ANALYSIS OF DOSE-RESPONSE DATA FROM DEVELOPMENTAL TOXICITY STUDIES

PANG ZHEN

NATIONAL UNIVERSITY OF SINGAPORE

2005

ANALYSIS OF DOSE-RESPONSE DATA FROM DEVELOPMENTAL TOXICITY STUDIES

PANG ZHEN

(Master of Science, Beijing University of Technology)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

This thesis would not have been possible without the support and help of many people. I would like to take this opportunity to thank them warmly.

First of all, I owe my deep gratitude to my supervisor, Prof. KUK Yung Cheung, Anthony. It has been a great privilege and pleasure to study from you so many things, which have contributed not only to my scientific research but to other parts of my life as well. I can only hope that our collaboration will keep on going in the future.

At different stages of my stay at NUS I received help from all the academic and the secretarial staff at the Department of Statistics and Applied Probability. I am really grateful to all of them.

Finally, I am greatly indebted to my parents who never failed to encourage me and to support me whenever they could.

Contents

List of Tables

List of Figures

Summary

Existing distributions for modeling fetal response data in developmental toxicology have a tendency of understating the risk of having at least one malformed fetus within a litter. As opposed to a shared probability extra-binomial model, we advocate a shared response model that allows a random number of fetuses within the same litter to share a common response. An explicit formula is given for the probability function and graphical plots suggest that it does not suffer from the problem of assigning too much probability to the event of observing no malformed fetuses. The EM algorithm can be used to estimate the model parameters. Results of a simulation show that the EM estimates are nearly unbiased and the associated confidence intervals based on the usual standard error estimates have coverage close to the nominal level. Simulation results also suggest that the shared response model estimates of the marginal malformation probabilities are robust to misspecification of the distributional form, but not so for the estimates of intralitter correlation and the litter-level probability of having at least one malformed fetus. The proposed model is fitted to a set of dose-response data. For the same dose-response

relationship, the fit based on the shared response distribution is superior to that based on the beta-binomial, and comparable to the q-power distribution (Kuk, 2004, *Applied Statistics* **53**, 369-386). An advantage of the shared response model over the q-power distribution is that it is more interpretable and can be extended more easily to the multivariate case. To illustrate this, a bivariate shared response model is fitted to fetal response data involving visceral and skeletal malformation.

While the parametric distributions in the literature can be matched to have the same marginal probability and intra-cluster correlation, they can be quite different in terms of shape and higher order quantities. A sensible alternative is to fit a saturated model (Bowman and George, 1995, *Journal of American Statistical Association* **90**, 871-879) using the EM algorithm proposed by Stefanescu and Turnbull (2003, *Biometrics* **59**, 18-24). The assumption of marginal compatibility is often made to link up the distributions for different cluster sizes so that estimation can be based on the combined data. Stefanescu and Turnbull proposed a modified trend test to test this assumption. Their test, however, fails to take into account the variability of an estimated null expectation and as a result leads to much inflated p-values. This drawback is rectified in the thesis. When the data are sparse, the probability function estimated using a saturated model can be very jagged and some kind of smoothing is needed. We extend the penalized likelihood method (Simonoff, 1983, *Annals of Statistics* **11**, 208-218) to the present case of unequal cluster sizes and implement the method using an EM type algorithm. In the presence of covariates, we propose a penalized kernel method that performs

smoothing in both the covariate and response space. The proposed methods are illustrated using several data sets and the sampling and robustness properties of the resulting estimators are evaluated by simulations.

Chapter 1

Introduction

In this chapter, we first introduce clustered binary data and some of its applications. More details are given to their application to the developmental toxicity studies. Some special features of these data are then discussed. We finally give a review of the different approaches proposed in the literature.

1.1 Clustered Binary Data and Its Applications

Clustered binary data are very common in many scientific and social studies. This generally occurs in the situation where binary data are collected in clusters. For example, clinical trials are often carried out in centers or groups of individuals. The binary responses are then collected in clusters naturally. The clustering of binary responses can also be easily found in economics, psychology, ophthalmological,

otolaryngological and periodontal studies, genetic studies, complex surveys and developmental toxicity studies. Depending on the application, a cluster could mean a litter of animals, a household of individuals, or measurements of the same type taken from different locations of the same individual. Among these applications, developmental toxicity studies have received relatively more attention. The reason may be attributed to the fact that they deal with the reproductive ability of human beings. In this thesis, our emphasis is also on this application. Therefore, we will give a detailed introduction to developmental toxicity studies.

In modern society, we are exposed to many harmful chemical compounds and other environmental hazards, all of which can cause problems related to fertility and pregnancy, birth defects, and developmental abnormalities. Therefore, regulatory agencies such as the U.S. Environmental Protection Agency (EPA) and the Food and Drug Administration (FDA) are charged with the responsibility of protecting the public from drugs, chemical and other environmental exposures that may contribute to these risks.

For ethical reasons, we cannot deliberately expose human beings to some specific chemical compounds to measure the risk. Moreover, these chemical compounds in nature sometimes cannot be measured precisely. These difficulties make it necessary to find an alternative source of evidence essential for identifying potential developmental toxicants. Laboratory experiments in small mammalian species can be controlled strictly and the results can be extrapolated to humans. Therefore, a

series of developmental toxicity experiments developed quickly in the last several decades.

In a typical developmental toxicity study, pregnant laboratory animals are randomly assigned to receive a toxin at varying dose levels during the period of major organogenesis. These animals are then sacrificed prior to term and the uterus is removed and examined for resorptions, fetal deaths and fetal malformations, resulting in clustered binary or multinomial data. The aim of such a study is to assess the relationship between exposure to the toxic substance and the incidence of developmental problems. Another important task is risk assessment and the determination of an acceptable low-risk or safe dose level (Crump, 1984; Chen and Kodell, 1989; Ryan, 1992).

1.2 Special Features of Clustered Binary Data

One of the classical hypotheses of the modelling of the binary data is the independence between observations. However, this hypothesis is generally not valid for clustered binary data. The objects in the same cluster generally share some common characteristics. For example, in developmental toxicity studies, due to the genetic similarity and the same treatment conditions, fetuses within the same litter tend to behave more similarly than those from different litters. This has been termed *litter effect*. As a consequence, littermates are likely to be dependent.

Therefore, one distinguishing feature of clustered binary data is that responses in the same cluster are correlated. This introduces one more source of variation besides the variation assuming independence. This extra-binomial variation is often called *over-dispersion*. Failure to account for *litter effect* and the *over-dispersion* it induces will lead to estimates with overstated precision in the analysis of clustered binary data.

Another natural assumption of clustered binary data is exchangeability. This implies that each objective within a cluster has the same marginal probability and the associations of any order are also constant within the same cluster. We have known that for independent binomial modelling, the distribution is totally determined by the marginal probability. For many parametric models accounting for *over-dispersion*, the distributions are determined by marginal probability and intra-litter association parameter. The nonparametric procedure by George and Bowman (1995) models all orders of associations.

Exchangeability assumption makes it sufficient to report only the cluster sums rather than the individual binary responses within clusters. For example, in developmental toxicity studies, what is recorded is the number of malformed fetuses within a litter.

1.3 Different Approaches

The analysis of correlated binary data is less well developed than the case of correlated continuous data because a truly satisfactory multivariate discrete distribution with as many nice properties as the multivariate normal distribution is yet to be found. The different approaches proposed in the literature include the quasi-likelihood method, GEE, a whole host of parametric models and the nonparametric model. We will give a brief introduction to these approaches in this section. More details can be found in subsequent chapters.

1.3.1 Quasi-likelihood and GEE

The main idea behind quasi-likelihood method (Wedderburn, 1974) is to avoid a fully specified distribution for the response variable when one is uncertain about the random mechanism by which the data were generated. Liang and Hanfelt (1994) recommended quasi-likelihood with a common intra-litter correlation parameter be used in the analysis of clustered binary data when the number of litters is small or modest.

The generalized estimating equations (GEE) method is related to the quasi-likelihood method in that no parametric assumptions need be made. It was first proposed by Zeger and Liang (1986) and Liang and Zeger (1986). They only made the first order assumption and the approach is often referred to as GEE1.

It was then extended by incorporating second order assumptions (Liang, Zeger, and Qaqish, 1992). This resulted in the GEE2 method. Bowman, Chen and George (1995) used GEE to model jointly the mean parameters and the intra-litter correlation coefficients as functions of dose levels.

A limitation of quasi-likelihood and GEE is that they cannot be used in a litter-based approach to quantitative risk assessment. As pointed out by Faustman *et al.* (1994) and Geys, Molenberghs, and Ryan (1999), it is important from a biological perspective to take into account the health of the entire litter. Under the so-called litter-based approach to quantitative risk assessment, a litter is said to be affected if at least one fetus is adversely affected within a litter. Since quasi-likelihood and GEE typically model only up to the first two moments, they cannot estimate the risk that at least one litter-mate is affected.

As we are interested in assessing litter-based risk and these two methods can not do this for us, we will emphasize models that can fully determine the distribution of the fetal response data in this thesis. Some important parametric distributions will be introduced in the following section.

1.3.2 Parametric Models

A popular distribution in the analysis of clustered binary data is the beta-binomial distribution (Williams, 1975; Haseman and Kupper, 1979), under which the binomial parameter p follows a beta distribution. Another model proposed by Conaway

(1990) assumes that $\ln(-\ln(p))$ follows a log gamma distribution. Other distributions that have been proposed include the correlated binomial distribution with additive or multiplicative interactions (Kupper and Haseman, 1978; Altham, 1978), the folded-logistic model (George and Bowman, 1995), and the extended folded-logistic model (Kuk, 2004). Kuk (2004) also advocated a q-power distribution which is particularly well suited for a litter-based approach to quantitative risk assessment.

1.3.3 Nonparametric Model

Bowman and George (1995) proposed a saturated model for clustered binary data. We also call this saturated model the nonparametric model (even though the number of parameters in the saturated model is still finite). Xu and Prorok (2003) pointed out that in the case of varying cluster sizes, the maximum likelihood estimators (MLE) derived by Bowman and George (1995) are actually not the MLEs as claimed. Xu and Prorok then worked out what the MLEs should be and gave a detailed analysis when the maximum cluster size is two. However, even for this simple situation, there are five different scenarios and one of them still requires solution of a nonlinear equation. They recommended using "uniroot" in S+ to solve it numerically. For the general case, they recommend using the Newton-Raphson method. Taking advantage of the statistical structure of this problem, Stefanescu and Turnbull (2003) derived an EM algorithm for fitting the saturated model to

exchangeable binary data by augmenting the data to make the cluster sizes equal. This EM algorithm appears to be stable.

1.4 Aim and Organization of the Thesis

In this thesis, we propose a shared response model to analyze clustered binary data parametrically. A generalization to the bivariate case is then studied. The marginal compatibility assumption is very important for exchangeable binary data, we rectify the modified trend test by Stefanescu and Turnbull (2003). Due to the sparseness of the data, the saturated model by Bowman and George (1995) can exhibit a lot of roughness, we extend the penalized likelihood method (Simonoff, 1983, *Annals of Statistics* **11**, 208-218) to the present case of unequal cluster sizes and implement the method using an EM type algorithm. In the presence of covariates, we propose a penalized kernel method that performs smoothing in both the covariate and response space.

In chapter 2, we advocate a distribution first suggested by Lunn and Davies (1998) and interpret the resulting model as a shared response model. The emphasis of Lunn and Davies was to propose a method for generating exchangeable binary random variables. We work out explicitly the probability function for the number of affected fetuses within a litter as well as explore the shape of this probability function. The shared response model provides a very good fit to a real data set

and the results of a simulation study conducted to look into the bias of the maximum likelihood estimators of the shared response model, the bias of the standard error estimates and the coverage of the resulting confidence intervals are also provided. The effect of model misspecification is investigated too. We then consider dose-response modelling for both the marginal fetal response probability and the intra-litter association parameter. We derive an EM algorithm to be used to obtain maximum likelihood estimates of the model parameters. The shared response model was used to analyze a set of 2,4,5-trichlorophenoxiacetic acid data and to estimate the safe dose. Comparison is made with alternative analyses based on the beta-binomial and q-power distributions. In this chapter, we also generalize the beta-binomial and shared response model to the bivariate case. It should be noted that the method is not confined to the bivariate case. Both of these two models can be generalized to higher dimensions in similar manner. Some properties of these two bivariate models are proved. The methods are illustrated by fitting a real data set.

In chapter 3, we first give a detailed introduction of the saturated model by Bowman and George (1995) and the EM algorithm by Stefanescu and Turnbull (2003). We give a new proof of the formula that links up litters with different litter size via hypergeometric thinning. Not only is the new proof simpler and more intuitive than the existing one based on induction, hypergeometric sampling also provides us with a simple way to generate litter data with unequal litter sizes. By fitting the saturated model, we can test the goodness of fit of any parametric

model via the likelihood ratio test. This is illustrated by 6 real datasets. We conduct a simulation to illustrate the robustness of the distribution free property of the saturated model estimates and in contrast show the lack of robustness of the parametric estimates. Another simulation designed to study the behaviour of the estimates when the marginal compatibility assumption is violated suggests that the saturated model maximum likelihood estimates are somewhat robust to moderate departure from the marginal compatibility assumption. We also give a new nonparametric estimator of the intra-cluster parameter ρ based on the saturated model. A simulation study shows that this new nonparametric estimator is on par with the best estimators in the literature. Finally, we rectify the modified trend test by Stefanescu and Turnbull (2003). The p-value of our new test statistic is quite close to the bootstrap results.

In chapter 4, we find that the MLE of the saturated model can display a lot of jaggedness when the data are sparse. We extend the penalized likelihood method by Simonoff (1983) to the present case of unequal cluster sizes and implement the method using an EM type algorithm. The sampling properties of estimators are evaluated by a simulation study. The results show that penalized likelihood can reduce the variability considerably.

In chapter 5, we first use the kernel weighted saturated model to analyze the dose-response data from developmental toxicity studies. Data from different dose groups are linked by the kernel weight. In this way, we smooth our data in the

covariate space. A fit to the real data sets shows that the estimates of the marginal fetal response probability and intra-litter correlation obtained using the kernel method are fairly smooth functions of the dose level. However, the same fit reveals that the estimated probability functions are all very erratic and are in need of smoothing. Thus we finally smooth our estimates in the response space as well as across covariates by combining kernel smoothing with the penalty approach.

In chapter 6, we give the summary and conclusion of the thesis. Some possible directions of further research are also discussed.

Chapter 2

Shared Response Model

In this chapter, we first give a detailed literature review of the existing parametric models. Based on Lunn and Davies' (1998) method to generate exchangeable binary random variables, we derive the explicit form of the probability function and interpret the resulting model as a shared response model. Some basic properties of the model are then studied. We derive an EM algorithm to get the maximum likelihood estimates and apply the model to the risk assessment of the developmental toxicity studies. At the end of this chapter, we generalize the beta-binomial and shared response model to the bivariate case and prove some properties of these two bivariate models.

2.1 Introduction to Existing Models

A common way to account for the *litter effect* and extra-binomial variation in clustered binary data is to assume that the intra-litter correlation is induced by a random effect shared by all the fetuses within the same litter. Given this litter specific random effect, the outcomes of the litter-mates are assumed to be conditionally independent. The use of a beta distribution to model this random effect results in the famous beta-binomial distribution (Williams, 1975; Haseman and Kupper, 1979). Chen and Kodell (1989) used the beta-binomial distribution to model data from teratology studies.

Another model proposed by Conaway (1990) assumes that $\ln(-\ln(p))$ follows a log gamma distribution. This is essentially a random effect model with a log-gamma latent distribution and a log-log link function instead of the commonly used logistic function.

The above two models induced the positive intra-litter correlation indirectly via a shared random effect. Kupper and Haseman (1978) and Altham (1978) developed correlated binomial distribution by directly assuming that the interactions are additive. Altham (1978) also proposed a multiplicative generalization of the binomial distribution by assuming that the interactions are multiplicative. This gives rise to a two-parameter exponential family.

George and Bowman (1995) proposed a folded-logistic model. However, the

folded-logistic model does not have additional parameters to model the correlation structure. Kuk (2004) gave an extended folded-logistic model that allows more flexibility in the value of the intra-litter correlation.

The beta-binomial distribution has dominated much of the statistical literature of clustered binary data for many years. However, it has its limitations. As pointed out by George and Bowman (1995) and Kuk (2004), the shape of a beta-binomial probability function is often U-shaped, J-shaped or reverse J-shaped rather than unimodal with mode near the expected value $\mu = np$. Therefore, it could happen that most of the probability mass is assigned to the two ends 0 and n, whereas the supposedly "expected" value $\mu = np$ does not have much of the probability mass and become highly improbable. When this is applied to the litter-based quantitative risk assessment (Faustman *et al.*, 1994 and Geys, Molenberghs, and Ryan, 1999), the probability that no fetus within a litter is affected will tend to be over-estimated. As a consequence, the risk that at least one fetus is affected within a litter is often under-estimated under the beta-binomial model. Kuk (2004) demonstrated that U-shaped probability function is a common occurrence for other distributions as well and proposed a q-power distribution that is not prone to under-estimating the risk that at least one fetus is affected within a litter. The q-power distribution is particularly well suited for a litter-based approach to quantitative risk assessment. Specifically, the risk of observing at least one adverse response within a litter takes on a simple form under this distribution and can be reduced further to a generalized linear model if a complementary log-log link function is

used. However, the q-power distribution with parameters $q = 1 - p$ and γ for the number of affected fetuses S within a litter of size n, given by

$$\mathrm{P}(S = s) = \binom{n}{s} \sum_{i=0}^{s} (-1)^i \binom{s}{i} q^{(n-s+i)^\gamma}$$

is just a mathematical construction based on the theory of completely monotone functions and is not readily interpretable. Furthermore, it is not clear how the q-power distribution can be extended to model multiple types of malformation.

In the next section, we will propose a distribution for exchangeable binary data that has the same desirable property as the q-power distribution of not exaggerating the probability that no fetus is affected, but yet is more interpretable and can be extended more easily to the multivariate case. We advocate a distribution first suggested by Lunn and Davies (1998). The emphasis of Lunn and Davies was to propose a method for generating exchangeable binary random variables. As a result, they did not work out the probability function explicitly, nor have they considered dose-response modelling, estimation of parameters, or risk assessment; problems that we will deal with in this chapter.

2.2 Shared Response Model

In this section, we will first introduce Lunn and Davies's method and interpret the resulting model as a shared response model. We also work out explicitly the probability function for the number of affected fetuses within a litter as well as explore

the shape of this probability function. It is demonstrated that the shared response model provides a very good fit to a real data set and the results of a simulation study conducted to look into the bias of the maximum likelihood estimators of the shared response model, the bias of the standard error estimates and the coverage of the resulting confidence intervals are also provided. The effect of model misspecification is also investigated. Finally, we consider dose-response modelling for both the marginal fetal response probability and the intra-litter association parameter. We show how the EM algorithm can be used to obtain maximum likelihood estimates of the model parameters. The shared response model was used to analyze a set of 2,4,5-trichlorophenoxiacetic acid data and to estimate the safe dose. Comparison is made with alternative analyses based on the beta-binomial and q-power distributions.

2.2.1 Derivation of the Shared Response Distribution

Lunn and Davies (1998) proposed the following simple method to generate exchangeable binary random variables $X_1, X_2, \ldots, X_n$. Let $Y_1, Y_2, \ldots, Y_n$ be independently distributed as Bernoulli(p). Additionally, Z is also a Bernoulli(p) random variable independent of the Y's. Each X_j independently equals to Y_j with probability $1 - \pi$ and to Z with probability π. In other words,

$$X_j = (1 - U_j)\, Y_j + U_j\, Z \tag{2.1}$$

where $U_1, U_2, \ldots, U_n$ are distributed as Bernoulli(π) independently of one another and from $Y_1, Y_2, \ldots, Y_n$ and Z.

