Limburgs Universitair Centrum

Center for Statistics

Flexible Modelling of Correlated Multivariate Data with Applications in Animal Studies.

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, Groep Wiskunde
aan het Limburgs Universitair Centrum te verdedigen door

CHRISTEL FAES

Promotors
Prof. Dr. H. Geys
Prof. Dr. M. Aerts

Voor Sven

Dankwoord

Toen ik vier jaar geleden aan dit doctoraatswerk begon, wist ik nog niet helemaal wat me te wachten stond, maar ging ik met veel plezier deze grote uitdaging tegemoet. Nu ik terugkijk op de afgelopen jaren, heb ik daar alleen maar positieve herinneringen bij. Dit is grotendeels door de onmisbare steun en inzet van een aantal mensen die ik hiervoor heel graag wil bedanken.

In de eerste plaats wil ik mijn promotoren, prof. dr. Helena Geys en prof. dr. Marc Aerts, bedanken voor hun fantastische begeleiding en de aangename samenwerking, voor de zeer waardevolle discussies en de onmisbare steun. Helena, bedankt voor de jarenlange, fijne samenwerking. Marc, bedankt voor je luisterend oor en goede raad waar ik steeds op kon rekenen.

Ik ben erg blij dat ik terecht gekomen ben in het Centrum voor Statistiek, and would like to thank all my colleagues for the pleasant and stimulating environment they have provided. Ook wil ik prof. dr. Geert Molenberghs erg bedanken voor zijn betrokkenheid, zijn bijdragen en interesse in mijn doctoraat.

I gratefully thank all my co-authors for the fruitful cooperation and the many helpful discussions. Their comments on earlier drafts have greatly improved this manuscript. A special thank goes to the Santiago de Compostela team for their warm hospitality and pleasant collaboration.

Verder wil ik mijn familie en vrienden van harte bedanken. Op hun belangstelling en vriendschap heb ik steeds kunnen rekenen. Heel speciaal wil ik mijn moeder bedanken voor de liefde en steun die ze me gegeven heeft en het geduld dat ze met me gehad heeft, niet alleen tijdens de laatste 4 jaar, maar al 26 jaar lang. Mama, bedankt, voor alles! Een laatste woord heb ik bewaard voor Sven. Sven, jou gezelschap, vertrouwen en liefde waren mijn grootste steun. Bedankt omdat je er steeds was als ik je nodig had.

Christel Faes

Diepenbeek, 17 september 2004

Contents

List of Abbreviations

AIC	Akaike's Information Criterion
BB	Beta-Binomial (Model)
BF	Bayes Factor
BHV	Bovine Herpesvirus-1
BIC	Bayesian Information Criterion
BMD	Benchmark Dose
CSM	Conditional Synchrony Measure
CWGEE	Cluster Weighted Generalized Estimating Equations
DIC	Deviance Information Criterion
DYME	Diethylene Glycol Dimethyl Ether
EG	Ethylene Glycol
EPA	Environmental Protection Agency
FDA	Food and Drug Administration
FT	Fixation Target
GEEn	Generalized Estimating Equations (nth order)
GLMM	Generalized Linear Mixed Model
ISI	Inter-Stimulus Interval
KU	Key Up
LED	Lower Effective Dose
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood

MR	Molenberghs-Ryan (Model)
MSE	Mean Squared Error
NOAEL	No Observable Adverse Effect Level
NTP	National Toxicology Program
PB	Push Buttons
PSTH	Peristimulus Time Histogram
PQL	Penalized Quasi-Likelihood
QRA	Quantitative Risk Assessment
REML	Restricted Maximum Likelihood
THEO	Theophylline
WGEE	Weighted Generalized Estimating Equations

List of Tables

List of Figures

Chapter 1

Introduction

Correlated data are common in many health sciences studies, where clustered, multivariate, longitudinal, hierarchical and spatially organized data are frequently observed. In clustered data, subjects within the same group are likely to be more similar than subjects among different groups. When several outcomes of interest are measured on the same individual, these multivariate outcomes are likely to be correlated. Associations in spatial data are due to spatial proximity. In all these cases, the observations under study share some common characteristics and statistical analysis requires taking such associations into account. There exist many ways to deal with these correlation structures, ranging from the most naive one of ignoring the associations to approaches that correct for correlations or model the clustering. Failure to account for the effect of clustering can result in erroneous estimation of the variability of the parameter estimates, and hence in misleading inference. Therefore, appropriate statistical techniques for the analysis of correlated measurements are of interest.

An important consideration in the statistical modelling of correlated data concerns the type of outcome. Methods for clustered or multivariate continuous data are widely available, where the normal distribution with its elegant properties plays a prominent role. However, when the outcome variable is discrete or categorical, techniques for correlated data are less standard, because of the lack of a discrete analogue to the multivariate normal distribution. In general, one might record more than one outcome for each individual, with outcomes of a different type and individuals clustered within groups or spatially correlated.

In this thesis, several scientific disciplines where clustering of categorical data are encountered are topic of interest. Main focus is on risk analysis modelling in developmental toxicity studies. Each of the research areas discussed involves a different study

1

design, leading to unique and interesting statistical problems. In the next sections, we briefly describe the research areas considered in this thesis, where correlated data arise naturally in different ways.

1.1 Developmental Toxicity Studies

One of the most common worries for expectant parents is that something will be wrong with their baby. Across Europe, just over 2 per cent of the newborn babies have a defect that will affect their ability to survive or function normally. Such adverse effects include foetal death, altered foetal growth, and structural changes. Defects of limbs, heart and spinal cord represent about half of all abnormalities. Often, defects in the development of foetuses are consequent to exposure of the pregnant women to a toxic substance. Perhaps the most prominent example of a drug that caused defects on the developing foetus is Thalidomide. Thalidomide, a sedative specifically prescribed for pregnant women to combat symptoms associated with morning sickness, was brought on the market in 1957 after inadequate testing. It caused severe congenital malformation of the limbs of children whose mothers had used this drug in the early stages of pregnancy. Following this, much stricter rules about drug testing were introduced, and regulatory agencies, such as the U.S. Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA) stimulated reproductive and developmental toxicity research to better protect people against exposures to agents that cause developmental toxicity, such as drugs, harmful chemicals and other environmental hazards.

Developmental toxicity studies are necessary to investigate the potential risk of exposures on fertility, birth defects, and developmental abnormalities, and play an important role in the regulation of adverse exposures for human health. In this area however, it is unethical to deliberately expose humans to potentially damaging substances. In addition, the complexity and inherent variability of human populations complicates the evaluation of adverse environmental effects. For this reason, the field of toxicology has traditionally relied heavily on controlled studies in laboratory animals (Williams and Ryan 1996). Since laboratory studies involve considerable amounts of time and money, as well as huge numbers of animals, it is essential that the most appropriate and efficient statistical models be used. Standard experimental protocols for conducting developmental toxicity studies were established by the U.S. FDA. Teratology studies are specifically designed to investigate the effects of exposure on the development of the foetuses. In a typical teratology study with a Segment II

design, the pregnant dams are exposed to a compound of interest during the critical period of major organogenesis and structural development. Dose levels for this design typically consist of a control group and three or four exposed groups, each involving 20 to 30 pregnant animals. The dams are sacrificed just prior to normal delivery, at which time the uterus is removed and the contents are thoroughly examined for the occurrence of defects. The number of dead and resorbed foetuses is recorded. Viable foetuses are measured for birth weight and examined carefully for the presence of malformations. An overview of the available data from developmental toxicity studies is given in Figure 1.1.

An important issue in developmental toxicity is the risk assessment. Risk assessment is the process by which scientific judgments are made concerning the potential adverse health consequences of exposure to a toxic agent. The National Research Council (1983) has defined risk assessment as including some or all of the following components: hazard identification, dose-response assessment, exposure assessment, and risk characterization. Hazard identification and dose-response evaluation involves examining all available experimental data to determine if an agent causes developmental toxicity. One component of the risk assessment is the examination of the dose-response relationship, i.e., the dependence of a particular outcome (e.g. the number of dead foetuses or the risk of a malformed foetus) on the dose that is administered to the dam. Another important component is the determination a safe level of exposure, i.e., quantitative risk assessment. The exposure assessment identifies human populations exposed or potentially exposed to an agent and the magnitude, duration, and spatial extent of the exposure. In risk characterization, the hazard identification, dose-response evaluation and the exposure assessment for given populations are combined to estimate some measure of the risk for developmental toxicity.

We concentrate on the hazard identification and dose-response evaluation of developmental toxicity. However, the analysis of developmental toxicity data as described above, raises a number of challenges (Aerts *et al.* 2002).

- *Multivariate Outcomes.* The developmental toxicity studies typically result in multiple outcomes of interest. Developmental toxicity may record both malformation outcomes (binary or ordinal) and birth weight (continuous) on each embryo, as both have been found to be indicative of a toxic effect. In addition, correlation between these outcomes exists (Ryan *et al.* 1991). Thus, jointly modelling the outcomes might be an appropriate statistical analysis. This involves a multivariate method to jointly analyse continuous and discrete outcomes.

dose

dam

. . .   implant   . . .

viable   non-viable

malformation   weight
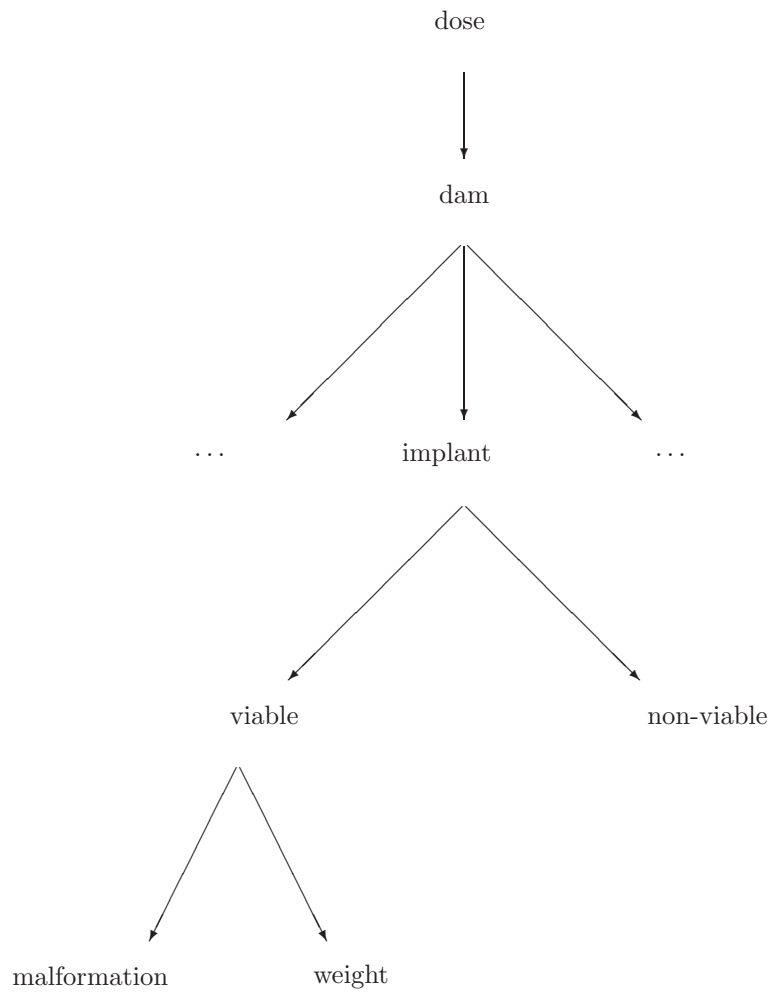
Figure 1.1: *Data structure of developmental toxicity studies.*

- *Litter Effect.* Correlation among the foetuses within litters is very likely to be present because of the genetic similarity and same treatment conditions for offspring of the same mother. The heterogeneity between clusters introduces an additional source of variation and complicates the analysis (Kupper *et al.* 1986, Chen and Kodell 1989).

- *Hierarchy.* Since the events of interest can occur at several point in development, methods for handling the hierarchically related outcomes (such as death and malformation) are also of interest. Ultimately, a model should take the hierarchical structure of the data into account: (i) a toxic insult early in gestation may result in an resorbed foetus; (ii) thereafter an implant is at risk of foetal death; (iii) foetuses that survived the entire gestation period are still at risk of low birth weight and/or malformations.

Accounting for the clustering of offspring within a litter and the multiplicity of outcomes has led to interesting statistical research topics, which are addressed in this thesis. Aside from the basic challenges of developing good multivariate models, particularly in the clustered data setting, additional technical challenges arise in applying these models in real data settings. For example, correlation structures are likely to change with exposure (Kupper *et al.* 1986). Determining how to use a complex model for risk assessment purposes, for example, calculating the benchmark dose, is another area of interest. In addition, proper flexible predictor models have to be used to fit the data appropriately.

1.2 Other Complex Studies

Challenging issues arise in various other settings as well. Each research area has its own specific purposes. The choice of the probability model used to analyse the data depends merely on the research questions and peculiarities of that setting. In every setting, flexible predictor models should be used to describe the data appropriately.

In *neurophysiology* one studies the function of the nervous system. One of the techniques used in neurophysiology is electrophysiology, which records the electrical activity produced by neurons. A version of this technique records the activity of single cells in the form of action potentials, or spikes, which typically last about 1 ms. Although the nature of the neural code is not clear at the present time, the firing of spike rates of neurons and its temporal evolution is considered a measurement of sensory, motor or behavioural activity. Since the frequency code is the way by which neurons communicate, it is important to the physiologist to assess the neural firing

rate. Another way to transmit information about the characteristics of a stimulus is via synchronization of the neurons. Therefore, it is also of interest to describe how 'synchronous' two spike trains are. This study results in multivariate binary observations that are correlated in time. In this study one needs an appropriate measure of synchrony of two neurons; a model that allows to analyse the firing rate for each neuron separately; a model that accounts for the correlation between the neurons; and a model that accounts for the correlation due to the longitudinal nature of the observations of neuronal activity during the experiment. In Chapter 8, we will propose a model that satisfies these properties.

In *veterinary epidemiology* one wants to investigate the occurrence of a disease in animals, to identify causes that may explain the occurrence of the disease, and ultimately to prevent or reduce the occurrence of that disease. When studying a specific infectious disease it is important to estimate the rate at which susceptible individuals acquire infection at different ages, i.e., the force of infection. Often however, animals in a study are grouped within herds. Once an infection is introduced in a herd, animals within the same herd have a high chance to get infected too. Thus, correlations between the animals arise naturally due to the grouping of animals in herds. In addition, not only the population force of infection can be important, but also the herd-specific force of infection is of interest. Chapter 9 describes possible models in this setting.

Gastroenterology is the branch of medicine dealing with the study of disorders affecting the stomach, intestines, and associated organs. The gastrointestinal tract is a tube, along which contents are pushed by the coordinated movements of its muscular wall. These movements are controlled mainly by the enteric nervous system, an extensive network of nerve cells that lies in the wall of the mammalian gastrointestinal tract. In *neuro-gastroenterology* one studies the functioning of the neurons in the intestines. Until now, there is no information as to the behaviour of spikes of neurons in successive slow waves. In Chapter 10, both the spatial and temporal determination of spikes are studied. In this setting, a flexible method for the spatial as well as longitudinal analysis of neural activity is investigated.

1.3   Organization of Subsequent Chapters

Throughout the present work, we focus mainly on quantitative risk analysis in developmental toxicity studies. Emphasis is on the choice of a flexible predictor model, on a correct choice of the probability model, and on efficient and computationally

attractive estimation method.

In Chapter 2, we present all data sets that are used throughout this work. The National Toxicology Program data are presented in Section 2.1. The electrophysiological experiment, investigating the synchrony of neurons in the brain of awake behaving animals, is introduced in Section 2.2. Section 2.3 describes the bovine herpesvirus study. The data investigating the spike patches in cat duodenum are presented in Section 2.4.

In Chapter 3 a review of possible statistical models for clustered binary data is given. Several types of model families exist. Possible probability models for clustered binary data are the marginal beta-binomial model (Williams 1975) and the conditional model of Molenberghs and Ryan (1999). Often however, complete specification of the joint distribution of the response vector can become problematic and fully likelihood-based estimation methods are awkward when dealing with complex data problems. Alternative estimation methods, such as pseudo-likelihood (Arnold and Strauss 1991) and generalized estimating equations (Liang and Zeger 1986, Zeger and Liang 1986), are described. There is also a wide class of possible predictor models describing the model parameters as function of variables of interest, ranging from linear to non-linear functions and from parametric to non-parametric models.

Two important components of the risk assessment process in developmental toxicity studies are the hazard identification and safe dose determination. Chapter 4 deals with the safe dose determination, whereas Chapter 5 focuses on the hazard identification. In both chapters, we restrict attention to clustered binary outcomes (malformation), and investigate the use of a flexible predictor model. In Chapter 4 the effect of misspecification of the dose-response model to assess a safe level of exposure is investigated. Although classical polynomials are very customary, they are often inadequate, especially when low dose extrapolation is envisaged. In order to get a reliable estimate of a safe level of exposure, the model should fit the data well in all respects. A very elegant alternative approach to classical polynomials, which falls within the realm of (generalized) linear methods, are fractional polynomials as proposed by Royston and Altman (1994). The behaviour of fractional polynomials in the context of quantitative risk assessment is investigated through extensive simulations. The fractional polynomials are much more flexible to attain the correct benchmark dose than conventional polynomials. It is shown how fractional polynomial predictors resolves possible probability model misspecification and may thus yield more reliable estimates.

Hazard identification in the risk assessment process deals with testing for dose trend. Also in this setting it is important that the model describes the data well.

Another alternative specification to the conventional linear predictor is given by a non-linear model (Davidian and Giltinan 1995). Nonlinear models pose non-trivial methodological challenges. For example, a classical power model $\alpha + \beta d^\gamma$, where d denotes dose and α, β, and γ are unknown parameters, suffers from lack of identifiability under the null hypothesis of no dose effect, since this null hypothesis corresponds to both $\beta = 0$ with γ arbitrary, as well as to $\gamma = 0$ with β arbitrary. A Bayesian method to test for no dose trend is topic of interest in Chapter 5.

Often however, the malformation type is recorded as an ordinal variable and the foetal birth weight is given, both being important measures of teratogenicity. A multivariate probability model reflecting the type of outcomes is required. In Chapter 6, a modelling approach for the joint analysis of ordinal and continuous outcomes in a clustered data setting is proposed. The model is motivated by dose-response modelling of malformation and foetal weight data from developmental toxicology experiments to be used for quantitative risk assessment. We introduce a likelihood-based method that is an extension of the Plackett-Dale approach. Specification of the full likelihood will be avoided using pseudo-likelihood methodology. The estimation of safe dose levels as part of quantitative risk assessment is illustrated based on a developmental toxicity experiment of diethylene glycol dimethyl ether in mice.

In previous chapters, focus was on the outcomes measured on the viable foetuses. However, also the number of viable foetuses in a dam, i.e., the litter size, is affected by the dose of a toxic agent. Therefore, a method that acknowledges the stochastic nature of the litter size is necessary. This is the topic of research in Chapter 7. A major problem in constructing a model for the full data structure is the intractability of the marginal likelihood. Due to this problem, formulation of the models is difficult. An interesting alternative is offered by Bayesian methods. A model to analyse the hierarchical data structure is proposed.

So far, methods were illustrated using developmental toxicity studies, but their relevance is far beyond this area. Other settings, each with their own research questions and design issues, are explored in the next chapters of this thesis. Chapter 8 deals with bivariate binary data, measured over time, in the context of electrophysiology. To describe how 'synchronous' two spike trains are, a variety of association measures can be used. We propose a new measure of synchrony, the conditional synchrony measure, which is the probability of firing together given that at least one of the two neurons is active. Focus is on the specification of a flexible marginal model for multivariate correlated binary data together with a pseudo-likelihood estimation approach, to adequately and directly describe the measures of interest. A joint model must allow different time- and covariate-depending firing rates for each neuron, and

must account for the association between them. The association between neurons might depend on covariates as well.

The next research area of interest is the veterinary epidemiology. Specific interest is in the force of infection in infectious diseases, or the rate at which susceptible individuals become infected. Often in this area, clustering arises due to the natural grouping of cattle within herds. So far however, one has always ignored the clustering when investigating the force of infection. The effect of clustering on the force of infection of the bovine herpesvirus-1 in Belgian cattle is investigated in Chapter 9. Both population-averaged and herd-specific patterns of disease are studied. Since the herdsize is related with the prevalence of the disease in this study, methods accounting for the non-ignorable cluster size are inspected.

Another type of correlated data is given by spatially organized data. In Chapter 10, the spatial location of neuron spikes in part of the small intestines is investigated. In addition, measurements are repeated in time. A flexible model describing the two-dimensional surface is investigated in Chapter 10. We explore the use of generalized linear mixed model for smoothing purposes. A major advantage of the mixed model approach is that it can handle smoothing together with grouping (or other types of correlations) in a unified model.

Chapter 2

Motivating Examples

2.1 National Toxicology Program Data

The National Toxicology Program (NTP) develops scientific information about potentially toxic chemicals that can be used for protection of public health and prevention of chemically induced diseases. One of the goals is to determine the developmental toxicity of chemicals to which a wide segment of the population is exposed. In this case, that population consists of pregnant women. Since it is difficult to assess exposure to potential developmental toxicants and outcomes of exposure within this group, well-controlled animals studies must provide a basis for human risk assessment.

In this section we introduce three developmental toxicity studies conducted by the Research Triangle Institute under contract to the National Toxicology Program of the U.S. These studies investigate the effects in mice of different chemicals: ethylene glycol (EG) (Price *et al.* 1985), diethylene glycol dimethyl ether (DYME) (Price *et al.* 1987) and theophylline (THEO) (Lindstrom *et al.* 1990).

2.1.1 Ethylene Glycol

Ethylene glycol (EG) is also called 1,2-ethanediol and can be represented by the chemical formula $HOCH_2CH_2OH$. EG is a high-volume industrial chemical with diverse applications. It is used to make antifreeze and de-icing solutions for cars, airplanes, and boats, to make polyester compounds, and is used as a solvent in the paint and plastics industries. EG is employed as an ingredient in photographic developing solutions, hydraulic brake fluids and in the formulation of several types of inks. Furthermore, one uses EG as a softening agent for cellophane and in the synthesis of

Table 2.1: Summary data from an EG experiment in mice.

Dose	Dams, ≥ 1		Viable		Adverse Effect	
(mg/kg/day)	impl.	viab.	Nr.	Mean	Malf.	Death
0	25	25	297	11.9	4.0	68.0
750	24	24	276	11.5	66.7	62.5
1500	23	22	229	10.4	81.8	73.9
3000	23	23	226	9.8	95.7	91.3

various chemical products, such as plasticizers, synthetic fibers and waxes (Windholz 1983).

While EG may not be hazardous to humans in normal industrial handling, it can become dangerous when used at elevated temperatures or when ingested. Exposure to large amounts of ethylene glycol can damage the kidneys, heart, and nervous system. In addition, ingestion of antifreeze products, which consist for approximately 95% of EG, is toxic and may result in death (Price *et al.* 1985).

Price *et al.* (1985) describe a study in which timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water. Dosing occurred during the period of major organogenesis and structural development of the foetuses (gestational days 6 through 15). The doses selected for the study were 0, 750, 1500 or 3000 mg/kg/day, with 25, 24, 23 and 23 timed-pregnant mice randomly assigned to each of these dose groups, respectively. Table 2.1 presents for each dose group the number of dams in the study containing at least one implant, as well as the number of dams having at least one viable foetus.

When interest in the risk of EG on the litter of a dam, the probability that at least one foetus in that litter has the adverse event under consideration is crucial. Table 2.1 summarizes the adverse event for a litter. For each dose group, the number of viable foetuses, the mean litter size, the percentage of malformed litters, and the percentage of dams with at least one dead foetus are displayed. It suggests clear dose-related trends for all outcomes.

Instead of considering a collapsed outcome, indicating whether a dam has at least one abnormal foetus, one can investigate the adverse event of having an abnormal foetus in a cluster. A summary of the data is presented in Figure 2.1. The dots represent the probability of an affected foetus per cluster. The stars represent the
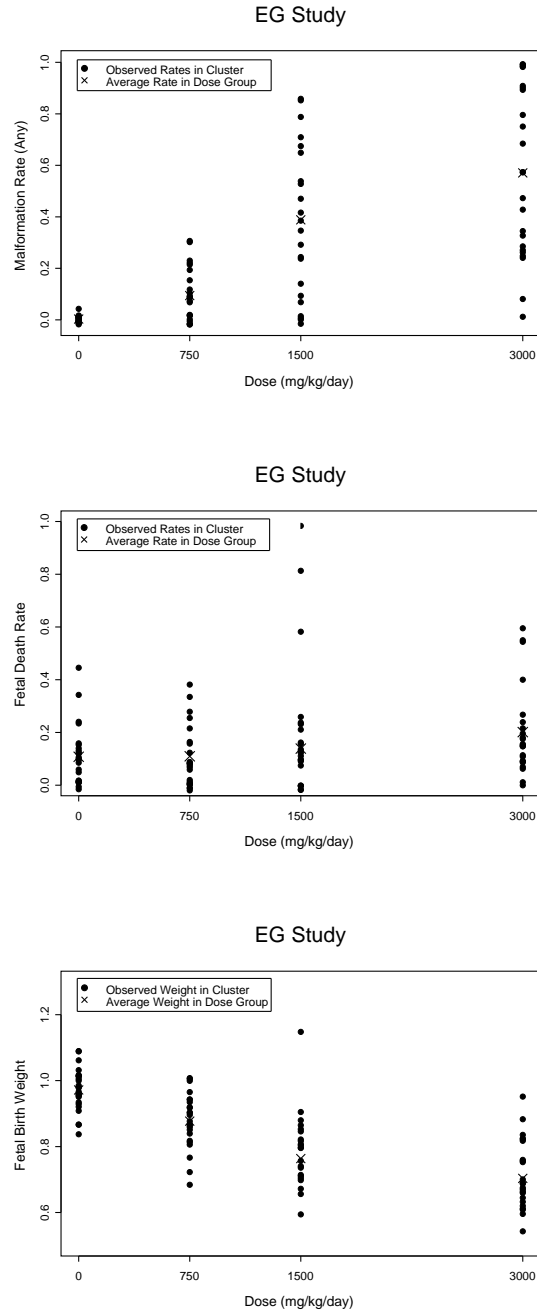
Figure 2.1: *EG study. Observed and averaged malformation rates, foetal death rates and average foetal birth weights.*

average risk of an affected foetus per dose group. The data show clear dose-related trends for both the malformation outcome, as well as for foetal death. The observed malformation rate in the dose groups increases quickly from 0.3% in the control group to 57.1% in the highest dose group. The risk of a dead foetus rises slowly with increasing dose, from 10.8% up to 20.1%. Also the foetal birth weight is pictured in Figure 2.1. The average foetal weight decreases monotonically with dose, declining from 0.972 g in the control group to 0.704 g in the highest dose group.

The EG data is used in Chapters 4 and 5, where alternative predictor models to the classical polynomials are studied, and in Chapter 7, where focus is on the specification of a model accounting for the hierarchical nature of the data.

2.1.2 Diethylene Glycol Dimethyl Ether

Diethylene glycol dimethyl ether (DYME), also referred to as diglyme, bis(2-methoxy-ethyl) ether or dimethyl carbitol, belongs to the group of ethylene glycol ethers. The molecular structure of DYME can be represented as $(CH_3OCH_2CH_2)_2O$. DYME is mainly used as a solvent, as an inert reaction medium in chemical synthesis, and as a separating agent in distillations. As a solvent it is widely used in the manufacture of textile dyes, cosmetics and protective coatings such as lacquers, metal coatings, baking enamels, etc. (Baumann and Muth 1997).

Although to date, several attempts have proven inadequate to evaluate the potential of glycol ethers to produce human reproductive toxicity, structurally related compounds have been identified as reproductive toxicants in several mammalian species, producing testicular toxicity and embryotoxicity. Price *et al.* (1987) describe a study in which DYME was administered by gavage in distilled water to timed-pregnant mice during major organogenesis (gestational days 6 through 15). The doses selected for the study were 0, 62.5, 125, 250, and 500 mg/kg/day with 21, 20, 24, 23 and 23 pregnant dams randomly assigned to each of these dose groups, respectively. In this experiment, each viable foetus is examined further for the occurrence of a malformation and for the birth weight. The malformation status is recorded on an ordinal scale: normal, a minor malformation or full malformation.

Table 2.2 shows, for each dose group, the number of dams having at least one implant, the number of dams having at least one viable foetus, the number of live foetuses, the mean litter size, the percentage of dams having at least one foetus with a full malformation, the percentage of dams having at least one foetus with a minor or full malformation and the percentage of dams with at least one dead foetus. The data show clear dose-related trends.
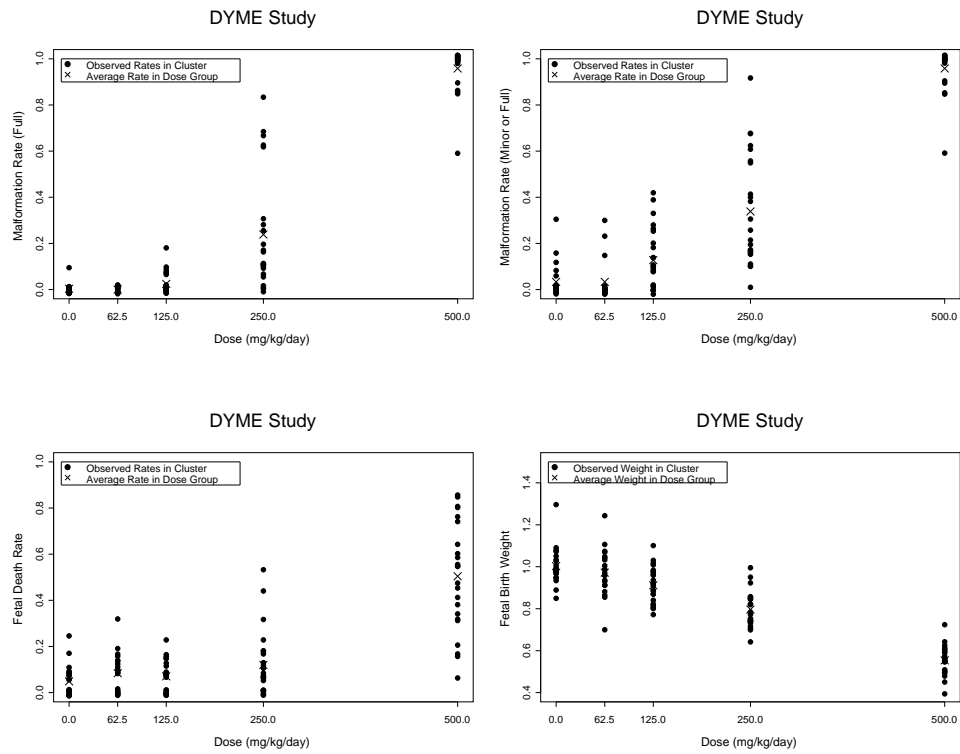
Figure 2.2: *DYME study. Observed and averaged malformation rates (left: full mal-formation, right: minor or full malformation), observed and averaged foetal death rates and average foetal birth weights.*

Table 2.2: Summary data from a DYME experiment in mice.

Dose	Dams, ≥ 1		Viable		Malformations		
(mg/kg/day)	impl.	viab.	Nr.	Mean	Full	Minor or Full	Death
0.0	21	21	282	13.4	4.8	23.8	47.6
62.5	20	20	225	11.3	0.0	15.0	60.0
125	24	24	290	12.1	25.0	66.7	62.5
250	23	23	261	11.3	82.6	95.7	69.6
500	23	23	141	6.1	100.0	100.0	100.0

The adverse effects of the possibly toxic chemical DYME on the foetus is presented in Figure 2.2. The data show clear dose-related trends for all outcomes. There is a pronounced dose-related reduction in foetal weight, with the highest administration of DYME resulting in roughly half of the mean weight in control animals. The average foetal birth weight declines from 1.000 g in the lowest dose group to 0.554 g in the highest dose group. The malformation data also exhibit trends with dose. The rate of a minor malformation increases at the lower doses, while the rate of full malformations shows strong increase at the highest doses. As a consequence, the rate of any malformation (minor or full) increases monotonically with dose, ranging from 2.9% in the control group to 93.6% in the highest dose group. The foetal death rate increases from 4.9% in the control group to 50.4% in the highest dose group.

The DYME data is used in Chapters 6, where focus is on the specification of a probability model which accounts for the complex data structure in developmental toxicity studies.

2.1.3   Theophylline

Theophylline is structurally classified as a methylxanthine. The molecular formula of anhydrous theophylline is $C_7H_8N_4O_2$. It occurs as a white, odourless, crystalline powder with a bitter taste. Theophylline is a natural component in tea, coffee and chocolate. Further, it is widely used in the pharmacy. It is one of the most commonly used medications for the treatment of the symptoms of chronic asthma. Its most important actions are to prevent the signs and symptoms of asthma, especially during the night, and to reduce the need for cortisone type medication. Theophylline has been shown to cross the human placenta and is secreted in breast milk.

Table 2.3: Summary data from a THEO experiment in mice.

Dose	Dams, ≥ 1		Viable		Adverse Effect	
(mg/kg/day)	impl.	viab.	Nr.	Mean	Malf.	Death
0	26	25	296	11.8	4.0	65.4
282	26	25	278	11.1	8.0	69.2
372	33	29	300	10.3	17.2	90.9
396	23	17	197	11.6	17.6	78.3

The oral intake of theophylline (THEO) was evaluated for toxic and teratogenic effects in mice. Lindstrom *et al.* (1990) expose the dams to the concentrations 0, 0.075, 0.15 and 0.2%, corresponding to 0, 282, 372 and 396 mg/kg/day respectively. THEO was administered continuously in the feed on gestational days 6 through 15.

The probability of an affected cluster, i.e., having a dam with at least one affected foetus, is investigated. Results are summarized in Table 2.3. The table suggests gradually increasing dose-related trends in the malformation and death rate.

Also the risk that a foetus exhibits the adverse event under investigation is crucial. The malformation rate, death rate and birth weights of the foetuses are displayed in Figure 2.3. There is a slight increase of the risk of a malformed foetus, from 0.3% in the control group to 2.5% in the highest dose group. The death rate rises gradually, from 8.4% until 23.6%. The foetal birth weight decreases from 1.036 g in the control group down to 0.917 g in the highest dose group.

This data set is used in Chapter 5, where it is investigated whether THEO has a toxic effect on the developing foetus.

2.2   Electrophysiological Data

Neurons can transmit information about the characteristics of a stimulus via the spike rate of neurons and via synchronization of the neurons. Therefore, it is of interest to describe how 'synchronous' two spike trains are. A fundamental methodology to investigate the activity of some neurons is electrophysiology which records the electrical signals produced by individual neurons within the brain of awake behaving animal.

This work is motivated by electrophysiological experiments carried out in two
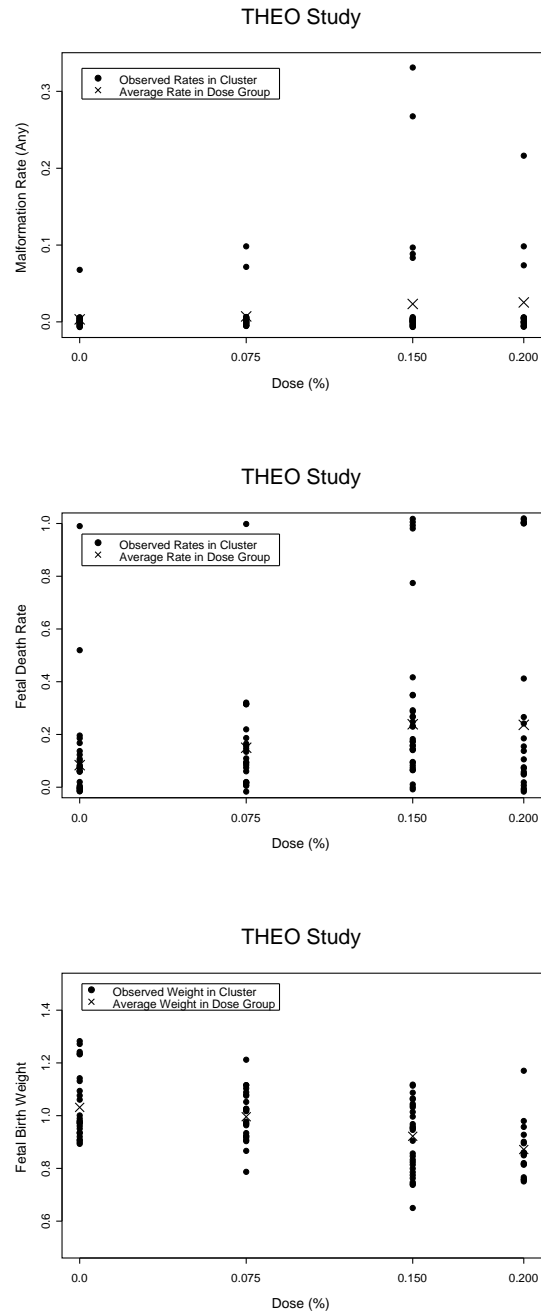
Figure 2.3: *THEO study. Observed and averaged malformation rates, foetal death rates and average foetal birth weights.*
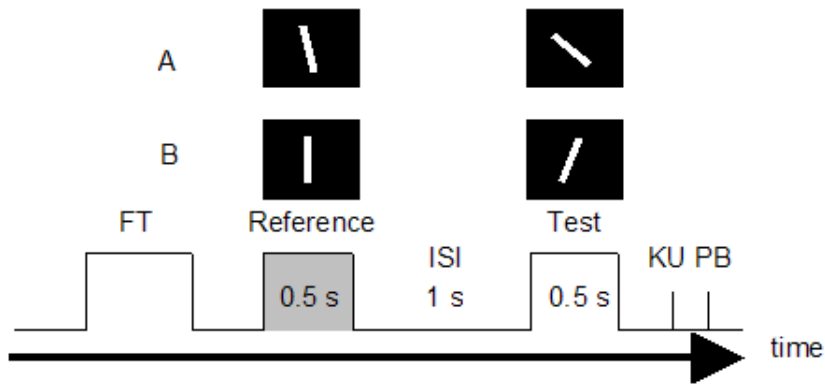
Figure 2.4: *Electrophysiological study. Temporal sequence of the discrimination task.*

male behaving monkeys (*Macaca mulatta*). The monkeys were trained to discriminate line orientations (Vazquez, Cano and Acuña 2000). Monkeys had their head fixed, and looked binocularly at a monitor screen placed at 114 cm from their eyes, and during the task their right arm operated a lever. A panel with two switches was in front of the monkey, at hand reach. The left and right switches were used in the discrimination task to signal orientations of the visual stimulus to the left and right, respectively. Figure 2.4 shows the temporal sequence of the discrimination task during the experiment.

The trial begins when the monkey presses a key lever, and visually fixates a short vertical bar (FT) at the center of the monitor screen. The monkey should maintain the fixation during the trial, otherwise the trial is aborted. Then two stimuli, reference and test, each of 500 ms duration, were presented in sequence, with a fixed inter-stimulus interval (ISI: 1000 ms). The stimuli were stationary bright lines, and three different reference orientations were used ($90°$, $95°$, $85°$). Different test lines, ten per reference stimulus, were presented clockwise or counter clockwise to the reference line in steps of $1°$. At the end of the second stimulus the monkey released the key (KU), and pressed one of two switches (PB) indicating whether the orientation of the second stimulus was clockwise or counter clockwise to the first stimulus. A and B are two examples of the stationary bright lines used as stimulus. Monkeys were rewarded with a drop of water for correct discrimination.

The activity of neurons in the primary visual cortex was recorded simultaneously, through the insertion of different microelectrodes into the monkey's brain, while the monkey performed the visual discrimination task. To study the synchrony between

cell pairs the analysis was restricted to the effect of the three reference stimuli (500 ms). The 200 ms before the reference stimulus comes on can be taken as control (baseline cell activity), due to the fact that during this period there was no visual stimulus whatsoever. Another period of 200 ms out of the 1000 ms of the Inter-Stimulus Interval (ISI) was also taken as control, i.e., recovery of the cell activity. We emphasized the fact that the monkey has to attend to the reference stimuli, otherwise the monkey could not solve the task (Vazquez *et al.* 2000). Per trial, every action potential is recorded during these 900 ms, resulting in 900 binary outcomes. We examine data from 19 trials, with stimuli at different orientations.

Data from two particular neurons are shown in Figure 2.5. On the time axes, 0 ms corresponds to the appearance of the stimulus and 500 ms to the removal of the stimulus. The raster plot shows the spike trains for each of the trials on separate lines. The peristimulus time histogram (PSTH) displays the number of spikes per second occurring within 10 ms bins, averaged over all trials. The solid curve is a smoothed version of the counts obtained with a kernel density estimator. After the appearance of the stimulus, the intensity of spikes increases rapidly up to some maximum, and then remains high. After removal of the stimulus, the intensity of spikes gradually drops back. The intensity of spikes in the first neuron is higher than in the second neuron.

In Chapter 8, we introduce a measure describing the 'synchrony' of two spike trains. Focus is on the specification of a flexible marginal model for multivariate correlated binary data together with a pseudo-likelihood estimation approach (Arnold and Strauss 1988), to adequately and directly describe the measures of interest.

2.3   Bovine Herpesvirus-1 Data

Bovine herpesvirus-1 (BHV) is a worldwide-distributed infection in cattle. To investigate BHV-1 prevalences in the Belgian cattle population, a survey was conducted from December 1997 to March 1998 in all provinces of Belgium. In each province, 1% of the total number of herds was randomly sampled, yielding a study sample of 309 unvaccinated herds. The number of cattle per herd varied from 1 to 294. In the selected herds, all animals were blood sampled. It was investigated whether the animal had, or had not, antibodies towards the gB antigen of the bovine herpesvirus-1. The age, sex and purchase (purchased or homebred) of the animal were recorded. Also the herdtype (dairy, mixed or beef) and size of herds were registered. As such, a total of 11284 cattle were investigated. An overview of the variables in the study is given
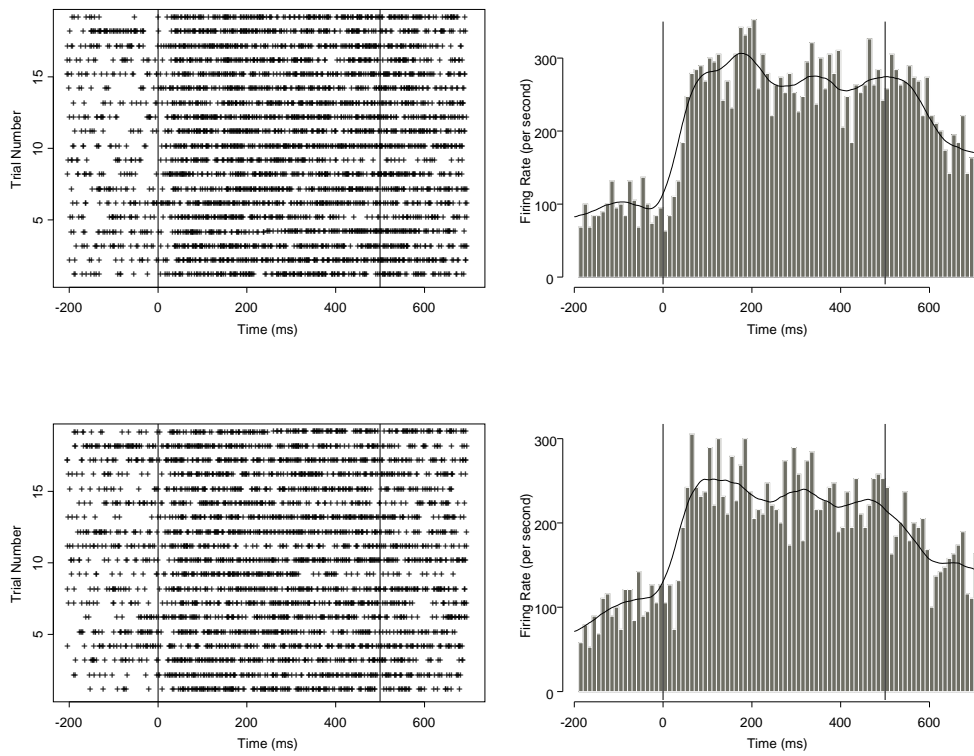
Figure 2.5: *Electrophysiological study. Raster plot (left) of spikes and corresponding peristimulus time histogram (right) for neuron 1 (top) and neuron 2 (bottom).*

Table 2.4: Bovine Herpesvirus-1 study. Overview of the different variables in the data.

Variable	Description
gb	infected or not
herd	identification umber of the herd
animal	identification number of the animal
province	province (Belgium has 10 provinces)
herdtype	dairy, mixed or beef
herdsize	size of the herd (number of cattle)
densanim	density of cattle population (number of cattle/km^2)
densherd	density of herds in the municipalities (number of herds/km^2))
age	age of the animal (in months)
sex	gender of the animal
purchase	purchased or homebred

Figure 2.6: *Bovine Herpesvirus-1 study. Prevalence of gB as a function of age (months).*

in Table 2.4.

In Figure 2.6 the prevalence of gB-antibodies against the age of the animal is displayed. There is a clear increase of the prevalence of the virus with age. In the analysis, only animals older than 6 months are considered, since high sero-prevalence in young animals are not necessarily due to high prevalence of the bovine herpesvirus-1. Animals less than 6 months old typically have high sero-prevalence of gB-antibodies because the maternal antibodies are hereditable. It is worthwhile to investigate this relationship in future research. We refer to Boelaert *et al.* (2000) and Speybroeck *et al.* (2003) for details about the study design.

In Chapter 9 we present methods to derive cluster-specific and population averaged force of infection estimates, while accounting for the complex study design. Also the prevalence of the bovine herpesvirus-1 is investigated using flexible predictor models such as fractional polynomials and penalized splines.

2.4   Isolated Cat Duodenum Data

The small intestine finishes the process of digestion, absorbs the nutrients, and passes the residue on to the large intestine. It is the longest section of the digestive tube. In most animals, the length of the small intestine is roughly 3.5 times body length. The small intestine of a cat is about 1 to 1.5 meters in length. It is divided into the duodenum, jejunum and ileum. The duodenum is the beginning portion of the small intestine, starting at the lower end of the stomach and extending to the jejunum.

Coordinated contractions of smooth muscle participate in several ways to facilitate digestion and absorption of food in the small intestine. This is controlled predominantly by signals from the enteric nervous system. Two basic patterns of electrical activity are important: the slow waves and spike potentials. Slow waves in the intestines are the basic electrical rhythm of the gut, sweeping along the digestive tube for long distances. The slow wave is thought to act as a pacemaker signal that induces the muscle layers to contract. Such slow waves occur 10 to 20 times per minute in the muscle of the small intestine. Action potentials superimposed on slow waves determine the force and duration of muscle contraction.

There is a considerable amount of information about the temporal and spatial organization of slow waves and action potentials (spikes) in the small intestine. In the spatial dimension, it is of interest to determine whether or not spikes are restricted to certain areas (i.e., spike patches). In addition, there is no information as to the behaviour of spikes in successive slow waves. If spikes had occurred during a particular
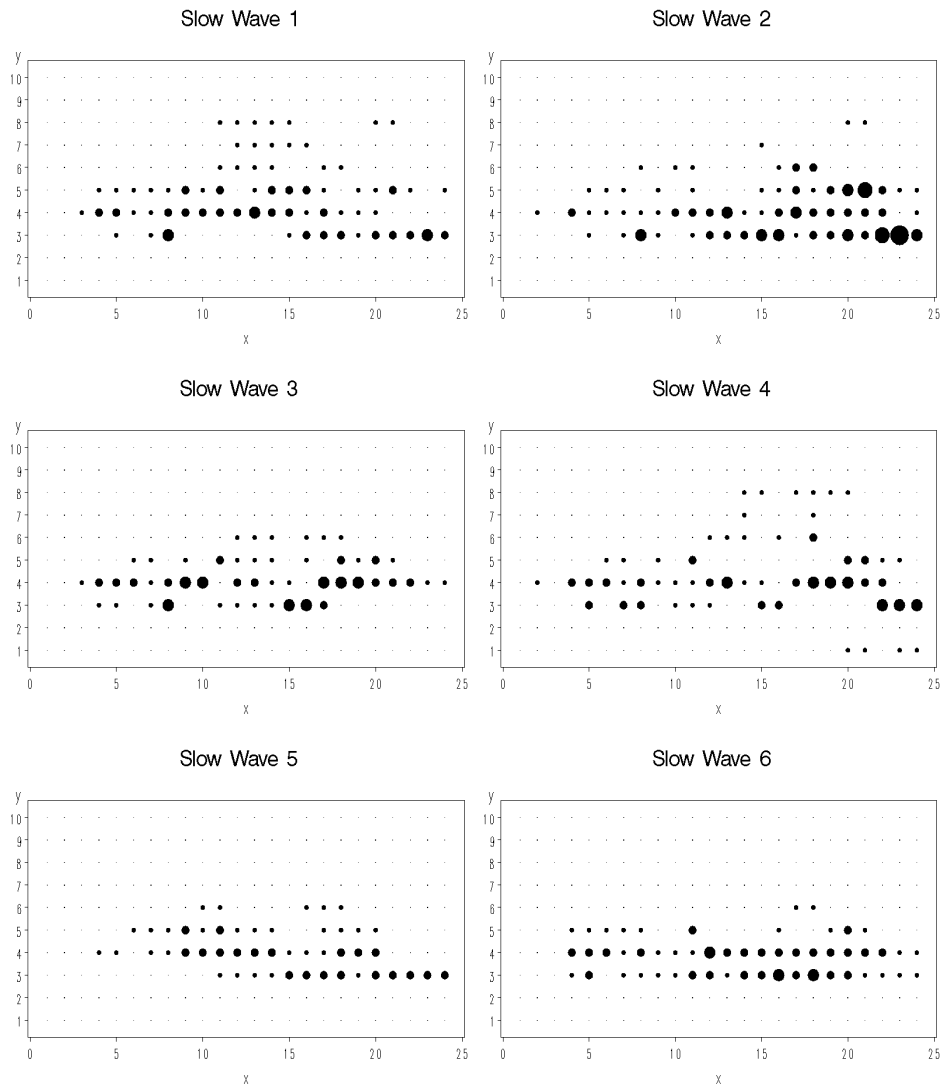
Figure 2.7: *A look at the number of spikes on a grid in the intestine of a cat, during six successive slow waves.*

slow wave, what are then the temporal and spatial characteristics of spikes and spike patches during the next slow waves? Understanding these patterns can help to clarify the mechanisms involved in their generation and propagation, and also the processes that underly the different types of small intestine movement.

In this study, a segment of the proximal small intestine from seven mongrel cats had been removed and transferred to a tissue bath. Spontaneous electrical activities on the surface were recorded. Simultaneous recordings were performed during one-minute periods using 240 extracellular electrodes (24×10 array; 2 mm interelectrode distance) positioned onto the serosal surface. The locations of successive spike patches occurring during 12 to 16 successive slow waves were recorded. Figure 2.7 shows the data for the grid of a cat at 6 successive slow waves. The size of the dots represents the number of spikes at that location. The number of spikes at a certain location varies between 0 and 4. Details of the experiment can be found in Lammers *et al.* (1996, 2000). In Chapter 10, the spatial and temporal characteristics of the spikes are investigated.

Chapter 3

Issues in Flexible Modelling

Each of the datasets described in the previous chapter deals with the analysis of correlated categorical data. In this chapter we present a toolbox of methods known in literature, which are used throughout this thesis.

In general, to build a flexible model for a complex data setting, some choices have to be made concerning the probability model, the estimation method and the predictor model. We will discuss each of them in turn.

- *Probability Model.* The probability model should reflect the nature of the data, i.e., the type of outcome, the data structure, etc. Different types of modelling approaches (marginal, conditional, random effects models) are available. The choice between the different model families should mainly depend on the research questions one is willing to answer. However, since each model family requires its own analysis and software tools, also the availability of computational algorithms is important in making the appropriate choice.

- *Estimation Method.* Based on the model formulation, the appropriate estimation method should be chosen. Estimation methods range from full likelihood to pseudo-likelihood, quasi-likelihood and generalized estimating equations. Likelihood methods enjoy many desirable properties, such as efficiency under appropriate regularity conditions and the ability to calculate functions of interest based on the proposed parametric model (Edwards 1971). However, not only the specification of the likelihood function can be cumbersome, but also estimation of the parameters can be computationally intensive. In addition, fully specifying the joint probability model comes at risk of possible misspecifications. Therefore, alternative estimation methods can be desirable.

- *Predictor Model.* The predictor model describes the relationship between the model parameters and some explanatory variables. High order polynomials offer a wide range of curves, but often fit badly at the extremes. A flexible predictor model is needed to describe the functional form. Both parametric and more non-parametrically inspired approaches can provide flexibly shaped curves.

A flexible predictor model, in combination with the model families and different estimation methods, yields a broad framework to model clustered binary data. All three aspects of the model will contribute to the flexibility of the model. In this chapter, we present some methodologies for the analysis of clustered binary data in terms of the developmental toxicity studies. However, the methodology is applicable in general clustered binary data settings.

Consider an experiment involving N litters (pregnant dams), the ith of which contains n_i individuals, each of whom are examined for the presence or absence of a malformation. Suppose $Y_{ij} = 1$ indicates whether the jth individual in cluster i is abnormal, and 0 otherwise. Then, define $Z_i = \sum_{j=1}^{n_i} Y_{ij}$, the total number of malformations in cluster i. Covariates of interest are the treatment or dosing d_i given to cluster i. Further, we assume exchangeability within a litter.

3.1 Probability Model

A dose-response model describing developmental toxicity data must take the structure of the data into account. Interest goes to the risk of observing a malformation, binary coded as absent/present. Further, we must account for the litter effect induced by the clustering of offspring within litters. Different types of probability models are available for correlated binary data, and can be grouped into conditional, marginal and cluster-specific models. A thorough review is given in Pendergast *et al.* (1996) and Aerts *et al.* (2002). As examples of full likelihood methods for univariate clustered binary data, we selected the beta-binomial model (Skellam 1948, Williams 1975), which can be viewed as a marginalized random effects model, and the conditional exponential family model (MR) of Molenberghs and Ryan (1999).

3.1.1 A Random Effects Likelihood Model

The beta-binomial model assumes that, conditional on litter size n_i and malformation probability of any foetus in litter i, the number of malformations Z_i in the ith cluster follows a binomial distribution. To account for the litter effect, i.e., the cluster effect,

the underlying malformation probabilities are assumed to vary within a litter according to a beta distribution with mean π_i. This leads to the beta-binomial distribution of the number of malformations Z_i in cluster i, and its probability density function is given by

$$f(z_i; \pi_i, \rho_i, n_i) = \binom{n_i}{z_i} \frac{B\big(\pi_i(\rho_i^{-1} - 1) + z_i, (1 - \pi_i)(\rho_i^{-1} - 1) + n_i - z_i\big)}{B\big(\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1)\big)}, \qquad (3.1)$$

where $B(.,.)$ denotes the beta function (Skellam 1948, Kleinman 1973). The association parameter ρ_i in this model indicates the correlation between two binary responses of litter i. The moments of this distribution can then be expressed as

$$\begin{aligned} \mathrm{E}(Z_i) &= n_i \pi_i, \\ \mathrm{Var}(Z_i) &= n_i \pi_i (1 - \pi_i)\big[1 + \rho_i(n_i - 1)\big]. \end{aligned}$$

It can be shown that the contribution of the ith cluster to the log-likelihood is given by

$$\binom{n_i}{z_i} + \sum_{r=0}^{z_i-1} \ln\Big(\pi_i + \frac{r\rho_i}{1 - \rho_i}\Big) + \sum_{r=0}^{n_i-z_i-1} \ln\Big(1 - \pi_i + \frac{r\rho_i}{1 - \rho_i}\Big) - \sum_{r=0}^{n_i-1} \ln\Big(1 + \frac{r\rho_i}{1 - \rho_i}\Big),$$

with $i = 1, \ldots, N$. Note that this expression reduces to the familiar binomial log-likelihood when $\rho_i = 0$. Further, the parameters π_i and ρ_i of the beta-binomial model have a marginal interpretation. A disadvantage of the beta-binomial model is that it incorporates only positive (extrabinomial) correlation. Prentice (1986), however, points out that it is possible to allow for slightly negative values of the correlation parameter as well.

To model the marginal parameters π_i and ρ_i we use a composite link function. An appropriate choice is given by the following generalized linear regression relations

$$\begin{pmatrix} \ln\big(\frac{\pi_i}{1 - \pi_i}\big) \\ \ln\big(\frac{1 + \rho_i}{1 - \rho_i}\big) \end{pmatrix} \equiv \eta_i = \boldsymbol{X_i}\boldsymbol{\beta}, \qquad (3.2)$$

where X_i is a design matrix and β is a vector of unknown parameters. However, other link functions, such as the probit link, the log-log link or the complementary log-log link for the mean parameter, could be chosen too. A frequently used model in literature is

$$\boldsymbol{X_i} = \begin{pmatrix} 1 & d_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_d \\ \beta_a \end{pmatrix} \qquad (3.3)$$

with a logit-linear dose trend for the mean parameter, and a constant association parameter ρ. Obviously, this model can be extended by adapting the design matrix and the vector of regression parameters, such that the logit of π_i depends on dose via, e.g., a quadratic or higher order polynomial function. Also, the association parameter ρ_i can be modelled as some function of dose, using a suitable link function.

3.1.2   A Conditional Likelihood Model

Molenberghs and Ryan (1999) proposed a likelihood-based model for clustered binary data, based on a multivariate exponential family model (Cox 1972). The model describes the probability of an outcome given values for the other outcomes, and therefore is conditional in nature. Molenberghs and Ryan (MR) considered $Y_{ij} = 1$ if the jth foetus in cluster i exhibits the adverse event of interest, and -1 otherwise. This coding is preferred above the 1/0 coding, since it provides a parameterisation that more naturally leads to desirable properties when the roles of success and failure are reversed (Cox and Wermuth 1994). They proposed the distribution of z_i, the number of individuals from cluster i with positive response, as

$$f(z_i; \theta_i, \delta_i, n_i) = \exp\left\{\theta_i z_i - \delta_i z_i (n_i - z_i) - A(\theta_i, \delta_i, n_i)\right\}, \qquad (3.4)$$

with θ_i the main parameter, δ_i the association parameter describing the association between pairs of individuals within the ith cluster and $A(\theta_i, \delta_i, n_i)$ the normalizing constant. The parameters θ_i and δ_i can be modelled as $(\theta_i, \delta_i)' = \boldsymbol{X}_i \boldsymbol{\beta}$ with $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ as in (3.3). Note that, as with the beta-binomial model, this model reduces to the logistic regression model in the absence of clustering.

Due to the popularity of marginal and random-effects models for correlated binary data, the conditional models have received little attention, especially in the context of multivariate clustered data. Molenberghs, Declerck and Aerts (1998) and Aerts, Declerck and Molenberghs (1997) have compared marginal, conditional and random-effects models. Their results are encouraging for the conditional models, since they are competitive for the dose effect testing and for benchmark dose estimation, and because they are computationally fast and stable. In addition, the conditional model provides a natural framework for quantitative risk assessment. Indeed, from a biological perspective, one might argue that it is important to take into account the health of the entire litter when modelling risk as a function of dose. The conditional model allows direct calculation of quantities such as the probability that at least one littermate is affected, i.e., the probability of an affected litter. The MR model benefits from the elegance and simplicity of exponential family theory. More details about

model properties and inference can be found in Molenberghs and Ryan (1999).

