\eta_{y_1y_2}$. By using the expression in (3.6) it is then shown, with the same method as for $\eta_{y_1+1y_2}$ above, that (6.4) is true for $\eta_{y_1y_2+1}$. Thus it has been shown that (6.4) is true for $y_1 = 2, 3, \ldots, k$ regardless of the value of y_2 and the other way around for y_2. In order for the proof to be complete, it must also be shown that (6.4) is true for η_{11}. This is done by using the third expression for the local odds ratio given in (3.7). Equating (3.7) to zero yields

$$\begin{aligned}
2\eta_{11} &= \eta_{20} + \eta_{02} - \ln\binom{k}{2 \quad 0} - \ln\binom{k}{0 \quad 2}\\
&\quad +2\ln\binom{k}{1 \quad 1}\\
&= 2\,(\eta_{10} - \ln k) + \ln\binom{k}{2 \quad 0} + 2\,(\eta_{01} - \ln k) + \ln\binom{k}{0 \quad 2}\\
&\quad -\ln\binom{k}{2 \quad 0} - \ln\binom{k}{0 \quad 2} + 2\ln\binom{k}{1 \quad 1}\\
&= 2\,(\eta_{10} - \ln k) + 2\,(\eta_{01} - \ln k) + 2\ln\binom{k}{1 \quad 1}.
\end{aligned}$$

Hence

$$\eta_{11} = (\eta_{10} - \ln k) + (\eta_{01} - \ln k) + \ln\binom{k}{1 \quad 1}$$

and the proof is complete.