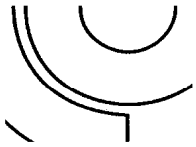

Limburgs Universitair Centrum

Faculteit Wetenschappen

Pseudo-likelihood Methods and Generalized Estimating Equations: Efficient Estimation Techniques for the Analysis of Correlated Multivariate Data

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

HELENA GEYS

Promotor:
Prof. dr. G. Molenberghs

1999

Voor Geert

Dankwoord

Bij het voltooien van dit werk wil ik graag al diegenen bedanken die rechtstreeks of onrechtstreeks bijgedragen hebben tot het tot stand komen ervan.

In de eerste plaats, mijn promotor Geert Molenberghs, die met zeer veel belangstelling deze thesis heeft begeleid en mij zoveel interessante suggesties deed. Zijn hulp en bijstand waarop ik voortdurend beroep mocht doen waren voor mij steeds aanmoedigingen om verder te werken. Geert, ik wil je nu dan ook van harte danken voor de jarenlange, openhartige samenwerking.

I gratefully acknowledge Tomasz Burzykowsky, Marc Buyse, Paul Catalano, Stuart Lipsitz, Meredith Regan, Didier Renard, Louise Ryan and Paige Williams for many helpful discussions and stimulating my interest in research. Their comments on earlier drafts have greatly improved this manuscript.

De aanhoudende blijken van belangstelling en de opbouwende kritiek van alle leden van het Centrum voor Statistiek waren een voortdurende stimulans bij de realisatie van deze thesis. Ik ben jullie allemaal heel veel dank verschuldigd.

Mijn dank gaat tevens uit naar Martien voor het scan-werk en Annemie, Conny, Hilde, Martine en Vivianne van het WNI secretariaat voor het uit handen nemen van vele administratieve werkjes.

Tenslotte, maar niet in het minst wens ik mijn ouders en Geert te bedanken voor hun onuitputtelijk geduld. Meer dan wie ook, waren zij voor mij een nooit aflatende morele steun, die soms wel heel erg noodzakelijk was ...

Helena Geys

Diepenbeek,
28 september 1999.

Contents

List of Abbreviations

ARE	Asymptotic Relative Efficiency
ARMD	Age Related Macular Degeneration
BMD	Benchmark Dose
CAP	cyclophosphmide plus adriamycin plus cisplatin
CS	Cluster-Specific
CSYM	Compound Symmetry Model
CONDLOG	Conditional Logistic Model
CP	cyclophosphamide plus cisplatin
DEHP	Di(-Ethyhexyl)-Phthalate
DYME	Diethylene Glycol Dimethyl Ether
EG	Ethylcnc Glycol
EPA	Environmental Protection Agency
FDA	Food and Drug Administration
GEEn	Generalized Estimating Equations (nth order)
GLMM	Generalized Linear Mixed Model
LED	Lower Effective Dose
MBN	Midbrain
MIXLOG	Mixed Effects Logistic Model
ML	Maximum Likelihood
MR	Molenberghs-Ryan (Model)
NOAEL	No Observable Adverse Effect Level
NTP	National Toxicology Progam

OLF	Olfactory System
OPT	Optic System
PA	Population-Averaged
PL	Pseudo (maximum) Likelihood
QRA	Quantitative Risk Assessment
RTG	Relative Time Gain
SD	Standard Deviation
TGDM	Triethylene Glycol Dimethyl Ether
THEO	Theophylline

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Developmental Toxicity Studies

1.1.1 Background

Lately, society has been increasingly concerned about problems related to fertility and pregnancy, birth defects, and developmental abnormalities. Consequently, regulatory agencies such as the U.S. Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA) have given increased priority to protection against drugs, harmful chemicals and other environmental hazards. As epidemiological evidence of adverse effects on fetal development may not be available for specific chemicals present in the environment, laboratory experiments in small mammalian species provide an alternative source of evidence essential for identifying potential developmental toxicants. For ethical reasons, animal studies afford a greater level of control than epidemiological studies. Moreover, they can be conducted in advance of human exposure. Unfortunately, there have been cases in which animal studies have not been run properly. The Thalidomide tragedy is a prominent example (Salsburg 1996). Thalidomide was present in at least 46 countries under many different brand names. In Belgium it is best known as "Softenon". The drug was described as being "safe" because it was not possible to develop toxic lesions in animal trials. Unfortunately, this was not the case. An estimated 10.000 children were born throughout the world as deformed, some with fin-like hands grown directly on the shoulders, with stunted or missing limbs, deformed eyes and ears, ingrown genitals, absence of a lung, a great many of them still-born or dying shortly after birth, etc. The animal

tests performed by the inventor of the drug were very superficial and incomplete. They did not carry out animal tests specifically to demonstrate teratogenetic effects. This runs contrary to the basic ideas behind such studies. According to Paracelsus all compounds are potential poisons: "Only the dose makes a thing not a poison". Malformations, like cancer, could occur when practically any substance, including sugar and salt, is given in excessive doses. A proper animal study should therefore always include a dose at which a toxic lesion happens.

As a consequence of the thalidomide tragedy, there has been a marked upsurge in the number of animals used in testing of new drugs. Also, drugs are now specifically tested on pregnant animals to safeguard against possible teratogenic effects on the human foetus. However, methods for extrapolating the results to humans are still being developed and refined. Differences in the physiological structure, function and biochemistry of the placenta, that exist between species make reliable predictions difficult.

Since laboratory studies further 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 are used (Williams and Ryan 1996). Three standard procedures (Segments I, II and III) have been established to assess specific types of effects.

- Segment I or fertility studies are designed to assess male and female fertility and general reproductive ability. Such studies are typically conducted in one species of animals and involve exposing males for 60 days and females for 14 days prior to mating.

- Segment II studies are also referred to as "teratology studies", since historically the primary goal was to study malformations (the origin of the word "teratology" lies in the Greek word "tera", meaning monster). In Section 1.1.2, I will describe standard teratology studies in greater detail.

- Segment III tests are focused on effects later in gestation and involve exposing pregnant animals from the 15th day of gestation through lactation.

In addition, I will describe alternative animal test systems, such as the so-called "heatshock studies" in Section 1.1.4. Throughout this work, we will focus on standard Segment II and heatshock studies.

1.1.2 The Segment II Study: a Standard Experimental Design

A Segment II experiment involves exposing timed-pregnant animals (rats, mice and occasionally rabbits) during major organogenesis (days 6 to 15 for mice and rats) and structural development. Administration of the exposure is generally by the clinical or environmental routes most relevant for human exposure. Dose levels consist of a control group and 3 or 4 dose groups, each with 20 to 30 pregnant dams. The dams are sacrificed just prior to normal delivery, at which time the uterus is removed and thoroughly examined (Figures 1.1 and 1.2).

An interesting aspect of Segment II designs is the hierarchical structure of the developmental outcomes. Figure 1.3 illustrates the data structure. An implant may be resorbed at different stages during gestation. If the implant survives being resorbed, the developing fetus is at risk of fetal death. Adding the number of resorptions and fetal deaths yields the number of non-viable fetuses. If the fetus survives the entire gestation period, growth reduction such as low birth weight may occur. The fetus may also exhibit one or more types of malformation. These are commonly classified into three broad categories:

- external malformations are those visible by naked eye, for instance missing limbs;
- skeletal malformations might include missing or malformed bones;
- visceral malformations affect internal organs such as the heart, the brain, the lungs etc.

Each specific malformation is typically recorded as a dichotomous variable (present or absent). Adding the number of resorptions, the number of fetal deaths and the number of viable fetuses yields the total number of implantations. Since exposure to the test agent takes place after implantation, the number of implants, a random variable, is not expected to be dose-related.

1.1.3 Challenges in Model Development

The analysis of developmental toxicity data as described above, raises a number of challenges (Molenberghs et al. 1998, Zhu and Fung 1996), summarized below.

Figure 1.1: *Dissected Mouse with Removed Uterus*

Figure 1.2: *Uterus with Removed Fetus*

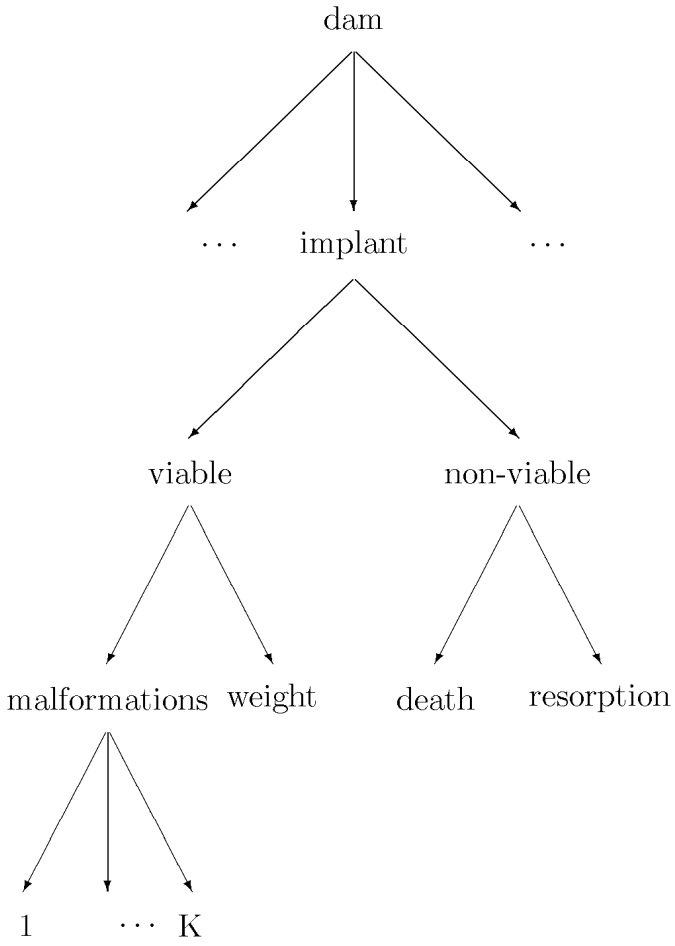

Figure 1.3: *Data Structure of Developmental Toxicity Studies.*

- Because of genetic similarity and the same treatment conditions, offspring of the same mother behave more alike than those of another mother. This has been termed *litter effect*. As a result, responses on different fetuses within a cluster are likely to be correlated, inducing extra variation in the data relative to those associated with the common binomial or multinomial distribution. This extra variation must be taken into account in statistical analyses (Chen and Kodell 1989; Kupper et al. 1986).

- Since deleterious events can occur at several points in development, an interesting aspect lies in the staging or *hierarchy* of possible adverse fetal outcomes (Williams and Ryan 1996). Ultimately, a model should take into account this hierarchical structure in the data: (i) a toxic insult early in gestation may result in a resorbed fetus; (ii) thereafter an implant is at risk of fetal death; (iii) fetuses that survived the entire gestation period are threatened by low birth weight and/or several types of malformation.

- While some attempts have been made for the joint analysis of prenatal death and malformation (Chen et al. 1991; Ryan 1992), the analysis of developmental toxicity data has usually been conducted on the number of viable fetuses alone. An appropriate statistical model should then account for possible *correlations among the different fetal endpoints*.

- As the number of viable fetuses can sometimes affect the chance of an adverse effect (in a large litter a larger number of animals have to compete for the same maternal resources and therefore the probability of malformation may be larger), a model should also be flexible enough to allow *litter size* to affect response probabilities.

- Finally, one may have to deal with outcomes of a *mixed* continuous (low birth weigth)/discrete (malformation indicator) nature.

1.1.4 Heatshock Studies

A unique type of developmental toxicity study was originally developped by Brown and Fabro (1981) to assess the impact of heat stress on embryonic development. Subsequent adaptations by Kimmel et al. (1994) allows the investigation of effects,

related to both temperature and duration of exposure. These heatshock experiments are described in Section 2.2. The embryos are explanted from the uterus of the maternal dam and cultured in vitro. Next, each embryo is exposed to a short period of heat stress by placing the culture vial into a warm water bath, involving an increase over body temperature of 4 to 5°C for a duration of 5 to 60 minutes. The embryos are examined 24 hours later for impaired or accelerated development. This type of developmental test system has several advantages over the standard Segment II design. First of all, the exposure is directly administered to the embryo, so that controversial issues regarding the unknown relationship between the exposure level to the maternal dam and that which is actually received by the embryo, need not be taken into account. Secondly, the exposure pattern can be easily controlled, since target temperature levels in the warm water baths can be achieved within 2 minutes. Further, information regarding the effects of exposure are quickly obtained, in contrast to the Segment II study which requires 8 to 12 days after exposure to assess impact. And finally, this animal test system provides a convenient mechanism for examining the joint effects of both duration of exposure and exposure levels.

1.2 Risk Assessment

Risk assessment can be defined as (Roberts and Abernathy 1996): "the use of available information to evaluate and estimate exposure to a substance and its consequent adverse health effects." An important goal in the risk assessment process is to determine a safe level of exposure. Traditionally, quantitative risk assessment in developmental toxicology has been based on the NOAEL, or No Observable Adverse Effect Level, which is the dose immediately below that deemed statistically or biologically significant when compared with controls. The NOAEL, however, has been criticized for its poor statistical properties (see for example, Williams and Ryan 1996). Therefore, interest in developing techniques for dose-response modeling of developmental toxicity data has increased, and new regulatory guidelines (U.S. EPA 1991) emphasize the need of quantitative methods for risk assessment. The standard approach requires the specification of an adverse event, along with $r(d)$ representing the probability that this event occurs at dose level d. 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 oc-

curred if a particular offspring is abnormal (fetus based). Alternatively, one might conclude that an adverse effect has occurred if at least one offspring from the litter is affected (litter based). Based on this probability, a common measure for the excess risk over background is defined as

$$r^*(d) = r(d) - r(0)$$

or as

$$r^*(d) = \frac{r(d) - r(0)}{1 - r(0)}, \tag{1.1}$$

where definition (1.1) puts greater weight on outcomes with large background risks. The benchmark dose (BMD_q) is then defined as the dose satisfying $r^*(d) = q$, where q corresponds to the pre-specified level of increased response and is typically specified as $0.01, 1, 5$ or 10% (Crump 1984).

In practice, calculation of the BMD follows several steps. After choosing and fitting an appropriate dose-response model, the excess risk function is solved for the dose, d, that yields $r^*(d) = q$. Since the dose-response curve is estimated from data and has inherent variability, the BMD is itself only an estimate of the true dose that would result in this level of excess risk. The final step therefore consists of acknowledging this sampling uncertainty for the model on which the BMD_q is based, by replacing the BMD_q by its lower confidence limit (Williams and Ryan 1996). Several approaches have been proposed.

Using the delta method, a Wald based method can be used:

$$\widehat{BMDL}_q = \widehat{BMD}_q - 1.645\sqrt{\widehat{\text{Var}}(\widehat{BMD}_q)}.$$

Assume that $\boldsymbol{\beta}$ is the vector of parameters included in the dose-response model, then the BMD_q variance can be obtained from the variance matrix of $\boldsymbol{\beta}$. Several authors have indicated that this method suffers from drawbacks, especially with low dose extrapolation (Aerts, Declerck and Molenberghs 1997; Crump 1984; Crump and Howe 1983; Krewski and Van Ryzin 1981) in which case the method may yield negative lower limits. Furthermore, Catalano, Ryan and Scharfstein (1994) have empirically found that this method can yield unstable estimates.

Alternatively, an upper limit for the risk function can be computed, and thus the dose that corresponds to a $q\%$ increased response above background is determined

from this upper limit curve by solving:

$$\hat{r}^*(d) + 1.645\sqrt{\widehat{\mathrm{Var}}(\hat{r}^*(d))} = q,$$

where the variance of the estimated increased risk function $\hat{r}^*(d)$ is estimated as:

$$\widehat{\mathrm{Var}}(\hat{r}^*(d)) = \left(\frac{\partial r^*(d)}{\partial \boldsymbol{\beta}}\right)^T \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial r^*(d)}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

and where $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. The resulting dose level is referred to as the lower effective dose (LED_q) (Kimmel and Gaylor 1988).

Crump and Howe (1983) recommend using the asymptotic distribution of the likelihood ratio (if available). According to this method, an approximate $100(1-\alpha)\%$ lower limit for the BMD, denoted by BMD(1), corresponding to an excess risk of q is defined as

$$\min\{d(\boldsymbol{\beta}) : r(d;\boldsymbol{\beta}) = q \text{ over all } \boldsymbol{\beta} \text{ such that } 2(\ell(\hat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta})) \leq \chi^2_p(1-\alpha)\},$$

where ℓ denotes the log-likelihood and p is the number of model parameters. A second approach, denoted BMD(2), is based on the profile likelihood method (Morgan 1992). First, construct a profile likelihood based confidence interval for the dose effect parameter β_d. Secondly, transform this interval into an interval for d and check that the transformation is monotonic. Aerts, Declerck and Molenberghs (1997) compare the different lower limits for the BMD and show that, in general, BMD(1) yields lower results than BMD(2). Furthermore, they note that for conditionally specified models, the transformation is not monotonic, and hence the BMD(2) should not be applied to such models. A variation on this theme, suggested by many authors (Chen and Kodell 1989; Ryan 1992; Gaylor 1989), first determines a lower confidence limit, e.g. corresponding to an excess risk of 1 per cent, and then linearly extrapolates it to a BMD. The main advantage quoted for this procedure is that the determination of a BMD is less model dependent.

1.3 Accounting for Litter Effects

In most developmental toxicity studies, exposure is administered to the dam, rather than directly to the developing fetuses. Because of genetic similarity and the same

treatment conditions, offspring of the same mother behave more similar than those of another mother. This has been termed "litter effect". Failure to account for the clustering in the data can lead to serious underestimation of the variances of dose effect parameters and, hence, inflated test statistics. The need for methods that appropriately account for the heterogeneity among litters, especially with regard to binary outcomes, has long been recognized. When the response is continuous and assumed to be approximately Gaussian, there is a general class of linear models that is suitable for analyses. However, when the response variable is categorical, fewer techniques are available. This is partly due to the lack of a discrete analogue to the multivariate normal distribution. The use of binomial or Poisson models in toxicological testing has frequently been criticized on the grounds that they generally poorly fit actual experimental data. This is caused by extra-binomial variation, i.e. more variability among litters than would be expected based on binomial or Poisson models. In an attempt to explain this variation, a number of generalized linear models have been proposed. Williams (1975) assumes that fetuses in the same litter provide a set of independent Bernoulli responses conditional on the litter-specific succes probability, and that the variation in this probability from litter to litter follows a beta-distribution. Haseman and Kupper (1979) provide an excellent survey of likelihood generalizations of standard distributions to account for clustering.

In general, models for multivariate correlated binary data can be grouped into the following different classes:

- conditionally specified models,
- marginal models,
- cluster-specific models

(Diggle, Liang and Zeger 1994). The answer to the question of which model family is to be preferred depends principally on the research question(s) to be answered. In conditionally specified models the probability of a positive response for one member of the cluster is modelled conditionally on other outcomes for the same cluster, while marginal models relate the covariates directly to the marginal probabilities. Cluster-specific models differ from the two previous models by the inclusion of parameters which are specific to the cluster. What *method* is used to fit the model, should not only depend on the assumptions the investigator is willing to make, but also (to

some extent) on the availability of computational algorithms. If one is willing to fully specify the joint probabilities, maximum likelihood methods can be adopted. Yet, if only a partial description in terms of marginal or conditional probabilities is given, one has to rely on non-likelihood methods such as:

- generalized estimating equations,
- pseudo-likelihood methods.

1.3.1 Conditional Modelling

In a conditional model the parameters describe a feature (probability, odds, logit, ...) of (a set of) outcomes, given values for the other outcomes (Cox 1972). The best known example is undoubtedly the log-linear model. Rosner (1984) described a conditional logistic model. Due to the popularity of marginal (especially generalized estimating equations) and random-effects models for correlated binary data, conditional models have received relatively little attention, especially in the context of multivariate clustered data. Diggle, Liang and Zeger (1994, pp. 147–148) criticized the conditional approach because the interpretation of the dose effect on the risk of one outcome is conditional on the responses of other outcomes for the same individual, outcomes of other individuals and the litter size. Molenberghs, Declerck and Aerts (1998) and Aerts, Declerck and Molenberghs (1997) have compared marginal, conditional and random-effects models for univariate clustered data. Their results are encouraging for the conditional model, since they are competitive for the dose effect testing and for benchmark dose estimation, and because they are computationally fast and stable. Molenberghs and Ryan (1999), henceforth abbreviated as MR, discuss the advantages of conditional models and how, with appropriate care, the disadvantages can be overcome. They constructed the joint distribution for clustered multivariate binary outcomes, based on a multivariate exponential family model. A slightly different approach, also based on the exponential family, is presented in Fitzmaurice, Laird, and Tosteson (1996). An advantage of such a likelihood-based approach is that, under correct model specification, efficiency can be gained over other procedures such as generalized estimating equations (GEE) methods. Furthermore, the model provides a natural framework for quantitative risk assessment (Chapter 5). Present approaches estimate benchmark doses (Crump 1984) based on the marginal probability of a single offspring being affected (Chen and Kodell

1989). 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 likelihood basis of the MR model allows calculation of quantities such as the probability that at least one littermate is affected (probability of an affected litter). In contrast, GEE based models do not provide a way to derive such quantities since they do not specify the joint probability between outcomes but only marginal probabilities and a working correlation matrix. While they could be calculated from a fully specified marginal model, fitting these models is hampered by lengthy computations and/or parameter restrictions (Molenberghs, Declerck and Aerts 1998 and Aerts, Declerck and Molenberghs 1997).

The flexibility of the MR model partly relies on the exponential family framework. However, maximum likelihood estimation can be unattractive, due to excessive computational requirements. For example, with multivariate exponential family models, the normalizing constant can have a cumbersome expression, rendering it hard to evaluate (Arnold and Strauss 1991). Several suggestions have been made to overcome this problem, such as Monte Carlo integration (Tanner 1991). For example, Geyer and Thompson (1992) use Markov Chain Monte Carlo simulations to construct a Monte Carlo approximation to the analytically intractable likelihood. Arnold and Strauss (1991) and Arnold, Castillo and Sarabia (1992) propose the use of a so-called *pseudo-likelihood* (PL). Pseudo-likelihood (or pseudo-maximum-likelihood) methods are alternatives to maximum likelihood estimation that retain the methodology and properties while trying to eliminate some of the difficulties such as strong distributional assumptions or intensive computations. The idea is that a parametric family of models is specified, to which likelihood methodology is applied; the method is denoted "pseudo", as there is no assumption that this family is the true distribution generating the data. Geys, Molenberghs and Ryan (1996, 1997, 1999) implemented a pseudo-likelihood method for the MR model that replaces the joint distribution of the responses, a multivariate exponential-family model, by a product of conditional densities that do not necessarily multiply to the joint distribution (see also Chapters 3 and 4). In this approach, the normalizing constant cancels, thus greatly simplifying computations, especially when litter sizes are large and variable (since the normalizing constant depends on litter size). In following chapters we will show that pseudo-likelihood estimation is an attractive alternative for maximum likelihood estimation in the context of clustered binary

data. Moreover, since the pseudo-likelihood still reflects the underlying likelihood it can be useful for dose-response modelling (e.g. to determine a benchmark dose). Pseudo-likelihood estimation turned out to be also extremely useful in the context of spatial statistics (Cressie 1991). Besag (1975) used pseudo-likelihood estimation in the context of a general Markov random field and established consistency of the estimators. A selection of other applications of this technique can be found in Connolly and Liang (1988), Liang and Zeger (1989) and Le Cessie and Van Houwelingen (1994).

1.3.2 Marginal Modelling

In marginal models, the parameters characterize the marginal probabilities of a subset of the outcomes, without conditioning on the other outcomes. Advantages and disadvantages of conditional and marginal modelling have been reviewed by Molenberghs (1992), pp. 24–25.

Bahadur (1961) proposed a marginal model, accounting for the association via marginal correlations. This model has also been studied by Cox (1972), Kupper and Hasemann (1978) and Altham (1978). Assuming exchangeability, in the sense that each fetus within a litter has the same malformation probability, and in addition setting all the three- and higher-way correlations equal to zero, Bahadur's representation can be simplified to give the following marginal distribution of Z_i, the number of malformations in cluster i:

$$\begin{aligned} f(z_i|\pi_i,\rho_i) &= \binom{n_i}{z_i}\pi_i^{z_i}(1-\pi_i)^{n_i-z_i} \\ &\quad \times\left[1+\rho_i\left\{\binom{z_i}{2}\frac{1-\pi_i}{\pi_i}+\binom{n_i-z_i}{2}\frac{\pi_i}{1-\pi_i}-z_i(n_i-z_i)\right\}\right], \end{aligned}$$

where π_i denotes the malformation probability in the ith cluster and n_i denotes the litter size. A drawback is the fact that the correlation ρ_i is highly constrained when the higher order correlations have been removed. Even when higher order parameters are included, the parameter space of marginal parameters and correlations has a very peculiar shape. Bahadur (1961) discusses restrictions on the parameter space in the case of a second order approximation. From these, it can be deduced that the lower bound approaches zero as the cluster size increases. However, it is important to note that also the upper bound for ρ_i is constrained. Indeed, even though it is one

for clusters of size two, the upper bound varies between $1/(n_i - 1)$ and $2/(n_i - 1)$ for larger clusters. Taking a (realistic) cluster of size 12, the upper bound is in the range $(0.09; 0.18)$. Kupper and Haseman (1978) present numerical values for the constraints on ρ_i for choices of π_i and n_i. Restrictions for a specific version where a third order association parameter is included as well, have been studied by Prentice (1988). A more general situation is discussed in Declerck, Aerts and Molenberghs (1997).

Molenberghs and Lesaffre (1994) and Lang and Agresti (1994) have proposed models which parameterize the association in terms of marginal odds ratios. Dale (1986) defined the bivariate global odds ratio model, based on a bivariate Plackett distribution (Plackett 1965). Molenberghs and Lesaffre (1994) extended this model to multivariate ordinal outcomes. They generalize the bivariate Plackett distribution in order to establish the multivariate cell probabilities. Their method involves solving polynomials of high degree and computing the derivatives thereof. Lang and Agresti (1994) exploit the equivalence between direct modelling and imposing restrictions on the multinomial probabilities, using undetermined Lagrange multipliers. Alternatively, the cell probabilities can be fitted using a Newton iteration scheme, as suggested by Glonek and McCullagh (1995).

However, even though a variety of flexible models exist, maximum likelihood can be unattractive due to excessive computational requirements, especially when high dimensional vectors of correlated data arise. As a consequence, alternative methods have been in demand. Liang and Zeger (1986) proposed so-called *generalized estimating equations* (GEE1) which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the association structure. They estimate the parameters associated with the 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. Prentice (1988) extended their results to allow joint estimation of probabilities and pairwise correlations. Lipsitz, Laird and Harrington (1991) modified the estimating equations of Prentice (1988) to allow modelling of the association through marginal odds ratios rather than marginal correlations. When adopting GEE1 one does not use information of the association structure to estimate the main effect parameters. As a result, it can be shown that GEE1 yields consistent main effect estimators, even when the association structure is misspeci-

fied. However, severe misspecification may seriously affect the efficiency of the GEE1 estimators. In addition, GEE1 should be avoided when some scientific interest is placed on the association parameters. A second order extension of these estimating equations (GEE2) that include the marginal pairwise association as well, has been studied by Liang, Zeger and Qaqish (1992). They note that GEE2 is nearly fully efficient though bias may occur in the estimation of the main effect parameters when the association structure is misspecified. A variation to this theme, using conditional probability ideas, has been proposed by Carey, Zeger and Diggle (1993). It is referred to as *alternating logistic regressions.*

Le Cessie and Van Houwelingen (1994) suggested to approximate the true likelihood by means of a pseudo-likelihood function that is easier to evaluate and to maximize. Both GEE2 and PL yield consistent and asymptotically normal estimators, provided an empirically corrected variance estimator, often referred to as the sandwich estimator, is used. However, GEE is typically geared towards marginal models, whereas PL can be used with both marginal (Le Cessie and Van Houwelingen 1994; Geys, Molenberghs and Lipsitz 1998) and conditional models (Geys, Molenberghs and Ryan 1997, 1999). In Chapter 6 we will discuss the relative merits of PL and GEE and illustrate them using data from developmental toxicity studies.

1.3.3 Cluster-specific Modelling

Cluster-specific models are differentiated from population-averaged models by the inclusion of parameters which are specific to the cluster. Unlike for correlated Gaussian outcomes, the parameters of the cluster-specific and population-averaged models for correlated binary data describe different types of effects of the covariates on the response probabilities (Neuhaus 1992). The choice between population-averaged and cluster-specific strategies may heavily depend on the scientific goals. Population-averaged models evaluate the overall risk as a function of covariates; the conditionally specified models and marginal models, described above, belong to this class. With the cluster-specific approach, the response rates are modelled as a function of covariates and parameters, specific to a cluster. In such models, interpretation of fixed-effect parameters is conditional on a constant level of the cluster-specific parameter (e.g. random effect). Population-averaged comparisons, on the other hand, make no use of within cluster comparisons for cluster varying covariates and substan-

tially underestimate within cluster risks. Neuhaus, Kalbfleisch and Hauck (1991) discuss parameter interpretations of these models. They also draw the analogy with omitted covariates; i.e. unless the included and omitted covariates are uncorrelated (conditional on the response), the effect of a randomly assigned treatment will be biased towards zero. Thus, from these papers, population-averaged effects would be expected to be closer to zero than cluster-specific effects.

Within the class of cluster-specific models, we will study a *mixed-effect logistic* model as an alternative way of accounting for intra-litter heterogeneity as well as a *conditional likelihood* method. In the mixed-effect logistic procedure cluster effects are removed by assuming that they are realizations of a random variable and integrating over their distribution. With conditional likelihood, one conditions on the sufficient statistics for the cluster-specific effects (Ten Have, Landis and Weaver 1995; Conaway 1989).

1.4 Joint Modelling of Continuous and Discrete Outcomes

Developmental toxicity studies may seek to determine the effects of dose on fetal weight (continuous) and malformation incidence (binary) simultaneously, as both have been found to be indicative of a toxic effect. This motivates the formulation of a joint distribution with mixed continuous and discrete outcomes. However, this is not standard.

Catalano and Ryan (1992) note that latent variable models provide a useful and intuitive way to motivate the distribution of the discrete outcome. Such models presuppose the existence of an unobservable, normally-distributed random variable, underlying the binary outcome. The binary event is then assumed to occur if the latent variable exceeds some threshold value. They further note that this notion of latent variables has much appeal to toxicologists, because it provides a natural and intuitive framework for the biological mechanism leading to adverse events such as malformation.

A flexible latent variable approach to model an arbitrary number of continuous and discrete outcomes, each of which follows an exponential family distribution, is proposed by Sammel, Ryan and Legler (1997). They introduce a modified EM al-

gorithm for parameter estimation with either a simple Monte Carlo expectation or a numerical integration technique based on e.g. Gauss-Hermite quadrature to approximate the E-step which is not necessarily available in closed form. The method allows for arbitrary covariate effects and estimates of the latent variable are produced as a by-product of the analysis. However, their approach does not extend to correlated (i.e. clustered) data.

In the context of developmental toxicity studies, the dose-response model is often characterized in each of the two outcomes (weight and malformation) separately, using appropriate methods to account for correlation induced by the clustering of fetuses within litters, or the well-known "litter-effect". The more sensitive of the two outcomes is determined based on the dose-response patterns and used for risk assessment purposes. However, because these outcomes are correlated (Ryan et al. 1991), jointly modelling the outcomes and using the bivariate outcome as a basis for risk assessment may be more appropriate (Regan and Catalano 1998a). A standard 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 discrete or continuous outcome (Catalano and Ryan 1992; Cox and Wermuth 1992; Cox and Wermuth 1994; Fitzmaurice and Laird 1995; Olkin and Tate 1961). Cox and Wermuth (1992, 1994) consider various factorization methods and tests for independence. Let us discuss some factorization methods.

Catalano and Ryan (1992) apply the latent variable concept to derive the joint distribution of a continuous and a discrete outcome and then extend the model, using GEE ideas, to incorporate clustering. They parametrize the model in a way that allows to write the joint distribution as the product of the marginal distribution of the continuous response, and the conditional distribution of the binary response given the continuous one. The marginal distribution of the continuous response is related to covariates, using a linear link function, while for the conditional distribution they use a probit link. Due to the non-linearity of the link function relating the conditional mean of the binary response to the covariates, the regression parameters in the probit model of Catalano and Ryan (1992) have no direct marginal interpretation. Furthermore, if the model for the mean has been correctly specified, but the model for the association between the binary and continuous outcomes is misspecified, the regression parameters in the probit model are not consistent. The

lack of marginal interpretation and lack of robustness may be considered unattractive features of this approach. An important advantage, however, is that it can be readily extended to allow for clustering. Fitzmaurice and Laird (1995) circumvent the difficulties in the approach of Catalano and Ryan (1992) by factorizing the joint distribution as the product of a marginal Bernoulli distribution for the discrete response, and a conditional Gaussian distribution for the continuous response given the discrete one. Under independence, their method yields maximum likelihood estimates of the marginal means that are robust to misspecification of the association between the binary and continuous response. They also consider an extension of their model that allows for clustering. By using GEE methodology, they avoid the computational complexity of maximum likelihood in this more elaborate setting. A conceptual difficulty with this model is the interpretation of the parameters, which depends on cluster size.

A drawback of mixed outcome models based on factorization (as above) is that they may be difficult to apply for quantitative risk assessment (Geys et al. 1999b, Regan and Catalano 1998a). While taking into account the dependence between weight and malformation, the intrafetus correlation itself cannot be directly estimated. Thus, an expression for the joint probability that a fetus is affected (i.e. malformed and/or of low birht weight) is difficult to specify. Catalano et al. (1993) used a factorization model for quantitative risk assessment, in which direct estimation of the bivariate correlation is approximated using a conditioning argument. To overcome this problem, one needs joint models that incorporate the correlation between outcomes directly. Thus, a desirable model should have three properties:

- it allows separate dose-response functions for each component of the bivariate outcome,
- it accounts for the correlations due to clustering within litters,
- it estimates the bivariate intrafetus association.

In Chapter 8, we will propose models that satisfy these properties (see also Molenberghs and Geys 1998 and Geys et al. 1999b).

1.5 Organization of Subsequent Chapters

In Chapter 2, we present an overview of the different data sets that will be used throughout this work. Essentially, the data refer to three completely different study types: (i) standard Segment II studies, (ii) heatshock studies and (iii) clinical trials. The last group of data will be tackled in Chapter 9.

Chapter 3 introduces the MR model for a single binary outcome and explores pseudo-likelihood as an alternative mode of inference to maximum likelihood. Consistency and asymptotic normality of the pseudo-likelihood estimators are established. The pseudo-likelihood equations are derived, the model is applied to the NTP data described in Chapter 2 and an asymptotic and small sample relative efficiency study is performed.

In Chapter 4 pseudo-likelihood estimating equations are derived for the general multivariate clustered setting of MR. In that setting the pseudo-likelihood procedure becomes extremely useful, especially for larger cluster sizes (three or higher), where full maximum likelihood is hampered by excessive computing time requirements. In contrast, the pseudo-likelihood estimation method converges quickly, with only minor losses in efficiency, especially for a range of realistic parameter settings. Whereas in the univariate case described in Chapter 3 there is only one "natural" formulation of the pseudo-likelihood estimating equations, several plausible routes can now be followed. In addition, pseudo-likelihood counterparts for classical inferential tools such as Wald, score and likelihood ratio test statistics are formulated. They are shown to have easy-to-compute expressions and their limiting distributions are intuitively appealing. In contrast, GEE type versions of likelihood ratio test statistics (Rotnitzky and Jewell 1990) take a slightly less appealing form. Next, likelihood and pseudo-likelihood test statistics are compared using asymptotic and small sample simulations, and exemplified using the NTP data.

While in Chapters 3 and 4, the NTP data are merely used to exemplify the pseudo-likelihood methodology, the true data analysis is addressed thoroughly in Chapter 5. We restrict ourselves to the DEHP and EG data described in Sections 2.1.1 and 2.1.5, which were collected to investigate the toxicity of di(2-ethylhexyl)-phthalate and ethylene glycol in mice. The primary goal of such studies is to perform risk assessment, i.e. to set safe limits for human exposure, based on the fitted model. For risk assessment to be reliable, models should fit the data well in all respects.

Although classical polynomial predictors are very customary, they are often of poor quality, especially when low dose extrapolation is envisaged. A very elegant alternative approach to classical polynomials, which falls within the realm of (generalized) linear methods, is given by fractional polynomials. This method has been advocated by Royston and Altman (1994), and was applied by Royston and Wright (1998) for the construction of age-specific reference intervals and by Sauerbrei and Royston (1999) for building prognostic and diagnostic indices for multivariate models. An attractive feature is that conventional polynomials are included as a subset of this extended family. Since fractional polynomials provide a much wider range of functional forms, we switched to this approach (see also Geys et al. 1999a). Estimation is by pseudo-likelihood rather than maximum likelihood, due to the latter's excessive computational requirements. In order to select an appropriate dose-response model, we need to use the test statistics developed in Chapter 4. Once a suitable model is selected, it serves as basis for quantitative risk assessment.

In Chapter 6 we consider generalized estimating equations and pseudo-likelihood as alternatives for maximum likelihood for the analysis of exchangeable clustered binary data, using a marginal odds ratio model. As mentioned earlier in Section 1.3.2, maximum likelihood estimation can become prohibitive in a marginally specified model due to excessive computational requirements, especially when high dimensional vectors of correlated data arise. The extension to longitudinal data, which typically require more complicated association designs, needs further investigation. First, we construct an appropriate pseudo-likelihood function and derive its corresponding estimating equations. Depending on whether scientific interest focuses mainly on the main effects or shifts towards the association parameters, different pseudo-likelihood versions can be considered. Next, we present an equivalent but more appealing representation of the pseudo-likelihood estimating equations in terms of contrasts between observed and expected frequencies. We discuss the relative merits of the pseudo-likelihood methodology and generalized estimating equations and illustrate them using data from the NTP studies.

The standard approach for many teratology applications is to use a population-averaged model with primary interest on evaluating dose-response effects and where the covariate level is considered to be constant over a litter of animals. Yet, recently there has been growing interest in evaluating effects of covariates, such as fetal weight and uterine position, which can vary between individuals within a cluster.

In that case, individual-level covariates for teratology data are well justified. Chapter 7 describes several population-averaged as well as cluster-specific models for the analysis of developmental toxicity studies, in which individual-level covariates play an important role and applies them on the heatshock studies. Within the class of population-averaged models, we show that conditionally specified models, such as the one described in Section 1.3.1, should be avoided since they lead to undesirable properties. In addition, we present a simple goodness-of-fit testing procedure for clustered binary data which allows us to compare several possible association structures. Indeed, the specific form of the heatshock study allows us to quantify the association between different embryos from the same initial dam in terms of genetic as well as environmental factors, in contrast to the more standard teratology studies where exposure occurs through the maternal dam (Geys, Molenberghs and Williams 1997, 1999).

Measurements of both continuous and discrete outcomes are encountered in many statistical problems. In Chapter 8 we consider the particular context of teratology studies, where quantitative risk assessment is aimed at determining the effect of dose on the probability that an individual is malformed (binary indicator) or of low birth weight (continuous), both being important measures of teratogenicity. We introduce two different joint marginal mean models for outcomes of a mixed nature. First, we introduce a probit approach (Regan and Catalano 1999), in which the existence of an underlying continuous variable is assumed for each binary outcome. Hence, the joint distribution of weight and malformation can be assumed to follow a multivariate normal distribution. The second approach that we consider is the Plackett-Dale approach. Here, the latent malformation outcomes are assumed to follow a Plackett distribution. In both cases, specification of the full distribution is avoided using generalized estimating equations and pseudo-likelihood methodology respectively. Quantitative risk assessment is illustrated using data from a developmental toxicology experiment of ethylene glycol in rats.

Chapter 9 is dedicated to an additional application of the previously described methods for analyzing correlated data, in the context of a multiple clinical trials study. In that case, the data have a similar structure as in developmental toxicity studies. Different trials (clusters) are assumed to be independent. Individuals within a trial (fetuses within a cluster) may however be correlated possibly yielding multiple associated outcomes of potentially mixed data types.

Chapter 2

Motivating Examples

2.1 NTP studies

In this section we introduce developmental toxicity studies conducted by the Research Triangle Institute under contract to the National Toxicology Program (NTP). The studies concerned the effects in mice (occasionally rats) of five different chemicals: di(2-ethylhexyl)-phthalate (DEHP) (Tyl et al. 1988), ethylene glycol (EG) (Price et al. 1987), triethylene glycol dimethyl ether (TGDM) (George et al. 1987), diethylene glycol dimethyl ether (DYME) (Price et al. 1985) and theophylline (THEO) (Lindstrom et al. 1990).

2.1.1 DEHP Study in Mice

The use of phtalic acid esters as plasticizers for numerous plastic devices is widespread. The most commonly used ester is di(2-ethylhexyl)-phthalate (DEHP), which may constitute as much as 40% by weight of the finished products, in order to provide them a desirable flexibility and clarity. It has been well documented that small quantities of phtalic acid esters may leak out of plastic containers in the presence of food, milk, blood, or various solvents. Due to their ubiquitous distribution and presence in human and animal tissues, the possible toxic effects of the phtalic acid esters have been the subject of considerable concern. In particular, the developmental toxicity study described by Tyl et al. (1988) has attracted much interest in the toxicity of DEHP. The doses selected for the study were 0, 0.025, 0.05, 0.10 and 0.15 % DEHP with 25, 26, 26, 17 and 9 timed-pregnant mice assigned to each of

Table 2.1: *Summary Data from a DEHP Experiment in Mice.*

Dose	Dams	Live	Litter Size	Malformations		
(%)			(mean)	Ext.	Visc.	Skel.
0.000	25	330	13.2	0.0	1.5	1.2
0.025	26	288	11.1	1.0	0.4	0.4
0.050	26	277	10.7	5.4	7.2	4.3
0.100	17	137	8.1	17.5	15.3	18.3
0.150	9	50	5.6	54.0	50.0	48.0

these dose groups, respectively. Females were observed daily during treatment, but no maternal deaths or distinctive clinical signs were observed. The dams were sacrificed, slightly prior to normal delivery and the status of uterine implantation sites recorded. A total of 1082 live fetuses were dissected from the uterus, anaesthetised, and examined for external, visceral and skeletal malformations. Table 2.1 shows for each dose group, the number of pregnant dams, the number of live fetuses, the mean litter size and the rate of malformation for three different classes: external malformations, visceral malformations and skeletal malformations. The table suggests clear dose-related trends in the malformation rates. The average litter size (number of viable animals) decreases with increased levels of exposure to DEHP, a finding that is attributable to the dose-related increase in fetal deaths.

2.1.2 DYME Study in Mice

Diethylene glycol dimethyl ether (DYME) is a component of industrial solvents. These are widely used in the manufacture of protective coatings such as lacquers, metal coatings, baking enamels, etc. 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 (1) testicular toxicity and (2) embryotoxicity. Price et al. (1987) describe a study in which timed-pregnant mice were dosed with DYME throughout major organogenesis (gestational days 8 through

Table 2.2: *Summary Data from a DYME Experiment in Mice.*

Dose	Dams	Live	Litter Size	Malformations		
(mg/kg/day)			(mean)	Ext.	Visc.	Skel.
0.0	21	282	13.4	0.0	0.0	0.0
62.5	20	225	11.3	0.0	0.0	0.0
125	24	290	12.1	1.0	0.0	1.0
250	23	261	11.3	2.7	0.1	20.0
500	23	141	6.1	66.0	19.9	79.4

15). The doses selected for the study were 0, 62.5, 125, 250 or 500 mg/kg/day with 21, 20, 24, 23 and 23 pregnant dams assigned to each of these dose groups, respectively. Table 2.2 summarizes the data.

2.1.3 THEO Study in Mice

The developmental toxicity of orally administered Theopylline (THEO) in mice has been described by Lindstrom et al. (1990). Theophylline belongs to the class of compounds, used in the treatment of respiratory diseases, as anti-asthmatics, diuretics, etc. Theophylline has been shown to cross the human placenta and is secreted in breast milk. Therefore, there has been an increased interest in the teratogenetic potential of Theophylline in rodents. Table 2.3 summarizes the data from a developmental toxicity study, investigating the effect of Theophylline in Mice. The doses selected for the study were 0, 0.075, 0.15 or 0.20 % THEO with 25, 25, 29 and 17 pregnant dams assigned to each of these dose groups, respectively. The table suggests small dose-related trends in the malformation rates.

2.1.4 TGDM Study in Mice

Similar to DEHP, Triethylene glycol dimethyl ether (TGDM) is an industrial solvent with diverse applications. Its potential developmental toxicity has been investigated by George et al. (1987). Table 2.4 summarizes the data from their study. The doses

Table 2.3: *Summary Data from a THEO Experiment in Mice.*

Dose	Dams	Live	Litter Size	Malformations		
(%)			(mean)	Ext.	Visc.	Skel.
0.00	25	296	11.8	0.003	0.000	0.000
0.075	25	278	11.1	0.007	0.000	0.000
0.15	29	300	10.3	0.017	0.003	0.003
0.20	17	197	11.6	0.020	0.005	0.000

Table 2.4: *Summary Data from a TGDM Experiment in Mice.*

Dose	Dams	Live	Litter Size	Malformations		
(mg/kg/day)			(mean)	Ext.	Visc.	Skel.
0.0	26	319	12.3	0.003	0.000	0.000
250	26	275	10.6	0.000	0.000	0.000
500	24	262	10.9	0.004	0.000	0.004
1000	26	286	11.0	0.042	0.003	0.073

selected for the study were 0, 250, 500 or 1000 mg/kg/day TGDM with 26, 26, 24 and 26 pregnant dams assigned to each of these dose groups, respectively. Visceral malformations are very infrequent with TGDM (only one malformation observed).

2.1.5 EG Study in Mice

Ethylene glycol (EG) is a high-volume industrial chemical with diverse applications. For instance, it can be used as an antifreeze, as a solvent in the paint and plastics industries, as a softener in cellophane, etc. While EG may not be hazardous to humans in normal industrial handling, it can become dangerous when used at elevated temperatures or when ingested. The potential reproductive toxicity of EG has been evaluated recently in several laboratories. Price et al. (1985) for example,

Table 2.5: *Summary Data from an EG Experiment in Mice.*

Dose	Dams	Live	Litter Size	Malformations		
(mg/kg/day)			(mean)	Ext.	Visc.	Skel.
0	25	297	11.9	0.0	0.0	0.3
750	24	276	11.5	1.1	0.0	8.7
1500	22	229	10.4	1.7	0.9	36.7
3000	23	226	9.8	7.1	4.0	55.8

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 organogenesis and structural development of the fetuses (gestational days 8 through 15). The doses selected for the study were 0, 750, 1500 or 3000 mg/kg/day with 25, 24, 22 and 23 pregnant dams assigned to each of these dose groups, respectively. Table 2.5 shows the rate of malformed litters for each dose group and suggests clear dose-related trends for all three classes of malformation. While skeletal malformations are substantial in the highest dose group, external and visceral malformations show only slight dose effects.

Figures 2.1–2.3 show for each of these studies and for each dose group, the observed and averaged malformation rates in mice.

2.1.6 EG Study in Rats

Price et al. (1985) also describe a developmental toxicity experiment, investigating the effect of EG in rats. The doses selected for the present teratology study were 0, 1.25, 2.50 and 5.0 g/kg/day. A total of 1368 live rat fetuses were examined for low birth weight (continuous) or defects (binary). This joint occurence of continuous and binary outcomes will provide additional challenges in model development. Table 2.6 summarizes the malformation and fetal weight data from this experiment. The data show clear dose-related trends for both outcomes. The rate of malformation increases with dose, ranging from 1.3% in the control group to 68.6% in the highest dose group. The mean fetal weight decreases monotonically with increasing dose,

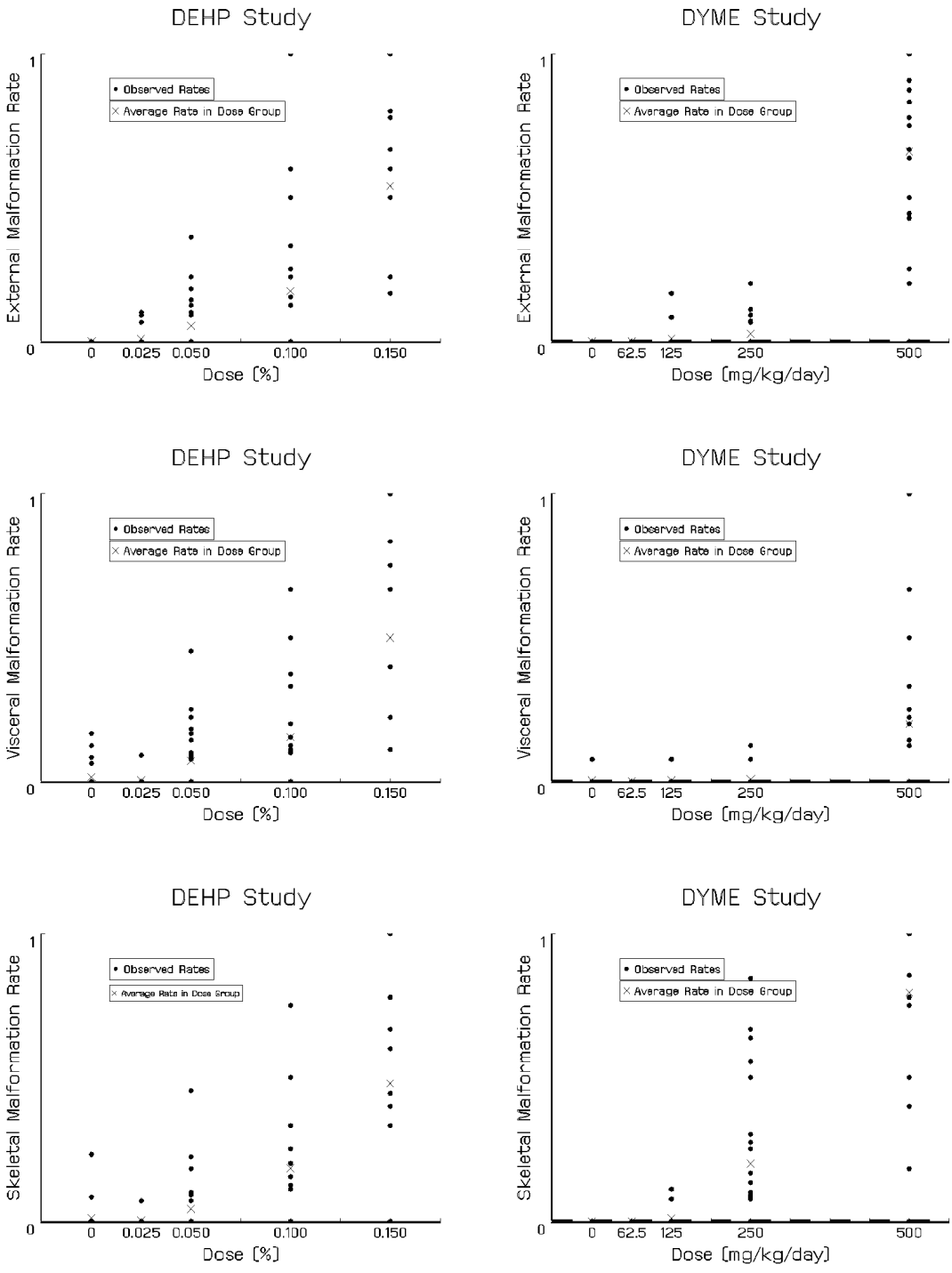

Figure 2.1: *DEHP and DYME Studies: Observed and Averaged Malformation Rates.*

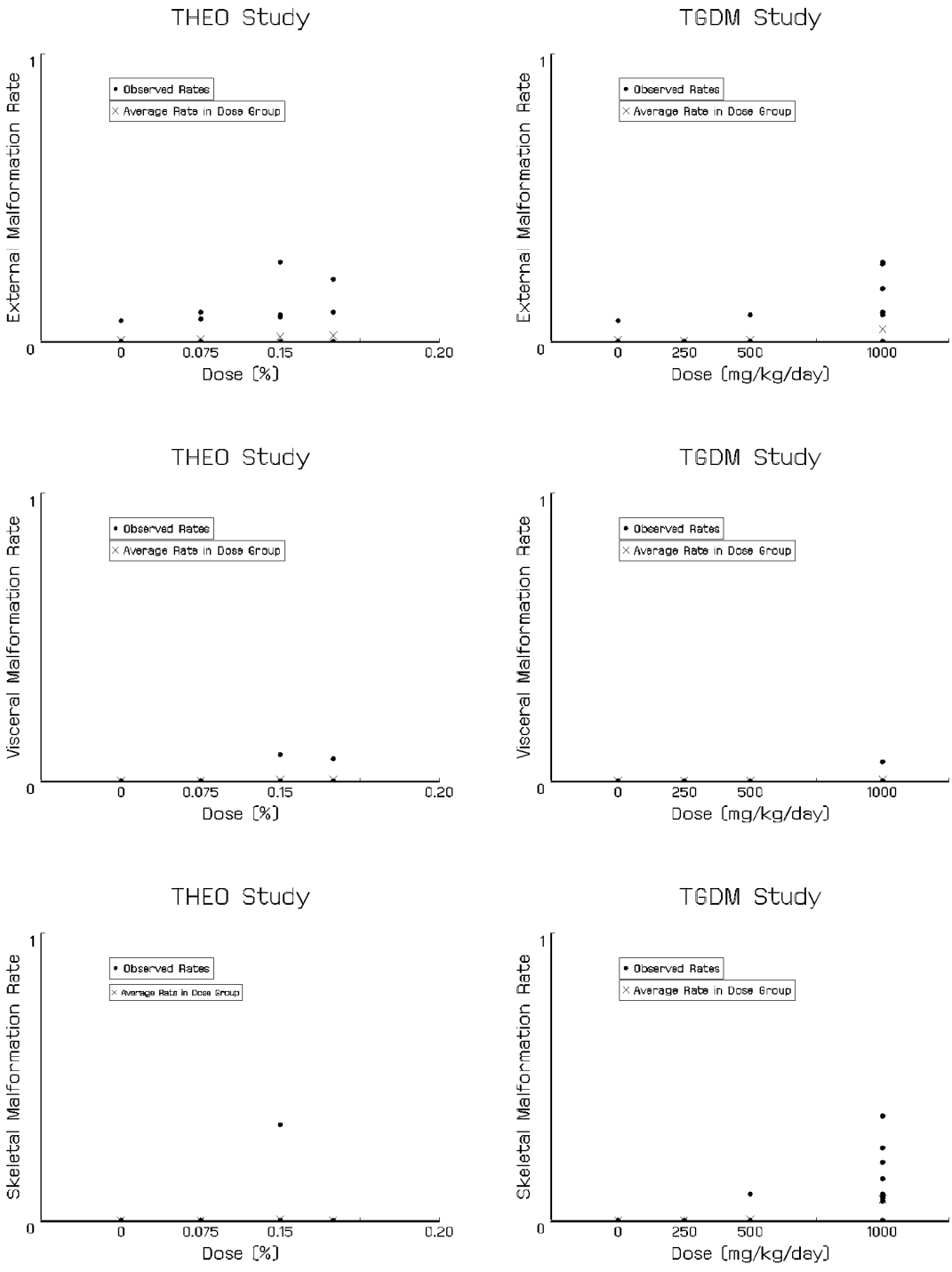

Figure 2.2: *THEO and TGDM Studies: Observed and Averaged Malformation Rates.*

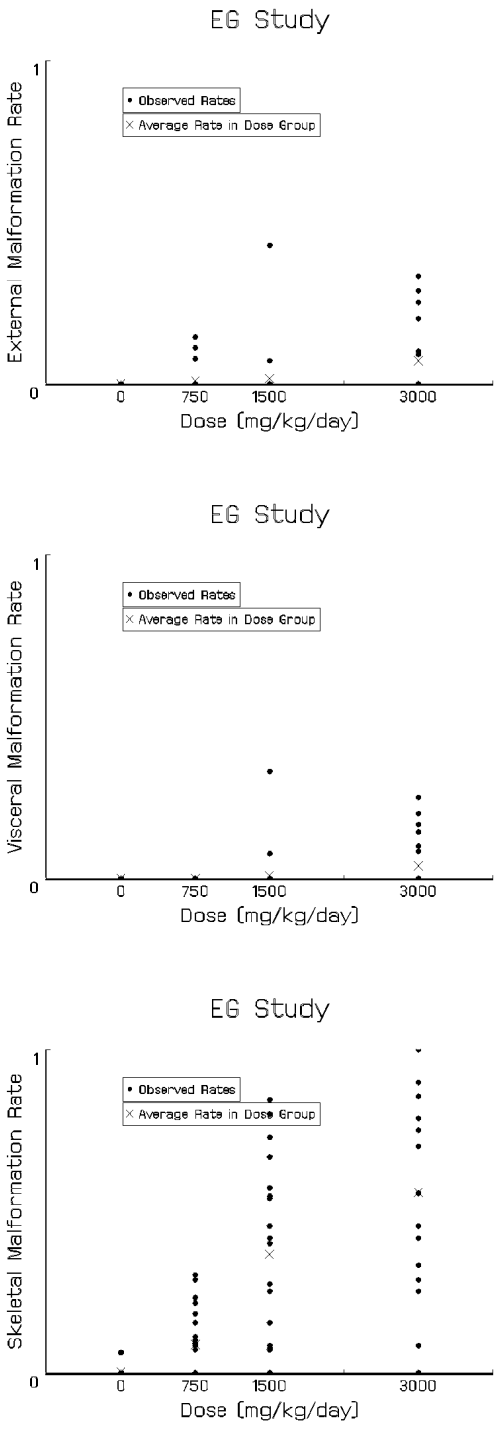

Figure 2.3: *EG Study: Observed and Averaged Malformation Rates.*

Table 2.6: *Summary Data from an EG Experiment in Rats.*

Dose	Dams	Live	Litter Size	Malf.		Weight		Pearson
(g/kg/day)			(mean)	Nr.	%	Mean	SD	Corr. (ρ)
0.00	28	379	13.50	5	1.3	3.40	0.38	0.07
1.25	28	357	12.75	21	5.8	3.30	0.37	0.00
2.50	29	345	11.89	86	24.9	2.90	0.36	-0.29
5.00	26	287	11.04	197	68.6	2.48	0.46	-0.37

ranging from 3.40 g to 2.48 g in control and highest dose group, respectively. The fetal weight variances, however, do not change monotonically with dose. In the lower dose groups, the variances remain approximately constant. However, in the highest dose group, the fetal weight variance is elevated. Further, it can be observed that simple Pearson correlation coefficients (ρ) between weight and malformation tend to strengthen with increasing doses. As doses increase, the correlation becomes more negative, because the probability of malformation is increasing and fetal weight is decreasing. This is illustrated in Figure 2.4, which shows the observed malformation rates for all clusters, the averaged malformation rates for each dose group, the average weight outcomes for all clusters and the average weight outcomes for each dose group.

2.2 Heatshock Studies

Heatshock studies have been described by Brown and Fabro (1981) and Kimmel et al. (1994). In these experiments, embryos are explanted from the uterus of a maternal dam (rats, mice or rabbits) during the gestation period and cultured *in vitro*. Each subject is subjected to a short period of heat stress by placing the culture vial into a water bath, usually involving an increase over body temperature of 4 to 5°C for a duration of 5 to 60 minutes. The embryos are examined 24 hours later for impaired and/or accelerated development. The studies collect measurements on 13 morphological variables. We will focus our attention on 3 of these (olfactory system (OLF), optic system (OPT), and midbrain (MBN)) and assess the effects of both

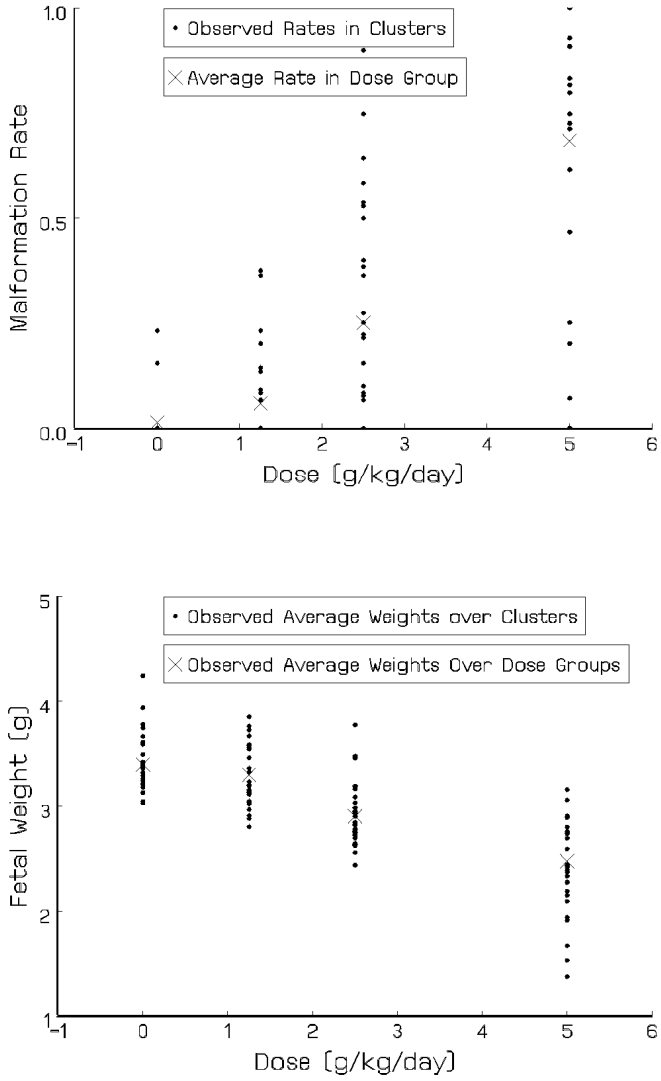

Figure 2.4: *EG (rats) Study: Observed Malformation Rates and Average Weights for all Clusters.*

Table 2.7: *Heatshock Studies: Number of (surviving) Embryo's Exposed to Each Combination of Duration and Temperature.*

Temperature	Duration of Exposure							Total
	5	10	15	20	30	45	60	
37.0	11	11	12	13	12	18	11	88
40.0	11	9	9	8	11	10	11	69
40.5	9	8	10	9	11	10	7	64
41.0	10	9	10	11	9	6	0	55
41.5	9	8	9	10	10	7	0	53
42.0	10	8	10	5	7	6	0	46
Total	60	53	60	56	60	57	29	375

Table 2.8: *Heatshock Studies: Distribution of Cluster Sizes.*

cluster size n_i	1	2	3	4	5	6	7	8	9	10	11
number of clusters of size n_i	6	3	6	12	13	11	8	5	2	3	2

duration and level of exposure on each morphological endpoint, coded as affected (1) versus normal (0).

While the heatshock studies do not represent a standard developmental toxicity test system (Tyl et al. 1988), they have several advantages. These include direct exposure to the embryo rather than the dam, easily controlled exposures, quick results, and a mechanism for exploring dose-rate effects.

The study design for the set of experiments conducted by Kimmel et al. (1994) is shown in Table 2.7, which indicates the number of embryos cultured in each temperature-duration combination. A total of 375 embryos, arising from 71 initial dams, survived the heat exposure. These were further examined for any affections and used for analysis.