3.2   Estimation Method

What method is used to fit the data depends on the assumptions one is willing to make. However, also the availability of computational algorithms needs to be incorporated. If one is willing to fully specify the joint probabilities, maximum likelihood methods can be adopted. In some situations however, full likelihood becomes very complex and computationally demanding, especially with large within-unit replications. As a consequence, alternative methods have been in demand and one has to rely on non-likelihood methods such as generalized estimating equations (Liang and Zeger 1986) or pseudo-likelihood methods (Arnold and Strauss 1991), which are easier and much less time consuming. In addition, since these methods allow certain aspects of the model to be unspecified, and hence not misspecified, often at a reasonable low cost of efficiency, they are more robust in nature.

3.2.1   Generalized Estimating Equations

Liang and Zeger (1986) and Zeger and Liang (1986) introduced generalized estimating equations (GEEs) to account for the correlation between observations in generalized linear regression models. Instead of specifying the full distribution, generalized estimating equations require only the correct specification of the univariate marginal distributions and a covariance structure of the correlated observations on a given subject.

Definition

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ be the vector of binary outcomes for subject i ($i = 1, \ldots, N$). Suppose that the mean and variance of each observation is thought to satisfy a generalized linear model

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{Y}_i) &= \boldsymbol{\mu}_i = g(\boldsymbol{X}_i \boldsymbol{\beta}), \\
\mathrm{Var}(\boldsymbol{Y}_i) &= v(\boldsymbol{\mu}_i),
\end{aligned}
$$

but that the observations within a subject are correlated. In above equations $\boldsymbol{X}_i$ is a design matrix, $\boldsymbol{\beta}$ is a vector of unknown parameters, g is a link function and $v(\boldsymbol{\mu}_i)$ is a known function of the marginal mean. Liang and Zeger (1986) proposed generalized

estimating equations of the form

$$S(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} (A_i^{1/2} R_i A_i^{1/2})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \;=\; 0, \qquad (3.5)$$

where A_i is a diagonal matrix with the marginal variances $v(\boldsymbol{\mu}_i)$ on the main diagonal and R_i a correlation matrix of $\boldsymbol{Y}_i$, often referred to as the 'working' correlation matrix. Usually, the marginal covariance matrix $V_i = A_i^{1/2} R_i A_i^{1/2}$ is a function of the pairwise association parameters $\boldsymbol{\alpha}$. There are a variety of possible working correlation structures. Some of the more popular choices are:

- *Independence*: The simplest choice is the working independence model, i.e.,

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = 0 \qquad (j \neq k).$$

- *Exchangeable*: For datasets with clustered observations such as in developmental toxicity studies, there may be no logical ordering for the observations within a cluster, and an exchangeable correlation structure (or equicorrelated structure) may be most appropriate:

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \alpha \qquad (j \neq k).$$

- *Unstructured*: A totally unspecified correlation matrix is given by

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk} \qquad (j \neq k).$$

As such, Liang and Zeger estimate the parameters associated with the marginal expected value of an individual's vector of binary responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations. There are two approaches to estimation, called GEE1 and GEE2 respectively. In the GEE1 approach the method combines estimating equations for the regression parameters $\boldsymbol{\beta}$ with moment-based estimation for the correlation parameters entering the working assumption. In the GEE2 approach, the problem is reformulated as a regression problem involving both the first two moments.

Consistency and Asymptotic Normality of the GEE Estimator

Assuming that the marginal mean $\boldsymbol{\mu}_i$ has been correctly specified as $h(\boldsymbol{\mu}_i) = \boldsymbol{X}_i \boldsymbol{\beta}$, then (under mild regularity conditions):

- the GEE estimator $\widehat{\boldsymbol{\beta}}$, obtained from solving (3.5) converges in probability to the true parameter $\boldsymbol{\beta}_0$

- the GEE estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and with covariance matrix $I_0^{-1} I_1 I_0^{-1}$, where

$$I_0 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right), \tag{3.6}$$

$$I_1 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \mathrm{Var}(\boldsymbol{y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right). \tag{3.7}$$

In practice, $\mathrm{Var}(\boldsymbol{y}_i)$ in (3.7) is replaced by $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'$, which is unbiased on the sole condition that the mean was again correctly specified. Valid inferences can now be obtained for the mean structure, only assuming that the model assumptions with respect to the first-order moments are correct. The model based estimator of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ is given by I_0^{-1}. The empirically corrected variance estimator (Liang and Zeger 1986) takes the form $I_0^{-1} I_1 I_0^{-1}$. Note also that, although arising from a likelihood approach, the GEE equations in (3.5) cannot be interpreted as score equations corresponding to some full likelihood for the data vector $\boldsymbol{y}_i$.

Despite the robustness and flexibility of generalized estimating equations, there are some situations where care is needed. When dealing with missing observations in the data, GEE might be no longer valid. Only if the missingness can be thought of as being completely at random in the sense of Little and Rubin (1987), then the consistency results established by Liang and Zeger (1986) hold. When data are missing at random, one can reduce bias in the parameter estimates by the use of a weighted GEE (Robins *et al.* 1995, Fitzmaurice *et al.* 1995). A closely related problem occurs when the cluster size is related with the response. The GEE method implicitly assumes that the size of the cluster is unrelated to the parameters under study, i.e., is noninformative (Hoffman *et al.* 2001). However, in many applications this may not be true. This problem is discussed in Chapter 9.

3.2.2 Pseudo-Likelihood

The principal idea of the pseudo-likelihood methodology is to replace the numerically challenging joint density by a simpler function. The method is well described by Arnold and Strauss (1991), and also found in Conolly and Liang (1988), Liang and Zeger (1986) and Le Cessie and Van Houwelingen (1994). While the method achieves important computational economies by changing the method of estimation, it provides consistent and asymptotical normal estimates (Arnold and Strauss 1991) and it does not affect model interpretation. An overview of the pseudo-likelihood methodology is given in this section.

Definition

Let $\boldsymbol{Y}_i$ be the vector of binary outcomes for subject i ($i = 1, \ldots, N$). Without loss of generality we can assume that $\boldsymbol{Y}_i$ has constant dimension L. The extension to variable lengths of $\boldsymbol{Y}_i$ is straightforward. Define S as the set of all $2^L - 1$ vectors of length M, consisting solely of zeros and ones, with each vector having at least one nonzero entry. Denote by $\boldsymbol{y}_k^{(s)}$ the subvector of $\boldsymbol{y}_k$ corresponding to the components of $s \in S$ that are nonzero. The associated joint density is written as $f_s(\boldsymbol{y}_k^{(s)}; \boldsymbol{\Theta}_k)$, with $\boldsymbol{\Theta}_k = \boldsymbol{X}_k \boldsymbol{\beta}$. Specify a set $\delta = \{\delta_s | s \in S\}$ of $2^L - 1$ real numbers, with at least one nonzero component and define the log pseudo-likelihood as:

$$p\ell = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \ln f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\Theta}_i), \tag{3.8}$$

where some (thought not all) of the δ_s's may be negative. This must correspond to a product of marginal and conditional densities. Some examples of possible pseudo-likelihood functions are given:

- Classical maximum likelihood corresponds to $\delta_s = 1$ for $s = 1_L$ and zero otherwise, where 1_L is a vector of ones.

- Another typical choice is $\delta_{1_L} = l$ and $\delta_{s_k} = -1$ for $k = 1, \ldots, l$ where s_k consists of ones every where, except for the kth entry. This particular choice is referred to as the 'full conditional' pseudo-likelihood function. It has the effect of replacing the joint density function by a product of l univariate conditional density functions, thus avoiding the incorporation of a possibly complicated normalizing constant which typically arises in exponential family models.

- Another convenient pseudo-likelihood function is found by replacing the joint density by the product of all pairwise likelihoods. This idea can be put into the pseudo-likelihood framework by choosing $\delta_{s_{k,l}}$ as a vector consisting of zeros everywhere, except for the entries k and l, for each pair $k \neq l$. For all other vectors s, δ_s equals zero. We refer to this choice as the pairwise pseudo-likelihood.

The pseudo-likelihood methodology is very general and flexible. It can be found in many applications and fields of interest. It has been most advantageously used in the spatial data context, where the full likelihood distribution is typically cumbersome (Hjort 1993, Guyon 1995). But also in the context where maximum likelihood methods are not feasible, e.g. due to excessive computational requirements, the pseudo-likelihood is an appealing methodology.

Consistency and Asymptotic Normality of the Pseudo-Likelihood Estimator

Let $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ be independent and identically distributed with common density with parameter $\boldsymbol{\Theta}_0$, then (under regularity conditions):

- the pseudo-likelihood estimator $\widehat{\boldsymbol{\Theta}}_N$, defined as the maximizer of (3.8), converges in probability to $\boldsymbol{\Theta}_0$.

- $\sqrt{N}(\widehat{\boldsymbol{\Theta}}_N - \boldsymbol{\Theta}_0)$ converges in distribution to $N_p\big(0, J(\boldsymbol{\Theta}_0)^{-1} K(\boldsymbol{\Theta}_0) J(\boldsymbol{\Theta}_0)^{-1}\big)$ with $J = J(\boldsymbol{\Theta}_0)$ defined by

$$J_{kl} = -\sum_{s \in S} \delta_s E_{\boldsymbol{\Theta}} \Big(\frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial \theta_k \theta_l} \Big)$$

and $K = K(\boldsymbol{\Theta}_0)$ by

$$K_{kl} = \sum_{s,t \in S} \delta_s \delta_t E_{\boldsymbol{\Theta}} \Big(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial \theta_k} \frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial \theta_l} \Big)$$

While the idea of the GEE approach is to simplify the estimation method by replacing the complicated first derivatives of the log-likelihood function, the pseudo-likelihood approach replaces the complicated log-likelihood function directly. Geys, Molenberghs and Lipsitz (1998) compared pairwise pseudo-likelihood with other estimating equations approaches (GEE1 and GEE2) in marginally specified odds ratio models with exchangeable association structure. The efficiency of the pseudo-likelihood estimators for the main effects is comparable to the efficiency of GEE estimators. Yet, pseudo-likelihood allows the estimation of both main effect parameters and association parameters, whereas GEE1 is restricted to main effect parameters. Even though GEE1 produces estimates of the working correlation structure, they should be handled with caution. Indeed, the strength of GEE1 is that the working correlation structure can be misspecified, without jeopardizing inferences about the mean structure. However, the consequence of this asset is that no formal inferences should be undertaken about the working correlation structure, and even interpretation ought to be done carefully, unless there is enough evidence that the working structure is sufficiently trustworthy. While GEE2 includes second order association parameters as well and is slightly more efficient than both GEE1 and PL, it is computationally much more complex and becomes cumbersome for large cluster sizes. In contrast, pseudo-likelihood can be used with very large clusters.

Pseudo-Likelihood Ratio Test Statistic

An important advantage of the pseudo-likelihood approach is the close connection with likelihood, which enabled Geys, Molenberghs and Ryan (1999) to construct pseudo-likelihood ratio test statistics that have easy-to-compute expressions and intuitively appealing limiting distributions.

Extensions of the Wald, score or likelihood ratio test statistics to the pseudo-likelihood framework were proposed by Geys, Molenberghs and Ryan (1999). We restrict attention to the pseudo-likelihood ratio test statistic. Suppose we are interested in testing the null hypothesis $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, where γ is an r-dimensional subvector of the p-dimensional vector of regression parameters $\boldsymbol{\beta}$ and write $\boldsymbol{\beta}$ as $(\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$. Then, the pseudo-likelihood ratio test statistic, defined by

$$G_a^{*2} = 2\big[p\ell(\widehat{\boldsymbol{\beta}}_N) - p\ell(\boldsymbol{\gamma}_0, \widehat{\boldsymbol{\delta}}(\boldsymbol{\gamma}_0))\big]/\bar{\boldsymbol{\lambda}},$$

is approximately χ_r^2 distributed. In this definition, $\widehat{\boldsymbol{\beta}}_N$ is the pseudo-likelihood parameter estimate of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\delta}}(\boldsymbol{\gamma}_0)$ denotes the maximum pseudo-likelihood estimator in the subspace where $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. Further, $\bar{\boldsymbol{\lambda}}$ is the mean of the eigenvalues of $(J^{\gamma\gamma})^{-1}\Sigma_{\boldsymbol{\gamma\gamma}}$, where $J^{\gamma\gamma}$ is the $r \times r$ submatrix of the inverse of J and $\Sigma_{\boldsymbol{\gamma\gamma}}$ is the submatrix of $\Sigma = J^{-1}KJ^{-1}$.

3.2.3   Bayesian Method

When random-effects models are used, the likelihood function involves the integration over the random-effects distribution. In case responses are normally distributed, the marginal likelihood can be derived analytically. However, this property of normal models does not extend to the case of non-normal distributions, where in general, no closed forms are available. Estimation methods then either employ an approximation of the integrand or uses numerical integration techniques. The need for complex numerical integration can be avoided by casting the random-effects model into a Bayesian framework, and resort to the Gibbs sampler (Zeger and Karim 1991). Gibbs sampling offers a Monte Carlo approach.

Definition Gibbs Sampler

Within the Bayesian framework, the unknown parameters are estimated by the posterior mean. In order to obtain such estimates, we can approximate the full conditional distributions using Markov Chain Monte Carlo (MCMC) methods (Gilks *et al.* 1996) and generate samples from the full conditional distributions using the Gibbs sampler. The sample averages are taken as the posterior means of the parameters of interest.

Suppose that we partition $\boldsymbol{\beta}$ into r subvectors $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_r)$, where for each k it is easy to sample from the full conditional distribution $p(\beta_k | \beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_r, \boldsymbol{y})$ with $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$. Iteration i of the Gibbs sampler starts with $\boldsymbol{\beta}^{(i)} = (\beta_1^{(i)}, \beta_2^{(i)}, \ldots, \beta_r^{(i)})$ and makes the transition to $\boldsymbol{\beta}^{(i+1)}$ via the following scheme:

Sample $\beta_1^{(i+1)}$ from $p(\beta_1 | \beta_2^{(i)}, \beta_3^{(i)}, \ldots, \beta_r^{(i)}, \boldsymbol{y})$

Sample $\beta_2^{(i+1)}$ from $p(\beta_2 | \beta_1^{(i+1)}, \beta_3^{(i)}, \ldots, \beta_r^{(i)}, \boldsymbol{y})$

$\vdots$

Sample $\beta_r^{(i+1)}$ from $p(\beta_r | \beta_1^{(i+1)}, \beta_2^{(i+1)}, \ldots, \beta_{r-1}^{(i+1)}, \boldsymbol{y})$

The process of updating each of the r blocks, as indicated above, produces one complete iteration of the Gibbs sampler. The sequence of vector produced by this scheme, $\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(t)}, \ldots$, are a Markov chain. Under weak assumptions, this chain converges in distribution to a stationary distribution that is the posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$ (Tierney 1994).

Deviance Information Criterion

A model selection procedure is needed in order to compare between models and to select the best fitting model. Goodness-of-fit and complexity of the models can be assessed using the deviance information criterion (DIC) as proposed by Spiegelhalter *et al.* (1998, 2002) and recently used by Erkanli *et al.* (2000), Rahmann *et al.* (1999) and Gelfand *et al.* (2000) for model selection within the Bayesian framework.

The deviance is defined as the posterior distribution of the log likelihood:

$$D = -2\log(p(\boldsymbol{y}|\boldsymbol{\beta})) + 2\log f(\boldsymbol{y}),$$

with $f(\boldsymbol{y})$ a standardizing term that does not affect model comparison. The goodness of fit of the model is then summarized by the posterior expectation of the deviance:

$$E_{\boldsymbol{\beta}|\boldsymbol{y}}[D].$$

Spiegelhalter *et al.* (1998, 2002) suggested to measure the complexity of the model by the difference between the posterior expectation of the deviance and the deviance evaluated at the posterior expectation of $\boldsymbol{\beta}$, that is

$$
\begin{aligned}
p_D &= E_{\boldsymbol{\beta}|\boldsymbol{y}}[D] - D(E_{\boldsymbol{\beta}|\boldsymbol{y}}[\boldsymbol{\beta}]) & (3.9) \\
&= \bar{D} - \hat{D}, & (3.10)
\end{aligned}
$$

where p_D can be interpreted as the effective number of parameters in the model. These are combined to give the overall DIC:

$$\text{DIC} = \bar{D} + p_D,$$

where the first term represents the goodness of fit and the second term represents the model complexity (the effective number of parameters). Smaller values of DIC indicate a better fitting model. The attraction of using this measure is that it is trivial to compute when performing MCMC on the model. All that needs to be done is to take samples of the deviance $D(\beta)$ along with samples of β and the terms can be calculated to give the DIC.

3.3   Predictor Model

Parametric statistical models are widely used to describe the relationship between a response variable and several factors or explanatory variables. Also clustered binary data have been analysed mainly in a parametric way. However, the selection of the proper functional forms describing the dependence of all main and association parameters in a specific probability model is not always an easy task. In general, parametric models assume that certain model assumptions hold. Because of these assumptions, parametric models are always prone to possible misspecification. Therefore, there is a clear need for flexible parametric models and, in case the design allows, for semi- and nonparametric approaches.

3.3.1   Fractional Polynomials

Although classical polynomial predictors are still very customary, they are often in-adequate. A very elegant alternative approach to classical polynomials, which falls within the realm of (generalized) linear methods, is given by fractional polynomials. They provide a much wider range of functional forms. Let us briefly describe this procedure, advocated by Royston and Altman (1994).

For a given degree m and an argument $d > 0$ (e.g., dose), fractional polynomials are defined as

$$\eta_m(d; \boldsymbol{\beta}, \boldsymbol{p}) = \beta_0 + \sum_{j=1}^{m} \beta_j d^{p_j}, \qquad (3.11)$$

where the $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ are regression parameters and $d^0 \equiv \ln(d)$ and the powers $p_1 < \ldots < p_m$ are positive or negative integers or fractions. The full definition includes possible "repeated powers" which involve powers of $\ln(d)$. For example, a

fractional polynomial of degree $m = 3$ with powers (-1,-1,2) is of the form $\beta_0 + \beta_1 d^{-1} + \beta_2 d^{-1} \ln(d) + \beta_3 d^2$ (Royston and Altman 1994, Sauerbrei and Royston 1999). Royston and Altman (1994) argue that polynomials with degree higher than 2 are rarely required in practice and further restrict the powers of d to a small predefined set of possibly non-integer values: $\Pi = \{-2, -1, -1/2, 0, 1/2, 1, 2, \ldots, \max(3, m)\}$. For example, setting $m = 2$ generates:

(1) 4 "quadratics" in powers of d, represented by

- $(1/d, 1/d^2) : \beta_0 + \beta_1 1/d + \beta_2 1/d^2,$

- $(1/\sqrt{d}, 1/d) : \beta_0 + \beta_1 1/\sqrt{d} + \beta_2 1/d,$

- $(\sqrt{d}, d) : \beta_0 + \beta_1 \sqrt{d} + \beta_2 d,$

- $(d, d^2) : \beta_0 + \beta_1 d + \beta_2 d^2,$

(2) a quadratic in $\ln(d)$: $\beta_0 + \beta_1 \ln(d) + \beta_2 \ln(d)^2$, and

(3) other curves which have shapes different from those of conventional low degree polynomials.

For given m, we consider as the best set of transformations, the one producing the highest log (pseudo)-likelihood. For example, the best first degree fractional polynomial is the one with the highest log (pseudo)-likelihood among the eight models with one regressor $(d^{-2}, d^{-1}, \ldots, d^3)$. As with conventional polynomials, the degree m is selected either informally on *a priori* grounds or by increasing m until no worthwhile improvement in the fit of the best fitting fractional polynomial occurs. In the above discussion, it is assumed that d is strictly positive. If d can take zero values, a preliminary transformation of x is needed to ensure positivity (e.g., $d + 1$).

A particular feature of the fractional polynomials is that they provide a wide class of functional forms, with only a small number of terms (Royston and Altman 1994, Sauerbrei and Royston 1999). Even with one predictor they provide a rich class of possible forms, in many cases leading to a reasonable fit to the data. Moreover, the conventional polynomials are included as a subset of this extended family.

In Chapter 4, the fractional polynomials are the topic of interest, where they are used for dose-response modelling as a basis for quantitative risk assessment. This chapter illustrates how fractional polynomials offer great flexibility in modelling clustered data.

3.3.2 Power Models

Fractional polynomial models are linear in the parameters. Next to fractional polyno-
mial models, there are other interesting parametric approaches. One such technique
is the nonlinear power model (Cox and Hinkley 1978, Davidan and Giltinan 1995):

$$\eta(d; \boldsymbol{\beta}) = \beta_0 + \beta_1 d^{\beta_2}, \tag{3.12}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ the unknown regression coefficients. Such nonlinear models form
another interesting family of models, offering a lot of flexibility. Notice that by setting
$\beta_2 \equiv 1$, this model simplifies to a linear model.

The use of power models invokes some interesting statistical issues. First, the
effect of dose d can be investigated via testing the null hypothesis $H_0 : \beta_1 \beta_2 = 0$. If
the dose effect is absent, then the regression parameters of the power model under
consideration are non-identifiable. Secondly, fitting models with power predictors
might be complicated if the dose effect is weak, since convergence problems can be
expected in that case.

In Chapter 5 the power model is used to test for trend in the context of develop-
mental toxicity studies.

3.3.3 Penalized Splines

The previously defined models are flexible but essentially parametric and user-defined.
More nonparametrically inspired approaches, fully data driven, can also be very useful.
In an explorative way, a parametric model can be graphically compared with its
nonparametric alternative. A relatively straightforward extension of linear models is
the penalized splines model.

A piecewise linear smoother proposed by Freedman and Silverman (1989) has the
form

$$\eta(d_i; \boldsymbol{\beta}) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(d_i), \tag{3.13}$$

where d_i are the design points, and $\phi_k(.), k = 1, \ldots, K$, are known functions. Note
that for $K = 1$ and $\phi_1(d) = d$ the model reduces to a simple linear regression model.
The piecewise linear model assumes that the basis function $\phi_k(d)$ has the form

$$\phi_k(d) = (d - \kappa_k)_+ = \begin{cases} 0, & d \leq \kappa_k \\ d - \kappa_k, & d > \kappa_k. \end{cases} \tag{3.14}$$

The linear piecewise model consists of K knots where κ_k, $k = 1, 2 \ldots, K$ is the location of the kth knot. The basis function represents a broken line with the knots κ_k as a joint point. Note that for $\kappa_1 = 0$ the piecewise linear model can be written as

$$\eta(d_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 d_i + \sum_{k=2}^{K} \beta_k (d_i - \kappa_k)_+. \tag{3.15}$$

Next, define two design matrices, an $n \times 2$ design matrix for which the ith row is $\boldsymbol{X}_i = [1, d_i]$ and an $n \times (K - 1)$ matrix for which the ith row is $Z_i = [(d_i - \kappa_2)_+, \ldots, (d_i - \kappa_K)_+]$. Then, the model in (3.15) can be rewritten as

$$\eta(d_i; \boldsymbol{\beta}, \boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \tag{3.16}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{b} = (\beta_2, \ldots, \beta_K)$. When using a large set of knots, this model has much flexibility. However, when fitting the model using ordinary least squares, this method might overfit the data, following random fluctuations in the data as well as the main features. To overcome this problem, the spline function (3.16) can be fit as a mixed model (Ruppert, Wand and Carroll 2003), with normally distributed random effects $\boldsymbol{b} \sim N(0, \sigma_b^2 I_{(K-1) \times (K-1)})$. Ordinary least squares correspond to $\sigma_b = \infty$, where the b_k are unrestricted. When σ_b is finite, this tends to shrink the b_k with a smooth fit as a result. Note that in this model the unknown smooth function is modelled with two components. The linear part, which is the fixed effects $\boldsymbol{X}\boldsymbol{\beta}$, and the smooth part $\boldsymbol{Z}\boldsymbol{b}$, being the random part. The conditional mean of $\boldsymbol{y}$, given the fixed and the random effects, is $E(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$.

The representation of the penalizes spline as a mixed model, with ML or REML used to select the amount of smoothing, is very appealing since it allows smoothing to be done using mixed model methodology and software. Moreover, they allow a Bayesian way of inference, which makes it also very attractive to be used in this setting.

In Chapter 10 a general design of mixed models is investigated, with penalized splines as one building block of the model.

Chapter 4

Use of Fractional Polynomials for Quantitative Risk Assessment in Developmental Toxicity Studies

Developmental toxicity studies are designed to assess the potential adverse effects of an exposure on developing foetuses. One major objective in the risk assessment process is to determine a safe dose level of exposure, i.e., quantitative risk assessment. Safe dose levels can be determined using dose-response modelling. However, dose-response modelling is complicated by the hierarchical, clustered and multivariate nature of the data. As a consequence, a multitude of modelling strategies have been proposed in literature, as demonstrated in previous chapter. Such choices are often subjective and can affect the quantitative risk assessment based on the fitted models. Therefore it is important to study the possible effects of misspecifying the dose-response model on quantitative risk assessment.

In this chapter, it is shown how important it is to select an adequate predictor model in the context of quantitative risk assessment. Section 4.1 describes a possible method for safe dose determination in developmental toxicity studies, based on the dose-response model. Since quantitative risk assessment is based on the fitted dose-response relationship, the model should fit the data well. This has implications for both the model family, as well as for the form of the predictor model (Section 4.2).

Classical polynomial predictors are often of poor quality, thus there is a clear need for alternative specifications of the predictors. Here, focus is on fractional polynomials (Royston and Altman 1994), which provide a wide class of functional forms. In the remaining sections of this chapter, we discuss the use of fractional polynomials to obtain a proper estimation of the safe dose level. In particular, Sections 4.4 and 4.5 show, by means of simulations, that fractional polynomial predictors may resolve possible model misspecifications and may thus yield more reliable estimates of the benchmark dose. Focus is on clustered binary data, concentrating on the malformation outcome in a typical developmental toxicity study. The content of this chapter is mainly based on the paper of Faes *et al.* (2003a).

4.1   Quantitative Risk Assessment

An important issue in developmental toxicity is the risk assessment. Risk assessment can be defined as "the use of available information to evaluate and estimate exposure to a substance and its consequent adverse health effects" (Roberts and Abernathy 1996), and thus deals with safety issues and regulation of exposures with potential adverse effects. An important goal in this risk assessment process is the determination of a safe level of exposure, i.e., quantitative risk assessment (QRA).

Different approaches to estimate a safe dose are used in literature. Quantitative risk assessment can be performed via the "No Observable Adverse Effect Level" (NOAEL) approach, which is the dose immediately below that deemed statistically or biologically significant when compared with controls. This methodology assumes that the substance induces toxicity only after a particular dose level is exceeded. However, many substances appear to have no threshold of effect, meaning that any exposure can cause an adverse effect. In addition, the NOAEL has been criticized for its poor statistical properties (Leisenring and Ryan 1992, Williams and Ryan 1996). The estimation of the NOAEL depends on the design of the experiment, on the sample size and on the number of dose groups, and it does not allow calculating a measure of variability of the estimation.

An alternative strategy is to base the QRA on a fitted dose-response model (Crump 1984). The standard approach to quantitative risk assessment based on dose-response modelling requires the specification of an adverse event, along with its risk expressed as a function of dose. For developmental toxicity studies where offspring are clustered within litters, there are several ways to define the concept of an adverse effect. First, one can state that an adverse effect has occurred if a particular offspring is abnormal

(foetus based). Alternatively, one might conclude that an adverse effect has occurred if at least one offspring from the litter is affected (litter based). From a biological perspective one might argue that it is important to take into account the health of the entire litter when modelling risk as a function of dose (Declerck *et al.* 2000). In this chapter, we focus on the risk function $r(d)$ representing the probability of observing a malformation at dose level d for at least one foetus within the litter, i.e., the litter based approach. Based on this probability, a common measure for the excess risk over background is given by

$$r^*(d) = \frac{r(d) - r(0)}{1 - r(0)}, \tag{4.1}$$

where greater weights are given to outcomes with larger background risk. Assuming that the chemical results in more adverse effects at non-zero dose d compared to dose level 0, the excess risk ranges from 0 to 1. This definition of the excess risk measures the relative increase in risk above background. The benchmark dose is then defined as the dose corresponding to a very small increase in risk over background. More formally, the benchmark dose (BMD_q) is defined as the dose satisfying

$$r^*(d) = q, \tag{4.2}$$

where q corresponds to a pre-specified level of increased response and is typically specified as $0.01, 1, 5$ or 10% (Crump 1984). Of course, the use of dose-response models to set a safe limit of exposure is far more complicated than determining a NOAEL, but it offers a number of important advantages. It allows adding a measure of variability to the point estimation of a safe dose, it can incorporate special features of the structure of developmental toxicity studies, etc. (Williams and Ryan 1996). Because of the disadvantages of the NOAEL approach, and because of the benefits of basing quantitative risk assessment on dose-response modelling, the latter approach will be considered here.

Because the dose-response curve is estimated from the data and has inherent variability, the benchmark dose itself is only an estimate of the true dose that would result in the specified level of excess risk q. This sampling uncertainty for the model on which the benchmark dose is based can be acknowledged, by replacing the benchmark dose by a lower confidence limit. Several approaches exist (Williams and Ryan 1996, Kimmel and Gaylor 1988, Crump and Howe 1983). A well-known approach is the use of the lower effective dose, where an upper limit for the risk function is used to determine a safe dose level. The lower effective dose (LED_q) is thus defined as the solution of

$$\widehat{r}^*(d) + 1.645\sqrt{\widehat{\text{Var}}\big(\widehat{r}^*(d)\big)} = q, \tag{4.3}$$

where q corresponds with the pre-specified level of increased response. Assume that $\boldsymbol{\beta}$ is the vector of parameters included in the dose-response model, then the variance of the estimated increased risk function $\widehat{r}^*(d)$ is estimated as

$$\widehat{\mathrm{Var}}\big(\widehat{r}^*(d)\big) = \Big(\frac{\partial r^*(d)}{\partial \boldsymbol{\beta}}\Big)^T \widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}})\Big(\frac{\partial r^*(d)}{\partial \boldsymbol{\beta}}\Big)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \qquad (4.4)$$

with $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}})$ the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$.

Of course, to get trustworthy results, models should fit the data well in all respects. Therefore, it is important to investigate the effect of misspecifications of the dose-response relationship on the estimation of a safe level of exposure.

4.2 Misspecification of the Dose-Response Model

When performing risk assessment based on the fitted dose-response model, the model should fit the data well. This has implications for both the model family chosen (the probability model), as well as for the form of the predictors. While the probability model can take special features of the data into account, the predictor model must take care of the flexibility of the model.

4.2.1 Probability Model

Developmental toxicity studies in rodents are faced with the fairly involved data structure of clustered binary outcomes. As a consequence, it is often not clear which probability model is generating the particular data. Different types of models (marginal, conditional, random effects models) are available to model clustered binary data (see Chapter 3). In this chapter we restrict attention to full likelihood models, and select the beta-binomial model, which can be viewed as a marginalized random effects model, and the conditional model (MR) of Molenberghs and Ryan (1999). We investigate, when using a misspecified probability model, the estimation of a safe level of exposure. Note that although models of quite different structure are being contrasted, the problem we are looking at really makes a lot of sense. It is exactly what happens in daily practice, since one (almost) never can be sure about the probability model generating the set of data at hand.

Both the beta-binomial model and the conditional model of Molenberghs and Ryan (1999) can easily handle litter-based risks, as shown below.

The Beta-Binomial Model

For the beta-binomial model, the probability that at least one foetus in a litter of size n_i is abnormal $(i = 1, \ldots, N)$, is

$$q(n_i; d) = 1 - \frac{B\big(\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1) + n_i\big)}{B\big(\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1)\big)}.$$

It can be shown that this expression equals

$$q(n_i; d) = 1 - \prod_{k=0}^{n_i - 1} \left(1 - \pi_i + \frac{k\pi_i\rho_i}{1 + (k-1)\rho_i}\right).$$

Now, consider all possible values of n_i with corresponding probability $P(n_i)$. The litter-based risk, corresponding to some specified dose d, is given by

$$r(d) = \sum_{n_i=0}^{\infty} P(n_i) q(n_i; d),$$

which is an average of the probabilities $q(n_i; d)$ with weights $P(n_i)$. The excess risk can be computed as

$$r^*(d) = 1 - \frac{\sum\limits_{n_i=0}^{\infty} P(n_i) \prod\limits_{k=0}^{n_i-1} \big(1 - \pi_i(d) + k\pi_i(d)\rho_i/(1 + (k-1)\rho_i)\big)}{\sum\limits_{n_i=0}^{\infty} P(n_i) \prod\limits_{k=0}^{n_i-1} \big(1 - \pi_i(0) + k\pi_i(0)\rho_i/(1 + (k-1)\rho_i)\big)}.$$

A Conditional Model

The exponential model of Molenberghs and Ryan (1999) also allows easy calculation of quantities such as the probability that at least one littermate is affected. Given the number of viable foetuses n_i, the probability of observing at least one abnormal foetus in a cluster is

$$q(n_i; d) = 1 - \exp\big(-A_{n_i}(\boldsymbol{\Theta_i})\big),$$

with $A_{n_i}(\boldsymbol{\Theta_i})$ the normalizing constant as is 3.4. Integrating over all possible values of n_i, we obtain the risk function

$$r(d) = \sum_{n_i=0}^{\infty} P(n_i) \big[1 - \exp\big(-A_{n_i}(\boldsymbol{\Theta_i})\big)\big],$$

where $P(n_i)$ is the probability of observing n_i viable foetuses in a pregnant dam. Using this equation, calculation of the excess risk $r^*(d)$ is straightforward.

4.2.2   Predictor Model

A frequently used predictor model in literature is the classical polynomial model. However, higher order conventional polynomials often fit the data badly at the extremes. Moreover, they do not have asymptotes and fit the data poorly whenever asymptotic behaviour is expected. Since low dose extrapolation is of primary interest in developmental toxicity studies, there is a clear need for alternative specifications in this setting. A very elegant alternative approach to classical polynomials, which falls within the realm of (generalized) linear methods, is given by fractional polynomials (Royston and Altman 1994). They are defined in Section 3.3.1. Fractional polynomials provide much more flexibly shaped curves than conventional polynomials, but in cases where the extension is not necessary, this family essentially reduces to conventional polynomials. It can be expected that the predictor model might have an important impact on the estimation of a safe dose of exposure. Thus, use of a flexible method is strongly recommended for an important public health matter such as the determination of safe limits for human exposure to potentially hazardous agents.

4.3   Use of Fractional Polynomials: An Example

In this chapter, fractional polynomials are used for dose-response modelling as basis for quantitative risk assessment. A similar application is illustrated in Aerts *et al.* (2002).

4.3.1   EG Study

Consider the EG study, as described in Section 2.1.1. In order to analyse the data from the EG study different modelling strategies are available such as the beta-binomial probability model or the conditional model of Molenberghs and Ryan. Traditionally one relies on a linear dose-trend for the main effect parameter, and a constant association. Figure 4.1 shows the observed frequencies of malformed litters at the selected dose levels (dots) and the dose-response curves based on the beta-binomial and conditional model with a linear d trend on the main effect (dotted and dashed line, respectively). These models are clearly too restrictive to adequately describe the dose-response relationship. The beta-binomial model gives an estimate of the benchmark dose equal to 375 mg/kg/day (corresponding to a 10% increase of risk over background). The MR model with a conventional linear polynomial results in a benchmark dose of 366 mg/kg/day. It can be questioned what the consequences are of using such restrictive models to assess a safe limit of the dosing. In this respect,
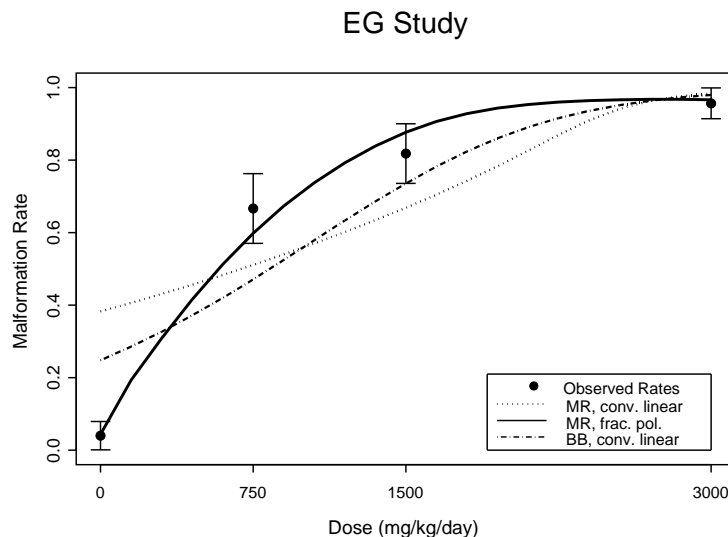
Figure 4.1: *EG study. Dose-response curves based on models 1, 2 and 3.*

one can try to further improve the fit by using a more flexible model with a fractional polynomial predictor in the MR model.

We select a suitable set of dose transformations on the main effect parameter, based on the fractional polynomial approach. We consider as the best set of transformations, the one producing the smallest Akaike's Information Criterion (AIC, Akaike 1974). Table 4.1 shows the best first and second degree models on the main effect parameters of the MR model. Among all models, the quadratic represented by $(\sqrt{d}, d)$ yields the smallest AIC. The clustering parameter is kept constant. The corresponding dose-response model is pictured in Figure 4.1 (full line). The MR model with the fractional polynomial predictor yields a benchmark dose of 94 mg/kg/day. Clearly, there is a large difference between the safe dose levels based on the linear predictor and on the fractional polynomial predictor. The fractional predictor yields a substantially lower (and thus more conservative) estimate of the safe dose compared with the conventional linear predictors in d, thanks to the better fitting dose-response curve.

4.3.2 DYME Study

Consider the DYME study from the NTP developmental experiments in mice, as described in Section 2.1.2. We restrict attention to the effect of DYME on the probability of having a full malformation. We select a suitable dose transformation to model

Table 4.1: EG study. Selection procedure for the main effect parameter in the MR model, describing the trend of the malformation outcome.

$m = 1$		$m = 2$	
transformation	AIC	transformation	AIC
$1/d^2$	737.49	$(1/d, d^2)$	722.99
$1/d$	745.66	$(1/d, 1/d^2)$	720.95
$1/\sqrt{d}$	749.80	$(1/d, d)$	722.23
$\ln(d)$	753.82	$(\sqrt{d}, d)$	720.30
$\sqrt{d}$	741.76	(d, d^2)	724.38
d	761.14	$(\ln(d), \ln(d)^2)$	722.37
d^2	776.79	$(\sqrt{d}, d^2)$	720.49
d^3	782.73	$(1/\sqrt{d}, 1/d^2)$	721.101

Table 4.2: DYME study. Selection procedure for the main effect parameter in the MR model, describing the trend of the malformation outcome.

$m = 1$		$m = 2$	
transformation	AIC	transformation	AIC
$1/d^2$	399.67	$(1/d, d^2)$	400.24
$1/d$	398.24	$(1/d, 1/d^2)$	400.23
$1/\sqrt{d}$	398.52	$(1/d, d)$	400.24
$\ln(d)$	399.47	$(\sqrt{d}, d)$	402.05
$\sqrt{d}$	400.22	(d, d^2)	400.15
d	403.39	$(\ln(d), \ln(d)^2)$	400.21
d^2	426.14	$(\ln(d), d^2)$	400.17
d^3	443.02	$(d, \ln(d))$	400.19

the developmental outcome. Table 4.2 shows the best first and second degree models on the main effect parameters of the MR model. Using the MR model, the polynomial represented by $(1/d)$ yields the smallest AIC among all first and second degree models. The clustering parameter is kept constant. The benchmark dose BMD_{10} from the beta-binomial model is equal to 107 mg/kg/day. The MR model with a conventional linear predictor yields 76 mg/kg/day, and the fractional polynomial predictor gives a benchmark dose of 85 mg/kg/day. Here, the difference between the use of a linear predictor in d or a transformed $1/d$ in the MR model, is much smaller than in the previous study. Use of the fractional polynomial predictor yields a lower, and thus more conservative, benchmark dose than the beta-binomial model with a linear predictor. The MR model with a linear predictor yields an overly conservative safe dose estimate, compared with the fractional polynomial model.

4.4 Asymptotic Simulation Study

In this section we perform an asymptotic simulation study to investigate the effect of model misspecifications on quantitative risk assessment. In addition, we investigate to what extent the use of flexible predictor models based on fractional polynomials can correct for such misspecification.

In order to obtain asymptotic information on the effect of model misspecification, we follow the particular recommendations of Rotnitzky and Wypij (1994). An artificial sample is constructed, where each possible realization is weighted according to its true probability under a given true model. In our case, we need to consider all realizations of the form (n_i, z_i, d_i), and have to specify: (1) $f(d_i)$, the relative frequencies of the dose groups, as prescribed by the design; (2) $f(n_i|d_i)$, the probability with which each cluster size can occur, possibly depending on the dose level (we will assume $f(n_i|d_i) = f(n_i)$), and (3) $f(z_i|n_i, d_i)$, the actual model probabilities. We assume that there are 4 dose groups, with one control group $(d_i = 0)$ and three exposed groups $(d_i = 0.25, 0.5, 1.0)$, and that each dose group has an equal probability of occurrence (i.e., $f(d_i) = 1/4$). A typical distribution of the number n_i of viable foetuses per cluster is given in Kupper *et al.* (1986), and can be considered representative of that encountered in actual experimental situations. Here, the cluster frequencies are assumed to follow a local linear smoothed version of these relative frequencies, as described in Aerts, Declerck and Molenberghs (1997) (Table 4.3).

Data are generated from the beta-binomial model with a given non-linear predictor for the mean parameter. Different dose trends on the mean parameter π of the true

Table 4.3: Local linear smoothed cluster frequencies.

n_i	$f(n_i)$	n_i	$f(n_i)$
1	0.0046	11	0.1179
2	0.0057	12	0.1529
3	0.0099	13	0.1605
4	0.0139	14	0.1424
5	0.0147	15	0.0975
6	0.0148	16	0.0542
7	0.0225	17	0.0207
8	0.0321	18	0.0086
9	0.0475	19	0.0030
10	0.0766		

model can be considered. Here, we look at two different models. The first model (Model A) is defined as

$$\text{logit}(\pi) = \beta_0 + \beta_d \sinh^2(d),$$

the second model (Model B) as

$$\text{logit}(\pi) = \beta_0 + \beta_d \cos(d).$$

In both models, the association parameter ρ is kept constant

$$\ln\Big(\frac{1+\rho}{1-\rho}\Big) = \beta_a.$$

We use parameter settings that were encountered in real data sets (Price *et al.* 1985, 1987). The parameter settings are summarized in Table 4.4. In the beta-binomial model the baseline risk is a function of both the intercept and association parameter. In model A, an association parameter β_a of 0.1, 0.3 and 0.5 correspond to a background rate of respectively 3%, 2% and 1% (for intercept -6), and 16%, 11%, 9% (for intercept -4). The Fisher's z-transformed correlation of 0.1, 0.3 and 0.5 corresponds to respectively a correlation ρ of about 0.05, 0.15 and 0.24. The parameters of model B correspond with a baseline malformation rate of about 15% and 11% for an association parameter β_a of 0.1 and 0.3 respectively. The Fisher's z-transformed correlation of 0.1 and 0.3 is used (correlation of respectively 0.05 and 0.15). Figure

Table 4.4: Parameter settings of the true (beta-binomial) model.

parameter	Model A $\sinh^2(d)$	Model B $\cos(d)$
intercept β_0	-4,-6	$2k$
dose effect β_d	3, 4, ..., 8	$-4 - 2k$
		$(k = 0, 1, \ldots, 7)$
association β_a	0.1, 0.3, 0.5	0.1, 0.3
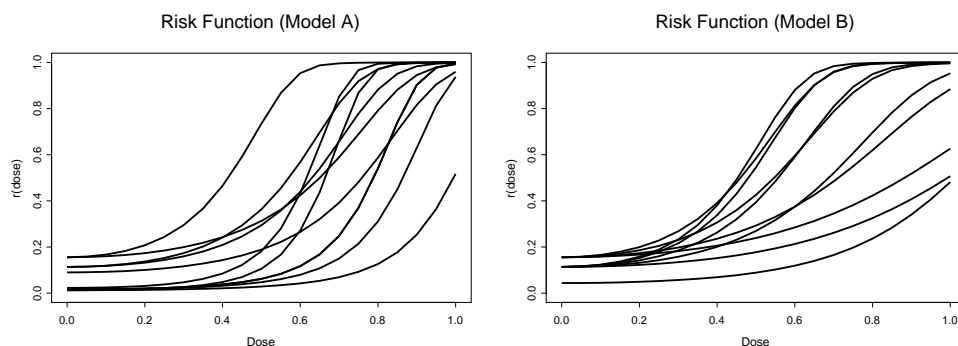
Figure 4.2: *Simulation setting. Dose-response models with $\sinh^2(d)$ or $\cos(d)$ trend on the main parameter.*

4.2 shows examples of both models to give some idea of the variety of dose-response models considered here.

For each model under study, we estimate the benchmark dose corresponding to a 10% extra risk over background ($q = 0.10$). The "true" benchmark dose is found by fitting the correct model, i.e., the model under which the data were generated, and by calculating the purely model-based benchmark dose. In real situations, the true benchmark dose can only be estimated based on an assumed model. Here, the technique introduced by Rotnitzky and Wypij (1994) is tailored to compute "asymptotic" values of the estimated benchmark dose. The benchmark dose is determined for the artificial sample under three different models:

- Model 1: the beta-binomial model (BB), with a conventional linear predictor for the mean π and a constant association ρ;

- Model 2: the conditional model of Molenberghs and Ryan (MR), with a conventional linear predictor for the main parameter θ and a constant association parameter δ;

- Model 3: the conditional model of Molenberghs and Ryan (MR), with the best fitting fractional polynomial predictor for the main parameter θ and a constant association parameter δ.

Different misspecifications occur in the above models. In the first model, the predictor model is misspecified. This often occurs in practice, when one uses a linear polynomial predictor. In the second model, the probability model is misspecified. Because the true probability model is unknown in general, this is what often happens in daily practice. And also in the third model, the predictor model is misspecified, but here we try to correct for misspecification of the probability model using a fractional polynomial predictor. Finding a suitable fractional polynomial predictor for the beta-binomial model is hard to accomplish. This involves the fitting of several different models, which in the case of a beta-binomial model often fail to converge (Molenberghs *et al.* 1998, Aerts *et al.* 1997). Therefore, we only use fractional polynomials for the conditional model and not for the beta-binomial model.

In choosing the optimal fractional polynomial, we follow the ideas of Royston and Altman (1994). Fractional polynomials of degree one and two, with powers of dose restricted to the set of values $\Pi = \{-2, -1, -1/2, 0, 1/2, 1, 2, 3\}$, are investigated. After fitting this set of dose transformations, only models that show a monotonic behaviour are kept as candidate models (assuming that the chemical results in more adverse effects when dose increases). Finally, the best fractional polynomial is chosen as the one producing the smallest value of Akaike's Information Criterion. Note that the penalizing constant cancels out in case of an asymptotic setting. Although we restrict attention to the AIC, alternative test statistics, such as the Bayesian Information Criterion (BIC, Schwarz 1978), could be investigated too. Some other selection procedures are described in Sauerbrei and Royson (1999).

Results of the simulation study are summarized in Tables 4.5 (Model A) and 4.6 (Model B), and visualized in Figures 4.3 and 4.4. Let us first have a look at the results when data are generated under Model A. It can be noted that none of the estimated benchmark doses are equal to the true benchmark dose. This of course is due to the misspecification of the model. The estimated benchmark doses based on the MR model with a conventional linear polynomial (Model 2, dashed line) are very small compared with the true benchmark doses (dots). However, when the dose-parameter increases, there is also a small decrease in the asymptotic bias. But still,

Table 4.5: Asymptotic estimation of benchmark dose (q=0.10), when true model is Model A.

$(\beta_0, \beta_d, \beta_a)$	True Model	Model 1	Model 2	Model 3
(-4,4,0.1)	0.344	0.228	0.178	0.310
(-4,6,0.1)	0.283	0.241	0.169	0.257
(-4,8,0.1)	0.246	0.216	0.160	0.229
(-4,4,0.3)	0.390	0.262	0.186	0.354
(-4,6,0.3)	0.321	0.263	0.173	0.276
(-4,8,0.3)	0.279	0.239	0.160	0.242
(-6,4,0.1)	0.616	0.479	0.455	0.563
(-6,6,0.1)	0.513	0.446	0.414	0.481
(-6,8,0.1)	0.449	0.433	0.380	0.420
(-6,4,0.3)	0.660	0.523	0.465	0.585
(-6,6,0.3)	0.550	0.474	0.422	0.526
(-6,8,0.3)	0.482	0.453	0.379	0.435

Table 4.6: Asymptotic estimation of benchmark dose (q=0.10), when true model is Model B.

$(\beta_0, \beta_d, \beta_a)$	True Model	Model 1	Model 2	Model 3
(0,-4,0.1)	0.502	0.298	0.261	0.484
(2,-6,0.1)	0.408	0.229	0.204	0.401
(4,-8,0.1)	0.353	0.204	0.179	0.351
(6,-10,0.1)	0.315	0.197	0.165	0.315
(8,-12,0.1)	0.288	0.199	0.155	0.283
(0,-4,0.3)	0.573	0.380	0.289	0.517
(2,-6,0.3)	0.466	0.287	0.222	0.431
(4,-8,0.3)	0.402	0.251	0.191	0.381
(6,-10,0.3)	0.360	0.237	0.173	0.286
(8,-12,0.3)	0.328	0.233	0.161	0.266
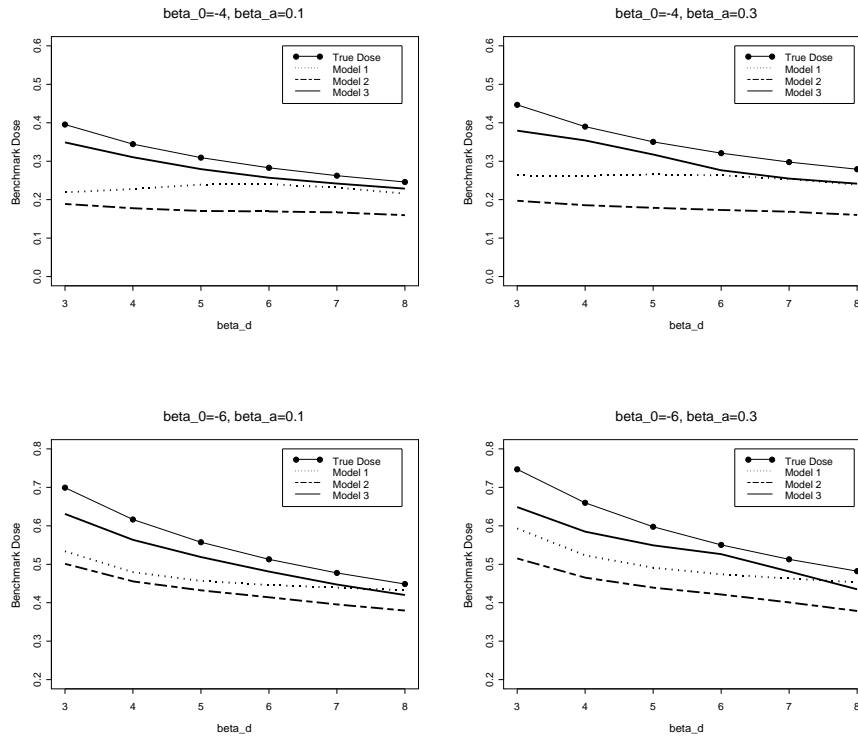
Figure 4.3: *Asymptotic estimation of benchmark dose (q=0.10), when true model is Model A.*
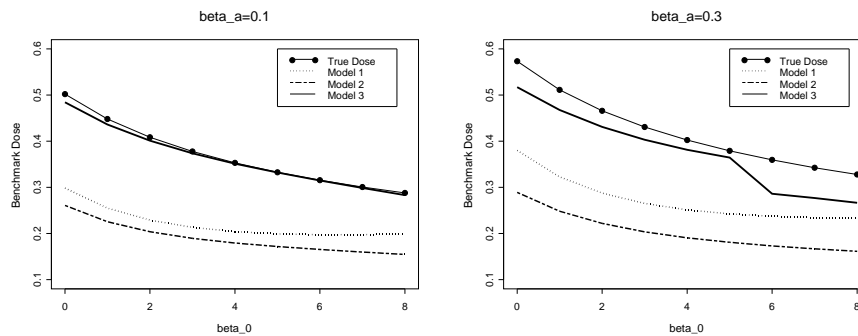
Figure 4.4: *Asymptotic estimation of benchmark dose (q=0.10), when true model is Model B.*

the estimated doses stay far apart from the true benchmark dose. Also the beta-binomial model (Model 1, dotted line) yield too small benchmark doses for most of the parameter settings. In contrast, benchmark dose estimates based on the MR model with a fractional polynomial predictor (Model 3, full line) are much closer to the true benchmark dose. This seems to indicate that fractional polynomials are much more flexible to attain the correct benchmark dose than conventional polynomials. While the use of small doses seems cautious, Morgan (1992) warns that safe dose determination should be tempered by common sense. For example, blind use of an overly conservative procedure has been regarded as scientifically indefensible by the Scientific Committee of the British Food Safety Council (1980), since it may produce unrealistically low safe doses.

When data are generated under Model B, the conclusions are similar and even more encouraging for the fractional polynomial model. The estimated benchmark doses for the beta-binomial model with a linear predictor are far from the true benchmark doses. This is due to the misspecification of the polynomial predictor. And also the estimated benchmark doses for the model of Molenberghs and Ryan with a conventional linear predictor are at a distance. Again, the fractional polynomials seem to partly correct for the model misspecification.

In order to investigate whether these conclusions also hold for classical random samples, a small sample simulation study was performed.

4.5   Small Sample Simulation Study

In the small sample simulation study, we reconstruct some realistic situations from developmental toxicity studies. The same models and parameter combinations as in the asymptotic study are investigated (Table 4.4), all of these might occur in practical experiments. For each of the selected models, 1000 datasets of 30 observations per dose group were generated. The estimated benchmark doses were averaged at the end of the run. Results are summarized in Tables 4.7 and 4.8, and pictured in Figures 4.5 to 4.6.

Similar to the asymptotic simulation results, benchmark dose estimates based on the conventional polynomial predictor are far from the true benchmark dose, for both the beta-binomial as well as the conditional model. In contrast, the fractional polynomial predictor yields benchmark dose estimates that are much closer to the true benchmark dose. Although, compared to the asymptotic sample setting, an increase in bias can be observed. This is especially true when the true model has a higher

Table 4.7: Small sample estimation of benchmark dose (MSE), when true model is Model A.

$(\beta_0, \beta_d, \beta_a)$	True Model	Model 1	Model 2	Model 3
(-4,4,0.1)	0.344	0.229 (0.014)	0.186 (0.026)	0.295 (0.005)
(-4,6,0.1)	0.283	0.241 (0.003)	0.179 (0.012)	0.259 (0.002)
(-4,8,0.1)	0.246	0.217 (0.002)	0.167 (0.007)	0.235 (0.001)
(-4,4,0.3)	0.390	0.267 (0.017)	0.194 (0.040)	0.288 (0.019)
(-4,6,0.3)	0.321	0.263 (0.004)	0.186 (0.024)	0.260 (0.012)
(-4,8,0.3)	0.279	0.243 (0.003)	0.170 (0.013)	0.241 (0.003)
(-6,4,0.1)	0.616	0.487 (0.017)	0.469 (0.022)	0.554 (0.004)
(-6,6,0.1)	0.513	0.448 (0.004)	0.421 (0.008)	0.495 (<0.001)
(-6,8,0.1)	0.449	0.433 (<0.001)	0.395 (0.003)	0.434 (<0.001)
(-6,4,0.3)	0.660	0.516 (0.021)	0.470 (0.036)	0.577 (0.007)
(-6,6,0.3)	0.550	0.485 (0.004)	0.477 (0.011)	0.515 (0.001)
(-6,8,0.3)	0.482	0.453 (0.001)	0.404 (0.006)	0.432 (0.003)

Table 4.8: Small sample estimation of benchmark dose (MSE), when true model is Model B.

$(\beta_0, \beta_d, \beta_a)$	True Model	Model 1	Model 2	Model 3
(0,-4,0.1)	0.502	0.400 (0.015)	0.274 (0.054)	0.400 (0.029)
(2,-6,0.1)	0.408	0.233 (0.032)	0.209 (0.040)	0.337 (0.014)
(4,-8,0.1)	0.353	0.207 (0.022)	0.185 (0.029)	0.308 (0.008)
(6,-10,0.1)	0.315	0.201 (0.014)	0.170 (0.022)	0.269 (0.007)
(8,-12,0.1)	0.288	0.200 (0.009)	0.161 (0.017)	0.245 (0.006)
(0,-4,0.3)	0.573	0.401 (0.030)	0.316 (0.066)	0.479 (0.009)
(2,-6,0.3)	0.466	0.296 (0.030)	0.234 (0.054)	0.384 (0.007)
(4,-8,0.3)	0.402	0.258 (0.021)	0.200 (0.041)	0.326 (0.006)
(6,-10,0.3)	0.360	0.243 (0.014)	0.181 (0.032)	0.268 (0.008)
(8,-12,0.3)	0.328	0.243 (0.007)	0.172 (0.024)	0.215 (0.013)
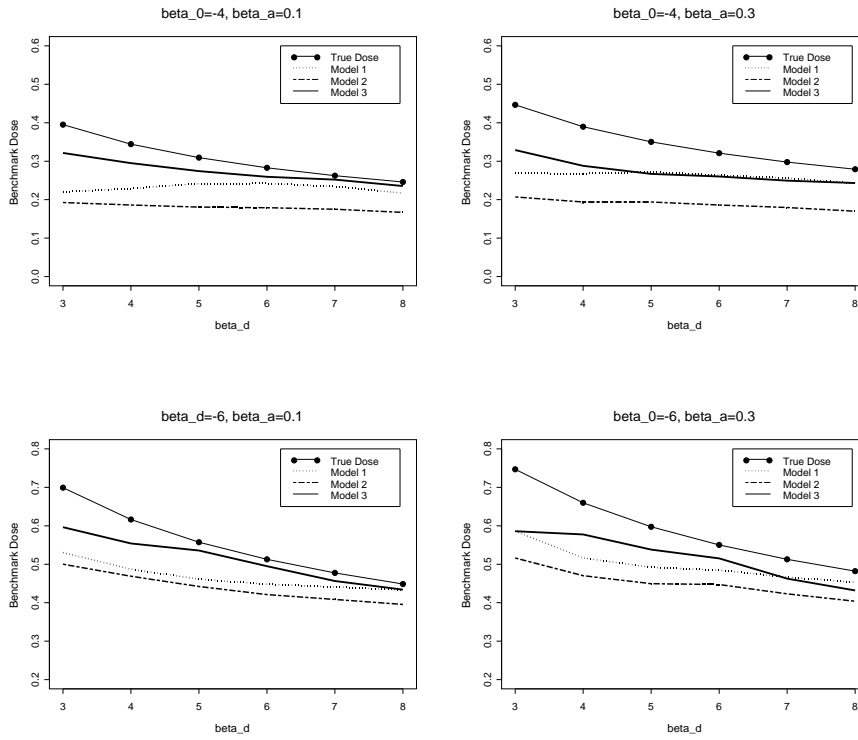
Figure 4.5: *Small sample estimation of benchmark dose (q=0.10), when true model is Model A.*
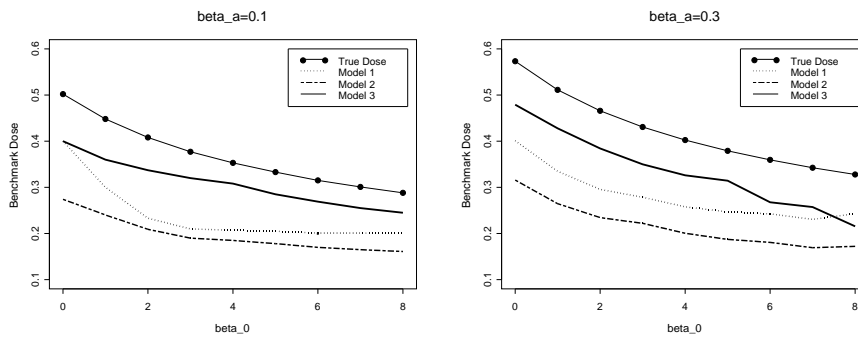
Figure 4.6: *Small sample estimation of benchmark dose (q=0.10), when true model is Model B.*

correlation, which is not so surprising because in this case the asymptotic results already indicated a misspecification bias.