We call this a shared response model for the following reason. Unlike the standard beta-binomial or other extra-binomial models where fetuses within the same litter share the same random probability p, it is the response Z that is shared by a random subset of the fetuses. This model is more interpretable than the q-power distribution because we can attribute the shared response to the combined effect of all factors, both genetic and environmental, shared by the litter-mates. Obviously, the fact that some of the X's may actually share the same Z with certain probability induces a positive correlation between them. It is straightforward to show that $\mathrm{P}(X_j = 1) = p$, $\mathrm{Var}(X_j) = p\,(1-p)$ and the pairwise correlation between $X_1, X_2, \ldots, X_n$ is given by $\rho = \pi^2$.

Let $S = X_1 + X_2 + \cdots + X_n$ be the number of affected fetuses within a litter of size n, and $T = U_1 + U_2 + \cdots + U_n \sim \mathrm{Bin}(n, \pi)$ the number of fetuses sharing Z, the probability function of S is given by

$$\begin{aligned}
\mathrm{P}(S=s) &= \mathrm{P}(Z=0)\mathrm{P}(S=s \mid Z=0) + \mathrm{P}(Z=1)\mathrm{P}(S=s \mid Z=1) \\
&= (1-p)\sum_{t=0}^{n-s} \mathrm{P}(T=t)\mathrm{P}(S=s \mid T=t, Z=0) \\
&\quad + p\sum_{t=0}^{s} \mathrm{P}(T=t)\mathrm{P}(S=s \mid T=t, Z=1) \\
&= (1-p)\sum_{t=0}^{n-s} \binom{n}{t}\pi^t(1-\pi)^{n-t}\binom{n-t}{s}p^s(1-p)^{n-t-s} \\
&\quad + p\sum_{t=0}^{s} \binom{n}{t}\pi^t(1-\pi)^{n-t}\binom{n-t}{s-t}p^{s-t}(1-p)^{n-s} \qquad (2.2)
\end{aligned}$$

In particular,

$$\mathrm{P}(S=0) = (1-p)\sum_{t=0}^{n}\binom{n}{t}\pi^t(1-\pi)^{n-t}(1-p)^{n-t} + p(1-\pi)^n(1-p)^n \qquad (2.3)$$

and $\mathrm{P}(S \geq 1) = 1 - \mathrm{P}(S=0)$ is the risk that at least one fetus is adversely affected within a litter.

2.2.2 Comparison with Other Distributions

The probability function of S under the shared response model is plotted for $p = .1, .2$ and $\rho = .1, .15, .2$ in Figure 2.1. These are typical values in toxicological experiments. Also shown are the probability functions under the beta-binomial, Conaway's log gamma random effects and the q-power models with the same marginal probability and pairwise correlation. It can be seen that the probability functions for Conaway's and beta-binomial models are almost identical. The probability of observing no adversely affected fetuses is much larger under these two distributions than the other two distributions. The probabilities of zero response are comparable under the shared response model and the q-power distribution. Between the two, the shared response model has the advantage of being more interpretable as the q-power distribution is just a mathematical construction.

We compare next the fits provided by the four distributions to a real data set, the E1 data (Brooks *et al.*, 1997) for the numbers of dead fetuses in litters of mice from untreated experimental animals. The maximum likelihood estimates of the

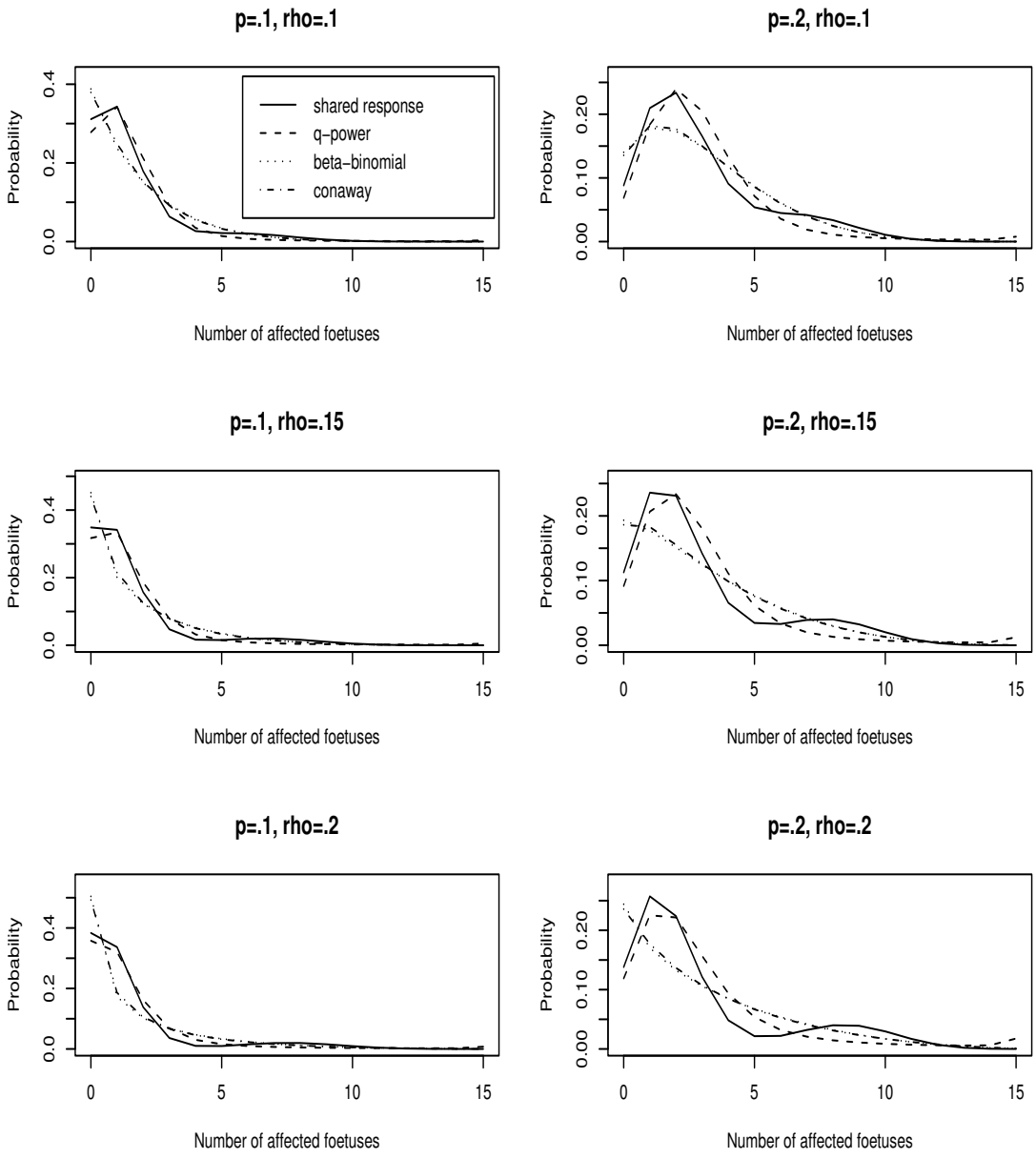


Figure 2.1: *A comparison of the probability function for litter size 15 under the shared response, q-power, beta-binomial and Conaway's model*

Table 2.1: *Comparing the fits of four distributions to the E1 data*

				affected fetuses		affected litters	
Model	$\hat{p}$	$\hat{\rho}$	Log-lik	Obs.	Exp.	Obs.	Exp.
beta-binomial	.0896	.0666	-282.65	211	211.12	115	111.29
Conaway	.0893	.0688	-282.04	211	210.25	115	111.26
q-power	.0935	.1449	-282.59	211	220.11	115	116.30
Shared response	.0898	.0820	-278.53	211	211.57	115	116.22

marginal response probability p and intra-litter correlation ρ under the four models are given in Table 2.1, together with the maximized log-likelihood, as well as the observed and expected numbers of affected fetuses and litters. Recall that a litter is said to be affected if at least one of the fetuses in the litter is affected and so the expected number of affected litters is

$$\sum_{n=1}^{17} m_n \mathrm{P}(S \geq 1 \mid n; \hat{p}, \hat{\rho}) = \sum_{n=1}^{17} m_n \{1 - \mathrm{P}(S = 0 \mid n; \hat{p}, \hat{\rho})\},$$

where m_n is the number of litters of size n in the E1 data set, and $\mathrm{P}(S = 0 \mid n; \hat{p}, \hat{\rho})$ is the probability of observing no dead fetuses in a litter of size n under the respective model evaluated at the maximum likelihood estimates of p and ρ. The maximum likelihood estimates for the shared response model are obtained by the EM algorithm, which will be described in detail later in the more general setting of dose-response modelling. It can be seen from Table 2.1 that the shared response model provides the best fit to the E1 data in terms of the likelihood value as well as matching the expected numbers of affected fetuses and litters to that actually observed. As expected, the beta-binomial distribution and Conaway's model give similar fits and both under-estimate the number of affected litters because they

assign too much probability to zero. The q-power distribution fits the number of affected litters well, but at the expense of over-estimating the number of affected fetuses. The shared response model does well in both.

2.2.3 Simulation Results

To look into the bias of the maximum likelihood estimators of the shared response model, the bias of the standard error estimates and the coverage of the resulting confidence intervals, a simulation study is conducted. We consider the cases of L=50, 100 and 200 litters, with litter sizes generated according to the distribution given in Table 6.5 of Aerts *et al.* (2002). For each combination of $p = .1, .15, .2$ and $\rho = .1, .2$. L litters of data are generated according to the shared response model. For each set of data, the maximum likelihood estimates $\hat{p}$ and $\hat{\rho}$ of p and ρ are computed together with the estimated standard errors $\widehat{SE}_p$ and $\widehat{SE}_\rho$, which are obtained by inverting the observed information matrix (Louis, 1982). This is replicated 200 times. Table 2.2 reports the bias of $\hat{p}$ and $\hat{\rho}$, the averages of $\widehat{SE}_p$ and $\widehat{SE}_\rho$, as well as the coverage of the confidence intervals $\hat{p} \pm 1.96\,\widehat{SE}_p$ and $\hat{\rho} \pm 1.96\,\widehat{SE}_\rho$. Assuming asymptotic normality, the nominal coverage should be 0.95. From Table 2.2, we can see that the estimated bias of $\hat{p}$ and $\hat{\rho}$ are quite small relative to their standard errors $\widehat{SE}_p/\sqrt{200}$ and $\widehat{SE}_\rho/\sqrt{200}$. We can also see that the bias tends to decrease as the number of litters increases, particularly for $\hat{\rho}$. The estimated standard errors of $\hat{p}$ and $\hat{\rho}$ obtained from Louis's formula appear to

Table 2.2: *Bias of maximum likelihood estimators under shared response model and coverage of confidence intervals*

	L=50		L=100		L=200	
	$\hat{p}$	$\hat{\rho}$	$\hat{p}$	$\hat{\rho}$	$\hat{p}$	$\hat{\rho}$
True values	.10	.10	.10	.10	.10	.10
Bias	-.00022	-.00247	.00078	-.00262	-.00032	-.00018
SE	.01607	.05000	.01331	.03261	.00795	.02292
Ave($\widehat{SE}$)	.01686	.04420	.01189	.03114	.00855	.02319
Coverage	.945	.895	.915	.920	.945	.930
True values	.10	.20	.10	.20	.10	.20
Bias	-.00086	-.00339	.00086	-.00301	-.00036	.00047
SE	.01852	.06742	.01489	.04700	.00917	.03196
Ave($\widehat{SE}$)	.01912	.06131	.01345	.04295	.00961	.03136
Coverage	.930	.930	.935	.930	.945	.955
True values	.15	.10	.15	.10	.15	.10
Bias	-.00083	-.00416	.00039	-.00085	-.00018	.00031
SE	.02070	.03890	.01463	.02758	.01012	.01964
Ave($\widehat{SE}$)	.02021	.03793	.01431	.02675	.01028	.01970
Coverage	.925	.910	.940	.910	.970	.945
True values	.15	.20	.15	.20	.15	.20
Bias	-.00159	-.00498	.00056	-.00162	-.00050	.00048
SE	.02351	.05704	.01691	.03800	.01127	.02768
Ave($\widehat{SE}$)	.02284	.05098	.01617	.03571	.01154	.02595
Coverage	.950	.900	.955	.925	.955	.905
True values	.20	.10	.20	.10	.20	.10
Bias	-.00076	-.00384	.00037	-.00134	-.00006	.00091
SE	.02286	.03359	.01620	.02484	.01147	.01662
Ave($\widehat{SE}$)	.02277	.03409	.01612	.02407	.01159	.01761
Coverage	.950	.925	.960	.940	.965	.955
True values	.20	.20	.20	.20	.20	.20
Bias	-.00140	-.00348	.00124	-.00012	-.00012	.00146
SE	.02526	.04328	.01938	.03426	.01301	.02255
Ave($\widehat{SE}$)	.02598	.04540	.01835	.03166	.01306	.02279
Coverage	.965	.945	.950	.935	.960	.945

do well and the resulting confidence intervals for p have reasonable coverage. The confidence interval for ρ slightly undercovers for the case of L=50 litters but the coverage improves as L increases.

It is also interesting to look at the performance of the maximum likelihood estimates obtained under the assumption of a shared response model when in fact the data are generated from another distribution. To facilitate this, we simulate data from the beta-binomial and q-power distribution using the same six configurations for p and ρ as in Table 2.2 and L=100 . In addition to p and ρ, we also estimate the probability that at least one fetus is affected, $\mathrm{P}(S \geq 1)$, for a litter of size 15. Regardless of which model we used to generate the data, the estimates are obtained by assuming a shared response model. In particular, $\mathrm{P}(S \geq 1) = 1 - \mathrm{P}(S = 0)$ is estimated by substituting the maximum likelihood estimates of p and ρ into (2.3). The results based on 200 replications are shown in Table 2.3. It can be seen that the bias in estimating p is quite small even though the data are generated from the beta-binomial and q-power distribution rather than the assumed shared response model. The bias in estimating p is typically no more than 5% of the true value when $\rho = .1$, and around 10% when $\rho = .2$. As for the estimation of ρ, Table 2.3 shows that there is a negative estimation bias, and the bias is more severe when the true distribution is beta-binomial. This is consistent with Figure 2.1, which shows that for the same values of p and ρ, the shared response model is closer to the q-power than the beta-binomial distribution. We consider finally the estimation of $\mathrm{P}(S \geq 1)$. Generally speaking, $\mathrm{P}(S \geq 1)$ increases with p just as we expected.

Table 2.3: *Bias of maximum likelihood estimators for shared response model under model misspecification*

True model	q-power			beta-binomial		
	$\hat{p}$	$\hat{\rho}$	$\hat{\mathrm{P}}(S \geq 1)$	$\hat{p}$	$\hat{\rho}$	$\hat{\mathrm{P}}(S \geq 1)$
True values	.10	.10	.7221	.10	.10	.6120
Bias	.00503	-.01457	-.0031	-.00622	-.02725	.0726
SE	.01910	.06658	.0371	.01214	.02298	.0477
True values	.10	.20	.6422	.10	.20	.4954
Bias	.00971	-.02074	.0184	-.01423	-.06295	.1033
SE	.02370	.09177	.0443	.01423	.03474	.0577
True values	.15	.10	.8597	.15	.10	.7645
Bias	.00633	-.01407	-.0059	-.00406	-.01932	.0703
SE	.02135	.05193	.0252	.01368	.02100	.0314
True values	.15	.20	.7917	.15	.20	.6471
Bias	.01533	-.01689	.0162	-.01602	-.05238	.1106
SE	.02871	.07548	.0318	.01898	.03595	.0490
True values	.20	.10	.9314	.20	.10	.8597
Bias	.00783	-.01365	-.0053	-.00300	-.01681	.0552
SE	.02350	.04333	.0147	.01691	.02152	.0195
True values	.20	.20	.8811	.20	.20	.7560
Bias	.02077	-.01364	.0118	-.01231	-.04184	.1065
SE	.03270	.06436	.0206	.02116	.03239	.0328

It decreases with ρ as a result of $\mathrm{P}(S = 0)$ increasing with ρ when the responses of litter-mates become more and more similar. Since $\mathrm{P}(S \geq 1)$ is a higher order probability that depends on the distributional form in addition to p and ρ, the estimation of $\mathrm{P}(S \geq 1)$ is expected to be model-sensitive. A clue is given in Figure 2.1, which shows that when p and ρ are matched, the shared response and the q-power probability functions pretty much start at the same $\mathrm{P}(S = 0)$, whereas the corresponding beta-binomial distribution typically has a much larger $\mathrm{P}(S = 0)$, and hence smaller $\mathrm{P}(S \geq 1)$. This explains why the shared response model tends to over-estimate $\mathrm{P}(S \geq 1)$ when the true model is actually the beta-binomial distribution, but there is not much bias if the data are generated from the q-power distribution.

2.2.4 Dose Response Modelling and EM Algorithm

In a developmental toxicity study, there are typically a control group and 3 or 4 dose groups, with 20 to 30 litters in each. The observed data are $n_i, s_i, d_i\ (i = 1, \ldots, m)$, where n_i is the number of fetuses in litter i, s_i the number of affected fetuses in litter i, d_i the dose level, and m the total number of litters. A typical dose response model specifies how the marginal fetal response probability p and the intra-litter association parameter ψ, which could be the pairwise correlation or odds ratio, depend on the dose level d. A popular choice is the generalized linear relationships $g(p) = \beta_0 + \beta_1 d$ and $h(\psi) = \alpha_0 + \alpha_1 d$, where $g(.)$ and $h(.)$ are appropriately chosen

link functions. As far as estimation via the EM algorithm is concerned, we do not need to confine ourselves to generalized linear relationships. We can assume more generally that

$$p = p(d; \beta) \tag{2.4}$$

and $\psi = \psi(d; \alpha)$ are arbitrary parametric functions of dose. To fit the shared response model (2.1), which is parameterized in terms of p and π, we also need to express π as a function of dose. Since $\rho = \pi^2$ under the shared response model, we have $\pi = \sqrt{\psi(d; \alpha)}$ if $\psi = \rho$ is the pairwise correlation. If $\psi(d; \alpha)$ is the pairwise odds ratio, then

$$\pi = \pi(d; \alpha, \beta) \tag{2.5}$$

will depend on β as well as α, because the pairwise correlation ρ, and hence also π, is a function of both the marginal response probability and the odds ratio. In what follows, we will assume the more general functional form (2.5).

We now describe how the EM algorithm can be used to obtain the maximum likelihood estimates of the shared response model given by (2.2), (2.4) and (2.5), based on the observed data $n_i, s_i, d_i \, (i = 1, \ldots, m)$. To apply the EM, which is an algorithm for obtaining maximum likelihood estimates based on the observed "incomplete" data, we define the "complete" data as $n_i, s_i, d_i, z_i, t_i \, (i = 1, \ldots, m)$, where z_i is the value of the unobserved Z in (2.1) for litter i, and $t_i = U_{i1} + \cdots + U_{in_i}$ is the number of fetuses in litter i that share the response z_i. The fact that t_i fetuses share the same response z_i means that there must be $s_i - t_i z_i$ 1's among the

remaining $n_i - t_i$ fetuses that do not share z_i. It follows that the "complete data" likelihood is simply a product of binomial likelihoods and hence the "complete data" log-likelihood is

$$\begin{aligned}\ell_c &= \sum_{i=1}^{m}\{t_i \log(\pi_i) + (n_i - t_i)\log(1-\pi_i) + z_i \log(p_i) + (1-z_i)\log(1-p_i) \\ &\quad +(s_i - t_i z_i)\log(p_i) + (n_i - t_i - s_i + t_i z_i)\log(1-p_i)\}\end{aligned}$$

where $p_i = p(d_i;\beta)$ and $\pi_i = \pi(d_i;\alpha,\beta)$.