The distribution of cluster sizes, ranging between 2 and 11, is given in Table 2.8.

The mean cluster size is 5. Since only surviving fetuses were included, cluster sizes are smaller than those observed in most other developmental toxicity studies and do not reflect the true original litter size.

Figure 2.5 shows the actual percentages of affected embryos for each experimental temperature-duration combination.

Historically, the strategy for comparing responses among exposures of different durations to a variety of environmental agents relies on a conjecture called Haber's Law, which states that adverse response levels should depend only on cumulative exposure (dose $\times$ exposure) (Haber 1924). We will return to this subject in Chapter 7.

2.3 Macular Degeneration Study

The data arise from a randomized multicentric clinical trial comparing an experimental treatment (Interferon-α) to a corresponding placebo administered to patients with age-related macular degeneration (ARMD). We focus on the comparison between placebo and the highest dose (6 million units daily) of Interferon-α, but the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group 1997). Patients with ARMD progressively loose vision. In the trial, the patients' visual acuity was assessed at different time points through their ability to read lines of letters on standardized vision charts. These charts display lines of 5 letters of decreasing size, which the patient should try to read from top (largest letters) to bottom (smallest letters). Each line with at least 4 letters correctly read is called one "line of vision". The patient's visual acuity is the total number of letters correctly read. The primary endpoint of the trial is the proportion of patients having lost at least 3 lines of vision in 1 year, compared to their baseline performance. The secondary endpoint of the trial is the mean visual acuity at 1 year. In Chapter 9, we examine whether visual acuity at 6 months can be used as a surrogate for visual acuity at 1 year with respect to the effect of Interferon-α. The data are shown graphically in Figure 2.6.

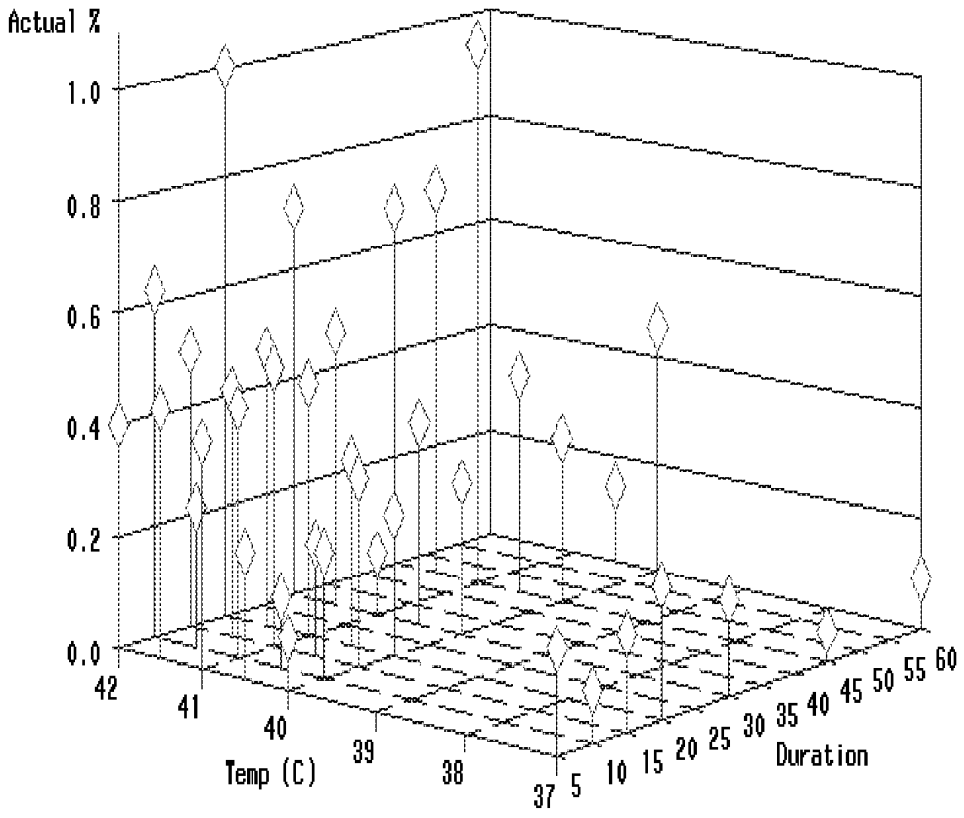

Figure 2.5: *Heatshock Studies: Actual Percentage of Affected Embryos (Experimental Data Points Only).*

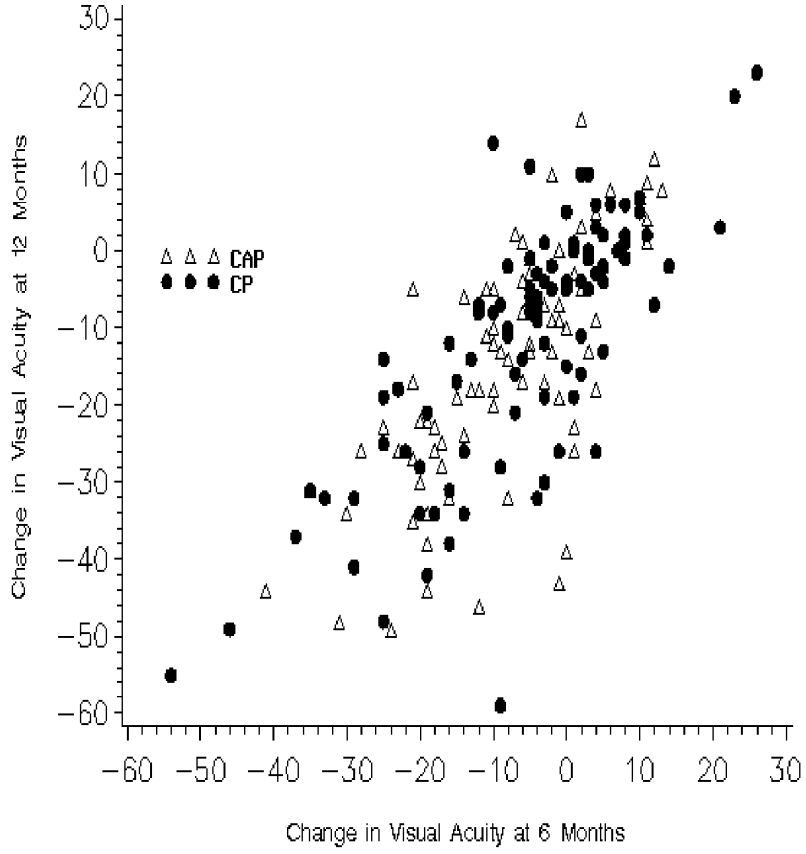

Figure 2.6: *ARMD Study: True Endpoint (change in visual acuity at 1 year) versus Surrogate Endpoint (change in visual acuity at 6 months) for all Individual Patients, Raw Data.*

2.4 Advanced Ovarian Cancer Study

Recently, there has been increased interest in the chemotherapy of ovarian carcinoma, and several large-scale, randomized trials have been conducted of various drug combinations. Here, the data come from a meta-analysis of four randomized multicenter trials in advanced ovarian cancer. Individual patient data are available in these four trials for the comparison of two treatment modalities: cyclophosphamide plus cisplatin (CP) versus cyclophosphamide plus adriamycin plus cisplatin (CAP). The full results of this meta-analysis were published with a mininimum follow-up of 5 years in all trials (Ovarian Cancer Meta-Analysis Project 1991). The dataset was subsequently updated to include a mininum follow-up of 10 years in all trials (Ovarian Cancer Meta-Analysis Project 1998). After such a long follow-up, most patients have had a disease progression or have died (952 of 1194 patients, i.e., 80%). Although methods that account for censoring would admittedly be preferable, censoring will be ignored in our analyses. The ovarian cancer dataset contains only four trials. This will turn out to be insufficient to apply the meta-analytic methods of Chapter 9. In the two larger trials, information is also available on the centers in which the patients had been treated. We can then use center as the unit of analysis for the two larger trials, and the trial as the unit of analysis for the two smaller trials. A total of 50 "units" are thus available for analysis, with a number of individual patients per unit ranging from 2 to 274.

Chapter 3

Pseudo-likelihood Estimation in Exponential Family Models with a Single Clustered Binary Outcome

3.1 Introduction

Molenberghs and Ryan (1999) proposed a likelihood-based model for clustered binary data, based on a multivariate exponential family model (Cox 1972). Their model is conditional in nature: it describes a feature of (a set of) outcomes condional on the other outcomes. This conditional interpretation is often seen as a drawback (Diggle, Liang and Zeger 1994, pp. 147–148), however in the Introduction we indicated that this difficulty is not a major issue. In this chapter we apply their model to the special case of a univariate clustered outcome, adopting exchangeability. While the model benefits from the elegance and simplicity of exponential family theory and is flexible in terms of allowing response rates to depend on cluster size, a main problem (particularly with large clusters) is the evaluation of the normalizing constant. Therefore, we introduce pseudo-likelihood as an alternative estimation method. Strictly speaking this is a non-likelihood method. 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. For example, when a joint density contains a computationally intractable normalizing constant, one might calculate a suitable product of conditional densities which does not involve

such a complicated function. While the method achieves important computational economies by changing the method of estimation, it does not affect model interpretation. Model parameters can be chosen in the same way as with full likelihood and retain their meaning.

Notation, model formulation and classical likelihood inference for the model proposed by Molenberghs and Ryan (1999) are introduced in Section 3.2. Section 3.3 defines pseudo-likelihood estimation. Section 3.4 describes the pseudo-likelihood concept for unclustered data. Pseudo-likelihood estimation for clustered binary outcomes is considered in Section 3.5 and its relative merits are assessed by means of some examples from developmental toxicity studies in Section 3.6. In addition, asymptotic as well as small sample relative efficiencies are studied in Sections 3.7 and 3.8.

3.2 Model Formulation

Consider an experiment involving N clusters, the ith of which contains n_i individuals, each of whom are examined for the presence or absence of M different responses. Suppose for the moment that $Y_{ijk} = 1$ when the kth individual in cluster i exhibits the jth response and 0 otherwise. Let $\boldsymbol{Y}_i$ represent the vector of outcomes for the ith cluster, and $\boldsymbol{x}_i$ an associated vector of cluster level covariates.

3.2.1 No Clustering

Let us first suppose there is no clustering ($n_i = 1; i = 1, \ldots, N$). Because $k \equiv 1$ in this setting, we drop this index temporarily from our notation. The observable outcome is thus $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iM})^T$. Let $\boldsymbol{W}_i$ be a $q \times 1$ vector containing the components of $\boldsymbol{Y}_i$, as well as their bivariate and higher order cross-products. (Hence the dimension of $\boldsymbol{W}_i$ is $\binom{M}{1} + \binom{M}{2} + \cdots + \binom{M}{M} = 2^M - 1$.) Next, consider the following probability mass function:

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i) = \exp\left\{\boldsymbol{\Theta}_i^T \boldsymbol{w}_i - A(\boldsymbol{\Theta}_i)\right\}, \tag{3.1}$$

where $\boldsymbol{\Theta}_i$ is a vector of natural parameters, having the same dimension as $\boldsymbol{W}_i$, and $A(\boldsymbol{\Theta}_i)$ is the normalizing constant. This model was proposed by Cox (1972).

Expanding the components explicitly leads to

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i;\boldsymbol{\Theta}_i) = \exp\left\{\sum_{j=1}^{M}\theta_{ij}y_{ij} + \sum_{j<j'}\omega_{ijj'}y_{ij}y_{ij'} + \ldots + \omega_{i1\ldots M}y_{i1}\ldots y_{iM} - A(\boldsymbol{\Theta}_i)\right\}$$

The θ parameters can be thought of as "main effects", whereas the ω parameters are association parameters or interactions. Models that do not include all interactions are derived by replacing $\boldsymbol{W}_i$ by one of its subvectors. A useful special case is found by setting all three and higher order parameters equal to zero:

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i;\boldsymbol{\Theta}_i) \propto \exp\left\{\sum_{j=1}^{M}\theta_{ij}y_{ij} + \sum_{j<j'}\omega_{ijj'}y_{ij}y_{ij'}\right\}, \tag{3.2}$$

which is a member of the quadratic exponential family discussed by Zhao and Prentice (1990). Thélot (1985) studied the case where $M = 2$. If $M = 1$, the model reduces to ordinary logistic regression.

We will briefly outline standard procedures for likelihood based parameter estimation in this setting. Modelling in terms of a parsimonious parameter vector of interest can be achieved using a linear model of the form $\boldsymbol{\Theta}_i = X_i\boldsymbol{\beta}$, where X_i is a $q \times p$ design matrix and $\boldsymbol{\beta}$ a $p \times 1$ vector of unknown regression coefficients. Let the mean parameter be $\boldsymbol{\pi}_i = E(\boldsymbol{W}_i)$. Then it is a basic property of exponential families (e.g. Brown 1986, p. 36) that $\boldsymbol{\pi}_i$ is related to the natural parameter $\boldsymbol{\Theta}_i$ by $\boldsymbol{\pi}_i = \partial A(\boldsymbol{\Theta}_i)/\partial\boldsymbol{\Theta}_i$. Next, the log-likelihood can be written as

$$\ell = \sum_{i=1}^{N}\ln f(\boldsymbol{y}_i;\boldsymbol{\Theta}_i) = \sum_{i=1}^{N}\left\{\boldsymbol{\beta}^T X_i^T\boldsymbol{w}_i - A(X_i\boldsymbol{\beta})\right\},$$

and the score function is

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N}X_i^T(\boldsymbol{w}_i - \boldsymbol{\pi}_i).$$

The maximum likelihood estimator for $\boldsymbol{\beta}$ is defined as the solution to $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$. It is usually found by applying a Newton-Raphson procedure, which coincides with a Fisher scoring algorithm for exponential family models with canonical link functions.

3.2.2 Clustered Outcomes

Let us now consider a single clustered outcome. Because the index j always equals 1, we drop it temporarily from our notation. We re-introduce however the subscript k to indicate an individual within a cluster.

Similarly to model (3.2), Molenberghs and Ryan (1999) derived the joint distribution of the clustered binary data $\boldsymbol{Y}_i$ as:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}_i; {\boldsymbol{\Theta}_i}^*, n_i) = \exp\left\{\sum_{k=1}^{n_i} \theta_i^* y_{ik} + \sum_{k<k'} \delta_i^* y_{ik} y_{ik'} - A(\boldsymbol{\Theta}_i^*)\right\}, \tag{3.3}$$

with δ_i^* describing the association between pairs of individuals within the ith cluster.

They code $Y_{ijk} = 1$ when the kth individual in cluster i exhibits the jth response and -1 otherwise. They use this coding rather than 1 and 0 since it provides a parameterization that more naturally leads to desirable properties when the roles of success and failure are reversed (see Cox and Wermuth 1994). Defining the number of individuals from cluster i with positive response to be z_i, (3.3) then becomes

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i^*, n_i) &= \exp\Bigg\{\theta_i^* z_i - \theta_i^*(n_i - z_i) \\ &\quad + \delta_i^*\left[\binom{z_i}{2} + \binom{n_i - z_i}{2} - z_i(n_i - z_i)\right] - A(\boldsymbol{\Theta}_i^*)\Bigg\} \\ &= \exp\left\{\theta_i^*(2z_i - n_i) + \delta_i^*\left[\binom{n_i}{2} - 2z_i n_i + 2z_i^2\right] - A(\boldsymbol{\Theta}_i^*)\right\}. \end{aligned}$$

Upon absorbing constant terms into the normalizing constant and using the reparametrization $\theta_i = 2\theta_i^*$ and $\delta_i = 2\delta_i^*$ this becomes

$$f_{\boldsymbol{Y}}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i, n_i) = \exp\left\{\theta_i z_i^{(1)} + \delta_i z_i^{(2)} - A(\boldsymbol{\Theta}_i)\right\}. \tag{3.4}$$

with $z_i^{(1)} = z_i$ and $z_i^{(2)} = -z_i(n_i - z_i)$. For this model, independence corresponds to $\delta_i = 0$. A positive δ_i corresponds to classical clustering or overdispersion, whereas a negative parameter value occurs in the underdispersed case. It is worthwhile to note that even for underdispersion, no restrictions are required on the parameter space. As discussed in Section 1.3.2, this feature is in contrast to other models for clustered data such as the Bahadur (1961) model. Molenberghs and Ryan (1999) show that model (3.4) has several additional desirable properties. First, the model is clearly invariant to interchanging the codes of successes and failures, whence

both estimation and testing will be invariant for this change as well. Second, the conditional probability of observing a positive response in a cluster of size n_i, given that the remaining littermates yield $z_i - 1$ successes is given by:

$$P(y_{ik} = 1 | z_i - 1, n_i) = \frac{\exp[\theta_i - \delta_i(n_i - 2z_i + 1)]}{1 + \exp[\theta_i - \delta_i(n_i - 2z_i + 1)]}, \tag{3.5}$$

which decreases to zero when n_i increases and z_i is bounded, and approaches unity for increasing n_i and bounded $n_i - z_i$, whenever there is a positive association between outcomes. From (3.5) it is clear that the conditional logit of an additional success, given $z_i - 1$ successes, equals $\theta_i - \delta_i(n_i - 2z_i + 1)$. Thus, upon noting that the second term vanishes if $z_i - 1 = (n_i - 1)/2$, θ_i is seen to be the conditional logit for an additional success when about half of the littermates exhibit a success already. Similarly, the log odds ratio for the responses between two littermates is equal to $2\delta_i$, confirming the association parameter interpretation of the δ-parameter. Finally, the marginal succes probability in a cluster of size n_i is clearly a (non-linear) function of n_i:

$$E\left(\frac{Z_i}{n_i}\right) = \frac{\sum_{z=0}^{n_i} z\binom{n_i}{z} \exp\{\theta_i z - \delta_i z(n_i - z)\}}{\sum_{z=0}^{n_i} n_i\binom{n_i}{z} \exp\{\theta_i z - \delta_i z(n_i - z)\}}.$$

Because this model is conditional in nature, this marginal quantity does not simplify in general. Nevertheless, this expectation can be easily calculated and plotted to explore the relationship between cluster size and response probability.

Although model (3.4) is very flexible and has several desirable properties, maximum likelihood estimation can become cumbersome due to the evaluation of the normalizing constant. Therefore we propose an alternative estimation method in Section 3.3.

3.3 Pseudo-likelihood: Definition and Asymptotic Properties

To introduce pseudo-likelihood formally, we will use the convenient general definition given by Arnold and Strauss (1991). Without loss of generality we can assume that the vector $\boldsymbol{Y}_i$ of binary outcomes for subject i (i=1, ... ,N) has constant dimension L. The extension to variable lengths of $\boldsymbol{Y}_i$ is straightforward.

3.3.1 Definition

Define S as the set of all 2^L-1 vectors of length L, consisting solely of zeros and ones, with each vector having at least one non zero entry. Denote by $\boldsymbol{y}_i^{(s)}$ the subvector of $\boldsymbol{y}_i$ corresponding to the components of s that are non zero. The associated joint density is $f_s(\boldsymbol{y}_i^{(s)};\boldsymbol{\Theta}_i)$. In order to define a pseudo-likelihood function, one chooses a set $\delta = \{\delta_s | s \in S\}$ of real numbers, with at least one non zero component. The log of the pseudo-likelihood is then defined as

$$p\ell = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \ln f_s(\boldsymbol{y}_i^{(s)};\boldsymbol{\Theta}_i). \tag{3.6}$$

In our development we will assume adequate regularity conditions to ensure that (3.6) can be maximized by solution of the pseudo-likelihood (score) equations, the latter obtained by differentiation of the logarithm of the pseudolikelihood and the setting of the derivative to zero.

The classical log-likelihood function is found by setting $\delta_s = 1$ if s is the vector consisting solely of ones, and 0 otherwise. A convenient pseudo-likelihood function for exponential family models such as described in Section 3.2.1, is found by replacing the joint density $f_{\boldsymbol{Y}}(\boldsymbol{y}_i;\boldsymbol{\Theta}_i)$ by the product of univariate "full" conditional densities $f(y_{ij}|\{y_{ij'}\}, j' \neq j;\boldsymbol{\Theta}_i)$ for $j = 1,\ldots,L$, obtained by conditioning each observed outcome on all others. This idea can be put into the framework (3.6) by choosing $\delta_{1_L} = L$ and $\delta_{s_j} = -1$ for $j = 1,\ldots,L$ where 1_L is a vector of ones and s_j consists of ones everywhere, except for the jth entry. For all other vectors s, δ_s equals zero. We refer to this particular choice as the *full conditional pseudo-likelihood function.* This pseudo-likelihood has the effect of replacing a joint mass function with a complicated normalizing constant by L univariate functions. Other types of pseudo-likelihood functions, that also fit into (3.6), will be considered in Chapter 6.

3.3.2 Consistency and Asymptotic Normality

Before stating the main asymptotic properties of the PL estimators, we first list the required regularity conditions.

A0 The densities $f_s(\boldsymbol{y}^{(s)};\boldsymbol{\Theta})$ are distinct for different values of the parameter $\boldsymbol{\Theta}$.

A1 The densities $f_s(\boldsymbol{y}^{(s)};\boldsymbol{\Theta})$ have common support, which does not depend on $\boldsymbol{\Theta}$.

A2 The parameter space Ω contains an open region ω of which the true parameter value $\boldsymbol{\Theta}_0$ is an interior point.

A3 ω is such that for all s, and almost all $\boldsymbol{y}^{(s)}$ in the support of $\boldsymbol{Y}^{(s)}$, the densities admit all third derivatives

$$\frac{\partial^3 f_s(\boldsymbol{y}^{(s)};\boldsymbol{\Theta})}{\partial\theta_j\partial\theta_k\partial\theta_\ell}$$

.

A4 The first and second logarithmic derivatives of f_s satisfy

$$E_{\boldsymbol{\Theta}}\left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)};\boldsymbol{\Theta})}{\partial\theta_k}\right) = 0, \qquad k = 1,\ldots,q,$$

and

$$0 < E_{\boldsymbol{\Theta}}\left(\frac{-\partial^2 \ln f_s(\boldsymbol{y}^{(s)};\boldsymbol{\Theta})}{\partial\theta_k\partial\theta_\ell}\right) < \infty, \qquad k,\ell = 1,\ldots,q.$$

A5 The matrix J, defined in (3.7) is positive definite.

A6 There exist functions M_{klr} such that

$$\sum_{s\in S}\delta_s E_{\boldsymbol{\Theta}}\left|\frac{\partial^3 \ln f_s(\boldsymbol{y}^{(s)};\boldsymbol{\Theta})}{\partial\theta_k\partial\theta_\ell\partial\theta_r}\right| < M_{k\ell r}(\boldsymbol{y})$$

for all $\boldsymbol{y}$ in the support of f and for all $\boldsymbol{\theta}\in\omega$ and $m_{k\ell r} = E_{\boldsymbol{\Theta}_0}(M_{k\ell r}(Y)) < \infty$.

Arnold and Strauss (1991) have shown consistency and asymptotic normality of the pseudo-likelihood estimator in the single parameter case. We will present the theorem for a vector valued parameter. Without loss of generality, we can assume $\boldsymbol{\Theta}$ is constant. Replacing it by $\boldsymbol{\Theta}_i$, and modelling it as a function of covariates is straightforward.

Theorem 3.3.1 (Consistency and Asymptotic Normality) *Let* $(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_N)$ *be iid with common density that depends on* $\boldsymbol{\Theta}_0$ *then under regularity conditions (A1)–(A6):*

1. *the pseudo-likelihood estimator* $\tilde{\boldsymbol{\Theta}}_N$*, defined as the maximizer of (3.6) converges in probability to* $\boldsymbol{\Theta}_0$.

2. $\sqrt{N}(\tilde{\boldsymbol{\Theta}}_N - \boldsymbol{\Theta}_0)$ *converges in distribution to* $N_p(\mathbf{0}, J(\boldsymbol{\Theta}_0)^{-1}K(\boldsymbol{\Theta}_0)J(\boldsymbol{\Theta}_0)^{-1})$ *with* $J(\boldsymbol{\Theta})$ *defined by*

$$J_{k\ell}(\boldsymbol{\Theta}) = -\sum_{s\in S} \delta_s E_{\boldsymbol{\Theta}} \left(\frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial\theta_k \partial\theta_\ell} \right) \tag{3.7}$$

and $K(\boldsymbol{\Theta})$ *by*

$$K_{k\ell}(\boldsymbol{\Theta}) = \sum_{s,t\in S} \delta_s \delta_t E_{\boldsymbol{\Theta}} \left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial\theta_k} \frac{\partial \ln f_t(\boldsymbol{y}^{(t)}; \boldsymbol{\Theta})}{\partial\theta_\ell} \right). \tag{3.8}$$

The proofs of consistency and asymptotic normality are based upon those presented by Lehmann (1983, p. 430–434) in the context of likelihood estimation.

Proof of Consistency

Consider the behaviour of the log pseudo-likelihood on a sphere Q_a with center $\boldsymbol{\Theta}_0$ and radius a. If it can be shown that for any sufficiently small a:

$$P(p\ell(\boldsymbol{\Theta}) < p\ell(\boldsymbol{\Theta}_0)) = 1 \qquad \text{for all } \boldsymbol{\Theta} \text{ on } Q_a, \tag{3.9}$$

then the $p\ell$ has a local maximum in the interior of Q_a. At this local maximum the pseudo-likelihood equations are satisfied. Hence, for any $a > 0$, the pseudo-likelihood equations have (with probability tending to one) a solution $\tilde{\boldsymbol{\Theta}}_N(a)$ within Q_a. To ensure the existence of a consistent root that does not depend on a, we can then define $\tilde{\boldsymbol{\Theta}}_N^*$ as the root closest to $\boldsymbol{\Theta}_0$.

Let us now prove (3.9). First, note that

$$\frac{1}{N} P_j(\boldsymbol{Y}) := \frac{1}{N} \sum_{i=1}^{N} \sum_{s\in S} \delta_s \frac{\partial}{\partial\theta_j} \ln(f_s(\boldsymbol{Y}_i^{(s)}; \boldsymbol{\Theta}_0)) \quad (j = 1, \ldots, q) \tag{3.10}$$

converges in probability to zero by (A4) and the weak law of large numbers and similarly that

$$\frac{1}{N} Q_{jk}(\boldsymbol{Y}) := \frac{1}{N} \sum_{i=1}^{N} \sum_{s\in S} \delta_s \frac{\partial^2}{\partial\theta_j \theta_k} \ln(f_s(\boldsymbol{Y}_i^{(s)}; \boldsymbol{\Theta}_0)) \tag{3.11}$$

converges in probability to $-J_{jk}(\boldsymbol{\Theta}_0)$ by (A4), (A5) and the weak law of large numbers. Next, using (A6), we can expand the log pseudo-likelihood around $\boldsymbol{\Theta}_0$

such that:

$$\begin{aligned}
&\tfrac{1}{N}p\ell(\boldsymbol{\Theta}) - \tfrac{1}{N}p\ell(\boldsymbol{\Theta}_0) \\
&= \tfrac{1}{N}\textstyle\sum_{j=1}^{q} P_j(y)(\theta_j - \theta_{0j}) \\
&+\tfrac{1}{2N}\textstyle\sum_{j=1}^{q}\sum_{k=1}^{q} Q_{jk}(y)(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \\
&+\tfrac{1}{6N}\textstyle\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q}(\theta_j - \theta_{0j})(\theta_k - \theta_{0k})(\theta_\ell - \theta_{0\ell})\sum_{i=1}^{N}\gamma_{jk\ell}(\boldsymbol{y}_i)M_{jk\ell}(\boldsymbol{y}_i),
\end{aligned}$$

with $0 \leq |\gamma_{jk\ell}(y)| \leq 1$. Further using (A6) and the weak law of large numbers we know that $1/N\sum_{i=1}^{N} M_{jk\ell}(\boldsymbol{Y}_i)$ converges in probability to $m_{jk\ell}$. In addition, if we define:

$$\begin{aligned}
S_1 &= \frac{1}{N}\sum_{j=1}^{q} P_j(y)(\theta_j - \theta_{0j}) \\
S_2 &= \frac{1}{2N}\sum_{j=1}^{q}\sum_{k=1}^{q} Q_{jk}(y)(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \\
S_3 &= \frac{1}{6N}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q}(\theta_j - \theta_{0j})(\theta_k - \theta_{0k})(\theta_\ell - \theta_{0\ell})\sum_{i=1}^{N}\gamma_{jk\ell}(\boldsymbol{y}_i)M_{jk\ell}(\boldsymbol{y}_i),
\end{aligned}$$

then using (3.10) and (3.11) it can be shown that there exists a $c > 0$ such that:

$$\max(S_1 + S_2 + S_3) < -ca^2 + (b+q)a^3$$

with probability tending to one and with b defined by:

$$b = \frac{1}{3}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} m_{jk\ell}.$$

This completes the proof since this is less than zero if $a < c/(b+q)$.

Lemma 3.3.1 *Suppose that* $(T_{1N}, \ldots, T_{qN})$ *converges in distribution to* $(T_1, \ldots, T_q)$, *where* T_{jN} $(j = 1, \ldots, q)$ *is defined by:*

$$T_{jN} = \sum_{k=1}^{q} A_{jkN}Y_{kN}.$$

Assume further that for each fixed j *and* k, A_{jkN} *converges in probability to* a_{jk} *for which the matrix* $A = (a_{jk})$ *is nonsingular and let* B *be* A^{-1}.

Then $(Y_{1N}, \ldots, Y_{qN})$ *converges in probability to* $(Y_1, \ldots, Y_q)$ *with* Y_j *defined by:*

$$Y_j = \sum_{k=1}^{q} b_{jk} T_k$$

.

A proof for this lemma can be found in Lehmann (1983, p. 432–433).

Proof of Asymptotic Normality

Denote

$$\begin{aligned} p\ell'_j(\boldsymbol{\Theta}) &= \frac{\partial}{\partial\theta_j} p\ell(\boldsymbol{\Theta}) \quad (j = 1, \ldots, q), \\ p\ell''_{jk}(\boldsymbol{\Theta}) &= \frac{\partial^2}{\partial\theta_j \partial\theta_k} p\ell(\boldsymbol{\Theta}) \quad (k = 1, \ldots, q), \\ p\ell'''_{jk\ell}(\boldsymbol{\Theta}) &= \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_\ell} p\ell(\boldsymbol{\Theta}) \quad (\ell = 1, \ldots, q). \end{aligned}$$

Since we already know that, with probability tending to one, there exists a consistent solution $\tilde{\boldsymbol{\Theta}}_N$ of the pseudo-likelihood equations, we can expand:

$$-\frac{1}{\sqrt{N}} p\ell'_j(\boldsymbol{\Theta}_0) = \sqrt{N} \sum_{k=1}^{q} (\tilde{\theta}_{Nk} - \theta_{0k}) \left(\frac{1}{N} p\ell''_{jk}(\boldsymbol{\Theta}_0) + \frac{1}{2N} \sum_{\ell=1}^{q} (\tilde{\theta}_{N\ell} - \theta_{0\ell}) p l'''_{jk\ell}(\boldsymbol{\Theta}^*) \right)$$

with $\boldsymbol{\Theta}^*$ on the line segment connecting $\tilde{\boldsymbol{\Theta}}_N$ and $\boldsymbol{\Theta}_0$. This expansion can be rewritten as:

$$T_{jN} = \sum_{k=1}^{q} A_{jkN} Y_{kN},$$

with

$$\begin{aligned} T_{jN} &= \frac{-1}{\sqrt{N}} p\ell'_j(\boldsymbol{\Theta}_0), \\ Y_{kN} &= \sqrt{N}(\tilde{\boldsymbol{\theta}}_{Nk} - \theta_{0k}), \\ A_{jkN} &= \frac{1}{N} p\ell''_{jk}(\boldsymbol{\Theta}_0) + \frac{1}{2N} \sum_{\ell=1}^{q} (\tilde{\theta}_{N\ell} - \theta_{0\ell}) p\ell'''_{jk\ell}(\boldsymbol{\Theta}^*). \end{aligned}$$

Because of (A4) and the multivariate central limit theorem we know that $(T_{1N}, \ldots, T_{qN})$ converges in distribution to $(T_1, \ldots, T_q)$ which follows a q-dimensional multivariate normal distribution with mean vector 0 and variance-covariance matrix $K(\boldsymbol{\Theta}_0)$. Using the weak law of large numbers and (A6) one can show that A_{jkN} converges in

probability to $-J_{jk}(\boldsymbol{\Theta}_0)$. By lemma 3.3.1 we know that $(Y_{1N}, \ldots, Y_{qN})$ converges in probability to $(Y_1, \ldots, Y_q)$ where

$$Y_j = -\sum_{k=1}^{q} J_{jk}^{-1}(\boldsymbol{\Theta}_0) T_k, \qquad j = 1, \ldots, q.$$

Hence, the distribution of $\boldsymbol{Y}$ is multivariate normal with mean zero and covariance matrix given by $J^{-1}(\boldsymbol{\Theta}_0)K(\boldsymbol{\Theta}_0)J^{-1}(\boldsymbol{\Theta}_0)$, completing the proof.

Similar in spirit to generalized estimating equations (Liang and Zeger 1986), the asymptotic normality result provides an easy way to estimate consistently the asymptotic covariance matrix. Indeed, the matrix J is found from evaluating the second derivative of the log PL function at the PL estimate. The expectation in K can be replaced by the cross-products of the observed scores. We will refer to J^{-1} as the model based variance estimator (which should not be used since it overestimates the precision), to K as the empirical correction, and to $J^{-1}KJ^{-1}$ as the empirically corrected variance estimator. In the context of generalized estimating equations, this is also known as the sandwich estimator.

As discussed by Arnold and Strauss (1991), the Cramèr-Rao inequality implies that $J^{-1}KJ^{-1}$ is greater than the inverse of I (the Fisher information matrix for the maximum likelihood case), in the sense that $J^{-1}KJ^{-1} - I^{-1}$ is positive semi-definite. Strict inequality holds if the PL estimator fails to be a function of a minimal sufficient statistic. Therefore, a PL estimator is always less efficient than a ML estimator.

3.4 Application to the Thélot Model

To clarify the pseudo-likelihood concepts, consider the special case of $M = 2$ in (3.2). This model was studied in detail by Thélot (1985). The log likelihood contribution for the ith cluster has the form:

$$\ell_i = \ln\left(\frac{e^{\theta_{i1}y_{i1} + \theta_{i2}y_{i2} + \omega_i y_{i1} y_{i2}}}{1 + e^{\theta_{i1}} + e^{\theta_{i2}} + e^{\theta_{i1}+\theta_{i2}+\omega_i}}\right).$$

Using definition (3.6), the $p\ell$ contribution on the other hand can be calculated from

$$\begin{aligned} f(y_{i1}|y_{i2}) &= \frac{e^{(\theta_{i1}+\omega_i y_{i2})y_{i1}}}{1 + e^{\theta_{i1}+\omega_i y_{i2}}} \\ f(y_{i2}|y_{i1}) &= \frac{e^{(\theta_{i2}+\omega_i y_{i1})y_{i2}}}{1 + e^{\theta_{i2}+\omega_i y_{i1}}} \end{aligned}$$

and equals

$$p\ell_i = \ln\left(f(y_{i1}|y_{i2})f(y_{i2}|y_{i1})\right) = \ln\left(\frac{e^{(\theta_{i1}+\omega_i y_{i2})y_{i1}}}{1+e^{\theta_{i1}+\omega_i y_{i2}}}\frac{e^{(\theta_{i2}+\omega_i y_{i1})y_{i2}}}{1+e^{\theta_{i2}+\omega_i y_{i1}}}\right).$$

Thus, the use of pseudo-likelihood translates a non-standard bivariate problem into one that can be tackled with standard logistic regression software. As an illustration, we consider the score equations for both ML and the full conditional PL estimation in the Thélot case. For simplicity, the cluster index i is kept fixed and dropped from notation. Further, we define $\pi_{jk} = P(y_1 = j, y_2 = k)$ $(j,k = 0,1)$, such that

$$\begin{aligned}
\pi_{11} &= \frac{e^{\theta_1+\theta_2+\omega}}{1+e^{\theta_1}+e^{\theta_2}+e^{\theta_1+\theta_2+\omega}},\\
\pi_{1+} &= \frac{e^{\theta_1}+e^{\theta_1+\theta_2+\omega}}{1+e^{\theta_1}+e^{\theta_2}+e^{\theta_1+\theta_2+\omega}},\\
\pi_{+1} &= \frac{e^{\theta_2}+e^{\theta_1+\theta_2+\omega}}{1+e^{\theta_1}+e^{\theta_2}+e^{\theta_1+\theta_2+\omega}}.
\end{aligned}$$

Contributions to the score equations for ML can then be written as:

$$\frac{\partial\ell}{\partial\theta_1} = y_1 - \pi_{1+}, \qquad \frac{\partial\ell}{\partial\theta_2} = y_2 - \pi_{+1}, \qquad \frac{\partial\ell}{\partial\omega} = y_1y_2 - \pi_{11}.$$

Contributions to the first derivatives of the $p\ell$ function are:

$$\frac{\partial p\ell}{\partial\theta_1} = y_1 - \mu_1(y_2), \qquad \frac{\partial p\ell}{\partial\theta_2} = y_2 - \mu_2(y_1), \qquad \frac{\partial p\ell}{\partial\omega} = y_2y_1 - \mu_1(y_2) + y_1y_2 - \mu_2(y_1),$$

with

$$\mu_1(y_2) = \frac{e^{(\theta_1 \,|\, \omega y_2)}}{1+e^{\theta_1+\omega y_2}}, \qquad \mu_2(y_1) = \frac{e^{(\theta_2 \,|\, \omega y_1)}}{1+e^{\theta_2+\omega y_1}}. \tag{3.12}$$

Clearly, the above equations for the main effect parameters are similar in form to the corresponding score equations for ML. Moreover, they are identical in the independence case. This model will be explored further in Section 3.7.1.

3.5 Application to Clustered Outcomes

We will now apply the pseudo-likelihood ideas to the specific context of exchangeable clustered binary outcomes.

In order to define a pseudo-likelihood function based on model (3.4) we will consider the n_i conditional probabilities of observing the outcome for littermate k, given the outcomes for the other $n_i - 1$ littermates. Due to the exchangeable nature of the littermates, there are only two types of contributions: (1) the conditional probability of an additional success, given there are $z_i - 1$ successes and $n_i - z_i$ failures (this contribution occurs with multiplicity z_i):

$$p_{is} = \frac{\exp\{\theta_i - \delta_i(n_i - 2z_i + 1)\}}{1 + \exp\{\theta_i - \delta_i(n_i - 2z_i + 1)\}}.$$

and (2) the conditional probability of an additional failure, given there are z_i successes and $n_i - z_i - 1$ failures (with multiplicity $n_i - z_i$):

$$p_{if} = \frac{\exp\{-\theta_i + \delta_i(n_i - 2z_i - 1)\}}{1 + \exp\{-\theta_i + \delta_i(n_i - 2z_i - 1)\}}.$$

The log PL contribution for cluster i can be expressed as $p\ell_i = z_i \ln p_{is} + (n_i - z_i) \ln p_{if}$. The contribution of cluster i to the pseudo-likelihood score vector is of the form

$$\begin{pmatrix} z_i(1 - p_{is}) - (n_i - z_i)(1 - p_{if}) \\ -z_i(n_i - 2z_i + 1)(1 - p_{is}) + (n_i - z_i)(n_i - 2z_i - 1)(1 - p_{if}) \end{pmatrix}.$$

Note that, if $\delta_i \equiv 0$, then $p_{is} \equiv 1 - p_{if}$ and the first component of the score vector is a sum of terms $z_i - n_i p_{is}$, i.e. standard logistic regression follows. In the general case, we have to account for the association, but this non-standard system of equations can be solved using logistic regression software as follows. Represent the contribution for cluster i by two separate records, with repetition counts z_i for the "success case" and $n_i - z_i$ for the "failure case" respectively. All interaction covariates need to be multiplied by $-(n_i - 2z_i + 1)$ in the success case and $-(n_i - 2z_i - 1)$ in the failure case.

3.6 Examples

To illustrate our findings, we apply the proposed method to the five developmental toxicity studies in mice (DEHP, EG, TGDM, DYME, THEO) conducted by the Research Triangle Institute under contract to the National Toxicology Program (NTP). These studies were described in Chapter 2.

Table 3.1: *NTP Studies: Maximum Likelihood Estimates (model based standard errors; empirically corrected standard errors) of Univariate Outcomes.*

Study	Par.	External	Visceral	Skeletal	Collapsed
DEHP	β_0	-2.81 (0.58;0.52)	-2.39 (0.50;0.52)	-2.79 (0.58;0.77)	-2.04 (0.35;0.42)
	β_d	3.07 (0.65;0.62)	2.45 (0.55;0.60)	2.91 (0.63;0.82)	2.98 (0.51;0.66)
	β_a	0.18 (0.04;0.04)	0.18 (0.04;0.04)	0.17 (0.04;0.05)	0.16 (0.03;0.03)
EG	β_0	-3.01 (0.79;1.01)	-5.09 (1.55;1.51)	-0.84 (0.17;0.18)	-0.81 (0.16;0.16)
	β_d	2.25 (0.68;0.85)	3.76 (1.34;1.20)	0.98 (0.20;0.20)	0.97 (0.20;0.20)
	β_a	0.25 (0.05;0.06)	0.23 (0.09;0.09)	0.20 (0.02;0.02)	0.20 (0.02;0.02)
TGDM	β_0	-6.19 (1.62;1.48)		-7.43 (2.00;1.72)	-5.24 (1.03;1.03)
	β_d	3.79 (1.10;1.31)		6.23 (1.88;1.67)	4.47 (0.94;1.01)
	β_a	0.08 (0.12;0.11)		0.16 (0.07;0.05)	0.17 (0.05;0.04)
DYME	β_0	-5.78 (1.13;1.23)	-3.32 (0.98;0.89)	-1.62 (0.35;0.48)	-2.90 (0.43;0.51)
	β_d	6.25 (1.25;1.41)	2.88 (0.93;0.83)	2.45 (0.51;0.82)	5.08 (0.74;0.96)
	β_a	0.09 (0.06;0.06)	0.29 (0.05;0.05)	0.25 (0.03;0.03)	0.19 (0.03;0.03)
THEO	β_0	-4.82 (1.52;1.55)	-10.50 (4.84;3.66)	-2.80 (2.79;1.00)	-4.14 (1.26;1.37)
	β_d	1.75 (0.94;1.06)	4.31 (3.56;2.05)	2.19 (2.92;0.96)	1.97 (0.88;0.93)
	β_a	0.07 (0.13;0.13)	-0.10 (0.36;0.18)	0.81 (0.32;0.09)	0.13 (0.11;0.12)

Table 3.2: *NTP Studies: Pseudo-likelihood Estimates (standard errors) of Univariate Outcomes.*

Study	Par.	External	Visceral	Skeletal	Collapsed
DEHP	β_0	-2.85 (0.53)	-2.30 (0.50)	-2.41 (0.73)	-1.80 (0.35)
	β_d	3.24 (0.60)	2.55 (0.53)	2.52 (0.81)	2.95 (0.56)
	β_a	0.18 (0.04)	0.20 (0.04)	0.21 (0.05)	0.20 (0.03)
EG	β_0	-2.61 (0.88)	-5.10 (1.55)	-1.18 (0.14)	-1.11 (0.14)
	β_d	2.14 (0.71)	3.79 (1.18)	1.43 (0.19)	1.41 (0.19)
	β_a	0.30 (0.06)	0.23 (0.10)	0.21 (0.01)	0.21 (0.01)
TGDM	β_0	-4.75 (1.06)		-7.10 (1.70)	-4.69 (0.97)
	β_d	3.52 (1.24)		6.10 (1.65)	4.13 (0.99)
	β_a	0.22 (0.07)		0.19 (0.06)	0.22 (0.03)
DYME	β_0	-5.04 (0.94)	-3.34 (0.99)	-2.20 (0.27)	-3.08 (0.47)
	β_d	5.52 (1.01)	2.91 (0.91)	3.22 (0.49)	5.20 (0.97)
	β_a	0.13 (0.05)	0.29 (0.06)	0.25 (0.02)	0.19 (0.02)
THEO	β_0	-3.51 (1.26)	-10.58 (3.66)	-4.33 (1.34)	-3.36 (1.08)
	β_d	1.65 (1.07)	4.30 (2.04)	4.69 (1.72)	1.92 (0.94)
	β_a	0.20 (0.12)	-0.11 (0.18)	0.84 (0.14)	0.22 (0.10)

We fitted model (3.4) to 4 outcomes in each of the 5 datasets: external, visceral, and skeletal malformation, as well as a collapsed outcome, defined to be 1 if any malformation occured and -1 otherwise. Parameters were estimated by both maximum likelihood (Table 3.1) and pseudo-likelihood (Table 3.2). The empirically corrected standard errors are commonly referred to as "robust" standard errors and introduced by Liang and Zeger (1986). The fitting procedure has been implemented in GAUSS. The natural parameters were modelled as follows: $\theta_i = \beta_0 + \beta_d d_i$ where d_i is the dose level applied to the ith cluster, and $\delta_i = \beta_a$, i.e. a constant association model.

An attractive feature of the proposed approach is that the parameters can also be obtained using standard and readily available software, such as SAS PROC LOGISTIC or SAS PROC GENMOD. As an illustration, the parameters for the external outcome in the DEHP study were also determined with SAS PROC LOGISTIC. An implementation and selected output is presented in Figures 3.1 and 3.2. Each cluster is represented by a two-line record. The first line corresponds with the "success" case so that the variable ASSOC represents $-(n_i - 2z_i + 1)$; the second line corresponds with the "failure" case so that ASSOC represents $-(n_i - 2z_i - 1)$.

While the estimates are identical to those obtained in Table 3.2, the standard errors are incorrect since they are based on the assumption of independence. To obtain a correct estimate of the variability, a short macro could be written.

Since visceral malformations are very infrequent with TGDM (only one malformation observed) a fit could not be obtained with either estimation technique.

The methods can be compared based on the parameter estimates, their standard errors (model based likelihood, empirically corrected likelihood, and pseudo-likelihood), or a combination of both (e.g. the Z statistic, defined as the ratio of estimate and standard error). Obviously, the development of methods to assess the fit of the proposed methods is necessary. However, classical tools cannot be used within the pseudo-likelihood framework without modification. Of course, one can always assess the fit by fitting an extended model and testing whether the additional parameters are significant. The extension of flexible tools such as likelihood ratio and score tests to the PL framework has been proposed by Geys, Molenberghs and Ryan (1999) and will be described in Chapter 4.

ML and PL dose parameter estimates agree fairly closely, except for the EG outcomes skeletal and collapsed and more dramatically for THEO skeletal. No method

```
data pseudo;
input success failure dose assoc;
total=success+failure;
cards;
    0.0000     0.0000     0.0000   -10.0000
    0.0000     9.0000     0.0000    -8.0000
    0.0000     0.0000     0.1667   -11.0000
    0.0000    10.0000     0.1667    -9.0000
    1.0000     0.0000     0.3333    -6.0000
    0.0000     6.0000     0.3333    -4.0000

    ...

    2.0000     0.0000     0.6667   -10.0000
    0.0000    11.0000     0.6667    -8.0000
;
run;

proc logistic data=pseudo;
model success/total = dose assoc;
run;
```

Figure 3.1: *DEHP Study: Implementation using the SAS procedure PROC LOGISTIC.*

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	-2.8520	0.5621	25.7456	0.0001	.
DOSE	1	3.2369	0.6501	24.7921	0.0001	0.474261
ASSOC	1	0.1833	0.0429	18.2737	0.0001	0.393847

Figure 3.2: *DEHP Study: Selected Output of the SAS procedure PROC LOGISTIC.*

systematically leads to larger parameter estimates (each one yields the largest value in about half of the cases).

Rather than comparing estimated standard errors directly, one could also consider the derived Z statistics (not shown) and their associated significance levels. The only non-significant dose effects are found for the THEO dataset: external is non-significant, independent of the method; visceral and skeletal are non-significant with the model based likelihood version only. Pairwise comparisons of the test statistics reveal again that no procedure systematically yields larger values. Indeed, in all three comparisons, the magnitude of one statistic is larger than the other in approximately 50% of the cases.

These results are promising because a loss of efficiency of pseudo-likelihood versus maximum likelihood could be anticipated. However, even though in Section 3.7 it will be shown that the asymptotic relative efficiency (ARE) is in general strictly less than 1 (except for saturated models), the data analysis suggests that the efficiency loss is moderate. To explore the extent to which this conjecture can be generalized, we calculate asymptotic and small sample relative efficiency (Geys, Molenberghs and Ryan 1997).

3.7 Asymptotic Relative Efficiency of Pseudo-likelihood versus Maximum Likelihood

3.7.1 Asymptotic Relative Efficiency for the Thélot Model

The price for computational ease usually consists of some efficiency loss. To illustrate this statement, we will consider two simple forms of the Thélot model, both without covariates (i.e. the sample comprises a single two by two table).

In the first case, all three parameters θ_1, θ_2, and ω are estimated. Pseudo-likelihood is then as efficient as maximum likelihood, in the sense that $I^{-1} = J^{-1}KJ^{-1}$. Indeed, consider the contributions to the expected information for both ML and PL. For the likelihood, this contribution is

$$I = \begin{pmatrix} \pi_{1+}(1-\pi_{1+}) & \pi_{11}-\pi_{1+}\pi_{+1} & \pi_{11}(1-\pi_{1+}) \\ \pi_{11}-\pi_{1+}\pi_{+1} & \pi_{+1}(1-\pi_{+1}) & \pi_{11}(1-\pi_{+1}) \\ \pi_{11}(1-\pi_{1+}) & \pi_{11}(1-\pi_{+1}) & \pi_{11}(1-\pi_{11}) \end{pmatrix}.$$

The negative second derivative matrix of the log PL is given by

$$\mathcal{J} = \begin{pmatrix} \mu_1(y_2)\{1-\mu_1(y_2)\} & 0 & y_2\mu_1(y_2)\{1-\mu_1(y_2)\} \\ 0 & \mu_2(y_1)\{1-\mu_2(y_1)\} & y_1\mu_2(y_1)\{1-\mu_2(y_1)\} \\ y_2\mu_1(y_2)\{1-\mu_1(y_2)\} & y_1\mu_2(y_1)\{1-\mu_2(y_1)\} & y_2^2\mu_1(y_2)\{1-\mu_1(y_2)\} \\ & & +y_1^2\mu_2(y_1)\{1-\mu_2(y_1)\} \end{pmatrix},$$

with expected value

$$J = \begin{pmatrix} R_1+R_0 & 0 & R_1 \\ 0 & S_1+S_0 & S_1 \\ R_1 & S_1 & R_1+S_1 \end{pmatrix},$$

where

$$\begin{aligned} R_1 &= \pi_{+1}\mu_1(1)(1-\mu_1(1)), \\ R_0 &= (1-\pi_{+1})\mu_1(0)(1-\mu_1(0)), \\ S_1 &= \pi_{1+}\mu_2(1)(1-\mu_2(1)), \\ S_0 &= (1-\pi_{1+})\mu_2(0)(1-\mu_2(0)). \end{aligned}$$

The entries of K are

$$\begin{aligned}
k_{11} &= \pi_{1+} - 2\pi_{10}\mu_1(0) + \pi_{11}\mu_1(1) + (1-\pi_{+1})\mu_1(0)^2 + \pi_{+1}\mu_1(1)^2,\\
k_{12} &= \pi_{11} - \pi_{1+}\mu_2(1) - \pi_{+1}\mu_1(1)\\
&\quad +\pi_{11}\mu_1(1)\mu_2(1) + \pi_{10}\mu_1(0)\mu_2(1) + \pi_{01}\mu_1(1)\mu_2(0) + \pi_{00}\mu_1(0)\mu_2(0),\\
k_{22} &= \pi_{+1} - 2\pi_{01}\mu_2(0) + \pi_{11}\mu_2(1) + (1-\pi_{1+})\mu_2(0)^2 + \pi_{1+}\mu_2(1)^2,\\
k_{13} &= 2\pi_{11} - 3\pi_{11}\mu_1(1) + \pi_{+1}\mu_1(1)^2 - \pi_{1+}\mu_2(1) + \pi_{10}\mu_1(0)\mu_2(1) + \pi_{11}\mu_1(1)\mu_2(1),\\
k_{23} &= 2\pi_{11} - 3\pi_{11}\mu_2(1) + \pi_{1+}\mu_2(1)^2 - \pi_{+1}\mu_1(1) + \pi_{01}\mu_1(1)\mu_2(0) + \pi_{11}\mu_1(1)\mu_2(1),\\
k_{33} &= 2\pi_{11}\{1-\mu_1(1)-\mu_2(1)+1-\mu_1(1)1-\mu_2(1)\} + \pi_{+1}\mu_1(1)^2 + \pi_{1+}\mu_2(1)^2.
\end{aligned}$$

Now, straightforward but tedious matrix manipulations establish the desired equality.

In the second case, the true value of both main effect parameters is assumed to be known (reduced Thélot model). In order to obtain a formula for the ARE of the remaining association parameter, all matrices derived in the first case are replaced by their (3,3) entry for some choices of main effect parameters. The ML and PL variances are then respectively given by:

$$\frac{1}{\pi_{11}(1-\pi_{11})},$$

and

$$\frac{2\pi_{11} - 2\pi_{11}\mu_1(1) + \pi_{1+}\mu_2(1)^2 - 2\pi_{11}\mu_2(1) + \pi_{+1}\mu_1(1)^2 + 2\pi_{11}\{1-\mu_1(1)\}\{1-\mu_2(1)\}}{\left[\mu_1(1)\{1-\mu_1(1)\}\pi_{+1} + \mu_2(1)\{1-\mu_2(1)\}\pi_{1+}\right]^2}.$$

Applying some algebra to these expressions, the ARE is found to be

$$\frac{\{2\pi_{10}\pi_{01} + \pi_{11}(\pi_{10}+\pi_{01})\}^2}{(1-\pi_{11})\left\{\pi_{11}^2 + \pi_{11}(\pi_{10}+\pi_{01}) + \pi_{10}\pi_{01}\right\}\left\{4\pi_{10}\pi_{01} + \pi_{11}(\pi_{10}+\pi_{01})\right\}}.$$

The condition for ARE≤ 1 implies

$$\pi_{11}\left\{\pi_{10}\pi_{01}(\pi_{1+}-\pi_{+1})^2 + \pi_{00}\left(\pi_{10}\pi_{+1}^2 + \pi_{01}\pi_{1+}^2\right)\right\} \geq 0.$$

Clearly, this condition is always satisfied. Equality holds solely in trivial boundary cases, when one or more cell probabilities equal zero. It is interesting to observe that this holds even in the independence case, i.e. when ω is estimated 0. Figure 3.3 shows the ARE of the association parameter, in that case, as a function of the first and second marginal probabilities: π_{1+} and π_{+1}.

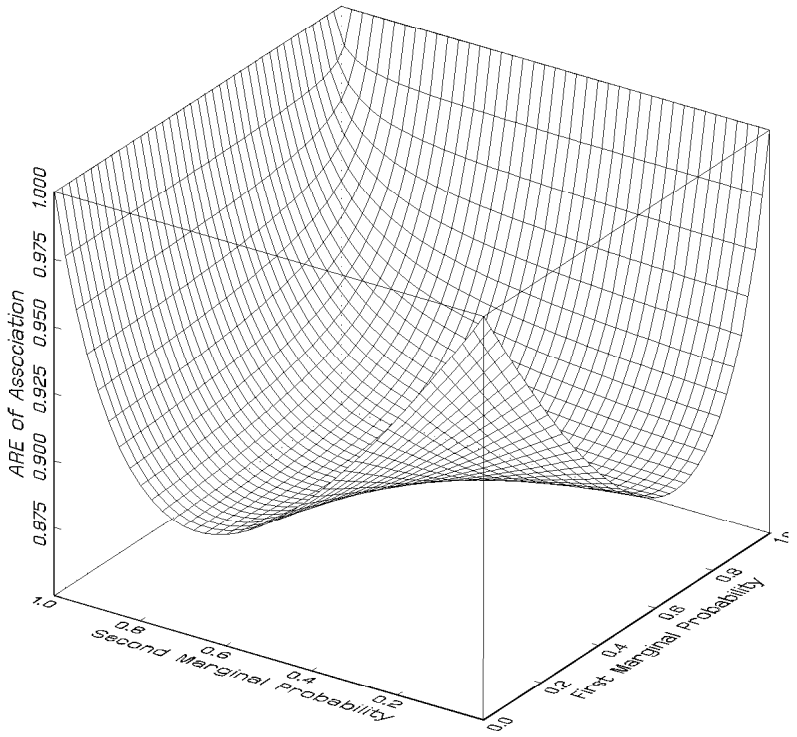

Figure 3.3: *Asymptotic Relative Efficiency of the Association in the Reduced Thélot Model (Independence Case).*

3.7.2 Asymptotic Relative Efficiency for the Saturated Model

The observation from Section 3.7.1, that the ARE in the full Thélot model equals 1 holds more generally. In fact, it holds for all saturated models, i.e. models of the form (3.1) without covariates and where all subvectors of $\boldsymbol{W}$ are included. The ARE for non-saturated models will be discussed in Section 3.7.3.

Consider the PL contribution for a single cluster, consisting of the product of all univariate conditional densities. Like in Section 3.2.1 the cluster index i is kept fixed and dropped from notation:

$$PL = \prod_{j=1}^{M} f_j(y_j|\boldsymbol{y}_{(j)}),$$

where $\boldsymbol{y}_{(j)}$ indicates omission of the jth component. Extending the notation introduced in (3.12), the logit of the conditional probability that y_j equals 1 given all others can be written as:

$$\begin{aligned}&\text{logit}\left(\mu_j(y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_M)\right) = \theta_j + \textstyle\sum_{k\neq j}\omega_{jk}y_k \\ &+ \textstyle\sum_{k<k';k,k'\neq j}\omega_{jkk'}y_k y_{k'} + \cdots + \omega_{12\ldots M}y_1\ldots y_{j-1}y_{j+1}\ldots y_M.\end{aligned} \tag{3.13}$$

In short we denote the logit in (3.13) by logit μ_j. In general, the pseudo-likelihood score contributions of the rth $(r = 1,\ldots,M)$ association parameter for a single subject can then be derived as:

$$\sum_{\ell=1}^{r}(y_{k_\ell}-\mu_{k_\ell})y_{k_1}\ldots y_{k_{\ell-1}}y_{k_{\ell+1}}\ldots y_{k_r} \qquad (1\leq k_1<k_2<\cdots<k_r\leq M). \tag{3.14}$$

For the main effect and the pairwise interactions, these contributions reduce to

$$\begin{aligned} y_j-\mu_j, &\qquad 1\leq j\leq M,\\ (y_j-\mu_j)y_k+(y_k-\mu_k)y_j, &\qquad 1\leq j<k\leq M.\end{aligned}$$

We will now show that the maximum likelihood estimator satisfies (3.14).

Organise the data into an M dimensional contingency table with cell counts

$$z_{j_1\ldots j_M}(j_p=0,1;p=1,\ldots,M). \tag{3.15}$$

Obviously, it may be more convenient to introduce an alternative notation for these cell counts. Rather than giving a sequence of M zeros and ones like in (3.15), we can present the subscripts for which $j_p - 1$. Thus, $z_{\cdot}$ is the number of individuals with failures on all variables, z_j refers to those having a success on outcome j and a failure on all others, $z_{j_1 j_2}$ refers to those having successes on both outcomes j_1 and j_2 and a failure on all others, etc. With straightforward notation, the maximum likelihood estimates for the corresponding cell probabilities are easily obtained:

$$\begin{aligned}
\hat{\pi}_{\cdot} &= \frac{z_{\cdot}}{N}, \\
\hat{\pi}_{j} &= \frac{z_{j}}{N}, && (3.16) \\
\hat{\pi}_{j_1 j_2} &= \frac{z_{j_1 j_2}}{N}, && (3.17) \\
&\vdots \\
\hat{\pi}_{j_1 \ldots j_p} &= \frac{z_{j_1 \ldots j_p}}{N}.
\end{aligned}$$

Now, simple relations exist between these cell probabilities and the natural parameters. For example:

$$\begin{aligned}
\hat{\pi}_{\cdot} &= \frac{1}{A(\widehat{\boldsymbol{\Theta}})}, \\
\hat{\pi}_{j} &= \frac{e^{\hat{\theta}_j}}{A(\widehat{\boldsymbol{\Theta}})}, && (3.18) \\
\hat{\pi}_{j_1 j_2} &= \frac{e^{\hat{\theta}_{j_1} + \hat{\theta}_{j_2} + \hat{\omega}_{j_1 j_2}}}{A(\widehat{\boldsymbol{\Theta}})}, && (3.19) \\
&\vdots
\end{aligned}$$

with $A(\widehat{\boldsymbol{\Theta}}) = 1 + e^{\hat{\theta}_1} + \cdots + e^{\hat{\theta}_1 + \cdots + \hat{\omega}_{12\ldots M}}$. Combining for example (3.16) and (3.18) we can rewrite $e^{\hat{\theta}_j} = z_j / z_{\cdot}$, which is the classical relationship between the main effect parameters and the conditional odds associated with outcome j, given failures on all others. Similarly, it follows from (3.17) and (3.19) that $e^{\hat{\omega}_{j_1 j_2}} = (z_{j_1 j_2} z_{\cdot}) / (z_{j_1} z_{j_2})$.

Using the notation introduced above, the PL score contribution for the main

effect θ_j, combined over all subjects can be written as:

$$\sum_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)} z_{t_1\ldots t_{j-1}1t_{j+1}\ldots t_M}\{1-\mu_j(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)\}$$
$$+\sum_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)} z_{t_1\ldots t_{j-1}0t_{j+1}\ldots t_M}\{-\mu_j(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)\} = 0.$$

where the summation is over all $M-1$ vectors (no jth component) of zeros and ones. Rewriting this equation as

$$\textstyle\sum_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)} z_{t_1\ldots t_{j-1}1t_{j+1}\ldots t_M} -$$
$$\textstyle\sum_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)} z_{t_1\ldots t_{j-1}+t_{j+1}\ldots t_M}\mu_j(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M) = 0,$$

it is easily seen that the MLE satisfies this equation, since on the one hand $\mu_j(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M)$ is the probability of observing a success on outcome j, given the value of the other outcomes, and on the other hand its MLE is given by

$$\hat{\mu}_j(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_M) = \frac{z_{t_1\ldots t_{j-1}1t_{j+1}\ldots t_M}}{z_{t_1\ldots t_{j-1}+t_{j+1}\ldots t_M}}.$$

Similar calculations can be carried out for the equations pertaining to the association parameters. This shows that the maximum likelihood estimator and the pseudo-likelihood estimator coincide in this case. A trivial consequence of this result is that ARE≡1.

3.7.3 Asymptotic Relative Efficiency for Clustered Outcomes

Although explicit formulae for the ARE were derived for unclustered outcomes in previous sections, similar expressions in the clustered case are difficult to obtain. Therefore, to study the ARE, we will follow the recommendations of Rotnizky and Wypij (1994). In order to compute asymptotic bias or efficiency, an artificial sample can be constructed, where each possible realization is weighted according to its true probability. 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, and (3) $f(z_i|n_i,d_i)$, the actual model probabilities.