In Figure 4.7, we present a scatterplot matrix of the estimated benchmark doses, for all generated datasets under Model A with parameters (-4,6,0.1). In this graph, the benchmark doses based on the beta-binomial model, the conventional MR model and the fractional MR model are compared, and also the true benchmark dose is marked (by a "T"). For this parameter setting, a dose trend d^2 is chosen as the best fractional polynomial in 42% of the cases, a $\sqrt{d}$ trend in 34% of the cases, $d^2 + d^2 \ln(d)$ in 16% of the cases, and several other trends. Since powers are restricted to a finite set of values, we can see a subgroup in the cloud of benchmark dose estimates based on the fractional polynomials. Most benchmark doses in this small group correspond to fractional polynomials with a $\sqrt{d} + d^2$ trend. Although the correlation between benchmark dose estimates of model 1 and 2 is high, these models yield unrealistically low benchmark dose estimates. It is clear that Model 3 is the most flexible model in attaining the correct benchmark dose.

The price that we have to pay for higher flexibility is precision. Use of a more flexible model yields higher standard errors. However, the increase in variability is less severe than the increase in bias. To acknowledge the sampling uncertainty for the model on which the benchmark dose is based, we replace the benchmark dose (BMD) by the lower effective dose (LED). Table 4.9 summarizes the LED estimates for model B. Results for Model A are similar. For both conventional and fractional polynomial predictors in the conditional model of Molenberghs and Ryan (Model 2 and 3), we show the (mean) estimated LED, the percentage of the LED's smaller than the true BMD, the mean difference of the LED's smaller than the true BMD and the mean distance of the LED's larger than the true BMD.

The estimated LED's using fractional polynomial predictors seem to behave quite well for realistic datasets. Around 95% of the lower effective doses of the 1000 generated samples are smaller than the true BMD, while the distance from the true safe dose stays small. Also the small percentage of estimated doses that are larger than the true benchmark dose are very close to the true BMD. When using Model 2, all estimated doses are smaller than the true benchmark dose; moreover the distance with the true safe dose is large. This confirms the conclusion that the estimated dose is too small when using the conventional linear predictor. In contrast, the fractional polynomials provide satisfactory results.

In summary, it is very important that the dose-response model fits the data well in order to determine a safe limit of exposure. This has implications for both the model family chosen as well as for the form of the predictors. Even when the prob-
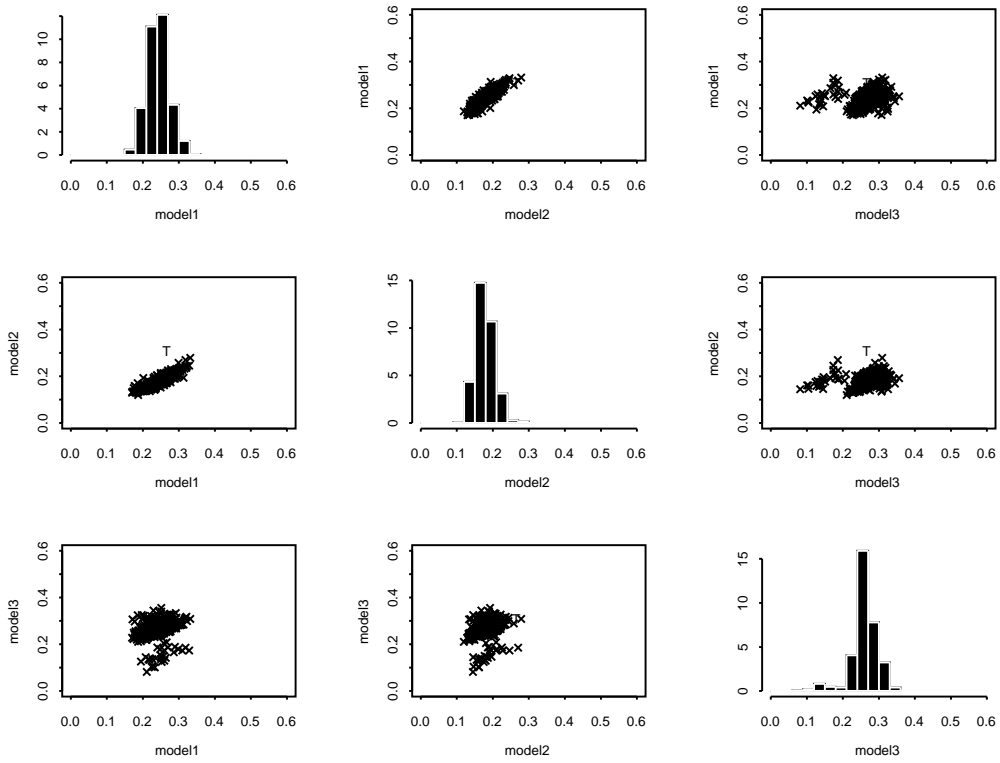
Figure 4.7: *Scatterplot matrix of the 1000 BMD's, estimated under Model 1, 2 and 3.*

Table 4.9: Lower effective dose estimation when true model is Model B.

True Model	Model 2			Model 3			
parameters	LED	perc <	dist <	LED	perc <	dist <	dist >
(0,-4,0.1)	0.229	100	0.273	0.357	97.4	0.150	0.018
(2,-6,0.1)	0.174	100	0.235	0.302	96.4	0.111	0.013
(4,-8,0.1)	0.153	100	0.200	0.275	94.2	0.084	0.013
(6,-10,0.1)	0.140	100	0.175	0.238	94.6	0.082	0.013
(8,-12,0.1)	0.134	100	0.154	0.215	94.8	0.078	0.016
(10,-14,0.1)	0.153	100	0.139	0.195	98.2	0.073	0.010
(12,-16,0.1)	0.127	100	0.122	0.184	98.7	0.066	0.009
(14,-18,0.1)	0.126	100	0.109	0.176	99.2	0.059	0.005

ability model is known, unreliable and unrealistically safe doses can be found. This demonstrates the importance of the shape of the predictors. In practice however, the true dose-response model is not known. Moreover, the choice between different dose-response models is often subjective and can affect the quantitative risk assessment. Use of a flexible polynomial predictor, such as a fractional polynomial, can partly solve the effects of model misspecification on quantitative risk assessment.

4.6   Discussion

Quantitative risk assessment is based on the relationship between dose and response to derive a safe dose. In this chapter, the effect of misspecifying the dose-response model on safe dose determination for clustered binary data was investigated. Both the probability model and the predictor model are important building blocks of a dose-response model, and are possibly subject to misspecification.

Simulation studies show that blind use of conventional linear predictors in the dose-response model can yield unrealistically low or unreliable safe doses, even when the probability model is well specified. As an alternative, fractional polynomial models were investigated. These flexible parametric models cannot only correct for misspecification of the predictor model, they can even partly correct for possible misspecification of the probability model. Therefore, the fractional polynomial approach is

important when searching for safe limits for human exposure to hazardous agents.

One concern, often raised for developmental toxicity studies is the danger of potential overfitting. Indeed, a standard teratology study typically involves no more than 4 or 5 different dose levels. Therefore, we have restricted ourselves to a (small) discrete set of fractional polynomials, with degree one or two. In general however, more design points would be desirable, but, from a practical point of view, such experiments are hard to manage in the developmental toxicity context. Other flexible parametric models could be considered too, such as models based on non-linear predictors and penalized splines. In contrast to the fractional polynomials, which are easy to handle with, nonlinear methods pose non-trivial methodological challenges. They are subject of interest in Chapter 5.

In this chapter, focus was on the risk of malformed litters. However, as observed from the data sets, the dose level also influences the foetal birth weight and number of viable foetuses in a dam, i.e., the litter size. Thus, extensions of the probability model, accounting for the multivariate outcomes and acknowledging the stochastic nature of the litter size would be more appropriate. These are topics of interest in Chapters 6 and 7. Also there, fractional polynomials are used to provide an appropriate fit to the data in order to obtain a safe level of exposure.

Chapter 5

Bayesian Testing for Trend in a Power Model for Clustered Binary Data

In the United States, more than 80000 chemicals are registered for commercial use and an estimated 2000 new ones are introduced annually for use in everyday items such as foods, personal care products, prescription drugs, household cleaners, etc. The effects of many of these chemicals on our health are still unknown (NTP 2003). However, the importance of hazard identification should not be disregarded. For safeguarding public health, one needs to identify the potential toxic exposures that cause adverse effects on developing foetuses. Secondly, if a specific substance is found toxic to the developing foetus, one should determine a safe dose of exposure. Recent techniques for risk assessment in this area are based on fitting dose-response models. As outlined before, the complexity of such studies implies a number of non-trivial challenges for model development and also the construction of dose-related trend tests, including the hierarchical structure of the data, litter effects inducing extra variation, the functional form of the dose-response curve, the adverse event at dam or at foetus level, the inference paradigm, etc. Whereas Chapter 4 dealt with estimation of a safe level of exposure, this chapter concentrates on testing whether a chemical exhibits a dose trend.

Tests for trend are often applied to toxicological data in order to assess possible

dose effects. The dose effect is investigated via testing the null hypothesis

$$H_0 : \text{no dose effect} \tag{5.1}$$

versus the alternative hypothesis that there is a dose effect. Results can only be
reliable when the alternative hypothesis model fits the data well (Williams and Ryan
1996, Aerts *et al.* 2002). Linear predictors of the type $g(\xi_i) = \beta_0 + \beta_1 d_i$ are widely used
to describe the model parameters, with g some link function, ξ_i a model parameter
and d_i a covariate associated with individual i. However, in the context of dose-
related trend tests, linear predictors can be too restrictive to adequately describe
the relationship of interest in real applications. As an alternative to a polynomial
model, the class of predictors can be enlarged by the use of fractional polynomials,
as described in Chapter 4, or by including non-linear predictors. Since the latter
approach invokes some interesting statistical issues when testing for trend, we will
concentrate on the family of non-linear models in this chapter. We focus on the
subgroup of power trend models. More specifically, power predictors of the following
type are studied in this chapter:

$$g(\xi_i) = \alpha + \beta d_i^\gamma, \tag{5.2}$$

with (α, β, γ) the unknown regression parameters. These power models allow very
flexible modelling, in spite of the parsimonious number of parameters (Davidan and
Giltinan 1995). Section 5.1 illustrates the use of power model in dose-response mod-
elling and quantitative risk assessment procedure.

    However, power models invoke some critical statistical issues when used for testing
purposes. The key item is linked with the effect of the covariate, e.g., dose given to a
dam, on the non-linear predictor. The power model suffers from lack of identifiability
under the null hypothesis of no dose effect, turning the testing problem into a non-
trivial one. Performing a test of no dose effect can be approached from a frequentist
point of view. Section 5.2 shows that this approach leads to complications that are due
to the non-identifiability of the parameters. Therefore, the feasibility of a Bayesian
approach to perform a trend test of no dose effect is investigated. Section 5.3 describes
a Bayesian framework for testing the null hypothesis of no dose effect on the predictor
$g(\xi_i)$. In case of clustered data, a modification of the Schwarz criterion (Schwarz
1978) is needed. This is presented in Section 5.4. We study and compare dose-related
trend tests on litter and on foetus level, taking into account clustering within litters.
Simulations show the performance of the method over a number of samples generated
under typical experimental conditions (Sections 5.5 and 5.6). In addition, a data
example shows the applicability of the Bayesian approach in developmental toxicity

studies. As in previous chapter, focus is on binary data from a developmental toxicity study. Most of the presented results can be found in Faes *et al.* (2004b).

5.1 Use of Power Models: An Example

In this section, the potential dose effect of ethylene glycol (EG) and of theophylline (THEO), as discussed in Section 2.1, are examined. When focus is on the risk of having an affected litter, in the sense that at least one of the foetuses in the litter has an adverse effect, the units under investigation are independent and binary observations. However, when focus is on the risk of having an affected foetus, units are clustered within litters. Here, the use of power models is investigated for both settings.

Independent Data

First, the adverse event of having a dam with at least one abnormal (i.e., dead or malformed) foetus will be taken as an illustration. In this respect, focus is on a binomial likelihood with a logit link function. The power model can be represented by

$$\text{logit}(\pi_i) = \alpha + \beta d_i^\gamma.$$

Here, doses are rescaled first to the $[0, 1]$ interval and then shifted to $[1, 2]$. The latter recoding is done in order to avoid numerical problems when fitting the power model, arising from the evaluation of the non-linear predictor when the control group is considered and $\gamma \leq 0$. Based on the parameters of the power model, the probability that a dam has at least one abnormal foetus can be estimated for each dose level. These probabilities are represented in Table 5.1. The Pearson goodness-of-fit statistic is used to check model fit. This approach shows no evidence of model lack of fit. Often, one uses a linear dose trend instead of a more flexible model. The corresponding model fits are given in Table 5.1. The observed malformation rates are displayed in Tables 2.1 and 2.3. One can see that the power model has more flexibility to adapt to the data compared with the linear model, as expected.

Clustered Data

Let us now investigate the adverse event of having an affected foetus. In that case, methods that account for the heterogeneity among litters are needed. An attractive approach is the use of the exponential model of Molenberghs and Ryan (1999). This method is a likelihood-based model for clustered binary data, based on a multivariate

Table 5.1: NTP studies. Estimated frequency based on power model and linear model.

		EG		THEO	
		Malf.	Death	Malf.	Death
Power Model		4.4	65.3	3.9	64.3
		63.7	67.2	8.6	73.8
		86.6	71.8	15.8	82.9
		93.7	91.4	19.3	85.8
Pearson X^2:		0.683	0.372	0.087	2.860
(df=1)					
Linear Model		15.7	61.3	3.3	63.5
		49.4	69.8	10.9	78.3
		83.6	77.1	15.6	82.0
		99.3	87.8	17.1	82.9
Pearson X^2:		9.902	1.476	0.318	3.422
(df=2)					

exponential family model (Cox 1972), as described in Section 3.1.2. The main param-
eter θ_i is modelled using a power model (5.2). Results are displayed in Figure 5.1.
The dots are the observed malformation or death rates in a cluster. The stars are the
average malformation or death rates in a dose group. The line corresponds with the
fitted power model. It is clear that the power model allows much flexibility on the
dose-response curve.

As seen in previous examples, a power model is a flexible approach to fit the data
with only a small number of parameters. It seems that power models are commonly
implemented primarily in order to get a better fit to the data, rather than for testing
purposes. However, in this chapter, the emphasis is on the latter aspect. Section 5.5
discusses the results of a Bayesian trend test for these NTP data when the adverse
event is defined on the dam (litter) level. In Section 5.6, we test for a dose-related
trend in case the risk is defined on foetus level. In both cases the outcomes are binary,
but in the latter case, all outcomes of foetuses within the same litter are correlated.
This has to be taken into account in the statistical model and in the Bayesian trend
test (see Section 5.4).
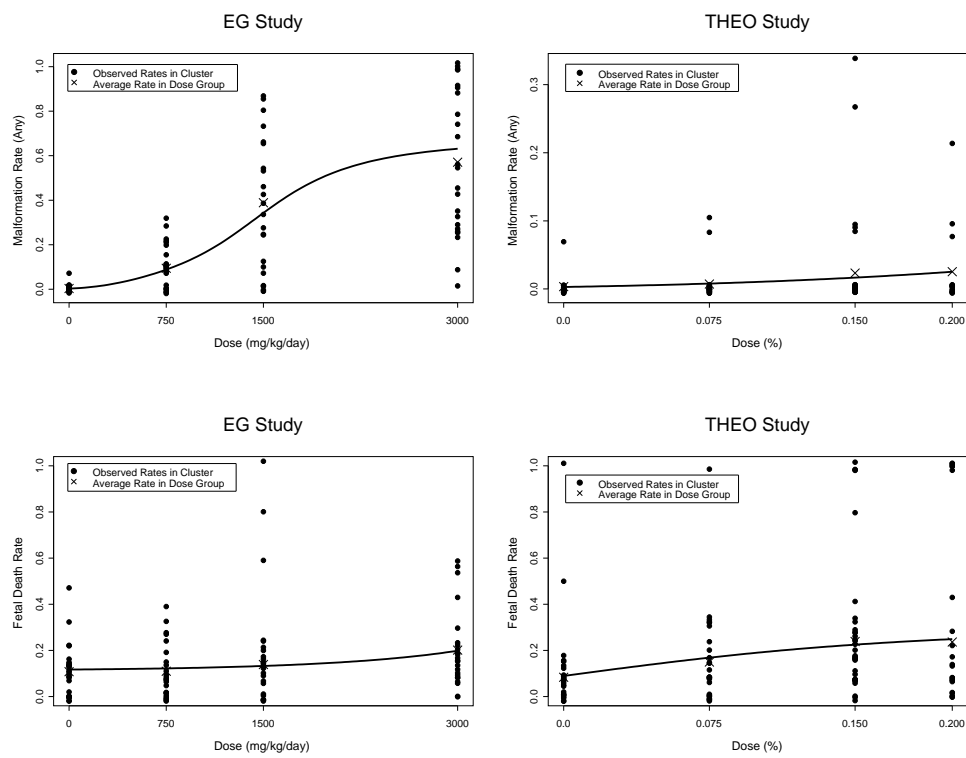
Figure 5.1: *NTP studies. Left: EG study. Right: THEO study. Observed and fitted malformation and foetal death rates, based on power model.*

5.2   Frequentist Testing

Use of power models leads to some critical statistical issues when testing hypotheses in a frequentist framework. Indeed, the case of no effect of dose d_i on the model parameter ξ_i can be rephrased as either $\beta = 0$ or $\gamma = 0$. This corresponds to the union of two planes in the parameter space of (α, β, γ), i.e., the planes with equation $\beta = 0$ or $\gamma = 0$. Furthermore, the condition that $\beta = 0$ or $\gamma = 0$ is equivalent with $\beta\gamma = 0$. As a consequence, the restriction put on the null model by implying no dose effect, is no longer linear in the parameters. Also, note that the regression parameters are not identifiable under the null hypothesis of no dose effect. The parameter γ is not identifiable if $\beta = 0$, since the model then reduces to $g(\xi_i) = \alpha$. If $\gamma = 0$, the model simplifies to $g(\xi_i) = \alpha + \beta$. In that case, one cannot identify α and β separately, although their sum is identifiable.

When performing a test of no dose effect from a frequentist point of view, there are severe complications due to the non-identifiability. Indeed, consider the likelihood ratio test

$$\Lambda_n = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}; y_1, \ldots, y_n)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}; y_1, \ldots, y_n)},$$

where $y_1, \ldots, y_n$ are the binary responses and $\boldsymbol{\theta} = (\alpha, \beta, \gamma)^T$. The (unconstrained) parameter space $\boldsymbol{\Theta}$ consists of all parameter vectors $\boldsymbol{\theta} \in I\!\!R^3$, while the (constrained) parameter space $\boldsymbol{\Theta}_0$ contains all $\boldsymbol{\theta} \in I\!\!R^3$ for which $\beta\gamma = 0$. The likelihood of the null model (numerator of Λ_n) reaches its maximum for some value of ξ, say ξ^*, or equivalently, some value of $g(\xi)$ equal to $g(\xi^*) \equiv \delta$. For regular cases, $-2\ln \Lambda_n$ has an asymptotic χ^2 null distribution. However, in the setting considered here, there is obviously a problem. Indeed, transforming the value ξ^* to the regression parameters, the maximum of the likelihood under the null model is obtained for

$$\text{all } \boldsymbol{\theta} \text{ for which } \alpha = g(\xi^*), \beta = 0$$

and

$$\text{all } \boldsymbol{\theta} \text{ for which } \alpha + \beta = g(\xi^*), \gamma = 0.$$

As a consequence, under the hypothesis of no dose effect, the likelihood is maximized at any parameter combination on these two intersecting lines. Thus, standard likelihood theory no longer holds and the asymptotic null distribution is unknown. For these reasons, the frequentist approach will not be pursued further. In contrast, a Bayesian approach leads to a more appealing approach.

5.3   Bayesian Testing

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote the binary response. Several Bayesian testing procedures have been proposed in the literature (Kass and Raftery 1995). Bayesian hypothesis testing can be performed by means of Bayes factors.

5.3.1   Bayes Factor

The Bayes factor can be defined as the ratio of the observed probability densities of $\boldsymbol{y}$ under both hypotheses:

$$B_{10} = \frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_0)}, \tag{5.3}$$

where H_0 and H_1 are defined in (5.1) and (5.2). From this definition, it follows that the Bayes factor can be viewed as measuring the relative success of H_0 and H_1 at predicting the data (Kass and Raftery 1995). Representing the a priori probabilities of the null and alternative hypotheses by $P(H_0)$ and $P(H_1)$, respectively, one notes that

$$B_{10}\frac{P(H_1)}{P(H_0)} = \frac{P(H_1, \boldsymbol{y})}{P(H_0, \boldsymbol{y})} = \frac{P(H_1|\boldsymbol{y})}{P(H_0|\boldsymbol{y})},$$

i.e., the Bayes factor multiplied by the prior odds results in the posterior odds. Assuming the prior information is 'neutral' between competing models, i.e., $P(H_0) = P(H_1) = 0.5$ (e.g., Berger 1985), the Bayes factor B_{10} equals $P(H_1|\boldsymbol{y})/P(H_0|\boldsymbol{y})$, the posterior odds in favor of H_1.

The two components of the Bayes factor, i.e., $P(\boldsymbol{y}|H_0)$ and $P(\boldsymbol{y}|H_1)$, are computed by integrating the joint density of the data and the regression parameters of the corresponding model, over its parameters. Hence, the probability $P(\boldsymbol{y}|H_0)$ is calculated by means of the expression

$$P(\boldsymbol{y}|H_0) = \int_{\boldsymbol{\Theta}_0} P(\boldsymbol{y}, \delta|H_0)d\delta = \int_{\boldsymbol{\Theta}_0} P(\boldsymbol{y}|\delta, H_0)\omega(\delta|H_0)d\delta,$$

where $\boldsymbol{\Theta}_0$ is the parameter space of δ and where $\omega(\delta|H_0)$ is the prior density of δ in the null model. The probability $P(\boldsymbol{y}|H_1)$ is found in an analogous way:

$$P(\boldsymbol{y}|H_1) = \int_{\boldsymbol{\Theta}_1} P(\boldsymbol{y}|\boldsymbol{\theta}_1, H_1)\omega(\boldsymbol{\theta}_1|H_1)d\boldsymbol{\theta}_1,$$

with $\boldsymbol{\theta}_1$ the parameter vector under the alternative model, i.e., $\boldsymbol{\theta}_1 = (\alpha, \beta, \gamma)^t$, with $\boldsymbol{\Theta}_1$ the parameter space of $\boldsymbol{\theta}_1$ and with $\omega(\boldsymbol{\theta}_1|H_1)$ the prior density of $\boldsymbol{\theta}_1$ in the alternative model. The two components of the Bayes factor are also called marginal

Table 5.2: Categories for the Bayes Factor, expressing the evidence against the null hypothesis (Kass and Raftery 1995).

B_{10}	$2 \ln B_{10}$	Evidence against H_0
< 1	< 0	negative (supports H_0)
1 to 3	0 to 2	not worth more than a bare mention
3 to 20	2 to 6	positive
20 to 150	6 to 10	strong
> 150	> 10	very strong

likelihoods or integrated likelihoods (Kass and Raftery 1995). The marginal likelihood is a weighted average of the likelihood, using the prior distribution as a weight function. In a limited number of cases, the marginal likelihood can be evaluated analytically. However, in most cases, the integrals are intractable and thus, numerical methods are needed to approximate the marginal likelihood. In the literature, an extensive number of numerical methods have been proposed (Kass and Raftery 1995, Carlin and Louis 1996, Gilks *et al.* 1996).

The Bayes factor is a summary of the evidence of one hypothesis as opposed to another hypothesis, provided by the data. Kass and Raftery (1995) provide categories for the Bayes factor, based on Jeffreys (1961), expressing the evidence against the null hypothesis. Table 5.2 lists classes for B_{10}, as well as for $2 \ln B_{10}$, which is on the same scale as, e.g., the likelihood ratio test statistic. These categories produce a rough descriptive statement about standards of evidence in scientific investigation (Kass and Raftery 1995). By comparing the computed Bayes factor of a data analysis with the classes of Table 5.2, one can draw a conclusion about the effect of dose d_i on the predictor $g(\xi_i)$.

For expressing evidence in favour of, or against, a model, the Bayes factor is desirable because of its direct interpretation. However, the Bayes factor requires specification of prior distributions on the parameters appearing in the null and alternative hypothesis models. In principle, these priors represent available information or some prior knowledge or belief. But in practice, proper priors must be chosen that may be perceived as subjective or ad hoc. For estimation problems 'flat' (uniform) priors on the parameters of interest are often used, but in testing such a prescription leads to serious difficulties. Thus, a first concern is how to choose a prior distribution to rep-

resent available information. Another important issue is the sensitivity of the Bayes factors to the prior distributions (Aitkin 1991, Kass and Wasserman 1995, Pauler 1998, Kass and Raftery 1995, Sinharay and Stern 2002). This will be illustrated in Section 5.5.

5.3.2 Schwarz Criterion

Because of the difficulties of assessing an appropriate prior under the nonlinear restrictions, and the computation of the required integrals, an approximation to the Bayes factor is attractive. The Schwarz criterion is easy to compute, and does not require explicit introduction of prior distributions. It is defined as

$$S = \ln \frac{P(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_1, H_1)}{P(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_0, H_0)} - \frac{1}{2}\big(\dim(\boldsymbol{\Theta}_0) - \dim(\boldsymbol{\Theta}_1)\big) \ln(n), \tag{5.4}$$

where $\hat{\boldsymbol{\theta}}_k$ is an ML estimate of $\boldsymbol{\theta}_k$ under H_k (with $k = 0, 1$), $\dim(\boldsymbol{\Theta}_k)$ is the dimension of $\boldsymbol{\Theta}_k$ and n is the sample size. Under the null hypothesis, there is only one non-redundant parameter, the constant success probability. Under the alternative hypothesis, there are three parameters of interest. The Schwarz criterion can be viewed as a rough approximation to the natural logarithm of the Bayes factor B_{10} (Kass and Raftery 1995). Hence, the Bayes factor B_{10} can be estimated by $\exp(S)$. In the setting considered here,

$$S = \ln \frac{P(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_1, H_1)}{P(\boldsymbol{y}|\hat{\boldsymbol{\delta}}, H_0)} - \ln(n), \tag{5.5}$$

with $\hat{\boldsymbol{\theta}}_1 = (\hat{\alpha}, \hat{\beta}, \hat{\gamma})^t$. Thus, it follows that

$$B_{10} \approx \frac{P(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_1, H_1)}{nP(\boldsymbol{y}|\hat{\boldsymbol{\delta}}, H_0)} \tag{5.6}$$

and that twice the Schwarz criterion equals the usual likelihood ratio test statistic minus twice the natural logarithm of the sample size. Keeping in mind the rough interpretation of the Bayes factor on the logarithmic scale of Table 5.2, it can be shown that in large samples, the Schwarz criterion should provide a reasonable indication of the evidence (Kass and Raftery 1995). Also, this procedure requires only the value of the likelihood ratio statistic and the number of parameters in both models. Furthermore, no prior distributions are needed. Nevertheless, several issues deserve further comments. An important point of concern regarding the use of Schwarz criterion is the sample size determination n in the case of clustered data. This is discussed in Section 5.4.

5.4   Effective Sample Size for Clustered Binary Data

When observations are not independent, it is not clear which sample size must be taken in the definition of the Schwarz criterion. In the context of developmental toxicity studies, exposure is often administered to the dam, rather than directly to the developing foetuses. Because of genetic similarity and the same treatment conditions, foetuses of the same mother behave more alike than those of another mother, inducing positive correlation (Chen and Kodell 1989, Kupper *et al.* 1986). One approach is to take the number of foetuses as the sample size in the definition of Schwarz criterion. As an alternative, one can take the number of clusters as the sample size. While the second one penalizes the effect too much, the first version of Schwarz does not sufficiently penalize. Discussions concerning the appropriate constant can be found in Kass and Wasserman (1995) and in Pauler (1997). Several extensions of Schwarz approximation are given by Haughton and Dudley (1992) and Kass and Vaidyanathan (1992). But, even with a supposedly well-specified method such as the Schwarz criterion, which depends only on the likelihood, the dimensions of the parameters and the sample size, there is considerable uncertainty as to how to define the sample size once one departs from i.i.d. situations. The determination of the effective sample size when dealing with more complex situations, such as in developmental toxicity studies, is a topic of interest.

Consider an experiment involving N litters (pregnant dams), the ith of which contains n_i viable foetuses. Suppose y_{ij} indicates whether the jth individual in dam i is malformed. Then, $z_i = \sum_{j=1}^{n_i} y_{ij}$ is the total number of malformations in cluster i. Let us focus on the information contributed by a cluster. Assuming the experiment was conducted with independence within clusters, observations can be seen as binomial counts with

$$
\begin{aligned}
\mathrm{E}\left(\frac{z_i}{n_i}\right) &= \pi_i \\
\mathrm{Var}\left(\frac{z_i}{n_i}\right) &= \frac{1}{n_i}\pi_i(1 - \pi_i).
\end{aligned}
$$

However in general, offspring of the same mother behave more alike than those of another mother. As a result, responses on different foetuses within a cluster are likely to be correlated, indicating extra variation in the data relative to those associated with the common binomial distribution. This extra variation must be taken into

account (Chen and Kodell 1989, Kupper *et al.* 1986). The variance equals

$$
\begin{aligned}
\mathrm{Var}\Big(\frac{z_i}{n_i}\Big) &= \frac{1}{n_i^2}\left[\sum_{j=1}^{n_i}\mathrm{Var}(y_{ij}) + 2\sum_{j<j'}\mathrm{Cov}(y_{ij}, y_{ij'})\right] \\
&= \frac{1}{n_i^2}\Big[n_i\pi_i(1-\pi_i) + n_i(n_i-1)\rho_i\pi_i(1-\pi_i)\Big] \\
&= \frac{1}{n_i}\pi_i(1-\pi_i)[1 + \rho_i(n_i-1)],
\end{aligned}
$$

with ρ_i the correlation among the observations in litter i. The effective sample size $\widetilde{n}_i$, or the number of independent samples contained in a litter, can be calculated by equating the variance under the assumption of independence with the true variance of a litter:

$$
\frac{1}{\widetilde{n}_i}\pi_i(1-\pi_i) = \frac{1}{n_i}\pi_i(1-\pi_i)[1 + \rho_i(n_i-1)],
$$

yielding

$$
\widetilde{n}_i = \frac{n_i}{1 + \rho_i(n_i-1)}. \tag{5.7}
$$

Note that the correction for the effective sample size is different for different litters. The effective sample size N_{eff} for the whole sample equals $\sum_{i=1}^{N}\widetilde{n}_i$, giving

$$
N_{\mathrm{eff}} = \sum_{i=1}^{N}\frac{n_i}{1 + \rho_i(n_i-1)}. \tag{5.8}
$$

There are some interesting special cases:

- When foetuses are independent within a litter ($\rho_i = 0$), the effective samples size equals

$$
N_{\mathrm{eff}} = \sum_{i=1}^{N} n_i.
$$

  This is equal to the number of foetuses.

- In case the foetuses are perfectly correlated ($\rho_i = 1$), the effective sample size equals the number of litters, since

$$
N_{\mathrm{eff}} = \sum_{i=1}^{N}\frac{n_i}{n_i} = N.
$$

- When the outcomes are continuous and normally distributed, and the compound-symmetry model

$$y_{ij} = \mu_{ij} + b_i + \epsilon_{ij},$$

  with $b_i \sim N(0, d)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ applies, exactly the same formula is obtained with $\rho = \frac{d}{d + \sigma^2}$.

The expression of the effective sample size yields insight into the amount of information contained in a cluster. When observations are correlated, a larger sample size is needed to achieve the same accuracy as with an independent sample. The larger the sample size, the more information and the better the accuracy. There is, however, a limit of information in a cluster:

$$\lim_{n_i \to +\infty} \frac{n_i}{1 + \rho(n_i - 1)} = \frac{1}{\rho} \tag{5.9}$$

Only when observations are independent, the information limit is infinite. When foetuses are correlated within a litter, there is a maximum amount of information that you can get for that cluster.

Above derivations are made for positive correlations. In general, the effective sample size is only positive for correlations $\rho > \frac{-1}{n-1}$. Thus, our argument can be used for mildly negative correlation, up to this bound. This is not dissimilar from negative correlation in the context of a positive definite variance-covariance matrix.

5.5   Independent Data

In this section, focus is on the risk of having a dam with at least one abnormal (i.e., dead or malformed) foetus. For each dam in the study, a binary response indicates whether there is an abnormality in the litter. Thus, the outcomes of interest are independent and binary observations. The effect of the choice of the prior on the Bayes factor and the performance of the Schwarz criterion are investigated in an analysis of the NTP data and in a simulation study.

5.5.1   Illustration: NTP Data

We investigate the effects in mice of the chemicals ethylene glycol and theophylline, as described in Section 5.1. For each adverse event under consideration, an approximation of the Bayes factor is computed in order to test for no dose effect in the power model.

Table 5.3: NTP studies. Three approximation of Bayes Factor (BF).

	EG		THEO	
	malf	death	malf	death
log-likel. (H_0)	-63.0	-54.8	-34.2	-58.4
log-likel. (H_1)	-35.3	-51.7	-32.5	-56.3
#dams	94	95	96	108
Schwarz	23.2	-1.53	-2.85	-2.54
BF(Schwarz)	$1.15*10^{10}$	0.216	0.058	0.079
BF(uniform, Var 1)	$3.66*10^{7}$	0.373	119.51	0.093
BF(normal, Var 1)	$2.81*10^{7}$	0.463	0.174	0.107
BF(uniform, Var 2)	$2.18*10^{8}$	0.557	0.236	0.275
BF(normal, Var 2)	$9.33*10^{7}$	2.278	0.559	0.702

Three methods are selected for the estimation of the Bayes factor. Besides the Schwarz criterion, the Bayes factor is approximated by computing the marginal likelihoods $P(\boldsymbol{y}|H_0)$ and $P(\boldsymbol{y}|H_1)$ numerically. Two types of prior distributions are considered: a uniform and a normal prior. The aim is to express some prior knowledge about the parameter estimates in this model. Therefore, the maximum likelihood estimates of a similar developmental toxicity study (Tyl *et al.* 1988) are used as a suggestion for the parameter estimates. The selected prior distribution for the parameter δ of the null model has mean 2, while the variance of δ equals 3. For the alternative model $\alpha + \beta d^\gamma$, the mean vector and the variance-covariance matrix of (α, β, γ), is (0.5,0.5,5) and 3 times the identity matrix respectively.

Table 5.3 displays the components of the Schwarz criterion, i.e., the log-likelihood of the power and null model and the number of dams. Also, the Schwarz criterion and the resulting estimate of the Bayes factor are given. Furthermore, the approximations of the Bayes factors obtained by computing the marginal likelihoods numerically using the aforementioned uniform or a normal prior distribution, are added to this table (Var 1). The results can be interpreted using Table 5.2.

For the effects of *EG on foetal death*, the results of the different approximations are similar. There seems to be no evidence against the null hypothesis of no dose effect of EG on the foetal death probability. For the effects of *EG on the presence*

of a malformation, there is very strong evidence against the null hypothesis. For the effects of *THEO on the presence of a malformation*, there is a striking discrepancy between the different methods. Depending on the method, there is negative to very strong evidence against the null hypothesis. Finally, there seems to be no evidence of an effect of *THEO on the foetal death* probability.

Other prior distributions could be considered. In order to express no prior knowledge, we could enlarge the variance of the prior distribution. As a sensitivity analysis, we calculated the Bayes factors under a uniform and normal prior with the same mean value, but with twice the variance of previously used priors (Var 2). Note that some of the conclusions change by the use of different prior distribution. This indicates that the Bayes factor can be very sensitive to the choice of the priors, and results should be interpreted with care.

Since the frequentist likelihood ratio test statistic is no longer appropriate to test the null hypothesis of no dose effect in case of a power model and the Bayes factor is very sensitive to the choice of the prior distribution, the Schwarz criterion based on the classes in Table 5.2 is an attractive alternative.

5.5.2   Simulation Setting

This section illustrates the finite sample behaviour of the Schwarz criterion and the Bayes Factor. We examine the performance of different approximations to the Bayes factor for the no dose effect null hypothesis. The adverse event under investigation is again *a dam with at least one abnormal foetus.*

In the simulation we reconstruct some realistic situations from developmental toxicity studies. A typical toxicological experiment includes one control group and some active dose groups. For the simulations we selected dose levels $0.000, 0.125,$ $0.250, 0.500$ and 1.000. Several parameter settings were investigated, all of these might occur in practical settings. For each of the selected models, the Schwarz criterion and Bayes factor approximation using a uniform and normal prior are considered. Two different variances of the prior distribution are chosen: (1) for the first prior we use 3 times the identity matrix (these prior distributions are referred to as uniform 1 and normal 1, respectively), (2) for the second prior we use 6 times the identity matrix (these prior distributions are displayed as uniform 2 and normal 2).

An equal number of 30 clusters were assigned to each dose group. Thus, for each dose level, 30 binary data values are generated, representing the health status of the group of dams. A value of zero indicates a dam without abnormal foetuses. For each parameter setting, 1000 datasets were generated.

5.5.3 Simulation Results

Several values of the parameter vector (α, β, γ) of the underlying power model are considered. In this study, a chosen combination of these parameters is $\alpha = 0.5, \beta = 0.5$ and $\gamma = 5$, which is comparable to the estimates of the EG study. Starting from the parameter vector $(\alpha, \beta, \gamma) = (0.5, 0.5, 5)$ one approaches the null hypothesis of no dose effect via three paths.

Results are presented in Tables 5.4 to 5.6. For each method, the percentages of simulations falling in the different categories of the Bayes factor, as indicated in Table 5.2, are given. For the estimation of the Bayes factor using the Schwarz criterion, the procedure for the estimation of the parameters in the power model did not always lead to convergence of the results. Careful choice of the starting points is needed in order for the model to converge. A tedious search for starting values is feasible for a single data analysis but it is impossible to always guarantee convergence in case of an extensive simulation study. Therefore the actual number of datasets for which convergence was reached is reported.

As expected, the distribution of the Bayes factor shifts to the left when approaching the null hypothesis. When generating data from the null hypothesis, i.e., $(\alpha, \beta, \gamma) = (0.5, 0, 5)$, $(0.5, 0.5, 0)$ or $(0.5, 0, 0)$, almost all runs lead to values of the Bayes factor smaller than one, implying that there is no evidence against the null hypothesis. When the underlying model is a power model, the distribution of the number of runs over the classes is similar for the uniform and the normal prior density. However, differences are seen between prior distributions with different variances. For the Schwarz criterion, the percentages of the number of runs in the classes with smaller values of the Bayes factor are larger than in case of the methods using prior distributions.

Use of Bayes factor and Schwarz criterion seems an attractive approach for testing the null hypothesis of no dose effect in a power model. However, there are some difficulties in choosing the prior distribution.

5.6 Correlated Data

In previous section, the litter effect issue is avoided by modelling the probability of an affected cluster via a logistic regression model. Here we consider the adverse event of having an affected foetus. Failure to account for the clustering in the data can seriously affect the test for trend. Thus, appropriate methods that account the heterogeneity among litters must be used. We restrict attention to the Schwarz criterion

Table 5.4: Distribution of the Bayes Factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes Factor, in addition to integrations over a grid using a uniform or a normal prior density function. Several values of the parameter β of the underlying power model are considered.

α	β	γ	method	# runs	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	> 150
0.5	0.5	5	Schwarz	579	7.9	5.4	19.9	21.1	45.8
			uniform 1	1000	0.4	1.4	10.9	17.8	69.5
			normal 1	1000	0.2	1.4	9.8	14.2	74.4
			uniform 2	1000	0.7	3.3	13.7	20.6	61.7
			normal 2	1000	0.1	1.8	11.5	21.8	64.8
0.5	0.4	5	Schwarz	640	13.0	6.1	18.3	20.2	42.5
			uniform 1	1000	0.5	2.0	9.1	17.2	71.2
			normal 1	1000	0.3	1.9	8.1	15.6	74.1
			uniform 2	1000	1.1	2.4	12.6	21.1	62.8
			normal 2	1000	0.2	2.3	8.8	21.3	67.4
0.5	0.3	5	Schwarz	790	12.0	5.1	19.5	20.8	42.7
			uniform 1	1000	0.6	1.9	9.6	20.9	67.0
			normal 1	1000	0.5	1.7	9.0	19.8	69.0
			uniform 2	1000	1.2	3.5	13.4	23.7	58.2
			normal 2	1000	0.0	1.3	8.9	23.8	66.0
0.5	0.2	5	Schwarz	898	22.0	3.3	15.1	24.2	35.3
			uniform 1	1000	0.5	2.1	11.6	24.7	61.1
			normal 1	1000	0.5	1.9	10.1	23.9	63.6
			uniform 2	1000	1.3	4.0	18.9	24.2	51.6
			normal 2	1000	0.2	0.6	7.9	30.1	61.2
0.5	0.1	5	Schwarz	871	34.6	10.6	17.8	16.9	20.2
			uniform 1	1000	7.8	10.4	22.7	23.6	35.5
			normal 1	1000	7.4	8.9	23.1	23.4	37.2
			uniform 2	1000	8.3	12.6	26.7	25.4	27.0
			normal 2	1000	2.3	4.9	20.8	27.5	44.5
0.5	0	5	Schwarz	552	99.8	0.2	0.0	0.0	0.0
			uniform 1	1000	99.8	0.2	0.0	0.0	0.0
			normal 1	1000	99.3	0.6	0.1	0.0	0.0
			uniform 2	1000	98.4	1.3	0.3	0.0	0.0
			normal 2	1000	95.5	3.9	0.6	0.0	0.0

Table 5.5: Distribution of the Bayes Factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes Factor, in addition to integrations over a grid using a uniform or a normal prior density function. Several values of the parameter γ of the underlying power model are considered.

α	β	γ	method	# runs	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	> 150
0.5	0.5	5	Schwarz	579	7.9	5.4	19.9	21.1	45.8
			uniform 1	1000	0.4	1.4	10.9	17.8	69.5
			normal 1	1000	0.2	1.4	9.8	14.2	74.4
			uniform 2	1000	0.7	3.3	13.7	20.6	61.7
			normal 2	1000	0.1	1.8	11.5	21.8	64.8
0.5	0.5	4	Schwarz	747	22.2	9.5	25.7	19.3	23.3
			uniform 1	1000	1.5	4.2	17.1	24.7	52.5
			normal 1	1000	1.2	3.7	17.0	23.2	54.0
			uniform 2	1000	2.3	5.8	23.1	26.0	42.8
			normal 2	1000	0.8	3.8	17.5	28.9	49.0
0.5	0.5	3	Schwarz	771	40.5	12.1	23.5	13.0	11.0
			uniform 1	1000	16.4	13.5	22.4	22.5	25.2
			normal 1	1000	15.0	12.7	22.6	22.4	27.3
			uniform 2	1000	15.1	16.6	25.9	22.7	19.7
			normal 2	1000	9.0	12.8	25.3	28.5	24.4
0.5	0.5	2	Schwarz	809	84.1	6.4	5.1	3.0	1.5
			uniform 1	1000	72.5	10.8	9.8	4.1	2.8
			normal 1	1000	71.0	10.0	11.6	4.3	3.1
			uniform 2	1000	64.5	15.3	13.7	4.3	2.2
			normal 2	1000	57.3	17.2	17.3	5.2	3.0
0.5	0.5	1	Schwarz	549	98.0	0.5	1.1	0.4	0.0
			uniform 1	1000	96.7	1.8	1.1	0.3	0.1
			normal 1	1000	96.2	2.1	1.3	0.3	0.1
			uniform 2	1000	94.8	3.0	1.8	0.3	0.1
			normal 2	1000	87.4	8.4	3.3	0.6	0.3
0.5	0.5	0	Schwarz	607	99.2	0.5	0.2	0.2	0.0
			uniform 1	1000	99.1	0.5	0.3	0.1	0.0
			normal 1	1000	99.1	0.6	0.3	0.0	0.0
			uniform 2	1000	98.8	0.9	0.2	0.1	0.0
			normal 2	1000	96.8	2.5	0.5	0.2	0.0

Table 5.6: Distribution of the Bayes Factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes Factor, in addition to integrations over a grid using a uniform or a normal prior density function. Several values of the parameter β and γ of the underlying power model are considered.

α	β	γ	method	# runs	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	> 150
0.5	0.5	5	Schwarz	579	7.9	5.4	19.9	21.1	45.8
			uniform 1	1000	0.4	1.4	10.9	17.8	69.5
			normal 1	1000	0.2	1.4	9.8	14.2	74.4
			uniform 2	1000	0.7	3.3	13.7	20.6	61.7
			normal 2	1000	0.1	1.8	11.5	21.8	64.8
0.5	0.4	4	Schwarz	837	25.6	7.8	22.2	20.0	24.5
			uniform 1	1000	2.4	6.7	14.9	26.2	49.8
			normal 1	1000	1.8	6.0	15.0	25.0	52.2
			uniform 2	1000	4.2	7.6	20.9	26.1	41.2
			normal 2	1000	0.8	4.4	15.9	30.6	48.3
0.5	0.3	3	Schwarz	852	61.7	11.2	13.7	7.7	5.6
			uniform 1	1000	44.3	12.8	20.1	12.4	10.4
			normal 1	1000	40.4	14.7	21.0	12.8	11.1
			uniform 2	1000	35.1	19.4	24.3	12.6	8.6
			normal 2	1000	25.7	20.2	25.0	17.3	11.8
0.5	0.2	2	Schwarz	837	97.0	1.4	1.4	0.1	0.0
			uniform 1	1000	95.5	2.6	1.5	0.3	0.1
			normal 1	1000	94.6	3.0	2.0	0.3	0.1
			uniform 2	1000	91.9	5.2	2.3	0.6	0.0
			normal 2	1000	83.2	10.5	5.3	0.9	0.1
0.5	0.1	1	Schwarz	837	99.1	0.6	0.1	0.1	0.0
			uniform 1	1000	99.2	0.5	0.2	0.1	0.0
			normal 1	1000	99.0	0.7	0.2	0.1	0.0
			uniform 2	1000	98.2	1.3	0.4	0.1	0.0
			normal 2	1000	97.7	1.7	0.4	0.2	0.0
0.5	0	0	Schwarz	552	99.8	0.2	0.0	0.0	0.0
			uniform 1	1000	99.8	0.2	0.0	0.0	0.0
			normal 1	1000	99.3	0.6	0.1	0.0	0.0
			uniform 2	1000	98.4	1.3	0.3	0.0	0.0
			normal 2	1000	95.5	3.9	0.6	0.0	0.0

and investigate the effect of the modified Schwarz criterion, as defined in Section 5.4. The NTP data are reconsidered and a simulation study shows finite sample power characteristics.

5.6.1 Illustration: NTP Data

In this section, interest is in the effect of the chemicals EG and THEO on the probability of an abnormal foetus, instead of on the probability of an abnormal cluster.

In a first step, we *ignore the clustering* by treating the littermates as if they were independent and focus on independent binary responses. Results of the analysis are tabulated in Table 5.7. Again, we see that the Bayes factor is very sensitive to the choice of the prior distribution. In addition, it is not clear what prior distribution is an appropriate choice. Therefore, in what follows, we advocate the use of the Schwarz criterion for Bayesian testing in this setting. Based on Schwarz criterion, there is strong evidence that the *chemical EG* has an effect on the developing foetus, in the sense that more foetuses will be malformed. In contrast, there is no evidence that EG will affect the risk of foetal death. For the *effect of THEO*, there is very strong evidence that this chemical has an effect on the risk of foetal death, but not on the malformation risk. But, treating the littermates as being independent is in general a very strong assumption. Therefore, results must be interpreted with care.

In a second step, the *clustering is taken into account* using the exponential model of Molenberghs and Ryan (1999), as described in Section 3.1.2. Note that it is not immediately clear which sample size must be taken in the definition of the Schwarz criterion. The number of foetuses can be taken as the sample size, but also the number of clusters. While the second one penalizes the effect too much, the first version of Schwarz does not penalize enough. When dealing with clustered binary data, the effective sample size discussed in Section 5.4 can be used.

Results of the analysis are presented in Table 5.7. For the malformations, the information obtained from the EG experiment is similar to what would be obtained from 596 independent foetuses. For the death outcome, the information obtained from the EG experiment is similar to what would be obtained from 782 independent foetuses. The larger the correlation in the litters, the smaller the effective sample size of the experiment. This is also illustrated in Figure 5.2 for the EG (death) data. For the EG (death) experiment, the information limit is equal to 2196.6, corresponding to approximately 23.4 per cluster. Hence, in the EG (death) data, a litter can never contribute more information, regarding the average (probability of success) within the cluster, than would be obtained from about 23.4 independent littermates. For the

Table 5.7: NTP Studies. Approximations of Bayes Factor.

Account	Method	Penalizing	EG		THEO	
Cluster		constant	malf	death	malf	death
no	Schwarz	$\sum_{i=1}^{N} n_i$	158.33	-0.465	-3.25	12.48
	BF (Schwarz)		5.79×10^{68}	0.628	0.0388	2.63×10^5
	BF (unif, var1)		6.35×10^{55}	10.36	202.79	3.11×10^6
	BF (norm, var1)		1.99×10^{59}	29.24	1.32	4.29×10^6
	BF (unif, var2)		2.38×10^{64}	47.52	1.28	3.36×10^6
	BF (norm, var2)		5.01×10^{66}	0.46	4.54	3.54×10^{11}
yes	Schwarz1	N	43.41	-1.57	-1.42	0.746
	Bayes		7.23×10^{18}	0.208	0.242	2.1090
	Schwarz2	$\sum_{i=1}^{N} n_i$	41.02	-4.10	-3.83	-1.74
	Bayes		6.51×10^{17}	0.017	0.022	0.175
	Schwarz3	N_{eff}	41.56	-3.68	-3.79	-0.655
	Bayes		1.12×10^{18}	0.025	0.022	0.519

	EG		THEO	
	malf	death	malf	death
$N_{\text{eff}}(\rho = 0)$	1028	1192	1071	1302
$N_{\text{eff}}(\rho = \widehat{\rho})$	596	782	1028	438
$N_{\text{eff}}(\rho = 1)$	94	95	96	108
Information Limit	22666.5	2196.6	44231.7	771.2

Effective Sample Size
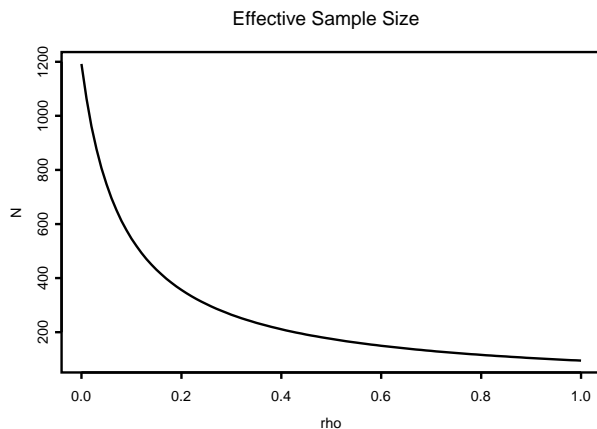
Figure 5.2: *EG Study (death outcome). Relationship between the effective sample size and the correlation in a litter.*

THEO (death) data, the information limit of a cluster is approximately equal to 7.1 independent littermates.

For the *EG experiment*, the three versions of the Schwarz criterion give similar results. There is no evidence that the chemical EG has an effect on the foetal death rate. But, there is very strong evidence against the null hypothesis of no dose effect of EG on the probability of a malformations in a litter. For the *chemical THEO*, there seems to be no evidence for an effect on the malformation risk. For the effects of THEO on the foetal death rate however, the differences in the penalizing constant affect the conclusions. The Schwarz criterion based on the effective sample size lies in between the naive values of the Schwarz criterion. Based on the effective sample size, there is no evidence that THEO affects the foetal death rate.

Comparing these results (accounting for clustering) with previous results (ignoring the clustering), we see that the Bayes factor is much smaller when accounting for clustering. When ignoring the clustering, the Bayes factor seems to favour the alternative hypothesis. Indeed, failure to account for the clustering in the data can lead to serious inflation of the test statistic.

5.6.2   Simulation Setting

A small sample simulation is performed, generating clustered binary data, to investigate the effect of the Bayes factor when approaching the null hypothesis in case of clustering, and to compare the Bayes factors under different probability models.

Data are generated following the ideas of Rotnitzky and Wypij (1994). An artifical sample is constructed where each possible realization of dose d, clustersize n, number of abnormal foetuses z is weighted according to the probability in the underlying model. An equal number of 30 clusters were assigned to each dose group. Again, the doses 0.000, 0.125, 0.250, 0.500 and 1.000 are considered when generating the asymptotic sample. The cluster sizes are random, according to a local linear smoothed version of the relative frequencies given in Kupper *et al.* (1986) (see Table 4.3).

5.6.3   Simulation Results

The parameters α and γ are kept fixed, equal to -2 and 4 respectively. We use the MR model (taking clustering into account) and the foetus-based logistic model (ignoring clustering). Because of the problems related to the choice of the prior distribution, we considered only the Schwarz criterion. Results are summarized in Tables 5.8 to 5.10.

When dose effect decreases, the distribution of the Bayes factor shifts to the left. This indicates that a smaller dose trend yields smaller Bayes factors, in line with expectation. The logistic model often yields larger Bayes factors compared with the Bayes factor when accounting for clustering. Thus, when testing for dose effect, it is important to use the most correct model. We should not ignore clustering when it is present, since it might influence the results. When accounting for clustering, different penalizing constants could be chosen. Note however that the second version of Schwarz (based on number of foetuses) penalizes the effect too much, while the first version of Schwarz (based on number of litters) does not sufficiently penalize. The distribution of the Bayes factors based on the effective sample size is close to the one based on the number of foetuses. However, when the clustering parameter increases, the differences get larger. This is in agreement with our expectations. Thus, use of the effective sample size in case of dependent data is an easy solution with attractive features.

5.7   Discussion

In this chapter, focus has been on testing for a dose-related trend based on a power model. The null hypothesis of no dose effect is equivalent with setting the product of

Table 5.8: Distribution of the Bayes Factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes Factor. Simulated data are clustered binary observations. Different penalizing factors are used in the Schwarz criterion. The association is only minor.

β	δ	model	penaliz.	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	$150 \geq .$
0.3	0.15	MR	litter	0	0	0	0	100
		MR	foetus	0	0	0	0	100
		MR	eff.ss.	0	0	0	0	100
		LOG	foetus	0	0	0	0	100
0.1	0.15	MR	litter	2	1	0	0	97
		MR	foetus	3	0	0	0	97
		MR	eff.ss.	3	0	0	0	97
		LOG	foetus	1	1	0	0	97
0.05	0.15	MR	litter	58	10	18	8	6
		MR	foetus	82	5	7	5	1
		MR	eff.ss.	82	5	7	5	1
		LOG	foetus	71	7	8	5	8

two regression parameters equal to zero. This non-linear restriction of the parameters in the null model results in parameter unidentifiability in case the effect of dose is absent. To avoid the computation of the distribution of the likelihood ratio test statistic in this setting, a Bayesian approach using Bayes factors is considered.

In the context of developmental toxicity studies, we examined the effect of dose on an adverse event, which can be defined on foetus or litter level. The results of both approaches, as given in Tables 5.3 and 5.7, are in line with each other. The results at foetus and litter level are consistent. However, failure to account for the clustering in the data can lead to serious inflation of the test statistic.

One of the methods applied to approximate the Bayes factor is based on the Schwarz criterion. Another method estimated the Bayes factor by integrating the marginal likelihoods numerically, using a uniform or normal prior distribution. However, the Bayes factor is very sensitive to the choice of the prior densities. Obviously, many other techniques for the calculation of the Bayes factor can be applied. One

Table 5.9: Distribution of the Bayes Factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes Factor. Simulated data are clustered binary observations. Different penalizing factors are used in the Schwarz criterion. The association in the data is moderate.

β	δ	model	penaliz.	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	$150 \geq .$
0.3	0.30	MR	litter	0	1	1	6	92
		MR	foetus	2	3	3	1	91
		MR	eff.ss.	0	1	4	4	91
		LOG	foetus	0	0	0	0	100
0.1	0.30	MR	litter	26	11	15	11	37
		MR	foetus	50	3	13	16	18
		MR	eff.ss.	45	8	10	18	19
		LOG	foetus	32	7	6	7	48
0.05	0.30	MR	litter	98	2	0	0	0
		MR	foetus	100	0	0	0	0
		MR	eff.ss.	100	0	0	0	0
		LOG	foetus	100	0	0	0	0

can also approximate the marginal likelihood by the simple Monte Carlo estimate or apply the technique of importance sampling. Furthermore, Kass and Raftery (1995) propose other methods for the calculation of the Bayes factor, e.g., Laplace's method.

The Schwarz criterion is a nice alternative to the Bayes factor, to test the null hypothesis of no dose effect in case of a power model. Although there are some issues concerning the penalizing constant, it is a good approximation. The investigation about the sample size and penalizing constant in case of dependent data gave us an easy solution, with attractive features, and might be used in other settings as well.

Table 5.10: Distribution of the Bayes Factor obtained from small sample simulations in which the Schwarz criterion is used to approximate the Bayes Factor. Simulated data are clustered binary observations. Different penalizing factors are used in the Schwarz criterion. The association in the data is high.