The E-step of the EM algorithm involves taking conditional expectation of the "complete data" log-likelihood given the observed data $\mathrm{D} = \{n_i, s_i, d_i\,(i = 1,\ldots,m)\}$ to get

$$\begin{aligned}\mathrm{E}(\ell_c \mid \mathrm{D}) &= \sum_{i=1}^{m}\Big[\mathrm{E}(t_i \mid \mathrm{D})\log(\pi_i) + \{(n_i - \mathrm{E}(t_i \mid \mathrm{D})\}\log(1-\pi_i) \\ &\quad +\{\mathrm{E}(z_i \mid \mathrm{D}) + s_i - \mathrm{E}(t_i z_i \mid \mathrm{D})\}\log(p_i) + \{1 - \mathrm{E}(z_i \mid \mathrm{D})\}\log(1-p_i) \\ &\quad +\{n_i - \mathrm{E}(t_i \mid \mathrm{D}) - s_i + \mathrm{E}(t_i z_i \mid \mathrm{D})\}\log(1-p_i)\Big]\end{aligned}$$

All the conditional expectations $\mathrm{E}(t_i \mid \mathrm{D})$, $\mathrm{E}(z_i \mid \mathrm{D})$ and $\mathrm{E}(t_i z_i \mid \mathrm{D})$ that appear in $\mathrm{E}(\ell_c \mid \mathrm{D})$ are evaluated at the current parameter estimates $\hat{\alpha}, \hat{\beta}$ and can be computed using the conditional probabilities

$$\begin{aligned}&\mathrm{P}(Z_i = z_i, T_i = t_i \mid \mathrm{D}) = \mathrm{P}(Z_i = z_i, T_i = t_i \mid S_i = s_i) \\ &= \frac{p_i^{z_i}(1-p_i)^{(1-z_i)}\binom{n_i}{t_i}\pi^{t_i}(1-\pi_i)^{(n_i-t_i)}\binom{n_i-t_i}{s_i-t_i z_i}p_i^{(s_i-t_i z_i)}(1-p_i)^{(n_i-t_i-s_i+t_i z_i)}}{\mathrm{P}(S_i = s_i)}\end{aligned}$$

if $0 \le s_i - t_i z_i \le n_i - t_i$ and zero otherwise. These conditional probabilities are evaluated at the current estimates $p_i = p_i(d_i;\hat{\beta})$ and $\pi_i = \pi_i(d_i;\hat{\alpha},\hat{\beta})$. Note that

the denominator of the above expression is just $\mathrm{P}(S_i = s_i)$ given by (2.2) and the constraint $0 \leq s_i - t_i z_i \leq n_i - t_i$ has already been incorporated.

At the M-step of the EM algorithm, the imputed log-likelihood $\mathrm{E}(\ell_c \mid \mathrm{D})$ is maximized to update the values of $\hat{\alpha}$ and $\hat{\beta}$. This can be implemented using the Newton-Raphson algorithm. Beginning with a set of initial estimates, the EM algorithm is iterated until convergence is reached. Standard errors can be computed by inverting the observed information matrix which is obtained by subtracting the "missing information" from the "complete information" (Louis, 1982).

We now consider quantitative risk assessment. Let $r(d)$ be a suitably chosen function that relates the risk of observing an adverse effect, such as death, resorption or malformation, to the exposure level of a toxic substance. In a litter-based approach, interest is focused on $\mathrm{P}(S \geq 1) = 1 - \mathrm{P}(S = 0)$, the probability that at least one fetus is affected, where $\mathrm{P}(S = 0)$ for a litter of size n is given by (2.3) under the shared response model with parameters p and π. Since p and π are parametric functions of the dose level d according to (2.4) and (2.5), so is $\mathrm{P}(S \geq 1) = \mathrm{P}(S \geq 1 \mid d, n; \alpha, \beta)$. Because $\mathrm{P}(S \geq 1)$ depends on the litter size n in addition to the exposure level d, it is customary to weight $\mathrm{P}(S \geq 1 \mid d, n; \alpha, \beta)$ according to the empirical relative frequency $f(n)$ of the litter sizes across all dose groups. Thus a suitable risk function is

$$r(d; \alpha, \beta) = \sum_{n=1}^{\infty} f(n) \mathrm{P}(S \geq 1 \mid d, n; \alpha, \beta).$$

Consider the excess risk over background,

$$\begin{aligned} r^*(d;\alpha,\beta) &= r(d;\alpha,\beta) - r(0;\alpha,\beta) \\ &= \sum_{n=1}^{\infty} f(n)\{\mathrm{P}(S \geq 1 \mid d,n;\alpha,\beta) - \mathrm{P}(S \geq 1 \mid 0,n;\alpha,\beta)\}. \quad (2.6) \end{aligned}$$

Crump (1984) defined the benchmark dose, BMD_ε as the dose level that produces an excess risk of ε. Typical choices of ε are .0001, .001, .01, and .05, depending on how big an excess risk is regarded as tolerable. In this chapter, we use $\varepsilon =$ .01. A point estimate $\widehat{\mathrm{BMD}}_\varepsilon$ of the benchmark dose is obtained by solving the equation $\hat{r}^*(d) = \varepsilon$, where $\hat{r}^*(d) = r^*(d;\hat{\alpha},\hat{\beta})$ is the estimated excess risk function that results from replacing the parameters α and β in $r^*(d;\alpha,\beta)$ by the estimates $\hat{\alpha}$ and $\hat{\beta}$. In the presence of sampling uncertainty, it is more meaningful to construct a 95% lower confidence limit for BMD_ε than to calculate just a point estimate. The conventional lower confidence limit based on asymptotic normality is given by $\widehat{\mathrm{BMD}}_\varepsilon - 1.645\left\{\widehat{\mathrm{Var}}(\widehat{\mathrm{BMD}}_\varepsilon)\right\}^{1/2}$, where $\widehat{\mathrm{Var}}(\widehat{\mathrm{BMD}}_\varepsilon)$ is the estimated variance of $\widehat{\mathrm{BMD}}_\varepsilon$. A drawback of this approach is that it might yield unstable (Catalano, Ryan, and Scharfstein, 1994) as well as negative estimates. Kimmel and Gaylor (1988) proposed an alternative way to obtain a lower confidence limit for BMD_ε via test inversion. To be specific, the confidence interval consists of all those dose levels d such that the hypothesis $H : r^*(d) = \varepsilon$ is not rejected in favour of the one-sided alternative $H_a : r^*(d) < \varepsilon$ at level 0.05. A little algebra shows that the resulting 95% confidence interval for the benchmark dose BMD_ε consists of all those d such that

$$\hat{r}_U(d) = r^*(d;\hat{\alpha},\hat{\beta}) + 1.645\left[\widehat{\mathrm{Var}}\{r^*(d;\hat{\alpha},\hat{\beta})\}\right]^{1/2} \geq \varepsilon \quad (2.7)$$

where $\widehat{\text{Var}}\{r^*(d;\hat{\alpha},\hat{\beta})\}$ can be obtained from the variance-covariance matrix of α and β using delta method. Solving $\hat{r}_U(d) = \varepsilon$ for d leads to the so-called lower effective dose LED_ε. Since $\hat{r}_U(d)$ is the 95% upper confidence limit for the excess risk $r^*(d)$, (2.7) tells us that a 95% confidence interval for the benchmark dose BMD_ε can be obtained by taking all those dose levels d such that the 95% upper confidence limit for $r^*(d)$ covers ε. A graphical illustration of this is given in Kuk (2003).

2.2.5 Analysis of the 2,4,5-T Data

In a study conducted at the U.S. National Center for Toxicological Research, pregnant mice from several strains were given daily doses of the herbicide 2,4,5-T from day 6 to day 14 of gestation. For each female mouse, the number of implantation sites, fetal deaths, resorptions and cleft palate malformations were recorded. Further details of this study can be found in Holson *et al.* (1991). In keeping with most published analyzes of the data set, we consider only data obtained from the out-bred strain CD-1 and use a combined endpoint of death, resorption or malformation. For this strain, there were six dose groups corresponding to exposure levels of 0, 30, 45, 60, 75 and 90 mg/kg of 2,4,5-T. A listing of the data can be found in George and Bowman (1995). As noted by Dominici and Parmigiani (2001), this data set is quite hard to model due to the presence of zero inflation, n-inflation, over-dispersion and large kurtosis. Furthermore, the extent of departure from the

binomial model varies significantly with dose.

Kuk (2004) gave a new analysis of the 2,4,5-T data based on the q-power distribution which resulted in superior fit when compared with other distributions such as the beta-binomial, fold-logistic (George and Bowman, 1995), as well as the generalized estimating equations approach (Bowman *et al.*, 1995) used previously to analyze this set of data. It was shown (Kuk, 2004, Figure. 3) that a reasonable dose-response relationship is

$$\log\{-\log(1-p)\} = \begin{cases} \beta_{00} & \text{if } d = 0 \\ \beta_0 + \beta_1 d & \text{if } d \geq 30 \end{cases} \tag{2.8}$$

so that p is linear on the complementary log-log scale for all dose groups used in the study except for the control group. One might notice that the complementary log-log link is also the natural link function for Conaway's model with log-gamma random effects and claim that the latter model is simpler and analytically more tractable. Note, however, that in Conaway's model, it is the conditional probabilities given the random effects that are linear on the log-log scale rather than the marginal probabilities. Furthermore, we have seen from Figure 2.1 and Table 2.1 that Conaway's model behaves very much like the beta-binomial model. For these reasons, we will not fit Conaway's model to the 2,4,5-T data. As for the dose-response modelling of the association parameter, we refer again to Figure 3 of Kuk (2004) which indicates that the log odds ratio is approximately linear in the dose level, hence

$$\log(\psi) = \alpha_0 + \alpha_1 d. \tag{2.9}$$

Table 2.4: *Generalized estimating equations estimates of the response probabilities and intra-litter correlations under dose-response relationships (2.8) and (2.9) for the 2,4,5-T data.*

			number of affected fetuses	
Dose group	$\hat{p}$	$\hat{\rho}$	Observed	Expected
Control	.0731	.0137	59	58.65
30 mg/kg	.1374	.2279	124	130.84
45 mg/kg	.2696	.4346	338	303.57
60 mg/kg	.4870	.5870	383	392.55
75 mg/kg	.7579	.6385	372	365.32
90 mg/kg	.9509	.5277	242	241.53

In order to compare the beta-binomial, the q-power, and the shared response model on equal footing, we obtain parameter estimates assuming only the dose-response relationships (2.8) and (2.9) without making further distributional assumptions. This is done using the method of generalized estimating equations proposed by Lipsitz, Laird, and Harrington (1991) specifically for the case where odds ratio is used as the measure of association. The estimates obtained in this way do not favor any particular distribution. The estimates are $\hat{\beta}_{00} = -2.578\,(.127)$, $\hat{\beta}_0 = -3.419\,(.206)$, $\hat{\beta}_1 = 0.0502\,(.00364)$, $\hat{\alpha}_0 = 0.189\,(.641)$ and $\hat{\alpha}_1 = 0.0417\,(.0151)$. Table 2.4 displays the estimates of the fetal response probability p, the intra-litter correlation ρ, as well as the expected number of affected fetuses for various dose groups. Since the same estimates of p are used for all distributions, the expected numbers of affected fetuses remain the same. When it comes to estimating the number of affected litters, however, the difference in distributional assumptions begins to show because the probability that a litter is affected is a "union" probability that cannot

Table 2.5: *Estimated number of affected litters for the 2,4,5-T data.*

Dose group	Observed	beta-binomial	q-power	shared response
Control	40	38.80	39.94	39.21
30 mg/kg	56	45.42	55.10	52.93
45 mg/kg	80	61.58	72.44	73.17
60 mg/kg	69	57.69	62.54	64.31
75 mg/kg	42	40.30	41.41	42.13
90 mg/kg	24	24.80	24.87	24.90

be determined from the first two moments alone. Table 2.5 displays the expected number of affected litters for each dose group under the various models. Just like the case of the E1 data, the beta-binomial distribution underestimates the numbers of affected litters because it assigns too much probability to zero. The q-power and shared response models give better and comparable estimates of the number of affected litters.

We now turn to the determination of safe dose. A complication is that (2.8) does not really give the dose-response relationship in the range $0 \leq d < 30$. While the data seem to indicate that p is linear on the complementary log-log scale for $d \geq 30$, we have no data to tell us how far we can extend the linear relationship to the range between 0 and 30. A plausible solution is to assume that $\log\{-\log(1-p)\}$ is piecewise linear with a changepoint somewhere between 0 and 30. Figure 2.2 demonstrates the effect of altering the changepoint from 30 (solid line) to points less than 30 (broken lines). By comparing the slope of the solid line with that of the broken lines, it can be seen that p increases most rapidly with dose in the

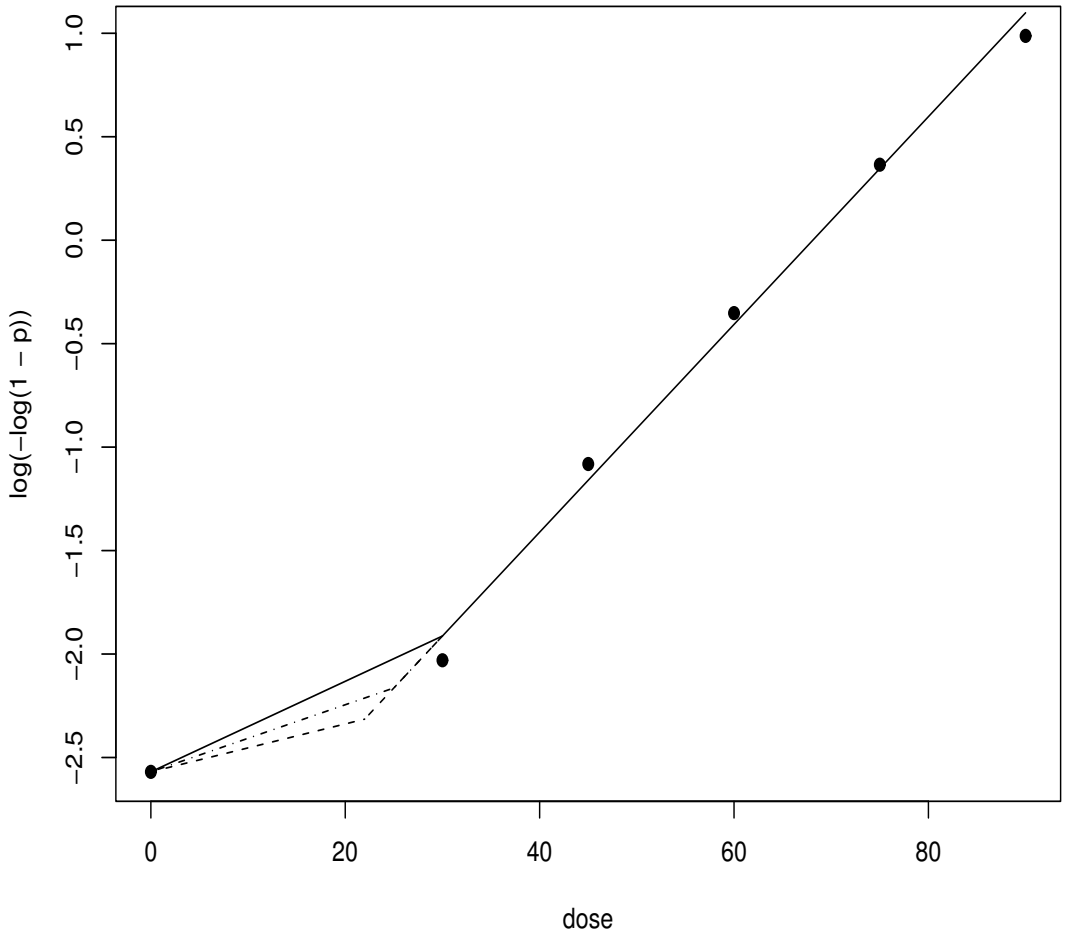


Figure 2.2: *Group-specific GEE estimates in filled circles and piecewise linear GEE fits of the fetal response probabilities on the complementary log-log scale with different changepoints for the 2,4,5-T data*

Table 2.6: *Litter-based determination of benchmark and lower effective dose in mg/kg from the 2,4,5-T data*

	q-power	shared response	beta-binomial
$\text{BMD}_{.01}$	1.86	2.82	N.A.
$\text{LED}_{.01}$	1.25	1.34	N.A.

neighborhood of zero when the changepoint is set at $d = 30$, and this should lead to conservative estimate of the safe dose. Hence we will set the changepoint at 30 to result in the following piecewise linear relationship

$$\log\{-\log(1-p)\} = \begin{cases} \beta_{00} + \beta_{01} d & \text{if } d < 30 \\ \beta_0 + \beta_1 d & \text{if } d \geq 30 \end{cases} \tag{2.10}$$

where

$$\beta_{01} = \frac{\beta_0 + 30\beta_1 - \beta_{00}}{30}$$

is the slope of the line connecting the point $(0, \beta_{00})$ to $(30, \beta_0 + 30\beta_1)$ in the first segment. Note that β_{01} is a function of $\beta = (\beta_{00}, \beta_0, \beta_1)$ rather than a free parameter to ensure continuity of the two line segments. With (2.10) in place of (2.4), and (2.9) in place of (2.5), the procedure for finding benchmark and lower effective dose described at the end of last section can be applied. The procedure remains applicable if we assume another distribution other than the shared response model. The only difference is that the form of $\text{P}(S \geq 1)$ is changed and hence $\hat{r}^*(d) = r^*(d; \hat{\alpha}, \hat{\beta})$ given by (2.6) is another function of $\hat{\alpha}$ and $\hat{\beta}$. Table 2.6 shows the litter-based benchmark dose and lower effective dose for $\varepsilon = .01$ estimated from the 2,4,5-T data using the q-power and shared response models when

the dose-response relationships (2.9) and (2.10) are assumed. The LED estimates given by the two distributions are quite comparable. It is interesting to note that estimates of $\text{BMD}_{.01}$ and $\text{LED}_{.01}$ are not available under the beta-binomial model, as can be seen from Figure 2.3, where the estimated excess risk is seen to increase to a maximum of around .008 at $d = 10$ without ever reaching .01 and then begin to decrease. In fact, the litter-based risk at $d = 23$ onwards is even lower than the baseline risk. This counter intuitive result can be explained by the following. As the dose level increases, the fetus response probability increases, but so is the intra-litter correlation, see Table 2.4. As noted before, the beta-binomial distribution has the tendency of inflating the probability of zero, given by

$$P(S_n = 0) = \frac{\prod_{r=0}^{n-1}(1 - p + r\theta)}{\prod_{r=0}^{n-1}(1 + r\theta)}$$

when the intra-litter correlation $\rho = \theta/(1 + \theta)$ is large. Thus the effect of the increase in p in reducing $\text{P}(S = 0)$ under the beta-binomial model is offset by the increase in ρ so that eventually $\text{P}(S = 0|d) > \text{P}(s = 0|0)$ and hence the litter-based excess risk

$$\begin{aligned} r^*(d) &= \sum_{n=1}^{\infty} f(n)\{\text{P}(S \geq 1 \mid d, n) - \text{P}(S \geq 1 \mid 0, n)\} \\ &= \sum_{n=1}^{\infty} f(n)\{\text{P}(S = 0 \mid 0, n) - \text{P}(S = 0 \mid d, n)\}. \end{aligned}$$

becomes negative as shown in Fig 2.3.