Table 3.3: *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		

Throughout we assume that there are 4 dose groups, with one control ($d_i = 0$) and three exposed groups ($d_i = 0.25, 0.5, 1.0$). The number n_i of viable fetuses per cluster is chosen at random, using a local linear smoothed version of the relative frequency distribution given in Table 1 of Kupper et al. (1986) (which is considered representative of that encountered in actual experimental situations). Least squares cross-validation has been used to choose the bandwidth. The smoothed frequencies are presented in Table 3.3. Guided by the analysis of the examples, we identified three values for each of the three parameters: $\beta_0 = -5, -3, 0$, $\beta_d = 0, 3, 5$, and $\beta_a = 0, 0.15, 0.30$, with notation as defined in Section 3.6. The full grid of 27 parameter combinations has been explored. Results are displayed in Table 3.4.

No AREs are exactly equal to one, although some appear to be due to rounding. The AREs are very high for the lowest background rate ($\beta_0 = -5$) and they are almost all above 90% for the medium background rate ($\beta_0 = -3$). We can notice the nonmonotone relationship of the ARE with β_d and β_a. While still high in some areas of the (β_d, β_a) space for $\beta_0 = 0$, a dramatic decrease is observed when β_a increases and/or β_d decreases. PL performs very poorly when there is no dose effect together with a reasonably high association. Unless background malformation

Table 3.4: *Simulation Results: Asymptotic Relative Efficiencies of Pseudo-likelihood versus Maximum Likelihood.*

		β_a		
β_0	β_d	0.00	0.15	0.30
-5	0	1.000	1.000	1.000
	3	0.982	0.999	1.000
	5	0.940	0.978	0.966
-3	0	1.000	1.000	1.000
	3	0.938	0.938	0.897
	5	0.921	0.959	0.907
0	0	1.000	0.725	0.055
	3	0.958	0.895	0.792
	5	0.943	0.928	0.890

probabilities or dose effects are extreme, large associations diminish the contribution to the information of a full conditional. As a limiting case it can even be reduced to zero when the association parameter approaches infinity. This phenomenon is further illustrated in Figure 3.4. The parameter estimates found from the data analysis are all in regions of the parameter space with a high ARE.

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

3.8 Small Sample Relative Efficiency of Pseudo-likelihood versus Maximum Likelihood

The same 27 parameter combinations of the previous sections are investigated, for samples of size 30. For each setting, 500 simulations were run. The estimated covariance matrices were kept and averaged at the end of the run. The relative efficiencies for the dose effect parameters are displayed in Table 3.5. For the maximum likelihood procedure, both the purely model based as well as the empirically

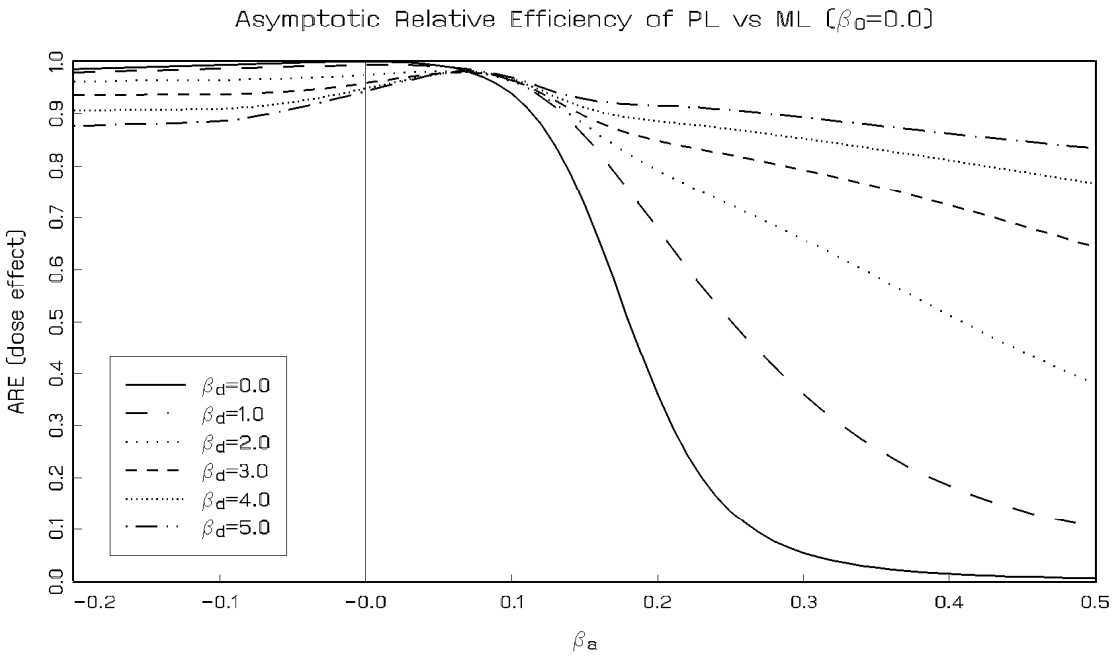

Figure 3.4: *Simulation Results: Asymptotic Relative Efficiency of Pseudo-likelihood versus Maximum Likelihood for the Dose Effect Parameter in the Clustered Data Model.*

Table 3.5: *Simulation Results: Small Sample Relative Efficiencies (500 replications) of Pseudo-likelihood versus Maximum Likelihood.*

		β_a								
		0.00			0.15			0.30		
β_0	β_d	model	emp.	runs	model	emp.	runs	model	emp.	runs
-5	0	1.073	0.973	500	2.592	0.990	319	8.919	0.945	93
	3	0.976	0.955	500	1.086	0.995	485	1.613	1.017	254
	5	0.959	0.918	500	1.003	0.982	499	1.123	0.986	411
-3	0	1.027	0.994	500	1.116	1.018	500	1.966	1.009	425
	3	0.929	0.915	500	0.970	0.938	500	1.005	0.921	500
	5	0.931	0.905	500	0.979	0.938	500	1.058	0.951	498
0	0	1.000	0.995	500	0.746	0.732	500	0.055	0.055	500
	3	0.948	0.942	500	0.925	0.903	500	0.912	0.831	500
	5	0.934	0.909	500	0.951	0.932	500	1.064	0.950	500

corrected version are considered. The results of the asymptotic study and the small sample study are remarkably well in agreement, except for the small sample relative efficiency (SSRE), which tends to be slightly higher in certain regions of the grid, such as $\beta_0 = -3$ or 0 and $\beta_a = 0.3$. Also, the SSRE is larger for the model based than for the empirically corrected likelihood version, which is in line with knowledge about the sandwich estimator. The only major discrepancies, deserving further explanation, are seen for $\beta_0 = -5$, no dose effect ($\beta_d = 0$) and $\beta_a \neq 0$. First, observe that these parameter settings correspond to a very low background rate (the background probability of observing no malformation in a single fetus being 0.9933). It can be calculated that the marginal probability of sampling a cluster without malformations is 0.9229, 0.9851, and 0.9966 for the respective association parameters 0.0, 0.15, and 0.30. Correspondingly, the number of datasets without malformations (and thus with parameters at infinity) in a batch of 500 runs is on average 0.03, 83, and 332 respectively. In our simulation study, we actually encountered 0, 83, 331 of such datasets. All 83 respectively 331 of these datasets were ignored,

along with 98 respectively 76 other problematic sets of data, mainly because the latter contain merely a single malformation, which renders the association parameter inestimable. Still, the remaining 319 and 93 datasets are not free of difficulties. Let us consider variances and relative efficiencies for the dose effect in the 0.30 association case. The asymptotic variances are all about 17.8, while the simulation result for the small sample variances are smaller (8.84 for model based likelihood, 0.94 for empirically corrected likelihood, and 0.99 for PL). This might be due to the fact that omitting the problematic datasets truncates the sampling space and effectively reduces the variability. In particular, these problematic datasets contain no events, yielding an estimate for the intercept of $-\infty$, the dose effect being inestimable. Typically, samples with extreme parameter values are excluded, leading to still smaller sample variances. This effect is more pronounced in the empirically corrected estimators than in the purely model based one.

For the other, often more realistic parameter settings the asymptotic and small sample variances are in fairly good agreement. This leads to very close SSREs and AREs. Further, the observed variances in these settings, whether asymptotic or small sample, are much smaller than in the problematic settings described earlier. E.g., when $\beta_0 = -3.0$, $\beta_d = 3.0$, and $\beta_a = 0.15$, the asymptotic variances of the dose effect are all close to 0.13, while the small sample versions are about 0.14.

3.9 Conclusion

In this chapter we have shown that pseudo-likelihood estimation, in the sense of Arnold and Strauss (1991), is a viable and attractive alternative to maximum likelihood in the case of a single clustered binary outcome, analyzed with the exponential family model of Cox (1972), and applied to clustered data by Molenberghs and Ryan (1999).

The method yields consistent and asymptotically normal estimates of the parameters of interest. It avoids the need to calculate complex normalizing constants, yielding substantial gains of computing time. This is an important issue when data contain large and variable sized clusters. Another problem arises when multivariate outcomes are recorded for each littermate. Formulating appropriate pseudo-likelihood functions for multivariate clustered data is the topic of the next chapter.

It was thus shown that the loss of (asymptotic and small sample) efficiency, even

when theoretically problematic, is not an issue for realistic parameter combinations in the models for clustered data considered here. This is closely connected to the fact that the ARE equals one for a family of saturated models. These findings are supported by the analysis of five developmental toxicity studies.

Chapter 4

Pseudo-likelihood Inference for Clustered Multivariate Binary Outcomes

4.1 Introduction

In this chapter we re-introduce the exponential family model of Molenberghs and Ryan (1999), this time in its general multivariate clustered setting. As indicated before, this model benefits from the elegance and simplicity of exponential family theory and is flexible in terms of allowing response rates to depend on cluster size. With large clusters a main problem is however, the evaluation of the normalizing constant. Especially for trivariate and higher-order clustered outcomes, this exceeds the capacity of state-of-the-art computing. Therefore, we explore pseudo-likelihood as an alternative inferential procedure. This non-likelihood method yields a considerable gain of computation time, shows minimal efficiency loss and provides a flexible modelling framework.

Section 4.2 presents the extended exponential family model of Molenberghs and Ryan (1999), which allows for clustering as well as multiple outcomes. In Section 4.3, we will derive pseudo-likelihood estimating equations for this general multivariate setting. Whereas in the univariate case, there turned out to be only one "natural" formulation of the pseudo-likelihood estimating equations, it is now indicated that several plausible routes can be followed. In general, it is not guaranteed that a

pseudo-likelihood function corresponds to an existing and uniquely defined probability mass function. In Section 4.3 we will show that for our proposals, both existence and uniqueness are guaranteed. Therefore, the pseudo-likelihood as proposed here, still reflects the underlying likelihood so that it can be useful for dose-response modelling and quantitative risk assessment. This will be further illustrated in Chapter 5. Section 4.4 explores pseudo-likelihood as an alternative mode of inference for clustered multivariate binary outcomes. While point estimation and asymptotic normality have already been established in Chapter 3, this section is devoted to the construction of pseudo-likelihood counterparts for classical inferential tools such as ratio test statistics and score tests statistics. Section 4.5 further explores the performance of the pseudo-likelihood test statistics using asymptotic and small sample simulations. In Section 4.6 our findings are exemplified, using data from developmental toxicology experiments. Since the results of that section seem to imply that the efficiency loss of pseudo-likelihood over maximum likelihood is minor, a limited asymptotic relative efficiency study is performed in Section 4.7.

4.2 Model Formulation

Consider the notation introduced in Section 3.2, i.e. the experiment involves N clusters, the ith of which contains n_i individuals, each of whom are examined for the presence or absence of M responses and $y_{ijk} = 1$ when the kth individual in cluster i exhibits response j and -1 otherwise. It is convenient to group the outcomes for the ith cluster in an Mn_i vector $\boldsymbol{Y}_i = (Y_{i11}, \ldots, Y_{i1n_i}, \ldots, Y_{iM1}, \ldots, Y_{iMn_i})^T$. Molenberghs and Ryan (1999) proposed the following model for the joint distribution of clustered multivariate binary data:

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i^*) = \exp\left\{ \sum_{j=1}^{M} \sum_{k=1}^{n_i} \theta_{ij}^* y_{ijk} + \sum_{j=1}^{M} \sum_{k<k'} \delta_{ij}^* y_{ijk} y_{ijk'} + \sum_{j<j'} \sum_{k=1}^{n_i} \omega_{ijj'}^* y_{ijk} y_{ij'k} + \sum_{j<j'} \sum_{k \neq k'} \gamma_{ijj'}^* y_{ijk} y_{ij'k'} - A(\boldsymbol{\Theta}_i^*) \right\}, \tag{4.1}$$

where $A(\boldsymbol{\Theta}_i^*)$ is the normalizing constant, resulting from summing (4.1) over all 2^{Mn_i} possible outcomes. The building blocks of this model are clearly the "main effects" (θ^*) and three types of association parameters, reflecting three different types of association. E.g., δ_{ij}^* refers to the association between two different individuals

from the same cluster on the same outcome j, $\omega^*_{ijj'}$ refers to the association between outcomes j and j' for a single individual within cluster i and $\gamma^*_{ijj'}$ gives the association between outcomes j and j' for two different individuals in the same cluster. The three different types of associations captured in the model are depicted in Figure 4.1.

The absence of individual specific subscripts reflects the implicit exchangeability assumption between any two individuals within the same cluster. This assumption will now be used to simplify the model. Let z_{ij} be the number of individuals from cluster i positive on outcome j and $z_{ijj'}$ as the number of individuals in cluster i, positive on both outcomes j and j'. For the ith cluster, these can be thought of as arising from the set of two-by-two tables obtained by cross-classifying every pair of outcomes. This is illustrated in Table 4.1.

Using these summary statistics, Molenberghs and Ryan (1999) derived (after reparameterization):

$$\begin{aligned} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i;\boldsymbol{\Theta}_i) &= \exp\left\{\sum_{j=1}^{M}\theta_{ij}z_{ij}^{(1)}+\sum_{j=1}^{M}\delta_{ij}z_{ij}^{(2)}+\sum_{j<j'}\omega_{ijj'}z_{ijj'}^{(3)}\right. \\ &\quad \left.+\sum_{j<j'}\gamma_{ijj'}z_{ijj'}^{(4)}-A(\boldsymbol{\Theta}_i)\right\}, \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} z_{ij}^{(1)} &= z_{ij} \\ z_{ij}^{(2)} &= -z_{ij}(n_i-z_{ij}) \\ z_{ijj'}^{(3)} &= 2z_{ijj'}-z_{ij}-z_{ij'} \\ z_{ijj'}^{(4)} &= -z_{ij}(n_i-z_{ij'})-z_{ij'}(n_i-z_{ij})-z_{ijj'}^{(3)}. \end{aligned} \tag{4.3}$$

In the sequel, this will be referred to as the MR model. Its advantages are the flexibility, with which both main effects and associations can be modelled, and the absence of constraints on the parameter space, which eases interpretability. Further, the fact that the probability model depends explicitly (see (4.3)) and implicitly on the cluster size is an advantage since it is in line with the observation that litter size itself may depend on the level of exposure. Note that model (4.2) is conditional in nature, since it describes a feature of (a set of) outcomes conditional to the other outcomes. In particular, it implies conditional odds and conditional odds ratios that are log-linear in the natural parameters. E.g., Molenberghs and Ryan (1999)

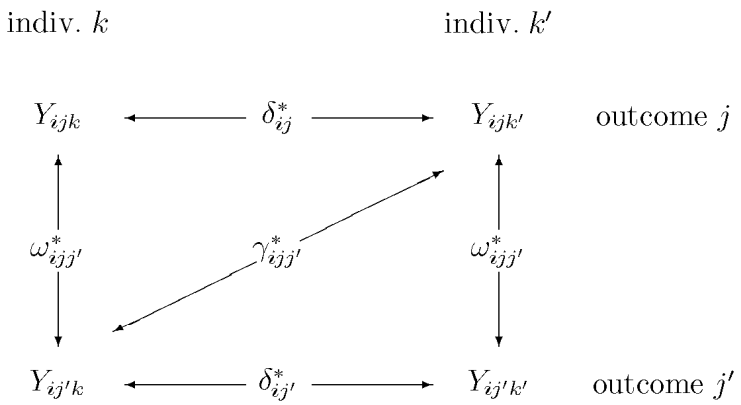

Figure 4.1: *Association Structure for Outcomes j and j' on Individuals k and k' in Cluster i.*

Table 4.1: *Cross-classification of Individuals in Cluster i with Respect to a Pair of Outcome Variables j and j'.*

	Outcome j	
Outcome j'	Absent Present	
Absent		
Present	$z_{ijj'}$	$z_{ij'}$
	z_{ij}	n_i

construct the conditional logit associated with the presence and absence of outcome j for an individual k in cluster i, given all other outcomes in the same cluster, and show that this function depends on cluster size and on the observed pattern of the remaining outcomes. Let $\kappa_{ijk} = 1$ if the kth individual exhibits a success on the jth variable and 0 otherwise. Then

$$\ln \frac{\text{pr}(Y_{ijk} = 1 | y_{ij'k'}, j' \neq j \text{ or } k' \neq k)}{\text{pr}(Y_{ijk} = -1 | y_{ij'k'}, j' \neq j \text{ or } k' \neq k)} - \theta_{ij} + \delta_{ij}(2z_{ij} - n_i - 1) + \sum_{j' \neq j} \omega_{ijj'}(2\kappa_{ij'k} - 1) + \sum_{j' \neq j} \gamma_{ijj'}(2z_{ij'} - n_i - 2\kappa_{ij'k} + 1). \tag{4.4}$$

As noted in Section 3.2.2, marginal quantities are fairly complicated functions of the parameters and are best represented graphically.

4.3 Pseudo-likelihood Estimation

The MR model, introduced in the previous section, is based on an exponential family model for multivariate binary data and exhibits a high flexibility to capture different patterns of non-linear dependencies of the marginal probabilities on the cluster size. Like most exponential family models, (4.2) enjoys well known properties, such as linearity of the log-likelihood in the minimal sufficient statistics, unimodality, etc. This implies a high numerical stability of iterative procedures to determine maximum likelihood estimators. In multivariate settings (with 3 or more outcomes) however, where the normalizing constant takes a complicated form, all of these advantages can be lost as this leads to excessive computational requirements. This is especially true for clusters of variable length, because the normalizing constant depends on the cluster size. Hence, alternative estimation methods, that do not require the explicit calculation of the normalizing constant, are in demand.

We explore the pseudo-likelihood estimation method, which is now indispensable. Again, the main idea is to replace the numerically intractable joint density by a simpler function that is a product of conditional densities that do not necessarily multiply to the joint distribution, but have the advantage that they do not involve that complicated normalizing constant. A bivariate distribution $f(y_1, y_2)$, for example, can be replaced by the product of both conditionals $f(y_1|y_2)f(y_2|y_1)$. This method converges quickly with only minor efficiency losses, especially for a range of realistic parameter settings. In Chapter 3 we presented a formal and more general

definition, initially proposed by Arnold and Strauss (1991), and proved consistency and asymptotic normality of the pseudo-likelihood estimator. We also showed that with a *single* clustered outcome, there is only one natural formulation of the pseudo-likelihood. It replaces the joint likelihood for the ith cluster by the product of n_i conditional probabilities of observing the outcome for littermate k, given the outcomes for the other $n_i - 1$ littermates. However, with the present model for clustered *multivariate* binary data, several formulations can be adopted.

One convenient PL function is found by replacing the joint density (4.2) by the product of Mn_i univariate conditional densities describing outcome j for the kth individual in a cluster, given all other outcomes in that cluster:

$$PL(1) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{k=1}^{n_i} f(y_{ijk} | y_{ij'k'}, j' \neq j \text{ or } k' \neq k; \boldsymbol{\Theta}_i). \tag{4.5}$$

This fits into framework (3.6) by choosing $\delta_{1_{Mn_i}} = Mn_i$ and $\delta_{s_{kj}} = -1$ for $k = 1, \ldots, n_i$ and $j = 1, \ldots, M$ where 1_{Mn_i} is a vector of ones and s_{kj} is a $Mn_i \times 1$ vector, obtained by applying the vec operator to an $n_i \times M$ matrix, consisting of ones everywhere, except for entry (k, j), which is 0. Since the members of each cluster are assumed to be exchangeable on every outcome separately, there are only $M2^M$ different contributions. For example, for clustered trivariate binary data, the logit of the conditional probability of observing a response of type 1 for the kth individual in cluster i, given there are responses of the two remaining types and given all outcomes for all other cluster members, is:

$$\theta_{i1} - \delta_{i1}(n_i - 2z_{i1} + 1) + \omega_{i12} + \omega_{i13} - \gamma_{i12}(n_i - 2z_{i2} + 1) - \gamma_{i13}(n_i - 2z_{i3} + 1),$$

with similar expressions for all other cases. The log pseudo-likelihood contribution for cluster i can now be written as a sum of such contributions, with appropriate multiplicities. Subsequently, one can model components of $\boldsymbol{\Theta}_i$ as a function of covariates, and take derivatives with respect to the regression parameters $\boldsymbol{\beta}$ to derive the score functions.

Equation (4.5) is one convenient definition of the PL function but certainly not the only one. E.g., one might want to preserve the multivariate nature of the data on each cluster member by considering the product of n_i conditional densities of the M outcomes for subject k, given the outcomes for the other subjects:

$$PL(2) = \prod_{i=1}^{N} \prod_{k=1}^{n_i} f(y_{ijk}, j = 1, \ldots, M | y_{ijk'}, k \neq k', j = 1, \ldots, M). \tag{4.6}$$

This satisfies the definition of Arnold and Strauss (1991) by taking $\delta_{1_{Mn_i}} = n_i$ and $\delta_{s_k} = -1$ for $k = 1, \ldots, n_i$. Here, 1_{Mn_i} denotes the Mn_i dimensional vector of ones, while s_k is the $(Mn_i \times 1)$ vector, obtained by applying the vec operator to an $(n_i \times M)$ matrix, consisting of ones everywhere, except for the kth row which consists of zeros.

Computational convenience may be the primary reason for choosing one PL definition over another. Let us discuss the relative merits of definitions (4.5) and (4.6). The former procedure is straightforward and natural when interest is focused on the estimation of main effect parameters. Furthermore, it is slightly easier to evaluate. If however, interest lies in the estimation of multivariate associations then approach (4.6) would be more natural. In Section 4.7 it is shown that both procedures are roughly equally efficient.

Further, it should be noted that in general, it is not guaranteed that a $p\ell$ function corresponds to an existing and uniquely defined probability mass function. However, since PL(1) and PL(2) are derived from (4.2), existence is guaranteed. In addition, both definitions (4.5) and (4.6) satisfy the conditions of the theorem presented in Gelman and Speed (1993), and hence uniqueness is guaranteed as well.

Since for toxicology data primary interest goes to the estimation of dose effects, which are usually incorporated into the main effects, we will focus on the use of the full conditional approach (4.5). In the context of the MR model, the notation PL will therefore refer to that approach. Only in cases where confusion might arise, it will be spelled out as PL(1).

4.4 Test Statistics

In Chapter 5, the data from two developmental toxicity studies will be used for quantitative risk assessment. One of the primary goals of quantitative risk assessment is to determine a safe level of exposure, based on an appropriate dose-response model. In the case of maximum likelihood estimation, several tools can be used to select such a model (e.g., Wald, score or likelihood ratio test statistics). Here we proposed pseudo-likelihood estimation as an attractive alternative to maximum likelihood estimation in the case of multivariate (e.g., clustered) binary outcomes, analyzed with the MR model. Therefore, in order to perform a flexible model selection, one needs extensions of the Wald, score or likelihood ratio test statistics to

the pseudo-likelihood framework.

Rotnitzky and Jewell (1990) examined the asymptotic distributions of generalized Wald and score tests, as well as likelihood ratio tests, for regression coefficients obtained by generalized estimating equations for a class of marginal generalized linear models for correlated data. Following a similar line of thought, we derive test statistics, as well as their asymptotic distributions for the pseudo-likelihood framework. Liang and Self (1996) have considered a test statistic, for one specific type of pseudo-likelihood function, which is similar in form to one of the tests we will derive below.

Suppose we are interested in testing the null hypothesis $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, where $\boldsymbol{\gamma}$ is an r-dimensonal 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 following test statistics can be used.

4.4.1 Wald Statistic

Because of the asymptotic normality of the PL estimator $\tilde{\boldsymbol{\beta}}_N$,

$$W^* = N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T \Sigma_{\gamma\gamma}^{-1} (\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)$$

has an asymptotic χ^2_r distribution under the null hypothesis, where $\Sigma_{\gamma\gamma}$ denotes the $r \times r$ submatrix of $\Sigma = J^{-1}KJ^{-1}$, with J and K shorthand notations for the matrices defined in (3.7) and (3.8). In practice, the matrix Σ can be replaced by a consistent estimator, obtained by substituting the PL estimator $\tilde{\boldsymbol{\beta}}_N$. Although the Wald test is usually simple to apply, it is notoriously sensitive to changes in parameterization (Fears, Benichou and Gail 1996). For example, using the delta method it is easy to show that the Wald test statistic for $H_0 : \gamma = 0$ is two times the Wald statistic for $H_0 : \gamma^2 = 0$. Therefore, the Wald test statistic is particularly unattractive for conditionally specified models, since marginal effects are likely to depend in a complex way on the model parameters (Diggle, Liang and Zeger 1994, pp. 148).

4.4.2 Pseudo-score Statistics

As an alternative to the Wald statistic, we propose pseudo-score statistics. A score test has the advantage to be obtained by fitting the null model. Furthermore, it is invariant to reparameterization.

Definitions

Let $\boldsymbol{U}(\boldsymbol{\beta})$ be the pseudo-score vector obtained by differentiation of the log of the pseudo-likelihood, and $\boldsymbol{U}_\gamma(\boldsymbol{\beta})$ its r-dimensional subvector corresponding to the components of $\boldsymbol{\gamma}$. Then, we can define an "empirically corrected" pseudo-score statistic as follows:

$$S^*(e.c) = \frac{1}{N}\boldsymbol{U}_\gamma(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0))^T J^{\gamma\gamma}\Sigma_{\gamma\gamma}^{\ 1} J^{\gamma\gamma}\boldsymbol{U}_\gamma(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)), \tag{4.7}$$

where $\tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)$ denotes the maximum pseudo-likelihood estimator of $\boldsymbol{\delta}$ in the subspace where $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, $J^{\gamma\gamma}$ is the $r \times r$ submatrix of the inverse of J, and $J^{\gamma\gamma}\Sigma_{\gamma\gamma}^{-1}J^{\gamma\gamma}$ is evaluated under H_0. In the following paragraph we show that, under mild regularity conditions, the pseudo-score statistic $S^*(e.c)$ is asymptotically χ^2_r distributed under H_0. As discussed by Rotnitzky and Jewell (1990) in the context of generalized estimating equations, the score statistic (4.7) may suffer from computational stability problems. A "model based" test that may be computationally simpler is:

$$S^*(m.b) = \frac{1}{N}\boldsymbol{U}_\gamma(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0))^T J^{\gamma\gamma}\boldsymbol{U}_\gamma(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)). \tag{4.8}$$

Its asymptotic distribution under H_0, however, is complicated and given by $\sum_{j=1}^r \lambda_j \chi^2_{1(j)}$ where the $\chi^2_{1(j)}$ are independently distributed as χ^2_1 variables and $\lambda_1 \geq \cdots \geq \lambda_r$ are the eigenvalues of $(J^{\gamma\gamma})^{-1}\Sigma_{\gamma\gamma}$, evaluated under H_0. The score statistic $S^*(m.b)$ in (4.8) can be adjusted such that it has an approximate χ^2_r distribution, which is much easier to evaluate. Several types of adjustments have been proposed in the literature (Rao and Scott 1987; Roberts, Rao and Kumar 1987). Similar to Rotnitzky and Jewell (1990), we propose an adjusted pseudo-score statistic

$$S^*_a(m.b) = S^*(m.b)/\overline{\lambda},$$

where $\overline{\lambda}$ is the arithmetic mean of the eigenvalues λ_j. Note that there is no distinction between $S^*(e.c)$ and $S^*_a(m.b)$ for $r = 1$. Moreover, in the likelihood-based case, all eigenvalues reduce to one and thus all three statistics coincide with the model based likelihood score statistic.

Derivation of the Asymptotic Distributions of the Pseudo-score Statistics

A Taylor series expansion of the pseudo-score $\boldsymbol{U}(\tilde{\boldsymbol{\beta}}_N)$ around $\boldsymbol{\beta}$ leads to:

$$(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) = \frac{1}{N}J^{-1}U(\boldsymbol{\beta}) + o_p^{p\times 1}(N^{-1/2}), \tag{4.9}$$

where $o_p^{k\times 1}(a_n)$ stands for a sequence of k-dimensional random variables that converges to zero in probability faster than a_n, as N tends to infinity. In partitioned matrix notation (4.9) can be rewritten as :

$$\begin{pmatrix} \tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma} \\ \tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta} \end{pmatrix} = \frac{1}{N} \left(\begin{array}{c|c} J^{\gamma\gamma} & J^{\gamma\delta} \\ \hline J^{\delta\gamma} & J^{\delta\delta} \end{array}\right) \begin{pmatrix} \boldsymbol{U}_\gamma(\boldsymbol{\beta}) \\ \boldsymbol{U}_\delta(\boldsymbol{\beta}) \end{pmatrix} + o_p^{p\times 1}(N^{-1/2})$$

Therefore,

$$(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0) = \frac{1}{N} J^{\gamma\gamma} \left[\boldsymbol{U}_\gamma(\boldsymbol{\gamma}_0, \boldsymbol{\delta}) + (J^{\gamma\gamma})^{-1} J^{\gamma\delta} \boldsymbol{U}_\delta(\boldsymbol{\gamma}_0, \boldsymbol{\delta})\right] + o_p^{r\times 1}(N^{-1/2}), \quad (4.10)$$

in which J is evaluated at $(\boldsymbol{\gamma}_0, \boldsymbol{\delta})$.

Next, we already know (by Theorem 3.3.1) that $\sqrt{N}(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)$ converges in distribution to a multivariate normal with zero mean and variance $\Sigma_{\gamma\gamma}$ evaluated at $(\boldsymbol{\gamma}_0, \boldsymbol{\delta})$. Therefore $N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T \Sigma_{\gamma\gamma}^{-1} (\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)$ converges in distribution to a χ^2-distribution with r degrees of freedom and thus, using (4.10),

$$\frac{1}{N} \left[U_\gamma(\boldsymbol{\gamma}_0, \boldsymbol{\delta}) + (J^{\gamma\gamma})^{-1} J^{\gamma\delta} U_\delta(\boldsymbol{\gamma}_0, \boldsymbol{\delta})\right]^T J^{\gamma\gamma} \Sigma_{\gamma\gamma}^{-1} J^{\gamma\gamma} \left[U_\gamma(\boldsymbol{\gamma}_0, \boldsymbol{\delta}) + (J^{\gamma\gamma})^{-1} J^{\gamma\delta} U_\delta(\boldsymbol{\gamma}_0, \boldsymbol{\delta})\right]. \quad (4.11)$$

converges to a χ^2-distribution with r degrees of freedom too. Under H_0, (4.11) simplifies to:

$$\frac{1}{N} (U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)))^T J^{\gamma\gamma} \Sigma_{\gamma\gamma}^{-1} J^{\gamma\gamma} (U_\gamma(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0))),$$

which completes the derivation of the asymptotic distribution of (4.7).

Next, following Johnson and Kotz (1970, p. 150) the model based score statistic

$$S^*(m.b) = \frac{1}{N} (\boldsymbol{U}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)))^T J^{\gamma\gamma} (\boldsymbol{U}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)))$$

is aymptotically distributed as $\sum_{j=1}^r \lambda_j \chi_j^2$, where χ_j^2 are independently distributed according to a χ^2-distribution with 1 degree of freedom and $\lambda_1 \geq \cdots \geq \lambda_r$ are the eigenvalues of $\Sigma_{\gamma\gamma}(J^{\gamma\gamma})^{-1}$, evaluated under H_0.

4.4.3 Pseudo-likelihood Ratio Statistic

Another alternative is provided by the pseudo-likelihood ratio test statistic, which requires comparison of full and reduced model.

Definition

We define:

$$G^{*2} = 2\left[p\ell(\tilde{\boldsymbol{\beta}}_N) - p\ell(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}(\boldsymbol{\gamma}_0))\right].$$

In the next paragraph, we show that the asymptotic distribution of G^{*2} can be written as a weighted sum $\sum_{j=1}^{r} \lambda_j \chi^2_{1(j)}$, where the $\chi^2_{1(j)}$ are independently distributed as χ^2_1 variables and $\lambda_1 \geq \cdots \geq \lambda_r$ are the eigenvalues of $(J^{\gamma\gamma})^{-1}\Sigma_{\gamma\gamma}$. Alternatively, the adjusted pseudo-likelihood ratio test statistic, defined by

$$G^{*2}_a = G^{*2}/\overline{\lambda}$$

is approximately χ^2_r distributed. The proof shows that G^{*2} can be rewritten as an approximation to a Wald statistic. The covariance structure of the Wald statistic can be calculated under the null hypothesis, but also under the alternative hypothesis. Both versions of the Wald tests are asymptotically equivalent under H_0 (Rao 1973, p. 418). It can therefore be argued that the adjustments in G^{*2}_a can also be evaluated under the null as well as under the alternative hypothesis. These adjusted statistics will then be denoted by $G^{*2}_a(H_0)$ and $G^{*2}_a(H_1)$ respectively. In analogy with the Wald test statistic, we expect $G^{*2}_a(H_1)$ to have high power. A similar reasoning suggests that the score test $S^*_a(m.b)$ might closely correspond to $G^{*2}_a(H_0)$, since both depend strongly on the fitted null model. Analogous results were obtained by Rotnitzky and Jewell (1990). Sections 4.5.1 and 4.5.2 briefly compare the asymptotic and small sample behaviours of the different test statistics.

Derivation of the Asymptotic Distribution of the Pseudo-likelihood Ratio Test Statistic

Using a Taylor expansion of the log pseudo-likelihood function around $\boldsymbol{\beta}$, we obtain:

$$p\ell(\tilde{\boldsymbol{\beta}}_N) = p\ell(\boldsymbol{\beta}) + (\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta})^T U(\boldsymbol{\beta}) + \frac{1}{2}(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta})^T(-NJ)(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) + o_p^{p\times 1}(1). \tag{4.12}$$

Using a Taylor expansion of the pseudo-score function around $\boldsymbol{\beta}$, we obtain:

$$0 = U(\tilde{\boldsymbol{\beta}}_N) = U(\boldsymbol{\beta}) + (-NJ)(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) + o_p^{p\times 1}(N^{1/2}), \tag{4.13}$$

or

$$U(\boldsymbol{\beta}) = NJ(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) + o_p^{p\times 1}(N^{1/2}). \tag{4.14}$$

Substituting (4.13) in (4.12), we find :

$$p\ell(\tilde{\boldsymbol{\beta}}_N) = p\ell(\boldsymbol{\beta}) + \frac{N}{2}(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta})^T J(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) + o_p^{p\times 1}(1). \tag{4.15}$$

In partitioned matrix notation (4.15) can be rewritten as:

$$\begin{aligned} &p\ell(\tilde{\boldsymbol{\gamma}}_N, \tilde{\boldsymbol{\delta}}_N) = p\ell(\boldsymbol{\gamma}_N, \boldsymbol{\delta}) \\ &+\tfrac{N}{2}\left((\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma})^T, (\tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta})^T\right) \left(\begin{array}{c|c} J_{\gamma\gamma} & J_{\gamma\delta} \\ \hline J_{\delta\gamma} & J_{\delta\delta} \end{array}\right) \left(\begin{array}{c} (\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ \hline (\tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta}) \end{array}\right) + o_p^{p\times 1}(1). \end{aligned} \tag{4.16}$$

Assuming the null hypothesis is true, (4.16) reduces to:

$$U_\delta(\boldsymbol{\gamma}_0, \boldsymbol{\delta}) = N J_{\delta\delta}(\tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0) - \boldsymbol{\delta}) + o_p^{(p-r)\times 1}(N^{1/2}) \tag{4.17}$$

in which J is evaluated at $(\boldsymbol{\gamma}_0, \boldsymbol{\delta})$. Equating (4.17) with the last $(p-r)$ rows of the score vector $U(\boldsymbol{\gamma}_0, \boldsymbol{\delta})$, obtained from (4.14), leads to:

$$(\tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0) - \boldsymbol{\delta}) = J_{\delta\delta}^{-1} J_{\delta\gamma}(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0) + (\tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta}) + o_p^{(p-r)\times 1}(N^{-1/2}).$$

Therefore,

$$\begin{aligned} 2\left[p\ell(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0) - p\ell(\boldsymbol{\gamma}_0, \boldsymbol{\delta})\right] &= N(\tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0) - \boldsymbol{\delta})^T J_{\delta\delta}(\tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0) - \boldsymbol{\delta}) + o_p^{p\times 1}(1) \\ &= N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T (J_{\delta\gamma}^T J_{\delta\delta}^{-1} J_{\delta\gamma})(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0) \\ &\quad +2N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T J_{\delta\gamma}^T(\tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta}) \\ &\quad +N(\tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta})^T J_{\delta\delta}(\tilde{\boldsymbol{\delta}}_N - \boldsymbol{\delta}) + o_p^{p\times 1}(1). \end{aligned}$$

Using the expression for the inverse of a partitioned matrix it follows that:

$$\begin{aligned} 2\left[p\ell(\tilde{\boldsymbol{\gamma}}_N, \tilde{\boldsymbol{\delta}}_N) - p\ell(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0))\right] &= 2\left[p\ell(\tilde{\boldsymbol{\gamma}}_N, \tilde{\boldsymbol{\delta}}_N) - p\ell(\boldsymbol{\gamma}_0, \boldsymbol{\delta})\right] \\ &\quad -2\left[p\ell(\boldsymbol{\gamma}_0, \tilde{\boldsymbol{\delta}}_N(\boldsymbol{\gamma}_0)) - p\ell(\boldsymbol{\gamma}_0, \boldsymbol{\delta})\right] \\ &= N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T (J_{\gamma\gamma} - J_{\delta\gamma}^T J_{\delta\delta}^{-1} J_{\delta\gamma})(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0) + o_p^{p\times 1}(1) \\ &= N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T (J^{\gamma\gamma})^{-1}(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0) + o_p^{p\times 1}(1) \end{aligned}$$

Since $\sqrt{N}(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)$ converges in distribution to a multivariate normal with mean zero and variance $\Sigma_{\gamma\gamma}$, it follows from Johnson and Kotz (1970, p. 150) that the distribution of

$$N(\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)^T (J^{\gamma\gamma})^{-1} (\tilde{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}_0)$$

is the same as that of $\sum_{j=1}^{r} \lambda_j \chi^2_{1(j)}$, where $\chi^2_{1(j)}$ are independently distributed according to a χ^2-distribution with 1 degree of freedom and $\lambda_1 \geq \cdots \geq \lambda_r$ are the eigenvalues of $\Sigma_{\gamma\gamma}(J^{\gamma\gamma})^{-1}$.

4.5 Simulation Results

4.5.1 Asymptotic Simulations

To explore more thoroughly the performance of the pseudo-likelihood estimator and pseudo-likelihood test statistics, we will show a few simulation results with asymptotic considerations similar to the ideas of Rotnitzky and Wypij (1994). Remember from Section 3.7.3 that these constitute an artificial sample, where each possible realization is weighted according to its true probability. E.g., in a univariate setting, they would consider all realizations of the form (d_i, n_i, z_i). So, we need 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 a cluster size can occur, possibly depending on the dosing (we assume here $f(n_i|d_i) = f(n_i)$) and (3) $f(z_i|n_i, d_i)$, the actual model probabilities. This approach can easily be adapted to a multivariate context. As above, we assume that there are 4 dose groups, with one control group ($d_i = 0$) and three active groups ($d_i = 0.25, 0.5, 1.0$) and that the number of viable fetuses (n_i) per cluster is chosen at random from a local linear smoothed version of the relative frequency distribution given in Table 1 of Kupper et al. (1986) (which is considered representative of that encountered in actual experimental situations). The smoothed frequencies were presented in Table 3.3.

The present study is restricted to clusters of *bivariate* binary data with maximum cluster size of 10, due to prohibitive time requirements of ML. The main effects are modelled as $\theta_{ij} = \beta_{0j} + \beta_d d_i (j = 1, 2)$, i.e. a common main dose effect is assumed, and all association parameters are assumed to be constant. Data are generated from a bivariate model with background rate parameters $(\beta_{01}, \beta_{02}) = (-3, -3)$ and a zero association vector $(\delta_1, \delta_2, \omega_{12}, \gamma_{12})$. Positive associations yield similar results.

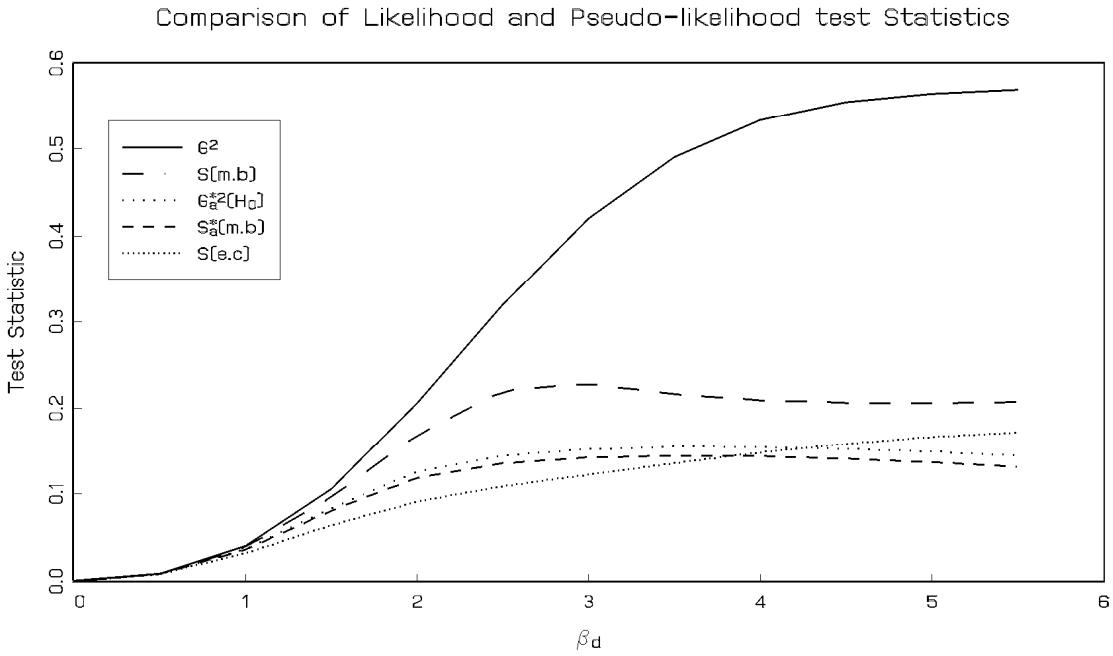

Figure 4.2: *Simulation Results: Comparison of Likelihood and Pseudo-likelihood Test Statistics for a Common Dose Trend in the Bivariate MR Model*

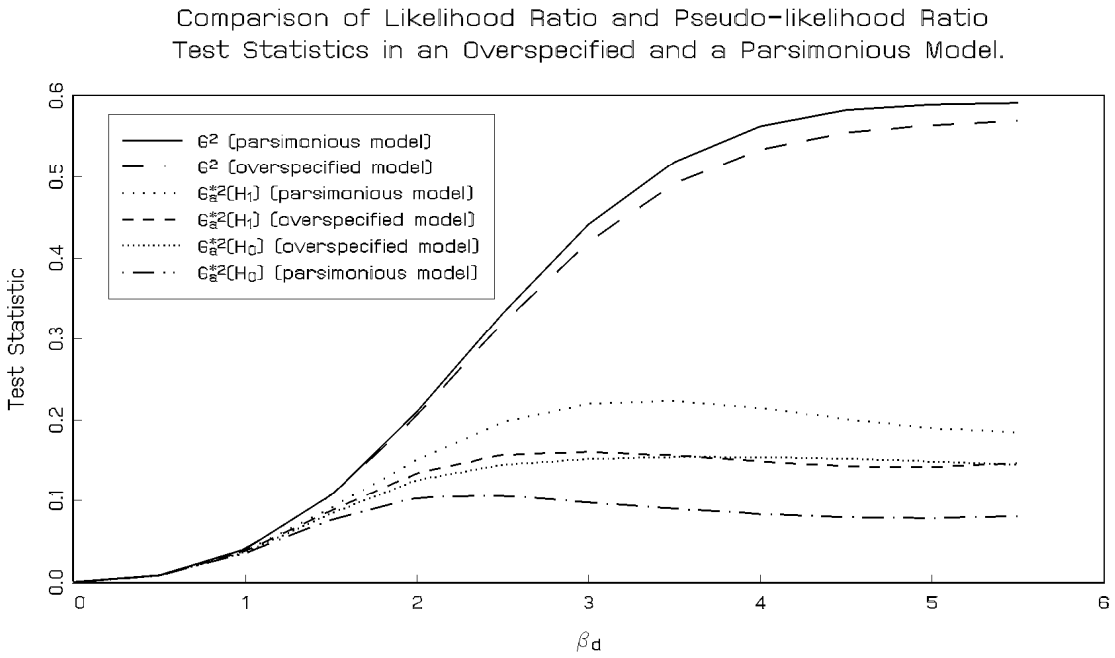

Figure 4.3: *Simulation Results: Comparison of Likelihood Ratio* (G^2) *and Adjusted Pseudo-likelihood Ratio* G_a^{*2} *Test Statistics for a Common Dose Trend in an Overspecified and a Parsimonious Bivariate MR Model. The Adjustments are Calculated under the Alternative* ($G_a^{*2}(H_1)$) *and under the Null Model* ($G_a^{*2}(H_0)$)

We want to assess the effect of β_d. Since the Wald test is known to depend on the particular parameterization, it might be a less relevant measure to use. We will therefore concentrate mainly on score and ratio statistics.

Figure 4.2 shows the adjusted pseudo-score and pseudo-likelihood ratio statistics $S_a^*(m.b)$ and $G_a^{*2}(H_0)$, as well as the model based, $S(m.b)$, and empirically corrected, $S(e.c)$, likelihood score tests and the likelihood ratio statistic G^2. We restrict to $G_a^{*2}(H_0)$, since it is similar to $G_a^{*2}(H_1)$ in this case. Note that $S_a^*(m.b)$ is identical to $S^*(e.c)$, since we are testing for the effect of a single parameter. In the absence of both a true dose effect and an association between outcomes or between cluster members, likelihood and pseudo-likelihood are equivalent. However, a substantial discrepancy arises between G^2 and $G_a^{*2}(H_0)$ for positive dose effects. Indeed, by ignoring an important effect, we introduce an apparent association, which is given too much weight in the pseudo-likelihood. This leads to a pseudo-likelihood value that is too large under the null. Therefore, the pseudo-deviance is much smaller than the likelihood deviance. As a consequence of the misspecification, the matrix $\Sigma_{\gamma\gamma}(J^{\gamma\gamma})^{-1}$, and hence also the corresponding adjustment, is overestimated, rendering an even greater discrepancy between the test statistics G^2 and $G_a^{*2}(H_0)$. A similar argument explains the discrepancy with the pseudo-score statistic, since this statistic is fully obtained from the null model. As follows from theory, $S_a^*(m.b)$ and $G_a^{*2}(H_0)$ are comparable. For small to moderate dose effects, both these statistics are situated between $S(m.b)$ and $S(e.c)$. However, for larger dose effects, the pseudo-statistics $S_a^*(m.b)$ and $G_a^{*2}(H_0)$, as well as their adjustments, show a non-monotone behaviour, in contrast to the likelihood ratio which increases monotonically with dose. This issue deserves further research.

We will now study an anomaly that was observed when fitting a more parsimonious model, excluding the zero clustering parameters δ_1 and δ_2. Results are summarized in Figure 4.3.

The likelihood ratio statistic, calculated under the parsimonious model is larger than when it is obtained from the overspecified model. This is in agreement with known properties of the likelihood function. The pseudo-likelihood ratio statistic $G_a^{*2}(H_0)$, on the contrary, becomes smaller. Again, we might argue that a model under the null hypothesis is misspecified, thus introducing an apparant association that cannot be captured by the clustering parameters. For reasons, similar to the ones in our previous discussion, this might lead to inflated variances and correspond-

Table 4.2: *Simulation Results: Type I Error Probabilities for* $\beta_0 = -2.5$ *and Dose Levels* $0, .25, .50, 1$ *(NC is the number of clusters per dose level).*

		Likelihood			Pseudo-likelihood		
β_a	NC	G^2	$S(m.b)$	$S(e.c)$	$G^*_a(H_0)$	$G^*_a(H_1)$	$S^*(e.c) = S^*_a(m.b)$
0.10	5	5.09	4.21	3.21	4.29	14.29	2.80
	30	6.00	6.20	5.00	5.40	6.60	5.40
0.25	15	3.63	3.68	1.23	4.70	18.37	2.25
	30	6.01	4.60	5.00	6.63	10.84	5.20

ing adjustments. However, the pseudo-likelihood ratio statistic $G^{*2}_a(H_1)$ behaves in agreement with intuition, since it is based on the (correct) alternative model. A similar feature was observed by Rotnitzky and Jewell (1990).

4.5.2 Small Sample Simulations

In this section we perform a small sample simulation study for a single clustered outcome, based upon 500 replications, to illustrate the finite sample behaviour of the pseudo-likelihood test statistics with respect to type I error probability and power. The number n_i of viable fetuses per cluster is again assumed to follow a local linear smoothed version of the relative frequency distribution in Table 1 of Kupper et al. (1986). Data are generated and fitted using a model where the main effect is modelled as $\theta_i = \beta_0 + \beta_d d_i$ and the association parameter is held constant ($\delta_i = \beta_a$). The hypothesis of interest is $\beta_d = 0$.

The simulation results are shown in Tables 4.2 and 4.3.

The pseudo-score test statistics as well as $G^{*2}_a(H_0)$ have satisfactory type I error probabilities, in good agreement with their likelihood counterparts. Since we are in the single parameter case, $S^*(e.c)$ and $S^*_a(m.b)$ yield identical results. The rejection probabilities for the pseudo-score test statistics tend to be somewhat smaller than for pseudo-likelihood ratio test statistics, which is often observed in the likelihood setting as well. The pseudo-likelihood ratio statistic $G^{*2}_a(H_1)$ shows inflated type I error probabilities, especially for small samples. Consequently its power may be misleadingly high. This feature is commonly observed for the Wald statistic

Table 4.3: *Simulation Results: Powers for* $\beta_0 = -2.5$, $\beta_a = 0.1$ *and Dose Levels* $0, .25, .50, 1$ *(NC is the number of clusters per dose level).*

		Likelihood			Pseudo-likelihood		
β_d	NC	G^2	$S(m.b)$	$S(e.c)$	$G_a^*(H_0)$	$G_a^*(H_1)$	$S^*(e.c) = S_a^*(m.b)$
1.0	5	25.05	24.80	16.80	20.04	29.66	20.20
2.0		96.42	95.79	79.79	90.18	91.38	90.40
2.5		100.00	100.00	92.46	98.20	97.80	98.00
1.0	30	88.40	88.60	83.40	83.80	87.60	84.40
2.0		100.00	100.00	100.00	100.00	100.00	100.00
2.5		100.00	100.00	100.00	100.00	100.00	100.00

(which is also based on the alternative model). The power of $G_a^{*2}(H_0)$ closely corresponds to that of the pseudo-score statistics. For realistic parameter settings such as $(\beta_0, \beta_d, \beta_a) - (-2.5, 2.5, 0.1)$ (based on analyses of National Toxicology Program data; Price, Kimmel, George and Marr 1987) and/or large samples, $G_a^{*2}(H_1)$ behaves similarly to the other pseudo-likelihood test statistics. In that case, powers are then very high anyway for all pseudo-likelihood statistics and comparable to their likelihood counterparts.

4.5.3 Summary

Wald tests can have poor properties for conditional models. Therefore we advocate the use of score and ratio test statistics. The pseudo-score test statistics have the advantage to need evaluation under the null model only. Moreover $S^*(e.c)$ has an appealing asymptotic distribution. On the other hand, $S^*(e.c)$ may be computationally unstable. In that case, the use of working score statistic $S^*(m.b)$, even though its asymptotic distribution is more complicated, should be preferred. To avoid this problem, an adjusted score statistic $S_a^*(m.b)$ was proposed. The distribution of the pseudo-likelihood ratio test statistic can be adjusted similarly. Our simulations suggest that the pseudo-score statistics as well as $G_a^2(H_0)$ may have lower power than their likelihood counterparts. Calculating the adjusted pseudo-likelihood ratio test

under the alternative, $G_a^2(H_1)$ may increase the power, but tends to inflate type I error probabilities in small samples. For realistic parameter settings however, the pseudo-likelihood ratio tests produce high powers, especially for large samples. To conclude, we suggest the use of the adjusted pseudo-likelihood ratio tests, but we recommend caution for small sample sizes.

4.6 Examples

In this section we illustrate the developed pseudo-likelihood techniques to the data from the NTP developmental experiments DEHP, EG and DYME in mice. We discuss parameter estimation and hypothesis testing. First, we consider bivariate analyses, choosing pairs out of the three possible malformation outcomes: external, visceral and skeletal (respectively indexed by 1,2 and 3). Secondly, we illustrate the derived test statistics in the bivariate setting. Finally, we perform trivariate data analyses, and we demonstrate some important advantages of PL estimation.

4.6.1 Bivariate Analyses

Parameters were estimated by means of ML and PL. The main effects are modelled as either: $(1)\theta_{ij} = \beta_{0j} + \beta_{dj}d_i$ $(j = 1, 2)$, i.e. a different dose effect parameter is included for each outcome, or $(2)\theta_{ij} = \beta_{0j} + \beta_d d_i$ $(j = 1, 2)$, i.e. a common dose effect is assumed. All association parameters are assumed to be constant. Tables 4.4 through 4.7 give a detailed picture of the results obtained for all three NTP studies.

The tables reveal that ML and PL estimates are fairly similar. Dose effect parameters are statistically highly significant for all analyses, independent of the estimation technique. The next chapter will try to provide with more insight in the scientific relevance of these dose effects.

The clustering parameters (δ_1 and δ_2) are significant, except for EG (External-Visceral). From Table 2.5 in Chapter 2 it follows that the malformation frequencies for external and visceral outcomes in the EG study are rather small. The remaining association parameters often do not reach the 5% significance level.

Although ML and PL estimates and standard errors are numerically different, they have similar magnitude and direction, and no clear ordering is seen between them. Further, the Wald test statistic for dose effect is higher with PL than with

Table 4.4: *NTP Studies: Maximum Likelihood Estimates (model based standard errors;empirically corrected standard errors) of Bivariate Outcomes (different main dose effects).*

Study	Par.	External-Visceral	External-Skeletal	Visceral-Skeletal
DEHP	β_{01}	-2.62 (0.60;0.60)	-2.66 (0.60;0.53)	-2.57 (0.54;0.55)
	β_{d1}	2.92 (0.68;0.69)	2.89 (0.67;0.62)	2.59 (0.58;0.62)
	β_{02}	-2.18 (0.54;0.65)	-2.63 (0.60;0.76)	-2.94 (0.61;0.81)
	β_{d2}	2.22 (0.59;0.75)	2.73 (0.66;0.83)	3.02 (0.65;0.83)
	δ_1	0.14 (0.06;0.06)	0.22 (0.04;0.04)	0.21 (0.04;0.03)
	δ_2	0.14 (0.06;0.05)	0.21 (0.05;0.05)	0.20 (0.04;0.04)
	ω_{12}	0.09 (0.20;0.27)	0.54 (0.20;0.20)	0.34 (0.20;0.26)
	γ_{12}	0.04 (0.05;0.06)	-0.08 (0.03;0.03)	-0.08 (0.03;0.02)
EG	β_{01}	-2.96 (0.96;1.10)	-2.74 (0.80;0.97)	-4.75 (1.56;1.38)
	β_{d1}	2.28 (0.69;0.88)	1.84 (0.76;0.84)	3.14 (1.43;0.89)
	β_{02}	-5.12 (1.61;1.60)	-0.53 (0.31;0.32)	-0.37 (0.47;0.43)
	β_{d2}	3.75 (1.35;1.20)	0.94 (0.20;0.20)	0.94 (0.20;0.19)
	δ_1	0.22 (0.10;0.10)	0.25 (0.05;0.06)	0.23 (0.09;0.09)
	δ_2	0.18 (0.14;0.13)	0.20 (0.02;0.02)	0.21 (0.02;0.02)
	ω_{12}	-0.09 (0.56;0.59)	0.43 (0.26;0.28)	0.91 (0.43;0.33)
	γ_{12}	0.05 (0.10;0.09)	-0.01 (0.02;0.03)	-0.04 (0.03;0.03)
DYME	β_{01}	-5.26 (1.15;1.27)	-7.09 (1.23;1.32)	-2.89 (1.06;1.21)
	β_{d1}	5.88 (1.22;1.37)	8.01 (1.38;1.56)	2.18 (1.14;1.37)
	β_{02}	-2.87 (0.98;0.90)	-3.59 (0.66;0.75)	-1.25 (0.53;0.83)
	β_{d2}	2.37 (0.99;0.95)	4.54 (0.80;0.94)	2.30 (0.53;0.97)
	δ_1	0.09 (0.06;0.07)	0.11 (0.05;0.04)	0.29 (0.05;0.05)
	δ_2	0.29 (0.05;0.05)	0.23 (0.02;0.02)	0.25 (0.03;0.03)
	ω_{12}	-0.29 (0.24;0.21)	0.01 (0.22;0.34)	0.42 (0.32;0.29)
	γ_{12}	0.06 (0.04;0.04)	-0.09 (0.02;0.03)	-0.02 (0.03;0.06)

Table 4.5: *NTP Studies: Pseudo-likelihood Estimates (standard errors) of Bivariate Outcomes (different main dose effects).*

Study	Par.	External-Visceral	External-Skeletal	Visceral-Skeletal
DEHP	β_{01}	-2.52 (0.61)	-2.80 (0.53)	-2.73 (0.53)
	β_{d1}	2.99 (0.67)	3.23 (0.58)	3.03 (0.58)
	β_{02}	-1.90 (0.65)	-2.41 (0.70)	-2.80 (0.71)
	β_{d2}	2.12 (0.70)	2.46 (0.79)	2.77 (0.75)
	δ_1	0.14 (0.05)	0.23 (0.05)	0.25 (0.04)
	δ_2	0.15 (0.05)	0.25 (0.06)	0.26 (0.05)
	$\omega 12$	0.12 (0.25)	0.56 (0.20)	0.36 (0.28)
	$\gamma 12$	0.07 (0.05)	-0.10 (0.05)	-0.12 (0.04)
EG	β_{01}	-2.39 (0.94)	-2.37 (0.88)	-4.69 (1.63)
	β_{d1}	2.14 (0.76)	1.64 (0.74)	3.16 (0.98)
	β_{02}	-5.04 (1.62)	-0.77 (0.30)	-0.69 (0.43)
	β_{d2}	3.73 (1.17)	1.39 (0.20)	1.44 (0.19)
	δ_1	0.24 (0.11)	0.28 (0.06)	0.24 (0.11)
	δ_2	0.16 (0.14)	0.20 (0.01)	0.21 (0.01)
	ω_{12}	-0.05 (0.58)	0.14 (0.31)	0.81 (0.33)
	γ_{12}	0.08 (0.11)	0.03 (0.03)	-0.03 (0.03)
DYME	β_{01}	-4.74 (0.90)	-5.87 (1.26)	-3.03 (1.19)
	β_{d1}	5.35 (0.92)	6.55 (1.45)	2.38 (1.19)
	β_{02}	-2.85 (1.06)	-3.36 (0.79)	-1.90 (0.55)
	β_{d2}	2.46 (1.02)	4.40 (1.02)	3.11 (0.56)
	δ_1	0.12 (0.05)	0.16 (0.05)	0.28 (0.06)
	δ_2	0.30 (0.06)	0.25 (0.02)	0.25 (0.02)
	ω_{12}	-0.41 (0.20)	0.23 (0.29)	0.34 (0.31)
	γ_{12}	0.07 (0.04)	-0.10 (0.05)	-0.01 (0.05)

Table 4.6: *NTP Studies: Maximum Likelihood Estimates (model based standard errors;empirically corrected standard errors) of Bivariate Outcomes (common main dose effects).*

Study	Par.	External-Visceral	External-Skeletal	Visceral-Skeletal
DEHP	β_{01}	-2.31 (0.43;0.57)	-2.59 (0.46;0.49)	-2.74 (0.46;0.54)
	β_{02}	-2.46 (0.43;0.60)	-2.70 (0.46;0.52)	-2.74 (0.46;0.56)
	δ_1	0.16 (0.05;0.06)	0.22 (0.04;0.04)	0.20 (0.03;0.03)
	δ_2	0.12 (0.06;0.05)	0.20 (0.04;0.04)	0.22 (0.03;0.03)
	ω_{12}	0.08 (0.20;0.27)	0.54 (0.20;0.20)	0.34 (0.20;0.26)
	γ_{12}	0.04 (0.05;0.06)	-0.08 (0.03;0.03)	-0.08 (0.03;0.02)
	β_d	2.54 (0.44;0.65)	2.81 (0.49;0.56)	2.79 (0.47;0.57)
EG	β_{01}	-3.33 (0.91;1.00)	-2.01 (0.42;0.43)	-2.86 (0.79;0.69)
	β_{02}	-4.02 (1.01;1.12)	-0.45 (0.31;0.34)	-0.12 (0.46;0.50)
	δ_1	0.20 (0.10;0.10)	0.26 (0.05;0.05)	0.27 (0.09;0.08)
	δ_2	0.19 (0.14;0.13)	0.20 (0.02;0.02)	0.20 (0.02;0.02)
	$\omega 12$	-0.08 (0.56;0.59)	0.52 (0.26;0.31)	1.16 (0.41;0.41)
	$\gamma 12$	0.06 (0.11;0.09)	-0.01 (0.02;0.03)	-0.04 (0.02;0.03)
	β_d	2.68 (0.61;0.75)	1.02 (0.20;0.21)	1.02 (0.20;0.20)
DYME	β_{01}	-3.93 (0.78;0.80)	-4.49 (0.63;0.72)	-2.99 (0.51;0.82)
	β_{02}	-4.81 (0.79;0.81)	-4.02 (0.64;0.77)	-1.25 (0.52;0.83)
	δ_1	0.16 (0.04;0.04)	0.22 (0.03;0.03)	0.29 (0.05;0.05)
	δ_2	0.25 (0.05;0.06)	0.21 (0.02;0.02)	0.25 (0.03;0.03)
	ω_{12}	-0.31 (0.23;0.19)	0.06 (0.23;0.37)	0.41 (0.30;0.31)
	γ_{12}	0.05 (0.03;0.04)	-0.10 (0.02;0.03)	-0.01 (0.03;0.05)
	β_d	4.30 (0.77;0.80)	5.26 (0.77;0.95)	2.28 (0.51;0.96)

Table 4.7: *NTP Studies: Pseudo likelihood Estimates (standard errors) of Bivariate Outcomes (common main dose effects).*

Study	Par.	External-Visceral	External-Skeletal	Visceral-Skeletal
DEHP	β_{01}	-2.15 (0.58)	-2.48 (0.47)	-2.64 (0.50)
	β_{02}	-2.21 (0.61)	-2.72 (0.50)	-2.90 (0.55)
	δ_1	0.16 (0.05)	0.25 (0.05)	0.25 (0.05)
	δ_2	0.13 (0.05)	0.23 (0.05)	0.25 (0.04)
	ω_{12}	0.12 (0.25)	0.56 (0.20)	0.36 (0.28)
	γ_{12}	0.06 (0.05)	-0.10 (0.05)	-0.12 (0.04)
	β_d	2.51 (0.60)	2.84 (0.53)	2.91 (0.52)
EG	β_{01}	-2.79 (0.85)	-2.18 (0.45)	-3.05 (0.74)
	β_{02}	-3.88 (1.08)	-0.76 (0.30)	-0.58 (0.47)
	δ_1	0.23 (0.11)	0.28 (0.06)	0.27 (0.09)
	δ_2	0.18 (0.14)	0.20 (0.01)	0.21 (0.01)
	ω_{12}	-0.04 (0.58)	0.15 (0.31)	0.90 (0.36)
	γ_{12}	0.09 (0.11)	0.03 (0.03)	-0.02 (0.03)
	β_d	2.59 (0.64)	1.42 (0.20)	1.52 (0.19)
DYME	β_{01}	-3.83 (0.78)	-4.50 (0.80)	-3.63 (0.59)
	β_{02}	-1.57 (0.83)	-3.87 (0.77)	-1.89 (0.55)
	δ_1	0.18 (0.04)	0.22 (0.03)	0.27 (0.06)
	δ_2	0.26 (0.06)	0.24 (0.02)	0.26 (0.02)
	ω_{12}	-0.42 (0.19)	0.23 (0.30)	0.32 (0.31)
	γ_{12}	0.05 (0.04)	-0.10 (0.05)	-0.01 (0.05)
	β_d	4.21 (0.74)	5.08 (0.99)	3.00 (0.57)

Table 4.8: *NTP Studies: Relative Time Gains (RTG) of Pseudo-likelihood Compared to Maximum Likelihood (in seconds).*

Study	Parameter Combination	ML	PL	RTG
DEHP	External-Visceral	73.17	14.06	5.20
	External-Skeletal	75.30	14.01	5.37
	Visceral-Skeletal	72.77	14.06	5.18
EG	External-Visceral	75.97	15.10	5.03
	External-Skeletal	73.55	13.84	5.31
	Visceral-Skeletal	76.02	15.11	5.03
DYME	External-Visceral	98.65	15.76	6.26
	External-Skeletal	98.42	15.77	6.24
	Visceral-Skeletal	95.35	15.77	6.04

ML in about 60% of the cases considered.