β	δ	model	penaliz.	≤ 1	$1 < . \leq 3$	$3 < . \leq 20$	$20 < . \leq 150$	$150 \geq .$
0.3	0.45	MR	litter	0	0	8	2	90
		MR	foetus	6	3	1	1	89
		MR	eff.ss.	0	0	10	1	89
		LOG	foetus	0	0	0	0	100
0.1	0.45	MR	litter	76	12	7	2	3
		MR	foetus	94	2	2	1	1
		MR	eff.ss.	92	3	3	1	1
		LOG	foetus	70	0	11	6	13
0.05	0.45	MR	litter	98	0	2	0	0
		MR	foetus	99	1	0	0	0
		MR	eff.ss.	99	1	0	0	0
		LOG	foetus	98	2	0	0	0

Chapter 6

Modelling Combined Continuous and Ordinal Outcomes in a Clustered Setting

In previous chapters we have concentrated on the clustered binary data setting. However, in the study of teratology and developmental toxicity, both the probability that a live foetus is malformed (ordinal) or of low birth weight (continuous) are important measures of teratogenicity. While multivariate methods for the analysis of continuous outcomes are well known (Johnson and Wichern 1992), methods for joint continuous and discrete outcomes are less familiar, especially with clustering. Nevertheless, measurements of both continuous and categorical outcomes appear in many statistical problems.

Some attempts have been made towards a joint distribution of binary and continuous responses within a litter. A frequent approach is to apply a conditioning argument that allows the joint distribution to be factorized in a marginal component and a conditional component, where the conditioning can be done on either the binary or continuous response (Catalano and Ryan 1992, Cox and Wermuth 1992, Cox and Wermuth 1994, Fitzmaurice and Laird 1995, Olkin and Tate 1961). Catalano (1997) extended the idea to model simultaneously ordinal and continuous outcomes. A description of this model is given in Section 6.1.

Alternative methods for joint binary and continuous outcomes were proposed by Regan and Catalano (1999a,b) and Geys *et al.* (2001). The first ones introduce a probit approach, where an underlying continuous variable is assumed for each binary outcome, following a normal distribution. The second approach is based on a Plackett-Dale approach (Plackett 1965), assuming a Plackett latent variable to model bivariate endpoints in which one component is continuous and the other is binary. The bivariate latent variable models are fundamentally different in the way the association between both variables is described. The probit approach uses a correlation coefficient, while the Plackett-Dale approach makes use of an odds ratio. The correlation coefficient of the bivariate normal induces constant local association (Holland and Wang 1987), while the odds ratio is a measure of constant global association (Dale 1986, Lapp, Molenberghs and Lesaffre 1998). However, extensions to joint ordinal and continuous outcomes are in demand. We propose a likelihood-based method that is an extension of the Plackett-Dale approach. Specification of the full likelihood will be avoided using pseudo-likelihood methodology. An advantage of the Plackett distribution is the flexibility with which the marginal densities can be chosen. In this case, a multinomial and a normal distribution are used, both being a common choice for the ordinal and continuous variable, respectively. Other related methods, using latent variable and shared random effect models, can be found in Sammel, Ryan and Legler (1997), Dunson (2000) and Gueorguieva and Agresti (2001).

In this chapter, we propose a new method for modelling combined continuous and ordinal outcomes in a clustered setting, as presented in Faes *et al.* (2004a). Section 6.1 introduces a factorization model for the joint analysis of an ordinal and a continuous outcome. In Section 6.2 a joint continuous-ordinal model based on a Plackett distribution is proposed. In a first step, the model is described under an independence assumption, and is then extended to deal with the clustering of outcomes within litters. The ultimate goal of developmental toxicity studies is to perform risk assessment, i.e., to set safe limits of exposure, based on the fitted model (Crump 1984). This will be discussed in Section 6.3. In Section 6.4, our method is applied to the developmental toxicity data as introduced in Chapter 2.

Let us extend previous notations. Consider an experiment involving N clusters, the ith of which contains n_i individual foetuses. Each viable individual is examined for two outcomes, the degree of malformation (e.g. none, minor, severe) and the foetal weight. Let M_{ij} be the random variable representing the status of malformation ($m = 1, 2, \ldots, c$) of the jth individual in litter i, and W_{ij} the continuous weight outcome. Together with this vector of two responses $\boldsymbol{Y}_{ij} = (W_{ij}, M_{ij})^T$, a vector of covariates $\boldsymbol{X}_{ij}$ is observed.

6.1 Factorization Model

Catalano (1997) proposed a joint model for combined ordinal and continuous outcomes. He assumes that the ordinal outcome M_{ij} has an associated unobservable continuous latent variable M_{ij}^* such that

$$M_{ij} = \begin{cases} 1 & \text{if } M_{ij}^* \leq h_1, & \text{normal} \\ 2 & \text{if } h_1 < M_{ij}^* \leq h_2, & \text{minor malformation} \\ 3 & \text{if } M_{ij}^* > h_2, & \text{full malformation} \end{cases}$$

where h_1 and h_2 ($h_1 < h_2$) are some threshold values. Then, it is assumed that the continuous outcome and the latent variable share a joint normal distribution. Let $W_i = (W_{i1}, \ldots, W_{in_i})^T$ and $M_i^* = (M_{i1}^*, \ldots, M_{in_i}^*)^T$ be the $n_i \times 1$ vectors of observed continuous and unobserved latent variables in the ith cluster. The $2n_i \times 1$ vector of observed and latent outcomes $(W_i^T, M_i^{*T})^T$ is assumed to follow a multivariate normal distribution with means $\mu_{w_{ij}}$ and $\mu_{m_{ij}}$ among the foetal weights and latent malformations, respectively. The assumed covariance structure allows for a constant correlation ρ between observations on the same foetus (intrafoetus) and separate correlations between observations on littermates (intralitter), resulting in the following block equicorrelated covariance matrix:

$$\text{Cov}(W_i^T, M_i^{*T})^T = \begin{pmatrix} \sigma_w^2[(1-\rho_w)I_i + \rho_w J_i] & \sigma_w\sigma_m[(\rho - \rho_{wm}I_i + \rho_{wm}J_i)] \\ \sigma_w\sigma_m[(\rho - \rho_{wm}I_i + \rho_{wm}J_i)] & \sigma_m^2[(1-\rho_m^*)I_i + \rho_m^* J_i] \end{pmatrix},$$

with ρ_w and ρ_m^* the intra-litter correlations for the continuous and unobserved latent outcomes, and ρ_{wm} the correlation between weight and the latent malformation variable for different animals in the same litter.

The bivariate distribution of the combined outcomes is derived by factorizing the joint density as the product of the marginal density for the continuous outcome and the conditional density of the ordinal outcome given the continuous outcome:

$$f_{W_{ij}, M_{ij}}(w, m) = f_{W_{ij}}(w) f_{M_{ij}|W_{ij}}(m|w).$$

Both the marginal and conditional distribution can be conveniently characterized. First, following from multivariate normal theory, the conditional distribution of the malformation latent variables given the foetal weight vector $M_{ij}^*|W_{ij}$ is also normal with mean:

$$\mu_{m|w_{ij}} = \mu_{m_{ij}} + \left(\frac{\sigma_m}{\sigma_w}\right)\left(\frac{\rho + (n_i - 1)\rho_{wm}}{1 + (n_i - 1)\rho_w}\right)\bar{e}_{w_i} + \left(\frac{\sigma_m}{\sigma_w}\right)\left(\frac{\rho + \rho_{wm}}{1 - \rho_w}\right)(e_{w_{ij}} - \bar{e}_{w_i})$$

which depends upon the average litter weight residual, $\bar{e}_{w_i} = (\bar{W}_i - \mu_{w_{ij}})$, and the individual foetal weight residuals, $e_{w_{ij}} - \bar{e}_{w_i}$, where $e_{w_{ij}} = (W_{ij} - \mu_{w_{ij}})$. As a result, the conditional distribution for the observed malformation indicator given foetal weights can be described by a correlated probit model:

$$
\begin{aligned}
P(M_{ij} = 3 | W_{ij}) &= P(M_{ij}^* > h_2 | W_{ij}) \\
&= 1 - \Phi\Big(\frac{h_2 - \mu_{m|w_{ij}}}{\sigma_{m|w_i}}\Big), \\
P(M_{ij} = 2 \text{ or } M_{ij} = 3 | W_{ij}) &= P(M_{ij}^* > h_1 | W_{ij}) \\
&= 1 - \Phi\Big(\frac{h_1 - \mu_{m|w_{ij}}}{\sigma_{m|w_i}}\Big),
\end{aligned}
$$

where $\mu_{m|w_{ij}}$ and $\sigma_{m|w_i}$ are the mean and variance of $M_{ij}^* | W_{ij}$. Not all parameters of this model are estimable, but the model can be reparameterised to a fully estimable form:

$$
\begin{aligned}
P(M_{ij} = 3 | W_{ij}) &= \Phi\Big(\beta_{01} + \beta_1 d_i + \beta_2 \bar{e}_{w_i} + \beta_3 (e_{w_{ij}} - \bar{e}_{w_i})\Big), \\
P(M_{ij} = 2 \text{ or } M_{ij} = 3 | W_{ij}) &= \Phi\Big(\beta_{02} + \beta_1 d_i + \beta_2 \bar{e}_{w_i} + \beta_3 (e_{w_{ij}} - \bar{e}_{w_i})\Big).
\end{aligned}
$$

where average litter weight and individual weight residuals are included naturally as covariates in the model. The interpretation of this model is that both the cluster average and the animal-specific residuals from the continuous outcome determine the ordinal probabilities if one allows for separate intra-cluster and intra-animal correlations in the latent variable formulation. The parameters β in the conditional probit model are directly related to variance and correlation parameters in the underlying latent variable model. A drawback, however, is that the β parameters do not have a marginal interpretation in the probit model.

Parameter estimation of the bivariate model proceeds in two steps, applying the GEE methodology to each component of the conditional model separately. First, a correlated regression of foetal weight on dose and other covariates is fit. Next, a correlated probit regression of malformation conditional on weight with dose, average and individual foetal weight residuals and other covariates is fit.

6.2   Model for Bivariate Data of a Mixed Nature

In this section we describe a model for simultaneously observed continuous and ordinal endpoints, based on a Plackett-Dale approach. The Plackett-Dale idea has been used by Molenberghs, Geys and Buyse (2001) to assess the validation of surrogate endpoints

in randomized experiments with a binary surrogate and a continuous true endpoint or vice versa. Geys *et al.* (2001) extended this idea to the context of teratology experiments where rodents are clustered within litters. We propose an extension of the Plackett-Dale idea to jointly model ordinal and continuous outcomes in a clustered framework.

6.2.1 No Clustering

First, suppose that all littermates are independent. Let us denote the continuous cumulative distribution of the weight outcome as $F_{W_{ij}}$, and the discrete cumulative distribution of the malformation outcome as $F_{M_{ij}}$. We assume a normal distribution for the continuous outcome W_{ij} with mean μ_{ij} and variance σ_{ij}^2, and a multinomial distribution for the ordinal outcome M_{ij} with $\pi_{l,ij}$ the cumulative probability $P(M_{ij} \le l)$ of observing a malformation of degree smaller or equal to l. The dependence between malformation status and foetal weight can be defined using a global cross-ratio at cutpoint (w, m):

$$\psi_{ij}(w, m) = \frac{F_{W_{ij}, M_{ij}}(w, m)\left[1 - F_{W_{ij}}(w) - F_{M_{ij}}(m) + F_{W_{ij}, M_{ij}}(w, m)\right]}{\left[F_{W_{ij}}(w) - F_{W_{ij}, M_{ij}}(w, m)\right]\left[F_{M_{ij}}(m) - F_{W_{ij}, M_{ij}}(w, m)\right]}.$$

Using this relationship, the joint cumulative distribution $F_{W_{ij}, M_{ij}}$ can be written as function of the marginal distributions and the global cross-ratio (Plackett 1965):

$$F_{W_{ij}, M_{ij}}(w, m) = \begin{cases} \frac{1 + \left[F_{W_{ij}}(w) + F_{M_{ij}}(m)\right][\psi_{ij} - 1] - S\left(F_{W_{ij}}(w), F_{M_{ij}}(m), \psi_{ij}\right)}{2[\psi_{ij} - 1]} & \psi_{ij} \ne 1, \\ F_{W_{ij}}(w) F_{M_{ij}}(m) & \psi_{ij} = 1, \end{cases}$$

with

$$S(F_{W_{ij}}, F_{M_{ij}}, \psi_{ij}) = \sqrt{\left[1 + (\psi_{ij} - 1)(F_{W_{ij}} + F_{M_{ij}})\right]^2 + 4\psi_{ij}(1 - \psi_{ij})F_{W_{ij}} F_{M_{ij}}}.$$

Note that, for every cutpoint (w, m), a global cross-ratio ψ_{ij} is obtained $(m = 1, \ldots, c - 1)$. And thus, for every cutpoint (w, m), a different underlying Plackett distribution is assumed. Assuming a constant odds ratio $\psi_{ij}(w, m) \equiv \psi_{ij}$, there is a single underlying Plackett distribution.

Based on the cumulative distribution function $F_{W_{ij}, M_{ij}}(w, m)$, a bivariate Plackett density function $g_{ij}(w, m)$ for joint continuous-ordinal outcomes is derived. Let $f_{W_{ij}|M_{ij}}(w|m) = \partial P(W_{ij} \le w | M_{ij} = m)/\partial w$, for every $m = 1, \ldots, c$, be the conditional density of the continuous outcome given the ordinal outcome. Defining

$g_{ij}(w, m) = f_{W_{ij}|M_{ij}}(w|m) \times P(M_{ij} = m)$, or

$$g_{ij}(w,m) = \begin{cases} \frac{\partial}{\partial w}\left(F_{W_{ij},M_{ij}}(w,1)\right) & m = 1, \\\\ \frac{\partial}{\partial w}\left(F_{W_{ij},M_{ij}}(w,m)\right) - \frac{\partial}{\partial w}\left(F_{W_{ij},M_{ij}}(w,m-1)\right) & m = 2,\ldots,c-1, \\\\ f_{W_{ij}}(w) - \frac{\partial}{\partial w}\left(F_{W_{ij},M_{ij}}(w,c-1)\right) & m = c, \end{cases}$$

leads to specifying the density function $g_{ij}(w, m)$ by:

$$g_{ij}(w,m) = \begin{cases} \frac{f_{W_{ij}}(w)}{2}\left[1 - d(w,m)\right] & m = 1, \\\\ \frac{f_{W_{ij}}(w)}{2}\left[d(w,m-1) - d(w,m)\right] & m = 2,\ldots,c-1, \\\\ \frac{f_{W_{ij}}(w)}{2}\left[1 + d(w,m-1)\right] & m = c. \end{cases}$$

with

$$d(w,m) = \frac{1 + F_{W_{ij}}(w)(\psi_{ij} - 1) - F_{M_{ij}}(m)(\psi_{ij} + 1)}{S(F_{W_{ij}}(w), F_{M_{ij}}(m), \psi_{ij})}. \tag{6.1}$$

One can show that the function $g_{ij}(w, m)$ satisfies the classical density properties:

(i) $g_{ij}(w, m) \geq 0$ for all possible values of w and m,

(ii) $\int \sum_{m=1}^{c} g_{ij}(w, m)dw = \int f_{W_{ij}}(w)dw = 1$.

Further, note that the density function $g_{ij}(w, m)$ also factorizes as a product of the marginal density of the continuous variable $f_{W_{ij}}(w)$ and the conditional density of the ordinal outcome given the continuous variable $f_{M_{ij}|W_{ij}}(m|w)$. For illustration, we displayed the joint density function $g_{ij}(w, m)$ in Figure 6.1 for different values of the odds ratio ψ, in case the continuous outcome is normally distributed with mean $\mu = 1.0$ and standard deviation $\sigma = 0.02$ and the ordinal outcome has three levels ($m = 1, 2, 3$) with probability $0.3, 0.5$ and 0.2, respectively. Some interesting special cases are obtained when the two outcomes are independent ($\psi_{ij} = 1$), perfectly negatively associated ($\psi_{ij} = 0$) and perfectly positively associated ($\psi_{ij} = \infty$).

1. In the first case, when weight and malformation are independent, the function $g_{ij}(w, m)$ reduces to:

$$g_{ij}(w, m) = f_{W_{ij}}(w) f_{M_{ij}}(m).$$
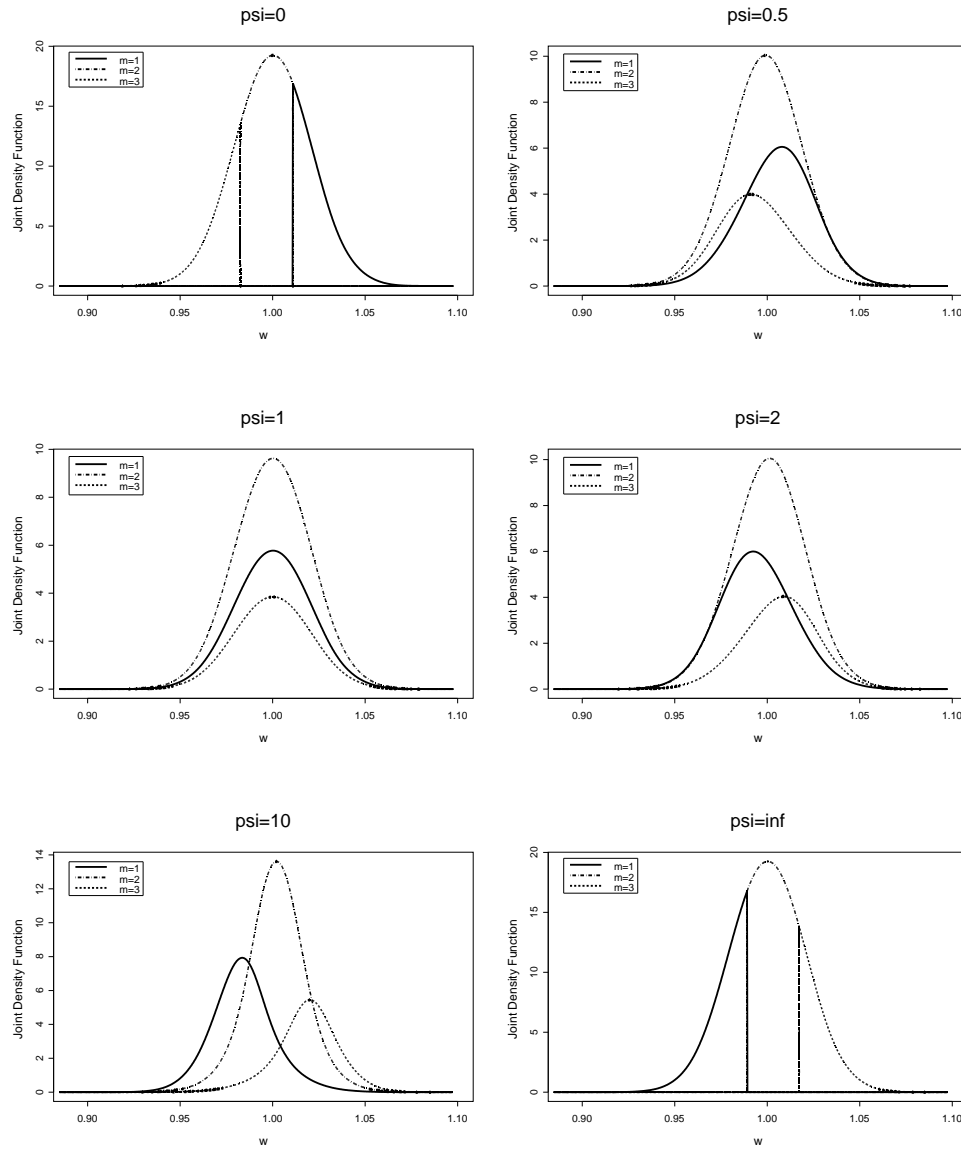
Figure 6.1: *Joint density function of a normally distributed continuous outcome W with mean $\mu = 1.0$ and standard deviation $\sigma = 0.02$ and an ordinal outcome M (m = 1, 2, 3) with probabilities $P(M = 1) = 0.3$, $P(M = 2) = 0.5$, $P(M = 3) = 0.2$, for different dependencies between the two outcomes.*

2. When weight and malformation are perfectly negative correlated, then the function $S(F_{W_{ij}}(w), F_{M_{ij}}(m), \psi_{ij})$ reduces to $|1 - F_{M_{ij}}(m) - F_{W_{ij}}(w)|$ and as a result:

$$g_{ij}(w,1) = \begin{cases} 0 & \text{if } 1 - F_{M_{ij}}(1) \geq F_{W_{ij}}(w), \\ f_{W_{ij}}(w) & \text{if } 1 - F_{M_{ij}}(1) < F_{W_{ij}}(w), \end{cases}$$

$$g_{ij}(w,m) = \begin{cases} 0 & \text{if } 1 - F_{M_{ij}}(m) \geq F_{W_{ij}}(w), \\ f_{W_{ij}}(w) & \text{if } 1 - F_{M_{ij}}(m) < F_{W_{ij}}(w) \\ & \quad \text{and } 1 - F_{M_{ij}}(m-1) > F_{W_{ij}}(w), \\ 0 & \text{if } 1 - F_{M_{ij}}(m-1) \leq F_{W_{ij}}(w), \end{cases}$$

$$g_{ij}(w,c) = \begin{cases} f_{W_{ij}}(w) & \text{if } 1 - F_{M_{ij}}(c-1) > F_{W_{ij}}(w), \\ 0 & \text{if } 1 - F_{M_{ij}}(c-1) \leq F_{W_{ij}}(w). \end{cases}$$

Thus, foetuses with no malformation have high probability to have a high birth weight, while foetuses with a full malformation are more likely to have a low birth weight.

3. In the last case, when weight and malformation are perfectly positive correlated ($\psi_{ij} = \infty$), define $\psi^* = 1/\psi$. The function $S(F_{W_{ij}}, F_{M_{ij}}, \psi_{ij})$ can be rewritten using this reparameterization as:

$$\begin{aligned} S(F_{W_{ij}}, F_{M_{ij}}, \psi_{ij}) &= \sqrt{\left(\psi_{ij}^* + (1 - \psi_{ij}^*)(F_{W_{ij}} + F_{M_{ij}})\right)^2 + 4(\psi_{ij}^* - 1)F_{W_{ij}}F_{M_{ij}}}/\psi_{ij}^* \\ &= S^*(F_{W_{ij}}, F_{M_{ij}}, \psi_{ij}^*)/\psi_{ij}^* \end{aligned}$$

Using this expression, we can now calculate

$$\begin{aligned} g_{ij}(w,1) &= \lim_{\psi_{ij}^* \to 0} \frac{f_{W_{ij}}(w)}{2}\left(1 - \frac{\psi_{ij}^* + F_{W_{ij}}(w)(1 - \psi_{ij}^*) - F_{M_{ij}}(1)(1 + \psi_{ij}^*)}{S^*(F_{W_{ij}}(w), F_{M_{ij}}(1), \psi_{ij}^*)}\right) \\ &= \frac{f_{W_{ij}}(w)}{2}\left(1 - \frac{F_{W_{ij}}(w) - F_{M_{ij}}(1)}{|F_{W_{ij}}(w) - F_{M_{ij}}(1)|}\right) \\ &= \begin{cases} 0 & \text{if } F_{W_{ij}}(w) > F_{M_{ij}}(1), \\ f_{W_{ij}}(w) & \text{if } F_{W_{ij}}(w) \leq F_{M_{ij}}(1). \end{cases} \end{aligned}$$

And analogously,

$$
g_{ij}(w,m) = \begin{cases}
0 & \text{if } F_{W_{ij}}(w) > F_{M_{ij}}(m), \\
f_{W_{ij}}(w) & \text{if } F_{W_{ij}}(w) \leq F_{M_{ij}}(m) \\
& \quad \text{and } F_{M_{ij}}(m-1) \leq F_{W_{ij}}(w), \\
0 & \text{if } F_{W_{ij}}(w) \leq F_{M_{ij}}(m-1),
\end{cases}
$$

$$
g_{ij}(w,c) = \begin{cases}
f_{W_{ij}}(w) & \text{if } F_{W_{ij}}(w) > F_{M_{ij}}(c-1), \\
0 & \text{if } F_{W_{ij}}(w) \leq F_{M_{ij}}(c-1).
\end{cases}
$$

In this case, the joint distribution reflects that small values for the ordinal outcome correspond with small values for the continuous outcome, while high values for the ordinal outcome are very likely linked with high values for the continuous outcome.

Dose-response models that incorporate litter- and foetus-specific covariates can be considered for each of the parameters by using appropriate link functions. The parameters $\mu_{ij}, \sigma_{ij}^2, \pi_{l,ij}$ and ψ_{ij} for individual j in cluster i can be modelled by:

$$
\eta_{ij} \equiv \begin{pmatrix}
\mu_{ij} \\
\ln(\sigma_{ij}^2) \\
\text{logit}(\pi_{1,ij}) \\
\vdots \\
\text{logit}(\pi_{c-1,ij}) \\
\ln(\psi_{ij})
\end{pmatrix} = \mathbf{X}_{ij}\boldsymbol{\beta}, \tag{6.2}
$$

where $\mathbf{X}_{ij}$ is a design matrix for the jth foetus in the ith cluster and $\boldsymbol{\beta}$ is a vector of unknown regression parameters. A key difference between the factorization model and the plackett-dale model is that one gets an estimate of the association in the latter approach. In addition, this method allows to directly model the bivariate intrafoetus association as function of exposure and other covariates of interest. The generality of the design matrix is an important advantage of this approach, as the assumption of constant variance and constant association are often not tenable. Indeed, in real data settings, correlation structures are likely to change with exposure (Kupper, Portier, Hogan and Yamato 1986) and ignoring this can lead to bias in the estimates or loss of efficiency (Ryan 2000).

6.2.2   Clustered Outcomes

Often however, littermates are not independent, but clustered within litters. In the case of clustering, we use a pseudo-likelihood function, rather than considering the full likelihood. The pseudo-likelihood approach was proposed by Arnold and Strauss (1991), and also found in Connolly and Liang (1988), Liang and Zeger (1986) and Le Cessie and Van Houwelingen (1994). The principal idea is to replace a numerically challenging joint density by a simpler function that is a suitable product of ratios of likelihoods of subsets of the variables. As such, a working correlation structure is introduced and we avoid the computational complexity of the full likelihood distribution of each cluster i, i.e., $f(w_{i1}, \ldots, w_{in_i}, m_{i1}, \ldots, m_{in_i})$. In addition, the pseudo-likelihood method provides a way to deal with nuisance parameters (Liang and Zeger 1989, Arnold and Strauss 1991). Arnold and Strauss (1991) established consistency and asymptotic normality of the pseudo-likelihood estimator. Thus, valid inference can be obtained from such models.

First Definition

In a first step, the association between weight and malformation outcomes for an individual foetus is modelled explicitly, but for outcomes from different littermates independence is taken as a working assumption.

$$p\ell_1 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ln g(w_{ij}, m_{ij}). \tag{6.3}$$

This leads to consistent estimates (Arnold and Strauss 1991, Geys, Molenberghs and Lipsitz 1998, le Cessie and Van Houwelingen 1994). A sandwich variance estimator is then used to adjust for potential overdispersion due to intralitter correlation. This approach acknowledges the fact that, while the association between different outcomes on the same littermate is often of scientific interest, the association due to clustering within litters is usually considered a nuisance. Indeed, in quantitative risk assessment primary interest lies in the probability that a foetus is affected, either by malformation or by low birth weight. This probability is a function only of the mean parameters and the bivariate intrafoetus association.

Estimates of the regression parameters are obtained by solving the estimating equations $\boldsymbol{U}(\boldsymbol{\beta}) = 0$. Grouping all parameters $\mu_{ij}, \sigma_{ij}^2, \pi_{l,ij}$ and ψ_{ij} for individual j in cluster i in a vector $\boldsymbol{\theta}_{ij}$, and grouping all vectors $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\eta}_{ij}$ for the ith cluster in

$\boldsymbol{\theta}_i$ and $\boldsymbol{\eta}_i$, respectively, the estimating equations can be written as:

$$
\begin{aligned}
\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\theta}_i}\right)^{-T} \left(\frac{\partial \ln g_{ij}(w,m)}{\partial \boldsymbol{\theta}_i}\right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{X}_{ij}^T (\boldsymbol{T}_{ij}^{-1})^T \left(\frac{\partial \ln g_{ij}(w,m)}{\partial \boldsymbol{\theta}_i}\right) = 0,
\end{aligned}
$$

where

$$
\boldsymbol{T}_{ij}^{-1} = \begin{pmatrix}
1 & 0 & 0 & \ldots & & 0 & 0 \\
0 & \sigma^2 & 0 & \ldots & & 0 & 0 \\
0 & 0 & \pi_1(1-\pi_1) & 0 & & \ldots & 0 \\
\vdots & \vdots & & \ddots & \ddots & & \vdots \\
0 & 0 & \ldots & & 0 & \pi_{c-1}(1-\pi_{c-1}) & 0 \\
0 & 0 & 0 & \ldots & & 0 & \psi_i
\end{pmatrix}.
$$

The first derivatives of $\ln g_{ij}(w,m)$ with respect to the parameters $\boldsymbol{\theta}_i$ are obtained by combining the derivatives of $d(w,m)$, as defined in (6.1), with those of $S(w,m,\psi)$. Expressions for the derivatives are given below. For simplicity, we have omitted the cluster-level index i and the foetus-level index j:

$$
\begin{aligned}
\frac{\partial f_W(w)}{\partial \mu} &= f_W(w)\frac{(w-\mu)}{\sigma^2} \\
\frac{\partial f_W(w)}{\partial \sigma^2} &= \frac{f_W(w)}{2}\left(\frac{(w-\mu)^2 - \sigma^2}{\sigma^4}\right) \\
\frac{\partial F_W(w)}{\partial \mu} &= \int_{-\infty}^{x} f_W(w)\frac{(w-\mu)}{\sigma^2} dw \\
&= -f_W(w) \\
\frac{\partial F_W(w)}{\partial \sigma^2} &= \int_{-\infty}^{x} \frac{f_W(w)}{2}\left(\frac{(w-\mu)^2 - \sigma^2}{\sigma^4}\right) dw \\
&= -\left(\frac{(w-\mu)}{\sigma^2}\right)\frac{f_W(w)}{2}
\end{aligned}
$$

$$\frac{\partial F_M(m)}{\partial \pi_k} = \begin{cases} 1 & \text{if } k \leq m \\ 0 & \text{if } k > m \end{cases}$$

$$\frac{\partial S}{\partial \mu} = \frac{1}{2S}\left[2[1 + (\psi - 1)(F_W + F_M)](\psi - 1)\frac{\partial F_W}{\partial \mu} + 4\psi(1 - \psi)\frac{\partial F_W}{\partial \mu}F_M\right]$$

$$\frac{\partial S}{\partial \sigma^2} = \frac{1}{2S}\left[2[1 + (\psi - 1)(F_W + F_M)](\psi - 1)\frac{\partial F_W}{\partial \sigma^2} + 4\psi(1 - \psi)\frac{\partial F_W}{\partial \sigma^2}F_M\right]$$

$$\frac{\partial S}{\partial p_k} = \frac{1}{2S}\left[2[1 + (\psi - 1)(F_W + F_M)](\psi - 1)\frac{\partial F_M}{\partial p_k} + 4\psi(1 - \psi)F_W\frac{\partial F_M}{\partial p_k}\right]$$

$$\frac{\partial S}{\partial \psi} = \frac{1}{2S}\Big[2[1 + (\psi - 1)(F_W + F_M)](F_W + F_M)$$
$$+4(1 - \psi)F_W F_M - 4\psi F_W F_M\Big]$$

$$\frac{\partial d(w,m)}{\partial \mu} = \frac{1}{2}\frac{\partial f_W(w)}{\partial \mu}\left\{1 - \frac{1 + F_W(w)(\psi - 1) - F_M(m)(\psi + 1)}{S(F_W(w), F_M(m), \psi)}\right\}$$
$$-\frac{f_W(w)}{2S}\left\{\frac{\partial F_W(w)}{\partial \mu}(\psi - 1)\right\}$$
$$+\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \mu}\left\{1 + F_W(w)(\psi - 1) - F_M(m)(\psi + 1)\right\}$$

$$\frac{\partial d(w,m)}{\partial \sigma^2} = \frac{1}{2}\frac{\partial f_W(w)}{\partial \sigma^2}\left\{1 - \frac{1 + F_W(w)(\psi - 1) - F_M(m)(\psi + 1)}{S(F_W(w), F_M(m), \psi)}\right\}$$
$$-\frac{f_W(w)}{2S}\left\{\frac{\partial F_W(w)}{\partial \sigma^2}(\psi - 1)\right\}$$
$$+\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \sigma^2}\left\{1 + F_W(w)(\psi - 1) - F_M(m)(\psi + 1)\right\}$$

$$\frac{\partial d(w,m)}{\partial \pi_k} = \frac{f_W(w)}{2S}\left\{\frac{\partial f_M(m)}{\partial \pi_k}(\psi + 1)\right\}$$
$$+\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \pi_k}\left\{1 + F_W(w)(\psi - 1) - F_M(m)(\psi + 1)\right\}$$

$$\frac{\partial d(w,m)}{\partial \psi} = -\frac{f_W(w)}{2S}\left\{F_W(w) - F_M(m)\right\}$$
$$+\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \psi}\left\{1 + F_W(w)(\psi - 1) - F_M(m)(\psi + 1)\right\}$$

Arnold and Strauss (1991) showed that under regularity conditions, the pseudo-likelihood estimator $\hat{\boldsymbol{\beta}}$, obtained by maximizing the log-pseudo likelihood function (6.3) is consistent and asymptotically normal with estimated covariance matrix:

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \Big(\sum_{i=1}^{N} \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}}\Big)^{-1} \Big(\sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i(\boldsymbol{\beta})^T\Big) \Big(\sum_{i=1}^{N} \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}}\Big)^{-1}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

Obviously, the advantage of the proposed pseudo-likelihood function lies in the fact that the association parameters of interest to the researchers are included while others can be ignored, and as such, adds to the computational simplicity of the methodology. Using ideas of robust inference, the proposed method is consistent and asymptotically normal.

Second Definition

If one is interested in the amount of clustering as well, the pseudo-likelihood function (6.3) can be extended by including the products of the bivariate probabilities of (i) two weight outcomes for two different individuals in the same cluster, (ii) two malformation outcomes for two different individuals in the same cluster and (iii) a weight and malformation outcome for two different individuals in the same cluster. This leads to the following log-pseudo likelihood function:

$$
\begin{aligned}
p\ell_2 =\ & \sum_{i=1}^{N}\sum_{j=1}^{n_i} \ln g_1(w_{ij}, m_{ij}) + \sum_{i=1}^{N}\sum_{j<k} \ln g_2(w_{ij}, w_{ik}) \\
& + \sum_{i=1}^{N}\sum_{j<k} \ln g_3(m_{ij}, m_{ik}) + \sum_{i=1}^{N}\sum_{j\neq k}^{n_i} \ln g_4(w_{ij}, m_{ik}),
\end{aligned}
\tag{6.4}
$$

with for example g_1, g_3, g_4 bivariate Plackett distributions, characterized by potentially different odds ratios, and g_2 a bivariate normal distribution. The four different types of contributions captured in the model are depicted in Figure 6.2. A non-equivalent specification of pseudo-likelihood (6.4) is

$$
\begin{aligned}
p\ell_2^* =\ & \sum_{i=1}^{N} \frac{1}{3}\sum_{j=1}^{n_i} \ln g_1(w_{ij}, m_{ij}) + \sum_{i=1}^{N} \frac{1}{3(n_i-1)}\sum_{j<k} \ln g_2(w_{ij}, w_{ik}) \\
& + \sum_{i=1}^{N} \frac{1}{3(n_i-1)}\sum_{j<k} \ln g_3(m_{ij}, m_{ik}) + \sum_{i=1}^{N} \frac{1}{3(n_i-1)}\sum_{j\neq k}^{n_i} \ln g_4(w_{ij}, m_{ik}).
\end{aligned}
\tag{6.5}
$$

The factors before each contribution in the log-pseudo-likelihood function (6.5) corrects for the fact that each response occurs several times in the ith contribution of

indiv. j indiv. k

w_{ij} (g_2) w_{ik} Weight

(g_1) (g_4) (g_1)

m_{ij} (g_3) m_{ik} Malformation
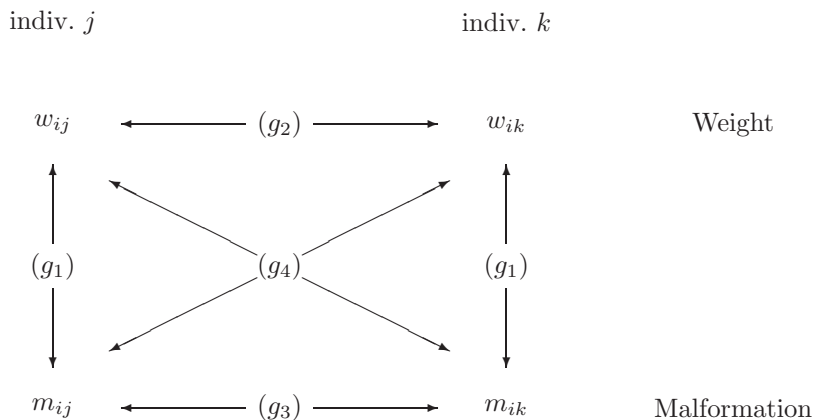
Figure 6.2: *Four different types of contributions for the pseudo-likelihood.*

the pseudo-likelihood. It ensures that the pseudo-likelihood reduces to full likelihood under independence. Indeed, under independence, (6.5) simplifies to:

$$p\ell_2^* = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ln f_W(w_{ij}) f_M(m_{ij}).$$

When all clusters are equal in size, the pseudo-likelihood estimator β and its variance-covariance matrix remain the same, no matter whether we use the pseudo-likelihood function $p\ell_2$ or $p\ell_2^*$. However, when n_i is random, we have to assume that n_i and (w_{ij}, m_{ij}) are independent, given dose level d_i for the ith cluster, to ensure that the expected value of $\boldsymbol{U}_i^*(\boldsymbol{\beta}) = \frac{\partial p\ell_{2i}^*}{\partial \boldsymbol{\beta}}$ equals zero. Geys, Molenberghs and Lipsitz (1998) recommend to correct for the number of times each response occurs in the pseudo-likelihood definition, and thus to use $p\ell_2^*$ rather than $p\ell_2$, when interest lies in the marginal parameters. Asymptotically, $p\ell_2^*$ is more efficient than $p\ell_2$ for estimating main effects. For estimating the association parameter however, the use of $p\ell_2$ is advised. If interest is combined, and one type of analysis should be chosen, $p\ell_2$ might be preferable above $p\ell_2^*$. A simulation study to investigate the efficiency of the pseudo-likelihood in this setting is interest of further research.

6.3 Application to Quantitative Risk Assessment

In the area of developmental toxicity, an important goal is the quantitative risk assessment, i.e., deriving a safe dose of exposure. Recent techniques for risk assessment are based on fitting dose-response models and estimating the dose corresponding to a certain increase in risk of an adverse effect over background, i.e., benchmark dose. In case of multiple outcomes, the outcomes are often examined individually, using appropriate methods to account for correlation, and regulation of exposure is then based on the most sensitive outcome. This approach assumes that protecting against the most sensitive outcomes protects against all adverse outcomes. It has been found, however, that a clear pattern of correlation exists between weight and malformation outcomes (Ryan *et al.* 1991), so that risk assessment based on a joint model may be more appropriate. For risk assessment purposes, the joint probability that an individual foetus is malformed and/or of low foetal weight must be characterized.

The standard approach to quantitative risk assessment based on dose-response modelling requires the specification of an adverse event, along with its risk expressed as a function of dose. The risk function $r(d)$ can be defined as the probability that a foetus has a high malformation level or a low birth weight at dose level d. In other words, for the jth foetus in the ith cluster:

$$r(d) = P(W_{ij} \leq W_c \text{ or } M_{ij} \geq M_c | d),$$

where W_c and M_c, respectively, denote some cutoff values that determines foetal weight low enough and malformation severe enough to be considered adverse. This expression can be rewritten using the univariate discrete distribution function $F_{M_{ij}}$ and the joint continuous-discrete distribution function $F_{W_{ij}, M_{ij}}$:

$$r(d) = 1 - F_{M_{ij}}(M_c - 1) + F_{W_{ij}, M_{ij}}(W_c, M_c - 1).$$

Based on this probability, a common measure for the excess risk over background can be calculated, which is used to estimate a safe level of exposure, i.e., the benchmark dose or lower effective dose estimation.

The main advantage of the Plackett-Dale model is that it lends itself in a natural way to quantitative risk assessment, since it allows separate dose-response models for the malformation and weight outcomes, while taking into account the correlation due to clustering as well as the intrafoetus association.

6.4   Illustration: DYME Data

In this section we illustrate the methods described through the analysis of developmental toxicity data from a study of diethylene glycol dimethyl ether (DYME) in mice. The data are described in Chapter 2. Scientific interest lies in the effects of dose on the overall risk due to malformation and low birth weight, i.e., the probability that an individual foetus is malformed or of low birth weight.

   The methodology used to analyse the data is as follows. The joint distribution of malformation and low birth weight is described using the Plackett-Dale model. Parameter estimates are obtained through the pseudo-likelihood function (6.3), to account for clustering within litters. To fit the data well, a selection procedure to obtain proper functional forms to describe the dependence of all main and association parameters is considered. The final model is used to estimate the univariate and joint risks due to the dose exposure. Based on this risk estimate, safe dose levels of exposure are calculated.

Model Selection

For risk assessment to be reliable, models should fit the data well. A stepwise approach for model selection is a standard methodology. To select the most parsimonious model for the DYME data, we rely on the adjusted pseudo-likelihood ratio test statistic, defined by Geys, Molenberghs and Ryan (1999) and described in Chapter 3. The most complex model that we considered is:

$$
\begin{aligned}
\mu_i &= \alpha_0 + \alpha_1 d_i + \alpha_2 (n_i - \bar{n}), \\
\mathrm{logit}(\pi_{1,i}) &= \beta_0 + \beta_1 d_i + \beta_2 (n_i - \bar{n}), \\
\mathrm{logit}(\pi_{2,i}) &= \gamma_0 + \gamma_1 d_i + \gamma_2 (n_i - \bar{n}), \\
\ln(\sigma_i^2) &= \delta_0 I(d_i = 0) + \delta_1 I(d_i = 62.5) + \\
&\quad\; \delta_2 I(d_i = 125) + \delta_3 I(d_i = 250) + \delta_4 I(d_i = 500), \\
\ln(\psi_i) &= \kappa_1 I(m = 1) + \kappa_2 I(m = 2) + \kappa_3 d_i,
\end{aligned}
$$

where μ_i is the mean weight outcome and $\pi_{m,i} = P(M_{ij} \leq m)$, $m = 1, 2$, the cumulative malformation probabilities. For model fitting, doses d_i are rescaled to the $[0, 1]$ interval. Possibly, foetal weight and malformation are affected by litter size. Therefore, to adjust for the litter size, a covariate for the deviation of the overall average litter size $(n_i - \bar{n})$ is incorporated into the model. Note that the considered model is more flexible than the standard proportional odds model. However, a careful

Table 6.1: DYME study. Model selection procedure for DYME study in mice.

Model	Description	# pars.	pl
1.	Different weight variances across doses; Common d trend on ψ; Different ψ depending on cutpoint m	17	-553.48
2.	Linear d trend on σ^2; Common d trend on ψ; Different ψ depending on cutpoint m	14	-551.09
3.	Constant σ^2; Common d trend on ψ; Different ψ depending on cutpoint m	13	-540.73
4.	Linear d trend on σ^2; Different ψ depending on cutpoint m	13	-549.99
5.	Linear d trend on σ^2; Constant ψ	12	-546.53

Comparison	df	$\bar{\lambda}$	G_a^{2*}	(p-value)
1-2	4	6.46	0.74	(0.946)
2-3	1	6.64	3.12	(0.077)
2-4	1	0.74	0.73	(0.393)
4-5	1	2.97	9.16	(0.002)

modelling approach has to be considered with constraints on the parameters in order to ensure that $\pi_{1,i} \leq \pi_{2,i}$. In order to estimate the parameters, the pseudo-likelihood function is maximized using a general nonlinear optimisation method with general nonlinear constraints on the parameters.

The model not only specifies dose trends on the mean parameters, but also allows dose effects on weight variability (σ_i^2) and odds ratio (ψ_i) between weight and malformation. Different models on σ_i^2 and ψ_i are considered, in order to find the most appropriate model. In Table 6.1, a summary of model selection is presented. Initially, different weight variances for each of the dose groups were assumed, but a more parsimonious model with a linear dose trend on the weight variance is accepted (Model 2: $\ln(\sigma_i^2) = \delta_0 + \delta_1 d_i$). The dose effect parameter on the variances is only borderline significant (Model 3 vs Model 2), but will be kept in the model. The linear

dose trend on the log odds ratio can be deleted without substantial decrease in fit (Model 4). In contrast, the use of different log odds ratios, depending on the malformation cutpoint is a significant improvement of the fit $(\ln(\psi_i) = \zeta_0 + \zeta_1 I(m = 2))$. Hence, the model can be reduced to Model 4. Other parameters cannot be removed without a substantial decrease in fit. Therefore, we accept Model 4 as the final model on the foetal weight variance σ_i^2 and odds ratio ψ_i.

Final Model

The results of fitting the clustered bivariate model to the DYME data, using the final Model 4, are summarized in Table 6.2, in the column labelled *Plackett-Dale*. The table displays the parameter estimates, standard errors and Z scores for the average weight (μ_i) and cumulative malformation probabilities $(\pi_{m,i})$, the weight variance (σ_i^2) and odds ratio (ψ_i). For foetal weight, the dose coefficient is significantly negative, but there appears to be little effect of litter size on weight. The foetal weight variance also decreases with dose. The dose coefficient for the cumulative malformation probabilities is significantly negative, and the significantly positive coefficient of litter size suggests that larger litters have a smaller malformation risk. The estimated odds ratios are less than 1, indicating a negative association between weight and malformation; the small value reflects the strength of the association. Having no malformation is about 2.5 times as likely as having a minor or full malformation, for foetuses with a normal birth weight. Having a full malformation is about 4 times as likely as having no or a minor malformation, for foetuses with a low birth weight.

Observed and predicted values in Figure 6.3 (a)-(b) show how the model fits the dose-specific averages for both univariate outcomes. The predicted weight outcome fits the data quite well. The lower points and curve denote, respectively, the observed and predicted probability of full malformation $(M_c = 3)$ and the upper points and curve correspond to the probabilities of minor or full malformation $(M_c = 2)$. Estimates for both malformation outcomes are in strong agreement with the observed probabilities.

Comparison with Factorization Model

Results can be compared with these of Catalano (1997), which are displayed in Table 6.2, in the column labeled *Cond-C*. Catalano (1997) used a factorization model that conditions on the continuous outcome. Three main differences are noted. First, in the factorization model, different intercepts and common slope parameters for the malformation model are assumed, as in the proportional odds model. Different slopes on the probit scale would complicate the model. Secondly, the regression parameters

Table 6.2: DYME study. Model fitting results from different approaches. Estimates (standard errors; Z values).

|	Coefficient	Plackett-Dale		Cond-C	
		Estim.	(s.e.; Z-score)	Estim.	(s.e.; Z-score)
Weight Model:					
	Intercept	1.014	(0.014; 72.4)	1.024	(0.019;55.3)
	Dose	-0.444	(0.028;-15.9)	-0.472	(0.035;-13.6)
	Litter Size	0.0002	(0.003; 0.1)	-0.002	(0.004;-0.6)
Malformation Status Model:					
π_1:	Intercept	3.462	(0.357; 9.7)	2.30	(0.207;11.12)
	Dose	-5.677	(0.987;-5.8)	-4.05	(0.452;-4.47)
	Litter Size	0.106	(0.060; 1.8)	0.076	(0.039;1.91)
π_2:	Intercept	5.021	(0.323; 15.5)	2.78	(0.215;12.90)
	Dose	-7.236	(0.623;-11.6)	-4.05	(0.452;-4.47)
	Litter Size	0.106	(0.100; 1.1)	0.076	(0.039;1.91)
Mean weight resid.		-	-	3.93	(1.200;3.27)
Weight resid.		-	-	2.12	(1.091;1.94)
Mean weight resid. $\times$ litter size		-	-	0.068	(0.390;0.17)
Foetal Weight Variance:					
Dose 0.000		0.014	(0.002; 6.1)	-	-
0.125		0.013	(0.002; 7.6)	-	-
0.250		0.012	(0.001; 9.4)	-	-
0.500		0.010	(0.001; 9.4)	-	-
1.000		0.007	(0.002; 4.1)	-	-
Foetal Weight / Malformation Association:					
ψ_1:		0.405	(0.222; 1.8)	-	-
ψ_2:		0.250	(0.188; 1.3)	-	-

[a] The dose unit is mg/kg/day/500
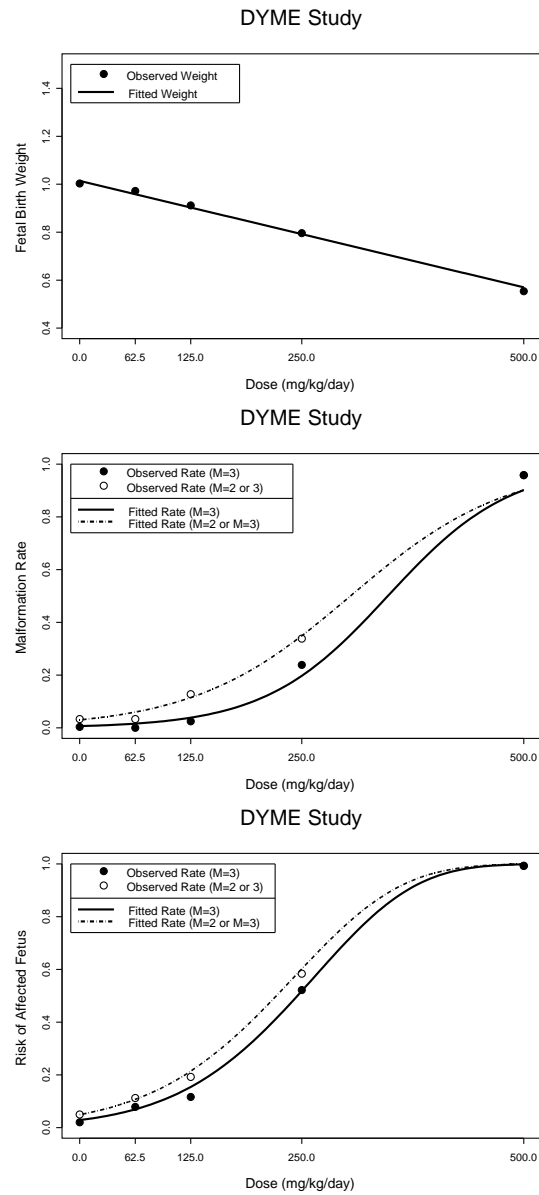
Figure 6.3: *DYME study. Observed and estimated outcomes. From top to bottom, (a) foetal weight model, (b) malformation status model, (c) joint weight-malformation risk model.*

of the malformation model do not have a direct marginal interpretation because of the non-linearity of the link function relating the conditional mean of the ordinal response to the covariates. The Plackett-Dale model allows separate dose-response models for malformation and weight outcomes, while taking into account the correlation due to clustering as well as the intrafoetus association. Finally, the model induces association by adding the residuals from the marginal model as covariates in the conditional model, however these association parameters are not directly interpretable. Thus, while taking into account the dependence between weight and malformation, the conditional models do not directly specify a measure of association. To overcome this problem, one needs a joint model that incorporates the correlation between outcomes directly, such as the proposed Plackett-Dale model. The Plackett-Dale model uses the odds ratio as intrafoetus measure of association, which is readily interpretable.

Extended Pseudo-Likelihood

Note that because of the working independence assumption, there is no estimated intralitter correlation for foetal weight or for malformation. Although, when interested in these associations, the model could be extended by use of the pseudo-likelihood function (6.4) or (6.5). In these expression, the function g_1 is defined as before. The function g_2 is assumed a bivariate normal density. We define g_3 as a Plackett density for two ordinal outcomes, with a constant odds ratio ψ_2. Also g_4 is specified as a bivariate Plackett distribution for a mixed continuous/ordinal outcome, with a constant odds ratio ψ_3. Parameter estimates with robust standard errors, arising from the different pseudo-likelihood definitions (6.3,6.4,6.5), are displayed in Table 6.3.

All parameter estimates arising from the extended pseudo-likelihood function, are similar to the ones obtained with the simpler version that assumes independence between the littermates $(p\ell_1)$. In general, we gain some efficiency when using the extended pseudo-likelihood, which is computationally more complex. In line with the conclusions of Geys, Molenberghs and Lipsitz (1998), we see that the $p\ell_2^*$ is more efficient for main effects in comparison with $p\ell_2$. On the contrary, $p\ell_2$ turns out to be superior for the estimation of the association parameters. With the extended version, some additional parameter estimates are obtained: the odds ratio between two malformation outcomes for two different individuals from the same cluster is estimated as 2.397; the correlation ρ between two weight outcomes for two different individuals in the same litter is 0.569; and the odds ratio between a weight and malformation outcome for two different individuals in the same cluster is estimated as 0.527. Note that the assumption of a constant odds ratio in g_3 and g_4 can be

Table 6.3: DYME study. Parameter estimates with robust standard errors, obtained from different pseudo-likelihood functions.

Effect	Param.	pl_1		pl_2		pl_2^*	
		Estim.	(s.e.)	Estim.	(s.e.)	Estim.	(s.e.)
μ	α_0	1.014	(0.014)	1.008	(0.018)	1.014	(0.005)
	α_1	-0.444	(0.028)	-0.430	(0.032)	-0.446	(0.012)
	α_2	0.0002	(0.003)	0.002	(0.004)	0.0002	(0.001)
σ^2	δ_0	-4.293	(0.164)	-4.363	(0.144)	-4.301	(0.060)
	δ_1	-0.659	(0.354)	-0.648	(0.309)	-0.639	(0.137)
π_1	β_0	3.462	(0.357)	3.378	(0.220)	3.480	(0.207)
	β_1	-5.677	(0.987)	-5.461	(0.586)	-5.720	(0.505)
	β_2	0.106	(0.060)	0.111	(0.051)	0.100	(0.043)
π_2	γ_0	5.021	(0.323)	5.051	(0.306)	5.056	(0.302)
	γ_1	-7.236	(0.623)	-7.134	(0.410)	-7.295	(0.600)
	γ_2	0.106	(0.100)	0.111	(0.078)	0.100	(0.059)
ψ_1	κ_1	-0.904	(0.348)	-0.867	(0.746)	-0.899	(0.282)
	κ_2	-0.482	(0.151)	-0.495	(0.579)	-0.481	(0.297)
ρ	ζ	-	-	1.216	(0.135)	1.292	(0.128)
ψ_2	ν	-	-	0.869	(0.114)	0.874	(0.507)
ψ_3	ι	-	-	-0.613	(0.227)	-0.641	(0.269)

relaxed by assuming the odds ratio to depend on the malformation cutpoint.

Quantitative Risk Assessment

In order to calculate a benchmark dose based on the joint model, we first need to specify the risk of an adverse effect, i.e., the probability that an individual foetus is malformed or of low birth weight. Therefore, we need to define a weight below which a foetus can be considered as being of "low foetal weight". However, there is no standard definition. A common approach is to specify the cutoff point W_c as two standard deviations below the control average foetal weight. As such, the cutoff level for determining low foetal weight is equal to $W_c = 0.7816$ g, corresponding to a 1.77% low birth weight rate in control animals. Further, we consider two definitions of risk, depending on the cutpoint M_c for what is considered as a "malformed" foetus. Either we define it as the probability that a foetus has a minor or full malformation ($M_c = 2$), or a low foetal weight. Alternatively, we define it as the probability that a foetus has a full malformation ($M_c = 3$), or has a low foetal weight. These two risk functions are displayed in Figure 6.3 (c). The risks are evaluated at the average litter size ($n_i = \bar{n}$). The lower points and curve in this graph correspond to the probability $P(W \leq W_c, M \geq 2)$. The higher points and curve correspond to the probability $P(W \leq W_c, M \geq 3)$. In case of the first definition, the estimated risk function is higher and steeper compared with risk when defined under the second definition. But, for both definitions of "risk", the risk functions seem to fit the data very well. Note that the definition of the risk is very important for the estimation of a safe dose. In Figure 6.4 it is displayed how different definitions of the risk, based on different cutpoints for the weight and malformation outcome, can affect the estimation of the benchmark dose.

Table 6.4 shows the benchmark doses corresponding to the 1% and 10% excess risk for Model 4, as well as the 1% and 10% lower limit LED. We also added the corresponding quantities, calculated from univariate versions of the model. We can compare the joint modelling approach with the traditional approach for multiple outcomes in which the lower of the individual malformation and foetal weight LEDs is used as an overall LED. The minimum of the two LEDs is more than 20% higher than those obtained using the bivariate methods that incorporate the relationship between the two outcomes. Since both univariate outcomes suffer from a substantial risk, focusing attention to a single response or a collapsed outcome would overestimate the safe dose. The joint model yields higher risks, since it accounts for the correlation between both outcomes. Thus, ignoring the correlation between the two outcomes
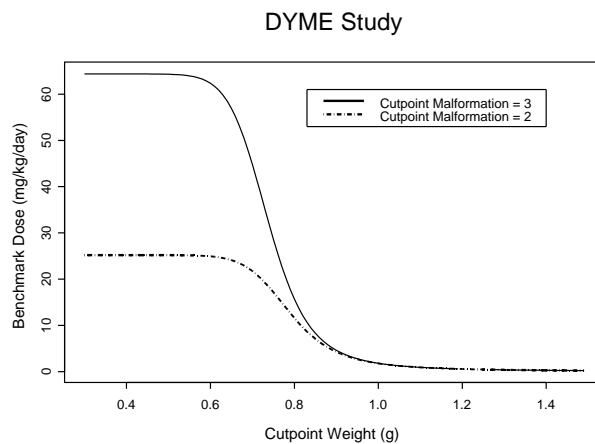
DYME Study

Figure 6.4: *DYME study. Benchmark dose estimation (q=1%) as function of the cutpoints for weight and malformation in the definition of the risk.*

leads to too high and hence inappropriate safe doses.

Note that joint outcome models based on factorization, such as the model of Catalano (1997), do not provide a clear way to compute joint benchmark doses (Geys *et al.* 2001, Regan and Catalano 1999) because there is no direct access to the marginal distributions. This is the major drawback of the conditional models, since the joint benchmark doses are of primary interest. In contrast, the Plackett-Dale approach models the marginal distributions directly, lending itself in a natural way to quantitative risk assessment.

6.5   Discussion

In this chapter, a modelling approach for the analysis of clustered data with both continuous and ordinal outcomes has been considered. The model was applied to a developmental toxicity study (DYME in mice), and used for quantitative risk assessment. The Plackett-Dale method uses a global odds ratio as intrafoetus measure of association. The association between foetal weight and malformation are directly modelled, and can be modelled in a general way, including covariate information.