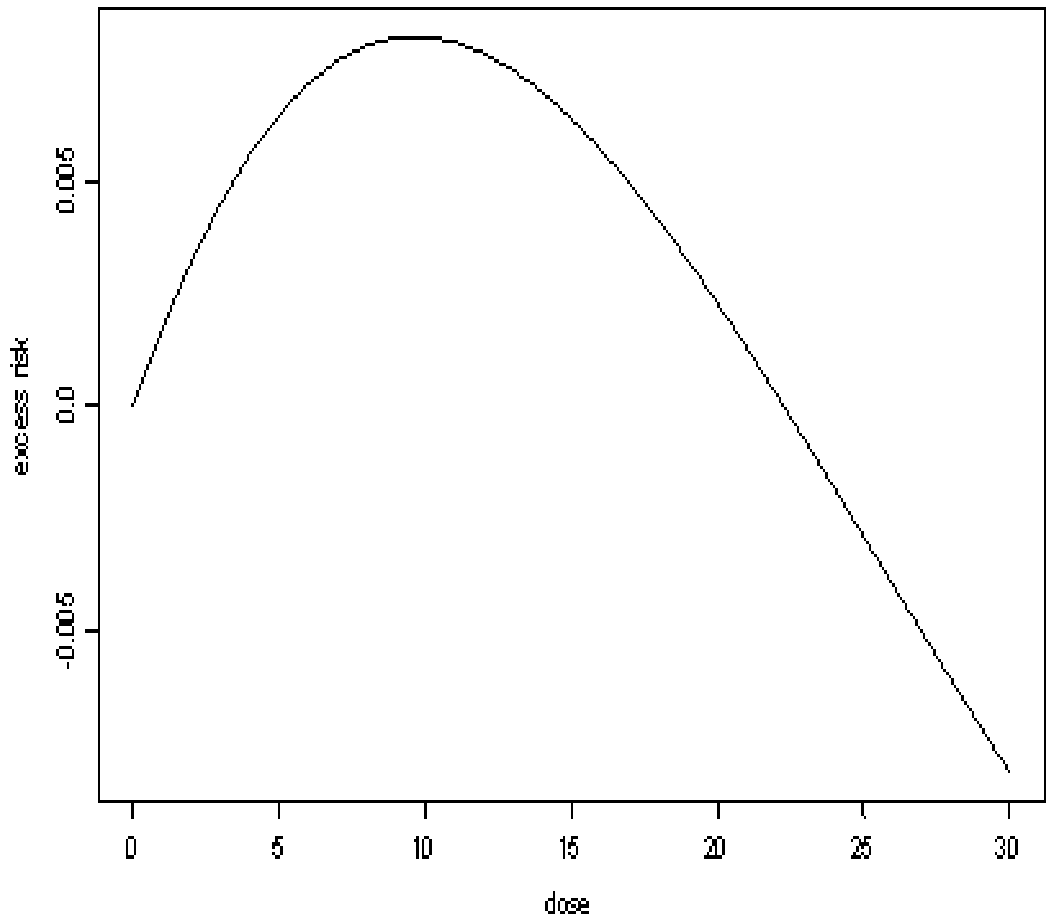


Figure 2.3: *Estimated litter-based excess risk under the beta-binomial model for the 2,4,5-T data*

2.3 Bivariate Models

Until now, we have proposed methods to model the univariate clustered binary data. However, as we have mentioned, clustered multinomial data are also very common in developmental toxicity studies. In this section, we will generalize the Beta-binomial and shared response models to model the bivariate clustered binary data. It should be noted that the the methods proposed here are not confined to the bivariate case. The bivariate beta-binomial model is just a special case of the Dirichlet-multinomial distribution and the shared response distribution can also be easily generalized to higher dimensions in similar manner.

2.3.1 Bivariate Beta-binomial Model

Recall that beta-binomial distribution assumes that marginal probability follows a beta distribution. A natural idea is to find a multivariate analogue of the beta distribution to generalize the beta-binomial distribution to the multivariate case. Mosimann (1962) did this generalization.

In the multinomial distribution, given by

$$\begin{aligned} &m(v_1, \ldots, v_{k-1}; n, \pi_1, \ldots, \pi_{k-1}) \\ &= \binom{n}{v_1, \ldots, v_{k-1}, n - \sum_{i=1}^{k-1} v_i} \pi_1^{v_1} \cdots \pi_{k-1}^{v_{k-1}} (1 - \sum_{i=1}^{k-1} \pi_i)^{n - \sum_{i=1}^{k-1} v_i} \end{aligned}$$

Mosimann assumed that the probabilities $\pi_1, \ldots, \pi_{k-1}$ are now positive random variables and the distribution function is given by

$$b(\pi_1, \ldots, \pi_{k-1}; l_1, \ldots, l_k) = \frac{\Gamma(\sum_{i=1}^{k} l_i)}{\prod_{i=1}^{k} \Gamma(l_i)} \pi_1^{l_1 - 1} \cdots \pi_{k-1}^{l_{k-1} - 1} (1 - \sum_{i=1}^{k-1} \pi_i)^{l_k - 1}$$

where all l's are constants greater than 0. Mosimann called this distribution the multivariate β-distribution, which is also known as the Dirichlet distribution. The Dirichlet-Multinomial distribution is then given by

$$\begin{aligned}
& f(v_1, \ldots, v_{k-1}; n, l_1, \ldots, l_k) \\
&= \int \cdots \int m(v_1, \ldots, v_{k-1}; n, \pi_1, \ldots, \pi_{k-1}) b(\pi_1, \ldots, \pi_{k-1}; l_1, \ldots, l_k) \mathrm{d}\pi_1 \cdots \mathrm{d}\pi_{k-1} \\
&= \binom{n}{v_1, \ldots, v_{k-1}, n - \sum_{i=1}^{k-1} v_i} \frac{\left\{\Gamma(\sum_{i=1}^{k} l_i)\right\} \left\{\prod_{i=1}^{k-1} \Gamma(v_i + l_i)\right\} \left\{\Gamma(n - \sum_{i=1}^{k-1} v_i + l_k)\right\}}{\left\{\prod_{i=1}^{k} \Gamma(l_i)\right\} \left\{\Gamma(n + \sum_{i=1}^{k} l_i)\right\}}
\end{aligned}$$

An alternative parametrization is to let

$$p_i = \frac{l_i}{\sum_{j=1}^{k} l_j}, i = 1, \ldots, k-1 \quad \text{and} \quad \theta = \frac{1}{\sum_{j=1}^{k} l_j}.$$

The Dirichlet-Multinomial distribution can be written as

$$\begin{aligned}
& f(v_1, \ldots, v_{k-1}; n, p_1, \ldots, p_{k-1}, \theta) \\
&= \binom{n}{v_1, \ldots, v_{k-1}, n - \sum_{i=1}^{k-1} v_i} \frac{\left[\prod_{i=1}^{k-1} \left\{\prod_{r=0}^{v_i - 1} (p_i + r\theta)\right\}\right] \left\{\prod_{r=0}^{n - \sum_{i=1}^{k-1} v_i - 1} (1 - \sum_{i=1}^{k-1} p_i + r\theta)\right\}}{\prod_{r=0}^{n-1} (1 + r\theta)}
\end{aligned}$$

Let $\rho = \frac{\theta}{\theta+1}$. It is easy to see that

$$\begin{cases} \mathrm{E}(V_i) = np_i \\ \mathrm{Var}(V_i) = np_i(1-p_i)\{1+(n-1)\rho\} \\ \mathrm{Cov}(V_i, V_j) = -np_ip_j\{1+(n-1)\rho\} \end{cases}$$

It is easy to show that the marginal distribution of a Dirichlet-Multinomial distribution is a beta-binomial distribution. An even more general result is that $W_m = \sum_{i=1}^{m} V_i$, for any $m = 1, \ldots, k$, follows a beta-binomial distribution too. This is because $\sum_{i=1}^{m} \pi_i$ follows a beta distribution and the conditional distribution of $\sum_{i=1}^{m} V_i$ given $\sum_{i=1}^{m} \pi_i$ follows a binomial distribution.

Now let's focus on the bivariate case. Let $n_{00}, n_{01}, n_{10}, n_{11}$ be the number of fetuses within a litter of size n that are non-affected, type II malformed only, type I malformed only and affected by both, and $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ be their respective probabilities. If $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})^T$ follows a Dirichlet distribution, then $(n_{00}, n_{01}, n_{10}, n_{11})^T$ follows a Dirichlet-Multinomial distribution. Let $S_1 = n_{10} + n_{11}$ be the number of fetuses that are affected by type I malformation. Assuming $E(\pi_{ij}) = p_{ij}$ for $i, j = 0, 1$, $p_1 = p_{10} + p_{11}$ is the probability that a fetus is affected by type I malformation and S_1 follows a beta-binomial distribution with parameter p_1. Similarly, $S_2 = n_{01} + n_{11}$ follows a beta-binomial distribution with parameter $p_2 = p_{01} + p_{11}$ and so we have a bivariate beta-binomial model. To take into account the bivariate nature of the problem, a more meaningful set of parameters are (p_1, p_2, ψ, ρ), where $\psi = \dfrac{p_{11}p_{00}}{p_{01}p_{10}}$ is the odds ratio, which is preferred to the cor-

Table 2.7: *Estimated number of affected litters for the DEHP data by malformation type based on bivariate beta-binomial model.*

	Visceral		Skeletal		Either	
Dose group	Observed	Expected	Observed	Expected	Observed	Expected
Control	4	N.A.	2	N.A.	6	N.A.
44 mg/kg	1	N.A.	1	N.A.	2	N.A.
91 mg/kg	11	10.45	6	5.90	15	13.58
191 mg/kg	11	10.66	13	11.76	15	13.99
292 mg/kg	9	8.27	7	8.18	9	8.91

relation because it is not constrained by the values of the marginal probability p_1 and p_2. In the presence of a covariate x, one may consider, for example, a logistic regression model

$$\begin{cases} logit(p_1) = \alpha_1 + \beta_1 x \\ logit(p_2) = \alpha_2 + \beta_2 x \\ log(\psi) = \alpha_3 + \beta_3 x \\ log(\frac{1+\rho}{1-\rho}) = \alpha_4 + \beta_4 x \end{cases}$$

for the dose-response data.

Finally, we use a real data set to illustrate the bivariate beta-binomial model. We use the DEHP data set listed in Table 3 of Lefkopoulou and Ryan (1993). At each dose level of di(2-ethylhexyl)-phthalate (DEHP), the table presents the numbers of fetuses that were found to have various combinations of three malformation types. Here we consider two malformation types: visceral and skeletal. At each dose level, we fit a bivariate beta-binomial model using the *vglm* command in the VGAM package, which is described in Yee and Wild (1996). Table 2.7 reports the

observed and expected numbers of litters affected with at least one viscerally malformed fetus, at least one skeletally affected fetus, and at least one malformed fetus of either types. It can be seen that the bivariate beta-binomial model provides a satisfactory fit to the observed number of affected litters for the high dose group. As for the control group and the first dose group, no estimation can be obtained. In fact, no fetus is affected by both the skeletal and visceral malformation in any of these two groups and the VGAM package can not fit the Dirichlet-Multinomial distribution when there are cells with zero counts. We can not figure out how to modify the VGAM package to cope with zero counts. However, in the next section, when we fit the same data set using the bivariate shared response model, we write our own code and get estimations for all the groups.

2.3.2 Bivariate Shared Response Model

Let X_{j1}, X_{j2} indicate whether the j^{th} fetus in a litter suffers from, say, visceral and skeletal malformation. Let $X_j = (X_{j1}, X_{j2})^T$, the bivariate analogue of (2.1) is evidently

$$X_j = (1 - U_j)\, Y_j + U_j\, Z$$

where $Y_j = (Y_{j1}, Y_{j2})^T$, $Z = (Z_1, Z_2)^T$, U_j are mutually independent random variables and $U_1, U_2, \ldots, U_n$ are identically distributed as Bernoulli(π) . Thus π is the probability that a fetus within the litter will take on the shared response $Z = (Z_1, Z_2)^T$. Assuming that the vectors $Y_1, Y_2, \ldots, Y_n$ and Z are

independent and identically distributed with common mean $p = (p_1, p_2)^T$ and correlation $corr(Y_{j1}, Y_{j2}) = corr(Z_1, Z_2) = \phi$, it is straightforward to show that $\mathrm{P}(X_{j1}) = p_1\,, \mathrm{P}(X_{j2}) = p_2$ and the correlation structure induced on the observed $X_1, X_2, \ldots, X_n$ are

$$\begin{cases} corr(X_{j1}, X_{j2}) = \phi \\ corr(X_{j1}, X_{k1}) = corr(X_{j2}, X_{k2}) = \pi^2 \qquad \text{for } j \neq k \\ corr(X_{j1}, X_{k2}) = \phi\pi^2 \end{cases}$$

Let $n_{00}, n_{01}, n_{10}, n_{11}$ be the number of fetuses within a litter of size n that are non-affected, type II malformed only, type I malformed only and affected by both, and $p_{00}, p_{01}, p_{10}, p_{11}$ be their respective probabilities. It is easy to see that $n_{00} + n_{01} + n_{10} + n_{11} = n$ and $p_{00} + p_{01} + p_{10} + p_{11} = 1$. We also define $T = U_1 + U_2 + \cdots + U_n \sim \mathrm{Bin}(n, \pi)$ as the number of fetuses sharing Z. Then the

probability function of $n_{00}, n_{01}, n_{10}, n_{11}$ is given by

$$\begin{aligned}
\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11}) =& \mathrm{P}(Z = (0,0)^T)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid Z = (0,0)^T) \\
&+\mathrm{P}(Z = (0,1)^T)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid Z = (0,1)^T) \\
&+\mathrm{P}(Z = (1,0)^T)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid Z = (1,0)^T) \\
&+\mathrm{P}(Z = (1,1)^T)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid Z = (1,1)^T) \\
=& \mathrm{P}(Z = (0,0)^T) \sum_{t=0}^{n_{00}} \mathrm{P}(T = t)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid T = t, Z = (0,0)^T) \\
&+\mathrm{P}(Z = (0,1)^T) \sum_{t=0}^{n_{01}} \mathrm{P}(T = t)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid T = t, Z = (0,1)^T) \\
&+\mathrm{P}(Z = (1,0)^T) \sum_{t=0}^{n_{10}} \mathrm{P}(T = t)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid T = t, Z = (1,0)^T) \\
&+\mathrm{P}(Z = (1,1)^T) \sum_{t=0}^{n_{11}} \mathrm{P}(T = t)\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11} \mid T = t, Z = (1,1)^T) \\
=& p_{00} \sum_{t=0}^{n_{00}} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{n_{00}-t, n_{01}, n_{10}, n_{11}} p_{00}^{n_{00}-t} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}} \\
&+ p_{01} \sum_{t=0}^{n_{01}} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{n_{00}, n_{01}-t, n_{10}, n_{11}} p_{00}^{n_{00}} p_{01}^{n_{01}-t} p_{10}^{n_{10}} p_{11}^{n_{11}} \\
&+ p_{10} \sum_{t=0}^{n_{10}} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{n_{00}, n_{01}, n_{10}-t, n_{11}} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}-t} p_{11}^{n_{11}} \\
&+ p_{11} \sum_{t=0}^{n_{11}} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{n_{00}, n_{01}, n_{10}, n_{11}-t} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}-t}
\end{aligned} \tag{2.11}$$

In particular,

$$\mathrm{P}(n, 0, 0, 0) = p_{00} \sum_{t=0}^{n} \binom{n}{t} \pi^t (1-\pi)^{n-t} p_{00}^{n-t} + (1 - p_{00})(1-\pi)^n p_{00}^n$$

and $1 - \mathrm{P}(n, 0, 0, 0)$ is the risk that at least one fetus is adversely affected by at least one type of malformation.

It is easy to show that the marginal distribution of a bivariate shared response distribution is a univariate shared response distribution. Let $S = n_{10} + n_{11}$ be the number of fetus that is affected by type I malformation and $p = p_{10} + p_{11}$ be the probability that a fetus is affected by type I malformation. S follows a univariate shared response distribution too. This can be shown by

$$
\begin{aligned}
\mathrm{P}(S = s) &= \mathrm{P}(n_{10} + n_{11} = s) \\
&= \mathrm{P}(Z = (1,0)^T \text{ or } (1,1)^T)\mathrm{P}(n_{10} + n_{11} = s \mid Z = (1,0)^T \text{ or } (1,1)^T) \\
&\quad + \mathrm{P}(Z = (0,0)^T \text{ or } (0,1)^T)\mathrm{P}(n_{10} + n_{11} = s \mid Z = (0,0)^T \text{ or } (0,1)^T) \\
&= (p_{10} + p_{11}) \sum_{t=0}^{s} \mathrm{P}(T = t)\mathrm{P}(n_{10} + n_{11} = s \mid T = t, Z = (1,0)^T \text{ or } (1,1)^T) \\
&\quad + (p_{00} + p_{01}) \sum_{t=0}^{n-s} \mathrm{P}(T = t)\mathrm{P}(n_{10} + n_{11} = s \mid T = t, Z = (0,0)^T \text{ or } (0,1)^T) \\
&= (p_{10} + p_{11}) \sum_{t=0}^{s} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{s-t} (p_{10} + p_{11})^{s-t} (p_{00} + p_{01})^{n-s} \\
&\quad + (p_{00} + p_{01}) \sum_{t=0}^{n-s} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{s} (p_{10} + p_{11})^{s} (p_{00} + p_{01})^{n-t-s} \\
&= p \sum_{t=0}^{s} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{s-t} p^{s-t} (1-p)^{n-s} \\
&\quad + (1-p) \sum_{t=0}^{n-s} \binom{n}{t} \pi^t (1-\pi)^{n-t} \binom{n-t}{s} p^{s} (1-p)^{n-t-s}
\end{aligned}
$$

By treating $Z = (Z_1, Z_2)^T$ and the number $T = \sum_{j=1}^{n} U_j$ of fetuses that share Z within each litter as the missing data, the EM algorithm can again be used to estimate the parameters of the bivariate shared response model in much the same way as in the univariate case. Let

$$
(Z_{00}, Z_{01}, Z_{10}, Z_{11}) = \big(I\{Z = (0,0)^T\}, I\{Z = (0,1)^T\}, I\{Z = (1,0)^T\}, I\{Z = (1,1)^T\}\big)
$$

we have $(Z_{00}, Z_{01}, Z_{10}, Z_{11}) \sim m(x_1, x_2, x_3; 1, p_{00}, p_{01}, p_{10})$. The "Complete data" log-likelihood for one litter is

$$\begin{aligned}\ell_c &= t\log(\pi) + (n-t)\log(1-\pi) + z_{00}\log(p_{00}) + z_{01}\log(p_{01}) + z_{10}\log(p_{10}) \\ &\quad +(1 - z_{00} - z_{01} - z_{10})\log(1 - p_{00} - p_{01} - p_{10}) + (n_{00} - tz_{00})\log(p_{00}) \\ &\quad +(n_{01} - tz_{01})\log(p_{01}) + (n_{10} - tz_{10})\log(p_{10}) \\ &\quad +(n - t + tz_{00} + tz_{01} + tz_{10})\log(1 - p_{00} - p_{01} - p_{10})\end{aligned}$$

The E-step of the EM algorithm is done by, given the observed data $\mathrm{D} = \{n, n_{00}, n_{01}, n_{10}\}$, evaluating the conditional expectation of $\mathrm{E}(t \mid \mathrm{D})$, $\mathrm{E}(z_{00} \mid \mathrm{D})$, $\mathrm{E}(z_{01} \mid \mathrm{D})$, $\mathrm{E}(z_{10} \mid \mathrm{D})$, $\mathrm{E}(tz_{00} \mid \mathrm{D})$, $\mathrm{E}(tz_{01} \mid \mathrm{D})$ and $\mathrm{E}(tz_{10} \mid \mathrm{D})$ at the current parameter estimates and conditional probabilities

$$\begin{aligned}&\mathrm{P}(Z_{00} = z_{00}, Z_{01} = z_{01}, Z_{10} = z_{10}, T = t \mid \mathrm{D}) \\ &= \frac{\mathrm{P}(Z_{00} = z_{00}, Z_{01} = z_{01}, Z_{10} = z_{10}, T = t, \mathrm{D})}{\mathrm{P}(n_{00}, n_{01}, n_{10}, n_{11})}\end{aligned}$$

where the numerator is

$$\begin{aligned}&p_{00}^{z_{00}} p_{01}^{z_{01}} p_{10}^{z_{10}} (1 - p_{00} - p_{01} - p_{10})^{(1 - z_{00} - z_{01} - z_{10})} \binom{n}{t} \pi^t (1-\pi)^{(n-t)} \\ &\cdot \binom{n-t}{n_{00} - tz_{00}, n_{01} - tz_{01}, n_{10} - tz_{10}, n - n_{00} - n_{01} - n_{10} - t + tz_{00} + tz_{01} + tz_{10}} \\ &\cdot p_{00}^{n_{00} - tz_{00}} p_{01}^{n_{01} - tz_{01}} p_{10}^{n_{10} - tz_{10}} (1 - p_{00} - p_{01} - p_{10})^{n - n_{00} - n_{01} - n_{10} - t + tz_{00} + tz_{01} + tz_{10}}\end{aligned}$$

and the denominator is given by (2.11).