From a computational perspective, Table 4.8 shows that the PL estimation procedure needs only between 14 and 16 seconds to converge, while ML needs 73 to 99 seconds. This translates into relative time gains of 5 to 6 seconds. Especially, in the trivariate case, PL will become really superior.

4.6.2 Tests for Trend

Tests for trend are often applied to toxicological data in order to assess dose effects. Lefkopoulou, Rotnitzky and Ryan (1996) explain the need for computationally simple trend tests. We compute the test statistics, proposed in Section 3 for main dose effects in the NTP data (bivariate case only). In particular, we test $H_{0j} : \beta_{dj} = 0$ $(j = 1, 2)$ and $H_0 : \beta_d = 0$. Results are shown in Tables 4.9 and 4.10. The notations $\overline{\lambda}(H_0)$ and $\overline{\lambda}(H_1)$ refer to the arithmetic means of the eigenvalues calculated under H_0 or H_1 respectively.

Based on the tabulated observations, Figure 4.4 informally shows the relative positions of score and likelihood ratio tests. As is known for the likelihood setting the empirically corrected score statistic S(e.c) is often much smaller than the model

Table 4.9: *NTP Studies: Likelihood Wald, Score and Ratio Tests for Dose Trends (empirically corrected (e.c) and model based (m.b)).*

		W(e.c)	W(m.b)	S(e.c)	S(m.b)	G^2
		common main dose effect				
DEHP	Ext-Vis	15.24	33.11	21.93	35.05	61.48
	Ext-Skel	25.69	33.42	15.93	36.76	67.07
	Vis-Skel	24.36	34.95	21.35	41.52	67.60
EG	Ext-Vis	12.73	19.36	7.46	23.17	27.39
	Ext-Skel	24.40	25.65	23.68	46.06	55.70
	Vis-Skel	25.64	24.61	25.53	47.35	55.60
DYME	Ext-Vis	29.01	31.32	17.19	31.50	86.31
	Ext-Skel	30.82	46.11	25.93	53.27	134.98
	Vis-Skel	5.69	20.02	25.11	38.99	65.61
		different main dose effects				
DEHP	Ext-Vis	17.69	33.09	22.61	35.15	62.08
	Ext-Skel	28.87	33.41	17.51	36.76	67.10
	Vis-Skel	24.20	34.99	24.05	41.55	67.89
EG	Ext-Vis	15.72	18.69	7.79	24.09	28.43
	Ext-Skel	28.52	27.57	26.19	46.66	57.06
	Vis-Skel	36.67	26.33	25.94	48.79	58.62
DYME	Ext-Vis	21.03	27.89	17.38	36.53	91.75
	Ext-Skel	29.39	45.96	28.53	53.39	145.56
	Vis-Skel	5.74	20.02	25.19	45.37	65.62

Table 4.10: *NTP Studies: Pseudo-likelihood Wald, Score and Ratio Tests for Dose Trends.*

		$\overline{\lambda}(H_0)$	$\overline{\lambda}(H_1)$	W^*	$S^*(e.c)$	$S_a^*(m.b)$	$G_a^{*2}(H_0)$	$G_a^{*2}(H_1)$
		common main dose effect						
DEHP	Ext-Vis	1.55	1.79	17.49	20.83	20.83	21.12	18.34
	Ext-Skel	1.54	1.02	29.24	19.91	19.91	20.81	31.65
	Vis-Skel	1.58	1.08	31.78	22.67	22.67	23.71	34.47
EG	Ext-Vis	2.46	1.02	16.36	8.14	8.14	8.62	8.43
	Ext-Skel	0.77	0.43	50.23	28.99	28.99	27.94	64.59
	Vis-Skel	0.68	0.35	63.62	33.75	33.75	32.39	91.98
DYME	Ext-Vis	2.41	0.87	32.57	13.35	13.35	15.20	41.92
	Ext-Skel	1.47	1.69	26.26	33.36	33.36	36.96	32.13
	Vis-Skel	1.26	0.99	27.95	23.07	23.07	22.98	29.34
		different main dose effects						
DEHP	Ext-Vis	1.07	1.19	26.72	21.02	31.25	31.73	28.21
	Ext-Skel	1.08	0.99	30.42	23.11	29.02	30.35	32.98
	Vis-Skel	1.07	1.02	33.75	24.49	33.65	35.14	36.58
EG	Ext-Vis	1.69	0.90	17.77	8.48	12.24	13.22	24.79
	Ext-Skel	0.73	0.59	36.44	29.02	30.77	29.72	36.22
	Vis-Skel	0.57	0.38	59.56	36.02	42.11	41.36	61.80
DYME	Ext-Vis	1.58	0.85	37.02	13.68	22.66	25.50	47.63
	Ext-Skel	1.14	1.34	34.64	34.25	45.27	50.46	42.79
	Vis-Skel	0.96	0.88	31.87	23.52	31.15	30.56	33.27

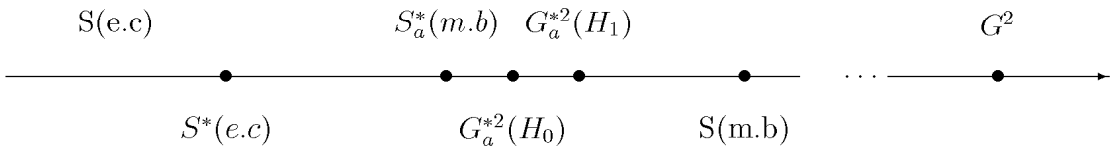

Figure 4.4: *NTP Studies: Informal Comparison of Score and Ratio Test Statistics in the Bivariate MR Model.*

based version S(m.b). While a similar observation can be made for the pseudo-score statistics $S^*(e.c)$ (playing the role of the empirically corrected statistic) and the adjusted statistic $S_a^*(m.b)$, the gap is considerably narrower and, as appears from the theory, it vanishes in the common dose effect model. Whereas the maximum likelihood ratio statistic is usually much larger than any other statistic, the same does not hold for the adjusted pseudo likelihood ratio statistic $G_a^{*2}(H_0)$, which in fact closely agrees with the adjusted pseudo-score $S_a^*(m.b)$. It should be noticed that these results correspond with our findings in Section 4.4.

As expected, the pseudo-Wald test W^* corresponds fairly well with $G_a^{*2}(H_1)$. Further note that the eigenvalues take unusually small values for the EG data (External-Skeletal;Visceral-Skeletal). This is probably due to rare events, leading to an inflated Hessian matrix.

4.6.3 Trivariate Analyses

When considering all three outcomes jointly, ML becomes prohibitively difficult to fit. Some analyses are very sensitive to initial values and take more than 10 hours to converge. Therefore, we abandoned ML and concentrated solely on the PL method, which took less than 3 minutes to converge.

For all three NTP studies, we considered (1) a model with a different dose effect per outcome and (2) a common dose effect model, both of which are tested for the null hypothesis of no dose effect. In both cases all association parameters are

Table 4.11: *NTP Studies: Pseudo-likelihood Estimates (standard errors) for Trivariate Outcomes (different main dose effects).*

Par.	DEHP	EG	DYME
β_{01}	-2.13 (0.64)	-1.64 (1.04)	-5.67 (1.16)
β_{02}	-2.38 (0.63)	-5.04 (1.75)	-2.34 (1.26)
β_{03}	-2.76 (0.72)	-0.39 (0.51)	-2.97 (0.90)
δ_1	0.14 (0.07)	0.18 (0.13)	0.15 (0.04)
δ_2	0.18 (0.04)	0.12 (0.17)	0.30 (0.06)
δ_3	0.29 (0.06)	0.20 (0.01)	0.25 (0.02)
ω_{12}	0.06 (0.25)	-0.05 (0.57)	-0.45 (0.20)
ω_{13}	0.60 (0.20)	0.11 (0.31)	0.25 (0.31)
ω_{23}	0.36 (0.29)	0.86 (0.34)	0.35 (0.31)
γ_{12}	0.11 (0.06)	0.14 (0.13)	0.07 (0.04)
γ_{13}	-0.06 (0.05)	0.08 (0.04)	-0.11 (0.05)
γ_{23}	-0.14 (0.06)	-0.09 (0.04)	0.01 (0.05)
β_{d1}	2.70 (0.66)	1.12 (0.86)	6.48 (1.26)
β_{d2}	2.63 (0.66)	3.63 (1.04)	1.66 (1.36)
β_{d3}	2.70 (0.76)	1.42 (0.19)	4.29 (0.99)

Table 4.12: *NTP Studies: Pseudo-likelihood Estimates (standard errors) for Trivariate Outcomes (common main dose effects).*

Par.	DEHP	EG	DYME
β_{01}	-2.10 (0.51)	-1.97 (0.56)	-3.89 (0.83)
β_{02}	-2.42 (0.50)	-2.96 (0.87)	-4.77 (0.87)
β_{03}	-2.74 (0.49)	-0.27 (0.55)	-3.21 (0.81)
δ_1	0.14 (0.07)	0.18 (0.13)	0.22 (0.03)
δ_2	0.18 (0.04)	0.17 (0.17)	0.25 (0.06)
δ_3	0.29 (0.05)	0.20 (0.01)	0.25 (0.02)
ω_{12}	0.06 (0.24)	-0.05 (0.57)	-0.46 (0.19)
ω_{13}	0.60 (0.20)	0.11 (0.30)	0.29 (0.30)
ω_{23}	0.36 (0.28)	0.97 (0.37)	0.28 (0.31)
γ_{12}	0.11 (0.06)	0.13 (0.13)	0.05 (0.04)
γ_{13}	-0.06 (0.05)	0.06 (0.04)	-0.09 (0.04)
γ_{23}	-0.14 (0.06)	-0.07 (0.03)	-0.03 (0.05)
β_d	2.67 (0.48)	1.50 (0.20)	4.31 (0.85)

held constant. Results of these analyses are tabulated in Tables 4.11 and 4.12 and indicate, based on Wald tests, that all dose effect parameters are significant (except for External outcomes in EG and for Visceral malformations in DYME). In addition, Tables 4.11 and 4.12 show that by fitting a relatively simple model with different dose effects for each outcome and constant association parameters, the three different main dose effect parameters in the DEHP study all seem to be relevant and of similar magnitude. This suggests that the use of a common main dose parameter is desirable, hereby increasing the efficiency (Lefkopoulou and Ryan 1993). The estimated clustering parameters δ_j $(j = 1, 2, 3)$ are all significant, except for External and Visceral malformation outcomes in the EG study. In contrast, the other association parameters often do not reach the 5% significance level.

4.6.4 Model Selection

In this section we merely exemplify how to select appropriate models for the EG, DYME and DEHP studies, using the test statistics developed in Section 4.4. The EG and DEHP studies will be re-analyzed more thoroughly in the next chapter.

Because of the indicated drawbacks of the Wald test statistic, specifically for conditional models, we concentrate on score and likelihood ratio test statistics only. The different trivariate models considered are described in Table 4.13. A summary of the model selection strategy for the two studies is given in Table 4.14.

The EG Study

For the EG chemical, the number of "events" tends to be small, especially for external and visceral malformation types. Therefore, the data do not support very complicated models. We start from a model that contains linear dose effects on all parameters (model 7). The linear dose effects on the association parameters are all non significant. Hence the model can be reduced to model 4. According to the statistics, it seems not necessary (sometimes borderline) to consider different dose effects for the main parameters. One common dose effect may be sufficient. However, since our model is conditionally specified, it is important to assess its fit to the observed malformation rates. This will be further investigated in Chapter 5.

We tabulated the adjusted pseudo-likelihood ratio test statistics, as well as the adjusted and unadjusted pseudo-score statistics, described earlier in this sec-

Table 4.13: *NTP Studies: Model Descriptions (l=linear ; q=quadratic).*

Model	Description	Par.
1	Null Model	12
2	Common (l) dose effect on main pars θ	13
3	Common (l+q) dose effect on main pars θ	14
4	Different (l) dose effects on main pars θ	15
5	Different (l) dose effects on θ, δ	18
6	Different (l) dose effects on θ, δ, ω	21
7	Different (l) dose effects on θ, δ, ω, γ	24
8	Different (l+q) dose effects on θ, δ and (l) dose effect on ω, γ pars	30
9	Different (l+q) dose effects on θ, δ, ω and (l) dose effect on γ pars	33
10	Different (l+q) dose effects on all parameters	36

Table 4.14: *NTP Studies: Model Selection.*

Model Comp.	df	$\overline{\lambda}(H_0)$	$\overline{\lambda}(H_1)$	$S^*(e.c)$	$S_a^*(m.b)$	$G_a^{*2}(H_0)$	$G_a^{*2}(H_1)$
EG							
6-7	3	1.27	0.94	7.71	5.25	5.33	7.16
5-6	3	1.32	1.13	5.06	4.56	6.38	7.45
4-5	3	1.34	1.33	5.45	2.89	2.81	2.84
2-4	2	0.66	0.72	6.21	3.82	4.19	3.85
DEHP							
1-2	1	1.83	1.26	22.09	22.09	22.62	32.88
2-3	1	1.01	0.92	5.58	5.58	5.73	6.29
DYME							
9-10	3	0.22	0.16	1.35	0.93	0.97	1.32
8-9	3	0.32	0.00	-1.07	2.61	1.43	63269.42

tion. Note that the $G_a^{*2}(H_0)$ and $S_a^*(m.b)$ statistics are similar, while $G_a^{*2}(H_1)$ is in most cases slightly less conservative. This confirms earlier results. The unadjusted pseudo-score statistic $S^*(e.c)$ yields comparable results in this case. However, Rotnitzky and Jewell (1990) note that it may suffer from computational stability problems, similar to those of the Wald test statistic. Our experience with the DYME study is in line with this statement. This will be further illustrated below.

The DYME Study

Let us start from the very complicated model with linear and quadratic dose effects on all parameters (model 10). From Table 4.14, it follows that the quadratic dose effects on the γ parameters can be omitted. However, the eigenvalues are very small, indicating a numerical stability problem. This is even more extreme, when we compare model 9 with model 8. Since events are rare, and the occurence of several malformations simultaneously is even rarer, the Hessian matrix is inflated. The adjustment $\overline{\lambda}(H_1)$ is therefore nearly zero, leading to an inflated $G_a^{*2}(H_1)$. This feature is well known for and shared with the Wald statistic, and is the reason why one often prefers score statistics. However, $G_a^{*2}(H_1)$ can be regarded as an internal diagnostic as well. Besides, $S^*(e.c)$ takes a negative value, which confirms our former statement that the compounded matrix $J^{\gamma\gamma}\Sigma_{\gamma\gamma}^{-1}J^{\gamma\gamma}$ can suffer badly from instabilities and that even the sign of the eigenvalues is affected. Our discussion suggests that these models are too complex, given the structure of the data and that more parsimonious models ought to be considered.

The DEHP Study

All test statistics in Table 4.14 lead to the conclusion that the null model (model 1), containing no dose effect, is clearly unacceptable, compared to model 2 which assumes a common dose effect on the main effect parameters. This could have been anticipated from the data in Table 2.1 which suggest clear exposure effects of DEHP on the three malformation outcomes: external, visceral and skeletal. Adding a common quadratic dose effect (model 3) ameliorates the fit significantly. In an attempt to improve the fit even further we might consider more complicated models. This will be further discussed in Chapter 5.

4.7 Asymptotic Relative Efficiency

One might expect that a loss of efficiency is the price to pay for computational ease. However, in Chapter 3 we showed that the ARE of PL versus ML equals 1 for all saturated exponential models, i.e. models of the form proposed by Cox (1972). In order to study the ARE for the dose trend in multivariate clustered outcomes, we follow the suggestion of Rotnitzky and Wypij (1994) and consider the same settings as in Section 4.5.1. Again, we restrict attention to bivariate binary data with maximum cluster size of 10, due to prohibitive time requirements of ML. The main effects are modelled as $\theta_{ij} = \beta_{0j} + \beta_d d_i$ $(j = 1, 2)$, i.e., a common main dose effect is assumed, and all association parameters are assumed to be constant.

A grid of several parameters is explored. The AREs of PL and ML are displayed in Table 4.15. The table shows that the AREs are very high for the lower values of the common background rate parameter β_0 (e.g., -5 to -3) and decrease with increasing dose effect when β_0 is held constant. When β_0 approaches zero, the ARE shows a non-monotonic behaviour as function of β_d, except for the zero association vector, and decreases with strength of association. In the case of a zero association vector, the ARE always decreases with increasing β_d, independent of the value of the common background rate. These conclusions are well in agreement with those found earlier in Chapter 3 for univariate clustered binary data. There, the ARE was found to be very high for low background rates, which are frequently observed. In addition it was found that the ARE decreases, when the dose effect increases for either a zero association or a low background rate.

When β_0 equals zero, the ARE decreases with strength of association. Table 4.16 shows the AREs for a zero background rate parameter and $(\delta_1, \delta_2) = (.2, .2)$ The ARE steeply decreases when ω or γ increase. This is especially true in the absence of a dose effect. The ARE increases as β_d increases within the range of positive ω and γ parameters. When γ equals 0, the ARE decreases for an increasing association parameter ω. Again, this is very marked in the case of a zero dose parameter. When γ is negative, the ARE is fairly high for all values of ω. Again, these conclusions are well in agreement with those found earlier for univariate clustered binary data. In that setting it was shown that the AREs are very high for low background rates and pseudo-likelihood performs relatively poorly for a background rate of 50%, no dose effect and a high within-cluster association. Furthermore, the ARE decreases, when

Table 4.15: *Simulation Results: Asymptotic Relative Efficiencies of Pseudo-likelihood versus Maximum Likelihood for the bivariate MR model.*

	δ_1	0	.20	.20	.20	.20	.20
	δ_2	0	.20	.20	.20	.20	.20
	ω_{12}	0	0	.10	.50	.10	.50
	γ_{12}	0	0	.05	.05	.10	.10
(β_{01}, β_{02})	β_d						
$(-5, -5)$	0.0	1.000	1.000	1.000	1.000	1.000	1.000
	1.5	0.999	1.000	1.000	1.000	1.000	1.000
	3.0	0.990	0.999	0.999	0.999	0.999	0.999
$(-3, -3)$	0.0	1.000	0.999	0.999	0.999	0.999	0.999
	1.5	0.979	0.997	0.998	0.998	0.998	0.998
	3.0	0.942	0.949	0.887	0.867	0.865	0.855
$(-2, -2)$	0.0	1.000	0.999	0.999	0.998	0.999	0.998
	1.5	0.959	0.966	0.971	0.973	0.974	0.973
	3.0	0.936	0.949	0.897	0.863	0.844	0.822
$(-1, -1)$	0.0	1.000	0.980	0.989	0.986	0.987	0.979
	1.5	0.976	0.942	0.879	0.842	0.818	0.790
	3.0	0.936	0.949	0.897	0.863	0.844	0.822
$(0, 0)$	0.0	1.000	0.856	0.551	0.312	0.252	0.127
	1.5	0.967	0.947	0.869	0.815	0.809	0.768
	3.0	0.943	0.949	0.887	0.867	0.865	0.855

Table 4.16: *Simulation Results: Asympotic Relative Efficiencies of Pseudo-likelihood versus Maximum Likelihood for the Bivariate MR Model with a Zero Background Rate Parameter Vector.*

ω	β_d	γ			
		-.1	0	.05	.1
0	0.0		0.857		
	1.5		0.948		
	3.0		0.949		
.1	0.0	0.984	0.830	0.551	0.252
	1.5	0.936	0.939	0.869	0.810
	3.0	0.966	0.937	0.887	0.865
.5	0.0	0.969	0.669	0.312	0.127
	1.5	0.971	0.886	0.815	0.768
	3.0	0.982	0.893	0.868	0.856

the dose effect increases for either a zero association or a low background rate. For a 50% background rate, the ARE decreases rather dramatically when the association becomes stronger or when dose effect decreases.

In conclusion, the efficiency loss of PL is very mild, especially for commonly encountered parameter values. In most settings, such a slight loss of efficiency will be well worth the gain in computational ease. A similar study confirmed that, based on their ARE, PL(1) and PL(2) can be considered roughly equivalent for most practical purposes. Table 4.17 shows that their AREs all vary around one. As can be anticipated, the AREs of PL(1) versus PL(2) decrease slighly with the strength of the association parameter ω, but in the worst case considered of $(\boldsymbol{\beta}_0 = 0, \beta_d = 0)$ and a positive association vector $(\delta_1, \delta_2, \omega_{12}, \gamma_{12}) = (.2, .2, 1.0, .05)$ no values smaller than .83 were observed.

4.8 Conclusion

In this chapter we have shown that pseudo-likelihood estimation is a very attractive alternative for maximum likelihood in the case of clustered multivariate binary outcomes, analyzed with the exponential family model of Molenberghs and Ryan (1999). The procedure becomes particularly useful for larger cluster sizes, where full maximum likelihood estimation is hampered, due to computing time requirements. In contrast, the pseudo-likelihood estimation method converges quickly with only minor efficiency losses, especially for a range of realistic parameter settings. Moreover, it is a natural estimation procedure. Often it can be derived directly from a probability mass function, as was the case here. Should one choose to specify a set of conditional densities directly, then compatibility conditions (Arnold and Strauss 1991) can be imposed to ensure existence of an underlying density.

To overcome the absence of inferential test procedures, we also proposed score and likelihood ratio tests within the pseudo-likelihood framework. They are easy to calculate, exhibit nice satisfactory behaviour and provide the necessary tools for model selection. These desirable properties were exemplified using data from the NTP developmental toxicity studies. In the next chapter we will perform analyses of developmental toxicity data in their own right.

Table 4.17: *Simulation Results: Asymptotic Relative Efficiencies of PL(1) versus PL(2) for the Bivariate MR Model.*

	δ_1	0	.2	.2	.2	.2
	δ_2	0	.2	.2	.2	.2
	ω_{12}	0	0	.1	.5	1.0
	γ_{12}	0	0	.05	.05	.05
(β_{01}, β_{02})	β_d					
$(-5, -5)$	0.0	1.000	1.000	1.000	1.000	1.000
	1.5	1.000	1.000	1.000	1.000	1.000
	3.0	0.999	1.000	1.000	0.997	0.999
$(-3, -3)$	0.0	1.000	1.000	1.000	0.999	0.999
	1.5	0.999	1.000	0.999	0.999	0.997
	3.0	0.997	1.001	0.996	0.973	0.938
$(-2, -2)$	0.0	1.000	1.000	1.000	0.999	0.997
	1.5	0.999	1.000	0.999	0.994	0.981
	3.0	0.999	0.997	0.991	0.978	0.953
$(-1, -1)$	0.0	1.000	1.000	0.999	0.993	0.975
	1.5	1.000	0.998	0.996	0.985	0.950
	3.0	0.999	0.997	0.991	0.978	0.953
$(0, 0)$	0.0	1.000	1.000	0.997	0.943	0.829
	1.5	0.999	0.999	0.997	0.978	0.943
	3.0	0.997	1.001	0.996	0.973	0.938

Chapter 5

Risk Assessment and Fractional Polynomials

5.1 Introduction

In this chapter we investigate the toxicity of DEHP and EG in mice. The data are described in Sections 2.1.1 and 2.1.5. Our primary goal in this chapter will be to determine safe levels of exposure for these studies, based on appropriate dose-response models. Since the data involve a vector of malformation indicators, flexible models for clustered, multivariate binary data are required. Williams and Ryan (1996) summarize a variety of reasons why multivariate methods for dose-response modelling are important, thereby controlling for several adverse events simultaneously. Incorporating multiple outcomes helps: (1) to control the Type I error rate, which can become inflated when several tests for dose effects are conducted across several univariate models, (2) to investigate relationships among adverse outcomes, (3) to more realistically quantify overall risk of "any adverse event" that can be used for the purpose of risk assessment. The exponential family likelihood model of Molenberghs and Ryan (1999) easily deals with multivariate outcomes. Other advantages of this model are the flexibility with which both main effects and associations can be modelled, and the absence of constraints on the parameter space which eases interpretability. Further it provides a natural framework for quantitative risk assessment. Present approaches based on marginal models estimate benchmark doses based on the marginal probability of a single offspring being affected, although litter

based versions can be considered as well. Declerck et al. (1999) compare litter based versus fetus based risks in the univariate case. 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 likelihood basis of the MR model allows calculation of quantities such as the probability that at least one littermate is affected. While they could in principle be calculated from a fully specified marginal model, computations tend to be involved, since fitting these models is hampered by lengthy computations and/or parameter restrictions (Molenberghs, Declerck and Aerts 1998 and Aerts, Declerck and Molenberghs 1997). A detailed comparison of litter-based versus fetus-based risk assessment for the multivariate MR model is the subject of ongoing research .

An important goal of developmental toxicity studies is to perform risk assessment, i.e. to set safe limits for human exposure, based on the fitted model (Crump 1984). To this end, models should fit the data well. This has implications for both the model family chosen, as well as for the form of the linear predictors. Since classical polynomial predictors are often of poor quality, especially when low dose extrapolation is envisaged, there is a clear need for alternative specifications of the predictors describing main effects and associations.

We describe how to find appropriate fractional polynomial predictors in Section 5.2. In Section 5.3, we construct several candidates for dose-response models by modelling the natural parameters $\boldsymbol{\Theta}$ in model (4.2) as fractional polynomial functions of dose (Royston and Altman 1994). We use fractional polynomials, since they provide more flexibly shaped curves than conventional polynomials. Estimation is by pseudo-likelihood rather than maximum likelihood, because of the latter's excessive computational requirements. Once a suitable model is selected, it can serve as basis for quantitative risk assessment. This is illustrated in Section 5.4.

5.2 Fractional Polynomial Predictors

For risk assessment to be reliable, models should fit the data well in all respects. Although classical polynomial predictors are still very customary, they are often inadequate. Perhaps an obvious alternative are non-linear predictor functions (Davidian and Giltinan 1995). Such 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. The advantages and disadvantages of Bayesian methods in this context are currently under investigation (Declerck 1999). 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

$$\beta_0 + \sum_{j=1}^{m} \beta_j d^{p_j},$$

where the β_j are regression parameters and $d^0 \equiv \ln(d)$ and the powers $p_1 < \cdots < p_m$ are positive or negative integers or fractions (Royston and Altman 1994). Royston and Altman (1994) argue that polynomials with degree higher than 2 are rarely required in practice and further restrict the powers of dose 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

- $(1/d^2, 1/d) : \beta_0 + \beta_1 1/d + \beta_2 1/d^2$,
- $(1/d, 1/\sqrt{d}) : \beta_0 + \beta_1 1/\sqrt{d} + \beta_2 1/d$,
- $(\sqrt{d}, d) : \beta_0 + \beta_1 \sqrt{d} + \beta_2 d$, and
- $(d, d^2) : \beta_0 + \beta_1 d + \beta_2 d^2$;

(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.

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).

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 d is needed to ensure positivity (e.g., $d+1$).

5.3 Modelling the Dose-response Relationship

As suggested earlier in Section 5.1, it is important to incorporate multiple outcomes at the same time. In that case, maximum likelihood becomes prohibitive and we are restricted to the pseudo-likelihood estimation method. In order to select appropriate dose-response models, we can rely on the test statistics, introduced in Chapter 4 and proposed by Geys, Molenberghs and Ryan (1999).

5.3.1 EG Study

For the EG data it was shown in Section 4.6.4 that a model with different linear dose trends on all parameters can be reduced to a model with a common linear dose trend on the main effect parameters (θ) only. Hence, the association parameters seem to be unaffected by dose. Now, the question arises whether these provisional models provide an adequate fit to the data. This is important when quantitative risk assessment is envisaged. A key tool to gain insight in a model is the qualitative study of the dose-response relationship. Given the number of viable fetuses n_i, the probability of observing at least one abnormal fetus in a cluster is $1-\exp(-A_{n_i}(\mathbf{\Theta}_i))$. Integrating over all possible values of n_i, we obtain the following *risk function* (introduced in Section 1.2):

$$r(d) = \sum_{n_i=0}^{\infty} P(n_i)[1 - \exp(-A_{n_i}(\mathbf{\Theta}_i))], \tag{5.1}$$

where $P(n_i)$ is the probability of observing n_i viable fetuses in a pregnant dam (we use the empirical distribution of $P(n_i)$). One of the major challenges of a teratology study lies in characterizing the relationship between dose and event probability

(5.1) by means of a dose-response curve. Figures 5.1(a)–(c) show the observed frequencies of malformed litters at the selected dose levels for external, visceral and skeletal malformations and the three (univariate) dose-response curves for models with constant association and a linear d trend on the main effects. Figure 5.1(d) shows the trivariate dose-response curve based on all three outcomes jointly and with a common linear dose trend on the main effect parameters. Clearly, the fit is not satisfactory. Of course, one can try to further improve the fit by imposing quadratic (or higher order) dose effects. Often, however, this is inadequate. Alternatively, the fractional polynomial approach can be adopted. Let us contrast both ways of extending the simple model for the skeletal malformation indicator. Consider (1) a conventional quadratic polynomial in dose ($\theta = \beta_0 + \beta_1 d + \beta_2 d^2$), and (ii) a fractional polynomial ($\theta = \beta_0 + \beta_1 \sqrt{d} + \beta_2 d$). The clustering parameter is kept constant in both cases. Figure 5.2 plots the fitted malformation rates for both these models on the observed ones. Note that the conventional polynomial (quadratic) approach is clearly inferior.

Therefore, in the analysis of the EG data, the following strategy is adopted (see also Geys, Molenberghs and Ryan 1999 and Geys et al. 1999a). First, we select a suitable set of dose transformations for each of the three developmental outcomes (skeletal, visceral and external) separately. The resulting set of transformations is then used to construct more elaborate (multivariate) models that can be scrutinized further by means of formal test statistics.

Table 5.1 shows that for skeletal malformation outcomes, the fractional polynomials approach suggests that a single effect of dose ($m = 1$), whether represented by $1/d^2$, $1/d$, $1/\sqrt{d}$, $\ln(d)$, $\sqrt{d}$, d, d^2 or d^3 is unacceptable as opposed to two effects simultaneously ($m = 2$) to model the main effect parameter (θ). Table 5.1 tabulates only the four quadratics in powers of d and a quadratic in $\ln(d)$. None of the other combinations provided a substantial improvement. The quadratic represented by $(\sqrt{d}, d)$ yields the highest pseudo-likelihood (no fit could be obtained for the quadratic represented by $(1/d, 1/\sqrt{d})$). A similar approach, applied to the clustering parameter, suggests that no dose effect needs to be incorporated. For both the external and visceral malformation outcomes, the main effect parameters are best modelled in a linear fashion ($m = 1$) in $\sqrt{d}$, while the clustering can be assumed constant. These univariate findings then served as a basis to construct more elaborate, trivariate models, presented in Table 5.2. The subscripts 1, 2 and

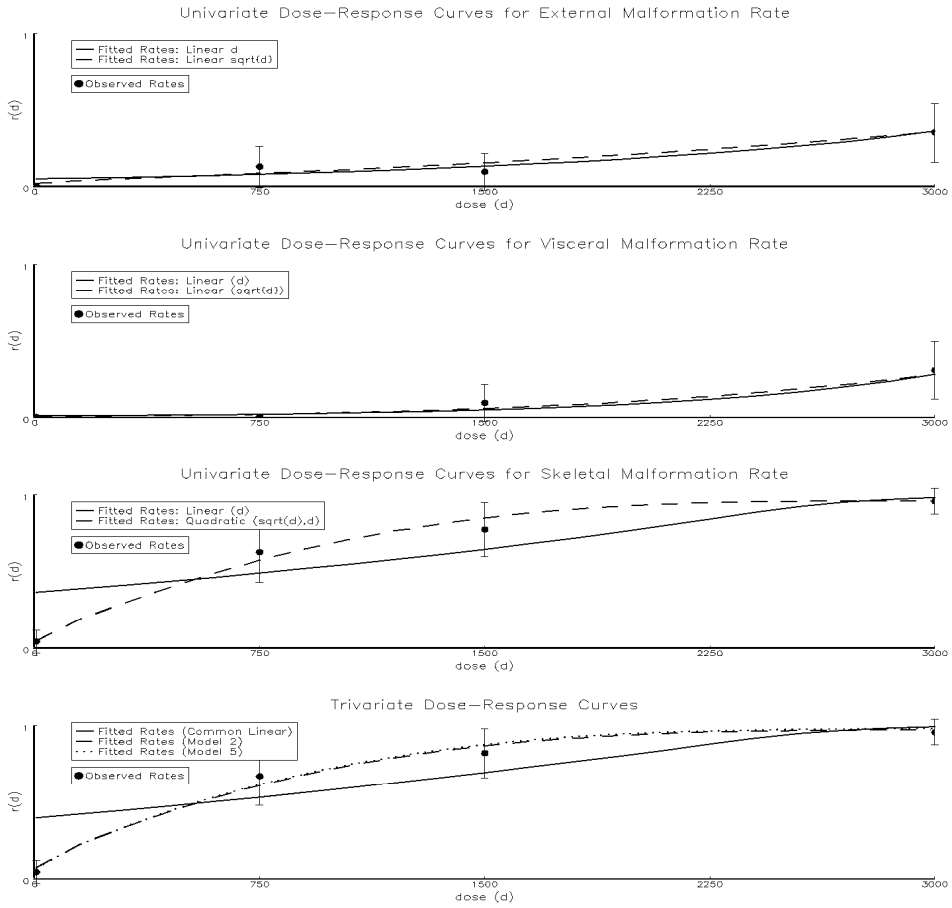

Figure 5.1: *EG Study: From Top to Bottom, (a) Univariate dose response curves for external malformations based on models with d and $\sqrt{d}$ trends on main effect parameters θ and constant clustering parameters δ, (b) Univariate dose response curves for visceral malformations based on models with d and $\sqrt{d}$ trends on main effect parameters θ and constant clustering parameters δ, (c) Univariate dose response curves for skeletal malformations based on models with a linear d and quadratic $(\sqrt{d}, d)$ trend on main effect parameters θ and constant clustering parameters δ, (d) Trivariate dose response curves based on model with common linear dose trend and models 2 and 5.*

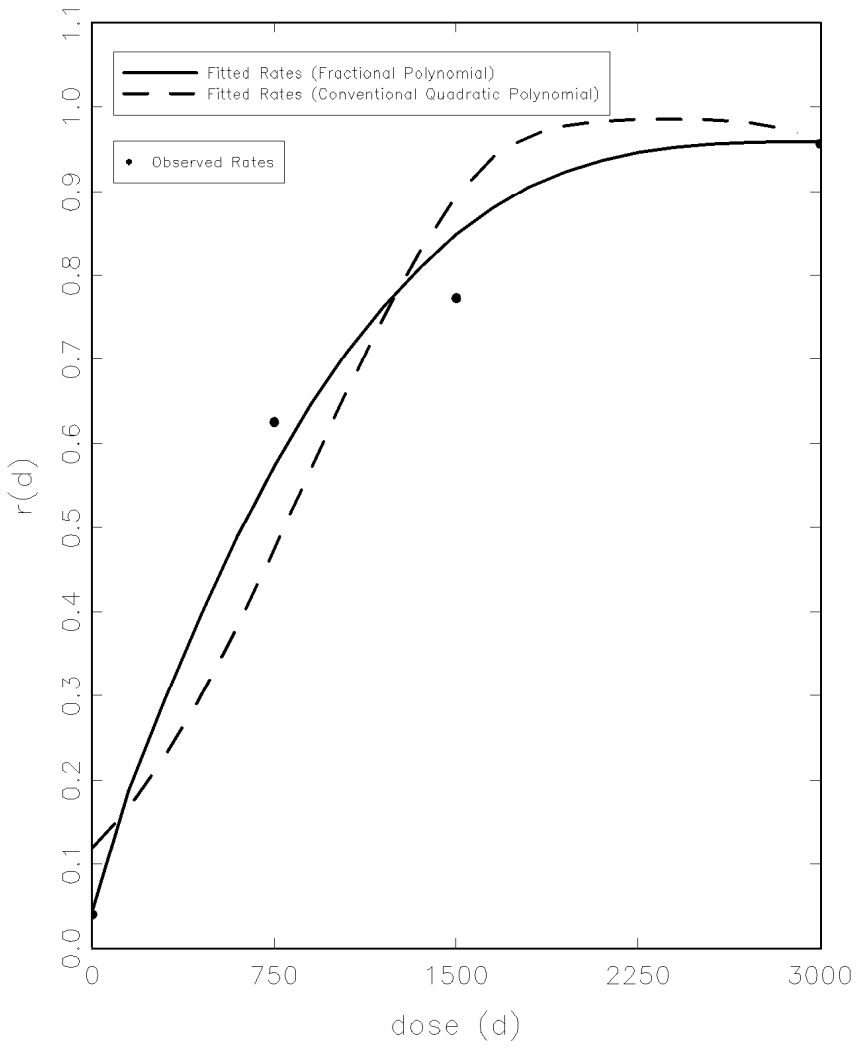

Figure 5.2: *EG Study: Observed and Fitted Skeletal Malformation Rates using a Univariate MR Model with the Main Effect Parameter Modelled as Function of Dose by (i) a Conventional Quadratic Polynomial, and (ii) a Fractional Polynomial.*

Table 5.1: *EG Study: Log Pseudo-likelihood Values for the Univariate MR Model, with Given Fractional Polynomial Dose Trends on the Skeletal Main Effect Parameter. The Clustering Parameter is Assumed Constant.*

$m = 1$		$m = 2$	
transformation	$p\ell$	transformation	$p\ell$
$1/d^2$	-337.59		
$1/d$	-337.48	$(1/d, 1/d^2)$	-332.44
$1/\sqrt{d}$	-338.47	$(1/d, 1/\sqrt{d})$	
$\ln(d)$	-339.46	$(\sqrt{d}, d)$	-331.96
$\sqrt{d}$	-335.99	(d, d^2)	-333.79
d	-341.29	$(\ln(d), \ln^2(d))$	-333.56
d^2	-345.28		
d^3	-346.51		

3 refer to external, visceral and skeletal malformations, respectively.

Since the EG data do not support really complicated models, the most complex model we considered (Model 1) allows different $\sqrt{d}$ trends on the external, visceral, and skeletal main effect parameters, an additional d trend on the skeletal main effect parameter:

$$\begin{aligned} \theta_1 &= \beta_{01} + \beta_{\sqrt{d}1}\sqrt{d}, \\ \theta_2 &= \beta_{02} + \beta_{\sqrt{d}2}\sqrt{d}, \\ \theta_3 &= \beta_{03} + \beta_{\sqrt{d}3}\sqrt{d} + \beta_{d3}d, \end{aligned}$$

and different $\sqrt{d}$ trends on the clustering parameters (δ). All other association parameters (ω and γ) are kept constant. This model was then further scrutinized by means of formal test statistics. From Table 5.2 it is clear that the clustering parameters do not depend on $\sqrt{d}$ (confirming the preliminary findings). Hence, Model 2 is now selected. The d trend on the skeletal main effect parameter cannot be removed (Model 2 vs Model 3), nor can the different $\sqrt{d}$ trends on the external, visceral and skeletal main effects be replaced by a common trend (Model 2 vs Model

Table 5.2: *EG Study: Model Selection (All effects are constant except the ones mentioned).*

Model	Description	no. Pars.
1	Different $\sqrt{d}$ trends on θ_1, θ_2, θ_3	19
	+ d trend on θ_3	
	+ different $\sqrt{d}$ trends on δ_1, δ_2, δ_3	
2	Different $\sqrt{d}$ trends on θ_1, θ_2, θ_3	16
	+ d trend on θ_3	
3	Different $\sqrt{d}$ trends on θ_1, θ_2, θ_3	15
4	Common $\sqrt{d}$ trend on θ_1, θ_2, θ_3	14
	+ d trend on θ_3	
5	Different $\sqrt{d}$ trends on θ_1, θ_2, θ_3	10
	+ d trend on θ_3	
	+ no ω and γ pars.	

Comparison	$S^*(e.c.)$ (p-value)	$S_a^*(m.b.)$ (p-value)	$G_a^{*2}(H_0)$ (p-value)	$G_a^{*2}(H_1)$ (p-value)
1–2	3.77 (0.29)	2.84 (0.42)	2.84 (0.42)	4.06 (0.26)
2–3	15.19 (0.00)	15.19 (0.00)	18.55 (0.00)	10.68 (0.00)
2–4	5.76 (0.05)	8.03 (0.02)	8.05 (0.02)	9.09 (0.01)
2–5	7.71 (0.26)	9.18 (0.16)	9.68 (0.14)	10.01 (0.12)

4). Therefore, Model 2 was selected at this point. Table 5.3 shows PL parameter estimates for this model.

Table 5.3: *EG Study: Pseudo-likelihood Estimates (standard errors) for Two Selected Models.*

Effect	Outcome	Parameter	Estimate (s.e.)	
			Model 2	Model 5
θ Main	Ext.	β_{01}	-2.27 (1.16)	-3.58 (1.10)
		$\beta_{\sqrt{d}1}$	1.71 (0.99)	3.07 (0.97)
	Visc.	β_{02}	-6.98 (2.36)	-7.17 (2.26)
		$\beta_{\sqrt{d}2}$	5.54 (1.71)	5.83 (1.96)
	Skel.	β_{03}	-2.81 (0.95)	-3.61 (0.84)
		$\beta_{\sqrt{d}3}$	7.73 (2.32)	7.59 (2.22)
		β_{d3}	-4.01 (1.50)	-3.89 (1.43)
δ Clustering	Ext.	δ_1	0.18 (0.13)	0.29 (0.06)
	Visc.	δ_2	0.12 (0.17)	0.22 (0.09)
	Skel.	δ_3	0.18 (0.01)	0.19 (0.01)
ω Assoc.	Ext.-Visc.	ω_{12}	-0.06 (0.57)	
	Ext.-Skel.	ω_{13}	0.11 (0.29)	
	Skel.-Visc.	ω_{23}	0.81 (0.34)	
γ Assoc.	Ext.-Visc.	γ_{12}	0.14 (0.13)	
	Ext.-Skel.	γ_{13}	0.08 (0.04)	
	Skel.-Visc.	γ_{23}	-0.08 (0.04)	

Next, Model 2 was used to construct a dose-response curve representing the probability of observing an adverse event as a function of dose (d). The risk function $r(d)$ was calculated using PL parameter estimates (due to excessive computational requirements for ML). Figures 5.1(a) and (b) show the (univariate) dose-response curves for models with constant association and $\sqrt{d}$ trends on the main effects. The dose-response curve for skeletal malformation (Figure 5.1(c)) is based on the

quadratic $(\sqrt{d}, d)$-model for the main effect parameter and constant clustering. Figure 5.1(d) shows the trivariate dose-response curve based on all three outcomes simultaneously (Model 2). Both the univariate as well as the trivariate fits are excellent. All curves gradually increase when dams are exposed to larger quantities of the toxic substance, before finally reaching an asymptote. Note that there is a fundamental difference in the dose-response curves for external and visceral outcomes on the one hand, and skeletal malformation on the other, the latter of which shows a much more pronounced dose-response relationship. This is in line with the data in Table 2.5. Further, the joint dose-response curve is clearly driven by skeletal malformation.

These observations incite to explore additional model simplifications. Candidates for removal are the dose trends on the external and visceral outcomes, as well as one or more assocation parameters. Table 5.2 shows that the ω and γ association parameters are redundant (Model 2 vs Model 5). However, the clustering parameters could not be removed from the model without a substantial decrease in fit. Futhermore, the dose trends on the external and visceral main effects are also important. Since the goal of selecting a well-fitting model is to perform risk assessment, merely concentrating on formal model selection criteria is insufficient. Arguably, the excellent fit of the dose-response curves which have been achieved, should not be compromised. However, Figure 5.1 shows that the simplified Model 5 produces essentially the same dose-response curve as Model 2. Therefore, Model 5 will be treated as our final model. It can thus serve as basis for quantitative risk assessment, aiming at determination of a low-risk dose level. The parameter estimates are tabulated in Table 5.3. These have a conditional interpretation. For example it can be derived from (4.4) that, in Model 5, the main effect parameter θ_{ij} can be interpreted as the conditional logit, associated with an additional malformation of type j in the ith cluster, given the cluster contains already $z_{ij} = (n_i + 1)/2$ fetuses with malformations of that type. Similarly δ_{ij} can be interpreted as the conditional log odds ratio for a pair of fetuses, exhibiting malformation j, given all other outcomes. Thus, if interest is in marginal quantities such as the dose-response curve, they have to be obtained as non-linear functions of the parameters. Computationally, this is feasible. Conditional questions can, on the contrary, be answered in terms of linear functions of the parameters.

Table 5.4: *DEHP Study: Log Pseudo-likelihood Values for the Univariate MR Model, with Given Fractional Polynomial Dose Trends on the External Main Effect Parameter. The Clustering Parameter is Assumed Constant.*

$m = 1$		$m = 2$	
transformation	$p\ell$	transformation	$p\ell$
$1/d$	-164.26	$(1/d, 1/d^2)$	-162.91
$1/\sqrt{d}$	-164.79	$(1/d, 1/\sqrt{d})$	
$\ln(d)$	-165.39	$(\sqrt{d}, d)$	-162.42
$\sqrt{d}$	-163.51	(d, d^2)	-164.20
d	-166.82	$(\ln(d), \ln^2(d))$	-163.95

5.3.2 DEHP Study

We now consider the DEHP Study (see also Geys et al. 1999a). We first select a suitable set of dose transformations for each of the three developmental outcomes. Table 5.4 suggests that for external malformation outcomes the main effect of dose (θ) may be modelled using a first or second degree model ($m = 2$). Since our aim is to find a suitable trivariate model first, on which more formal model selection criteria can be applied later, we have chosen the second degree model. Among all second degree models, the quadratic represented by $(\sqrt{d}, d)$ yields the highest pseudo-likelihood for external malformations. However, since this model yields the lowest pseudo-likelihood for visceral and skeletal malformations (although the difference is not significant), we preferred to use the model represented by $(1/d, 1/d^2)$. Next, we consider a univariate MR model on the external malformation indicators in the DEHP study, and model the main effect parameter as: (1) a conventional quadratic polynomial in dose ($\theta = \beta_0 + \beta_1 d + \beta_2 d^2$), and (2) a fractional polynomial ($\theta = \beta_0 + \beta_1 \frac{1}{d+1} + \beta_2 \frac{1}{(d+1)^2}$). The clustering parameter is kept constant in both cases. Figure 5.3 (a) plots the fitted malformation rates for both these models on the observed ones. Clearly, the fractional polynomial approach yields only a marginally improved fit. It is comforting, however, that when the fractional polynomial approach is strictly speaking not necessary, it reduces to a standard polynomial approach. Still, the difference between both may become important when doing

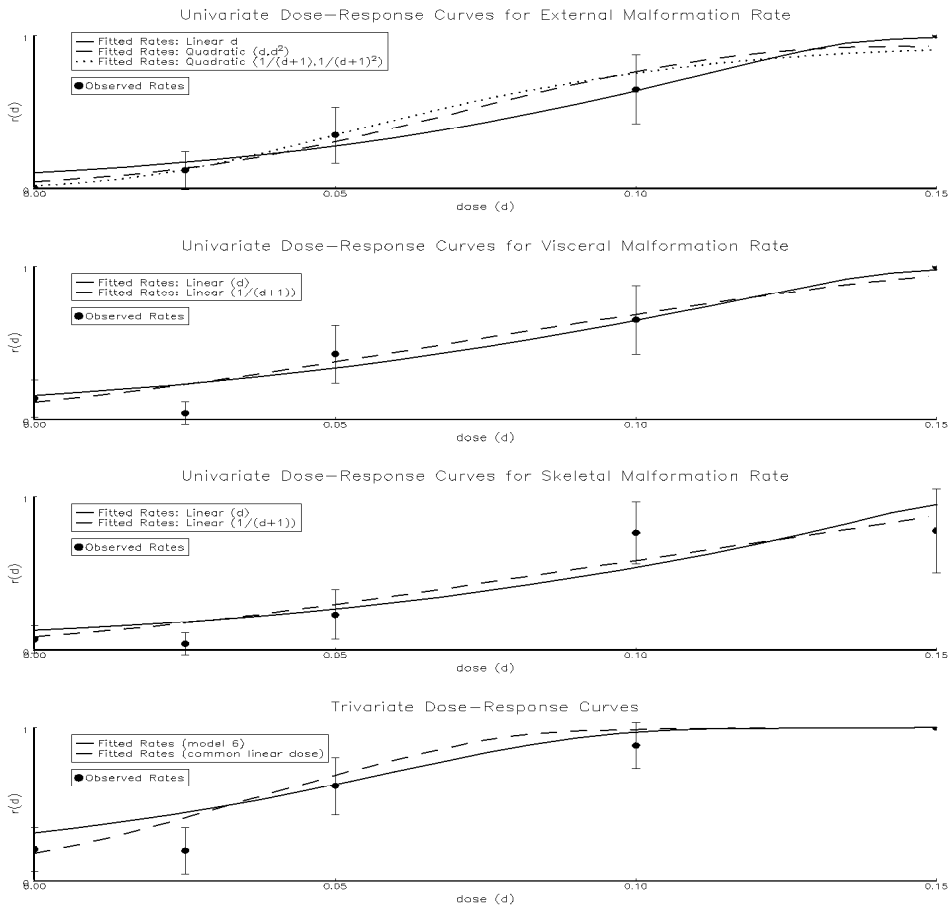

Figure 5.3: *DEHP Study: From Top to Bottom, (a) Univariate dose response curves for external malformations based on models with a linear* d*, a quadratic* (d, d^2) *and a quadratic* $(1/(d+1), 1/(d+1)^2)$ *trend on main effect parameters* θ *and constant clustering parameters* δ*, (b) Univariate dose response curves for visceral malformations based on models with* d *and* $1/(d+1)$ *trends on main effect parameters* θ *and constant clustering parameters* δ*, (c) Univariate dose response curves for visceral malformations based on models with* d *and* $1/(d+1)$ *trends on main effect parameters* θ *and constant clustering parameters* δ*, (d) Trivariate dose response curves based on model with common linear dose trend and model* 6.

extrapolation to very low doses. Therefore, we continue to use the fractional polynomial approach. For visceral and skeletal outcomes, a linear model in $1/(d+1)$ suffices. Moreover, Figures 5.3(b) and (c) show that this transformation results in a slightly better fit than a conventional linear model. The clustering parameters are most appropriately modelled as a constant.

Based on the above preliminary selection, we can construct more elaborate, trivariate models presented in Table 5.5. The most complex model we consider (Model 1) allows different $1/d$ and $1/d^2$ trends on all three main effect parameters and different $1/d$ trends on the clustering parameters. All other association parameters are assumed constant. This model can now be investigated further by means of formal test statistics. Table 5.5 suggests that the dose trends on δ_1, δ_2 are non-significant (Model 1 vs Model 2). After removal of the $\frac{1}{d^2}$ trends from the main effect parameters θ_2, θ_3 (Model 2 vs Model 4), the dose trend on δ_3 can be removed as well (Model 4 vs Model 6). The quadratic trend on the external main effect parameter, θ_1, cannot be removed without a substantial decrease in fit. Nor can any of the remaining dose trends on the other parameters. Therefore, we accept Model 6 as final model. Table 5.6 shows PL parameter estimates for this model. Figure 5.3(d) shows the resulting dose-response curve. For comparison, we have also plotted a model with a common linear dose trend on the main effect parameters.

5.4 Risk Assessment

Dose-response curves, such as the ones obtained in Section 5.3, can serve as basis for quantitative risk assessment, aiming at determination of a low-risk dose level, where risk is defined as the probability of an adverse outcome resulting from the dose. Several authors have discussed the use of dose-response models to characterize risk at low doses. Crump (1984) advocates fitting a reasonably flexible dose-response model to the data and then using the estimated model to find the dose corresponding to a specified level of increased risk over background, usually referred to as benchmark dose (BMD). The increased risk over background, $r^*(d)$, defines the actual level of risk to which extrapolation is targeted. It is therefore a very important quantity from a regulatory point of view (Williams and Ryan 1996). As mentioned in Section 1.2, two common safe dose levels that can derived from $r^*(d)$, by solving $r^*(d) = q\%$, are the benchmark dose (BMD_q) and a lower limit: the lower effective dose (LED_q).

Table 5.5: *DEHP Study: Model Selection (All effects are constant except the ones mentioned.)*

Model	Description	no. Pars.
1	Different $\frac{1}{d}$ and $\frac{1}{d^2}$ trends on θ_1, θ_2, θ_3	21
	+ different $\frac{1}{d}$ trends on δ_1, δ_2, δ_3	
2	Different $\frac{1}{d}$ and $\frac{1}{d^2}$ trends on θ_1, θ_2, θ_3	19
	+ $\frac{1}{d}$ trend on δ_3	
3	Different $\frac{1}{d}$ and $\frac{1}{d^2}$ trends on θ_1, θ_2, θ_3	18
4	Different $\frac{1}{d}$ trends on θ_1, θ_2, θ_3	17
	$\frac{1}{d^2}$ trend on θ_1	
	$\frac{1}{d}$ trend on δ_3	
5	Different $\frac{1}{d}$ trends on θ_1, θ_2, θ_3	16
	$\frac{1}{d}$ trend on δ_3	
6	Different $\frac{1}{d}$ trends on θ_1, θ_2, θ_3	16
	$\frac{1}{d^2}$ trend on θ_1	

Comparison	$S^*(e.c.)$ (p-value)	$S^*_a(m.b.)$ (p-value)	$G^{*2}_a(H_0)$ (p-value)	$G^{*2}_a(H_1)$ (p-value)
1–2	1.75 (0.42)	1.95 (0.38)	2.03 (0.36)	2.08 (0.35)
2–3	3.19 (0.07)	3.19 (0.07)	3.07 (0.08)	4.87 (0.03)
2–4	1.06 (0.59)	1.17 (0.56)	1.10 (0.58)	1.50 (0.47)
4–5	5.16 (0.02)	5.16 (0.02)	5.96 (0.01)	7.17 (0.01)
4–6	2.29 (0.13)	2.29 (0.13)	2.37 (0.12)	2.16 (0.14)

Table 5.6: *DEHP Study: Pseudo-likelihood Estimates (standard errors) for the Final Model.*

Effect	Outcome	Parameter	Estimate (s.e.)
θ Main	Ext.	β_{01}	-5.14 (3.85)
		$\beta_{(\frac{1}{d})1}$	19.93 (11.39)
		$\beta_{(\frac{1}{d^2})1}$	-18.68 (7.89)
	Visc.	β_{02}	2.72 (0.78)
		$\beta_{(\frac{1}{d})2}$	-5.29 (1.37)
	Skel.	β_{03}	2.56 (1.05)
		$\beta_{(\frac{1}{d})3}$	-5.50 (1.89)
δ Clustering	Ext.	δ_1	0.15 (0.06)
	Visc.	δ_2	0.18 (0.04)
	Skel.	δ_3	0.28 (0.05)
ω Assoc.	Ext.-Visc.	ω_{12}	0.09 (0.22)
	Ext.-Skel.	ω_{13}	0.59 (0.19)
	Skel.-Visc.	ω_{23}	0.37 (0.27)
γ Assoc.	Ext.-Visc.	γ_{12}	0.11 (0.05)
	Ext.-Skel.	γ_{13}	-0.04 (0.05)
	Skel.-Visc.	γ_{23}	-0.13 (0.06)

These will be used subsequently.

5.4.1 EG Study

For the EG study, the BMD_{05} and LED_{05} for Models 2 and 5 are displayed in Table 5.7. We also added the corresponding quantities, calculated from univariate versions of our model, applied to external, visceral, and skeletal malformation, as well as to the collapsed outcome defined as "any malformation". Finally, risk assessment based on logistic regression applied to an indicator for affected litter has been added. Clearly, there is very little difference between the safe dose levels based on Models 2 and 5. This is consistent with the virtual identity of the dose-response curves. However, these multivariate MR models provide LED's which are more conservative than the ones obtained from any other univariate MR model. This is an argument in favor of a joint approach, even though it is tempered by the fact that external and visceral outcomes show a very mild risk. In case the three outcomes would suffer from a substantial risk, then focusing attention to a single response or a collapsed outcome would overestimate the safe dose. Further, it is noteworthy how the square root transformed linear predictors yield substantially lower (and thus more conservative) BMDs and LEDs than the conventional linear predictors in d, thanks to a better fit of the dose-response curve to the data. Moreover, within skeletal or collapsed outcomes, LED_{05} is lowest when using the more complex linear predictor in $\sqrt{d}$ and d. The LED_{05} obtained from logistic regression equals 9 and is clearly the lowest of all. While this seems cautious, Morgan (1992, p. 175) 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, since it may produce unrealistically low safe doses.

5.4.2 DEHP Study

For the DEHP study, the BMD_{05} and LED_{05} for several univariate models, as well as the final model are displayed in Table 5.8. Here, the difference between the use of a linear predictor in d or a transformed $1/d$ is much smaller than in the previous study. The BMDs and LEDs calculated with the fractional polynomial approach are only slightly more conservative. Adding a quadratic term in $\frac{1}{d+1}$ to the predictor of the

Table 5.7: *EG (mice) Study: Estimated Values of the BMD_{05} and LED_{05} (mg/kg/day) under Different Models (functional form of linear predictor in dose d is indicated when necessary).*

	Univariate Models						
	External		Visceral		Skeletal		
	(d)	$(\sqrt{d})$	(d)	$(\sqrt{d})$	(d)	$(\sqrt{d})$	$(\sqrt{d}+d)$
BMD_{05}	1035	566	1672	1445	207	24	33
LED_{05}	788	313	1273	954	170	17	14

	Univariate Models		
	Collapsed		
	(d)	$(\sqrt{d})$	$(\sqrt{d}+d)$
BMD_{05}	197	22	31
LED_{05}	161	15	13

	Multivariate Models	
	Model2	Model5
BMD_{05}	28	27
LED_{05}	12	12

Table 5.8: *DEHP Study: Estimated Values of the BMD_{05} and LED_{05} (%) under Different Models (functional form of linear predictor in dose d is indicated when necessary).*

	Univariate Models								
	External			Visceral		Skeletal		Collapsed	
	(d)	$(\frac{1}{d+1})$	$(\frac{1}{(d+1)^2})$	(d)	$(\frac{1}{d+1})$	(d)	$(\frac{1}{d+1})$	d	$(\frac{1}{d+1})$
BMD_{05}	0.018	0.015	0.016	0.015	0.011	0.019	0.014	0.009	0.006
LED_{05}	0.014	0.011	0.011	0.012	0.008	0.015	0.009	0.007	0.005

	Final Multivariate Models
BMD_{05}	0.006
LED_{05}	0.004

external main effect does not seem to provide a further reduction of the calculated benchmark doses. As before, the multivariate model yields the most conservative results. However, in this case, it should be noted that the results are virtually identical to the ones obtained with a univariate model on a collapsed outcome.

5.5 Conclusion

We have studied risk assessment from developmental toxicity studies. Such studies combine clustering (fetuses within dams), multivariate outcomes (visceral, skeletal, and external malformation), and binary data indicator variables. Likelihood based models for this fairly involved data structure do not abound, due to the demanding computational requirements.

The model combines conditional logits for the main effects of malformation with pairwise conditional log odds ratios for the association structure. Each of these natural parameters needs to be specified as a realistic function of dose. Whereas linear models may be too simplistic, higher order polynomial extensions suffer from

well-known drawbacks, especially when low dose extrapolation is envisaged. Non-linear predictors pose particular and currently unresolved methodological challenges.

Therefore, we have advocated the use of the fractional polynomial approach, suggested by Royston and Altman (1994). This heuristic scheme of model selection has properties, superior to those of polynomial predictors, when both are different. In case the extension is not necessary, this family essentially returns to a polynomial structure. Thus, their use is strongly recommended and considering a polynomial and a fractional polynomial approach simultaneously, is certainly a worthwile sensitivity analysis in an important public health matter such as the determination of safe limits for human exposure to potentially hazardous agents.

Chapter 6

Comparison of Pseudo-likelihood and Generalized Estimating Equations for Marginally Specified Odds Ratio Models

6.1 Introduction

In the framework of a marginally specified odds ratio model (Lipsitz, Laird and Harrington 1991, Dale 1986, Molenberghs and Lesaffre 1994, Glonek and McCullagh 1995, Lang and Agresti 1994) for multivariate, clustered binary data, full maximum likelihood estimation can also become prohibitive, especially with large within-unit replication. In this chapter we describe two alternative estimation procedures, which are easier to fit: pseudo-likelihood and generalized estimating equations.

