



**ATHENS UNIVERSITY  
OF ECONOMICS AND BUSINESS**

**DEPARTMENT OF STATISTICS**

**POSTGRADUATE PROGRAM**

**RANDOM EFFECTS MODELS  
FOR BINARY DATA**

By

**Androniki P. Matsioulas**

A THESIS

Submitted to the Department of Statistics  
of the Athens University of Economics and Business  
in partial fulfilment of the requirements for  
the degree of Master of Science in Statistics

Athens, Greece  
2004



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ  
ΚΑΤΑΛΟΓΟΣ



0 000000 524285



ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ  
ΒΙΒΛΙΟΘΗΚΗ  
ΣΕΒ. 76396  
ΑΠΛ.  
ΤΟΣ



**ATHENS UNIVERSITY  
OF ECONOMICS AND BUSINESS**

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

**RANDOM EFFECTS MODELS  
FOR BINARY DATA**

By

Androniki P. Matsioulas

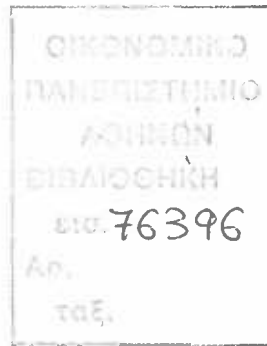


A THESIS

Submitted to the Department of Statistics  
of the Athens University of Economics and Business  
in partial fulfilment of the requirements for  
the degree of Master of Science in Statistics

Athens, Greece  
December 2003





# ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

## ΜΟΝΤΕΛΑ ΤΥΧΑΙΩΝ ΕΠΙΔΡΑΣΕΩΝ ΓΙΑ ΔΙΩΝΥΜΙΚΑ ΔΕΔΟΜΕΝΑ

Ανδρονίκη Π. Ματσιούλα

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής  
του Οικονομικού Πανεπιστημίου Αθηνών  
ως μέρος των απαιτήσεων για την απόκτηση  
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική



Αθήνα  
Δεκέμβριος 2003





**ATHENS UNIVERSITY  
OF ECONOMICS AND BUSINESS  
DEPARTMENT OF STATISTICS**

A Thesis submitted in partial fulfillment of  
the requirements for the degree of  
Master of Science

**RANDOM EFFECTS MODELS  
FOR BINARY DATA**

Androniki P. Matsioula



**Approved by the Graduate Committee**

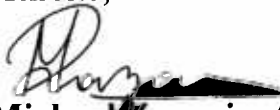
D. Karlis  
Lecturer  
Thesis Supervisor

V. Vasdekis  
Assistant Professor

P. Tsiamyrtzis  
Lecturer

Members of the Committee

**Athens, March 2004**

  
**Michael Lazanis, Associate Professor  
Director of the Graduate Program**



## ACKNOWLEDGEMENTS

I would like to thank my parents Paul and Helen, whose financial support made this contribution possible. I would also like to thank Dr. Dimitris Karlis, Lecturer of the Department of Statistics, for his supervision and assistance during the preparation of this dissertation. Finally, I thank my friends at the university for their psychological support and especially my fellow students. This class of MSc in Statistics will be unforgettable. I wish them all success in the future.





## VITA

I was born in Athens in March 1978. In 1996 I entered the Department of Mathematics in the University of Patras and in October 2001 I received my degree. In October 2001 I was accepted from Department of Statistics of Economic and Business University of Athens to follow the MSc Program in Statistics.





## ABSTRACT

Androniki Matsioulas

### RANDOM EFFECTS MODELS FOR BINARY DATA

December 2003

In this dissertation we present different ways for dealing with the problem of overdispersion in the generalized linear models. In order to accommodate the extra-binomial variation, which may be observed on fitting a binomial model, we can either add unobserved random effects to the linear predictor or consider binomial mixture distributions. Specifically, we perform the Penalized Quasi-Likelihood approach, the Gauss-Hermite technique and the Beta-Binomial model by using the SAS macro GLIMMIX, the statistical program MIXOR and the SAS procedure LOGISTIC, respectively. Applying the above methods to three different data sets, our aim is to explore any differences on the results between, or within, the three programs. We observe small differences between programs, because each program applies its own method of analysis. However, we underline that the number of quadrature points in the Gauss-Hermite algorithm plays a major role on whether a global maximum or local maxima are obtained at convergence for the log-likelihood, and therefore we must be cautious with their choice. Many authors suggest that we must run the algorithm several times with different starting values each time, hoping that if they all give the same solution, then it is likely to be the global maximum. Finally, in the last Chapter we apply the three above methods to data from twenty-three Greek industrial pig farms.





## ΠΕΡΙΛΗΨΗ

Ανδρονίκη Ματσιούλα

### ΜΟΝΤΕΛΑ ΤΥΧΑΙΩΝ ΕΠΙΔΡΑΣΕΩΝ ΓΙΑ ΔΙΩΝΥΜΙΚΑ ΔΕΔΟΜΕΝΑ

Δεκέμβριος 2003

Σε αυτήν τη διπλωματική παρουσιάζουμε διάφορους τρόπους αντιμετώπισης για το πρόβλημα της υπερδιασποράς (overdispersion) στα γενικευμένα γραμμικά μοντέλα. Προκειμένου να λάβουμε υπόψιν μας την έξτρα-διωνυμική μεταβλητότητα, η οποία ενδέχεται να παρατηρηθεί κατά την προσαρμογή ενός διωνυμικού μοντέλου, μπορούμε είτε να προσθέσουμε μη-παρατηρήσιμες τυχαίες επιδράσεις στη γραμμική σχέση ή να θεωρήσουμε μίξεις της διωνυμικής κατανομής. Συγκεκριμένα, παρουσιάζουμε την Penalized Quasi-Likelihood μέθοδο, την Gauss-Hermite τεχνική και το μοντέλο της Βήτα Διωνυμικής κατανομής χρησιμοποιώντας την SAS μακροεντολή GLIMMIX, το στατιστικό πρόγραμμα MIXOR και την SAS διαδικασία LOGISTIC, αντίστοιχα. Εφαρμόζοντας τις παραπάνω μεθόδους σε τρία διαφορετικά σετ δεδομένων, στόχος μας είναι να εξετάσουμε οποιεσδήποτε διαφορές στα αποτελέσματα μεταξύ, ή εντός, των τριών προγραμμάτων. Παρατηρούμε μικρές διαφορές μεταξύ των προγραμμάτων, οι οποίες οφείλονται στο ότι κάθε πρόγραμμα χρησιμοποιεί διαφορετική μέθοδο ανάλυσης. Παρόλα αυτά, τονίζουμε ότι ο αριθμός των quadrature σημείων στον αλγόριθμο Gauss-Hermite παίζει σημαντικό ρόλο στο εάν η λογαριθμική πιθανοφάνεια συγκλίνει σε ένα ολικό μέγιστο ή σε τοπικά μέγιστα, και για αυτό πρέπει να είμαστε προσεκτικοί όσον αφορά την επιλογή τους. Πολλοί συγγραφείς προτείνουν να τρέξουμε τον αλγόριθμο αρκετές φορές με διαφορετικές αρχικές τιμές κάθε φορά, ελπίζοντας ότι εάν όλες οι επαναλήψεις δώσουν την ίδια λύση, τότε είναι πολύ πιθανόν αυτή να αποτελεί και το ολικό μέγιστο. Τέλος, στο τελευταίο κεφάλαιο εφαρμόζουμε τις τρεις παραπάνω μεθόδους σε δεδομένα από εικοσιτρείς Ελληνικές βιομηχανικές φάρμες χοιριδίων.





## TABLE OF CONTENTS

<b>1 Introduction</b>	1
<b>2 The generalized linear mixed model</b>	7
2.1 Introduction	7
2.2 The generalized linear mixed model (GLMM)	8
2.3 Penalized Quasi-Likelihood	10
2.3.1 Estimation of fixed and random effects	10
2.3.2 Variance component estimation	14
2.3.3 Iterative procedure for estimating the mean and variance parameters	15
2.4 Marginal Quasi-Likelihood	16
2.4.1 Estimation of fixed and random effects	16
2.4.2 Variance component estimation	18
2.4.3 Iterative procedure for estimating the mean and variance parameters	19
2.5 PQL compared to MQL and some further remarks	19
2.6 The EM algorithm for estimating the finite mixture model	20
2.7 Other approaches for the GLM with normal random effects	25
2.8 NPML estimate of the mixing distribution	26
2.8.1 NPML estimation of the masses and mass-points	26
2.8.2 Further remarks on the NPML estimate	29
2.8.3 Computational issues	30
<b>3 Mixed Binomial Distributions in the Presence of Extra-Binomial Variation</b>	33
3.1 Introduction	33



3.2	Binomial mixture distributions for adjusting for Overdispersion	34
3.2.1	The specific case of the Beta-Binomial distribution	34
3.2.2	Estimating the parameters of the Beta-Binomial model	36
3.2.3	Other expansions of the Binomial distribution	40
3.3	Testing the goodness of fit of the Binomial distribution	42
<b>4</b>	<b>Effect of the number of quadrature points and comparison of the results between different programs: 3 examples</b>	<b>45</b>
4.1	Introduction	45
4.2	Description of the data for each example	46
4.3	Definition of the logistic random-intercept models	48
4.4	Detecting Overdispersion	49
4.5	Implication of the number K of quadrature points with MIXOR	50
4.6	Comparing the results between different programs	53
4.7	Conclusions	56
<b>5</b>	<b>Causes of mortality in piglets in Greek commercial pig farms</b>	<b>59</b>
5.1	Introduction	59
5.2	Some information about the way the sample of the pig farms and sows was selected	60
5.3	Description of the Data	60
5.4	Fitting the standard logistic linear model	62
5.5	Fitting the Beta-Binomial model (Williams' method)- Tarone's test	67
5.5.1	Results from estimating the Beta-Binomial model	67
5.5.2	Tarone's one-sided test of the Binomial distribution versus Beta-Binomial alternatives	71
5.6	Fitting the logistic random-intercept model	72
5.6.1	The Penalized Quasi-Likelihood approach	72
5.6.2	The Gauss-Hermite technique	75
5.7	Returning to the standard logistic regression model	76



<b>Appendix A: The exponential family and a review of the GLM theory</b>	81
Appendix A1: The Exponential family	81
Appendix A2: A brief review of the GLM theory	82
<b>Appendix B: How to use SAS for fitting the overdispersed GLM</b>	85
Appendix B1: The SAS procedure LOGISTIC (Williams' Method)	85
Appendix B2: The SAS macro GLIMMIX	86
<b>Appendix C: Gaussian Quadratures: The Gauss-Hermite formula</b>	91
<b>References</b>	95





## LIST OF TABLES

<b>Table 4.1:</b> <i>Data from Beitler and Landis (1985)</i>	47
<b>Table 4.2:</b> <i>Data from Efron (1986) and Francis et al. (1993)</i>	47
<b>Table 4.3:</b> <i>Data from Crowder (1978)</i>	47
<b>Table 4.4:</b> The residual deviances under the standard logistic linear model	50
<b>Table 4.5:</b> Effect of the number of quadrature points K on the output when fitting model (4.1) with MIXOR	50
<b>Table 4.6:</b> Effect of the number of quadrature points K on the output when fitting model (4.2) with MIXOR	51
<b>Table 4.7:</b> Effect of the number of quadrature points K on the output when fitting model (4.3) with MIXOR	53
<b>Table 4.8:</b> Effect of the number of quadrature points K on the output when fitting model (4.3) with MIXOR	53
<b>Table 4.9:</b> Fitting model (4.1) with different programs	54
<b>Table 4.10:</b> Fitting model (4.2) with different programs	55
<b>Table 4.11:</b> Fitting model (4.3) with different programs	56
<b>Table 4.12:</b> Fitting model (4.3) with different programs	56
<b>Table 5.1:</b> Categories of the response variable	62
<b>Table 5.2:</b> Analysis of Maximum Likelihood Estimates for model (5.1)	65
<b>Table 5.3:</b> Deviance and Pearson goodness-of-fit statistics for model (5.1)	66
<b>Table 5.4:</b> Model Fitting Information for model (5.2)	68
<b>Table 5.5:</b> Testing Global Null Hypothesis: BETA=0 for model (5.2)	69
<b>Table 5.6:</b> Analysis of Maximum Likelihood Estimates for model (5.2)	71
<b>Table 5.7:</b> Parameter Estimates for model (5.3)	74
<b>Table 5.8:</b> Random Effects Estimates for model (5.3)	75
<b>Table 5.9:</b> Deviance and Pearson goodness-of-fit statistics for model (5.4)	77



<b>Table 5.10:</b> Testing Global Null Hypothesis: $BETA=0$ for model (5.4)	77
<b>Table 5.11:</b> Analysis of Maximum Likelihood Estimates for model (5.4)	77
<b>Table 5.12:</b> 95% Confidence Intervals for the Odds Ratio	78



## LIST OF FIGURES

<b>Figure 5.1:</b> Pearson Residuals versus estimated linear predictors	66
<b>Figure 5.2:</b> Deviance Residuals versus Fitted values	79





## CHAPTER 1

### Introduction

Generalized Linear Models (GLMs) are fashionable and powerful modeling tools in statistical analysis. The GLM synthesizes many of the most commonly used statistical techniques for the analysis of both continuous and discrete data. Fitting of GLMs involve a number of assumptions that often are not adequately evaluated by the data analyst. The critical assumptions that underlie the GLM are: statistical independence of the observations, correct specification of the variance and link function, correct specification of the dispersion factor (for example, it is assumed to be equal to 1 for Poisson data), correct form for the explanatory variables and lack of influence of individual observations on the fit. Appendix A.2 contains a brief review of GLM theory and nomenclature. Failure of any of the above assumptions may lead to false conclusions. Moreover, another major assumption that underlies the use of GLMs is that the variance of the error distribution is completely determined by the mean. This assumption often fails. The first indication that something is wrong is that the deviance measure of goodness-of-fit of a model exceeds its degrees of freedom. This phenomenon is known as overdispersion.

The literature on overdispersion in GLMs is now quite extensive. The problem of overdispersion is easy to state but difficult to solve in any generality. McCullagh and Nelder (1989) say that overdispersion is the rule rather than the exception. Analysis of data via a single parameter family of distributions implies in particular that the variance is determined by the mean. Familiar examples are the Poisson, binomial and exponential distributions. A very common practical complication is the presence of overdispersion, or more rarely underdispersion, leading to a failure of the variance-mean relation. That is, given a standard exponential family GLM (detailed in Appendix A.1) with a specified variance-mean relationship, we observe on fitting the model that the variance is greater than that predicted by the mean,



observable in a large residual deviance or Pearson  $X^2$ . Thus, overdispersion (underdispersion) means that the data show evidence that the variance of the response  $y_i$  is greater (smaller) than expected under the given model, (e.g. binomial model), meaning that other (some) sources of variation, which have not been included (have been included) in the regression model, were present (were not present) in the data. Therefore, the methodology of the GLM, which allows only one source of randomness, (the random error term), is too restrictive to perform satisfactory data analyses of these more complex data. In contrast, the Generalized Linear Mixed Models (GLMM) are much more flexible, since they remove this restriction by adding unobserved random effects to the linear predictor, and thus allowing the 'extra variation' to be modeled, whenever it exists.

In order to explore further this limitation of the generalized linear models for analyzing more complex data, consider the case of the hierarchically structured data. In the social and other applied sciences, data have often a hierarchical structure in which 'units' at one 'level' are grouped within units at the next higher level. Thus, for example, students are nested within schools, citizens within nations, workers within workplaces and animals within litters. The existence of such 'clustering', presents particular problems of model specification due to lack of independence between measurements. The correlation among lower-level units from the same upper-level unit is higher than that from different upper-level units, (that is measured by the intra-class correlation coefficient). Consider as an example, the case of students (level 1) which are nested within schools (level 2). The correlation between variables measured on pupils from the same school will be higher than the correlation between variables measured on pupils from different schools. We would expect a positive correlation between exam scores for pupils in the same school, since pupils within a school share the same teachers, the same school environment and may live in the same or similar neighborhoods. By assuming independence among pupils, as the GLM do, the standard errors are underestimated and this might faulty lead to statistically significant results. In addition, considering independence between observations through the standard GLM, we do not take into account the 'extra variation' which is introduced due to the variation among upper-level



units, and thus a better mode of analysis should be chosen, in order to account for various sources of variability. We will discuss later in details that an appropriate mode of analysis under such circumstances of extra sources of variation is based on the framework of the GLMM.

As already mentioned, the nested structure of the responses induces an intra-class correlation between the lower-level responses on the same upper-level unit. A natural way of modeling this common variation is by adding a common unobserved random effect to the linear predictor for each lower-level unit in the same upper-level unit. Therefore, if binomial variability is assumed through the standard-independence GLM, then the random effect not only covers the correlation but also copes with overdispersion found in the data.

Motivated by the above discussion, we have restricted the scope of this dissertation to different ways for adjusting for overdispersion. It has long been recognized that the problem of overdispersion is usually confounded with the problem of omitted covariates (Orme, 1998; Fitzmaurice, Heath and Cox, 1997), or with measurement errors in the covariates for the experimental units (Prentice, 1986; Follmann and Lambert, 1989). For example, if Pearson's residuals tend to be too large on fitting a model, this does not mean necessarily that overdispersion is the cause. Omitting some important covariates may also cause large residuals. However, if we include all the available covariates related to the response in our model and it still does not fit, it might be because our regression function  $x_i^T \beta$  is incomplete. Or it might be due to overdispersion. Unless we collect more data, we cannot do anything about omitted covariates, but on the other hand we can attribute the lack of fit to overdispersion. Overdispersion can be handled in different ways. Some of them will be presented in this dissertation.

One way, is to add unobserved random effects to the linear predictor, as discussed in Chapter 2. A quite popular method for adjusting for overdispersion, assuming that the unobserved random effects are normally distributed, comes from the theory of quaslikelihood. Specifically, after a brief statement of the generalized linear mixed model (GLMM) in Section 2.2, the penalized quasi-likelihood (PQL) criterion is motivated in Section 2.3 by approximating the integrated quasi-likelihood. Quaslikelihood plays a very



important role in modern statistics, since it does not require specification of a full parametric model. In Section 2.4 an approximate GLM for the marginal distribution of the data is developed and the marginal quasi-likelihood (MQL) analysis is proposed in order to estimate the mean and variance parameters. Motivated by the above discussion, Section 2.5 gives a simple comparison between PQL and MQL. Section 2.6 is devoted to the description of an EM algorithm, which is initially derived as a form of Gaussian quadrature assuming a normal mixing distribution of the unobserved random effects. The algorithm provides an alternative analysis to approximate MQL and PQL analyses. Finally, in Section 2.7 we present some further approaches for fitting the GLM with normal random effects.

However, a disadvantage of the above approaches using specified parametric form for the mixing distribution of the unobserved random effects is the possible sensitivity of the conclusions to this specification. Heckman and Singer (1984) showed substantial changes in parameter estimates with quite small changes in mixing distribution specification; Davies (1987) showed similar effects. This difficulty can be avoided by nonparametric maximum likelihood (NPML) estimation of the mixing distribution concurrently with the structural model parameters. Section 2.8 gives detailed discussion of this approach. It considers the EM algorithm for the more general case. Particularly, the NPML method is applied for models where the mixing distribution cannot be specified in advance.

Apart from adding unobserved random effects to the linear predictor, another way to handle the problem of overdispersion is to specify a richer parametric model, where the distribution of the response variable is assumed to be something more dispersed than usual. The two most common examples of this are:

- Changing a binomial model to beta-binomial.
- Changing a Poisson model to negative binomial.

In Chapter 3, we concentrate on the first example. As known, when the response variable is a proportion, the standard logistic-linear model is well established for analyzing regression data, or data from a designed experiment. Even when all available explanatory variables have been fitted, the residual variation may be greater than can be attributed to the binomial sampling



variation assumed by the model, (this may be taken as an evidence of overdispersion). In this event we can either seek additional explanatory variables, or postulate a source of extra-binomial random variation between observations. The second choice is usually more realistic. For example, in toxicological studies, the experimental unit is often a litter of animals, and the response, is the proportion of animals in a litter exhibiting a certain abnormality. The incidence of the abnormality will vary between litters treated identically because they differ with respect to unrecorded genetic and environmental influences. Methods for analyzing such data are reviewed by Haseman and Kupper (1979).

Chapter 3 is organized as follows. Section 3.2 presents some binomial mixture distributions for handling the problem of overdispersion. Section 3.2.1 is devoted to the most popular binomial mixture distribution, the Beta-Binomial distribution, in order to accommodate extra-binomial variation. Specifically, to allow modeling this extra-binomial variation, we take the parameters of the binomial distribution  $p_i$ , as beta-distributed variables on  $(0,1)$ . The main difference from the ordinary binomial distribution is that the probability of a success is not constant. In contrast, we make probabilities  $p_i$  fluctuate by allowing their distributions to vary between the sets, meaning that  $p_i \sim \text{Beta}(a_i, b_i), a_i > 0, b_i > 0$ , where  $p_i$  are the same for all trials within a particular set. The obtainable response variable is beta-binomial distributed with variance that consists of two terms; the first term may be thought as the binomial component of the variance and the second term as the extra-binomial component. In Section 3.2.2, we go on presenting the maximum quasi-likelihood equations for estimating the Beta-Binomial models. Considering the Beta-Binomial model for adjusting for overdispersion is also referred as the Williams' method. Section 3.2.3 suggests other expansions of the binomial distribution and Section 3.3 proposes different ways in order to test the goodness of fit of the standard logistic regression model for binomial data.

Chapter 4 is devoted to the effect of the number  $K$  of quadrature points with the Gauss-Hermite method on the results. Furthermore, we apply the Gauss-Hermite technique, the Penalized Quasi-Likelihood (PQL) analysis and Williams' method for analysing three different data sets. Specifically, we



perform the Gauss-Hermite technique, the PQL analysis and Williams' method through the statistical package MIXOR, the SAS (Statistical Analysis System) macro GLIMMIX and the SAS procedure LOGISTIC, respectively. The results obtained from the three methods are then compared.

In Chapter 5, some causes of mortality in piglets are examined on twenty-three industrial pig farms all over Greece. Section 5.3 gives description of all the available variables. The mortality rate is considered as the variable of interest. The purpose of the statistical proceeding is the determination of the factors that may influence the mortality rate through a mathematical model. In order to take into account the extra variation that is introduced due to the variation among the piggeries, we perform the Beta-Binomial model (Section 5.5), the penalized quasi-likelihood analysis (Section 5.6.1) and the Gauss-Hermite method (Section 5.6.2).



## CHAPTER 2

### The generalized linear mixed model

#### 2.1 Introduction

We have already stated that overdispersion is not uncommon in practice. In this Chapter, we present different ways in order to solve the problem of overdispersion by adding unobserved random effects to the linear predictor. Actually, we could separate this Chapter into two basic parts; Sections 2.2-2.7, which assume a normal mixing distribution and Section 2.8, which refers to the non-parametric maximum likelihood (NPML) estimation of the mixing distribution, i.e. without assuming any parametric form for the mixing distribution. More specifically, after introducing the form of the generalized linear mixed model (GLMM) in Section 2.2, we propose to use the quasi-likelihood functions. Sections 2.3 and 2.4 are devoted to the description of the penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) method, respectively, for handling the problem of overdispersion. As far as the PQL (MQL) criterion, in Section 2.3.1 (2.4.1) we describe how to estimate the fixed and random effects (mean parameters) in a GLMM, in Section 2.3.2 (2.4.2) we present the method for estimating the vector of the unknown variance components, in Section 2.3.3 (2.4.3) we give a brief description of the iterative algorithm and in Section 2.5 we mention some essential differences between the PQL and MQL analysis. Section 2.6 presents another algorithm, which is based on the Gauss-Hermite technique, based on Gaussian quadrature points. In particular, in Section 2.6 a normal mixing distribution is assumed, but with only slight variation the algorithm can be used for a completely unknown mixing distribution, giving a straightforward method for the fully NPML estimation of this distribution (Section 2.8). Moreover, Section 2.7 proposes other available approaches for estimating the GLMM with normal random effects.



## 2.2 The generalized linear mixed model (GLMM)

A generalized linear model with random effects is defined as follows: We denote the  $n$  observations of the response variable by  $y = (y_1, y_2, \dots, y_n)'$ . These are assumed to be observations of the random variables  $Y = (Y_1, Y_2, \dots, Y_n)'$ , which have the same distribution from the exponential family. For each  $y_i$  there is a vector  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$  of  $p$  explanatory variables for the fixed effects, and a vector  $z_i = (z_{i1}, z_{i2}, \dots, z_{iq})'$  of  $q$  explanatory variables for the random effects. Given a  $q$ -dimensional vector  $b$  of random effects, the  $y_i$  are conditionally independent with means  $E(y_i | b) = \mu_i$  (or  $\mu_i^b$ ) and variances  $\text{var}(y_i | b) = \phi m_i \nu(\mu_i)$ , where  $\nu(\cdot)$  is a specified variance function,  $m_i$  is a known constant (e.g., for a random variable  $Y_i \sim \text{Binomial}(n_i, p_i)$ ,  $m_i = 1/n_i$ , where  $n_i$  is assumed known and represents the number of trials) and  $\phi$  is a dispersion parameter that may or may not be known. Usually the dispersion parameter is assumed to be the same for all observations. The relationship between the distribution of the response variable  $y_i$  and the linear predictor  $\eta_i$  is provided by the link function  $g$ ,  $g(\mu_i) = \eta_i, i = 1, \dots, n$ . Hence, the dependence of the distribution of the response on  $x_i$  and  $z_i$  is established as:

$$g(E(y_i | b)) = g(\mu_i) = \eta_i = x_i' a + z_i' b, \quad i = 1, \dots, n \quad (2.1)$$

where  $a$  is a  $p$  vector of fixed effects and  $b$  a  $q$  vector of random effects with  $b_j \sim N(0, \sigma_j^2), j = 1, \dots, q$ . Denoting the design matrices with rows  $x_i'$  and  $z_i'$  by  $X$  and  $Z$ , with dimensions  $n \times p$  and  $n \times q$  respectively, the conditional mean satisfies:

$$E(y | b) = h(Xa + Zb) \quad (2.2)$$

where  $h$  is the inverse of the link function  $g$ , i.e.  $h = g^{-1}$ .