At the M-step of the EM algorithm, the imputed log-likelihood $\mathrm{E}(\ell_c \mid \mathrm{D})$ is maximized to update the values of current estimates. Beginning with a set of initial estimates, the EM algorithm is iterated until convergence is reached.

Table 2.8: *Estimated number of affected litters for the DEHP data by malformation type based on bivariate shared response model.*

	Visceral		Skeletal		Either	
Dose group	Observed	Expected	Observed	Expected	Observed	Expected
Control	4	4.37	2	2.07	6	6.14
44 mg/kg	1	0.98	1	0.98	2	1.93
91 mg/kg	11	11.84	6	6.53	15	14.89
191 mg/kg	11	10.73	13	12.19	15	14.13
292 mg/kg	9	8.21	7	8.30	9	8.91

Again, we use the same data set as last section to illustrate the bivariate shared response model. It is the DEHP data set listed in Table 3 of Lefkopoulou and Ryan (1993). At each dose level of di(2-ethylhexyl)-phthalate (DEHP), the table presents the numbers of fetuses that were found to have various combinations of three malformation types. Here we consider two malformation types: visceral and skeletal. At each dose level, we fit a bivariate shared response model using the EM algorithm. Table 2.8 reports the observed and expected numbers of litters affected with at least one viscerally malformed fetus, at least one skeletally affected fetus, and at least one malformed fetus of either types. It can be seen that the bivariate shared response model can estimate all groups of data and provides a good fit to the observed number of affected litters. As compared with the bivariate beta-binomial model, it gives comparable estimates to the single type of malformation and performs much better for the malformation of either types.

Chapter 3

Saturated model

In this chapter, we first introduce the saturated model by Bowman and George (1995) and the EM algorithm by Stefanescu and Turnbull (2003). We give a new proof of the formula in the E-step of the EM algorithm and our idea behind the proof provides a way for simulating data with unequal litter sizes. By fitting the saturated model, we test the goodness of fit of some commonly used parametric models via the likelihood ratio test and propose a new nonparametric estimator of the intra-litter correlation parameter ρ. Finally, we rectify the modified trend test by Stefanescu and Turnbull and show that the p-value of our new test statitic is quite close to the bootstrap results.

3.1 Introduction to Existing Work

As we have discussed in the last chapter, a whole host of distributions accounting for the *litter effect* and extra-binomial variation have been proposed to model exchangeable binary data in the literature. These distributions, even when they are matched to have the same marginal probability and intra-cluster correlation, can have very different shapes (George and Bowman, 1995; Kuk, 2004) and higher order joint probabilities for the underlying binary variables that could be of interest in certain applications. For example, in teratology risk assessment, the probability of having at least one malformed fetus within a litter of size n, denoted by $P(S_n \geqslant 1) = 1 - P(S_n = 0)$, is a measure of risk at the litter level. Interest in higher order joint probabilities such as $P(S_n = 0)$ necessitates the use of a fully parametric approach instead of approaches like quasi-likelihood (Liang and Hanfelt, 1994) or generalized estimating equations (Bowman, Chen, and George, 1995) that typically model only the first two moments. However, $P(S_n = 0)$ takes on different functional forms under different parametric distributions. For example,

$$P(S_n = 0) = \frac{\prod_{r=0}^{n-1} (1 - p + r\theta)}{\prod_{r=0}^{n-1} (1 + r\theta)}$$

under the beta-binomial model, where p is the marginal fetal response probability, $\theta = \rho/(1-\rho)$, and ρ is the intra-litter correlation. Under the correlated binomial distribution with additive interactions (Kupper and Haseman, 1978; Altham, 1978),

$$P(S_n = 0) = (1-p)^n \left\{ 1 + \frac{\rho n(n-1)p}{2(1-p)} \right\}.$$

Under the shared response model (Pang and Kuk, 2005),

$$P(S_n = 0) = (1-p)\sum_{t=0}^{n} \binom{n}{t} \pi^t (1-\pi)^{n-t}(1-p)^{n-t} + p(1-\pi)^n(1-p)^n,$$

where π is the probability of sharing a response and is related to the intra-litter correlation ρ by $\rho = \pi^2$. The simplest functional form for $P(S_n = 0)$ is

$$P(S_n = 0) = q^{n^\gamma}$$

for the q-power distribution, where $q = 1 - p$ and γ is a parameter that can be expressed in terms of p and the intra-litter correlation ρ (Kuk, 2004).

With so many distributions to choose from, model selection becomes an important issue. An alternative approach is to fit a saturated model as proposed by Bowman and George (1995). Xu and Prorok (2003) pointed out that in the case of varying cluster sizes, the maximum likelihood estimators (MLE) derived by Bowman and George (1995) are actually not the MLEs as claimed. Xu and Prorok then worked out what the MLEs should be and gave a detailed analysis when the maximum cluster size is two. However, even for this simple situation, there are five different scenarios and one of them still requires the solution of a nonlinear equation. They recommended using "uniroot" in S+ to solve it numerically. For the general case of cluster size greater than two, they recommend the Newton-Raphson method. Unfortunately, the algorithm fails to converge when applied to the six data sets reported in Brooks et al. (1997). Taking advantage of the statistical structure of the problem, Stefanescu and Turnbull (2003) derive an EM

algorithm for fitting the saturated model to exchangeable binary data by augmenting the data to make the cluster sizes equal. The algorithm appears to be stable and we encounter no convergence problem in using it to fit the saturated model to all six data sets. In deriving the EM algorithm, the assumption of compatibility of marginal distributions is made to link up the distributions for different cluster sizes so that estimation can be based on the combined data. Stefanescu and Turnbull (2003) proposed a modified trend test to test this assumption. Their test, however, fails to take into account the variability of an estimated null expectation and as a result leads to much inflated p-values. This drawback is rectified in this chapter.

In the next section, we give a detailed introduction to the saturated model and suggest a new proof of the formula that links up litters with different litter size via hypergeometric thinning. Not only is the new proof simpler and more intuitive than the existing one based on induction (Stefanescu and Turnbull, 2003), hypergeometric sampling also provides us with a simple way to generate litter data with unequal litter sizes.

3.2 The Saturated Model

As pointed out by Bowman and George (1995), if $X_1, ..., X_n$ are exchangeable binary variables, then the distribution of their sum $S_n = X_1 + ... + X_n$ can be

parameterized in terms of

$$\lambda_k = P(X_1 = \cdots = X_k = 1) = E(X_1 \cdots X_k), \quad 1 \leqslant k \leqslant n.$$

This is because for $0 \leqslant s \leqslant n$,

$$\begin{aligned} P(S_n = s) = p_n(s) &= \binom{n}{s} P(X_1 = \cdots = X_s = 1,\ X_{s+1} = \cdots = X_n = 0) \\ &= \binom{n}{s} E\left\{X_1 \cdots X_s (1 - X_{s+1}) \cdots (1 - X_n)\right\} \\ &= \binom{n}{s} \sum_{k=0}^{n-s} (-1)^k \binom{n-s}{k} \lambda_{s+k} \end{aligned} \tag{3.1}$$

by expanding the product inside the expectation sign and making use of exchangeability to group terms together. There is also an inversion formula

$$\lambda_k = \sum_{j=0}^{n-k} \binom{n-k}{k} \frac{P(S_n = n - j)}{\binom{n}{n-j}} \tag{3.2}$$

that expresses λ_k, $1 \leqslant k \leqslant n$, in terms of $P(S_n = s)$, $0 \leqslant s \leqslant n$. When the cluster sizes are all equal to n, the MLE $\hat{P}(S_n = s)$ of $P(S_n = s)$ is obviously just the observed proportion of clusters with $S_n = s$. Bowman and George (1995) then substituted $\hat{P}(S_n = s)$ into the inversion formula (3.2) to obtain the MLE of λ_k. When the clusters sizes are unequal, Bowman and George basically repeated the above procedure for each cluster size and weight the size-specific estimates of λ_k according to the empirical frequencies of the cluster sizes. As pointed out by Xu and Prorok (2003), the resulting estimates of λ_k are not the MLEs. Introducing double subscript notation for the present discussion, if $X_{n,1}, ..., X_{n,n}$ denote the

n exchangeable binary variables in a cluster of size n, Bowman and George were proceeding as if there is a new set of parameters

$$\lambda_{n,k} = P(X_{n,1} = \cdots = X_{n,k} = 1), \quad 1 \leqslant k \leqslant n$$

for every cluster size n , resulting in a large triangular array of parameters. Throughout this chapter, we make the common assumption of marginal compatibility, meaning that the marginal distribution of $X_{m,1}, \ldots, X_{m,n}$ in a cluster of size $m > n$ should be the same as that of $X_{n,1}, \ldots, X_{n,n}$. As a result, $\lambda_{n,k}$ no longer depends on n and can be written simply as λ_k, leading to a more parsimonious sequence rather than a triangular array of parameters. This assumption links up the distributions for different cluster sizes so that estimation can be based on the combined data across all cluster sizes. Other authors have called this the reproducibility assumption (Prentice, 1988), or the interpretability assumption (Stefanescu and Turnbull, 2003), but we find marginal compatibility to be a self-explanatory name.

The observed data D consists of (n_i, s_i), $1 \leqslant i \leqslant C$, where n_i is the size of cluster i, s_i the sum of the exchangeable binary variables in cluster i, and C the total number of clusters. The log-likelihood function can be written down as

$$\ell = \sum_{i=1}^{C} \log P\left(S_{n_i} = s_i\right) = \sum_{i=1}^{C} \log p_{n_i}(s_i),$$

where $P(S_{n_i} = s_i)$ is given by equation (3.1). It is obvious that the likelihood is a function of the parameters $\lambda_1, \ldots, \lambda_m$, where m is the maximum cluster size. However, $\lambda_1, \ldots, \lambda_m$ are not good parameters to work with because they have to form a completely monotone sequence, satisfying $(-1)^k \Delta^k \lambda_j \geqslant 0$ for integers $k \geqslant 1$,

where Δ^k is the k^{th} forward difference, in order for (3.1) to define a bona fide probability function (Feller, 1971, p.224). These conditions of alternating signs for the finite differences are difficult to enforce during the iterations of any numerical maximization procedure. A more convenient parameterization is given by

$$p_m(0) = P(S_m = 0), \quad p_m(1) = P(S_m = 1), \quad \ldots \quad , \quad p_m(m) = P(S_m = m),$$

where S_m is the sum of the exchangeable binary variables in a cluster of maximum size m. Even though $0 \leqslant p_m(s) \leqslant 1$ for every s and they have to sum up to one, these constraints are automatically satisfied if all the clusters are of the same size m.

Note that once we know $p_m(s)$ for $0 \leqslant s \leqslant m$, we can determine the probabilities $p_n(s)$ in the smaller clusters as well by using the inversion formula (3.2) to obtain $\lambda_1, \ldots, \lambda_m$, which can be substituted back into (3.1) for $n < m$. By induction on $m - n$, Stefanescu and Turnbull (2003) gave a representation of $p_n(s)$ in terms of $p_m(s)$ for $0 \leqslant s \leqslant m$ and derived an EM algorithm for obtaining the maximum likelihood estimates of $p_m(0), \cdots, p_m(m)$ via data augmentation.

A convenient way to embed the observed data within the conceptual "complete" data is to assume that all the clusters are of the same size m but for cluster i, we only observe the sum s_i of the first n_i binary variables whereas the sum u_i of the last $m - n_i$ variables is unobserved. Thus the "complete" data are $r_i = s_i + u_i$, $1 \leqslant i \leqslant C$, and only the s_i are observed. In other words, we are augmenting the data to make the cluster sizes equal. The log-likelihood based on the complete data

is evidently

$$\ell_c = \sum_{i=1}^{C} \log P\left(S_m = r_i\right) = \sum_{i=1}^{C} \log p_m(r_i) = \sum_{s=0}^{m} f(s) \log p_m(s), \tag{3.3}$$

where

$$f(s) = \sum_{i=1}^{C} z_{i,s} = \sum_{i=1}^{C} I\{r_i = s\} \tag{3.4}$$

is the observed frequency of s in the complete data. The E-step of the EM algorithm is to reconstruct the complete data likelihood from the observed data by taking conditional expectation to get

$$Q(\theta\ ; \hat{\theta}^{(t)}) = E\left\{ \ell_c(\theta) \,\middle|D;\ \hat{\theta}^{(t)} \right\},$$

where θ is a generic symbol for the vector of all the model parameters, and the conditional expectation is evaluated at the current parameter estimate $\hat{\theta}^{(t)}$. At the M step of the EM algorithm, $Q(\theta\ ; \hat{\theta}^{(t)}) = E\left\{ \ell_c(\theta) \,\middle|D;\ \hat{\theta}^{(t)} \right\}$ is maximized with respect to θ to obtain the updated estimate $\hat{\theta}^{(t+1)}$. The procedure is iterated until convergence. A standard result is that

$$\ell'(\theta) = E\left\{ \ell_c'(\theta) \,|D;\ \theta \right\},$$

where $\ell'(\theta)$ and $\ell_c'(\theta)$ denote the derivative of the observed and complete data log-likelihood with respect to the parameters. Thus at convergence of the EM algorithm, $\hat{\theta} = \lim_{t\to\infty} \hat{\theta}^{(t)}$ satisfies $E\left\{ \ell_c'(\hat{\theta}) \,\middle|D;\ \hat{\theta} \right\} = \ell'(\hat{\theta}) = 0$, which is the likelihood equation for finding the observed data MLE. The multinomial form (3.3) of the complete data log-likelihood leads to some simplifications of the EM algorithm.

Firstly, since the clusters are all of equal size m under the complete data setup, the complete data MLEs of $p_m(s)$, $0 \leqslant s \leqslant m$, are simply the multinomial proportions $f(s)/C$. Secondly, it is straightforward to show that the updated estimates obtained from the M step are simply conditional expectations of the complete data MLE evaluated at the current parameter estimates. Specifically,

$$\hat{p}_m^{(t+1)}(s) = \frac{E\left\{ f(s) \,\middle|D;\ \hat{\theta}^{(t)} \right\}}{\mathrm{C}} = \frac{\sum_{i=1}^{C} E(z_{i,s}|s_i)}{C} = \frac{\sum_{i=1}^{C} P(s_i + u_i = s|s_i)}{C}, \qquad (3.5)$$

where

$$P(s_i + u_i = s|s_i) = \frac{P(s_i, u_i = s - s_i)}{P(s_i)}.$$

Now, there are $\binom{n_i}{s_i}\binom{m-n_i}{s-s_i}$ strings of s 1's and $m-s$ 0's with s_i 1's in the first n_i positions, and by exchangeability, the probability of any such string is $p_m(s)\Big/\binom{m}{s}$. Thus

$$P(s_i + u_i = s|s_i) = \frac{\binom{n_i}{s_i}\binom{m-n_i}{s-s_i}}{\binom{m}{s}} \frac{p_m(s)}{p_{n_i}(s_i)} \qquad (3.6)$$

if $s_i \leqslant s \leqslant s_i + m - n_i$, and zero otherwise. This result has also been proven independently by Stefanescu and Turnbull (2003),

To evaluate (3.6) and hence (3.5) at the current estimates $\hat{p}_m^{(t)}(0), \cdots, \hat{p}_m^{(t)}(m)$, we need a formula that relates the probabilities $p_n(s)$ for smaller cluster sizes n to the probabilities $p_m(s)$ for the maximum cluster size m. The naive way to do this is to use the inversion formula (3.2) to obtain $\lambda_1, ..., \lambda_m$, which can be substituted

back into (3.1) for $n < m$. This approach leads to a double summation and is not the simplest formula relating $p_n(s)$, for $n < m$, to $p_m(s)$. We are able to derive a simpler formula by arguing as follows. In order to get s malformed fetuses in the observed litter, there must be $t = s,\ s+1,\ \ldots,\ s+(m-n)$ malformed fetuses in the "completed" litter of size m, with precisely s malformed ones among the first n fetuses. Assuming marginal compatibility, the probability of t malformations in the "completed" litter is given by $p_m(t)$, and the probability that there are s malformations among the first n fetuses of a litter of size $m > n$ with t malformations follows a hypergeometric distribution. Thus,

$$\begin{aligned} p_n(s) = P(S_n = s) = P(S_n = s, S_m \geqslant s) &= \sum_{t=s}^{s+(m-n)} P(S_m = t) P(S_n = s | S_m = t) \\ &= \sum_{t=s}^{s+(m-n)} p_m(t) \frac{\binom{t}{s}\binom{m-t}{n-s}}{\binom{m}{n}} \end{aligned} \tag{3.7}$$

Note that

$$\frac{\binom{t}{s}\binom{m-t}{n-s}}{\binom{m}{n}} = \frac{\binom{n}{s}\binom{m-n}{t-s}}{\binom{m}{t}}$$

and hence (3.7) is the same as the formula proven by Stefanescu and Turnbull (2003) using induction on $m - n$. Not only is our proof of (3.7) more intuitive, it also provides us with a way of simulating data with unequal litter sizes, namely, by simulating data from the probability distribution $p_m(0),\ \ldots\ , p_m(m)$ first which is appropriate for litters of size m, followed by trimming to the observed litter size n by means of hypergeometric thinning.

The EM algorithm for obtaining MLE based on the observed incomplete data

is stable in our experience. If there is a maximum cluster size in the population, which seems reasonable for litter data in developmental toxicity studies, the usual asymptotic theory for the MLE should hold as the number of clusters increases. In particular, the likelihood ratio test of comparing the fit of the saturated model with a parametric model is asymptotically chi-square.

3.3 Goodness of Fit Test of Parametric Models

In this section, the saturated model is fitted to the six data sets used by Brooks *et al.* (1997) where the observed outcomes are the number of dead fetuses or implants in each litter. As commented previously, one way to assess the goodness of fit of a parametric model is to compare the parametric fit versus the saturated fit via the likelihood ratio test, which follows a chi-square distribution under the null hypothesis. We tested the goodness of fit of two parametric models here: the beta-binomial distribution which is perhaps the most widely used distribution for modelling litter data and the q-power distribution which possesses a lot of nice properties (Kuk, 2004). The results are shown in Table 3.1. It can be seen that the beta-binomial distributed is rejected at level 0.05 for the two data sets HS2 and HS3, whereas the likelihood ratio test of the q-power distribution is not significant for all six data sets.

Table 3.1: *Minus log-likelihood of saturated, beta-binomial and q-power distributions for six data sets.*

	Data sets					
	E1	E2	HS1	HS2	HS3	AVSS
Saturated	276.95	338.01	764.21	1628.13	681.54	166.84
Beta-binomial	282.65	344.88	777.79	1657.30	701.33	168.93
Likelihood ratio	11.39	13.72	27.15	58.33*	39.58*	4.17
(d.f.)	(15)	(17)	(18)	(11)	(16)	(18)
q-power	282.59	345.85	776.40	1636.40	685.18	172.90
Likelihood ratio	11.30	15.67	24.38	16.54	7.28	12.13
(d.f.)	(15)	(17)	(18)	(11)	(16)	(18)

* Significant at level 0.05

3.4 Simulation Results for the Saturated Model

To illustrate and contrast the lack of robustness of the parametric estimates with the distribution free property of the saturated model estimates, we simulate data for 200 litters 200 times using either a beta-binomial or the q-power distribution with marginal probability $p = 0.2$ and intra-litter correlation $\rho = 0.2$. The litter sizes are generated according to the distribution given in Table 6.5 of Aerts *et al.* (2002). The true probability function for a litter of size 16 is shown in Figure 3.1, together with the average estimates based on the saturated as well as a misspecified model (beta-binomial instead of q-power, or vice versa). It can be seen clearly that the parametric fit is biased when the model is misspecified whereas the fit based on the saturated model is generally valid.