The latter is generally well known, but typically aimed at marginal models. In contrast, PL can be used with both marginal (Le Cessie and Van Houwelingen 1994, Geys, Molenberghs and Lipsitz 1998) and conditional models. For conditionally specified models, PL is often seen as the most natural choice, as it exhibits several desirable properties (Geys, Molenberghs and Ryan 1997, 1999), studied in length in Chapters 3 and 4. Here, we discuss the relative merits of PL and GEE, which will be illustrated using data from NTP studies. As before, we will only pay attention to exchangeable association structures and cluster-level covariates, since this simplifies

comparison and covers the setting encountered in the data. While our findings can be applied to some longitudinal settings, the assumption of exchangeability is frequently not tenable, so that more complex association structures are needed. The extension to longitudinal data therefore needs further investigation.

In Section 6.2 we construct an appropriate PL function and derive its corresponding estimating equations. Next, we present an equivalent but more appealing representation of the PL estimating equations in terms of contrasts between observed and expected frequencies. Section 6.3 deals with a GEE1 and GEE2 approach for this setting. We discuss the relative merits of full likelihood, first and second order GEE and PL in Section 6.4. Finally, in Section 6.5 we apply GEE and PL procedures on data from NTP studies. Evaluating the likelihood for this kind of data can become computationally very demanding. This is especially true for large clusters and motivates the exploration of alternative estimation procedures such as PL and GEE.

6.2 Pseudo-likelihood Estimating Equations

In this section we first present a general PL form, accomodating clustered responses. Next, we concentrate on the special case of exchangeability leading to an elegant formulation of the PL. Again, we assume there are N clusters with $k = 1, \ldots, n_i$ indexing the individuals in the ith cluster. If we denote the binary outcome for subject k in cluster i by Y_{ik} then the exchangeability assumption allows us to introduce the summary statistic $Z_i = \sum_{k=1}^{n_i} Y_{ik}$: the total number of successes within the ith cluster.

6.2.1 Classical Representation

Definition 1

Le Cessie and Van Houwelingen (1994) replace the true contribution $f(y_{i1}, \ldots, y_{in_i})$ of a vector of correlated binary data to the full likelihood by the product of all pairwise contributions $f(y_{ij}, y_{ik})$ $(1 \leq j < k \leq n_i)$, to obtain a *pseudo-likelihood function*. Grouping the outcomes for subject i into a vector $\boldsymbol{Y}_i$, the contribution of

the ith cluster to the log pseudo-likelihood is

$$p\ell_i = \sum_{j<k} \ln f(y_{ij}, y_{ik}), \tag{6.1}$$

if it contains more than one observation. Otherwise $p\ell_i = f(y_{i1})$. In the sequel we restrict our attention to clusters of size larger than 1. Clusters of size 1 contribute to the marginal parameters only.

Using a bivariate Plackett distribution (Plackett 1965) the joint probabilities $f(y_{ij}, y_{ik})$, denoted by π_{ijk}, can be specified in terms of marginal probabilities and pairwise odds ratios. For individuals j and k (or for measurement occasions j and k in a longitudinal study), the pairwise odds ratio ψ_{ijk} is defined as (Fitzmaurice, Molenberghs and Lipsitz 1995):

$$\psi_{ijk} = \frac{P(Y_{ij}=1, Y_{ik}=1)P(Y_{ij}=0, Y_{ik}=0)}{P(Y_{ij}=1, Y_{ik}=0)P(Y_{ij}=0, Y_{ik}=1)}.$$

Dale (1986) refers to this quantity as the *global cross ratio*.

The univariate marginal means π_{ij}, as well as the pairwise odds ratios ψ_{ijk}, can be modelled in terms of regression parameters, using (for example) logit and log links respectively, whence the bivariate marginal means π_{ijk} satisfy:

$$\begin{aligned} \pi_{ijk} &= \frac{1+(\pi_{ij}+\pi_{ik})(\psi_{ijk}-1)-S(\pi_{ik},\pi_{ij},\psi_{ijk})}{2(\psi_{ijk}-1)} && \text{if } \psi_{ijk} \neq 1 \\ \pi_{ijk} &= \pi_{ij}\pi_{ik} && \text{if } \psi_{ijk} = 1, \end{aligned}$$

with

$$S(\pi_{ij}, \pi_{ik}, \psi_{ijk}) = \sqrt{[1+(\pi_{ij}+\pi_{ik})(\psi_{ijk}-1)]^2 + 4\psi_{ijk}(1-\psi_{ijk})\pi_{ij}\pi_{ik}}.$$

Under Exchangeability

For binary data and taking the exchangeability assumption into account, the log pseudo-likelihood contribution $p\ell_i$ can be formulated as:

$$p\ell_i = \binom{z_i}{2} \ln \pi^*_{i11} + \binom{n_i - z_i}{2} \ln \pi^*_{i00} + z_i(n_i - z_i) \ln \pi^*_{i10}. \tag{6.2}$$

In this formulation, π^*_{i11} and π^*_{i00} denote the bivariate probabilities of observing two *successes* or two *failures* respectively, while π^*_{i10} is the probability for the first component being 1 and the second being 0. Under exchangeability, this is identical

to the probability π^*_{i01} for the first being 0 and the second being 1. If we consider the following reparameterization: $\pi_{i11} = \pi^*_{i11}$, $\pi_{i10} = \pi^*_{i11} + \pi^*_{i10} = \pi_{01}$ and $\pi_{i00} = \pi^*_{i11} + \pi^*_{i10} + \pi^*_{i01} + \pi^*_{i00} - 1$, then this one-to-one reparameterization maps the three, common within-cluster, two-way marginal probabilities $(\pi^*_{i11}, \pi^*_{i10}, \pi^*_{i00})$ to two one-way marginal probabilities (which under exchangeability are both equal to π_{i10}) and one two-way probability $\pi_{i11} = \pi^*_{i11}$. Hence, equation (6.2) can be reformulated as:

$$p\ell_i = \binom{z_i}{2} \ln \pi_{i11} + \binom{n_i - z_i}{2} \ln(1 - 2\pi_{i10} + \pi_{i11}) + z_i(n_i - z_i) \ln(\pi_{i10} - \pi_{i11}), \tag{6.3}$$

and the pairwise odds ratio ψ_{ijk} reduces to:

$$\psi_i = \frac{\pi_{i11}(1 - 2\pi_{i10} + \pi_{i11})}{(\pi_{i10} - \pi_{i11})^2}.$$

In order to enable model specification, we assume a composite link function $\boldsymbol{\eta_i} = (\eta_{i1}, \eta_{i2})^T$ with a mean and an association component:

$$\begin{aligned} \eta_{i1} &= \ln(\pi_{i10}) - \ln(1 - \pi_{i10}), \\ \eta_{i2} &= \ln(\psi_i) = \ln(\pi_{i11}) + \ln(1 - 2\pi_{i10} + \pi_{i11}) - 2ln(\pi_{i10} - \pi_{i11}). \end{aligned}$$

From these links, the univariate and pairwise probabilities are easily derived (Plackett 1965):

$$\pi_{i10} = \frac{\exp(\eta_{i1})}{1 + \exp(\eta_{i1})}$$

and

$$\pi_{i11} = \begin{cases} \frac{1 + 2\pi_{i10}(\psi_i - 1) - S_i}{2(\psi_i - 1)} & \text{if } \psi_i \neq 1 \\ \pi^2_{i10} & \text{if } \psi_i = 1, \end{cases}$$

with

$$S_i = \sqrt{[1 + 2\pi_{i10}(\psi_i - 1)]^2 + 4\psi_i(1 - \psi_i)\pi^2_{i10}}.$$

Next, we can assume a linear model $\boldsymbol{\eta_i} = X_i\boldsymbol{\beta}$, with X_i a known design matrix and $\boldsymbol{\beta}$ a vector of unknown regression parameters.

The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is then defined as the solution to the pseudo-score equations:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}.$$

Using the chain rule, $\boldsymbol{U}(\boldsymbol{\beta})$ can be written as:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T (T_i^{-1})^T \frac{\partial p\ell_i}{\partial \boldsymbol{\pi}_i} \tag{6.4}$$

with $\boldsymbol{\pi}_i = (\pi_{i10}, \pi_{i11})^T$ and $T_i = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\pi}_i$. Writing the different components in (6.4) in full yields:

$$\frac{\partial p\ell_i}{\partial \boldsymbol{\pi}_i} = \begin{pmatrix} \dfrac{\partial p\ell_i}{\partial \pi_{i10}} \\ \dfrac{\partial p\ell_i}{\partial \pi_{i11}} \end{pmatrix}$$

with

$$\begin{aligned} \frac{\partial p\ell_i}{\partial \pi_{i10}} &= \frac{-2}{1 - 2\pi_{i10} + \pi_{i11}} \binom{n_i - z_i}{2} + \frac{z_i(n_i - z_i)}{\pi_{i10} - \pi_{i11}} \\ \frac{\partial p\ell_i}{\partial \pi_{i11}} &= \frac{1}{\pi_{i11}} \binom{z_i}{2} + \frac{1}{1 - 2\pi_{i10} + \pi_{i11}} \binom{n_i - z_i}{2} - \frac{z_i(n_i - z_i)}{\pi_{i10} - \pi_{i11}} \end{aligned}$$

and

$$T = \begin{pmatrix} \dfrac{1}{\pi_{i10}} + \dfrac{1}{1 - \pi_{i10}} & 0 \\ -\dfrac{2}{1 - 2\pi_{i10} + \pi_{i11}} - \dfrac{2}{\pi_{i10} - \pi_{i11}} & \dfrac{1}{\pi_{i11}} + \dfrac{1}{1 - 2\pi_{i10} + \pi_{i11}} + \dfrac{2}{\pi_{i10} - \pi_{i11}} \end{pmatrix}.$$

Two frequently used fitting algorithms are the Newton-Raphson and the Fisher scoring algorithms. Newton-Raphson starts with a vector of initial estimates $\boldsymbol{\beta}^{(0)}$ and updates the current value of the parameter vector $\boldsymbol{\beta}^{(s)}$ by:

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + W(\boldsymbol{\beta}^{(s)})^{-1} \boldsymbol{U}(\boldsymbol{\beta}^{(s)}).$$

Here, $W(\boldsymbol{\beta})$ is the matrix of the second derivatives of the log pseudo-likelihood with respect to the regression parameters $\boldsymbol{\beta}$:

$$W(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T \left(\boldsymbol{F}_i + (T_i^{-1})^T \frac{\partial^2 p\ell_i}{\partial \boldsymbol{\pi}_i \partial \boldsymbol{\pi}_i^T} (T_i^{-1}) \right) X_i,$$

and $\boldsymbol{F}_i$ is defined by (McCullagh 1987, p. 5; Molenberghs and Lesaffre 1999):

$$(F_i)_{jk} = \sum_s \sum_{r,t,u} \frac{\partial^2 \eta_{ir}}{\partial \pi_{it} \partial \pi_{iu}} \frac{\partial \pi_{is}}{\partial \eta_{ir}} \frac{\partial \pi_{it}}{\partial \eta_{ij}} \frac{\partial \pi_{iu}}{\partial \eta_{ik}} \frac{\partial p\ell_i}{\partial \pi_{is}}.$$

The Fisher scoring algorithm is obtained by replacing the matrix $W(\boldsymbol{\beta})$ by its expected value:

$$W(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T (T_i^{-1})^T A_i (T_i^{-1}) X_i,$$

with A_i the expected value of the matrix of second derivatives of the log pseudo-likelihood $p\ell_i$ with respect to $\boldsymbol{\pi}_i$. Calculating the three relevant elements of A_i, it can easily be demonstrated that:

$$E\left[\binom{z_i}{2}\right] = \binom{n_i}{2}\pi_{i11}, \tag{6.5}$$

$$E\left[\binom{n_i - z_i}{2}\right] = \binom{n_i}{2} + \binom{n_i}{2}\pi_{i11} - n_i(n_i-1)\pi_{i10}, \tag{6.6}$$

$$E\left[z_i(n_i - z_i)\right] = n_i(n_i-1)(\pi_{i10} - \pi_{i11}). \tag{6.7}$$

Using (6.5), (6.6) and (6.7) yields:

$$\begin{aligned}
A_{i11} &= E\left[\frac{\partial^2 p\ell_i}{\partial \pi_{i10}^2}\right] \\
&= \frac{-4}{(1-2\pi_{i10}+\pi_{i11})^2}\left[\binom{n_i}{2}(1+\pi_{i11}) - n_i(n_i-1)\pi_{i10}\right] - \frac{n_i(n_i-1)}{(\pi_{i10}-\pi_{i11})}, \\
A_{i12} &= E\left[\frac{\partial^2 p\ell_i}{\partial \pi_{i10}\partial \pi_{i11}}\right] \\
&= \frac{2}{(1-2\pi_{i10}+\pi_{i11})^2}\left[\binom{n_i}{2}(1+\pi_{i11}) - n_i(n_i-1)\pi_{i10}\right] + \frac{n_i(n_i-1)}{(\pi_{i10}-\pi_{i11})}, \\
A_{i22} &= E\left[\frac{\partial^2 p\ell_i}{\partial \pi_{i11}^2}\right] \\
&= -\binom{n_i}{2}\frac{1}{\pi_{i11}} - \frac{1}{(1-2\pi_{i10}+\pi_{i11})^2}\left[\binom{n_i}{2}(1+\pi_{i11}) - n_i(n_i-1)\pi_{i10}\right] \\
&\quad -\frac{n_i(n_i-1)}{(\pi_{i10}-\pi_{i11})}
\end{aligned}$$

Similar in spirit to generalized estimating equations, the asymptotic covariance matrix of the regression parameters $\hat{\boldsymbol{\beta}}$ is consistently estimated by (Arnold and Strauss

1991, Geys, Molenberghs and Ryan 1997):

$$W(\hat{\boldsymbol{\beta}})^{-1}\left(\sum_{i=1}^{N}\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})^T\right)W(\hat{\boldsymbol{\beta}})^{-1}.$$

In the context of generalized estimating equations, this estimator is also known as the empirically corrected or sandwich estimator.

Definition 2

A non-equivalent specification of the pseudo-likelihood contribution (6.1) is:.

$$p\ell_i^* = p\ell_i/(n_i - 1).$$

The factor $1/(n_i - 1)$ corrects for the fact that each response Y_{ij} occurs $n_i - 1$ times in the ith contribution to the PL and it ensures that the PL reduces to full likelihood under independence. Indeed, under independence, (6.3) simplifies to:

$$p\ell_i = (n_i - 1)\left[z_i \ln(\pi_{i10}) + (n_i - z_i)\ln(1 - \pi_{i10})\right].$$

We can replace $p\ell_i$ by $p\ell_i^*$ everywhere in this discussion. However, if $(n_i - 1)$ is random it is not obvious that the expected value of $U_i(\boldsymbol{\beta})/(n_i - 1)$ equals zero. To ensure that the solution to the new pseudo-score equation is consistent, we have to assume that n_i is independent of z_i given the dose level d_i for the ith cluster. Then we have:

$$E\left[U_i(\boldsymbol{\beta})/(n_i - 1)|d_i\right] = E\left[U_i(\boldsymbol{\beta})|d_i\right]E\left[1/(n_i - 1)|d_i\right] = 0.$$

When all clusters are equal in size, the PL estimator $\boldsymbol{\beta}$ and its variance-covariance matrix remain the same, no matter whether we use $p\ell_i$ or $p\ell_i^*$ in the definition of the log pseudo-likelihood.

6.2.2 Generalized Linear Model Representation

To obtain the pseudo-likelihood function described in Section 6.2.1 we replaced the true contribution $f(y_{i1}, \ldots, y_{in_i})$ of the ith cluster to the full likelihood, by the product of all pairwise contributions $f(y_{ij}, y_{ik})$ with $1 \leq j < k \leq n_i$. This implies that a particular response y_{ij} occurs $n_i - 1$ times in $p\ell_i$. Therefore, it is useful to construct for each response y_{ij}, $n_i - 1$ replicated $y_{ij}^{(k)}$ with $k \neq j$. The dummy response $y_{ij}^{(k)}$ is to be interpreted as the particular replicate of y_{ij} that is paired

with the replicate $y_{ik}^{(j)}$ of y_{ik} in the pseudo-likelihood function. Using this specific device we are able to rewrite the gradient of the log pseudo-likelihood $p\ell$ in an appealing generalized linear model type representation. With notation introduced in the previous section the gradient can now be written as:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T (T_i^{-1})^T V_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{\pi}_i)$$

or, using the second representation $p\ell_i^*$, as:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T (T_i^{-1})^T V_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{\pi}_i)/(n_i - 1),$$

where

$$\boldsymbol{Z}_i = \begin{pmatrix} \sum_{j=1}^{n_i} \sum_{k \neq j} Y_{ij}^{(k)} \\ \frac{1}{2} \sum_{j=1}^{n_i} \sum_{k \neq j} Y_{ij}^{(k)} Y_{ik}^{(j)} \end{pmatrix}, \quad \pi_i = \begin{pmatrix} n_i(n_i - 1)\pi_{i10} \\ \binom{n_i}{2} \pi_{i11} \end{pmatrix}$$

The components of V_i, the covariance matrix of $\boldsymbol{Z}_i$, can be obtained after some straightforward calculations. We illustrate them on the (1,1) element only. If we denote the components of $\boldsymbol{Z}_i$ by (Z_{i1}, Z_{i2}), then:

$$E(Z_{i1}) \quad = \quad n_i(n_i - 1)\pi_{i10}$$

and

$$\begin{aligned}
E(Z_{i1}^2) &= E\left[\left(\sum_{j=1}^{n_i}\sum_{t\neq j} Y_{ij}^{(t)}\right)\left(\sum_{k=1}^{n_i}\sum_{s\neq k} Y_{ik}^{(s)}\right)\right] \\
&= E\left[\sum_{j=1}^{n_i}\sum_{t\neq j}\sum_{s\neq j} Y_{ij}^{(t)}Y_{ij}^{(s)}\right] + E\left[\sum_{j=1}^{n_i}\sum_{t\neq j}\sum_{k\neq j}^{n_i}\sum_{s\neq k} Y_{ij}^{(t)}Y_{ik}^{(s)}\right] \\
&= E\left[\sum_{j=1}^{n_i}\sum_{t\neq j}(Y_{ij}^{(t)})^2\right] + E\left[\sum_{j=1}^{n_i}\sum_{t\neq j}\sum_{s\neq t,j} Y_{ij}^{(t)}Y_{ij}^{(s)}\right] \\
&\quad + E\left[\sum_{j=1}^{n_i}\sum_{k\neq j} Y_{ij}^{(k)}Y_{ik}^{(j)}\right] + 2E\left[\sum_{j=1}^{n_i}\sum_{k\neq j}\sum_{t\neq j,k} Y_{ij}^{(t)}Y_{ik}^{(j)}\right] \\
&\quad + E\left[\sum_{j=1}^{n_i}\sum_{k\neq j}\sum_{t\neq j,k}\sum_{s\neq j,k} Y_{ij}^{(k)}Y_{ik}^{(s)}\right] \\
&= n_i(n_i-1)\pi_{i10} + n_i(n_i-1)(n_i-2)\pi_{i10}^2 + n_i(n_i-1)\pi_{i11} \\
&\quad + 2n_i(n_i-1)(n_i-2)\pi_{i10}^2 + n_i(n_i-1)(n_i-2)^2\pi_{i10}^2.
\end{aligned}$$

This leads to:

$$\begin{aligned}
\mathrm{Var}(Z_{i1}) &= E(Z_{i1}^2) - E(Z_{i1})^2 \\
&= n_i(n_i-1)(\pi_{i10} + \pi_{i11} - 2\pi_{i10}^2).
\end{aligned}$$

Under independence, this reduces to $Var(Z_{i1}) = n_i(n_i-1)\pi_{i10}(1-\pi_{i10})$. Similar calculations lead to:

$$\begin{aligned}
\mathrm{Cov}(Z_{i1}, Z_{i2}) &= n_i(n_i-1)\pi_{i11}(1-\pi_{i10}) \\
\mathrm{Var}(Z_{i2}) &= \binom{n_i}{2}\pi_{i11}(1-\pi_{i11}).
\end{aligned}$$

Clearly, the elements of V_i take appealing expressions and are easy to implement. One only needs to evaluate first and second order probabilities. Under independence, the variances reduce to well-known quantities.

To obtain a suitable PL estimator for $\boldsymbol{\beta}$ we can use the Fisher-scoring algorithm where the matrix A_i in the previous section is now replaced by the inverse of V_i. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated in a similar fashion as before.

6.3 Generalized Estimating Equations

When we are mainly interested in first order marginal mean parameters and the pairwise interactions, a full likelihood procedure can be replaced by quasi-likelihood methods (McCullagh and Nelder 1989). In quasi-likelihood, the mean response is expressed as a parametric function of covariates; the variance is assumed to be a function of the mean up to possibly unknown scale parameters. Wedderburn (1974) first noted that likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood "estimating equations" provide consistent estimates of the regression parameters $\boldsymbol{\beta}$ in any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families.

In the introduction we did already mention that Liang and Zeger (1986) introduced first order estimating equations, GEE1, which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the association structure.

Prentice (1988) extended these results to allow joint estimation of probabilities and pairwise correlations, using a pair of estimating equations. Williamson, Lipsitz and Kim (1997) wrote a SAS macro for GEE1, based on Prentice's approach. Lipsitz, Laird and Harrington (1991) modified the estimating equations of Prentice (1988) to allow modelling of the association through marginal odds ratios rather than marginal correlations.

Adopting the ideas of Prentice (1988) and Lipsitz, Laird and Harrington (1991), we first consider a GEE1 approach that allows joint estimation of regression parameters $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ in respectively the marginal means and pairwise associations, using two sets of estimating equations. Both extended the GEE1 approach of Liang and Zeger (1986), where estimators for $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ were obtained using iteratively reweighted least squares calculations and moment-based estimation of $\boldsymbol{\alpha}$. If we let the marginal means π_{i10} and pairwise probabilities π_{i11} depend on a vector of regression parameters $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ through the following generalized linear model:

$$\boldsymbol{\eta}_i = \begin{pmatrix} \ln(\pi_{i10}) - \ln(1-\pi_{i10}) \\ \ln(\pi_{i11}) + \ln(1-2\pi_{i10}+\pi_{i11}) - 2\ln(\pi_{i10}-\pi_{i11}) \end{pmatrix} = X_i \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix},$$

then the two sets of estimating equations for, respectively, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be combined

into:

$$\sum_{i=1}^{N} \begin{pmatrix} D_i^T & 0 \\ 0 & C_i^T \end{pmatrix} \begin{pmatrix} \text{Var}(Z_i) & 0 \\ 0 & \text{Var}(\binom{Z_i}{2}) \end{pmatrix}^{-1} \begin{pmatrix} Z_i - n_i\pi_{i10} \\ \binom{Z_i}{2} - \binom{n_i}{2}\pi_{i11} \end{pmatrix},$$

where $D_i = n_i\partial\pi_{i10}/\partial\boldsymbol{\beta}$ and $C_i = \binom{n_i}{2}\partial\pi_{i11}/\partial\boldsymbol{\alpha}$. An iterative procedure for calculating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ begins with starting values $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ and produces updated values $\boldsymbol{\beta}_{s+1}$,$\boldsymbol{\alpha}_{s+1}$ from values $\boldsymbol{\beta}_s$,$\boldsymbol{\alpha}_s$ by means of

$$\begin{aligned} \boldsymbol{\beta}_{s+1} &= \boldsymbol{\beta}_s + \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1} \sum_{i=1}^{N} D_i^T\ V_i^{-1}\left(Z_i - n_i\pi_{i10}\right) \\ \boldsymbol{\alpha}_{s+1} &= \boldsymbol{\alpha}_s + \left(\sum_{i=1}^{N} C_i^T W_i^{-1} C_i\right)^{-1} \sum_{i=1}^{N} C_i^T\ W_i^{-1}\left(\binom{Z_i}{2} - \binom{n_i}{2}\pi_{i11}\right), \end{aligned}$$

where $V_i = \text{Var}(Z_i)$ and $W_i = \text{Var}(\binom{Z_i}{2}) = \text{Var}(\sum_{j<k} Y_{ij}Y_{ik})$. Here, W_i is a function of third and fourth order probabilities, which are nuisance parameters we would rather not estimate. Assuming three- and higher order independence, in the spirit of Lipsitz, Laird and Harrington (1991), and taking into account the exchangeability assumption, W_i reduces to:

$$\binom{n_i}{2}\pi_{i11}(1-\pi_{i11}).$$

Prentice (1988) and Lipsitz, Laird and Harrington (1991) have shown that the joint asymptotic covariance matrix of $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T)^T$ equals:

$$V_{\beta,\alpha} = \lim_{N\to\infty} \begin{pmatrix} B_{11}^{-1} & 0 \\ B_{21} & B_{22}^{-1} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \begin{pmatrix} B_{11}^{-1} & 0 \\ B_{21} & B_{22}^{-1} \end{pmatrix}^T,$$

where

$$\begin{aligned} B_{11} &= N^{-1}\textstyle\sum_{i=1}^{N} D_i^T V_i^{-1} D_i, \\ B_{22} &= N^{-1}\textstyle\sum_{i=1}^{N} C_i^T W_i^{-1} C_i, \\ B_{21} &= B_{22}^{-1}(\textstyle\sum_{i=1}^{N} C_i^T W_i^{-1}\partial(\binom{n_i}{2}\pi_{i11})/\partial\boldsymbol{\beta})B_{11}^{-1}, \\ \Sigma_{11} &= N^{-1}\textstyle\sum_{i=1}^{N} D_i^T V_i^{-1}\text{Var}(Z_i)V_i^{-1} D_i, \\ \Sigma_{22} &= N^{-1}\textstyle\sum_{i=1}^{N} C_i^T W_i^{-1}\text{Var}(\binom{Z_i}{2})W_i^{-1} C_i, \\ \Sigma_{12} &= N^{-1}\textstyle\sum_{i=1}^{N} D_i^T V_i^{-1}\text{Cov}(Z_i, \binom{Z_i}{2})W_i^{-1} C_i. \end{aligned}$$

The matrix $V_{\beta,\alpha}$ can be consistently estimated by replacing $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by their estimates, and also

$$\text{Var}(Z_i) \text{ by } (Z_i - n_i\pi_{i10})(Z_i - n_i\pi_{i10})^T,$$

$$\text{Var}(\tbinom{Z_i}{2}) \text{ by } (\tbinom{Z_i}{2} - \tbinom{n_i}{2}\pi_{i11})(\tbinom{Z_i}{2} - \tbinom{n_i}{2}\pi_{i11})^T,$$

$$\text{Cov}(Z_i, \tbinom{Z_i}{2}) \text{ by } (Z_i - n_i\pi_{i10})(\tbinom{Z_i}{2} - \tbinom{n_i}{2}\pi_{i11})^T.$$

Note that GEE1 operates as if $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal to one another even when they actually are not. The effect is that GEE1 gives consistent estimators of $\boldsymbol{\beta}$ whether or not the association structure is correctly specified. On the other hand, GEE1 can be extremely inefficient for the estimation of $\boldsymbol{\alpha}$.

A second order extension of these estimating equations that includes marginal pairwise associations as well has been studied by Liang, Zeger and Qaqish (1992), Molenberghs and Ritter (1996) and Heagerty and Zeger (1996). Liang, Zeger and Qaqish (1992) point out the connection of the quasi-likelihood theories with second order generalized estimating equations, GEE2. In fact, GEE2 can be simply regarded as a multivariate extension of quasi-likelihood. As in quasi-likelihood, GEE2 requires specification of first and second order moments, which are usually of great scientific interest. Indeed, even when there is considerable association between outcomes, three-way and higher order interactions tend to be negligible and are certainly more difficult to interpret. Therefore, a working higher order independence assumption is often plausible. We will develop a second-order estimating equations procedure (GEE2), following the ideas of Liang, Zeger and Qaqish (1992) and adopting a working higher order independence assumption. It is very appealing that such a procedure closely corresponds to the way in which the pseudo-likelihood function was represented. Recall that the pseudo-likelihood function also limits its attention to pairwise interactions, since it is constructed as a product of pairwise probabilities. In the GEE2 framework the following set of estimating equations can be considered:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T (T_i^{-1})^T V_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{\pi}_i) = 0$$

with

$$\boldsymbol{Z}_i = \begin{pmatrix} Z_i \\ \begin{pmatrix} Z_i \\ 2 \end{pmatrix} \end{pmatrix} \text{ and } \pi_i = \begin{pmatrix} n_i\pi_{i10} \\ \begin{pmatrix} n_i \\ 2 \end{pmatrix} \pi_{i11} \end{pmatrix}.$$

Furthermore, $T_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\pi}_i$ and V_i is the covariance matrix of $\boldsymbol{Z}_i$. The computation of T_i presents no difficulties and is analogous to the calculations performed in Section 6.2.1. We obtain:

$$T_i = \begin{pmatrix} \frac{1}{n_i}\left(\frac{1}{\pi_{i10}} + \frac{1}{1-\pi_{i10}}\right) & 0 \\ \frac{1}{n_i}\left(\frac{-2}{1-2\pi_{i10}+\pi_{i11}} - \frac{2}{\pi_{i10}-\pi_{i11}}\right) & \frac{2}{n_i(n_i-1)}\left(\frac{1}{\pi_{i11}} + \frac{1}{1-2\pi_{i10}+\pi_{11}} + \frac{2}{\pi_{i10}-\pi_{i11}}\right). \end{pmatrix}$$

However, the matrix V_i contains third and fourth order probabilities, which can be found using either the iterative proportional fitting (IPF) algorithm, outlined in Molenberghs and Lesaffre (1999) or alternatively by the procedure given in Molenberghs and Lessaffre (1994), which we use here. This is an important difference with both PL and GEE1. Indeed, these only need first and second order probabilities, which are straightforward to implement. Probabilities of order n can be computed, provided all lower-dimensional probabilities together with the odds-ratio of dimension n are known. At this point we introduce the higher order independence working assumption. Let us denote the so-obtained three and four way probabilities $P(y_{ij} = 1, y_{ik} = 1, y_{il} = 1)$ and $P(y_{ij} = 1, y_{ik} = 1, y_{il} = 1, y_{im} = 1)$ by $\mu_{i1}^{(3)}$ resp. $\mu_{i1}^{(4)}$, then we can calculate the different components of V_i:

$$\begin{aligned} \text{Var}(Z_{i1}) &= E(Z_i^2) - E(Z_i)^2 \\ &= 2E\left[\sum_{j=1}^{n_i}\sum_{k>j} Y_{ij}Y_{ik}\right] + E\left[\sum_{j=1}^{n_i} Y_{ij}^2\right] - E\left[\sum_{j=1}^{n_i} Y_{ij}\right]^2 \\ &= 2\binom{n_i}{2}\pi_{i11} + n_i\pi_{i10}(1 - n_i\pi_{i10}) \end{aligned} \tag{6.8}$$

Note that (6.8) reduces to $n_i\pi_{i10}(1 - \pi_{i10})$, under independence. Similarly, we cal-

culate:

$$\begin{aligned}
\text{Cov}(Z_{i1}, Z_{i2}) &= 3\sum_{j=1}^{n_i}\sum_{k>j}\sum_{l>k} E\left[Y_{ij}Y_{ik}Y_{il}\right] + 2E\left[\sum_{j=1}^{n_i}\sum_{l>k} Y_{ik}^2 Y_{il}\right] \\
&\quad -n_i\binom{n_i}{2}\pi_{i10}\pi_{i11} \\
&= 3\binom{n_i}{3}\pi_1^{(3)} + 2\binom{n_i}{2}\pi_{i11} - n_i\binom{n_i}{2}\pi_{i10}\pi_{i11}, \\
\text{Var}(Z_{i2}) &= E\left[\sum_{j=1}^{n_i}\sum_{k>j} Y_{ij}Y_{ik}\sum_{r=1}^{n_i}\sum_{s>r} Y_{ir}Y_{is}\right] - \binom{n_i}{2}^2\pi_{i11}^2 \\
&= 6\binom{n_i}{4}\pi_1^{(4)} + 6\binom{n_i}{3}\pi_1^{(3)} + 2\binom{n_i}{2}\pi_{i11} - \binom{n_i}{2}^2\pi_{i11}^2.
\end{aligned}$$

A Fisher scoring algorithm can now be applied to calculate the parameter estimates. The empirically corrected version of the asymptotic covariance matrix proposed by Liang and Zeger (1986) is similar to the one described in Section 6.2.1 and is estimated by:

$$\left(\sum_{i=1}^{N} X_i^T \hat{T}_i^{-T}\hat{V}_i^{-1}\hat{T}_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})^T\right)\left(\sum_{i=1}^{N} X_i^T \hat{T}_i^{-T}\hat{V}_i^{-1}\hat{T}_i^{-1}X_i\right)^{-1}. \tag{6.9}$$

Thus, provided the model is correctly specified, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and is asymptotically normally distributed with covariance matrix estimated by (6.9). If the model for the association structure is misspecified, bias may follow in first order parameters (Liang, Zeger and Qaqish 1992). This contrasts with the classical first order estimating equations, GEE1, which yield consistent estimates even if the association structure is misspecified.

6.4 Comparison

In the previous sections we described two alternative estimating procedures for full maximum likelihood estimation in the framework of a marginally specified odds ratio model, which are easier and much less time consuming. In this section we provide insight in both these methods. Several questions arise such as to how the

different methods compare in terms of efficiency and in terms of computing time and what the mathematical differences and similarities are. At first glance, there is a fundamental difference. A pseudo-likelihood function is constructed by modifying a joint density. Parameters are estimated by setting the first derivatives of this function equal to zero. On the contrary, generalized estimating equations follow from specification of the first few moments and by adopting assumptions about the higher order moments. One could also consider them as resulting from modifying the score equations from the likelihood function. In that respect McCullagh and Nelder (1989) note that these estimating equations need not necessarily integrate to a so-called *quasi-likelihood*.

The close connection of PL to likelihood is an attractive feature. Indeed, it enabled Geys, Molenberghs and Ryan (1999) to construct pseudo-likelihood ratio test statistics that have easy-to-compute expressions and intuitively appealing limiting distributions. In contrast, likelihood ratio test statistics for GEE (Rotnitzky and Jewell 1990) are slightly more complicated.

In Section 6.2.2 we have rewritten the PL score equations as contrasts of observed and fitted frequencies, herewith showing some agreement between PL and GEE2. Both procedures lead to similar estimating equations. The most important difference is in the evaluation of the matrix $V_i = \mathrm{Cov}(Z_i)$. This only involves first and second order probabilities for the pseudo-likelihood procedure. In that respect, PL resembles GEE1. In contrast, GEE2 also requires evaluation of third and fourth order probabilities. This makes the GEE2 score equations harder to evaluate and also more time consuming.

Both pseudo-likelihood and generalized estimating equations yield consistent and asymptotically normally distributed estimators, provided an empirically corrected variance estimator is used and provided the model is correctly specified. This variance estimator is similar for both procedures, the main difference being the evaluation of V_i.

If we define the log of the pseudo-likelihood contribution for clusters with size larger than one as $p\ell_i^* = p\ell_i/(n_i - 1)$, the first component of the PL vector contribution $\boldsymbol{S}_i = \boldsymbol{Z}_i - \boldsymbol{\pi}_i$ equals that of GEE2. On the contrary, the association component, differs by a factor of $1/(n_i - 1)$. Yet, if we would define the log pseudo-likelihood as $p\ell = \sum_{i=1}^{N} p\ell_i$, then the second components would be equal, while the first components would differ by a factor of $n_i - 1$. Therefore, in studies where the main interest

lies in the marginal mean parameters one would prefer $p\ell^*$ over $p\ell$. However, if main interest lies in the estimation of the association parameters we advocate the use of $p\ell$ instead. GEE1 in that case should be avoided, since its goal is limited to estimation of the mean model parameters, while GEE2 is computationally more complex.

The price to pay for computational ease is usually efficiency. Therefore we will study the asymptotic relative efficiencies (AREs) of the different estimation procedures. For clusters of fixed size, $p\ell$ and $p\ell^*$ are equally efficient. For variable sized clusters, the loss of efficiency for main effects of $p\ell$ will turn out to be very small compared to $p\ell^*$. On the contrary, $p\ell$ will turn out to be superior for estimation of association parameters. We follow the suggestion of Rotnitzky and Wypij (1994), described in Section 3.7.3. In our case, we need to consider all realizations of the form $(n_i, d_i, y_{i1}, \ldots, y_{in_i})$, 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) $f(y_{i1}, \ldots, y_{in_i}|n_i, d_i)$, the actual model probabilities. These can be derived from the cumulative Dale model probabilities. For instance, let $\pi^{(k)}$ denote the cumulative Dale probability of observing at least k successes and $\pi^{(k)*}$ the probability of observing exactly k successes, then

$$\pi^{(k)*} = \binom{n_i}{k} \sum_{j=0}^{(n_i-k)} (-1)^j \binom{n_i - k}{n_i - k - j} \pi^{(k+j)}.$$

As before, we assume that there are 4 dose groups, with one control ($d_i = 0$) and three exposed groups ($d_i = 0.25, 0.5, 1.0$). The number n_i of viable fetuses per cluster can be chosen at random using a local linear smoothed version of the relative frequency distribution given in Table 3.3. Due to excessive time requirements for the maximum likelihood procedure, the calculations are restricted to clusters of size 4. The ML estimating equations are:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\beta}} V_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{\pi}) = 0,$$

where

Table 6.1: *Simulation Studies: Asymptotic Relative Efficiencies for Dose Effect Parameter of GEE1, GEE2 and PL versus ML.*

β_0	β_d	β_a	PL	GEE1	GEE2
-5	5	0.0	1.000	1.000	1.000
		0.3	0.999	0.999	0.999
		1.0	0.995	0.999	0.999
-5	3	0.0	1.000	1.000	1.000
		0.3	0.999	0.999	0.999
		1.0	0.998	0.999	0.999
-5	0	0.0	1.000	1.000	1.000
		0.3	0.999	0.999	0.999
		1.0	0.999	0.999	0.999
0	0	0.0	1.000	1.000	1.000
		0.3	0.999	0.999	0.999
		1.0	0.999	0.999	0.999

$$\boldsymbol{Z}_i = \begin{pmatrix} Z_i \\ \binom{Z_i}{2} \\ \binom{Z_i}{3} \\ \binom{Z_i}{4} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\pi}_i = \begin{pmatrix} n_i\pi_{i10} \\ \binom{n_i}{2}\pi_{i11} \\ \binom{n_i}{3}\pi_i^{(3)} \\ \binom{n_i}{4}\pi_i^{(4)} \end{pmatrix}.$$

This involves the evaluation of third and fourth order probabilities, which is computationally laborious, though feasible. Data are generated from a univariate model where the parameters of interest are modelled as follows: $\text{logit}(\pi_{i10}) = \beta_0 + \beta_d d_i$ with d_i, the dose level applied to the ith cluster, and $\ln \psi_i = \beta_a$, i.e. a constant marginal odds ratio model. The background rate parameters (β_0) equal either 0 or -5 and dose effect parameters (β_d) are chosen from $0, 3, 5$. The second order association parameters (β_a) are chosen from $0, 0.3, 1$. The third and fourth order associations are assumed to be zero. The AREs will decrease for increasing higher order associations.

Table 6.1 shows that, when main interest lies in the estimation of the dose effect,

the AREs are highest for GEE2, followed by GEE1 and PL. Since the cluster sizes are assumed to be constant and equal to 4, it does not matter whether we use $p\ell$ or $p\ell^*$ to define the log of the pseudo-likelihood. This result shows that GEE1 has some advantage when interest lies in the estimation of main effect parameters. ML and GEE2 are computationally more complex. GEE1 is the easiest one to fit and the loss of efficiency for the main effect parameters is very small compared to GEE2 and ML. Similar results were found by Liang, Zeger and Qaqish (1992). The PL estimation procedure proposed by Le Cessie and Van Houwelingen (1994) is also computationally easy but is slightly less efficient than GEE1. The differences in ARE between GEE1 and PL are minor.

Table 6.2: *Simulation Studies: Asymptotic Relative Efficiencies for Association Parameter of GEE1, GEE2 and PL versus ML.*

β_0	β_d	β_a	PL	GEE1	GEE2
-5	5	0.0	1.000	0.865	1.000
		0.3	0.998	0.888	0.999
		1.0	0.995	0.862	0.999
-5	3	0.0	1.000	0.992	1.000
		0.3	0.999	0.992	0.999
		1.0	0.993	0.992	0.999
-5	0	0.0	1.000	1.000	1.000
		0.3	1.000	1.000	1.000
		1.0	1.000	1.000	1.000
0	0	0.0	1.000	1.000	1.000
		0.3	1.000	1.000	1.000
		1.0	1.000	1.000	1.000

When main interest lies in the estimation of the association parameters, Table 6.2 shows that GEE1 can loose considerable efficiency. Moreover, in general, one should not use GEE1 for estimating association parameters, unless confidence in the working assumption is great. Therefore, we would advocate the use of PL. ML and GEE2 are again the most efficient procedures, but computationally intensive. In

case of no dose effect, the three procedures are equally efficient with respect to the association parameter.

As Liang, Zeger and Qaqish (1992) suggested, GEE1, GEE2 and PL may be less efficient when the cluster sizes are unequal. Figures 1 and 2 show the efficiencies of $p\ell$ and $p\ell^*$ and GEE1 versus GEE2 for *varying* cluster sizes. In that case $p\ell$ and $p\ell^*$ behave differently. Since maximum likelihood is prohibitive, we calculated the AREs of several methods versus the GEE2 method. Since even data generation from the assumed true distribution is rather time consuming, we restricted the calculations to clusters of size less than or equal to 6. Association parameters of order three and higher are assumed to be zero.

Figure 6.1 shows that $p\ell^*$ is much more efficient than $p\ell$ for estimating dose effects. Furthermore, it has the desirable property that the ARE equals 1 under independence. For estimating the second order association parameter however, Figure 6.2 suggests the use of $p\ell$ rather than $p\ell^*$. Therefore, if main interest lies in the marginal mean parameters we would suggest to use $p\ell^*$ rather than $p\ell$. However, if main interest lies in the estimation of association parameters, the use of $p\ell$ is advised. If interest is combined, and one type of analysis should be chosen, $p\ell$ might be preferable. When using $p\ell^*$ the ARE increases for increasing association. Furthermore, in all cases, AREs are highest for the lowest background rate parameters. This is in agreement with our findings in Chapter 3.

6.5 Examples

We apply the PL and first and second order GEE estimating procedures to data from the DEHP and DYME studies, described in Chapter 3. Malformations are classified as being external, visceral and skeletal. However, we fit the marginal odds ratio model described in the previous sections to a collapsed outcome, defined as 1 if at least one malformation was found and 0 otherwise. The parameters of interest are modelled as follows:

$$\text{logit}(\pi_{i10}) = \beta_0 + \beta_d d_i,$$

with d_i the dose level applied to the ith cluster, and

$$ln\psi_i = \beta_a,$$

i.e. a constant marginal odds ratio model is assumed.

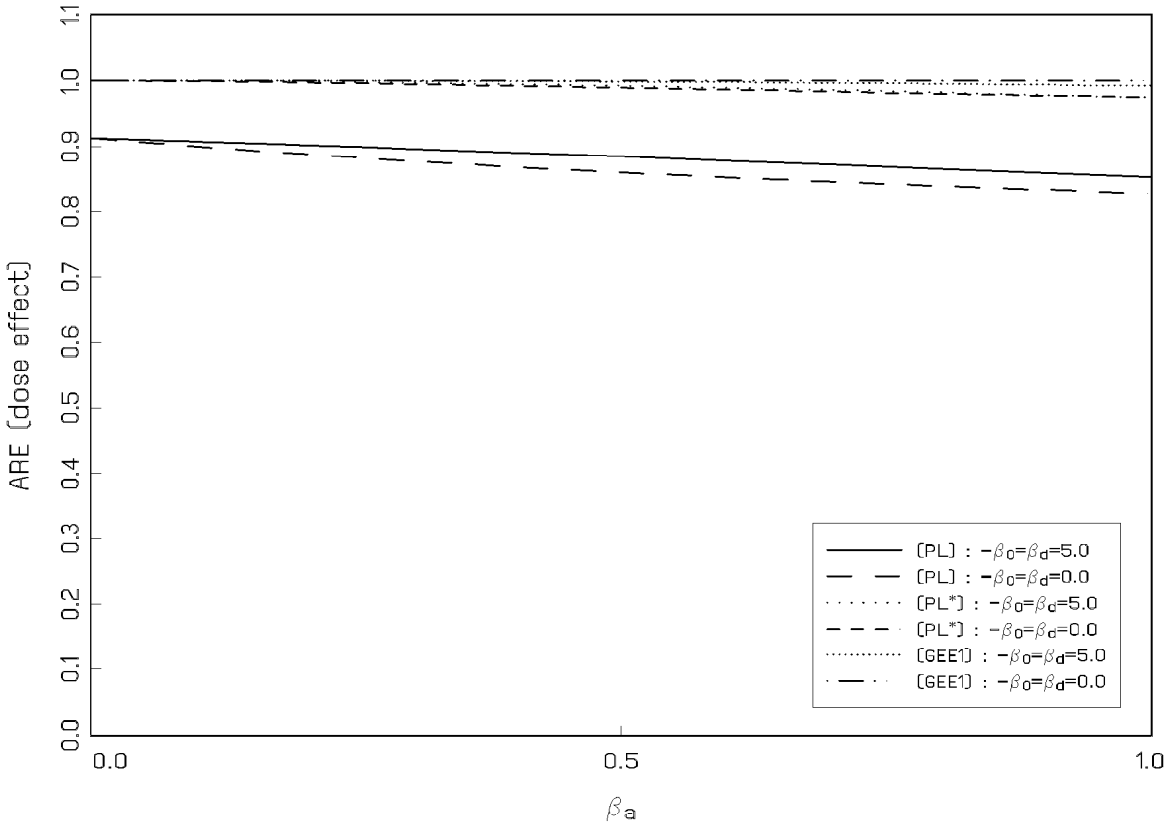

Figure 6.1: *Simulation Results: Asymptotic Relative Efficiency of GEE2 versus PL and GEE1 for the Dose Effect Parameter in a Marginally Specified Odds Ratio Model*

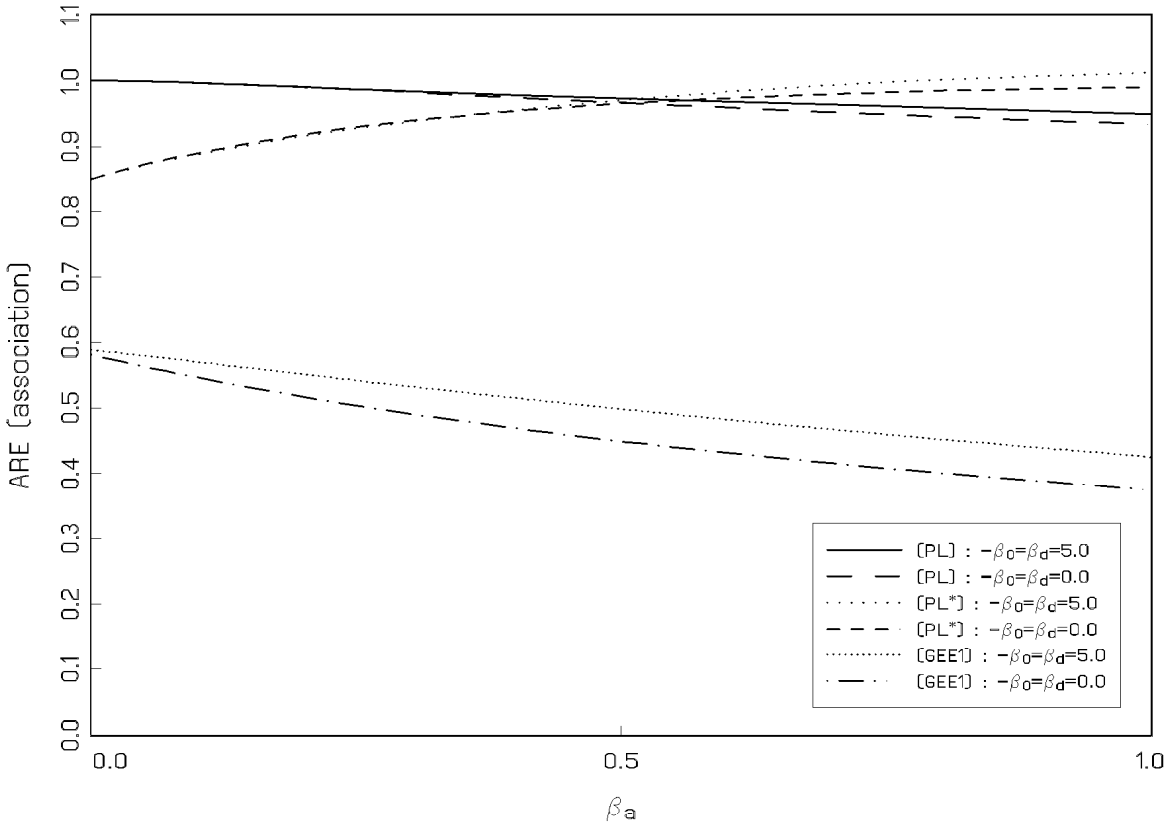

Figure 6.2: *Simulation Results: Asymptotic Relative Efficiency of GEE2 versus PL and GEE1 for the Association Parameter in a Marginally Specified Odds Ratio Model*

Table 6.3: *NTP Studies: Parameter Estimates (standard errors) for a Marginal Odds Ratio Model fitted with PL, GEE1 and GEE2.*

Study	β_0	β_d	β_a
	Newton Raphson PL Estimates		
DEHP	-3.98 (0.30)	5.57 (0.61)	1.10 (0.27)
DYME	-5.73 (0.46)	8.71 (0.94)	1.42 (0.31)
	Fisher scoring PL Estimates		
DEHP	-3.98 (0.30)	5.57 (0.61)	1.11 (0.27)
DYME	-5.73 (0.47)	8.71 (0.95)	1.42 (0.35)
	GEE2 Estimates		
DEHP	-3.69 (0.25)	5.06 (0.51)	0.97 (0.23)
DYME	-5.86 (0.42)	8.96 (0.87)	1.36 (0.34)
	GEE1 Estimates		
DEHP	-4.02 (0.31)	5.79 (0.62)	0.41 (0.34)
DYME	-5.89 (0.42)	8.99 (0.87)	1.46 (0.75)

Table 6.3 shows that the parameter estimates, obtained by either the pseudo-likelihood or generalized estimating equations approach, are comparable. Because main interest is focused on the dose effect, we used $p\ell^*$ rather than $p\ell$. Dose effects and association parameters are always significant, except for the GEE1 association estimates. For this procedure β_a is not significant for the DEHP study and marginally significant for the DYME study. The GEE1 standard errors for β_a are much larger than for their PL and GEE2 counterparts. The GEE2 standard errors are the smallest among the different estimating approaches, which is in agreement with findings in previous sections. Furthermore, it is observed that the standard errors of the Newton-Raphson PL algorithm are generally slightly smaller than those obtained using Fisher scoring, which is in line with other empirical findings. On the other hand, the Newton-Raphson procedure is computationally slightly more complex in this case. The time gain of Fisher scoring however is negligible.

Table 6.4 presents the time (in seconds) needed for each procedure. As was ex-

Table 6.4: *NTP Studies: Time (in seconds) needed for the PL, GEE1 and GEE2 Procedures.*

Study	GEE2	PL (Fisher Scoring)	PL (Fisher Scoring)	GEE1
		Classical representation	GLM representation	
DEHP	1280	116	72	25
DYME	801	110	76	26

pected, GEE2 is relatively time consuming. Then comes the PL estimating approach in its classical form, followed by the generalized linear model type representation, which is be computationally less complex. As anticipated, GEE1 is the least complicated fitting procedure.

6.6 Conclusion

We considered both generalized estimating equations (GEE1 and GEE2) and pseudo-likelihood (Le Cessie and Van Houwelingen 1994) as alternatives for maximum likelihood for the analysis of exchangeable clustered binary data, using a marginal odds ratio model. The applicability to longitudinal data needs further investigation, since they usually do not satisfy the exchangeability assumption on which this work heavily relies: they beg for more complex association structures. First, we have shown that, upon rewriting the pseudo-likelihood and its corresponding score equations, both GEE and PL are similar in spirit. Pseudo-likelihood allows the estimation of both main effect parameters and association parameters, whereas GEE1 is restricted to main effect parameters. In addition, a nice and intuitively appealing class of inferential tools has been proposed by Geys, Molenberghs and Ryan (1999) for the PL case. Depending on whether scientific interest focuses mainly on the main effects or shifts towards the association parameters, different PL versions can be considered. When the main interest lies in the marginal mean parameters, GEE1 has some advantage. Compared to GEE1, PL has a nearly equal asymptotic relative efficiency performance, while the additional computational burden is minor. In contrast, when some interest lies in the estimation of the association parameters as well, we advocate the use of PL. GEE1 can become very inefficient and should not be used for

estimating association parameters, unless strong confidence can be placed in the working assumption. 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, GEE1 and PL can be used with very large clusters.

Chapter 7

Analysis of Toxicology Data with Individual-level Covariates

7.1 Introduction

Over the last decades, a large number of models have been suggested for clustered binary data. Such models are typically considered to fall into one of two classes: cluster-specific (CS) or population-averaged (PA). Section 1.3.3 described the differences between both classes. The PA approach is most appropriate for assessing effects of cluster-level covariates. Cluster-level covariates take on the same value for every unit in the cluster. The effects of within-cluster covariates can also be estimated from these models, but their interpretations are based on the overall population. In a CS model, covariate effects are measured conditional on a cluster-specific parameter.

The standard approach for many teratological applications is to use a population-averaged model with primary interest on evaluating dose-response effects and where the covariate level is considered to be constant over a litter of animals. Lately however, interest for potential effects of individual-specific covariates, such as for example, the position of a fetus within the uterine horn, has been growing. This may also affect the probability of malformation. In this chapter we present population-averaged and cluster-specific modelling strategies that can adjust for these effects and we apply them to the heatshock studies, described in Section 2.2.

For the heatshock studies, described in Section 2.2, the vector of exposure co-

variates must incorporate both exposure level, d_{ik}, and duration, t_{ik}, for the kth embryo of the ith dam. Furthermore, models must be formulated in such a way that departures from Haber's premise of the same adverse response level for any equivalent multiple of dose times duration can easily be assessed. The exposure metrics in these models are the cumulative heat exposure, $d_{ik} \times t_{ik}$, which will be denoted by dt_{ik}, and the effect of duration of exposure at temperatures above normal body temperature, $t_{ik} \times \delta_{d_{ik}}$ (in short t^*_{ik}), where

$$\begin{aligned} \delta_{d_{ik}} &= 1 \text{ if } d_{ik} > 37^{\circ}\text{C} \\ &\quad\ 0 \text{ otherwise.} \end{aligned}$$

Williams, Molenberghs and Lipsitz (1996) applied a maximum likelihood estimation procedure and two approaches, based on generalized estimating equations, to investigate the effects of heat stress exposure on the joint distribution of multiple ordinally measured developmental outcomes. They argue that, while genetic factors are still expected to exert an influence on the vulnerability of embryos from a common dam, direct exposure to individual embryos reduces the need to account for such litter effects. However, they note that this may be too strong an assumption. In this chapter we will concentrate on models, that allow for individual-specific covariates, while clustering induced by litter effects is also taken into account. Furthermore, interest is focused on the comparison of several possible association structures. The specific form of the heatshock study allows us to quantify the association between different embryos from the same initial dam in terms of genetic as well as environmental factors, in contrast to the more standard teratology studies where such a decomposition is not possible.

Within the class of PA models, we consider conditionally and marginally specified models. While the conditionally specified MR model, introduced in Chapters 3 and 4, was very flexible for exchangeable clustered binary data, we show here that limitations are severe, as soon as there are individual level covariates. Therefore, in this chapter, we mainly focus on marginal models, which can further be subdivided into likelihood based and non-likelihood (e.g. generalized estimating equations) approaches. The likelihood based model proposed by Bahadur (1961) readily extends to the context of individual-specific covariates. Associations for this model are measured in terms of correlations. However, in order to obtain a valid probability distribution, more stringent restrictions must be imposed on the correlations

as cluster size increases (Kupper and Haseman 1978; Molenberghs, Declerck and Aerts 1998). Alternatively, odds ratio models such as the one proposed by Molenberghs and Lesaffre (1994) could be used. For their model, only mild constraints on the association parameters apply. However, no closed form expressions can be obtained for third and higher order probabilities, rendering the model computationally cumbersome. Prentice (1988) advocates an extension of the first order generalized estimating equations approach, proposed by Liang and Zeger (1986) that allows for joint estimation of marginal response probabilities and pairwise correlations. However, GEE1 has only limited applicability when some interest is placed on the estimation of correlation parameters. In the context of the heatshock studies, the comparison of different association structures is of particular interest; we therefore also consider a set of second order generalized estimating equations, as proposed by Liang, Zeger and Qaqish (1992).

Finally, within the class of CS models, we study a mixed-effects logistic model as an alternative way of accounting for intra-litter heterogeneity as well as a conditional logistic method. In the mixed-effect logistic procedure cluster effects are removed by assuming that they are manifestations of a random variable and integrating over their distribution. With conditional likelihood, one conditions on the sufficient statistics for the cluster effects (Ten Have, Landis and Weaver 1995; Conaway 1989).

Sections 7.2 and 7.3 respectively describe population-averaged and cluster-specific modelling approaches. Section 7.4 presents a simple goodness-of-fit testing approach for clustered binary data. In Section 7.5 we apply the different methods to the heatshock data. Several possible association structures are studied, allowing an evaluation of both generic and environmental components of the intralitter correlation. An overview of conclusions and remarks is given in Section 7.6.

7.2 Population-averaged Models

7.2.1 Conditionally Specified Models

In Chapter 3, we introduced the likelihood-based model (3.3), proposed by Molenberghs and Ryan (1999), for exchangeable clustered binary data. Extending their model to individual-level covariates is, at the formal level, straightforward. With

notation introduced in Chapter 3, the model becomes:

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i;\boldsymbol{\Theta}_i,n_i) \;=\; \exp\left\{\sum_{k=1}^{n_i}\theta_{ik}y_{ik}+\sum_{k<k'}\delta_{ikk'}y_{ik}y_{ik'}-A(\boldsymbol{\Theta}_i)\right\}.$$

However, while this model benefits from the elegance and simplicity of exponential family theory, it turns out that it is not entirely appropriate within the context of clustered binary data with covariates specific to each observation. To illustrate this point, consider a cluster of size 2 yielding two outcomes (Y_1, Y_2). The conditional probability of observing $Y_1 = y_1$ given the other animal in the cluster is malformed is:

$$P(Y_1|Y_2=1)=\frac{\exp\left[(\theta_1+\delta_{12})y_1\right]}{\exp(-\theta_1-\delta_{12})+\exp(\theta_1+\delta_{12})}. \tag{7.1}$$

Assuming the association parameter δ_{12} is constant, (7.1) does not depend on the covariates for the second individual. In addition, the marginal malformation probability of the first individual is:

$$P(Y_1=1)=\exp(\theta_1-\theta_2-\delta_{12}-A(\boldsymbol{\Theta}))+\exp(\theta_1+\theta_2+\delta_{12}-A(\boldsymbol{\Theta})),$$

which depends on the covariates of the second individual, as reflected by $A(\boldsymbol{\Theta})$. Both properties are undesirable when fetuses are randomized to a certain dose group after sacrifice of the maternal dam. This is a strong warning against the use of conditionally specified models. Hence, they will be ignored henceforth.

7.2.2 Likelihood-based Marginal Models

The Bahadur model has been used by several authors in the context of toxicological experiments (Kupper and Haseman 1978; Altham 1978) and can thus be considered an important representative of the marginal family. Bahadur (1961) describes the joint distribution of clustered binary data for a single outcome in terms of marginal means $\boldsymbol{\pi_i} = (\pi_{i1},\ldots,\pi_{in_i})^T$ and marginal correlations $\boldsymbol{\rho_i} = (\rho_{i12},\rho_{i13},\ldots,\rho_{i12\ldots n_i})^T$. The closed form probability mass function is given as:

$$\begin{aligned} f(\boldsymbol{y_i},\boldsymbol{\pi_i},\boldsymbol{\rho_i}) \;=\; & \prod_{k=1}^{n}\pi_{ik}^{y_{ik}}(1-\pi_{ik})^{(1-y_{ik})}\times\left(1+\sum_{k_1<k_2}\rho_{ik_1k_2}r_{ik_1}r_{ik_2}\right.\\ & \left.+\sum_{k_1<k_2<k_3}\rho_{ik_1k_2k_3}r_{ik_1}r_{ik_2}r_{ik_3}+\cdots+\rho_{i1\ldots n}r_{i1}\ldots r_{in}\right). \end{aligned}$$

Unfortunately, the association parameters may be subject to severe restrictions. In practice, three-way and higher order associations are often difficult to interpret and can be neglected. Therefore, a working higher order independence assumption is often plausible. If we set all three- and higher way correlations equal to zero, Bahadur's representation simplifies to:

$$f(\boldsymbol{y}_i|\boldsymbol{\pi}_i, \boldsymbol{\rho}_i) = \prod_{k=1}^{n_i} \pi_{ik}^{y_{ik}}(1-\pi_{ik}^{1-y_{ik}}) \left(1 + \sum_{k<l} \rho_{ikl} e_{ik} e_{il}\right),$$

with

$$e_{ik} = \frac{y_{ik} - \pi_{ik}}{\sqrt{\pi_{ik}(1-\pi_{ik})}}.$$

In the simple case of equicorrelation and a constant mean, Bahadur gives ranges in the parameter space within which this second order approximation is a valid probability distribution. Declerck, Aerts and Molenberghs (1998) have shown that the range of positive second order associations is markedly enlarged in a four-way Bahadur model. But fitting higher order Bahadur models is difficult, due to the increasingly complicated nature of the restrictions on the parameter space. Using appropriate link functions, the marginal mean parameters π_{ik} $(k = 1, \ldots, n_i)$, as well as the marginal correlations $\rho_{ikl}(k < l)$, can be modelled as a function of a $(n_i\,(n_i+1)\,/2 \times p)$ covariate matrix X_i and a parsimonious $(p \times 1)$ vector of regression parameters $\boldsymbol{\beta}$. The logistic link function is a natural choice for π_{ik}, while Fisher's z-transform is convenient to model ρ_{ikl}. This leads to the following generalized linear model:

$$\boldsymbol{\eta}_i = \begin{pmatrix} \ln\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right)_{k=1}^{n_i} \\ \ln\left(\frac{1+\rho_{ikl}}{1-\rho_{ikl}}\right)_{k<l} \end{pmatrix} = X_i\boldsymbol{\beta}. \tag{7.2}$$

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is defined as the solution to $\boldsymbol{U}(\hat{\boldsymbol{\beta}}) = 0$ with $\boldsymbol{U}(\boldsymbol{\beta})$ the score function. A Fisher scoring or Newton-Raphson algorithm can be used to obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

7.2.3 Generalized Estimating Equations

For marginal odds ratio models, generalized estimating equations have been introduced in Chapter 6. Here, we model the association in terms of correlations, in order to enable an easy comparison with the Bahadur model.