The model is completed by the assumption that  $b$  has a multivariate normal distribution with mean 0 and covariance matrix  $D = D(\theta)$ ,  $b \sim N(0, D(\theta))$ .

For simplicity, we assume that the random effects  $b_j$  are independent, and

therefore the covariance matrix  $D$  is a  $q \times q$  diagonal matrix with diagonal elements  $\theta_j = \{\sigma_j^2\}$  ( $j = 1, \dots, q$ ), and  $\theta = (\sigma_1^2, \sigma_2^2, \dots, \sigma_q^2)$  is the vector of the unknown variances components. The random effects  $b$  are also assumed to be uncorrelated with the vector of the random errors  $e = y - E(y|b)$ . Moreover, as for Normal, Poisson or Binomial data the variance function contains a constant nuisance parameter equal to 1, we also consider here that the dispersion parameter  $\phi$  is fixed at unity. This assumption is necessary in order to attribute the extra variation that may be observed only to the unobserved random effects, which are added to the linear predictor. In other applications, however,  $\phi$  may be estimated together with  $\theta$  as a parameter in the covariance matrix of the marginal distribution of  $y$ .

Next, in order to understand how the GLMM is appropriate for analysing hierarchically structured data, we consider Crowder's study (1978) on the proportion of seeds that germinated on each of 21 plates arranged according to a 2x2 factorial layout by seed variety and type of root extract. He noted that the within-group variation exceeded the one predicted by the binomial sampling theory, which made him conclude that the logistic regression analysis of treatment and interaction effects should account appropriately for such overdispersion. One way of accounting for the extraneous plate-to-plate variability in this situation is by means of a GLMM that has linear predictor:

$$\eta_i = \text{logit}(\Pr[y_i = 1 | x_i, b_i]) = \text{logit}(p_i) = \ln(p_i/(1 - p_i)) = x_i' a + b_i, \quad i = 1, \dots, 21$$

where  $p_i$  represents the proportion of seeds that germinated on each of 21 plates,  $a$  represents the fixed effects associated with seed, root extract and their interaction and  $b_i$  represent random effects associated with each plate. The random effects  $b_i$  are assumed to be independently normally distributed with common variance, i.e.  $b_i \sim N(0, \sigma^2)$ . Further details for Crowder's data are given later in Sections 4.2 and 4.3 of Chapter 4.



### 2.3 Penalized Quasi-Likelihood

#### 2.3.1 Estimation of fixed and random effects

In the previous section, we defined the relationship between the mean and variance of  $y_i | b$  as given below:

$$E(y_i | b) = \mu_i = g^{-1}(x_i' a + z_i' b) \text{ and } \text{var}(y_i | b) = \phi m_i v(\mu_i)$$

The distribution of  $y_i | b$  is not fully specified, but the structure of its mean and variance allows us to define quasi-likelihood equations. Particularly, the marginal likelihood, or integrated quasi-likelihood function used to estimate the parameter of interest  $l = (a, \theta)'$  is defined by:

$$Q(a, \theta) = \int Q(y | z_i' b, a) \phi(b | \theta) db \tag{2.3}$$

where  $\phi(b | \theta)$  is the prior distribution of the unobserved random effects  $b = (b_1, \dots, b_q)'$  and  $Q(y | z_i' b, a)$  is the conditional quasi-likelihood contribution of observed  $y$  given the unobserved random effects  $b$ . Since observations  $y_i$  are assumed to be independent conditional on  $z_i' b$ , we have:

$$Q(y | z_i' b, a) = \prod_{i=1}^n Q(y_i | z_i' b, a) \tag{2.4}$$

It holds that (see, for example, McCullagh and Nelder 1989, sec. 9.2, p.327):

$$\log\{Q(y_i | z_i' b, a)\} = -d_i(y_i; \mu_i)/2\phi \Rightarrow Q(y_i | z_i' b, a) = \exp\left(-\frac{d_i(y_i; \mu_i)}{2\phi}\right) \tag{2.5}$$

where  $d_i(y_i; \mu_i)$  is the quasi-deviance function corresponding to a single observation. Applying Equation (2.5) to (2.4) we have:

$$Q(y | z_i' b, a) = \prod_{i=1}^n Q(y_i | z_i' b, a) = \exp\left(-\sum_{i=1}^n \frac{d_i(y_i; \mu_i)}{2\phi}\right) \tag{2.6}$$

Furthermore,  $b_{q \times 1} \sim N(0, D = D(\theta)) \Rightarrow$

$$\phi(b | \theta) = (2\pi)^{-q/2} |D|^{-1/2} \exp\left(-\frac{b' D^{-1} b}{2}\right) \tag{2.7}$$

Applying now Equations (2.6) and (2.7) to (2.3) gives:



$$Q(a, \theta) = (2\pi)^{-q/2} |D|^{-1/2} \int \exp \left[ -\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i) - \frac{1}{2} b' D^{-1} b \right] db$$

or equivalently

$$e^{q(a, \theta)} \propto |D|^{-1/2} \int \exp \left[ -\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i) - \frac{1}{2} b' D^{-1} b \right] db \tag{2.8}$$

where  $q(a, \theta) = \log Q(a, \theta)$  and

$$d_i(y, \mu) = -2 \int_y^\mu \frac{y-u}{m_i \nu(u)} du$$

denotes the quasi-deviance measure of fit. We remind that the deviance statistic measures the distance between the observed values and predicted values under the family of distribution being tested. Furthermore, as the so-called saturated model fits the data perfectly (fitted and observed values are the same), we consider the deviance as a measure of comparison between the saturated model and the model that is tested (e.g. binomial model). If the saturated model does not provide a significantly better fit than the model being tested, we can conclude that the particular model is an acceptable fit to the data.

If conditionally on  $b$ , the observations are drawn from an exponential family distribution with variance function  $\nu(\cdot)$ , then the deviance is well known to be equal to the scaled difference  $2\phi[l(y; y, \phi) - l(y; \mu, \phi)]$ , where  $l(y; \mu, \phi)$  denotes the conditional log-likelihood of  $y$  given its mean  $\mu$ , (see, for example, McCullagh and Nelder 1989, sec. 2.1.2, p.24), and  $l(y; y, \phi)$  denotes the corresponding log-likelihood for the saturated model. We should however note that the difficulty in applying full likelihood inference lies in the integrations needed to evaluate the logarithm of the integrated quasi-likelihood function,  $q$ , and its partial derivatives. Since these integrations cannot be analytically calculated, we go on numerical integration using Laplace's approximation (Barndorff-Nielsen and Cox 1989, sec. 3.3; Tierney and Kadane 1986). Writing Equation (2.8) in the form  $c|D|^{-1/2} \int e^{-\kappa(b)} db$ , where:

$$\kappa(b) = \frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i) + \frac{1}{2} b' D^{-1} b,$$



we apply Laplace's method in order to overcome the problem of integration. More specifically, let  $\kappa'$  and  $\kappa''$  vector of dimension  $q$  and  $q \times q$  dimensional matrix of first- and second-order partial derivatives of  $\kappa$  with respect to  $b$ . Ignoring the constant  $c$ , the approximation yields:

$$q(a, \theta) \approx -\frac{1}{2} \log|D| - \frac{1}{2} \log|\kappa''(\tilde{b})| - \kappa(\tilde{b}) \tag{2.9}$$

where  $\tilde{b} = \tilde{b}(a, \theta)$  denotes the solution to

$$\kappa'(b) = -\sum_{i=1}^n \frac{(y_i - \mu_i)z_i}{\phi_{m_i, \nu}(\mu_i)g'(\mu_i)} + D^{-1}b = 0$$

that minimizes  $\kappa(b)$ . Differentiating again with respect to  $b$ , we have:

$$\kappa''(b) = \sum_{i=1}^n \frac{z_i z_i'}{\phi_{m_i, \nu}(\mu_i)[g'(\mu_i)]^2} + D^{-1} + R \approx Z'WZ + D^{-1} \tag{2.10}$$

where  $W$  is the  $n \times n$  diagonal matrix with diagonal terms  $w_i = \{\phi_{m_i, \nu}(\mu_i)[g'(\mu_i)]^2\}^{-1}$  that are recognizable as the GLM iterated weights (Firth 1991, p. 63; McCullagh and Nelder 1989, sec. 2.5). The remainder term,

$$R = -\sum_{i=1}^n (y_i - \mu_i)z_i \frac{\partial}{\partial b} \left[ \frac{1}{\phi_{m_i, \nu}(\mu_i)g'(\mu_i)} \right]$$

has expectation 0, since  $E(y_i - \mu_i) = 0$ , and is thus, in probability as a function of  $n$ , of lower order than the two leading terms in Equation (2.10). Furthermore,  $R$  equals 0 for the canonical link functions, since  $g'(\mu) = \nu^{-1}(\mu)$  and  $\partial/\partial b[(\phi_{m_i})^{-1}] = 0$ . Ignoring  $R$  and applying Equation (2.10) to (2.9) leads to:

$$\begin{aligned} q(a, \theta) &\approx -\frac{1}{2} \log|D| - \frac{1}{2} \log|Z'WZ + D^{-1}| - \kappa(\tilde{b}) = -\frac{1}{2} \log|Z'WZ + D^{-1}| |D| - \kappa(\tilde{b}) \Rightarrow \\ q(a, \theta) &\approx -\frac{1}{2} \log|I + Z'WZD| - \frac{1}{2\phi} \sum_{i=1}^n d_i(y_i, \mu_i^{\tilde{b}}) - \frac{1}{2} \tilde{b}' D^{-1} \tilde{b} \end{aligned} \tag{2.11}$$

where  $\tilde{b}$  is chosen to maximize the sum of the last two terms.

Assuming that the GLM iterative weights  $w_i$  vary slowly (or not at all) as a function of the mean, we ignore the first term in this expression and choose  $a$  to maximize the second. Thus,  $(\hat{a}, \hat{b}) = (\hat{a}(\theta), \hat{b}(\theta))$ , where



$\hat{b}(\theta) = \tilde{b}(\hat{a}(\theta))$  is evaluated at  $\hat{a}$ , jointly maximize Green's (1987) Penalized Quasi-Likelihood (PQL):

$$-\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i, \mu_i) - \frac{1}{2} b' D^{-1} b \tag{2.12}$$

Differentiation with respect to  $a$  and  $b$  leads to the score equations for the mean parameters:

$$\sum_{i=1}^n \frac{(y_i - \mu_i)x_i}{\phi m_i \nu(\mu_i) g'(\mu_i)} = \sum_{i=1}^n (y_i - \mu_i)x_i w_i g'(\mu_i) = 0 \tag{2.13}$$

$$\sum_{i=1}^n \frac{(y_i - \mu_i)z_i}{\phi m_i \nu(\mu_i) g'(\mu_i)} = \sum_{i=1}^n (y_i - \mu_i)z_i w_i g'(\mu_i) = D^{-1}b \tag{2.14}$$

Equation (2.13) and (2.14) are non-linear and have no analytic solution. Therefore, we rely on numerical methods to solve them. Specifically, Green (1987) developed the Fisher scoring algorithm for solution of Equations (2.13) and (2.14) as an iterated weighted least squares (IWLS) problem involving a working dependent vector  $Y$  with components  $Y_i = \eta_i + (y_i - \mu_i)g'(\mu_i)$  and a diagonal weight matrix  $W$  that are updated at each iteration. Using Taylor's Theorem, the link function  $g(\cdot)$  applied to the original data  $y$  is linearized, giving to the first order:

$$g(y_i) = Y_i \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) = \eta_i + (y_i - \mu_i)g'(\mu_i) \tag{2.15}$$

Now, from (2.1) and (2.15) a linear random effect model for the working dependent variable  $Y$  is:

$$Y = Xa + Zb + eg'(\mu) = Xa + Zb + \varepsilon$$

Here, we have:

$$e = y - E(y|b), E(Y) = Xa, \text{cov}(b) = D, \text{cov}(\varepsilon) = \text{cov}\{eg'(\mu)\} = [g'(\mu)]^2 \text{cov}(e) = W^{-1} \text{ and consequently } \text{cov}(Y) = V = W^{-1} + ZDZ'$$

Green (1987) used the Fisher scoring algorithm in order to solve Equations (2.13) and (2.14). Particularly, the solution to (2.13) and (2.14) via Fisher scoring may be expressed as the iterative solution to the system:

$$\begin{bmatrix} X'WX & X'WZD \\ Z'WX & I + Z'WZD \end{bmatrix} \begin{pmatrix} a \\ v \end{pmatrix} = \begin{bmatrix} X'WY \\ Z'WY \end{bmatrix} \tag{2.16}$$



where  $b = Dv$  and  $v$  is a  $q$ -dimensional vector. Harville (1977) derived (2.16) for the best linear unbiased estimation (BLUE) of  $a$  and  $b$  assuming that the working dependent variable  $Y = Xa + Zb + \varepsilon$  is normally distributed with  $\varepsilon \sim N(0, W^{-1})$ ,  $b \sim N(0, D)$  and  $\text{cov}(b, \varepsilon) = 0$ . Therefore, solving system (2.16) is the same as applying iteratively weighted least squares on the concept of the linear model, using as dependent variable the transformed observations  $g(y_i) = Y_i$ . Equivalently, one may first solve for  $a$  in:

$$(X'V^{-1}X)a = X'V^{-1}Y \tag{2.17}$$

and then set:

$$\hat{b} = D\hat{v} = DZ'V^{-1}(Y - X\hat{a}) \tag{2.18}$$

where  $\hat{v} = Z'V^{-1}(Y - X\hat{a})$ . This suggests that approximately  $\text{cov}(\hat{a}) = (X'V^{-1}X)^{-1}$ . This is true for the normal theory linear model, under the assumption that  $\theta$  is fixed. Standard errors for  $\hat{b}$  may be calculated from (2.18) as  $\text{cov}(\hat{b}) = D\text{cov}(Z'V^{-1}(Y - X\hat{a}))D'$ . However, as we previously mentioned, by computing standard errors for the estimated mean parameters  $\hat{a}$  and  $\hat{b}$  this way, we do not take into account the additional variability introduced from the need to estimate the unknown variance components in  $\theta$ . Kackar and Harville (1984) have proposed normal theory approximations that account for this extra variability. Moreover, it would be of interest to explore further the suitability of the above procedure for the more general models, that do not restrict themselves on the special case of normally distributed random effects and working variable  $Y$ , (see e.g. Schall, 1991). Breslow and Clayton (1993) give further details about the PQL analysis.

### 2.3.2 Variance component estimation

Substitution of the maximized value of (2.12) into (2.11) and evaluation of  $W$  at  $(\hat{a}(\theta), \hat{b}(\theta))$  generates an approximate profile quasi-likelihood function for inference on the vector  $\theta$  of the unknown variance components. Some further approximations are made in order to motivate



standard estimating equations in terms of the working vector  $Y$ , the iterated weights  $W$  and the design matrices  $X$  and  $Z$ . Ignoring the dependence of  $W$  on  $\theta$  and replacing the deviance  $\sum_{i=1}^n d_i(y_i, \mu_i)$  by the Pearson chi-squared statistic  $\sum_{i=1}^n \frac{(y_i - \mu_i)^2}{m_i \nu(\mu_i)}$ , we have:

$$q(\hat{a}(\theta), \theta) \approx -\frac{1}{2} \log|V| - \frac{1}{2} (Y - X\hat{a})' V^{-1} (Y - X\hat{a}) \quad (2.19)$$

This quantity may be recognized as the profile likelihood based on the associated normal theory model for  $Y$ . To account for the loss of degrees-of-freedom in estimating the fixed-effects vector  $a$  by  $\hat{a}$ , and the fact that  $\hat{a}$  rather than  $a$  appears in (2.19), we use in practice the Restricted/residual Maximum Likelihood (REML) version (Patterson and Thomson, 1971):

$$q_1(\hat{a}(\theta), \theta) \approx -\frac{1}{2} \log|V| - \frac{1}{2} \log|X'V^{-1}X| - \frac{1}{2} (Y - X\hat{a})' V^{-1} (Y - X\hat{a}) \quad (2.20)$$

Following Harville (1977), by defining  $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$  and differentiating (2.20) with respect to the components of  $\theta$ , we obtain estimating equations for the variance parameters. Thus,  $\hat{\theta}$  satisfies:

$$-\frac{1}{2} \left[ (Y - X\hat{a})' V^{-1} \frac{\partial V}{\partial \theta_j} V^{-1} (Y - X\hat{a}) - \text{tr} \left( P \frac{\partial V}{\partial \theta_j} \right) \right] = 0 \quad (2.21)$$

The corresponding Fisher information matrix  $I$  has components:

$$I_{jk} = -\frac{1}{2} \text{tr} \left( P \frac{\partial V}{\partial \theta_j} P \frac{\partial V}{\partial \theta_k} \right) \quad (2.22)$$

### 2.3.3 Iterative procedure for estimating the mean and variance parameters

As we have already discussed, we derived the penalized quasi-likelihood (2.12) in order to estimate the mean parameters  $a$  and  $b$  and the profile quasi-likelihood (2.20) for estimating vector  $\theta$  of the unknown



variance components. We give a brief description of the iterative algorithm, as follows:

1. Given the current estimates  $\hat{\theta}^{(t)}, \hat{a}^{(t)}, \hat{b}^{(t)}$ , compute  $\hat{\theta}^{(t+1)}$  from (2.21).
2. Obtain  $\hat{a}^{(t+1)}$  from (2.17) and  $\hat{b}^{(t+1)}$  from (2.18).
3. Calculate  $\hat{\mu}_i^{(t+1)} = g^{-1}(x_i^t \hat{a}^{(t+1)} + z_i^t \hat{b}^{(t+1)})$ ,  $w_i^{(t+1)} = \{\phi m_i \nu(\hat{\mu}_i^{(t+1)}) [g'(\hat{\mu}_i^{(t+1)})]^2\}^{-1}$ ,  $Y_i^{(t+1)} = \hat{\eta}_i^{(t+1)} + g'(\hat{\mu}_i^{(t+1)})(y_i - \hat{\mu}_i^{(t+1)})$  and  $V_{(t+1)} = W_{(t+1)}^{-1} + ZD(\hat{\theta}^{(t+1)})Z'$ .

Return to step 1 and iterate between 1-3 steps until convergence.

Thus, since the Equations (2.17) and (2.18) for the mean parameters have been solved, the variance scores given from (2.21) are used to estimate a Newton step towards the unknown variance components in  $\theta$ . One then returns to Equations (2.17) and (2.18) to reestimate the mean parameters  $a$  and  $b$  respectively. Finally, zero or small positive values can be considered as initial estimates for the components of  $\theta = (\sigma_1^2, \sigma_2^2, \dots, \sigma_q^2)$ . However, initial estimates can be also obtained from other convenient methods.

## 2.4 Marginal Quasi-Likelihood

### 2.4.1 Estimation of fixed and random effects

As already mentioned in Section 2.2, in the GLMM observations are assumed to be conditionally independent given an unobserved vector of random effects  $b$ , with means that depend on the linear predictor through a specified link function  $g$ . Because sometimes the interest is focused more on the marginal relationship between covariables and outcome, it is often more appropriate (Heagerty and Zeger, 2000; see also the following example that refers to Crowder's data) to specify the GLM in terms of the marginal mean as:

$$E(y_i) = \mu_i = h(x_i^t a) \tag{2.23}$$

The marginal mean as defined in Equation (2.23), does not generally coincide with the marginal mean calculated from Equation (2.2), unless the link function is the identity.

In order to understand the difference between the marginal and the conditional model given from Equations (2.23) and (2.2) respectively,



consider Crowder’s study, which has been discussed in Section 2.2. For the agriculturalist interested in the fixed effects associated with seed variety and root extract treatment on germination rates, it is more appropriate to model the marginal probabilities of germination (averaged over plates). In contrast, the hierarchical (conditional) model is of interest in selecting plates containing subgroups of seeds that may have particularly high germination rates; therefore estimation of random effects is of interest too.

One may think of (2.23) as derived from a rather crude first-order approximation to the hierarchical model that is valid as the components of dispersion approach 0. Writing the model in the form  $y_i = \mu_i + \varepsilon_i$  with  $\text{var}(\varepsilon_i) = \phi m_i \nu(\mu_i)$  and  $b \sim N(0, D)$ , one has  $y_i \approx h(x_i^t a) + h'(x_i^t a) z_i^t b + \varepsilon_i$  (Goldstein 1991). Defining  $V_0$  and  $\Delta$  to be the  $n \times n$  diagonal matrices with diagonal elements  $\phi m_i \nu(\mu_i)$  and  $g'(\mu_i)$  respectively, the corresponding first-order variance approximation is:

$$\text{var}(y) = V_0 + \Delta^{-1} Z D Z' \Delta^{-1} \tag{2.24}$$

It has been shown that the true marginal mean for the hierarchical model with normally distributed random effects often could be approximately expressed in the form of (2.23), but with different values for the regression coefficients. With the *logit* link, for example, one finds:

$$E(y_i) \approx \exp(c_i x_i^t a) / (1 + \exp(c_i x_i^t a))$$

where  $c_i = (1 + c^2 z_i^t D z_i)^{-1/2}$ ,  $c = 16\sqrt{3} / (15\pi)$ , and thus we obtain altered (reduced) values for the regression coefficients  $a$ . However, these values are different when assuming non-normal random effects.

In the marginal model we estimate the fixed effects  $a$ , using the quasi-likelihood equations appropriate for dependent outcomes (McCullagh and Nelder 1989, sec. 9.3). Denoting the marginal mean vector by  $\mu = (\mu_1, \dots, \mu_n)'$  and considering the vector of the unknown variance parameters  $\theta$  as fixed, the score equation for the mean parameter  $a$ :

$$U(a, \theta) = \frac{\partial \mu}{\partial a'} \text{var}^{-1}(y)(y - \mu) = 0$$

takes the form



$$X'(\Delta V_0 \Delta + ZDZ')^{-1} \Delta(y - \mu) = 0 \tag{2.25}$$

Fisher scoring leads to IWLS regression of the working vector  $Y = \eta + \Delta(y - \mu)$  on  $X$  with weight matrix:

$$V^{-1} = (\Delta V_0 \Delta + ZDZ')^{-1} = (W^{-1} + ZDZ')^{-1} \tag{2.26}$$

where  $\eta = (\eta_1, \dots, \eta_n)'$  denotes the vector of linear predictors  $\eta_i = x_i' a$  and  $W = (\Delta V_0 \Delta)^{-1}$  as before the  $n \times n$  diagonal matrix with the GLM iterated weights  $w_i = \{\phi m_i v(\mu_i) [g'(\mu_i)]^2\}^{-1}$  as diagonal elements. At each step in the iteration, the problem is equivalent to that of estimating  $a$  in the associated normal theory model  $Y = Xa + Zb + \varepsilon$ , with  $\varepsilon \sim N(0, W^{-1})$  and  $b \sim N(0, D)$ . Shrinkage estimates for the random effects are again obtained from (2.18).

### 2.4.2 Variance component estimation

Variance parameters may be estimated using Carroll and Ruppert's (1982) method of pseudolikelihood. Assuming that  $E(y)$  is known, we consider the normal theory likelihood based on the variance approximation (2.24) and take logarithmic derivatives with respect to  $\theta$ . This leads to the same REML Equation (2.20) for the variance parameters as was derived previously for PQL. Thus, differentiating once again Equation (2.20) with respect to the components of  $\theta$ , we obtain  $\hat{\theta}$ , which satisfies Equation (2.21). Instead of iterating back and forth between (2.13)-(2.14) or (2.17)-(2.18), and (2.21), to obtain new values of  $b$  at each iteration for use in  $Y$  and  $V$ , ( $Y$  and  $V$  depend on  $b$  through  $\mu_i = g^{-1}(x_i' a + z_i' b)$ ), we iterate between (2.25) for  $a$  and (2.21) for  $\theta$ , delaying the estimation of the random effects  $b$  from (2.18) until convergence.



### 2.4.3 Iterative procedure for estimating the mean and variance parameters

The iterative algorithm for estimating the fixed and random effects, as well as the variance components, proceeds as follows:

1. Given the current estimates  $\hat{a}^{(t)}, \hat{\theta}^{(t)}, \hat{b}^{(t)}$ , compute  $\hat{a}^{(t+1)}$  from (2.25).
2. Calculate  $\hat{\mu}_i^{(t+1)} = g^{-1}(\hat{\eta}_i^{(t+1)}) = g^{-1}(x_i' \hat{a}^{(t+1)})$ ,  
 $w_i^{(t+1)}$  which depends on  $\hat{\mu}_i^{(t+1)}$ ,  
 $\Delta^{(t+1)}$  with diagonal elements  $g'(\hat{\mu}_i^{(t+1)})$ ,  
 $Y_i^{(t+1)} = \hat{\eta}_i^{(t+1)} + g'(\hat{\mu}_i^{(t+1)})(y_i - \hat{\mu}_i^{(t+1)})$  and  
 $V_{(t+1)}^{-1} = (W_{(t+1)}^{-1} + ZD(\hat{\theta}^{(t)})Z')^{-1}$  from (2.26).
3. Obtain  $\hat{\theta}^{(t+1)}$  from (2.21) and  $\hat{b}^{(t+1)}$  from (2.18) respectively.  
 Return then to step 1, and iterate between 1-3 steps until convergence.