To study the behaviour of the estimates when the marginal compatibility as-

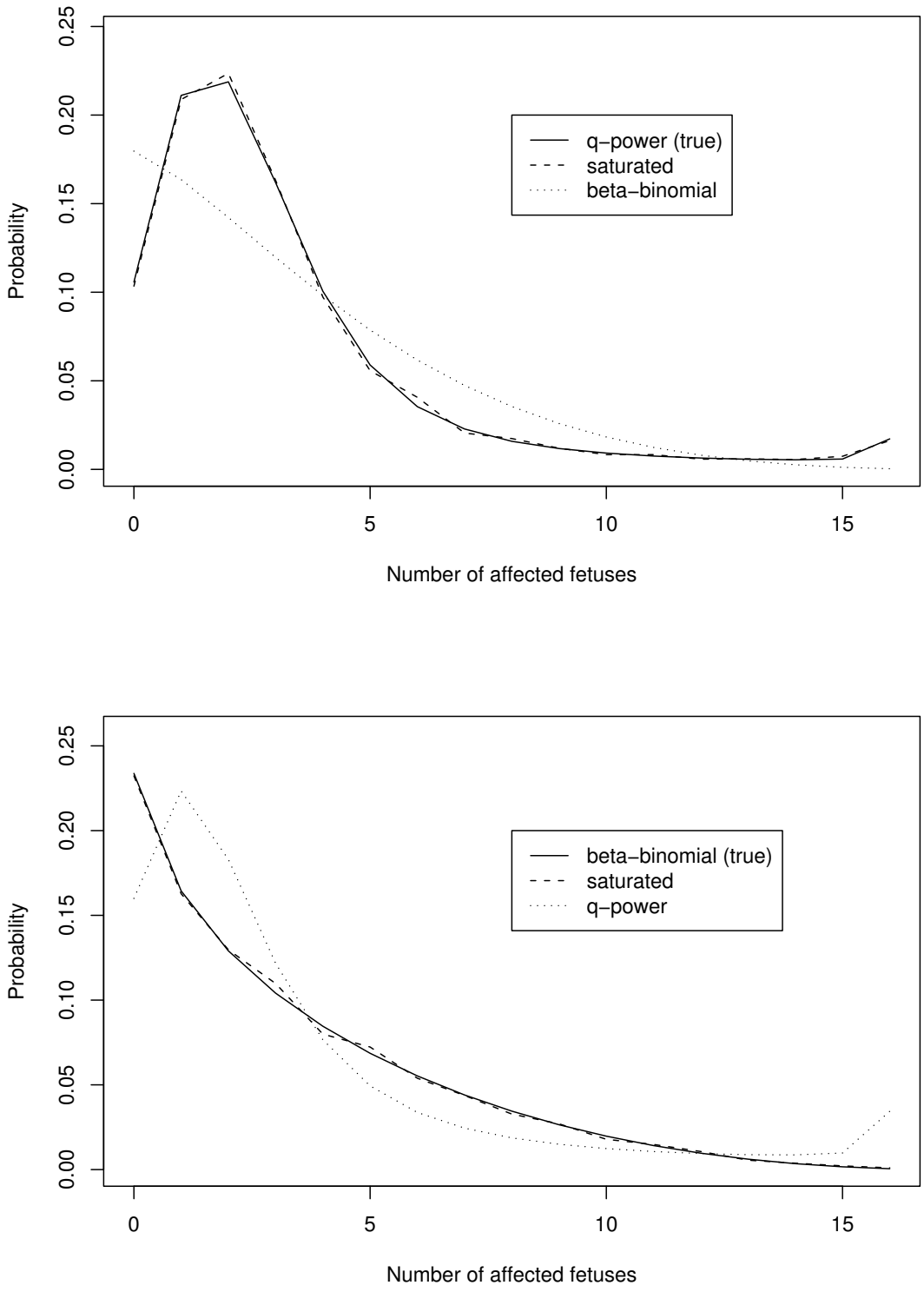


Figure 3.1: *Averages of maximum likelihood estimates under the saturated model and a misspecified parametric model.*

sumption is violated, we conduct the following simulation study. As before, we simulate data for 200 litters from both the beta-binomial and $q-$power distributions, with litter sizes generated according to Table 6.5 of Aerts *et al.* (2002). Unlike the previous simulations where the marginal response probability $p \equiv .2$ for litters of all sizes, here we let the marginal probability depends linearly on the log litter size n in the logit scale

$$\text{logit}(p) = \text{logit}(.15) + \left\{ \frac{\text{logit}(.25) - \text{logit}(.15)}{log(19)} \right\} log(n),$$

so that p increases from .15 to .25 as the litter size increases from 1 to the maximum size of 19. The pairwise odds ratio within litters is kept constant at the value 2.953 chosen to make the pairwise correlation equal to .2 when $p = .2$. Note that the intra-litter correlation is also changing with litter size as it depends on the marginal probability. Since p is actually litter size dependent, we conjecture that the MLE $\hat{p}$ of p obtained by assuming marginal compatibility is actually estimating

$$p^* = \sum_{n=1}^{19} f_n p_n,$$

where f_n is the relative frequency of litter size n from Table 6.5 of Aerts *et al.* (2002), and p_n is the marginal response probability given by the logistic regression above. We can interpret p^* as the probability that a randomly selected fetus from a randomly selected litter is malformed. Based on 1000 simulations, Table 3.2 gives the bias of $\hat{p}$ as an estimator of p^*, the empirical standard error of $\hat{p}$, as well as the average of the estimated standard error $\hat{SE}(\hat{p})$ obtained using the information matrix of the misspecified saturated model. Also shown in Table 3.2 are the

Table 3.2: *Bias of estimator and coverage of confidence interval when the marginal compatibility assumption is violated.*

Distribution	p^*	Bias	SE	Ave($\hat{SE}$)	$p^* < \hat{p} - 1.96\hat{SE}$	$p^* > \hat{p} + 1.96\hat{SE}$
Beta-binomial	.2301	.0014	.0157	.0151	.027	.039
q-power	.2301	.0016	.0163	.0161	.024	.031

proportions of times $p^* < \hat{p} - 1.96\,\hat{SE}(\hat{p})$ and $p^* > \hat{p} + 1.96\,\hat{SE}(\hat{p})$. Ideally, they should both be close to the nominal value of .025. It can be seen from Table 3.2 that even though p is now ranging from .15 to .25 rather than constant, the MLE $\hat{p}$ that assumes marginal compatibility is an almost unbiased estimator of p^*, and the conventional 95% confidence interval for p turns out to cover p^* almost 95% of the times. These results suggest that the saturated model maximum likelihood estimates are somewhat robust to moderate departure from the marginal compatibility assumption.

3.5 Estimation of Intra-litter Correlation Parameter

There is considerable interest in estimating the intraclass correlation from clustered binary data (Ridout, Dométrio and Firth, 1999; Zou and Donner, 2004). Twenty estimators are compared in the study by Ridout *et al.* (1999). They conclude that the asymptotically equivalent estimators $\hat{\rho}_{AOV}, \hat{\rho}^*_{AOV}, \hat{\rho}_{FC}, \hat{\rho}^*_{KPR}, \hat{\rho}^*_W$ and $\hat{\rho}_{UB}$

all performed well in their simulations, and none of them appeared to have any consistent small-sample advantage. Note that the equation defining $\hat{\rho}_{UB}$ in Ridout *et al.* (1999) contains one typographical error.

Let $\hat{\lambda}_1$ and $\hat{\lambda}_2$ be the MLE of λ_1 and λ_2 under the saturated model, another nonparametric estimator of ρ is

$$\hat{\rho}_{NP} = \frac{\hat{\lambda}_2 - \hat{\lambda}_1^2}{\hat{\lambda}_1(1 - \hat{\lambda}_1)}.$$

We investigate the performance of $\hat{\rho}_{NP}$ using the same study design as Ridout *et al.* (1999) together with the MLE based on beta-binomial distribution, alternating logistic regression (ALR) estimator and six estimators recommended by Ridout *et al.* The plot resembling figure 1 of Ridout *et al* is given in Figure 3.2, which gives a summary of the performance of the 9 estimators over 180 simulations. In Figure 3.2, the plotted point is the median. Lower and upper end-points of the vertical lines indicate the 5^{th}, 25^{th}, 75^{th} and 95^{th} percentiles of the distribution. We can see from Figure 3.2 that $\hat{\rho}_{NP}$ is almost an unbiased estimator and the standard deviation is comparable to the six estimators recommended by Ridout *et al.* This shows that the performance of $\hat{\rho}_{NP}$ is on par with the six estimators recommended by Ridout *et al.* However, it is also clear that the improvement is not substantial. Figure 3.2 also shows that the ALR estimator is another good estimator that can be used to estimate the intraclass correlation from clustered binary data.

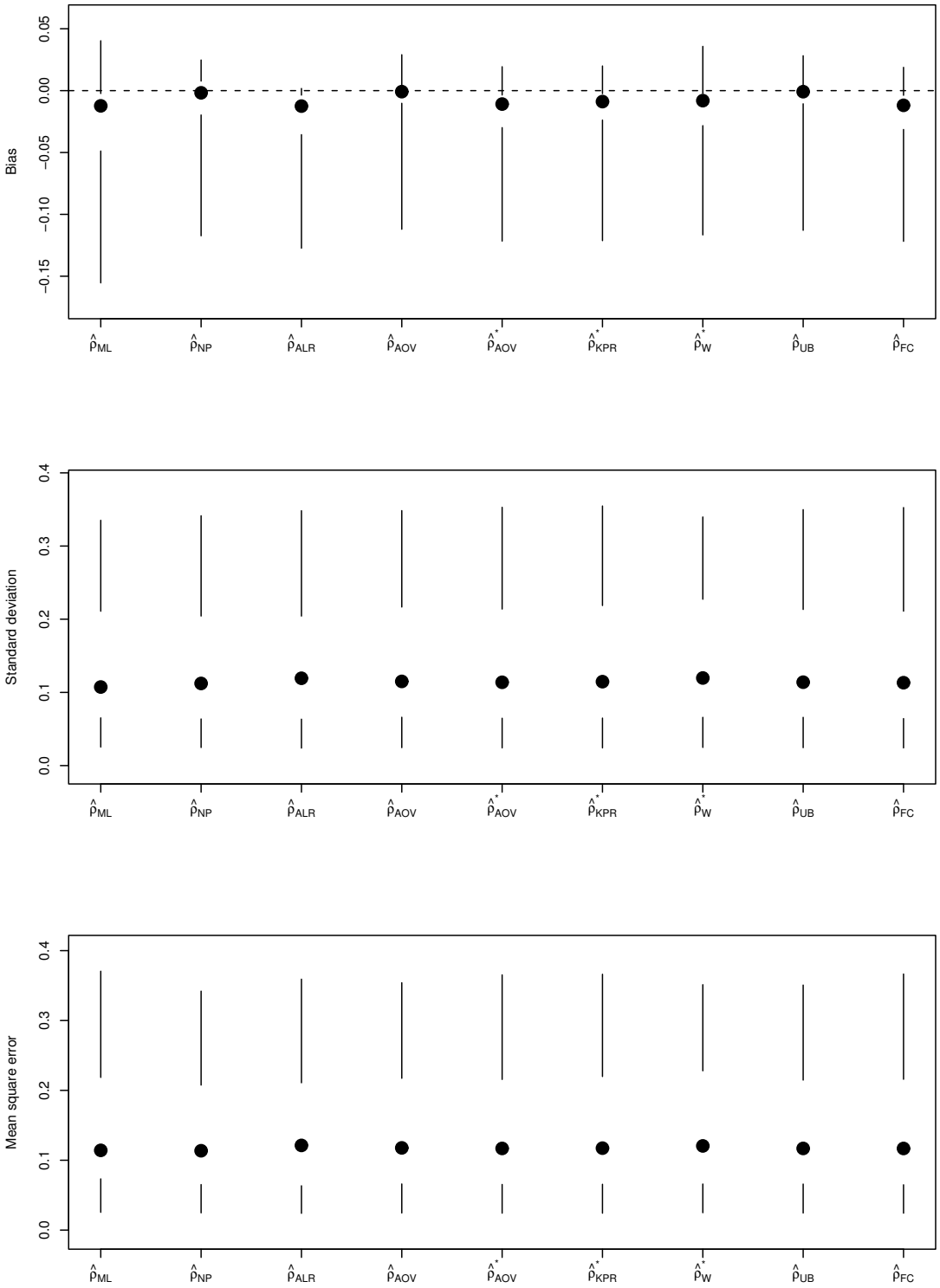


Figure 3.2: *Bias, standard deviation and square root mean square error of 9 estimators of* ρ

3.6 Testing the Marginal Compatibility Assumption

The assumption of marginal compatibility links up the fetal response distributions for different litter sizes so that estimation can be based on the combined data across all litter sizes. Note that, in the FDA stipulation of the Segment II design of rodent teratology experiments, treatments to dams are applied after the fetuses have been implanted and so the number of implantation should not be affected by treatment. It follows that if the toxicity endpoint is combined resorption, death, or malformation, then the "litter size" may be taken as the number of implanted fetuses which is not dose-related and marginal compatibility is a reasonable assumption to make. If the toxicity endpoint of interest is malformation, then the litter size is usually taken as the number of live fetuses which may be dose related, and the marginal compatibility assumption may not be appropriate. Stefanescu and Turnbull (2003) mentioned competition or cooperation between litter-mates as potential sources of violations from the marginal compatibility assumption, but remarked that the assumption may be reasonable in situations such as familial aggregation studies of disease, or grouped randomized trials. The bottom line is that the assumption has to be checked against the observed data. Stefanescu and Turnbull (2003) generalized Armitage's trend test (Armitage, 1955) for independent data to the present case of clustered data by taking intracluster correlation into account in the variance calculation. For $1 \leqslant n \leqslant m$, let t_n be the total number of malformed fetuses in all

C_n clusters of size n. An estimate of the marginal fetal response probability p_n for clusters of size n is evidently $\hat{p}_n = t_n/nC_n$, the sample proportion of malformed fetuses in the C_n clusters of size n. To verify marginal compatibility, one would obviously start by checking whether the marginal response probabilities are equal or not by testing the hypothesis $H_0 : p_1 = \ . \ . \ . \ = p_m = p$. Armitage's test is a test of H_0 against a possible trend in $p_1, \ . \ . \ . \ , p_m$. In deriving this test, a score w_n is assigned to the total t_n over clusters of size n to yield $T = \sum_{n=1}^{m} w_n t_n = \sum_{n=1}^{m} a_n \hat{p}_n$, where $a_n = nC_n w_n$. We follow Stefanescu and Turnbull (2003) in using the scores $w_1 = -(m-1)/2$, $w_2 = -(m-3)/2$, . . ., $w_m = (m-1)/2$. Under H_0, the expected value of T is $E_0(T) = \sum_{n=1}^{m} a_n p$ and the null variance of T is

$$var_0(T) = p(1-p) \sum_{n=1}^{m} w_n^2 C_n n \{1 + (n-1)\rho\}, \tag{3.8}$$

where ρ is the common intra-cluster correlation. A plausible test statistic of H_0 is the standardized difference

$$Z_0 = \frac{T - E_0(T)}{\sqrt{\hat{var}_0(T)}},$$

where $\hat{var}_0(T)$ is an estimate of $var_0(T)$ obtained by replacing the parameters p and ρ in (3.8) by their maximum likelihood estimates $\hat{p}$ and $\hat{\rho}$. It follows from standard theory that the null distribution of Z_0 is asymptotically standard normal. However, Z_0 is not a usable test statistic because the unknown parameter p is involved in $E_0(T)$. Stefanescu and Turnbull (2003) proposed to overcome this problem by replacing p with $\hat{p}$ to yield $\hat{E}_0(T) = \sum_{n=1}^{m} a_n \hat{p}$,

$$\hat{Z}_0 = \frac{T - \hat{E}_0(T)}{\sqrt{\hat{var}_0(T)}},$$

Table 3.3: *Nominal and bootstrap p-values for two versions of Armitage's trend test for seven data sets*

	Data sets						
	COPD	E1	E2	HS1	HS2	HS3	AVSS
$\hat{Z}_0$	1.744	.206	.583	.322	1.859	1.745	1.117
Nominal p-value	.081	.837	.560	.747	.063	.081	.264
Bootstrap p-value	.068	.722	.325	.596	.013	.009	.066
Z	1.842	.342	.940	.588	2.386	2.641	2.040
Nominal p-value	.065	.733	.347	.556	.017	.008	.041
Bootstrap p-value	.070	.705	.332	.544	.016	.009	.043

and it is claimed that $\hat{Z}_0$ is asymptotically standard normal just like Z_0. What has been overlooked is the fact that the null variance of $T - \hat{E}_0(T)$ is not asymptotically the same as that of T and so $\hat{var}_0(T)$ is not a consistent estimator of $var_0\left\{T - \hat{E}_0(T)\right\}$. Therefore, the use of standard normal distribution as reference should lead to misleading $p-$values. This is confirmed when we compare the nominal $p-$values of $\hat{Z}_0$ based on the claimed standard normal distribution with the bootstrap $p-$values obtained by simulating litter data 1000 times from the estimated probability distribution $\hat{p}_m(0), \cdots, \hat{p}_m(m)$ using the hypergeometric method described in the comments that follow equation (3.7). We do this for the COPD data that Stefanescu and Turnbull (2003) used to illustrate their method, as well as the six data sets used by Brooks et al. (1997). As can be seen from Table 3.3, the nominal $p-$values based on standard normal approximation are all greater than the bootstrap $p-$values, and substantially so except for the COPD data. We conjecture that $var_0\left\{T - \hat{E}_0(T)\right\}$ is less than $\hat{var}_0(T)$ due to positive

correlation between T and $\hat{E}_0(T)$. As a result, the denominator of $\hat{Z}_0$ overestimates the standard deviation of $T - \hat{E}_0(T)$.

For a standard normal approximation to be valid, we need to divide $T - \hat{E}_0(T)$ by the correct standard deviation estimate. Now,

$$var_0\left\{T - \hat{E}_0(T)\right\} = var_0(T) + var_0\left\{\hat{E}_0(T)\right\} - 2cov_0\left(T,\ \hat{E}_0(T)\right). \tag{3.9}$$

To simplify the calculation of the covariance term, we propose to estimate p by the overall sample proportion of malformed fetuses $\hat{p} = \sum_{n=1}^{m} t_n/N$, where $N = \sum_{n=1}^{m} nC_n$ is the total number of fetuses over all clusters. Now, $\hat{E}_0(T) = \hat{p}\sum_{n=1}^{m} a_n = \sum_{n=1}^{m} b_n t_n$, where $b_n \equiv A/N$, with $A = \sum_{n=1}^{m} a_n$. The important thing to note is that $T = \sum_{n=1}^{m} w_n t_n$ and $\hat{E}_0(T) = \sum_{n=1}^{m} b_n t_n$ are just two linear combinations of the t_n and hence the variance and covariance terms can be written down easily using the fact $var_0(t_n) = C_n np(1-p)\{1 + (n-1)\rho\}$ and the fact that t_n and t_k are independent for $n \neq k$ because they are totals over non-overlapping litters. Hence,

$$var_0\left\{\hat{E}_0(T)\right\} = \sum_{n=1}^{m} b_n^2 var_0(t_n) = p(1-p)\sum_{n=1}^{m} b_n^2 C_n n\{1 + (n-1)\rho\} \tag{3.10}$$

and

$$cov_0\left(T,\ \hat{E}_0(T)\right) = \sum_{n=1}^{m} w_n b_n var_0(t_n) = p(1-p)\sum_{n=1}^{m} w_n b_n C_n n\{1 + (n-1)\rho\}. \tag{3.11}$$

Substituting (3.8), (3.10) and (3.11) into (3.9) and replacing the unknown p and ρ by consistent estimators $\hat{p}$ and $\hat{\rho}$ will lead to a consistent estimator $\hat{var}_0\left\{T - \hat{E}_0(T)\right\}$ of $var_0\left\{T - \hat{E}_0(T)\right\}$. Recall that we are using $\hat{p} = \sum_{n=1}^{m} t_n/N$ rather than the maximum likelihood estimator to obtain closed form variance and covariance formulae.