GEE1

GEE1 were first proposed by Liang and Zeger (1986) and require only the correct specification of the univariate marginal distributions provided one is willing to adopt working assumptions about the association structure. However, the appropriateness of GEE1 is reduced when the association parameters themselves are of scientific interest. Assuming that a function of the mean can be written as a linear function of regression parameters $\boldsymbol{\beta}$, the generalized estimating equations for $\boldsymbol{\beta}$ are given by:

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\pi}_i) = 0,$$

where $V_i = \phi A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}$, $R_i(\boldsymbol{\alpha})$ is a $n_i \times n_i$ working correlation matrix fully specified by the vector of parameters $\boldsymbol{\alpha}$, and A_i is an $(n_i \times n_i)$ diagonal matrix with diagonal elements $\pi_{ik}(1 - \pi_{ik})$. Usually, the working correlation matrix depends on unknown parameters which have to be estimated. Standard procedures, such as the SAS/STAT procedure GENMOD (1997) and the Oswald functions in Splus (Smith, Robertson and Diggle 1996), that include GEE1 capabilities use an iterative fitting process, where estimation of the parameters $\boldsymbol{\alpha}$ is based on standardized residuals. The model based estimator of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ is given by I_0^{-1}, where

$$I_0 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\pi}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}}.$$

The empirically corrected variance estimator (Liang and Zeger 1986), is $I_0^{-1} I_1 I_0^{-1}$, where

$$I_1 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\pi}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i) V_i^{-1} \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}}.$$

Williamson, Lipsitz and Kim (1997) wrote a SAS macro for GEE1 which is based on Prentice's approach. The latter considered an extension of the GEE1 approach of Liang and Zeger (1986) that allows joint estimation of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in both the marginal response probabilities and the pairwise correlations. A GEE1 estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ may be defined as a solution to:

$$\begin{aligned} \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\pi}_i) &= 0 \\ \sum_{i=1}^{N} \boldsymbol{E}_i^T \boldsymbol{W}_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{\delta}_i) &= 0, \end{aligned}$$

where

$$Z_{ijk} = \frac{(Y_{ij} - \pi_{ij})(Y_{ik} - \pi_{ik})}{\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})}}$$

and $\delta_{ijk} = E(Z_{ijk})$. Under exchangeability we have $\delta_{ijk} = \rho_i$, the correlation between any two outcomes of the same cluster i. This can be reparametrized in terms of α, using Fisher's z-transformation: $\alpha = \ln(1+\rho) - \ln(1-\rho)$. The joint asymptotic distribution of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ is Gaussian with mean zero and with variance-covariance matrix consistently estimated by N times

$$\begin{pmatrix} \boldsymbol{A} & 0 \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ 0 & \boldsymbol{C} \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{A} &= (\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i)^{-1}, \\
\boldsymbol{B} &= (\sum_{i=1}^{N} \boldsymbol{E}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{E}_i)^{-1} (\sum_{i=1}^{N} \boldsymbol{E}_i^T \boldsymbol{W}_i^{-1} \frac{\partial \boldsymbol{Z}_i}{\partial \boldsymbol{\beta}}) (\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i)^{-1}, \\
\boldsymbol{C} &= (\sum_{i=1}^{N} \boldsymbol{E}_i^T \boldsymbol{W}_i^{-1} \boldsymbol{E}_i)^{-1}, \\
\Lambda_{11} &= \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i, \\
\Lambda_{12} &= \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Z}_i) \boldsymbol{W}_i^{-1} \boldsymbol{E}_i, \\
\Lambda_{21} &= \Lambda_{12}, \\
\Lambda_{22} &= \sum_{i=1}^{N} \boldsymbol{E}_i^T \boldsymbol{W}_i^{-1} \mathrm{Cov}(\boldsymbol{Z}_i) \boldsymbol{W}_i^{-1} \boldsymbol{E}_i,
\end{aligned}$$

and $\mathrm{Var}(\boldsymbol{Y}_i)$, $\mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Z}_i)$ and $\mathrm{Var}(\boldsymbol{Z}_i)$ respectively estimated by $(\boldsymbol{Y}_i - \boldsymbol{\pi}_i)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i)^T$, $(\boldsymbol{Y}_i - \boldsymbol{\pi}_i)(\boldsymbol{Z}_i - \boldsymbol{\delta}_i)^T$ and $(\boldsymbol{Z}_i - \boldsymbol{\delta}_i)(\boldsymbol{Z}_i - \boldsymbol{\delta}_i)^T$. The SAS macro defines:

$$\boldsymbol{Z}_i = \begin{pmatrix} Y_{i1}Y_{i2} \\ Y_{i1}Y_{i3} \\ \vdots \\ Y_{in_i}Y_{i(n_i-1)} \end{pmatrix}$$

Hence, under exchangeability,

$$\begin{aligned} E(Z_{ijk}) &= \pi_{ijk} = \rho\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})} + \pi_{ij}\pi_{ik}, \\ \mathrm{Var}(Z_{ijk}) &= \pi_{ijk}(1-\pi_{ijk}), \\ \frac{\partial E(Z_{ijk})}{\partial \alpha} &= \frac{2\exp(\alpha)}{(\exp(\alpha)+1)^2}\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})}. \end{aligned}$$

The matrix $\boldsymbol{C}$ then reduces to:

$$C = \left(\frac{2\exp(\alpha)}{(\exp(\alpha)+1)^2}\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})}\right)^2 \frac{1}{\pi_{ijk}(1-\pi_{ijk})}.$$

To obtain the variance-covariance matrix of the correlation parameters $\boldsymbol{\rho}$, one can apply the delta method. In the case of exhangeability we multiply the standard error of α with a factor $2\exp(\alpha)/(\exp(\alpha)+1)^2$ to obtain the standard error of ρ.

GEE2

The GEE2 approach naturally accomodates individual-level covariates in the estimation of marginal response probabilities. For each cluster, define

$$\boldsymbol{w}_i = (y_{i1}, \ldots, y_{in_i}, y_{i1}y_{i2}, \ldots, y_{in_i-1}y_{in_i})^T,$$

a vector of $n_i + \binom{n_i}{2}$ components. Further, let $\boldsymbol{\Theta}_i = (\boldsymbol{\pi}_i^T, \boldsymbol{\rho}_i^T)^T$ which depends on a $p \times 1$ vector of regression parameters $\boldsymbol{\beta}$ through the generalized linear model (7.2). Estimation of $\boldsymbol{\beta}$ is accomplished by solving the following second order estimating equations:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T \left(T_i^{-1}\right)^T V_i^{-1}(\boldsymbol{W}_i - E(\boldsymbol{W}_i)) = \boldsymbol{0},$$

with $X_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\beta}$, $T_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\Theta}_i$ and $V_i = \mathrm{Cov}(\boldsymbol{W}_i)$. Calculation of all matrices involved is straightforward with the exception of the covariance matrix, which contains third and fourth order probabilities. To this end, the three-way and higher order correlations are set equal to zero. As before, the parameter estimates $\hat{\boldsymbol{\beta}}$ can then be calculated using, for example, a Fisher scoring algorithm. Provided the first and second order models have been correctly specified, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and has an asymptotic multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and variance-covariance matrix consistently estimated by:

$$V(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} X_i^T \hat{T}_i^{-T} \hat{V}_i^{-1} \hat{T}_i^{-1} X_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{U}_i(\hat{\boldsymbol{\beta}})\boldsymbol{U}_i(\hat{\boldsymbol{\beta}})^T \left(\sum_{i=1}^{N} X_i^T \hat{T}_i^{-T} \hat{V}_i^{-1} \hat{T}_i^{-1} X_i\right)^{-1}.$$

7.3 Cluster-specific Models

To introduce cluster-specific approaches, consider the generalized linear mixed model (Breslow and Clayton 1993):

$$\boldsymbol{\eta}_i = g(\boldsymbol{\pi}_i) = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i,$$

where g is an appropriate link function, $\boldsymbol{\beta}$ is a vector of unknown fixed effects with known model matrix X_i, $\boldsymbol{b}_i$ is a vector of unknown random effects with known model matrix Z_i and $\boldsymbol{\pi}_i$ is the conditional mean of the observations, given the random effects $\boldsymbol{b}_i$. The random effects $\boldsymbol{b}_i$ can have any distribution (Lee and Nelder 1996). Standard approaches assume them to be normally distributed with mean $\mathbf{0}$ and variance D. As pointed out by Zeger and Karim (1991), the beta-binomial and Poisson-gamma models were among the earliest extensions of random-effects models to the generalized linear context. However, these models only allow for cluster-level covariates. Within the mixed-effects models framework, several approaches can be used for estimating the parameters of interest ($\boldsymbol{\beta}$).

7.3.1 Marginal Likelihood Approach

In a marginal likelihood approach, the random intercept terms are integrated out. Likelihood inference for generalized linear mixed models requires evaluation of integrals, the dimension of which is equal to the number of random effects. Zeger and Karim (1991) avoid this need for numerical integration by casting the generalized linear mixed-effects model in a Bayesian framework and using the Gibbs sampler. Breslow and Clayton (1993) apply Laplace's method for integral approximation, whereas Wolfinger and O'Connell (1993) propose a pseudo-likelihood (PL) procedure. Starting with an initial estimate of the conditional mean vector $\boldsymbol{\pi}_i$, they construct a vector of pseudo-data (i.e. a linearized version of the link function) and fit a weighted linear mixed model to them. Iteratively solving the mixed model equations, they obtain an updated estimate of the mean. This process then iterates until convergence. This procedure has been implemented in SAS using the GLIMMIX macro which iteratively calls the MIXED procedure. Using this macro without random effects, but with a compound symmetry covariance structure for the correlated error terms describes a PA model, where the individuals within a cluster are assumed to be exchangeable (compound symmetry model). This procedure is related

to the GEE concept, where an exchangeable working correlation matrix is assumed. A criticism of GEE is that it generally does not correspond directly to a likelihood (which could be used to calculate deviances), even though some approximations to a likelihood ratio statistic have been proposed (Rotnitzky and Jewell 1990). Random-effects models, on the contrary, correspond to a likelihood (or pseudo-likelihood) function. But very little research has been done on the form of the deviance reported in the output of the GLIMMIX macro (Littell et al. 1996). Moreover, Littell et al. (1996) warn that one should be aware that relatively little research has been done on the small-sample properties of inference statistics for the generalized linear mixed model. The test statistics are basically reasonable-looking extensions of standard tests for mixed models and generalized linear models. More work is needed to either validate or modify these procedures. Finally, Neuhaus and Segal (1997) warn that approximate methods based on Laplace or Taylor series approximations (cf. Breslow and Clayton 1993; Wolfinger and O'Connell 1993) may require modification before their asymptotic bias can be competitive with mixed-effects model methods such as the Gibbs sampler that provide consistent estimation.

In this section we will focus on cluster-specific mixed-effects logistic models, where the intercept terms b_i are allowed to vary from cluster to cluster, according to a normal distribution:

$$\text{logit}P(Y_{ik} = 1|b_i, \boldsymbol{x}_{ik}) = \boldsymbol{x}_{ik}\boldsymbol{\beta} + b_i. \tag{7.3}$$

In this formulation, $\boldsymbol{x}_{ik}$ denotes the kth row of the design matrix X_i. The regression parameters ($\boldsymbol{\beta}_{CS}$) in this CS mixed-effects logistic model measure the change in the conditional logit of the probability of response with a unit increase in the corresponding covariates for individuals at the same random-effects level (e.g. within a cluster with only individual-level covariates). The association between littermates is induced by the random intercept. Because cluster sizes for developmental toxicology studies are relatively small, more complex random-effect structures can seldom be addressed from a practical perspective. Further, while mixed-effects models uniquely specify marginal models by integrating (7.3) over the random-effect terms $\boldsymbol{b}$, the reverse is not true since there is a many-to-one mapping between CS and PA models. Integration of (7.3) leads to the following PA model for Y_{ik}:

$$P(Y_{ik} = 1|\boldsymbol{x}_{ik}) = \int \frac{\exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + \boldsymbol{b})}{1 + \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + \boldsymbol{b})} f(\boldsymbol{b})d\boldsymbol{b}$$

In contrast to the cluster-specific regression parameters $\boldsymbol{\beta}_{CS}$, regression parameters in a PA model ($\boldsymbol{\beta}_{PA}$) measure the change in the logit of the "success" probability for a unit increase in the corresponding covariates. Neuhaus, Kalbfleisch and Hauck (1991) have derived an interesting approximate relationship between PA and CS parameters in the above described context:

$$\boldsymbol{\beta}_{PA} \simeq \boldsymbol{\beta}_{CS}[1 - \rho(0)], \tag{7.4}$$

where $\rho(0)$ refers to the intracluster correlation when the covariate has no effect. Since $0 \leq \rho \leq 1$, this suggests that, at least for small $\boldsymbol{\beta}_{CS}$, $|\boldsymbol{\beta}_{PA}| < |\boldsymbol{\beta}_{CS}|$, so that the population-averaged effect is smaller than the cluster-specific effect.

7.3.2 Conditional Likelihood Approach

The conditional likelihood approach eliminates the random intercept terms $\boldsymbol{b}$ by conditioning on sufficient statistics. It can be written as:

$$\prod_{i=1}^{N} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\beta}, b_i)}{P(S_i = s_i | \boldsymbol{x}_i, \boldsymbol{\beta}, b_i)},$$

where $S_i = \sum_{k=1}^{n} Y_{ik}$ is a sufficient statistic. By the sufficiency of S_i, the conditional likelihood does not depend on the random-effects distributions (Neuhaus and Kalbfleisch 1997). One drawback of this conditional likelihood method, particularly with binary data, is that the method only uses data from clusters that are discordant on both the outcome and the covariates. The conditioning also removes effects from cluster level covariates together with the random effects. Hence, it is not possible to estimate effects from cluster level covariates. Further, the conditional likelihood approach can also lose efficiency because S_i depends on $\boldsymbol{\beta}$. Neuhaus and Lesperance (1996) have investigated the ARE of the conditional likelihood estimators relative to the estimators obtained with a mixed-effects logistic model. They show that the ARE is a decreasing function of within-cluster covariate correlation. However, for a fixed covariate correlation, the ARE increases as cluster size increases. This is a consequence of fewer discarded clusters, resulting from the decreasing probability of concordance.

7.4 Goodness-of-Fit for Likelihood Based Models with Clustered Binary Data

In order to evaluate how effective models are in describing the outcome variable, we need to assess the quality of their fit. Lipsitz, Fitzmaurice and Molenberghs (1996) note that for the special case of a binary response, several methods for assessing the goodness-of-fit of binary logistic regression models have been proposed. All these methods are based on the notion of partitioning the covariate space into groups or regions. Tsiatis (1980) proposed a goodness-of-fit statistic for the logistic regression model for a given partition of the covariate space, but he did not provide a method for partitioning the covariate space into suitable regions. Hosmer and Lemeshow (1989) proposed the partition of subjects into groups or regions on the basis of the percentiles of the predicted probabilities from the fitted logistic regression model. To construct a goodness-of-fit measure for clustered binary data, we adapted the methods proposed by Hosmer and Lemeshow (1989) and Tsiatis (1980). Following these authors, groups are constructed according to deciles of the predicted malformation probabilities in each temperature-duration combination. Given this partition, the goodness-of-fit statistic is formulated by defining $G-1$ group indicators (in our example, $G=10$):

$$I_{ik}^{g} = \begin{cases} 1 & \text{if } \hat{\pi}_{ik} \text{ is in region } g \ (g = 1, \ldots, G-1) \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{\pi}_{ik}$ is the estimated malformation probability of the kth individual within the ith cluster, calculated from the model that takes into account the clustering between the individuals. For example, in the context of the heatshock studies, the following model is considered:

$$\ln\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \beta_0 + \beta_{t^*} t_{ik}^{*} + \beta_{dt} dt_{ik} + \sum_{g=1}^{G-1} I_{ik}^{g} \gamma_g .$$

The association is modelled similarly as in the model for which the goodness-of-fit is assessed. One possible choice would be an exchangeable correlation structure. Other more complicated structures will be described in the following section. If the mean structure in the original model is correctly specified, then $\gamma_1 = \cdots = \gamma_{G-1} = 0$. Moore and Spruill (1975) note that, even though I_{ik}^{g} is based on random quantities

$\hat{\pi}_{ik}$, the partition can be treated asymptotically as if it were based on the true π_{ik}. To test the goodness-of-fit of the model, one can use either a likelihood ratio, Wald or score statistic to test $H_0 : \gamma_1 = \cdots = \gamma_{G-1} = 0$. For large samples, each of these statistics has approximately a χ^2 distribution with $G-1$ degrees of freedom, if the model under the null hypothesis is correctly specified. We suggest the use of the likelihood ratio statistic, since it is simple to calculate and is fairly powerful. For large samples, all estimated expected frequencies should typically be greater than 1 and at least 80 percent should be greater than 5. Otherwise, one can collapse some frequencies, reducing the number of groups G (Lipsitz, Fitzmaurice and Molenberghs 1996). Hosmer and Lemeshow (1989) noted that $G = 6$ should be a minimum, since a test statistic calculated from fewer than 6 groups will usually have low power and thus indicates that the model fits well. Note that in the goodness-of-fit assessment described above, correlation is essentially treated as a nuisance parameter and interest is focused on the relationship between the covariates and the probability of response. Recent work has shown there may be disadvantages in the use of goodness-of-fit tests based on the ones proposed by Hosmer and Lemeshow (Hosmer, Hosmer, Lemeshow, Le Cessie 1997). Decisions on model fit may depend more on choice of cutpoints than on lack-of-fit and their test statistic may have relatively low power with small sample sizes. Developing improved goodness-of-fit test statistics for likelihood based models for clustered binary data is a topic of further research.

7.5 Analysis of Heatshock Study

As mentioned in Section 7.1, the specific form of the heatshock study allows us to quantify the association between different embryos from the same initial dam in terms of genetic as well as environmental factors. Therefore, we can consider several possible designs for the association structure.

Design 1

The mean parameters are assumed to be a linear function of dt and t^*, while the pairwise associations are assumed to equal the constant value ρ. Hence, the design matrix $\boldsymbol{X}_i$ for the ith cluster is a matrix with $n_i + \binom{n_i}{2}$ rows and 4

columns:

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & t^*_{i1} & dt_{i1} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t^*_{in_i} & dt_{in_i} & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix}, \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_{t^*} \\ \beta_{dt} \\ \rho \end{pmatrix}$$

Design 2

As in the previous case, the mean parameters are assumed to be a linear function of dt and t^*, but the pairwise associations are modelled as:

$$\delta_{ikl} = \begin{cases} \alpha + \gamma & \text{if } t^*_{ik} = t^*_{il} \text{ and } dt_{ik} = dt_{il}, \\ \alpha & \text{otherwise.} \end{cases}$$

A significant α would then mean a large association within all clusters, while a significant γ would indicate an extra association within the same duration-temperature group.

Design 3

This design is analogous to the previous one, except that the pairwise association is now modelled by specifying an extra association parameter within each cumulative exposure group:

$$\delta_{ikl} = \begin{cases} \alpha + \gamma & \text{if } dt_{ik} = dt_{il}, \\ \alpha & \text{otherwise.} \end{cases}$$

Design 4

Here, the main difference with design 3 is that we model the mean parameters as a linear function of the exposure level, d, and duration t. The pairwise association is modelled in a similar fashion as in the previous design:

$$\delta_{ikl} = \begin{cases} \alpha + \gamma & \text{if } d_{ik} = d_{il} \text{ and } t_{ik} = t_{il}, \\ \alpha & \text{otherwise.} \end{cases}$$

Design 5

Analogous to designs $1-3$, we assume that the mean parameters are a linear function of dt and t^*, but the pairwise associations are modelled as a linear function of the "quadratic distances" between any two cumulative exposure values, i.e.

$$\delta_{ikl} = \alpha + \gamma(dt_{ik} - dt_{il})^2.$$

For the design matrix $\boldsymbol{X}_i$ and the vector of regression parameters $\boldsymbol{\beta}$, this implies the following choices:

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & t^*_{i1} & dt_{i1} & 0 & 0 \\ \vdots & & & & \\ 1 & t^*_{in_i} & dt_{in_i} & 0 & 0 \\ 0 & 0 & 0 & 1 & (dt_{i1} - dt_{i2})^2 \\ 0 & 0 & 0 & 1 & (dt_{i1} - dt_{i3})^2 \\ \vdots & & & & \\ 0 & 0 & 0 & 1 & (dt_{i(n_i-1)} - dt_{in_i})^2 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_{t^*} \\ \beta_{dt} \\ \alpha \\ \gamma \end{pmatrix}.$$

7.5.1 Population Averaged Models

In Section 7.2.1 it was shown that conditionally specified models are less suitable for developmental toxicity studies with individual-level covariates. Therefore, we will restrict attention to the Bahadur model and generalized estimating equations as representatives of PA models.

Bahadur Model

Restrictions on the parameter space in the Bahadur representation present no problem for our data. As mentioned in Section 7.1, severe restrictions may arise for larger cluster sizes, but cluster sizes are relatively small for the heatshock studies. Table 7.1 gives the parameter estimates (standard errors) using the Bahadur model for the different designs.

Table 7.1: *Heatshock Study: Parameter Estimates (standard error) for the Bahadur Model, Applying Different Designs for the Association Structure.*

Outcome	Par.	Design				
		1	2	3	4	5
MBN	β_0	-1.84 (0.21)	-1.85 (0.21)	-1.85 (0.20)	-1.85 (0.21)	-1.64 (0.23)
	β_{t^*}	-3.83 (1.69)	-3.75 (1.69)	-3.77 (1.67)	-3.90 (1.68)	-4.03 (1.41)
	β_{dt}	5.88 (1.71)	5.80 (1.70)	5.82 (1.68)	5.96 (1.70)	5.49 (1.46)
	α	0.13 (0.08)	0.13 (0.08)	0.13 (0.08)	0.13 (0.08)	0.22 (0.08)
	γ		-0.05 (0.21)	-0.04 (0.20)	0.07 (0.24)	-1.26 (0.58)
OPT	β_0	-2.50 (0.23)	-2.41 (0.25)	-2.41 (0.24)	-2.49 (0.24)	-2.43 (0.25)
	β_{t^*}	-3.69 (1.62)	-4.26 (1.72)	-4.27 (1.74)	-4.17 (1.57)	-3.93 (1.64)
	β_{dt}	5.66 (1.58)	6.13 (1.70)	6.13 (1.70)	6.14 (1.58)	5.61 (1.58)
	α	-0.06 (0.07)	-0.10 (0.07)	-0.10 (0.07)	-0.10 (0.07)	0.00 (0.09)
	γ		0.72 (0.28)	0.73 (0.28)	0.59 (0.29)	-0.54 (0.43)
OLF	β_0	-1.47 (0.19)	-1.40 (0.20)	-1.39 (0.20)	-1.44 (0.19)	-1.43 (0.21)
	β_{t^*}	-4.91 (2.00)	-5.61 (1.69)	-5.63 (1.81)	-5.65 (1.84)	-5.52 (1.80)
	β_{dt}	6.69 (1.97)	7.35 (1.69)	7.37 (1.84)	7.49 (1.87)	7.12 (1.74)
	α	0.26 (0.07)	0.22 (0.08)	0.22 (0.08)	0.22 (0.08)	0.29 (0.08)
	γ		0.49 (0.27)	0.50 (0.25)	0.57 (0.32)	-0.37 (0.51)

Midbrain (MBN)

For the MBN response, we observe both an important cumulative exposure effect and an important additional effect of duration of exposure to temperatures above normal body temperature. This indicates a departure from Haber's premise that the probability of observing an adverse effect is the same for each exposure $\times$ duration combination. Further, the coefficients for β_{t^*} are consistently negative, indicating that shorter exposures of the same cumulative exposure cause more developmental damage than longer ones. As expected, the malformation probability tends to increase with increasing cumulative exposures. Figure 7.1 shows how the fetus-level risk surface (i.e. plot of the probability that a fetus is malformed at a given cumulative exposure and a given duration of exposure to a temperature above 37^0C) changes with dt and t^*. For this visual representation, we used the constant association design model and rescaled the covariates between 0 and 1.

Based on the first association design model, there is no evidence of a significant intracluster correlation for MBN responses. Direct exposure to the individual embryos seems to reduce the need to account for litter effects on midbrain malformations. More complex association structures, such as designs 2, 3 and 4 lead to similar conclusions. Furthermore, according to these models, there is no evidence for an extra association contribution for MBN responses on individuals within the same duration-temperature group. Comparing the likelihoods of models obtained with designs 2, 3 and 4 to the constant association model (design 1) yields deviances (D) of resp. 0.052, 0.032 and 0.087 (in the remainder, all deviances will refer to a comparison with the constant association model). In contrast, design 5 yields a significant quadratic distance effect parameter γ (D=3.9028). Hence, the association between any two individuals decreases with the "distance" between their cumulative exposures. The goodness-of-fit deviances ($G = 10$) for all fitted models are tabulated in Table 7.2. All expected malformation frequencies are larger than 1 and the collapsed frequencies within groups are all larger than 5. None of the deviances indicates a lack of fit for the MBN predicted probabilities, compared to a χ^2 distribution with 9 degrees of freedom.

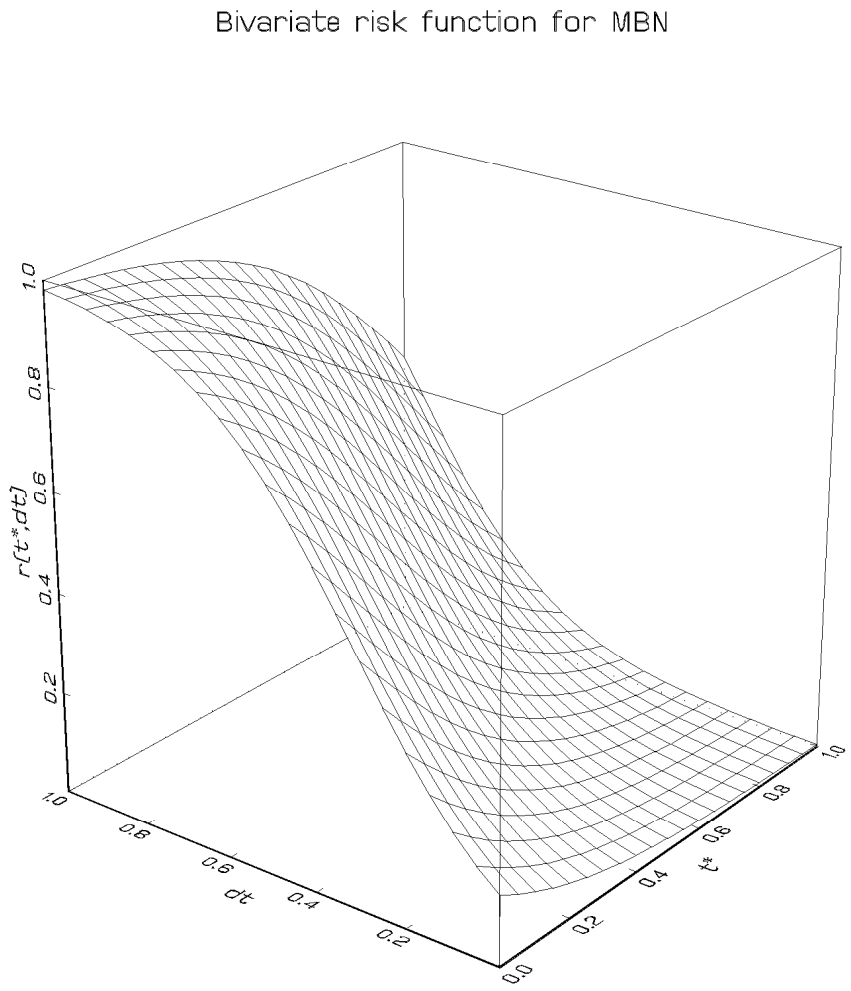

Figure 7.1: *Heatshock Study: Fetus-level Risk Surface for MBN.*

Table 7.2: *Heatshock Study: Goodness-of-fit Deviances (p-values).*

Outcome	Design				
	1	2	3	4	5
MBN	6.78 (0.66)	6.91 (0.65)	6.91 (0.65)	6.75 (0.66)	6.36 (0.70)
OPT	17.77 (0.04)	14.89 (0.09)	14.89 (0.09)	16.93 (0.05)	18.09 (0.03)
OLF	25.18 (0.00)	22.65 (0.01)	22.65 (0.01)	23.00 (0.01)	25.45 (0.00)

Optic System (OPT)

Again we observe an important effect of cumulative exposure and an additional effect of duration of exposure to temperatures above normal body temperature. The clustering parameter α is never significant, indicating that there is no important intra-litter correlation working on the optic system. However, there is evidence for an extra association between animals within the same duration-temperature group (design 2) (D=6.021). Parameter estimates for the two designs (2 and 1) are similar but dt and t^* tend to be slightly more significant for the more complex design 2. Designs 3 and 4 lead to similar results as design 2. There seems to be no evidence for a quadratic distance effect (D=1.622). Further, Table 7.2 shows that designs 4 and 5 give a rather poor fit to the data. Based on these results, the third association design might be the most preferable.

Olfactory System (OLF)

For the OLF outcome, dt and t^* are again highly significant. Design 1 now shows evidence of a significant intra-litter correlation. Furthermore, it follows from design 2 that there might be a (borderline significant) extra contribution of association for individuals within the same duration-temperature group (D=3.6836). The simpler design 3 leads to similar results (D=3.877) and might therefore be preferable. The results of design 4 are comparable with those obtained by the second design (D=3.868). The quadratic distance effect parameter γ is apparently superfluous (D=0.564). However, Table 7.2 shows that all these models fit the data poorly. Including a quadratic main dt effect improves the fit substantially. Adopting one of the two-level association designs rather than assuming a constant association im-

proves the fit of the mean model even further. Again, the best fit was obtained for design 3, yielding a goodness-of-fit deviance of 12.31. The intra-litter correlation becomes even more significant, but the extra association for individuals within the same duration-temperature group is now reduced to non-significance.

Generalized Estimating Equations

Table 7.3 gives the parameter estimates together with model-based and empirically corrected standard errors of the second order generalized estimating equations. In many cases, GEE2 models were hard to fit. For instance, in the case of the more complicated design 5, we were not able to fit GEE2 for any of the outcomes, and therefore we exclude it from the table. For outcome MBN, the results of GEE2 are similar to those obtained using the Bahadur model. The model-based standard errors correspond closely with those calculated by the likelihood method. Furthermore, model-based and empirically corrected (robust) standard errors are close to each other, indicating that complex association designs need not be considered. In contrast, for the OPT outcome, there is a larger gap between model-based and empirically corrected standard errors, especially for design 1. This might indicate that other more complex designs should be considered. In some cases, it may be worthwhile to consider association structures that include a genetic association component (α), a serial association component (γ) and a random covariate effect. Full likelihood methods, as well as second order generalized estimating equations, can only handle the first two components.

Importance of Complex Association Patterns

To illustrate the importance of addressing complex association patterns, Table 7.4 tabulates the parameter estimates (standard errors) for each of the three binary outcomes (MBN, OPT and OLF) respectively for the logistic model, the standard GEE1 procedure and Prentice's extended GEE1 approach under an exchangeable working correlation assumption, and the generalized linear mixed model fitting procedure of Wolfinger and O' Connell (1993), fitted without random effects but assuming a compound symmetry (CSYM) covariance structure. All are PA approaches where individuals within a cluster are assumed to be either independent (logistic) or exchangeable.

Table 7.3: *Heatshock Study: Parameter estimates (model based standard error;empirically corrected standard error) for GEE2, Applying Different Designs for the Association Structure.*

	Par.	Design			
Outcome		1	2	3	4
MBN	β_0	-1.81 (0.21;0.22)	-1.83 (0.21;0.22)	-1.83 (0.21;0.22)	-1.81 (0.21;0.22)
	β_{t^*}	-3.93 (1.66;1.96)	-3.68 (1.62;1.97)	-3.72 (1.62;1.97)	-3.96 (1.66;1.95)
	β_{dt}	5.93 (1.68;1.94)	5.70 (1.64;1.96)	5.74 (1.64;1.96)	5.97 (1.69;1.94)
	α	0.10 (0.08;0.06)	0.11 (0.08;0.07)	0.11 (0.08;0.07)	0.09 (0.08;0.07)
	γ		-0.16 (0.18;0.17)	-0.14 (0.18;0.17)	0.03 (0.25;0.24)
OPT	β_0	-2.49 (0.24;0.22)	-2.42 (0.25;0.23)		
	β_{t^*}	-3.69 (1.70;2.22)	-3.71 (1.79;2.22)		
	β_{dt}	5.64 (1.67;2.27)	5.64 (1.77;2.23)		
	α	-0.05 (0.05;0.04)	-0.07 (0.06;0.06)		
	γ		0.76 (0.25;0.27)		
OLF	β_0	-1.52 (0.22;0.28)			-1.55 (0.21;0.28)
	β_{t^*}	-3.70 (1.49;1.56)			-4.51 (1.44;1.54)
	β_{dt}	5.61 (1.50;1.53)			6.47 (1.40;1.60)
	α	0.51 (0.08;0.14)			0.41 (0.07;0.09)
	γ				-0.09 (0.22;0.23)

Table 7.4: *Heatshock Study: Parameter Estimates (standard errors (model based; empirically corrected)) for Logistic Regression, Two Different GEE1 Procedures and the Generalized Linear Mixed Model (using GLIMMIX Macro).*

Outcome	Par.	Model			
		LOGISTIC	GEE1 (standard)	GEE1 (Prentice)	GLIMMIX (CSYM)
MBN	β_0	-1.80 (0.20)	-1.81 (0.20;0.23)	-1.82 (0.21;0.22)	-1.82 (0.21;0.22)
	β_{t^*}	-3.88 (1.58)	-3.92 (1.60;1.95)	-3.98 (1.66;1.98)	-3.97 (1.66;1.97)
	β_{dt}	5.85 (1.60)	5.91 (1.62;1.95)	5.99 (1.69;1.97)	5.99 (1.69;1.97)
	ρ	0	.02	.05	.05
OPT	β_0	-2.48 (0.25)	-2.47 (0.24;0.22)	-2.47 (0.24;0.22)	-0.47 (0.24;0.22)
	β_{t^*}	-3.68 (1.77)	-3.77 (1.69;2.26)	-3.75 (1.71;2.26)	-3.73 (1.67;2.26)
	β_{dt}	5.61 (1.73)	5.68 (1.65;2.30)	5.66 (1.67;2.30)	5.65 (1.66;2.30)
	ρ	0	-.03	-.03	-.02
OLF	β_0	-1.44 (0.18)	-1.54 (0.22;0.22)	-1.57 (0.23;0.21)	-1.56 (0.22;0.22)
	β_{t^*}	-5.69 (1.65)	-4.85(1.74;2.11)	-4.50 (1.73;2.07)	-4.71 (1.74;2.09)
	β_{dt}	7.19 (1.66)	6.65 (1.77;2.15)	6.40 (1.77;2.12)	6.55 (1.77;2.13)
	ρ	0	.15	.24	.19

From Table 7.4 a clear distinction between the logistic and correlated models can be deduced. For outcome MBN we find that the logistic standard error is smaller than the model-based standard errors of any of the other three procedures, which are smaller than any of the empirically corrected standard errors. More complex association designs which were previously fitted using the second order generalized estimating equations (e.g. Table 7.2, design 2) do not reduce the difference beween model-based and empirically corrected standard errors. This may be due to the fact that the association parameters are not significant for MBN. Fitting more complex designs will therefore not be very helpful. In contrast, for the OPT outcome, model based and empirically corrected standard errors tend to lie closer to each other for the GEE2 estimates (e.g. Table 7.3, design 2) than for the exchangeably correlated PA procedures in Table 7.4. Indeed, the association parameter γ results in a significant amelioration of the association model. Hence, it might be important to consider more complicated designs for the association structure. Further, one observes that the correlation parameter is always estimated negative (but not significantly). For OLF, the model based standard error for the logistic procedures are considerably smaller than the model based standard errors for the correlated procedures. Furthermore, the discrepancy between model based and empirically corrected standard errors for the correlated procedures is rather high. Unfortunately, GEE2 was hard to fit for complex association designs, such as designs 2 and 3.

In conclusion, although GEE1 is much easier to fit than GEE2, it presents more difficulties when coping with complex association designs. In the discussion of Fitzmaurice, Laird and Rotnitzky (1993), Drum and McCullagh note that "ideally, one should calculate both model based and empirically corrected standard errors and aim to understand any differences that occur". Whenever model based and empirically corrected standard error estimates are similar, GEE1 is trustworthy and might be helpful in finding crude effects like a dominant compound symmetry component. Even GEE2 or full likelihood methods may sometimes be unsatisfactory for addressing complex association patterns. Whenever more complex designs are needed where all three components (genetic, serial and random covariates) are important, generalized linear mixed models would be necessary.

7.5.2 Cluster-specific Approaches

In this section we used the SAS/STAT GLIMMIX macro (Littell et al. 1996) to fit mixed effects logistic models such as (7.3) to the binary outcomes MBN, OPT and OLF. The responses from the ith cluster are now correlated by virtue of their sharing a common intercept. For continuous outcomes a random intercept model and a compound symmetry model yield equivalent results. In contrast, parameter estimates obtained from a compound symmetry model (PA) for discrete outcomes tend to be smaller than those obtained from a random intercept model (CS).

Table 7.5: *Heatshock Study: Parameter Estimates (standard errors; p–values) for the Mixed Effects Logistic (MIXLOG), Compound Symmetry (CSYM) and Conditional Logistic (CONDLOG) models.*

Outcome	Par.	Model		
		MIXLOG	CSYM	CONDLOG
MBN	β_0	-1.87 (0.21;0.00)	-1.82 (0.21;0.00)	
	β_{t^*}	-4.08 (1.65;0.01)	-3.97 (1.66;0.02)	-4.64 (2.55;0.07)
	β_{dt}	6.16 (1.67;0.00)	5.99 (1.69;0.00)	6.84 (2.63;0.01)
OPT	β_0	-2.48 (0.25;0.00)	-2.47 (0.24;0.00)	
	β_{t^*}	-3.67 (1.75;0.04)	-3.73 (1.67;0.03)	-1.46 (3.04;0.63)
	β_{dt}	5.60 (1.71;0.00)	5.65 (1.66;0.00)	3.96 (3.01;0.19)
OLF	β_0	-1.84 (0.24;0.00)	-1.56 (0.22;0.00)	
	β_{t^*}	-4.99 (1.90;0.01)	-4.71 (1.74;0.01)	-3.40 (2.96;0.25)
	β_{dt}	7.25 (1.94;0.00)	6.55 (1.77;0.00)	6.30 (3.04;0.04)

Table 7.5 shows the parameter estimates, with p–values, for the mixed–effects logistic model (MIXLOG) and the compound symmetry model (CSYM). Fixed effects can be easily tested using the ratio of the parameter estimate of interest over its standard error. However, in finite samples, this statistic is only approximately t-distributed, and the appropriate number of degrees of freedom needs to be estimated from the data. The SAS procedure MIXED (that is called recursively within the GLIMMIX macro) provides several methods for estimating the appropriate number

of degrees of freedom, including a Satterthwaite (1946) approximation.

The observed "shrinkage effect" is, in most cases, in agreement with (7.4) and with findings of Neuhaus and Jewell (1993), who state that PA parameters will be closer to zero than CS parameters for convex link functions. One exception is formed by the OPT outcome, for which the correlation parameter was estimated negative.

For all outcomes there is evidence of a significant effect of the cumulative exposure (dt) and a significant effect of duration of exposure at temperatures above normal body temperature (t^*). Furthermore, the parameter estimate for t^* is again negative, which is in agreement with earlier results. One could also try to include a random dt parameter in addition to a random intercept. However, for MBN, this parameter estimate was very small (.00005). For the other outcomes convergence problems occured when fitting this model. This covariance structure might be too complex and requires substantial binary data per cluster to provide accurate information. Therefore, we restricted attention to the mixed-effects logistic model.

Table 7.5 also shows the conditional logistic regression (CONDLOG) parameter estimates. All cluster-level effects are conditioned out. Therefore we cannot obtain parameter estimates for the intercepts. Where we found strong significant effects for dt and t^* by the MIXLOG and CSYM approaches, we now observe a severe reduction in statistical significance of the CONDLOG estimates. This is in agreement with the results of Neuhaus and Lesperance (1996) summarized in Section 7.3.2. The cumulative exposure and duration of exposure at "positive increases" of temperature are highly correlated (correlation coefficient=97%) and moreover the cluster sizes in the heatshock study are relatively small (mean cluster size is 5).

7.6 Conclusion

We have described population-averaged and cluster-specific models for the analysis of developmental toxicity studies, in which individual-level covariates play an important role.

Within the class of population-averaged models, we have illustrated that conditionally specified models, such as that described in Section 7.2.1, should be avoided since they lead to undesirable properties. Alternatively, one can use marginal models. The likelihood based method proposed by Bahadur (1961), readily extends to individual-level covariates. But the correlation parameters may be subject to

severe restrictions. Likelihood-based odds ratio models (Dale 1986; Molenberghs and Lesaffre 1994; Lang and Agresti 1994; Glonek and McCullagh 1995) exhibit less constraints, but they are more involved to fit when cluster sizes are moderate to large. Therefore, generalized estimating equations are a very viable alternative marginal approach. In addition to the classical first order GEE, second order estimating equations can be considered as well when the association structure is also of scientific interest. GEE1 models are relatively straightforward to fit even with large clusters, but the second order version appeared to be more cumbersome. The lack of a likelihood base is generally viewed as a disadvantage of GEE, since it prevents standard calculation of joint and union probabilities, even though such calculations are often needed in risk assessment. Further, a likelihood ratio test statistic cannot be computed, even though some approximations have been proposed (Rotnitzky and Jewell 1990).

Random-effects models, on the contrary, correspond to a likelihood function. Unfortunately, this function is difficult to evaluate for many realistic applications. Therefore, a variety of approximations, as reviewed by McCulloch (1997) have been proposed. As already stated in previous sections, the best known approximations are based on Laplace transforms (Breslow and Clayton 1993) and on pseudo-likelihood (Wolfinger and O'Connell 1993) already stated in previous sections. Caution should be used with the GLIMMIX macro since, e.g., the deviance such as reported in the output and also mentioned in Lee and Nelder (1996) requires additional work. Arguably, a more appropriate form needs to be proposed. In addition, Neuhaus and Segal (1997) warn that such approximate methods may require modification before their asymptotic bias can be competitive with sampling based methods (e.g., the Gibbs sampler).

An important difference between the categorical data setting considered here and clustered normally distributed data is that, unlike in the latter case, the interpretation of parameters depends crucially on the choice between marginal and random-effects models (Neuhaus, Kalbfleisch and Hauck 1991). A marginal model describes the average evolution within a certain subpopulation, whereas the fixed effects in a mixed model describe the evolution, conditional on values for the random effects. This implies that one should be guided not only by computational convenience, but in particular by the nature of the scientific question. In some cases, however, it is reasonable to consider both approaches as equally valid.

The conditional likelihood approach described in Section 7.3.2 lacks efficiency with small cluster sizes and large within-cluster correlations. Therefore we cannot recommend it.

Chapter 8

GEE and PL Risk Assessment Approaches for Combined Continuous and Discrete Outcomes from Developmental Toxicity Studies

8.1 Introduction

Measurements of both continuous and discrete outcomes are encountered in many statistical problems. However, methods that jointly analyze discrete and continuous outcomes and adequately account for the correlation structure in the data are not widely available and remain a topic of statistical research. In this chapter we consider the particular context of teratology studies, where quantitative risk assessment is aimed at determining the effect of dose on the probability that a live fetus is malformed (binary) or of low birth weight (continuous), both being important measures of teratogenicity. Although 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, we have previously (see Introduction) promoted the use of joint models that:

- allow separate dose-response functions for each component of the bivariate

outcome,

- account for the association due to clustering within litters,
- estimate the bivariate intra-fetus association.

Here, we describe two modelling approaches that satisfy these criteria and apply them on data from a developmental toxicity study (see also Geys et al. 1999b).

Regan and Catalano (1999a) introduced a probit approach, based on the Ochi and Prentice (1984) method. They assume an underlying continuous variable for each binary outcome. Hence, the joint distribution of the vector of weight and latent malformation outcomes can be assumed to follow a multivariate normal distribution. Their full likelihood approach can easily be used for quantitative risk assessment and model checking. It provides marginal dose-response models for each outcome and allows for estimation of the joint risk to a fetus due to malformation and low birth weight. A difficulty, however, is the computational intractability of their full likelihood. In Section 8.2.1 we show how this problem can be avoided by adopting GEE methodology (Regan and Catalano 1999b). Since in quantitative risk assessment clustering is a nuisance, Regan and Catalano (1999b) argue that one can avoid fully specifying the distribution within a litter by specifying only the marginal distribution of the bivariate outcome and using GEE ideas to account for correlations due to clustering. Their GEEs are derived from the marginal distribution of the bivariate outcome, which defines the mean and association parameters that are of interest for the application to quantitative risk assessment. Those association parameters of interest are estimated via second order GEEs. The nuisance parameters that account for clustering are estimated via the method of moments as in first order GEE.

In Section 8.2.2, we describe the Plackett-Dale approach that has been used by Molenberghs, Geys and Buyse (1999) and Molenberghs and Geys (1998) to model independent bivariate endpoints in which one component is continuous and the other is binary (see also Chapter 9). It provides an alternative to frequently used multivariate normal latent variables. Main advantages are the flexibility with which the marginal densities can be chosen (normal, logistic, complementary log-log, etc.) and the familiarity of the odds ratio which is used as a measure of association, providing an alternative to correlation. Geys et al. (1999b) extend this method to allow for within-cluster association as well, while specification of the full likelihood is avoided by using pseudo-likelihood methodology.

Thus, in this chapter we concentrate on models that satisfy the three criteria listed above, while specification of the full distribution is avoided using GEE and PL methodologies. Section 8.3 describes tools for quantitative risk assessment with these models. Section 8.4 applies the methods to data, introduced in Section 2.1.6, that investigate the toxicity of ethylene glycol in rats.

8.2 Models for Bivariate Data of a Mixed Nature

Let us formalize the setting of this chapter. Consider, as before, an experiment involving N clusters, the ith of which contains n_i individuals, each of whom are examined for the presence ($M_{ik} = 1$) or absence ($M_{ik} = 0$) of a certain malformation indicator (M_{ik}) and measured for fetal weight (W_{ik}) ($k = 1, \ldots, n_i$). The binary outcome is assumed to arise from an unobservable continuous random variable, denoted M^*_{ik}; M_{ik} represents an indicator of whether this underlying variable exceeds some threshold, arbitrarily assumed to be 0. In the remainder of this chapter we assume a normal distribution for W_{ik} with mean μ_{ik} and variance σ^2_{ik}. The malformation probability for the kth individual in the ith cluster will be denoted by π_{ik}.

8.2.1 Probit Model

Independence

To derive the marginal distribution of the bivariate response (W_{ik}, M_{ik}), first assume that littermates are independent. Under a probit model for the binary response M_{ik}, the latent variable M^*_{ik} is assumed normally distributed with mean γ_{ik} and unit variance so that

$$\begin{aligned} \pi_{ik} &= \Pr(M_{ik} = 1) \\ &= \Pr(M^*_{ik} > 0) \\ &= \Phi(\gamma_{ik}), \end{aligned}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. The probability of malformation is related to covariates by expressing γ_{ik} as some linear combination of predictors. For the bivariate response, we assume the observed fetal weight and unobserved continuous malformation variables for fetus k in litter i to share a bivariate

normal distribution,

$$\boldsymbol{\phi}_2(W_{ik}, M^*_{ik};\, \mu_{ik}, \gamma_{ik}, \sigma_{ik}, 1, \rho_{ik}) \;=\; (2\pi\sigma_{ik})^{-1}(1-\rho^2_{ik})^{-\frac{1}{2}} \times \tag{8.1}$$

$$\exp\left\{ \frac{-1}{2(1-\rho^2_{ik})}\left[\left(\frac{W_{ik}-\mu_{ik}}{\sigma_{ik}}\right)^2 - 2\rho_{ik}\left(\frac{W_{ik}-\mu_{ik}}{\sigma_{ik}}\right)(M^*_{ik}-\gamma_{ik}) + (M^*_{ik}-\gamma_{ik})^2\right]\right\}.$$

To arrive at a convenient form of the bivariate distribution of the mixed outcomes, this density (8.1) is rewritten as a product of the marginal density for fetal weight and conditional density of latent malformation given weight, so the joint distribution of the bivariate fetal weight and binary malformation outcome for fetus k in litter i can be written

$$f(W_{ik}, M_{ik}) \;=\; \boldsymbol{\phi}(W_{ik};\, \mu_{ik}, \sigma^2_{ik})\;\; \Phi(\gamma_{m|w_{ik}})^{M_{ik}}\;\; [1-\Phi(\gamma_{m|w_{ik}})]^{1-M_{ik}}, \tag{8.2}$$

where $\boldsymbol{\phi}(\cdot)$ denotes the univariate standard normal density and from bivariate normal theory,

$$\gamma_{m|w_{ik}} = \frac{\gamma_{ik} + \rho_{ik}\,\frac{W_{ik}-\mu_{ik}}{\sigma_{ik}}}{(1-\rho^2_{ik})^{1/2}}.$$

The probit $\Phi(\gamma_{m|w_{ik}})$ represents the mean of the conditional binary malformation outcome $\mathrm{E}(M_{ik}\,|\,W_{ik})$, and the marginal expectation of M_{ik} is $\pi_{ik} = \Phi(\gamma_{ik})$.

Dose–response models are specified for all four parameters of the bivariate normal density to allow the fetal weight mean and variance, the probability of malformation, and the bivariate correlation to vary as functions of dose and other covariates. In the light of restrictions on the respective parameter spaces, we use an exponential link function for the fetal weight variance and the inverse of Fisher's Z–transformation as a link function for the correlation. The dose–response models can be written generally as

$$\begin{aligned} \gamma_{ik} &= \boldsymbol{X}'_{b_{ik}}\boldsymbol{\beta}, & \rho_{ik} &= \{\exp(\boldsymbol{X}'_{t_{ik}}\boldsymbol{\tau}) - 1\}/\{\exp(\boldsymbol{X}'_{t_{ik}}\boldsymbol{\tau}) + 1\}, \\ \mu_{ik} &= \boldsymbol{X}'_{a_{ik}}\boldsymbol{\alpha}, & \sigma^2_{ik} &= \exp(\boldsymbol{X}'_{s_{ik}}\boldsymbol{\varsigma}), \end{aligned} \tag{8.3}$$

where the ikth fetus has $\{b\times 1, t\times 1, a\times 1, s\times 1\}$ vectors $\{\boldsymbol{X}_{b_{ik}}, \boldsymbol{X}_{t_{ik}}, \boldsymbol{X}_{a_{ik}}, \boldsymbol{X}_{s_{ik}}\}$ of covariates that may be both fetus– and litter–specific. These vectors correspond to $\{b\times 1, t\times 1, a\times 1, s\times 1\}$ vectors of fixed regression parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\varsigma}\}$.

From the log–likelihood based on the bivariate distribution (8.2), score equations for the regression parameters $\boldsymbol{\theta}$ can be written,

$$\sum_{i=1}^{N}\sum_{k=1}^{n_i}\left(\frac{\partial \ell_{ik}}{\partial \boldsymbol{\beta}'} \quad \frac{\partial \ell_{ik}}{\partial \boldsymbol{\tau}'} \quad \frac{\partial \ell_{ik}}{\partial \boldsymbol{\alpha}'} \quad \frac{\partial \ell_{ik}}{\partial \boldsymbol{\varsigma}'}\right)'$$

$$=\sum_{i=1}^{N}\sum_{k=1}^{n_i}\begin{pmatrix} \Delta_{b_{ik}}\boldsymbol{X}_{b_{ik}} & \mathbf{0} & \mathbf{0} \\ \Delta_{t_{ik}}\boldsymbol{X}_{t_{ik}} & \mathbf{0} & \mathbf{0} \\ \Delta_{a_{ik}}\boldsymbol{X}_{a_{ik}} & \boldsymbol{X}_{a_{ik}} & \mathbf{0} \\ \Delta_{s_{ik}}\boldsymbol{X}_{s_{ik}} & \mathbf{0} & \sigma_{ik}^2\boldsymbol{X}_{s_{ik}} \end{pmatrix}\begin{pmatrix} \left(\Phi(\gamma_{m|w_{ik}})[1-\Phi(\gamma_{m|w_{ik}})]\right)^{-1} & 0 & 0 \\ 0 & \sigma_{ik}^{-2} & 0 \\ 0 & 0 & \frac{1}{2}\sigma_{ik}^{-4} \end{pmatrix}$$

$$\times\begin{pmatrix} M_{ik}-\Phi(\gamma_{m|w_{ik}}) \\ W_{ik}-\mu_{ik} \\ S_{ik}-\sigma_{ik}^2 \end{pmatrix} \tag{8.4}$$

where

$$\begin{aligned} S_{ik} &= (W_{ik}-\mu_{ik})^2, \\ \Delta_{b_{ik}} &= \partial\Phi(\gamma_{m|w_{ik}})/\partial\gamma_{ik}, \\ \Delta_{t_{ik}} &= \partial\Phi(\gamma_{m|w_{ik}})/\partial\rho_{ik}, \\ \Delta_{a_{ik}} &= \partial\Phi(\gamma_{m|w_{ik}})/\partial\mu_{ik}, \\ \Delta_{s_{ik}} &= \partial\Phi(\gamma_{m|w_{ik}})/\partial\sigma_{ik}^2. \end{aligned}$$

Clustering

In the case of clustering, we avoid fully specifying the joint distribution of the n_i bivariate outcomes in litter i by using the score equations (8.4) of the bivariate distribution derived under independence to motivate a set of generalized estimating equations (GEEs) for the clustered setting; the GEE methodology of Liang and Zeger (1986) and Zeger and Liang (1986) is the basis of estimation. Now assume that each of N independent litters has measurements on n_i bivariate response vectors (W_{ik}, M_{ik}), $(i = 1, \ldots, N;\ k = 1, \ldots, n_i)$; thus the responses within a litter are no longer assumed independent. For litter i, $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{in_i})'$ and

$\boldsymbol{M}_i = (M_{i1}, \ldots, M_{in_i})'$ will denote the $n_i \times 1$ vectors of fetal weights and malformation outcomes. Let $\{\boldsymbol{X}_{b_i}, \boldsymbol{X}_{t_i}, \boldsymbol{X}_{a_i}, \boldsymbol{X}_{s_i}\}$ represent the $\{n_i \times b, n_i \times t, n_i \times a, n_i \times s\}$ matrices of covariates for the ith litter whose rows are defined by the covariate vectors in (8.3) above.

First note that the score equations (8.4) can be rewritten at the level of the litter as:

$$\sum_{i=1}^{N}\sum_{k=1}^{n_i}\left(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}'} \quad \frac{\partial \ell_i}{\partial \boldsymbol{\tau}'} \quad \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}'} \quad \frac{\partial \ell_i}{\partial \boldsymbol{\varsigma}'}\right)' \tag{8.5}$$

$$= \sum_{i=1}^{N}\begin{pmatrix} \Delta_{b_i}\boldsymbol{X}_{b_i} & \mathbf{0} & \mathbf{0} \\ \Delta_{t_i}\boldsymbol{X}_{t_i} & \mathbf{0} & \mathbf{0} \\ \Delta_{a_i}\boldsymbol{X}_{a_i} & \boldsymbol{X}_{a_i} & \mathbf{0} \\ \Delta_{s_i}\boldsymbol{X}_{s_i} & \mathbf{0} & \boldsymbol{\Sigma}_{w_i}\boldsymbol{X}_{s_i} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{m_i} & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{w_i} & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_{s_i} \end{pmatrix}^{-1}\begin{pmatrix} \boldsymbol{M}_i - \boldsymbol{\Phi}(\boldsymbol{\gamma}_{m|w_i}) \\ \boldsymbol{W}_i - \boldsymbol{\mu}_i \\ \boldsymbol{S}_i - \boldsymbol{\sigma}_i^2 \end{pmatrix},$$

where $\Delta_{b_i}, \Delta_{t_i}, \Delta_{a_i}, \Delta_{s_i}$ are $n_i \times n_i$ diagonal matrices with elements $\Delta_{b_{ik}}, \Delta_{t_{ik}}, \Delta_{a_{ik}}, \Delta_{s_{ik}}$, respectively, and $\boldsymbol{\Phi}(\boldsymbol{\gamma}_{m|w_i}) = \big(\Phi(\gamma_{m|w_{i1}}), \ldots, \Phi(\gamma_{m|w_{in_i}})\big)'$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})'$, and $\boldsymbol{\sigma}_i^2 = (\sigma_{i1}^2, \ldots, \sigma_{in_i}^2)'$. The three $n_i \times n_i$ diagonal covariance matrices are $\boldsymbol{\Sigma}_{m_i} = \text{diag}\{\Phi(\gamma_{m|w_{ik}})\,[1 - \Phi(\gamma_{m|w_{ik}})]\}$, $\boldsymbol{\Sigma}_{w_i} = \text{diag}\{\sigma_{ik}^2\}$ and $\boldsymbol{\Sigma}_{s_i} = \text{diag}\{2\,\sigma_{ik}^4\}$; $\boldsymbol{S}_i = (S_{i1}, \ldots, S_{in_i})'$ is the vector of squared weight residuals. We assume the marginal distribution of the bivariate outcome is defined by (8.2) and use the form of the score equations (8.5) to construct a set of GEEs for the regression parameters by replacing the block–diagonal covariance matrix $\{\boldsymbol{\Sigma}_{m_i}, \boldsymbol{\Sigma}_{w_i}, \boldsymbol{\Sigma}_{s_i}\}$ by a working covariance matrix that incorporates correlation between littermates. Thus the regression parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\varsigma}\}$ are estimated in the clustered setting from the following set of GEEs:

$$\sum_{i=1}^{N}\begin{pmatrix} \Delta_{b_i}\boldsymbol{X}_{b_i} & \mathbf{0} & \mathbf{0} \\ \Delta_{t_i}\boldsymbol{X}_{t_i} & \mathbf{0} & \mathbf{0} \\ \Delta_{a_i}\boldsymbol{X}_{a_i} & \boldsymbol{X}_{a_i} & \mathbf{0} \\ \Delta_{s_i}\boldsymbol{X}_{s_i} & \mathbf{0} & \boldsymbol{\Sigma}_{w_i}\boldsymbol{X}_{s_i} \end{pmatrix}\begin{pmatrix} \boldsymbol{V}_{m_i} & \boldsymbol{V}_{wm_i} & 0 \\ \boldsymbol{V}_{wm_i} & \boldsymbol{V}_{w_i} & 0 \\ 0 & 0 & \boldsymbol{V}_{s_i} \end{pmatrix}^{-1}\begin{pmatrix} \boldsymbol{M}_i - \boldsymbol{\Phi}(\boldsymbol{\gamma}_{m|w_i}) \\ \boldsymbol{W}_i - \boldsymbol{\mu}_i \\ \boldsymbol{S}_i - \boldsymbol{\sigma}_i^2 \end{pmatrix} = 0, \tag{8.6}$$

where $\boldsymbol{V}_{m_i}, \boldsymbol{V}_{w_i}, \boldsymbol{V}_{s_i}$ and $\boldsymbol{V}_{wm_i}$ are equicorrelated working covariance matrices. For

developmental toxicity data, the assumption that littermates are exchangeable is reasonable and generally implemented. Note, when accounting for the correlation between the weight and malformation outcomes among littermates, $\boldsymbol{V}_{wm_i}$ has zeros on the diagonal because of the conditional independence of the outcomes of an individual fetus in the joint distribution (8.2); the correlation within a fetus is still characterized by ρ_{ik}. These equations follow the familiar form

$$\sum_{i=1}^{N} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})\right) = 0$$

with an obvious correspondence to the three major matrices in (8.6), to which the approach of Liang and Zeger (1986) and Zeger and Liang (1986) is implemented for estimation.

8.2.2 Plackett-Dale Model

The Plackett-Dale idea has been used by Molenberghs, Geys and Buyse (1999) to assess the validation of surrogate endpoints in randomized experiments with a binary surrogate endpoint and a continuous true endpoint or vice versa (see Chapter 9). In their work, the bivariate outcomes on different subjects are independent. In contrast, teratology experiments in rodents entail clustering between littermates, which has to be incorporated (Geys et al. 1999b).

Independence

First, suppose that all littermates are independent. Assume that the cumulative distributions of the weight (W_{ik}) and malformation (M_{ik}) outcome of the kth individual in the ith cluster are given by $F_{W_{ik}}$ and $F_{M_{ik}}$. Their dependence can be defined using a global cross-ratio at cutpoint (w, m) $(m = 0, 1)$:

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

This expression can be solved for the joint cumulative distribution $F_{W_{ik},M_{ik}}$ (Plackett 1965):

$$F_{W_{ik},M_{ik}}(w,m) = \begin{cases} \frac{1+(F_{W_{ik}}(w)+F_{M_{ik}}(m))(\psi_{ik}-1)-S(F_{W_{ik}}(w),F_{M_{ik}}(m),\psi_{ik})}{2(\psi_{ik}-1)} & \text{if } \psi_{ik} \neq 1, \\ F_{W_{ik}}(w)F_{M_{ik}}(m) & \text{if } \psi_{ik} = 1, \end{cases}$$

where

$$S(F_{W_{ik}}, F_{M_{ik}}, \psi_{ik})$$
$$= \sqrt{[1 + (\psi_{ik} - 1)(F_{W_{ik}}(w) + F_{M_{ik}}(m))]^2 + 4\psi_{ik}(1 - \psi_{ik})F_{W_{ik}}(w)F_{M_{ik}}(m)}$$

Based upon this distribution function, we can derive a bivariate Plackett *density* function $g_{ik}(w, m)$ for mixed continuous-binary outcomes. With the success probability for M_{ik} denoted by π_{ik} and the density function for W_{ik} by $f_{W_{ik}}$, we define $g_{ik}(w, m)$ by specifying $g_{ik}(w, 0)$ and $g_{ik}(w, 1)$ such that they sum to $f_{W_{ik}}(w)$. If we define $g_{ik}(w, 0) = \partial F_{W_{ik},M_{ik}}(w, 0)/\partial w$, then this leads to specifying g_{ik} by:

$$g_{ik}(w, 0) = \begin{cases} \frac{f_{W_{ik}}(w)}{2}\left[1 - \frac{1+F_{W_{ik}}(w)(\psi_{ik}-1)-(1-\pi_{ik})(\psi_{ik}+1)}{S\left(F_{W_{ik}},1-\pi_{ik},\psi_{ik}\right)}\right] & \text{if } \psi_{ik} \neq 1, \\ f_{W_{ik}}(w)(1 - \pi_{ik}) & \text{if } \psi_{ik} = 1, \end{cases} \quad (8.7)$$

and

$$g_{ik}(w, 1) = f_{W_{ik}}(w) - g_{ik}(w, 0).$$

Note how $g_{ik}(w, m)$ in (8.7) satisfies the classical density properties:

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

(ii) $\int \{g_{ik}(w, 0) + g_{ik}(w, 1)\}\, dw = \int f_{W_{ik}}(w)dw = 1$

Further, $g_{ik}(w, 0)$ naturally factorizes as a product of the marginal density $f_{W_{ik}}(w)$ and the conditional density $f_{M_{ik}|W_{ik}}(0|w)$ (and similarly for $g_{ik}(w, 1)$). Some interesting special cases are obtained by putting $\psi_{ik} = 1$ (independence), $\psi_{ik} = 0$ (perfect negative association) and $\psi_{ik} = \infty$ (perfect positive association).