### 2.5 PQL compared to MQL and some further remarks

The essential difference between the MQL estimating equations for the marginal model and the PQL equations for the hierarchical model is that the latter incorporate the random effect terms  $z_i' b$  in the linear predictor  $\eta_i = x_i' a + z_i' b$ . As we have previously stated, PQL is the method of choice for estimating parameters in the hierarchical model, especially when attention is focused on shrinkage estimation of the random effects  $b$  (Robinson 1991). In contrast, MQL is the method of choice when interest is focused on the marginal relationship between covariables and response, while the random effects are mainly useful for calculating the covariance matrix  $V^{-1}$  as expressed in Equation (2.26). That enables one to get efficient estimated equations for the mean value parameters.

Another important distinction between PQL and MQL is the fact that the regression coefficients of the former, but not of the latter, depend strongly on the estimated variance components when the link function is not the



identity, even in large samples. Furthermore, through a series of worked examples, it appears that PQL is of practical value for approximate inference on parameters and realizations of random effects in the hierarchical model. However, the success of the approximation depends on the closeness to normality of the observations and might fail badly, e.g., for binary response data (Rodriguez and Goldman, 1995). Thus, the approximation is likely to improve as observations become more normally distributed, for example as the denominators of binomial observations or the means of Poisson observations increase.

Moreover, PQL tends to underestimate somewhat the variance components  $\theta$  and (in absolute value) fixed effects  $a$  when applied to clustered binary data, but the situation improves as the denominators increase. Particularly, very often inference on variance parameters under PQL is not satisfactory, due to the tendency of the procedure to converge to a non-positive definite variance matrix  $D$ . This might happen when the response probabilities are small and only limited information is present for estimating random effects and their associated variances and covariances. Following that, the procedure may underestimate the covariances of the random effects.

## 2.6 The EM algorithm for estimating the finite mixture model

As already stated, one natural way of representing the ‘extra variation’ which may be observed on fitting the standard exponential family GLM, is by adding an unobserved random effect to the linear predictor. By this way, the extra variation is modeled on the same scale as the linear predictor. This choice seems to be quite natural, considering that overdispersion very often arises due to the omission of important explanatory variables. An obvious choice and certainly traditional, is to assume that this random effect is normally distributed. This is particularly reasonable for link functions giving an unbounded parameter space for the linear predictor. For example, for Binomial data we use the *logit* link that maps the range  $[0,1]$  of the response probabilities onto the real line  $(-\infty, +\infty)$ . Therefore, if the random effect can



take any real value as being normally distributed, the parameter space for the linear predictor is unbounded, lying into the interval  $(-\infty, +\infty)$ .

Before giving further details on estimating the parameters of interest, we give a brief statement of the overdispersed GLM. Let us suppose that we have a random sample  $y = (y_1, \dots, y_n)'$  from an exponential family distribution. The vector of the canonical parameters is given by  $\theta = (\theta_1, \dots, \theta_n)'$  and the mean vector by  $\mu = (\mu_1, \dots, \mu_n)'$ . Once again, we consider the scale parameter to be fixed at unity. In addition, we have a vector  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$  of  $p$  explanatory variables for the fixed effects and an unobserved random effect  $b_i$  for the  $i$ th observation,  $i = 1, \dots, n$ . The random effects  $b_i$  are assumed independently normally distributed with common variance  $\sigma^2$ ,  $b_i \sim N(0, \sigma^2)$ . Conditional on  $b_i$ ,  $y_i$  has the GLM with linear predictor  $\eta_i$ , such that:

$$g(E(y_i | b_i)) = g(\mu_i) = \eta_i = x_i' a + b_i \tag{2.27}$$

The latter equation can be also expressed as:

$$g(E(y_i | b_i)) = g(\mu_i) = \eta_i = x_i' a + \sigma b_i \tag{2.28}$$

where now  $b_i$  is the standardized value of  $b_i$ , that is  $b_i \sim N(0, 1)$ , and the unknown standard deviation  $\sigma$  enters the linear predictor as an additional regression parameter. We also note that in Equation (2.28) explanatory variables and random effects are both related to mean  $\mu$ , through the link function  $g$ .

The observed marginal likelihood is given by:

$$L(\theta_1) = L(a, \sigma) = f(y; \theta_1) = \int f(y, b; \theta_1) db = \int f(y | a, b) \phi(b | \sigma) db \tag{2.29}$$

where  $\theta_1 = (a, \sigma)'$  is our parameter of real interest,  $f$  is the density function from the exponential family distribution,  $f(y, b; \theta_1)$  is the joint density of  $y$  and  $b$ ,  $\phi(b | \sigma)$  is the prior distribution of the unobserved random effects  $b = (b_1, \dots, b_n)'$  and  $f(y | \alpha, b)$  is the conditional distribution of observed  $y$  given the unobserved random effects  $b$  and the mean parameter  $\alpha$ . Since observations  $y_i$  are assumed to be independent conditional on  $b_i$ , we have:



$$f(y | a, b) = \prod_{i=1}^n f(y_i | a, b_i) \tag{2.30}$$

Applying Equation (2.30) to (2.29) and considering also the independence assumption of the random effects  $b_i$ , Equation (2.29) gives the likelihood:

$$L(a, \sigma) = \int \prod_{i=1}^n f(y_i | a, b_i) \phi(b_i | \sigma) db_i$$

The log-likelihood is:

$$l(a, \sigma) = \log\{L(a, \sigma)\} = \sum_{i=1}^n \log \int f(y_i | a, b_i) \phi(b_i | \sigma) db_i \tag{2.31}$$

In general, the integral in (2.31) does not have a closed form and its analytic calculation is often not easily possible. An exception occurs for normal responses  $y$  and normal random effects  $b$ . To overcome the problem of integration, we approximate the integral by a quadrature formula. This means we substitute the integral by a finite weighted sum, where  $b$  is replaced by some mass points  $b_k$  with corresponding masses  $p_k = P(b_k)$  for  $k = 1, \dots, K$ . Since we assume a normal mixing distribution for the unobserved random effect, this allows us to apply a Gauss-Hermite (GH) quadrature (see Appendix C for further details), where the masses and mass points are known and available from the relevant tables (e.g. Abramowitz and Stegun, 1964). Using the approximation:

$$\int f(y_i | a, b_i) \phi(b_i | \sigma) db_i = \sum_{k=1}^K p_k f'(y_i | b_k; a, \sigma),$$

Equation (2.31) gives the log-likelihood:

$$l(a, \sigma) = \sum_{i=1}^n \log \sum_{k=1}^K p_k f'(y_i | b_k; a, \sigma) \tag{2.32}$$

The log-likelihood is thus approximately the log-likelihood of a finite mixture of exponential family densities,  $f'$ , with known mixture proportions  $p_k$  at known mass-points  $b_k$ , with the linear predictor for the  $i$ th observation in the  $k$ th mixture component being  $\eta_{ik} = x_i' a + \sigma b_k$ . For compactness we write:

$$f'_{ik} = f'(y_i | b_k; a, \sigma) = \exp\{\theta_{ik} y_i - b(\theta_{ik}) + c(y_i)\}$$

where  $E(y_i | b_k; a, \sigma) = \mu_{ik} = b'(\theta_{ik})$  and  $\eta_{ik} = g(\mu_{ik}) = x_i' a + \sigma b_k$  denote the  $i$ th conditional mean and linear predictor respectively, given  $b_k$ ,  $V_{ik} = b''(\theta_{ik})$ .



the variance function and  $g'_{ik} = g'(\mu_{ik})$  is the first derivative of the link function  $g$  with respect to  $\mu_{ik}$ . The log-likelihood is then,

$$l(a, \sigma) = \sum_{i=1}^n \log \sum_{k=1}^K p_k f'(y_i | b_k; a, \sigma) = \sum_{i=1}^n \left( \log \sum_{k=1}^K p_k f'_{ik} \right) \Rightarrow$$

$$\frac{\partial l(a, \sigma)}{\partial a} = \sum_{i=1}^n \frac{\sum_{k=1}^K p_k \frac{\partial f'_{ik}}{\partial a}}{\sum_{k=1}^K p_k f'_{ik}} = \sum_{i=1}^n \frac{\sum_{k=1}^K p_k f'_{ik} \frac{\partial \log f'_{ik}}{\partial a}}{\sum_{k=1}^K p_k f'_{ik}} = \sum_{i=1}^n \sum_{k=1}^K w_{ik} s_{ik}(a)$$

and equating this to 0 we obtain

$$\frac{\partial l(a, \sigma)}{\partial a} = 0 \Leftrightarrow \sum_{i=1}^n \sum_{k=1}^K w_{ik} s_{ik}(a) = 0 \tag{2.33}$$

with weights

$$w_{ik} = f(b_k | y_i; a, \sigma) = \frac{p_k f'_{ik}}{\sum_{l=1}^K p_l f'_{il}} \tag{2.34}$$

which are the posterior masses for  $b_k$  given the data, or allocation of individual  $y_i$  into component  $k$  and

$$s_{ik}(a) = \frac{\partial \log f'_{ik}}{\partial a} = (y_i - \mu_{ik}) x_i / V_{ik} g'_{ik} \tag{2.35}$$

the  $a$ -component of the score (the log-likelihood derivative with respect to  $a$ ) for observation  $i$  in component  $k$ .

Similarly,

$$\frac{\partial l(a, \sigma)}{\partial \sigma} = \sum_{i=1}^n \sum_{k=1}^K w_{ik} s_{ik}(\sigma)$$

and equating this to zero we obtain

$$\frac{\partial l(a, \sigma)}{\partial \sigma} = 0 \Leftrightarrow \sum_{i=1}^n \sum_{k=1}^K w_{ik} s_{ik}(\sigma) = 0 \tag{2.36}$$

where now

$$s_{ik}(\sigma) = \frac{\partial \log f'_{ik}}{\partial \sigma} = (y_i - \mu_{ik}) b_k / V_{ik} g'_{ik} \tag{2.37}$$

is the  $\sigma$ -component of the score and thus  $b_k$  becomes another observable variable in the regression, with regression coefficient  $\sigma$ . Solving Equations



(2.33) and (2.36) for given weights  $w_{ik}$  and updating these weights from the current parameter estimates for  $a$  and  $\sigma$ , is an EM algorithm (Dempster *et al.*, 1977; McLachlan and Peel, 2000; Karlis, 2003). The EM algorithm is used as a standard framework to deal with missing data. Regarding the random effects as missing variables, the algorithm can also be used for estimating the GLMM. Initial estimates for the first E-step for  $a$  are conveniently obtained from the ordinary independence GLM fit ( $\sigma = 0$ ) and for  $\sigma$  by arbitrary specification other than zero, e.g.  $\sigma = 1$ .

Next, we describe the EM algorithm, which consists of an expectation and a maximization step, (E- and M-step respectively).

1. E-step. Given the current estimates  $a^{(t)}, \sigma^{(t)}$  compute

$$w_{ik}^{(t+1)} = \frac{p_k f_{ik}^{(t)}}{\sum_{l=1}^K p_l f_{il}^{(t)}}$$

with  $f_{ik}^{(t)} = f'(y_i | b_k; a^{(t)} \sigma^{(t)})$ .

2. M-step. Solving Equations (2.33) and (2.36) for given weights  $w_{ik}^{(t+1)}$ , the M-step determines  $a^{(t+1)}, \sigma^{(t+1)}$  respectively.

One then iterates between the E and M steps until convergence.

The main drawbacks of the EM algorithm are its typically slow rate of convergence. The double iterative structure of many implementations adds to the problem. Further, the algorithm does not automatically provide precision estimators. To obtain correct standard errors for the parameter estimates in the final model, we use the property (Dietz and Böhning, 1995) that in large samples from regular models for which the log-likelihood is quadratic in the parameters, the likelihood ratio and Wald tests for the significance of an individual parameter are equivalent, so that the deviance change (d.c.) on omitting the variable is equal to the square of the t-statistic. That is, if  $\theta$  is the parameter of interest, then:

$$d.c. = \left( \frac{\hat{\theta}}{se(\hat{\theta})} \right)^2 \Rightarrow se(\hat{\theta}) = \frac{|\hat{\theta}|}{\sqrt{d.c.}}$$



This requires the fitting of a set of reduced models in which each variable is left out at a time from the final model. By this way, we can also assess the significance of each variable by its deviance change. However, the above property of regular models may not hold in small samples with skewed log-likelihoods. In the latter case, the proposed standard error estimate is a more appropriate reflection of the significance of the variable than that based on the inverse information matrix, since it gives the correct likelihood ratio test statistic value for each variable rather than the misleading Wald test value.

## **2.7 Other approaches for the GLM with normal random effects**

So far, our main concern was the special case of the normal mixing distribution. Exponential family models other than the normal with a normal random effect have been difficult and slow to fit by ML, because the resulting likelihood does not have a closed form. Particularly, we showed that the likelihood can be integrated numerically using some form of Gaussian quadrature to give full ML estimation. Current quadrature methods use the EM algorithm for estimating the finite mixture model (Aitkin, 1996 and 1999). Moreover, penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) can be used for estimating GLM with normal random effects (Breslow and Clayton, 1993).

However, there are different approaches that can be followed in order to deal with the problem of the no closed form of the likelihood. First, the integrals required in the E-step of the EM algorithm can be avoided by Laplace approximations (Steele, 1996) or by Monte Carlo integration (Walker, 1996; McCulloch, 1997).

Second, the problem can be handled by the generalized estimating equation approach (GEE) (Liang and Zeger, 1986; Diggle, Liang and Zeger, 1994). Here the marginal distribution of the response is assumed to be exponential family and the repeated-measures structure is represented by a covariance matrix model whose parameters are estimated by a form of quasi-likelihood (MQL), which does not require a full parametric specification for the random effect distribution. Because the estimating equations for MQL and GEE are identical as regards the mean parameters, one might anticipate



similar results for MQL. Once again, GEE is the method of choice when attention is focused on the marginal relationship between covariables and response.

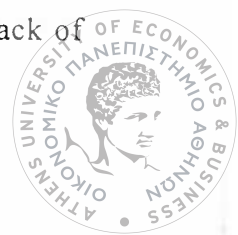
Third, recent Bayesian procedures avoid the need for numerical integration by taking repeated samples from the marginal posterior distributions of the parameters, using Markov chain Monte Carlo methods (Walker, 1996; McCulloch, 1997; Gelman *et.al*, 1995). An attractive feature of the Bayesian approach is its flexibility for full assessment of the uncertainty in the estimated random effects, through the additional structure of a prior distribution on all the model parameters. As already mentioned, sometimes there is not enough information to estimate random effects and their associated covariances. The Bayesian formulation enjoys an advantage here because of the information on variance components contributed by the prior distribution.

Finally, Booth and Hobert (1999) propose two new implementations of the EM algorithm for maximum likelihood estimating of the GLMM. They propose two different implementations of the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) in which simulation methods are used to evaluate the integral at the E-step. Booth and Hobert (1999) use three different data sets to illustrate their techniques. The results show that their methods can be considerably more efficient than those based on Markov chain Monte Carlo algorithms. However, they recommend using Markov chain Monte Carlo algorithms in problems involving very high dimensional integrals.

## **2.8 NPML estimate of the mixing distribution**

### **2.8.1 NPML estimation of the masses and mass-points**

Although it is quite convenient and certainly traditional to assume a normal mixing distribution, it is very often not appropriate to use such a specified parametric form for the distribution of the unobserved random effects, especially for a completely unknown mixing distribution. A particular disadvantage of the modeling approach described in Section 2.6 is the lack of



information in the data about the mixing distribution, since this can come only from the marginal distribution of the data. Generally, a disadvantage of any approach considering the normal mixing distribution as described in Sections 2.2-2.7, is the possible sensitivity of conclusions to the choice of such an assumption for the distribution of random effects. Another disadvantage is the need to expand the data vector to length  $Kn$ . Particularly, the double summation over  $i$  and  $k$  in Equations (2.33) and (2.36), can be handled by replicating  $y$  and  $x$   $K$  times and the Gaussian quadrature variable  $b$   $n$  times. Model fitting is then identical to that of a single sample of size  $Kn$ . If however  $K$  is large, the required time for model fitting increases substantially. A third disadvantage is the possible inaccuracy of Gaussian quadrature, where even the number  $K$  of mass points is large.

In Section 2.6 we assumed that the mixing distribution is normal and we invoked the EM algorithm in order to estimate the parameters of the model. However, there is often little information about the random effects and therefore we do not want to assume a specific parametric form for the prior distribution function,  $H$ , of the random effects. In such cases we consider as a preferable modeling strategy the non-parametric maximum likelihood (NPML) estimation of the mixing distribution  $H$  together with the GLM parameters. We note that the NPML estimate of the mixing distribution is well known (e.g. Laird, 1978; Lindsay, 1983; Böhning, 1999, sec. 2.5) to be a discrete distribution on a finite number of mass-points. At this point we should emphasize that our aim is not to estimate the mixing distribution, which is regarded as a nuisance parameter, but to avoid possibly misleading inferences from an inappropriate model assumption.

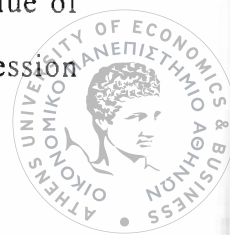
More specifically, we seek an algorithm for maximizing the log-likelihood  $l(H)$  over the family of all distribution functions of the random effects, which yields the NPML estimate of  $H$ . Laird (1978), Lindsay (1983 and 1995) and Böhning (1999) guarantee the discreteness of the NPML estimate of the mixing distribution function  $H$  in full generality. Discreteness is not a serious restriction, if  $H$  is estimated by maximum likelihood, since no matter what the true  $H$  is, there is a maximum likelihood estimate  $\hat{H}$  with few steps. Follman and Lambert (1988) discuss how many steps  $H$  may have



without losing identifiability. According to Laird (1978), the NPML estimate of the mixing distribution  $H$  is self-consistent, a property that characterizes the NPML estimate of a distribution function in incomplete data problems. Under various conditions the estimate  $\hat{H}$  is a step function with a finite number of steps (mass-points), which depend in general on the sample size  $n$ , the conditional density  $f(y_i|a, b_i)$  of observed  $y_i$  given the unobserved random effects  $b_i$  and the mean parameter  $a$  and the spacing among  $y_1, \dots, y_n$ . In fact for any  $0 < k < \infty$ , where  $k$  denotes the number of mass-points, the maximum likelihood estimate of the discrete distribution,  $H_k$ , putting probability  $p_k$  at  $\beta_k$  ( $p_k = P(\beta_k)$ ), is self-consistent (see, e.g., Hasselblad, 1969 and Wolfe, 1970). Unfortunately, it is in general not possible to count the number of mass-points. However, this is not a practical problem, since it is a by-product of the computation. Furthermore, conditions can be established for a finite number of mass-points (Laird, 1978). Follman and Lambert (1989) and Lesperance and Kalbfleisch (1992) also discuss about the NPML estimate of the mixing distribution.

In this more general case, we treat the masses  $p_k$  and mass-points  $b_k$  ( $k = 1, \dots, K$ ) as unknown parameters that both have to be estimated from the data. The number  $K$  of mass-points is also unknown and it is sequentially increased until the likelihood is maximized. Since the variance of the mixing distribution is a function of the unknown parameters, we drop the scale parameter  $\sigma$  and define the mass-point parameters as  $\beta_k$  with unknown proportions  $p_k = P(\beta_k)$  and the linear predictor for the  $i$ th observation in the  $k$ th component being  $\eta_{ik} = x_i' a + \beta_k$  with  $i = 1, \dots, n$  and  $k = 1, \dots, K$ . Thus,  $\beta_k$  is like an intercept parameter for the  $k$ th component and it can be simply estimated by including a 'component factor' in the model with  $K$  levels.

Let us suppose that  $\beta = (\beta_1, \dots, \beta_K)'$  is the vector of the unknown mass-point parameters and  $p = (p_1, \dots, p_K)'$  is the vector of the unknown masses. Given  $\beta_k$ , we can write the (conditional) model as  $\mu_{ik} = g^{-1}(x_i' a + e_k' \beta)$ , with  $e_k$  a  $K$ -dimensional vector of zeros, except the  $k$ th position with a value of one. Hence, the unknown mass-points enter the model as additional regression



coefficients. We now denote with  $\theta_2 = (a, \beta)'$  and  $p$  the unknown parameters. Differentiating again the log-likelihood with respect to  $a$  and  $\beta$ , we obtain estimating equations similar to (2.33) and (2.36), which derived when estimating the finite mixture model in Section 2.6. These equations are once again used in each M-step, in order to compute one step towards the values of  $a$  and  $\beta$ . Moreover, if we differentiate the log-likelihood with respect to  $p_k$

and use the restrictions,  $0 < p_k < 1$  and  $\sum_{k=1}^K p_k = 1$ , we have directly:

$$\frac{\partial l}{\partial p_k} = \sum_{i=1}^n \frac{f'_{ik} - f'_{iK}}{\sum_{l=1}^K p_l f'_{il}} = \sum_{i=1}^n \left\{ \frac{w_{ik}}{p_k} - \frac{w_{iK}}{p_K} \right\} \quad (2.38)$$

Equating this to zero gives simply:

$$\hat{p}_k = \sum_{i=1}^n w_{ik} / n \quad (2.39)$$

Thus, the same EM algorithm, as described in Section 2.6, applies with the additional estimate of  $p_k$  in each M-step from the current weights. Initial estimates of the  $\beta_k$  can be taken as the standard normal values  $b_k$ .

### 2.8.2 Further remarks on the NPML estimate

Hypothesis testing or model comparisons can be carried out through the likelihood ratio test in the usual way. Particularly, the log likelihood ratio statistic for comparing two nested models is the difference of their deviances and the asymptotic distribution of the statistic under the null hypothesis is a chi-squared distribution with known degrees of freedom given by the difference in the degrees of freedom of the two models. However, theoretical justifications of this approach seem to be lacking and it would be better not to be used. Actually, the main reason for the appropriateness of the log likelihood ratio statistic, is that the residual deviance from the overdispersed model relative to a saturated model has however no interpretation as a goodness-of-fit test for this model. This is true, since the concept of the saturated model does not apply to a two-parameter model, for which the variation parameter would be unidentifiable. Therefore, when we have to



confront the overdispersed model, we should not use the residual deviance as a measure of goodness-of-fit, or to compare models using the log likelihood ratio statistic. AIC and other penalized likelihood ratio criteria could also be used for model comparisons.

A remarkable consequence of the NPML estimation is that it sometimes allows for, and corrects, possible associations between the random effects and the explanatory variables, even though these are not allowed for in the model specification (Aitkin, 1996). For example, one may observe that large positive residuals tend to occur for large values of the explanatory variable, meaning that the random effect estimated by NPML is positively correlated with the explanatory variable. Thus, if we include the random effect to the model, the estimated slope of this variable will be reduced due to its positive correlation with the random effect. The normal distribution for the random effects however forces the random effects to be symmetric about the regression.

Finally, the discrete nature of the NPML estimate might be found unattractive if one believes *a priori* in the existence of a continuous mixing distribution. Davidian and<sup>1</sup>Gallant (1993) described an alternative approach that assumed a smooth mixing density. In addition, as we previously emphasized, the mixing distribution that has been estimated non-parametrically, is treated as a nuisance function. Thus, the distribution of the change in deviance on fitting the overdispersed model is not of direct interest: the regression model coefficients are the parameters of interest instead.

### 2.8.3 Computational issues

The EM algorithm for the overdispersed models is very stable and converges rapidly in many cases, though for the normal mass-point models the deviances and parameter estimates sometimes fluctuate considerably for different numbers  $K$  of mass-points (Lesaffre and Spiessens, 2001). Through a series of examples, the NPML estimate has been impressively stable too. Unfortunately, the procedure may result in a local rather than a global maximum of the likelihood. It is therefore recommended that the EM algorithm is run several times with different starting values, hoping that if



they all give the same solution, then it is likely to be the global maximum. Experience suggests that local maxima could lead us to unreliable results. In many examples, a considerable difference in local maxima sometimes occurs, depending on whether the number  $K$  of mass-points is odd or even. In such cases, we can find a reliable estimate of the true maximum by overfitting the number of mass-points and identifying the location of the reduced number of points actually required for the NPML estimate. Moreover, even though past experience with mixture modeling for overdispersion may have left the impression that the problem is computationally quite intensive, with nowadays speeds on personal computers this no longer seems such a serious issue. Therefore, the general EM algorithm for full NPML estimation in overdispersed exponential family models has become a powerful modeling tool.



2



## CHAPTER 3

### Mixed Binomial Distributions in the Presence of Extra-Binomial Variation

#### 3.1 Introduction

In the previous Chapter, we mentioned certain methods in order to account for the extra-binomial variation, which may be observed on fitting a model. A major assumption underlying the use of logistic regression analysis for binomial data is that the variance of the error distribution is completely determined by the mean. In practice this assumption often fails. The first indication that something is wrong is that the deviance measure of goodness-of-fit for ‘full’ models exceeds its degree of freedom and the data are then said to have extra-binomial variation. A number of different methods have been developed to deal with extra-binomial variation. For example, the basic idea as described in Chapter 2 was to model the extra variation on the same scale as the linear predictor, or in other words, to add unobserved random effects to the linear predictor and then use different approximations in order to evaluate the integrations that arise numerically.