A problem in the quantitative risk assessment, that has received only minor attention, is the choice of suitable cutpoints. For binary outcomes, the definition of an

Table 6.4: DYME study. Benchmark dose and lower effective dose estimation (mg/kg/day).

q	Model	$M_C = 2$		$M_C = 3$	
		BMD_q	LED_q	BMD_q	LED_q
1%	Joint	13.91	12.66	20.05	17.72
	Continuous	23.70	23.37	23.70	23.37
	Ordinal	26.26	25.52	66.90	60.26
10%	Joint	88.64	83.88	108.13	101.85
	Continuous	118.99	118.26	118.99	118.26
	Ordinal	136.00	133.71	199.59	193.29

adverse health effect is intuitively clear. This is less the case for continuous and/or ordinal responses. We considered a dichotomised version of the continuous outcomes to determine a benchmark dose, after fitting the dose-response model based on the continuous outcome. The risk of low foetal weight was based on a cutoff level for determining a low weight extreme enough to be considered an adverse event. Because of arbitrariness of the cutpoint, estimating a BMD from a continuous response has led to much discussion (Bosch *et al.* 1996, Crump 1984, 1995, Gaylor and Slikker 1990, Kavlock *et al.* 1995, Kodell and West 1993). Several efforts have been made to develop risk assessment for continuous outcomes (West and Kodell 1993). However, the used definition of risk, defined in terms of the tail of the background (control) distribution, seems quite plausible.

Although the method is presented in the specialized field of developmental toxicity, the methodology is applicable in a general clustered or even correlated data setting with a continuous and ordinal outcome. Thus, use of the proposed modelling approach is far beyond the developmental toxicity context.

Chapter 7

A Hierarchical Modelling Approach for Risk Assessment in Developmental Toxicity Studies

In the previous chapter, interest lied in the joint analysis of clustered multiple outcome data, motivated by developmental toxicity applications (Fitzmaurice and Laird 1995, Gueorguieva and Agresti 2001, Molenberghs and Ryan 1999, Regan and Catalano 1999, Aerts *et al.* 2002). So far however, we have restricted ourselves to the outcomes on the viable foetuses only. Ideally, a model should take the complete correlated hierarchical structure of the data into account. A hierarchical Bayesian method is proposed to this effect. Such a model can serve as a basis for quantitative risk assessment.

The analysis of developmental toxicity data raises a number of challenges (Molenberghs *et al.* 1998), as indicated before. Since deleterious events can occur at several points in development, an interesting aspect lies in the staging or hierarchy of possible adverse foetal outcomes (Williams and Ryan 1996). Figure 1.1 illustrates the data structure. Because of the toxic insult, the developing foetus is at risk of foetal death. If the foetus survives the entire gestation period, growth reduction such as low birth weight may occur. The foetus may also exhibit one or more types of malformation. In addition, as mentioned before, offspring of the same mother behave more alike than

those of another mother, i.e., the litter or cluster effect. Thus, responses on different foetuses within a cluster are likely to be correlated.

Ultimately, analysis of the developmental toxicity data must account for the entire hierarchical, multivariate and clustered nature of the data. So far, one has tackled the challenges in this setting only partly each time making different restrictive assumptions, e.g., restricting to viable foetuses only. However, litters with a lot of malformed foetuses are likely to have more death foetuses than litters with good foetal health. As a result, litter size (number of viable foetuses) may be informative. A classical way to account for the litter size is to include it as a covariate in modelling the response rates (Williams 1987, Rai and Van Ryzin 1985, Catalano *et al.* 1993) and then calculating a safe dose at an "average" litter size, thereby avoiding the need for direct adjustment. However, several perspectives for modelling these data in a direct way can be considered. One may look at the hierarchical structure, and consider cluster size as a random variable. Xu and Prorok (2003) developed a non-parametric procedure for the analysis of exchangeable clustered binary data when the cluster size is a random variable. As such, one acknowledges the stochastic nature of the litter size. Indeed, variation in the litter size is an extra source of variability in the data that must be accounted for. We may also consider a missing data model, because the unformed foetuses are not observable. Dunson (1998) proposed a multiple imputation scheme to estimate the number of missing foetuses. In this context, the random cluster size perspective seems more natural than does the missing data perspective. Alternatively, Williamson *et al.* (2003) proposed a weighted generalized estimating equation approach for fitting marginal models to clustered data when litter size is informative. Although this method accounts for the cluster size, it does not allow for modelling the cluster size as a function of covariates of interest.

In this chapter, we propose a Bayesian model dealing with the hierarchical structure in two stages. At the first stage, we express the probability that a foetus is non-viable. At the second stage, we model the probability that a viable foetus has a malformation and/or suffers from low birth weight. At each stage we account for the intralitter correlation. The intractability of the likelihood function has led various authors to propose a host of alternative estimation methods rather than carrying out maximum likelihood estimation. A full likelihood procedure can be replaced by quasi-likelihood methods (McCullagh and Nelder 1989), pseudo-likelihood (Arnold and Strauss 1991) or generalized estimating equations (Liang and Zeger 1986). Generalized linear random-effects models or hierarchical Bayesian models (McCulloch and Searle 2001) are attractive alternative modelling appraoches. We opted for the latter approach and used Gibbs sampling (Zeger and Karim 1991) to deal with complex

integrations. The contents of this chapter can also be found in Faes *et al.* (2004c).

7.1 Modelling Approach

We propose a Bayesian hierarchical modelling framework for the joint analysis of foetal death and malformation/weight among the viable foetuses. Let L denote the total number of dams, and hence litters, in the study. For the ith litter ($i = 1, \ldots, L$), let k_i be the number of implants. Let r_i indicate the number of foetal deaths in cluster i. The number of viable foetuses, i.e., the litter size, is $n_i \equiv k_i - r_i$. The outcome measured on the viable foetuses is denoted $\boldsymbol{y}_{ij} = (w_{ij}, m_{ij}), j = 1, \ldots, n_i$, with w_{ij} the foetal birth weight and $m_{ij} = 1$ when foetus j in cluster i has a malformation, 0 otherwise.

To define a model for the developmental toxicity data, the underlying hierarchy of the data is used. At the bottom level, the foetuses surviving the entire gestation period are at risk for low birth weight and/or malformation. Assume that $\boldsymbol{y}_{ij}$ satisfies

$$\boldsymbol{y}_{ij}|n_i \sim F(\boldsymbol{y}_{ij}|\zeta, n_i),$$

i.e., conditional on the litter size, $\boldsymbol{y}_{ij}$ follows a pre-specified distribution F, possibly depending on covariates, such as the dose level, and parameterised through a vector ζ of unknown parameters. Further, the litter size n_i is a random variable, possibly depending on the dose level and other covariates of interest. Indeed, a toxic insult may result in a foetal death. The litter size n_i can be modelled through modelling the number of non-viable foetuses $r_i \equiv k_i - n_i$. Assume the number of non-viable foetuses r_i to follow a distribution G depending on a vector ψ of unknown parameters, i.e.,

$$r_i \sim G(r_i|\psi, k_i).$$

Let $f(\boldsymbol{y}_{ij}|\zeta, n_i)$ and $g(r_i|\psi, k_i)$ denote the density functions corresponding to the distributions F and G, respectively.

Because of the hierarchy in the model, it lends itself naturally to estimate the parameters using Bayesian techniques (Box and Tiao 1992, Gelman 1995). In the Bayesian framework, unknown parameters are also considered as random, and all inference is based on their distribution conditional on the observed data, i.e., the posterior distribution.

It is obvious that different choices for F and G will lead to different models. The distribution G is crucial in the calculation of the marginal model for $\boldsymbol{y}_{ij}$. Next, a possible choice for the distributions F and G in the developmental toxicity setting is given.

7.1.1 Stage 1: A Model for the Death Outcome

In the first step, a toxic insult early in gestation may result in a foetal death. This effect of dose d_i on cluster i with k_i implants can be described using the density $g(r_i|\psi, k_i)$. Considering the foetuses within a litter as independent, one could assume that r_i satisfies a binomial density

$$\binom{k_i}{r_i} \pi_{R_i}^{r_i} (1 - \pi_{R_i})^{k_i - r_i},$$

with π_{R_i} the probability of a dead foetus in litter i, depending on the dose. To account for clustering, a random effects model in which each litter has a random parameter is considered. Skellam (1948), Kleinman (1973) and Williams (1975) assume the probability of death π_{R_i} of any foetus in litter i to come from a beta distribution with parameters a_i and b_i:

$$\frac{\pi_{R_i}^{a_i - 1}(1 - \pi_{R_i})^{b_i - 1}}{B(a_i, b_i)}, \qquad (0 \leq \pi_{R_i} \leq 1),$$

where $B(.,.)$ denotes the beta function. This leads to the well-known beta-binomial distribution.

The probability mass function $g(r_i|\psi, k_i)$ can be expressed directly in terms of the mean and correlation parameters, i.e., $g(r_i|\pi_{R_i}, \rho_{R_i}, k_i)$. The mean of this distribution is

$$\mu_{R_i} = k_i \pi_{R_i} = k_i \frac{a_i}{a_i + b_i},$$

and the variance is

$$\sigma_{R_i}^2 = k_i \pi_{R_i}(1 - \pi_{R_i})[1 + \rho_{R_i}(k_i - 1)],$$

with ρ_{R_i} the intra-litter correlation, which is the correlation between two binary responses of litter i.

7.1.2 Stage 2: A Model for Malformation and Weight

When a foetus survives the entire gestation period, it is still at risk for low foetal weight and malformation. A distribution for the combined continuous and binary outcomes, i.e., $f(w_{ij}, m_{ij}|\zeta, n_i)$ must be specified. Based on the mixed outcome probit model of Regan and Catalano (1999), we propose the following model.

First, assume that littermates are independent. Under a probit model for the binary response M_{ij}, one assumes a latent variable M_{ij}^* to be normally distributed with mean $\gamma_{m_{ij}}$ and unit variance, so that

$$\pi_{m_{ij}} = P(M_{ij} = 1) = P(M_{ij}^* > 0) = \Phi(\gamma_{m_{ij}}),$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. The probability of malformation is related to covariates by expressing $\gamma_{m_{ij}}$ as some parameterised function of the predictors, e.g., the dose level, and the litter size.

For the bivariate response (W_{ij}, M_{ij}), a bivariate normal distribution is assumed for the observed weight and the latent malformation variable for foetus j in litter i:

$$f(w_{ij}, m_{ij}^*) = \phi_2(w_{ij}, m_{ij}^* | \mu_{w_{ij}}, \sigma_{w_{ij}}^2, \gamma_{m_{ij}}, 1, \rho_{mw_{ij}}),$$

where $\rho_{mw_{ij}}$ is the intrafoetus correlation between the latent malformation and the weight outcomes. As a result, the joint distribution of the bivariate foetal weight and binary malformation outcome can be written as

$$
\begin{aligned}
f(w_{ij}, z_{ij}) &= f_w(w_{ij}) \times f_{m|w}(m_{ij} | w_{ij}) \\
&= \phi(w_{ij} | \mu_{w_{ij}}, \sigma_{w_{ij}}^2) \times \pi_{m|w_{ij}}^{M_{ij}} (1 - \pi_{m|w_{ij}})^{1 - M_{ij}},
\end{aligned}
$$

where $\pi_{m|w_{ij}} = \Phi(\gamma_{m|w_{ij}})$ is the conditional expectation of the binary malformation outcome $E(M_{ij} | W_{ij})$. From bivariate normal theory,

$$\gamma_{m|w_{ij}} = \frac{\gamma_{m_{ij}} + \rho_{mw_{ij}} \frac{w_{ij} - \mu_{w_{ij}}}{\sigma_{w_{ij}}}}{(1 - \rho_{mw_{ij}}^2)^{1/2}},$$

with $\pi_{m_{ij}} = \Phi(\gamma_{m_{ij}})$ the marginal expectation $E(M_{ij})$.

In case of clustering, litter-specific parameters are considered to account for the correlation among the outcomes. Random effects on the mean foetal birth weight $\mu_{w_{ij}}$ and on the malformation parameter $\gamma_{m_{ij}}$ are introduced

$$
\begin{aligned}
\mu_{w_{ij}} &\sim N(\mu_{W_i}, \sigma_{\mu_i}^2) \\
\gamma_{m_{ij}} &\sim N(\gamma_{M_i}, \sigma_{\gamma_i}^2),
\end{aligned}
$$

such that the bivariate distribution for foetal weight and binary malformation equals

$$
\begin{aligned}
f(w_{ij}, m_{ij}) &= f_w(w_{ij}) \times f_{m|w}(m_{ij} | w_{ij}) \\
&= \phi(w_{ij} | \mu_{W_i}, \sigma_{w_{ij}}^2 + \sigma_{\mu_i}^2) \times \pi_{M|W_{ij}}^{M_{ij}} (1 - \pi_{M|W_{ij}})^{1 - M_{ij}},
\end{aligned}
$$

with $\pi_{M|W_{ij}} = \Phi(\gamma_{M|W_{ij}})$ the conditional expectation for the binary malformation outcome $E(M_{ij} | W_{ij})$. We can derive that

$$\gamma_{M|W_{ij}} = \frac{\frac{\gamma_{M_i}}{\sqrt{1 + \sigma_{\gamma_i}^2}} + \rho_{MW_{ij}} \frac{w_{ij} - \mu_{W_i}}{\sqrt{\sigma_{w_{ij}}^2 + \sigma_{\mu_i}^2}}}{(1 - \rho_{MW_{ij}}^2)^{1/2}},$$

with $\rho_{MW_{ij}}$ the intrafoetus correlation between the latent malformation and the weight outcomes. The marginal expectation for the binary malformation outcome $E(M_{ij})$ equals $\pi_{Z_i} = \Phi(\gamma_{M_i}/(1+\sigma^2_{\gamma_i}))$. The intra-litter correlation among the weight outcomes equals $\rho_{W_i} = \sigma^2_{\mu_i}/(\sigma^2_{\mu_i}+\sigma^2_{w_{ij}})$. The intra-litter correlation among the latent malformation outcomes equals $\rho_{M_i} = \sigma^2_{\gamma_i}/(1+\sigma^2_{\gamma_i})$.

7.1.3   Dose-Response Model

Dose-response models are specified for the marginal outcomes of interest, i.e., the foetal weight, the probability of malformation, and the probability of death. Each of the univariate outcomes are allowed to vary as functions of dose and other covariates. The dose-response models can generally be written as

$$
\begin{aligned}
\mu_{W_i} &= \boldsymbol{X}'_{a_{ij}}\boldsymbol{\alpha} + (n_i - \bar{n})\gamma_a, \\
\gamma_{M_i} &= \boldsymbol{X}'_{b_{ij}}\boldsymbol{\beta} + (n_i - \bar{n})\gamma_b, \\
\pi_{R_i} &= \exp(\boldsymbol{X}'_{c_{ij}}\boldsymbol{\delta})/(1 + \exp(\boldsymbol{X}'_{c_{ij}}\boldsymbol{\delta})),
\end{aligned}
$$

where $\{X_{a_{ij}}, X_{b_{ij}}, X_{c_{ij}}\}$ are the foetus- and/or litter-specific covariates with regression parameters $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}\}$. Often with developmental toxicity data, the assumption that variances and correlations are constant across dose groups is not appropriate. Therefore we allow the variances and correlations to vary with dose and possible other covariates as well. Thus, dose-response models for the parameters $\{\rho_R, \rho_{MW}, \sigma_w, \sigma_\mu, \sigma_\gamma\}$ can be written as well, using appropriate transformations:

$$
\begin{aligned}
\rho_{ij} &= (\exp(\boldsymbol{X}'_{t_{ij}}\boldsymbol{\tau}) - 1)/(\exp(\boldsymbol{X}'_{t_{ij}}\boldsymbol{\tau}) + 1), \\
\sigma_{ij} &= \exp(\boldsymbol{X}'_{s_{ij}}\boldsymbol{\zeta}),
\end{aligned}
$$

with $\rho_{ij} = \{\rho_R, \rho_{MW}\}$ and $\sigma_{ij} = \{\sigma_w, \sigma_\mu, \sigma_\gamma\}$.

The need for numerical integration can be avoided by casting the model into a Bayesian framework and by resorting to the Gibbs sampler (Zeger and Karim 1991). In addition to the specified model, hyperprior distributions for the regression parameters need to be selected. We follow the recommendations of Besag, Green, Higdon and Mengersen (1995) in using proper but highly dispersed hyperprior distributions. The hyperpriors chosen on the regression parameters for this analysis were $N(0, 10^6)$. We expect these priors to have minimal influence on the final conclusions of our analysis.

7.2 Application to Quantitative Risk Assessment

The primary goal of these studies is to determine a safe level of exposure. We define the combined risk due to a toxic effect as the probability that a foetus is dead or a viable foetus is malformed and/or suffers from low birth weight. This risk can be expressed as

$$
\begin{aligned}
r(d) &= P(\text{death}|d) + P(\text{viable}|d) \times P(\text{malformed or low weight}|\text{viable}, d) \\
&= P(R=1|d) + (1 - P(R=1|d)) \times P(M=1 \text{ or } W < W_c|N \geq 1, d) \\
&= \pi_R + (1 - \pi_R) \times P(M=1 \text{ or } W < W_c|N \geq 1, d).
\end{aligned}
$$

For simplicity, we have omitted the cluster-level index i and the foetus-level index j. The joint probability of a malformation or low birth weight is equal to:

$$
\begin{aligned}
P(M=1 \text{ or } &W < W_c|N \geq 1, d) \\
&= \int_1^\infty P(M=1 \text{ or } W < W_c|N = n, d) P(N = n|d) dn
\end{aligned}
$$

and

$$
\begin{aligned}
P(M=1 \text{ or } &W < W_c|N = n, d) \\
&= 1 - \int_{-\infty}^{-\tau} \int_{W_c}^{\infty} \phi_2(W_{ij}, M_{ij}^*; \mu_W(d), 0, \sigma_w(d)^2 + \sigma_\mu^2, 1 + \sigma_\gamma^2, \rho_{MW}(d)) dW_{ij} dM_{ij}^* \\
&= \Phi(\tau) + \Phi_2(-\tau, \omega; \rho_{MW}(d)),
\end{aligned}
$$

where $\tau = \gamma_M(d)/\sqrt{1 + \sigma_\gamma^2(d)}$ and $\omega = (W_c - \mu_W(d))/\sqrt{\sigma_w(d)^2 + \sigma_\mu(d)^2}$ and Φ_2 is the standard bivariate normal distribution function.

The benchmark dose is defined as the level of exposure corresponding to an acceptably small excess risk over background, i.e., the dose satisfying

$$
r^*(d) = \frac{r(d) - r(0)}{1 - r(0)} = q,
$$

with q the prespecified level of increased risk over background, typically specified as 0.01, 1, 5, or 10% (Crump 1984). In the frequentist framework, the benchmark dose calculation is based on the estimated dose-response curve. In the Bayesian approach one could choose to base the benchmark dose calculation on the mean posterior risk curve. This method is illustrated in Figure 7.1. The full line corresponds with the mean posterior risk. But, benchmark dose calculations are no more precise than the data on which they are based. Therefore, rather than calculating a point estimate of the safe dose, one might be interested in the entire posterior distribution of the safe
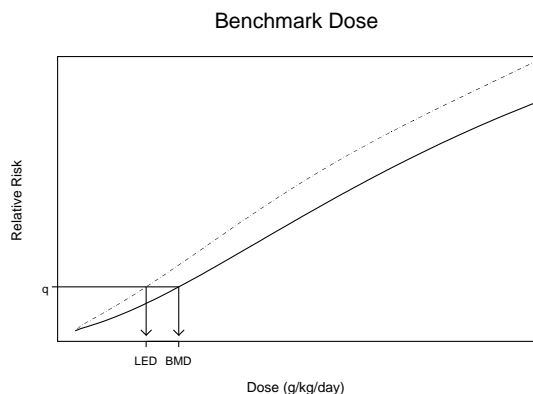
Benchmark Dose

Figure 7.1: *Definition of benchmark dose (BMD) and lower effective dose (LED) at q% of increased risk over background.*

dose. In this way, the researcher could get an idea of the precision of the estimate. Often, one is interested in an upper bound of the benchmark dose to set a safe level of exposure. One can construct a 95% upper credibility limit of the risk function and base the safe dose calculation upon this upper limit. In analogy with the frequentist approach, the lower effective dose is defined as the dose such that the 95% upper credibility limit of the excess risk is equal or greater than the predefined level q. This is illustrated in Figure 7.1. The dashed line corresponds to the 95% upper credibility limit of the posterior risk.

7.3   Illustration: EG Data

Dose-Response Modelling

For risk assessment to be reliable, the dose-response model should fit the data well in all respects. As described in Chapter 4, fractional polynomials are very useful both in the context of dose-response modelling and quantitative risk assessment (Geys *et al.* 1999, Faes *et al.* 2003). With the aim on low dose extrapolation, fractional polynomials are used in this setting.

In order to select a parsimonious model for the data we select a suitable set of dose transformations for each of the three outcomes separately. Model selection is performed using the deviance information criterion (DIC) as proposed by Spiegelhalter *et al.* (1998, 2002) and described in Chapter 3. Smaller values of DIC indicate a better fitting model. Table 7.1 shows that a fractional polynomial of degree $m = 1$,

Table 7.1: EG study. Deviance Information Criterion for the best first and second order fractional polynomials to model the malformation parameter γ_m.

	$m = 1$		$m = 2$	
tranformation	DIC	transformation	DIC	
$1/d^2$	678.4	$(\sqrt{d}, d)$	677.8	
$1/d$	679.9	(d, d^2)	678.4	
$1/\sqrt{d}$	681.0	$(\ln(d), \ln^2(d))$	678.0	
$\ln(d)$	682.2	$(1/\sqrt{d}, 1/d)$	677.9	
$\sqrt{d}$	678.1	$(1/d, 1/d^2)$	678.1	
d	685.3	$(d, d\ln(d))$	678.5	
d^2	692.0	$(\sqrt{d}, \ln(d))$	677.2	
d^3	693.6			

whether represented by $1/d^2, 1/d, 1/\sqrt{d}, \ln(d), d, d^2$ or d^3, is unacceptable as opposed to a fractional polynomial of degree $m = 2$ to model the malformation parameter γ_{M_i}. Table 7.1 tabulates only a selection of the considered two-degree fractional polynomials. None of the other combinations provided a substantial improvement. The fractional polynomial represented by $(\sqrt{d}, \ln(d))$ yields the smallest DIC. A similar approach, applied to the death outcome and weight outcome, suggest a d^2 trend on π_{R_i} and a $\ln(d)$ trend on μ_{W_i}.

The resulting set of transformations is then used to construct more elaborate models that can be scrutinized further by means of the DIC. The most complex model we considered (Model 1) allows the following trends on the malformation, weight and death outcomes:

$$
\begin{aligned}
\mu_{W_i} &= \beta_{0W} + \beta_{1W} \ln(d+1) + \beta_{2W}(n - \bar{n}), \\
\gamma_{M_i} &= \beta_{0Z} + \beta_{1Z}\sqrt{d} + \beta_{2Z}\ln(d+1) + \beta_{3Z}(n - \bar{n}), \\
\text{logit}(\pi_{R_i}) &= \beta_{0R} + \beta_{1R}d^2,
\end{aligned}
$$

where d is the dose, rescaled to the $[0, 1]$ interval. Further, linear d trends on the association parameters

$$\rho_{R_i}, \sigma^2_{\gamma_i}, \sigma^2_{\mu_i}, \rho_{MW},$$

are considered. Table 7.2 summarizes the model selection procedure on the associa-

Table 7.2: EG study. Model selection on the association parameters. A '*' indicates a linear d trend on that parameter. All other effects are kept constant.

Model	ρ_{dth}	σ^2_μ	σ^2_γ	ρ_{MW}	DIC
1	*	*	*	*	-1356.870
2	*	*	*	·	-1358.310
3	*	*	·	·	-1358.490
4	*	·	·	·	-1357.950
5	·	*	·	·	-1360.470
6	·	·	·	·	-1359.540

tion parameters. Based on the deviance information criterion, there is evidence for choosing a model with a constant association between weight and malformation and a constant malformation variance (Model 3). In contrast, there is evidence for choosing a model with a d trend on the weight variance (Model 4). Finally, there seems to be no evidence for the linear d trend on the correlation among death outcomes. As such, we choose Model 5.

Final Model

Parameter estimates obtained from fitting the final model are displayed in Table 7.3. The dose coefficient is significantly negative for foetal birth weight, and the negative coefficient of litter size suggests that larger litters had a higher risk of low foetal birth weight, which is not unexpected due to competition for food resources. The intralitter correlation for weight is substantial, and increases from 0.441 in the control group to 0.644 in the highest dose group. For malformation, there is an increasing dose effect, and there appears to be little effect of litter size on malformation. The intralitter correlation for malformation is also large. The correlation between malformation and birth weight appears to be negative, indicating that foetal malformations are associated with lower foetal weights. For foetal death, there is a significantly positive effect with dose. The intralitter correlation for foetal death is also significantly positive. The 2-dimensional plots in Figures 7.2 to 7.4 show the posterior mean curves together with the 95% credibility intervals of the univariate dose-response curves. All the univariate fits are acceptable. The 3-dimensional plots show the posterior density of the univariate dose-response curves.

Table 7.3: EG study. Posterior mean and standard deviation of the parameters in the final model.

	Effect	Mean	(StDev)
Foetal Weight:			
Mean:	intercept	0.980	(0.013)
	log(dose)	-0.433	(0.037)
	$n_i - \bar{n}$	-0.011	(0.003)
Correlation: 0.000		0.441	(0.063)
0.250		0.492	(0.047)
0.500		0.545	(0.043)
1.000		0.644	(0.069)
Malformation:			
Mean:	intercept	-3.857	(0.631)
	$\sqrt{\text{dose}}$	6.062	(2.529)
	log(dose)	-2.664	(2.995)
	$n_i - \bar{n}$	-0.002	(0.049)
Correlation:		0.691	(0.048)
Foetal Weight / Malformation:			
Correlation:		-0.018	(0.005)
Foetal Death:			
Mean:	intercept	-2.099	(0.153)
	dose^2	0.730	(0.258)
Correlation:		0.069	(0.024)

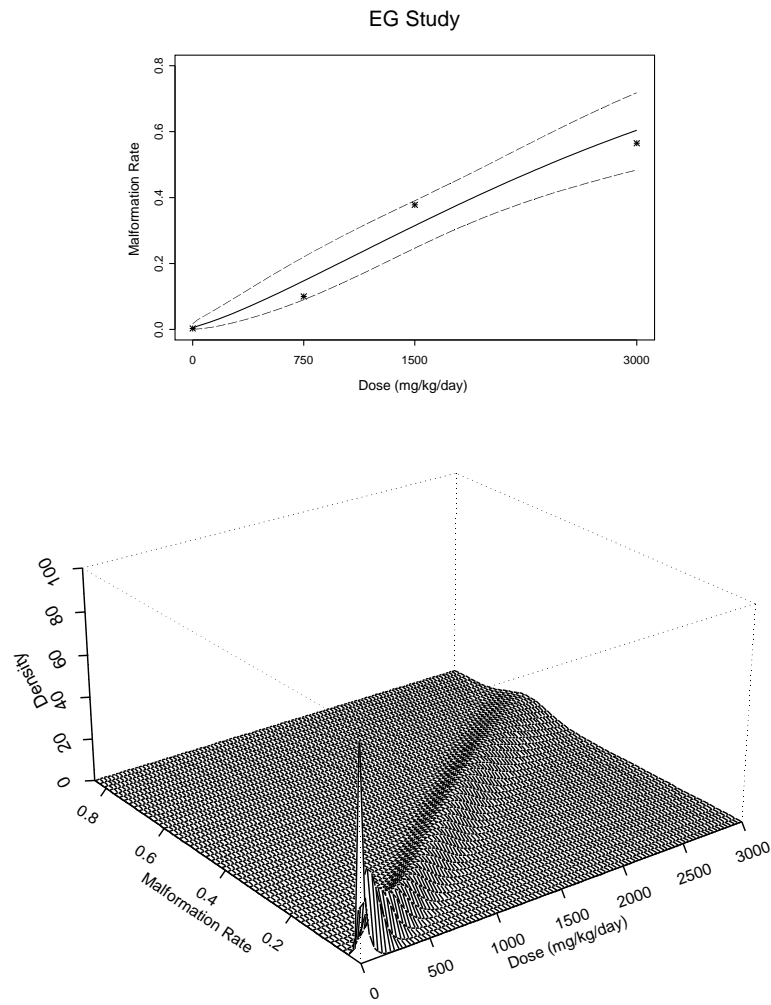[a] The dose unit is mg/kg/day/3000

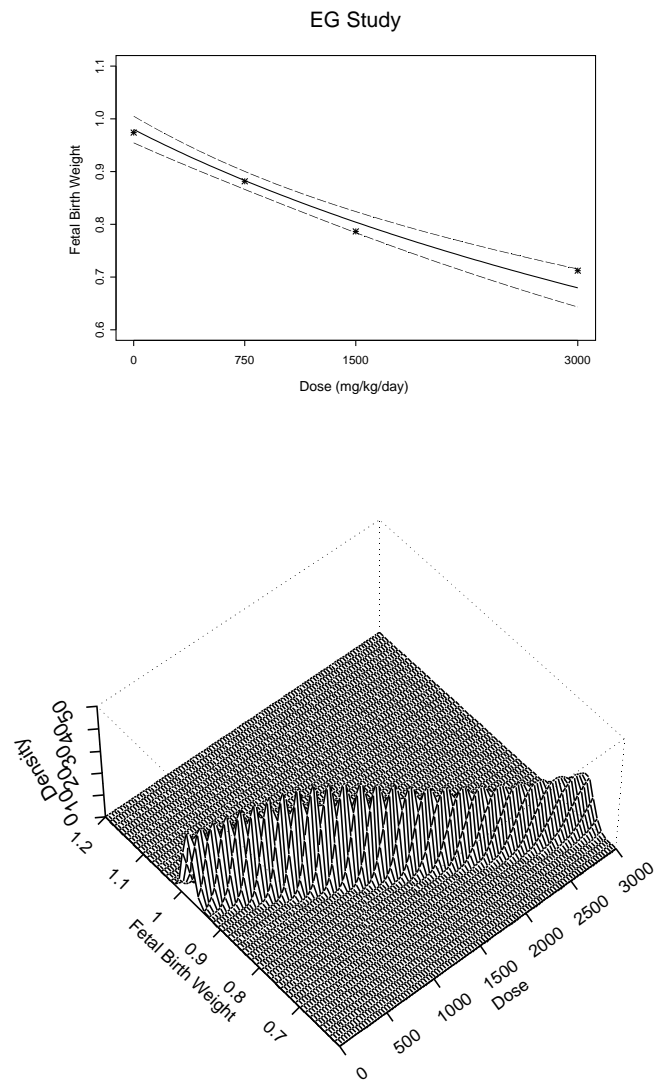Figure 7.2: *EG study. Estimated malformation rates.*
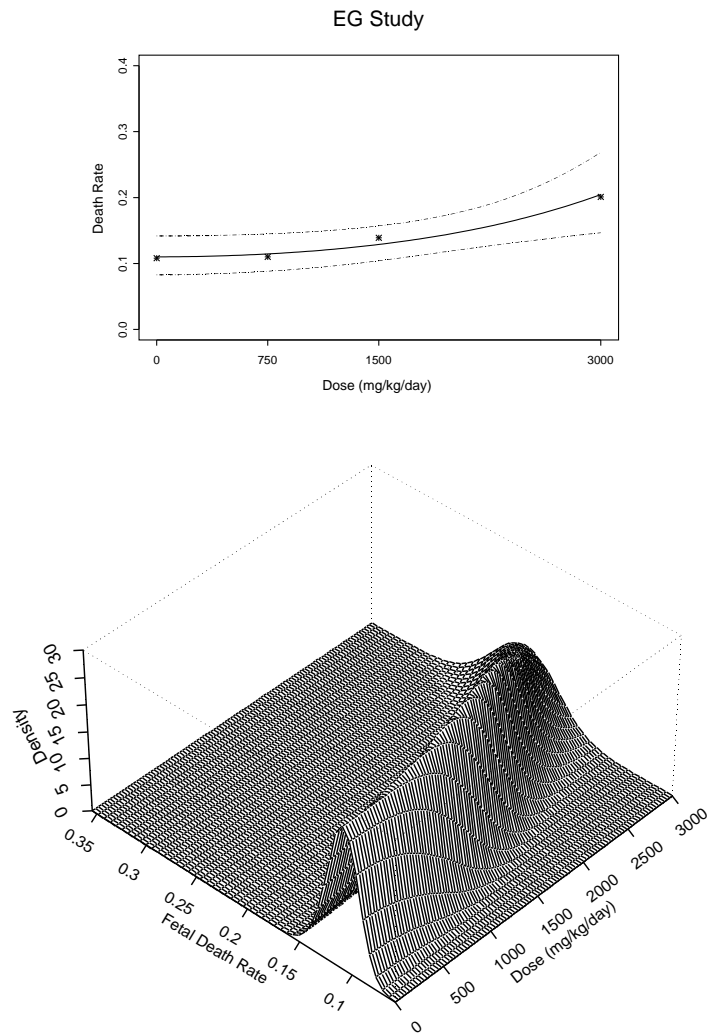
Figure 7.3: *EG study. Estimated foetal weight.*

Figure 7.4: *EG study. Estimated death rate.*

Table 7.4: EG study. Benchmark dose and lower effective dose estimation (mg/kg/day).

Model	BMD		LED	
	$q = 0.01$	$q = 0.05$	$q = 0.01$	$q = 0.05$
Joint	37	190	17	126
Malf	56	299	19	163
Weight	106	383	77	312
Death	1055	2209	878	1843

Quantitative Risk Assessment

To calculate the risk of low birth weight, we need to define a weight below which a foetus can be considered as being of "low foetal weight". Because of the arbitrariness of the cutpoint, estimating a benchmark dose from a continuous response has led to much discussion (Bosch *et al.* 1996, Crump 1984). We specify the cutoff point W_c as two standard errors below the control average foetal weight (Catalano and Ryan 1992). By means of this definition, foetuses that weighed less than 0.777 g are considered to be of low foetal weight, which corresponds to a 3.4% rate in the control animals. The posterior density of the combined risk due to a foetal death, a malformation or low foetal weight is pictured in Figure 7.5. The risk gradually increases when dams are exposed to larger quantities of the toxic substance, before finally reaching an asymptote.

Table 7.4 shows the benchmark dose and lower effective dose corresponding to a 1% and 5% excess risk over background, respectively based upon the posterior mean and 95% upper credibility limit of the risk curve. We also added the corresponding quantities, calculated from univariate risks. The joint model yields more conservative doses. Often, a safe level of exposure is determined separately for each outcome and the lower of the individual outcomes is used as an overall benchmark dose. It is clear that this approach would yield too high estimated safe doses. Therefore, it is necessary to model the full hierarchical data structure when searching for a safe level of exposure.

Instead of calculating a point estimate of the benchmark dose, we can derive the full posterior distribution of the benchmark dose. The posterior distribution of the benchmark dose corresponding to a 1% increase in risk over background is pictured in
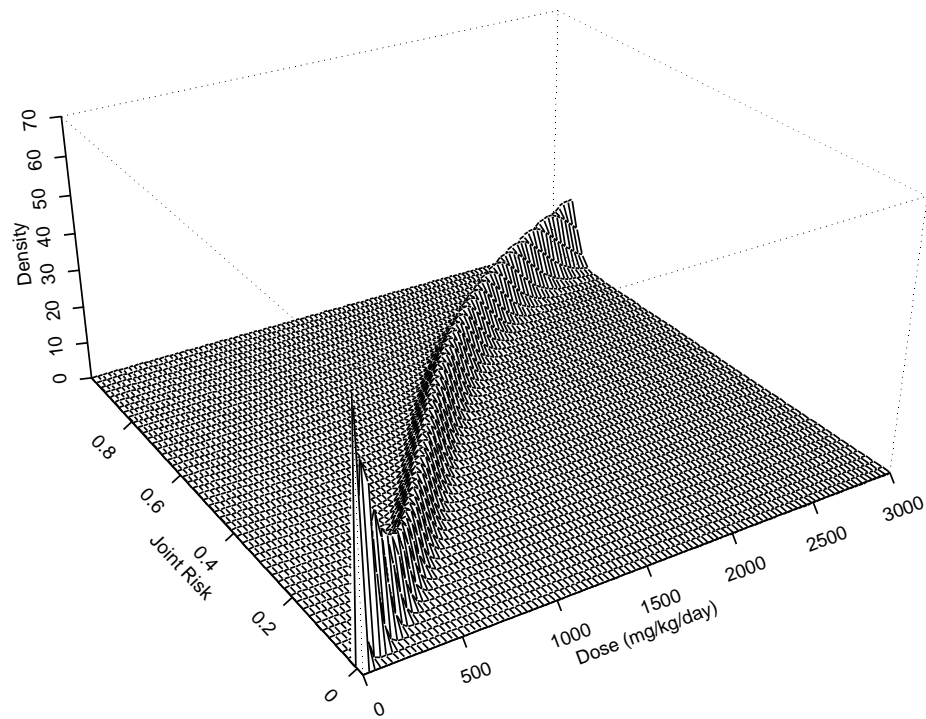
Figure 7.5: *EG study. Posterior density of combined risk due to a foetal death, a malformation or low foetal weight.*
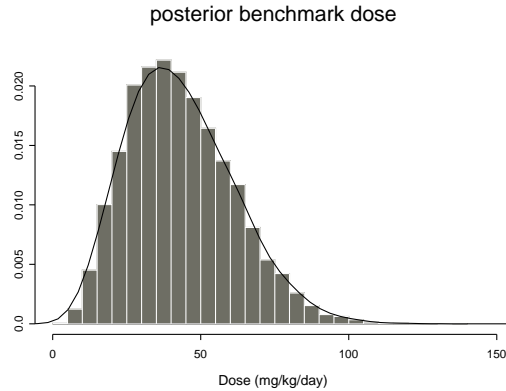
Figure 7.6: *EG study. Posterior distribution of benchmark dose, corresponding to a 1% increase of risk over background.*

Figure 7.6. The posterior mode equals 35 mg/kg/day. Previously obtained estimate of the benchmark dose based on the posterior mean of the risk curve (37 mg/kg/day) lies in between the mode (35 mg/kg/day) and mean (43 mg/kg/day) of the posterior distribution of the benchmark dose. The 95% lower credibility limit of the posterior benchmark dose is equal to 17 mg/kg/day. Calculation of the posterior distribution of the benchmark dose does not only give information about the estimated safe level of exposure, but also its uncertainty and shape of the distribution.

7.4   Discussion

Developmental toxicity studies are complicated by the hierarchical, clustered and multivariate nature of the data. As a consequence, a multitude of modelling strategies have been proposed in literature. Often, focus is only on the outcomes measured on the viable foetuses. However, as observed from the data sets, the number of viable foetuses in a dam, i.e., the litter size, also decreases with increasing dose levels. Thus, a method that acknowledges the stochastic nature of the litter size is in demand. A major problem in constructing a model for the full data structure is the intractability of the marginal likelihood. Due to this problem, formulating the models is difficult. An interesting alternative lies in the use of Bayesian methods. Markov Chain Monte Carlo methods are very flexible with respect to the structure of the models that can be considered. In this setting, a Bayesian random effects model was proposed. All

outcomes of the developmental toxicity study were analysed simultaneously. The main advantage of the proposed methodology is the flexibility in which all stages of the data can be modelled.

The model was applied to a developmental toxicity study (EG in mice), and used for quantitative risk assessment. When interested in a safe level of exposure, it is important to account for all possible adverse effects. Often however, focus is only on the outcome that is most sensitive to the exposure when performing quantitative risk assessment. But, use of univariate methods to determine a safe dose level can yield unreliable, and thus unsafe, dose levels. This acknowledges the importance of a model that accounts for the full data structure.

A Bayesian estimation of a safe level of exposure provides an attractive alternative to the commonly used frequentist approaches. The posterior distribution of the benchmark dose does not only give a point estimate, but reflects also the uncertainty associated with this estimate.

Chapter 8

Modelling Synchrony in Neuronal Firing

As indicated in Chapter 1, many sets of data collected in biological sciences have a correlated data structure. In this chapter, we consider a study in the area of neurophysiology.

Neural assemblies can transmit information about the characteristics of a stimulus via the spike rate of neurons, and there is evidence for association between spike rate and psychophysical performance (Newsome *et al.* 1989, Mountcastle *et al.* 1990, Romo and Salinas 1999). Another information processing strategy in the brain is the temporal structure, or relative timing of spike trains, between functional assemblies (Abeles 1982, Shadlen and Newsome 1994, Singer 1999). Synchronization refers to near simultaneous neural activity (spikes) provoked by a stimulus, and it is hypothesized that it should be correlated to perceptual grouping of the feature (Abeles 1982, Malsburg 1999, Usrey and Reid 1999).

There is a growing interest in the search for statistical methods for detecting and testing synchronized neural activity. Conventional approaches are based on the use of cross-correlation techniques, usually applied to the activity of pairs of neurons recorded under appropriate stimulus conditions. In this context, the basic tool is the cross-correlogram, representing a time-averaged correlation among the spiking events of the participating neurons. Extensions of this analysis are the Gravitational Clustering (Gerstein and Aertsen 1985) and the Joint-Peristimulus Time Histogram (JPSTH; Aertsen *et al.* 1989), which address the dynamics of the correlation between cells on a very short time scale. However, although the dynamics of synchronicity can

be observed as a function of time by averaging over trials, with these approaches it is not possible to analyse individual spike coincidences on a trial-by-trial basis.

Recent investigations have focused on the detection of individual instances of synchronized activity between groups of two or more neurons. One of the most commonly used methods for this task is the Unitary-Event (UE) analysis (Grün 1996, Grün *et al.* 1999, Grün, Diesmann, and Aertsen 2002a, 2000b, Riehle *et al.* 1997). This approach allows us to ascertain the statistical significance of brief epochs of synchronous spiking. The statistical null-hypothesis is formulated in terms of the individual firing probabilities of the participating neurons. This method searches recordings from multiple single neurons for epochs with distinctly more (approximately-)coincident spikes than expected from independent neurons obeying Poissonian spike statistics. The core of UE analysis consists of computing the probabilities (joint p-values) for the occurrence of a given minimum number of coincident spikes in short time segments, under the null hypothesis of independence. Segments with a joint p-value below a fixed level of significance are identified as significant epochs where the null hypothesis is rejected. Gütig, Aertsen and Rotter (2002) reformulated the statistical test underlying this method using a coincidence count distribution based on empirical spike counts rather than on estimated spike probabilities. Recently, Kass, Ventura and Cai (2003) considered a more general framework that could handle problems of alternative structures avoiding the assumption that spike trains are Poisson processes and suggest a (simulation-based) significance test for synchrony, in which p-values are calculated by using bootstrap techniques.

Alternative procedures to assess synchrony were suggested recently and lie within the general regression framework. Models for multivariate correlated binary data can be grouped into different classes along the distinction between conditionally specified models, marginal models and cluster-specific models. The answer to the question which model family is to be preferred principally depends on the research question(s) to be answered. Conditional models describe the distribution of the outcomes conditional on (a subset of) the other outcomes. Well-known members of this class of models are log-linear models. Martignon *et al.* (2002) introduced a log-linear model for representing firing rates on a set of neurons and showed that nonzero coefficients or effects of these models are a natural measure for synchronous firing. Advantages of the log-linear model are that it provides a relatively simple representation of association: the main effect terms may be taken to be smooth functions of time using regression splines and then the model may be fitted by using standard software for generalized linear models. In this way, standard likelihood-based estimation and testing of the interaction coefficients provide an assessment for association. Generalization to more

than two neurons involving two-way interactions is easy. Higher order interactions may also be examined, and in a sense this is an attractive feature of the approach. On the other hand there are obvious complexities in including large numbers of terms. Another disadvantage of this approach is that the main effect terms are not interpretable as marginal firing rates. Indeed, as many authors have pointed out (e.g. McCullagh and Nelder 1989) the lack of compatibility between marginal models and joint models is a general feature of loglinear models.

In this chapter, we will consider an alternative approach based on a marginal model for multivariate correlated binary data. This model was designed to (i) describe the individual activity of the neurons involved and (ii) detect correlations of any order in a unified way. This method allows to jointly analyse the firing rates of the individual neurons as well as the synchrony between the neurons. Both the temporal evolution, as well as trial-specific covariates, can be investigated in a flexible way. In addition, the extension to more than two neurons is also possible.

Section 2.2 introduces the electrophysiological experiment. The chapter is structured as follows. In Section 8.1 we describe some measures of synchrony that are used in the literature and offer a new methodology to measure the neuronal synchrony. Section 8.2 discusses specific and general issues in modelling the data. In Section 8.3, specific tools for analysis will be exemplified. This work is also presented in Faes *et al.* (2004d).

8.1 Measure of Synchrony

Synchrony refers to the observation that action potentials emitted from different neurons are emitted at the same time, or very close in time. In this paper, two neurons are considered to discharge in synchrony if they fire together in a 1 ms window. However, methods can easily be extended to less precise coincidences. First, focus is on the synchrony of two neurons only. Later, an extension of the methodology to more than two neurons is proposed. To describe how 'synchronous' two spike trains are, a variety of methods can be used. Let us first introduce the necessary notation.

Let Y_{itj} be the binary outcome of the ith neuron at time t ($t = 1, 2, \ldots, T$) for trial j ($j = 1, 2, \ldots, N$). The frequency of matches and mismatches over the different trials for neurons i and k at time point t can be written in the form of a contingency table, such as displayed in Table 8.1. In this table, n_{11} represents the frequency of $1 - 1$ matches (at time t), n_{10} is the frequency of $1 - 0$ matches, and so forth. The matching rate for neuron 1 and neuron 2 at different time points are pictured

Table 8.1: A contingency table of the number of matches and mismatches for two neurons at a certain time point.

		Neuron k		
		1	0	Totals
Neuron i	1	n_{11}	n_{10}	$n_{11} + n_{10}$
	0	n_{01}	n_{00}	$n_{01} + n_{00}$
Totals		$n_{11} + n_{01}$	$n_{10} + n_{00}$	$N = n_{11} + n_{10} + n_{01} + n_{00}$
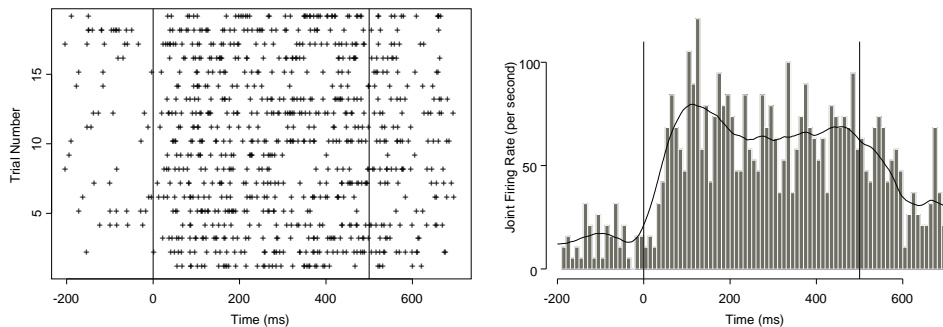
Figure 8.1: *Raster plot (left) of spike coincidence of neurons 1 and 2, and corresponding joint peristimulus time histogram with smoothed (kernel) version of the counts (right).*

in Figure 8.1. The probability of coincidence increases rapidly after appearance of the stimulus and decreases gradually after removal of the stimulus. Note however that one has to be careful in interpreting the increase in the number of synchronous events, since a certain number of synchronous events would always occur due to purely random coincidence.

In the next section, we give an overview of possible synchrony measures.

8.1.1   Odds Ratio

One possible measure for synchrony is the odds ratio, indicating the association between two neurons. The odds ratio is often a measure of choice to capture the asso-
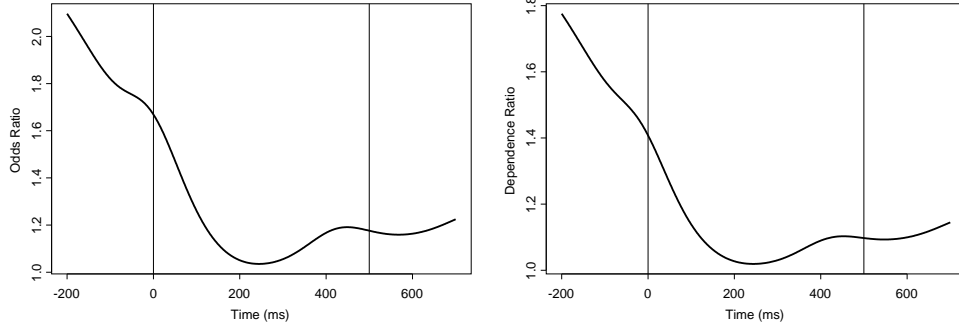
Figure 8.2: *Plot of odds ratio (left panel) and dependence ratio (right panel) between neuron 1 and neuron 2.*

ciation in a contingency table. The odds ratio at time t is defined as

$$\psi(t) = \frac{n_{11}n_{00}}{n_{10}n_{01}} = \frac{\pi_{11}(t)\big[1 - \pi_1(t) - \pi_2(t) + \pi_{11}(t)\big]}{\big[\pi_{11}(t) - \pi_1(t)\big]\big[\pi_{11}(t) - \pi_2(t)\big]}, \tag{8.1}$$

with $\pi_{11}(.)$ the joint probability of 2 neurons to fire and $\pi_1(.)$ and $\pi_2(.)$ the marginal probabilities. Using this relationship, the joint distribution of neuron 1 and 2 can be written as function of the marginal probabilities and the odds ratio (Plackett 1965):

$$\pi_{11}(t) = \begin{cases} \frac{1 + [\pi_1(t) + \pi_2(t)][\psi(t) - 1] - R(\pi_1(t), \pi_2(t), \psi(t))}{2[\psi(t) - 1]}, & \text{if } \psi \neq 1, \\ \pi_1(t)\pi_2(t), & \text{otherwise,} \end{cases} \tag{8.2}$$

with

$$R(\pi_1, \pi_2, \psi) = \sqrt{[1 + (\pi_1 + \pi_2)(\psi - 1)]^2 + 4\psi(1 - \psi)\pi_1\pi_2}. \tag{8.3}$$

The Plackett-Dale distribution has been used extensively in the modelling of many biomedical applications. For example, it turned out to be very useful in the modelling of clustered binary data together with a pseudo-likelihood estimation approach (Geys, Molenberghs, and Lipsitz 1998). In Chapter 6 it is shown how the Plackett-Dale model can be utilized to jointly model a continuous and an ordinal outcome.

A plot of the odds ratio between neurons 1 and 2 as function of time is given in the left panel of Figure 8.2. All values are higher than 1, indicating a positive association between the two neurons. But, in contrast with the probability of joint firing, there is a decrease at the start of the stimulus, and a small increase at the end of the stimulus. The elevated odds ratios before the appearance of the stimulus

might be explained by the low activity rate, or high number of $0 - 0$ matches, in this time interval (Figure 2.5). Note however that synchrony is a process happening only during activity of the neurons. Thus, although the odds ratio is an attractive measure to describe the association between neurons, as a measure of synchrony it suffers from being symmetric, treating the $1 - 1$ matches and $0 - 0$ matches equally, inflating the odds ratio at the start of the trial. In this context, a $1 - 1$ match is a stronger indication of synchrony than a $0 - 0$ match. This must be reflected in the measure of synchrony.

8.1.2   Dependence Ratio

As an alternative, synchrony can be measured as the proportion of observed number of synchronous events to the expected number from two independent stochastic processes. This measure expresses to which extent the probability of having a spike simultaneously for two neurons is different from the product of the marginal probabilities. This idea was exploited by several authors (Gütig, Aertsen and Rotter 2002, Grün *et al.* 1999, Riehle *et al.* 1997).

Based on the same idea, Ekholm (1995) defined the dependence ratio as

$$\tau(t) = \frac{\pi_{11}(t)}{\pi_1(t)\pi_2(t)}. \tag{8.4}$$

It is easy to see that $(\tau - 1) \times 100$ indicates the increase (as %) in probability for both neuron 1 and neuron 2 to have a spike, compared to what it would be under independence; that $\tau = 1$ if and only if neurons 1 and 2 are independent; and that the following relationship between τ and ψ holds:

$$\psi = \frac{\tau - 1}{(1 - \tau\pi_1)(1 - \tau\pi_2)} + 1. \tag{8.5}$$

In Figure 8.2, the dependence ratio between neurons 1 and 2 is depicted. The τ-curve is almost identical to the plot of the odds ratio. There is a downward shift of about 0.2. Similar to the odds ratio, it seems that the dependence ratio is highly affected by the small number of spikes at the start and end of the experiment. Further, note that when neurons are dependent, they are not necessarily in synchrony. For example, consider two neurons with the following spikes during several trials:

$$\text{neuron 1:}\quad 110011101001,$$
$$\text{neuron 2:}\quad 001100010110.$$

Table 8.2: Some possible similarity measures in terms of frequencies.

Measure	Rationale
$\frac{n_{11}}{n_{11}+n_{10}+n_{01}}$	No $0-0$ matches in numerator or denominator. The $0-0$ matches are treated as irrelevant.
$\frac{2n_{11}}{2n_{11}+n_{10}+n_{01}}$	No $0-0$ matches in numerator or denominator. Double weight for $1-1$ matches.
$\frac{n_{11}}{n_{11}+2(n_{10}+n_{01})}$	No $0-0$ matches in numerator or denominator. Double weight for unmatched pairs.
$\frac{n_{11}}{n_{10}+n_{01}}$	Ratio of matches to mismatches with $0-0$ matches excluded.

These neurons are asynchronous (none of the events collapse), although the neurons are dependent ($\pi_{11} \neq \pi_1\pi_2$). Thus, one should be cautious with the use of independence as a basis for a measure of synchrony.

8.1.3   Conditional Synchrony Measure

A measure of synchrony is to be regarded as a specific measure of 'closeness', or 'similarity', and should treat a $1-1$ match as a stronger indication of similarity than a $0-0$ match. Indeed, the evidence that two neurons react is stronger evidence of synchrony than the absence of a spike in both neurons. Thus, it might be reasonable to discount the $0-0$ matches or even disregard them entirely. Table 8.2 lists some similarity measures in terms of the frequencies. The first three measures are monotonically related (Johnson and Wichern 1998).

The different measures of similarity are pictured in Figure 8.3. These plots all give the same idea of synchrony. After the stimulus is given, there is an increase in synchrony. After the end of the stimulus, there is a decrease.

An attractive measure of similarity, which could be used in the context of neuronal synchrony, is the first one from Table 8.2. The main advantage is the absence of the nuisance $(0,0)$ pairs. The conditional probability of firing together, given there is a spike in one of the neurons, could be used to measure the 'strength' of synchrony. This measure, which we call the Conditional Synchrony Measure (CSM), reflects how much the neurons fire together during activity in one of the two neurons. It can be written
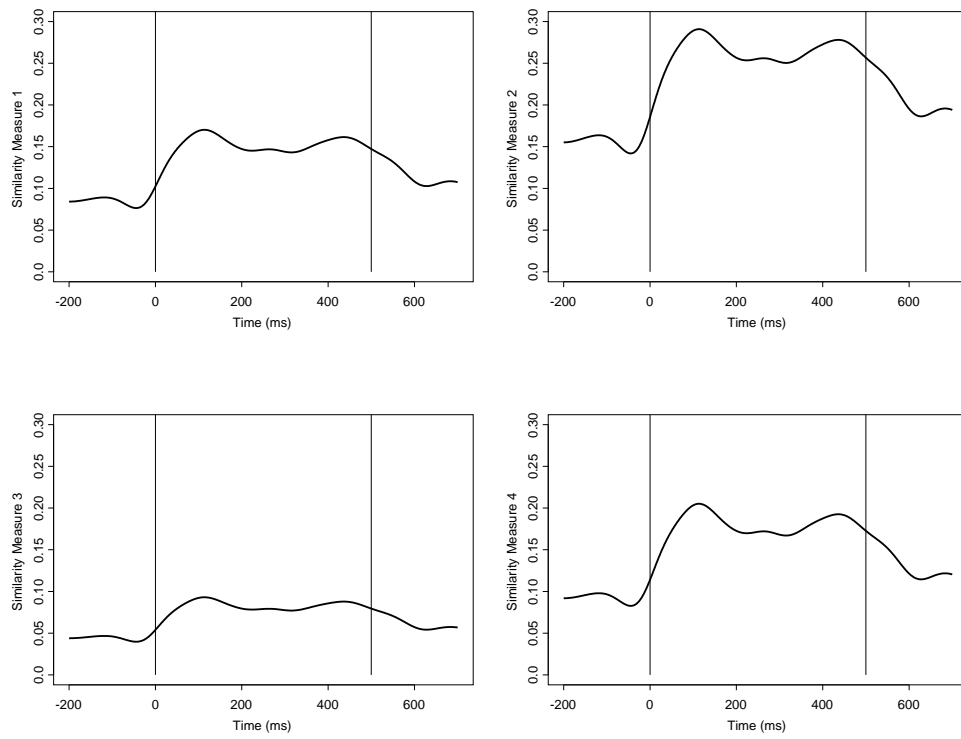
Figure 8.3: *From left to right and from top to bottom: a plot of similarity measures proposed in Table 1, for neuron 1 and neuron 2.*

as a function of the joint probability of firing and the two marginal probabilities:

$$\text{CSM}(t) = \frac{\pi_{11}(t)}{\pi_1(t) + \pi_2(t) - \pi_{11}(t)}, \tag{8.6}$$

estimated by

$$\widehat{\text{CSM}}(t) = \frac{n_{11}}{n_{10} + n_{01} + n_{11}}. \tag{8.7}$$

The CSM has an easy to understand interpretation, and is based on the definition of synchrony. The top left panel in Figure 8.3 shows a plot of the CSM for neurons 1 and 2. This plot shows an increase of synchrony after the stimulus, and a decrease of synchrony after the end of stimulus.

In the sequel, we will use the CSM as a measure of synchrony. In Section 8.2, a model based on the conditional probability is described, and is combined with a pseudo-likelihood estimation method. Section 8.3 describes the analysis of the experiment.

8.2 Model Formulation and Pseudo-Likelihood

8.2.1 The Model

Suppose that for each neuron i $(i = 1, 2)$ under study a vector of binary responses $\{Y_{it}, t = 1, \ldots, T\}$ is observed, together with a vector of covariates $\boldsymbol{x}$. First, assume that observations at different time points are independent. We want to establish the dependence of each of the two neurons on the covariate vectors, taking the correlation between both responses into account.