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

$$\begin{aligned} g_{ik}(w, 0) &= f_{W_{ik}}(w)(1 - \pi_{ik}), \\ g_{ik}(w, 1) &= f_{W_{ik}}(w)\pi_{ik}. \end{aligned}$$

2. Suppose weight and malformation are perfectly negatively correlated ($\psi_{ik} = 0$),

then the function $S(F_{W_{ik}}, 1-\pi_{ik}, 0)$ reduces to $|\pi_{ik} - F_{W_{ik}}(w)|$ and as a result:

$$\begin{aligned} g_{ik}(w,0) &= \begin{cases} 0 & \text{if } \pi_{ik} - F_{W_{ik}}(w) > 0 \\ f_{W_{ik}}(w) & \text{if } \pi_{ik} - F_{W_{ik}}(w) \leq 0 \end{cases} \\ g_{ik}(w,1) &= \begin{cases} f_{W_{ik}}(w) & \text{if } \pi_{ik} - F_{w_{ik}}(w) > 0 \\ 0 & \text{if } \pi_{ik} - F_{W_{ik}}(w) \leq 0 \end{cases} \end{aligned}$$

3. If weight and malformation are perfectly negatively correlated ($\psi_{ik} = \infty$), then one can define $\psi^* = 1/\psi$ whence $S(F_{W_{ik}}, 1-\pi_{ik}, \psi_{ik})$ can be rewritten as:

$$\begin{aligned} & S(F_{W_{ik}}, 1-\pi_{ik}, \psi_{ik}) \\ = & \sqrt{(\psi^*_{ik} + (1-\psi^*_{ik})(F_{W_{ik}}(w) + 1 - \pi_{ik}))^2 + 4(\psi^*_{ik} - 1)F_{W_{ik}}(w)(1-\pi_{ik})}/\psi^*_{ik} \\ = & S^*(F_{W_{ik}}, 1-\pi_{ik}, \psi^*_{ik})/\psi^*_{ik}. \end{aligned}$$

As a result we can now calculate:

$$\begin{aligned} g_{ik}(w,0) &= \lim_{\psi^*_{ik}\to 0} \frac{f_{W_{ik}}(w)}{2}\left(1 - \frac{\psi^*_{ik} + F_{W_{ik}}(w)(1-\psi^*_{ik}) - (1-\pi_{ik})(1+\psi^*_{ik})}{S^*(F_{W_{ik}}, 1-\pi_i k, \psi^*_{ik})}\right) \\ &= \frac{f_{W_{ik}}(w)}{2}\left(1 - \frac{F_{W_{ik}}(w) - (1-\pi_{ik})}{|F_{W_{ik}}(w) - (1-\pi_{ik})|}\right) \\ &= \begin{cases} 0 & F_{W_{ik}}(w) - (1-\pi_{ik}) > 0 \\ f_{W_{ik}}(w) & F_{W_{ik}}(w) - (1-\pi_{ik}) \leq 0 \end{cases} \end{aligned}$$

and

$$g_{ik}(w,1) = \begin{cases} 0 & F_{W_{ik}}(w) - (1-\pi_{ik}) > 0 \\ f_{W_{ik}}(w) & F_{W_{ik}}(w) - (1-\pi_{ik}) \leq 0 \end{cases}.$$

Clustering

In the case of clustering, rather than considering the full likelihood contribution for cluster i, i.e. $f(w_{i1}, \ldots, w_{in_i}, m_{i1}, \ldots, m_{in_i})$, we avoid computational complexity by replacing the full likelihood by a pseudo-likelihood function that is easier to evaluate.

We define the following log pseudo-likelihood function:

$$p\ell = \sum_{i=1}^{N}\sum_{k=1}^{n_i} \ln g_{ik}(w_{ik}, m_{ik}). \tag{8.8}$$

The contribution $p\ell_i$ for the ith cluster equals $\sum_{k=1}^{n_i} \ln g_{ik}(w_{ik}, m_{ik})$. With this approach, weight and malformation outcomes for a given littermate are allowed to be correlated, but for outcomes from different littermates independence is taken as a working assumption. This leads to consistent estimates (Arnold and Strauss 1991; Geys, Molenberghs and Lipsitz 1998; Le Cessie and Van Houwelingen 1994). We then correct for potential bias in the variance estimator by using a sandwich estimator, formulated in (8.10). As was the case in Section 8.2.1, this approach acknowledges the fact that, while the association between different outcomes on the same littermate is often of scientific interest, the clustering (association between different littermates) is usually considered a nuisance. Indeed, in quantitative risk assessment primary interest lies in the probability that an individual is affected, either by malformation or by low birth weight. This probability is a function only of the mean parameters and the association of the bivariate outcome. Nevertheless, if one is interested in the amount of clustering as well, pseudo-likelihood (8.8) 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.

Let us group all parameters μ_{ik}, σ^2_{ik}, π_{ik} and ψ_{ik} for individual $k = 1, \ldots, n_i$ in cluster $i = 1, \ldots, N$ in a vector $\boldsymbol{\theta}_{ik}$. Linear dose-response models can be considered on each of the parameters in θ_{ik} by using appropriate link functions:

$$\boldsymbol{\eta}_{ik} = \begin{pmatrix} \mu_{ik} \\ \ln(\sigma^2_{ik}) \\ \text{logit}(\pi_{ik}) \\ \ln(\psi_{ik}) \end{pmatrix} = \boldsymbol{X}_{ik}\boldsymbol{\beta}, \tag{8.9}$$

where $\boldsymbol{X}_{ik}$ is a design matrix for the kth fetus in the ith cluster and $\boldsymbol{\beta}$ is a vector of regression parameters. The generality of (8.9) is an important advantage. For example, for developmental toxicity data, a constant variance assumption is often not tenable.

Grouping all vectors $\boldsymbol{\theta}_{ik}$ and $\boldsymbol{\eta}_{ik}$ for the ith cluster in $\boldsymbol{\theta}_i$ and $\boldsymbol{\eta}_i$ respectively, estimates are obtained by solving the estimating equations $\boldsymbol{U}(\boldsymbol{\beta}) = 0$, where $\boldsymbol{U}(\boldsymbol{\beta})$

can be written as:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\theta}_i}\right)^{-T} \left(\frac{\partial \ln g_{ik}(w,m)}{\partial \boldsymbol{\theta}_i}\right).$$

Expressions for the derivatives are given below (occasionally dropping indices for ease of notation). The derivatives of $f_W(w)$ and $F_W(w)$ are combined with those of $S(F_W, 1-\pi, \psi)$ to obtain the first derivatives of $\ln g_{ik}$ with respect to the regression parameters:

$$\begin{aligned}
\frac{\partial f_W(w)}{\partial \mu} &= f_W(w)\frac{(w-\mu)}{\sigma^2} \\
\frac{\partial f_W(w)}{\partial \sigma^2} &= \frac{f_W(w)}{2}\left(\frac{(w-\mu)^2 - \sigma 2}{\sigma^4}\right) \\
\frac{\partial F_W(w)}{\partial \mu} &= -f_W(w) \\
\frac{\partial F_W(w)}{\partial \sigma^2} &= -\left(\frac{w-\mu}{\sigma^2}\right)\frac{f_W(w)}{2} \\
\frac{\partial S}{\partial \mu} &= \frac{2\left[1+(\psi-1)(F_W(w)+1-\pi)\right](\psi-1)\frac{\partial F_W(w)}{\partial \mu}}{2S} \\
&\quad + \frac{4\psi(1-\psi)\frac{\partial F_W(w)}{\partial \mu}(1-\pi)}{2S} \\
\frac{\partial S}{\partial \sigma^2} &= \frac{2\left[1+(\psi-1)(F_W(w)+1-\pi)\right](\psi-1)\frac{\partial F_W(w)}{\partial \sigma^2}}{2S} \\
&\quad + \frac{4\psi(1-\psi)\frac{\partial F_W(w)}{\partial \sigma^2}(1-\pi)}{2S} \\
\frac{\partial S}{\partial \pi} &= \frac{2\left[1+(\psi-1)(F_W(w)+1-\pi)\right](1-\psi) - 4\psi(1-\psi)F_W(w)}{2S} \\
\frac{\partial S}{\partial \psi} &= \frac{2\left[1+(\psi-1)(F_W(w)+1-\pi)\right](F_W(w)+1-\pi)}{2S} \\
&\quad + \frac{4(1-\psi)F_W(w)(1-\pi) - 4\psi F_W(w)(1-\pi)}{2S}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial g}{\partial \mu}(w,0) &= \frac{1}{2}\frac{\partial f_W(w)}{\partial \mu}\left[1-\frac{1+F_W(w)(\psi-1)-(1-\pi)(\psi+1)}{S}\right] \\
&\quad -\frac{\partial F_W(w)}{\partial \mu}\frac{f_W(w)(\psi-1)}{2S} \\
&\quad +\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \mu}\left[1+F_W(w)(\psi-1)-(1-\pi)(1+\psi)\right] \\
\frac{\partial g}{\partial \sigma^2}(w,0) &= \frac{1}{2}\frac{\partial f_W(w)}{\partial \sigma^2}\left[1-\frac{1+F_W(w)(\psi-1)-(1-\pi)(\psi+1)}{S}\right] \\
&\quad -\frac{\partial F_W(w)}{\partial \sigma^2}\frac{f_W(w)}{2S}(\psi-1) \\
&\quad +\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \sigma^2}\left[1+F_W(w)(\psi-1)-(1-\pi)(1+\psi)\right] \\
\frac{\partial g}{\partial \pi}(w,0) &= -\frac{f_W(w)}{2S}(\psi+1)+\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \pi}\left[1+F_W(\psi-1)-(1-\pi)(\psi+1)\right] \\
\frac{\partial g}{\partial \psi}(w,0) &= -\frac{f_W(w)}{2S}[F_W(w)-(1-\pi)] \\
&\quad +\frac{f_W(w)}{2S^2}\frac{\partial S}{\partial \psi}\left[1+F_W(\psi-1)-(1-\pi)(\psi+1)\right]
\end{aligned}$$

Arnold and Strauss (1991) showed that the PL estimator $\hat{\boldsymbol{\beta}}$, obtained by maximizing (8.8) is consistent and asymptotically normal with covariance matrix estimated by:

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N}\frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}}\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i(\boldsymbol{\beta})^T\right)\left(\sum_{i=1}^{N}\frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}}\right)^{-1}\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}. \quad (8.10)$$

An advantage of this approach is the close connection of pseudo-likelihood 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 (see Chapter 4).

8.3 Application to Quantitative Risk Assessment

The theory of quantitative risk assessment has been introduced before in Section 5.4. The standard approach requires the specification of an adverse event, along with $p(d)$ representing the probability that this event occurs at dose level d. In developmental toxicity studies, the choice exists between fetus and litter based risks. In this chapter we focus on the first approach. The fetus based approach focuses on the risk of a fetus as a function of the dose that was administered to the maternal dam. Here, $p(d)$ represents the probability that a fetus is malformed or of low birth weight. In other words, for the kth fetus in the ith cluster:

$$p(d) = Pr(W_{ik} < W_c \text{ or } M_{ik} = 1|d),$$

where W_c denotes some cutoff value that determines fetal weight low enough to be considered adverse. Hence, $p(d)$ does not depend on correlation parameters that account for clustering among littermates. Formulas (8.11) and (8.12) show the explicit expressions for $p(d)$ with respectively the probit approach and the Plackett-Dale approach.

$$p(d) = 1 - \int_{-\infty}^{-\gamma(d)} \int_{W_c}^{\infty} \boldsymbol{\phi}_2(W, M^*;\ \mu_{(d)}, 0, \sigma_{(d)}, 1, \rho_{(d)})\, dW\, dM^*, \tag{8.11}$$

$$p(d) = \pi + F_{W,M}(W_c, 0), \tag{8.12}$$

where (8.11) is based on (8.1).

8.4 Analysis of EG (Rats) Data

In this section we apply both methods introduced in Section 8.2 to the ethylene glycol data described in Section 2.1.6. Scientific interest lies in the probability of an overall adverse effect, i.e. the probability that an individual fetus is malformed or of low birth weight. Only one cluster-level covariate is available for each fetus, namely the dose that was administered to the maternal dam. The cutoff level for determining low fetal weight is specified as two standard errors below the control average fetal weight $W_c = 2.644$, corresponding to a 1.6% low birth weight rate in control animals.

In order to select a parsimonious model for these data we rely on the pseudo-likelihood procedure for the Plackett-Dale method, described in Section 8.2.2 , for

which we can use the pseudo-likelihood ratio test statistic, defined in (4.12). Model comparisons in the probit framework are done using a robust Wald test statistic. Although the Wald test is in general simpler to apply, it is well known to have some unattractive features such as sensitivity to changes in parameterization (Hauck and Donner 1977). Table 8.1 shows the different fitted models and the results from the model selection with the Plackett-Dale approach.

Model 1, the most complicated one we consider, assumes different quadratic dose trends on the mean weight outcome and on the logit of the estimated malformation probability in a cluster, a linear dose trend on the weight-malformation log odds ratio and separate fetal weight variances within each dose group because of the non-monotone relationship of these variances with dose (cf. Table 2.6):

$$\begin{aligned} \mu_i &= \gamma_0 + \gamma_1 d_i + \gamma_2 d_i^2, \\ \text{logit}(\pi_i) &= \alpha_0 + \alpha_1 d_i + \alpha_2 d_i^2, \\ \ln(\psi_i) &= \zeta_0 + \zeta_1 d_i. \end{aligned}$$

From Table 8.1 follows that the quadratic dose trends on the mean weights (Model 1–Model 2) and on the logits of the malformation probabilities (Model 2–Model 3) are not significant and can be removed from the model. So far, similar results were obtained with the robust Wald test statistic in the probit modelling approach. The linear dose trend on the log odds ratio between weight and malformation is borderline not significant (Model 3–Model 4) in the Plackett-Dale approach. However, for the probit model, a Wald test shows the linear dose trend on the weight–malformation correlation is borderline significant (p-value=0.02). Therefore, we chose not to remove this trend from the model. Linear dose trends on the mean weight outcomes and the logits of the malformation probabilities cannot be removed without an important decrease in fit. Therefore, we propose Model 3 as our final model.

Table 8.2 presents the results of fitting the probit and the Plackett-Dale approach to the data, using the more complex Model 1 and the final Model 3. The table shows the average weight ($\mu(d)$) and standard error of the weight outcomes ($\sigma(d)$) per dose group, as well as the malformation probabilities ($\pi(d)$) and the association between weight and malformation outcomes (either $\rho(d)$ or $\psi(d)$, depending on whether the probit or Plackett-Dale model is being used).

The results of both modelling approaches are in remarkable agreement. The fitted values are close to the observed ones, shown in Table 2.6. Both models suggest

Table 8.1: *EG Study in Rats: Model Selection. All models assume separate fetal weight variances within each dose group. A $*$ indicates inclusion of the corresponding effect on the mean weight outcome (μ), the logit of the malformation probability (logit(π)) or the log odds ratio* $\ln(\psi)$ *between weight and malformation.*

Model	μ			$\text{logit}(\pi)$			$\ln\psi$	
	1	d	d^2	1	d	d^2	1	d
1	*	*	*	*	*	*	*	*
2	*	*		*	*	*	*	*
3	*	*		*	*		*	*
4	*	*		*	*		*	
5	*			*			*	

Comparison	G_a^{2*}	(p-value)
1–2	0.15	(0.698)
2–3	2.73	(0.148)
3 4	2.73	(0.097)
4–5	139.49	(0.000)
1–3	1.23	(0.541)

Table 8.2: *EG Study in Rats: Correlated Probit and Plackett-Dale Model Fits.*

Dose (g/kg/day)	$\mu_w(d)$	$\sigma_w(d)$	$\pi(d)$	$\rho(d)$	$\mu_w(d)$	$\sigma_w(d)$	$\pi(d)$	$\psi(d)$
				Model 1				
	Probit				Plackett-Dale			
0.00	3.439	0.378	0.013	-0.046	3.435	0.381	0.012	0.965
1.25	3.216	0.387	0.072	-0.154	3.208	0.383	0.066	0.632
2.50	2.967	0.367	0.234	-0.263	2.969	0.367	0.231	0.414
5.00	2.386	0.497	0.739	-0.480	2.454	0.459	0.701	0.178
				Model 3				
	Probit				Plackett-Dale			
0.00	3.464	0.382	0.017	-0.035	3.448	0.382	0.024	0.938
1.25	3.203	0.391	0.075	-0.156	3.202	0.385	0.073	0.617
2.50	2.940	0.363	0.227	-0.273	2.957	0.365	0.199	0.406
5.00	2.417	0.496	0.736	-0.483	2.466	0.458	0.713	0.176

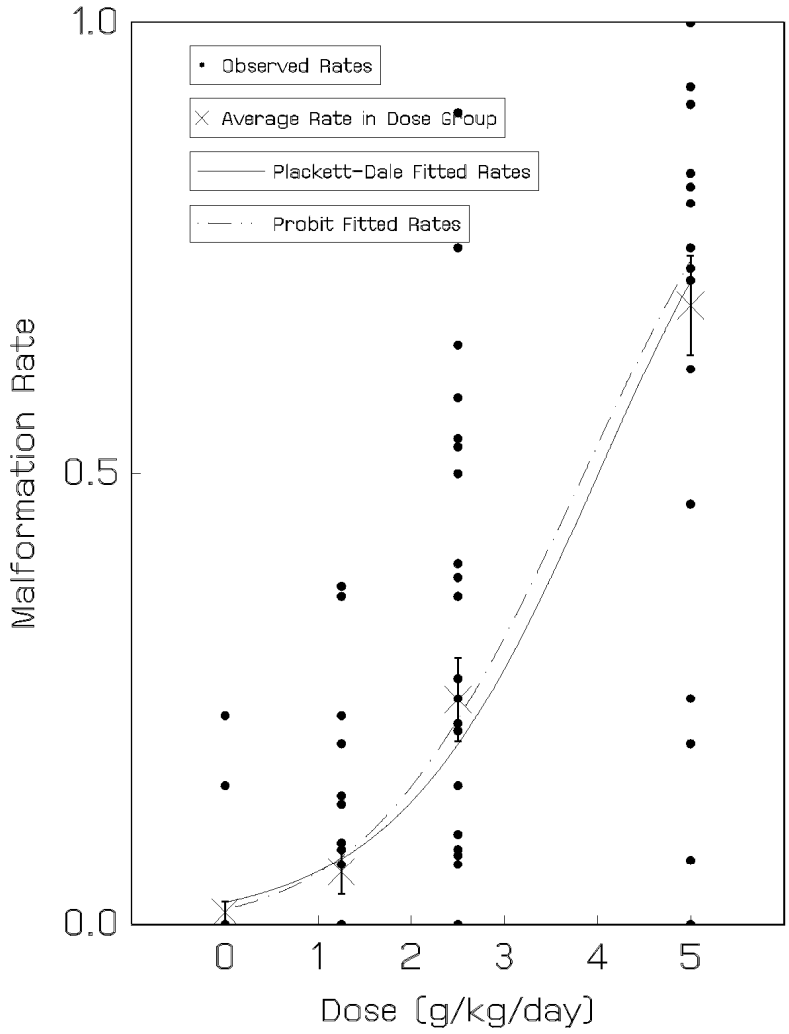

Figure 8.1: *EG Study in Rats: Observed and Fitted Malformation Probabilities for the Correlated Probit and Plackett-Dale Approach.*

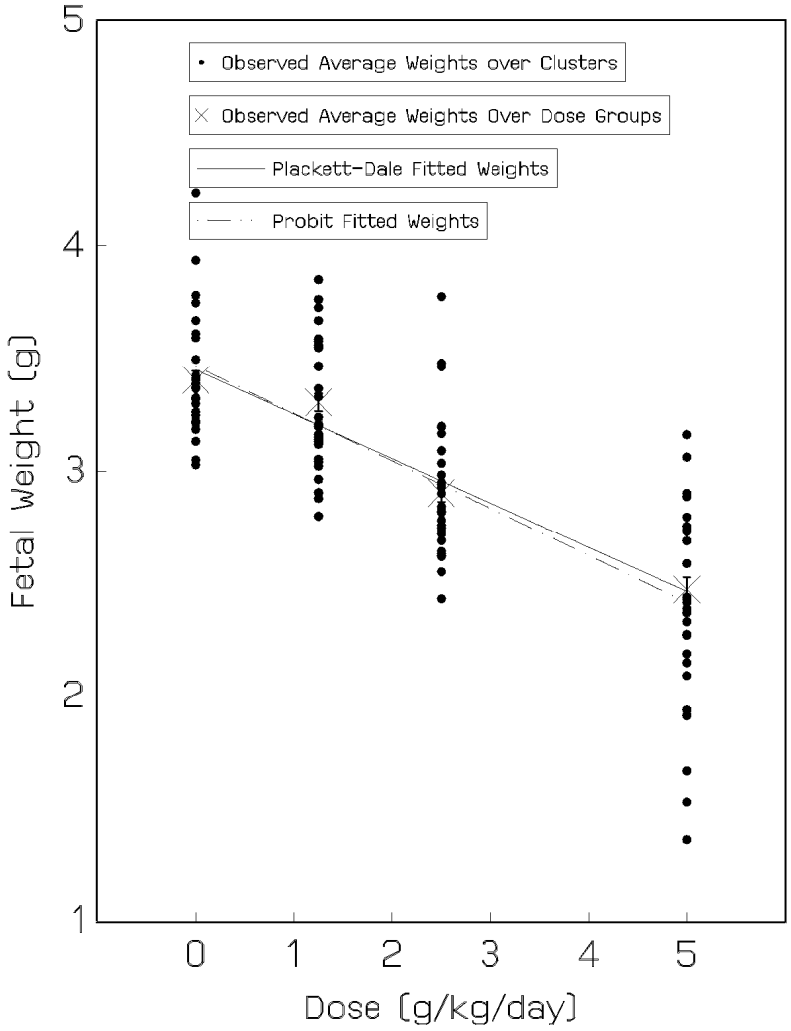

Figure 8.2: *EG Study in Rats: Observed and Fitted Average Weights for the Correlated Probit and Plackett-Dale Approach*

that the average fetal weights drop with increasing doses. The fetal weight variances show a non-monotone behaviour with dose, but are definitely higher in the largest dose group. The malformation probabilities clearly increase with increasing dose effects. Further, the models indicate that the weight–malformation association becomes stronger when the amount of EG administered increases. As expected, the correlation parameter ρ becomes more negative, while the odds ratio parameter becomes much smaller than 1.

Figures 8.1 and 8.2 show the fitted and observed malformation rates, respectively fetal weights as a function of dose for both approaches based on our final dose-response model (Model 3). The observed averaged weights and malformation rates in each dose group are supplemented with pointwise 95% confidence intervals.

Table 8.4 shows the benchmark doses corresponding to the 10% excess risk for Models 1 and 3, as well as the 10% lower limits $BMDL_{10}$ and LED_{10}.

Table 8.3: *EG Study in Rats: Risk Assessment*

Model	Probit			Plackett-Dale		
	BMD_{10}	$BMDL_{10}$	LED_{10}	BMD_{10}	$BMDL_{10}$	LED_{10}
1	1.25	1.00	1.05	1.27	1.05	1.08
3	1.23	1.05	1.07	1.29	1.10	1.12

Again, both modelling approaches agree closely. Furthermore, there is only a small difference between the results for Models 1 and 3. Both models yield benchmark doses of approximately 1.3 with lower limits approximately 1.1. Morgan (1992, p. 175) 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, since it may produce unrealistically low VSDs. Here, similar safe doses were obtained, using two radically different models and different modelling approaches.

8.5 Conclusion

We have considered two latent variable approaches for mixed continuous-discrete outcomes from clustered data. The probit approach uses GEE ideas to incorporate

the clustering, while the Plackett-Dale model is based on pseudo-likelihood ideas. The bivariate latent variables are fundamentally different in the way the association between both variables is described. The correlation coefficient of the bivariate normal induces constant local association in the sense of Holland and Wang (1987) while the odds ratio is a measure of constant global association (Dale 1986; Lapp, Molenberghs and Lesaffre 1998). This is important since it allows us to consider fitting both models simultaneously as a sensitivity analysis. It is comforting to see that not only the model fits are virtually identical, but also that the risk assessment based on both models produces very similar results.

Chapter 9

Validation of Surrogate Endpoints in Clinical Trials

9.1 Introduction

Surrogate endpoints are loosely referred to as endpoints that can be used in lieu of other endpoints in the evaluation of experimental treatments or other interventions. Surrogate endpoints are useful when they can be measured earlier, more conveniently, or more frequently than the endpoints of interest, which are referred to as the "true" or "final" endpoints (Ellenberg and Hamilton 1989). The need to evaluate treatment benefits as fast as possible on easily measurable endpoints has always been a preoccupation in clinical research. In most clinical trials, several endpoints are measured over the course of the disease, and treatment benefits can be evaluated on all of them. In general, however, one endpoint is pre-specified as being of primary interest, and serves to determine the significance of any observed treatment benefit. Ideally, the primary endpoint should be the one which is most clinically relevant, but considerations of time and cost may force the investigators to use some other endpoint instead. Examples abound, particularly in chronic diseases in which the duration of survival is the ultimate endpoint which clinicians would like to affect, but cannot always afford to observe due to the prolonged period of follow-up needed. Alternative endpoints must then be considered as surrogates for survival: for instance, disease recurrence after surgical removal of early cancers, tumor shrinkage (usually called "response") in advanced cancers, progression to AIDS

in HIV positive subjects, lymphocyte T4 (CD4) counts in AIDS patients, etc. The true endpoint may also be a rare event, such as a disease or unexpected side-effect of treatment, which occurs so infrequently as to make a study unrealistically large. Finally, surrogate endpoints may be needed when competing risks and secondary treatments contaminate the impact of an experimental treatment or intervention upon the true endpoint (Wittes, Lagakos and Probstfield 1989). In some cases, the surrogate endpoint directly affects the patient's condition, and is therefore itself of clinical relevance, in other cases it is merely a biological marker of the disease process leading to the final endpoint. In the latter case the term "surrogate marker" may be preferred, and the endpoint of interest is then referred to as "the clinical endpoint".

While the practice of looking at multiple endpoints is by no means recent in clinical research, the validity of using one endpoint as a surrogate for another has been raised and studied only over the last few years. The dramatic surge of the AIDS epidemic, the pressure for an accelerated evaluation of new therapies, etc. have all played a major role in focusing attention on the need for a formal definition of surrogate endpoints, along with practical methods to validate them. Much applied research on surrogate endpoints has concentrated on evaluating the possible value of changes in CD4 counts as surrogates for time to clinical events in asymptomatic HIV-infected persons and in AIDS patients (Machado, Gail and Ellenberg 1990; Lin, Fischl and Schoenfeld 1993, De Gruttola et al. 1993, De Gruttola and Tu 1995). This research has revealed that, although CD4 counts were useful to monitor the disease process, they were only of limited value as a surrogate marker for clinically relevant endpoints (Lagakos and Hoth 1992). In cardiovascular disease, the unsettling discovery that the two major antiarrhythmic drugs encanaide and flecanaide reduced arrhythmia but cause a more than 3-fold increase in overall mortality stressed the need for caution in using non-validated surrogate markers in the evaluation of the possible clinical benefits of new drugs (The Cardiac Arrhythmia Suppression Trial (CAST) Investigators 1989).

The validation of surrogate endpoints is a controversial issue (Boissel et al. 1992; Fleming and DeMets 1996; De Gruttola et al. 1997) and should be rigorously established. In a landmark paper, Prentice (1989) proposed a formal definition of surrogate endpoints and outlined how potential surrogate endpoints could be validated. Much debate ensued, for the criteria set out by Prentice are too stringent

(Fleming et al. 1994) and neither necessary nor sufficient for his definition to be fulfilled, except in the special case of binary outcomes (Buyse and Molenberghs 1998). In addition, Freedman, Graubard and Schatzkin (1992) showed that these criteria were not straightforward to verify through statistical hypothesis tests. They introduced the *proportion explained* (PE) to quantify how much of the treatment effect is captured by the surrogate endpoint. The latter proposal is itself surrounded with difficulties, the most dramatic one being that it is not confined to the unit interval. Another major drawback of PE is that it is not estimated accurately unless treatment has a highly significant effect on the true endpoint, a rare situation in which the need for a surrogate is questionable. As a consequence, use of this measure should no longer be recommended in practice (Flandre and Saidi 1998, Molenberghs and Buyse 1999).

Buyse and Molenberghs (1998), henceforth called BM, proposed to replace PE by two new measures to assess the quality of a surrogate. The first one, termed *relative effect* (RE) is the (population-averaged) effect of the treatment on the true endpoint relative to that on the surrogate endpoint. The second one is the *adjusted association* between both endpoints, an individual measure of agreement between both endpoints after accounting for the effect of treatment. Their methodology focuses on surrogate and true endpoints which are both binary or both normally distributed. Technically, a joint model for both endpoints is required. In the binary case, the Dale (1986) model is used, whereas for continuous endpoints, a bivariate normal model is considered. The association parameters are then the log odds ratio and the Pearson correlation coefficient respectively. In this chapter we first extend the BM proposals to cases where the surrogate and the true endpoints are of a different data type (mixed binary–continuous). In this situation, the choice of a joint model is less straighforward. Following Geys et al. (1999b), both a hybrid probit model and a hybrid Dale model are considered, where one latent variable is observed directly, and the other latent variable is recorded in dichotomized form. For the sake of presentation, we consider a binary surrogate and a continuous true endpoint, but the reverse case is entirely similar (see also Molenberghs, Geys and Buyse 1999).

In order to be informative and of practical value, however, the validation of a surrogate endpoint will typically require large numbers of observations. It is therefore useful to consider extensions of the foregoing quantities to situations in which

data are available from several randomized experimens, where the experimental unit can be center in a multicentric trial or trial in a meta-analysis of several trials. In that case, the data have a similar structure as in developmental toxicity studies. Different trials or centers are assumed to be independent. Individuals within a trial (center) may however be correlated possibly yielding multiple associated outcomes of potentially mixed data types.

Buyse et al. (1999) show that the individual-level association between the surrogate and final endpoints carries over naturally to this setting. The notion of relative effect, on the other hand, can be extended to a trial-level measure of association between the effects of treatment on both endpoints. Their approach suggests a new definition of validity in terms of the quality of both trial-level and individual-level associations between the surrogate and true endpoints. The quality of a surrogate at the trial level is assessed by means of a coefficient of determination R^2_{trial}. At the individual level the squared correlation R^2_{indiv} between the surrogate and true endpoint, after adjustment for both the trial effects and the treatment effects is used. A surrogate will be said to be valid when it is both trial-level valid ($R^2_{trial} \approx 1$) and individual-level valid ($R^2_{indiv} \approx 1$). From a modelling perspective, a two-stage hierarchical model is required. This can be fitted using a variety of methods, such as linear mixed-effects models methodology (Verbeke and Molenberghs 1997), a two-stage approach, or pseudo-likelihood (Geys et al. 1999b). As Buyse et al. (1999) centered solely on the case of normally distributed endpoints, it is necessary to explore other settings, often more complicated due to the absence of a unifying framework such as the multivariate normal distribution.

Section 9.2 gives a brief history on validation criteria for a single trial. In Section 9.3 we extend the BM proposals for normally and binary endpoints in a single trial case to situations where the surrogate and true endpoints are of different data types (binary/continuous). In Section 9.4, the proposed methodologies are then exemplified on data from a clinical trial, described in Section 2.3. Section 9.5 looks at meta-analytic extensions, which are exemplified in Sections 9.6 and 9.7.

9.2 A Brief History on Validation Criteria in a Single Trial

Let us first introduce some notation. Throughout this chapter we assume that T and S are random variables that denote the true and surrogate endpoints respectively and Z is an indicator variable for treatment. We restrict attention to a binary treatment indicator ($Z = 0$ or 1).

BM have given an overview, with discussion, of common practice for validation of surrogate endpoints. In this section, we summarize their main arguments.

9.2.1 Prentice's Criteria

Prentice (1989) defined a surrogate endpoint as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint":

$$f(S|Z) = f(S) \Leftrightarrow f(T|Z) = f(T) \tag{9.1}$$

where $f(X)$ denotes the probability distribution of a random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the value of Z. Note that this definition involves the triplet (T, S, Z), hence the endpoint S is a surrogate for T only with respect to the effect of some specific treatment Z. Based on his definition, Prentice (1989) proposed the following 4 criteria to validate a surrogate endpoint:

$$f(T|Z) \neq f(T), \tag{9.2}$$

$$f(S|Z) \neq f(S), \tag{9.3}$$

$$f(T|S) \neq f(T), \tag{9.4}$$

$$f(T|S) = f(T|S, Z). \tag{9.5}$$

Criteria (9.2) and (9.3) measure departures from the null hypothesis, implicit in (9.1). Criterion (9.4) implies that the surrogate endpoint has prognostic value for the true endpoint. Criterion (9.5) requires S to fully capture the effect of treatment on the true endpoint, that is: there is no effect of treatment on the true endpoint after correction for the surrogate endpoint. Of course, this last condition is so restrictive that it rarely holds in practice. Moreover, it is hard to verify since it would formally require equivalence testing.

BM have shown that criteria (9.2)–(9.5) are necessary and sufficient to establish the validity of *binary* surrogate endpoints. Let us first show that ($\Rightarrow$) holds in (9.1). By definition, we have

$$f(T|Z) = \int f(T,S|Z)dS = \int f(T|S,Z)f(S|Z)dS.$$

By (9.1), $f(S|Z) = f(S)$ and

$$f(T|Z) = \int f(T|S,Z)f(S)dS.$$

By (9.5), this can be written as:

$$f(T|Z) = \int f(T|S)f(S)dS = f(T).$$

Next, we show that ($\Leftarrow$) holds in (9.1) for binary surrogate endpoints. First note that

$$f(T|Z) = \int f(T|S,Z)f(S|Z)dS = \int f(T|S)f(S|Z)dS. \tag{9.6}$$

Similarly,

$$f(T) = \int f(T|S)f(S)dS. \tag{9.7}$$

Since $f(T|Z) = f(T)$, by subtraction of (9.7) from (9.6),

$$\int f(T|S)\left[f(S|Z) - f(S)\right]dS = 0.$$

For a binary surrogate endpoint, this reduces to

$$\left[f(T|S=0) - f(T|S=1)\right]\left[f(S=1|Z) - f(S=1)\right] = 0.$$

By (9.4), ($\Leftarrow$) holds.

These criteria however do not necessarily establish the validity of other than binary surrogate endpoints. The simplest counterexample is found by considering a multi-categorical surrogate endpoint, as illustrated in Table 9.1.

Table 9.1: *Relationship between T (true endpoint) and S (surrogate endpoint), and Z (treatment) in an artificial set of data for which $f(T|S) \neq f(T)$, $f(S|Z) \neq f(S)$, and $f(T|S,Z) = f(T|S)$ yet $f(T|Z) = f(T)$. Cell counts represent numbers of patients.*

		Z	
S	T	0	1
0	0	40	120
	1	10	30
1	0	150	50
	1	150	50
2	0	30	50
	1	120	200

9.2.2 Freedman's Proportion Explained

Freedman et al. (1992) argued that criterion (9.5) raises a conceptual difficulty in that it would require the statistical test for treatment effect on the true endpoint to be *non*-significant after adjustment for the surrogate. The non-significance of this test does not prove that the effect of treatment upon the true endpoint is *fully* captured by the surrogate. Therefore, they supplemented these criteria with the so-called *proportion explained*, the proportion of the treatment effect explained by the surrogate. In this paradigm, it is *suggested* that a good surrogate is one which explains a large proportion (but not necessarily everything) of that effect. Let $PE = PE(T,S,Z)$ stand for the proportion of the effect of Z on T which can be explained by S. An estimate of $PE(T,S,Z)$ is then as follows:

$$PE(T,S,Z) = 1 - \frac{\beta_S}{\beta}$$

where β and β_S are the estimates of the effect of Z on T without and with adjustment for S. The PE is large if β_S is small in comparison to β. Prentice's criterion (9.5) requires that $\beta_S = 0$, or equivalently $PE = 1$. A surrogate endpoint for which $PE < 1$ explains only part of the treatment effect on the true endpoint (Choi et

al. 1993). Hence, Freedman et al. suggested that a good surrogate is one for which the PE is close to unity. However, this reasoning is not generally valid. Several conceptual difficulties surrounding the PE have been outlined in the literature (Choi et al. 1993; Lin, Fleming and De Gruttola 1997; Buyse and Molenberghs 1998; Flandre and Saidi 1998; Buyse et al. 1999) For example, Lin, Fleming and De Gruttola (1997) argue that a value of PE near 1 is not sufficient for inferring that a marker is a good surrogate for a clinical endpoint, since a variety of factors such as drug toxicity, non-compliance with study medications, and incomplete marker information can artificially raise this value to 1, even for poor surrogates. In addition, Molenberghs and Buyse (1999) and Flandre and Saidi (1998) note that there are serious conceptual difficulties with the PE. First, PE is not a proportion since it can lie anywhere on the real line, which makes its interpretation problematic. For example, it is possible for PE to be greater than 1 if β_S and β have opposite signs, i.e., if the adjustment for S changes the *direction* of the effect of Z on T. It is even possible for PE to be negative, which can hardly be justified for a proportion. Secondly, its confidence limits tend to be very wide, unless trial sizes are very large and the treatment effect on the true endpoint is strong enough (in which case the need for a surrogate is questionable). When Fieller confidence intervals (Herson 1995) are used instead of the more common but less performing delta intervals, the confidence interval would in many reported cases be unbounded. These arguments have lead Flandre and Saidi (1998) and Molenberghs and Buyse (1999) to simply recommend that use of the measure be discontinued.

9.2.3 New Validation Measures for a Single Trial

The previously described problems with the PE have lead BM to replace this measure by two other quantities: the relative effect (RE) and the treatment-adjusted association between the surrogate and the true endpoint (γ_Z).

Relative Effect

The relative effect (RE) links the surrogate and the true endpoint at the population-averaged level. Let $RE(T, S, Z)$ stand for the effect of Z on T relative to that of Z on S. An intuitively appealing way of defining $RE(T, S, Z)$ is as follows:

$$RE(T, S, Z) = \frac{\beta}{\alpha},$$

where β and α are the estimates of the effect of treatment on T and S. They are estimated in the verification process of (9.2) and (9.3), respectively. Clearly, RE connects the treatment effects at the population-averaged levels. Therefore, BM call a surrogate for which $RE \approx 1$ *valid at the population level.* To be of practical value, that is to enable prediction of the effect of Z on T based on an observed effect of Z on S, the RE must be estimated with good precision. This requires large numbers of observations. Clearly, in order to be meaningful the validation process will have to be based on large-scale randomized evidence. Such evidence is not always available from individual trials. Therefore, BM suggest that meta-analyses based on individual patient data from several randomized trials may be the best way to validate a surrogate endpoint. We will return to this matter in Section 9.5.

Adjusted Association

The treatment-adjusted association (γ_Z) is the subject-specific association between the surrogate and true endpoints, adjusting for treatment. For binary endpoints γ_Z takes the form of a log odds ratio (see BM). Large (infinite) γ_Z means that the surrogate and true endpoints are very similar (the same), possibly up to a deterministic transformation. For normal endpoints, γ_Z is the correlation between the error terms of the surrogate and true endpoints or, equivalently, the coefficient of S in the regression of T on Z and S simultaneously. When $\gamma_Z \approx \infty$ (binary case) or $\gamma_Z \approx 1$ (normal case), the surrogate is said to be *valid at the individual level.* Even if either α or β would be small (RE far from 1), a surrogate for which γ_Z is large, would be useful to predict the outcome at the individual level.

Remark

Let us indicate the link between PE, RE, and adjusted association, as presented by Molenberghs and Buyse (1999). For ease of exposition, we assume both S and T to be normally distributed:

$$\begin{aligned} S_i|Z_i &= \mu_S + \alpha Z_i + \varepsilon_{S_i}, \\ T_i|Z_i &= \mu_T + \beta Z_i + \varepsilon_{T_i}, \end{aligned}$$

where $i = 1, \ldots, n$ indicates patients, and the error terms have a joint zero-mean normal distribution with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}.$$

Formally,

$$\begin{aligned} RE &= \beta/\alpha \\ \rho_Z &= \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}}. \end{aligned}$$

Rewriting

$$\sigma_{SS} = \sigma^2, \quad \sigma_{ST} = \rho_Z \lambda \sigma^2, \quad \sigma_{TT} = \lambda^2 \sigma^2,$$

we deduce $\beta_S = \beta - \rho_Z \lambda \alpha$ and

$$PE = \rho_Z \lambda \frac{\alpha}{\beta} = \lambda \rho_Z \frac{1}{RE}.$$

Therefore, in addition to the earlier mentioned problems with the PE, the quantity is hard to interpret since it is an amalgamation of three sources of information:

- the adjusted association ρ_Z, which is an individual-level measure of agreement;
- the RE, expressing the relation between the two treatment effects at the trial level;
- the variance ratio λ, which is a nuissance characteristic.

Clearly, this is less attractive to interpret than the couple RE and γ_Z. RE connects the treatment effects at the population-averaged level, while γ_Z connects them at the individual-specific level. In the remaining of this chapter we will abandon the PE approach.

9.3 Validation of Surrogate Markers with Mixed Continuous and Binary Endpoints in a Single Trial

In this section, we propose joint models for a binary surrogate and a continuous true endpoint, in the single trial case (Molenberghs, Geys and Buyse 1999). The reverse case is entirely similar.

Let $\tilde{S}_i$ be a latent variable of which S_i is the dichotomized version. In Section 9.3.1 we describe a bivariate normal model for $\tilde{S}_i$ and T_i, resulting in a probit-linear model for S_i and T_i. Section 9.3.2 presents an alternative formulation based on the bivariate Plackett (1965) density and resulting in a Plackett-Dale model.

9.3.1 A Probit Formulation

In this formulation, we assume the following model:

$$T_i = \mu_T + \beta Z_i + \varepsilon_{Ti}, \tag{9.8}$$

$$\tilde{S}_i = \mu_S + \alpha Z_i + \varepsilon_{Si}, \tag{9.9}$$

where μ_S and μ_T are fixed intercepts and α and β are the fixed effects of the treatment Z on the surrogate and true endpoints respectively. Further, ε_{Si} and ε_{Ti} are correlated error terms, assumed to satisfy:

$$\begin{pmatrix} \varepsilon_{Ti} \\ \varepsilon_{Si} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \frac{\rho\sigma}{\sqrt{1-\rho^2}} \\ \frac{\rho\sigma}{\sqrt{1-\rho^2}} & \frac{1}{1-\rho^2} \end{pmatrix} \right).$$

Model (9.8)–(9.9) specifies a bivariate normal density. The variance of $\tilde{S}_i$ is chosen for reasons that will be made clear in what follows. From this model, it is easily seen that the density of T_i is univariate normal with mean $\mu_T + \beta Z_i$ and variance σ^2, implying that the parameters μ_T, β, and σ^2 can be determined using linear regression software with response T_i and single covariate Z_i. Similarly, the conditional density of $\tilde{S}_i$, given Z_i and T_i is

$$\tilde{S}_i | T_i, Z_i \sim N\left[\left(\mu_S - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mu_T \right) + \left(\alpha - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\beta \right) Z_i + \frac{\rho}{\sigma\sqrt{1-\rho^2}} T_i; 1 \right],$$

motivating our earlier choice for the covariance matrix of T_i and $\tilde{S}_i$. The corresponding probability

$$P(S_i = 1 | T_i, Z_i) = \Phi(\lambda_0 + \lambda_Z Z_i + \lambda_T T_i), \tag{9.10}$$

where

$$\lambda_0 = \mu_S - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mu_T, \tag{9.11}$$

$$\lambda_Z = \alpha - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\beta, \tag{9.12}$$

$$\lambda_T = \frac{\rho}{\sigma\sqrt{1-\rho^2}}, \tag{9.13}$$

and Φ is the standard normal cumulative distribution function. The λ parameters can be found by fitting model (9.10) to S_i with covariates Z_i and T_i. This can be done with standard logistic regression software if it allows to specify the probit rather than the logit link (e.g., the LOGISTIC procedure in SAS). Given the parameters from the linear regression on T_i (μ_T, β, and σ^2) and the probit regression on S_i (λ_0, λ_Z, and λ_T), the parameters from the linear regression on $\tilde{S}_i$ can now be obtained from (9.11)–(9.13):

$$\mu_S = \lambda_0 + \lambda_T\mu_T, \tag{9.14}$$

$$\alpha = \lambda_Z + \lambda_T\beta, \tag{9.15}$$

$$\rho^2 = \frac{\lambda_T^2\sigma^2}{1+\lambda_T^2\sigma^2}. \tag{9.16}$$

The asymptotic covariance matrix of the parameters (μ_T, β) can be found from standard linear regression output. The estimated variance of σ^2 equals $2\hat{\sigma}^4/N$. The asymptotic covariance of $(\lambda_0, \lambda_Z, \lambda_T)$ follows from logistic (probit) regression output. These three statements yield the covariance matrix of the six parameters upon noting that it is block-diagonal. In order to derive the asymptotic covariance of (μ_S, α, ρ) it suffices to calculate the derivatives of (9.14)–(9.16) with respect to the six original parameters and apply the delta method. They are:

$$\frac{\partial(\mu_S, \alpha, \rho)}{\partial(\mu_T, \beta, \sigma^2, \lambda_0, \lambda_Z, \lambda_T)} = \begin{pmatrix} \lambda_T & 0 & 0 & 1 & 0 & \mu_T \\ 0 & \lambda_T & 0 & 0 & 1 & \beta \\ 0 & 0 & h_1 & 0 & 0 & h_2 \end{pmatrix},$$

where

$$h_1 = \frac{1}{2\rho}\frac{\lambda_T^2}{(1+\lambda_T^2\sigma^2)^2},$$

$$h_2 = \frac{1}{2\rho}\frac{2\lambda_T\sigma^2}{(1+\lambda_T^2\sigma^2)^2}.$$

In addition, we developed a program in GAUSS that performs the joint estimation directly by maximizing the likelihood based on contributions (9.8) and (9.10).

The adjusted association is given by ρ (which is slightly different from γ_Z since we have to correct for the variances σ^2 and $1/(1-\rho^2)$). The relative effect, $RE = \beta/\alpha$, can be determined directly from the output.

9.3.2 A Plackett-Dale Formulation

Let us now consider an alternative approach. Assume that the cumulative distributions of S_i and T_i are given by F_{S_i} and F_{T_i}. The joint cumulative distribution of both these quantities has been studied by Plackett (1965):

$$F_{T_i,S_i} = \begin{cases} \dfrac{1 + (F_{T_i} + F_{S_i})(\psi_i - 1) - S(F_{T_i}, F_{S_i}, \psi_i)}{2(\psi_i - 1)} & \text{if } \psi_i \neq 1, \\ F_{T_i} F_{S_i} & \text{if } \psi_i = 1, \end{cases}$$

where

- ψ_i is the global odds ratio (see Chapter 6),
- $S(q_1, q_2, \psi) = \sqrt{[1 + (q_1 + q_2)(\psi - 1)]^2 + 4\psi(1 - \psi)q_1 q_2}$.

Based upon this distribution function, we can derive a bivariate Plackett "density" function $G_i(t, s)$ for mixed continuous- binary outcomes (see also Chapter 8). Suppose the success probability for S_i is denoted by π_i, then we can define $G_i(t, s)$ by specifying $G_i(t, 0)$ and $G_i(t, 1)$ such that they sum to $f_{T_i}(t)$. If we define $G_i(t, 0) = \partial F_{T_i,S_i}(t, 0)/\partial t$, then this leads to specifying G_i by:

$$G_i(t,0) = \begin{cases} \dfrac{f_{T_i}(t)}{2}\left(1 - \dfrac{1+F_{T_i}(t)(\psi_i-1)-F_{S_i}(s)(\psi_i+1)}{S\left(F_{T_i},1-\pi_i,\psi_i\right)}\right) & \text{if } \psi_i \neq 1, \\ f_{T_i}(t)(1 - \pi_i) & \text{if } \psi_i = 1, \end{cases}$$

and

$$G_i(t, 1) = f_{T_i}(t) - G_i(t, 0).$$

In this formulation we assume $T_i \sim N(\mu_i, \sigma^2)$, with $\mu_i = \mu_T + \beta Z_i$ and $\text{logit}(\pi_i) = \mu_S + \alpha Z_i$ with similar notation as in the probit case. The global odds ratio is assumed

to be constant. If we denote

$$\boldsymbol{\theta}_i = \begin{pmatrix} \mu_i \\ \sigma^2 \\ \pi_i \\ \psi \end{pmatrix} \text{ and } \boldsymbol{\eta}_i = \begin{pmatrix} \mu_i \\ \ln(\sigma^2) \\ \text{logit}(\pi_i) \\ \ln(\psi) \end{pmatrix},$$

estimates of the regression parameters $\boldsymbol{\tau} = (\mu_T, \mu_S, \beta, \alpha, \ln\sigma^2, \ln\psi)$ are easily obtained by solving the estimating equations $\boldsymbol{U}(\boldsymbol{\tau}) = 0$, using a Newton-Raphson iteration scheme, where $\boldsymbol{U}(\boldsymbol{\tau})$ is given by:

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\tau}}\right)^T \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\theta}_i}\right)^{-T} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln G_i(t_i, s_i)\right).$$

Note that the adjusted association γ_Z is given by ψ in this case and the relative effect $RE = \beta/\alpha$ can be readily determined.

9.4 An Example in Ophthalmology

In this section, we analyze the data described in Section 2.3. The binary indicator for treatment (Z_{ij}) is set to 0 for placebo and to 1 for Interferon-α.

Let us first present the results of Buyse and Molenberghs (1998). They assume that visual acuity is a continuous, normally distributed variable and investigate whether the change in visual acuity at 6 months after starting treatment (S_{ij}) can be used as a surrogate for the change in visual acuity at 1 year (T_{ij}). Table 9.2 shows the visual acuity (mean and standard error) by treatment group at baseline, at 6 months and at 1 year.

The results of Buyse and Molenberghs (1998) are summarized in Table 9.3. The analyses were carried out using the SAS procedure MIXED (Littel et al. 1996). The first three Prentice Criteria (9.2)–(9.4) are provided by tests of significance of α, β and γ in the following models:

$$\begin{aligned} S_{ij}|Z_{ij} &= \mu_S + \alpha Z_{ij} + \varepsilon_{S_{ij}} \\ T_{ij}|Z_{ij} &= \mu_T + \beta Z_{ij} + \varepsilon_{T_{ij}} \\ T_{ij}|S_{ij} &= \mu + \gamma S_{ij} + \varepsilon_{ij} \end{aligned}$$

Table 9.2: *ARMD Study: Mean (standard error) of Visual Acuity at Baseline, at 6 Months and at 1 Year According to Randomized Treatment Group (P=Placebo, I=Interferon-α)*

Time point	P	I	Total
Baseline	55.3 (1.4)	54.6 (1.3)	55.0 (1.0)
6 months	49.3 (1.8)	45.5 (1.8)	47.5 (1.3)
1 year	44.4 (1.8)	39.1 (1.9)	42.0 (1.3)

Table 9.3: *ARMD Study: The quantities of interest for the validation of a surrogate endpoint (T: true endpoint, S: surrogate endpoint, Z: treatment, f(.): density function, PE: proportion explained, RE: relative effect)*

Quantity of interest	Estimate	Test
Effect of treatment on true endpoint	$\beta = 4.12$ (s.e. 2.32)	$H_0 : f(T\|Z) = f(T)$
Effect of treatment on surrogate endpoint	$\alpha = 2.83$ (s.e. 1.86)	$H_0 : f(S\|Z) = f(S)$
Effect of surrogate on true endpoint	$\gamma = 0.95$ (s.e. 0.06)	$H_0 : f(T\|S) = f(T)$
Proportion of treatment effect on true endpoint explained by surrogate	$PE = 0.65$ (95% confidence interval [-0.22;1.51])	
Effect of treatment on true endpoint relative to that on surrogate endpoint	$RE = 1.45$ (95% confidence interval [-0.48;3.39])	
Adjusted effect of surrogate on true endpoint	$\rho = 0.75$ (95% confidence interval [0.69;0.82])	

Here, $\alpha = 2.83$ (s.e. 1.86), $\beta = 4.12$ (s.e. 2.32), $\gamma = 0.95$ (s.e. 0.06). Thus, there is little evidence for an effect of Z on either endpoint but overwhelming evidence that the surrogate is strongly correlated with the true endpoint. Therefore the validation procedure has to stop inconclusively. Note, however, that the lack of statistical significance of α and β could merely be due to the insufficient number of observations available in this trial. Freedman's proportion explained was calculated as 0.65 with 95 % confidence interval [-0.22;1.51] and the relative effect was 1.45 with 95% confidence interval [-0.48;3.39]. It is noteworthy that, while the confidence intervals for PE and RE are too wide to convey any useful information, $\gamma_Z = 0.75$ with confidence interval [0.69;0.82], which implies that a very large part of the variability of the surrogate is shared with the true endpoint.

Let us now apply the methodology for mixed continuous and discrete outcomes. First, we consider dichotomized visual acuity at 6 months as the surrogate and (continuous) visual acuity at 12 months as the true endpoint. 0 , is achieved by setting a binary variable to 1 if visual acuity at 6 months is larger than the value at baseline and to 0 otherwise. Let us first consider the probit model. The parameter estimates for the true endpoint are $\mu_T = 11.04$ (s.e. 1.57), $\beta = 4.12$ (s.e. 2.32), and $\sigma = 15.95$ (s.e. 0.82). The parameter estimates for the surrogate endpoint are $\mu_S = 0.64$ (s.e. 0.20) and $\alpha = 0.39$ (s.e. 0.28) and the correlation is $\rho = 0.74$ (s.e. 0.05). Note that the parameter estimates for the true endpoint coincide with those in BM, who employed a bivariate normal model for the case where both outcomes are continuous. The relative effect is estimated to be $RE = 10.44$ (95% confidence interval $[-1.77; 22.65]$) and the adjusted correlation $\rho = 0.74$ (95% confidence interval $[0.64; 0.84]$). While care has to be taken with the RE since both numerator and denominator are non-significant (leading to a Fieller confidence interval equal to the whole real line), the adjusted correlation is estimated very precisely and there is clearly a strong correlation between both endpoints. BM found an adjusted correlation of 0.75 (95% confidence interval $[0.69; 0.81]$) which agrees remarkably well with our results. The slightly wider standard error results from the loss of information through dichotomizing the surrogate endpoint. Let us now analyze the same data using the Plackett-Dale model. The parameter estimates for the true endpoint are $\mu_T = 10.89$ (s.e. 1.56), $\beta = 4.02$ (s.e. 2.32), and $\sigma = 16.04$ (s.e. 0.81). These results are relatively close to the ones obtained with the probit model since in both cases a linear regression of T on Z is assumed. The binary regression of S on T and Z

contains additional information about the true endpoint parameters as well, which is why the results are not exactly equal. The values for the surrogate endpoint are $\mu_S - 0.74$ (s.e. 0.19) and $\alpha - 0.45$ (s.e. 0.30) and the log odds ratio, $\ln \psi - 2.85$ (s.e. 0.37) with corresponding odds ratio 17.29. The relative effect is estimated to be $RE = 8.92$ (95% confidence interval $[-0.41; 18.25]$), in close agreement with the above estimate. For sake of presentation and comparison, these results are summarized in Table 9.4.

Table 9.4: *ARMD Study: The quantities of interest for the validation of the surrogate endpoint*

Quantity of interest	Estimate	
	Probit	Plackett-Dale
Effect of treatment on true endpoint	$\beta = 4.12$ (s.e. 2.32)	$\beta = 4.02$ (s.e. 2.32)
Effect of treatment on surrogate endpoint	$\alpha = 0.39$ (s.e. 0.28)	$\alpha = 0.45$ (s.e. 0.30)
Effect of treatment on true endpoint relative to that on surrogate endpoint	RE – 10.44 (95% confidence interval [-1.77;22.65])	RE – 8.92 (95% confidence interval [-0.41;18.25])
Adjusted effect of surrogate on true endpoint	$\rho = 0.74$ (95% confidence interval [0.64;0.84])	$\ln(\psi) = 2.85$ (95% confidence interval [2.12;3.58])

Next, we consider the binary indicator for loss of at least 2 lines of vision at 6 months as a surrogate for (continuous) visual acuity at 12 months. With the probit model, the regression coefficients (standard errors) for the true endpoint are $\mu_T = 11.04$ (s.e. 1.57), $\beta = 4.12$ (s.e. 2.32), and $\sigma = 15.95$ (s.e. 0.82). The values for the surrogate endpoint are $\mu_S = -0.43$ (s.e. 0.19) and $\alpha = 0.58$ (s.e. 0.28) and the correlation is $\rho = 0.75$ (s.e. 0.05). The regression parameters for the true endpoint are once again in agreement with earlier findings. The relative effect is estimated to be $RE = 7.08$ (95% confidence interval $[5.77; 19.93]$) and the adjusted correlation $\rho = 0.75$ (95% confidence interval $[0.66; 0.84]$). With the Plackett-Dale model, the regression coefficients (standard errors) for the true endpoint are $\mu_T = 10.95$ (s.e. 1.56), $\beta = 3.74$ (s.e. 2.30), and $\sigma = 15.97$ (s.e. 0.82). The values for the surrogate

endpoint are $\mu_S = -0.38$ (s.e. 0.20) and $\alpha = 0.63$ (s.e. 0.29) and the log odds ratio $\ln \psi = -2.78$ (s.e. 0.30) with corresponding odds ratio 16.18. The relative effect is estimated to be $RE - -5.93$ (95% confidence interval $[-17.17; 5.32]$).

Finally, we consider the more interesting situation of (continuous) visual acuity at 6 months as a surrogate for the binary indicator for loss of at least 3 lines of vision lost at one year. With the probit model, the regression coefficients (standard errors) for the true endpoint are $\mu_T = -0.36$ (s.e. 0.21), $\beta = 0.60$ (s.e. 0.30). The values for the surrogate endpoint are $\mu_S = 5.53$ (s.e. 1.26), $\alpha = 2.83$ (s.e. 1.87), and $\sigma = 12.80$ (s.e. 0.66). The correlation is $\rho = 0.81$ (s.e. 0.04). The relative effect is estimated to be $RE = 4.75$ (95% confidence interval $[-5.11; 14.61]$). With the Plackett-Dale model , the regression coefficients (standard errors) for the true endpoint are $\mu_T = -0.36$ (s.e. 0.19), $\beta = 0.58$ (s.e. 0.28), and $\sigma = 12.90$ (s.e. 0.65). The values for the surrogate endpoint are $\mu_S = 5.89$ (s.e. 1.24) and $\alpha = 2.72$ (s.e. 1.84) and the log odds ratio $\ln \psi = 2.83$ (s.e. 0.29) with corresponding odds ratio 16.93. The relative effect is estimated to be $RE = 4.67$ (95% confidence interval $[-5.00; 14.35]$).

Conclusion

The examples show that the two approaches yield very comparable results, so that in practice one approach can be regarded as a sensitivity analysis for the other. It is interesting to note that the discretization of a continuous endpoint into a binary variable does not lead to a great loss of information for the purposes of validation of surrogate endpoints. The reliability of the analyses is primarily driven by the number of observations, rather than by the data type of the endpoints considered.

The example used in this section underscores one of the greatest practical difficulties of surrogate validation, i.e., the need for very large datasets from randomized experiments. As mentioned above, the confidence limits of the proportion explained tend to be hopelessly wide regardless of the number of observations, which casts doubts on the value of this quantity. The confidence limits of the relative effect will also be wide unless the number of observations is large. Here, for instance, with only 190 patients the confidence limits of RE are too wide to be useful. In contrast, the confidence limits of the adjusted association will generally be narrow enough to be of practical interest even with small numbers of observations. This is because the surrogate endpoint and the true endpoint are generally strongly correlated (at the

individual level). For the validation to be complete, however, a strong association between the surrogate and the true endpoint is not sufficient: the relative effect must also be estimated with good precision to permit the reliable prediction of a treatment effect on the true endpoint based on the observation of the treatment effect on the surrogate endpoint.

9.5 Validation from Multiple Trials

Buyse et al. (1999) have shown how further progress can be made by conducting the validation process within a meta-analytic framework. They developed their methodology for surrogate and true endpoints which are jointly normally distributed. They considered two distinct modelling strategies, based on a two-stage fixed effects representation on the one hand and random effects on the other hand. However, it may also be necessary to explore other settings, often more complicated due to the absence of a unifying framework such as the multivariate normal distribution. Here, we will present approaches for surrogate and true endpoints which are

- both continuous,
- both binary,
- of a mixed binary–continuous nature.

First, we extend the setting and notation by supposing we now have data from $i = 1, \ldots, N$ trials, in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled. Let T_{ij} and S_{ij} denote the true and surrogate endpoints respectively, and let Z_{ij} be an indicator variable for treatment.

9.5.1 Continuous Endpoints

In this section we focus on surrogate and true endpoints which are assumed to be jointly normally distributed. Two distinct modelling strategies will be followed, based on a two-stage fixed effects representation on the one hand and random effects on the other hand.

Let us first consider the two-stage fixed effects representation. At first, we consider the fixed-effects model

$$S_{ij}|Z_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}}, \tag{9.17}$$

$$T_{ij}|Z_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, \tag{9.18}$$

where α_i and β_i are trial-specific effects of treatment Z on the endpoints in trial i, μ_{S_i} and μ_{T_i} are trial-specific intercepts, and ε_{S_i} and ε_{T_i} are correlated error terms, assumed to be mean-zero normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}.$$

Due to the replication at the trial level, we can impose a distribution on the trial-specific parameters. At the second stage, we assume

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix} \tag{9.19}$$

where the second term on the right hand side of (9.19) is assumed to follow a zero-mean normal distribution with dispersion matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{TS} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{aS} & d_{aT} & d_{aa} & d_{ab} \\ d_{bS} & d_{bT} & d_{ba} & d_{bb} \end{pmatrix}.$$

Next, the random-effects representation is based upon combining the above two steps.

$$S_{ij}|Z_{ij} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}}, \tag{9.20}$$

$$T_{ij}|Z_{ij} = \mu_T + m_{T_i} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{T_{ij}}, \tag{9.21}$$

In the above formulation, μ_S and μ_T are now fixed intercepts, α and β are the fixed effects of treatment on the endpoints, m_{S_i} and m_{T_i} are random intercepts

and a_i and b_i are the random effects of treatment on the endpoints in trial i. The vector of random effects $\boldsymbol{b_i} = (m_{S_i}, m_{T_i}, a_i, b_i)$ is assumed to be mean-zero normally distributed with covariance matrix D. The error terms ε_{S_i} and ε_{T_i} follow the same assumptions as in fixed-effects model (9.17)–(9.18). Suppose the fixed effects are grouped in a vector $\boldsymbol{\beta}$ and the outcomes S_{ij}, T_{ij} in a $2n_i$ dimensional response vector $\boldsymbol{Y_i} = (S_{i1}, T_{i1}, \ldots, S_{in_i}, T_{in_i})$, then it can be easily shown that $\boldsymbol{Y_i}$ follows a $2n_i$ dimensional normal distribution with mean vector $X_i\boldsymbol{\beta}$ and with covariance matrix $V_i = \Sigma_i^* + Z_i D Z_i^T$ (Verbeke and Molenberghs 1997). In the above formulation, $\Sigma_i^* = I_{n_i} \otimes \Sigma$, where I_{n_i} denotes the identity matrix of dimension n_i, and X_i and Z_i are suitable design matrices for the fixed and random effects, respectively.