In this Chapter, another approach is proposed for overcoming the problem of overdispersion. In particular, binomial mixture distributions can be used with the special case, and the best known, of the Beta-Binomial distribution. In Sections 3.2.1 and 3.2.2 we present the Beta-Binomial (BB) model, how to estimate its parameters, under what circumstances it seems more appropriate than the standard binomial model and generally why it is recommended whenever extra-binomial variation is observed. Closing, Section 3.2.3 suggests further generalizations of the Binomial distribution and Section 3.3 presents some methods to test the goodness of fit of the Binomial distribution.



### 3.2 Binomial mixture distributions for adjusting for Overdispersion

#### 3.2.1 The specific case of the Beta-Binomial distribution

In many laboratory experiments the response of interest is binary, with values of 0 or 1 representing failure and success respectively, and the outcome for each experimental ‘group’ or ‘litter’ can be expressed as a proportion or several proportions. Consider an example of a toxicology study in which rabbits are in litters of various sizes and the response measured is dichotomous; the rabbits being either alive or dead at the end of the study. Moreover, the outcomes can be expressed as the proportions of the survived rabbits from each experimental group. For a particular litter of size  $n$ , let  $R$  denote the number of rabbits in that litter, which are alive at the end of the experiment. Often the Bernoulli response for the individual rabbit in a litter will be not observable, so that the statistician may only have the litter totals  $R$ . It is often convenient to treat the number of successes  $R$  in a litter, and consequently the proportions  $R/n$ , as binomial distributed random variables. In fact, the binomial may not be a good fit, possibly because the observed variance of  $R$  exceeds the nominal variance predicted under the binomial model. In addition, in such studies the responses of the individual rabbits within each litter may be correlated, and hence the outcome  $R/n$  for a given litter of size  $n$  may not follow a binomial distribution. Analysis of such studies can be based on two-parameter generalizations of the binomial model, which allow for the presence of dependent responses within litters, (e.g. Altham, 1978). Special attention is given to the expansion of the Binomial model to a Beta-Binomial (BB) model, which is one of the most popular models to account for overdispersion.

In many applications it is natural to assume that the binary response probabilities have a common value within an experimental group, and that pairs of binary observations within a group have a common correlation. As we have already stated in Chapter 1, this common correlation is measured by the well-known intra-cluster correlation coefficient. When the data structure in the population is hierarchical, different methods are proposed in order to analyze these more complex data. The general concept however, is that



individuals interact with the contexts to which they belong. Considering the previous example, where rabbits are nested within litters, individual rabbits are influenced by the litter to which they belong and the properties of those litters are in turn influenced by the individual rabbits who make up that litter. With no doubt, we expect a positive correlation between rabbits from the same litter, since rabbits within a litter share the same environmental conditions or may be fed in the same way.

Motivated by the above discussion, we conclude that by assuming independence among individuals from the same experimental group through the standard logistic linear model, we do not take into account the presence of extra-binomial variation. Using the Binomial model for data in which there is inter-cluster variation may lead to seriously misleading conclusions. If however we still use the Binomial model for overdispersed Binomial data, underestimated standard errors can be obtained - the confidence intervals would be too narrow - and misleading conclusions may be drawn. In other words, an unacceptably high Type I error may occur, while the experimenter thinks it is equal to 0.05. In contrast, the BB model was developed to fit overdispersed Binomial data and is valid even if there is more than one source of variation in the data. There is a long history of research on the mixed Binomial models (see, for example, Brooks 2001) and specifically on the BB distribution. The basic theoretical properties of the distribution have been discussed by Skellam (1948), Ishii and Hayakawa (1960), Moran (1968), Johnson, Kotz and Kemp (1992), and Kleinman (1973). The BB model has been applied successfully in the study of chromosomes (Skellam 1948), in market research (Chatfield and Goodhardt 1970), in toxicology (Haseman and Kupper 1979), in disease incidence (Griffiths 1973), in teratology (Williams 1975; Paul 1982), in a dominant lethal study (Aeschbacher *et al.* 1978), in mutagenesis (Otake and Prentice 1984) and to study the effect of policy changes on the appropriateness of hospital stays (Gange *et al.* 1996). Generally, it has been shown that the potential applications of the BB model are numerous in a wide variety of fields.



### 3.2.2 Estimating the parameters of the Beta-Binomial model

We begin by defining the BB model as follows. Let us suppose that the units under study can be classified into  $k$  groups. Furthermore, let  $n_i$  denote the number of observations in group  $i$ , and let  $R_i$  denote the number of successes in that group. Assume also that conditional on  $p_i$ ,  $R_i | p_i \sim$  Binomial  $(n_i, p_i)$ . Associated with the  $i$ th response ( $1 \leq i \leq k$ ), which is a count of  $R_i$  successes and  $n_i - R_i$  failures, is a vector  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^t$  of  $p$  explanatory variables. We denote the  $k \times p$  design matrix with rows  $x_i^t$  by  $X$  and the  $p$ -dimensional vector of regression coefficients by  $\beta$ . To allow for extra-binomial variation, the parameters  $p_i \in (0,1)$  in the Binomial distribution are assumed to follow a Beta distribution with parameters  $a_i$  and  $b_i$ , meaning that  $p_i \sim \text{Beta}(a_i, b_i)$ ,  $a_i > 0$ ,  $b_i > 0$ , so that unconditionally the  $R_i$  have a Beta-Binomial distribution with density:

$$\Pr(R_i = r_i) = \binom{n_i}{r_i} \frac{\Gamma(a_i + r_i)\Gamma(b_i + n_i - r_i)\Gamma(a_i + b_i)}{\Gamma(a_i + b_i + n_i)\Gamma(a_i)\Gamma(b_i)} \tag{3.1}$$

where  $\Gamma(v) = (v - 1)!$  denotes a gamma function and  $r_i = 0, \dots, n_i$  is the number of successes out of  $n_i$  trials. The mean and variance of the Beta-distributed  $p_i$ , are given respectively from:

$$E(p_i) = \frac{a_i}{a_i + b_i} = \pi_i \tag{3.2}$$

$$\text{var}(p_i) = \frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)} = \rho_i \pi_i (1 - \pi_i) \quad (= \rho \pi_i (1 - \pi_i)) \tag{3.3}$$

with

$$\rho_i = \frac{1}{a_i + b_i + 1} \tag{3.4}$$

By analogy with ANOVA, an assumption is made concerning homogeneity of variance between data sets or groups. Through (3.4), we take  $\rho_i = \rho$  here for illustration. One could say that factor  $\rho$  in (3.3) is a scale



parameter which measures the variation of  $p_i$  and as usually is considered to be the same for all groups  $i$ . In addition, since  $a_i$  and  $b_i$  are positive for a Beta distribution,  $0 < \rho < 1$ . Actually,  $\rho$  plays the role of a positive correlation between the binary components of a proportion, or in other words, corresponds to the common positive correlation, which is observed between pairs of binary responses within a group. Considering then the constraint  $\rho_i = \rho$ , it is the same as assuming that the pairwise correlation among binary observations within a group is common for all groups. If however, there is a need to treat differently dependencies within and dependencies between groups, a more elaborate discussion is necessary.

Moreover, unconditionally, the mean and variance of the Beta-Binomial distributed  $R_i$ , are given as follows:

$$E(R_i) = n_i \pi_i \tag{3.5}$$

$$\text{var}(R_i) = v_i d_i^{-1} \tag{3.6}$$

where  $v_i = n_i \pi_i (1 - \pi_i)$  and  $d_i^{-1} = 1 + \rho(n_i - 1)$ . At this point we remind that the Binomial distribution has mean and variance  $n_i \pi_i$  and  $n_i \pi_i (1 - \pi_i)$  respectively, so the term  $1 + \rho(n_i - 1)$  in the variance of the BB distribution acts as a multiplier of the Binomial variance. The term is always greater than or equal to 1, and models the overdispersion due to the variation of  $p_i$ . The case of pure Binomial variation, in which parameters  $p_i$  are constant, corresponds to  $\rho = 0$ . The case of BB variation with maximum variance corresponds to  $\rho = 1$  in which the inflation factor  $1 + \rho(n_i - 1)$  increases up to a maximum of  $n_i$ .

Expressing the outcome as the proportion of the successes in group  $i$  by  $Y_i = R_i / n_i$ , we have:

$$E(Y_i) = \pi_i \tag{3.7}$$

$$\text{var}(Y_i) = \frac{\pi_i(1 - \pi_i)}{n_i} + \frac{n_i - 1}{n_i} \rho \pi_i (1 - \pi_i) = \frac{\pi_i(1 - \pi_i)}{n_i} (1 + \rho(n_i - 1)) \tag{3.8}$$

From Equation (3.8), we deduce that the variance of the Beta-Binomial distributed  $Y_i$  is once again greater than that under Binomial data and that



factor  $\rho$  is incorporated in the second term in order to account for overdispersion. Expressions (3.7) and (3.8) also imply that it would be convenient to reparameterize the Beta-Binomial in terms of  $\pi_i$  and  $\rho$ .

Suppose now that we do not specify that the  $p_i$  have a Beta or any other particular distribution, but we have, following (3.2) and (3.3), that:

$$E(p_i) = \pi_i \tag{3.9}$$

$$\text{var}(p_i) = \rho \pi_i(1 - \pi_i) \tag{3.10}$$

Then, using the properties  $E(Y_i) = E[E(Y_i | p_i)]$  and  $\text{var}(Y_i) = \text{var}[E(Y_i | p_i)] + E[\text{var}(Y_i | p_i)]$ , we obtain under these weaker conditions, Equations (3.7) and (3.8) respectively, as follows:

- $E(Y_i) = E[E(Y_i | p_i)] = E[E\{(R_i / n_i) | p_i\}] = (1/n_i)E(n_i p_i) = \pi_i$

- $\text{var}(Y_i) = \text{var}[E(Y_i | p_i)] + E[\text{var}(Y_i | p_i)] =$

$$\begin{aligned} & \text{var}[(1/n_i)E(R_i | p_i)] + E[(1/n_i)^2 \text{var}(R_i | p_i)] = \text{var}[(1/n_i)n_i p_i] + (1/n_i)^2 E[n_i p_i(1 - p_i)] = \\ & \text{var}(p_i) + (1/n_i)[E(p_i) - \text{var}(p_i) - (E(p_i))^2] = \rho \pi_i(1 - \pi_i) + (1/n_i)[\pi_i - \rho \pi_i(1 - \pi_i) - \pi_i^2] = \\ & [\pi_i(1 - \pi_i)]/n_i + [(n_i - 1)/n_i]\rho \pi_i(1 - \pi_i) \end{aligned}$$

We also model  $\pi_i = E(Y_i)$  as:

$$g(\pi_i) = \text{logit}(\pi_i) = \ln\{\pi_i / (1 - \pi_i)\} = \eta_i = x_i^t \beta \tag{3.11}$$

where  $i=1, \dots, k$  and  $g$  is the familiar link function. Particularly, Equation (3.11) defines a logistic regression model, in which the *logit* of  $\pi_i$  is a linear function of the explanatory variables. The parameters for the *i*th group are  $(\pi_i, \rho)$ . We note that considering the constraint  $\rho_i = \rho$ , the number of parameters is reduced from  $2k$  to  $k + 1$ . Furthermore, since  $\pi_i(\beta) = g^{-1}(x_i^t \beta)$ , we conclude that  $\pi_i$  is a function of the coefficients  $\beta$ , and therefore the parameters of real interest are the  $p$ -dimensional vector  $\beta$  and  $\rho$ .

The method of estimation described in this section depends only on the relationship between the expectation and variance of  $p_i$  as given from Equations (3.9) and (3.10), while the need to choose between different distributions which exhibit this relationship is avoided. Maximum likelihood cannot now be used because the distribution of the  $p_i$  is not fully specified.



but the structure of its mean and variance allows us to define quasi-likelihood equations (Wedderburn, 1974; McCullagh and Nelder, 1989). For fixed  $\rho$ , the variance of  $Y_i$  given by (3.8), is a function of  $\pi_i$  only, which depends on  $\beta$  through the inverse of the link function,  $g^{-1}$ . The score equation for the parameter  $\beta$  is:

$$\sum_{i=1}^k \frac{(Y_i - \pi_i(\beta))g'(\pi_i)}{\text{var}(Y_i)[g'(\pi_i)]^2} x_i = \sum_{i=1}^k (Y_i - \pi_i(\beta))x_i w_i g'(\pi_i) = 0 \quad (3.12)$$

where  $w_i = \{\text{var}(Y_i)[g'(\pi_i)]^2\}^{-1}$  are the GLM iterated weights. The estimate  $\hat{\beta}$  satisfies Equation (3.12) and the solution is called maximum quasi-likelihood estimate of  $\beta$  that may be obtained by iterated reweighted least squares as discussed by McCullagh and Nelder (1989). In the special case of  $\rho = 0$ , the Equations (3.12) reduce to the ordinary maximum likelihood equations for Binomial data. A reasonable estimate for  $\rho$  may be obtained by equating the Pearson chi-square statistic  $X^2 = \sum (Y_i - \hat{\pi}_i)^2 / \text{var}(Y_i)$  to its expected value. The asymptotic distribution of the statistic  $X^2$  under the null hypothesis is a chi-squared distribution with  $k - p$  degrees of freedom. Therefore, equating the statistic  $X^2$  to its expected value  $E(X^2) = k - p$  leads to:

$$\sum_{i=1}^k \frac{(Y_i - \hat{\pi}_i)^2}{\text{var}(Y_i)} = k - p \quad (3.13)$$

Summarizing all these results together, we derive the following iterative procedure for estimating the parameters  $\beta$  and  $\rho$  of the Beta-Binomial model (3.11) with response  $Y_i$ :

1. Given the current estimates  $\hat{\beta}^{(t)}$ ,  $\hat{\rho}^{(t)}$ ,  $w_i^{(t)}$  compute  $\hat{\beta}^{(t+1)}$  from (3.12).
2. Calculate  $\hat{\pi}_i^{(t+1)} = g^{-1}(\hat{\eta}_i^{(t+1)}) = g^{-1}(x_i' \hat{\beta}^{(t+1)})$  and obtain  $\hat{\rho}^{(t+1)}$  from (3.13) for fixed  $\hat{\pi}_i^{(t+1)}$ .
3. Compute  $w_i^{(t+1)} = \{\text{var}(Y_i^{(t+1)})[g'(\hat{\pi}_i^{(t+1)})]^2\}^{-1}$ , with  $\text{var}(Y_i^{(t+1)}) = [\hat{\pi}_i^{(t+1)}(1 - \hat{\pi}_i^{(t+1)})] / n_i + [(n_i - 1) / n_i] \hat{\rho}^{(t+1)} \hat{\pi}_i^{(t+1)}(1 - \hat{\pi}_i^{(t+1)})$ , as given from (3.8).

Return to step 1 and iterate between 1-3 steps until convergence.



Thus, using new weights, we estimate  $\beta$  iteratively from (3.12) and recalculate the statistic  $X^2$  in order to equate it to its expected value.

The parameters of the Beta-Binomial model can be estimated using different methods. For example, Tripathi, Gupta and Gurland (1994) propose some alternative methods for estimating the parameters in the Beta-Binomial (BB) and Truncated Beta-Binomial (TBB) models. The estimates of the parameters that are obtained from the different methods are then compared on the basis of their Asymptotic Relative Efficiency (ARE).

### 3.2.3 Other expansions of the Binomial distribution

The binomial distribution with parameters  $n$  and  $p$  may be generalized by permitting either or both to be realizations of random variables. Most applications consider  $n$  fixed and  $p$  varying according to a continuous probability density function defined on the interval  $(0,1)$ . For example, we mentioned that by allowing the parameters  $p_i$  of the binomial distribution to be beta-distributed, this leads to the Beta-Binomial model. The Beta-Binomial distribution is one of the most popular mixed binomial distributions to account for extra-binomial variation. However, the Binomial model can be expanded in many ways in order to adjust for overdispersion. In the following, we briefly discuss some of them.

In Section 3.2.2, the Beta-Binomial (BB) distribution was mentioned as a possible basis for correlated binary regression. It was pointed out that the common correlation  $\rho$  among pairs of binary responses within a litter or group lies between  $0 < \rho < 1$ . This means that the BB distribution does not permit underdispersion, so with a BB distributed variable the variance always exceeds the corresponding binomial variance. In contrast, underdispersed data would arise from negative values of  $\rho$  allowing the observed variance of the response to be less than the corresponding binomial quantity. Altham (1978) proposed an 'additive' and a 'multiplicative' generalization of the Binomial model, which naturally include the Binomial model as a special case. As Altham indicated, the BB model allows only positive correlation between

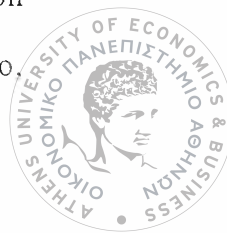


siblings of a litter, whereas the additive and the multiplicative generalization of the Binomial model allow negative or positive correlation. Generally, it is hard to say whether the multiplicative generalization is better or worse than the additive generalization of the Binomial. If the dependence between siblings of a litter is not strong, then the distinctions between these two generalizations are probably not of great practical importance. However, both generalizations are more flexible than the BB model, in the sense that they permit either overdispersion or underdispersion. Since the scope of this dissertation is to present certain ways in order to deal with overdispersion, we do not discuss in details these two generalizations of the Binomial Distribution. For those who want to further explore the additive and multiplicative generalization of the Binomial model, we propose to see Altham (1978).

Moreover, Equation (3.11) which defines the Beta-Binomial model assumes that the  $R_i | p_i \sim \text{Binomial}(n_i, p_i)$ ,  $i=1, \dots, k$ , and that the extra-binomial variation is variation in  $p_i$  about an expectation  $E(p_i) = \pi_i$  whose logit is  $\text{logit}(\pi_i) = x_i' \beta$ . An alternative model assumes that the logit of  $p_i$  varies about an expectation  $E\{\text{logit}(p_i)\} = x_i' \beta$  with a constant variance  $\text{Var}\{\text{logit}(p_i)\} = \sigma^2$ . This model was investigated by Pierce and Sands (1975) for analyzing data with extra-binomial variation. Following Pierce and Sands (1975), the model can be fitted using the procedure developed in Section 3.2.2 after some modifications for the weights  $w_i$  and the positive correlation  $\rho$ . If additionally,  $\text{logit}(p_i)$  is assumed to be normally distributed,  $\text{logit}(p_i) \sim N(x_i' \beta, \sigma^2)$ , then  $p_i$  follows a logistic-normal distribution whose properties and applications were investigated by Aitchison and Shen (1980). The logistic-normal model is defined as follows:

- $R_i | p_i \sim \text{Binomial}(n_i, p_i) \quad i=1, \dots, k$
- $\text{logit}(p_i) \sim N(x_i' \beta, \sigma^2) \Rightarrow$
- $\text{logit}(p_i) = \ln(p_i/(1-p_i)) = x_i' \beta + \varepsilon_i \quad i=1, \dots, k$  (3.14)

where the random variables  $R_i$  are assumed to be independent conditional on  $p_i$  and  $\varepsilon_i$  are independent random error terms with  $\varepsilon_i \sim N(0, \sigma^2)$ . Also



Goutis (1993) proposed a method of approximate maximum likelihood estimation of the parameters for the hierarchical logistic-normal model (3.14). He uses the EM algorithm with some approximations in order to facilitate its implementation.

Alanko and Duffy (1996) develop the class of binomial mixtures arising from transformation of the binomial parameter  $p_i$  as  $1 - \exp(-\lambda_i)$ . Here,  $\lambda_i$  is treated as a random variable and is mixed by a probability distribution defined on  $(0, \infty)$ . Particularly, they present Sichel's (1975, 1982) generalized inverse Gaussian (GIG) family. The GIG family is a three-parameter family of mixing distributions and contains many distributions as special cases. For example, both the Grassia II-binomial distribution (Grassia, 1977) and the transformed inverse Gaussian (IG) binomial distribution (Tweedie, 1957) belong to the GIG family. Grassia examines the transforms  $p_i = 1 - \exp(-\lambda_i)$  and  $p_i = \exp(-\lambda_i)$  when  $\lambda_i$  follows the gamma distribution. On the other hand, the transformed IG binomial distribution assumes that  $\lambda_i$  follows the inverse Gaussian distribution. For further details see Alanko and Duffy (1996).

Finally, Booth and Caffo (2002) propose to treat the beta-distributed parameters  $p_i$  of the binomial distribution as missing data and to perform the Monte Carlo EM algorithm for estimating the corresponding model.

### 3.3 Testing the goodness of fit of the Binomial distribution

We can use Tarone's  $Z$  statistic to test the goodness of fit of the Binomial distribution against the Beta-Binomial distribution (Tarone, 1979). The null hypothesis is that the underlying distribution is a Binomial distribution while the alternative hypothesis is that the underlying distribution is a Beta-Binomial distribution. Since the case of the Binomial distribution corresponds to  $\rho = 0$ , testing the goodness of fit of the Binomial distribution is equivalent to test the null hypothesis  $H_0 : \rho = 0$ . Note however, that in the Beta-Binomial model the parameter  $\rho$  cannot take negative values and thus the alternative hypothesis is necessarily one sided. Therefore, a test of the



goodness of fit of the Binomial distribution is obtained by testing the null hypothesis  $H_0 : \rho = 0$  against the alternative  $H_1 : \rho > 0$ . Tarone's standardized one-sided test of the Binomial distribution versus Beta-Binomial alternatives, is based on the statistic:

$$Z = \left( S - \sum_{i=1}^k n_i \right) / \left\{ 2 \sum_{i=1}^k n_i (n_i - 1) \right\}^{\frac{1}{2}} \tag{3.15}$$

where

$$S = \sum_{i=1}^k (R_i - n_i \hat{p})^2 / (\hat{p}\hat{q}) \tag{3.16}$$

$$\hat{p} = \sum_{i=1}^k R_i / \sum_{i=1}^k n_i \tag{3.17}$$

$\hat{q} = 1 - \hat{p}$  and  $n_i, R_i$ , denote the number of observations and the number of successes in group  $i = 1, \dots, k$ , respectively.

The statistic  $Z$  in (3.15) has an asymptotic standard normal distribution under the null hypothesis  $H_0 : \rho = 0$ , meaning that  $Z \sim N(0,1)$ . We can use this statistic to test the goodness of fit of a Binomial distribution without estimating the parameter  $\rho$  using the BB model. When the actual  $\rho$  is zero, it is difficult to estimate the parameters  $\beta$  and  $\rho$  of the BB model, since the iterative maximization process fails to converge at this point (Vuataz and Sotek 1978; Crowder 1978). The  $Z$  statistic is recommended as a goodness of fit test statistic because of its higher power and computational simplicity.

In addition, Fowlkes (1987) proposes some statistical techniques based on smoothing for the assessment of the fit of binary logistic models. He describes a battery of diagnostic tools associated with the smoothed binary responses. However, he suggests that, as with any smoothing technique, we must be cautious with the size of the smoothing window that we choose.

Finally, Fitzmaurice, Heath and Cox (1997) propose a simple general method for detecting overdispersion. Their approach consists of grouping the data into a number of stratum of approximately equal size. However, they underline that the choice of stratum cannot be made through any automatic procedure and therefore we must be cautious. Under the assumption that



observations are independent, meaning under the standard logistic regression model, there is a direct relationship between the nominal standard errors and the empirical or sample standard deviation of the parameters estimates obtained from each of the separate strata. Any departure from this relationship indicates that overdispersion is present and consequently the standard logistic regression model does not fit the data adequately.



## CHAPTER 4

### Effect of the number of quadrature points and comparison of the results between different programs: 3 examples

#### 4.1 Introduction

Before the description and the analysis of our data that we will widely discuss in Chapter 5, in this Chapter we present 3 data sets with 3 different models respectively. Furthermore, we apply 3 methods for estimating each model. In particular, we perform the Gauss-Hermite technique, the penalized quasi-likelihood method and the Beta-Binomial model, meaning that we assume that the distribution of the response variable is something more dispersed than usual. Our aim through these examples is to compare the results derived from the different methods and find out whether the interpretation of them is similar or not and to what extent. Furthermore, we examine the effect of the number of quadrature points with the Gauss-Hermite method on the results. For estimating the models, we have used the statistical packages MIXOR and SAS. As far as the SAS statistical program, we have used the SAS macro GLIMMIX and the SAS procedure LOGISTIC. The SAS macro GLIMMIX and the SAS procedure LOGISTIC require SAS version 6.12, or later, to run.

The organization of this Chapter is as follows: In Section 4.2 we describe the 3 data sets and in Section 4.3 we define the 3 logistic random-intercept models. Before fitting the models, Section 4.4 explains the need to account for the variation among level-2 units. Section 4.5 is focused on the effect of the number of quadrature points on the results and Section 4.6 compares the results obtained from the 3 methods. Eventually, in Section 4.7 we point out some important conclusions.