The model arises from the decomposition of the joint probabilities

$$\pi_{j_1 j_2}(t) = P(Y_{1t} = j_1, Y_{2t} = j_2 | \boldsymbol{x}), \qquad (j_1, j_2 = 0, 1),$$

into 'main effects' and 'effect of synchrony'. Let the marginal probabilities of neurons 1 and 2 at time t be $\pi_1(t)$ and $\pi_2(t)$, respectively. The Conditional Synchrony Measure, or conditional probability of observing two spikes at time t, given there is a spike in at least one of the two neurons equals

$$\text{CSM}(t) = \frac{\pi_{11}(t)}{\pi_1(t) + \pi_2(t) - \pi_{11}(t)}. \tag{8.8}$$

Formally, the decomposition of the joint probabilities $\pi_{j_1 j_2}$ is given by

$$\begin{aligned}
h_1\big(\pi_1(t)\big) &= \boldsymbol{\beta}_1^T \boldsymbol{x}, \\
h_2\big(\pi_2(t)\big) &= \boldsymbol{\beta}_2^T \boldsymbol{x}, \\
h_3\big(\text{CSM}(t)\big) &= \boldsymbol{\beta}_3^T \boldsymbol{x},
\end{aligned}$$

with h_1, h_2, and h_3 link functions in the generalized linear model sense. As such, the marginal structure can be modelled in a flexible fashion: the marginal probabilities can be fitted within the generalized linear models framework. And also the synchrony can be modelled in a general way, including time-varying covariates as well as trial-specific covariates.

Solving these equations for the joint probability $\pi_{j_1 j_2}(t)$ yields:

$$\pi_{11}(t) = \frac{\text{CSM}(t)}{1 + \text{CSM}(t)} \big[\pi_1(t) + \pi_2(t)\big].$$

Based upon this probability, we can derive the joint density function

$$g(y_{1t}, y_{2t}) = \begin{cases} \pi_{11}(t) & \text{if } y_1 = 1 \text{ and } y_2 = 1, \\ \pi_1(t) - \pi_{11}(t) & \text{if } y_1 = 1 \text{ and } y_2 = 0, \\ \pi_2(t) - \pi_{11}(t) & \text{if } y_1 = 0 \text{ and } y_2 = 1, \\ 1 - \pi_1(t) - \pi_2(t) + \pi_{11}(t) & \text{if } y_1 = 0 \text{ and } y_2 = 0. \end{cases} \tag{8.9}$$

The above model can be generalized to model M neurons simultaneously in the presence of explanatory variables $\boldsymbol{x}$. Let $\{Y_{it}, t = 1, \ldots, T\}$ be the random vector for neuron i $(i = 1, 2, \ldots, M)$. The synchrony of 3 neurons can be defined as the conditional probability of joint firing in the three neurons, given there is activity in at least one of the neurons. The joint probability

$$\pi_{j_1 j_2 j_3} = P(Y_{1t} = j_1, Y_{2t} = j_2, Y_{3t} = j_3 | \boldsymbol{x}),$$

can be decomposed into 'main effects', 'synchrony between two neurons' and 'synchrony between three neurons':

$$\begin{aligned} \pi_{j_1 j_2 j_3} &= \frac{\text{CSM}_{123}}{1 - \text{CSM}_{123}} \Big[\pi_1 \frac{1 - \text{CSM}_{12}\text{CSM}_{13}}{(1 + \text{CSM}_{12})(1 + \text{CSM}_{13})} \\ &\quad + \pi_2 \frac{1 - \text{CSM}_{12}\text{CSM}_{23}}{(1 + \text{CSM}_{12})(1 + \text{CSM}_{23})} + \pi_3 \frac{1 - \text{CSM}_{13}\text{CSM}_{23}}{(1 + \text{CSM}_{13})(1 + \text{CSM}_{13})} \Big]. \end{aligned}$$

As such, one can jointly investigate the firing rates in the three neurons and the synchronous activity between two or three neurons. A generalization to more than three neurons is also possible.

8.2.2 Pseudo-Likelihood Estimation Method

Arguably, observations at different time points are not independent. Three different types of associations, depicted in Figure 8.4, can be present: the association between
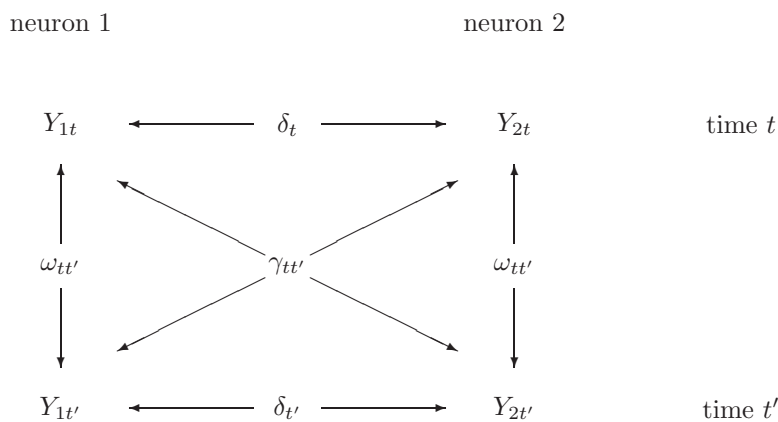
neuron 1 neuron 2

Y_{1t} ⟵ δ_t ⟶ Y_{2t} time t

$\omega_{tt'}$ $\gamma_{tt'}$ $\omega_{tt'}$

$Y_{1t'}$ ⟵ $\delta_{t'}$ ⟶ $Y_{2t'}$ time t'

Figure 8.4: *Four different types of contributions for the pseudo-likelihood.*

two different time points from the same neuron (ω), the association between two neurons at the same time point (δ), and the association between two neurons at two different time points (γ). Although there is only one association of direct interest, namely the association between two neurons at the same time point, the other associations cannot be neglected. Indeed, ignoring the associations in the data overestimates precision and hence underestimates standard errors and lengths of confidence intervals. One can treat such associations as a nuisance, correcting for them to obtain suitable variance estimates. In such a case, we can use pseudo-likelihood, rather than considering full likelihood. The full likelihood function for trial j ($j = 1, \ldots, N$), i.e., $f_j(y_{11j}, \ldots, y_{1Tj}, y_{21j}, \ldots, y_{2Tj})$, can be replaced by

$$p\ell_j \;=\; \sum_{t=1}^{T} \ln g(y_{1tj}, y_{2tj}). \tag{8.10}$$

As such, we avoid the computational complexity of the full likelihood distribution of each trial. The value of the parameters that maximize the log pseudo-likelihood function $p\ell = \sum_{j=1}^{N} p\ell_j$ are the pseudo-likelihood estimates.

Arnold and Strauss (1991) presented a formal and more general definition of the pseudo-likelihood estimation method, and established consistency and asymptotic normality of the pseudo-likelihood estimator. Similar in spirit to generalized estimat-

ing equations (Liang and Zeger 1986), the asymptotic normality result provides an easy way to consistently estimate the asymptotic covariance matrix:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = J^{-1}KJ^{-1} = \Big(\sum_{j=1}^{N}\frac{\partial \boldsymbol{U}_j}{\partial \boldsymbol{\beta}}\Big)^{-1}\Big(\sum_{j=1}^{N}\boldsymbol{U}_j(\boldsymbol{\beta})\boldsymbol{U}_j(\boldsymbol{\beta})^T\Big)\Big(\sum_{j=1}^{N}\frac{\partial \boldsymbol{U}_j}{\partial \boldsymbol{\beta}}\Big)^{-1}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \quad (8.11)$$

with $\boldsymbol{U}_j$ the pseudo-likelihood estimating equations

$$\boldsymbol{U}_j(\boldsymbol{\beta}) = \frac{\partial \ln g(y_{1tj}, y_{2tj})}{\partial \boldsymbol{\beta}}. \tag{8.12}$$

This approach acknowledges the fact that, while the synchrony between different neurons on the same time point is often of scientific interest, the association between different time points is usually considered a nuisance. The sandwich variance estimator (8.11) is then used to adjust for potential bias in the variance estimator. Geys, Molenberghs and Lipsitz (1998) compared pairwise likelihood with other estimation equation approaches in marginally specified odds ratio models with exchangeable association structure, and showed that the efficiency of pseudo-likelihood estimators was comparable to the efficiency of GEE estimators.

8.2.3 Test for Synchrony

A possible information processing strategy in the nervous system is the use of synchrony between neurons. In this mechanism, the selection of specific sensory information is implemented by increasing the synchrony between neurons that represent the information. Therefore, it is of interest to test whether there is an increase of synchrony during the task.

A sensible approach is to compare the CSM during the experiment with the CSM at baseline (CSM^0):

$$H_0 : \text{CSM}(t) = \text{CSM}^0.$$

The period before the stimulus comes on (time -200 ms until -1 ms) can be used to estimate the baseline cell activity. The condition to be checked is very simple: is the $\widehat{\text{CSM}}(t)$ different from the estimated baseline $\widehat{\text{CSM}}^0$:

$$H_0 : \text{CSM}(t) - \text{CSM}^0 = 0.$$

Often, one thinks about synchrony in terms of the probability of joint firing. However, one needs to be careful in interpreting the mechanisms of joint firing. An increase of joint firing might be induced by an increase of synchronous action potentials or might be a byproduct of the increased firing rates. We can translate the above

methodology to test for synchrony in terms of the joint firing probability. Assume the stimulus has no effect on the Conditional Synchrony Measure, and is equal to the background CSM^0. Then, the probability of joint firing at time t, under the null hypothesis of a constant CSM, equals

$$\pi_{11}^{H_0}(t) = \frac{\text{CSM}^0}{1 + \text{CSM}^0} \big[\pi_1(t) + \pi_2(t)\big].$$

By comparing this joint probability $\pi_{11}^{H_0}(t)$ with the true probability of joint firing $\pi_{11}(t)$, one can test whether the increase of joint firing is due to an increase of synchronous action potentials. The null hypothesis, in terms of the probability of joint firing, is

$$H_0 : \tau(t) = \pi_{11}(t) - \pi_{11}^{H_0}(t) = 0. \tag{8.13}$$

Based on the estimated values $\tau(t)$, it can be checked whether there is a difference. Of course, the sampling uncertainty of the estimate should be acknowledged. Therefore, a $(1 - \alpha)100\%$ (e.g., $\alpha = 0.05$) confidence interval around $\tau(t)$ is derived, and it is checked whether the 0 lies within this interval. Using the delta method, the pointwise confidence intervals of $\tau(t)$ under the null are determined by

$$\widehat{\tau}(t) \pm z_{1-\alpha} \sqrt{\widehat{\text{Var}}\big(\widehat{\tau}(t)\big)}, \tag{8.14}$$

and the variance of the joint probability of firing is estimated as

$$\widehat{\text{Var}}\big(\widehat{\tau}(t)\big) = \Big(\frac{\partial \tau(t)}{\partial \boldsymbol{\beta}}\Big)^T \widehat{\text{Cov}}(\boldsymbol{\beta}) \Big(\frac{\partial \tau(t)}{\partial \boldsymbol{\beta}}\Big)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \tag{8.15}$$

with $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. Since multiple comparisons are of interest in this situation (one test for each time point), the Bonferroni method can be used. Here, the use of α/m is recommended instead of α, with m the number of comparisons to be made. This methodology will be illustrated in the next section.

8.3 Data Analyses

In neurophysiology, interest is in the temporal evolution of the firing rate and the synchrony between certain neurons, both being important neural mechanisms to transmit information about the characteristics of a stimulus. Further, one wants to investigate the effect of the stimulus properties (in this case, different orientations of the stimulus) on the firing rate and synchrony.

Table 8.3: A description and comparison of models. The pseudo-likelihood value is displayed, together with the number of parameters in the model and the pseudo-likelihood AIC value.

Model	Description	degree	$p\ell$	# par	AIC
1	orthonormal polynomials	5	-16661.13	18	-16679.13
2	orthonormal polynomials	10	-16623.70	33	-16656.70
3	piecewise-cubic splines	5	-16646.77	18	-16664.77
4	piecewise-cubic splines	10	-16604.55	33	-16637.55

8.3.1   Time Course

As a first application of the model, we jointly investigate the temporal structure of spike trains and spike coincidences of neurons 1 and 2. Flexible models to describe the time trends are needed. For both the marginal probabilities as well as the synchrony, the model will be fit in a generalized linear models framework

$$\text{logit}\big(\pi_1(t)\big) \quad = \quad f_1(t), \tag{8.16}$$

$$\text{logit}\big(\pi_2(t)\big) \quad = \quad f_2(t), \tag{8.17}$$

$$\text{logit}\big(\text{CSM}(t)\big) \quad = \quad f_3(t), \tag{8.18}$$

with $f_1(t)$, $f_2(t)$ and $f_3(t)$ some functions of time. Both parametric and semi-parametric methods are used: a parametric polynomial with an orthonormal basis and a semi-parametric piecewise cubic spline are considered. As a model selection criterion, we suggest to use the pseudo-likelihood $AIC = p\ell - r$, with r the number of parameters in the model. Larger values of AIC indicate a better fitting model. A summary of model fit is given in Table 8.3. The piecewise-cubic spline of degree 10 (corresponding with 9 knots) seems to give the best fit.

The estimated firing rate of a spike, i.e., the number of spikes per seconde, and the conditional synchrony measure, based on the piecewise cubic spline of degree 10, is shown in Figure 8.5. The solid lines correspond to the estimated model. The smoothed observed curve is represented by the dotted line. From the plots it is seen that the natural splines has much flexibility to closely follow the data.

Based on this model, several quantities of interest can be investigated. Both the firing rates in the two neurons, as well as the firing synchrony change during the discrimination task. Reactions on the stimulus are reflected in growing firing rates
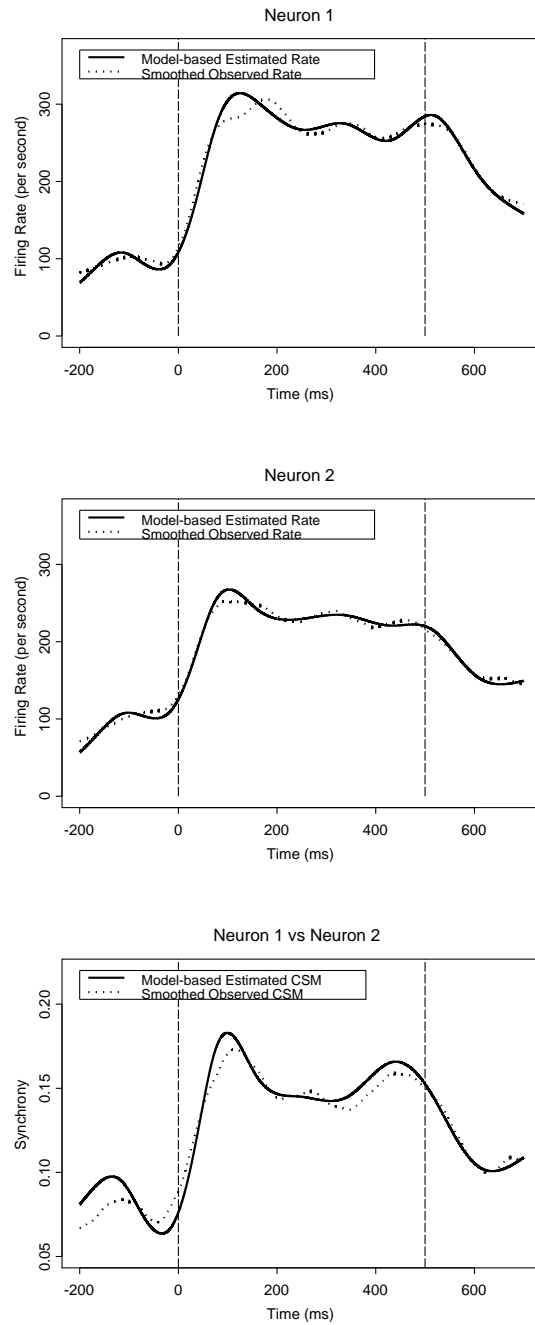
Figure 8.5: *Model 4. The top and middle panel presents the estimated firing rate for neuron 1 and 2, respectively. The bottom panel presents the estimated synchrony between neurons 1 and 2.*
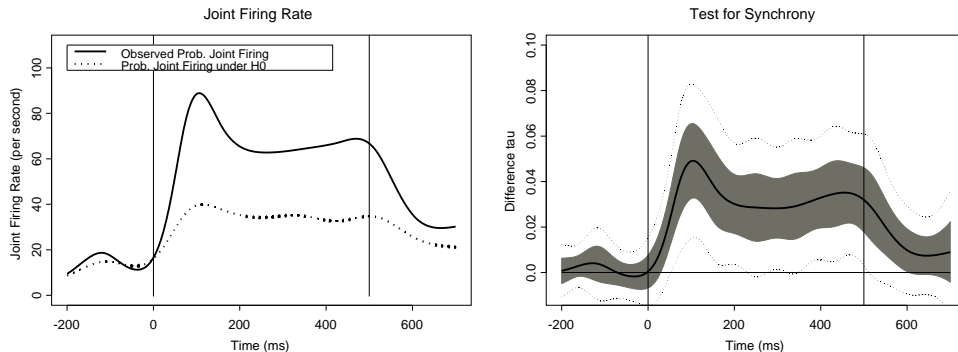
Figure 8.6: *Test for synchrony. Left: joint firing rates (per second) under the null and alternative hypothesis. Right: test based on the probability of joint firing.*

and increasing synchrony between the two neurons. After removal of the stimulus, the firing rates of both neurons and the synchrony drop back to the initial state. The maximal firing rate of neuron 1 appears 125 ms after the start of the stimulus, with a firing rate of 0.31. The maximal firing rate of neuron 2 appears already at 100 ms after the start of the stimulus, with a firing rate of 0.27.

To test whether the neurons are firing in synchrony, we use the test as proposed in Section 8.2.3. In the left panel of Figure 8.6, the joint firing rates (per second) under the null and alternative hypothesis are displayed. The full line is the estimated joint firing rate in the experiment. The dotted line represents the joint firing rate assuming the CSM stays constant during the experiment, equal to the baseline CSM. The joint probability of firing, under the null hypothesis of a constant conditional synchrony measure $CSM=CSM^0$ changes over time, due to the varying firing probabilities of the two neurons individually.

To test for synchrony, we compare the true probability of joint firing with the null hypothesis of joint firing with a background CSM, as displayed in the right panel of Figure 8.6. The bold line is the estimated difference ($\tau(t)$) in probability of joint firing under the null and alternative of a constant CSM, as defined in (8.13). The 95% confidence bounds (grey band) and Bonferroni corrected confidence bounds (dotted lines) are displayed. By comparing the confidence bounds with 0, one can see that there is a there is a significant increase of synchrony almost immediately after the stimulus is shown (after 53 ms), which disappears again at 11 ms after removal of the stimulus. Note however that the Bonferroni method is highly conservative

and may miss real differences, since the multiple tests are possibly highly correlated. Therefore we will consider both the corrected and uncorrected confidence bounds. Other multiple testing procedures are subject of further research.

8.3.2 Orientation

The orientation of the stimulus (85°, 90° or 95°) can have an important effect on the firing rate and on the synchronization of the neurons. One goal of this study is to quantitatively determine and formally compare how the temporal patterns of neuronal activity are affected by the different orientations of the stimuli.

Consider a model with inclusion of the orientation (Model 5):

$$
\begin{aligned}
f(t, \alpha) &= \beta_1 + \sum_{i=1}^{5} \beta_{1+i} f_1(t) + \beta_7 I_{85} + \sum_{i=1}^{5} \beta_{7+i} I_{85} f_2(t) \\
&\quad + \beta_{13} I_{95} + \sum_{i=1}^{5} \beta_{13+i} I_{95} f_3(t),
\end{aligned} \tag{8.19}
$$

where I_{85} and I_{95} are dummy variables corresponding to orientations of 85° and 95°, respectively, and $f_1(t)$, $f_2(t)$ and $f_3(t)$ are natural splines in time t of degree 5. Results are presented in Figure 8.7.

The solid line corresponds to 90°, the dotted line with 85°, and the dashed line with 95°. The firing rates in neurons 1 and 2 are slightly different for stimuli with orientations of 85° or 95°, but they are quite different for the 90° oriented line. In the primary visual cortex the receptive field position and its orientation preference changes across the visual space. For that reason, a long oriented line, such as the ones used in the experiment, will provoke different responses in each cell. This explains the differences seen in the graph.

A pseudo-likelihood ratio test statistic, as proposed in Section 3.2.2, is calculated to test whether the effects of different orientations are different. Table 8.4 displays the pseudo-likelihood ratio test statistic G_a^2 and the corresponding p-value for the null hypothesis that the orientations have the same effect on π_1, π_2 and CSM. The effect of a 90° stimulus on the firing rate of the neurons is significantly different from the effect of a 85° or 95° stimulus (all p-values < 0.001). In contrast, there is no evidence for a difference between a 85° or 95° stimulus (p-value > 0.05). The differences on the synchrony are less pronounced, with only a different effect between stimuli of 85° and 90°.

In Figure 8.8, it is displayed whether there is a significant increase in the number of joint coincidences due to the stimulus, for each of the orientations separately. Based
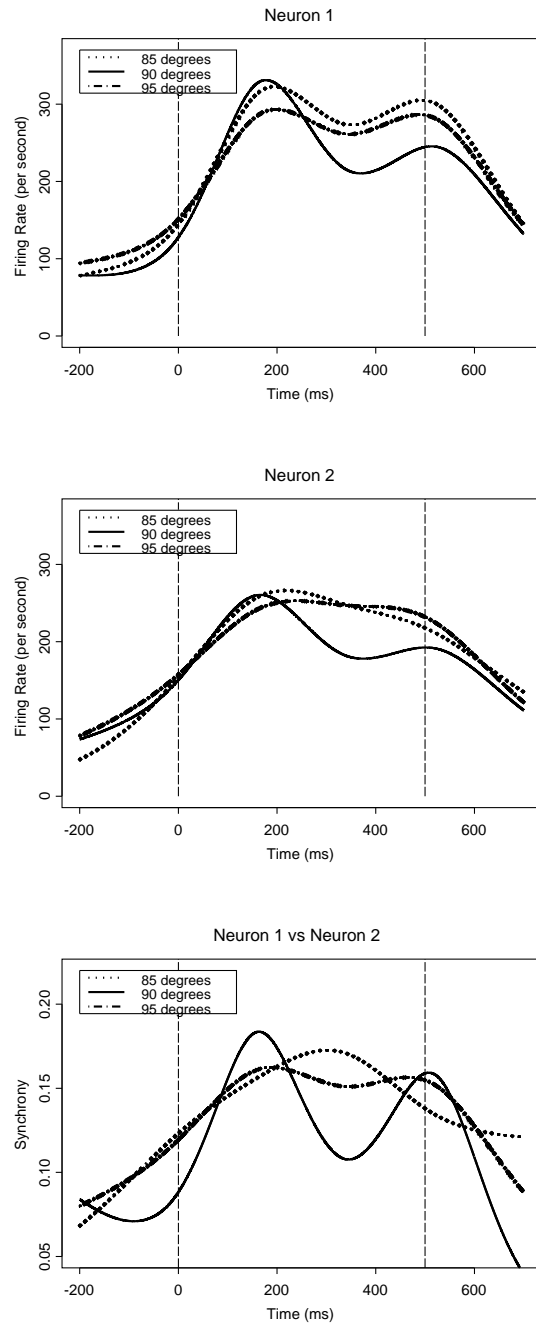
Figure 8.7: *Model 5. The top and middle panel presents the estimated firing rate of neuron 1 and 2, respectively. The bottom panel presents the estimated synchrony between neurons 1 and 2.*
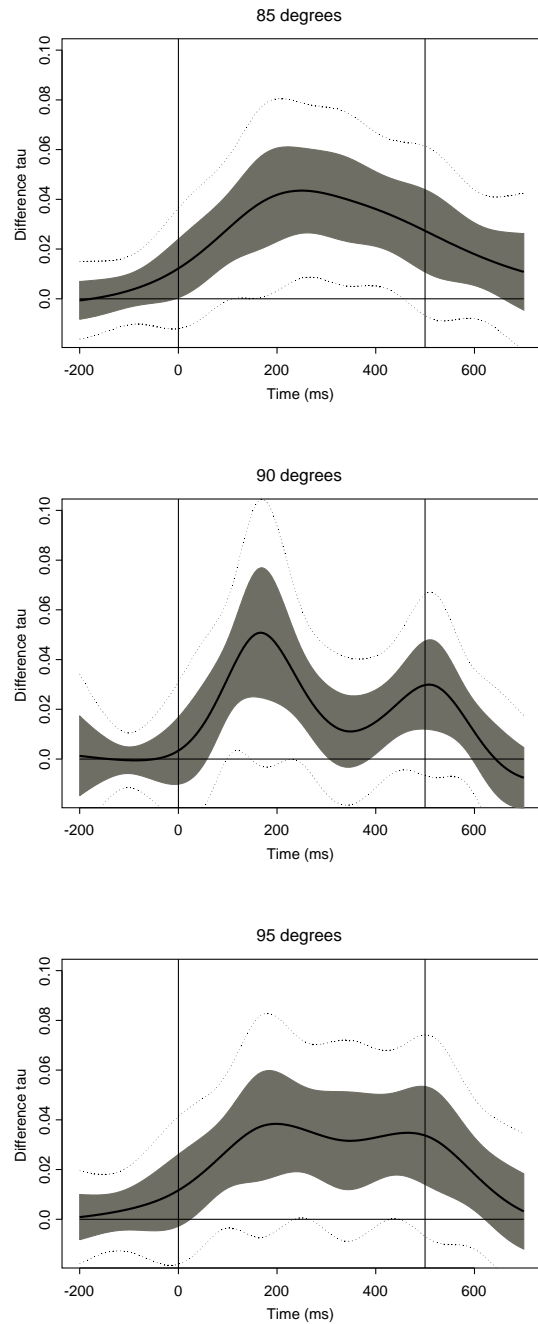
Figure 8.8: *Test for synchrony, based on the probability of joint firing. From bottom to top: orientations 85, 90 or 95 degrees.*

Table 8.4: Pseudo-likelihood ratio test statistic G_a^2 and corresponding p-value for all pairwise comparisons on π_1, π_2 and CSM.

Comparison	π_1	π_2	CSM
$85° = 90°$	25.58 ($<$0.001)	21.06 (0.002)	19.49 (0.003)
$95° = 90°$	17.08 (0.009)	20.03 (0.003)	7.76 (0.256)
$85° = 95°$	4.59 (0.597)	5.05 (0.538)	1.44 (0.964)
$85° = 95° = 90°$	32.04 (0.001)	31.13 (0.002)	16.36 (0.175)

on the 95% confidence bound, we see that there is an immediate increase of joint firing when a stimulus of 85° or 95° is given. For a stimulus of 90°, the reaction is somewhat later in time. The difference between observed and expected probability of synchronous spikes may be explained by postulating a common input or some other form of functional connectivity due to the stimulus. When we carry out the Bonferroni correction, a significant increase in joint firing is seen for a 85° stimulus, whereas there is almost no evidence for a difference when a 90° or 95° stimulus is given. However, note that the power of the Bonferroni test is diminished due to the lack of independence between the tests, and thus possibly does not detect true differences in the population.

8.4   Discussion

In this chapter, we defined synchrony as the observation that action potentials emitted from different neurons are emitted at the same time. But, there is no "true" definition of synchrony, this term is dependent on the question and experimental situation. Synchrony can be defined at different temporal resolutions, e.g., 1, 3 and 5 ms. That is, spikes in one neuron were considered synchronous if they occurred at 0, ± 1, or ± 2 ms, relative to spikes in the second neuron, respectively. The proposed conditional synchrony measure is flexible enough to allow extensions of the definition.

The conditional synchrony measure is the probability of firing together, given that at least one of the two neurons is active. The advantage of the CSM is that it is robust against the high number of $(0,0)$ matches, since these do not provide information about synchrony. While most association measures treat both $(0,0)$ and $(1,1)$ matches as synchronous events, the Conditional Synchrony Measure only uses

the $(1,1)$ match as relevant synchronous event.

If one assumes that two spike trains have a certain background CSM, then the number of coincidences will depend on the firing rates of the neurons (this number will increase as the rates increase). Therefore, one cannot just compare the probability of joint firing with the baseline probability of joint firing, but one needs to include the firing rates of the neurons separately. By use of the CSM, one corrects for the probability that there is activity at a certain time point. Thus, comparing the CSM with the baseline value is a justifiable comparison.

Neurons of the visual cortex respond to oriented lines by modifying their firing rate, and the reaction to the stimulus depends on the position of the line over the receptive field of the cell. A long oriented line that falls on several receptive fields provokes different reactions. However, although each cell might respond differently and transmit different information, neighboring cells might synchronize their activity to such common stimulus and therefore constitute an assembly (Abeles 1982). CSM analysis of activity of two simultaneously recorded neurons reveals that the strength and temporal course of synchrony is modulated by the orientation of a common stimulus. This dynamic engagement of neural assemblies might transmit additional information about common features of a stimulus (Singer and Gray 1995, Shadlen and Movshon 1999).

Although the method is presented in the specialized field of neurophysiology, the methodology is applicable in other medical and epidemiological areas where the similarity of a rare binary outcome among subjects is of interest (for example, twins studies).

Chapter 9

Modelling Force of Infection in Veterinary Study

Another research area where one is often confronted with the collection of clustered data is veterinary epidemiology. Veterinary epidemiology deals with the investigation of diseases in animals. It is used to identify factors that may affect the disease occurrence. Ultimately, this information is then used to prevent or reduce the disease problem. Quantification of the disease pattern can be based on the prevalence, describing the probability that an animal from the population has the disease. Another measure of interest is the age-specific force of infection, i.e., the rate at which susceptible animals acquire infection at different ages.

Clustering occurs, for example, when data from different European countries are collected or due to the grouping of animals within herds. So far, one has always ignored the clustering in the data to calculate the force of infection. However, one cannot ignore the possibility of animals within herds or individuals within countries being more alike than between herds or countries, respectively. In this case, both the cluster-specific and the population-averaged prevalence and force of infection will give some information about the pattern of the disease.

In this chapter focus is on a sero-prevalence survey of the bovine herpesvirus-1 (BHV-1) in Belgian cattle, as introduced in Section 2.3. The BHV-1 infection is a transmissible disease in cattle, which is of economic importance and significance to international trade. It is distributed worldwide. We focus on the description of the age-specific disease pattern. In Section 9.1, the force of infection is introduced. As is done frequently, a logistic regression is performed in Section 9.2. Nevertheless, once an

infection is introduced in a herd, animals within the same herd have an increased risk to get infected too. Thus, individual responses are more homogeneously distributed within herds than in the whole population. There are several ways to deal with such clustering (Aerts *et al.* 2002), some of which estimate marginal, population-averaged measures of effect (Pendergast *et al.* 1996) and some of which estimate the subject-specific measures of effect, e.g., random effects models (Laird and Ware 1982). We will discuss each of them in turn. In Section 9.3, a marginal model is considered. A random effects approach is considered in Section 9.4. A definition of the herd-specific force of infection is proposed in Section 9.4.1.

9.1   Force of Infection

Primary interest in the analysis of infectious diseases is the force of infection, or the rate at which susceptible animals become infected, at different ages. The force of infection can be estimated from an age-specific cross-sectional prevalence sample, where instead of observing the age of infection, we observe a binary response indicator Y_i taking the value 1 if subject i had experienced infection before age a, and 0 otherwise. Let $\pi(a)$ be the probability to be infected before age a. In general, the sero-prevalence $\pi(a)$ is modelled as

$$\pi(a) = h^{-1}(\eta(a)) = g(\eta(a)), \tag{9.1}$$

where $\eta(a)$ is a linear predictor and h is a link function. If it is assumed that the disease is in a steady state, i.e., time independent, then the age-dependent force of infection $\ell(a)$ can be modelled according to equation (Anderson and May 1991)

$$\frac{d}{da}q(a) = -\ell(a)q(a), \tag{9.2}$$

with $q(a) = 1 - \pi(a)$. The differential equation (9.2) describes the change in the fraction of susceptible individuals with the age of the host. Consequently, the force of infection can be expressed as

$$\ell(a) = \frac{\pi'(a)}{1 - \pi(a)}. \tag{9.3}$$

When a logit link is considered, the force of infection reduces to

$$\ell(a) = \pi(a)\eta'(a). \tag{9.4}$$

In literature, the predictor $\eta(a)$ is often chosen to be a linear or higher order polynomial as function of age a. However, selection of the proper functional forms

describing the dependence of the parameters in the probability model is crucial to obtain correct estimates of the force of infection. Although classical polynomial predictors are still very customary, they are often inadequate. In this setting, there is a clear need for more flexible models, describing the relationship between age and infection of the animal (Shkedy, Aerts and Molenberghs 2003). In this chapter we illustrate how fractional polynomials offer great flexibility in modelling the force of infection based on clustered data. For the fractional polynomials, we consider as the best set of transformations, the one producing the lowest Akaike's Information Criterion (AIC, Akaike 1974), while being monotone in age. The latter is necessary to ensure that the force of infection is positive. The estimate for the force of infection is negative whenever $\eta'(a) < 0$, since $\pi(a)$ is strictly positive. Therefore, one should fit the fractional polynomials subject to the constraint that $\eta'(a) \geq 0$, for all ages a in the predefined range.

The model for the force of infection as discussed above is derived for a fixed effects model. In Section 9.4.1 we discuss equivalent models for the force of infection in case that random effects are included in the model.

9.2 Logistic Regression

First, we ignore the clustering by treating the animals as if they were all independent. In this case, we can assume a logistic regression model for the binary variable Y_{ij}

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernouilli}(\pi_{ij}) \\
\eta_{ij} &= \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = f(a_{ij}),
\end{aligned}
\tag{9.5}
$$

with $Y_{ij} = 1$ when animal j of herd i has antibodies to glycoprotein B of BHV-1, and 0 otherwise, and $f(a_{ij})$ a functional form describing the dependence of the parameters in the probability model with age, e.g., a fractional polynomial as defined in (3.11).

For the logistic regression model, the fractional polynomial $f(a_{ij}) = \beta_0 + \beta_1 a_{ij} + \beta_2 a_{ij}^{1;2}$ is selected as the best fitting predictor model. The fractional polynomial is fit subject to the constraint of monotonicity. Parameter estimates are summarized in Table 9.1 ($p_1 = 1, p_2 = 1.2$). The estimated model, with 95% confidence intervals, are shown in the left panel of Figure 9.1. The dots represent the observed prevalence of an infected animal, per age class (month), ignoring the clustering of animals within herds. The prevalence of BHV-1 antibodies increases with age. The force of infection can be derived from equation (9.3), and is depicted in the right panel of Figure 9.1. The maximal force of infection (0.089) is obtained at 37 months of age.

Table 9.1: Analysis of maximum likelihood estimates of a logistic regression using a fractional polynomial with $p_1 = 1, p_2 = 1.2$.

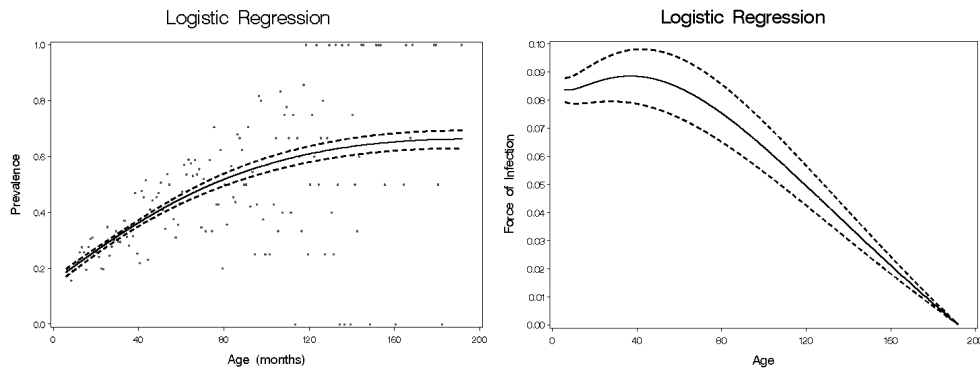| Parameter	Estimate	S.E.	t Value	Pr > |t|
Intercept	-1.7526	0.05871	-29.85	<0.0001
age^{p_1}	0.9083	0.04557	19.93	<0.0001
age^{p_2}	-0.4347	0.02181	-19.93	<0.0001

Figure 9.1: *From left to right: The fitted prevalence and force of infection according to the logistic regression model.*

Although this model seems to fit the data quite well, we have overlooked certain aspects in the data. First, the assumption that animals within herds are independent will in general be too strong. While this typically leaves the consistency of point estimation intact, the same is not true for measures of precision. In case of a "positive" clustering effect (i.e., animals within a herd are more alike than between herds), then ignoring this aspect of the data overestimates precision and hence underestimates standard errors and lengths of confidence intervals. This might result in spuriously 'significant' results. Secondly, the outcome of interest is related to the cluster size. This is termed 'informative cluster size' (Hoffman *et al.* 2001). Animals selected from a large herd have higher probability to be infected compared to animals selected from a small herd, as discussed in Section 9.3.2 and displayed in Figure 9.3. The logistic regression model weighs each animal equally. As a result, large clusters have more impact on the analysis in comparison with small clusters. Thus, the prevalence of

Table 9.2: Estimates according to the generalized estimating equations model, using an independence or exchangeable correlation matrix and a fractional polynomial with powers $p = (1.0, 1.2)$. Model-based and empirical standard error estimates.

Parameter	Estimate	Empirical S.E.	Model-based S.E.
Independence Working Correlation:			
Intercept	-1.775	0.278	0.089
age^{p_1}	0.954	0.312	0.142
age^{p_1}	-0.464	0.187	0.089
Exchangeable Working Correlation:			
Intercept	-2.519	0.261	0.176
age^{p_1}	1.306	0.296	0.113
age^{p_2}	-0.651	0.175	0.068

an infected animal will be estimated as relatively high, compared to a method that weighs each cluster equally. In the next sections, we present alternative methods that incorporate both aspects of the clustering.

9.3   Population-Averaged Model

A marginal model evaluates the overall, or population-averaged, trend as function of covariates while accounting for the correlations in the data. The existence of clustering is recognized but considered a nuisance characteristic. Marginal models based on generalized estimating equations can be used for this purpose.

9.3.1   Generalized Estimating Equations

Liang and Zeger (1986) proposed the generalized estimating equations (see Section 3.2.1). The GEE method requires only the correct specification of the univariate marginal distributions provided one is willing to adopt working assumptions about the association structure.

For binary data, one can fit the marginal logistic regression model (9.5), while correcting the estimated standard errors for clustering. For simplicity, we assume
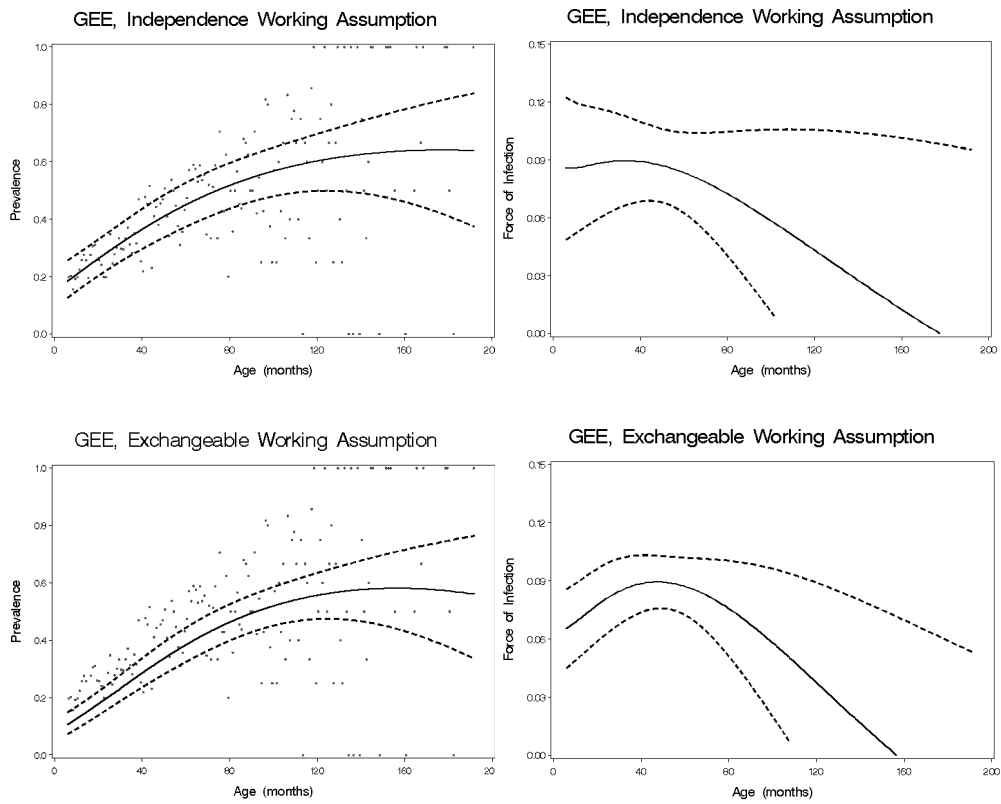
Figure 9.2: *From left to right: The fitted prevalence and force of infection according to the generalized estimating equations model. Top panel presents the GEE model with independence working correlation. The bottom panel presents the GEE model with exchangeable correlation matrix.*

an independence working correlation matrix. This choice is justified since the GEE method is robust against misspecification of the working correlation structure, at the cost of efficiency of the parameter estimates. The same fractional polynomial as in the logistic regression was fit (powers $p = (1.0, 1.2)$), but without the constraint of monotonicity because of computational complexity. The results are shown in the top panel of Figure 9.2. The dashed lines are the 95% Wald-based confidence intervals. The maximal force of infection (0.090) occurs at 34 months of age. The model-based and empirically corrected variance estimator, as defined in Section 3.2.1, are given in Table 9.2. The parameter estimates are very close to the logistic regression model, but account for the correlation in the data yielding higher standard errors. This was expected, since ignoring correlations in the data can lead to serious underestimation of the variances. The wide confidence intervals at the higher age levels, as seen in Figure 9.2, are also a result of ignoring the constraint of monotonicity.

Note that the working correlation structure does not need to hit the true correlation structure to obtain valid inferences. However, to increase the efficiency of the parameter estimates it is better to choose a working correlation matrix that is close to the true one. As another typical choice, we consider the working correlation matrix to be exchangeable,

$$\text{Corr}(Y_{ij}, Y_{ik}) = \alpha \qquad (j \neq k), \tag{9.6}$$

hypothesizing that the correlation between any two animals within a herd is constant. The corresponding model fit is shown in the bottom panel of Figure 9.2, and parameter estimates are displayed in Table 9.2. The maximal force of infection (0.089) occurs at 47 months of age. The intercept is much lower in the GEE model with exchangeable correlation compared with the GEE model with an independence working correlation or the logistic regression model. Indeed, also Figure 9.2 seems to display an underfitting of the prevalence of infection. This difference needs further explanation.

9.3.2 Informative Cluster Size

The cluster size is informative when the cluster size is related with the outcome of interest. In the BHV-1 study, cluster sizes vary from 1 to 264 animals per herd. For simplicity, we categorize the herdsizes in three groups:

$$\text{type} = \begin{cases} \text{small} & \text{herdsize } \leq 30; \\ \text{medium} & 30 < \text{ herdsize } < 60; \\ \text{large} & 60 \leq \text{ herdsize }. \end{cases} \tag{9.7}$$
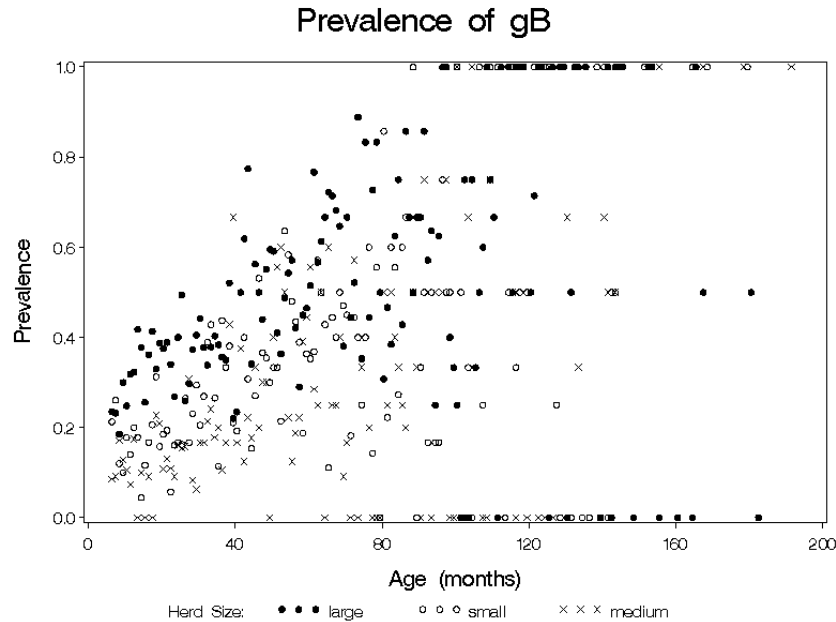
Figure 9.3: *Observed prevalence of a randomly selected animal from a herd of a specific herdsize (small/medium/large)*

The average herd size is 40. Farms with more than 60 cattle are economically viable farms, profiting of the large investments (equipment, machinery, ...) needed. The small farms, with less than 30 cattle, typically are hobby farms. In between these two categories, there are family farms, where family members provide the majority of the labor. The prevalence of a randomly sampled animal from the set of all animals per age-category (month) and per type is displayed in Figure 9.3. It can be seen that animals from large herds have a higher prevalence than animals from small herds.

When cluster size is informative, two marginal analyses can be of interest. First, we might be interested in the probability of a randomly sampled animal from the set of all animals. Secondly, interest can be in the probability of a random animal from a randomly selected herd. These two marginal analyses will have the same asymptotic parameter estimates, except when cluster size is related to the outcome.

To clarify this issue, let us give some other examples. Suppose the cluster of interest is a set of pregnancy outcomes for each of a random sample of women (Hoffman *et al.* 2001). The probability of spontaneous abortions is of interest. The cluster size will be related to risk, because women at high risk of abortion need to have more pregnan-

cies on average to achieve their desired family size. Thus, cluster size is informative. In this example, the cluster-based analysis corresponds to the risk for a randomly-sampled woman. In the context of the developmental toxicity studies, litters with more foetuses are weighted more than litters with fewer foetuses (Williamson *et al.* 2003). When the baseline risk for a birth defect is negatively related with the cluster size, then the GEE method will estimate the risk for a birth defect as relatively low, compared with a method that weights each litter equally. It is clear that when cluster size is informative, sometimes the cluster-based marginal model may be more relevant than the observation-based model.

This also means that when interest in a marginal analysis with the herd as the sampling unit, we may not be too confident in the plotted observed prevalences in Figures 9.1-9.2. They do not present the probability that an animal from a randomly chosen herd is infected, but rather the probability that an animal from the set of all animals is infected. The latter probability is highly affected by the large herds, in which the prevalence is higher. Both the logistic regression model and the generalized estimating equations with independence working correlation matrix give each animal the same weight, and consequently, provide estimates that should be interpreted as related to a randomly sampled animal from the set of all animals. An alternative method is of interest.

9.3.3 Cluster Weighted Generalized Estimating Equations

Williamson *et al.* (2003) and Hoffman *et al.* (2001) discuss the problem of informative cluster size, and present alternative marginal methods with the cluster as sampling unit. Williamson *et al.* (2003) present a modification of the generalized estimating equations (GEE) for handling binary response data with informative cluster size. They propose the use of weighted generalized estimating equations, where the contribution to the estimating equation from a herd is weighted by the inverse of the cluster size, with an independence working correlation matrix. As a result, all herds are given equal weight and individuals in large clusters are no longer over-weighted. The marginal parameter in the cluster-weighted GEE (CWGEE) will have a cluster-based interpretation. This is in direct contrast to GEE, where large clusters are weighted more than small clusters.

Results are displayed in Table 9.3 and Figure 9.4. This model present the probability to be infected, for an animal from a randomly selected herd. Thus, this model accounts for the different selection probabilities of an animal within a herd. As expected, the resulting prevalence is much lower than the one obtained from the GEE

Table 9.3: Estimates according to the cluster weighted generalized estimating equations model, using an independence correlation matrix. Model-based and empirical standard error estimates.

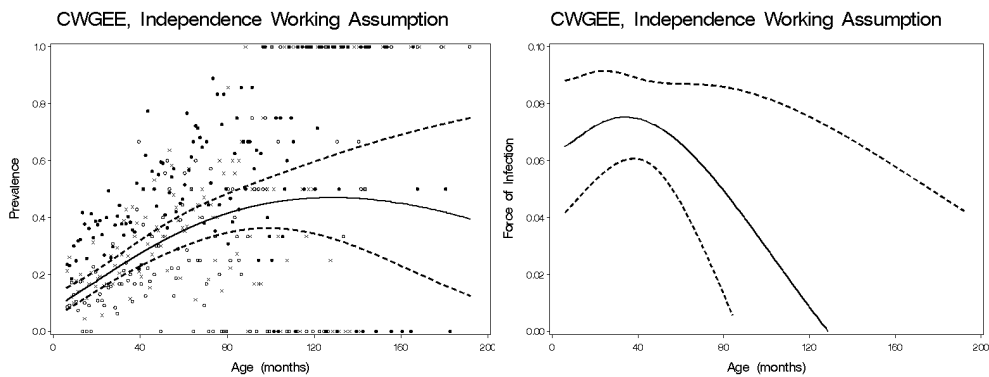| | | Empirical | Model-based |
Parameter	Estimate	S.E.	S.E.
Intercept	-2.4804	0.280	0.523
age^{p_1}	1.3189	0.383	0.793
age^{p_2}	-0.6839	0.235	0.488

Figure 9.4: *From left to right: The fitted prevalence and force of infection according to the cluster weighted generalized estimating equations model, using an independence working correlation matrix.*

approach with independence structure. This also has an effect on the population averaged force of infection, as shown in Figure 9.4. The maximal force of infection is obtained at 31 months of age.

9.3.4 Behaviour of GEE with Exchangeable Working Correlation

An explanation of the behaviour of the GEE model with an exchangeable working correlation structure, as was shown in Figure 9.2, still needs to be given. The generalized estimating equations implicitly presume that the size of the cluster is unrelated to the parameters under study. Under the assumption that the marginal mean is correctly specified, the GEE will produce unbiased estimates through the sandwich estimator, even when the working correlation matrix is not equal to the true covariance. Although, when an important covariate is missing in the marginal mean model, the choice of the working correlation seems to have an important effect, even on the estimated mean model.

In the GEE approach, the correlations in the working covariance matrix can be seen as weights that are assigned to the data from each cluster. However, when cluster size is informative, then weighting the data in different ways, as is done by choosing different working correlations, can result in different marginal models. In a GEE model with independence working correlations, each observation is given the same weight. However, as seen in Figure 9.2, it seems that, when using an exchangeable working correlation, large herds are down-weighted such that large herds have less influence on the overall trend. Intuitively, one might explain the behaviour of the GEE with exchangeable working correlation as follows. When accounting for the correlation of animals within a herd, the effective sample size is reduced (Chapter 5). The working correlation in the BHV study was estimated as 0.6294. According to (5.7) a herd of size 1, 10 or 100 corresponds with an effective sample size of 1, 1.50 and 1.58, respectively. Thus, by assuming a working correlation matrix that is different from the independence working correlation, the GEE method seems to partly account for the informative herdsize. Indeed, the parameter estimates of the GEE with exchangeable correlation matrix (Table 9.2) are close to the ones obtained from the cluster-weighted GEE (Table 9.3). An extensive simulation study is needed to fully understand the behaviour of the GEE when dealing with informative cluster sizes.

The CWGEE seems to be a nice alternative to the GEE in case of informative cluster sizes. However, the performance of the CWGEE with a working correlation matrix different from the independence working correlation, is not yet understood.
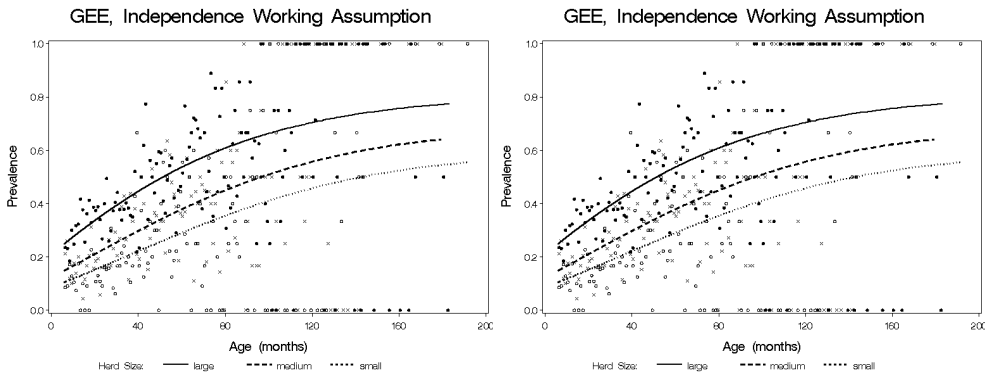
Figure 9.5: *The fitted prevalence and force of infection according to the generalized estimating equations model, corrected for herdsize (large, medium and small herdsize, respectively). Use of independence working correlations.*

The weighting scheme resulting from the inclusion of a working correlation matrix may changes the weights given to each cluster. Therefore, another strategy to account for the herdsize is investigated.

9.3.5   Herdsize as Covariate

Informative cluster sizes can be defined as any violation of the property that $E[Y_{ij}|X_{ij}, n_i] = E[Y_{ij}|X_{ij}]$. Thus, one could take the herdsize as a covariate in the model. As a first exploratory analysis, we include the herdsize as a categorical covariate in the GEE model. The model fit is shown in Figure 9.5 and parameter estimates are summarized in Table 9.4. There are clear differences between the small, medium and large herdsize. The GEE model fits the data very well. The force of infection is maximal at 37, 56 and 66 months of age, for large, medium and small herds, respectively.

The cluster-weighted GEE and the inclusion of herdsize as a covariate are two methods to account for the nonignorable cluster size. Both approaches yield results that seem to be consistent in interpretation. Weighting the GEE pulls the marginal probability downwards. The same thing happens when including cluster size as a covariate in the model. In the BHV-1 study, we prefer the use of the herdsize as a covariate over the cluster-weighted GEE method. Whereas the CWGEE gives us an estimate of the probability of a randomly selected animal from a randomly selected herd, the second method gives us an estimate of the probability of an infected animal

Table 9.4: Estimates according to the generalized estimating equations model corrected for herdsize and using an independence correlation matrix. Model-based and empirical standard error estimates.

Parameter	Estimate	Empirical S.E.	Robust S.E.
Intercept	-2.396	0.107	0.274
large	1.039	0.072	0.291
medium	0.402	0.079	0.260
age^{p_1}	0.820	0.146	0.323
age^{p_2}	-0.377	0.091	0.195

from a herd with a specific herdsize.

9.3.6   Overview Methods

When cluster size is a random variable or is related with the response, one needs to be careful with a marginal analysis. There are several ways to account for an informative herdsize. Depending on the data we consider, one method can be more useful compared to another method. A short overview of possible approaches is given:

- A first method, as proposed by Williamson *et al.* (2003), is the use of a cluster weighted generalized estimating equations. This approach has a cluster-based interpretation. In the example of the pregnancy outcomes, a cluster-based model may be more generalizable than an observation-based model.

- An alternative method, as proposed in Chapter 7, is to build a hierarchical model by assuming herdsize as a random variable. In the situation of the toxicological experiment, the litter size depends on the dose level and can be seen as an additional response of interest. In this case, the hierarchical model approach may be most suitable.

- Cluster size could also be included as a covariate in the model. An appealing property of using it as a covariate is, that it gives the opportunity to predict different probability curves for clusters with different cluster sizes. In the example of the BHV-1 study, where there is interest in finding the risk factors for

the disease, this method might be most appropriate.

- In some situations, clusters have a maximum number of units. In this case, the varying cluster size can be thought of as a missing data problem. Weighted estimating equations, such as those proposed by Robins *et al.* (1995) may be used in this situation. Dunson (1998) also proposed a multiple imputation scheme to estimate the number of missing foetuses.

9.4   Random Effects Model

In contrast to the previous view-point, one can have a genuine scientific interest in the clustering itself. Animals belonging to a herd share the same environment (physical location), as well as characteristics such as the type of farm (milk- or meat-oriented) and other unobserved factors (Speybroeck *et al.* 2003). A random-effects, or cluster-specific, model describes the dependencies between responses because of shared factors in a herd. For example, one can account for intra-cluster correlation by inducing cluster-specific effects, and study the herd-specific profiles. In this section, a generalized linear mixed model is used to model the infection rate as function of covariates and parameters specific to a herd.

First, consider the case where the random effects represent herd-specific intercepts. Conditionally on the random intercepts u_i, it is assumed that the individual-level outcomes are independently distributed as:

$$
\begin{aligned}
Y_{ij}|u_i &\sim \text{Bernouilli}(\pi_{ij}) \\
\eta_{ij} &= \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = f(a_{ij}) + u_i.
\end{aligned}
\tag{9.8}
$$

The random effects u_i are assumed to be sampled from a normal distribution with mean 0 and variance σ_u^2. The herd-specific profiles are displayed in Figure 9.6. A fractional polynomial of the form $f(a_{ij}) = \beta_0 + \beta_1 a_{ij}^{0.4} + \beta_2 a_{ij}^{0.6}$ was chosen as the best fitting model. The value of σ_u^2, the between-herd variance, was estimated as 8.12 (SE=1.03), indicating that there are large differences among herds. The intra-herd correlation coefficient ρ equals 0.712 (SE=0.026). Thus, animals within herds are highly correlated.

In case one is interested in the population-averaged risk of infection, one needs to integrate over the random effects:

$$
\int_{-\infty}^{+\infty} \frac{\exp\{\beta_0 + \beta_1 a_{ij}^{p_1} + \beta_2 a_{ij}^{p_2} + u_i\}}{1 + \exp\{\beta_0 + \beta_1 a_{ij}^{p_1} + \beta_2 a_{ij}^{p_2} + u_i\}} \frac{1}{\sqrt{2\pi}\sigma_u} e^{-u^2/(2\sigma_u^2)} du.
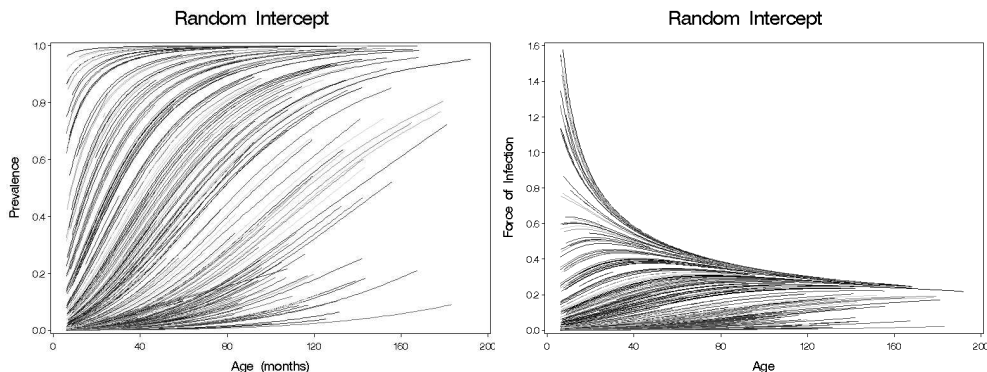\tag{9.9}
$$

Figure 9.6: *From left to right: The herd-specific profiles of prevalence and force of infection according to a generalized random intercept model.*

In Figure 9.7, the marginal fit is shown. The dots represent the probability that an animal is infected, per age and per herdsize. The line represents the marginal effect, obtained by integrating over the random effect. Note that this model has a cluster-based interpretation, since each herd gets the same weight in the calculation of the overall trend. This marginalized model is very close to the result obtained from the cluster-weighted GEE. For ages above 120 months, the differences are probably due to the limited amount of data at higher age levels.