By the same token, we suggest to take $\hat{\rho}$ as the Fleiss-Cuzick estimator (Fleiss and Cuzick, 1979), which is found to perform well in the studies by Ridout, Dométrio and Firth (1999) and Zou and Donner (2004), since the maximum likelihood estimator of ρ has no closed form. The test statistic that we propose is

$$Z = \frac{T - \hat{E}_0(T)}{\sqrt{\hat{var}_0\left\{T - \hat{E}_0(T)\right\}}}$$

which should be asymptotically standard normal under the hypothesis. The results based on this test are also shown in Table 3.3. It can be seen that the nominal $p-$values are now much closer to the bootstrap $p-$values which lends support to the validity of the standard normal approximation. We can conclude from Table 3.3 that the marginal compatibility assumption is not satisfied for data sets HS2, HS3 and AVSS.

Chapter 4

Smoothing the Nonparametric Estimates

Due to the sparseness of data, the MLE of the probability function under the saturated model, which we also call the nonparametric MLE (even though the number of parameters in the saturated model is still finite), can exhibit a lot of roughness. We extend the penalized likelihood approach proposed by Simonoff (1983) for smoothing the nonparametric MLE to the case of unequal cluster sizes and again use an EM type algorithm for its implementation.

4.1 Penalized Saturated Model

Considering the data sets E1, E2 and HS1 with insignificant $p-$values in Table 3.3, the nonparametric probability function estimates displayed in Figure 4.1 are quite jagged for some data sets and are in need of smoothing. For the case of equal cluster size $n_i \equiv m$, so that $r_i = s_i$ (no missing data) in the notation that we used in chapter 3, Simonoff (1983) proposed the following penalized log-likelihood

$$\begin{aligned} \ell_{c,\beta} &= \sum_{i=1}^{C} \log p_m(r_i) - \beta \sum_{s=0}^{m-1} \{\log p_m(s+1) - \log p_m(s)\}^2 \\ &= \sum_{s=0}^{m} f(s) \log p_m(s) - \beta \sum_{s=0}^{m-1} \{\log p_m(s+1) - \log p_m(s)\}^2, \end{aligned} \tag{4.1}$$

where $f(s)$ given by (3.4) is the frequency or number of clusters with $r_i = s$. In the general case, we observe s_i positive responses from cluster i which is of size n_i, where the n_i are unequal and m is the maximum litter size. We extend the penalized log-likelihood to the general case by using the same penalty term as in (4.1) and subtract it from the log-likelihood $\ell = \sum_{i=1}^{C} \log p_{n_i}(s_i)$ of the observed data to get

$$\ell_\beta = \sum_{i=1}^{C} \log p_{n_i}(s_i) - \beta \sum_{s=0}^{m-1} \{\log p_m(s+1) - \log p_m(s)\}^2. \tag{4.2}$$

Note again that $p_n(s)$ for $n < m$ can be determined from $p_m(0), \cdots, p_m(m)$ via (3.7) and hence the penalized log-likelihood ℓ_β is a complex function of the parameters $\theta = \{p_m(0), \cdots, p_m(m)\}$. To maximize ℓ_β with respect to θ for a fixed β, we can again augment every cluster to size m to get the "complete" data $r_i = s_i + u_i$ and use an EM type algorithm.

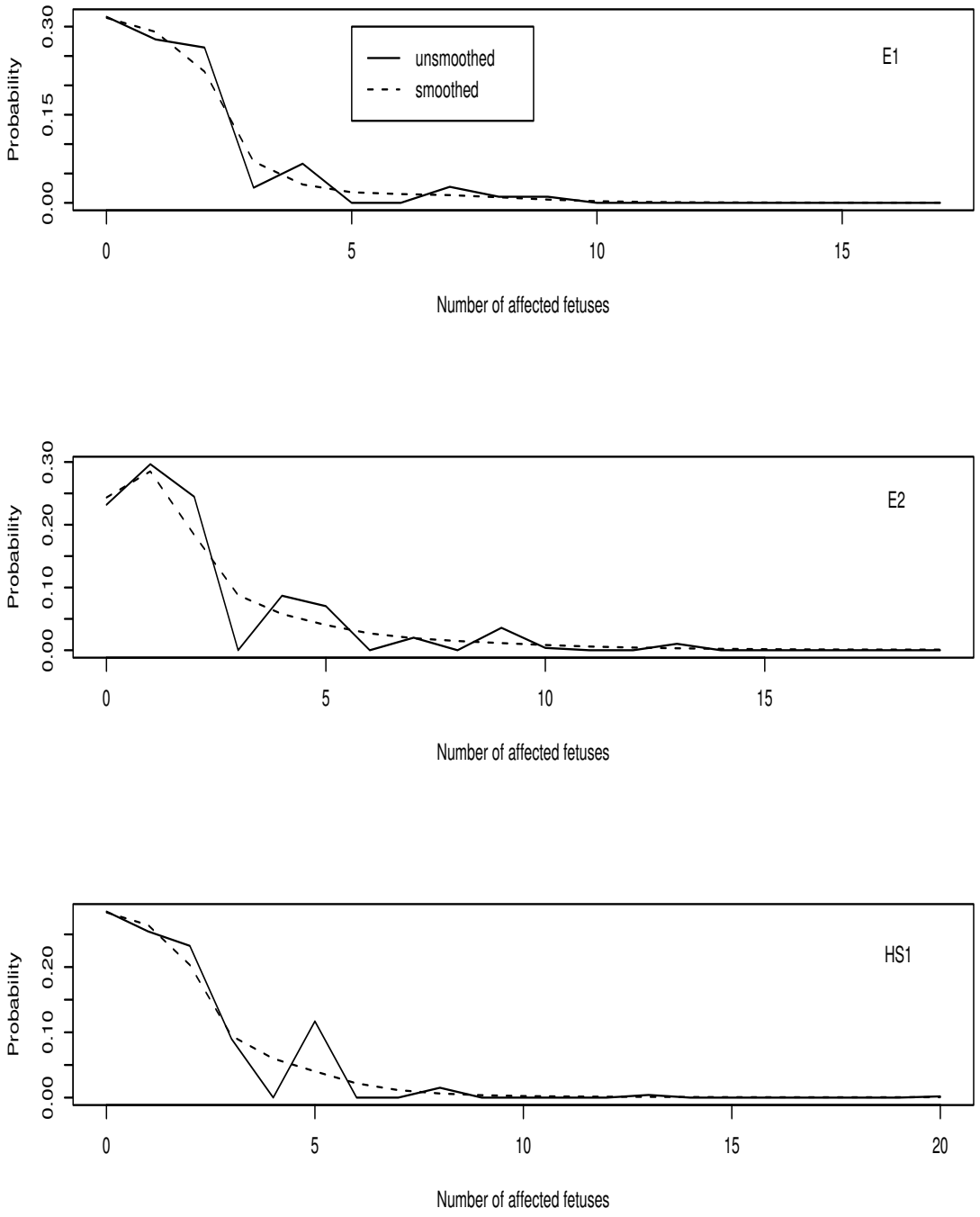


Figure 4.1: *Maximum likelihood and penalized likelihood estimates for three data sets under the saturated model*

At the E step, we take conditional expectation of (4.1), the complete data version of the penalized log-likelihood to get

$$\begin{aligned} E\left\{\ell_{c,\beta}(\theta) \;\middle|\; D;\hat{\theta}^{(t)}\right\} &= \sum_{s=0}^{m} E\left\{f(s)\,\middle|\, D;\hat{\theta}^{(t)}\right\}\log p_m(s) \\ &\quad -\beta\sum_{s=0}^{m-1}\left\{\log p_m(s+1)-\log p_m(s)\right\}^2, \end{aligned}$$

where $E\left\{f(s) \;\middle|\; D;\hat{\theta}^{(t)}\right\}$ can be obtained as before using (3.5) and (3.6).

At the M step, we maximize $E\left\{\ell_{c,\beta}(\theta) \;\middle|\; D;\hat{\theta}^{(t)}\right\}$ with respect to

$\theta = \{p_m(0),\cdots,p_m(m)\}$ subject to the constraint $\sum_{s=0}^{m} p_m(s) = 1$. This can be done using the method of Lagrange multiplers by defining

$$\begin{aligned} L &= E\left\{\ell_{c,\beta}(\theta)\,\middle|\, D;\hat{\theta}^{(t)}\right\} - \lambda\left(\sum_{s=0}^{m} p_m(s) - 1\right) \\ &= \sum_{s=0}^{m} E\left\{f(s) \;\middle|\; D;\hat{\theta}^{(t)}\right\}\log p_m(s) - \beta\sum_{s=0}^{m-1}\left\{\log p_m(s+1)-\log p_m(s)\right\}^2 \\ &\quad -\lambda\left(\sum_{s=0}^{m} p_m(s) - 1\right) \end{aligned}$$

and setting the derivatives of L with respect to $p_m(0),\cdots,p_m(m)$ and λ equal to zero. Writing $E\left\{f(s) \;\middle|\; D;\hat{\theta}^{(t)}\right\}$ as $E\{f(s) \mid D\}$ to save space, the resulting equations are

$$E\{f(0) \mid D\} + 2\beta\{\log p_m(1) - \log p_m(0)\} = \lambda\, p_m(0),$$

$$E\{f(s)\,|D\} - 2\beta\{\log p_m(s) - \log p_m(s-1)\} + 2\beta\{\log p_m(s+1) - \log p_m(s)\} = \lambda p_m(s)$$

for $1 \leqslant s \leqslant m-1$, and

$$E\{f(m) \mid D\} - 2\beta\{\log p_m(m) - \log p_m(m-1)\} = \lambda\, p_m(m).$$

Summing these equations, all the terms involving β add up to zero and we are left with $\sum_{s=0}^{m} E\{f(s) \mid D\} = \lambda \sum_{s=0}^{m} p_m(s)$, or in other words, $\lambda = C$, the total number of clusters. Solving the above set of equations with $\lambda = C$ leads to the updated estimates $\hat{p}_m^{(t+1)}(0), \cdots, \hat{p}_m^{(t+1)}(m)$. To solve these equations, we can use the Newton-Raphson method or we can just carry out one Newton-Raphson iteration in the spirit of the gradient EM algorithm proposed by Lange (1995) to save computation.

To choose β which controls the amount of smoothing, we can use likelihood cross validation. Specifically, we choose β to maximize

$$\ell_{cv}(\beta) = \sum_{i=1}^{C} \log p_{n_i}\left(s_i; \theta = \hat{\theta}_{(-i)}(\beta)\right),$$

where for a fixed β, $\hat{\theta}_{(-i)}(\beta)$ is the penalized likelihood estimate of θ obtained after deleting cluster i. In other words, $\hat{\theta}_{(-i)}(\beta)$ is the maximizer of

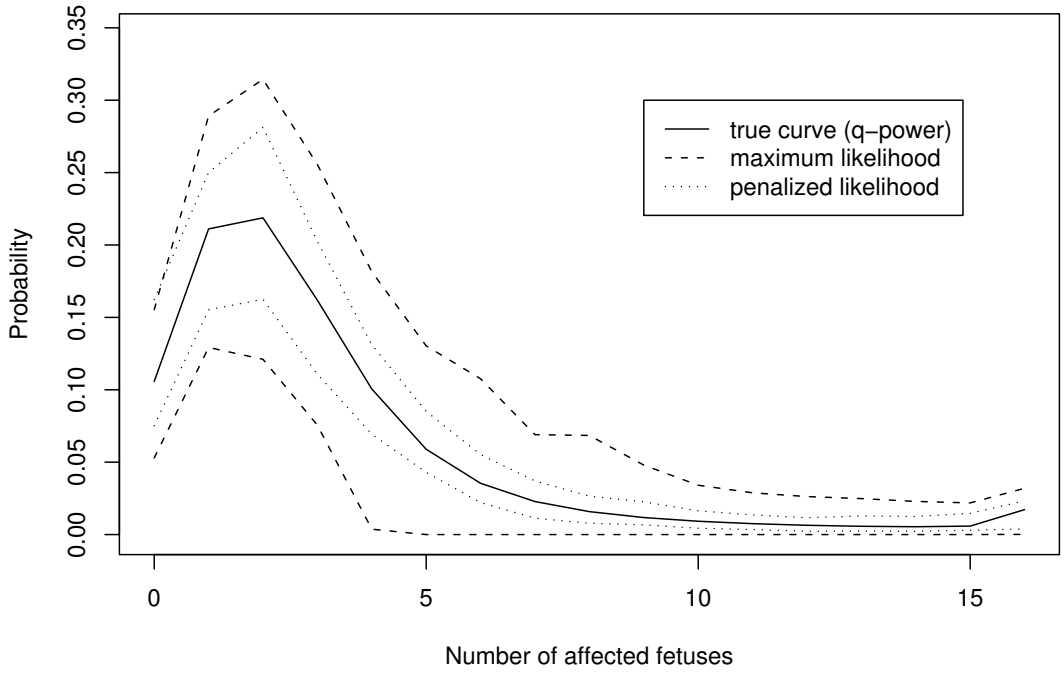
$$l_\beta^{(-i)} = \sum_{j \neq i} \log p_{n_j}(s_j) - \beta \sum_{s=0}^{m-1} \{\log p_m(s+1) - \log p_m(s)\}^2.$$

4.2 Numerical and Simulation Results

The penalized likelihood estimates with β chosen by cross validation for the three data sets used by Brooks *et al.* (1997) are included in Figure 4.1. It can be seen

from Figure 4.1 that the penalized likelihood estimates succeed in smoothing out the jaggedness and irregularity that are apparent in the nonparametric maximum likelihood estimates.

To compare the sampling distributions of the smoothed and unsmoothed estimates, a simulation study was conducted using the same design as the one reported in chapter 3, i.e., data were simulated for 200 litters 200 times using either a beta-binomial or q-power distribution with marginal probability $p = 0.2$ and intra-litter correlation $\rho = 0.2$. The lower and upper curves in Figure 4.2 depict the lower and upper 5^{th} percentile of the 200 sample estimates of $P(S_{16} = s)$, $s = 0, \ldots, 16$, for a litter of size 16. It is clear that the band for the smoothed estimates is much narrower which is a manifestation of how smoothing can reduce variability. It is also clear that without smoothing, the estimates are too rough.



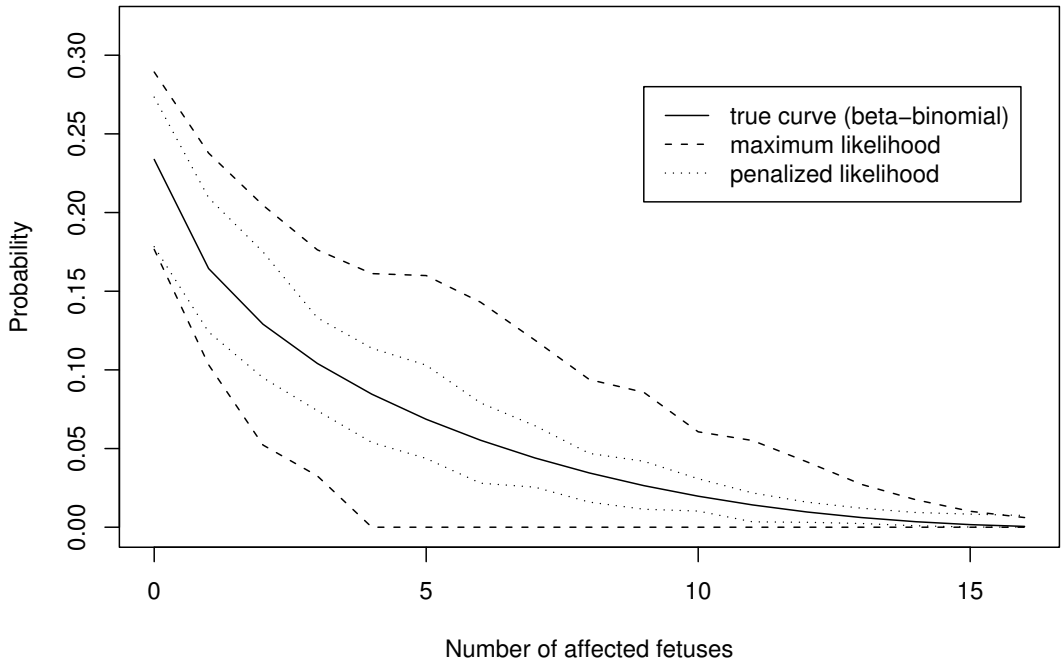


Figure 4.2: *Empirical upper and lower 5-percentiles of the saturated model maximum likelihood and maximum penalized likelihood estimates.*

Chapter 5

Combining Kernel Smoothing with Penalized Likelihood

In chapter 3, we applied the saturated model to the analysis of clustered binary data. It is illustrated by six no dose data sets. However, typical data sets from developmental toxicity studies are composed of one control group and several dose groups. If a covariate such as dose level is present, one would be interested in modelling how the response depends on the covariate. One means to analyze these dose response data is to use the local likelihood estimation. The concept of local likelihood estimation was first introduced by Tibshirani and Hastie (1987). Staniswalis (1989) used a kernel weighted likelihood to get the estimators. For the choosing of underlying likelihood function, one can either use the beta-binomial, q-power, shared response, or other parametric distributions. However, from Table 3.1, we

can see that although the beta-binomial distribution was rejected two times out of the six data sets, the log-likelihood of the q-power distribution is not always bigger than that of the beta-binomial distribution. This suggests that no parametric models can perform uniformly well over other parametric models. Therefore, we shall adopt the saturated model, the most general probability function assuming marginal compatibility, as the underlying distribution in this chapter. The robustness of the saturated model has been shown in Figure 4.2.

In the next section, we shall introduce the kernel weighted saturated model, which smooths the saturated model in the covariate space.

5.1 Kernel Weighted Saturated Model

Denote the observed data D by (n_i, s_i, x_i), $1 \leqslant i \leqslant C$, where n_i is the size of cluster i, s_i the sum of the exchangeable binary variables in cluster i, x_i the covariate value associated with cluster i, and C the total number of clusters. The aim here is to obtain the distribution of the response as a smooth function of the covariate without making parametric assumptions. We will fit a saturated model, for a given x value, on the basis of the observed data, by maximizing the following kernel-weighted local log-likelihood

$$\ell_h = \sum_{i=1}^{C} K\left(\frac{x - x_i}{h}\right) \log p_{n_i}(s_i; \theta), \tag{5.1}$$

where $K(.)$ is a kernel function, which is taken to be standard normal in this thesis. The maximization is with respect to the parameters $\theta = \{p_m(0), \ldots, p_m(m)\}$ which we have argued in Chapter 3 to be the appropriate parameterization for the saturated model. Note also that we have adopted the notation $p_{n_i}(s_i;\theta)$ to emphasize that $p_{n_i}(s_i)$ is a function of $p_m(0), \ldots, p_m(m)$ via (3.7). Again, an EM type algorithm can be used to maximize (5.1) by augmenting the data from s_i to $r_i = (s_i, u_i)$, with u_i unobserved, so that all the litters are of size m after augmentation. Using (3.6), we can evaluate $P(r_i = s|s_i;\ \hat{\theta}^{(t)})$, the conditional probability that $r_i = s$ given the observed s_i, evaluated at the current estimate $\hat{\theta}^{(t)}$. The updated estimates are given by

$$\hat{p}_m^{(t+1)}(s) = \frac{\sum_{i=1}^{C} K\left(\frac{x-x_i}{h}\right) P(r_i = s|s_i;\ \hat{\theta}^{(t)})}{\sum_{i=1}^{C} K\left(\frac{x-x_i}{h}\right)},$$

for $s = 0, \ldots, m$.