## 4.2 Description of the data for each example

A brief description of the data for each example is given as follows:

- 1) The first example analyzes the data from Beitler and Landis (1985), which represent results from a multi-center clinical trial investigating the effectiveness of two topical cream treatments (active drug, control) in curing an infection. For each of eight clinics, the number of trials  $n_{ij}$  and favorable cures  $x_{ij}$  are recorded for each treatment, (Table 4.1).
- 2) The second example is the overdispersed logistic regression model for incidence of toxoplasmosis in 32 cities of El Salvador, discussed by Efron (1986) and Francis *et al.* (1993). Table 4.2 shows the number  $n_i$  of people tested and the number  $r_i$  with a positive test for toxoplasmosis in each city, together with the annual city rainfall  $x_i$  in metres.
- 3) The last example analyzes data from Crowder (1978). In particular, Crowder presented data on the proportion of seeds  $r_{ijk}/n_{ijk}$  that germinated on each of 21 plates arranged according to a 2x2 factorial layout by seed variety and type of root extract. The two types of seed are *O.aegyptiaca* 75 and *O.aegyptiaca* 73, and the two root extracts are Bean and Cucumber, (Table 4.3).

As we have previously mentioned, MIXOR and SAS are the statistical packages that have been used for analyzing the above examples. In particular, MIXOR gives results according to the Gauss-Hermite method as described in Section 2.6. Through the SAS macro GLIMMIX we perform the penalized quasi likelihood (PQL) method with the dispersion parameter fixed at unity, as discussed in Section 2.3. Finally, we use the SAS procedure LOGISTIC, which corresponds to the theory developed in Section 3.2.2, in order to deal with overdispersion.



Clinic	$x_{ij}$	$n_{ij}$	$t_j$	Clinic	$x_{ij}$	$n_{ij}$	$t_j$
1	11	36	1	5	6	17	1
1	10	37	0	5	0	12	0
2	16	20	1	6	1	11	1
2	22	32	0	6	0	10	0
3	14	19	1	7	1	5	1
3	7	19	0	7	1	9	0
4	2	16	1	8	4	6	1
4	1	17	0	8	6	7	0

Table 4.1: Data from Beitler and Landis (1985)

City	$r_i$	$n_i$	$x_i$	City	$r_i$	$n_i$	$x_i$
1	2	4	1.735	17	33	54	1.770
2	3	10	1.936	18	4	9	2.240
3	1	5	2.000	19	5	18	1.620
4	2	2	1.750	20	0	1	1.650
5	3	5	1.800	21	8	11	2.250
6	2	8	1.750	22	41	77	1.796
7	7	19	2.077	23	24	51	1.890
8	3	6	1.920	24	7	16	1.871
9	8	10	1.800	25	46	82	2.063
10	7	24	2.050	26	9	13	2.100
11	0	1	1.830	27	23	43	1.918
12	15	30	1.650	28	53	75	1.834
13	4	22	2.200	29	8	13	1.780
14	0	1	2.000	30	3	10	1.900
15	6	11	1.770	31	1	6	1.976
16	0	1	1.920	32	23	37	2.292

Table 4.2: Data from Efron (1986) and Francis et al. (1993)

O.aegyptiaca 75				O.aegyptiaca 73			
Bean		Cucumber		Bean		Cucumber	
$r_{ijk}$	$n_{ijk}$	$r_{ijk}$	$n_{ijk}$	$r_{ijk}$	$n_{ijk}$	$r_{ijk}$	$n_{ijk}$
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

Table 4.3: Data from Crowder (1978)



### 4.3 Definition of the logistic random-intercept models

For the Gauss-Hermite technique by MIXOR and the PQL method by the SAS macro GLIMMIX, we next give the 3 following models that have been fitted for analyzing the 3 preceding examples respectively:

1) Suppose  $n_{ij}$  denotes the number of trials for the  $i$ th clinic and the  $j$ th treatment ( $i = 1, \dots, 8$   $j = 0, 1$ ), and  $x_{ij}$  denotes the corresponding number of favorable cures.

Then, a reasonable model for the first data set is the following logistic model with random effects:

$$\eta_{ij} = g(\mu_{ij}) = \log(p_{ij}/(1-p_{ij})) = \beta_0 + \beta_1 t_j + b_{1i} \quad (4.1)$$

where  $x_{ij} | p_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$ . The notation  $t_j$  indicates the  $j$ th treatment and the  $b_{1i}$  are assumed to be independently normally distributed  $b_{1i} \sim N(0, \sigma_1^2)$ .

2) The logistic random-intercepts model for the second example outlined above is given by:

$$\eta_i = \log(p_i/(1-p_i)) = \beta_0 + \beta_1 x_i + b_{2i} \quad (4.2)$$

where  $r_i | p_i \sim \text{Binomial}(n_i, p_i)$ . The notation  $x_i$  corresponds to the annual rainfall in metres in the  $i$ th city ( $i = 1, \dots, 32$ ). Once again, we assume that the random effects  $b_{2i}$  are independently normally distributed  $b_{2i} \sim N(0, \sigma_2^2)$ .

3) In order to handle the plate-to-plate variability in Crowder's example, the following GLMM is proposed:

$$\eta_{ijk} = g(\mu_{ijk}) = \beta_0 + \beta_1 s_j + \beta_2 e_k + \beta_3 s_j * e_k + b_{3i}, \quad i = 1, \dots, 21 \quad (4.3)$$

where  $s_j, e_k$  and  $s_j * e_k$  represent the seed, root extract and their interaction respectively. In addition, the symbol  $j = 0, 1$  represent the seed variety and the symbol  $k = 0, 1$  the type of root extract. Particularly, the *O.aegyptiaca* 75 seed corresponds to  $j = 1$  and the Bean root extract corresponds to  $k = 1$ . The independent random effects associated with each plate are denoted by  $b_{3i} \sim N(0, \sigma_3^2)$ .

For technical details about the maximization part of MIXOR, the SAS macro GLIMMIX and the SAS procedure LOGISTIC, we refer the reader to



the manual of the statistical packages. Briefly, MIXOR uses marginal maximum likelihood estimation, utilizing a Fisher-scoring solution. For the scoring solution, the Cholesky factor of the random effects variance-covariance matrix is estimated, along with the effects of model covariates. Multi-dimensional quadrature is used to numerically integrate over the distribution of the random effects. The SAS macro GLIMMIX uses, by default, restricted/residual pseudo likelihood (REPL) to find the parameter estimates of the generalized linear mixed model we specify. The macro was originally written to estimate the pseudo-likelihood function of Wolfinger and O'Connell, (1993). For more details about the macro see Appendix B.2. The SAS procedure LOGISTIC (detailed in Appendix B.1) provides two iterative maximum likelihood algorithms. The default is the Fisher-scoring method, which is equivalent to fitting by Iteratively Reweighted Least Squares (IRLS). The alternative algorithm is the Newton-Raphson method. Both algorithms give the same parameter estimates. However, the estimated covariance matrix of the parameter estimators may differ slightly. In the case of a binomial family with logit link (the canonical link), the Fisher-scoring and Newton-Raphson methods are equivalent, resulting in identical estimated covariance matrices for both algorithms. Finally, we note that the theory in the manual of the SAS procedure LOGISTIC (referred to the manual as the Williams' method) is similar to the theory discussed in Section 3.2.2, where the variance of the response variable is greater than that under Binomial data and a scale parameter  $\rho$  is incorporated in its variance in order to account for overdispersion.

#### 4.4 Detecting Overdispersion

Next, we consider the standard logistic model for the above examples, where  $\sigma_1^2, \sigma_2^2$  and  $\sigma_3^2$  are assumed to be equal to 0. The results are shown in Table 4.4. Fitting the models, it is obvious from Table 4.4 that the residual deviances show substantial overdispersion, since the deviances measures of goodness-of-fit exceed their degrees of freedom (d.f). For example, the first model gives a deviance of 90.9602 on 14 d.f. Moreover, the  $-2 \log$  Likelihood



value for this model is equal to 358.234 and the p-value which is equal to 0.0001 (from a Wald test) suggests that the fitted model is not an acceptable fit to the data at 5% significance level.

Data from:	Deviance	Df	Deviance/Df	p-value	-2logL
<i>Beitler &amp; Landis</i>	90.9602	14	6.4972	0.0001	358.234
<i>Efron &amp; Francis et al.</i>	65.6198	30	2.1873	0.0002	934.302
<i>Crowder</i>	33.2778	17	1.9575	0.0104	1086.221

**Table 4.4:** The residual deviances under the standard logistic linear model

#### 4.5 Implication of the number K of quadrature points with MIXOR

Even though MIXOR indicates that K=10 quadrature points are often appropriate for one random effect, we will examine the dependence of the results on K. As an illustration of the number of quadrature points on the calculations, we analyzed the logistic random-effects models (4.1), (4.2) and (4.3) with MIXOR for K ranging from 10 to 30 in steps of 5 (Tables 4.5, 4.6 and 4.7-4.8 respectively).

Data from:	K	Starting value	loglik	$\hat{\beta}_1$	SE	p-value	$\hat{\sigma}_1^2$	$\hat{\rho}_1$
<i>Beitler</i>	10	Program	-158.091	0.687	0.530	0.195	3.620	0.524
		User	-152.270	0.680	0.414	0.100	2.674	0.448
	15	Program	-151.782	0.680	0.492	0.167	4.117	0.556
		User	-150.558	0.759	0.321	0.018	1.860	0.361
<i>and</i>	20	Program	-150.741	0.758	0.347	0.028	2.480	0.430
		User	-150.741	0.758	0.347	0.028	2.480	0.430
<i>Landis</i>	25	Program	-150.700	0.754	0.297	0.011	2.724	0.453
		User	-150.700	0.754	0.297	0.011	2.724	0.453
	30	Program	-151.476	0.745	0.412	0.070	1.636	0.332
		User	-151.210	0.748	0.383	0.050	2.873	0.466

**Table 4.5:** Effect of the number of quadrature points K on the output when fitting model (4.1) with MIXOR

For model (4.1) (Table 4.5), we observe that for K=20 and K=25 the treatment effect is significant at  $\alpha=0.05$  significant level for both program and user-defined starting values. At this point, we note that the user-defined starting values for the fixed effects were obtained from the ordinary GLM fits



and that we considered as an initial estimate for all the variances  $\hat{\sigma}_1^2$ ,  $\hat{\sigma}_2^2$  and  $\hat{\sigma}_3^2$  the value of 1. In addition, we observe that with  $K=10, 15$  and  $30$ , and with user-defined starting values, higher log-likelihoods (loglik) are obtained at convergence than with program-defined initial values. Hence, the analyses with the starting values defined by the program converge to a local maximum. For example, at  $K=15$ , the program MIXOR converges to a value of  $-151.782$  for the log-likelihood with starting values defined by the program and to a value of  $-150.558$  with user-defined starting values. This discrepancy produces the different estimates for the treatment effect, for its standard error (SE) and consequently has a considerable effect on its p-value, whether the program or user starting values are defined. For instance, with  $K=15$  and with program-defined starting values the estimated treatment effect is equal to  $0.68$  ( $SE=0.492$ ;  $p\text{-value}=0.167$ ), whereas with  $K=15$  and with user-defined starting values it is  $0.759$  ( $SE=0.321$ ;  $p\text{-value}=0.018$ ). However, the estimated treatment effects are all positive, and therefore these small differences between the estimates are not material. Finally, in the last two columns of Table 4.5, the estimated variance  $\hat{\sigma}_1^2$  of the random intercept and the estimated intra-clinic correlation coefficient  $\hat{\rho}_1$  are given. It is obvious from the last two columns that the estimates  $\hat{\sigma}_1^2$  and  $\hat{\rho}_1$  differ depending on the number  $K$  of the quadrature points (actually expected) and on whether program or used starting values are defined.

Data from:	K	Starting value	loglik	$\hat{\beta}_1$	SE	p-value	$\hat{\sigma}_2^2$	$\hat{\rho}_2$
	10	Program	-461.648	-0.124	0.728	0.864	0.249	0.070
		User	-461.648	-0.124	0.729	0.864	0.248	0.070
<i>Efron</i>	15	Program	-461.730	-0.180	0.680	0.790	0.223	0.063
		User	-461.730	-0.180	0.681	0.791	0.223	0.063
<i>and</i>	20	Program	-461.728	-0.168	0.685	0.805	0.224	0.064
		User	-461.728	-0.168	0.686	0.806	0.224	0.064
<i>Francis et al.</i>	25	Program	-461.728	-0.171	0.685	0.802	0.224	0.064
		User	-461.728	-0.171	0.685	0.802	0.224	0.064
	30	Program	-461.728	-0.171	0.685	0.802	0.224	0.064
		User	-461.728	-0.171	0.685	0.802	0.224	0.064

**Table 4.6:** Effect of the number of quadrature points  $K$  on the output when fitting model (4.2) with MIXOR



As far as the data from Efron & Francis et al., (Table 4.6), we observe that the outcome does not depend on the starting values, since identical values of the log-likelihood are obtained at convergence whether program or user-defined starting values are defined. Furthermore, from  $K=20$  to  $K=30$ , MIXOR converges to the same value of  $-461.728$  for the log-likelihood and all the results in columns 5 to 9 are equal or almost equal depending on whether  $K=20, 25$  or  $30$ . Therefore, we suppose that a good choice for the required number  $K$  of quadrature points is  $K=20$ . However, we must be cautious with the choice of  $K$  in order to obtain a global maximum rather than local maxima. To overcome this problem, it is recommended to use different starting values for the maximization routine and compare the results each time. We also note that all the estimated fixed effects in column 5 are negative. We keep on emphasizing this comment due to its importance on the interpretation of the results.

At last, fitting model (4.3) gives results that are summarized in Tables 4.7 and 4.8. We again observe that the results do not depend on the defined starting values (by the program or user). There is just one exception at  $K=10$  with user-defined starting values, where it is apparent that MIXOR either converges to a local maximum with a value of  $-563.961$  ( $< -541.931$ ) for the log-likelihood (Table 4.7), or the program stops before reaching the maximum. We note that the estimates  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  and  $\hat{\beta}_3$ , correspond to the estimated fixed effects of seed, root extract and their interaction respectively and that their p-values are denoted by  $p_1$ -value,  $p_2$ -value and  $p_3$ -value respectively (columns 4-6; Table 4.8). At last, from Table 4.7 we conclude that convergence is probably obtained to the global maximum from  $K=15$  onwards.



Data from:	K	Starting value	loglik	$\hat{\beta}_1$ (s.e)	$\hat{\beta}_2$ (s.e)	$\hat{\beta}_3$ (s.e)
<i>Crowder</i>	10	Program	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
		User	-563.961	0.628 (0.190)	-88.263 (199.732)	-0.815 (0.262)
	15	Program	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
		User	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
	20	Program	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
		User	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
	25	Program	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
		User	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
	30	Program	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)
		User	-541.931	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)

**Table 4.7:** Effect of the number of quadrature points K on the output when fitting model (4.3) with MIXOR

Data from:	K	Starting value	p <sub>1</sub> -value	p <sub>2</sub> -value	p <sub>3</sub> -value	$\hat{\sigma}_3^2$	$\hat{\rho}_3$
<i>Crowder</i>	10	Program	0.014	0.096	0.038	0.056	0.017
		User	0.0009	0.658	0.0019	81.885	0.961
	15	Program	0.014	0.096	0.038	0.056	0.017
		User	0.014	0.096	0.038	0.056	0.017
	20	Program	0.014	0.096	0.038	0.056	0.017
		User	0.014	0.096	0.038	0.056	0.017
	25	Program	0.014	0.096	0.038	0.056	0.017
		User	0.014	0.096	0.038	0.056	0.017
	30	Program	0.014	0.096	0.038	0.056	0.017
		User	0.014	0.096	0.038	0.056	0.017

**Table 4.8:** Effect of the number of quadrature points K on the output when fitting model (4.3) with MIXOR

#### 4.6 Comparing the results between different programs

Next, the results from fitting models (4.1)-(4.3) are summarized in Tables 4.9, 4.10 and 4.11–4.12 respectively. This time our aim is not to compare the results within the same program (as we previously did with MIXOR), but to compare the results obtained from the different programs MIXOR, the SAS macro GLIMMIX and the SAS procedure LOGISTIC. We have already mentioned that with K=15 and with program-defined starting values, the analysis for model (4.1) converges to a local maximum. Therefore, in Table 4.9 we keep our attention only on user-defined starting values at K=15.



Program	K	Starting value	Method of Analysis	$\hat{\beta}_1$	SE	p-value	$\hat{\sigma}_1^2$	$\hat{\rho}_1$
<i>MIXOR</i>	15	Program		0.680	0.492	0.167	4.117	0.556
	20	Program	Gauss-	0.758	0.347	0.028	2.480	0.430
	15	User	Hermite	0.759	0.321	0.018	1.860	0.361
	20	User		0.758	0.347	0.028	2.480	0.430
<i>GLIMMIX</i>	–	User	PQL	0.724	0.296	0.014	2.032	–
<i>LOGISTIC</i>	–	Program	Williams'	0.525	0.680	0.440	–	0.364

**Table 4.9:** Fitting model (4.1) with different programs

Table 4.9 shows clearly that the estimated treatment effect do not vary considerably between programs, except of the value of 0.525 from the SAS procedure LOGISTIC. However, the positive values of the estimates indicate that the treatment effect increases the chance of a favorable cure. Furthermore, at K=15 (user) and 20 with MIXOR and with the SAS macro GLIMMIX, we observe that the treatment effect is statistical significant at  $\alpha = 5\%$  significant level, since  $\alpha = 0.05 > p\text{-value}$  (from the Wald statistic). On the other hand, this is not true for the SAS procedure LOGISTIC with a p-value equal to 0.44. In addition, we observe that the estimated variance of the random intercept  $\hat{\sigma}_1^2$  varies with the higher (smaller) value of 2.48 (1.86). As far as the estimated intra-clinic correlation coefficient  $\hat{\rho}_1$ , it has the higher (smaller) value of 0.43 (0.361).

Summarizing the above results, we conclude that there are small differences between MIXOR and the SAS macro GLIMMIX. In contrast, the results from the LOGISTIC procedure differ considerably from the other results, since it produces a treatment effect, which is not statistically significant. We also note that the outputs of the SAS macro GLIMMIX and the SAS procedure LOGISTIC do not produce the estimated intra-cluster correlation coefficient and the estimated variance of the random effect respectively. Moreover, we remind that we performed the PQL analysis provided by the SAS macro GLIMMIX with the scale parameter fixed at unity. In particular, the user-starting values for the macro GLIMMIX are defined in such a way so as to correspond to the ordinary GLM fit with an initial value for the cluster variance equal to 0. Therefore, by assuming a compound symmetry (CS) covariance structure for the independent random



effects with an initial value for the cluster variance equal to 0, and by keeping the common covariance of the cs structure fixed at 0 and the residual variance fixed at 1, a diagonal covariance matrix is produced for the random effects with equal diagonal elements.

Afterwards, we present the results from fitting model (4.2) as shown in Table 4.10. The results are more straight-out for this example, since they are quite similar between the different programs.

Program	K	Starting value	Method of Analysis	$\hat{\beta}_1$	SE	p-value	$\hat{\sigma}_2^2$	$\hat{\rho}_2$
MIXOR	20	Program	Gauss-	-0.168	0.685	0.805	0.224	0.064
	20	User	Hermite	-0.168	0.686	0.806	0.224	0.064
GLIMMIX	-	User	PQL	-0.164	0.736	0.823	0.248	-
LOGISTIC	-	Program	Williams'	-0.145	0.723	0.841	-	0.062

**Table 4.10:** Fitting model (4.2) with different programs

More specifically, the estimates  $\hat{\beta}_1$  of the fixed effect of the annual city rainfall as produced by the 3 programs are very close to each other, apart from the estimate (-0.145) by the SAS procedure LOGISTIC. However, this does not concern us a lot, since all the estimates in column 5 have negative values and all the p-values in column 7 suggest that the annual city rainfall does not affect the proportion of the people testing positively for toxoplasmosis in the 32 cities. We only note that MIXOR gives smaller standard errors than the other two programs. In the last two columns the estimated variance of the random intercept and the estimated intra-city correlation coefficient are given respectively. In particular, the estimated intra-city correlation coefficient has a value of 0.064 or 0.062, meaning that 6.4% or 6.2% respectively in the variation of the response is due to variation among cities.

At last, Tables 4.11 and 4.12 show the results obtained from fitting model (4.3). We observe that there are small differences between the 3 programs. There is just one point that concerns us due to its importance on the interpretation of the results. Even though MIXOR suggests that the interaction term of the seed and root extract is statistically significant ( $p_3$ -value=0.038),



the other two programs do not ( $p_3=0.054$  and  $p_3=0.059$  for GLIMMIX and LOGISTIC respectively).

Program	K	Starting value	Method of Analysis	$\hat{\beta}_1$ (s.e)	$\hat{\beta}_2$ (s.e)	$\hat{\beta}_3$ (s.e)	$\hat{\rho}_3$
MIXOR	15	Program	Gauss-	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)	0.017
	15	User	Hermite	0.713 (0.290)	-0.526 (0.316)	-0.810 (0.392)	0.017
GLIMMIX	-	User	PQL	0.748 (0.299)	-0.513 (0.334)	-0.825 (0.429)	-
LOGISTIC	-	Program	Williams'	0.749 (0.304)	-0.510 (0.334)	-0.819 (0.435)	0.024

Table 4.11: Fitting model (4.3) with different programs

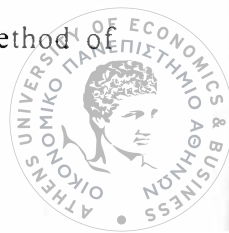
Program	K	Starting value	Method of Analysis	$p_1$ -value	$p_2$ -value	$p_3$ -value	$\hat{\sigma}_3^2$
MIXOR	15	Program	Gauss-	0.014	0.096	0.038	0.056
	15	User	Hermite	0.014	0.096	0.038	0.056
GLIMMIX	-	User	PQL	0.012	0.124	0.054	0.097
LOGISTIC	-	Program	Williams'	0.013	0.127	0.059	-

Table 4.12: Fitting model (4.3) with different programs

#### 4.7 Conclusions

In this Chapter we shortly analyzed 3 different data sets by using MIXOR, the SAS macro GLIMMIX and the SAS procedure LOGISTIC. Our goal was to explore any differences between, or within, the programs on fitting the logistic random-intercept models (4.1), (4.2) and (4.3). Summarizing the above results, we conclude that:

- The choice of the number K of quadrature points plays a major role on whether a global maximum or local maxima are obtained at convergence for the log-likelihood (see also Lesaffre and Spiessens, 2001). It is usually suggested that local maxima may require variations in starting values for the iterative algorithm to locate all the local maxima.
- We observed that results may vary considerably both between programs as well as within the same program. Even though small differences between programs are expected, because each program applies its unique method of



analysis, large differences should concern us. Large differences between the programs are very disturbing, especially when they affect the interpretation of the results.

In order to overcome such computational difficulties some authors suggest using the adaptive Gaussian quadrature approach (Lesaffre and Spiessens, 2001), which has been recently released by the SAS procedure NLMIXED. However, it is their experience that, even with adaptive Gaussian quadrature and with relatively simple models, convergence to a global maximum may be again difficult to obtain. This emphasizes the computational difficulties with random-effects models for categorical outcome data.



2



## CHAPTER 5

### Causes of mortality in piglets in Greek commercial pig farms

#### 5.1 Introduction

This Chapter presents some causes of mortality in piglets in 23 Greek industrial pig farms. The variables that have been used in this study include characteristics of the commercial pig farms, the sows and the piglets. The statistical unit was the sow and its litter. Our aim is to investigate variables that have an effect to the mortality rate and to discard unnecessary variables.

Specifically, in Section 5.2 we briefly discuss about the way the sample of the pig farms and sows was selected. Section 5.3 is focused on the description of the data and in Section 5.4 we estimate the standard logistic regression model and explain the need to account for the variation among the twenty-three piggeries. In Section 5.5 we estimate the Beta-Binomial (BB) model through the SAS procedure LOGISTIC and we perform Tarone's one-sided test of the Binomial distribution versus Beta-Binomial alternatives, as discussed in Section 3.3. In Section 5.6 we perform the penalized quasi-likelihood (PQL) method and the Gauss-Hermite technique, using the SAS macro GLIMMIX and MIXOR, respectively. Unfortunately, the BB model and the PQL analysis do not give satisfactory results. In particular, all the variables were not found to be statistically significant for the mortality rate. As far as MIXOR, did not produce any results at all. In such a case, Hedeker and Gibbons who wrote MIXOR recommend in the manual of the program that we should try to fit a model with no random effects and therefore in Section 5.7 we turn back to the standard logistic linear model.



## **5.2 Some information about the way the sample of the pig farms and sows was selected**

Before the description of the data in the next section, it appears to us that it is necessary to give a brief description about the way the sample of the piggeries and sows was selected and generally the way this survey has been done. Even though the way of collecting the data or the sampling scheme that has been used are not the purpose of this dissertation, they seem always very helpful to any further statistical analysis one may want to perform. More specifically, in this study only veterinarians completed the relevant questionnaires. The whole study completed after one year. Furthermore, twenty-three industrial pig farms all over Greece were selected. A multistage sampling scheme was used in order to select the twenty-three piggeries. At the first stage, the farms of industrial type were stratified according to their size. Finally, from all the strata, twenty-three farms were selected in total. This choice was based on some criteria, such as the occupation of veterinary staff on a permanent basis, the size of the unit relatively to the number of the sows, the unit's data collection system, the geographical density and the geographical position of the piggeries. At the second stage, sows were randomly selected from each farm. The number of sows per farm was proportional to the farm size. They represent a mean percentage of 27.7% of the total number of sows of each farm included in the sample, this number varying from 12.42% to 40%. Sows that have realized parturition could not be included again in the survey.