Second, suppose that the random effects represent unit-specific intercepts as well as linear age effects. The corresponding model is of the form:

$$
\begin{aligned}
gB_{ij}|u_i &\sim \text{Bernouilli}(\pi_{ij}) \\
\eta_{ij} &= \log\Big(\frac{\pi_{ij}}{1-\pi_{ij}}\Big) = f(a_{ij}) + u_{0i} + u_{1i}a_{ij},
\end{aligned}
\tag{9.10}
$$

with the random effects parameters sampled from a bivariate normal distribution. The same fractional polynomial as in the random intercepts model is used. Results are shown in Figure 9.8. Each herd has its own age-specific profile, with very large differences among herds.

9.4.1 Cluster Specific Force of Infection

Both the logistic regression model and the marginal model discussed in the previous sections include only fixed effects (the fractional polynomial of the age). Therefore, the population force of infection is $\ell(a) = g(\eta(a))\eta(a)'$. The generalized linear mixed model specified in (9.8) account for possible intra-cluster correlation by including a
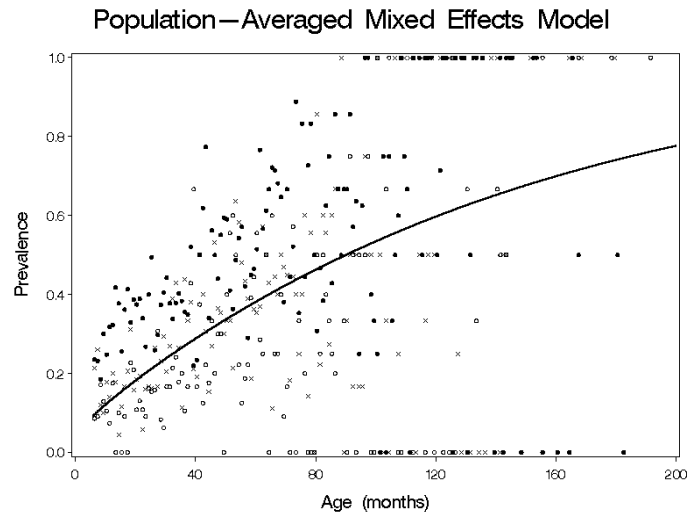
Figure 9.7: *Population-averaged profile of prevalence, based on a generalized mixed model with random intercept.*
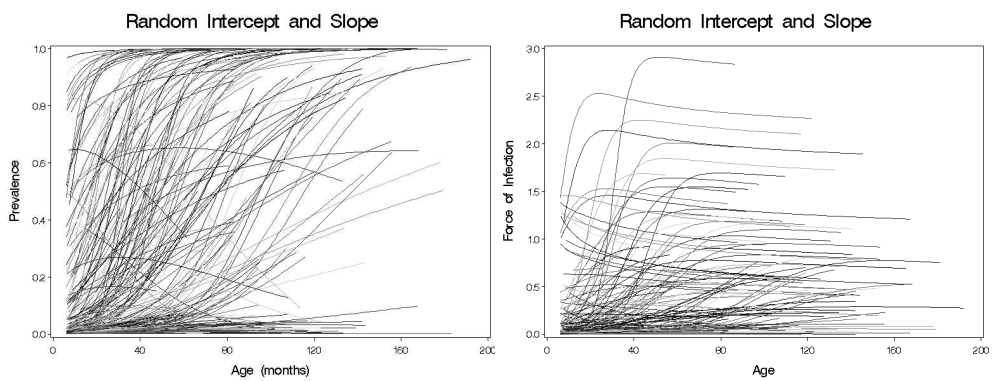

Figure 9.8: *From left to right: The herd-specific profiles of prevalence and force of infection according to a generalized mixed model with random intercept and age.*

random intercept effect for the cluster. As a consequence, each cluster has a specific force of infection. Let $f(a) = \beta_0 + \sum_{j=1}^{m} \beta_j a^{p_j}$, and consider a fractional polynomial model with random intercept of the form

$$h(\pi(a)) = f(a) + u_i, \quad i = 1 \ldots, K,$$

where K is the number of clusters. Note that $f(a)$ represent the fixed part of the linear predictor so that $\eta(a|u_i) = f(a) + u_i$. The cluster-specific force of infection can be calculated by $\ell(a|u_i) = g(\eta(a|u_i))\eta(a|u_i)'$. For a random intercept model,

$$\eta(a|u_i)' = \frac{\partial \eta(a|u_i)}{\partial a} = \frac{\partial f(a)}{\partial a} = f(a)',$$

and it does not involve the random intercept u_i. Hence, in the general case the cluster specific force of infection is $\ell(a|u_i) = g(\eta(a|u_i))f(a)'$ and for a model with logit link function, $\ell(a|u_i) = \pi(a|u_i)f(a)'$. Thus, for a logit model with random intercept and for two clusters with the same age

$$\frac{\ell(a|u_j)}{\ell(a|u_k)} = \frac{\pi(a|u_j)}{\pi(a|u_k)}.$$

The cluster-specific force of infections, based on the random intercept model, are displayed in Figure 9.6.

The population force of infection (or the marginal force of infection) can be derived by

$$\ell(a) = f(a)' \int g(\eta(a|u))du. \tag{9.11}$$

Note that for a model with a logit link function the second term in the right hand side of (9.11) is equal to the population average probability given in (9.9).

For a model with random intercept and slope as presented in (9.10), i.e., $\eta(a|u_{0i}, u_{1i}) = f(a) + u_{0i} + u_{1i}a$, we have

$$\frac{\partial \eta(a|u_{0i}, u_{1i})}{\partial a} = \frac{\partial f(a)}{\partial a} + u_{1i},$$

and the cluster-specific force of infection is

$$\ell(a|u_{0i}, u_{1i}) = \left(\frac{\partial f(a)}{\partial a} + u_{1i} \right) \pi(a|u_{0i}, u_{1i}).$$

Results are shown in the right panel of Figure 9.8. Large differences between herds are seen. Both the magnitude of the force of infection, as well as the age at which the force of infection is maximal differ among herds. These differences between herds can be due to both observed or unobserved variables. One possible factor is the herdsize.

In Figure 9.9, the force of infection within a herd is displayed per type: large, medium or small herds, as defined in (9.7). Small farms show much lower force of infections, in comparison with large herds. An extensive investigation of possible risk factors is topic of further research.

As mentioned above, for fixed effect model, the force of infection is non negative as long as $\eta(a)' \geq 0$. For a random intercept model, $\eta(a|u_i)' = \partial f(a)/\partial a$, so the force of infection will be non negative as long as $\partial f(a)/\partial a \geq 0$ since $\pi(a|u_j)$ is a monotone non decreasing function. The case of random intercept and slope is somewhat more problematic since even if $\partial f(a)/\partial a \geq 0$ the first derivative of the linear predictor $\left(\frac{\partial f(a)}{\partial a} + u_{1i}\right)$ is not necessarily positive, and there is no guaranty that the cluster specific prevalence $\pi(a|u_{0i}, u_{1i})$ is a monotone non decreasing function with respect to age. This can be seen in Figure 9.6 (for the random intercept model) and Figure 9.8 (for the model which include random intercept and random slope).

9.5   Discussion

The choice between population-averaged and random effects strategies may heavily depend on the scientific goals. Population-averaged models are appropriate to evaluate the overall prevalence as a function of covariates. The population averaged force of infection can be investigated by the use of the (cluster weighted) GEE model. With a herd-specific approach, the prevalence is modelled as a function of covariates and parameters, specific to a herd, so that herd-specific profiles can be investigated.

So far, we have investigated the age-specific disease pattern. However, it is important to investigate which factors are of risk for the prevalence of BHV-1. We performed a first model selection procedure for the random-effects model. An overview of the model selection procedure is given in Table 9.5. Both Akaike's Information Criterion (AIC) and Bayesian Information Criteria (BIC) values are given. Since the BIC accounts for the sample size in the penalizing constant, this method is preferred for model selection. Indeed, when dealing with large surveys, weak effects can show significant results because the large sample provides small standard errors. Thus, a statistically significant effect need not be important in a practical sense in such case. In addition, use of the effective sample size, or the number of independent samples contained in each herd such as introduced in Chapter 5, might be more appropriate in the definition of BIC (BIC$^{(E)}$, Faes *et al.* 2004b). In Table 9.5, BIC$^{(E)}$ represents the BIC with the effective sample size as penalizing constant, BIC$^{(A)}$ is the BIC using the number of animals in the study, and BIC$^{(H)}$ corresponds with the BIC corrected
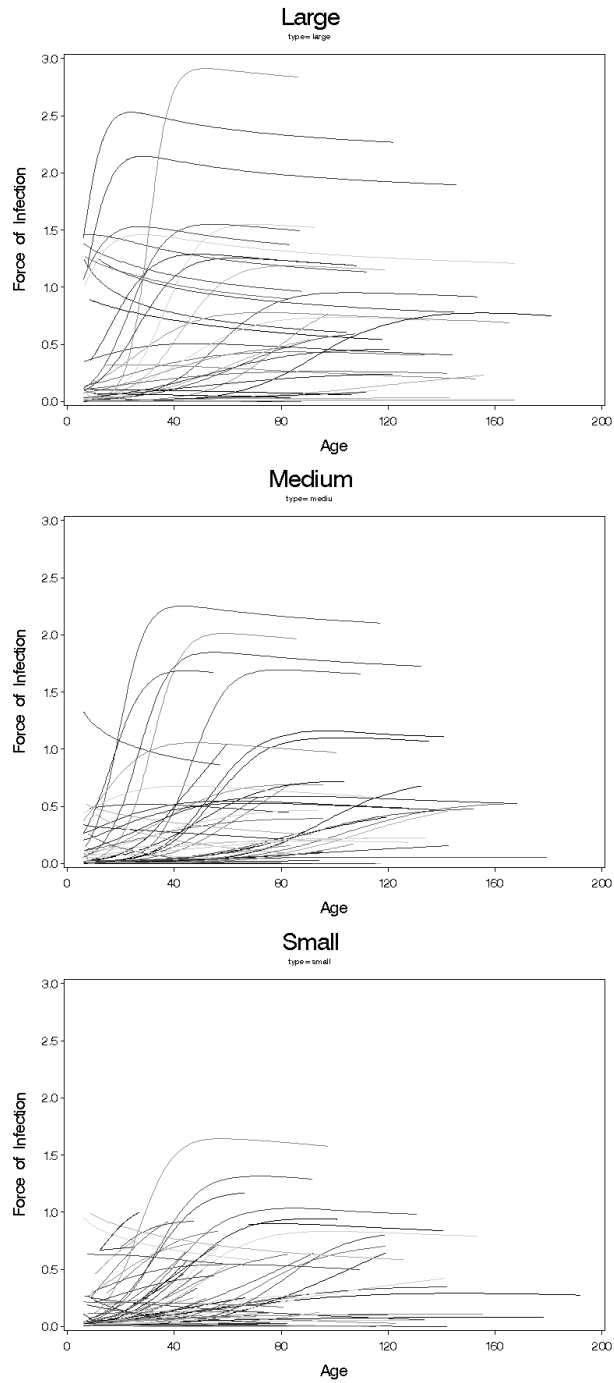
Figure 9.9: *From top to bottom: The herd-specific profiles of force of infection according to a generalized mixed model with random intercept and age for different herdsize: large, medium and small, respectively.*

Table 9.5: Overview stepwise model selection

effect	AIC	BIC$^{(E)}$	BIC$^{(A)}$	BIC$^{(H)}$
f(age)	5352.2	9575.8	5394.4	5374.1
+ herdsize	5343.1	5370.1	5392.3	5368.6
+ purchase	5337.5	5369.0	5393.7	5366.6
+ herdsize^2	5331.8	5367.2	5395.0	5364.6
+ age*herdsize	5322.2	5361.6	5392.5	5358.6
+ age*herdsize^2	5317.0	5360.3	5394.3	5357.1

for the number of herds. The final model contains the animal-level factors age and origin (purchased or homebred), and the herd-level covariate herdsize. The model is summarized in Table 9.6. The prevalence of gB-antibodies in cattle increases with the age of the animal, and also with the size of the herd. Purchased animals also have higher probability to be infected than homebred animals. An extensive analysis of possible risk factors is topic of current research.

In the data, 2148 records of the 11284 records have at least one missing value in response and covariates. There is a substantial amount of missingness in the variable purchase, being an important risk factor as indicated in previous analysis. The first issue in dealing with the problem is determining whether the missing data mechanism has distorted the observed data. So far, we considered the complete cases only. However, some first analysis showed that the missingness has a noticeable effect on the estimation of prevalence and force of infection. Dealing with the missingness in the data, together with the clustering, is under investigation.

Some other issues in this study need further attention. First, the generalized estimating equations should be fit under the constraint of monotonicity, in order to get a positive force of infection. Standard statistical software for fitting a GEE model do not allow the inclusion of constraints on the parameters. Secondly, the behaviour of the cluster-weighted GEE under a non-independence working correlation should be investigated further.

Table 9.6: Estimates of final logistic random-effects model.

Parameter	Estimate	S.E.	t Value	Pr > | t |
Intercept	-6.953	0.925	-7.52	<0.0001
$age^{0.4}$	4.655	2.040	2.28	0.0232
$age^{0.6}$	-2.546	1.371	-1.86	0.0644
herdsize	0.018	0.013	1.34	0.1803
$herdsize^2$	-0.00004	0.00007	-0.60	0.5498
age*hsize	0.01556	0.00398	3.91	0.0001
$age*hsize^2$	-0.00006	0.00002	-2.77	0.0060
purchase	0.280	0.098	2.86	0.0046
s11	7.921	1.345	5.89	<0.0001
s12	-0.450	0.255	-1.76	0.0793
s22	0.569	0.127	4.49	<0.0001

Chapter 10

Spatial Modelling of Spikes in Small Intestines

Mixed models can be applied in a wide range of settings. Probably, they are most commonly used to handle grouping in the data. Examples for this type of data include longitudinal data, clustered data, hierarchical data, etc. In addition, mixed models can be used for smoothing purposes as well. Speed (1991) explicitly made the connection between nonparametric regression and mixed models. Since then, the use of smoothing splines within the mixed model framework is becoming more and more appreciated and the number of applications in literature is growing. Ruppert *et al.* (2003), Wand (2003), Ngo and Wand (2004) all used mixed models for smoothing purposes. In case when dealing with non-normal data, the adoption of smoothing methods within generalized linear mixed models (GLMM) is less familiar. The main difficulty in case of GLMMs is the high dimensional integral over the unobserved random effects, which, in general, cannot be computed exactly. A key feature of the mixed model approach is that it allows us to handle both correlations and smoothing in a unified framework.

In this chapter, we consider the cat duodenum data, as described in Chapter 2. The numbers of spikes occurring at specific locations on an isolated segment of the small intestines of a cat are recorded during 12 to 16 successive slow waves. So far, in both the temporal and spatial dimension, there is no information available about the behaviour of spikes in successive slow waves. One question of interest is whether there are areas with high spike incidence, compared with other areas. Some preliminary analysis were performed in this context in Lammers *et al.* (2004) and

showed important spatial effects. Another question is about the temporal and spatial characteristics of spikes during successive slow waves. In this chapter, we summarized the general mixed model design for discrete data as proposed by Ruppert, Wand and Carroll (2003), and investigate how it can be used for the analysis of the cat duodenum data.

In Sections 10.1 and 10.2 we give a brief introduction on generalized linear mixed models and penalized quasi-likelihood. In Section 10.3, it is explained how smoothing methods can be incorporated within the mixed models framework. First, the method is explained for the penalized regression splines in one dimension. Later, it is extended for smoothing purposes in two dimensions. Next to smoothing, the mixed model is used to account for correlations in the data. The general design of GLMMs is emphasized in Section 10.5. An illustration of the use of generalized linear mixed models for smoothing purposes, and in addition accounting for correlations in the data, is given in Section 10.6.

10.1　Generalized Linear Mixed Model

For a generalized response, such as a binary or count variable, the generalized linear mixed model (GLMM) is a commonly used random effects model (McCullagh and Nelder 1989). The GLMM is a very powerful tool, allowing many complications to be handled. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ denote the vector of all measurements for the ith group $(i = 1, \ldots, N)$. Conditional on the random effects $\boldsymbol{u}_i$, it is assumed that the measurements Y_{ij} of $\boldsymbol{Y}_i$ are independent with density function of an exponential family form:

$$f_i(y_{ij}|\boldsymbol{u}_i) = \exp\Big\{\phi^{-1}\big[y_{ij}\eta_{ij} - a(\eta_{ij})\big] + c(y_{ij}, \phi)\Big\},$$

with η_{ij} the natural parameter, ϕ an overdispersion parameter and a a function satisfying

$$\begin{aligned}
\mathrm{E}\big(Y_{ij}|\boldsymbol{u}_i\big) &= a'(\eta_{ij}), \\
\mathrm{Var}\big(Y_{ij}|\boldsymbol{u}_i\big) &= \phi a''(\eta_{ij}).
\end{aligned}$$

A general design for GLMM, describing the relation between the conditional mean and the covariates, can be formulated as

$$h\big(\mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{u}_i)\big) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i,$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters, called the fixed effects, and h is

some link function. The random effects $\boldsymbol{u}_i$ are assumed to follow a normal distribution

$$\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{G}),$$

with $\boldsymbol{G}$ a general $(q \times q)$ covariance matrix.

Usually, the canonical link function is used, i.e., $h = (a')^{-1}$, such that $\eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{u}_i$. Commonly used GLMMs are the logistic-normal mixed model for binary data (Williams 1982) which corresponds to

$$a(x) = \log(1 + e^x), c(x) = 0, \phi = 1,$$

and the Poisson-normal mixed model for counts (Breslow 1984) corresponding to

$$a(x) = e^x, c(x) = -\log(x!), \phi = 1.$$

The linear mixed model is a special case of the generalized linear mixed model, with identity link function. Several other distributions (e.g., gamma, inverse Gaussian) fit into this exponential family structure (McCullagh and Nelder 1989). In this chapter, it is assumed that the canonical link function is used.

10.2 Penalized Quasi-Likelihood

The non-linear nature of the GLMM implies that model fitting is, in general, not straightforward. Maximum likelihood estimation requires the marginal likelihood of $\boldsymbol{Y}$, which is obtained by integrating over the random effects. The contribution of the ith group to the log marginal likelihood can be written as

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{G}) = \log f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{G}) = \log \int_{\mathbb{R}^{n_i}} \prod_{j=1}^{n_i} f_i(y_{ij}|\boldsymbol{u}_i)\phi(\boldsymbol{u}_i; \boldsymbol{G})d\boldsymbol{u}_i, \tag{10.1}$$

where $\phi(\boldsymbol{u}_i; \boldsymbol{G})$ is the normal density function $N(\boldsymbol{0}, \boldsymbol{G})$. The log marginal likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{G}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}, \boldsymbol{G})$$

can then be maximized to obtain estimates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{G}$.

In general, the above equation cannot be derived analytically, and maximum likelihood estimation is hindered by the presence of this n_i-dimensional integral. Several alternative methods to estimate parameters in a random effects model are available. The integral (10.1) can be computed numerically. Different integral approximations are available, the principal one being (adaptive) Gaussian quadrature (Anderson and

Aitkin 1985). Alternatively, the generalized linear random effects model can be cast into a Bayesian framework, avoiding the need for numerical integration by taking repeated samples from the posterior distributions using Gibbs sampling techniques (Zeger and Karim 1991). Other methods use approximations of the likelihood, to circumvent the computational burden caused by numerical integration. A possible method involves Laplace approximation of integrals (Breslow and Clayton 1993, Wolfinger and O'Connell 1993), which is commonly referred to as penalized quasi-likelihood (PQL).

PQL is a relatively simple method for fitting GLMM. Approximation of the marginal likelihood using Laplace's method leads to estimating $(\boldsymbol{\beta}, \boldsymbol{u})$ by treating the random effects $\boldsymbol{u}$ as fixed parameters, but penalizing the likelihood according to the distribution of $\boldsymbol{u}$. As such, $(\boldsymbol{\beta}, \boldsymbol{u})$ are obtained by maximizing the penalized log-likelihood (Green 1987)

$$\log f(\boldsymbol{y}; \boldsymbol{u}) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u},$$

and thus

$$\left[\begin{array}{c} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{array}\right] = \operatorname*{argmax}_{\boldsymbol{\beta}, \boldsymbol{u}} \left\{ \phi^{-1}\left[\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - 1^T a(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right] - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u} \right\}.$$

Implementation of the PQL algorithm involves repeatedly fitting of linear mixed models (Wolfinger and O'Connell 1993), just as generalized linear models may be fitted by repeated calls to weighted least squares procedures (Nelder and Wedderburn 1972). In this procedure, parameter estimates are obtained using a linearized version $\boldsymbol{y}^*$ of the responses $\boldsymbol{y}$. The PQL algorithm proposed by Wolfinger and O'Connel (1993) has been implemented in the %GLIMMIX macro in SAS.

10.3   Smoothing Discrete Data using Mixed Models

A regression curve describes a general relationship between an explanatory variable and a response variable. It is of great interest to have some knowledge about this relation as it may indicate whether a special sort of dependence between the two variables is present. In this section we explain why smoothing methods that use basis functions with penalization, can be formulated as estimators in a generalized mixed model framework.

Consider the observed pairs (y_i, x_i) for each individual i $(i = 1, \ldots, N)$, with y_i a generalized response and x_i a given covariate. Assume that the responses y_i are

drawn independently, and that y_i has density function

$$f(y_i; \eta_i, \phi) = \exp\left\{\phi^{-1}\left[y_i\eta_i - a(\eta_i)\right] + c(y_i, \phi)\right\},$$

where η_i depends on x_i (McCullagh and Nelder 1989). Parametric generalized linear models are constructed by assuming that there exist a link function h such that

$$h\big(E(Y_i|x_i)\big) = \boldsymbol{X}_i\boldsymbol{\beta},$$

with, e.g., $\boldsymbol{X}_i = [\ 1 \quad x_i\ ]$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. In nonparametric regression, the aim is to replace the parameterised predictor $\eta_i = \boldsymbol{X}_i\boldsymbol{\beta}$ by a general predictor $\eta_i = f(x_i)$, where f is a smooth function. In the nonparametric framework the shape of the functional relationship is determined by the data, whereas in the parametric framework the shape is determined by the model.

In this chapter, we focus on semiparametric regression models using piecewise regression splines or radial basis splines as described by Ruppert, Wand and Carroll (2003). We discuss each of them in turn.

10.3.1 Penalized Regression Spline

The unknown smooth function f can be modelled as a piecewise linear smoother, as discussed by Friedman and Silverman (1989). The piecewise linear spline model is given as

$$\eta_i = f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k(x_i - \kappa_k)_+,$$

where x_i are design points $(i = 1, \ldots, N)$, κ_k is the location of the kth knot $(k = 1, \ldots, K)$ and

$$(x_i - \kappa_k)_+ = \begin{cases} 0, & x \leq \kappa_k \\ x_i - \kappa_k, & x > \kappa_k. \end{cases}$$

The basis functions represent broken lines with the knots κ_k as a joint point. Figure 10.1 shows an example of a linear spline basis with equally spaced knots. In matrix notation, with $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\boldsymbol{u} = (u_1, \ldots, u_K)^T$, $\boldsymbol{X}$ the matrix with the ith row $\boldsymbol{X}_i = [1 \quad x_i]$ and $\boldsymbol{Z}$ the matrix with the ith row $\boldsymbol{Z}_i = [(x_i - \kappa_1)_+, \ldots, (x_i - \kappa_K)_+\ ]$, the piecewise linear spline function can be written as

$$f(x_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}.$$

Within the parametric regression framework, f can be estimated by maximizing the likelihood. However, it is always possible to choose f sufficiently complicated that
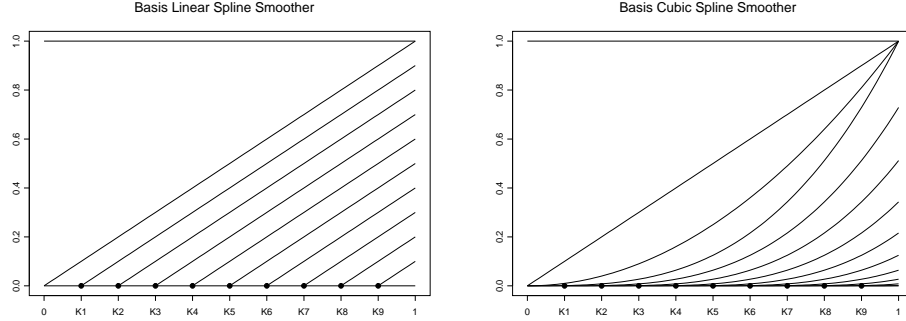
Figure 10.1: *Piecewise regression spline basis function, with 9 equidistant knots. Left: linear spline basis; Right: Cubic spline basis.*

it interpolates the data, in the sense that the fitted values agree with the observed responses $y_i = a'(f(x_i))$. Following Ruppert *et al.* (2003), we consider a number of knots that is large enough (typically 5 to 20) to ensure the desired flexibility. But, to overcome the problem of overfitting the data due to there being too many knots, we constrain their influence. A smooth fit is obtained by maximizing

$$\phi^{-1}\big[\boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Zu}) - 1^T a(\boldsymbol{X\beta} + \boldsymbol{Zu})\big] - \frac{1}{2}\lambda^2 \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \boldsymbol{D} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \qquad (10.2)$$

with $\boldsymbol{D}$ a known positive semi-definite penalty matrix (Wahba 1978, Green and Silverman 1994). The first term measures the goodness-of-fit while the second term is the roughness penalty. The parameter λ is the smoothing parameter. Large values of λ produce smoother curves while smaller values produce wiggly curves. A possible constrain on the parameters $\boldsymbol{u}$ is $\sum_k u_k^2 < C$, for some positive value C. This is equivalent to choosing $(\boldsymbol{\beta}, \boldsymbol{u})$ to maximize the penalized log-likelihood (10.2) with $\boldsymbol{D} = diag(0, 0, 1, \ldots, 1)$. In this case, the penalized spline solution is

$$\hat{\eta}_i = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{u}},$$

where

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{u}}{\operatorname{argmax}} \left\{ \phi^{-1}\big[\boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Zu}) - 1^T a(\boldsymbol{X\beta} + \boldsymbol{Zu})\big] - \frac{1}{2}\lambda^2 ||\boldsymbol{u}||^2 \right\}. \qquad (10.3)$$

10.3.2 Connection with mixed models

An equivalent model representation of the penalized spline, in the form of a generalized linear mixed model, can be found (Wand 2002, Ngo and Wand 2004). Specifically, for a fixed value of λ, (10.3) is equivalent to fitting the generalized linear mixed model

$$
\begin{aligned}
f_i(y_i|\boldsymbol{u}) &= \exp\left\{\phi^{-1}\left[y_i^T(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}) - a(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u})\right] + 1^T c(y_i)\right\}, \\
\boldsymbol{u} &\sim N(0, \boldsymbol{G}),
\end{aligned}
$$

by using the penalized quasi-likelihood approach as described in Section 10.2 (Breslow and Clayton 1993, Wolfinger and O'Connell 1993), with $\boldsymbol{G} = \sigma_u^2 \boldsymbol{I}_{K \times K}$. Given this equivalence, the penalized spline model can be fit using existing statistical software for generalized linear mixed models, such as the SAS macro %GLIMMIX (Wolfinger 1993). The amount of smoothing, $\lambda = 1/\sigma_u$, is automatically selected via the PQL and REML algorithm.

10.3.3 Higher Order Spline Bases

The piecewise linear spline model is conceptually very simple, and will fit the data sufficiently in many applications. However, other spline bases are also possible. Smoother fits can be obtained by the use of higher degree truncated polynomial bases. The truncated power spline function of degree p with fixed knots $\kappa_1, \ldots, \kappa_K$ is given by

$$
\eta_i = f(x_i) = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+^p.
$$

The right panel of Figure 10.1 shows an example of a cubic basis ($p = 3$), with equally spaced knots. When fitting this spline model by penalized least squares, the parameters in this model are estimated as in (10.3) with $\boldsymbol{X}$ the matrix with the ith row $\boldsymbol{X}_i = \begin{bmatrix} 1 & x_i & \ldots & x_i^p \end{bmatrix}$, $\boldsymbol{Z}$ with ith row $\boldsymbol{Z}_i = \begin{bmatrix} (x_i - \kappa_1)_+^p & \ldots & (x_i - \kappa_K)_+^p \end{bmatrix}$ and

$$
D = \left[\begin{array}{cc} \boldsymbol{0}_{(p+1)\times(p+1)} & \boldsymbol{0}_{(p+1)\times K} \\ \boldsymbol{0}_{K\times(p+1)} & \boldsymbol{I}_{K\times K} \end{array} \right].
$$

The same results are obtained by fitting the generalized linear mixed model

$$
\begin{aligned}
\eta_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}, \\
\boldsymbol{u} &\sim N(0, \boldsymbol{G}),
\end{aligned}
$$

with $\boldsymbol{G} = \sigma_u^2 \boldsymbol{I}_{K \times K}$ by a penalized quasi-likelihood approach. The amount of smoothing, $\lambda^{2p} = 1/\sigma_u^2$, is chosen automatically in the mixed model.
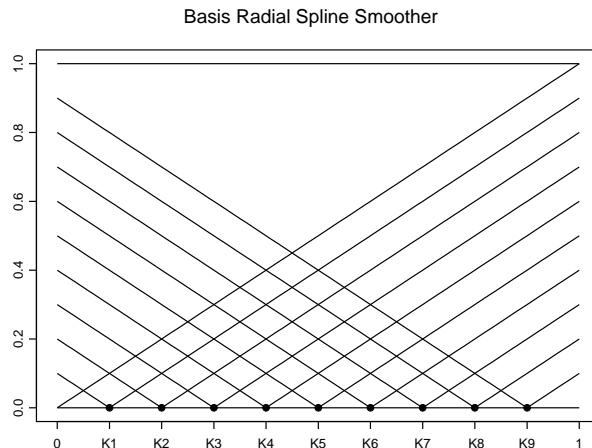
Figure 10.2: *Radial spline basis function (m=1), with 9 equidistant knots.*

10.3.4 Radial Smoothers

Another spline method is to work with radial basis functions. These are basis functions of the form $C(|x_i - \kappa_k|)$ for some univariate function C. For example, the basis function C can be defined as

$$C(r) = |r|^{2m-1},$$

with m a positive value. Figure 10.2 shows the basis functions for $m = 1$. A characteristic of radial basis functions is that they only depend on the distances between the data and the knots. As a consequence, the function is radially symmetric about this point. The advantage of this property is that it easily extends to higher-dimensional predictor variables.

As in previous sections, the spline function can be written as $\boldsymbol{X\beta} + \boldsymbol{Z}_K\boldsymbol{u}$ where (Nychka 2000)

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_i & \dots & x_i^{m-1} \end{bmatrix}_{1 \leq i \leq N} \text{ and } \boldsymbol{Z}_K = \left[|x_i - \kappa_k|^{2m-1}\right]_{1 \leq i \leq N, 1 \leq k \leq K}.$$

In order to obtain a smooth fit, an appropriate penalty factor must be chosen. Green and Silverman (1994) showed that maximizing the penalized log likelihood

$$\boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Z}_K\boldsymbol{u}) - 1^T a(\boldsymbol{X\beta} + \boldsymbol{Z}_K\boldsymbol{u}) - \frac{1}{2}\lambda^{2m-1}\boldsymbol{u}^T\boldsymbol{\Omega}_K\boldsymbol{u}$$

with

$$\boldsymbol{\Omega}_K = \left[|\kappa_k - \kappa_{k'}|^{2m-1}\right]_{1 \leq k,k' \leq K; 1 \leq i \leq N}$$

corresponds with the thin plate spline family of smoothers. The thin-plate spline is a generalization of the usual cubic smoothing spline with a roughness penalty function that is rotationally invariant. The theoretical foundations for the thin-plate smoothing spline are described in Duchon (1976, 1977) and Meinguet (1979).

Ruppert, Wand and Carroll (2003) noted that radial smoothers can be described by the following generalized linear mixed model:

$$
\begin{aligned}
h\big(E(\boldsymbol{Y}|\boldsymbol{u})\big) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_K\boldsymbol{u}, \\
\boldsymbol{u} &\sim N(0, \sigma_u^2(\boldsymbol{\Omega}_K^{-1/2})(\boldsymbol{\Omega}_K^{-1/2})^T).
\end{aligned}
$$

Using the transformation $\boldsymbol{Z} = \boldsymbol{Z}_K\boldsymbol{\Omega}_K^{-1/2}$, this results in

$$
\begin{aligned}
h\big(E(\boldsymbol{Y}|\boldsymbol{u})\big) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \\
\boldsymbol{u} &\sim N(0, \sigma_u^2\boldsymbol{I}_{K \times K}).
\end{aligned}
$$

This form allows fitting through standard mixed model software.

These radial bases are particularly useful for higher-dimensional smoothing, as will be shown in Section 10.4.2.

10.3.5 Illustration: BHV-1 Study

For illustration, consider the Bovine Herpesvirus-1 study as described in Chapter 2. In was noted that young animals have high sero-prevalence of gb-antibodies because of the maternal antibodies. Here, we investigate the relationship between the age of the animal and the prevalence of gb-antibodies, without taking the clustering into account.

We use the truncated power basis to illustrate the use of penalized splines as mixed models. Some choices have to be made when fitting the penalized spline function. The degree of the polynomial basis and the number and positioning of knots needs to be specified. Following the recommendations of Ruppert (2002) and Kammann and Wand (2003), we chose the knot location as

$$
\kappa_k = \frac{k+1}{K+2}\text{th sample quantile of the unique } x_i,
$$

for $k = 1\ldots, K$. The effects of the number of knots and the degree of the regression spline are shown in Figure 10.3. The dotted line represents the fit when the parameters are treated as fixed effect, the full line is the result of the penalized spline as a mixed model.

In this example, the penalized spline model is very flexible in describing the non-linear relationship between age and infection of the animal. Penalized estimation
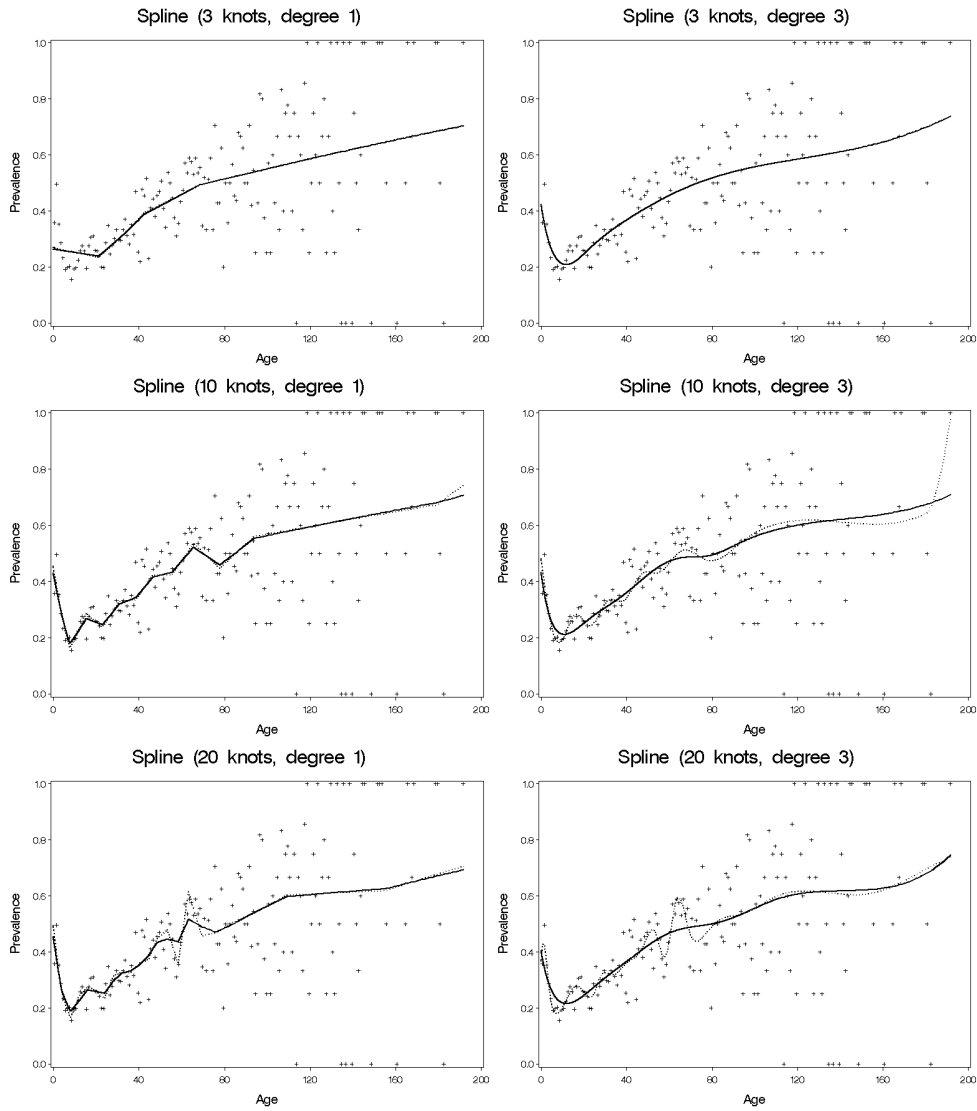
Figure 10.3: *The fitted logistic regression models with penalized splines function, with varying number of knots and degree of basis.*

based on a mixed model shrinks the coefficients of the spline basis towards zero, resulting in a smoother fit compared with the fixed effects estimation. The cubic fit smooths the non-linear relationship much more effectively than the linear spline fit, and there are no unsightly corners. Increasing the number of knots also increases the flexibility in the fit, as seen for the linear spline basis. When using the cubic spline, there are only minor differences between the models with 3, 10 or 20 knots. In summary, the penalized splines offer great flexibility in modelling nonlinear data.

10.4 Two-dimensional smoothing

Penalized spline regression relies on a set of basis functions that permit the handling of a nonlinear structure. Bivariate smoothing requires bivariate basis functions. Penalized spline regression can be extended to the bivariate situation in several ways. Suppose x_{1i} and x_{2i} are two continuous predictors of the response variable y. A general bivariate smoothing model is

$$\eta_i = f(x_{1i}, x_{2i}),$$

where f is a bivariate smooth function.

10.4.1 Tensor Product Bases

One possible methodology for constructing two-dimensional basis function is by taking products of one-dimensional basis functions. Penalized splines as presented in previous section can easily be extended to fit this model:

$$
\begin{aligned}
\eta_i &= \beta_0 + \beta_1 x_{1i} + \sum_{k=1}^{K_1} u_{1k}(x_{1i} - \kappa_{1k})_+ + \beta_2 x_{2i} + \sum_{k=1}^{K_2} u_{2k}(x_{2i} - \kappa_{2k})_+ \\
&\quad + \gamma x_{1i} x_{2i} + \sum_{k=1}^{K_2} \nu_{1k} x_{1i}(x_{2i} - \kappa_{2k})_+ + \sum_{k=1}^{K_1} \nu_{2k} x_{2i}(x_{1i} - \kappa_{1k})_+ \\
&\quad + \sum_{k=1}^{K_1} \sum_{k'=1}^{K_2} \nu_{kk'}(x_{1i} - \kappa_{1k})_+ (x_{2i} - \kappa_{2k'})_+
\end{aligned}
\tag{10.4}
$$

The corresponding basis is often referred to as a tensor product basis. Figure 10.4 shows the basis functions corresponding to (10.4) with knots $\kappa_{11} = \kappa_{21} = 0.3$ and $\kappa_{12} = \kappa_{22} = 0.6$. It is easy to see that this model can be fit in a generalized linear mixed model. A possible drawback of the tensor product splines is their dependence on the orientation of the coordinate axes.
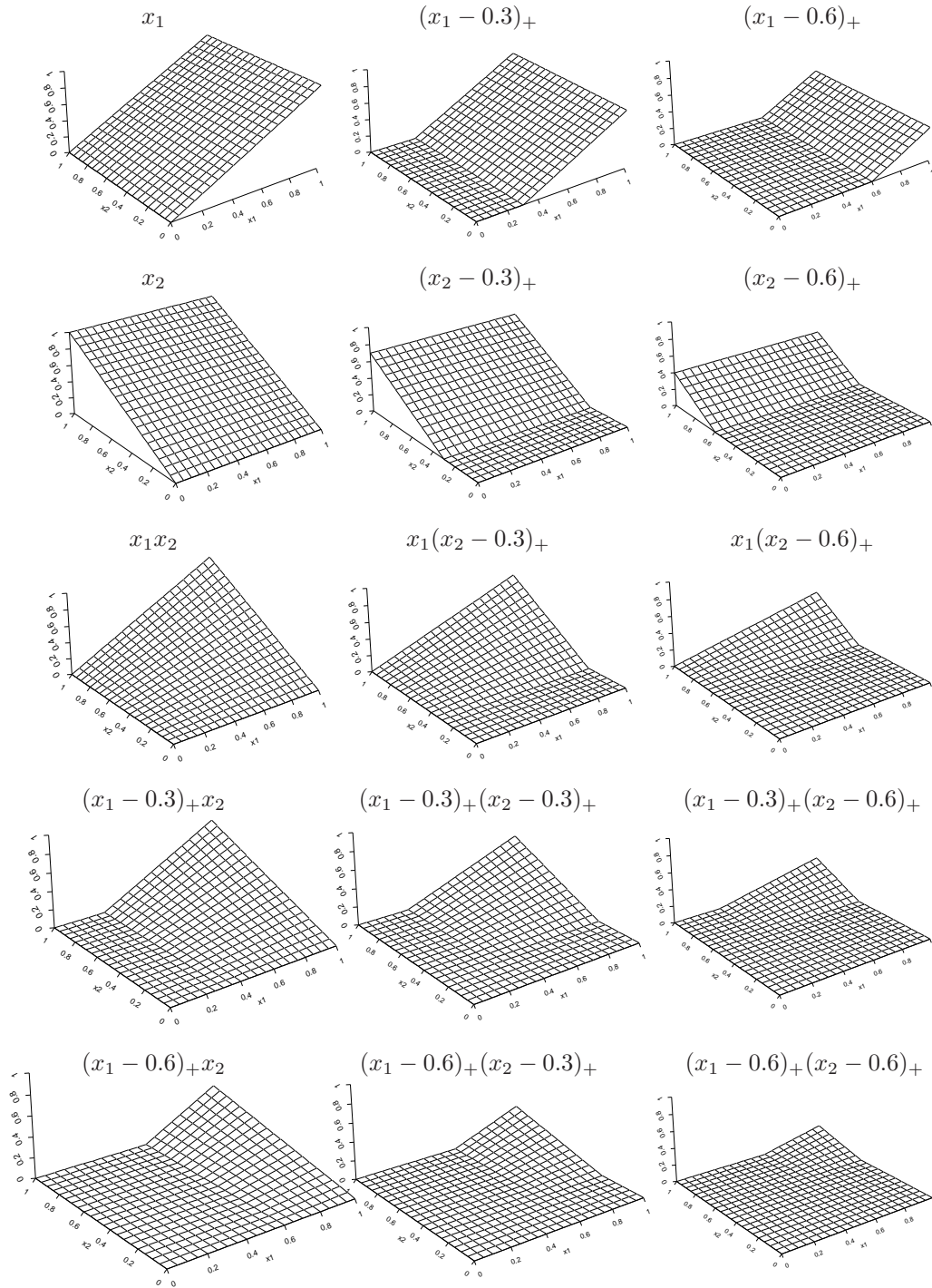
Figure 10.4: *Basis functions for tensor product model, with knots* $\kappa_{11} = \kappa_{21} = 0.3$ *and* $\kappa_{12} = \kappa_{22} = 0.6$.

10.4.2 Radial Smoothers

The extension of radial smoothers to higher dimensions is straightforward. For $\boldsymbol{x}_i \in \mathbb{R}^d$, $1 \leq i \leq N$, and $\boldsymbol{\kappa}_k \in \mathbb{R}^d$, $1 \leq k \leq K$, higher dimensional splines can be obtained by taking $\boldsymbol{X}$ to have columns spanning the space of all d-dimensional polynomials in the components of $\boldsymbol{x}_i$ with degree less than m and

$$\boldsymbol{Z} = \boldsymbol{Z}_K \boldsymbol{\Omega}_K^{-1/2} = \Big[C(\| \, \boldsymbol{x}_i - \boldsymbol{\kappa}_k \, \|) \Big]_{\substack{1 \leq k \leq K \\ 1 \leq i \leq N}} \Big[C(\| \, \boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'} \, \|) \Big]_{1 \leq k, k' \leq K}^{-1/2}$$

where

$$C(r) = \begin{cases} \| \, r \, \|^{2m-d} & \text{for } d \text{ odd} \\ \| \, r \, \|^{2m-d} \log \| \, r \, \| & \text{for } d \text{ even} \end{cases}, \tag{10.5}$$

(Nychka 2000). Note that the function $C(.)$ could be replaced by any other proper covariance function used in kriging (Cressie 1993, O'Connell and Wolfinger 1997, Stein 1999).

10.5 General Design GLMMs

An advantage of smoothing through mixed models is the possibility to handle other complications in the data at the same time. For example, the mixed model representation of smoothers allows for straightforward combinations of smoothing with other modelling tools such as random effects for longitudinal data (Ngo and Wand 2003). Consider the situation where we observe responses y_{ij} for the individuals j ($j = 1, \ldots, n_i$) from group i ($i = 1, \ldots, N$), together with a covariate x_{ij}. A possible model for such data is

$$h\big(E(Y_{ij}|x_{ij})\big) = \beta_0 + U_i + \beta_1 x_{ij},$$

where we model the covariate as a linear trend, and account for the grouping by assuming that $U_i \sim N(0, \sigma_U^2)$. Instead of assuming a parametric model, we can consider a smooth function f describing the trend of x.

$$h\big(E(Y_{ij}|x_{ij})\big) = U_i + f(x_{ij}).$$

For f, a linear spline basis or radial basis functions could be used. If one defines

$$
X = \begin{bmatrix}
1 & x_{11} \\
\vdots & \vdots \\
1 & x_{1n_1} \\
\vdots & \vdots \\
1 & x_{N1} \\
\vdots & \vdots \\
1 & x_{Nn_N}
\end{bmatrix},
$$

$$
Z = \begin{bmatrix}
1 & \ldots & 0 & (x_{11} - \kappa_1)+ & \ldots & (x_{11} - \kappa_K)+ \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & \ldots & 0 & (x_{1n_1} - \kappa_1)+ & \ldots & (x_{1n_1} - \kappa_K)+ \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 1 & (x_{N1} - \kappa_1)+ & \ldots & (x_{N1} - \kappa_K)+ \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 1 & (x_{Nn_N} - \kappa_1)+ & \ldots & (x_{Nn_N} - \kappa_K)+
\end{bmatrix},
$$

$\boldsymbol{\beta} = [\beta_0, \beta_1]^Y$ and $\boldsymbol{u} = [U_1, \ldots, U_N, u_1, \ldots, u_K]^T$ then one can simultaneously estimate variance components for the random intercept and the amount of smoothing for f by using the mixed model

$$
\eta = X\boldsymbol{\beta} + Z\boldsymbol{u}
$$

with

$$
\mathrm{Cov}(\boldsymbol{u}) = \begin{bmatrix}
\sigma_U^2 I & \mathbf{0} \\
\mathbf{0} & \sigma_u^2 I
\end{bmatrix}.
$$

Here, σ_U^2 measures the between-subject variability and σ_u^2 controls the amount of smoothing done to estimate f.

A similar combination is performed by Kammann and Wand (2003). They combine a geostatistical analysis with additive models to account for non-linear effects.

The key to full generality is to allow the random effects design matrix to have a general structure (Zhao *et al.* 2004). Breaking the fixed and random effects structure

into pieces to account for different covariance structures used in longitudinal data modelling, smoothing and spatial statistics, we get

$$X\beta + Zu = X^R\beta^R + Z^R u^R + X^S\beta^S + \sum_{l=1}^{L} Z_l^S u_l^S + Z^C u^C.$$

The longitudinal structure with random intercepts and slopes is characterized by

$$X^R \equiv \begin{bmatrix} X_1^R \\ \dots \\ X_m^R \end{bmatrix}, Z^R \equiv \underset{1 \leq i \leq m}{\text{blockdiag}}(X_i^R)$$

and

$$\text{Cov}(u^R) \equiv \underset{1 \leq i \leq m}{\text{blockdiag}}(\Sigma_i^R) \equiv I_m \otimes \Sigma^R.$$

The matrix Z^S contains the smoothing spline basis functions. As in previous sections, the smoothing spline is plugged in a generalized linear mixed model by assuming

$$\text{Cov}(u^S) = \boldsymbol{D}.$$

The component $Z^C u^C$ represents random effects with spatial correlation structure (Wakefield, Best and Waller 2000). As such, different complexities in the data can be translated into a unified generalized linear mixed model.

10.6   Data Analysis

The data depicted in Figure 2.7 correspond to the number of spikes measured during successive slow waves in a 24×10 array. We have data from 7 different cat, with 12 to 16 successive slow waves per cat. Denote y_{ijl} the number of spikes occurring at location $\boldsymbol{x}_i = (x_{1i}, x_{2i})$ during slow wave j for cat l.

10.6.1   One Cat

First, consider data from one cat only and assume that observations from successive slow waves are independent. The number of spikes on the surface is modelled using a Poisson distribution:

$$\begin{aligned} y_{ij1} &\sim Poisson(\lambda_i) \\ \log(\lambda_i) &= f(\boldsymbol{x}_i) \end{aligned}$$
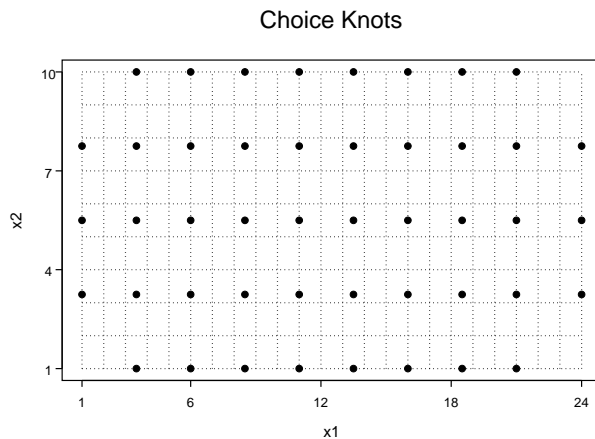
Figure 10.5: *Choice of knots on the surface.*

where $f(\boldsymbol{x}_i)$ is modelled non-parametrically using a cubic radial spline function:

$$\log(\lambda_i) \quad = \quad \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sum_{k_1=1}^{K} u_k C(\|(x_{1i}, x_{2i}) - (\kappa_{1k}, \kappa_{2k})\|) \quad (10.6)$$

with $C(.)$ as in (10.5) with $d = 2$ and $m = 2$. Defining $\boldsymbol{X}_i = [\ 1 \quad \boldsymbol{x}_i^T\ ]$ and $\boldsymbol{Z}$ as in Section 10.4.2, we can write the model as $\log(\lambda) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$. A regular lattice of 46 equidistant knots $\boldsymbol{\kappa}_k = (\kappa_{1k}, \kappa_{2k})$ is chosen (Figure 10.5). Figure 10.6 shows the fit $\widehat{f(\boldsymbol{x}_i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{u}}$ obtained by treating the coefficients of the knots as fixed effects. The top panel is a contour plot, showing the contour levels of the estimated λ. The dots represent the observed values for this cat, with the size proportional to the mean number of spikes at that location. The bottom panel is a three-dimensional graph of the fitted model. This plot seems to overfit the data, rather than smoothing it. By treating the above model as a mixed model with

$$u_k \sim N(0, \sigma_u^2),$$

a smooth fit is obtained. Indeed, fitting a smoothing spline as a generalized linear mixed model has the advantage that the amount of smoothing is selected automatically. In this example, the selected smoothing parameters is $\hat{\lambda} = (1/\sigma_u)^{1/3} = (1/0.003326)^{1/3} = 6.70$. Figure 10.7 shows the fit $\widehat{f(\boldsymbol{x}_i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{u}}$ obtained using PQL estimation. A smooth fit, indicating the location of spikes on the surface, is obtained.

It is of interest to test whether the predictor variable $\boldsymbol{x}_i$ has an effect on the
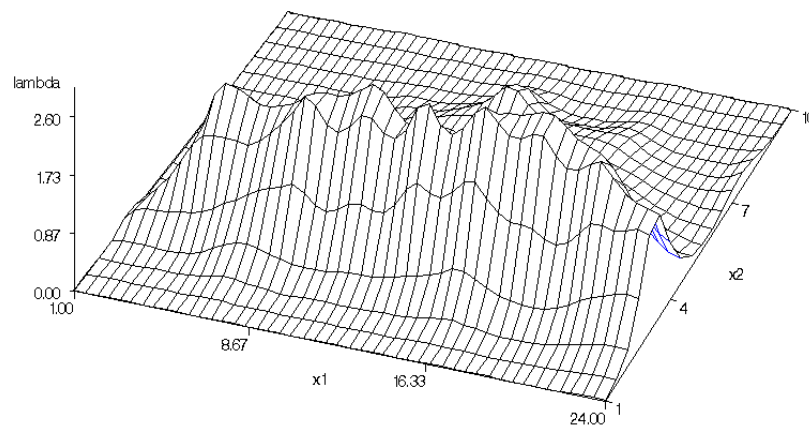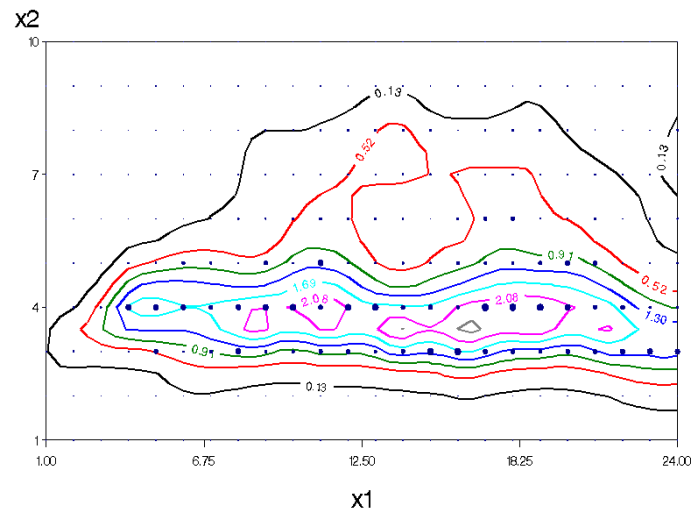
Figure 10.6: *Estimated number of spikes on the preparation of cat 1, based on Model 1 by treating the coefficients of the knots as fixed effects.*
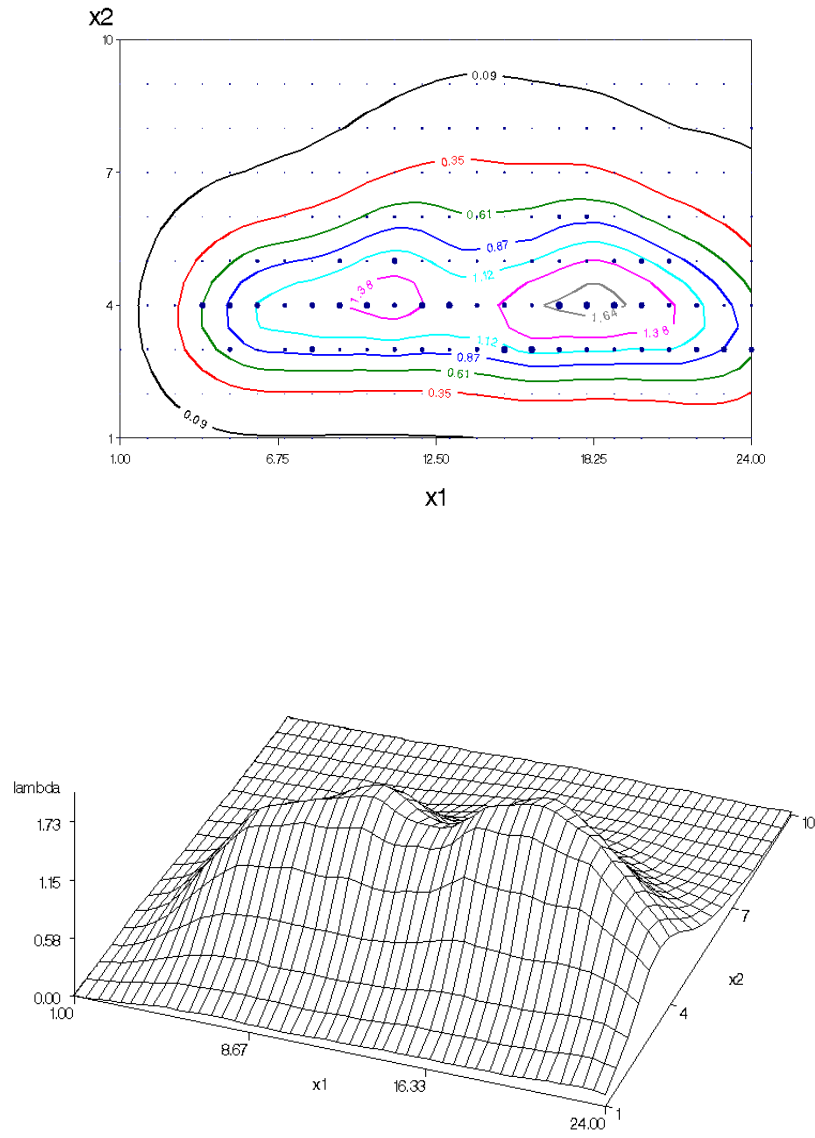
Figure 10.7: *Estimated number of spikes on the preparation of cat 1, based on Model 1, by treating the coefficients of the knots as random effects.*

response. In this case, the null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \sigma_u^2 = 0 \text{ versus } H_1 : \beta_1 \neq 0, \beta_2 \neq 0 \text{ or } \sigma_u^2 > 0.$$

As the null-hypothesis is on the boundary of the parameter space, classical inference no longer holds. Let us consider the likelihood ratio test. When testing the hypothesis $H_0 : \sigma_u^2 = 0$ versus $H_1 : \sigma_u^2 > 0$, the asymptotic null distribution of the likelihood ratio test is a mixture of two chi-squared distributions, $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Self and Liang 1987). More generally, the asymptotic null distribution for the likelihood ratio test statistic for testing the null hypothesis which sets the variance component and s regression coefficients equal to zero, is a mixture of χ_s^2 and χ_{s+1}^2, with equal probability $1/2$. Note however that large-sample theory for likelihood ratio test assumes that the number of subjects approaches infinity (Stram and Lee 1994). In addition, maximum likelihood estimates of the variance components can be biased. Using REML estimates might improve the accuracy of the test. However, restricted likelihood can only be used to compare models with the same mean structure (Verbeke and Molenberghs 2000). For these reasons, likelihood ratio test based on a mixture of chi-squared distributions might perform poor when applied to penalized splines (Crainiceanu *et al.* 2003). As an alternative, simulation methods can be used to estimate the appropriate null distribution. This is topic of current research.