The settings described above naturally lend themselves to introduce surrogacy at both the trial level as well as the individual level. We will discuss them in turn.

Trial-Level Surrogacy

The key motivation for validating a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint based on the observed effect of treatment on the surrogate endpoint. It is essential, therefore, to explore the quality of the prediction of the treatment effect on the true endpoint in trial i by:

(a) information obtained in the validation process based on trials $i = 1, \ldots, N$,

(b) the estimate of the effect of Z on S in a new trial $i = 0$.

To this end, observe that $(\beta + b_0 | m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0 | m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} \quad (9.22)$$

$$\mathrm{Var}(\beta + b_0 | m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \quad (9.23)$$

This has lead Buyse et al. (1999) to call a surrogate *perfect at the trial level* if the conditional variance (9.23) is equal to zero. A measure to assess the quality of the

surrogate at the trial level is the coefficient of determination

$$R^2_{trial(f)} = R^2_{b_i|m_{Si},a_i} = \frac{1}{d_{bb}} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \tag{9.24}$$

(9.24) is unitless and ranges in the unit interval, two desirable features for its interpretation. Intuition can be gained by considering the special case where the prediction of b_0 can be done independently of the random intercept m_{S0}. Expressions (9.22) and (9.23) then reduce to

$$\begin{aligned} E(\beta + b_0|a_0) &= \beta + \frac{d_{ab}}{d_{aa}}(\alpha_0 - \alpha), \\ \text{Var}(\beta + b_0|a_0) &= d_{bb} - \frac{d^2_{ab}}{d_{aa}} \end{aligned}$$

with corresponding

$$R^2_{trial(r)} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}.$$

Now, $R^2_{trial(r)} = 1$ if the trial level treatment effects are simply multiples of each other. We will refer to this simplified version as the reduced random effects model, while the orginal expression (9.24) will be said to derive from the full random effects model.

Individual-Level Surrogacy

To validate a surrogate endpoint, Buyse and Molenberghs (1998) suggested to consider the association between the surrogate and the final endpoints after adjustment for the treatment effect. To this end, we need to construct the conditional distribution of $\tilde{T}$, given $\tilde{S}$ and Z. From (9.17)–(9.18) we derive

$$\begin{aligned} \tilde{T}_{ij}|Z_{ij},\tilde{S}_{ij} \sim\ & N\left\{\mu_{Ti} - \sigma_{TS}\sigma^{-1}_{SS}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma^{-1}_{SS}\alpha_i)Z_{ij} + \sigma_{TS}\sigma^{-1}_{SS}S_{ij}\ ; \right. \\ & \left. \sigma_{TT} - \sigma^2_{TS}\sigma^{-1}_{SS}\right\}. \end{aligned}$$

A similar expression can be found for the random effects model (9.20)–(9.21), but conditioning is then also on the random effects. The association between both endpoints after adjustment for the treatment effect is then captured by:

$$R^2_{indiv} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}},$$

the squared correlation between S and T after adjustment for both the trial effects as well as for the treatment effect.

Summary

The above developments suggest to term a surrogate *trial-level valid* if R^2_{trial} is sufficiently close to one, and to call it *individual-level valid* if R^2_{indiv} is sufficiently close to one. Finally, a surrogate is termed *valid* if it is both trial-level and individual-level valid. In order to replace the words "valid" with "perfect" the corresponding R-squared values are required to equal one.

Buyse et al. (1999) note that the validation criteria proposed here do not require the treatment to have a significant effect on either endpoint. In particular, it is possible to have $\alpha \equiv 0$ and yet have a perfect surrogate. Indeed, even though the treatment may not have any effect on the surrogate endpoint as a whole, the fluctuations around zero in individual trials (or other experimental units) can be very strongly predictive of the effect on the true endpoint. However such a situation is unlikely to occur since the heterogeneity between the trials is generally small compared to that between individual patients.

9.5.2 Binary Endpoints

Let $\tilde{S}_{ij}$ and $\tilde{T}_{ij}$ represent unobserved latent variables that are related to the actual (binary) responses S_{ij}, T_{ij} through a threshold value. Without loss of generality, we can assume this threshold value to be zero. Hence:

$$S_{ij} = \begin{cases} 1 & \text{if } \tilde{S}_{ij} > 0 \\ 0 & \text{if } \tilde{S}_{ij} < 0 \end{cases},$$

and

$$T_{ij} = \begin{cases} 1 & \text{if } \tilde{T}_{ij} > 0 \\ 0 & \text{if } \tilde{T}_{ij} < 0 \end{cases}.$$

We can then easily update the previously described two-stage fixed effects model and random effects model.

Two-stage Fixed Effects Representation

In the first step of the two-stage approach, we consider the following fixed-effects model:

$$\tilde{S}_{ij}|Z_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}}, \quad (9.25)$$

$$\tilde{T}_{ij}|Z_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, \quad (9.26)$$

with the same assumptions as in (9.17)-(9.18). At the second stage, we consider the same model as in (9.19).

random-effects Representation

Similar to Section 9.5.1, combining the above two steps gives rise to a random-effects regression model, which itself now leads to a joint probit regression model with random effects for the response probabilities on the surrogate and true endpoints:

$$\tilde{S}_{ij}|Z_{ij} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}}, \quad (9.27)$$

$$\tilde{T}_{ij}|Z_{ij} = \mu_T + m_{T_i} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{T_{ij}}, . \quad (9.28)$$

The above equations (9.27) and (9.28) can be seen from a "generalized linear mixed model" (GLMM) viewpoint (Breslow and Clayton 1993) with a probit link. In theory, any link function could be used to model both endpoints. However, we prefer probit link based approaches, the main advantage being that the measures R^2_{trial} and R^2_{indiv} to assess the quality of the surrogate generalize immediately to this setting. Parameter estimation can proceed by maximizing the likelihood function, but unfortunately, to obtain the unconditional likelihood from such model, intractable expressions need to be evaluated. The joint marginal probability for an entire trial can be obtained from:

$$\int \prod_{j=1}^{n_i} P(S_{ij} = s_{ij}, T_{ij} = t_{ij}|Z_{ij}, \boldsymbol{\beta}, \boldsymbol{b_i}) f(\boldsymbol{b_i}|D) d\boldsymbol{b_i}, \quad (9.29)$$

where $\boldsymbol{\beta}$ groups all the fixed effects. Clearly, the integration in (9.29) can hardly be accomplished, in view of the generally large size of the clusters (trials), even though an explicit expression exists when a probit formulation is adopted. Zeger, Liang and Albert (1988) presented such a closed form expression in the univariate case. One consequently needs to resort on some kind of approximation to the likelihood

function (Breslow and Clayton 1993, Wolfinger and O'Connel 1993). When applied, such procedures can however lead to substantial downward bias for the estimates of variance components.

Pseudo-likelihood Approach

A useful alternative to likelihood-based methods can rest on a pseudo-likelihood approach, where we replace the likelihood contribution $f(S_{i1}, \ldots, S_{in_i}, T_{i1}, \ldots, T_{in_i})$ of cluster i by the product of all possible pairwise contributions (see Chapter 6). Hence, the contribution of the ith cluster to the log pseudo-likelihood function can be written as:

$$p\ell_i = \sum_{j<k} \ln f(y_{ij}, y_{ik})$$

where y_{ij} and y_{ik} are taken from $S_{i1}, \ldots, S_{in_i}, T_{i1}, \ldots, T_{in_i}$. These pairwise contributions $p\ell_i$ reflect four different association types:

(i) the association between the surrogate and true endpoint for a certain individual,

(ii) the association between two surrogate endpoints for two different individuals,

(iii) the association between two true endpoints for two different individuals,

(iv) the association between a surrogate and true endpoint for two different individuals.

This is illustrated in Figure 9.1.

Each of these pairwise contributions can then be written in terms of bivariate probits. For example,the probability that both the surrogate and the true endpoint are zero can be written as:

$$\begin{aligned} Pr(S_{ij} = 0, T_{ij} = 0 | \boldsymbol{\beta}, \boldsymbol{b_i}, Z_{ij}) &= P(\tilde{S}_{ij} < 0, \tilde{T}_{ij} < 0 | \boldsymbol{\beta}, \boldsymbol{b_i}, Z_{ij}) \\ &= \Phi_2 \left(-\frac{\mu_S + \alpha Z_{ij}}{\text{Var}(\tilde{S}_{ij})}, -\frac{\mu_T + \beta Z_{ij}}{\text{Var}(\tilde{T}_{ij})}, \rho_{ij} \right). \end{aligned}$$

Similar expression (in terms of univariate and bivariate probits) can be obtained for all other pairwise probabilities. In this formulation, the variance terms, $\text{Var}(\tilde{S}_{ij})$ and $\text{Var}(\tilde{T}_{ij})$, are obtained by selecting the appropriate 2×2 submatrix of the covariance

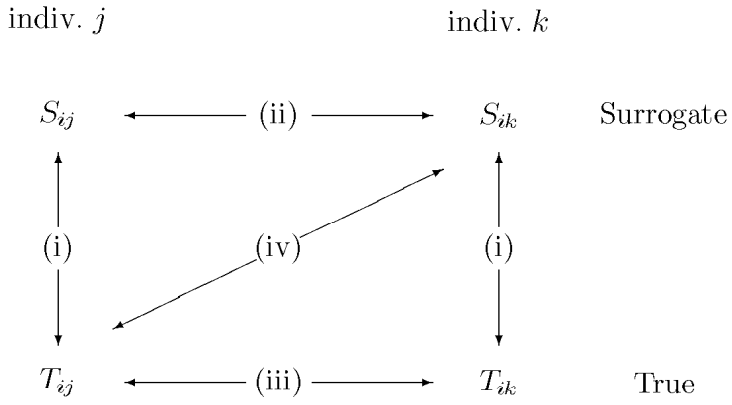

Figure 9.1: *Association Structure for the Surrogate and True Endpoints on Individuals j and k in Cluster i.*

matrix, $V_i = R_i + Z_i^T D Z_i$, where R_i is the correlation matrix of the measurement errors $(\varepsilon_{S_{i1}}, \ldots, \varepsilon_{S_{in_i}}, \varepsilon_{T_{i1}}, \ldots, \varepsilon_{T_{in_i}})$ and Z_i is a suitable design matrix for the random effects. ρ_{ij} is the correlation parameter that corresponds with this submatrix.

Estimates for $\boldsymbol{\beta}$ and the components of D can then be obtained by maximizing the log pseudo-likelihood $p\ell = \sum_{i=1}^{N} p\ell_i$. Similar to Chapter 6, if main interest lies in the main effect parameters, we might prefer maximizing:

$$p\ell^* = \sum_{i=1}^{N} p\ell_i/(2n_i - 1), \tag{9.30}$$

where (9.30) corrects for the fact that each response occurs $2n_i - 1$ times in the ith contribution to the pseudo-likelihood.

9.5.3 Mixed Binary-Continuous Outcomes

In this section we concentrate on a two-stage fixed-effects model only. Generalized linear mixed models, such as obtained from (9.27)–(9.28), have the advantage that they are a straightforward generalization of the linear model discussed in Molenberghs and Buyse (1998), however they cannot be applied on endpoints that are of

different data type. Whereas a pseudo-likelihood approach might be possible theoretically, it can become very involved in practice. Indeed, we would need to consider different contributions, based on two continuous outcomes, two binary outcomes, and a binary and continuous outcome.

Let us now describe the two-stage fixed-effects model, with a binary surrgate and a continuous true endpoint. The reverse case is entirely similar. Let $\tilde{S}_{ij}$ be a latent variable of which S_{ij} is a dichotomized version.

In the first step, we can assume the following model:

$$\begin{aligned} \tilde{S}_{ij}|Z_{ij} &= \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}}, \\ T_{ij}|Z_{ij} &= \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, \end{aligned}$$

with similar notation as before. But, the error terms, ε_{S_i} and ε_{T_i}, are assumed to be mean-zero normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \frac{1}{(1-\rho^2)} & \frac{\rho\sigma}{\sqrt{1-\rho^2}} \\ \frac{\rho\sigma}{\sqrt{1-\rho^2}} & \sigma^2 \end{pmatrix}.$$

In short, we use the probit formulation, described in Section 9.3.1. At the second stage, we again assume model (9.19). The trial-level and individual-level measures for surrogacy (R^2_{trial} and R^2_{indiv}) easily extend to this setting.

9.6 An Example in Cancer

In this section we will illustrate our methods using data from the Ovarian Cancer Meta-Analysis Project (1998), described in Section 2.4. The treatment indicator Z is set to 0 for CP and to 1 for CAP. The primary endpoint is "survival", defined as the time from randomization to death from any cause. To assess sensitivity, all analyses have been performed with and without the two smaller trials. Excluding the two smaller trials has very little impact on the estimates of interest and therefore the results reported are those obtained with all four trials.

9.6.1 Continuous Outcomes

In this subsection, we present the results obtained by Buyse et al. (1999). The surrogate endpoint S_{ij} is the logarithm of time to progression, defined as the time

(in weeks) from randomization to clinical progression of the disease or death due to the disease, while the final endpoint T_{ij} is the logarithm of survival, defined as the time (in weeks) from randomization to death from any cause. Two-stage fixed-effects models could be fitted, as well as a reduced version of the mixed-effects model (9.20)–(9.21). Figure 9.2 shows a plot of the treatment effects on the true endpoint by the treatment effects on the surrogate endpoints. The size of each point is proportional to the number of patients in the corresponding unit. Clearly, the treatment effects are highly correlated.

When the sample size of the trial are used to weight the pairs (a_i, b_i), the reduced fixed-effects model provides $R^2_{trial(r)} = 0.92$ (s.e. 0.08). The full fixed-effects model yields $R^2_{trial(f)} = 0.94$ (s.e. 0.07). In the reduced random effects model, $R^2_{trial(r)} = 0.95$ (s.e. 0.10).

At the individual level, $R^2_{indiv} = 0.89$ (s.e. 0.01) in the fixed-effects model, and $R^2_{indiv(r)} = 0.89$ (s.e. 0.01) in the reduced random effects model.

Thus, Buyse et al. (1999) conclude that time to progression is a valid surrogate for survival in advanced ovarian cancer. Hence, the effect of treatment can be observed earlier if time to progression is used instead of survival.

9.6.2 Binary Outcomes

To illustrate our methods when both surrogate and true endpoints are binary, we dichotomize the true endpoint as 1 if the patient survived the first 5 years of follow-up and 0 if it died during this period. The clinical response, dichotomized as 1 if the patient exhibits no progression during treatment and 0 otherwise, is used as surrogate endpoint. For the purpose of this analysis we omit trials with no centers and centers for which all clinical responses are unknown. Unfortunately, this drastically reduces the total number of units of analysis from 50 to 13. In all other cases, "unknown" clinical responses are considered as "progressions".

Let us first consider Table 9.5, which shows the fixed parameter estimates for the full and reduced two-stage fixed-effects models and the reduced random effects model, fitted with the GLIMMIX macro in SAS (Wolfinger and O'Connell 1993).

In the two-stage fixed-effects model, sample sizes of the experimental units are used to weigh the pairs (a_i, b_i). The reduced two-stage fixed-effects model provides $R^2_{trial(r)} = 0.09$ (s.e. 0.15) and $R^2_{indiv(r)} = 0.45$ (s.e. 0.07). In addition, the full two-

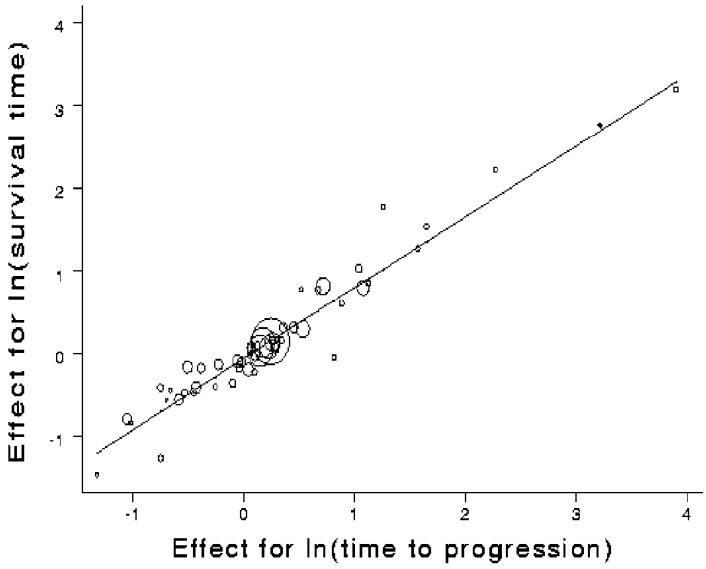

Figure 9.2: *Ovarian Cancer Trial: Treatment Effects on the True Endpoint versus Treatment Effects on the Surrogate Endpoint for all Units of Analysis. The Size of Each Point is Proportional to the Number of Patients in the Corresponding Unit (Buyse et al. 1999).*

Table 9.5: *Advanced Ovarian Cancer Trial: Parameter Estimates (standard error) for the Full and Reduced Two-stage Fixed Effects Models, as well as for the Reduced Random Effects Model.*

	2-Stage	2-Stage	Random Effects
	full	reduced	reduced
μ_S	1.33 (0.51)	0.47 (0.07)	0.47 (0.07)
α	-0.32 (0.66)	0.52 (0.36)	0.18 (0.10)
μ_T	-1.02 (0.38)	-0.80 (0.07)	-0.80 (0.07)
β	0.08 (0.13)	-0.12 (0.36)	0.17 (0.10)

stage fixed-effects model provides $R^2_{trial(f)} = 0.79$ (s.e. 0.10) and $R^2_{indiv(f)} = 0.45$ (s.e. 0.07). Based on these results, it is clear that the assessment of the clinical response is not a good surrogate for survival. It should however be noted that $R^2_{trial(f)}$ is much larger than $R^2_{trial(r)}$. This might indicate important effects for the intercepts. Individual level measures of surrogacy are comparable.

Clearly, the parameters, obtained with the two-stage fixed-effects models, are not estimated with great precision. This is especially true for the full model. Alternatively, we tried to fit full and reduced random effects models, as in (9.27)–(9.28), using the GLIMMIX macro in SAS or using the pseudo-likelihood approach. However, none of these methods could provide us with trustworthy estimates for the variance components in D. The GLIMMIX approach yielded negative variance components for the reduced model, and could not be fitted for the full model. The pseudo-likelihood approach yielded a non-positive definite information matrix in both cases. Of course, problems that occur with fitting mixed-effects models have long been recognized. Already in the case of jointly normally distributed surrogate and true endpoints, Buyse et al. (1999) observed that, in many practical instances, convergence of the Newton-Raphson algorithm could hardly be achieved. Simulation studies revealed that there should be enough variability at the "cluster" level, and a sufficient number of "clusters" to obtain convergence. When these requirements are not fulfilled, they argue that one must rely on simpler models, such as the two-stage fixed effecs models.

9.6.3 Mixed Binary-Continuous Outcomes

In addition to the analyses presented in Section 9.6.2, we can consider clinical response as a binary surrogate and log survival time as a continuous response. Analyzing these data with the full two-stage fixed-effects probit model, we find the following measurements for the trial-level and individual-level validities: $R^2_{trial(f)} = 0.65$(s.e.0.17) and $R^2_{indiv(f)} = 0.53$(s.e.0.03). The reduced two-stage fixed-effects model provides $R^2_{trial(r)} = 0.15$(s.e.0.19) and $R^2_{indiv(r)} = 0.53$(s.e.0.04). These results are in close agreement with those obtained in Section 9.6.2. The values for R^2_{indiv} are well in agreement for the reduced and full models, however $R^2_{full(r)}$ is much smaller than $R^2_{full(f)}$, indicating a potential important effect from the trial-specific intercepts. In addition, we have also investigated the potential of pathological response, dichotomized as progression or no progression, as a valid surrogate. The full model yielded only a $R^2_{trial(f)}$ of only 0.14(s.e.0.09) and a $R^2_{indiv(f)}$ of 0.26(s.e.0.03).

9.7 An Example in Ophthalmology: Revisited

The ARMD data come from a single multicentric trial. Therefore, it is natural to consider the center in which the patients were treated as the unit of analysis. A total of 36 centers were thus available for analysis, with a number of individual patients per center ranging from 2 to 18.

Table 9.6 summarizes the R^2-values of interest to validate the surrogate endpoints in each of the following subsections.

Table 9.6: *ARMD Study: R^2 Values of Interest for the Validation of a Surrogate Endpoint. See Text for Details.*

Section	$R^2_{indiv(r)}$	$R^2_{trial(r)}$	$R^2_{indiv(f)}$	$R^2_{trial(f)}$
9.7.1	0.48 (0.05)	0.69 (0.16)	.	.
9.7.2	0.64 (0.13)	0.22 (0.12)	.	.
9.7.3	0.44 (0.09)	0.42 (0.13)	0.56 (0.08)	0.36 (0.13)

9.7.1 Continuous Outcomes

Buyse et al. (1999) applied their meta-analytic methods in the case of jointly continuous surrogate (visual acuity at 6 months) and true (visual acuity at 1 year) endpoints. Irrespective of the software used, random effects were difficult to obtain and they had to restrict to the reduced two-stage fixed-effects model. Let us briefly summarize their results. At the individual level, they found $R^2_{indiv(r)} = 0.48$ (s.e. 0.05). Note that $R_{indiv(r)} = 0.69$ is close to $\rho = 0.74$, estimated in Section 9.4. At the trial-level, $R^2_{trial(r)} = 0.69$ (s.e. 0.16). Clearly, the coefficients of determination are both too low to make visual acuity at 6 months a reliable surrogate for visual acuity at 12 months. This is in contrast with the inconclusive analysis in Section 9.4.

9.7.2 Binary Outcomes

We illustrate our method for binary outcomes using dichotomized visual acuity at 6 months as the surrogate and dichotomized visual acuity at 1 year as the true endpoint. Obtaining a fit for these data was very difficult. A random effects model could not be fitted. The PL approach yielded a non positive definite information matrix. The only feasible approach was to use the reduced two-stage fixed-effects model. This resulted in $R^2_{indiv(r)} = 0.64$ (s.e. 0.13) and $R^2_{trial(r)} = 0.22$ (s.e. 0.12). The latter quantity is much smaller than in the normal case. This is probably due to the loss of information, resulting from the discretization of the continuous endpoints.

9.7.3 Mixed Binary-Continuous Outcomes

Let us now consider a situation where dichotomized visual acuity at 6 months acts as surrogate for the continuous visual acuity at 12 months. Table 9.7 shows the parameter estimates for the full and reduced two-stage fixed-effects probit model. In Section 9.4, we found similar parameter estimates for μ_T and β. Here, the full model yields $R^2_{trial(f)} = 0.36$ (s.e. 0.13) and $R^2_{indiv(f)} = 0.56$ (s.e. 0.08). The square root of the latter quantity equals 0.75 (s.e. 0.05), which is almost identical to the adjusted association $\rho_Z = 0.74$ (s.e. 0.05), estimated in Section 9.4. The reduced model yields $R^2_{indiv(r)} = 0.44$ (s.e. 0.09) and $R^2_{trial(r)} = 0.42$ (s.e. 0.13). This confirms our earlier results of a poor surrogate.

Table 9.7: *Macular Degeneration Trial: Parameter Estimates (standard errors) for the Full and Reduced Two-stage Fixed Effects Probit Model*

	Full	Reduced
μ_S	1.46 (0.68)	0.67 (0.15)
α	1.10 (0.98)	1.75 (0.69)
μ_T	11.13 (1.69)	11.82 (1.00)
β	4.40 (2.94)	3.72 (2.38)
σ	11.43 (0.60)	13.60 (0.71)
ρ	0.75 (0.05)	0.66 (0.07)

9.8 Conclusion

In this chapter, we have first extended the single-trial approach proposed in BM for the validation of surrogate endpoints when the surrogate and true endpoint are of a mixed continuous and discrete nature. In that case, a latent variable approach is a natural extension of the likelihood based approach. Such an approach discretizes one latent response variable and assumes the other one is measured directly. We have presented two approaches, one based on a probit-linear model, the other on a Plackett-Dale model.

Next, we have looked into meta-analytic extensions for cases where both endpoints are binary or of a mixed continuous-binary data type. The approach presented here is an extension of the methodology described by Buyse et al. (1999), who assumed jointly normally distributed endpoints. In contrast with the Prentice-Freedman criteria, it is not based on tests of hypothesis, but evaluates the validity of a surrogate in terms of coefficients of determination, which are intuitively appealing quantities in the unit interval. Such an approach is more informative than a mere dichotomization of surrogate endpoints as being "valid" or "invalid". Moreover, the validation procedure no longer requires statistical tests to be statistically significant: for instance, an endpoint with a low individual-level coefficient of determination ($R^2_{indiv} << 1$) is unlikely to be a good surrogate (even if $R^2_{trial} = 1$), a conclusion that may be reached with a limited number of observations.

References

Aerts, M., Declerck, L., and Molenberghs, G. (1997) "Likelihood Misspecification and Safe Dose Determination for Clustered Binary Data," *Environmetrics*, **8**, 613–627.

Altham, P.M.E. (1978) "Two Generalizations of the Binomial Distribution," *Applied Statistics*, **27**, 162-167.

Arnold, B.C., Castillo, E., and Sarabia, J.-M. (1992) *Conditionally Specified Distributions*, Lecture Notes in Statistics, **73**, New York: Springer-Verlag.

Arnold, B.C. and Strauss, D. (1991) "Pseudolikelihood Estimation : Some Examples," *Sankhya B*, **53**, 233-243.

Bahadur, R.R. (1961) "A Representation of the Joint Distribution of Responses to n Dichotomous Items," in *Studies in Item Analysis and Prediction*, H. Solomon (ed.), *Stanford Mathematical Studies in the Social Sciences VI,* Stanford, California: Stanford University Press.

Besag, J. (1975) "Statistical Analysis of Non-lattice Data," *The Statistician*, **24**, 179–195.

Boissel, J.P., Collet, J.P., Moleur, P., and Haugh, M. (1992) "Surrogate Endpoints: A Basis for a Rational Approach," *European Journal of Clinical Pharmacology*, **43**, 235–244.

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

Brown, L.D. (1986) *Fundamentals of Statistical Exponential Families.* California : Institute of Mathematical Statistics.

Brown, N.A. and Fabro, S. (1981) "Quantitation of Rat Embryonic Development in Vitro: A Morphologhical Scoring System," *Teratology*, **24**, 65–78.

Buyse, M., and Molenberghs, G. (1998) "Criteria for the Validation of Surrogate Endpoints in Randomized Experiments," *Biometrics*, **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (1999) "The Validation of Surrogate Endpoints in Meta-analyses of Randomized Experiments," submitted.

Carey, V.C., Zeger, S.L., and Diggle, P.J. (1993) "Modelling Multivariate Binary Data with Alternating Logistic Regressions," *Biometrika*, **80**, 517-526.

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," *Risk Analysis*, **14**, 611–619.

Catalano, P.J., Scharfstein, D.O., Ryan, L.M., Kimmel, C.A. and Kimmel, G.L. (1993), "Statistical Model for Fetal Death, Fetal Weight, and Malformation in Developmental Toxicity Studies," *Teratology*, **47**, 281–290.

Chen, J.J., and Kodell, R.L. (1989) "Quantitative Risk Assessment for Teratologic Effects," *Journal of the American Statistical Association*, **84**, 966–971.

Chen, J.J., Kodell, R.B., Howe, R.B. and Gaylor, D.W. (1991) "Analysis of Trinomial Responses from Reproductive and Developmental Toxicity Experiments," *Biometrics*, **47**, 1049–1058.

Choi, S., Lagakos, S., Schooley, R.T., and Volberding, P.A. (1993) "CD4+ Lymphocytes Are an Incomplete Surrogate Marker for Clinical Progression in Persons with Asymptomatic HIV Infection Taking Zidovudine," *Annals of Internal Medicine,* **118**, 674–680.

Conaway, M. (1989) "Analysis of Repeated Categorical Measurements with Conditional Likelihood Methods," *Journal of the American Statistical Association*, **84**, 53–62.

Connolly, M.A., and Liang, K.Y. (1988) "Conditional Logistic Regression Models for Correlated Binary Data," *Biometrika*, **75**, 501–506.

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.

Cressie, N. (1991) *Statistics for Spatial Data.* New York: Wiley.

Crump, K. (1984) "A New Method for Determining Allowable Daily Intakes," *Fundamental and Applied Toxicology*, **4**, 854–871.

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 Munro, I. (eds.), *Toxicological Risk Assessment. Volume I: Biological and Statistical Criteria,* Boca Raton: CRC Press, pp. 187–203.

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. (1999) "Likelihood Based Analysis of Clustered Binary Data with Applications in Developmental Toxicity Studies," unpublished Ph.D. dissertation, Limburgs Universitaire Centrum, Center for Statistics.

Declerck, L., Aerts, M. and Molenberghs, G. (1998) "Behaviour of the Likelihood Ratio Test Statistic Under A Bahadur Model For Exchangeable Binary Data," *Journal of Statistical Computation and Simulation*, **61**, 15–38.

Declerck, L., Molenberghs, G., Aerts, M. and Ryan, L. (1999) "Litter-based Methods in Developmental Toxicity Assessment," *Environmental and Ecological Statistics*, **00**, 000–000.

De Gruttola, V., Fleming, T.R., Lin, D.Y. and Coombs, R. (1997) "Validating Surrogate Markers–Are We Being Naive?," *Journal of Infectious Diseases,* **175**, 237–246.

De Gruttola, V., and Tu, X.M. (1995) "Modelling Progression of CD-4 Lymphocyte Count and Its Relationship to Survival Time," *Biometrics*, **50**, 1003–1014.

De Gruttola, V., Wulfsohn, M., Fischl, M.A., and Tsiatis, A. (1993) "Modelling the Relationship Between Survival and CD4 Lymphocytes in Patients with AIDS and AIDS-Related Complex," *Journal of Acquired Immune Deficiency Syndromes*, **6**, 359–365.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994) *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Ellenberg, S.S., and Hamilton, J.M. (1989 "Surrogate Endpoints in Clinical Trials: Cancer," *Statistics in Medicine*, **8**, 405–413.

Fears, T.R., Benichou, J. and Gail, M.H. (1996) "A Reminder of the Fallibility of the Wald Statistic," *The American Statistician*, **50**, 226–227.

Fitzmaurice, G.M., and Laird, N.M. (1993) "A Likelihood-based Method for Analysing Longitudinal Binary Responses," *Biometrika*, **80**, 141–151.

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.

Fitzmaurice, G., Laird, N., and Tosteson, T. (1996), "Polynomial Exponential Models for Clustered Binary Outcomes," unpublished manuscript.

Fitzmaurice, G.M., Molenberghs, G. and Lipsitz, S.R. (1995) "Regression Models for Longitudinal Binary Responses with Informative Drop-outs," *Journal of the Royal Statistical Society, B*, **57**, 691–704.

Fitzmaurice, G.M., Laird, N.M. and Rotnitzky, A.G. (1993) "Regression Models for Discrete Longitudinal Responses," *Statistical Science*, **8**, 284–309.

Flandre, P. and Saidi, Y. (1999) "Letters to the Editor: Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker," *Statistics in Medicine*, **18**, 107–115.

Fleming, T.R., Prentice, R.L., Pepe, M.S. and Glidden, D. (1994) "Surrogate and Auxiliary Endpoints in Clinical Trials, with Potential Applications in Cancer and AIDS research," *Statistics in Medicine*, **13**, 955–968.

Fleming, T.R. and DeMets, D.L. (1996) "Surrogate Endpoints in Clinical Trials: Are we Being Misled?," *Annals of Internal Medicine*, **125**, 605–613.

Freedman, L.S., Graubard, B.I., Schatzkin, A. (1992) "Statistical Validation of Intermediate Endpoints for Chronic Diseases," *Statistics in Medicine*, **11**, 167–178.

Gaylor, D.W. (1989) "Quantitative Risk Analysis for Quantal Reproductive and Developmental Effects," *Environmental Health Perspectives,* **79**, 243–246.

Gelman, A., and Speed, T.P. (1993) "Characterizing a Joint Probability Distribution by Conditionals," *Journal of the Royal Statistical Society, Series B*, **70**, 185–188.

George, J. D., Price, C. J., Kimmel, C. A., and Marr, M. C. (1987) "The Developmental Toxicity of Triethylene Glycol Dimethyl Ether in Mice," *Fundamental and Applied Toxicology*, **9**, 173-181.

Geyer, C.J., and Thompson, E.A. (1992) "Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion)," *Journal of the Royal Statistical Society, Series B*, 657–699.

Geys, H., Molenberghs, G., Declerck, L., and Ryan, L. (1999) "Flexible Quantitative Risk Assessment for Developmental Toxicity Based on Fractional Polynomial Predictors," submitted.

Geys, H., Molenberghs, G. and Lipsitz, S.R. (1998) "A Note on the Comparison of Pseudo-likelihood and Generalized Estimating Equations for Marginal Odds Ratio Models," *Journal of Statistical Computation and Simulation*, **62**, 45–72.

Geys, H., Molenberghs, G. and Ryan, L. (1996) "Pseudo-likelihood Estimation for Clustered Binary Data," in Forcina, A., Marchetti, G., Hatzinger, R. and Galmacci, G. (eds.), *Statistical Modelling: Proceedings of the 11th International Workshop on Statistical Modelling*, Orvieto, pp. 176–183.

Geys, H., Molenberghs, G. and Ryan, L. (1997) "Pseudo-likelihood Inference for Clustered Binary Data," *Communications in Statistics: Theory and Methods*, **26**,2743–2767.

Geys, H., Molenberghs, G. and Ryan, L. (1999) "Pseudo-likelihood Modelling of Multivariate Outcomes in Developmental Toxicology," *Journal of the American Statistical Association,* **00**,000–000.

Geys, H., Molenberghs, G. and Williams, P. (1997) "Analysis of Clustered Binary Data with Covariates Specific to Each Observation," in Minder, C. and Friedl, H. (eds.), *Good Statistical Practice: Proceedings of the 12th International Workshop on Statistical Modelling*, Wien: Schriftenreihe der Osterreichischen Statistischen Gesellschaft, pp. 170–174.

Geys, H., Molenberghs, G. and Williams, P. (1999) "Analysis of Toxicology Data with Covariates Specific to Each Observation," revised.

Geys, H., Regan, M., Catalano, P. and Molenberghs, G. (1999) "Two Latent Variable Risk Assessment Approaches for Combined Continuous and Discrete Outcomes from Developmental Toxicity Data," submitted.

Glonek, G.F.V. and McCullagh, P. (1995) "Multivariate Logistic Models," *Journal of the Royal Statistical Society B*, **57**, 533-546.

Haber, F. (1924) "Zur Geschichte des Gaskrieges" ("On the History of Gas Warfare"), in *Funf Vortrage aus den Jahren 1920-1923 (five Lectures from the years 1920-1923)*, Berlin: Springer, pp. 76–92.

Haseman, J.K. and Kupper, L.L. (1979)"Analysis of Dichotomous Response Data from Certain Toxicological Experiments," *Biometrics*, **35**, 281–293.

Hauck, K.W. and Donner, A. (1977) "Wald's test as applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association*, **72**, 851–853.

Heagerty, P. and Zeger, S. (1996) "Marginal Regression Models for Clustered Ordinal Measurements," *Journal of the American Statistical Association*, **91**, 1024–1036.

Herson, J. (1975) "Fieller's Theorem versus the Delta Method for Significance Intervals for Ratios," *Journal of Statistical Computation and Simulation,* **3**, 265–274.

Holland, P.W. and Wang, Y.J. (1987) "Dependence Function for Continuous Bivariate Densities," *Communications in Statistics-Theory and Methods*, **2**, 863-876.

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression*, New York: Wiley.

Hosmer, D.W., Hosmer, T., Lemeshow, S. and Le Cessie, S. (1997) "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model," *Statistics in Medicine*, **16**, 965–980.

Johnson, N.L., and Kotz, S. (1970) *Distributions in Statistics, Continuous Univariate Distributions,* **2**, Boston : Houghton-Mifflin.

Kimmel, G.L., Williams, P.L., Kimmel, C.A., Claggett, T.W., and Tudor, N. (1994) "The Effects of Temperature and Duration of Exposure on in Vitro Development and Response-Surface Modelling of Their Interaction," *Teratology*, **49**, 366–367.

Kimmel, C.A. and Gaylor D.W. (1988) "Issues in Qualitative and Quantitative Risk Analysis for Developmental Toxicology," *Risk Analysis*, **8**, 15–20.

Krewski, D. and Van Ryzin, J. (1981) "Dose-response Models for Quantal Response Toxicity Data," in Csorgo, M., Dawson, D., Rao, J.N.K. and Saleh, E. (eds.), *Statistics and Related Topics*, North-Holland, New York, pp. 201–231.

Kupper, L.L. and Haseman, J.K. (1978) "The Use of a Correlated Binomial Model for the Analysis of certain Toxicology Experiments," *Biometrics*, **34**, 69-76.

Kupper, L. L., Portier, C., Hogan, M. D. and Yamamoto, E. (1986) "The Impact of Litter Effects on Dose-Response Modeling in Teratology," *Biometrics*, **42**, 85–98.

Lagakos, S.W., and Hoth, D.F. (1992) "Surrogate Markers in AIDS: Where Are We? Where Are We Going?," *Annals of Internal Medicine*, **116**, 599–601.

Lang, J.B. and Agresti A. (1994) "Simultaneously Modeling of Joint and Marginal Distributions of Multivariate Categorical Responses," *Journal of the American Statistical Association*, **89**, 625-632.

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.

Lee, Y. and Nelder, J.A. (1996) "Hierarchical Generalized Linear Models," *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

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

Lefkopoulou, M., and Ryan, L. (1993) "Global Tests for Multiple Binary Outcomes," *Biometrics*, **49**, 975–988.

Lefkopoulou, M., Moore. D., and Ryan, L. (1989) "The Analysis of Multiple Correlated Binary Outcomes: Application to Rodent Teratology Experiments," *Journal of the American Statistical Association*, **84**, 810–815.

Lefkopoulou, M., Rotnitzky, A., and Ryan, L.M. (1995) "Trend Tests for Clustered Data". In: Morgan, B.T. (ed.) *Statistics in Toxicology*, Oxford University Press.

Lehmann, E.L. (1983) *Theory of Point Estimation*, Wiley: New York.

Liang, K.Y., and Self, S. (1996) "On the Asymptotic Behavior of the Pseudolikelihood Ratio Test Statistic," *Journal of the Royal Statistical Society, Series B*, **58**, 785–796.

Liang, K.Y. and Zeger,S. (1986) "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika*, **73**, 13–22.

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

Lin, D.Y., Fischl, M.A. and Schoenfeld, D.A. (1993) "Evaluating the Role of CD4-Lymphocyte Change as a Surrogate Endpoint in HIV Clinical Trials," *Statistics in Medicine*, **12**, 835–842.

Lin, D.Y., Fleming, T.R. and De Gruttola, V. (1997) "Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker," *Statistics in Medicine,* **16**, 1515–1527.

Lindstrom, P., Morrisey, R.E., George, J.D., Price, C.J., Marr, M.C., Kimmel C.A., and Schwetz, B.A. (1990) "The Developmental Toxicity of Orally Administered Theophylline in Rats and Mice," *Fundamental and Applied Toxicology*, **14**, 167–178.

Lipsitz, SR., Fitzmaurice, G.M. and Molenberghs, G. (1996) "Goodness-of-fit Tests for Ordinal Response Regression Models," *Applied Statistics*, **45**, 175-190.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991) "Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association, *Biometrika*, **78**, 153–160.

Littell, S.R., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996) *SAS system for Mixed Models*, SAS Institute Inc: Cary, NC, USA.

Machado, S.G., Gail, M.H. and Ellenberg, S.S. (1990) "On the Use of Laboratory Markers as Surrogates for Clinical Endpoints in the Evaluation of Treatment for HIV Infection," *Journal Acquired Immune Deficiency Syndromes*, **3**, 1065–1073.

McCullagh, P. (1987) *Tensor Methods in Statistics*. London New York: Chapman and Hall.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. London New York: Chapman and Hall.

McCulloch, C.E. (1997) "Maximum likelihood algorithms for generalized linear mixed models," *Journal of the American Statistical Association*, **92**, 162–170.

Molenberghs, G. (1992) "A Full Maximum Likelihood Method for the Analysis of Multivariate Ordered Categorical Data," unpublished Ph.D. dissertation, Universitaire Instelling Antwerpen, Dept. of Mathematics.

Molenberghs, G., and Buyse, M. (1999) "The Validation of Surrogate Endpoints in Randomized Experiments: Why the Proportion Explained is Misleading," submitted.

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., and Geys, H. (1998) "A Plackett-Dale Approach to Analyze Mixed Continuous and Discrete Data," in Proceedings of the XIXth International Biometric Conference, Invited Papers, pp. 31–42.

Molenberghs, G., Geys, H. and Buyse, M. (1999) "Validation of Surrogate Endpoints in Randomized Experiments with Mixed Discrete and Continuous Outcomes," submitted.

Molenberghs, G., Geys, H., Declerck, L., Claeskens, G. and Aerts, M. (1998) "Analysis of Clustered Multivariate Data from Developmental Toxicity Studies" in Payne, R. and Green, P. (eds.), *COMPSTAT: Proceeding in Computational Statistics* (keynote paper), Physica Verlag, pp. 3–14.

Molenberghs, G. and Lesaffre, E. (1994) "Marginal Modelling of Correlated Ordinal Data Using a Multivariate Plackett Distribution," *Journal of the American Statistical Association*, **89**,633–644.

Molenberghs, G. and Lesaffre, E. (1999) "Marginal Modelling of Multivariate Categorical Data," *Statistics in Medicine*, **00**, 000–000.

Molenberghs, G. and Ritter, L. (1996) "Methods for Analyzing Multivariate Binary Data, with Association Between Outcomes of Interest," *Biometrics*, **52**, 1121–1133.

Molenberghs, G., and Ryan, L.M. (1999) "An Exponential Family Model for Clustered Multivariate Binary Data," *Environmetrics*, **00**, 000–000.

Moore, D.S. and Spruill, M.C. (1975) "Unified Large-sample Theory of General Chi-squared Statistics for Tests of Fit," *Annals of Statistics*, **3**, 599-616.

Morgan, B.J.T. (1992) *Analysis of Quantal Response Data*, Chapman and Hall, London.

Neuhaus, J.M. (1992) "Statistical Methods for Longitudinal and Clustered Designs with Binary Responses," *Statistical Methods in Medical Research*, **1**, 249–273.

Neuhaus, J.M. and Jewell, N.P. (1993) "A Geometric Approach to Assess Bias due to Omitted Covariates in Generalized Linear Models," *Biometrika*, **80**, 807–815.

Neuhaus, J.M. and Lesperance, M.L. (1996) "Estimation Efficiency in a Binary Mixed-Effects Model Setting," *Biometrika*, **83**, 441–446.

Neuhaus, J.M. and Kalbfleisch, J.D. (1997) "Between- and within-cluster Covariate Effects in the Analysis of Clustered Data," submitted for publication.

Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W. (1991) "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, **59**, 25–35.

Neuhaus, J.M. and Segal, M.R. (1997) "An Assessment of Approximate Maximum Likelihood Estimators in Generalized Linear Mixed Models," in *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, Lecture Notes in Statistics, **122**, New York: Springer, pp. 11–22.

Ochi, Y. and Prentice, R.L. (1984) "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, **71**, 531–543.

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).

Ovarian Cancer Meta-Analysis Project (1991) "Cyclophosphamide Plus Cisplatin Versus Cyclophosphamide, Doxorubicin, and Cisplatin Chemotherapy of Ovarian Carcinoma: A Meta-Analysis," *Journal of Clinical Oncology,* **9**, 1668–1674.

Ovarian Cancer Meta-Analysis Project (1998) "Cyclophosphamide Plus Cisplatin Versus Cyclophosphamide, Doxorubicin, and Cisplatin Chemotherapy of Ovarian Carcinoma: A Meta-Analysis," *Classic Papers and Current Comments,* **3**, 237–243.

Pharmacological Therapy for Macular Degeneration Study Group (1997) "Interferon α-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial," *Archives of Ophthalmology*, **115**, 865–872.

Plackett, R.L. (1965) "A Class of Bivariate Distributions," *Journal of the American Statistical Association*, **60**, 516–522.

Prentice, R.L. (1988) "Correlated Binary Regression with Covariates Specific to Each Binary Observation," *Biometrics* **44**, 1033–1048.

Prentice, R.L. (1989) "Surrogate Endpoints in Clinical Trials: Definitions and Operational Criteria," *Statistics in Medicine* **8**, 431–440.

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 Dimethyl Ether in Mice," *Fundamental and Applied Toxicology*, **8**, 115-126.

Rao, C.R. (1973) *Linear Statistical Inferences and Its Applications*, New York: Wiley.

Rao, J.N.K., and Scott, A.J. (1987) "On Simple Adjustments to Chi-square Tests with Sample Survey Data," *The Annals of Statistics*, **15**, 385-397.

Regan, M.M. and Catalano, P.J. (1999a) "Likelihood Models for Clustered Binary and Continuous Outcomes: Application to Developmental Toxicology," submitted.

Regan, M.M. and Catalano, P.J. (1999b) "Bivariate Dose-Response Modeling and Risk Estimation in Developmental Toxicology," submitted.

Roberts, W.C. and Abernathy, C.O. (1996) "Risk assessment: principles and methodologies," in A. Fan and L.W. Chang (eds.), *Toxicology and Risk Assessment, Principles, M ethods and Applications,* New York: Marcel Dekker Inc. ,pp. 245–270.

Roberts, J., Rao, J.N.K., and Kumar, S. (1987) "Logistic Regression Analysis of Sample Survey Data," *Biometrika*, **74**,1-12.

Rosner, B. (1984) "Multivariate Methods in Ophtalmology with Applications to Other Paired-Data," *Biometrics*, **40**, 1025–1035.

Rotnitzky, A. and Jewell, P. (1990) "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Clustered Correlated Data," *Biometrika*, **77**, 485–97.

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.

Royston, P., and Wright, E. (1998) "A Method for Estimating Age-specific Reference Intervals ("normal ranges") based on fractional polynomials and exponential transformation," *Journal of the Royal Statistical Society, Series A,* **161**, 79–101.

Ryan, L. (1992) "Quantitative Risk Assessment for Developmental Toxicity," *Biometrics*, **48**, 163–174.

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.

Salsburg, D. (1996) "Estimating Dose-Response for Toxic Endpoints". In: Morgan, B.T. (ed.) *Statistics in Toxicology*, Oxford University Press.

Sammel, M.D., Ryan, L.M. and Legler, J.M. (1997) "Latent Variable Models for Mixed Discrete and Continuous Outcomes," *Journal of the Royal Statistical Society, Series B,* **59**, 667-678.

SAS Institute Inc., (1997) *SAS/STAT Software: Changes and Enhancements through Release 6.12*, Cary, NC: SAS Institute Inc.

Satterthwaite, F.E. (1946) "An Approximate Distribution of Estimates of Variance Components," *Biometric Bulletin*, **2**, 110–114.

Sauerbrei, W. and Royston, P. (1999) "Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials," *Journal of the Royal Statistical Society, Series A*, **162**, 71–94.

Smith, D.M., Robertson, B. and Diggle, P.J. (1996) "Object-Oriented Software for the Analysis of Longitudinal Data in S," Technical Report MA 96/192. Department of Mathematics and Statistics, University of Lancaster, LA1 4YF, United Kingdom.

Tanner, M.A. (1991) *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Berlin: Springer Verlag.

Ten Have, T.R., Landis, J.R. and Weaver, S.L. (1995) "Association Models for Periodontal Disease Progression: A Comparison of Methods for Clustered Binary Data," *Statistics in Medicine*, **14**, 413–429.

The Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989) "Preliminary Report: Effect of Encainide and Flecainide on Mortality in a Randomized Trial of Arrhythmia Suppression after Myocardial Infarction," *New England Journal of Medicine*, **321**, 406–412.

Thélot, C. (1985) "Lois logistiques à deux dimensions," *Annales de l'Insée*, **58**, 123–149.

Tsiatis, A.A. (1980) "A Note on a Goodness-of-fit Test for the Logistic Regression Model," *Biometrika*, **67**, 250-251.

Tyl, R. W., Price, M. C., Marr, M. C., and Kimmel, C. A. (1988) "Developmental Toxicity Evaluation of Dietary Di(2-ethylhexyl)phthalate 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.

Verbeke, G., and Molenberghs, G. (1997) *Lecture Notes in Statistics. Linear Mixed Models in Practice: A SAS–Oriented Approach*, New York: Springer.

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

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

Williams, P.L., Molenberghs, G. and Lipsitz, S.R (1996) "Analysis of Multiple Ordinal Outcomes in Developmental Toxicity Studies," *Journal of Agricultural, Biological, and Environmental Statistics*, **1**, 250-274.

Williams, P.L., and Ryan, L.M. (1996) "Dose-Response Models for Developmental Toxicology," in R.D. Hood (ed.), *Handbook of Developmental Toxicology,* New York: CRC Press,pp. 635–666.

Williamson, J.M., Lipsitz, S.R. and Kim, K.M. (1997) "GEECAT and GEEGOR: Computer Programs for the Analysis of Correlated Categorical Response Data," submitted.

Wittes, J., Lagakos, E. and Probstfield, J. (1989) "Surrogate Endpoints in Clinical Trials: Cardiovascular Diseases," *Statistics in Medicine*, **8**, 415–425.

Wolfinger, R. and O'Connell, M. (1993) "Generalized Linear Mixed Models: a Pseudo-likelihood Appoach," *Journal of Statistical Computations and Simulations,* **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**, 79-86.

Zeger, S.L. and Liang, K.Y. (1986) "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, **42**, 121–130.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equations Approach," *Biometrics*, **44**, 1049–1060.

Zhao, L. P. and Prentice, R. L. (1990) "Correlated Binary Regression using a Quadratic Exponential Model," *Biometrika*, **77**, 642–648.

Zhu and Fung (1996) "Statistical Methods in Developmental Toxicity Risk Assessment," in A. Fan and L.W. Chang (eds.), *Toxicology and Risk Assessment, Principles, Methods and Applications,* New York: Marcel Dekker Inc. ,pp. 413–446.

Samenvatting

Recent is er in de maatschappij een grote bezorgdheid ontstaan omtrent het effect van blootstelling aan mogelijk toxische verbindingen op de menselijke voortplanting en ontwikkeling. Vermits, om dit te onderzoeken nauwelijks of geen epidemiologische gegevens beschikbaar zijn, is men genoodzaakt toevlucht te nemen tot toxicologische studies uitgevoerd bij proefdieren, zoals muizen en ratten. Afhankelijk van het type effect dat men wenst te bestuderen zijn er een drietal standaard procedures voorhanden. Segment I studies zijn bedoeld om het effect van scheikundige stoffen op de vruchtbaarheid van zowel het mannetje als het wijfje te bestuderen. Segment II studies worden ook wel *teratologische* studies genoemd. Hier bestudeert men vooral het effect van een toxische stof op het al dan niet aanwezig zijn van verschillende types malformaties en/of laag geboortegewicht. Segment III studies bestuderen effecten later in de zwangerschap. In dit proefschrift bestuderen we vooral Segment II studies. Daarnaast komen ook de zogenaamde "heatshock" experimenten aan bod, die we later zullen beschrijven.

Een Segment II experiment bevat over het algemeen een controle groep en 3 tot 4 groepen, waarbij een 20 tot 30-tal zwangere dieren aan verschillende doses van een chemische stof worden blootgesteld. Net voor het baren worden de moederdieren gedissecteerd. Vervolgens wordt de baarmoeder grondig onderzocht. Men telt dan het aantal embryo's die nooit tot ontwikkeling zijn gekomen en terug in de baarmoeder zijn geabsorbeerd (dit kan waargenomen worden aan de hand van donkere plekken in de baarmoederwand), het aantal dode foetussen en het aantal levensvatbare foetussen. De levensvatbare foetussen worden verder onderzocht op de aanwezigheid van verschillende types malformaties en laag geboortegewicht. Deze malformaties kan men indelen in 3 klassen: (i) de externe malformaties welke met het blote oog kunnen waargenomen worden (b.v. ontbreken van ledematen), (ii) skeletmisvormingen en (iii) inwendige malformaties (b.v. aangetaste lever, lon-

gen, ...). Ieder type van misvorming wordt meestal weergegeven via een binaire variabele (aan- of afwezig).

De bedoeling van dergelijke experimenten is duidelijk. Enerzijds wenst men de relatie tussen de toegediende dosis en de respons (risico op een bepaald type afwijking, laag geboortegewicht, ...) te bestuderen. Anderzijds wil men op basis van een gepaste dosis-respons modellering een kwantitatieve risico-analyse uitvoeren, d.w.z. men wil een veilig niveau van blootstelling aan een bepaalde toxische stof schatten.

Het analyseren van de hierboven beschreven experimenten is methodologisch een grote uitdaging. Men dient immers verschillende deelaspecten in rekening te brengen. Ten eerste worden in de meeste toxicologische studies foetussen *via de moederdieren* blootgesteld aan een of andere toxische stof. Dit heeft tot gevolg dat de foetussen van eenzelfde moederdier zich gelijkaardiger gedragen dan foetussen die afkomstig zijn van verschillende moederdieren. Dit is het zogenaamde "nest-effect" of "cluster-effect" Om dergelijke gecorreleerde binaire gegevens te modelleren kan men terugvallen op verschillende klassen van modellen: (i) conditionele modellen, (ii) marginale modellen of (iii) cluster-specifieke modellen. Verder dient men ook de hiërarchische natuur van de data in rekening te brengen: (i) een blootstelling aan een chemische stof in een vroeg stadium van de zwangerschap kan leiden tot absorptie van het embryo in de baarmoederwand; (ii) eens dit stadium gepasseerd loopt de foetus alsnog het risico om te sterven in een later stadium; (iii) levensvatbare foetussen kunnen allerlei misvormingen en/of laag geboortegewicht vertonen. In dit proefschrift beperken we ons enkel tot levensvatbare foetussen. Vervolgens dient men rekening te houden met de mogelijke associaties tussen de verschillende types van misvormingen. Het is ook niet ondenkbaar dat de nestgrootte een belangrijke invloed heeft op de responskansen. In een groot nest concurreren immers een groot aantal foetussen voor de voedingsstoffen van dezelfde moeder. De kans op een misvorming in een groot nest kan daarom groter zijn dan in een kleiner nest. Ten slotte moet een statistisch model ook in staat zijn om binaire (malformaties) en continue (geboortegewicht) responsen gezamenlijk te modelleren.

In dit proefschrift laten we verschillende van deze deelaspecten aan bod komen.

In hoofdstuk 3 beschrijven we het conditionele model van Molenberghs en Ryan (1999), afgekort als MR, voor gecorreleerde binaire gegevens. Dit model is gebaseerd op een exponentiële familie en heeft dus ook alle daarmee verbonden voordelen.

Echter, een exponentieel familie model wordt gekenmerkt door een normaliseringsconstante welke in het geval van gecorreleerde (multivariate) binaire gegevens computationeel erg onaantrekkelijk kan worden. Vooral voor multivariate gegevens vergt het bepalen van deze constante enorm veel tijd. Om dit te vermijden introduceren we in hoofdstuk 3 de pseudo-likelihood schattingsmethode voor het MR model met een enkele binaire respons. Deze methode levert consistente en asymptotisch normale schatters op. Bovendien induceert ze (vooral voor multivariate responsen) een aanzienlijke tijdswinst, vermits de normaliseringsconstante met deze methode niet hoeft bepaald te worden. In ruil hiervoor boeten we een beetje aan efficiëntie in, echter het efficiëntieverlies blijkt erg klein te zijn voor realistische parameter combinaties.

In hoofdstuk 4 breiden we de pseudo-likelihood methode voor het MR model uit naar multivariate responsen. Vooral in deze situatie bewijst de pseudo-likelihood methode zijn nut. Reeds in het geval van 3 responsen blijkt de maximum likelihood methode computationeel te complex te zijn. De pseudo-likelihood methode daarentegen convergeert zeer snel. In dit hoofdstuk formuleren we eveneens enkele klassieke toetsingsgrootheden voor de pseudo-likelihood context, zoals Wald, score en pseudo-likelihood ratio test statistieken. Deze grootheden zijn eenvoudig te bepalen en hebben aantrekkelijke asymptotische verdelingen. Likelihood en pseudo-likelihood test statistieken worden in dit hoofdstuk met elkaar vergeleken via asymptotische en kleine steekproef simulaties. Hieruit blijkt dat het onderscheidingsvermogen voor de score statistieken in de pseudo-likelihood context slechts een weinig kleiner is dan in de maximum likelihood context. Voor de pseudo-likelihood ratio statistiek kunnen we twee versies construeren, respectievelijk geëvalueerd onder de nul of alternatieve hypothese. Het onderscheidingsvermogen van deze laatste kan hoger worden dan die van de klassieke likelihood ratio test statistiek, maar voor kleine steekproeven dient men dan een prijs te betalen in termen van type I fout. Immers, voor kleine steekproeven vertoont deze statistiek soms een veel te grote gesimuleerde type I fout.

Een van de doelstellingen van kwantitatieve risico-analyse ligt by het schatten van een veilige dosis, welke kan worden gedefinieerd als de dosis waarbij het extra risico op een bijwerking bovenop het achtergrond-risico, gelijk is aan een bepaalde kans, b.v. 10^{-4}. Het is bijgevolg van groot belang dat de gekozen modellen goed aanpassen aan de data. In de statistische literatuur wordt nog steeds veel aandacht besteed aan klassieke veelterm predictoren. Deze zijn echter vaak ontoereikend voor

kwantitatieve risico-analyse. In hoofdstuk 5 bestuderen we fractionele veelterm predictoren, welke een veel grotere verscheidenheid van functionele vormen kunnen bieden dan de klassieke veelterm predictoren.

In hoofdstuk 6 vergelijken we veralgemeende schattingsvergelijkingen (generalized estimating equations, GEE) met pseudo-likelihood in de context van marginale odds ratio modellen. We construeren eerst een geschikte pseudo-likelihood functie en de bijhorende schattingsvergelijkingen. Afhankelijk van het feit of de wetenschappelijke interesse meer bij de hoofdeffecten ligt dan wel bij de associatie, beschouwen we verschillende types van pseudo-likelihood. De resultaten van dit hoofdstuk suggereren het gebruik van eerste orde veralgemeende schattingsvergelijkingen wanneer de interesse voornamelijk bij hoofdeffecten ligt. Desondanks blijkt de pseudo-likelihood methode bijna even efficiënt. Wanneer echter de interesse voornamelijk bepaald wordt door de associatie parameters, stellen we voor om de pseudo-likelihood methode te gebruiken. In dat geval kan GEE1 buitengewoon inefficiënt worden. Hoewel men zich in theorie ook zou kunnen beroepen op tweede orde veralgemeende schattingsvergelijkingen (GEE2), welke lichtjes meer efficiënt zijn dan GEE1 en pseudo-likelihood, zijn deze computationeel veel minder aantrekkelijk (vooral voor grote nesten). We raden dan ook het gebruik van GEE2 niet aan.

In hoofdstuk 7 bestuderen we modellen voor toxicologische studies, waar iedere foetus gekenmerkt wordt door een eigen set van covariaten. In het bijzonder bestuderen we de zogenaamde "heatshock" studies, waar foetussen op een gegeven tijdstip uit het moederdier worden gedissecteerd en zich vervolgens verder *in vitro* ontwikkelen. Iedere foetus wordt daarna voor een bepaalde tijd in een warm waterbad gedompeld van een bepaalde temperatuur. Enkele uren later kan men dan het effect van deze behandeling op de foetus bestuderen. Dergelijke studies stellen ons in staat om de associatie tussen verschillende embryo's van hetzelfde moederdier te quantificeren in termen van zowel genetische als omgevingsfactoren. In dit hoofdstuk tonen we aan dat de hoger beschreven conditionele modellen best niet gebruikt worden voor dergelijke studies. Als alternatieven kunnen we wel gebruik maken van marginale modellen of random effect modellen. Verder introduceren we een eenvoudige aanpassingstoets voor het analyseren van gecorreleerde binaire gegevens.

In toxicologische studies treft men naast verschillende types van malformaties ook vaak het lage geboortegewicht aan als mogelijke respons. In hoofdstuk 8 stellen we twee verschillende methodes voor die het gezamenlijk modelleren mogelijk maken van

zowel binaire (malformaties) als continue (laag geboortegewicht) responsen. Eerst introduceren we een probit benadering. Hier veronderstellen we dat er voor iedere binaire respons een onderliggende continue variabele bestaat, die normaal verdeeld is. De gezamenlijke verdeling van laag geboortegewicht en malformatie kan bijgevolg beschreven worden via een multivariate normale verdeling. Een tweede benadering is de Plackett-Dale methode. Hier veronderstellen we dat de latente malformatie variabelen een Plackett verdeling volgen. In beide gevallen is het specifiëren van de volledige verdeling echter computationeel veel te complex. Voor de probit benadering maken we daarom gebruik van veralgemeende schattingsvergelijkingen; voor de Plackett-Dale benadering maken we gebruik van de pseudo-likelihood methode. Beide methodes zijn fundamenteel verschillend in de manier waarop de associatie tussen beide soorten variabelen wordt geschat. De probit benadering maakt hiervoor gebruik van een correlatiecoëfficiënt. De Plackett-Dale benadering daarentegen maakt gebruik van een odds ratio. De eerste is veeleer een maat voor "lokale" associatie, terwijl de tweede eerder een maat is voor "globale" associatie. Het is dan ook heel verrassend dat beide benaderingen toch bijna identieke resultaten geven.

Hoofdstuk 9 is gewijd aan het valideren van surrogaat responsen in klinische studies. Data afkomstig van dergelijke studies hebben een gelijkaardige structuur als in toxicologische studies. Verschillende studies (clusters) worden onafhankelijk verondersteld. Patiënten binnen eenzelfde studie zijn echter mogelijks gecorreleerd en kunnen aanleiding geven tot meerdere responsen van verschillende types (binair, continu, enz.). Hoewel in een klinische studie meestal wel meerdere respons data opgemeten worden, is men vaak slechts geïnteresseerd in 1 enkele respons (de hoofdrespons), waarop men zich vervolgens baseert om het effect van een bepaalde behandeling na te gaan. Normaliter is de hoofdrespons de klinisch meest relevante respons variabele. Maar vaak is het moeilijk of zelfs onmogelijk om die te observeren. In dat geval probeert men het effect van een behandeling te evalueren aan de hand van alternatieve responsen, zogenaamde "surrogaten". Surrogaten kan men definiëren als variabelen die gebruikt kunnen worden in de plaats van een hoofdrespons bij het evalueren van het effect van een behandeling, omdat ze vroeger, gemakkelijker of vaker opgemeten kunnen worden. Vooral het onderzoek in verschillende vormen van kanker maar ook de recente ontwikkelingen in het AIDS onderzoek hebben de interesse in surrogaat responsen nieuw leven ingeblazen. Uiteraard stelt zich de vraag of dergelijke surrogaat responsen "goed" zijn, d.w.z. of het effect van de behandeling

op de surrogaat respons een getrouwe weergave is van het effect van de behandeling op de hoofdrespons. De validatie van surrogaten is tot op heden nog steeds een controversieel onderwerp. In hoofdstuk 9 wordt dit in detail besproken.