## **5.3 Description of the Data**

The variables that have been used in this research include characteristics of the pig farms (neighboring with other farms, size, geographical position etc.), the sows (whether the sows are vaccinated for the Aujeszky virus or not) and the litters (the number of piglets born, the number of dead piglets after one year etc.). These variables represent some of the factors that may affect the mortality rate, such as environmental or medical factors etc. We are interested in finding which of the variables have an effect



to the dependent variable. Firstly, we give a description of all the available variables, as follows:

- **FARMS:** the serial number of the commercial pig farms (1 to 23).
- **PIGSNUMB:** the number of piglets total born in each farm.
- **DIED:** the total number of piglets being dead at the end of the study.
- **SOWSNUMB:** the number of the sows in each farm. Actually, this variable shows the size of each unit-piggery.
- **AREAPRO:** the cell's area for the pre-fattening period of the piglets (measured in m<sup>2</sup>). The relevant measurements refer to the average of the cell's area for all the piglets in a particular farm.
- **AREA:** the geographical position of the piggeries. Six geographical positions have been selected in total all over Greece: Greek Mainland, Epirus, Thessaly, Macedonia, Crete and Peloponnese.
- **AIRVOLUM:** the air's volume in the fattening cells per piglet.
- **WORKERS:** the piggery's quotient numbers of sows/workers. These numbers express the average number of sows that are supervised by one worker in each farm during the study.
- **TAFROS:** a trench with a disinfectant material located in the unit's entry.
- **DISTAN:** the presence of other piggeries in the neighborhood of a concrete pig farm. The relevant measurements are based on whether the distance of the other enterprises is less or more than 3 kilometers from a particular farm.
- **PRRS:** the medical history of each enterprise due to Porcine Reproduction and Respiratory Syndrome (PRRS). It is about a disease that causes many losses in the piggeries.
- **MIKOTOX:** whether the piglet's feeding is examined for toxins or not. The pig farms may never, or rarely, check the piglet's feedings, or may examine them on regular, or not, periods.
- **FLIES:** whether there is a protection system in the farms for the flies or not.
- **VET:** whether the occupation of the veterinary staff is on a permanent basis, or it is organized so as only some visits are taken place.



- **AUJE**: whether the sows are vaccinated during their pregnancy for the Aujeszky virus or not.

The response of interest is binary with values of 1 or 0, indicating whether the individual-piglet has been dead or alive, respectively, at the end of the experiment. We note that we consider as a ‘success’ the death of a pig. From Table 5.1, we observe that only 2.22% of the piglets were dead at the end of the study and 97.8% were alive.

Category	Frequency	Proportion
1	558	0.0222
0	24522	0.9778

**Table 5.1:** Categories of the response variable

However, we have only the total number of the piglets being dead in each farm and therefore the outcomes are expressed as the proportions of the dead piglets from each farm. Considering the mortality rate as the variable of interest, we perform statistical analysis in order to study how the dependent variable, the mortality rate, is related to a set of variables and to explore any explanatory variables that may affect the mortality rate.

#### 5.4 Fitting the standard logistic linear model

As known, the standard logistic regression model is a powerful modeling tool for analysing data with the response variable expressed as a proportion. However, there is typically a choice of link functions according to the family of distributions. Specifically, for the binomial family, the standard link functions may be either the *logit* ( $\log(p/(1-p))$ ), or *probit* ( $\phi^{-1}(p)$ ), or complementary *log-log* ( $\log(-\log(1-p))$ ), or *log-log* ( $-\log(-\log(p))$ ) link function. The results of plotting these link functions versus the explanatory variables did not differ a lot. Meaning that, the assumption of linearity between the link function and the independent variables does not depend to a great extent on the choice of the link function, and therefore we use the *logit* transformation for the mathematical convenience.

For our data, the individual units-piglets are classified into  $k = 1, \dots, 23$  farms. Let  $n_i$  denote the number of piglets total born in farm  $i$  ( $1 \leq i \leq k$ ), and let  $R_i$  denote the number of piglets being dead in that farm. Assume also that conditional on  $p_i = R_i/n_i$ ,  $R_i | p_i \sim \text{Binomial}(n_i, p_i)$ . Considering the outcome as the mortality rate  $p_i$  in farm  $i$ , we have  $p_i = R_i/n_i \sim \text{Binomial}(n_i, p_i)/n_i$ . Associated with the  $i$ th proportion ( $i = 1, \dots, 23$ ) are seventeen explanatory variables. We note that the number of the variables, eighteen including the intercept, is quite large while the number of the observations is only twenty-three. The logistic model for our data set is given by:

- $$\eta_i = g(p_i) = \text{logit}(p_i) = \ln(p_i/(1-p_i)) = \beta_0 + \beta_1(X_{1i} - \bar{X}_1) + \beta_2(X_{2i} - \bar{X}_2) + \beta_3 ar_{1i} + \beta_4 ar_{2i} + \beta_5 ar_{3i} + \beta_6 ar_{4i} + \beta_7 ar_{5i} + \beta_8(X_{3i} - \bar{X}_3) + \beta_9(X_{4i} - \bar{X}_4) + \beta_{10} auje_i + \beta_{11} taf_i + \beta_{12} dis_i + \beta_{13} pr_i + \beta_{14} mik_{1i} + \beta_{15} mik_{2i} + \beta_{16} fl_i + \beta_{17} vet_i \quad (5.1)$$

where  $\eta_i$ ,  $g(p_i)$  are the linear predictor and the *logit* link for the  $i$ th farm, respectively,  $X_{1i} - \bar{X}_1$ ,  $X_{2i} - \bar{X}_2$ ,  $X_{3i} - \bar{X}_3$  and  $X_{4i} - \bar{X}_4$  correspond to the centered values of the continuous variables Sowsnumb, Areapro, Airvolum and Workers, respectively, for the  $i$ th farm. We use these variables than the original ones, in order to make easier the interpretation of the results. In addition,  $ar_{1, \dots, 5i}$ ,  $auje_i$ ,  $taf_i$ ,  $dis_i$ ,  $pr_i$ ,  $mik_{1, 2i}$ ,  $fl_i$  and  $vet_i$  represent the variables Area, Auje, Tafros, Distan, Prrs, Mikotox, Flies and Vet, respectively. Particularly,  $ar_{1, \dots, 5i}$  are five dummy variables with:

- $(ar_{1i}, ar_{2i}, ar_{3i}, ar_{4i}, ar_{5i}) = (1, 0, 0, 0, 0)$  if the  $i$ th farm comes from Greek

Mainland

= (0, 1, 0, 0, 0) from Epirus

= (0, 0, 1, 0, 0) from Thessaly

= (0, 0, 0, 1, 0) from Macedonia

= (0, 0, 0, 0, 1) from Crete

= (0, 0, 0, 0, 0) from Peloponnese (reference category)



Furthermore,  $auje_i, taf_i, dis_i, pr_i, fl_i$  and  $vet_i$  are dummy variables with:

- $auje_i = 1$  if the  $i$ th farm vaccinates the sows during their pregnancy for the Aujeszky virus  
 $= 0$  otherwise
- $taf_i = 1$  when a trench with a disinfectant material exists in the  $i$ th farm  
 $= 0$  otherwise
- $dis_i = 1$  if there are other piggeries near the  $i$ th farm in a distance of more than 3 km  
 $= 0$  if other piggeries are present in a distance of less, or equal, than 3km
- $pr_i = 1$  if there is a medical history in the  $i$ th farm regarding PRRS  
 $= 0$  otherwise
- $fl_i = 1$  if the  $i$ th farm protects its piglets from the flies  
 $= 0$  otherwise
- $vet_i = 1$  if veterinarians are occupied in the  $i$ th farm on a permanent basis  
 $= 0$  if their occupation is limited to some organized visits

Finally,  $mik_{1,2i}$  is a categorical variable too, with:

- $(mik_{1i}, mik_{2i}) = (1, 0)$  if the  $i$ th farm checks the piglet's feedings for toxins on not regular periods  
 $= (0, 1)$  if there is a check on regular periods  
 $= (0, 0)$  whether there is no check at all or the check is rare

Results of fitting model (5.1) are shown in Tables 5.2 and 5.3. Specifically, Table 5.2 gives the estimates of all the explanatory effects with their standard errors and the Wald Chi-Square statistics,  $(\hat{\beta}/(s\hat{e}(\hat{\beta})))^2$ , with the corresponding p-values. We observe that the p-values are quite large, meaning that none of the variables is statistically significant for the mortality rate. The odds ratios of the explanatory variables are calculated as shown in the last



Column of Table 5.2. They are obtained by simply exponentiating the estimates of the parameters.

Parameter	DF	Estimate	Std. Error	Chi-Square	p-value	Odds Ratio
Intercept	1	-3.3263	0.5702	34.0315	0.0001	-
Sowsnumb	1	0.000018	0.00029	0.0039	0.9504	1.000
Areapro	1	-0.0523	1.3760	0.0014	0.9697	0.949
Area1	1	-0.2418	0.3272	0.5459	0.4600	0.785
Area2	1	0.1243	0.3438	0.1307	0.7177	1.132
Area3	1	0.1081	0.2464	0.1925	0.6609	1.114
Area4	1	-0.0999	0.2547	0.1539	0.6948	0.905
Area5	1	-0.4097	0.6820	0.3610	0.5479	0.664
Airvolum	1	-0.0215	0.2538	0.0072	0.9324	0.979
Workers	1	0.0181	0.0167	1.1770	0.2780	1.018
Auje	1	-0.2728	0.2269	1.4460	0.2292	0.761
Tafros	1	-0.0593	0.2199	0.0727	0.7874	0.942
Distan	1	0.1519	0.3969	0.1465	0.7019	1.164
Prrs	1	-0.2736	0.5101	0.2878	0.5916	0.761
Mikotox1	1	-0.0756	0.2125	0.1265	0.7221	0.927
Mikotox2	1	-0.0408	0.2521	0.0261	0.8716	0.960
Flies	1	-0.1737	0.4292	0.1637	0.6858	0.841
Vet	1	-0.1878	0.3316	0.3208	0.5711	0.829

**Table 5.2:** Analysis of Maximum Likelihood Estimates for model (5.1)

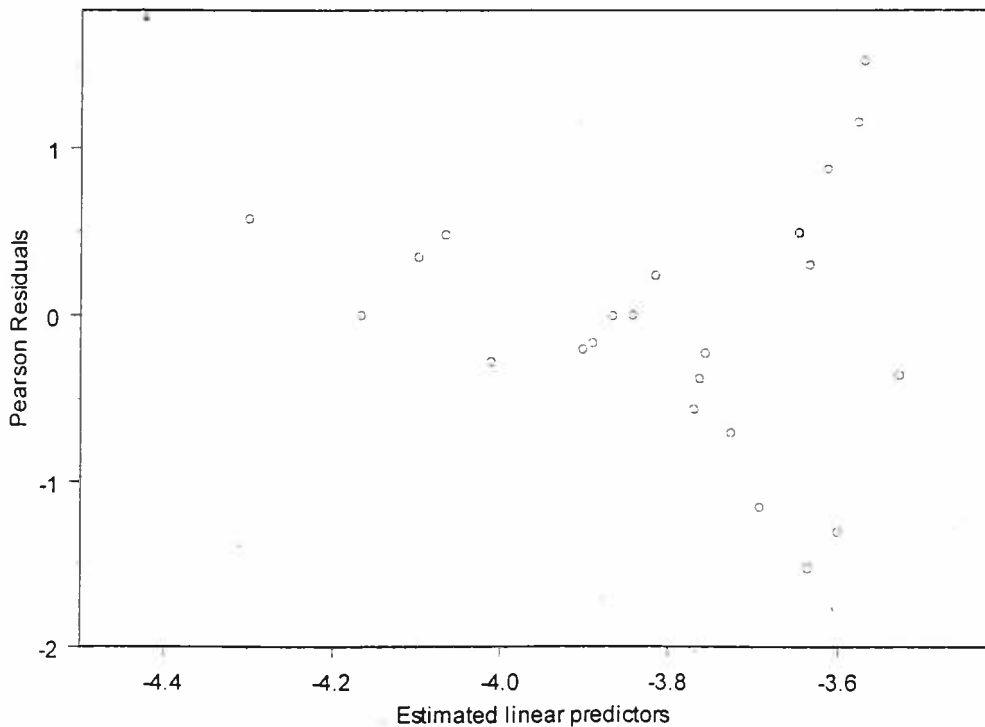
Moreover, in Table 5.3 we observe that both Pearson  $\chi^2$  and deviance are significant (p-value=0.033 and 0.0283, respectively from a Wald test), suggesting that the model does not fit well. Particularly, fitting the model gives a Pearson  $\chi^2$  and deviance of 12.1319 and of 12.5183, respectively, on 5 degrees of freedom. As both the Pearson  $\chi^2$  and the residual deviance show substantial overdispersion, we suspect that the lack of fit may be due to overdispersion. In addition, model (5.1) has a value of 5334.525 for the  $-2 \text{Log Likelihood}$  statistic.



Next, in Figure 5.1 we plot the Pearson Residuals against the linear predictors  $\hat{\eta}_i = g(\hat{p}_i) = \ln(\hat{p}_i/(1 - \hat{p}_i))$ . Figure 5.1 suggests that the logistic regression model (5.1) is not true. Even though, there are not any outliers as all the residuals fall within  $\pm 2$ , we do not observe a horizontal band, as we would expect if model (5.1) was true. Figure 5.1 shows non-constant variance as we move from left to right, meaning that the assumed form of the variance function, here  $V(\mu_i) = \mu_i(1 - \mu_i) = p_i(1 - p_i)$ , is inappropriate and thus the variance of the response variable  $V(p_i) = p_i(1 - p_i)/n_i$  is inappropriate too. In the following, we hope to get more satisfactory results by attributing the lack of fit to overdispersion.

Criterion	DF	Value	Value/DF	p-value
Deviance	5	12.5183	2.5037	0.0283
Pearson	5	12.1319	2.4264	0.0330

**Table 5.3:** Deviance and Pearson goodness-of-fit statistics for model (5.1)



**Figure 5.1:** Pearson Residuals versus estimated linear predictors



## 5.5 Fitting the Beta-Binomial model (Williams' method) -Tarone's test

### 5.5.1 Results from estimating the Beta-Binomial model

As we rejected the logistic model (5.1) in the previous section, we go on fitting a Beta-Binomial model for our data. We underline that model (5.1) assumes independence among individual-piglets, and therefore  $R_i | p_i \sim \text{Binomial}(n_i, p_i)$ . This assumption is not quite rational, since we expect that the correlation between piglets from the same upper-level unit, farm, is higher than that from different farms. This is quite reasonable, seeing that piglets from the same farm share the same piggery environment, or have the same medical care, etc. In order to take into account the extra-binomial variation that is introduced due to either the variation among the upper-level farms or the intra-class correlation inside farms and sows, we consider that the response variable is beta-binomial distributed. We therefore assume that the parameters  $p_i$  of the Binomial distribution are no longer constant, but are random variables with  $p_i \sim \text{Beta}(a_i, b_i)$ , where  $a_i > 0$ ,  $b_i > 0$ ,  $i = 1, \dots, 23$ . According to the theory developed in Section 3.2.2 of Chapter 3, we fit the following logistic regression model for the response variable  $p_i = R_i/n_i$ :

- $$\eta_i = g(\pi_i) = \ln(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1(X_{1i} - \bar{X}_1) + \beta_2(X_{2i} - \bar{X}_2) + \beta_3ar_{1i} + \beta_4ar_{2i} + \beta_5ar_{3i} + \beta_6ar_{4i} + \beta_7ar_{5i} + \beta_8(X_{3i} - \bar{X}_3) + \beta_9(X_{4i} - \bar{X}_4) + \beta_{10}auje_i + \beta_{11}taf_i + \beta_{12}dis_i + \beta_{13}pr_i + \beta_{14}mik_{1i} + \beta_{15}mik_{2i} + \beta_{16}fl_i + \beta_{17}vet_i \quad (5.2)$$

where

$$E(p_i) = \pi_i = a_i/(a_i + b_i)_i$$

$$\text{var}(p_i) = [\pi_i(1-\pi_i)]/n_i + [(n_i - 1)/n_i]\rho\pi_i(1-\pi_i)$$

$$\rho_i = 1/(a_i + b_i + 1), \quad 0 < \rho_i < 1$$

We remind that the mortality rate,  $p_i = R_i/n_i$ , follows a Beta-Binomial distribution with mean and variance as given above. Furthermore, estimates of the regression parameters,  $\beta_0$  to  $\beta_{17}$ , are obtained by iterative reweighted least squares and an estimate for  $\rho$  is obtained by equating the Pearson



chi-square statistic  $X^2 = \sum (p_i - \hat{\pi}_i)^2 / \text{var}(p_i)$  to its expected value. Particularly,  $\rho$  plays the role of the common positive correlation, which is observed between individual-piglets within a farm.

We use the SAS procedure LOGISTIC (detailed in Appendix B1), in order to deal with overdispersion. Even if PROC LOGISTIC has three SCALE=options, we prefer the SCALE=WILLIAMS option, since the number of piglets,  $n_i$ , total born in each piggery, differs between the twenty-three farms. Using Williams' method, the standard errors of the parameter estimates are adjusted for overdispersion, and consequently their significance tests. As we have already mentioned in Section 4.3, the theory discussed in Section 3.2.2 is similar to the manual's theory for the SAS procedure LOGISTIC, which is referred as the Williams' method. Results using Williams' method for estimating model (5.2) are shown in Tables 5.4, 5.5 and 5.6. The estimated intra-farm correlation coefficient is  $\hat{\rho}_1=0.001636$ , meaning that the positive correlation between pairs of piglets within the same farm is equal to 0.001636.

Criterion	Intercept Only	Intercept and Covariates
AIC	1800.142	1828.437
SC	1808.272	1974.774
-2 LOG L	1798.142	1792.437

**Table 5.4:** Model Fitting Information for model (5.2)

Table 5.4 shows the three model fitting criteria, *AIC*, *SC* and  $-2LOGL$ , displayed by the LOGISTIC procedure. The criteria are calculated as follows:

- $-2LOGL = -2 \sum_{i=1}^{23} d_i n_i (r_i \ln(\hat{p}_i) + (n_i - r_i) \ln(1 - \hat{p}_i))$

where  $n_i$  is the number of piglets total born in the *ith* farm,  $r_i$  is the number of the dead piglets,  $\hat{p}_i$  is the estimated probability that corresponds to a positive response and  $d_i = \{1 + \rho(n_i - 1)\}^{-1}$  is the weight value for the *ith* observation.



• **Akaike Information Criterion (AIC):**

$$AIC = -2LOGL + 2(k + s)$$

where  $k$  is the total number of response levels minus one and  $s$  is the number of explanatory effects. Here, we have  $k = 1$  and  $s = 17$ .

• **Schwarz Criterion (SC):**

$$SC = -2LOGL + (k + s) \ln \left( \sum_i n_i \right)$$

where  $k, s$  and  $n_i$  are as defined previously.

Particularly, Table 5.4 gives the three model fitting criteria for the model with the intercept only and the model, which includes the intercept and all the covariates. Actually, the statistics  $AIC$  and  $SC$  are used when comparing different models for the same data. Lower values of the statistics indicate a more desirable model. The  $-2LOGL$  statistic is equal to 1792.437 for model (5.2).

Test	Chi-Square	DF	p-value
Likelihood Ratio	5.704	17	0.995

**Table 5.5:** Testing Global Null Hypothesis: BETA=0 for model (5.2)

Table 5.5 gives the results of testing the null hypothesis:

- $H_0$ : the model with the intercept only fits adequately the data against the alternative
- $H_1$ : the model with the intercept and all the explanatory effects provides a better fit to the data than the model under  $H_0$

The likelihood ratio test statistic has a chi-square distribution with 17 degrees of freedom under the null hypothesis that all the explanatory variables in the model are zero. Its value is 5.704 and the procedure produces a p-value, which is equal to 0.995, suggesting that the model with the intercept only is more appropriate for the available data, or that we do not reject the global null hypothesis  $H_0: \beta_1 = \dots = \beta_{17} = 0$  at 5% significance level.



Furthermore, in order to test whether the intra-farm correlation coefficient  $\rho_1$  is statistically significant or not, we compare models (5.1) and (5.2) by using a generalized likelihood ratio test. The hypothesis being tested is:

- $H_0: \rho_1 = 0$  against the alternative  $H_1: \rho_1 \neq 0$

or equivalently

- $H_0$ : models (5.1) and (5.2) provide the same fit to the data  
against

$H_1$ : model (5.2) provides a better fit to the data than model (5.1)

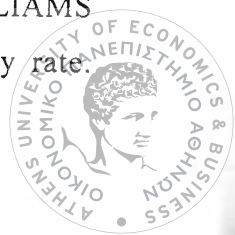
We note that model (5.1) is nested within model (5.2), since the explanatory variables under the null hypothesis are a subset of those under the alternative hypothesis. Specifically, the only extra term that is considered under model (5.2) is the intra-farm correlation coefficient  $\rho_1$ . The log-likelihood ratio test statistic for comparing the two models is calculated as follows:

- $L_{01} = L_1 - L_0 = -1792.437 + 5334.525 = 3542.088$

where  $L_0, L_1$  is the  $2LOGL$  statistic for model (5.1) and (5.2) respectively.

Under  $H_0$ ,  $L_{01}$  has an asymptotic chi-squared distribution with degrees of freedom (df) equal to the difference in df of the two models. Here, model (5.1) has 18 df and model (5.2) 19 df. Since  $L_{01} = 3542.088 > \chi_{1,0.95}^2$ , we reject  $H_0: \rho_1 = 0$  in favour of  $H_1: \rho_1 \neq 0$  and the intra-farm correlation coefficient  $\rho_1$  is statistically significant at 5% significance level. We keep this result for now.

Unfortunately, results of the 'Analysis of Maximum Likelihood Estimates' Table 5.6 are not satisfactory at all. The estimated standard errors are dramatically large. This of course affects the significance test of the explanatory effects by producing p-values that are very large. The smallest p-value (0.3995) is observed for the explanatory variable *Auje* and the largest p-value (0.9754) for *Airvolum*, and therefore all the variables are not statistically significant for the mortality rate. Even though we used the Forward, Backward and Stepwise methods, keeping the SCALE=WILLIAMS option, we did not find variables that have an effect to the mortality rate.



Particularly, the common result of the three effect selection processes was a model, which includes only the intercept term and none of the explanatory effects.

Parameter	DF	Estimate	Std. Error	Chi-Square	p-value	Odds Ratio
Intercept	1	-3.3969	0.8743	15.0936	0.0001	-
Sowsnumb	1	0.000058	0.00046	0.0160	0.8992	1.000
Areapro	1	-0.3591	2.1325	0.0284	0.8663	0.698
Area1	1	-0.1614	0.5165	0.0977	0.7546	0.851
Area2	1	0.1816	0.5278	0.1184	0.7308	1.199
Area3	1	0.1363	0.3779	0.1301	0.7183	1.146
Area4	1	-0.0500	0.3966	0.0159	0.8997	0.951
Area5	1	-0.3279	1.0583	0.0960	0.7567	0.720
Airvolum	1	0.0126	0.4074	0.0010	0.9754	1.013
Workers	1	0.0146	0.0260	0.3141	0.5752	1.015
Auje	1	-0.2913	0.3457	0.7099	0.3995	0.747
Tafros	1	-0.0226	0.3329	0.0046	0.9458	0.978
Distan	1	0.0523	0.5870	0.0079	0.9290	1.054
Prrs	1	-0.2695	0.7835	0.1184	0.7308	0.764
Mikotox1	1	-0.0867	0.3246	0.0713	0.7894	0.917
Mikotox2	1	-0.0726	0.4042	0.0322	0.8575	0.930
Flies	1	-0.0436	0.6311	0.0048	0.9450	0.957
Vet	1	-0.2239	0.5213	0.1844	0.6676	0.799

**Table 5.6:** Analysis of Maximum Likelihood Estimates for model (5.2)

### 5.5.2 Tarone’s one-sided test of the Binomial distribution versus Beta-Binomial alternatives.

In this section we perform Tarone’s standardized one-sided test of the Binomial distribution versus Beta-Binomial alternatives, which has already discussed in Section 3.3. In particular, we test the null hypothesis  $H_0 : \rho_1 = 0$  against the alternative  $H_1 : \rho_1 > 0$ , where  $\rho_1$  represents the intra-piggery correlation coefficient. Next, we compute:



$$\hat{p} = \sum_{i=1}^{23} R_i / \sum_{i=1}^{23} n_i = 558/25080 = 0.0222$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.02224 = 0.9778$$

$$S = \sum_{i=1}^{23} (R_i - n_i \hat{p})^2 / (\hat{p}\hat{q}) = 700.5653/0.0217 = 32284.1152$$

where  $n_i, R_i$  denote the number of piglets total born and the number of dead piglets in the  $i$ th piggery ( $i = 1, \dots, 23$ ), respectively. The Tarone's  $Z$  statistic is given by:

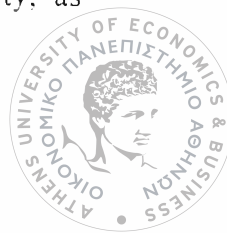
$$Z = \left( S - \sum_{i=1}^{23} n_i \right) / \left\{ 2 \sum_{i=1}^{23} n_i (n_i - 1) \right\}^{\frac{1}{2}} = (32284.1152 - 25080) / \left\{ 2 * 3.7 * 10^7 \right\}^{\frac{1}{2}} = 0.837$$

We reject the null hypothesis  $H_0$  when  $Z > Z_{1-\alpha}$ . Here we have  $Z = 0.837 < Z_{1-\alpha} = Z_{0.95} = 1.645$ , and therefore we do not reject  $H_0$  at 5% significance level and the intra-piggery correlation coefficient  $\rho_1$  is not statistically significant. At this point we note that while Tarone's one-sided test produces a non-statistically significant correlation coefficient  $\rho_1$ , the log-likelihood ratio test as performed in the previous section gave a statistically significant correlation. This is quite reasonable since Tarone's test is a more restricted test (one-sided) than the log-likelihood ratio test (two-sided). However, in our case it is Tarone's test that it is in matter and therefore we conclude that the intra-piggery correlation coefficient  $\rho_1$  is not statistically significant. Finally, we mention that many authors suggest not using the log-likelihood ratio test since the regularity conditions do not hold.