To choose the smoothing parameter h, we again use cross-validation by maximizing

$$\ell_{cv}(h) = \sum_{i=1}^{C} \log p_{n_i}\Big(s_i;\ \theta = \hat{\theta}_{(-i)}(x_i, h)\Big), \tag{5.2}$$

where, for a given h, $\hat{\theta}_{(-i)}(x_i, h)$ is the maximizer of the kernel likelihood

$$\ell_h^{(-i)} = \sum_{j\neq i} K\left(\frac{x_i - x_j}{h}\right) \log p_{n_j}(s_j;\theta)$$

evaluated at $x = x_i$ after deleting cluster i.

Note that the above kernel smoothing method is applicable to data with unequal cluster sizes. As an illustration, we consider the 2,4,5-T data analyzed previously by George and Bowman (1995), Dominici and Parmigiani (2001), Kuk (2004) and Pang and Kuk (2005), among others. For this data set, there are six dose groups corresponding to exposure levels of 0, 30, 45, 60, 75 and 90 mg/kg of the herbicide 2,4,5-T that was given to pregnant mice during day 6 to day 14 of gestation. In our analysis, the litter size is the number of implantation sites, and the toxicity endpoint is the number of fetal deaths, resorptions and cleft palate malformations. A listing of the data can be found in George and Bowman (1995). It can be seen from Figure 5.1 that the estimates of the marginal fetal response probability and intra-litter correlation obtained using the kernel method are fairly smooth functions of the dose level.

5.2 Penalized Kernel Method

In Chapter 4, we have already seen that the saturated model can exhibit a lot of roughness due to the sparseness of the data sets. This suggests that the kernel weighted saturated model may need some smoothing too. Figure 5.2 shows the estimated probability functions (for the number of response in a litter of size 21) at the 6 dose groups, we can see that they are all very erratic and are in need of smoothing. Thus we need to smooth in the response space as well as across covariates. This can be done by combining kernel smoothing (5.1) with the penalty

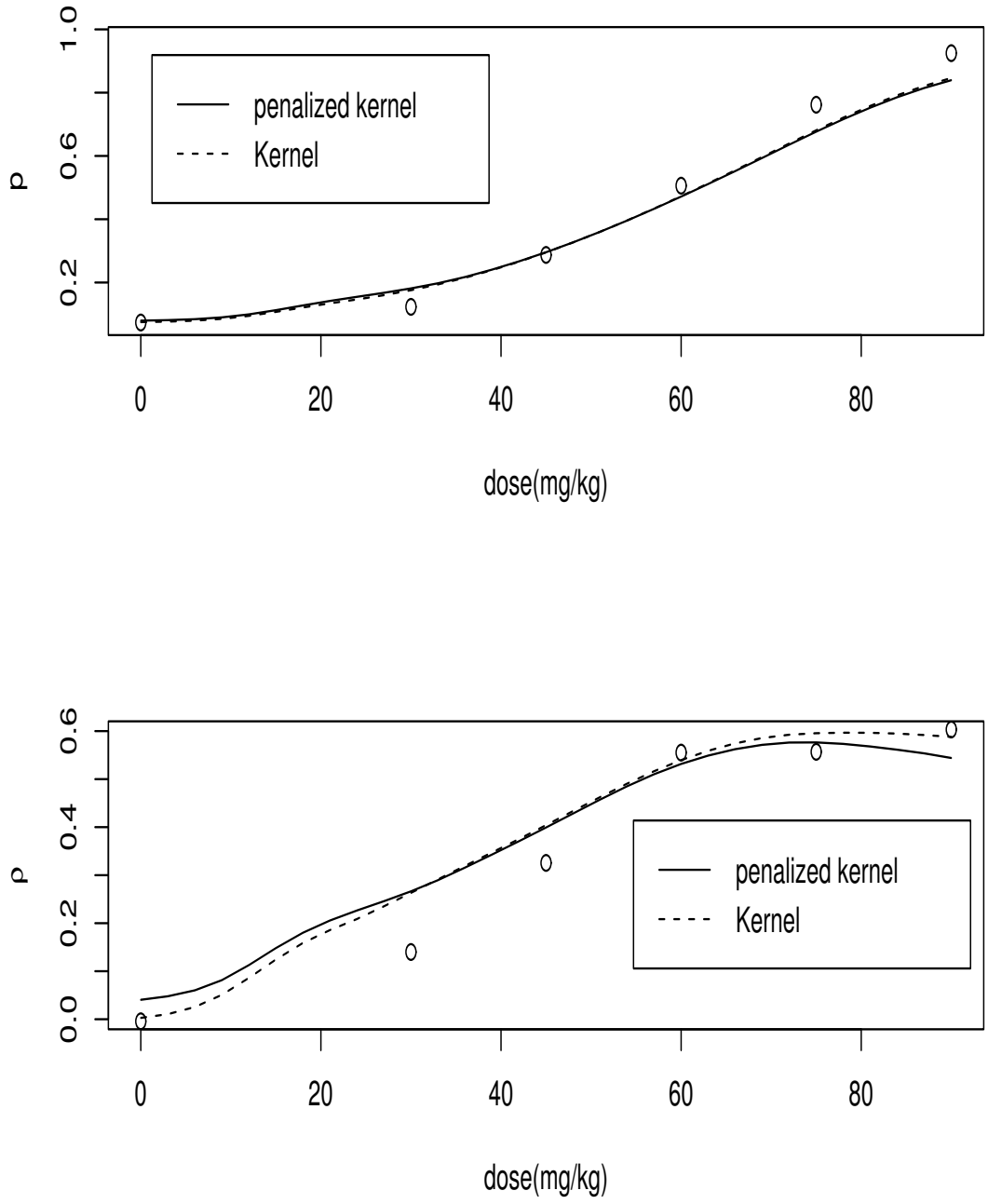


Figure 5.1: *Kernel likelihood and penalized kernel estimates of the marginal probability and intra-litter correlation for the 2,4,5-T data*

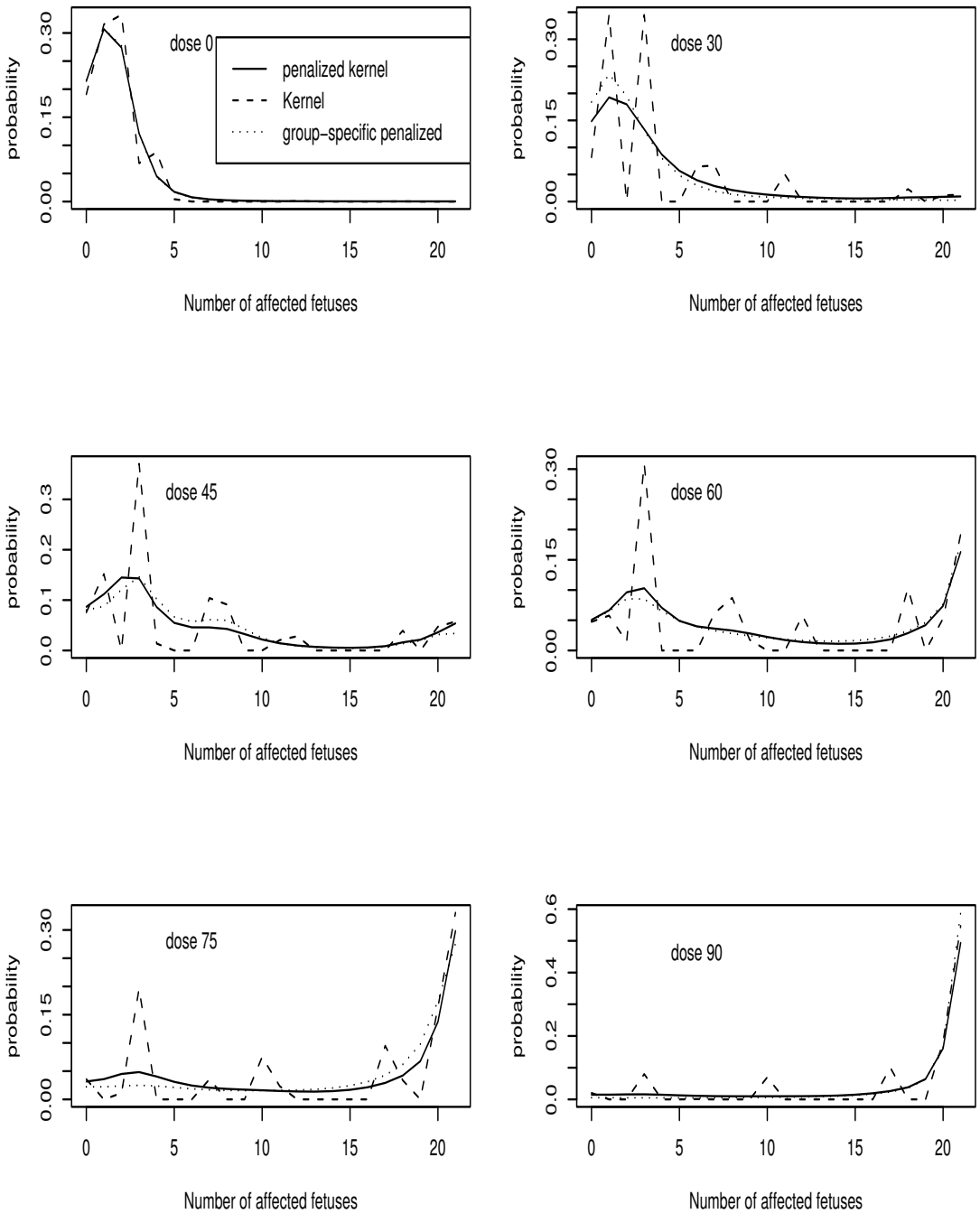


Figure 5.2: *Kernel likelihood, penalized kernel and group-specific penalized likelihood estimates of the probability function constructed from the 2,4,5-T data for a litter of size 21 at 6 different dose levels*

approach introduced in Chapter 4. The resulting penalized kernel method can be described as follows.

Begin by choosing the smoothing parameter h for kernel smoothing by cross-validation as in (5.2). With h fixed at the selected value, the probability function $\theta = \{p_m(0), \ldots, p_m(m)\}$ at a given x value can be estimated by maximizing the following penalized kernel-weighted log-likelihood

$$\ell_\beta = \sum_{i=1}^{C} K\left(\frac{x - x_i}{h}\right) \log p_{n_i}(s_i; \theta) - \beta \sum_{s=0}^{m-1} \{\log p_m(s+1) - \log p_m(s)\}^2,$$

where β controls the amount of smoothing along the response space. The maximization can again be done using an EM type algorithm similar to Chapter 4. The only difference is that the original frequencies become kernel weighted.

As illustrated by the 2,4,5-T example, the degree of sparseness of data can vary considerably between different dose groups and so β has to be chosen locally. Our suggestion is to choose β for a given x value by maximizing the following kernel-weighted cross validation criterion

$$\ell_{cv}(\beta) = \sum_{i=1}^{C} K\left(\frac{x - x_i}{h}\right) \log p_{n_i}\left(s_i;\ \theta = \hat{\theta}_{(-i)}(x_i, \beta)\right),$$

where, for a given β, $\hat{\theta}_{(-i)}(x_i, \beta)$ maximizes

$$\ell_\beta^{(-i)} = \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right) \log p_{n_j}(s_j; \theta) - \beta \sum_{s=0}^{m-1} \{\log p_m(s+1) - \log p_m(s)\}^2.$$

The results of applying the above penalized kernel method to the 2,4,5-T data are also shown in Figures 5.1 and 5.2. From Figure 5.1, we can see that as far as

the marginal probability and intra-litter correlation are concerned, the penalized kernel method leads to estimates that are as smooth in the dose level as the kernel method. However, when we look at the probability functions at the 6 dose groups, we can see in Figure 5.2 that the penalized kernel method manages to smooth away the jaggedness of the estimates produced by the kernel method alone and are in fact very close to the group-specific penalized likelihood estimates. Thus the penalized kernel method seems to enjoy the best of both worlds.

Chapter 6

Summary, Conclusion and Further Work

6.1 Summary and Conclusion

In this thesis, we have proposed a shared response model that, like the q-power distribution, is not prone to inflating the probability of observing no affected fetuses within a litter. Results of our simulation study show that the EM estimates are nearly unbiased and the associated confidence intervals based on the usual standard error estimates have coverage close to the nominal level. Simulation results also suggest that the shared response model estimates of the marginal malformation probabilities are robust to misspecification of the distributional form, but not so for the estimates of intralitter correlation and the litter-level probability of having

at least one malformed fetus. This is an inherent problem of the method of maximum likelihood and is not peculiar to the shared response model. When applied to the 2,4,5-T data, the shared response model gives results similar to the q-power model and both out-perform other models proposed in the literature. An advantage of the shared response model over the q-power distribution is that it is more interpretable. It can also be extended to the multivariate case more easily. We generalized the beta-binomial and shared response models to the bivariate case. A nice property of these two bivariate models is that the marginal distributions are just their respective univariate counterparts with corresponding parameters. These two models can also be easily generalized to higher dimensions in similar manner.

The marginal compatibility assumption is very crucial for exchangeable binary data, we give a rectified trend test statistic in this thesis. The p-value of our statistic is very close to the bootstrap results.

The shared response model adds one more option in the analysis of exchangeable binary data. Meanwhile, model selection becomes more urgent. By fitting the saturated model, we can assess the goodness of fit of these parametric models. A new nonparametric estimator of the intralitter correlation is also proposed based on the saturated model. Simulation studies show that this new estimator performs on par with the best estimators proposed in the literature. We also extend the penalized likelihood method to the case of varying cluster sizes and implement it using an EM type algorithm. Simulation shows that smoothing has reduced the

variation significantly.

In the presence of covariates, a kernel method is often adopted to smooth the data in the covariate space, and we finally combine the kernel smoothing with penalized likelihood to perform smoothing in both the covariate and response space. This penalized kernel method seems to do well in achieving smoothness in the response space as well as across covariates.

6.2 Further Work

There is much work to be done in the analysis of exchangeable binary data. An alternative parametric model not considered in this thesis is to use the exponential family model (Molenberghs and Ryan, 1999; Geys *et al.*, 1999). The advantages of this class of models are the unconstrained parameter space, the modelling flexibility, and the ease in estimation if one is willing to use pseudolikelihood to avoid the computation of normalizing constants. The exponential family model, however, is conditional in nature with no closed form formulae for the marginal response probability or the unconditional odds ratio. Moreover, the model is not "reproductive" (Prentice, 1988), in the sense that if $Y_1, Y_2, \ldots, Y_n$ follow the exponential family model, then the marginal distribution of a proper subset of $Y_1, Y_2, \ldots, Y_n$ will not be of the same form. The shared response model and other parametric models in this thesis focus on models that can be parameterized in terms of the marginal

response probability and unconditional odds ratio.

For the smoothing of the saturated model, we used penalized likelihood. Other methods for smoothing discrete data will also be investigated. A key assumption commonly made which allows us to link up the distributions for different cluster sizes so that estimation can be based on the combined data across all cluster sizes is the assumption of reproducibility or compatibility of marginal distributions. We have proposed a modified trend test in this thesis. That test is only a test that the marginal fetal response probability does not depend on cluster size. More generally, one may want to test whether the second and higher order marginal distributions depend on cluster size or not. Another ad hoc way to test the marginal compatibility assumption in general is to stratify the clusters into small and large clusters to see if there are significant differences between the stratum specific estimates. Further work is needed to develop a more systematic and optimal approach for testing the marginal compatibility assumption.

References

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. M. (2002). *Topics in Modelling of Clustered Data*. New York: Chapman and Hall.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics* **27**, 162–167.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

Bowman, D., Chen, J. J. and George, E. O. (1995). Estimating variance functions in developmental toxicity studies. *Biometrics* **51**, 1523-1528.

Bowman, D. and George, E. O. (1995). A saturated model for analyzing exchangeable binary data: Applications to clinical and developmental toxicity study. *Journal of the American Statistical Association* **90**, 871–879.

Brooks, S. P., Morgan, B. J. T., Ridout, M. S. and Pack, S. E. (1997). Finite mixture models for proportions. *Biometrics* **53**, 1097–1115.

Catalano, P. J., Ryan, L. M. and Scharfstein, D. (1994). Modelling fetal death and malformation in developmental toxicity. *Risk Analysis* **14**, 611–619.

Chen, J. J. and Kodell, R. L. (1989). Quantitative risk assessment for teratological effects. *Journal of the American Statistical Association* **84**, 966–971.

Conaway, M. R. (1990). A random effects model for binary data. *Biometrics* **46**, 317–328.

Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology* **4**, 854–871.

Dominici, F. and Parmigiani, G. (2001). Bayesian semiparametric analysis of developmental toxicology data. *Biometrics* **57**, 150–157.

Faustman, E. M., Allen, B. C., Kavlock, R. J. and Kimmel, C. A. (1994). Dose response assessment for developmental toxicity. I. Characterization of database and determination of no observed adverse effect levels. *Fundamental and Applied Toxicology* **23**, 478–486.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Volume II, 2nd ed. New York: Wiley.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgements: Unequal numbers of judges per subject. *Applied Psychological Measurement* **3**, 537–542.

George, E. O. and Bowman, D. (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics* **51**, 512–523.

Geys, H., Molenberghs, G. and Ryan, L. (1999). Pseudolikelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association* **94**, 734–745.

Haseman, J. K. and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **35**, 281–293.

Holston, J. F., Gaines, T. B., Nelson, C. J., LaBorde, J. B., Gaylor, D. W., Sheehan, D. M. and Young, J. F. (1991). Developmental toxicity of 2,4,5-trichlorophenoxiacetic acid I: Multireplicated dose response studies in four inbred strains and one outbred stock of mice. *Fundamental and Applied Toxicology* **19**, 286–297.

Kimmel, C. A. and Gaylor, D. W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Analysis* **8**, 15–20.

Kuk, A. Y. C. (2003). A generalised estimating equation approach to modelling foetal response in developmental toxicity studies when number of implants is dose-dependent. *Applied Statistics* **52**, 51–61.

Kuk, A. Y. C. (2004). A litter-based approach to risk assessment in developmental toxicity studies via a power family of completely monotone functions. *Applied Statistics* **53**, 369–386.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34**, 69–76.

Lange, K. (1995). A gradient EM algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B* **57**, 425–437.

Lefkopoulou, M. and Ryan, L. (1993). Global tests for multiple binary outcomes. *Biometrics* **49**, 975–988.

Liang, K. Y. and Hanfelt, J. (1994). On the use of Quasi-likelihood method in teratological experiments. *Biometrics* **50**, 872–880.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *B*iometrika, **7**3, 13–22.

Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

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.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

Lunn, A. D. and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika* **85**, 487–490.

Molenberghs, G. and Ryan, L. M. (1999) An exponential family model for clustered multivariate binary data. *Environmetrics*, **10**, 279–300.

Mosimann, J. E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49**, 65–82.

Pang, Z. and Kuk, A. Y. C. (2005). A shared response model for clustered binary data in developmental toxicity studies. *Biometrics* **61**, 1076–1084.

Pang, Z. and Kuk, A. Y. C. (2005). Test of marginal compatibility and smoothing methods for exchangeable binary data with unequal cluster sizes. under revision to *Biometrics*.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

Ridout, M. S., Demétrio, C. G. B. and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55**, 137–148.

Ryan, L. (1992). Quantitative risk assessment for developmental toxicity. *Biometrics* **48**, 163–174.

Simonoff, J. S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Annals of Statistics* **11**, 208–218.

Staniswalis, J. G. (1989). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559–568.

Stefanescu, C. and Turnbull, B.W. (2003). Likelihood inference for exchangeable binary data with varying cluster sizes. *Biometrics* **59**, 18–24.

Tibshirani, R. and Hastie, T. (1987). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* **84**, 276–283.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949–952.

Xu, J. and Prorok, P.C. (2003). Modelling and analyzing exchangeable binary data with random cluster sizes. *Statistics in Medicine* **22**, 2401–2416.

Yee, T. W. and Wild, C. J. (1996). Vector Generalized Additive Models. *Journal of the Royal Statistical Society, Series B* **58**, 481–493.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zou, G. and Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics* **60**, 807–811.