10.6.2 Cat-Specific Profiles

In previous section, we considered data from only one isolated segment of the duodenum. However, data from seven preparations of different cat are available. Both the number of spikes and the locations of the spike patches are very different among cats. Therefore, it seems sensible to allow the profiles to vary with the cats. Instead of the previous model, we specify the model

$$y_{ijl} \quad \sim \quad Poisson(\lambda_{il})$$

with

$$\begin{aligned}
\log(\lambda_{il}) \quad = \quad & (\beta_0 + b_{0l}) + (\beta_1 + b_{1l})x_{1i} + (\beta_2 + b_{2l})x_{2i} \quad (10.7) \\
& + \sum_{k=1}^{K}(u_k + \nu_{kl})C(\|\boldsymbol{x}_i - \boldsymbol{\kappa}_k\|),
\end{aligned}$$

with $u_k \sim N(0, \sigma_u^2)$. Further, $\boldsymbol{b}_l = (b_{0l}, b_{1l}, b_{2l})^T \sim N(0, \Sigma)$ for an unstructured covariance matrix Σ, allowing for complex departures from the overall structure, and

$\boldsymbol{\nu}_l = (\nu_{1l}, \dots, \nu_{kl})^T \sim N(0, \sigma_\nu^2 \boldsymbol{I})$. This is equivalent to modelling

$$\log(\lambda_{il}) \quad = \quad f(\boldsymbol{x}_i) + g_l(\boldsymbol{x}_i)$$

with $f(\boldsymbol{x}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}$ as in (10.6) and $g_l = \boldsymbol{X}_i\boldsymbol{b}_l + \boldsymbol{Z}_i\boldsymbol{\nu}_l$ random functions with mean zero, which are modelled as a penalized spline.

The reparameterization $\nu_{kl}^* = u_k + \nu_{kl}$ is used in order to fit the model. Note that the parameters ν_{kl}^* are distributed normally with mean zero and variance $\sigma_{\nu*}^2 = \sigma_u^2 + \sigma_\nu^2$. The covariance matrix Σ is estimated as

$$\begin{pmatrix} 1.278 & -0.250 & -1.232 \\ -0.250 & 0.060 & 0.233 \\ -1.232 & 0.233 & 1.441 \end{pmatrix},$$

and the variance $\sigma_{\nu*}^2$, which controls the amount of smoothing, is estimated as 0.040. Results are shown in figures 10.8 and 10.9 . Each plot shows the cat-specific profile. The contour plot shows the contour levels in steps of 0.1. It can be seen that the location of spikes on the surface are different among animals.

It can be argued that the use of one smoothing parameter to fit the cat-specific profiles is too restrictive, and that each cat requires its own smoothing parameter. This is possible by assuming $\nu_{kl}^* \sim N(0, \sigma_l^2)$. A disadvantage of this approach is that the number of parameters increases considerably, hampering the estimation.

10.6.3   Successive slow waves

In addition, the number of spikes on each preparation is repeatedly measured, during successive slow waves. Consequently, these repeated measurements might be correlated. We assume

$$y_{itl} \quad \sim \quad Poisson(\lambda_{itl})$$
$$\log(\lambda_i) \quad = \quad u_{tl} + f_l(\boldsymbol{x}_i)$$

with $u_{tl} \sim N(0, \sigma_t^2)$ and $f_l(\boldsymbol{x}_i)$ the cat-specific profiles as defined in (10.7). The between-subject variability σ_t^2 is estimated as 0.086 (s.e.= 0.006). The covariance matrix Σ is estimated as

$$\begin{pmatrix} 1.113 & -0.256 & -1.205 \\ -0.256 & 0.071 & 0.250 \\ -1.205 & 0.250 & 1.505 \end{pmatrix},$$
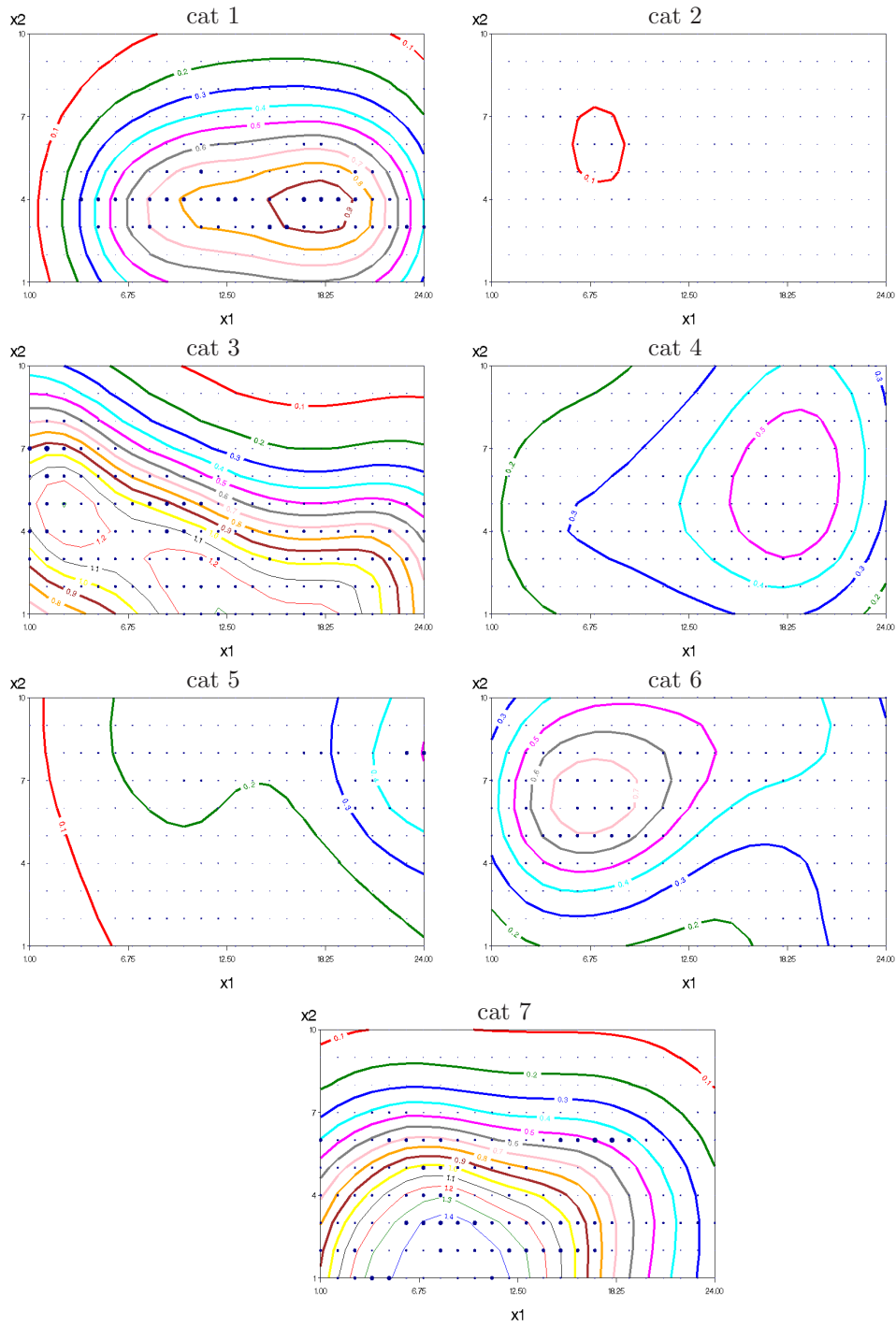
Figure 10.8: *Contour plot of the cat-specific profiles of estimated number of spikes.*
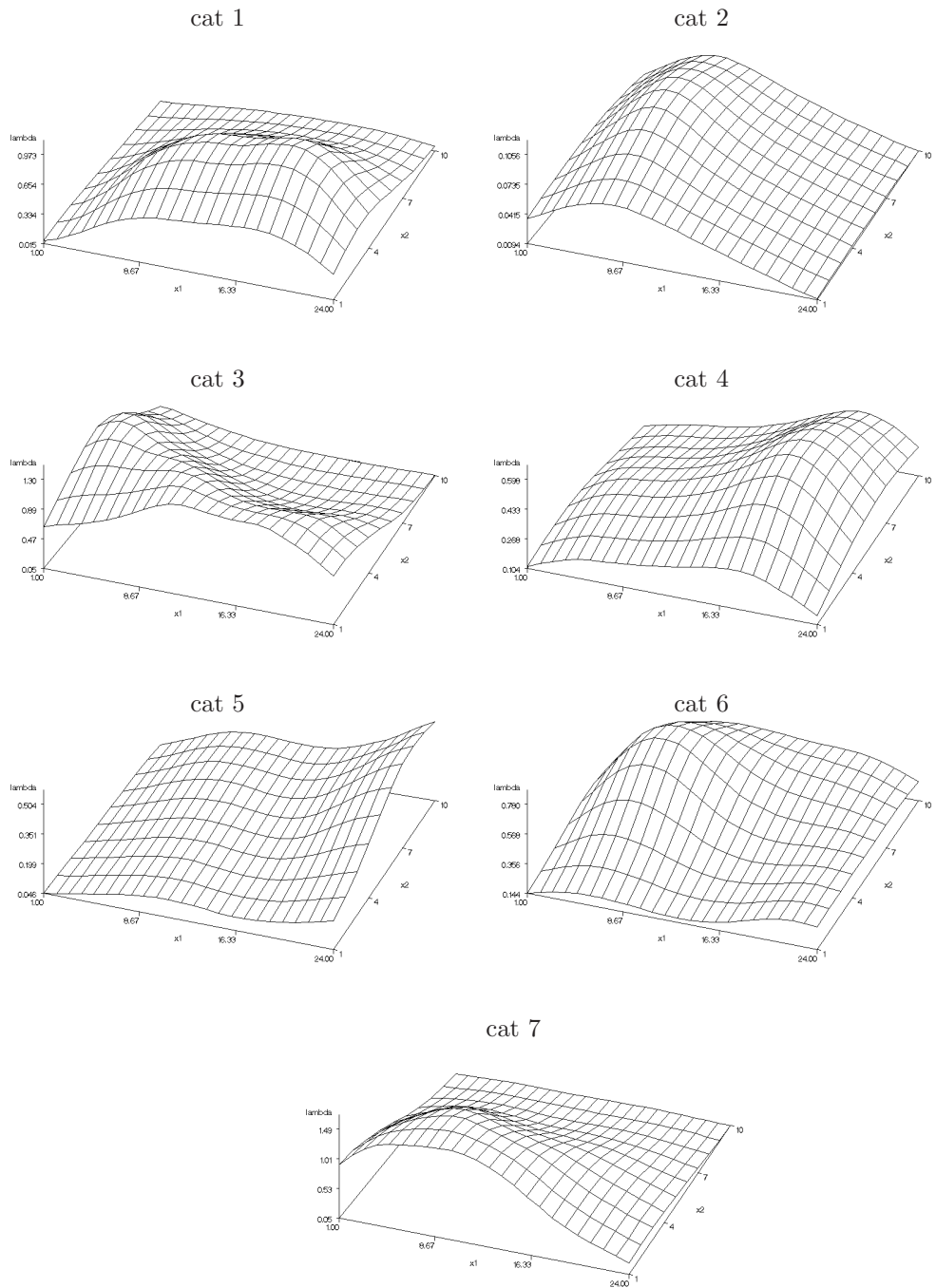
cat 1

cat 2

cat 3

cat 4

cat 5

cat 6

cat 7

Figure 10.9: *Three dimensional plot of the cat-specific profiles of estimated number of spikes.*
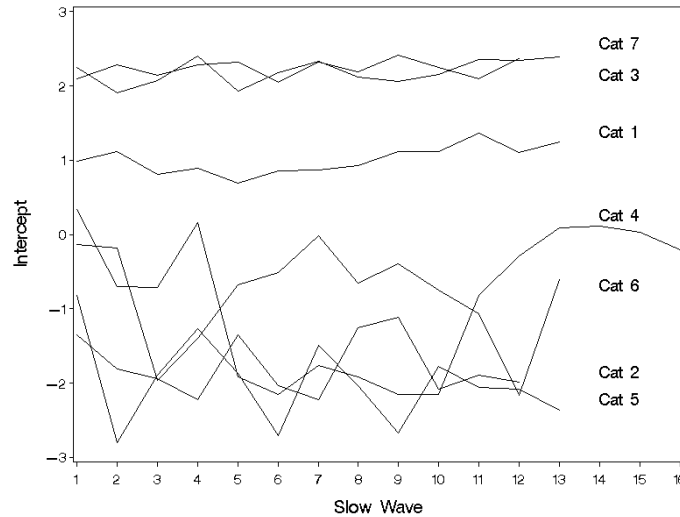
Figure 10.10: *Random intercept of successive slow waves for different cat.*

and the variance $\sigma_{\nu*}^2$, which controls the amount of smoothing, is estimated as 0.037 (s.e.=0.002).

The intercept $\beta_{0l} + u_{tl}$ for each cat during the successive slow waves is shown in Figure 10.10. It is seen that the intercept during the successive slow waves are similar for most preparations. The number of spikes occurring at a fixed point on the surface, but at different time points, are closely related. The effect of a random intercept for each slow wave on the estimated profile, is that the peaks in the profile become stronger or weaker, but the estimated location of spikes stays the same. As an example, the estimated number of spikes in the successive slow waves of cat 4 are shown in Figures 10.11 and 10.12.

It is also possible to assume that the covariance between the measurements of two slow wave is a decreasing function of the time lag between them. Thus, an autoregressive structure might be useful in this context. This is should be investigated further.

10.7 Discussion

In this chapter, the spatial organization of spike patches occurring in seven isolated segments of the duodenum were investigated. Spikes tend to occur in some areas and not in others. This spatial heterogeneity will play a role in determining some aspects
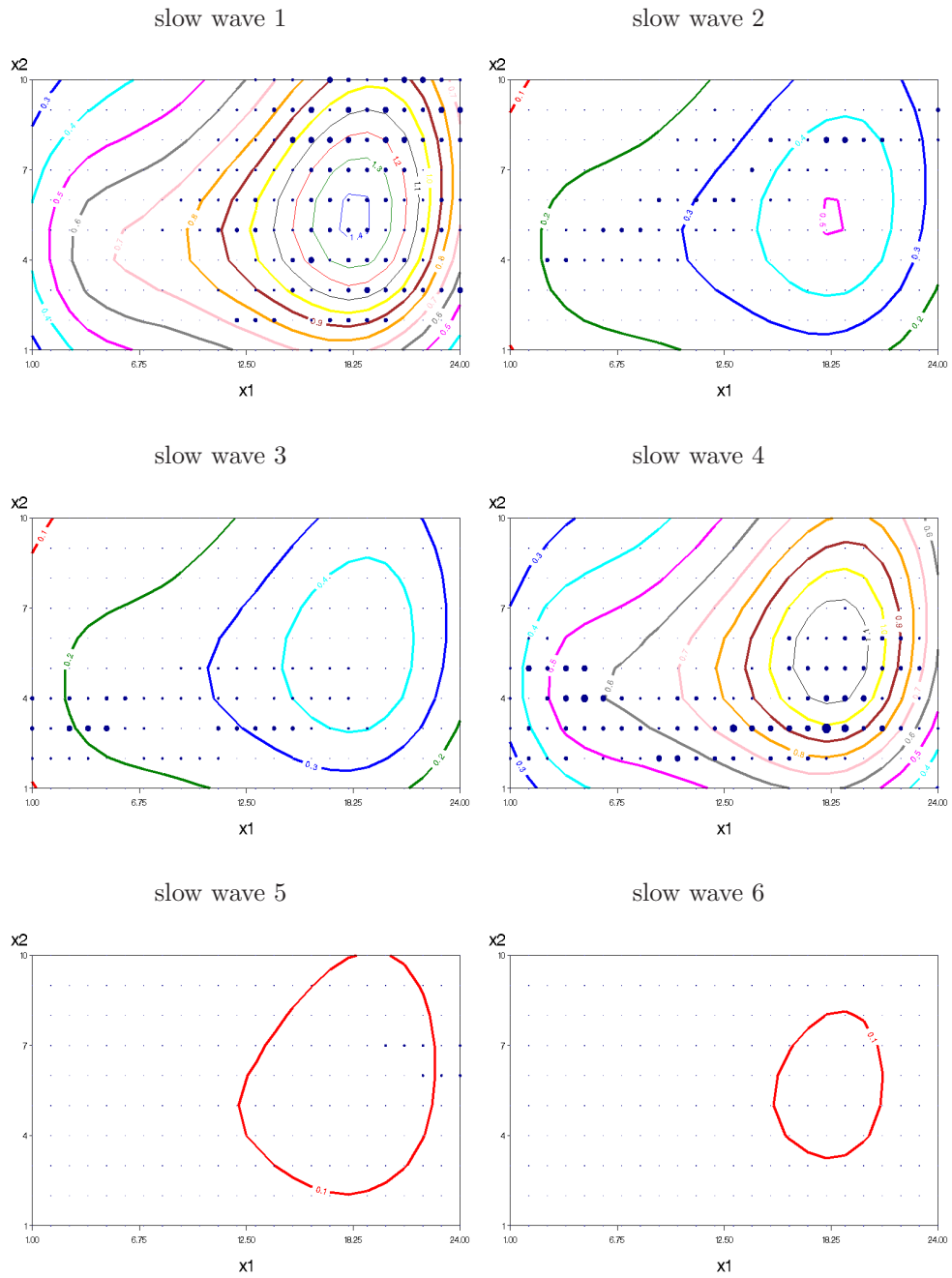
slow wave 1

slow wave 2

slow wave 3

slow wave 4

slow wave 5

slow wave 6

Figure 10.11: *Contour plot of estimated number of spikes during successive slow waves 1 to 6 of cat 4.*
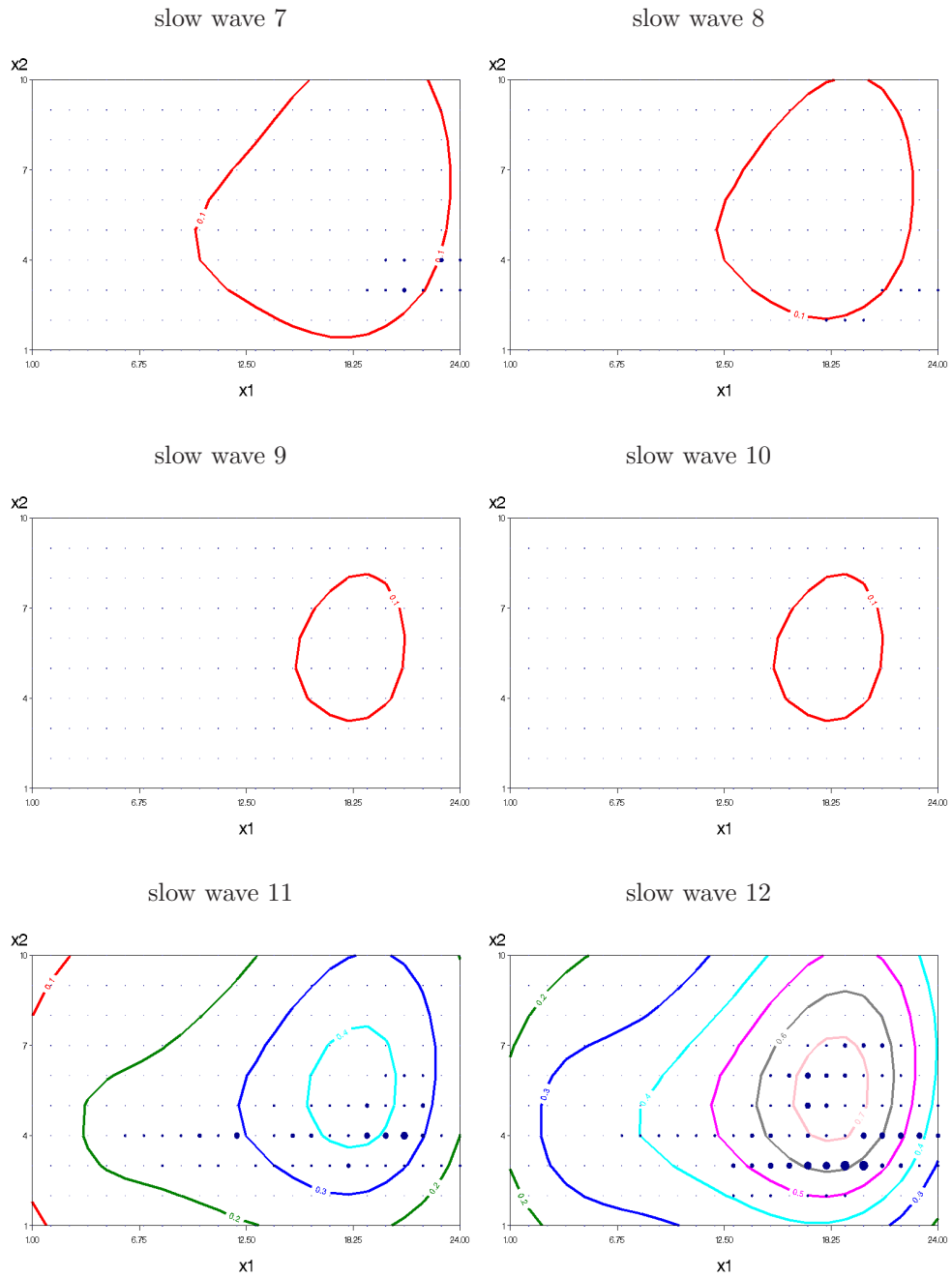
Figure 10.12: *Contour plot of estimated number of spikes during successive slow waves 7 to 12 of cat 4.*

of intestinal motility (Lammers *et al.* 2004).

The mixed model approach for non-parametric regression has important advantages. It can be used both for continuous data, as well as for non-Gaussian data. It can be easily extended for handling higher-dimensional smoothing problems. It can handle grouping or other types of correlations in the same model. In addition, it selects the amount of smoothing in the model automatically.

In this chapter, we used penalized quasi-likelihood methods for fitting the generalized linear mixed models. However, the approximation can be quite inaccurate in certain circumstances. A possible alternative is the PQL2 procedure available in MLwiN software. Another alternative is the use of a Bayesian approach, and the use of Markov Chain Monte Carlo for estimation and inference. The WinBUGS package is quite flexible to implement a GLMM. Further research is needed in this context.

Acknowledgment

References

Abeles, M. (1982). *Local cortical circuits. An electrophysiological study.* Springer.

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modeling of Clustered Data*, Chapmann and Hall.

Aerts, M., Declerck, L., and Molenberghs, G. (1997). Likelihood Misspecification and Safe Dose Determination for Clustered Binary Data. *Environmetrics*, **8**, 613–627.

Aertsen, A.M., Gerstein, G.L., Habib, M.K., and Palm, G. (1989). Dynamics of neuronal firing correlation: modulation of effective connectivity, *J Neurophysiol*, **61**, 900–917.

Aitkin, M. (1991). Posterior Bayes Factors. *Journal of the Royal Statistical Society*, Ser. B, **53**, 111–142.

Akaike, H. (1974). A new look at statistical model identification. *I.E.E.E. Transactions on Automatic Control*, **19**, 816–723.

Altham, P.M.E. (1978). Two Generalizations of the Binomial Distribution. *Applied Statistics*, **27**, 162–167.

Anderson, D.A., and Aitkin, M. (1985). Variance component models with binary responses: interviewer variability. *Journal of the Royal Statistical Society, Series B.*, **47**, 203–210.

Anderson, R.M., and May, R.M. (1991). *Infectious diseases of humans, dynamic and control.* Oxford University Press Inc. New York.

Arnold, B.C., and Strauss, D. (1991). Pseudolikelihood estimation : Some examples. *Sankhya B* **53**, 233–243.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New-York: Springer-Verlag.

Besag, J., Green, P.J., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**, 3–66.

Boelaert, F., Biront, P., Soumare, B., Dispas, M., Vanopdenbosch, E., Vermeersch, J.P., Raskin, A., Dufey, J., Berkvens, D., and Kerkhofs, P. (2000). Prevalence of bovine herpesvirus-1 in the Belgian cattle population. *Prev. Vet. Med.*, **45**, 285–295.

Bosch, R.J., Wypij, D., and Ryan, L.M. (1996). A semiparametric approach to risk assessment for quantitative outcomes. *Risk Analysis* **16**, 657–665.

Box, G.E.P., and Tiao, G.C. (1992). *Bayesian Inference in Statistical analysis*, John Wiley and Sons: New York.

Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Budtz-Jørgensen, E., Keiding, N., and Grandjean, P. (2001). Benchmark Dose Calculation from Epidemiological Data. *Biometrics*, **57**, 698–706.

Cadarso-Suarez, C., Roca-Pardinas, J., Molenberghs, G., Faes, C., Nacher, V., Ojeda, S., and Acuna, C. (2004). Flexible modeling of neuron firing rates across different experimental conditions. An application to neural activity in the prefrontal cortex during a discrimination task. Submitted.

Carlin, B.P. and Louis, T.S. (1996). *Bayes and empirical bayes methods for data analysis. Journal of the Royal Statistical Society*, London: Chapman and Hall.

Catalano, P.J. and Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* **87**, 651–658.

Catalano, P., Ryan, L., and Scharfstein, D. (1994). Modeling fetal death and malformation in developmental toxicity. *Journal of the American Statistical Association*, **87**, 651–658.

Catalano, P.J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine* **16**, 883–900.

Chen, J.J., and Kodell, R.L. (1989). Quantitative Risk Assessment for Teratologic Effects. *Journal of the American Statistical Association*, **84**, 966–971.

Connolly, M.A. and Liang, K.Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501–506.

Corcoran, C.D., Mehta, C.R., and Senchaudhuri, P. (2000). Power Comparisons for Tests of Trend in Dose-response Studies. *Statistics in Medicine,* **19**, 3037–3050.

Corcoran, C.D., Senchaudhuri, P., Ryan, L., Mehta, C., Patel, N., and Molenberghs, G. (2001). An Exact Trend Test for Correlated Binary Data. *Biometrics*, **57**, 941–948.

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

Cox, D.R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–461.

Cox, D.R., and Wermuth, N. (1994). A Note on the Quadratic Exponential Binary Distribution. *Biometrika*, **81**, 403–408.

Crainiceanu, C. M., Ruppert, D., Claeskens, G. and Wand, M. P. (2003) Exact likelihood ratio tests for penalized splines. To be published.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York, NY.

Crump, K.S., and Howe, R.B. (1983). A review of methods for calculating statistical confidence limits in low dose extrapolation. In Clayson, D.B., Krewki, D. and Mundro, I. eds. *Toxicological Risk Assessment. Volume I: Biological and Statistical Criteria*. Boca Raton: CRC Press, 187–203.

Crump, K. (1984). A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology* **4**, 854–871.

Crump, K. (1985). Calculation of benchmark doses from continuous data. *Risk Analysis* **15**, 79–89.

Dale, J. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.

Davidian, M., and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data.* London: Chapman and Hall.

Declerck, L., Molenberghs, G., Aerts, M., and Ryan, L. (2000). Litter-based methods in developmental toxicity risk assessment. *Environmental and Ecological Statistics*, **7**, 57–76.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal data.* Oxford: Oxford University Press.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la exion des plaques minces, *R.A.I.R.O. Analyse numerique*, **10**, 5–12.

Duchon, J. (1977). Splines Minimizing Rotation-Invariant Semi-Norms in Sovolev Spaces. In *Constructive Theory of Functions of Several Variables*, eds. W. Schempp and K. Zeller, 85 –100.

Dunson, D.B. (1998). Dose-dependent number of implants and implications in developmental toxicity. *Biometrics*, 54, 558–569.

Dunson, D.B. (2000). Bayesian laten variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355–366.

Ekholm, A., Smith, P. W. F., and McDonald, J. W. (1995). Marginal regression analysis of a multivariate binary response, *Biometrika*, **82**, 847–854.

Faes, C., Geys, H., Aerts, M., and Molenberghs, G. (2003a). Use of Fractional Polynomials for Dose-Response Modelling and Quantitative Risk Assessment in Developmental Toxicity Studies. *Statistical Modelling*, **3**, 109–125.

Faes, C., Geys, H., Aerts, M. and Molenberghs, G. (2003b). Hierarchical Modelling Approach for Risk Assessment in Developmental Toxicity Studies, in: *Proceedings of the 18th International Workshop on Statistical Modelling, Leuven, Belgium*, Verbeke G. (eds), 137–142.

Faes, C., Geys, H., Aerts, M., Molenberghs, G., Catalano, P. (2004a), 'Modelling combined continuous and ordinal outcomes in a clustered setting. *Journal of Agricultural, Biological and Environmental Statistics*, **00**, 000–000.

Faes, C., Aerts, M., Geys, H., Molenberghs, G., and Declerck, L. (2004b). Bayesian Testing for Trend in a Power Model for Clustered Binary Data. *Environmental and Ecological Statistics*, **00**, 000–000.

Faes, C., Geys, H., Aerts, M. and Molenberghs, G. (2004c). A Hierarchical Modelling Approach for Risk Assessment in Developmental Toxicity Studies. Submitted to Biostatistics.

Faes, C., Molenberghs, G., Geys, H., Aerts, M., Cadarso-Suarez, C. , and Acuna, C. (2004d). A Flexible Method to Measure Synchrony in Neuronal Firing. Submitted to JASA, Case Studies.

Faes, C., Aerts, M., and Claeskens, G. (2002) "Flexible Polynomial Models", in Aerts, M. (et al.), *Topics in Modelling of Clustered Data*, Chapman and Hall, pp. 115–125.

Fitzmaurice, G.M. and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**, 845–852.

Food and Drug Administration (1966). Guidelines for developmental toxicity and risk assessment. *Fed. Regist.* **56**, 63798.

Friedman, J. H., and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics*, **31**, 3–39.

Gaylor, D.W. and Slikker, W. (1990). Risk assessment for neurotoxic effects. *NeuroToxicology* **11**, 211–218.

Gerstein, G.L., and Aertsen, A.M. (1985). Representation of cooperative firing activity among simultaneously recorded neurons. *J Neurophysiology*, **54**, 1513–1528.

Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association* **94**, 734–745.

Geys, H., Regan, M., Catalano, P., and Molenberghs, G. (2001). Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *Journal of Agricultural Biological and Environment Statistics* **6**, 340–355.

Geys, H., Molenberghs, G. and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood a nd generalized estimating equations for marginally specified odds ratio models. *Journal of Statistical Computation and Simulation*, **62**, 45–72

Geys, H., Molenberghs, G. and Ryan, L. (1997). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics: Theory and Methods*, **26**, 2743–2767.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice.*, London: Chapman and Hall.

Gueorguieva, R.V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *JASA* **96**, 1102–1112.

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55, 245–259.

Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Grün, S. (1996). Unitary joint-events in multiple-neuron spiking activity: detection, significance, and interpretation. Reihe Physik, Band 60. Thun, Frankfurt/Main: Verlag Harri Deutsch.

Grün, S., Diesmann, M., Grammont, F., Riehle, A., and Aertsen, A. (1999). Detecting unitary events without discretization of time. *Journal of Neuroscience Methods*, **94**, 67–79.

Grün, S., Diesmann, M., and Aertsen, A. (2002a). Unitary Events in Multiple Single-Neuron Spiking Activity: I. Detection and significance. *Neural Computation*, **14**, 43–80.

Grün, S., Diesmann, M., and Aertsen, A. (2002b). Unitary events in multiple single-neuron spiking activity: II. Nonstationary data. *Neural Computation*, **14**, 81–119.

Gütig, R., Aertsen, A., and Rotter, S. (2002). Statistical significance of coincident spikes: count-based versus rate-based statistics. *Neural Computation*, **14**, 121–153.

Haughton, D. and Dudley (1992). Information Criteria for Multiple Data Sets and Restricted Parameters. Technical Report, Bentley College, Department of Mathematics.

Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001). Within-cluster resampling. *Biometrika*, **88**, 1121–1134.

Holland, P.W. and Wang, Y.J. (1987). Dependence function for continuous bivariate densities. *Communications in Statistics - Theory and Methods* **2**, 863–876.

Jeffreys, H. (1961). *Theory of Probability.*, London: Oxford University Press.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, 3rd Edn. New Jersey: Prentice Hall.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Applied Statistics*, **52**, 1-18.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R.E., and Vaidyanathan (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions. *Journal of the Royal Statistical Society*, **54**, 129–144.

Kass, R.E., Ventura, V., and Cai, C. (2003). Statistical smoothing of neuronal data. *Network*, **14**, 5–15.

Kass, R.E., and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypothesis and Its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Kavlock, R.J., Allen, B.C., Faustman, E.M., and Kimmel, C.A. (1995). Dose-response assessment for developmental toxicity. IV. Benchmark dose for fetal weight changes. *Fundamental and Applied Toxicology* **26**, 211–222.

Kimmel, C. (1990). Quantitative approaches to human risk assessment for non-cancer health effects. *Neurotoxicology* **11**, 201.

Kimmel, G.L., and Gaylor, D.W. (1988). Issues in Qualitative and Quantitative Risk Analysis for Developmental Toxicology. *Risk Analysis*, **8**, 15–20.

Kleinman, J.C. (1973). Properties with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, **68**, 46–54.

Kodell, R.L., and West, R.W. (1993). Upper confidence limits on excess risk for quantitative responses. *Risk Analysis* **13**, 177–182.

Kupper, L.L., Portier, C., Hogan, M.D., and Yamamoto, E. (1986). The Impact of Litter Effects on Dose-Response Modeling in Teratology. *Biometrics*, **34**, 69–76.

Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lammers, W. J. E. P., Stephen, B., Arafat, K., and Manefield, G.W. (1996). High resolution mapping in the gastrointestinal system: initial results. *Neurogastroenterology and Motility*, **8**, 207–216.

Lammers, W. J. E. P., Faes, C., Stephen, B., Bijnens, L., Ver Donck, L., and Schuurkes, J.A.J. (2004). Spatial Determination of Successive Spikes in the Isolated Cat Duodenum. *Neurogastroenterology and Motility*, **00**, 000–000.

Lapp, K., Molenberghs, G., and Lesaffre, E. (1998). Local and global cross ratios to model the association between ordinal variables. *Computational Statistics and Data Analysis* **28**, 387–412.

Le Cessie, S., and Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics* **43**, 95–108.

Leisenring, W., and Ryan, L. (1992). Statistical Properties of the NOAEL. *Regulatory Toxicology and Pharmacology*, **15**, 161.

Lefkopoulou, M., Rotnitzky, A. and Ryan, L. (1996). Trend Tests for Clustered Data. In: *Statistics in Toxicology*. B.J.T. Morgan (Ed). New York: Oxford University Press Inc., 179–197.

Liang, K.Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Malsburg, vdC. (1999). The what and why of binding: the modeler's perspective. *Neuron*, **24**, 95–25.

Mancuso, J. Y., Ahn, H. and Chen, J. J. (2001). Order-Restricted Dose-Related Trend Tests. Statistics in Medicine, 20, 2305–2318.

Martignon, L., Deco, G., Laskey, K., Diamond, M., Freiwald, W., and Vaadia, E. (2000). Neural coding: higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation*, **12**, 2621–2653.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C.E., and Searle, S.R. (1989). *Generalized, linear and mixed models*. Willey: New York.

Meinguet, J. (1979). Multivariate Interpolation at Arbitrary Points Made Simple. *J. Appl. Math. Phys. (ZAMP)*, **5**, 439–468.

Molenberghs, G., Declerck, L., and Aerts, M. (1998). Misspecifying the Likelihood for Clustered Binary Data. *Computational Statistics and Data Analysis*, **26**, 327–350.

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluations of surrogate enpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023–3038.

Molenberghs, G., and Lesaffre, E. (1999). Marginal Modeling of Multivariate Categorical Data. *Statistics in Medicine*, **18**, 2237–2255.

Molenberghs, G., and Ryan, L.M. (1999). An Exponential Family Model for Clustered Multivariate Binary Data. *Environmetrics*, **10**, 279–300.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. Chapman and Hall, London.

Mountcastle, V.B., Steinmetz, M.A., and Romo, R. (1990). Frequency discrimination in the sense of flutter: psychophysical measurements correlated with postcentral events in behaving monkeys. *J Neuroscience*, **10**, 3032–3044.

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Newsome, W.T., Britten, K.H., and Movshon, J.A. (1989). Neuronal correlates of a perceptual decision. *Nature*, **341**, 52–54.

Ngo, L., and Wand, M.P. (2004) Smoothing with mixed model software. *Journal of Statistical Software*, **9**.

Nychka, D.W. (2000). Spatial process estimates as smoothers. In *Smoothing and Regression* (M. Schimek, ed.). Heidelberg: Springer-Verlag.

O'Connell, M.A., and Wolfinger, R.D. (1997). Spatial regression models, response surfaces, and process optimization. *Journal of Computational and Graphical Statistics*, **6**, 224–241.

Olkin, I., and Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448–465 (with correction in **36**, 343–344).

Pauler, D. (1998). The Schwarz Criterion and Related Methods for Normal Linear Models. *Biometrika*, **85**, 13–27.

Pendergast, J., Gange, S.J., Newton, M.A., Lindstrom, M., Palta, M., and Fisher, M. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review*, **64**, 89–118.

Plackett, R.L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association* **60**, 516–522.

Price, C.J., Kimmel, C.A., Tyl, R.W., and Marr, M.C. (1985). The Developmental Toxicity of Ethylene Glycol in Rats and Mice. *Toxicology and Applied Pharmacology*, **81**, 113–127.

Price, C.J., Kimmel, C.A., George, J.D., and Marr, M.C. (1987). The Developmental Toxicity of Diethylene Glycol in Mice. *Fundamental and Applied Toxicology*, **8**, 115–126.

Price, C.J., Kimmel, C.A., George, J.D. and Marr, M.C. (1987). The developmental toxicity of diethylene glycol dimethyl ether in mice. *Fundamental and Applied Toxicology* **8**, 1033–1048.

Rai, K., and Van Ryzin, J. (1985). A dose-response model for teratological experiments involving quantal responses. *Biometrics*, **47**, 825–839.

Regan, M.M., and Catalano, P.J. (1999a). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics* **55**, 760–768.

Regan, M.M., and Catalano, P.J. (1999b). Bivariate dose-response modeling and risk estimation in developmental toxicology. *Journal of Agricultural, Biological and Environmental Statistics* **4**, 217–237.

Riehle, A., Grün, S., Diesmann, M., and Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* 7, **278**, 1950–1953.

Roberts, W.C., and Abernathy, C.O. (1996). Risk assessment: principles and methodologies. In Fan, A., and Chang, L.W. eds.. *Toxicology and Risk Assessment, Principles, Methods and Applications*, New York: Marcel Dekker Inc., 245–270.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rodricks, J., Tardiff, R., Brett, S., Putzrath, R., and St. Hilaire, C. (1986). Elements of Toxicology and Chemical Risk Assessment. *Environ Corporation*, Washington, D.C.

Romo, R., and Salinas, E. (1999). Sensing and deciding in the somatosensory system. *Curr Opin Neurobiol*, **9**, 487–493.

Rotnitzky, A., and Wypij, D. (1994). A Note on the Bias of Estimators with Missing Data. *Biometrics*, **50**, 1163–1170.

Royston, P., and Altman, D.G. (1994). Regression using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Applied Statistics*, **43**, 429–467.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735-757.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* New York: Cambridge University Press.

Ryan, L.M., Catalano, P.J., Kimmel, C.A., and Kimmel, G.L. (1991). Relationship between fetal weight and malformation in developmental toxicity studies, *Teratology* **44**, 215–223.

Sammel, M.D. Ryan, L.M., and Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. *JRSS-B* **59**, 667–678.

Sauerbrei, W., and Royston, P. (1999). Building Multivariate Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials. *Journal of the Royal Statistical Society, Series A*, **162**, 71–94.

Schwetz, B., and Harris, M. (1993). Developmental toxicology: Status of the field and contribution of the National Toxicology Program. *Environmental Health Perspective* **100**, 269.

Scientific Committee of the Food Safety Council (1980). Proposed system for food safety assessment. *Food and Cosmetic Toxicology*, **16**, Supplement 2, 1–136 (1978). Revised report published June 1980 by the Food Safety Council, Washington, DC.

Self, S. G., and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Shadlen, M.N., and Movshon, J.A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, **24**, 67–75.

Shadlen, M.N., and Newsome, W.T. (1994). Noise, neural codes and cortical organization. *Curr Opin Neurobiology*, **4**, 569–579.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P., and Damme, P.V. (2003). Modelling forces of infection by using monotone local polynomials. *Journal of the Royal Statistical Society: Series C*, **52**, 469–485.

Simonoff, J.S. (1996). *Smoothing methods in statistics*. New York: Springer.

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, **24**, 49–25.

Singer W., and Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 18: 555–586.

Sinharay, S., and Stern, H.S. (2002). On the Sensitivity of Bayes Factors to the Prior Distributions. *American Statistical Association*, **56**, 196–201.

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. *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

Speed, T. (1991). Comment on paper by Robinson. *Statistical Science*, **6**, 42-44.

Speybroeck, N., Boelaert, F., Renard, D., Burzykowski, T., Mintiens, K, Molenberghs, G., and Berkvens, D.L. (2003). Design-based analysis of surveys: a Bovine Herpesvirus 1 case study. *Epidemiology and Infection*, **13**, 991–1002.

Spiegelhalter, D.J., Best, N.G., and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research Report 98–009, Division of Biostatistics, University of Minisota*.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linder, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 1–34.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer.

Stram, D. O., and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.

Tyl, R.W., Price, C.J., Marr, M.C., and Kimmel, C.A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phtalate in Fisher 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, **10**, 395–412.

U.S. Environmental Protection Agency (1991). Guidelines for Developmental Toxicity Risk Assessment. *Federal Register*, **56**, 63798–63826.

Usrey, W.M., and Reid, R.C. (1999). Synchronous activity in the visual system. *Annu Rev Physiol*, **61**, 435–456.

Vazquez, P., Cano, M., and Acuña, C. (2000). Discrimination of line orientation in humans and monkeys. *Journal of Neurophysiology*, **83**, 2639–2648.

Verbeke, G., and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* Springer Series in Statistics. New York: Springer-Verlag.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, **40**, 364–372.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

Wakefield, J., Best, N., and Waller, L.A. (2000). Bayesian approaches to disease mapping. In: *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best, and D.J. Briggs, eds. Oxford University Press, Oxford., 106–127

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223-249.

West, R.W., and Kodell, R.L. (1993). Statistical methods of risk assessment for continuous variables. *Communications in Statistics - Theory and Methods* **22**, 3363–3376.

Williams, D.A. (1975). The analysis of binary responses from toxicology experiments involving reproduction and teratogenicity. *Biometrics*, **38**, 150.

Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144–148.

Williams, P.L., Molenberghs, G., and Lipsitz, S.R. (1996). Analysis of multiple ordinal outcomes in developmental toxicity studies. *Journal of Agricultural, Biological, and Developmental Toxicity Studies* **1**, 250–274.

Williams, P.L., and Ryan, L.M. (1996). Dose-Response Models for Developmental Toxicology. In Hood, R.D. eds. *Handbook of Developmental Toxicology*, New York: CRC Press, 635–666.

Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal Analyses of Clustered Data When Cluster Size is Informative. *Biometrics*, **59**, 36–42.

Wolfinger, R.D. (1993), Laplace's Approximation for Nonlinear Mixed Models. *Biometrika*, **80**, 791–795.

Wolfinger, R., and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computing and Simulation*, **48**, 233–243.

Zeger, S.L., and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–102.

Zhao, Y., Staudenmayer, J., Coull, B.A., and Wand, M. P. (2004). General Design Bayesian Generalized Linear Mixed Models. *Submitted for publication.*

Samenvatting

In deze thesis staat het gebruik van flexibele methoden voor het analyseren van gecorreleerde multivariate data afkomstig uit dierenexperimenten centraal. Correlaties in de data kunnen ontstaan doordat meerdere metingen worden opgenomen bij een zelfde dier, doordat dieren gegroepeerd zijn of metingen ruimtelijk aan elkaar verwant zijn. Bovendien zijn de metingen afkomstig uit deze studies vaak discreet van aard. De statistische analyse wordt gekenmerkt door de keuze van een kansmodel, een predictor model en een schattingsmethode. Het kansmodel reflecteert de structuur van de data (type metingen, correlaties in de data, ...), terwijl het predictormodel de relatie beschrijft tussen de respons en mogelijks verklarende variabelen. Bovendien trachten we steeds gebruik te maken van computationeel aantrekkelijke schattingsmethodes, waarbij de likelihood functie niet noodzakelijk moet gespecificeerd worden. Het grootste deel van deze thesis is gewijd aan de ontwikkeling van statistische procedures in het gebied van risicoanalyse bij ontwikkelingstoxicologische studies. Daarnaast worden andere toepassingsgebieden bekeken zoals electrofysiologie, veterinaire epidemiologie and gastro-enterologie. De gebruikte data worden beschreven in Hoofdstuk 2.

In de publieke opinie is er een grote bezorgdheid omtrent de blootstelling aan tal van toxische stoffen. Denken we maar aan de alarmerende geruchten omtrent de mogelijks schadelijke effecten van dioxine en PCBs in de voedingsketen of in uitstoten van verbrandingsovens. Andere voor de publieke opinie minder gekende haarden van toxiciteit kan men terugvinden in geneesmiddelen, die naast hun therapeutische werking ook bijwerkingen kunnen vertonen, additieven gebruikt in de voedingssector, enz. Hoewel een nulblootstelling van al de mogelijk toxische stoffen uiteraard ideaal zou zijn, is dit in de huidige moderne maatschappij niet haalbaar. Vandaar wordt er getracht de toxische stoffen te identificeren en veilige blootstellingslimieten voor deze stoffen te bepalen.

Ontwikkelingstoxicologische studies zijn ontwikkeld om het potentiële gevaar van een blootstelling op de ontwikkeling van een foetus te onderzoeken, en spelen bij-

221

gevolg een belangrijke rol op het vlak van de volksgezondheid. In deze context is
het echter ethisch onverantwoord mensen in studies bloot te stellen aan mogelijks
schadelijke stoffen. Bovendien zijn betrouwbare epidemiologische gegevens nauwelijks
of niet beschikbaar. Bijgevolg wordt er vaak geopteerd voor toxicologische, en in
het bijzonder teratologische experimenten op knaagdieren. Hierbij worden zwangere
moederdieren blootgesteld aan een mogelijks toxische verbinding gedurende de meest
kritische periode van de dracht. Net voor het baren worden de moederdieren gedissecteerd.
Vervolgens wordt de baarmoeder grondig onderzocht. Men telt het aantal
embryo's die nooit tot ontwikkeling zijn gekomen en terug in de baarmoederwand zijn
geabsorbeerd. Men telt ook het aantal dode foetussen en het aantal levensvatbare
foetussen. De levensvatbare foetussen worden verder onderzocht op de aanwezigheid
van een malformatie en laag geboortegewicht. Aan de hand van dergelijke experimenten
wenst men te onderzoeken of de toegediende stof toxisch is en wenst men een
veilig niveau van blootstelling aan de toxische stof te schatten. Dit behoort tot het
domein van de kwantitatieve risicoanalyse.

Risicoanalyse kan men baseren op de berekening van de NOAEL, of "No Observable
Adverse Effect Level", wat de dosis is juist onder de kleinste dosis die een
statistisch of biologisch significante toename veroorzaakt in vergelijking met de controle.
Gezien de vele nadelen van deze aanpak (Williams and Ryan 1996), wordt
veelal verkozen om de kwantitatieve risicoanalyse te baseren op dosis-respons modellering.
Hierbij stelt men een gepast dosis-respons op, welke dan gebruikt wordt bij
de bepaling van de BMD, of "benchmark dosis". Dit is de dosis waarbij de toename
van het risico ten opzichte van het achtergrond risico gelijk is aan een bepaalde, typisch
kleine, waarde, b.v. 10^{-4}. In tegenstelling tot de NOAEL benadering laat deze
aanpak toe om een maat voor variabiliteit van de geschatte veilige dosis te bepalen, alsook
om de structuur van een toxicologisch experiment in de analyse op te nemen. De
kwantitatieve risicoanalyse gesteund op dosis-respons modellering leidt echter tot een
aantal algemene, relevante onderzoeksvragen. Vooreerst dient men bij de statistische
analyse van gegevens uit toxicologische studies rekening te houden met de genetische
verwantschap van foetussen uit een zelfde nest en de gelijkaardige omstandigheden
voor die foetussen in de baarmoeder. Modellen die het complexe mechanisme
waaruit de data worden gegenereerd, trachten te benaderen, dienen bijgevolg rekening
te houden met dit zogenaamde *nest-effect*. Ook de mogelijke *associaties* tussen
verschillende types van afwijkingen die bij een foetus kunnen optreden zijn belangrijk.
Bij de levensvatbare foetussen worden verschillende ontwikkelingsstoornissen,
zoals malformaties en een te laag geboortegewicht, onderzocht. Vaak wordt hierbij
de graad van malformatie opgemeten (bijvoorbeeld normaal, lichte afwijking, zwaar

misvormd), terwijl het gewicht als een continue variabele wordt gegeven. Bijgevolg moeten ook statistische modellen ontwikkeld worden om *responsen van gemengde aard* (bijvoorbeeld ordinaal en continu) gezamenlijk te modelleren. Verder dient men de *hiërarchische structuur* met dode foetussen enerzijds en levensvatbare maar mogelijks afwijkende foetussen anderzijds mee in rekening te brengen. In dit onderzoeksdomein bekijken we verscheidene, specifieke deelaspecten: het schatten van een veilige blootstellingslimiet, het toetsen van de nulhypothese dat er geen dosiseffect aanwezig is, en het beschrijven van een dosis-respons relatie welke de structuur van de data zo volledig mogelijk in rekening brengt.

Modelleren van data afkomstig van een ontwikkelingstoxicologische studie is complex door de hiërarchische, geclusterde en multivariate structuur van de data. Vooreerst bekijken we het effect van de mogelijks toxische stof op de aanwezigheid van malformaties bij de ontwikkelende foetus. Het opstellen van een geschikt dosis-respons model is niet vanzelfsprekend. Echter, keuze van dit model kan invloed hebben op de kwantitatieve risicoanalyse. Indien we een veilige dosis wensen te bepalen is het zeer belangrijk dat het model goed aanpast aan de data. In de statistische literatuur wordt bij het opstellen van een geschikt dosisrespons model nog steeds veel aandacht besteed aan klassieke veelterm predictoren. Deze zijn echter vaak ontoereikend voor kwantitatieve risicoanalyse, te meer daar de extrapolatie naar lage dosissen gebaseerd op een verkeerd gespecificeerd dosisrespons model kan leiden tot onbetrouwbare virtueel veilige dosissen. Als alternatief kunnen fractionele veeltermen, zoals voorgesteld door Royston and Altman (1993), aangewend worden. Deze veeltermen passen nog steeds binnen de klasse van (veralgemeende) lineaire modellen, maar bieden een veel grotere verscheidenheid aan functionele vormen dan de klassieke veeltermen. In Hoofdstuk 4 tonen we aan de hand van uitgebreide simulatiestudies hoe misleidend het blindelings toepassen van klassieke veeltermen kan zijn, en hoe fractionele veeltermen een meer betrouwbare schatting geven van een veilige dosis, zelfs indien het kansmodel fout gespecificeerd is.

In voorgaande paragraaf werden fractionele veelterm predictoren voorgesteld als alternatief voor de conventionele lineaire predictoren voor de parameters van een dosis-respons model. We kunnen echter ook opteren voor een specifiek type van niet-lineaire predictoren, waar in plaats van een lineaire functie van de toegediende dosis een machtsfunctie $(\alpha + \beta \text{dosis}^\gamma)$ wordt beschouwd. Dit geeft echter aanleiding toch een niet-triviaal statistisch probleem, namelijk het toetsen van de nulhypothese dat er geen dosiseffect is. De nulhypothese dat er geen dosiseffect is, is in dit geval equivalent met het nul stellen van het product van twee regressie-parameters. Onder die nulhypothese zijn de regressie-parameters bijgevolg niet identificeerbaar. Dit resulteert in

problemen indien voor een frequentistische aanpak wordt gekozen. Een bayesiaanse benadering laat toe om aan de hand van Bayes-factoren deze nulhypothese te toetsen. In Hoofdstuk 5 wordt geïllustreerd hoe we de nulhypothese van geen trend in het machtsmodel kunnen testen in de setting van gecorreleerde binaire gegevens.

In de voorgestelde toxicologische studies worden de levensvatbare foetussen onderzocht op de aanwezigheid van malformaties en wordt het geboortegewicht opgemeten. Beide responsen zijn belangrijke indicatoren om de toxiciteit van een stof te bepalen. Vaak zal men voor elk van de 2 soorten responsen een apart dosis-respons model opstellen, waarna een "veilige limiet" wordt bepaald voor de meest gevoelige respons. Vermits de beide responsen echter gecorreleerd zijn is het meer aanvaardbaar om een gezamenlijke verdeling te formuleren voor responsen van gemengde aard (continu en discreet). Terwijl methodes voor het gezamenlijk modelleren van meerdere continue gegevens goed gekend zijn (Johnson en Wichern 1992), is de keuze van een verdeling voor gemengd continu en ordinale variabelen geen vanzelfsprekende keuze. Een mogelijke benadering die in de literatuur vaak gebruikt wordt bestaat erin om de gezamenlijke verdeling te factoriseren in een marginale en een conditionele component, waarbij het conditioneren ofwel op de discrete ofwel op de continue respons kan gebeuren (Catalano 1997). Echter, het bepalen van een veilige blootstellingslimiet is met behulp van deze methode niet evident, want hoewel deze methode wel corrigeert voor de associatie tussen laag geboortegewicht en misvorming kan een schatting voor deze maat niet rechtstreeks bekomen worden. Dus, een uitdrukking voor de gezamenlijke kans dat een foetus aangetast is (i.e., malformatie en/of laag geboortegewicht) is moeilijk te specificeren. In Hoofdstuk 6 beschrijven we een Plackett-Dale methode als alternatieve methode voor het analyseren van gemengd ordinale/continue gegevens (Plackett 1965), waar verondersteld wordt dat de latente malformatie variabelen een Plackett verdeling volgen. De Plackett-Dale benadering maakt gebruik van een odds ratio om de associatie tussen malformatie en gewicht te beschrijven. Deze methode laat toe om de bivariate intra-foetus associatie te schatten zodat de kans op malformatie en/of laag geboortegewicht kan gespecificeerd worden met het oog op kwantitatieve risicoanalyse. Verder worden aparte dosis-respons functies gespecificeerd voor zowel de ordinale als de continue variabele. En tot slot houden we rekening met de correlaties als gevolg van de clustering in het nest. Hierbij maken we gebruik van de pseudo-likelihood methode (Arnold and Strauss 1991).

Daar de verschillende gebeurtenissen kunnen plaatsvinden op verschillende tijdstippen in de ontwikkeling, is er een natuurlijke hiërarchie aanwezig in de verschillende responsen (Williams en Ryan 1996). Door een nadelig effect in een vroeg stadium in de ontwikkeling kan de implantatie opnieuw worden geabsorbeerd in de baarmoed-

erwand (resorptie) of loopt een foetus het risico dat hij niet-levensvatbaar is. Ver-
volgens hebben foetussen die de ganse zwangerschap hebben overleefd mogelijks een
laag geboortegewicht en/of vertonen een malformatie. In voorgaande hoofdstukken
gebeurde het bepalen van het risico van blootstelling steeds aan de hand van de metin-
gen bij de levensvatbare foetussen. Echter, ook het aantal foetussen in een worp kan
beïnvloed worden door de toegediende toxische stof. De worpgrootte is dus eveneens
een stochastische variabele, afhankelijk van de dosis. De klassieke manier om een do-
sisgerelateerde trend in de worpgrootte in rekening te brengen is door de worpgrootte
als covariaat in het model te incorporeren. Het is echter meer aangewezen om dit
dosis effect rechtstreeks te modelleren. In Hoofdstuk 7 beschrijven we een Bayesiaans
model welke de hiërarchische structuur beschrijft, en welke toelaat een veilige dosis
te bepalen door alle informatie in de data in rekening te brengen.

In de resterende hoofdstukken worden verschillende toepassingsgebieden bekeken.
De toepassing en ontwikkeling van statistische methodes voor de analyse van gecor-
releerde discrete data, met het oog op het beantwoorden van specifieke onderzoeksvra-
gen, staat centraal. In Hoofdstuk 8 wordt de functionaliteit van neuronen in de herse-
nen bestudeerd. Hier wordt aangenomen dat neuronen informatie over een stimulus
kunnen verspreiden door het geven van impulsen. Enerzijds gebeurt dit door het
aantal impulsen in een bepaalde neuron te verhogen, en anderzijds door de synchro-
nisatie van de impulsen in verschillende neuronen. Deze veronderstellingen wensen we
te toetsen aan de hand van een statistische methode welke de synchroniciteit tussen
neuronen meet. De CSM, of 'conditionele synchroniciteitmaat', beschrijft de kans
waarmee neuronen samen een impuls uitstoten, gegeven dat er activiteit is in één van
de neuronen. Een model voor gecorreleerde binaire data wordt voorgesteld, welke
zowel de activiteit in de neuronen afzonderlijk, als de synchroniciteit aan de hand van
de CSM beschrijft.

In Hoofdstuk 9 staat een dierenziekten survey uit België centraal. Gebaseerd
op serologische gegevens van een prevalentie studie van het boviene herpesvirus-1
bestuderen we de infectiedruk. De infectiedruk is een biologische parameter die de
snelheid weergeeft om op een bepaalde leeftijd geïnfecteerd te worden. Het is in feite
de afgeleide functie van de prevalentie curve van geïnfecteerde subjecten als func-
tie van de leeftijd. Ook hier speelt clustering een cruciale rol aangezien runderen
uit een zelfde bedrijf meer gelijkaardig zullen zijn in vergelijking met runderen uit
verschillende bedrijven door gelijke levensomstandigheden en overdraagbaarheid van
de ziekte. Bovendien is de kans op een infectie gerelateerd met de grootte van het
bedrijf, waardoor de probleemstelling zeer nauwkeurig moet worden omschreven. Het
belang van het gebruik van modellen welke rekening houden met de clustering wordt

getoond, alsook de mogelijkheid voor het schatten van de prevalentie en de infec-
tiedruk. Door op gepaste wijze de correlaties in rekening te brengen kunnen zowel
populatiegemiddelde als bedrijfsspecifieke patronen van de prevalentie en infectiedruk
bestudeerd worden.

Hoofdstuk 10 is gewijd aan een toepassing in de neuro-gastroenterologie. In deze
context wensen we het gedrag te bestuderen van impulsen gegenereerd door neu-
ronen in de darmwand gedurende opeenvolgende zogenaamde slow waves. Zowel de
ruimtelijk als tijdsafhankelijke trend werd gemodelleerd door het gebruik van flexibele
modellen. GLMM, of 'generalized linear mixed model', kunnen aangewend worden
om de ruimtelijk geordende data te 'smoothen' door het verband van mixed models
met penalized splines. Door het algemene kader van de GLMMs kan smoothing en
clustering op uniforme wijze in het model gebracht worden.