## 5.6 Fitting the logistic random-intercept model

### 5.6.1 The Penalized Quasi-Likelihood approach

Except of fitting a Beta-Binomial model, another way to handle the problem of overdispersion is to add unobserved random effects to the linear predictor, as discussed in Chapter 2. Specifically, we perform the penalized quasi-likelihood (PQL) method with the scale parameter fixed at unity, as



developed in Section 2.3. We next define the logistic random-intercept model for our data in order to account for the piggery-to-piggery variability:

$$\begin{aligned} \bullet \eta_i = g(p_i) = \ln(p_i/(1-p_i)) = & \beta_0 + \beta_1(X_{1i} - \bar{X}_1) + \beta_2(X_{2i} - \bar{X}_2) + \beta_3ar_{1i} + \beta_4ar_{2i} + \\ & \beta_5ar_{3i} + \beta_6ar_{4i} + \beta_7ar_{5i} + \beta_8(X_{3i} - \bar{X}_3) + \beta_9(X_{4i} - \bar{X}_4) + \beta_{10}auje_i + \beta_{11}taf_i + \\ & \beta_{12}dis_i + \beta_{13}pr_i + \beta_{14}mik_{1i} + \beta_{15}mik_{2i} + \beta_{16}fl_i + \beta_{17}vet_i + b_i \end{aligned} \quad (5.3)$$

where  $p_i = R_i/n_i$  ( $i = 1, \dots, 23$ ) and  $p_i | b_i \sim \text{Binomial}(n_i, p_i)/n_i$ . The notation  $b_i$  corresponds to the random effects associated with each piggery. We assume that the random effects  $b_i$  are independently normally distributed  $b_i \sim N(0, \sigma_b^2)$ . Using the SAS macro GLIMMIX (see in Appendix B.2 for the syntax of the macro), we apply the PQL analysis for estimating model (5.3). The results from fitting model (5.3) are presented in Tables 5.7 and 5.8.

Once again, the ‘Parameter Estimates’ Table 5.7 does not give satisfactory results, since all the covariates have large p-values (from the Wald statistic). Using again the Forward, Backward and Stepwise effect selection processes, none of the variables was found to be statistically significant for the mortality rate. In Table 5.8, the estimates  $\hat{b}_i$  of the random effects are produced. These estimates are the empirical best linear unbiased predictors (EBLUPs), which are given by Equation (2.18) of Section 2.3.1. The SE Pred column of Table 5.8, gives the standard errors of predictions  $\hat{b}_i - b_i$ , where  $\hat{b}_i$  is the  $i$ th EBLUP and  $b_i$  is the  $i$ th random-effect parameter. We note that in the first column of Table 5.8, Intercept is specified as the random effect.

With the SAS macro GLIMMIX, we find that the estimated common variance of the random effects is  $\hat{\sigma}_b^2 = 0.0751$ . Even though the output of the SAS macro GLIMMIX does not produce the estimated intra-piggery correlation coefficient, we can compute it as follows:

$$\bullet \hat{\rho}_2 = \hat{\sigma}_b^2 / (\hat{\sigma}_b^2 + \hat{\sigma}^2) = 0.0751 / (0.0751 + 1) = 0.0698$$

where  $\hat{\sigma}^2 = 1$  is the scale parameter or the residual variance, which we fixed at unity in order to request the estimation of the PQL model of Breslow and



Clayton (1993). We observe that the estimate  $\hat{\rho}_2 = 0.0698$  differs from  $\hat{\rho}_1 = 0.001636$ , which was previously obtained using Williams' method. However, since all the explanatory effects have large p-values we conclude that model (5.3) is not an acceptable fit to the data.

Parameter	DF	Estimate	Std. Error	Chi-Square	p-value	Odds Ratio
Intercept	1	-3.47201	0.8365	-	-	-
Sowsnumb	1	0.00004	0.0004	0.01	0.926	1
Areapro	1	-0.46433	2.1222	0.05	0.826	0.628
Area1	1	-0.15104	0.5079	0.09	0.766	0.859
Area2	1	0.17723	0.5215	0.12	0.734	1.193
Area3	1	0.12679	0.3765	0.11	0.736	1.135
Area4	1	-0.02602	0.3986	0.00	0.948	0.974
Area5	1	-0.23779	1.0268	0.05	0.816	0.788
Airvolum	1	0.00097	0.4064	0.00	0.998	1
Workers	1	0.01290	0.0247	0.27	0.602	1.012
Auje	1	-0.27769	0.3301	0.71	0.400	0.757
Tafros	1	-0.00710	0.3309	0.00	0.982	0.992
Distan	1	0.01840	0.5848	0.00	0.974	1.018
Prrs	1	-0.23071	0.7499	0.09	0.758	0.793
Mikotox1	1	-0.06783	0.3331	0.04	0.838	0.934
Mikotox2	1	-0.07670	0.4271	0.03	0.857	0.926
Flies	1	-0.00209	0.6115	0.00	0.997	0.997
Vet	1	-0.20855	0.5017	0.17	0.677	0.811

Table 5.7: Parameter Estimates for model (5.3)



Effect	Farms	Estimate	SE Pred
Intercept	1	-0.0510	0.2705
Intercept	2	0.0562	0.2591
Intercept	3	-0.0033	0.2531
Intercept	4	0.2197	0.2229
Intercept	5	-0.1686	0.2486
Intercept	6	0.1332	0.2546
Intercept	7	-0.1862	0.2411
Intercept	8	-0.0401	0.2558
Intercept	9	0.0401	0.2558
Intercept	10	-0.0364	0.2557
Intercept	11	0.0511	0.2614
Intercept	12	0.0400	0.2642
Intercept	13	-0.0547	0.2510
Intercept	14	0.1829	0.2555
Intercept	15	-0.0610	0.2421
Intercept	16	-0.0098	0.2621
Intercept	17	-0.1218	0.2649
Intercept	18	-0.0322	0.2650
Intercept	19	0.0419	0.2438
Intercept	20	0.0000	0.2741
Intercept	21	-0.0897	0.2561
Intercept	22	0.0897	0.2561
Intercept	23	-0.0000	0.2741

**Table 5.8:** Random Effects Estimates for model (5.3)

### 5.6.2 The Gauss-Hermite technique

Next, using MIXOR we apply the Gauss-Hermite technique for estimating model (5.3). Unfortunately, the program ‘blows up’ and does not produce any results at all. In such a case, as Hedeker and Gibbons mention in



the manual of MIXOR, the model that is specified may not be estimable. They then suggest that we should try to fit a less complicated model by specifying fewer random effects, or fewer explanatory effects, or collapsing some of the ordered categories of the response variable if these are very sparse. In our case, the model being fitted has only one random effect and the response of interest is binary; whether the individual-piglet has been dead or alive at the end of the experiment. However, even though we reduced the number of covariates from twenty-one to one, or two, the program still could not run. In addition, according to Hedeker and Gibbons, if with one random effect problems still exist, it may be that the estimated variance of the random effect is zero, and thus a model with no random effects may be more appropriate. We note that considering a model without random effects leads to the standard generalized linear model (5.1). Vuataz and Sotek (1978) and Crowder (1978) discuss about such cases where the estimated variance of the random effect cannot be estimated as being different from zero and consequently the iterative maximization process fails to converge.

### 5.7 Returning to the standard logistic regression model

Since we have rejected models (5.2) and (5.3) in Sections 5.5 and 5.6 respectively, we now turn back to the logistic regression model (5.1). Using the Forward, Backward and Stepwise effect selection processes for estimating model (5.1), two variables (Auje and Vet) were found to be statistically significant for the outcome, the mortality rate. In particular, the common result of the three above methods was the following model:

$$\eta_i = g(p_i) = \text{logit}(p_i) = \ln(p_i/(1-p_i)) = \beta_0 + \beta_1 \text{auje}_i + \beta_2 \text{vet}_i \quad (5.4)$$

Fitting model (5.4) gives results that are summarized in Tables 5.9, 5.10 and 5.11. In Table 5.9, we observe that model (5.4) gives a Pearson  $\chi^2$  and deviance of 18.854 and of 19.0716 respectively, on 20 degrees of freedom. The corresponding p-values from a Wald test for the Pearson  $\chi^2$  and deviance are 0.5313 and 0.5172 respectively, suggesting that model (5.4) is an acceptable fit to the data. Furthermore, the  $-2 \text{Log Likelihood}$  statistic is equal to  $-2 \text{LOGL} = 5341.078$  for model (5.4).



Criterion	DF	Value	Value/DF	p-value
Deviance	20	19.0716	0.9536	0.5172
Pearson	20	18.8540	0.9427	0.5313

**Table 5.9:** Deviance and Pearson goodness-of-fit statistics for model (5.4)

Test	Chi-Square	DF	p-value
Likelihood Ratio	9.315	2	0.0095

**Table 5.10:** Testing Global Null Hypothesis: BETA=0 for model (5.4)

In Table 5.10 we test the null hypothesis:

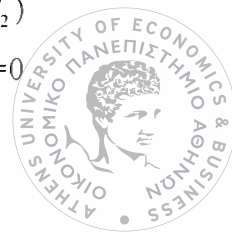
- $H_0: \beta_1 = \beta_2 = 0$  against the alternative  $H_1: \beta_i \neq 0$ , for at least one  $i = 1, 2$

The likelihood ratio test statistic has a chi-square distribution with 2 degrees of freedom under  $H_0$ . The statistic has a value of 9.315 and the corresponding p-value from a Wald Chi-square statistic is equal to 0.0095, meaning that we reject the null hypothesis  $H_0: \beta_1 = \beta_2 = 0$  at 5% significance level and at least one of the explanatory effects Auje and Vet is statistically significant.

Parameter	DF	Estimate	Std. Error	Chi-Square	p-value	Odds Ratio
Intercept	1	-3.6524	0.0589	3839.7030	0.0001	
Auje	1	-0.2157	0.1021	4.4619	0.0347	0.806
Vet	1	-0.2249	0.0927	5.8938	0.0152	0.799

**Table 5.11:** Analysis of Maximum Likelihood Estimates for model (5.4)

In Table 5.11 we give the maximum likelihood estimates from fitting model (5.4). Particularly, the regression coefficients  $\beta_1$  and  $\beta_2$  for the explanatory effects Auje and Vet, respectively, are estimated as  $\hat{\beta}_1 = -0.2157$  and  $\hat{\beta}_2 = -0.2249$ . Both  $\beta_1$  and  $\beta_2$  are statistically significant with p-values equal to 0.0347 and 0.0152, respectively. The corresponding estimated odds ratios are  $\psi_1 = e^{\hat{\beta}_1} = 0.806$  and  $\psi_2 = e^{\hat{\beta}_2} = 0.799$ . The parameter  $\beta_1$  ( $\beta_2$ ) associated with Auje (Vet) represents the change in the log-odds from Auje=0



to  $Auje=1$  (Vet=0 to Vet=1). So, the odds ratio ( $P(event | X)/P(nonevent | X)$ ) indicates how the odds of event (here a piglet’s death) change as we change  $Auje$  or  $Vet$  from 0 to 1. Consequently:

- $\psi_1 = e^{\hat{\beta}_1} = 0.806 \Rightarrow$  the odds of a piglet’s death are reduced by  $(1-0.806)=19.4\%$  as we change  $Auje$  from 0 to 1  $\Rightarrow$  the odds of a piglet’s death are reduced by 19.4% when the sows are vaccinated during their pregnancy for the Aujeszky virus ( $Auje=1$ ).
- $\psi_2 = e^{\hat{\beta}_2} = 0.799 \Rightarrow$  the odds of a piglet’s death are reduced by  $(1-0.799)=20.1\%$  as we change  $Vet$  from 0 to 1  $\Rightarrow$  the odds of a piglet’s death are reduced by 20.1% when the occupation of the veterinary staff is on a permanent basis ( $Vet=1$ ).

Wald Confidence Limits			
Variable	Odds Ratio	Lower	Upper
Auje	0.806	0.660	0.985
Vet	0.799	0.666	0.958

**Table 5.12:** 95% Confidence Intervals for the Odds Ratio

Table 5.12 contains the Odds Ratio estimates and the corresponding 95% Wald confidence intervals. From Table 5.12 we conclude that the odds ratio for both variables  $Auje$  and  $Vet$  are statistically significant since the two confidence intervals do not include the value of zero.

Eventually, in Figure 5.2 we plot the Deviance Residuals versus the Fitted values,  $g(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 auje_i + \hat{\beta}_2 vet_i$ , for model (5.4). There is an abnormality in the graph due to three points. Figure 5.2 shows that these points correspond to the 4<sup>th</sup>, 7<sup>th</sup> and 14<sup>th</sup> pig farm. In addition, all the residuals fall within  $\pm 2$ .



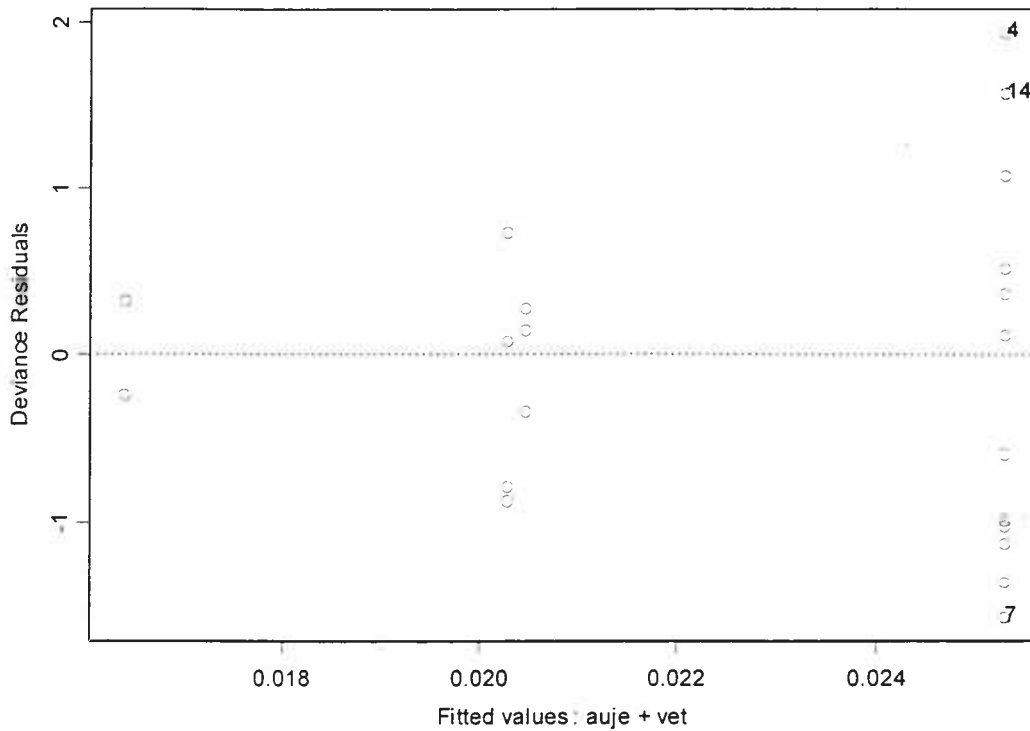


Figure 5.2: Deviance Residuals versus Fitted values

Summarizing the above results, we conclude that the data do not exhibit significant extra-binomial variation, and therefore we turned back to the standard logistic regression model (5.1). In order to estimate model (5.1), we used three effect selection processes. The common result of these processes was that the only explanatory variables that have an effect to the mortality rate of the piglets are Auje and Vet.





## Appendix A

### The exponential family and a review of the GLM theory

#### A1. The Exponential family

A probability distribution is said to be a member of the exponential family if its probability density function (or probability mass function, if discrete) can be written in the form:

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \quad (1)$$

The parameter  $\theta$  is called the natural or canonical parameter. The parameter  $\phi$  is usually assumed known. If it is unknown then it is often called the nuisance parameter. The density function (1) can be thought of as a likelihood resulting from a single observation  $y$ . Then,

$$\log f(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

The mean and variance of a random variable with probability density function (or probability function) of the form (1), are given by (see McCullagh and Nelder, 1989; sec. 2.2.2, p. 29):

$$E(Y) = \mu = b'(\theta)$$

$$\text{Var}(Y) = a(\phi)b''(\theta)$$

The variance is the product of two functions;  $b''(\theta)$  depends only on the canonical parameter  $\theta$  (and consequently on  $\mu$ , as  $\theta = b^{-1}(\mu)$ ) and is called the variance function ( $V(\mu) \equiv b''(\theta)$ );  $a(\phi)$  is sometimes of the form  $a(\phi) = \sigma^2/w$ , where  $w$  is a known weight and  $\sigma^2$  is called the dispersion parameter or scale parameter.

## A2. A brief review of the GLM theory

In practical applications, the aim is to determine the pattern of dependence of the response variable on a group of explanatory variables. We denote the  $n$  observations of the response by  $y = (y_1, y_2, \dots, y_n)'$ . In a generalized linear model (GLM) these are assumed to be observations of independent random variables  $Y = (Y_1, Y_2, \dots, Y_n)'$ , which have the same distribution from the exponential family (1). Meaning that, the functions  $a, b, c$  in Equation (1) and usually the scale parameter  $\phi$  are the same for all observations. In contrast, the canonical parameter  $\theta$  may differ between observations. Therefore, we write:

$$f_{Y_i}(y_i; \theta_i, \phi_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right), \quad i = 1, \dots, n$$

The joint density for  $Y = (Y_1, Y_2, \dots, Y_n)'$  is:

$$f_Y(y; \theta, \phi) = \prod_{i=1}^n f_{Y_i}(y_i; \theta_i, \phi_i) = \exp\left(\sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^n c(y_i, \phi_i)\right) \quad (2)$$

where  $\theta = (\theta_1, \theta_2, \dots, \theta_n)'$  is the vector of the canonical parameters and  $\phi = (\phi_1, \phi_2, \dots, \phi_n)'$  is the vector of the nuisance parameters, when they exist.

In a GLM, the distribution of the response variable  $Y_i$  depends on  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$  through the linear predictor  $\eta_i$ , where  $\eta_i = x_i' \beta$ ,  $x_i$  is the vector of the  $p$  explanatory variables and  $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$  is the vector of fixed but unknown parameters describing the dependence of  $Y_i$  on  $x_i$ . Particularly:

$$g(E[Y_i]) = g(\mu_i) = \eta_i = x_i' \beta, \quad i = 1, \dots, n$$

or in a matrix form

$$g(E[Y]) = g(\mu) = \eta = X\beta$$

where  $\mu = (\mu_1, \mu_2, \dots, \mu_n)'$ ,  $\eta = (\eta_1, \eta_2, \dots, \eta_n)'$ ,  $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$  and  $X' = (x_1', x_2', \dots, x_n')$  is the  $n \times p$  design matrix with  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ ,  $i = 1, \dots, n$ .



Furthermore, we remind that it is the parameters  $\beta_1, \beta_2, \dots, \beta_p$  of the linear predictor, which are of direct interest, and not the canonical parameters  $\theta_1, \theta_2, \dots, \theta_n$ . In order to find the maximum likelihood estimates (M.L.E.) of  $\beta_1, \beta_2, \dots, \beta_p$ , we maximize the log-likelihood function, which from (2), can be written:

$$\log f_Y(y; \beta, \phi) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^n c(y_i, \phi_i) \quad (3)$$

In Equation (3), we express the joint density (2) in terms of the coefficients  $\beta$ . This is true, since  $E(Y_i) = \mu_i = b'(\theta_i)$ ,  $g(\mu_i) = \eta_i = x_i' \beta$ , and therefore we can write:

$$\theta_i = b^{-1}(\mu_i) = b^{-1}(g^{-1}[x_i' \beta])$$

To find  $\hat{\beta}$ , we consider the scores:

$$u_k(\beta) = \frac{\partial}{\partial \beta_k} \log f_Y(y; \beta, \phi) = \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi) \text{Var}(Y_i)} \frac{x_k}{g'(\mu_i)}, \quad k = 1, \dots, p \quad (4)$$

and then

- $u_k(\hat{\beta}) = 0, \quad \forall k = 1, \dots, p \Rightarrow$

$$u(\hat{\beta}) = 0$$

where  $u = (u_1, u_2, \dots, u_p)'$ . In practice, we solve the  $p$  simultaneous equations  $u_k(\hat{\beta}) = 0, \quad \forall k = 1, \dots, p$ , to evaluate  $\hat{\beta}$ . These equations have usually no analytic solution, and therefore we rely on numerical methods to solve them.

The Hessian and Fisher information matrices ( $H(\beta)$  and  $I(\beta)$ , respectively) are matrices with components:

$$\{H(\beta)\}_{jk} = \frac{\partial}{\partial \beta_j} u_k(\beta) = \frac{\partial^2}{\partial \beta_j \partial \beta_k} \log f_Y(y; \beta, \phi) \quad k, j = 1, \dots, p$$

$$\{I(\beta)\}_{jk} = E(-\{H(\beta)\}_{jk}) \quad k, j = 1, \dots, p$$

It holds that (see McCullagh and Nelder, 1989; sec. 4.4.2 and 2.5.1):

$$I(\beta) = X'WX$$

and

$$u(\beta) = X'Wz$$



where  $X$  is the design matrix,  $W$  is a  $n \times n$  diagonal matrix with diagonal elements  $w_i = [Var(Y_i)\{g'(\mu_i)\}^2]^{-1}$ ,  $i = 1, \dots, n$  and  $z = (z_1, z_2, \dots, z_n)'$  is a  $n \times 1$  vector with components  $z_i = (y_i - \mu_i)g'(\mu_i)$ ,  $i = 1, \dots, n$ .

Next, we describe the Fisher-Scoring algorithm in order to solve the  $p$  simultaneous equations  $u(\hat{\beta}) = 0$  that give  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)'$ .

1. Choose an initial estimate  $\beta^{(l)}$  for  $\hat{\beta}$  at  $l = 0$ .

2. Evaluate:

$$g(\mu^{(l)}) = \eta^{(l)} = X\beta^{(l)} \Rightarrow \mu^{(l)} = g^{-1}(X\beta^{(l)}),$$

$$W^{(l)} = \text{diag}(w^{(l)}), \text{ where}$$

$$w_i^{(l)} = [Var(Y_i)\{g'(\mu_i^{(l)})\}^2]^{-1} = [a(\phi_i)V(\mu_i^{(l)})\{g'(\mu_i^{(l)})\}^2]^{-1}, i = 1, \dots, n$$

$$z^{(l)} = (y - \mu^{(l)})g'(\mu^{(l)})$$

3. Calculate  $\beta^{(l+1)} = [X'W^{(l)}X]^{-1} X'W^{(l)}[z^{(l)} - \eta^{(l)}]$

4. If  $\|\beta^{(l+1)} - \beta^{(l)}\| >$  some prespecified (small) tolerance then set  $l \rightarrow l + 1$  and go to step 2.

5. Use  $\beta^{(l+1)}$  as the solution for  $\hat{\beta}$ .

This algorithm is sometimes known as Iteratively Reweighted Least Squares (IRLS). McCullagh and Nelder (1989) give further details for the generalized linear models.



## Appendix B

### How to use SAS for fitting the overdispersed GLM

We have already stated that a crucial point in standard logistic regression analysis is that observations are independent of one another and that if the assumption of independence is violated this may result in invalid statistical inference. However, many study designs in applied sciences give rise to correlated data. For example, subjects are followed over time and responses are assessed at different time points, or are observed in logical units (e.g. clinics, litters, families). In the following we show two different options that SAS software offers for the analysis of binary responses with correlated data (LOGISTIC procedure, %GLIMMIX macro). The SAS macro GLIMMIX and the SAS procedure LOGISTIC require SAS version 6.12, or later, to run.

#### B1. The SAS procedure LOGISTIC (Williams' method)

The LOGISTIC procedure is the standard tool in SAS software for fitting logistic regression models. The corresponding SAS code for the overdispersed logistic regression model (5.2) in Section 5.5 is as follows:

```
proc logistic data=piggeries order=data simple;
model died/pigsnumb=sowsnumbc areaproc area1 area2 area3 area4 area5
airvolumc workerse auje tafros distan prrs mikotox1 mikotox2 flies vet
/ clodds=wald clparm=wald corrb covb scale=williams;
output out=estprob pred=p;
run;
proc print data=estprob;
run;
```

where the DATA=PIGGERIES option in the PROC LOGISTIC statement names the SAS data set containing the data "piggeries" to be analysed. In addition, the ORDER=DATA option sorts the levels of the response variable



'died/pigsnumb' as they appear in the "piggeries" data set and the SIMPLE option displays simple descriptive statistics (mean, standard deviation, minimum and maximum) for each explanatory variable in the MODEL statement.

In the MODEL statement we name the response variable, died/pigsnumb, which is the observed proportion of the dead piglets in each farm for various combinations of the explanatory variables. After the equal sign we name the explanatory effects of the model, sowsnumbc, areaproc, ...vet, and after the slash (/) we specify the following options of the MODEL statement: The CLODDS, CLPARM=WALD options, which compute confidence intervals based on individual Wald tests for the odds ratios and the parameters, respectively. The CORRB and COVB options, which display the correlation and the covariance matrix of the fixed parameter estimates, respectively. Moreover, in the MODEL statement there are three SCALE= options to accommodate overdispersion (DEVIANCE, PEARSON and WILLIAMS). Since we have unequal sample sizes for the observations in our example, the SCALE=WILLIAMS option is preferred. The Williams model estimates an overdispersion parameter by equating the value of Pearson  $\chi^2$  for the model to its approximate expected value.

Next, in the OUTPUT statement we create the new SAS data set ESTPROB that contains all the variables in the input data set and, optionally, the estimated probabilities of the positive responses by using the PRED=P option. We then request that the data set ESTPROB be printed, using the DATA=ESTPROB option in the PROC PRINT statement.

## B2. The SAS macro GLIMMIX

The %GLIMMIX macro was written by Russ Wolfinger from SAS Institute. An overview about the macro and the theory behind is given in Chapter 11 of Littell *et al.*, 1996. The macro fits our needs, since it is designed for the analysis of the Generalized Linear Mixed Model (GLMM) and our logistic regression model with random effects is a special case of that model. By default, the macro uses restricted/residual pseudo likelihood (REPL) to find the parameter estimates of the GLMM we specify. Actually,



the macro was originally written to estimate the pseudo-likelihood function of Wolfinger and O'Connell (1993), which extended the penalized quasi-likelihood approach of Breslow and Clayton (1993), by estimating an additional overdispersion parameter. Briefly, the estimating algorithm uses the principle of quasi-likelihood and an approximation to the likelihood function of the model results in an iterative procedure repeatedly fitting a linear mixed model to a pseudo response. Particularly, the macro calls the PROC MIXED iteratively until convergence in order to estimate the linear mixed model with the pseudo response. Another two approaches to fitting GLMM are the so-called penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) approaches (Breslow and Clayton, 1993). Implementation of both PQL and MQL estimation procedure may be achieved with the previous algorithm.

As an illustration, we give the following SAS code, which fits model (5.3) in Section 5.6.1 with the %GLIMMIX macro:

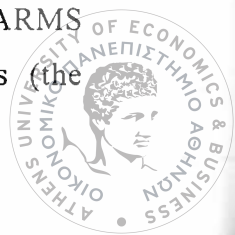
```
%include 'glimmix.sas';
%glimmix(data=piggeries,
  procopt=noprofile order=data,
  stmmts=%str(
class area1 area2 area 3 area4 area5 auje tafros distan prrs mikotox1
mikotox2 flies vet farms;
model died/pigsnumb= sowsnumbc areaproc area1 area2 area 3 area4
area5 airvolumc workeresc auje tafros distan prrs mikotox1 mikotox2
flies vet / chisq solution;
random intercept / solution type=cs subject=farms;
parms (0) (0) (1) / eqcons=2,3;
),
error=binomial,link=logit,
output=xbeta=Predmean pred=Linpred reschi=ResPearson
resraw=Resraw,
options=mixprintlast printdata
)
run;
```



```
proc print;  
run;
```

where the DATA= PIGGERIES option specifies the data set “piggeries” that we are using and the PROC OPT option specifies options appropriate for a PROC MIXED statement. Specifically, the NOPROFILE option includes the residual variance as part of the Newton-Raphson iterations and the value DATA of the ORDER=option sorts the levels of all CLASS variables according to the order of appearance in the input data set “piggeries”.

Furthermore, the %str() macro function may include any of the following PROC MIXED statements for the analysis: CLASS, MODEL RANDOM, REPEATED, PARMS, ID, CONTRAST, ESTIMATE and LSMEANS. Syntax and options for each statement are exactly as in the PROC MIXED documentation. As far as our example, in the CLASS statement we name the classification variables, area1, area2, ..., farms, that are used in the analysis and in the MODEL statement we name the dependent variable, died/pigsnumb, and the fixed effects, sowsnumbc, areaproc, ...vet, which determine the  $X$  matrix of the mixed model. Using the CHISQ and the SOLUTION options in the MODEL statement we request that  $\chi^2$ -tests be performed for all specified effects in addition to the F-tests and that estimates be produced for the fixed-effects parameters, respectively. The RANDOM statement defines the random effects in the mixed model. We specify INTERCEPT as a random effect in order to specify a random-intercept model. Moreover, as complete independence is assumed across the farms, we use the SUBJECT=FARMS option to produce a block-diagonal structure in the covariance matrix of the random effects with identical blocks. In addition, we specify a Compound Symmetry (CS) covariance structure for the covariance matrix of the random effects using the TYPE=CS option and we request the estimates for the random-effects parameters to be produced by the SOLUTION option. These estimates are the empirical best linear unbiased predictors (EBLUPs), which are obtained from Equation (2.18) in Section 2.3.1. Next, in the PARMS statement we specify initial values for the covariance parameters. Using the EQCONS=2, 3 option in the PARMS statement, we constrain the second and third covariance parameters (the



constant covariance of the CS structure and the residual variance, respectively) to equal 0 and 1, respectively. We note that the EQCONS=3 option fixes the residual variance at the given value of 1, and thus requests the estimation of the PQL model of Breslow and Clayton (1993). Leaving out the EQCONS=3 option gives the pseudo-likelihood estimates of Wolfinger and O'Connell (1993). Finally, in the PARMS statement we give an initial value for the first covariance parameter equal to 0, which is the constant variance of the CS covariance structure or the cluster variance.

Using the ERROR=BINOMIAL and LINK=LOGIT arguments, we request that the error distribution is binomial and that the link function is the logit (the default values). The options XBETA, PRED, RESCHI and RESRAW of the OUTPUT argument, create variables named Predmean, Linpred, ResPearson and Resraw respectively, and set them equal to the predicted mean, the linear predictor, the Pearson residual and the raw residual (response minus the predicted mean), respectively. At last, in the OPTIONS= argument the MIXPRINTLAST and PRINTDATA options print the final PROC MIXED run and the pseudo data after each iteration, respectively.





## Appendix C

### Gaussian Quadratures: The Gauss Hermite formula

Quadrature refers to the numerical integration of a function. The goal is to attain a given level of precision with the fewest possible function evaluations. The crucial factors that control the difficulty of this problem are:

- (i) the dimensionality, and
- (ii) the smoothness of the function.

Any method for numerically approximating  $\int W(x)f(x)dx$  relies on evaluating  $f$  on a finite set of points (called the abscissas or quadrature points), then processing these evaluations somehow to produce an approximation to the integral. Broadly speaking, there are two classes of methods. Methods for evenly spaced abscissas, which evaluate  $f$  on an evenly-spaced grid of points (e.g. the well known trapezoidal method) and methods for unevenly spaced abscissas, which may evaluate  $f$  at arbitrary locations. The Gaussian quadratures belong to the latter methods where the location of the abscissas at which the function  $f$  is to be evaluated is not equally spaced.

More specifically, the Gaussian quadratures provide the flexibility of choosing not only the weighting coefficients but also the locations (abscissas) where the functions are evaluated. Thus, by using Gaussian quadrature formulas, whose order is twice that of the formulas that consider evenly spaced abscissas, we will have twice the number of degrees of freedom at our disposal, meaning that the Gaussian quadratures yield twice as many places of accuracy with the same number of function evaluations. Another feature of Gaussian quadrature formulas that adds to their usefulness is: We can arrange the choice of weights and abscissas to make the integral exact for a class of integrands “polynomials times some known function  $W(x)$ ” rather than for the usual class of integrands “polynomials”. The function  $W(x)$  can then be chosen to remove integrable singularities from the desired integral. In other



words, given  $W(x)$  and given an integer  $K$ , we can find a set of weights  $w_j$  and abscissas  $x_j$  such that the approximation

$$\int_a^b W(x)f(x)dx \approx \sum_{i=1}^K w_i f(x_i) \tag{1}$$

is exact if  $f(x)$  is a polynomial.

We can find a set of polynomials (i) that includes exactly one polynomial of order  $j$ , called  $p_j(x)$ , for each  $j=0,1,2,\dots$ , and (ii) all of which are mutually orthogonal over the specified weight function  $W(x)$ . A constructive procedure for finding such a set is the recurrence relation:

- $p_{-1}(x) = 0$
- $p_0(x) = 1$
- $p_{j+1}(x) = (x - a_j)p_j(x) - b_j p_{j-1}(x) \quad j = 0,1,2,\dots$  (2)

where

$$a_j = \frac{\langle xp_j | p_j \rangle}{\langle p_j | p_j \rangle} \quad j = 0,1,\dots \tag{3}$$

$$b_j = \frac{\langle p_j | p_j \rangle}{\langle p_{j-1} | p_{j-1} \rangle} \quad j = 1,2,\dots \tag{4}$$

The coefficient  $b_0$  is arbitrary; we can take it to be zero. We note that in Equations (3) and (4), the notation  $\langle \cdot | \cdot \rangle$  corresponds to the scalar product. For example, the scalar product of two functions  $f$  and  $g$  over a weight function  $W$  can be defined as:

$$\langle f | g \rangle = \int_a^b W(x)f(x)g(x)dx$$

Two functions are said to be orthogonal if their scalar product is zero. A function is said to be normalized if its scalar product with itself is unity. Finally, a set of functions that are all mutually orthogonal and also all individually normalized is called an orthonormal set.



The polynomial  $p_j(x)$  can be shown to have exactly  $j$  distinct roots in the interval  $(a,b)$ . Moreover, we want to find all the roots of an orthogonal polynomial  $p_j(x)$ , because the abscissas of the  $K$ -point Gaussian quadrature formula (1) with weighting function  $W(x)$  in the interval  $(a,b)$  are precisely the roots of the orthogonal polynomial  $p_K(x)$  for the same interval and weighting function. This is the fundamental theorem of Gaussian quadratures and lets us find the abscissas for any particular case.

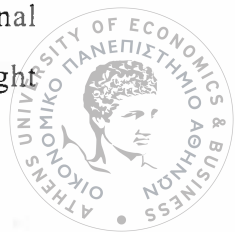
As we know the abscissas, we next find the weights  $w_j, j=1, \dots, K$ , by solving the set of linear equations

$$\begin{bmatrix} p_0(x_1) & \dots & p_0(x_K) \\ p_1(x_1) & \dots & p_1(x_K) \\ \vdots & & \vdots \\ p_{K-1}(x_1) & \dots & p_{K-1}(x_K) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} = \begin{bmatrix} \int_a^b W(x)p_0(x)dx \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (5)$$

Equation (5) computes those weights such that the quadrature (1) gives the correct answer for the integral of the first  $K$  orthogonal polynomials. The zeros of the right-hand side of (5) appear because  $p_1(x), \dots, p_{K-1}(x)$  are all orthogonal to  $p_0(x)$ , which is a constant. It can be shown that with those weights the integral of the next  $K-1$  polynomials is also exact, so that the quadrature is exact for all polynomials of degree  $2K-1$  or less.

Motivated by the above discussion, we next describe the steps that are required for computing the Gaussian quadrature formulas: (i) the generation of the orthogonal polynomials  $p_0, \dots, p_K$ , i.e., the computation of the coefficients  $a_j, b_j$  in (3) and (4) respectively; (ii) the determination of the zeros of  $p_K(x)$  and the computation of the associated weights. However, for the case of the “classical” orthogonal polynomials (for example, the Hermite polynomials), the coefficients  $a_j, b_j$  are known and step (i) can be omitted.

Afterwards, we focus on the Gauss-Hermite integration formula, which is one of the forms of Gaussian quadratures. The Gauss-Hermite formula uses the Hermite polynomials  $H_j(x)$  to deal with the integration interval of  $(-\infty, \infty)$ . The Hermite polynomials are classical, well-studied, orthogonal polynomials, and therefore, practically everything is known. The weight



function, interval, and recurrence relation that generate the Hermite polynomials and their corresponding Gauss-Hermite formula, are given as follows:

- $W(x) = e^{-x^2} \quad -\infty < x < \infty$

$$H_{j+1} = 2xH_j - 2jH_{j-1} \tag{6}$$

In practice, using Equation (6) we find that the computations overflow for large  $K$  because of various factorials that occur. One way to overcome this problem is to use the orthonormal set of polynomials  $\tilde{H}_j$ . The following recurrence relation generates them.

$$\tilde{H}_{-1} = 0, \quad \tilde{H}_0 = \frac{1}{\pi^{1/4}}, \quad \tilde{H}_{j+1} = x\sqrt{\frac{2}{j+1}}\tilde{H}_j - \sqrt{\frac{j}{j+1}}\tilde{H}_{j-1} \tag{7}$$

The weights with this normalization are given by

$$w_j = \frac{2}{[\tilde{H}'_K(x_j)]^2} \tag{8}$$

and the derivative by

$$\tilde{H}'_j = \sqrt{2j}\tilde{H}_{j-1} \tag{9}$$

The calculated Gauss-Hermite abscissas (corresponding to the zeros of the associated polynomials) and weights given in (8) are then used with the integration formula

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx = \sum_{i=1}^K w_i f(x_i) \tag{10}$$

Stroud and Secrest (1966), Golub and Welsch (1969), Stoer and Bulirsch (1980) and Johnson and Riess (1982), give further details for the Gaussian quadratures.



## References

- Abramowitz, M. and Stegun, I.A. (1964).** *Handbook of Mathematical Functions*. National Bureau of Standards, Washington, DC
- Aeschbacher, H.U., Milon, H. and Wurzner, H.P. (1978).** Caffeine concentrations in mice plasma and testicular tissue and the effect of caffeine on the dominant lethal test, *Mutation Research*, 57, 193-200
- Aitchison, J. and Shen, S.M. (1980).** Logistic-normal distributions: some properties and uses, *Biometrika*, 67, 261-272
- Aitkin, M. (1996).** A general maximum likelihood analysis of overdispersion in generalized linear models, *Statistics and Computing*, 6, 251-262
- Aitkin, M. (1999).** A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models, *Biometrics*, 55, 117-128
- Alanko, T. and Duffy, J.C. (1996).** Compound Binomial Distributions for Modelling Consumption Data, *The Statistician*, 45, 269-286
- Altham, P.M.E. (1978).** Two Generalizations of the Binomial Distribution, *Applied Statistics*, 27, 162-167
- Barndorff-Nielsen, O.E. and Cox, D.R. (1989).** *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London
- Beitler, P.J. and Landis, J.R. (1985).** A Mixed-effects Model for Categorical Data, *Biometrics*, 41, 991-1000
- Böhning, D. (1999).** *Computer-Assisted Analysis of Mixtures and Applications: Meta-analysis, disease mapping and others*. Chapman and Hall/CRC, United States of America
- Booth, J.G. and Caffo, B.S. (2002).** Unequal sampling for Monte Carlo EM algorithms, *Computational Statistics and Data Analysis*, 39, 261-270
- Booth, J.G. and Hobert, J.P. (1999).** Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society, Series B*, 61, 265-285



- Breslow, N.E. and Clayton, D.G. (1993).** Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, 88, 9-25
- Brooks, S.P. (2001).** On Bayesian Analyses and Finite Mixture Models for Proportions, *Statistics and Computing*, 11, 179-190
- Carroll, R.J. and Ruppert, D. (1982).** Robust Estimation in Heteroscedastic Linear Models, *The Annals of Statistics*, 10, 429-441
- Chatfield, C. and Goodhardt, G.J. (1970).** The Beta-Binomial model for consumer purchasing behaviour, *Applied Statistics*, 19, 240-250
- Crowder, M.J. (1978).** Beta-Binomial Anova for Proportions, *Applied Statistics*, 27, 34-37
- Davidian, M. and Gallant, A.R. (1993).** The nonlinear mixed effects model with a smooth random effects density, *Biometrika*, 80, 475-488
- Davies, R.B. (1987).** Mass point methods for dealing with nuisance parameters in longitudinal studies. In: R. Crouchley, ed. *Longitudinal Data Analysis*, Avebury, Aldershot, Hants
- Dempster, A.P., Laird, N.M. and Rubin, D. (1977).** Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1-38
- Dietz, E. and Böhning, D. (1995).** Statistical inference based on a general model of unobserved heterogeneity. Lecture Notes in Statistics 104, pp. 75-82 (In *Statistical Modelling: proceedings of the 10th International Workshop on Statistical Modelling*, Innsbruck, Austria, 10-14 July, 1995). Springer-Verlag, New York
- Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994).** *Analysis of Longitudinal Data*. Clarendon Press, Oxford
- Efron, B. (1986).** Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association*, 81, 709-721
- Firth, D. (1991).** *Statistical Theory and Modelling*. Chapman and Hall, London
- Fitzmaurice, G.M., Heath, A.F. and Cox, D.R. (1997).** Detecting Overdispersion in Large Scale Surveys: Application to a Study of Education and Social Class in Britain, *Applied Statistics*, 46, 415-432



- Follmann, D.A. and Lambert, D. (1988).** Identifiability for Non-parametric Mixtures of Logistic Regressions, unpublished manuscript
- Follmann, D.A. and Lambert, D. (1989).** Generalizing Logistic Regression by Nonparametric Mixing, *Journal of the American Statistical Association*, 84, 295-300
- Fowlkes, E.B. (1987).** Some Diagnostics for Binary Logistic Regression via Smoothing, *Biometrika*, 74, 503-515
- Francis, B.J., Green, M. and Payne, C. (1993).** *The GLIM System: Release 4 Manual*. Clarendon Press, Oxford
- Gange, S.J., Munoz, A., Saez, M. and Alonso, J. (1996).** Use of the Beta-Binomial distribution to model the effect of policy change on appropriateness of hospital stays, *Applied Statistics*, 45, 371-382
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995).** *Bayesian Data Analysis*. Chapman and Hall, London
- Goldstein, H. (1991).** Nonlinear Multilevel Models With an Application to Discrete Response Data, *Biometrika*, 78, 45-51
- Golub, G.H. and Welsch, J.H. (1969).** Calculation of Gauss Quadrature Rules, *Mathematics of Computation*, 23, 221-230
- Goutis, C. (1993).** Recovering extra-binomial variation, *Journal of Statistical Computation and Simulation*, 45, 233-242
- Grassia, A. (1977).** On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions, *Australian and New Zealand Journal of Statistics*, 19, 108-114
- Green, P.J. (1987).** Penalized Likelihood for General Semi-Parametric Regression Models, *International Statistical Review*, 55, 245-259
- Griffiths, D.A. (1973).** Maximum likelihood estimation for the Beta-Binomial distribution and an application to the household distribution of the total number of cases of a disease, *Biometrics*, 29, 637-648
- Harville, D.A. (1977).** Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association*, 72, 320-340
- Haseman, J.K and Kupper, L.L. (1979).** Analysis of dichotomous response data from certain toxicological experiments, *Biometrics*, 35, 281-293



- Hasselblad, V. (1969).** Estimation of Finite Mixtures of Distributions from the Exponential Family, *Journal of the American Statistical Association*, 64, 1459-1471
- Heagerty, P.J. and Zeger, S.L. (2000).** Marginalized Multilevel Models and Likelihood Inference, *Statistical Science*, 15, 1-26
- Heckman, J.J. and Singer, B. (1984).** A method for minimizing the impact of distributional assumptions in econometric models of duration, *Econometrica*, 52, 271-320
- Hedeker, D. and Gibbons, R.D. MIXOR: user's manual.**  
(Available in: <http://tigger.uic.edu/~hedeker/mixwin.html>)
- Ishii, G. and Hayakawa, R. (1960).** On the compound Binomial distribution, *Annals of the Institute of Statistical Mathematics*, 12, 69-80
- Johnson, L.W. and Riess, R.D. (1982).** *Numerical Analysis*, 2<sup>nd</sup> edition. Addison-Wesley, Reading, Massachusetts
- Johnson, N.L., Kotz, S. and Kemp, A.W. (1992).** *Univariate discrete distributions*, second edition. John Wiley and Sons, New York
- Kackar, R.N. and Harville, D.A. (1984).** Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models, *Journal of the American Statistical Association*, 79, 853-862
- Karlis, D. (2003).** *EM algorithm for mixed Poisson and other discrete Distributions*, (to appear in ASTIN bulletin)
- Kleinman, J.C. (1973).** Proportions with extraneous variance: single and independent samples, *Journal of the American Statistical Association*, 68, 46-54
- Laird, N.M. (1978).** Nonparametric maximum likelihood estimation of a mixing distribution, *Journal of the American Statistical Association*, 73, 805-811
- Lesaffre, E. and Spiessens, B. (2001).** On the effect of the number of quadrature points in a logistic random-effects model: an example, *Applied Statistics*, 50, 325-335
- Lesperance, M.L. and Kalbfleisch, J.D. (1992).** An Algorithm for Computing the Nonparametric MLE of a Mixing Distribution, *Journal of the American Statistical Association*, 87, 120-126



- Liang, K.Y. and Zeger, S.L. (1986).** Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, 73, 13-22
- Lindsay, B.G. (1983).** The geometry of mixture likelihoods, part I: a general theory, *The Annals of Statistics*, 11, 86-94
- Lindsay, B.G. (1995).** *Mixture Models: Theory, Geometry and Applications*. In Regional Conference Series in Probability and Statistics, Vol5, (Institute of Mathematical Statistics and American Statistical Association), Hayward
- Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996).** *SAS System for Mixed Models*, Cary, NC, SAS Institute Inc.
- McCullagh, P. and Nelder, J.A. (1989).** *Generalized Linear Models*, Second Edition. Chapman and Hall, London
- McCulloch, C.E. (1997).** Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association*, 92, 162-170
- McLachlan, G.J. and Peel, D. (2000).** *Finite mixture models*. John Wiley and Sons, New York
- Moran, P.A.P. (1968).** *An Introduction to Probability Theory*. University Press, Oxford
- Orme, C.D. (1998).** On the insensitivity of the score test for heterogeneity to omitted covariates in multivariate failure time models, *Biometrika*, 85, 457-461
- Otake, M. and Prentice, R.L. (1984).** The analysis of chromosomally aberrant cells based on Beta-Binomial distribution, *Radiation Research*, 98, 456-470
- Patterson, H.D. and Thomson, R. (1971).** Recovery of Interblock Information When Block Sizes Are Unequal, *Biometrika*, 58, 545-554
- Paul, S.R. (1982).** Analysis of proportions of affected foetuses in teratological experiments, *Biometrics*, 38, 361-370
- Pierce, D.A. and Sands, B.R. (1975).** Extra-Bernoulli variation in binary data. *Technical Report 46*, Department of Statistics, Oregon State University
- Prentice, R.L. (1986).** Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation Induced by Covariate



Measurement Errors, *Journal of the American Statistical Association*, 81, 321-327

**Robinson, G.K. (1991).** That BLUP Is a Good Thing: The Estimation of Random Effects, *Statistical Science*, 6, 15-51

**Rodriguez, G. and Goldman, N. (1995).** An assessment of estimation procedures for multilevel models with binary responses, *Journal of the Royal Statistical Society, Series A*, 158, 73-89

**Schall, R. (1991).** Estimation in generalized linear models with random effects, *Biometrika*, 78, 719-727

**Sichel, H.S. (1975).** On a distribution law of word frequencies, *Journal of the American Statistical Association*, 70, 542-547

**Sichel, H.S. (1982).** Repeat-buying and the generalized inverse Gaussian-Poisson distribution, *Applied Statistics*, 31, 193-204

**Skellam, J.G. (1948).** A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials, *Journal of the Royal Statistical Society, Series B*, 10, 257-261

**Steele, B.M. (1996).** A modified EM algorithm for estimation in generalized mixed models, *Biometrics*, 52, 1295-1310

**Stoer, J. and Bulirsch, R. (1980).** *Introduction to Numerical Analysis*. Springer-Verlag, New York

**Stroud, A.H. and Secrest, D. (1966).** *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, New Jersey

**Tarone, R.E. (1979).** Testing the Goodness of Fit of the Binomial Distribution, *Biometrika*, 66, 585-590

**Tierney, L. and Kadane, J.B. (1986).** Accurate Approximation for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, 81, 82-86

**Tripathi, R.C., Gupta, R.C. and Gurland, J. (1994).** Estimation of parameters in the Beta-Binomial model, *Annals of the Institute of Statistical Mathematics*, 46, 317-331

**Tweedie, M.C.K. (1957).** Statistical properties of inverse Gaussian distributions I, *Annals of Mathematical Statistics*, 28, 362-377



- Vuataz, L. and Sotek, J. (1978).** Use of the Beta-Binomial distribution in dominant-lethal testing for “week mutagenic activity”, Part 2, *Mutation Research*, 52, 211-230
- Walker, S. (1996).** An EM algorithm for nonlinear random effects models, *Biometrics*, 52, 934-944
- Wedderburn, R.W.M. (1974).** Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, 61, 439-447
- Wei, G.C.G. and Tanner, M.A. (1990).** A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms, *Journal of the American Statistical Association*, 85, 699-704
- Williams, D.A. (1975).** The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, 31, 949-952
- Wolfe, J.H. (1970).** Pattern Clustering by Multivariate Mixture Analysis, *Multivariate Behavioral Research*, 5, 329-350
- Wolfinger, R. and O’Connell, M. (1993).** Generalized Linear Mixed Models: A Pseudo-Likelihood Approach, *Journal of Statistical Computation and Simulation*, 48, 233-243





Δωρεά

